Dynamic and Thermodynamic Signatures of Native and Non-Native Protein States with Application to the Improvement of Protein Structures

Da-Wei Li and Rafael Brüschweiler*

Chemical Sciences Laboratory, Department of Chemistry and Biochemistry and National High Magnetic Field Laboratory, Florida State University, Tallahassee, Florida 32306, United States

S *Supporting Information*

ABSTRACT: Accurate knowledge of the 3D structural ensemble of proteins is important for understanding of their biological function. We report here the application of microsecond all-atom molecular dynamics (MD) simulations in explicit solvent for the improvement of the quality of low-resolution structures obtained by protein structure prediction (decoys). Seventy MD simulations of ~1 μs average duration were performed on 13 different protein systems starting from X-ray crystal structures and decoys. Their behavior can be divided into three groups: 22 trajectories converged toward the native state, 27 trajectories displayed a quasi-equilibrium by populating mainly a single non-native free energy basin, and 21 trajectories drifted away from their initial decoy structure transiently visiting multiple free energy minima. To determine whether the native structure can be identified among non-native ensembles, the free energy was determined for each basin by the MM/GBSA method together with the von Mises entropy estimator in dihedral angle space. For the proteins studied here, it is found that the ensembles belonging to free energy basins with the lowest free energies and the longest residence times are most native-like. The results demonstrate that explicit solvent microsecond MD simulations using the latest generation of protein force fields and free energy metrics are sufficiently accurate to permit positive identification of native state ensembles against low-resolution structural models and decoys. The approach can be applied to the direct refinement of predicted or experimental low-resolution protein structures.

INTRODUCTION

Atomic-detail three-dimensional structural ensembles of proteins provide crucial information about their function. While the vast majority of protein structures in the Protein Databank (PDB)[1] have been determined by X-ray crystallography or by NMR spectroscopy, protein structure prediction using *ab initio* or homology modeling methods[2] has made important in-roads as a complementary approach to traditional experiment-based protein structure determination. When no homologous template structure is available, protocols that use knowledge-based energy potentials assemble secondary structures into tertiary structures using information of proteins with known structures.[3] In this way, a large number of structural candidates or decoys is typically produced. The similarity between decoys and the corresponding experimental, i.e. native, protein structure can greatly vary, whereby some decoys can be close to the native structure. The reliable identification of those decoys that are closest to the native structure is therefore key for the success of *ab initio* protein structure determination. In a recent Critical Assessment of Structure Prediction (CASP) competition,[2] several software packages, such as Rosetta[4] and I-TASSER,[5] demonstrated for some proteins the ability to build and correctly identify structural models within a few Ångstroms of the native structures, while for other proteins the predicted structures were considerably off.

In a parallel development, all-atom molecular dynamics (MD) simulations of proteins in explicit solvent have made significant strides in their ability to quantitatively reproduce a wide range of experimental structural dynamic parameters of proteins. These advances were made possible by improvements of all-atom molecular mechanics force fields and the availability of ever more powerful computer hardware, which have made microsecond time scale simulations in explicit solvent practically feasible.[6] Recently developed force fields, such as AMBER ff99SB,[7] AMBER ff03,[8] and CHARMM CMAP,[9] stimulated the use of various types of experimental protein data for the quantitative certification of MD simulations. These include a wide range of solution NMR data of proteins and peptides, such as residual dipolar couplings (RDCs),[10−12] scalar J-couplings,[11,13,14] spin relaxation order parameters,[15−17] and chemical shifts.[18,19] The performance of 12 different commonly used protein force fields has been compared recently,[20] which showed that for a benchmark of 524 protein NMR parameters, the ff99SBnmr1-ILDN force field[12,21,22] has the highest accuracy among the force fields tested.

This development opens up the prospect of critically advancing the improvement of the accuracy of low- to medium-resolution protein structures that have been determined by experiment or by protein structure prediction methods.[23,24] More than a decade ago, Kollman and coworkers put forward a strategy, which uses thermodynamic free energy calculations to identify the native structure of a protein.[25,26] These early attempts were hampered by limitations in both the force field accuracy and trajectory lengths (~1 ns), which prevented decoys from fully relaxing, causing artificially high free energies.[27] More recently, the use of longer MD simulations combined with accelerated sampling technologies

Received: May 7, 2012
Published: June 5, 2012

ACS Publications © 2012 American Chemical Society

using physics-based force fields has shown partial success in protein structure refinement[28−30] and prediction.[31]

Here, we describe how recent methodological advances permit the assessment and improvement of low-resolution protein structural models. Ideally, an accurate MD force field should allow computation of trajectories that spontaneously reach the native energy basin starting from an arbitrary initial low-resolution decoy structure in the absence of experimental constraints. Due to the presence of kinetic bottlenecks, however, this type of structure refinement presently only works in a subset of all cases (see Results section). Therefore, we use here, in addition, a thermodynamic approach by comparing computed free energies of microsecond MD ensembles of candidate structures, both from experiments and decoys. The free energies are determined by the Molecular Mechanics-Generalized Born/Surface Area (MM-GBSA) method[32] averaged over the extended MD ensembles, augmented by a recently developed dihedral-angle space configurational entropy estimator.[33,34] Application of this approach, referred to as PRESTO for Protein Refinement by Sampling and Thermodynamics, to 13 different proteins finds that the thermodynamically most stable models are the ones closest to their native X-ray crystal structure.

METHODS

Protein Selection. The proteins used in this study were taken from the Rosetta decoy library.[35] A list of the proteins together with their PDB codes, lengths, and topologies is given in Table 1, and their ribbon diagrams are displayed in Figure S1. To rigorously test our approach, all proteins in the decoy library for which Rosetta failed to identify the native conformation[36] were included, with the exception of proteins with more than 100 residues (to restrict the computational costs) and calbindin D_{9k} (1IG5; as it binds multiple divalent ions). For the decoy set, initial structures were randomly picked from the first 100 lowest energy decoys according to the Rosetta energy function, with priority given to those structures whose RMSD with respect to the native structure is within 3−8 Å.

