JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

Significant Refinement of Protein Structure Models Using a Residue-Specific Force Field

Sangni Xun,[†] Fan Jiang,*[,†] and Yun-Dong Wu*[,†,‡]

[†]Laboratory of Computational Chemistry and Drug Design, Laboratory of Chemical Genomics, Peking University Shenzhen Graduate School, Shenzhen, 518055, China

[‡]College of Chemistry and Molecular Engineering, Peking University, Beijing, 100871, China

Ⓢ *Supporting Information*

ABSTRACT: An important application of all-atom explicit-solvent molecular dynamics (MD) simulations is the refinement of protein structures from low-resolution experiments or template-based modeling. A critical requirement is that the native structure is stable with the force field. We have applied a recently developed residue-specific force field, RSFF1, to a set of 30 refinement targets from recent CASP experiments. Starting from their experimental structures, 1.0 μs unrestrained simulations at 298 K retain most of the native structures quite well except for a few flexible terminals and long internal loops. Starting from each homology model, a 150 ns MD simulation at 380 K generates the best RMSD improvement of 0.85 Å on average. The structural improvements roughly correlate with the RMSD of the initial homology models, indicating possible consistent structure refinement. Finally, targets TR614 and TR624 have been subjected to long-time replica-exchange MD simulations. Significant structural improvements are generated, with RMSD of 1.91 and 1.36 Å with respect to their crystal structures. Thus, it is possible to achieve realistic refinement of protein structure models to near-experimental accuracy, using accurate force field with sufficient conformational sampling.

Homology model
Refined structure
MD RSFF1
RMSD = 5.34 Å
RMSD = 1.36 Å

■ INTRODUCTION

The determination of protein three-dimensional (3D) structure is of vital importance in many aspects of modern biological and medical research.[1−3] Although the number of high-resolution crystal structures determined by X-ray diffraction increases rapidly in recent years, an increasing number of protein structures with lower resolutions are generated with small-angle X-ray diffractions, cryo-electron microscopy (cryo-EM), and NMR experiments.[4−8] These structures need to be further refined. In another front, a tremendous gap between the number of experimental protein structures and that of known sequences strongly motivates the prediction of protein structures from their amino acid sequences.[9−11] The most successful method is the template-based modeling, utilizing the information from an increasing number of high-resolution structures in the Protein Data Bank (PDB).[12−17] These template-based protein structure predictions normally generate structures of root-mean-square deviations (RMSDs) of 4−8 Å with respect to experimental structures.[13,18,19] Further refinement of these predicted structures becomes an important area of protein modeling because higher resolution is often needed for the detailed study of protein functions.[2,20] Both knowledge-based energy functions and physics-based force fields were used in structure refinement.[21−36] Early refinement efforts were mainly carried out in implicit solvent, with simple energy minimizations, Monte Carlo simulations, or the molecular dynamics (MD) simulations. Recent studies focus on using the MD simulations with all-atom force fields, especially with explicit solvent.[29−31,37]

A category of the Critical Assessment of Structure Prediction (CASP) experiment is to evaluate current protein structure refinement methodologies in a blind way.[38−40] In these studies, usually only limited improvements of structures can be achieved. For example, Zhang et al. have shown that, using fragment-guided MD simulations with restraint from template structures, global distance test high accuracy (GDT-HA) score[41] of models from CASP8−9 could be improved by 0.6% on average; however no obvious improvement was obtained in terms of the RMSD to experimental structures.[28] The best reported results comes from Feig's group, who performed MD simulations using the CHARMM36 force field at 298 K to refine models from the CASP8−10. In their work, the best RMSD improvement from their 200 ns trajectories averaged at −0.33 Å (CASP8−9) and −0.29 Å (CASP10).[29,31] And much less improvements were achieved in blindly predicted structures. When restraint was applied to structures, homology models were more likely to be refined.[28,29,31] But strong restraint may confine the conformations in the region near the initial model and decrease the extent of low-energy conformation sampling. Indeed, in the recent CASP10 model refinement category, there is still no prediction more similar to the experimental structure than to the initial model.[42]

The success of structural refinement highly relies on the accuracy of the potential energy functions (force field) as well as sufficient conformational sampling. Utilizing their unique

Received: January 13, 2015
Published: March 3, 2015

DOI: 10.1021/acs.jctc.5b00029

Table 1. 30 CASP Targets Used in This Study, with Their Experimental Information and the Simulation Results with the RSFF1 and the CHARMM36 Force Fields

