How fast can fast-folding proteins autonomously fold in silico?

Yuan-Ping Pang

Computer-Aided Molecular Design Laboratory, Mayo Clinic, Rochester, MN 55905, USA

Correspondence should be addressed to pang@mayo.edu.

Keywords: Protein folding; Protein simulation; Folding kinetics; Folding rate; Folding time; Molecular dynamics simulation; CLN025; Trp-cage.

In microcanonical molecular dynamics simulations, fast-folding proteins CLN025[1] and Trp-cage[2] can autonomously fold to conformations with Cα root mean square deviations (RMSDs) of 1.0–1.4 Å from the experimentally determined native conformations[3]. However, the folding times of CLN025 and Trp-cage predicted from the simulations[3] are more than 4–10 times longer than the experimental values[4,5], indicating an accuracy gap between experiment and simulation for folding speed. Here I report how combining a new protein simulation method[6] and a revised AMBER forcefield[7] results in accurate folding of CLN025 and Trp-cage in 40 distinct, independent, unrestricted, unbiased, and isobaric–isothermal molecular dynamics simulations. According to a survival analysis of these simulations using a Cα-and-Cβ RMSD cutoff of 0.98 Å, the simulated folding times of CLN025 at 293 and 300 K and Trp-cage at 280 and 300 K are 279 ns (95% CI: 204–380 ns), 198 ns (95% CI: 146–270 ns), 2.4 μs (95% CI: 1.8–3.3 μs), and 0.8 μs (95% CI: 0.6–1.0 μs), respectively. The corresponding experimental values are 261 ns, 137 ns, 2.4 μs, and 1.4 μs, respectively[4,5]. These results show that CLN025 and Trp-cage now can autonomously fold in silico as fast as they do in experiments, indicating that the accuracy of folding simulations begins to overlap with the accuracy of folding experiments. This represents a step forward in combining simulation with experiment to develop algorithms that predict structure and dynamics of a globular protein from its sequence for artificial intelligence of biomedical research.

How fast can fast-folding proteins autonomously fold in silico? This question is important because experimental folding times (τs)[4,5,8] are rigorous benchmarks for evaluating the accuracy of protein folding simulations[9-11]. Presumably due to approximations in the empirical potential energy functions with a set of parameters that are used in simulations of protein folding, the simulated τs reported to date have been considerably longer than the experimental τs. For example, early molecular dynamics (MD) simulations of fast-folding proteins using a distributed computing implementation with implicit solvation yielded τs that were reportedly consistent

with the corresponding experimental values[12,13]. Those simulated τs were derived by using cutoffs for Cα RMSD (CαRMSD) of 2.5–3.0 Å[13] and 3.622 Å[12] from the experimentally determined native conformations to identify conformations that constitute the native structural ensembles. If CαRMSD cutoffs of <2.0 Å were used, the τs estimated from those simulations would be considerably longer than the experimental values according to the reported sensitivities of the simulated τs to CαRMSD cutoffs[12,13]. For another example, advanced microcanonical MD simulations predicted τs of fast-folding proteins CLN025[1] and Trp-cage[2] to be 600 ns at 343 K and 14 μs at 335 K, respectively[3]. These τs were derived from the microcanonical simulations with the most populated conformations of CLN025 and Trp-cage that have respective stringent CαRMSDs of 1.0 Å and 1.4 Å from the experimentally determined native conformations[3]. Based on the predicted τs at 335 K and 343 K, the simulated τs of CLN025 and Trp-cage at 300 K are conceivably more than 4–10 times longer than the experimental τs (137 ns for CLN025 and 1.4 μs for Trp-cage) at 300 K because the experimental τs reportedly increase as temperature decreases[4,5]. Therefore, how fast can fast-folding proteins fold in silico equates to how accurate protein folding simulations can be. The reported τs to date suggest that fast-folding proteins cannot autonomously fold in silico as fast as they do in experiments and there is an accuracy gap between simulation and experiment for protein folding speed.

