Communication

Folding of Fourteen Small Proteins with a Residue-Specific Force Field and Replica-Exchange Molecular Dynamics

Fan Jiang, and Yun-Dong Wu

J. Am. Chem. Soc., **Just Accepted Manuscript** • Publication Date (Web): 23 Jun 2014

Folding of Fourteen Small Proteins with a Residue-Specific Force Field and Replica-Exchange Molecular Dynamics

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, Peking University, Beijing, 100871, China

Supporting Information Placeholder

ABSTRACT: Ab initio protein folding via physical-based all-atom simulation is still quite challenging. Using a recently developed residue-specific force field (RSFF1) in explicit solvent, we are able to fold a diverse set of 14 model proteins. The obtained structural features of unfolded state are in good agreement with previous observations. The replica-exchange molecular dynamics simulation is found to be efficient, resulting in multiple folding events for each protein. Transition path time is found to be significantly reduced under elevated temperature.

Atomistic simulation of protein folding can provide rich information about structures and mechanisms, and remains an active research area.[1] It places high demands on both the accuracy of force field and the adequacy of conformational sampling.[2] Recently, Lindorff-Larsen et al. successfully folded a set of 12 model proteins using CHARMM22* force field.[3a] Ubiquitin[1e] and a dimeric protein were also successfully folded[3b]. They were able to use special purpose computer ANTON to perform millisecond molecular dynamics (MD) simulations in explicit water. This remarkable achievement enabled further theoretical studies of folding.[4] However, currently such time scale can hardly be reached using commonly accessible computing resources.

Efficiency of MD simulations may be increased by using enhanced conformational sampling methods.[5] One attractive method is replica exchange molecular dynamics (REMD).[6] However, the efficiency of REMD in folding simulation have been questioned, mainly due to the entropic nature of the major folding free energy barrier.[7] Thus, large-scale all-atom folding simulation using normal REMD is still limited.

The development of accurate protein force fields remain highly demanding[8] and challenging.[9] Recently, we developed a residue-specific force field (RSFF1) based on the local conformational preferences of the twenty amino acid residues obtained from coil library of protein crystal structures.[10] We have shown that (1) statistical analysis of the coil library may indeed provide intrinsic conformational features of residues in solution;[11] (2) with a small set of residue-specific torsion and local non-bonded parameters, the coil library Ramachandran plots and side-chain conformational distributions of each residue can be reproduced accurately.[10] Thus, we hope that the RSFF1 may give balanced secondary structure preferences of various sequences, and be able to consistently fold proteins.

Here we report that combining the RSFF1 and REMD, a variety of fast-folding small proteins can be folded into their native structures. The simulations also reveal useful information about the features of folding landscape and indicate that REMD is efficient for folding simulations.

Table 1 summarizes the simulated proteins, simulation conditions and some results. Our systems include the set of 12 fast-folding proteins studied by Lindorff-Larsen et al., along with the original Trp-cage (TC5b)[10,12] and wild-type Engrailed Homeodomain (EnHD) which native structure could not be well stabilized by Charmm22* force field.[3a] The simulations were carried out with the *GROMACS* 4.5.4.[13] Each protein was solvated in a truncated octahedron box (36-49 Å in length) with 1100-2600 TIP4P/Ew water molecules depending on the size of the protein (Table S1 for details). After energy minimization and equilibrium for the box volume, a 600 K NVT MD simulation of 5-20 ns was carried out to obtain the initial structures for REMD simulation, which are well unfolded (Figures S1–S14). For each protein we used 12-36 replicas. Initially, the temperature ranges for some small proteins (CLN025, BBA, Villin, and protein B) were set to about 290-460 K. For other proteins the lowest T was chosen to be about 380 K (except for BBL) after we realized that RSFF1 tends to stabilize the folded state.[10] For all proteins except for α3D, each replica was simulated for less than 2.0 μs (t_{trj}) with a step size of 3 fs.