Table 1. Names and Other Information about Proteins Used in This Study

protein name	PDB code	length[a]	SCOP class	N_1[b]	N_2[c]
ribosomal protein S15	1A32	65	α	1	3
potassium channel KV1.1	1A68	87	$\alpha+\beta$	1	7
AOP-RANTES	1B3A	55	$\alpha+\beta$	1	7
DNA-binding protein 7A	1C8C	62	β	1	3
cold-shock protein	1C90	66	β	1	5
ribosomal protein L7/L12	1CTF	68	$\alpha+\beta$	1	7
B1 domian of protein L	1HZ6	61	$\alpha+\beta$	1	7
repressor protein CI	1R69	61	α	1	3
subtilisin	1SCJ	66	$\alpha+\beta$	1	3
tyrosine kinase SH3 domain	1SHF	59	β	1	3
translation initiation factor	1TIG	88	$\alpha+\beta$	1	3
dihydrofolate reductase	1VIE	56	β	1	3
Cro repressor protein	5CRO	55	α	1	3

[a]Number of amino acids. [b]Number of native MD simulations (starting from X-ray crystal structure). [c]Total number of MD simulations starting from decoys.

MD Simulations. All MD simulations were performed using the Gromacs 4.5 package.[37−40] Water molecules were explicitly included using the TIP3P water model.[41] The integration time step was set to 2 fs with all bond lengths involving hydrogen atoms constrained by the SETTLE algorithm. Na^+ and Cl^- ions were added as needed to obtain a system with neutral total charge. Electrostatic interactions were cut off at 10 Å, and the long-range electrostatic interactions were calculated using the PME algorithm with 1.2 Å spacing. All van der Waals interactions were cut off at 8 Å. To minimize possible finite size effects, periodic boundary conditions were applied in all three dimensions. A cubic simulation box with a length equal to the largest interatomic distance of the protein plus 16 Å was employed. After 50 000 steps of steepest descent energy minimization, the systems were simulated for 100 ps at a constant temperature of 300 K and constant volume with all protein heavy atoms positionally constrained. Next, pressure coupling at 1 atm was turned on, and the systems were simulated for another 100 ps.[42] The final production runs were performed at a constant temperature and pressure (NPT ensemble) of 330 K and 1 atm, respectively. The slightly elevated temperature was chosen to improve sampling efficiency. Unfolding of the native state during the course of the MD simulation was not observed for any protein studied here. Nearly all MD simulations were run for 1000 ns, resulting in a cumulative length of all simulations of over 60 μs. For a 100 amino acid protein, 1 μs MD simulation took about 25 days on Florida State University's High Performance Computer Cluster using 48 cores per simulation. The recent protein force field ff99SB_$\varphi\psi$(g24;CS)[43] was used for all simulations, in which the φ,ψ dihedral angle potential had been improved by the inclusion of cross terms through the addition of 24 bivariate Gaussian functions of variable depth, width, and tilt angle. This force field, which reproduces a wide range of experimental NMR parameters of full-length proteins, was further enhanced by the addition of the recently developed ILDN side chain corrections.[22] The combination of the NMR-optimized force field ff99SBnmr1 with the ILDN correction has been shown recently to reproduce experimental NMR parameters with remarkably high accuracy.[12,20] Since the backbone potential of ff99SB_$\varphi\psi$(g24;CS) is improved over ff99SBnmr1, ff99SB_$\varphi\psi$(g24;CS) + ILDN is expected to be at least as accurate as ff99SBnmr1 + ILDN. It should be noted that none of the proteins studied here were used for the optimization of the ff99SB_$\varphi\psi$(g24;CS) force field.

Free Energy Calculations. For the free energy calculations, we used the Molecular Mechanics/Generalized Born Solvent Accessible Surface Area (MM/GBSA) method[32] with the OBC model,[44] together with a configurational entropy estimator S:

$$G = \langle E^{\text{bonded}} \rangle + \langle E^{\text{Coulomb}} \rangle + \langle E^{\text{GB}} \rangle + \langle E^{\text{vdW}} \rangle + \sigma \langle A^{\text{solv}} \rangle - TS \tag{1}$$

where the angular brackets represent ensemble averages over MD snapshots after the removal of all solvent molecules and counterions. The first term describes the dependence of the internal energy on the protein geometry (bond lengths, bond angles, dihedral angles). The second term describes Coulombic interaction energies between the atomic charges of the protein. The third term represents the electrostatic contribution to the solvation free energy, which is computed using the generalized Born approximation at zero salt concentration.[45−47] The fourth term describes the van der Waals interactions between protein atoms. No cutoff was used for the nonbonded interactions. The fifth term describes nonpolar contributions to the solvation free energy, which is assumed to be proportional to the solvent-accessibility surface area, A^{solv}, of the protein. The proportionality coefficient σ was set to 0.005 kcal/mol/Å^2. Since this term is relatively small, its precise value has only little effect on the results reported in this work. The final term, $-TS$, contains the entropy estimator in dihedral angle space, which includes dihedral angle pair correlation effects:

$$S = \sum_{k=1}^{I} S_{1d,k} + \sum_{m=1}^{M} I_m \tag{2}$$

where $I_m = S_{2d,ij} - S_{1d,i} - S_{1d,j}$ and S_{1d} and S_{2d} are the 1D and 2D von Mises kernel density estimation based entropies introduced previously.[33,34] The first sum in eq 2 extends over all (mobile) dihedral angles of the protein, and the second sum extends over M dihedral angle pairs ij that are next neighbors along the backbone ($\psi_{n-1}\varphi_n$ and $\varphi_n\psi_n$) or belong to the same side chain ($\chi_1\chi_2$, etc.). Equation 2 accounts for the non-Gaussian nature of the distributions of many dihedral angles and includes dihedral angle correlations up to second order (for more details, see the Supporting Information).[48]

RESULTS

Spontaneous Approach of Free MD to Native State. A useful indicator of protein behavior during the microsecond-MD simulation is the backbone root-mean-square deviation (RMSD) with respect to the native reference structure, which in this work is the X-ray crystal structure. For the two proteins 1C8C and 1R69 out of 13 proteins, all six decoys approach the native structure during a 20−300 ns time window (Figure 1A,B). Their final RMSDs to the native structures are 1−1.5 Å, which agree well with the RMSDs obtained when starting the MD trajectories directly from the native structures (black curves). The two proteins of Figure 1 have distinct folds (see ribbon diagrams in Figure 1), and they display variable transition times to the native basin. For protein 1C8C, which is a mixed α/β protein (Figure 1A), a wrongly packed β-hairpin reshapes into the correct structure within 125−175 ns. For the five-helix-bundle protein 1R69 (Figure 1B), the rearrangement of the wrongly packed helix in all three decoys to the native state takes only a few tens of nanoseconds.