To narrow the gap I developed a new protein simulation method that uses uniformly scaled atomic masses to compress or expand MD simulation time for improving configurational sampling efficiency or temporal resolution, respectively[6,14,15]. As explained in Ref. [6], uniformly reducing all atomic masses of a simulation system by tenfold can compress the simulation time by a factor of $\sqrt{10}$ and hence improve the configurational sampling efficiency of the low-mass

simulations at temperatures of ≤340 K. This method consequently enables miniprotein folding simulations to be performed on personal computers such as Apple Mac Pros under isobaric–isothermal conditions under which most experimental folding studies are performed. Also the kinetics of the low-mass simulation system can be converted to the kinetics of the standard-mass simulation system by simply scaling the low-mass time with a factor of $\sqrt{10}$. Aided by the low-mass simulation method, I subsequently developed a revised AMBER forcefield—an empirical potential energy function with a set of revised parameters—that has shown improvements in (*i*) autonomously folding fast-folding proteins, (*ii*) simulating genuine localized disorders of folded globular proteins, and (*iii*) refining comparative models of monomeric globular proteins[7,16,17]. Hereafter the combination of the revised AMBER forcefield with the low-mass simulation method is termed FF12MC[7].

I also reported the use of the open-source R survival package, which has been widely used in preclinical and clinical studies[18], to predict τs of fast-folding proteins from their sequences[6,7]. In performing zebrafish toxicology experiments, I observed that even though all fish with nearly the same body weights received an intraperitoneal injection of the same dose of the same toxin, the times-to-death of 20 toxin-treated fish varied widely in each experiment. In some experiments a few fish did not even succumb to the toxin. Yet, the mean times-to-death and their 95% confidence intervals (95%CIs) obtained from the R survival package varied slightly among different experiments. The resemblance of the live and dead states of the zebrafish to the unfolded and folded states of a protein inspired me to use the R survival package to predict τs from sequences as follows[7]: Perform 1) 20–40 distinct and independent MD simulations of a fast-folding protein sequence using FF12MC to obtain 20–40 sets of instantaneous conformations in time, wherein the sequence adopts a fully extended backbone conformation, 2) a cluster analysis of all instantaneous conformations from the 20–40 sets to obtain the average conformation of the largest conformation cluster as the native conformation of the protein, and 3) a survival

analysis using the 20–40 sets of instantaneous conformations in time and the average conformation to determine τ and its 95%CI. One advantage of this method is the rigorous estimation of the mean and 95%CI of τ from a set of simulations—a few of which did not capture a folding event. Another advantage is that the τ prediction does not require the assumption that the protein must follow a two-state folding mechanism. By examining a plot of the natural logarithm of the nonnative state population of the protein versus time-to-folding, one can determine the hazard function for the nonnative state population of the protein. If the plot reveals an exponential decay of the nonnative state population over simulation time, then the protein follows a two-state folding mechanism in the simulations.

To determine how fast CLN025 autonomously folds in silico using the methods outlined above, 40 distinct, independent, unrestricted, unbiased, isobaric–isothermal, and 3.16-μs (of the standard-mass time) classical MD simulations of CLN025 were performed at 300 K using FF12MC. A fully extended backbone conformation of CLN025 was used as the initial conformation of the 40 simulations. All simulations described hereafter are 40 distinct, independent, unrestricted, unbiased, isobaric–isothermal, and classical MD simulations with FF12MC using a fully extended backbone conformation as the initial conformation of the protein for the 40 simulations. Also all simulation times described hereafter are the standard-mass simulation times. A cluster analysis of the simulations revealed that the average conformation in the largest cluster adopted a β-hairpin conformation (Fig. 1*B*). This average conformation had CαRMSD of 0.87 Å and Cα and Cβ RMSD (CαβRMSD) of 0.94 Å relative to the average conformation (Fig. 1*A*) of 20 NMR-determined conformations of CLN025 (PDB ID: 2RVD)[1]. It is worth noting that CαβRMSD includes main-chain and side-chain structural information and is hence more stringent to measure structural similarity than CαRMSD. Using the average conformation of the largest cluster as the predicted native conformation of CLN025,