For each protein, clustering analysis was carried out on the structures sampled in the second half of the trajectory near 300 K or at lowest T. Except for BBL, the most populated cluster (folded structures) is > 50% (F%, Table 1). Since the lowest T in most simulations is higher than corresponding experimental melting temperature (T_m), the RSFF1 indeed consistently over-stabilizes these proteins. On the other hand, CHARMM22* was reported to under-stabilize some of the proteins.[3a]

We define the predicted structure from a simulation as the center structure of the most populated cluster. The superpositions of predicted and experimental structures are shown in Figure 1. The predicted structures for 13 out of 14 proteins have the root mean square deviations (RMSD) of Cα from corresponding experimental structures (R_{pred}) < 4.0 Å, and 7 out of 14 proteins have R_{pred} < 2.0 Å (Table 1). The simulations of 10 proteins sampled the structures with Cα-RMSD < 2.0 Å (Table 1, R_{min}). For BBL, both our and previous simulations[3a] give highest R_{pred}, although we are able to sample structures of R_{min} < 3.0 Å of NMR structure. Experiments suggest that BBL is a downhill (one-state) folder.[14] The R_{pred} of some proteins is considerably reduced if some flexible terminal residues are removed (in parenthesis).

Table 1. Summary of the simulations for the fourteen fast-folding proteins.

#	protein	N^a	expt. T_m (K)	range of T (K)^b	$N_{repl}{}^b$	$t_{trj}{}^b$ (μs)	F%^{c}	$R_{pred}{}^d$ (Å)	$R_{mim}{}^e$ (Å)	$R_{G,F}{}^f$ (Å)	$R_{G,U}{}^f$ (Å)	$\alpha_U\%^g$	$\beta_U\%^g$	$N_F{}^h$	$N_U{}^h$
1	CLN025	10	343	320-450	16	0.4	>99	0.8	0.2	6.0	7.1	1	1	30	15
2	TC10b	20	335	370-451	12	1.2	95	1.6 (1.2)	0.4	7.0	7.9	7	4	57	54
3	TC5b	20	317	350-454	16	1.1	97	1.3	0.4	7.3	7.9	9	5	71	63
4	BBA	28	<298	280-460	36	1.7	64	2.7 (1.9)	1.2	9.4	9.5	26	12	35	22
5	Villin	35	361	290-460	36	2.0	>99	1.1	0.3	9.5	9.6	41	2	48	20
6	WW	35	371	380-491	16	2.0	98	1.5 (1.1)	0.7	9.9	10.0	11	13	7	1
7	NTL9	39	355	380-491	16	2.0	97	0.5	0.3	9.1	9.9	16	33	5	1
8	BBL	47	327	270-437	36	1.5	27	6.2 (5.0)	3.0	10.5	10.3	28	6	9	5
9	protein B	47	372	300-460	36	2.0	91	3.1 (1.3)	1.2	10.0	10.1	57	1	18	12
10	UVF	52	>372	380-488	16	1.8	82	2.3 (2.0)	1.7	10.8	11.2	52	1	60	51
11	EnHD	54	325	330-455	24	1.8	80	3.2 (1.7)	2.1	10.6	10.9	53	2	3	2
12	protein G	56	>323	380-474	16	1.9	77	3.2 (2.9)	2.2	10.9	11.1	25	26	3	1
13	α3D	73	>363	380-484	16	3.3	52	3.8 (3.2)	2.8	12.8	12.2	49	5	5	1
14	λ-repressor	80	347	380-474	16	1.2	83	2.0 (1.3)	1.2	12.0	12.1	54	1	3	0

[a]Number of amino acid residues. [b]Settings of the REMD simulations: temperature range, number and trajectory length of each replica. [c]Percentage of the most populated cluster. [d]Cα-RMSD of the center structure from the most populated cluster, values in parenthesis are without a few terminal residues. [e]Miminal Cα-RMSD seen in the simulation. [f]Average radius of gyration of folded (F) and unfolded (U) structures, respectively. [g]Average percentage of residues forming α-helix (α_U%) and β-sheet (β_U%) in the unfolded state. [h]Number of folding (F) and unfolding (U) events observed from all continuous replica trajectories.