Sampling of Native and Non-Native Free Energy Basins. Internal reorganization of the secondary structure can be a slow process. Hence, it cannot be expected that brute force MD simulations can convert all decoys to the native basin within 1 μs. It is therefore important to have the means to reliably identify native-like structural ensembles generated by MD from decoys or low-resolution experimental structures. The pairwise RMSD of all backbone Cα atoms between different MD snapshots provides direct insight into the sampling of free energy basins during a trajectory. Figure 2 and Figures S2 and S3 (Supporting Information) show pairwise RMSD maps of MD simulations of three representative protein systems 1TIG, 1VIE, and 1SHF comparing snapshots stored every 1 ps with each other, where the MD simulations belonging to each panel either started from the native structures (panel A) or decoys (panels B−D). Below each panel, the backbone RMSD with respect to the native structure is plotted as a function of trajectory length. Comparison of the pairwise RMSD maps demonstrates that for all four simulations

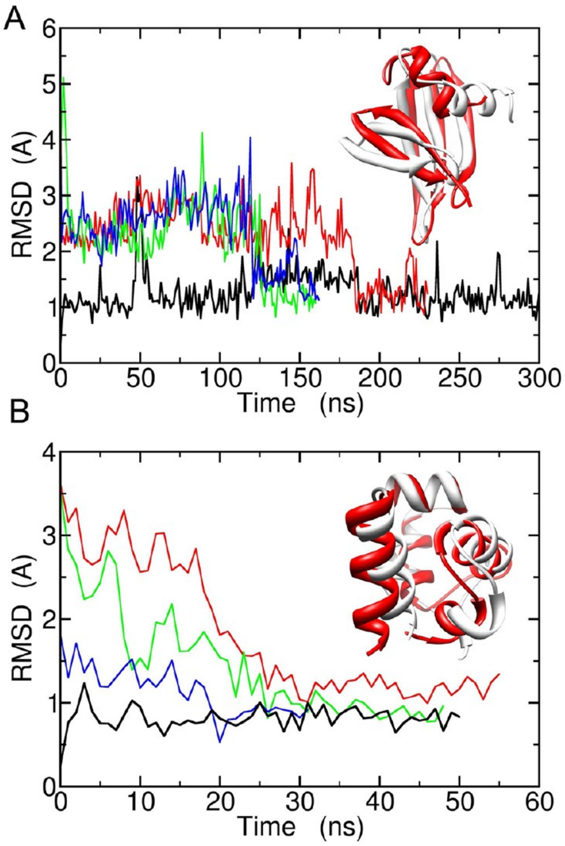

Figure 1. Backbone RMSD with respect to native state as a function of simulation time for proteins (A) 1C8C and (B) 1R69. The black and colored lines belong to the native state and different decoy simulations, respectively. The insets show ribbon representations of one of the initial decoy structures (red) superimposed on the structure of the native state (gray).

performed for each of these proteins the native simulation was the most stable simulation (panel A) with RMSDs < 2 Å (large blue areas), whereas the simulations starting from a decoy displayed conformational transitions between different free energy minima or basins. A basin is characterized by the blue square-shaped regions along the diagonal reflecting low pairwise RMSD between a series of consecutive snapshots. This behavior is exemplified in Figure 2B for protein 1TIG. The decoy starting structure was stable for only about 80 ns before the protein drifted into a new basin, in which it stayed for 120 ns. The system then proceeded to the next basin, which is distinct from the first two as evidenced by the relatively large RMSDs (yellow and red color), where it resided during the rest of the simulation (>800 ns). The third basin is native-like, as can be seen from the low RMSD (<2 Å) with respect to the native structure (lower graph in panel B). For the other two decoys (panels C and D), the protein underwent multiple conformational changes without reaching the native state. In panel C, a metastable energy basin was reached after 220 ns, where the system resided for 400 ns. This energy basin is clearly non-native, as is reflected by the large RMSD of ~4.5 Å to the native state. The other two proteins, 1VIE and 1SHF (Figures S2 and S3), displayed similar behavior to 1TIG. While the native simulations were stable over the whole simulation length (panel A), the simulations starting from decoy structures sampled a multitude of non-native energy basins with variable residence times. For each decoy existed at least one energy

