Lattice simulations of co-translational protein folding

Yuanchi Zhao*
Department of Chemistry & Chemical Biology, Harvard University, Cambridge, MA

(Dated: May 16, 2016)

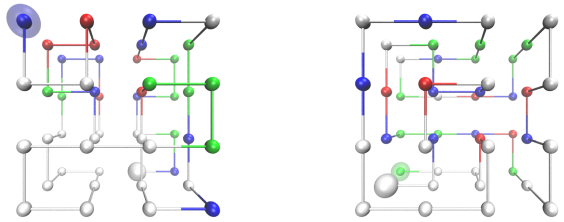
Understanding *in vivo* protein folding mechanisms requires understanding the milieu of cellular processes governing protein biogenesis, one of which is co-translational folding. Co-translational folding occurs when a protein folds as it is being synthesized. We study co-translational folding using lattice protein models, focusing on two previously characterized lattice proteins that differ in their contact topologies. Equilibrium simulations of the full length proteins at various temperatures reproduces a known temperature dependence on folding kinetics. The energy spectra of the two proteins as a function of protein chain length are characterized and found not to favor folding to the native state without nearly the entire chain present. Non-equilibrium co-translational folding simulations are run on the two proteins at two different temperatures. Remarkably, it is observed that more simulations at the lower temperature lead to the folded state than simulations at the higher temperature, despite the higher temperature being the optimal temperature for fast-folding of the full length proteins. It is likely that near-native conformations at intermediate chain lengths are not stable at the higher temperature.

I. INTRODUCTION

Protein biogenesis occurs by stepwise synthesis of a polypeptide chain within the ribosome. The peptide polymer emerges from the ribosome and adopts a functional three-dimensional structure. Understanding how proteins adopt their bioactive conformation is known as the protein folding problem. While *in vitro* experiments and computational investigations of isolated proteins have unveiled much about the mechanism of folding, *in vivo* protein folding is complicated by a crowded hetereogenous environment that includes chaperones and other protein quality control agents, protein trafficking machinery, and post-translational processors [1].

One process that can occur for some proteins during their synthesis *in vivo* is co-translational folding, the adoption of the native fold at some point prior to the complete synthesis of the protein by the ribosome [2–4]. A computational and bioinformatic study that considered translation and folding kinetics estimated that one-third of *Escherichia coli* cytosolic proteins fold co-translationally [5].

With polypeptide elongation rates of 15 amino acids/s in prokaryotes and 5 amino acid/s in eukaryotes [1], and folding timescales from microseconds (for small proteins) to milliseconds, co-translational folding may just be a simple consequence of the kinetics of protein synthesis. On the other hand, evidence for selective pressure for particular synonymous codons at certain positions in genes (rarer codons having slower translation rates) suggests that protein synthesis and folding is a coordinated process [2, 6]. There is evidence for slower translation rates improving protein folding efficiency of certain proteins [7] as well as for faster translation rates improving protein folding efficiency of other proteins through avoidance of misfolding [8, 9]. Thus a key part of understanding *in vivo* protein folding is characterizing co-translational folding: identifying signature features of proteins that fold by this mechanism such as conservation of specific synonymous codons, the dependence of cotranslational folding on protein topology, or the impact of translation rate on folding efficiency.



(a) P1 (b) P2

FIG. 1. The native folds of the two lattice proteins under study in this work. P1 (a) has native contacts involving monomers that are close together in the 1D sequence. P2 (b) has native contacts involving monomers that are distant in the 1D sequence. N-termini are indicated by the larger, solid spheres, and C-termini are indicated by the larger, translucent spheres. Residues are colored by their type: nonpolar (white), basic (blue), acidic (red), and polar (green).

Monte Carlo (MC) simulations of lattice protein models can provide insight into the protein folding problem. Although a lattice polymer is a simplified model of a protein, MC lattice simulations have low computational cost, enabling generation of accurate statistical samples while capturing aspects of the protein folding problem such as the folding of a polymer into a single native state out of

* yzhao01@g.harvard.edu

numerous possible compact states [10].