the first time-instant at which CαβRMSD of the full-length CLN025 sequence reached ≤0.98 Å was recorded as an individual folding time for each of the 40 simulations (Table S1A). Using the 40 individual folding times as times-to-folding, a survival analysis predicted the τ of CLN025 to be 198 ns (95%CI = 146–270 ns; n = 40) at 300 K (Table 1). Plotting the natural logarithm of the nonnative state population of CLN025 versus time-to-folding revealed a linear relationship with r^2 of 0.97 (Fig. 2), which indicates that CLN025 follows the two-state folding mechanism. These results agree with the experimental studies showing that the folding of CLN025 follows a two-state folding mechanism with a τ of 137 ns at 300 K, which was obtained from Fig. 6 of Ref. [4]. To substantiate the agreement between the experimental and computational τs of CLN025 at 300 K, the 40 CLN025 simulations were repeated at 293 K. Using the same CαβRMSD cutoff and the same predicted native conformation, a survival analysis showed that CLN025 followed the two-state folding mechanism (r^2 = 0.94; Fig. 2) with a τ of 279 ns (95%CI = 204–380 ns; n = 40) at 293 K (Table 1), showing again an agreement with the experimental τ of 261 ns at 293 K that was also obtained from Fig. 6 of Ref. [4].

To determine how fast Trp-cage autonomously folds in silico, 40 9.48-μs MD simulations of the Trp-cage TC10b sequence were performed at 280 K. The average conformation of the largest cluster of the simulations (Fig. 1*D*) had CαRMSD of 1.69 Å and CαβRMSD of 1.86 Å from the average conformation (Fig. 1*C*) of 28 NMR-determined conformations (PDB ID: 2JOF)[2]. Using the CαβRMSD cutoff of 0.98 Å and the average conformation as the predicted native conformation, a survival analysis showed that TC10b followed the two-state folding mechanism (r^2 = 0.94; Fig. 2) in the simulations with a τ of 2.4 μs (95% CI = 1.8–3.3 μs; n = 40) at 280 K (Table 1). This is consistent with the τ of 2.4 μs at 280 K that was obtained from Fig. 4 of NMR ln Kf in Ref. [5]. Repeating the 40 simulations of TC10b at 300 K, using the same simulation conditions and the same criteria to define the native structural ensemble, revealed

the two-state folding mechanism (r^2 = 0.96; Fig. 2) and led to a shortened τ of 0.8 μs (95% CI = 0.6–1.0 μs; n = 40; Table 1), which is also consistent with the experimental τ of 1.4 μs at 300 K[5].

In the above studies, the average conformation of the largest cluster was used as the predicted native conformation, and a full-length CαβRMSD cutoff of 0.98 Å from the average conformation was used to identify conformations that constitute the native structural ensemble. The use of the average rather than the representative conformation of the largest cluster may unrealistically shorten the simulated τ, whereas the use of an "overly" stringent CαβRMSD cutoff of 0.98 Å may unrealistically lengthen the simulated τ. To address these concerns, all τs in Table 1 were re-estimated from the same simulation data using both the average and representative conformations with RMSD cutoffs varying from 0.98 Å to 1.40 Å. As apparent from Table S2, the τs of CLN025 and Trp-cage are insensitive to the change from the average to the representative conformation, and these τs are also insensitive to the variation of the CαβRMSD cutoff within 0.98–1.40 Å. Therefore, the concerns about the definition of the native structural ensemble are unwarranted.

In this context, it is evident from the present data that an agreement between the experimental and computational τs was achieved within a factor of 0.69–1.75 (Table 1) when folding simulations were performed using FF12MC. While additional folding simulations and 95%CIs for experimental τs are needed to conclusively answer the question of how fast can fast-folding proteins fold in silico, the present data show that CLN025 and Trp-cage now can autonomously fold in simulations as fast as they do in experiments, indicating that the accuracy of folding simulations begins to overlap with the accuracy of folding experiments. This finding represents an important step forward in combining computation with experiment to develop algorithms that predict structure and dynamics of a globular protein from its sequence for artificial intelligence of biomedical research.