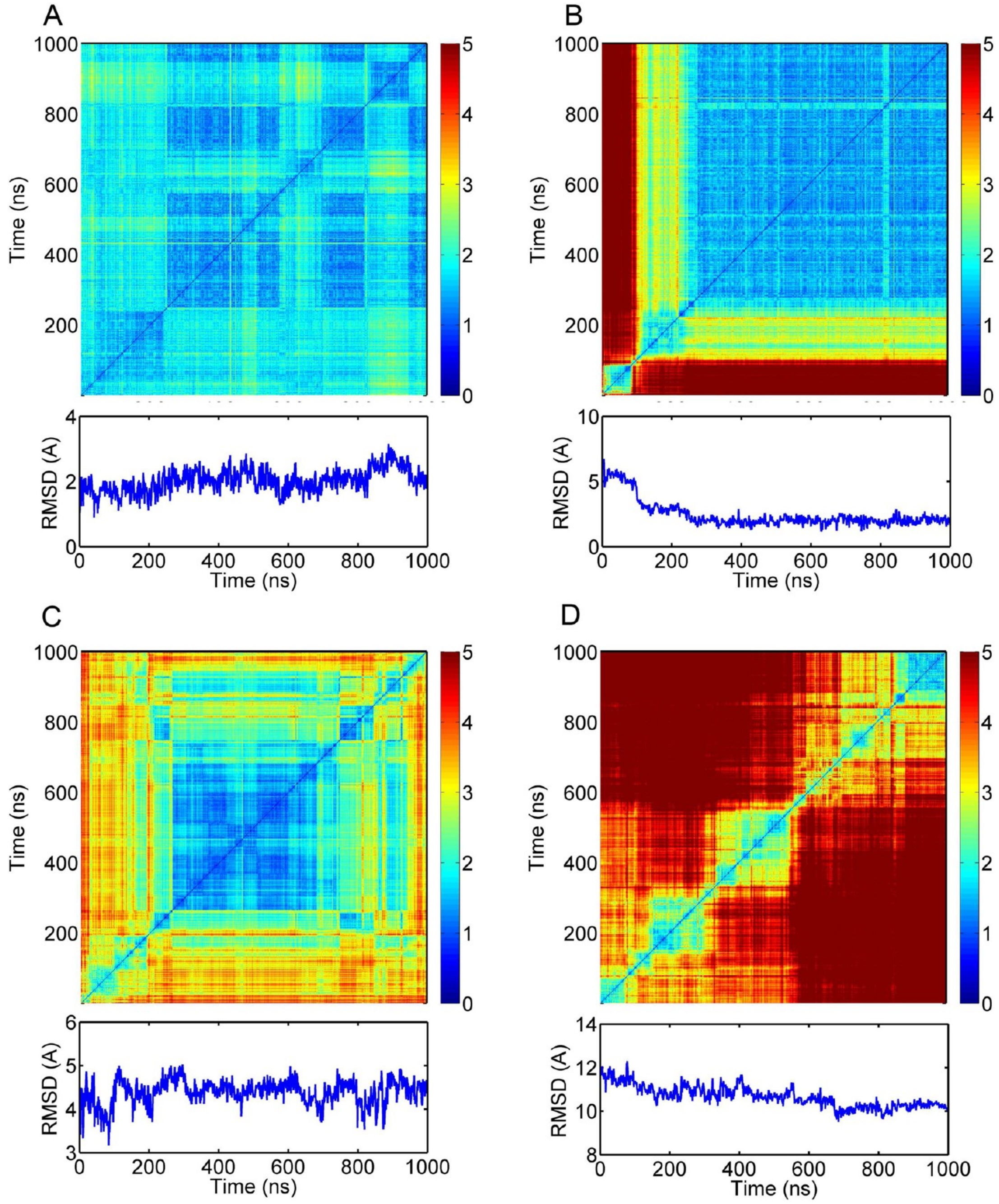

Figure 2. Backbone Cα pairwise RMSD (top panel) and backbone RMSD with respect to native X-ray crystal structure (bottom panel) of MD ensembles of protein 1TIG starting (A) from the native state and (B–D) from different decoys. The unit of the color bar is Ångstroms. Pairwise RMSDs > 5 Å are colored in dark red. The subtrajectories selected for free energy calculations are (A) 0–1000 ns, (B) 300–1000 ns, (C) 250–700 ns, and (D) 900–1000 ns. The RMSD behavior for the different starting structures is representative also for other proteins (see Supporting Information).

basin that was continuously occupied for 100 ns or longer. This property is important for the meaningful computation of the configurational entropy component (eq 2) of the free energy of such a basin.

Native State Similarity and Free Energies. For simulations for which no conformational transitions were observed to the native state, it is not *a priori* clear whether this is due to deficiencies of the molecular mechanics force field or because the transition is kinetically hindered on the time scale of the MD trajectory. To compute a free energy, it is necessary to define a subensemble that corresponds to a macrostate that is stable or metastable. This is because the configurational entropy contribution S of eq 2 to the free energy (eq 1) is sensitive to structural drifts, such as the ones depicted in Figure 2B–D, which may lead to an overestimation of the configurational entropy. To address this issue, for each simulation, a subensemble was selected consisting of successive snapshots that showed stable or metastable behavior as reflected by a sizable blue/yellow square along the diagonal in the 2D RMSD maps of the kind seen in Figure 2. The snapshots of these subtrajectories usually have pairwise RMSDs < 3.5 Å. For example, in Figure 2A–D, the selected subtrajectories extended

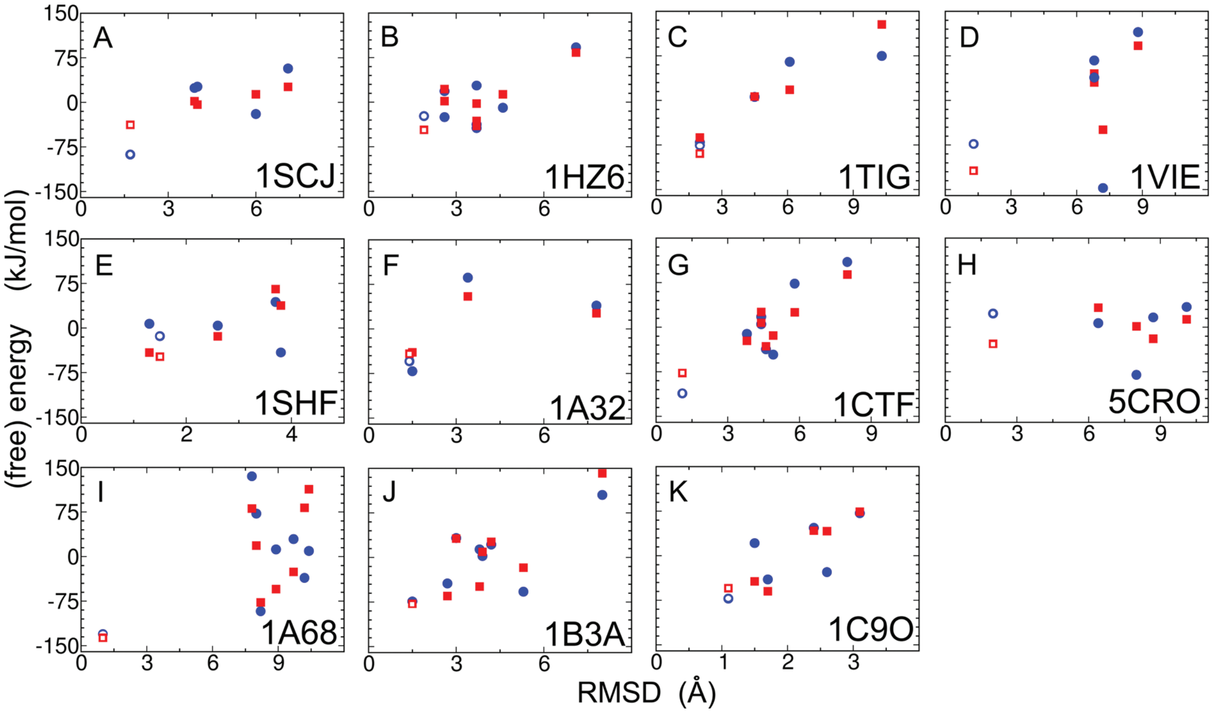

Figure 3. Free energies (red squares, calculated according to eq 1) and average potential energies (blue circles) of microsecond MD protein ensembles as a function of the average RMSD with respect to their X-ray crystal reference structures. (A–K) The MD trajectories started either from the native X-ray crystal structures (open symbols) or selected Rosetta decoys (filled symbols). From panels A to K, the crystal structures used belong to PDB entries 1SCJ, 1HZ6, 1TIG, 1VIE, 1SHF, 1A32, 1CTF, 5CRO, 1A68, 1B3A, and 1C9O with the full protein names given in Table 1.

between 0 and 1000 ns, 300 and 1000 ns, 300 and 700 ns, and 900 and 1000 ns, respectively. These subtrajectories statistically sampled their respective basins in an adequate manner, and their free energies could be calculated by eqs 1 and 2. Figures analogous to Figure 2 are shown for all other proteins in the Supporting Information (Figures S2–S11) with the selected subtrajectories specified in the figure captions. For the two proteins 1C8C and 1R69 of Figure 1 that converge to the native state, the corresponding 2D RMSD maps are fully analogous to the one in Figure 2B, except that convergence was faster.

