Modeling the effect of codon translation rates on co-translational protein folding mechanisms of arbitrary complexity

Luca Caniparoli¹ and Edward P. O'Brien^{2,a)}

¹*International School for Advanced Studies (SISSA), via Bonomea 265, 34136 Trieste, Italy*

²*Department of Chemistry, Pennsylvania State University, University Park, University Park, Pennsylvania 16802, USA*

(Received 11 November 2014; accepted 24 March 2015; published online 9 April 2015)

In a cell, the folding of a protein molecule into tertiary structure can begin while it is synthesized by the ribosome. The rate at which individual amino acids are incorporated into the elongating nascent chain has been shown to affect the likelihood that proteins will populate their folded state, indicating that co-translational protein folding is a far from equilibrium process. Developing a theoretical framework to accurately describe this process is, therefore, crucial for advancing our understanding of how proteins acquire their functional conformation in living cells. Current state-of-the-art computational approaches, such as molecular dynamics simulations, are very demanding in terms of the required computer resources, making the simulation of co-translational protein folding difficult. Here, we overcome this limitation by introducing an efficient approach that predicts the effects that variable codon translation rates have on co-translational folding pathways. Our approach is based on Markov chains. By using as an input a relatively small number of molecular dynamics simulations, it allows for the computation of the probability that a nascent protein is in any state as a function of the translation rate of individual codons along a mRNA's open reading frame. Due to its computational efficiency and favorable scalability with the complexity of the folding mechanism, this approach could enable proteome-wide computational studies of the influence of translation dynamics on co-translational folding. © 2015 AIP Publishing LLC. [<http://dx.doi.org/10.1063/1.4916914>]

I. INTRODUCTION

Far-from-equilibrium processes can govern various aspects of cellular life. In such cases, kinetics rather than thermodynamics can determine the structures formed by self-organizing systems, and the pathways by which they assemble.¹ One fundamental example of such phenomena in cells is the folding of a protein molecule concomitant with its synthesis by the ribosome^{2–5} (Fig. 1). The ribosome translates each codon position along a messenger RNA (mRNA) molecule into a specific amino acid that is covalently attached to the C-terminus of the elongating nascent protein. The rate at which this reaction occurs is referred to as the codon translation rate. It has been found that altering this rate at specific codon positions within a transcript's open reading frame (ORF) can dramatically influence the extent of co-translational folding and what structures are formed,⁶ demonstrating experimentally that for some proteins, co-translational folding is a far-from-equilibrium self-assembly process.

Cells benefit by taking advantage of the coupling between codon translation rates and co-translational folding. Experiments have revealed that converting naturally occurring slow-translating codons to fast translating codons near domain boundaries can decrease the probability that a protein will co-translationally fold^{7–10}—suggesting that evolutionary selective pressures have optimized translation-rate profiles along an ORF in some cases to maximize co-translational folding.

Analyses of synonymous codon usage across transcriptomes reveal systematic biases between different species, and that rare codons that are assumed to translate more slowly are often found in α -helical and β -strand structural motifs,^{11–13} further supporting the idea that the pattern of codon translation rates along a mRNA's ORF can have an important role in determining aspects of an organism's phenotype. When these patterns of translation rates are altered, the process of co-translational folding can go awry, with the misfolding and malfunction of a nascent protein ensuing.^{14–16}

For these reasons, a crucial challenge to understanding protein behavior in cells is to be able to model the coupling between individual codon translation rates and the states (conformations) that a nascent protein populates during its co-translational folding process. However, a comprehensive theoretical framework describing this phenomenon is still lacking. Attempts at addressing this challenge via a probabilistic approach¹⁷ were successful in modeling the influence of codon translation rates on co-translational folding involving a single pathway and up to three states,^{3,18} while coarse-grained molecular dynamics simulations of a two-state folding domain allowed molecular aspects of co-translational folding to be explored.^{19,20} In cells, more complex situations can occur involving multiple folding pathways²¹ and multiple metastable intermediate or misfolded states.^{9,22} In these cases, determining the functional relationship between the codon translation rate and co-translational folding is mathematically challenging as are the molecular dynamics simulations of such situations.

Here, we introduce a general approach for predicting the influence of codon translation rates on co-translational folding

^{a)} Author to whom correspondence should be addressed. Electronic mail: epo2@psu.edu

FIG. 1. Co-translational protein folding. (a) The ribosome translates codons contained in the open reading frame of a mRNA molecule into a nascent protein. (b) Starting from the 5' end (codon 1), the ribosome uni-directionally slides (large gray arrow) along the mRNA molecule and converts the genomic information in the ORF into a nascent protein (blue), which emerges through a channel known as the ribosome exit tunnel. (c) At a given nascent chain length L , the nascent chain has the potential to form tertiary structure; such states may include folded, intermediate, and unfolded conformations. The arrows indicate that these states may be able to interconvert at the given chain length.

mechanisms of arbitrary complexity. We first model the co-translational folding process as a Markov chain, for which we are able to analytically solve the probability that a nascent protein is in any one of an arbitrarily large number of states during translation as a function of the nascent chain length, the translation rates of individual codon positions along an ORF, and the rates of inter-conversion between the states of the protein molecule. Combining this mathematical model with conformational inter-conversion rates calculated from simulations of arrested ribosomes allows for the accurate prediction of co-translational folding behavior as a function of any translation-rate profile along a mRNA's open reading frame. The process is as follows: A set of MD simulations of a particular protein are run at different nascent chain lengths on translationally arrested ribosomes. From these simulations, the inter-conversion rates between different nascent protein states are extracted and used as input parameters for the Markov chain model, leaving the individual codon translation rates as the only free parameters in the mathematical model. This allows for the rapid and accurate prediction of the influence of different translation-rate profiles on the co-translational folding of the simulated protein, without the need for running explicit molecular dynamics simulations of continuous translation at those rates. A key distinction between our approach and many others^{23–26} is that our model uses a pathway-probability equation instead of a master equation.