							MD (homology model)					
			experimental structure		MD (native)[a] RSFF1		initial model		RSFF1[b]		CHARMM36[c]	
ID	CASP	no. res.	method	resolution (Å)	average RMSD (Å)	average GDT-TS	RMSD	GDT-HA	best ΔRMSD[d]	best ΔGDT-HA[d]	best ΔRMSD[d]	best ΔGDT-HA[d]
TR389	8	134	X-ray	2.20	1.50	86.31	2.63	63.30	−0.92	7.22	−0.62	−1.9
TR432	8	130	X-ray	1.95	1.40	90.77	1.65	77.50	−0.65	4.62	−0.31	6.4
TR453	8	87	X-ray	2.14	1.58	86.58	1.47	71.30	−0.49	9.67	−0.09	5.8
TR454	8	192	X-ray	1.75	2.91	71.82	3.24	42.30	−0.26	4.95	−0.20	4.0
TR469	8	63	NMR	N/A	1.84	90.76	2.18	63.50	−0.94	20.63	−0.19	3.2
TR488	8	95	X-ray	1.30	3.25	85.96	2.11	75.00	−0.36	6.84	−0.26	7.1
TR517	9	159	X-ray	1.92	1.69	85.42	4.64	54.70	−0.82	6.13	−0.12	3.0
TR530	9	80	X-ray	2.15	1.66	87.53	1.99	70.90	−1.08	13.44	−0.35	3.1
TR557	9	125	NMR	N/A	1.55	84.23	4.06	48.60	−0.74	10.80	−0.84	9.0
TR567	9	142	X-ray	2.80	1.59	86.09	3.44	59.00	−1.09	6.87	−0.20	5.3
TR568	9	97	X-ray	1.50	1.58	88.36	6.15	35.80	−0.89	9.54	−0.43	4.6
TR569	9	79	NMR	N/A	1.74	83.44	3.01	52.90	−1.01	11.39	−0.69	7.6
TR574	9	102	X-ray	1.50	1.73	85.94	3.58	39.70	−0.38	16.67	−0.40	6.4
TR576	9	138	X-ray	2.29	2.60	79.27	6.85	46.90	0.22	−2.72	0.00	2.9
TR592	9	105	X-ray	2.50	2.48	78.49	1.26	74.00	−0.61	14.76	−0.20	9.1
TR594	9	140	X-ray	2.50	1.07	93.55	1.82	67.50	−0.71	14.11	0.00	4.1
TR606	9	123	X-ray	1.60	1.92	87.88	4.85	53.20	−1.74	9.97	−1.51	5.7
TR614	9	105	X-ray	2.00	2.74	89.93	5.34	58.81	−1.86	9.53	−0.13	6.3
TR622	9	122	X-ray	1.90	1.57	88.73	7.47	49.40	−1.35	11.68	−0.32	7.4
TR624	9	69	X-ray	1.90	1.06	94.07	5.19	36.60	−2.00	13.77	−0.89	6.2
TR655	10	175	NMR	N/A	3.34	76.27	4.65	49.28	−0.42	4.43	−0.15	1.0
TR663	10	152	X-ray	1.85	1.14	92.24	3.37	49.34	−0.84	10.36	−0.37	3.1
TR671	10	88	X-ray	2.07	4.68	79.50	7.72	36.36	−1.07	−1.70	−0.12	0.3
TR674	10	132	X-ray	2.23	1.76	90.55	3.44	71.40	−0.72	9.28	−0.39	7.0
TR679	10	199	X-ray	1.80	1.37	88.76	3.95	51.63	−0.61	5.66	−0.23	2.3
TR696	10	100	X-ray	1.50	2.64	75.44	3.52	52.00	−0.93	8.25	−0.38	9.0
TR698	10	119	X-ray	N/A	2.57	77.96	4.65	45.38	−0.32	2.52	−0.32	3.6
TR704	10	235	X-ray	1.60	1.85	80.77	2.78	49.15	−0.96	19.96	−0.45	11.8
TR705	10	96	X-ray	1.91	1.23	90.17	4.71	44.79	−1.06	10.68	−0.37	10.9
TR708	10	196	X-ray	1.60	2.56	84.81	4.63	72.83	−0.94	2.94	−0.32	0.0
avg.					2.02	85.39			−0.85	9.07	−0.37	5.1

[a] 1 μs unrestrained simulations at 298 K starting from experimental structures. [b] 150 ns weakly restrained simulations at 380 K starting from the homology models. [c] The results of 200 ns MD simulations from Feig's group.[29,31] [d] Cα-RMSD and GDT-HA changes of the best structure sampled in MD simulation compared with the initial model, negative ΔRMSD and positive ΔGDT-HA indicate improvement.

advantage of ANTON machine, Shaw's group recently applied 100 μs long-time simulations to 25 CASP8−9 refinement targets[30] with the CHARMM22* force field,[43,44] which has been shown to successfully fold a series of fast folding peptides/proteins.[45] Starting from their experimental structures, the long-time simulations showed that most of them drifted away from experimental structures. That means the force field cannot stabilize most of the native structures of these targets. They suggested that it is probably more urgent to improve the accuracy of force field than to improve conformational sampling in order to achieve protein structure refinement using all-atom force fields.[30]

We recently developed a residue-specific force field, RSFF1, by modifying the OPLS-AA/L force field with residue-specific torsional (ϕ, ψ, χ_1,χ_2, etc.) functions for the 20 amino acid residues and some 1−5 and 1−6 van der Waals parameters for Ser, Thr, Asp, and Asn residues.[46] The parameters were developed by fitting the torsional free energy distributions of the 20 dipeptide models from molecular dynamics simulations to those obtained from the statistical analysis of coil residues (coil library) of high resolution protein structures from PDB.[47,48] This force field has successfully folded a set of 16 mini-proteins.[49]

Here, we evaluate the applicability of the RSFF1 in protein structure refinement. We first carried out 1.0 μs unrestrained simulations at 298 K for 30 experimental structures from CASP8−10 refinement targets to see whether the force field can stabilize the native structures. Then MD simulations at 380 K was carried for 150 ns for the 30 homology models with weak restraints, to investigate whether RSFF1 can improve systematic refinement. Finally, for two homology models (TR614 and TR624) with RMSD > 5 Å, long-time replica exchange MD simulations were carried out to see whether structures close to experimental resolution can be generated.

■ MATERIALS AND METHODS

Residue-Specific Force Field RSFF1. The overall potential energy function of RSFF1 is within the framework of classical force field:

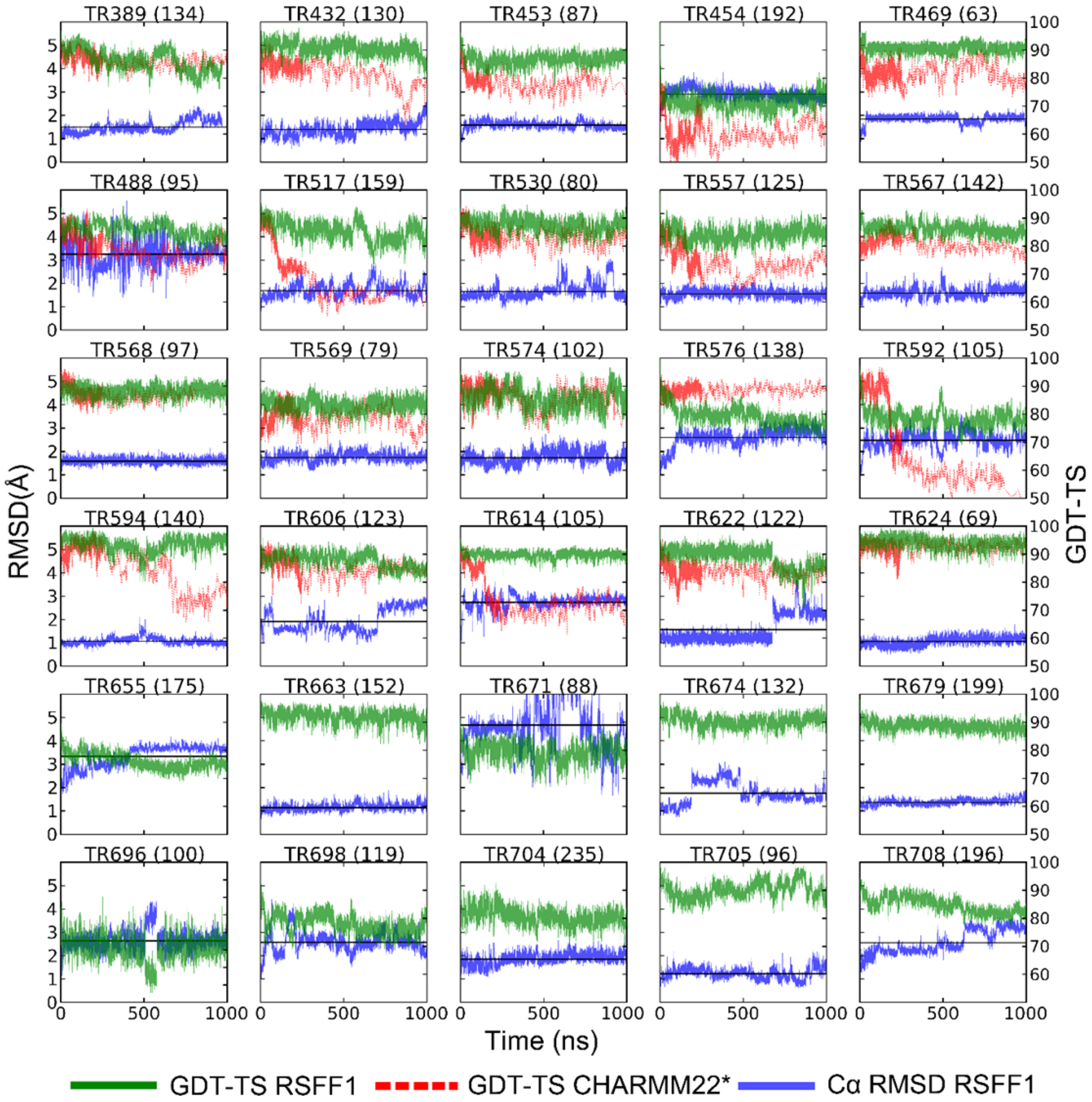

Figure 1. Stability of experimental structures in 1 μs unrestrained MD simulations. For each protein simulated using the RSFF1, Cα RMSD (blue line), and GDT-TS (green line) as a function of time are given. The GDT-TS (red line) from previous simulations[30] using CHARMM22* force field are also shown for comparison. The solid black line in each plot gives the average Cα RMSD during the entire RSFF1 simulations. The number of residues for each protein is given in parentheses after its target ID.

$$V_{\text{total}} = V_{\text{bond}} + V_{\text{angle}} + V_{\text{torsion}} + V_{\text{local-LJ}} + V_{\text{LJ}} + V_{\text{local-Coulomb}} + V_{\text{Coulomb}} \quad (1)$$

All the bond stretching (V_{bond}), bond angle bending (V_{angle}), van der Waals interactions (using Lennard-Jones potential V_{LJ}), and electrostatic interactions (Coulomb potential, V_{Coulomb}) are adopted from the OPLS-AA/L force field.[50,51] On the other hand, the dihedral angle potential (V_{torsion}) for all the rotable bonds were reparameterized using the statistical free energy surfaces (potential of mean force, PMF) from protein coil library as reference, such that the backbone ϕ, ψ, and side-chain χ PMF obtained for dipeptide simulations agree excellently with the reference data. The 1−4 Lennard-Jones interactions ($V_{\text{local-LJ}}$) are scaled by a factor of 0.50, whereas the 1−4 Coulomb interactions ($V_{\text{local-Coulomb}}$) are not scaled down to achieve more balanced local electrostatic interactions. Unlike common protein force fields, special parameters were used for some 1−5 and 1−6 van der Waals interactions (included in the $V_{\text{local-Coulomb}}$) to achieve better agreement with reference data.

Unlike most protein force fields, RSFF1 was parametrized using free energies as reference instead of potential energies. However, the torsion parameters can be efficiently optimized by adding corrections equal to the difference between the coil library PMF and the simulated PMF:

$$V_{\text{new}}(\theta) - V_{\text{old}}(\theta) = G_{\text{coil}}(\theta) - G_{\text{MD,old}}(\theta) \quad (2)$$

To fit the statistical ϕ, ψ PMF of different amino acid residues under different side-chain rotamers, we use residue-specific dihedral angle (torsion) parameters. The parameters for specially treated 1−5/1−6 van der Waals interactions were modified manually. The detailed procedure of parametrization can be found in our previous works.[46]