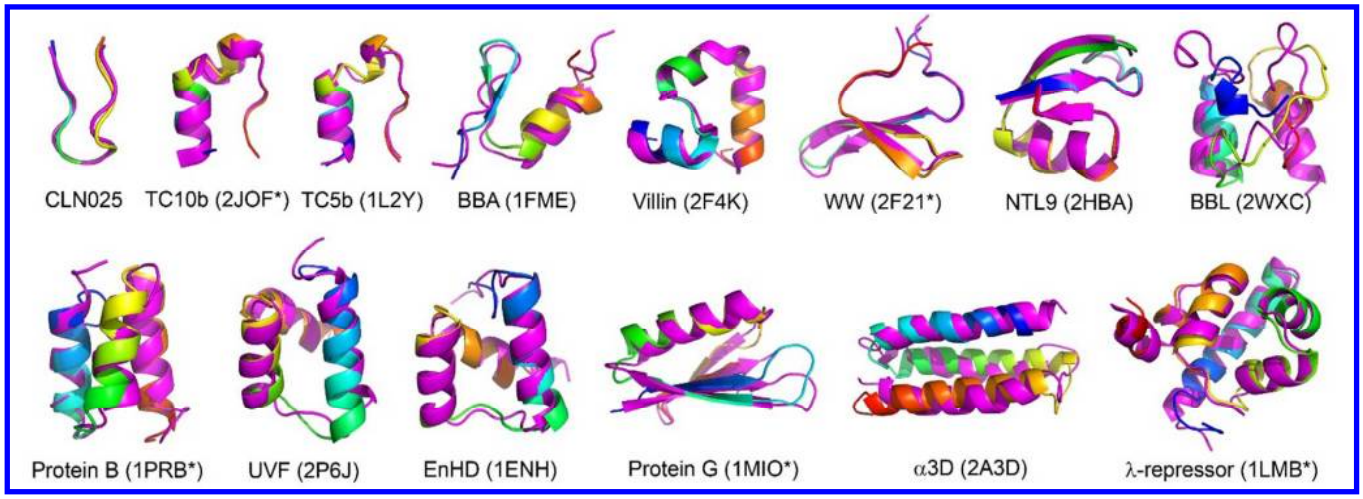


Figure 1. Superposition of the experimental (magenta) and predicted (rainbow) structures of the 14 proteins. PDB ID is given in parentheses. Simulations of 2JOF, 2F21, 1PRB, 1MIO and 1LMB used a slightly different sequences with faster folding.

Figure 2 shows the free energy landscape (FEL) of each protein, which is obtained by projecting onto folding reaction coordinates of Cα-RMSD and the fraction of native contacts[4b] (Q). Except for BBL,[14] each protein has a deep native state basin with relatively small Cα-RMSD and high Q. Among them, CLN025, TC10b, TC5B and WW show a clear two-state FEL. Villin, protein B and α3D are also approximately two-state. On the other hand, BBA shows three major basins, while BBL gives only one large basin. These features are in agreement with the one-dimensional free energy profiles observed in previous folding simulations.[3]We observe more than two major basins for NTL9, protein G, and λ-repressor. Multi-state models with heterogeneous folding pathways for NTL9 and λ-repressor were established in previous simulation studies.[1b,1c]

Both UVF and EnHD belong to the Homeodomain fold, but they show quite different FELs. EnHD is multi-state, with higher barrier between intermediate (I) and native (N) states than that between I and unfolded (U) state. Indeed, a faster conversion between U and I and slower conversion between I and N was observed experimentally.[15] UVF has very low folding free energy barrier. It is one-state in a previous folding simulation.[3a] A recent MD simulation showed that the native structure of UVF is more dynamic than EnHD.[16]