II. METHODS

A. Coarse-grained model details

To model the 50S subunit of the *E. coli* ribosome for molecular dynamics simulations, we used the coarse-grained model recently introduced in Refs. 3 and 19. In this model, proteins are represented using one interaction site per amino acid and nucleic acids by three or four interaction sites depending on whether they are a purine or a pyrimidine. There are five energy terms in the force field of this coarse-grained model: a bond energy term applied to two covalently linked interaction sites, a bond angle term applied to three covalently linked interaction sites, a dihedral angle term between four sequentially linked sites, a pair-wise electrostatic interaction energy term applied to interaction sites that are not covalently

linked, and a Lennard-Jones term that models non-covalent van der Waals interactions. The first four energy terms are fully transferable between different protein and RNA molecules, while the Lennard-Jones energy term is system-dependent. Specifically, we use Go's approach^{27–29} of treating native interactions as attractive and non-native interactions as repulsive in the Lennard-Jones interactions. This approximation produces many realistic features of protein folding and unfolding, can lead to accurate predictions of experimental observables,³⁰ and is an assumption supported from results from all-atom simulation models in which transferable force fields were used.^{31,32} We also note that in this model, there are non-specific electrostatic interactions between the ribosomal components and the nascent chain, with counter-ion screening accounted for using Debye-Huckel theory. Full details of the functional forms of the force-field terms and their parameters can be found in Refs. 3 and 19. The transferable force-field parameters reported in Ref. 20 were used to model the polyglycine linker that was fused to MIT (Microtubule Interacting and Trafficking molecule), a single domain protein (Fig. 4(a)).

To produce realistic thermodynamic properties of the folding process of the MIT domain, the Lennard-Jones well-depths for the native interactions between the MIT interaction sites were scaled to result in a native state stability of -3 kcal/mol at 310 K. This is the stability expected for a domain of MIT's size and structural class.³³

B. Langevin dynamic's simulations

The starting conformation of the ribosome nascent chain complex, containing the 65-residue MIT-polyglycine fusion construct, was produced as detailed previously.^{3,19} Three different global translation speeds were simulated that involved adding a new amino-acid to the growing nascent chain with characteristic simulation times of 0.15 ns, 1.5 ns, and 15 ns. At each codon position, new amino-acids were incorporated with time scales that were exponentially distributed and had an average time equal to the characteristic time. By using different random number seeds for the initial velocity distribution at 310 K, 300 independent synthesis trajectories were simulated starting from the initial ribosome-nascent chain complex (RNC) configuration for each of the three different global translation speeds. Charmm³⁴ version c35b5 was used

to run these Langevin dynamic simulations. In the simulations, the integration time step was 15 fs, the collision frequency was 0.05 ps^{-1} , and the ribosomal protein and RNA molecules were held rigid using the “cons fix” function of the constraints module in CHARMM. It has been shown previously²⁰ that holding the ribosome rigid does not alter the co-translational folding properties of the nascent chain, as there are no large-scale structural fluctuations in the ribosome tunnel or tunnel vestibule where co-translational folding takes place. Coordinates from these trajectories were saved for later analysis every 50 integration time steps for the 0.15 ns-per-codon simulations and every 500 time steps for the other synthesis times. The trajectories were stopped once they reached a nascent chain length of 121 residues. In the supplementary material, we provide the CHARMM script used to run the continuous translation simulations.⁴⁷

Arrested ribosome Langevin dynamic’s simulations at 310 K were run for nascent chain lengths between 65 and 120 residues. During these simulations, no new residues were added to the nascent chain. At each nascent chain length, 8 independent trajectories were simulated, each for 1.1×10^8 integration time-steps, with the first 2.2×10^4 step discarded to allow for equilibration. The calculated quantities did not change with longer equilibration periods. Coordinates from these trajectories were saved for later analysis every 500 time steps.

C. Analysis of the simulations

The time series of RNC structures saved from the simulation trajectories was first classified as corresponding either to the unfolded state, intermediate state, or folded state. To rigorously define these categories along various order parameters, we examined a two-dimensional free energy surface of the isolated MIT domain near its melting temperature of 324 K. The free energy (Fig. 2(a)) was projected onto one axis corresponding to the fraction of native contacts between helices 1 and 2 (Q_{12}), and the other axis corresponding to the fraction of native contacts of helix 3 with helices 1 and 2 (Q_{12-3}). Three free-energy basins were observed on this free-energy surface. A simulation structure of the MIT domain was classified as

being in the unfolded state when its $Q_{12} < 0.05$ and $Q_{12-3} < 0.05$, folded state when its $Q_{12} > 0.85$ and $Q_{12-3} > 0.85$, and intermediate state when its $Q_{12} > 0.75$ and $Q_{12-3} < 0.05$. We considered only transitions between these states and used the transition-based assignment method to assign conformations in the transition region to them.^{23,24} Specifically, structures that fell outside these defined regions were considered to be sampling the transition region and were classified as U, I, or F based on which of these states had most recently been sampled in the trajectory. That is, if a trajectory was sampling the unfolded region (i.e., $Q_{12} < 0.05$ and $Q_{12-3} < 0.05$) and it made an excursion to the transition region, those structures were still classified as being unfolded until the trajectory reached one of the other defined states (i.e., either I ($Q_{12} > 0.75$ and $Q_{12-3} < 0.05$) or F ($Q_{12} > 0.85$ and $Q_{12-3} > 0.85$)). These definitions rule out short-lived transitions, which lead to a deviation from the expected Markovian behavior. The resulting observed life-times of the individual states were single-exponentially distributed (data not shown).