DOI: 10.1021/acs.jctc.5b00029
J. Chem. Theory Comput. 2015, 11, 1949−1956

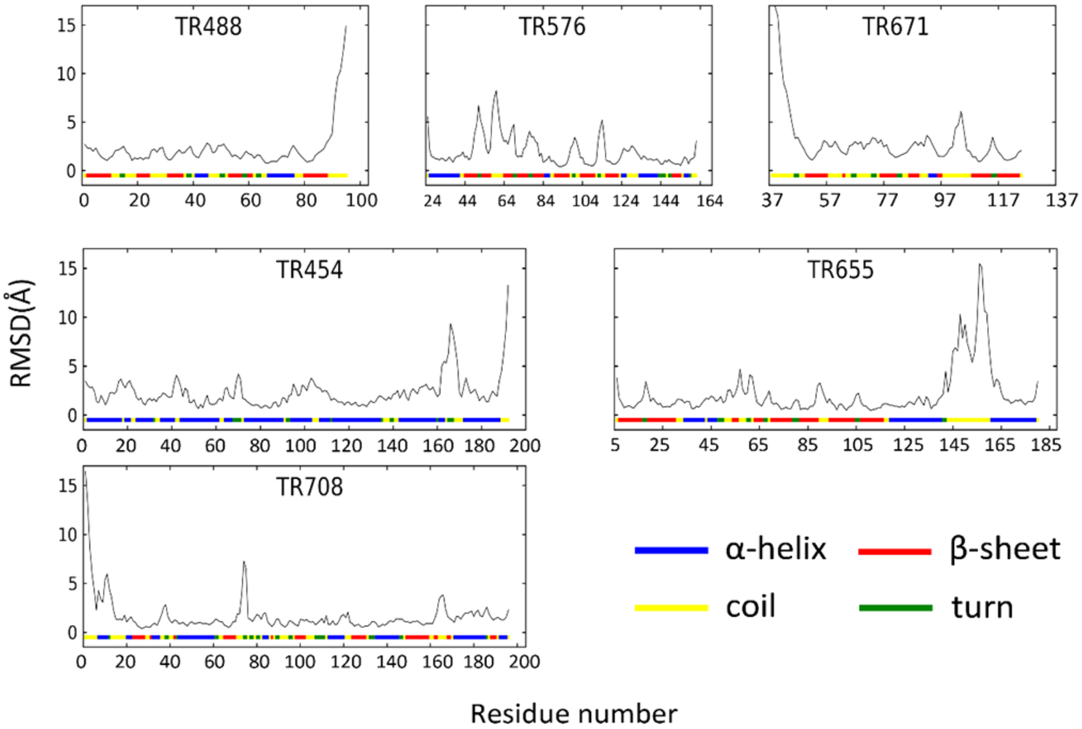

Figure 2. Residue-wise Cα RMSD between structures sampled in the 1 μs MD simulations and corresponding experimental ones. Only the proteins with average overall Cα RMSD > 2.50 Å are shown. The bottom of each plot shows the experimental secondary structures: α-helix (blue), β-sheet (red), turn (green), and coil (yellow).

Protein Targets for the Refinement. We chose 30 single-domain proteins from the refinement category of CASP8−10, excluding those with missing residues in the experimental structures. As shown in Table 1, their sizes range from 63 (TR469) to 235 (TR704) residues. Their sequences have quite low similarities, with pairwise identities between 8% and 25%. They have diverse secondary structures, and three of them contain disulfide bridges (TR574, TR592, TR698). The homology models provided by CASP have a large coverage range (1−8 Å) of Cα RMSD, which were also studied by other groups for protein structure refinement.[2,29−31] The diversity of the set of proteins is necessary for us to reach statistically meaningful and general conclusions. Their model structures and experimental structures were downloaded from the CASP Web site and Protein Data Bank,[14] respectively. For TR614 and TR698, the experimental structures provided by the CASP organizer are used, because residues 36−41 of TR614 are missing in its PDB structure (3VOQ) and the PDB structure of TR698 is unavailable. The hydrogen atoms of the proteins were added under neutral pH, except for TR671, for which the PROPKA3.0 server[52,53] was used under its experimental condition pH = 10.5.

Simulation Settings and Trajectory Analysis. All simulations were carried out using the Gromacs 4.5.4 software[54,55] with the RSFF1 force field.[46] Each protein was solvated in an octahedron box (length 48.4−68.8 Å) of TIP4P-Ew water.[56] Na^+ or Cl^- ions were added to neutralize the system at 0.05 mol/L. The cutoffs of electrostatic and van der Waals interactions were both at 9 Å. The long-range electrostatic interactions were calculated with the PME method,[57] and the energy and pressure were corrected to account for long-range dispersion interactions. The integration step time was set to 3 fs. All bonds involving hydrogens were constraint using the LINCS. These simulation settings are the same as in our pervious development and application of the RSFF1 force field.[46,49] The Berendsen method was used for temperature coupling and pressure coupling at 1 atm. After energy minimization, a 5 ns NPT simulation was performed with temperature increased from 10 to 298 K in the first ns. During this pre-equilibrium process, all Cα atoms were restrained by 10 kJ/(mol Å^2) harmonic potentials. The last snapshot was used as the initial structure for subsequent productive NVT simulation. Structures were stored every 0.6 ps.

All Cα RMSD were calculated by the MDAnalysis program.[58] The GDT-TS and the GDT-HA scores were generated using the LGA local program. The secondary structure assignments were done with the DSSP.[59] The PyMOL 1.2 was used to draw graphics of molecules.[60] For the clustering analysis, the "Gromos" method[61] with cutoff of 3.0 Å was applied on 8667 structures (every 150 ps) sampled in the lowest-temperature replica.

RESULTS AND DISCUSSION

Stability of the Experimental Structures. As shown in Figure 1, most of the 1.0 μs MD simulations with the RSFF1 force field can keep the proteins close to their experimental structures. Among the 30 proteins, 20 of them (TR389, TR432, TR453, TR469, TR517, TR530, TR557, TR567, TR568, TR569, TR574, TR594, TR606, TR622, TR624, TR663, TR674, TR679, TR704, TR705) have average Cα RMSD < 2.50 Å with low fluctuations throughout the simulations (Figure 1, Table 1). Thus, these structures (group 1) can be considered quite stable with the RSFF1 force field. The simulations of TR454, TR576, TR592, TR614, TR696, TR698, and TR708 give average Cα RMSD between 2.5 and 2.9 Å. Overall, the RMSD of these structures (group 2) are also stable

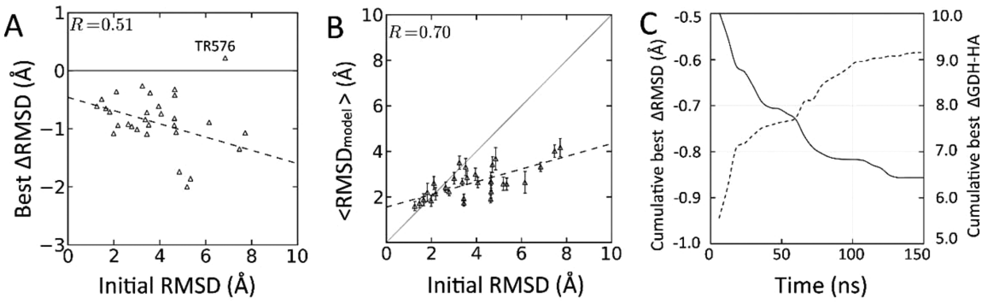

Figure 3. Refinement of the homology models through the 150 ns MD simulations at 380 K. (A) Best RMSD improvements (Best ΔRMSD) sampled in entire simulations versus the corresponding RMSD of initial models (initial RMSD). TR576 is not included in the linear fitting because of the crystal packing observed in its X-ray structure.[40] (B) Average RMSD of structures sampled in MD simulations with respect to the homology models (⟨RMSD$_{model}$⟩), against the corresponding initial RMSD. (C) Cumulative best ΔRMSD (solid line, left *y*-axis) and best ΔGDT-HA (dashed line, right *y*-axis), averaged over simulations of 30 models.