METHODS

MD simulations. A fast-folding protein in a fully extended backbone conformation was solvated with the TIP3P water[19] with surrounding counter ions and/or NaCls and then energy-minimized for 100 cycles of steepest-descent minimization followed by 900 cycles of conjugate-gradient minimization to remove close van der Waals contacts using SANDER of AMBER 11 (University of California, San Francisco). The resulting system was heated from 0 to a temperature of 280–300 K at a rate of 10 K/ps under constant temperature and constant volume, then equilibrated for 10^6 timesteps under constant temperature and constant pressure of 1 atm employing isotropic molecule-based scaling, and finally simulated in 40 distinct, independent, unrestricted, unbiased, isobaric–isothermal, and classical MD simulations using PMEMD of AMBER 11 with a periodic boundary condition at 280–300 K and 1 atm. The fully extended backbone conformations (*viz.*, anti-parallel β-strand conformations) were generated by MacPyMOL Version 1.5.0 (Schrödinger LLC, Portland, OR). The numbers of TIP3P waters and surrounding ions, initial solvation box size, and ionizable residues are provided in Table S3. The 40 unique seed numbers for initial velocities of Simulations 1–40 are listed in Table S4. All simulations used (*i*) a dielectric constant of 1.0, (*ii*) the Berendsen coupling algorithm[20], (*iii*) the Particle Mesh Ewald method to calculate electrostatic interactions of two atoms at a separation of >8 Å[21], (*iv*) Δt = 1.00 fs of the standard-mass time[7], (*v*) the SHAKE-bond-length constraints applied to all bonds involving hydrogen, (*vi*) a protocol to save the image closest to the middle of the "primary box" to the restart and trajectory files, (*vii*) a formatted restart file, (*viii*) the revised alkali and halide ions parameters[22], (*ix*) a cutoff of 8.0 Å for nonbonded interactions, (*x*) the atomic masses of the entire simulation system (both solute and solvent) were reduced uniformly by tenfold, and (*xi*) default values of all other inputs of the PMEMD module. The forcefield parameters of FF12MC are available in the Supporting Information of Ref. [6]. All simulations were performed on an in-house cluster of 100 12-core Apple Mac Pros with Intel Westmere (2.40/2.93 GHz).

Folding time estimation. The τ of a miniprotein was estimated from the mean time-to-folding in 40 distinct, independent, unrestricted, unbiased, isobaric–isothermal, and classical MD simulations using survival analysis methods[18] implemented in the R survival package Version 2.38-3 (http://cran.r-project.org/package=survival). A C$\alpha\beta$RMSD cutoff of 0.98 Å was used to identify conformations that constitute the native structural ensemble. For each simulation with conformations saved at every 10^5 timesteps, the first time-instant at which C$\alpha\beta$RMSD reached $\leq$0.98 Å was recorded as an individual folding time (Table S1). Using the Kaplan-Meier estimator[23,24] [the Surv() function in the R survival package], the mean time-to-folding was first calculated from a first set of simulations each of which captured a folding event. If a parametric survival function mostly fell within the 95% confidence interval (95% CI) of the Kaplan-Meier estimation for the first set of simulations, the parametric survival function [the Surreg() function in the R survival package] was then used to calculate the mean time-to-folding of the first set of simulations and the mean time-to-folding of a second set of simulations that were identical to the first set except that the simulation temperature of the second set was changed.

Cluster analysis and data processing. The conformational cluster analyses of CLN025 and TC10b were performed using CPPTRAJ of AmberTools 16 with the average-linkage algorithm[25], epsilon of 2.0 Å, and root mean square coordinate deviation on all Cα and Cβ atoms (see Table S5). No energy minimization was performed on the average conformation of any cluster. The linear regression analysis was performed using the PRISM 5 program.

References

1. Honda, S. *et al.* Crystal structure of a ten-amino acid protein. *J. Am. Chem. Soc.* **130**, 15327–15331 (2008).
2. Barua, B. *et al.* The Trp-cage: Optimizing the stability of a globular miniprotein. *Protein Eng. Des. Sel.* **21**, 171–185 (2008).