In a recent work, Krobath et al. used such models to investigate the effect of protein native topology and non-native interactions on co-translational folding. Two previously studied 48-mer cubic lattice proteins [11], one with a local native contact order (CO) (P1 in Fig. 1, CO = 0.23) and one with a non-local native contact topology (P2 in Fig. 1, CO = 0.45) were used to model ribosome-bound stalled nascent chains (RNCs) [12]. Varying lengths of the full lattice protein were simulated with one end tethered to a point one lattice spacing away from an inert flat surface, which represents a ribosome. The equilibrium properties of the ribosome-attached chains as a function of their length were studied. The key finding was that for both topologies, the native structure does not form until almost the entire protein chain is present (at around 44/48 residues), although P1 does form slightly more native contacts at shorter chain lengths than does P2 due to its shorter range native contacts [12].

The study by Krobath et al. suggests that these two particular lattice proteins do not exhibit co-translational folding, but the study does present arenas for further investigation. The simulations performed by Krobath et al. were equilibrium MC simulations that were initiated with a random conformation for each subsequence of the entire protein chain. Given sufficient sampling, the collected statistics are accurate for each chain length, but the kinetic properties of co-translational folding (in so far as MC steps approximate dynamics) such as the history dependence of folding on the current translation step are not captured. Since protein translation is an irreversible process, kinetic effects (i.e. non-equilibrium events) can affect the folding outcome. Krobath et al. performed their folding simulations at the optimal folding temperature for the full length chain. One possibility is that co-translational folding can occur in a folding simulation that incorporates protein translation and is run at a lower temperature[1]. Another question, is why did P1 and P2 behave similarly despite the favorable topology of P1 with local native contacts? Although Krobath et al. extensively investigated the energy spectrum and conformations of the 44-residue fragments of the 48-residue proteins, some further studies on the energies and conformations of shorter chain fragments could be useful.

In the present work, the lattice proteins studied by Krobath et al. were simulated using LatPack, a lattice protein simulation suite [13]. The folding kinetics of the full chains were first studied. The energies of the lattice proteins as a function of chain length were then studied. Based on the folding kinetics of the full length chain, co-translational folding simulations of the proteins were conducted.

[1] The rough hypothesis is that shorter chain segments, lacking the full complement of stabilizing native interactions, might have stable native-like conformations only at lower temperatures. Additionally, the kinetic slowdown for a lower temperature is less important because shorter chain segments have fewer available conformations. This project focused on simulations, but a good follow-up to the present work might examine in general the dependence of polymer properties on length and temperature.

II. MODEL AND METHODS

A. Lattice protein model

In a lattice protein representation, the residues of a protein chain are reduced to occupied sites on a cubic lattice, with successive residues connected by covalent bonds that link two neighboring sites. No two residues may occupy the same site simultaneously. The two proteins studied in this work are characterized by different native topologies, with P1 having a local contact topology and P2 a non-local contact topology. It can be seen in Fig. 1 that P1 consists of two subdomains whereas in the native structure of P2, the N-terminus and C-terminus are in contact.

The particular sequences of the two 48-mer proteins used in the present work are taken from Krobath et al [12] and are shown in Table I along with their melting temperatures (kT units). The sequences were designed for the two topologies under study via the Z-score optimization method [14].

B. Interaction potential

The sequence-specific potential governing lattice proteins is simple. Two residues are considered to be in contact if they occupy two neighboring lattice sites and they are not covalently bonded. The potential energy function may be defined as follows:

$$E(\{\vec{r}_i\},\{n_i\}) = \sum_{i>j}^{L} \epsilon(n_i,n_j)\Delta(\vec{r}_i - \vec{r}_j) \qquad (1)$$

where $\{\vec{r}_i\}$ are the coordinates of the residues, $\{n_i\}$ is the particular sequence of residues, and $\Delta(\vec{r}_i - \vec{r}_j)$ is unity for any pair of neighboring, non-covalently bonded residues. The interaction energy between residues is the Miyazawa-Jernigan statistical potential, a 20×20 matrix of residue-residue interaction energies (Table VI in [15]).