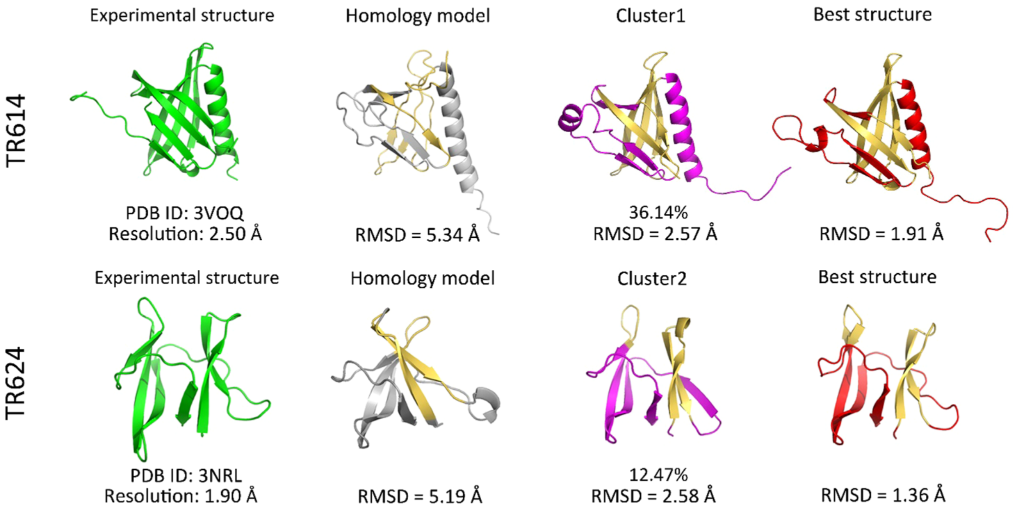

Figure 4. Experimental structures (green), initial (gray) and refined homology models of TR614 and TR624. For the refined models, both the representative structures of major clusters (magenta) and the best improved structures (red) from the lowest-temperature replica in the REMD simulations are shown. The regions in yellow undergo significant structure changes during the REMD simulations. Below these structures are their RMSDs to experimental structures.

during the 1 μs simulations, except for TR708, which starts to have increased RMSD near the end of the simulation.

Unlike RMSD, GDT-TS is less sensitive to the structural fluctuations of flexible terminals and long loops. It essentially measures the percentage of the residues that can be aligned with the reference structure within a set of distance cutoffs.[39,41] As shown in Figure 1, the group 1 targets have calculated average GDT-TS of these simulated structures >80%, consistent with the satisfactory average RMSD values. The group 2 targets have average GDT-TS around 80%, except for TR454, which is around 70%. Also, the GDT-TS is generally quite constant, indicating that the structures can be reasonably stable with the RSFF1.

For comparison, the GDT-TS of some targets from unrestrained MD simulations using CHARMM22* force field are also shown in Figure 1. In most cases, the calculated GDT-TS values by the CHARMM22* force field are smaller than those corresponding values by the RSFF1. The structures drifted away from the corresponding experimental structures in the CHARMM22* simulations of TR454, TR517, TR557, TR592, and TR614. For example, RSFF1 simulation of TR614 gives stable and high GDT-TS of about 90%, whereas its GDT-TS dropped to near 70% in previous CHARMM22* simulation.

However, RSFF1 simulations of TR488, TR655, and TR671 give relatively larger average RMSD values of 3.3, 3.3, and 4.7 Å, respectively. To gain more insight, these structures along with TR454, TR576, and TR708 were further analyzed. Shown in Figure 2 are the calculated residue-by-residue deviations of these structures to the corresponding experimental structures, averaged over the 1 μs simulation trajectories. For TR454, TR488, TR671, and TR708, their overall structures are actually quite close to experimental structures except for the C- or N-terminal, which have deviations over 10 Å. These terminal fluctuations are quite common in solution molecular dynamics, and often contribute significantly to the overall Cα RMSDs.[62] If these terminal residues are excluded, the calculated average Cα RMSDs are reduced to <2.0 Å.

As shown in Figure 2, the 3.3 Å RMSD of TR655 from our RSFF1 simulation is resulted from its long internal loop (residues 141 to 160) near the C-terminal, which is also very flexible in its NMR ensembles (Figure S1A). When that loop is excluded, the global RMSD fell below 2 Å. The TR576 is the only one that has considerable fluctuations in secondary structure regions. These happen in residues 58−69 and 96−101. The result is similar to that found by MacCallum et al.[40] Its experimental structure suffers from severe crystal packing (see Figure S1B, dark green).[40] Thus, the influenced part of the

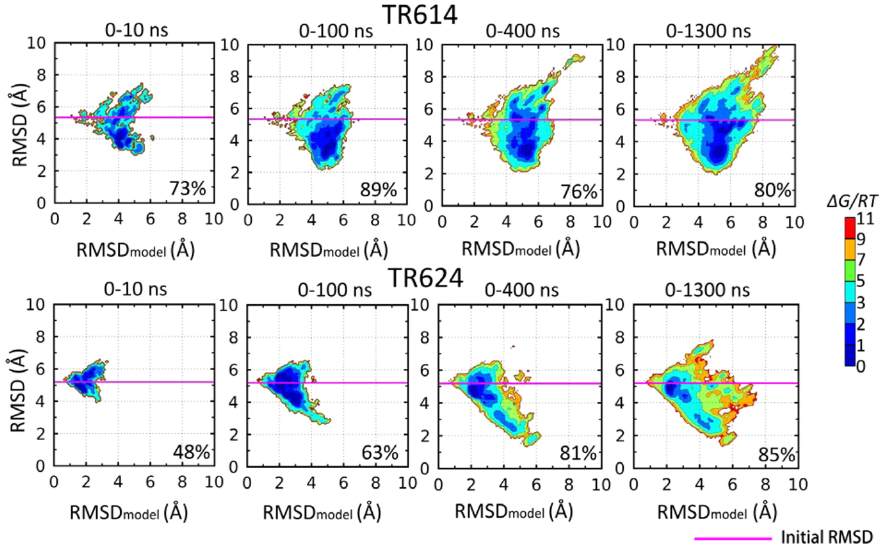

Figure 5. Free energy surfaces (FES) as a function of the Cα RMSD to the experimental structure (RMSD) and the Cα RMSD to the initial homology model ($\mathrm{RMSD_{model}}$), from the REMD simulations of TR614 and TR624. In each plot, magenta line marks the initial RMSD, and the percentage of snapshots with lower RMSD is also given. Results from four different time windows are given for each protein.