Figure 3 shows for 11 proteins (which are all the proteins studied here, except 1C8C and 1R69 whose decoys approach the native state during the MD simulation) the correlation between MD-derived free energies starting from Rosetta decoy structures and the native X-ray structures with the average backbone RMSDs relative to the X-ray crystal reference structures. The free energies correlate remarkably well with the RMSDs relative to the native structures, with the decoys further away from the native structure on average possessing a higher free energy. For 9 out of the 11 proteins, the state with the lowest free energy is the one closest to the native structure. Importantly, also for the remaining two proteins, 1SHF and 1C9O (Figure 3E,K), the lowest free energy structures still have a RMSD < 2 Å to the native state and hence represent accurate models of their native states.

For a number of proteins, the configurational entropy S plays a critical role in the discrimination of decoys from the native state. Interestingly, if S is not included in eq 1 (blue symbols), for five proteins the lowest free energy structure does not have the lowest RMSD. For example, for proteins 1VIE (Figure 3D) and 5CRO (Figure 3H), the lowest free energy structures deviate substantially from their native structures with RMSDs of 7.2 Å and 8.0 Å, respectively.

DISCUSSION

Different Categories of Protein Behavior. The 70 MD simulations of up to 1 μs length of the 13 proteins studied here, which started from experimental protein structures and decoys, display three different types of behavior, shown in Figure 4. For the first category (Figure 4A), which comprises the two proteins of Figure 1, all six MD simulations starting from the decoys converge toward the native state structure within tens to hundreds of nanoseconds. For these proteins, the application of current MD methodology to Rosetta-generated decoys or low-resolution experimental structures leads to ensembles that are native. The all-atom force field with explicit solvent reproduces their global free energy minima at the correct locations.

The decoys of the remainder of the proteins did not reach the native state within 1 μs. Decoys commonly display secondary structural swaps, β-sheet registry shifts, incorrect rotameric states, or incorrect secondary structures relative to the native structures. Similarly, low-resolution experimental structures may display incorrectly packed side chains and small segments of incorrect secondary structures. In order to overcome some of these shortcomings in a free MD simulation, a protein may need to undergo (partial) unfolding and refolding, which can take significantly longer than a microsecond.[49] This behavior is due to the existence of relatively high energy barriers that separate non-native basins from the native one (Figure 4B) or a drifting behavior on a relatively smooth free energy surface with large RMSD changes (Figure 4C). Of the 56 simulations starting from decoy structures shown in Figures 1 and 2 and Figures S2–S11, Supporting Information, eight fall into category 1 (spontaneous convergence to the

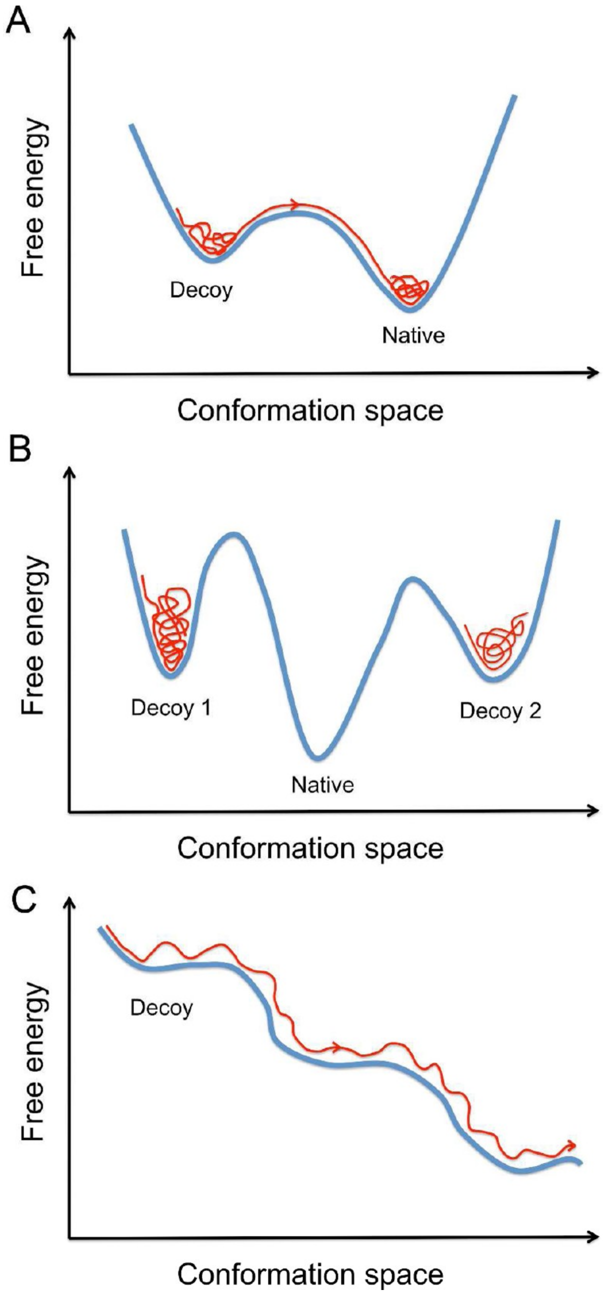

Figure 4. Schematic representation of three categories of behavior observed for 70 ~1 μs MD simulations that started from decoys, native structures, or low-resolution experimental models. Category A covers trajectories that spontaneously converge toward the native state. Category B covers trajectories that are stable and predominantly sample a non-native free energy basin. Category C includes trajectories that display a drifting behavior without convergence on the time scale of the trajectory length.

native state, Figure 4A), 27 into category 2 (well-defined, non-native free energy basins, Figure 4B), and 21 into category 3 (drifting behavior, Figure 4C). It is important to note that these three types of behavior are not exclusive—they reflect differential sampling properties for MD trajectories on the ~1 μs time scale of different proteins and their initial structures. In fact, a trajectory that drifts in the submicrosecond range can reach a well-defined global free energy minimum over longer time scales, occasionally visiting higher (i.e., excited) free energy basins.