Applying the above definitions to both the arrested- and continuous-translation simulations resulted in the time-series of the MIT domain being in states U, I, or F. The probability of being in states U, I, or F at nascent chain length L , immediately before adding the next amino acid, was computed from the simulations as the arithmetic mean $N_X^{(L)}/N_{\text{traj}}^{(L)}$, where $N_{\text{traj}}^{(L)}$ is the number of independent trajectories simulated at that length and translation speed, and $N_X^{(L)}$ is the number of those trajectories in which state X was found immediately preceding the addition of the next amino acid. The standard error about the mean was computed for these probabilities using the bootstrapping method in which 10 000 independent distributions were calculated for each probability value.

From the arrested ribosome simulations, the inter-conversion rates between states U, I, and F were computed as follows. First, from the time series of the states, we calculated the transition matrix $H^{(L)}$ whose elements $h_{i \rightarrow j}^{(L)}$ correspond to the number of times a transition from state i to state j was observed in a time interval Δt . The elements on the diagonal $h_{i \rightarrow i}^{(L)}$ are the number of times no transition was observed. From the counts, we obtained the empirical probability of transitioning from i to j in a time interval Δt ,

FIG. 2. A free-energy surface of the MIT domain in bulk solution near its melting temperature and implied time scales from the Markov analysis. (a) The free-energy contour surface is projected along the order parameters Q_{12} and Q_{12-3} at 320 K. The free energy at a given point on this surface is calculated as $-RT \ln[P(Q_{12}, Q_{12-3})]$, where R is the universal gas constant, T is the temperature K, and $P(Q_{12}, Q_{12-3})$ is the probability of the simulation sampling a point (Q_{12}, Q_{12-3}) during the simulations. The energy scale, on the right, is in units of kcal/mol. The three states (U, I, and F) can be seen as blue basins in this surface. (b) The largest implied time scales as a function of the lag time, Δt , for nascent chain lengths of 69 (circles), 75 (squares), 85 (diamonds), 95 (upward triangles), 105 (downward triangles), and 115 residues (stars). Lines are to guide the eye.

$$p_{i \rightarrow j}^{(L)}(\Delta t) = \frac{h_{i \rightarrow j}^{(L)}}{\sum_l h_{i \rightarrow l}^{(L)}}. \quad (1)$$

Supposing the underlying process to be Markovian, the time $t_{i \rightarrow j}$ between the transition events is exponentially distributed. The probability $p_{i \rightarrow i}^{(L)}(\Delta t)$ that no transition occurs before Δt is thus

$$p_{i \rightarrow i}^{(L)}(\Delta t) = \text{Prob}(t_{i \rightarrow j} > \Delta t, \forall j (\neq i)) = \prod_{j(\neq i)} e^{-k_{ij}^{(L)} \Delta t}, \quad (2)$$

where $k_{ij}^{(L)}$ is the transition rate from states i to j . The derivation of this equation relies on the fact that the timescale of the transitions, $(k_{ij}^{(L)})^{-1}$, is larger than Δt , i.e., that the probability of the 2-step transitions $i \rightarrow j \rightarrow i$ is negligibly small. The probability $p_{i \rightarrow j}^{(L)}(\Delta t)$ of transitioning from i to j can be written as the complement of the previous probability, multiplied by the probability $q_{i \rightarrow j}^{(L)} \equiv \frac{k_{ij}^{(L)}}{\sum_l k_{il}^{(L)}}$ of choosing the $i \rightarrow j$ transition instead of all the other possible transitions

$$\begin{aligned} p_{i \rightarrow j}^{(L)}(\Delta t) &= (1 - p_{i \rightarrow i}^{(L)}(\Delta t)) q_{i \rightarrow j}^{(L)} \\ &= \frac{k_{ij}^{(L)}}{\sum_l k_{il}^{(L)}} \left(1 - \prod_{l(\neq i)} e^{-k_{il}^{(L)} \Delta t}\right). \end{aligned} \quad (3)$$

By using Eqs. (2) and (3), the relations for the rates are

$$\begin{aligned} S &\equiv \sum_{j(\neq i)} k_{ij}^{(L)} = -\frac{\log p_{i \rightarrow i}^{(L)}(\Delta t)}{\Delta t}, \\ k_{ij}^{(L)} &= \frac{S}{1 - p_{i \rightarrow i}^{(L)}(\Delta t)} p_{i \rightarrow j}^{(L)}(\Delta t), \end{aligned} \quad (4)$$

and finally, by replacing the probabilities with their empirical estimates, we obtain

$$k_{ij}^{(L)} = -\frac{1}{\Delta t} \frac{\log \frac{h_{i \rightarrow j}^{(L)}}{\sum_l h_{i \rightarrow l}^{(L)}}}{\sum_{l(\neq i)} \frac{h_{i \rightarrow l}^{(L)}}{\sum_l h_{i \rightarrow l}^{(L)}}}. \quad (5)$$

Equation (5) was used to calculate $k_{ij}^{(L)}$ from the arrested ribosome simulations. In this calculation, Δt is set to 7.5 ps, the time interval at which coordinates were saved from the arrested ribosome simulations. To test if the calculated rates were robust to the value of Δt , we plotted the largest implied timescale $\tau = -\frac{\Delta t}{\ln \lambda_{\max}}$ (where $\lambda_{\max}$ is the largest Eigenvalue smaller than 1 of the transition probability matrix $\mathbf{T}$, see below) as a function of Δt at different nascent chain lengths of the MIT domain. We find that τ is constant with respect to Δt (Fig. 2(b)), demonstrating that $\Delta t = 7.5$ ps provides accurate rate estimates.

