Biochemical and Biophysical Research Communications 492 (2017) 135–139

Contents lists available at ScienceDirect

Biochemical and Biophysical Research Communications

journal homepage: www.elsevier.com/locate/ybbrc

How fast fast-folding proteins fold in silico

Yuan-Ping Pang*

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

ARTICLE INFO

Article history:
Received 25 July 2017
Accepted 2 August 2017
Available online 9 August 2017

Keywords:
Folding kinetics
Folding rate
Folding time
Protein folding
Molecular dynamics
Survival analysis

ABSTRACT

In reported microcanonical molecular dynamics simulations, fast-folding proteins CLN025 and Trp-cage autonomously folded to experimentally determined native conformations. However, the folding times of these proteins derived from the simulations were more than 4–10 times longer than their experimental values. This article reports autonomous folding of CLN025 and Trp-cage in isobaric–isothermal molecular dynamics simulations with agreements within factors of 0.69–1.75 between simulated and experimental folding times at different temperatures. These results show that CLN025 and Trp-cage can now autonomously fold in silico as fast as in experiments, and suggest that the accuracy of folding simulations for fast-folding proteins begins to overlap with the accuracy of folding experiments. This opens new prospects of developing computer algorithms that can predict both ensembles of conformations and their interconversion rates for a protein from its sequence for artificial intelligence on how and when a protein acts as a receiver, switch, and relay to facilitate various subcellular-to-tissue communications. Then the genetic information that encodes proteins can be better read in the context of intricate biological functions.

1. Introduction

How fast can fast-folding proteins autonomously fold in silico? This question is important because experimental folding times (τs) [1–3] are rigorous benchmarks for evaluating the accuracy of protein folding simulations. If accurate, such simulations offer not only insight into protein folding pathways and mechanisms [4–7] but also a means to determine ensembles of conformations and their interconversion rates for a protein, which are responsible for "proteins to act as receivers, switches, and relays and facilitate communication from the subcellular level through to the cell and tissue levels" [8]. Due to approximations in the empirical potential energy functions for the folding simulations, most simulated τs reported to date have been much longer than the corresponding 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 consistent with the corresponding experimental values if Cα root mean square deviation (CαRMSD) cutoffs of 2.5–3.0 Å or 3.622 Å (in combination with a set of secondary structure criteria) were used to identify conformations that constitute the native structural ensembles [9,10]. However, according to the reported sensitivities of the simulated τs to CαRMSD cutoffs [9,10], the τs would be considerably longer than the experimental values, if typical CαRMSD cutoffs of <2.0 Å were used. For another example, advanced microcanonical MD simulations predicted τs of fast-folding proteins CLN025 [11] and Trp-cage [12] to be 600 ns at 343 K and 14 μs at 335 K, respectively [13]. These τs are of high quality as the τs were derived from the microcanonical MD simulations that resulted in the most populated conformations of CLN025 and Trp-cage with CαRMSDs of 1.0 and 1.4 Å from the experimental native conformations, respectively [13]. However, because the experimental τs of the two proteins reportedly increase as temperature decreases [1,2], the simulated τs at 300 K are conceivably more than 4–10 times longer than the experimental τs. Therefore, how fast fast-folding proteins fold in silico equates to how accurate protein folding simulations can be. Most reported τs to date suggest that fast-folding proteins cannot autonomously fold in silico as fast as in experiments. This implies an accuracy gap between simulation and experiment for protein folding rate (1/τ) that is determined by folding mechanism or pathways [14].

To narrow the accuracy gap, a new protein simulation method was developed. This method uses uniformly scaled atomic masses to compress or expand MD simulation time for improving

Abbreviations: τ, folding time; MD, molecular dynamics; CαRMSD, Cα root mean square deviation; 95%CI, 95% confidence interval; CαβRMSD, Cα and Cβ root mean square deviation.

* Stabile 12-26, Mayo Clinic, 200 First Street SW, Rochester, MN 55905, USA.
E-mail address: pang@mayo.edu.

http://dx.doi.org/10.1016/j.bbrc.2017.08.010

configurational sampling efficiency or temporal resolution [15–17]. 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 [16]. As detailed in Refs. [15,16], this method facilitates protein folding simulations on personal computers (such as Apple Mac Pros) under isobaric–isothermal conditions at which most experimental folding studies are performed. As explained in Ref. [16], 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 $\sqrt{10}$. Subsequently, this low-mass simulation method led to the development of a revised AMBER forcefield 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 [18–20]. Hereafter the combination of the revised AMBER forcefield with the low-mass simulation method is termed FF12MC [18].