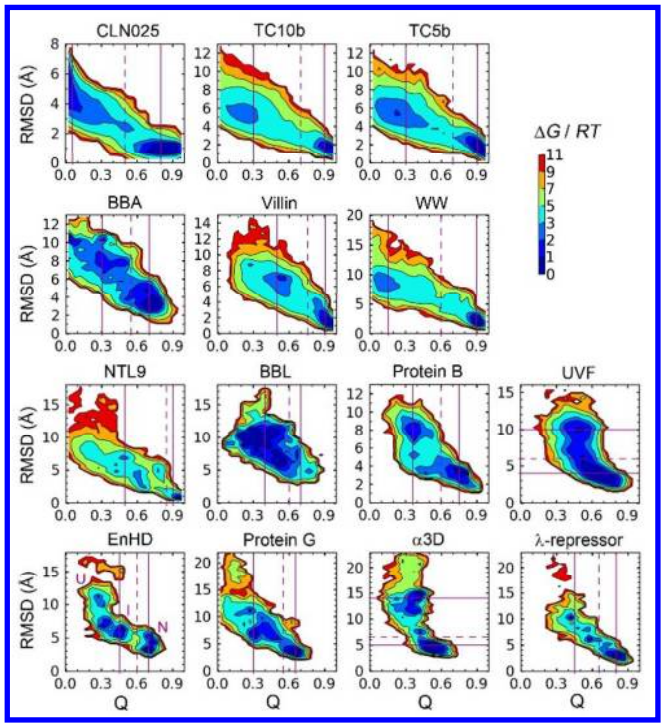


Figure 2. The folding free energy surfaces as the function of the Cα-RMSD to experimental structure and the fraction of native contacts (Q). In each plot, the dashed magenta line divides the folded and unfolded states, and the two solid magenta lines are the boarders of the transition region, upon which folding/unfolding events are defined.

As shown in Table 1, our simulations give highly compact unfolded state for each protein at the lowest simulation temperature, as indicated by quite similar radius of gyration between unfolded ($R_{G,U}$) and folded ($R_{G,F}$) states. There are also significant secondary structure contents in the unfolded state. This result is similar to other simulation results.[1b,1c] Highly compact unfolded states were also observed in previous Charmm22* simulations.[3a,17] A highly collapsed unfolded state was observed by NMR for Trp-cage even in denaturant,[17] with the hydrodynamic radius (R_h) of 7.4 and 8.0 Å for folded and unfolded states, respectively. There was a report of decreased R_G of an unfolded protein from 280 K to 320 K. [19] For BBA, Villin, and BBL, we indeed observed slight compactions (reduced R_G) of the unfolded state with increasing T near 300 K (Figures S20, S21). But at much higher T as for most of our simulations, we observe a considerable decrease of secondary structures and an expansion of the R_G for unfolded state.

It has been found that the native-like structures existing in the unfolded state are closely related to the folding mechanism. Figure 3 gives a residue-by-residue analysis on the formation of native-like local structures in the unfolded state of each protein, using the same method by Lindorff-Larsen et al.[3a] In general, our results are quite similar with the previous results except that our simulations gave somewhat more native-like structures for the second helix region in unfolded protein B and UVF. The regions corresponding to α-helix in the native structures are more native-like than loop regions. We also note that for the five three-helix bundle proteins, the middle helical region is less native-like than the two terminal helical regions. For β-sheet protein WW domain, as observed previously, the N-terminal region around Pro-5 and Pro-6 is most native-like. The same is true for the C-terminal poly-Pro region in the two Trp-cage proteins. The two homeodomain proteins UVF and EnHD differ mainly in the middle helical region, with UVF being more native-like. This might explain why UVF has a faster folding than EnHD.