Accurate Structure Selection from Free Energy Calculations. Although MD simulations on these longer time scales are currently not feasible on a routine basis, the results of this study suggest that MD simulations can help identify the lowest free energy basin sampled in a set of trajectories that started from different decoys. Figure 3 demonstrates that the free energy metric of eq 1 is sufficiently accurate to identify the native MD ensemble from decoy ensembles. For the two proteins 1HZ6 (Figure 3B) and 5CRO (Figure 3H), the free energies of the native states are only slightly lower than the best decoys, while for all other proteins the native state free energies are markedly lower than for any decoy ensemble. Since for proteins 1B3A and 1C9O the decoy structures with the lowest free energy have low RMSDs with respect to their native states (2.7 Å and 1.7 Å, respectively), a relatively small difference in free energy has only minor consequences with respect to structural selection.

The computation of the free energy requires identification of a (meta)stable subtrajectory corresponding to a free energy basin. For the proteins studied here, this is generally possible with the time spent in the selected free energy basin varying between a few tens of nanonseconds and a microsecond or longer. Only few trajectories that start from a decoy display extreme transient behavior. They do not reach any (meta)stable basin while drifting away from the initial structure. Because a free energy basin cannot be determined in these cases, they can be ruled out as representative models of the native ensembles of globular proteins.

Basins with the Longest Residence Times Are Most Native-Like. Because a low free energy alone does not determine protein stability, the protein residence times in the various free energy basins were assessed in both native and decoy trajectories. Figure 5 shows correlation plots for two representative proteins (1B3A, 1CTF) of the free energies of basins vs basin residence times (analogous figures for the other proteins are shown in Figure S12, Supporting Information). Both proteins show a pronounced anticorrelation between free energies and residence times. The longest residence times of 1 μs (or longer) belong to the native-state simulations, whereas short residence times are displayed by the basins with the highest free energies. Although the relatively sparse statistics preclude a more quantitative analysis, Figure 5 suggests that the native state basins, besides having the lowest free energy according to eq 1, possess increased stability by being separated from non-native basins by higher barriers. In native state simulations, loop regions have the highest degree of mobility, while regular secondary structures are most stable. By contrast, for many decoys, some well-formed α-helices or β-strands are prone to local unfolding and refolding. This is a clear signature that either the secondary structure is wrong or side-chain packing is nonoptimal in these regions. This type of MD information directly points to incorrect parts of a model, which is useful for subsequent structure optimization.

The other proteins show similar trends with the exception of proteins 1SHF and 1HZ6 (Figures S3 and S4, Supporting Information). A long 12 amino acid loop in the native state of 1SHF participates in a homodimeric interaction in the crystal structure, while the same loop is rather flexible during the MD simulation where it is represented as a monomer. It is possible that this protein behaves differently in solution and in its crystalline state, which could be tested, for example, by comparing experimental NMR residual dipolar couplings with the ones back-calculated from the PDB coordinates. In the

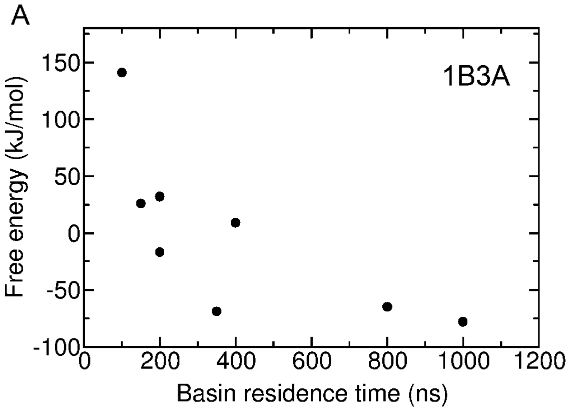

Figure 5. Relationship between protein free energies and residence times in free energy basins for both native and decoy trajectories for proteins (A) 1B3A and (B) 1CTF. For each trajectory, the basin with the longest residence time was selected. In both panels, the 1 μs residence times belong to the native trajectories (they correspond to a lower limit as the trajectories were limited to 1 μs).

native simulation of 1HZ6 (Figure S4A), a rapid local unfolding/refolding event was observed around 400 ns while retaining an average RMSD of ~2 Å with respect to the initial native structure. The decoy simulation of Figure S4B with an RMSD of ~3.75 Å from the native structure displayed a longer residence time in its decoy basin. This decoy is the only one of all 57 decoys studied in this work with a "native-like" 1 μs residence time.

Protein Structure Determination from Sparse Data or *ab Initio*. Over the years, a number of different strategies have been developed to improve protein structure refinement. In most standard NMR structure determination protocols, the protein structure is driven toward the native state by the use of pseudoenergy terms directly derived from the experimental data.[50,51] These constraints help overcome energy barriers and correct for inaccuracies in the underlying molecular mechanics force field. In the absence of sufficient experimental information for certain protein regions, e.g., the lack of proton−proton NOEs in NMR structure determination and the absence of electron density in X-ray crystallographic studies, the resulting structure increasingly depends on the inherent force-field quality and sampling efficiency of the underlying protocol, which is typically based on all-atom MD. Protein structure determination protocols using only chemical shift information, on the other hand, rely on knowledge-based potentials and the sampling efficiency of *ab initio* structure prediction algorithms.[52,53] Experimental chemical shift information is thereby used during the early stages, mainly for secondary structure definition, and during the final stage for the *selection* of the structural model with the highest consistency with chemical shift data.

In this work, we showed that both spontaneous convergence and thermodynamic selection can yield improved protein structures in the absence of experimental information. The use of free MD simulations yields native structures by spontaneously driving decoy sets for 1C8C and 1R69 into the native state. For the other proteins, the selection strategy using PRESTO proved successful to determine the native structure by thermodynamic selection. In this approach, the estimated thermodynamic free energy is used as a metric to discriminate between native structures and lower resolution structural models and decoys. In order to allow for sufficient relaxation of decoy structures, each simulation lasted ~1 μs. This simulation length was found to be adequate to allow the 76 amino acid globular protein ubiquitin to reversibly and statistically adequately sample its relevant conformational space.[12,54] In situations where the native structure is not reached during the MD simulation(s), the results presented here suggest that the structure with the lowest free energy provides the best approximation for the native state in terms of a low RMSD.