D. Predictions using Eq. (10)

From the analysis of the arrested ribosome simulations, the inter-conversion rates between states are known (Eq. (5)), and hence, the only free variable in the $\mathbf{A}_{(L)}$ and $\mathbf{T}_{(L)}$ matrices at nascent chain length L is the translation rate at codon $L + 1$. Thus, Eq. (10) can be used to predict the influence on the state probabilities by substituting in different values of $k_A^{(L+1)}$. In the supplementary material, we provide the Matlab code

that implements Eq. (10), that is, the code utilizes the inter-conversion rates measured on arrested ribosomes to predict how the state probabilities change with the translation rate of individual codons.⁴⁷

The standard error about the mean for these predictions was estimated by Metropolis Monte Carlo. Specifically, 10 000 random transition matrices $H^{(L)}$ were generated from a multinomial distribution. These matrices were used to compute an ensemble of predicted probabilities of being in the states U, I, or F, whose standard deviation gives the estimate of the error.

III. RESULTS

Below, we first develop the Markov chain formalism for co-translational folding and then present an application of the method to a domain that co-translationally folds via parallel pathways.

A. Markov chain model of arbitrarily complex co-translational folding mechanisms

Consider a mRNA molecule whose ORF consists of M codons, numbered from 1 (the start codon) to M (the stop codon), as in Fig. 1(a). A ribosome molecule translating this ORF converts the genomic information encoded in the sequence of codons by uni-directionally translocating along the ORF one codon at a time, decoding the information at the new codon, and covalently attaching the corresponding amino acid to the nascent polypeptide chain before the next translocation step (Fig. 1(b)). For a nascent chain that is L residues in length at a given point during its translation, the rate of translation of codon $L + 1$, which elongates the nascent chain by one residue, is denoted $k_A^{(L+1)}$. L is always less than or equal to M . At length L , the nascent chain can interconvert between $N^{(L)}$ distinct states, e.g., folded, intermediate, and misfolded states, which we denote as $S_i^{(L)}$, where $i = 1, \dots, N^{(L)}$ (Fig. 1(c)). The explicit dependence of the $N^{(L)}$ on the chain length L is due to the fact that the states that are accessible to the protein can vary while it is synthesized by the ribosome. At nascent chain length L , these states can directly and reversibly interconvert with one another. The rate of interconversion between states $S_i^{(L)}$ and $S_j^{(L)}$ is denoted $k_{i,j}^{(L)}$. As explained in Sec. II, $S_i^{(L)}$ and $k_{i,j}^{(L)}$ are obtained from the arrested-ribosome Langevin dynamic's simulations. The probability of the $S_i^{(L)} \rightarrow S_j^{(L)}$ transition to occur is $t_{i,j}^{(L)} = k_{i,j}^{(L)} / (k_A^{(L+1)} + \sum_{l=1}^{N^{(L)}} k_{i,l}^{(L)})$ (Figs. 3(a) and 3(b)). The rates $k_{i,j}^{(L)}$ and transition probabilities $t_{i,j}^{(L)}$ explicitly depend on L because the chemical environment experienced by a nascent chain segment changes as it elongates.

With probability $a_i^{(L)} = k_A^{(L+1)} / (k_A^{(L+1)} + \sum_{l=1}^{N^{(L)}} k_{i,l}^{(L)})$, the ribosome attaches an amino acid to this nascent chain and the state $S_i^{(L)}$ directly transitions to state $S_i^{(L+1)}$ (Fig. 3(b)). The parameters $t_{i,j}^{(L)}$ and $a_i^{(L)}$, due to conservation of probability, satisfy the relation $\sum_{j=1}^{N^{(L)}} t_{i,j}^{(L)} + a_i^{(L)} = 1$. The time scale of the chemical step of peptide bond formation (i.e., the transition path time) is much smaller than that associated with transitions between different states of the protein and therefore, it is not

FIG. 3. A parallel co-translational folding reaction scheme with (a) rates and (b) elementary transition probabilities indicated. Assuming state S_1 corresponds to the folded state, then a domain that folds via this mechanism can take parallel pathways to the folded state, either directly from S_2 or S_3 . At length L , these three states can reversibly and directly interconvert with one another with rates $k_{i,j}^{(L)}$ and elementary transition probabilities $t_{i,j}^{(L)}$. Addition of a residue to the nascent chain shifts the system irreversibly from length L to length $L+1$ with rate $k_A^{(L)}$ and elementary reaction probability $a_i^{(L)}$ that state $S_i^{(L)}$ transitions to state $S_i^{(L+1)}$ after one step on this reaction network.

possible for state $S_i^{(L)}$ to directly convert into state $S_j^{(L+1)}$, when $i \neq j$. States $S_i^{(L)}$ and $S_i^{(L+1)}$ are effectively equivalent, having the same conformational characteristics with regard to the nascent chain segment of interest, e.g., a domain, but differ in that they occur at nascent chain lengths L and $L+1$. Furthermore, the process of amino acid addition by the ribosome is irreversible under physiological conditions, and hence, the states at length $L+1$ act as absorbing states for those at length L . A nascent chain described by this model will thus, at length L , interconvert between states due to thermal fluctuations, performing a random walk in the space of $S_i^{(L)}$ states until a new amino acid is added, at which time-point state $S_i^{(L)}$ directly transitions to, and is absorbed by, state $S_i^{(L+1)}$.