C. Monte Carlo simulations

The lattice simulations in the present work were run using LatPack, a lattice protein software suite [13]. The `LatFold` program within LatPack explores lattice protein conformational space via the Metropolis MC algorithm [16] and supports two ergodic movesets: pull-moves [17] or pivot-moves [18]. In the present work, the pull moveset was used exclusively. The pivot moveset was found to

Protein	Sequence	T_m (from [12])	T_f
P1	FHNFKNGDRRATSHCHWFWDQSYPWWAMFLAVPISHKDLMRYEDDPPK	0.32	0.31
P2	PMHFDLKRYADHSYRDQPWFREVLNGKDPSAHTNIHAKCWFMFDWPWS	0.34	0.33

TABLE I. The two lattice proteins under study. Temperatures in kT units. T_f determined based on data in Fig. 2

cost more CPU time per MC step and was less efficient at folding the proteins under study.

LatPack also implements vectorial protein folding via the `LatFoldVec` program for simulations of co-translational folding. In between a fixed interval of MC steps, the protein chain is elongated by a single residue. Simulations of co-translational folding of lattice were studied using `LatFoldVec` in the present work.

In the studies by Krobath et al., lattice protein conformations were restricted by the presence of a chemically inert ribosome as well as tethering of one end of the chain near the ribosome. The LatPack software suite does not currently support such a restriction on protein conformations, and as such, the lattice proteins simulated in the present work were unrestrained.[2] Continuing studies should begin with implementation of conformational restrictions in LatPack.

1. Simulation procedures

The computations in the present work were run on the Odyssey cluster supported by the FAS Division of Science, Research Computing Group at Harvard University.

a. Folding simulations The dependence of folding speed on temperature was first analyzed. For each protein at each temperature (kT = 0.14, 0.17, 0.20, 0.23, 0.26, 0.29, 0.31, 0.32, 0.33, 0.34, 0.35, 0.37, 0.40), 20 folding simulations were run, each starting from a linear conformation. Simulations were stopped when the folded conformation was reached or after 10^8 simulation steps.

b. Energy sampling simulations To measure the equilibrium properties of the lattice proteins as a function of chain length, `LatFoldVec` was used, but with a large number of steps in between each chain elongation to ensure equilibration at each chain length. The number of steps per chain length was 50000*(chain length); e.g. 250000 steps for a 5-residue chain, 300000 steps for a 6-residue chain, etc.

c. Co-translational folding simulations Co-translational folding simulations were performed at kT=0.26 and kT=0.31 for P1 and kT=0.28 and kT=0.33 for P2. Two different fixed translation rates were tested: 150000 steps/residue and 300000 steps/residue (6600000 and 13200000 steps total, respectively). For each simulation configuration, 5 simulations were run. Co-translational simulations begin with the first five N-terminal residues in a linear conformation; successive residues are added to the growing C-terminus of the chain. To measure the similarity of a structure to the state, cRMSD (coordinate root mean square deviation) was used, as determined by the `latMap` utility within LatPack. RMSD is the average distance between corresponding lattice points after an optimal rigid body superposition is performed. For a protein chain shorter than the full length protein, the reference native state was considered to be the fragment of the native protein corresponding to the chain under consideration. The value of the spacing between lattice points was set to unity.

[2] LatPack is built upon the Bioinformatic Uility library. Altering the code to support restricted coordinates was not feasible for this project.

III. RESULTS

A. Kinetics of the full polymers

The literature-reported melting temperatures of the two proteins is reported in Table I. To determine the optimal folding temperature as well as characterize the kinetics of folding rate and temperature, folding simulations were performed for each protein for a range of temperatures (see part II C 1 a).

The first-passage time to the folded conformation was measured. A simulation was stopped if folding did not occur within 10^8 MC steps. Unfortunately, this limit was too low, and statistics could only be obtained for temperatures close to the optimal folding temperature. The average time to fold is shown in Fig. 2. For each simulation configuration, a maximum of 20 folding simulations were run within the allocated simulation time. This number of simulations is insufficient for the amount variance in folding times, so no error bars are depicted in the plots. Nevertheless, the general trend of the plot indicates that folding time is optimal at a temperature slightly below the melting temperature (kT=0.31 for P1 and kT=0.33 for P2). These optimal folding temperatures are in agreement with Krobath et al. [12].