Both types of structure refinement strategies benefit from recent progress in available computer power and the quality of molecular mechanics force fields. Although current MD simulation lengths do not yield convergence to the native state from any arbitrary chosen initial state, the thermodynamic free energy evaluation of energy basins used here suggests that the bottleneck is due to insufficient sampling and not due to limitations of the force field used. Hence, further advances in sampling efficiency will directly benefit the generation of more realistic, i.e., more accurate, protein structural ensembles from lower resolution input data.

Energy, Entropy, and Free Energy Calculations. The generalized Born model (MM/GBSA) is a widely used implicit solvation model for MD simulations and free energy calculations, although the accuracy for electrostatic interactions is being debated.[55,56] For free energy calculations, the configurational entropy is typically estimated by normal-mode analysis in Cartesian space, which, however, can be inadequate at room temperature.[57] The recently developed configurational entropy calculation method (eq 2) based on the von Mises kernel density estimation in dihedral angle space utilizes the separability of entropy contributions from hard degrees of freedom (bond lengths and bond angles) and soft degrees of freedom (dihedral angles).[33] Taking advantage of the accuracy and computational speed of this method,[33,34] it is applied here to the configurational entropy part of the MM/GBSA free energy estimation of eq 1. We previously found that during local ligand binding events that leave the overall protein shape unchanged, the correlations among dihedral angles are largely conserved,[34] which allows the determination of an entropy difference without inclusion of correlation effects. However, because decoys and native structures may have different secondary structures and shapes, dihedral angle correlations are not necessarily conserved when protein segments change their secondary structures. Therefore, correlations between neighboring dihedral angles along the backbone and correlations among intraresidue side chain dihedral angles were both included in S (eq 2). Long-range correlations on the microsecond time scale are expected to be rare[42,58] and were therefore neglected, also because the inclusion of a large number of weak correlations tends to introduce significant amounts of "noise" with adverse effects on the accuracy of *S*.[34] Accurate entropies play an important role in PRESTO, since for about one-third of the proteins studied here the inclusion of the entropy is key to identifying the native state.

Ab initio protein structure prediction has become increasingly powerful due to improved algorithms that efficiently sample conformational space and the development of knowledge-based potentials to discriminate native-like states from decoys. In fact, bioinformatics methods have dominated protein structure prediction through homology modeling and protein threading in recent years.[2] Because of the coarse-grained nature of these knowledge-based potentials, they have been limited in their ability to produce high-resolution structures with high confidence and a low error rate. All-atom physical force fields could play an important role, but they have not been successful in protein structure prediction and refinement as was noted at a previous CASP competition.[59] There is an ongoing debate on whether this is because of insufficient sampling or the limited quality of the underlying force fields, or both.[27,29] The results for the 13 proteins described here point to sampling as a main limitation for direct structure refinement, since the free energies with their strong dependence on the force field quality correctly discriminate between the native structures and the lower resolution decoys. This conclusion is consistent with a recent submillisecond MD study using a physics-based force field, which succeeded in the folding of 12 small proteins to their native or near-native states.[60] As long-time MD simulations of larger multidomain proteins are becoming available, however, careful monitoring of sampling efficiency vs force-field accuracy will remain important. This can be achieved by comparison with quantitative experimental data and by computational thermodynamic approaches such as PRESTO.

Recent progress in the accuracy and extent of *in silico* sampling of the conformational space of proteins has the potential to alter the way both experimental and predicted protein structures are generated, refined, and validated. The availability of a realistic time-resolved Boltzmann-weighted native state ensemble of a protein directly helps in the understanding of its function. This includes the assessment of dynamic regions and their time scales, the relative stability of substates, and the ability to transmit a perturbation, for example due to ligand binding or protein−protein interactions at an allosteric site, to an active site,[61] highlighting the close relationship between protein dynamics, computational thermodynamics, and quantitative structural biology.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Detailed description of von Mises entropy computation; ribbon diagrams of all proteins studied (1 figure); pairwise RMSD plots analogous to Figure 2 for other proteins (10 figures); correlation plots analogous to Figure 5 for other proteins (1 figure). This information is available free of charge via the Internet at http://pubs.acs.org/.

■ AUTHOR INFORMATION

Corresponding Author

*Tel.: 850-644-1768. Fax: 850-644-8281. E-mail: Bruschweiler@magnet.fsu.edu.

Notes

Competing Interests. The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the National Science Foundation (grant MCB-0918362).

■ REFERENCES

(1) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Res.* **2000**, *28*, 235−242.

(2) Moult, J.; Fidelis, K.; Kryshtafovych, A.; Rost, B.; Tramontano, A. *Proteins* **2011**, *79*, 1−4.

(3) Bowie, J. U.; Eisenberg, D. *Proc. Natl. Acad. Sci. U. S. A.* **1994**, *91*, 4436−4440.

(4) Bradley, P.; Misura, K. M. S.; Baker, D. *Science* **2005**, *309*, 1868−1871.

(5) Wu, S. T.; Skolnick, J.; Zhang, Y. *BMC Biol.* **2007**, *5*, ARTN 17.

(6) Klepeis, J. L.; Lindorff-Larsen, K.; Dror, R. O.; Shaw, D. E. *Curr. Opin. Struct. Biol.* **2009**, *19*, 120−127.

(7) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins* **2006**, *65*, 712−725.

(8) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.; Kollman, P. *J. Comput. Chem.* **2003**, *24*, 1999−2012.

(9) Buck, M.; Bouguet-Bonnet, S.; Pastor, R. W.; MacKerell, A. D. *Biophys. J.* **2006**, *90*, L36−L38.

(10) Showalter, S. A.; Johnson, E.; Rance, M.; Brüschweiler, R. *J. Am. Chem. Soc.* **2007**, *129*, 14146−14147.

(11) Lange, O. F.; van der Spoel, D.; de Groot, B. L. *Biophys. J.* **2010**, *99*, 647−655.

(12) Long, D.; Li, D. W.; Walter, K. F. A.; Griesinger, C.; Brüschweiler, R. *Biophys. J.* **2011**, *101*, 910−915.

(13) Wickstrom, L.; Okur, A.; Simmerling, C. *Biophys. J.* **2009**, *97*, 853−856.

(14) Markwick, P. R.; Showalter, S. A.; Bouvignies, G.; Brüschweiler, R.; Blackledge, M. *J. Biomol. NMR* **2009**, *45*, 17−21.