To model the influence of codon translation rates on co-translational folding, the central quantity we are interested in calculating is the probability $p_{i,final}^{(L)}$ of occupying the state $S_i^{(L)}$ when the new amino acid is added and the state is absorbed by $S_i^{(L+1)}$ as a function of the underlying codon translation rate $k_A^{(L)}$. We note that the probability $p_{i,final}^{(L)}$ is equal to the probability $p_{i,initial}^{(L+1)}$ that the nascent chain starts in state $S_i^{(L+1)}$ at length $L+1$ immediately after adding the $L+1$ residue

$$p_{i,initial}^{(L+1)} = p_{i,final}^{(L)} \quad (6)$$

The co-translational folding process from this perspective is a biased random walk on a reaction network consisting of subsets of reactions that can reversibly interconvert, connected by irreversible transitions between those subsets of states (Fig. 3). This problem can be mathematically defined as a Markov chain³⁵ and we can therefore use well established methods to derive an expression for the probability $p_{i,final}^{(L)}$ for the nascent chain to be in state $S_i^{(L)}$ immediately before attaching the next residue to the nascent chain. This approach is more general than recent approaches to analytically model co-translational

folding^{3,18} as it makes it possible to solve for folding mechanisms involving an arbitrarily large number of states.

Within the framework of Markov chains, we first define the matrices $\mathbf{T}_{(L)}$, whose elements are the transition probabilities $t_{i,j}^{(L)}$, and $\mathbf{A}_{(L)}$, which is a diagonal matrix with elements $a_i^{(L)}$. The initial probability distribution of nascent chain states at length L can be written as the vector $\mathbf{p}_{initial}^{(L)} = \{p_{1,initial}^{(L)}, p_{2,initial}^{(L)}, \dots, p_{N^{(L)},initial}^{(L)}\}$, where each $p_{i,initial}^{(L)}$ is the probability that the nascent chain is in state $S_i^{(L)}$ immediately after the L^{th} residue was added to the nascent chain. We can compute the probability $\mathbf{p}_{RW}^{(L)}(1) = \{p_{1,RW}^{(L)}(1), \dots, p_{N^{(L)},RW}^{(L)}(1)\}$ of being in each of the states after one step of the random walk, given that the step does not lead to absorption by the states at $L+1$, by taking the product $\mathbf{p}_{RW}^{(L)}(1) = \mathbf{p}_{initial}^{(L)} \mathbf{T}_{(L)}$. Iterating after n such steps, we have

$$\mathbf{p}_{RW}^{(L)}(n) = \mathbf{p}_{initial}^{(L)} \mathbf{T}_{(L)}^n \quad (7)$$

On the other hand, the probability $\mathbf{p}_A^{(L)}(n+1)$ of being absorbed by state $S_i^{(L+1)}$ at step $n+1$ can be computed from Eq. (7) by applying the matrix $\mathbf{A}_{(L)}$,

$$\mathbf{p}_A^{(L)}(n+1) = \mathbf{p}_{initial}^{(L)} \mathbf{T}_{(L)}^n \mathbf{A}_{(L)} \quad (8)$$

The probability $\mathbf{p}_{final}^{(L)}$ of being absorbed can be calculated exactly by summing Eq. (8) over all possible values of n ,

$$\begin{aligned} \mathbf{p}_{final}^{(L)} &= \mathbf{p}_{initial}^{(L)} \sum_{n=0}^{\infty} \mathbf{T}_{(L)}^n \mathbf{A}_{(L)}, \\ &= \mathbf{p}_{initial}^{(L)} [\mathbf{1} - \mathbf{T}_{(L)}]^{-1} \mathbf{A}_{(L)}, \end{aligned} \quad (9)$$

where $\mathbf{1}$ is the $N^{(L)} \times N^{(L)}$ identity matrix and the second equality in Eq. (9) uses the geometric sum $\sum_{n=0}^{\infty} \mathbf{T}_{(L)}^n = [\mathbf{1} - \mathbf{T}_{(L)}]^{-1}$.³⁵ Finally, using the equality $\mathbf{p}_{initial}^{(L)} = \mathbf{p}_{final}^{(L-1)}$ from Eq. (6), Eq. (9) yields

$$\mathbf{p}_{final}^{(L)} = \mathbf{p}_{final}^{(L-1)} [\mathbf{1} - \mathbf{T}_{(L)}]^{-1} \mathbf{A}_{(L)} \quad (10)$$

The matrix inversion operation $[\mathbf{1} - \mathbf{T}_{(L-1)}]^{-1}$ in Eq. (10) can be performed efficiently for matrices of very large dimensions (on a standard laptop, this can easily be done for matrices larger than 1000×1000), making it possible to use this equation to model the influence of codon translation rates on co-translational folding mechanisms on arbitrarily complex folding energy landscapes.³⁶

Equation (10) is the main theoretical result of this paper as it provides an exact expression for the probability of being in a given state (e.g., the folded state) during translation for arbitrarily complex folding mechanisms. Specifically, the i^{th} element of the vector $\mathbf{p}_{final}^{(L)}$ expresses the probability of finding the nascent chain in state $S_i^{(L)}$ at length L immediately before adding the next residue in terms of the codon translation rates and inter-conversion rates between states. We emphasize that once the parameters $k_{i,j}^{(L)}$ have been determined from arrested ribosome simulations, Eq. (10) enables the efficient computation of the probabilities $\mathbf{p}_{final}^{(L)}$ for arbitrary choices of codon translation rates along a transcript's ORF. It is, therefore, possible to rapidly study the effect of different codon translation-rate profiles on the process of co-translational folding.