For these two optimal folding temperatures, the folding kinetics were analyzed by fitting the folding time data to a single exponential (Fig. 2). An estimate of 3.77×10^-7/step and 1.11×10^-7/step was obtained for P1 and P2, respectively, which are comparable in magnitude to what Krobath et al. obtained ($2.18 \pm 2 \times 10^-7$/step and $0.523 \pm .001 \times 10^-7$/step) [11]. Differences from Krobath et al. can be attributed to an insufficient statistical sample or the different moveset employed by LatPack.

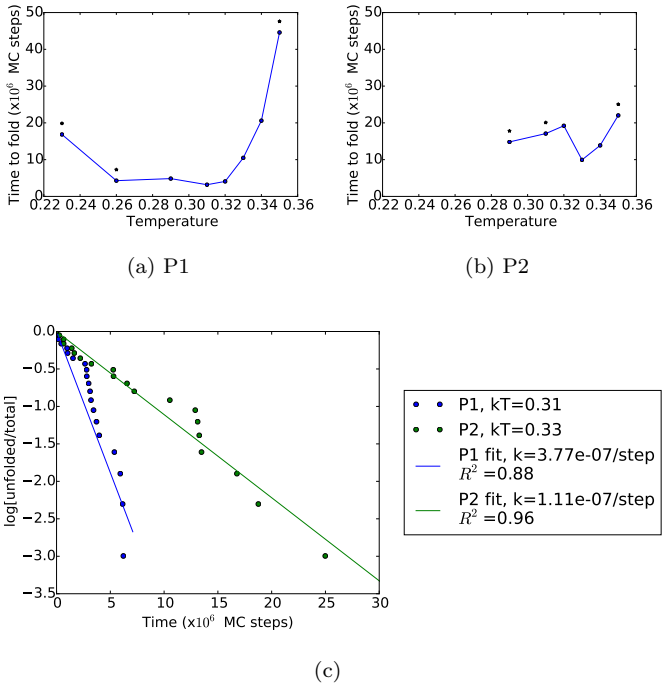


FIG. 2. Top: the average time to fold (in millions of MC steps) for P1 (a) and P2 (b) as a function of simulation temperature. Points marked with * indicate sets of simulations in which not all simulations ended up in a folded state after 10^8 steps. (Only the successful folding runs could be averaged.) Bottom: The folding times at the optimal folding temperature for each protein was fit to a single exponential.

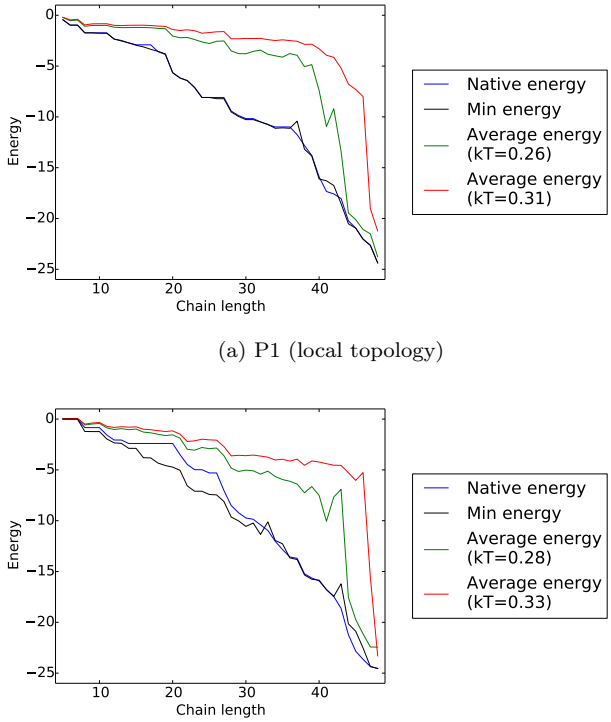


FIG. 3. Energy as a function of chain length for P1 (a) and P2 (b). Native energy is the energy of a particular chain length in the native conformation. Min energy is the minimum energy that was sampled at each chain length (using the simulations at the two different temperatures). Average energies at two different temperatures are shown for each protein. The statistical error in the average energies is of the same order as or less than the line thickness.