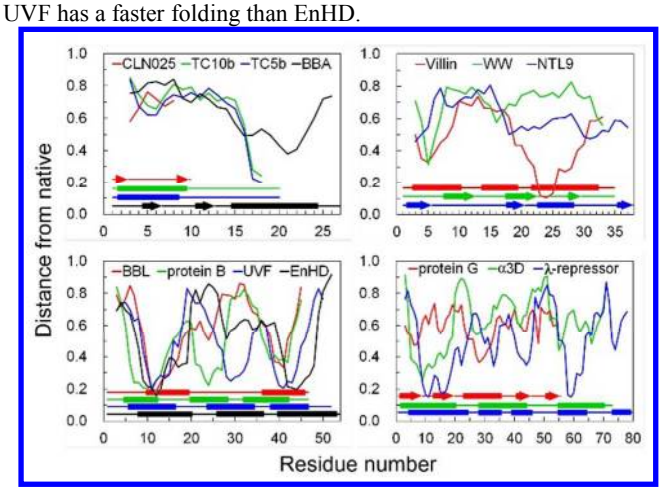


Figure 3. The average distance from the native structure in the unfolded state. Lower value indicates more native-like for a residue. The secondary structures in folded state are shown in the bottom of each plot.

To capture folding events, continuous trajectories (Figure S21) were obtained by tracking every replica exchange, each of which can experience a full range of Ts. Following the method used by Best et al,[4b] a folding/unfolding event is defined as a trajectory cross from the unfolded/folded basin to the folded/unfolded basin, or crossing the two solid magenta lines in Figure 2. As shown in Table 1, multiple folding events (N_F) are observed for many proteins. But several proteins only have small N_F numbers, and much smaller N_U numbers, indicating that simulations have not been long enough to reach convergence for these proteins.

Figure 4 shows transition path time, τ_{TP}, against the average temperature ($\langle T \rangle$) for folding events of four proteins. It is clear that τ_{TP} is considerably reduced as $\langle T \rangle$ is increased. Other proteins have the same feature (Figure S22). A similar trend was also observed in a recent MD simulation of villin.[20]

To gain some insight, we first applied a simple exponential relationship (Arrhenius-like) to fit the data in Figure 4 and Figure S22. Unrealistically high energy barriers of more than 10 kcal/mol are necessary to fit the observed strong T-dependence of τ_{TP}. We then used Zwanzig's super-exponential temperature dependence model for effective diffusion coefficient (D^*) on a rough energy landscape with many random small barriers.[21] Based on the assumption that τ_{TP} is inversely proportional to D^*, we obtain Eq. 1.

$$\tau_{TP}(T) = \tau_0 \exp[(\varepsilon/RT)^2] \qquad [1]$$

We assume a single ε (root-mean-squared energy roughness) for all proteins and that different proteins have different τ_0 (hypothetic transition path time if there is no roughness). We found that a single roughness of 2.5 kcal/mol fits well for most proteins. This roughness is comparable to the barriers of transitions between major backbone and side-chain conformations, and also the interaction energy between side-chains. The ε = 2.5 kcal/mol is somewhat larger than the experimental estimates of ~1 kcal/mol by Wensley et al.,[22a] but is within the range of 1 ~ 4.8 k_BT of the free energy barriers of Trpzip-2 folding obtained from simulations by Gruebele et al.[22b] The ε here measures the internal friction of protein conformational changes. It should be distinguished from the major folding free energy barrier, which is usually smaller for these fast folding proteins.[3a] Recently, experimental measurement

of the τ_{TP} becomes possible.[23] Thus, our theoretical prediction of its strong *T*-dependence can be verified.

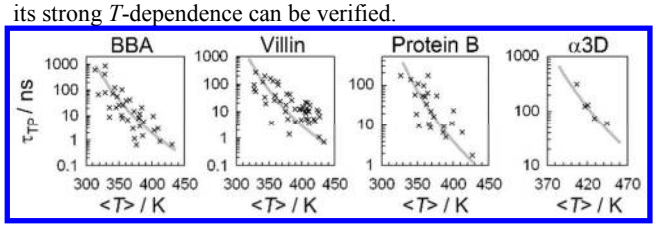


Figure 4. The transition path time (τ_{TP}) against the average temperature (<*T*>) for folding events of selected proteins (See Figure S22 for other proteins). The grey curves are from eq. 1 with ε = 2.5 kcal/mol.