B. An example: parallel-pathway co-translational folding

To illustrate this method in practice, consider the simple 3-state reaction scheme shown in Fig. 3. It represents a domain that can co-translationally fold via two-parallel pathways. In this situation, the matrices $\mathbf{T}_{(L)}$ and $\mathbf{A}_{(L)}$ used in Eq. (10) are

$$\mathbf{T}_{(L)} = \begin{pmatrix} 0 & t_{1,2}^{(L)} & t_{1,3}^{(L)} \\ t_{2,1}^{(L)} & 0 & t_{2,3}^{(L)} \\ t_{3,1}^{(L)} & t_{3,2}^{(L)} & 0 \end{pmatrix}, \quad \mathbf{A}_{(L)} = \begin{pmatrix} a_1^{(L)} & 0 & 0 \\ 0 & a_2^{(L)} & 0 \\ 0 & 0 & a_3^{(L)} \end{pmatrix}. \quad (11)$$

For example, the transition probability $t_{1,3}^{(L)} = k_{1,3}^{(L)} / [k_A^{(L+1)} + k_{1,2}^{(L)} + k_{1,3}^{(L)}]$, while $a_1^{(L)} = k_A^{(L+1)} / [k_A^{(L+1)} + k_{1,2}^{(L)} + k_{1,3}^{(L)}]$. The other elementary transition probabilities can be defined in a similar manner.¹⁷

Inserting the matrices in Eq. (11) into Eq. (10) and computing the matrix product $[\mathbf{1} - \mathbf{T}_{(L)}]^{-1} \mathbf{A}_{(L)}$ in Eq. (10) yields

$$[\mathbf{1} - \mathbf{T}_{(L)}]^{-1} \mathbf{A}_{(L)} = \frac{1}{\det(\mathbf{1} - \mathbf{T}_{(L)})} \begin{pmatrix} [1 - t_{2,3}^{(L)} t_{3,2}^{(L)}] a_1^{(L)} & [t_{1,2}^{(L)} + t_{1,3}^{(L)} t_{3,2}^{(L)}] a_2^{(L)} & [t_{1,3}^{(L)} + t_{1,2}^{(L)} t_{3,1}^{(L)}] a_3^{(L)} \\ [t_{2,1}^{(L)} + t_{2,3}^{(L)} t_{3,1}^{(L)}] a_1^{(L)} & [1 - t_{1,3}^{(L)} t_{3,1}^{(L)}] a_2^{(L)} & [t_{2,3}^{(L)} + t_{1,3}^{(L)} t_{2,1}^{(L)}] a_3^{(L)} \\ [t_{3,1}^{(L)} + t_{2,1}^{(L)} t_{3,2}^{(L)}] a_1^{(L)} & [t_{3,2}^{(L)} + t_{1,2}^{(L)} t_{3,1}^{(L)}] a_2^{(L)} & [1 - t_{1,2}^{(L)} t_{2,1}^{(L)}] a_3^{(L)} \end{pmatrix}, \quad (12)$$

where $\det(\mathbf{1} - \mathbf{T}_{(L)})$ is the determinant of the matrix $\mathbf{1} - \mathbf{T}_{(L)}$, which in this case equals $1 - t_{1,2}^{(L)} t_{2,1}^{(L)} - t_{1,3}^{(L)} t_{3,1}^{(L)} - t_{2,3}^{(L)} t_{3,2}^{(L)} - t_{1,2}^{(L)} t_{2,3}^{(L)} t_{3,1}^{(L)} - t_{3,2}^{(L)} t_{2,1}^{(L)} t_{1,3}^{(L)}$. The (i, j) element of the matrix in Eq. (12) is the probability of being absorbed by state $S_j^{(L+1)}$, starting from the initial state $S_i^{(L)}$. For example, matrix element (1,3) is the probability of being absorbed by state $S_3^{(L+1)}$ at length $L + 1$ having started from state $S_1^{(L)}$ at length L . Substituting the above equation into Eq. (10) and defining state $S_1^{(L)}$ as the folded state of a domain (which we denote F), the probability of the domain being folded at length L is then

$$p_{F,final}^{(L)} = p_{F,final}^{(L-1)} \left[k_{2,3}^{(L)} [k_{3,F}^{(L)} + k_A^{(L)}] + [k_{2,F}^{(L)} + k_A^{(L+1)}] \times [k_{3,F}^{(L)} + k_{3,2}^{(L)} + k_A^{(L+1)}] \right] / D + p_{2,final}^{(L-1)} [k_{2,3}^{(L)} k_{3,F}^{(L)} + k_{2,1}^{(L)} [k_{3,F}^{(L)} + k_{3,2}^{(L)} + k_A^{(L+1)}]] / D + p_{3,final}^{(L-1)} [k_{2,F}^{(L)} \times [k_{3,F}^{(L)} + k_{3,2}^{(L)}] + k_{3,F}^{(L)} [k_{2,3}^{(L)} + k_A^{(L+1)}]] / D. \quad (13)$$