B. Energy distributions

To characterize the proteins under study at various chain lengths, which are the states in mid-translation, long MC simulations that sequentially add residues to the C-terminus of the protein chain were run, with the number of steps proportional to the length of the lattice protein chain: the number of steps for each chain length was 50000 per residue. The energy of the system was recorded. Based on the kinetics measured in Part III A, simulations were run at two temperatures for each protein: kT = 0.31 and kT = 0.26 for P1 and kT = 0.33 and kT = 0.28 for P2. The temperatures selected correspond to the optimal folding temperature was well as some lower temperature (0.05 lower) that still maintained reasonable folding kinetics.

In Fig. 3, the average energies measured at each chain length are compared with the native conformation energy. The native conformation energy is the energy of the protein chain of a particular length in the conformation of the native, folded state. The native energy for both P1 and P2 monotonically decrases as a function of chain length, with the energy for P1 decreasing slightly more rapidly at intermediate chain lengths. The minimum energy at each chain length was determined by taking the minimum energy sampled by these MC simulations at each chain length. It can be observed that for P1, the protein with local topology, the minimum energy largely coincides with the native energy. This is not the case for P2, since the native contacts are distant along the protein chain.

The average energies as a function of chain length are higher than the minimum or native energy. This is to be expected since native conformation or other states close in energy to the native conformation represent only a fraction of the possible conformations. On the other hand, it also shows that, for both P1 and P2, the equilibrium conformation of the intermediate length chains is not the native conformation. Similar to what Krobath et al. found, the average energy (and conformation) only approaches the native state at chain lengths beyond 42 (out of 48) [12]. Note that in the design for the sequences of the two proteins, the sequences were optimized in the context of the full protein chain, so these results are reasonable. Average energies for both proteins seem to exhibit a kink around 40 residues that is more pronounced at the lower simulation temperatures. The reason for this was not investigated.

C. Co-translational folding

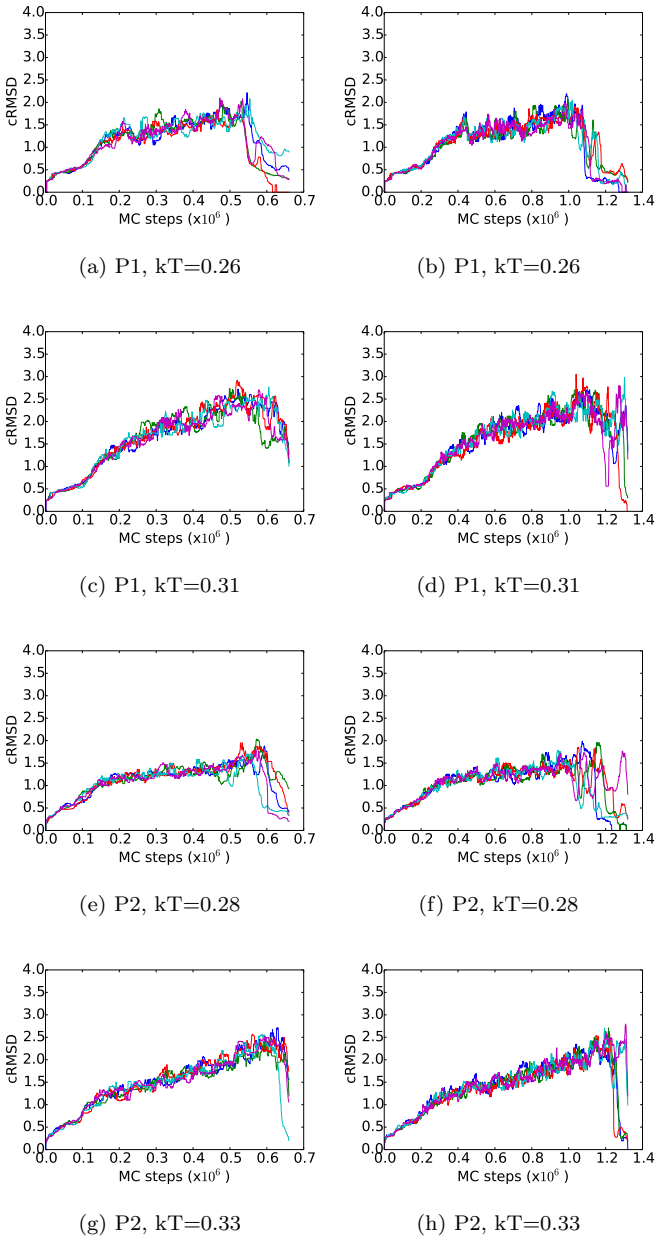
Although the analysis of equilibrium energy distributions in the previous section showed that the equilibrium energy/conformations at chain lengths shorter than around 42 residues do not match the native conformation, the question remained whether non-equilibrium co-translational simulations might exhibit different behavior.