3. Lindorff-Larsen, K., Piana, S., Dror, R. O. & Shaw, D. E. How fast-folding proteins fold. *Science* **334**, 517–520 (2011).

4. Davis, C. M., Xiao, S. F., Raeigh, D. P. & Dyer, R. B. Raising the speed limit for β-hairpin formation. *J. Am. Chem. Soc.* **134**, 14476–14482 (2012).

5. Byrne, A. *et al.* Folding dynamics and pathways of the trp-cage miniproteins. *Biochemistry* **53**, 6011–6021 (2014).

6. Pang, Y.-P. Low-mass molecular dynamics simulation for configurational sampling enhancement: More evidence and theoretical explanation. *Biochem. Biophys. Rep.* **4**, 126–133 (2015).

7. Pang, Y.-P. FF12MC: A revised AMBER forcefield and new protein simulation protocol. *Proteins* 84, 1490–1516 (2016).

8. Kubelka, J., Hofrichter, J. & Eaton, W. A. The protein folding 'speed limit'. *Curr Opin Struct Biol* **14**, 76–88 (2004).

9. Gillespie, B. & Plaxco, K. W. Using protein folding rates to test protein folding theories. *Annu. Rev. Biochem.* **73**, 837–859 (2004).

10. Snow, C. D., Sorin, E. J., Rhee, Y. M. & Pande, V. S. How well can simulation predict protein folding kinetics and thermodynamics? *Annu. Rev. Biophys. Biomol. Struct.* **34**, 43–69 (2005).

11. Gelman, H. & Gruebele, M. Fast protein folding kinetics. *Q Rev Biophys* **47**, 95–142 (2014).

12. Snow, C. D., Nguyen, N., Pande, V. S. & Gruebele, M. Absolute comparison of simulated and experimental protein-folding dynamics. *Nature* **420**, 102–106 (2002).

13. Snow, C. D., Zagrovic, B. & Pande, V. S. The Trp cage: folding kinetics and unfolded state topology via molecular dynamics simulations. *J. Am. Chem. Soc.* **124**, 14548–14549 (2002).

14. Pang, Y.-P. Low-mass molecular dynamics simulation: A simple and generic technique to enhance configurational sampling. *Biochem. Biophys. Res. Commun.* **452**, 588–592 (2014).

15. Pang, Y.-P. Use of multiple picosecond high-mass molecular dynamics simulations to predict crystallographic B-factors of folded globular proteins. *Heliyon* **2**, e00161 (2016).

16. Pang, Y.-P. At least 10% shorter C–H bonds in cryogenic protein crystal structures than in current AMBER forcefields. *Biochem. Biophys. Res. Commun.* **458**, 352–355 (2015).

17. Pang, Y.-P. Use of 1–4 interaction scaling factors to control the conformational equilibrium between α-helix and β-strand. *Biochem. Biophys. Res. Commun.* **457**, 183–186 (2015).

18. Therneau, T. M. & Grambsch, P. M. *Modeling Survival Data: Extending the Cox Model*. 350 (Springer-Verlag, 2000).

19. Jorgensen, W. L., Chandreskhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926–935 (1983).

20. Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., Di Nola, A. & Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **81**, 3684–3690 (1984).

21. Darden, T. A., York, D. M. & Pedersen, L. G. Particle mesh Ewald: An N log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **98**, 10089–10092 (1993).

22. Joung, I. S. & Cheatham, T. E. Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *J. Phys. Chem. B* **112**, 9020–9041 (2008).

23. Kaplan, E. L. & Meier, P. Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.* **53**, 457–481 (1958).

24. Rich, J. T. *et al.* A practical guide to understanding Kaplan-Meier curves. *Otolaryngol. Head Neck Surg.* **143**, 331–336 (2010).

25. Shao, J., Tanner, S. W., Thompson, N. & Cheatham III, T. E. Clustering molecular dynamics trajectories: 1. Characterizing the performance of different clustering algorithms. *J. Chem. Theory Comput.* **3**, 2312–2334 (2007).