protein may not maintain the secondary structure in solution. We also calculated the average fractions of α-helix and β-sheet residues of the 30 proteins during the 1.0 μs RSFF1 simulations. The results are essentially the same as those from the experimental structures (Figure S2 and Table S1). Thus, the RSFF1 satisfy the required accuracy for protein structure refinement.

Short-Time MD Simulations of the Homology Models. To test whether structural improvement can be quickly achieved, we carried out a 150 ns normal MD simulation starting from each of the 30 homology models. They have a large coverage range (1−9 Å) of Cα RMSD and contain a variety of secondary structures. Since it has been shown that the RSFF1 tends to over stabilize the folded structure,[49] the simulation temperature was set at 380 K in order to enhance conformational sampling. Harmonic potentials of 0.02 kcal/(mol Å^2) were applied to restrain the positions of all Cα atoms, which was much weaker than previously reported.[29,31] Previous unrestrained MD simulations initiated from homology models at 300 K usually deteriorate their structures (as compared with experimental ones) rapidly.[29,30]

During the 150 ns MD simulations at 380 K, only about 10 (TR432, TR454, TR488, TR517, TR576, TR594, TR671, TR655, TR696, TR698) out of all 30 homology models obviously drifted away from their experimental structures (Figure S3, S4). Noticeably, RMSD improvement of over 1.0 Å can be found for 11 targets (Table 1), and the best GDT-HA improvements sampled in 13 homology models are more than 10%. Averaged over all 30 targets, the sampled best structures have an improvement of −0.85 Å in RMSD (ΔRMSD) and 9.1% in GDT-HA (ΔGDT-HA) with respect to the experimental structures. These compare quite favorably to the corresponding values of −0.37 Å (ΔRMSD) and 5.1% (ΔGDT-HA) reported by Feig et al. for the same set of 30 targets using the CHARMM36 force field.[29,31]

Interestingly, as shown in Figure 3A, the structural improvements achieved (best ΔRMSD) correlate with the qualities of these homology models (initial RMSD from the experimental structures). In line with this, less deviations from the initial homology models (⟨$\mathrm{RMSD_{model}}$⟩) were observed in the simulations of the models more similar (smaller initial RMSD) to the experimental structures (Figure 4B). Therefore, RSFF1 simulations are unlikely to move them away from the native structure. Shown in Figure 4C are the average improvement of RMSD and GDT-HA as a function of simulation time. Although significantly structure improvement can be observed quite early, the simulations using the RSFF1 give constantly increasing structure refinement as the simulation time increases. These features indicate that longer simulation time is helpful to generate better refined structures.

More Realistic Refinement with Enhanced Conformational Sampling. Although the above short-time simulations using the RSFF1 force field indicate that improved structures can be generated efficiently, it is critical to investigate whether structures approaching the experimental resolution can be generated by better conformational sampling without any structural restraint. To that end, homology models of TR624 and TR614 were selected for further MD simulations. Their RMSD to corresponding experimental structures are 5.19 and 5.34 Å, respectively (Figure 4). Shaw and co-workers reported a successful refinement of TR624 to a RMSD of <2 Å using the CHARMM22* force field,[30] but there is no report on high-quality refinement of TR614.

We use the replica-exchange molecular dynamics (REMD) simulations to enhance the conformational sampling. Sixteen replicas were used for both targets, with temperatures ranging from 350 to 420 K and from 310 to 410 K for TR614 and TR624, respectively. Each replica was simulated for 1.3 μs. As shown in Figure 4, structures with RMSD < 2 Å were observed in the simulations of both targets, significantly improved upon their initial homology models (RMSD > 5 Å) and within their experimental resolutions. This is better than their best structures sampled in the 150 ns MD simulations (RMSD: 3.48 and 3.19 Å for TR614 and TR624, respectively). Significant structure changes were be observed in the REMD simulations. Especially, β-sheets very similar to experimental structures were formed. In TR614, the N-terminal region forms a β-sheet in the simulation, which is missing in the initial model. In TR624, the wrong positions of β-strands were corrected. Therefore, given a reliable force field, a good conformational sampling is also crucial to achieve significant structure refinement for real applications.

In real structure prediction, the best structure cannot be picked out based on its RMSD to the experimental structure, which is unknown *a priori*. A simplest blind prediction is to carry out clustering analysis and to choose the representative structures of the most populated clusters. As shown in Figure 4, the structure from the largest cluster of TR614 and the structure from the second-largest cluster of TR624 have RMSD of 2.57 and 2.58 Å, close to their experimental structures. Therefore, it is possible to achieve realistic protein structure refinement using explicit-solvent MD simulations.

As shown in Figure 5, the structure of TR614 quickly moves away from the initial homology model during the simulation. Within the first 10 ns, 73% of structures sampled are already better and RMSD as low as 3.03 Å can be reached from the initial 5.34 Å, and structures quite different from the initial model ($\mathrm{RMSD}_{\mathrm{model}}$ up to 6.0 Å) were sampled. However, the significantly refined structures around 2 Å were still not sampled within first 10 ns, which can be observed within 100 ns. The FES from the first 100 ns is already close to the FES from the whole simulation of 1.3 μs. After 100 ns, the REMD simulation sampled some structures very different from both initial homology model and the experimental structure, leading to a lower percentage of improved structures. Unlike that of TR614, the conformational sampling of TR624 seems much slower. The two basins of RMSD ~ 2.6 Å and < 2.0 Å (significantly improved structures) were only observed after 100 ns. There is also significant increase of the percentage of improved structures (from 63% to 81%). Similar to the situation of TR614, there is an expansion of reached conformational region later in the REMD simulation. Noticeably, TR624 is actually smaller (69 residues) than TR614 (105 residues). Therefore, the total simulation time needed to achieve converged structure refinement is system-dependent and difficult to be estimated from the protein size.