Co-translational simulations were run using `LatFoldVec` for the two proteins, at two temperatures, and using two chain elongation schedules (for a total of 8 simulation configurations). The two elongation schedules are 150000 MC steps/residue and 300000 MC steps/residue. The schedules were selected on the basis of the mean folding times for the two proteins such that the total simulation time is on order of the mean folding time for the two proteins at the optimal folding temperature. These folding simulations are non-equilibrium because elongation is a non-equilibrium event, and the lattice protein has only a short time (in steps) to equilibrate before the next monomer is added to the chain.

To measure the resemblance of intermediate chain lengths to the native structure, cRMSD (see Part II C 1 c) was used. Fig. 4 depicts simulation trajectories for the 8 different simulation configurations. Each plot contains 5 trajectories. P1 trajectories are shown in the top two rows, and P2 trajectories in the bottom two rows. The left column of plots are folding simulations with an elongation schedule of 150000 MC steps/residue, and the right column of plots show folding simulations with an elongation schedule of 30000 MC steps/residue. Although the bottom axis is demarcated in MC steps, it should be kept in mind that the chain length increases with the number of MC steps at a constant rate.

For each simulation condition, the set of five trajectories exhibit very similar behaviors, with some greater variance at the end of the simulations (higher chain lengths) where the probability of folding increases. It is apparent for all 8 simulation conditions that the equilibrium conformations at intermediate chain length are not close to native, as one might predict given Krobath et al.'s finding from their equilibrium simulations that the probability to fold is only significant past 44 out of 48 residues [12]. There is not much difference between the two elongation schedules. All configurations run at 300000 steps/residue show at least one or more trajectories reaching the native state at the end of the simulation, which is sensible given that there is more time to fold to the native state with the longer elongation schedule. This agrees with Ciryam et al.'s kinetic study [5].

Nonetheless there are some notable differences between simulations at lower and higher temperatures. Remarkably, more simulations for P1 and P2 that were run at the lower temperature ended in a folded state than simulations that were run at the temperature that is optimal for folding each of the full length proteins. In the higher temperature simulations, the folding trajectories reach higher peak RMSD values, indicating greater deviation from the native conformation. For example, the two P1 co-translational folding simulations following the 150000 steps/residue schedule can be compared (Fig. 4a can be compared with Fig. 4c). While RMSD at kT=0.26 reaches a maximum of about 2, maximum RMSD for trajectories kT=0.31 exceeds 2.5. The folding trajectories at kT=0.26 are able to go towards a native-like state (one



(a) P1, kT=0.26 (b) P1, kT=0.26

(c) P1, kT=0.31 (d) P1, kT=0.31

(e) P2, kT=0.28 (f) P2, kT=0.28

(g) P2, kT=0.33 (h) P2, kT=0.33

FIG. 4. cRMSD timeseries of co-translational folding simulation trajectories. Left: 150k MC steps/elongation. Right: 300k MC steps/elongation. Each plot shows 5 trajectories. cRMSDs were measured every 5000 MC steps and subject to a median filter with a window size of 101.

of them achieving 0 RMSD) towards the end of the simulation time, but this is not the case for the trajectories at kT=0.31. A similar result is the case for P2, although folding is slightly more delayed for P2. Even though P2 has a long-distance native contact topology, it follows the same pattern as P1: the lower temperature simulations are better able to reach native-like states at the end of the simulation. It is likely that native-like conformations of intermediate chain lengths are not stable at higher temperatures, and therefore co-translational folding is more likely at a lower temperature than the optimal folding temperature for the full protein.