Further, in performing zebrafish toxicology experiments for a different project, this author observed that the times-to-death of the 20 toxin-treated fish varied widely in each experiment, although all 20 fish with nearly the same body weights received an intraperitoneal injection of the same dose of the same batch of botulinum neurotoxin serotype A. Yet, the mean time-to-death and its 95% confidence interval (95%CI) calculated using the open-source R survival package [21] varied slightly from one experiment to another. The resemblance of the live and dead states of the zebrafish to the unfolded and folded states of a protein inspired the use of the R survival package to predict τ of a fast-folding protein from its sequence as follows [16,18]: Perform (*i*) ≥20 distinct and independent MD simulations to autonomously fold a fast-folding protein sequence using FF12MC, which results in ≥20 sets of instantaneous protein conformations in time, (*ii*) a cluster analysis of all instantaneous conformations from the ≥20 sets to obtain the average conformation of the largest cluster and use the average conformation as the predicted native conformation of the protein, and (*iii*) a survival analysis using the ≥20 sets of the instantaneous conformations in time and the predicted native conformation to determine the mean τ and its 95%CI. As exemplified in Refs. [16,18], one advantage of this survival analysis method is that the τ prediction does assume that the fast-folding protein must follow a two-state folding mechanism; another advantage is rigorous estimation of mean τ and 95%CI from ≥20 simulations that are relatively short so that a few of these simulations may not capture a folding event.

As demonstrated below, use of the methods and forcefield outlined above resulted in accurate prediction of τs for CLN025 and Trp-cage (TC10b) and an answer to the important question of how fast fast-folding proteins fold in silico. A total of 160 distinct, independent, unrestricted, unbiased, isobaric–isothermal, microsecond MD simulations with a total aggregated simulation time of 1011.2 μs were used for the prediction. All simulation times described hereafter have been converted to standard-mass simulation times.

2. Methods

2.1. Molecular dynamics simulations

A fast-folding protein in a fully extended backbone conformation was solvated with the TIP3P water [22] 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 5 K 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, and isobaric–isothermal 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 S1. The 40 unique seed numbers for initial velocities of Simulations 1–40 are listed in Table S2. All simulations used (*i*) a dielectric constant of 1.0, (*ii*) the Berendsen coupling algorithm [23], (*iii*) the Particle Mesh Ewald method to calculate electrostatic interactions of two atoms at a separation of >8 Å [24], (*iv*) Δt = 1.00 fs of the standard-mass time [18], (*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 [25], (*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. [16]. All simulations were performed on an in-house cluster of 100 12-core Apple Mac Pros with Intel Westmere (2.40/2.93 GHz).

2.2. Folding time estimation

The τ of a fast-folding protein was estimated from the mean time-to-folding in 40 distinct, independent, unrestricted, unbiased, and isobaric–isothermal MD simulations using survival analysis methods [21] implemented in the R survival package Version 2.38–3 (http://cran.r-project.org/package=survival). A Cα and Cβ root mean square deviation (CαβRMSD) 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 CαβRMSD reached ≤0.98 Å was recorded as an individual folding time (Table S3). Using the Kaplan-Meier estimator [26,27] [the Surv() function in the R survival package], the mean time-to-folding was first calculated from 40 simulations each of which captured a folding event at a low temperature of 280 K or 293 K. If a parametric survival function mostly fell within the 95%CI of the Kaplan-Meier estimation for these low-temperature simulations, the parametric survival function [the Surreg() function in the R survival package] was then used to calculate (*i*) the mean time-to-folding of the 40 low-temperature simulations and (*ii*) the mean time-to-folding of 40 new simulations, which were identical to the low-temperature simulations except that the temperature was increased to 300 K.

2.3. Cluster analysis and data processing

The conformational cluster analyses of CLN025 and TC10b were performed using CPPTRAJ of AmberTools 16 (University of California, San Francisco) with the average-linkage algorithm [28], epsilon of 2.0 Å, and root mean square coordinate deviation on all Cα and Cβ atoms (see Table S4). No energy minimization was performed on the average conformation of any cluster. The linear regression analysis was performed using the PRISM 5 program for Mac OS X, Version 5.0d (GraphPad Software, La Jolla, California).