The parameter D in Eq. (13) was introduced for the sake of compactness and is defined as

$$D = k_{2,3}^{(L)} k_{3,F}^{(L)} + k_A^{(L+1)} [k_{2,3}^{(L)} + k_{3,F}^{(L)} + k_{3,2}^{(L)}] + k_A^{(L+1)^2} + k_{F,3}^{(L)} [k_{2,F}^{(L)} + k_{2,3}^{(L)} + k_{3,2}^{(L)} + k_A^{(L+1)}] + k_{2,F}^{(L)} [k_{3,F}^{(L)} + k_{3,2}^{(L)} + k_A^{(L+1)}] + k_{F,2}^{(L)} [k_{2,3}^{(L)} + k_{3,F}^{(L)} + k_{3,2}^{(L)} + k_A^{(L+1)}]. \quad (14)$$

The terms $p_{F,final}^{(L-1)}$, $p_{2,final}^{(L-1)}$, and $p_{3,final}^{(L-1)}$ are obtained by recursively using Eq. (10) $L - 1$ times and starting from the initial condition $\mathbf{p}_{final}^{(L=0)} = \{0, 0, 1\}$ of a fully unfolded state at $L = 0$ (assuming that state 3 is the fully unfolded state). By using this initial condition, the result from Eq. (13) is a curve describing the probability of being in the folded state as a function of nascent chain length. This illustrates how Eq. (10) provides a means to derive analytic expressions describing the influence of codon translation rates on arbitrarily complex co-translational folding mechanisms.

To test the accuracy of this approach in being able to predict the results from continuous translation simulations,

we ran Langevin dynamics simulations on a coarse-grained model of the synthesis of the MIT protein domain. The MIT domain folds into a three-helix bundle structure and can do so via a three-state, parallel pathway mechanism, as modeled by Eq. (13). The three states that can be populated by the MIT domain are the unfolded state, an intermediate state comprised of natively structured helices 1 and 2 (Fig. 4(a)), and the fully folded state (Fig. 4(b)). The coarse-grained model and simulation protocol have been described in Ref. 3. Briefly, ribosomal protein residues are represented by one interaction site and ribosomal RNA by up to four interaction sites for each nucleotide; electrostatic interactions are modeled using Debye-Huckel theory. Additional details on this model and the simulation can be found in Sec. II.

Two sets of simulations were carried out on this system. The first set was used as part of the process of making the predictions and the second set was used to test those predictions. In the first set of simulations, a series of arrested ribosomes at nascent chain lengths ranging from 65 to 120 residues were simulated. Arrested ribosomes do not undergo translation, i.e., $k_A^{(L)} = 0$. The MIT domain is 77 residues in length; therefore, this domain emerges fully from the narrow ribosome exit tunnel (which can contain around 30 nascent chain residues) at a nascent chain length of around 110 residues in its fusion construct with polyglycine (Fig. 4(a)). At each length, the rates of inter-conversion between the folded, unfolded, and intermediate states were measured from these simulations by using the method outlined in Sec. II, and the matrices $\mathbf{T}_{(L)}$ and $\mathbf{A}_{(L)}$ computed, keeping $k_A^{(L)}$ as the only free variable in these matrices. We inserted these $\mathbf{T}_{(L)}$ and $\mathbf{A}_{(L)}$ matrices into Eq. (10) and predicted how the MIT domain should behave during continuous translation for open reading frames that translate each codon position at rates approximately equal to $0.01 \cdot k_F^{(bulk)}$, $0.1 \cdot k_F^{(bulk)}$, and $k_F^{(bulk)}$, where $k_F^{(bulk)} = 81 \mu s^{-1}$ is the simulated folding rate of the MIT domain at 310 K in bulk solution, i.e., when no ribosome is present. We note that coarse-grained models and low-friction Langevin dynamics significantly speed up this protein's folding rate relative to experimentally observed values²⁸ but preserve realistic

FIG. 4. Equation (10) accurately predicts the effect of codon translation rates on the probability of populating different states during co-translational folding in coarse-grained Langevin dynamics simulations. (a) The 77-residue MIT domain consists of three helices and was fused to the N-terminus of an unstructured 43-residue polyglycine linker. (b) The domain forms a helix bundle in the folded state. (c) The synthesis of this nascent chain was simulated using a coarse-grained model, a simulation structure of which is shown in which the intermediate is present. (d) The populations of unfolded, intermediate, and folded states are shown, respectively, in black, red, and green at different translation rates. The predictions from Eq. (10) are shown as solid lines (their width corresponds to the 68% confidence interval). The predictions were made at $k_A^{(L)}$ rates equal to $0.01 \cdot k_F^{(bulk)}$ ((d), top panel), $0.1 \cdot k_F^{(bulk)}$ ((d), middle panel), and $k_F^{(bulk)}$ ((d), bottom panel). The continuous translation results from the coarse-grained model are shown as symbols at the various $k_A^{(L)}$ values; error bars correspond to the standard error about the mean (computed from 300 independent trajectories at each nascent chain length).

thermodynamic properties. While these predictions are for ORFs with uniform translation-rate profiles, we emphasize that our model is applicable to non-uniform profiles as well. We tested these predicted state curves against explicit simulations of continuous translation. In this second set of simulations, residues were stochastically attached to the C-terminus of the ribosome-bound nascent chain with the rates $k_A^{(L)}$ that were used in Eq. (10) to make the predictions (see Sec. II for further details).

The results from predictions, from Eq. (10), and from the continuous translation simulations are plotted in Fig. 4(d). We find that Eq. (10), combined with the computationally cheaper arrested ribosome simulation results, yields accurate predictions of the effect of codon translation rates on the probability of the MIT domain being in the folded, intermediate, and unfolded states (Fig. 4(d)). This indicates that the

model (Eq. (10)) captures the essential features present in co-translational folding and can very rapidly make accurate predictions about the influence of codon translation rates on arbitrarily complex co-translational folding mechanisms.