Acknowledgments The author acknowledges the support of this work from the US Defense Advanced Research Projects Agency (DAAD19-01-1-0322), the US Army Medical Research Material Command (W81XWH-04-2-0001), the US Army Research Office (DAAD19-03-1-0318, W911NF-09-1-0095, and W911NF-16-1-0264), the US Department of Defense High Performance Computing Modernization Office, and the Mayo Foundation for Medical Education and Research. The contents of this article are the sole responsibility of the author and do not necessarily represent the official views of the funders.

Author Contributions Y.-P.P. designed and performed all computational studies, analyzed and interpreted the simulation results, and wrote the manuscript.

Author Information Y.-P.P. declares no competing financial interests.

Supplementary Information: Tables S1–S5

Table 1. Experimental and computational folding times of CLN025 and Trp-cage (TC10b)

	Folding time (μs)			
Fast-folding protein	Experimental	Computational		E/C
		Mean	95%CI	
CLN025 at 293 K	0.261	0.279	0.204–0.380	0.94
CLN025 at 300 K	0.137	0.198	0.146–0.270	0.69
TC10b at 280 K	2.4	2.4	1.8–3.3	1.00
TC10b at 300 K	1.4	0.8	0.6–1.1	1.75

The experimental folding times of CLN025 and TC10b were obtained from the Arrhenius plots of Refs. [4] and [5]. Each computational folding time was predicted from 40 distinct, independent, unrestricted, unbiased, isobaric–isothermal, and 3.16-μs (for CLN025) or 9.48-μs (for TC10b) molecular dynamics simulations with FF12MC using a parametric survival function and a Cα-and-Cβ root mean square deviation of 0.98 Å from the average conformation of the largest conformation cluster of the simulations to identify conformations that constitute the native structural ensemble. E/C: Experimental folding time divided by computational folding time. 95%CI: 95% confidence interval.

Fig. 1. Native conformations of CLN025 and Trp-cage (TC10b) derived from experiments and simulations. (*A*) The average of 20 CLN025 NMR structures. (*B*) The average CLN025 conformation of the largest cluster in the simulations using FF12MC. (*C*) The average of 28 Trp-cage NMR structures. (*D*) The average Trp-cage conformation of the largest cluster in the simulations using FF12MC.

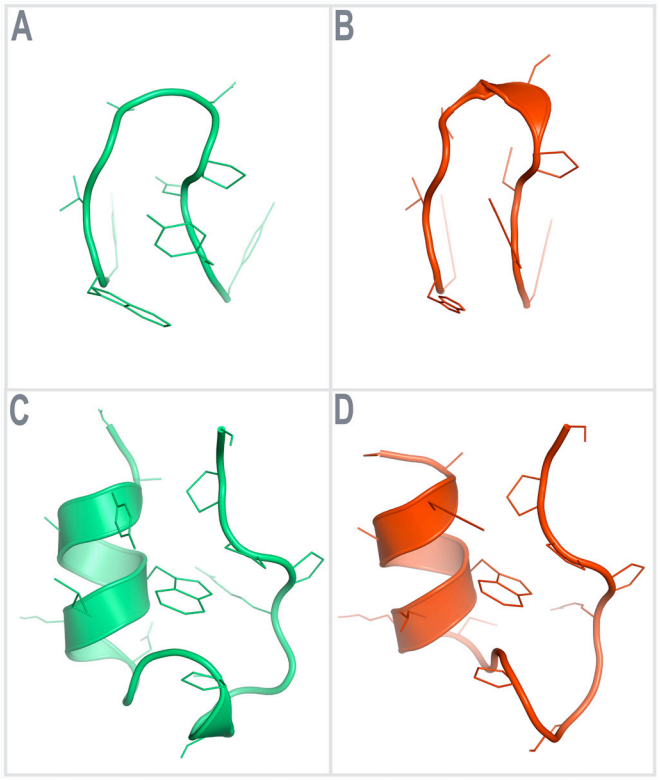

Fig. 2. Plots of the natural logarithm of the nonnative state population of CLN025 and Trp-cage (TC10b) over time-to-folding. The individual folding times were taken from Table S1A.