IV. CONCLUSIONS

Building off of results by Krobath et al., two lattice proteins with different contact topologies were simulated. Equilibrium simulations of full entire proteins matched literature results in terms of kinetics and folding dependence on temperature. The energy spectra of the proteins was studied as function of chain length. Because these proteins have been designed for folding, but not specifically for co-translational folding, the average energy is greater than the native conformation energy at all chain lengths until the protein chain is nearly full-length.

The full-length proteins fold fastest at some optimal folding temperature T_f, but co-translational simulations of the two proteins under study showed that co-translational folding is more probable at lower temperatures. It is debateable whether the folding observed in the co-translational MC simulations is truly co-translational folding since much of the folding comes toward the end of the simulations, but it is certainly the case that more trajectories run at the lower temperature ended up folded at the end of the simulation than trajectories run at T_f. Further simulations at different temperatures and elongation schedules should be explored.

The simulations reported in the present work represent a preliminary result that needs some further refinment. In addition to revising LatPack to support conformational restrictions, more and longer simulations should be run to converge statistical estimates. Additionally, more extensive analysis of the co-translational simulations should be carried out beyond RMSD.

V. ACKNOWLEDGMENTS

LatPack [13] may be downloaded from `http://www.bioinf.uni-freiburg.de/Software/LatPack/`. Lattice proteins were visualized in VMD and rendered using the Tachyon ray tracing library [19, 20]. Special thanks to William Jacobs for helpful discussions.

[1] K. S. Hingorani and L. M. Gierasch, Current opinion in structural biology **24**, 81 (2014).
[2] A. A. Komar, Trends in biochemical sciences **34**, 16 (2009).
[3] A. N. Fedorov and T. O. Baldwin, Journal of Biological Chemistry **272**, 32715 (1997).
[4] Y. Han, A. David, B. Liu, J. G. Magadán, J. R. Bennink, J. W. Yewdell, and S.-B. Qian, Proceedings of the National Academy of Sciences **109**, 12467 (2012).
[5] P. Ciryam, R. I. Morimoto, M. Vendruscolo, C. M. Dobson, and E. P. O'Brien, Proceedings of the National Academy of Sciences **110**, E132 (2013).
[6] E. P. O'Brien, P. Ciryam, M. Vendruscolo, and C. M. Dobson, Accounts of chemical research **47**, 1536 (2014).
[7] E. Siller, D. C. DeZwaan, J. F. Anderson, B. C. Freeman, and J. M. Barral, Journal of molecular biology **396**, 1310 (2010).
[8] E. P. O'Brien, M. Vendruscolo, and C. M. Dobson, Nature communications **5** (2014).
[9] E. Wang, J. Wang, C. Chen, and Y. Xiao, Scientific reports **5** (2015).
[10] L. Mirny and E. Shakhnovich, Annual review of biophysics and biomolecular structure **30**, 361 (2001).
[11] H. Krobath and P. F. Faísca, Physical biology **10**, 016002 (2013).
[12] H. Krobath, E. I. Shakhnovich, and P. F. Faísca, The Journal of chemical physics **138**, 215101 (2013).
[13] M. Mann, D. Maticzka, R. Saunders, and R. Backofen, HFSP journal **2**, 396 (2008).
[14] T.-L. Chiu and R. A. Goldstein, Protein engineering **11**, 749 (1998).
[15] S. Miyazawa and R. L. Jernigan, Macromolecules **18**, 534 (1985).
[16] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, The journal of chemical physics **21**, 1087 (1953).
[17] N. Lesh, M. Mitzenmacher, and S. Whitesides, in *Proceedings of the seventh annual international conference on Research in computational molecular biology* (ACM, 2003) pp. 188–195.
[18] N. Madras and A. D. Sokal, Journal of Statistical Physics **50**, 109 (1988).
[19] W. Humphrey, A. Dalke, and K. Schulten, Journal of Molecular Graphics **14**, 33 (1996).
[20] J. Stone, An Efficient Library for Parallel Ray Tracing and Animation, Master's thesis, Computer Science Department, University of Missouri-Rolla (1998).