In summary, using REMD simulation and RSFF1, a force field that stabilizes protein native state, it is possible to fold a diverse set of fast-folding proteins using common computers. The elevated temperature in REMD can facilitate crossing entropic barrier by increasing the diffusion on rough energy landscape. We expect that the force field can find many applications including the refinement of protein structures with low resolutions.

ASSOCIATED CONTENT

Supporting Information

Details about the RSFF1, simulation settings and trajectory analysis; Tables S1 and S2, and Figure S1-S22; A few additional simulations. This material is available free of charge via the Internet at http://pubs.acs.org. The simulation trajectories and related data can be provided upon request.

AUTHOR INFORMATION

Corresponding Author

jiangfan@pku.edu.cn; wuyd@pkusz.edu.cn

Notes

The authors declare no competing financial interests.

ACKNOWLEDGMENT

Financial supports from the National Natural Science Foundation of China (21133002, 21203004) and the Shenzhen Peacock Program (KQTD201103) are acknowledged. We thank Profs. Yi-Qin Gao and Xu-Hui Huang for helpful discussion.

REFERENCES

(1) (a) Duan, Y.; Kollman, P. A. *Science* **1998**, *282*, 740. (b) Voelz, V. A.; Bowman, G. R.; Beauchamp, K.; Pande, V. S. *J. Am. Chem. Soc.* **2010**, *132*, 1526. (c) Bowman, G. R.; Voelz, V. A.; Pande, V. S. *J. Am. Chem. Soc.* **2011**, *133*, 664. (d) Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y. B.; Wriggers, W. *Science* **2010**, *330*, 341. (e) Piana, S.; Lindorff-Larsen, K.; Shaw, D. E. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *110*, 5915.

(2) (a) Freddolino, P. L.; Harrison, C. B.; Liu, Y. X.; Schulten, K. *Nat. Phys.* **2010**, *6*, 751. (b) Lane, T. J.; Shukla, D.; Beauchamp, K. A.; Pande, V. S. *Curr. Opin. Struc. Biol.* **2013**, *23*, 58.

(3) (a) Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E. *Science* **2011**, *334*, 517. (b) Piana, S.; Lindorff-Larsen, K.; Shaw, D. E. *J. Phys. Chem. B* **2013**, *117*, 12935.

(4) (a) Dickson, A.; Brooks III, C. L. *J. Am. Chem. Soc.* **2013**, *135*, 4729. (b) Best, R. B.; Hummer, G.; Eaton, W. A. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *110*, 17874. (c) Deng, N.; Dai, W.; Levy, R. M. *J. Phys. Chem. B* **2013**, *117*, 12787.

(5) (a) Laio, A.; Parrinello, M. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 12562. (b) Hamelberg, D.; Mongan, J.; McCammon, J. A. *J. Chem. Phys.* **2004**, *120*, 11919. (c) Lei, H.; Duan, Y. *Curr. Opin. Struc. Biol.* **2007**, *17*, 187.

(6) (a) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314*, 141. (b) Ostermeir, K.; Zacharias, M. *Biochim. Biophys. Acta* **2013**, *1834*, 847.

(7) (a) Zuckerman, D. M.; Lyman, E. *J. Chem. Theory Comput.* **2006**, *2*, 1200. (b) Zheng, W.; Andrec, M.; Gallicchio, E.; Levy, R. M. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 15340. (c) Zhang, W. H.; Chen, J. H. *J. Chem. Theory Comput.* **2013**, *9*, 2849.