(15) Showalter, S. A.; Brüschweiler, R. *J. Chem. Theory Comput.* **2007**, *3*, 961−975.

(16) Markwick, P. R. L.; Bouvignies, G.; Blackledge, M. *J. Am. Chem. Soc.* **2007**, *129*, 4724−4730.

(17) Trbovic, N.; Kim, B.; Friesner, R. A.; Palmer, A. G. *Proteins* **2008**, *71*, 684−694.

(18) Li, D. W.; Brüschweiler, R. *J. Phys. Chem. Lett.* **2010**, *1*, 246−248.

(19) Markwick, P. R. L.; Cervantes, C. F.; Abel, B. L.; Komives, E. A.; Blackledge, M.; McCammon, J. A. *J. Am. Chem. Soc.* **2010**, *132*, 1220−1221.

(20) Beauchamp, K. A.; Lin, Y. S.; Das, R.; Pande, V. S. *J. Chem. Theory Comput.* **2012**, *8*, 1409−1414.

(21) Li, D. W.; Brüschweiler, R. *Angew. Chem., Int. Ed.* **2010**, *49*, 6778−6780.

(22) Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. *Proteins* **2010**, *78*, 1950−1958.

(23) Nilges, M.; Brunger, A. T. *Protein Eng.* **1991**, *4*, 649−659.

(24) Vieth, M.; Kolinski, A.; Brooks, C. L.; Skolnick, J. *J. Mol. Biol.* **1994**, *237*, 361−367.

(25) Lee, M. R.; Baker, D.; Kollman, P. A. *J. Am. Chem. Soc.* **2001**, *123*, 1040−1046.

(26) Lee, M. R.; Tsai, J.; Baker, D.; Kollman, P. A. *J. Mol. Biol.* **2001**, *313*, 417−430.

(27) Wroblewska, L.; Skolnick, J. *J. Comput. Chem.* **2007**, *28*, 2059−2066.

(28) Fan, H.; Mark, A. E. *Protein Sci.* **2004**, *13*, 211−220.

(29) Chen, J. H.; Brooks, C. L. *Proteins* **2007**, *67*, 922−930.

(30) Ishitani, R.; Terada, T.; Shimizu, K. *Mol. Simul.* **2008**, *34*, 327−336.

(31) Shell, M. S.; Ozkan, S. B.; Voelz, V.; Wu, G. H. A.; Dill, K. A. *Biophys. J.* **2009**, *96*, 917−924.
(32) Hou, T. J.; Wang, J. M.; Li, Y. Y.; Wang, W. *J. Chem. Inf. Model.* **2011**, *51*, 69−82.
(33) Li, D. W.; Brüschweiler, R. *Phys. Rev. Lett.* **2009**, *102*, 118108.
(34) Li, D. W.; Showalter, S. A.; Brüschweiler, R. *J. Phys. Chem. B* **2010**, *114*, 16036−16044.
(35) Tsai, J.; Bonneau, R.; Morozov, A. V.; Kuhlman, B.; Rohl, C. A.; Baker, D. *Proteins* **2003**, *53*, 76−87.
(36) Kim, D. E.; Blum, B.; Bradley, P.; Baker, D. *J. Mol. Biol.* **2009**, *393*, 249−260.
(37) Berendsen, H. J. C.; van der Spoel, D.; van der Unen, R. *Comput. Phys. Commun.* **1995**, *91*, 43−56.
(38) Lindahl, E.; Hess, B.; van der Spoel, D. *J. Mol. Model.* **2001**, *7*, 306−317.
(39) van der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. *J. Comput. Chem.* **2005**, *26*, 1701−1718.
(40) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. *J. Chem. Theory Comput.* **2008**, *4*, 435−447.
(41) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926−935.
(42) Li, D. W.; Meng, D.; Brüschweiler, R. *J. Am. Chem. Soc.* **2009**, *131*, 14610−14611.
(43) Li, D. W.; Brüschweiler, R. *J. Chem. Theory Comput.* **2011**, *7*, 1773−1782.
(44) Onufriev, A.; Bashford, D.; Case, D. A. *Proteins* **2004**, *55*, 383−394.
(45) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127−6129.
(46) Schaefer, M.; Karplus, M. *J. Phys. Chem.* **1996**, *100*, 1578−1599.
(47) Lee, M. S.; Salsbury, F. R.; Brooks, C. L. *J. Chem. Phys.* **2002**, *116*, 10606−10614.
(48) Killian, B. J.; Kravitz, J. Y.; Gilson, M. K. *J. Chem. Phys.* **2007**, *127*, 024107.
(49) Kubelka, J.; Hofrichter, J.; Eaton, W. A. *Curr. Opin. Struct. Biol.* **2004**, *14*, 76−88.
(50) Nilges, M.; Clore, G. M.; Gronenborn, A. M. *FEBS Lett.* **1988**, *229*, 317−324.
(51) Guerry, P.; Herrmann, T. *Q. Rev. Biophys.* **2011**, *44*, 257−309.
(52) Shen, Y.; Lange, O.; Delaglio, F.; Rossi, P.; Aramini, J. M.; Liu, G. H.; Eletsky, A.; Wu, Y. B.; 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.
(53) Shen, Y.; Vernon, R.; Baker, D.; Bax, A. *J. Biomol. NMR* **2009**, *43*, 63−78.
(54) Long, D.; Brüschweiler, R. *PLoS Comput. Biol.* **2011**, *7*, ARTN e1002035.
(55) Singh, N.; Warshel, A. *Proteins* **2010**, *78*, 1705−1723.
(56) Vicatos, S.; Roca, M.; Warshel, A. *Proteins* **2009**, *77*, 670−684.
(57) Li, D. W.; Khanlarzadeh, M.; Wang, J. B.; Huo, S. H.; Brüschweiler, R. *J. Phys. Chem. B* **2007**, *111*, 13807−13813.
(58) Fenwick, R. B.; Esteban-Martin, S.; Richter, B.; Lee, D.; Walter, K. F. A.; Milovanovic, D.; Becker, S.; Lakomek, N. A.; Griesinger, C.; Salvatella, X. *J. Am. Chem. Soc.* **2011**, *133*, 10336−10339.
(59) Zhang, Y. *Curr. Opin. Struct. Biol.* **2008**, *18*, 342−348.
(60) Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E. *Science* **2011**, *334*, 517−520.
(61) Long, D.; Brüschweiler, R. *J. Am. Chem. Soc.* **2011**, *133*, 18999−19005.