CONCLUSION

To achieve significant refinement of a low-quality protein structure, both accurate force field and sufficient conformational searching are necessary. In general, the new residue-specific force field RSFF1 can well stabilize the protein native structures. Applied on homology models, the new force field can give structure improvements better than previously reported, during short-time MD simulations at relatively high temperature. More significant refinements can be achieve by using long-time replica-exchange MD simulations with enhanced conformational sampling. Thus, low-quality homology models (RMSD > 5 Å) can be refined to near experimental resolution in real applications including structure predictions.

ASSOCIATED CONTENT

Ⓢ Supporting Information

Average fraction of different secondary structures from 1.0 μs RSFF1 simulations, results of the cluster analysis of REMD simulations, representing structures from 1.0 μs simulations initiated from experimental structures of TR655 and TR576, averaged fractions of the secondary structures from 1 μs MD simulations plotted against those from experimental structures; Cα-RMSD and GDT-HA as a function of time for 380 K weak-restrained simulations initiated from the homology models. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Authors

*E-mail: jiangfan@pku.edu.cn (F.J.).
*E-mail: wuyd@pkusz.edu.cn (Y.-D.W.).

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank Dr. Stefano Piana from D. E. Shaw Research for providing their GDT-TS data. Y.-D.W. and F.J. are supported by the National Natural Science Foundation of China (grant nos. 21133002 and 21203004). Y.-D.W. also thanks the Shenzhen Peacock Program (KQTD201103) and Peking University Shenzhen Graduate School for financial support.

REFERENCES

(1) Laskowski, R. A.; Watson, J. D.; Thornton, J. M. *Nucleic. Acids. Res.* **2005**, *33*, W89−W93.

(2) Zhang, Y. *Curr. Opin. Struct. Biol.* **2009**, *19*, 145−155.

(3) Ekins, S.; Mestres, J.; Testa, B. *Br. J. Pharmacol.* **2007**, *152*, 21−37.

(4) Chapman, H. N.; Fromme, P.; Barty, A.; White, T. A.; Kirian, R. A.; Aquila, A.; Hunter, M. S.; Schulz, J.; DePonte, D. P.; Weierstall, U.; Doak, R. B.; Maia, F. R. N. C.; Martin, A. V.; Schlichting, I.; Lomb, L.; Coppola, N.; Shoeman, R. L.; Epp, S. W.; Hartmann, R.; Rolles, D.; Rudenko, A.; Foucar, L.; Kimmel, N.; Weidenspointner, G.; Holl, P.; Liang, M.; Barthelmess, M.; Caleman, C.; Boutet, S.; Bogan, M. J.; Krzywinski, J.; Bostedt, C.; Bajt, S.; Gumprecht, L.; Rudek, B.; Erk, B.; Schmidt, C.; Homke, A.; Reich, C.; Pietschner, D.; Struder, L.; Hauser, G.; Gorke, H.; Ullrich, J.; Herrmann, S.; Schaller, G.; Schopper, F.; Soltau, H.; Kuhnel, K.-U.; Messerschmidt, M.; Bozek, J. D.; Hau-Riege, S. P.; Frank, M.; Hampton, C. Y.; Sierra, R. G.; Starodub, D.; Williams, G. J.; Hajdu, J.; Timneanu, N.; Seibert, M. M.; Andreasson, J.; Rocker, A.; Jonsson, O.; Svenda, M.; Stern, S.; Nass, K.; Andritschke, R.; Schroter, C.-D.; Krasniqi, F.; Bott, M.; Schmidt, K. E.; Wang, X.; Grotjohann, I.; Holton, J. M.; Barends, T. R. M.; Neutze, R.; Marchesini, S.; Fromme, R.; Schorb, S.; Rupp, D.; Adolph, M.; Gorkhover, T.; Andersson, I.; Hirsemann, H.; Potdevin, G.; Graafsma, H.; Nilsson, B.; Spence, J. C. H. *Nature* **2011**, *470*, 73−77.

(5) Zhao, G.; Perilla, J. R.; Yufenyuy, E. L.; Meng, X.; Chen, B.; Ning, J.; Ahn, J.; Gronenborn, A. M.; Schulten, K.; Aiken, C.; Zhang, P. *Nature* **2013**, *497*, 643−646.

(6) Zhang, X.; Settembre, E.; Xu, C.; Dormitzer, P. R.; Bellamy, R.; Harrison, S. C.; Grigorieff, N. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 1867−1872.

(7) Bax, A.; Grzesiek, S. *Acc. Chem. Res.* **1993**, *26*, 131−138.

(8) Topf, M.; Lasker, K.; Webb, B.; Wolfson, H.; Chiu, W.; Sali, A. *Structure* **2008**, *16*, 295−307.

(9) Roy, A.; Kucukural, A.; Zhang, Y. *Nat. Protocols.* **2010**, *5*, 725−738.

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

(11) Šali, A.; Blundell, T. L. *J. Mol. Biol.* **1993**, *234*, 779−815.

(12) Lance, B. K.; Deane, C. M.; Wood, G. R. *Bioinformatics* **2010**, *26*, 1849−1856.

(13) Moult, J. *Curr. Opin. Struct. Biol.* **2005**, *15*, 285−289.

(14) Berman, H.; Henrick, K.; Nakamura, H.; Markley, J. L. *Nucleic. Acids. Res.* **2007**, *35*, D301−D303.

(15) Dunbrack, R. L., Jr *Curr. Opin. Struct. Biol.* **2006**, *16*, 374−384.

(16) Schwede, T.; Kopp, J.; Guex, N.; Peitsch, M. C. *Nucleic. Acids. Res.* **2003**, *31*, 3381−3385.

(17) Misura, K. M. S.; Chivian, D.; Rohl, C. A.; Kim, D. E.; Baker, D. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 5361−5366.

(18) Huang, Y. J.; Mao, B.; Aramini, J. M.; Montelione, G. T. *Proteins* **2014**, *82*, 43−56.

(19) Mariani, V.; Kiefer, F.; Schmidt, T.; Haas, J.; Schwede, T. *Proteins* **2011**, *79*, 37−58.