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

(9) (a) Li, D. W.; Bruschweiler, R. *Angew. Chem.* **2010**, *122*, 6930. (b) Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E.; Mittal, J.; Feig, M.; Mackerell, A. D. Jr. *J. Chem. Theory Comput*. **2012**, *8*, 3257. (c) Best, R. B.; Hummer, G. *J. Phys. Chem. B* **2009**, *113*, 9004.

(10) Jiang, F.; Zhou, C. Y.; Wu, Y.-D. *J. Phys. Chem. B* ASAP DOI: 10.1021/jp5017449.

(11) Jiang, F.; Han, W.; Wu, Y. D. *Phys. Chem. Chem. Phys.* **2013**, *15*, 3413.

(12) Neidigh, J. W.; Fesinmeyer, R. M.; Andersen, N. H. *Nat. Struct. Mol. Biol.* **2002**, *9*, 425.

(13) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. *J. Chem. Theory Comput.* **2008**, *4*, 435-447.

(14) Sadqi, M.; Fushman, D.; Muñoz, V. *Nature* **2006**, *442*, 317.

(15) Mayor, U.; Guydosh, N. R.; Johnson, C. M.; Grossmann, J. G.; Sato, S.; Jas, G. S.; Freund, S. M.; Alonso, D. O.; Daggett, V.; Fersht, A. R. *Nature* **2003**, *421*, 863.

(16) McCully, M. E.; Beck, D. A.; Daggett, V. *Protein Eng. Des. Sel.* **2013**, *26*, 35.

(17) Piana, S.; Klepeis, J. L.; Shaw, D. E. *Curr. Opin. Struc. Biol.* **2014**, *24*, 98.

(18) Mok, K. H.; Kuhn, L. T.; Goez, M.; Day, I. J.; Lin, J. C.; Andersen, N. H.; Hore, P. J. *Nature* **2007**, *447*, 106.

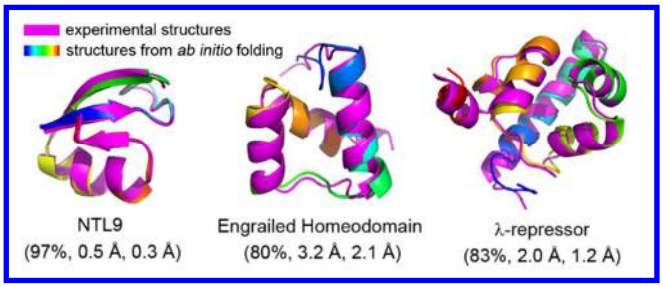
(19) Nettels, D.; Müller-Späth, S.; Küster, F.; Hofmann, H.; Haenni, D.; Rüegger, S.; Reymond, L.; Hoffmann, A.; Kubelka, J.; Heinz, B.; Gast, K.; Best, R. B.; Schuler, B. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 20740.

(20) Piana, S.; Lindorff-Larsen, K.; Shaw, D. E. *Proc. Natl. Acad. Sci. U.S.A* **2012**, *109*, 17845.

(21) Zwanzig, R. *Proc. Natl. Acad. Sci. U.S.A* **1988**, *85*, 2029.

(22) (a) Wensley, B. G.; Batey, S.; Bone, F. A. C.; Chan, Z. M.; Tumelty, N. R.; Steward, A.; Kwa, L. G.; Borgia, A.; Clarke, J. *Nature*, **2010**, *463*, 685. (b) Yang, W. Y.; Pitera, J. W.; Swope, W. C.; Gruebele, M. *J. Mol. Biol*. **2004**, *336*, 241.

(23) Chung, H. S.; McHale, K.; Louis, J. M.; Eaton, W. A. *Science* **2012**, *335*, 981.


experimental structures
structures from *ab initio* folding
NTL9
(97%, 0.5 Å, 0.3 Å)
Engrailed Homeodomain
(80%, 3.2 Å, 2.1 Å)
λ-repressor
(83%, 2.0 Å, 1.2 Å)