DOI: 10.1021/acs.jctc.5b00029

(20) Gront, D.; Kmiecik, S.; Blaszczyk, M.; Ekonomiuk, D.; Koliński, A. *WIREs. Comput. Mol. Sci.* **2012**, *2*, 479−493.

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

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

(23) Summa, C. M.; Levitt, M. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 3177−3182.

(24) Verma, A.; Wenzel, W. *BMC. Struct. Biol.* **2007**, *7*, 12.

(25) Jagielska, A.; Wroblewska, L.; Skolnick, J. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 8268−8273.

(26) Wroblewska, L.; Jagielska, A.; Skolnick, J. *Biophys. J.* **2008**, *94*, 3227−3240.

(27) Zhu, J.; Fan, H.; Periole, X.; Honig, B.; Mark, A. E. *Proteins* **2008**, *72*, 1171−1188.

(28) Zhang, J.; Liang, Y.; Zhang, Y. *Structure* **2011**, *19*, 1784−1795.

(29) Mirjalili, V.; Feig, M. *J. Chem. Theory. Comput.* **2012**, *9*, 1294−1303.

(30) Raval, A.; Piana, S.; Eastwood, M. P.; Dror, R. O.; Shaw, D. E. *Proteins* **2012**, *80*, 2071−2079.

(31) Mirjalili, V.; Noyes, K.; Feig, M. *Proteins* **2014**, *82*, 196−207.

(32) Park, H.; Seok, C. *Proteins* **2012**, *80*, 1974−1986.

(33) Mao, B.; Tejero, R.; Baker, D.; Montelione, G. T. *J. Am. Chem. Soc.* **2014**, *136*, 1893−1906.

(34) Bhattacharya, D.; Cheng, J. *Proteins* **2013**, *81*, 119−131.

(35) Raman, S.; Vernon, R.; Thompson, J.; Tyka, M.; Sadreyev, R.; Pei, J.; Kim, D.; Kellogg, E.; DiMaio, F.; Lange, O.; Kinch, L.; Sheffler, W.; Kim, B.-H.; Das, R.; Grishin, N. V.; Baker, D. *Proteins* **2009**, *77*, 89−99.

(36) Lin, M. S.; Head-Gordon, T. *J. Comput. Chem.* **2011**, *32*, 709−717.

(37) Larsen, A. B.; Wagner, J. R.; Jain, A.; Vaidehi, N. *J. Chem. Inf. Model.* **2014**, *54*, 508−517.

(38) Das, R.; Qian, B.; Raman, S.; Vernon, R.; Thompson, J.; Bradley, P.; Khare, S.; Tyka, M. D.; Bhat, D.; Chivian, D.; Kim, D. E.; Sheffler, W. H.; Malmström, L.; Wollacott, A. M.; Wang, C.; Andre, I.; Baker, D. *Proteins* **2007**, *69*, 118−128.

(39) MacCallum, J. L.; Hua, L.; Schnieders, M. J.; Pande, V. S.; Jacobson, M. P.; Dill, K. A. *Proteins* **2009**, *77*, 66−80.

(40) MacCallum, J. L.; Pérez, A.; Schnieders, M. J.; Hua, L.; Jacobson, M. P.; Dill, K. A. *Proteins* **2011**, *79*, 74−90.

(41) Zemla, A. *Nucleic. Acids. Res.* **2003**, *31*, 3370−3374.

(42) Nugent, T.; Cozzetto, D.; Jones, D. T. *Proteins* **2014**, *82*, 98−111.

(43) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiórkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem. B* **1998**, *102*, 3586−3616.

(44) Piana, S.; Lindorff-Larsen, K.; Shaw, D. E. *Biophys. J.* **2011**, *100*, L47−L49.

(45) Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E. *Science* **2011**, *334*, 517−520.

(46) Jiang, F.; Zhou, C.-Y.; Wu, Y.-D. *J. Phys. Chem. B* **2014**, *118*, 6983−6998.

(47) Jiang, F.; Han, W.; Wu, Y.-D. *J. Phys. Chem. B* **2010**, *114*, 5840−5850.

(48) Jiang, F.; Han, W.; Wu, Y.-D. *Phys. Chem. Chem. Phys.* **2013**, *15*, 3413−3428.

(49) Jiang, F.; Wu, Y.-D. *J. Am. Chem. Soc.* **2014**, *136*, 9536−9539.

(50) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225−11236.

(51) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. *J. Phys. Chem. B* **2001**, *105*, 6474−6487.

(52) Li, H.; Robertson, A. D.; Jensen, J. H. *Proteins* **2005**, *61*, 704−721.

(53) Olsson, M. H. M.; Søndergaard, C. R.; Rostkowski, M.; Jensen, J. H. *J. Chem. Theory. Comput.* **2011**, *7*, 525−537.

(54) Berendsen, H. J. C.; van der Spoel, D.; van Drunen, R. *Comput. Phys. Commun.* **1995**, *91*, 43−56.

(55) Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. *J. Comput. Chem.* **2005**, *26*, 1701−1718.

(56) Horn, H. W.; Swope, W. C.; Pitera, J. W.; Madura, J. D.; Dick, T. J.; Hura, G. L.; Head-Gordon, T. *J. Chem. Phys.* **2004**, *120*, 9665−9678.

(57) Darden, T.; Perera, L.; Li, L.; Pedersen, L. *Structure* **1999**, *7*, R55−R60.

(58) Michaud-Agrawal, N.; Denning, E. J.; Woolf, T. B.; Beckstein, O. *J. Comput. Chem.* **2011**, *32*, 2319−2327.

(59) Kabsch, W.; Sander, C. *Biopolymers* **1983**, *22*, 2577−2637.

(60) DeLano, W. L.; Lam, J. W. *Abstr. Pap. Am. Chem. Soc.* **2005**, *230*, U1371−U1372.

(61) Daura, X.; Gademann, K.; Jaun, B.; Seebach, D.; van Gunsteren, W. F.; Mark, A. E. *Angew. Chem., Int. Ed.* **1999**, *38*, 236−240.

(62) Shesham, R. D.; Bartolotti, L. J.; Li, Y. *Protein. Eng. Des. Sel.* **2008**, *21*, 115−120.