IV. DISCUSSION

We have introduced a model (Eq. (10)) that describes the influence of individual codon translation rates on co-translational folding mechanisms of arbitrary complexity, that is, mechanisms involving a nascent chain sampling any number of states during its synthesis. The rates used in Eq. (10) can be taken from any source: experiment, simulation, or theory. Thus, while we have focused in this paper on utilizing the inter-conversion rates reported from simulations, Eq. (10) can also be used in combination with experimentally measured rates.

By utilizing as input parameters the rates or inter-conversion between states obtained from arrested ribosome molecular dynamics simulations, the codon translation rates along the ORF are left as the only free parameters in Eq. (10). We are then able to compute the probability that a nascent chain segment is in a given conformational state as a function of the codon translation rates and nascent chain length. With this approach, an analysis of the effect of various translational-rate profiles on co-translational folding can be rapidly carried out. We tested predictions from this approach against results from molecular dynamics simulations of continuous translation of a protein that folds via parallel pathways. We found that Eq. (10) yielded highly accurate predictions at the three different global translation rates tested (Fig. 4(d)).

Equation (10) is not limited to co-translational folding mechanisms involving only three conformational states; it can be applied to proteins sampling thousands of states during their synthesis. Thus, Eq. (10) is a general solution for describing the influence of codon translation rates on co-translational folding mechanisms of arbitrary complexity.

For clarity, it is worth discussing what Eq. (10) is and what it is not, and how it contrasts with other approaches. Equation (10) mathematically represents the probability of taking a particular pathway out of the states that can be populated at nascent chain length L after performing an infinite random walk in that subset of Markov states. This is in the spirit of the approach Jacques Ninio introduced in his seminal work on calculating reaction rates from pathway probabilities.¹⁷ Equation (10) is not a master equation, which is commonly used in conjunction with Markov state analysis. Master equations calculate the probability of being in a state as a function of time, while here, we are interested in modeling the probability of being in a state as a function of nascent chain length. The steady-state master equation can calculate the probability of being in a state as a function of L ; however, those probabilities are averages over the entire time the ribosome dwells at codon L , whereas Eq. (10) is the state probability only at the instant before the next amino-acid is added to the nascent chain. Thus, the steady-state master equation and Eq. (10) are not equivalent. This is an important distinction because many approaches used to analyze molecular simulations couple a Markov state analysis with a master equation.^{23–26,37} Our method couples a Markov state analysis with a pathway-probability equation.

Two important questions regarding our approach are (i) what are its benefits to the experimental and simulation fields? and (ii) what are the practical challenges in implementing this model? For experimentalists, this model has the benefit of providing a means to utilize the results from arrested ribosomes, which are much easier to experimentally probe and to accurately predict what happens during continuous translation, which is more difficult to probe at high resolution. It is not currently possible to experimentally measure the inter-conversion rates for large proteins sampling a large number of states. However, as experimental methods advance, resolving ever finer spatial^{38,39} and temporal^{40,41} details of the translation process, the utility and applicability of Eq. (10) will increase concomitantly.

For the simulation field, combining Eq. (10) with arrested ribosome simulation results allows for the rapid exploration of various translation-rate profiles to examine how codon translation rates can govern the co-translational folding behavior. Even if one is interested only in a single translation-rate profile, it can still be a more efficient use of computer resources to utilize our approach than to perform the direct simulation of continuous synthesis as Eq. (10) can utilize a large number of short simulations that can be run simultaneously, rather than several long simulations of continuous synthesis. For example, if, on one central processing unit (CPU) core, it takes 20 days of wall time to simulate a trajectory of the continuous synthesis of a 200-residue protein, then a scientist who has simultaneous access to 200 cores can generate the arrested ribosome results at all 200 nascent chain lengths in ≈ 1 day of wall time and then use our method (see the supplementary material⁴⁷) to accurately predict the results from the continuous translation simulations. Thus, if a researcher has access to a large number of computers, our approach would allow them to produce accurate results in a shorter amount of wall time even for a single translation-rate profile. The time savings becomes greater the larger the protein or if more translation-rate profiles need to be examined.

A practical challenge in implementing this approach lies in identifying and defining the Markov states at a given nascent chain length. While for three-state systems (like the one studied here), it is fairly straightforward to identify the Markov states (see Sec. II); for a system with a larger number of states, the problem can become acute as the projection of the system dynamics onto lower dimensions can result in states being identified that exhibit non-Markovian behavior (e.g., a non-single exponential dwell time distribution). Extensive efforts by a number of research groups have resulted in automated Markov state identification algorithms,^{42–44} where in a large number of Markov states can be rapidly identified from simulations and their inter-conversion rates calculated. Utilizing these methods could provide a practical means to efficiently analyze arrested ribosome simulations sampling a large number of states and provide the rates needed for Eq. (10) to make accurate predictions.

In this study, we utilized Eq. (10) in combination with coarse-grained simulations of co-translational folding based on a Go-model force field,^{27,29} i.e., a force-field in which some parameters are system dependent and not transferable between molecules. This is in contrast to transferable force fields which

do not utilize native-state structural information. A Go force field was chosen because transferable coarse-grained force fields are currently unable to reliably fold proteins into their correct native structure. One limitation of the Go model we used is that any intermediates that are populated are native like, as opposed to forming non-native, misfolded tertiary structure. Thus, in this study, Eq. (10) has been proven accurate for a subset of states where all folding intermediates are assumed to be native like. This is not a serious limitation for this study for two reasons. There are very few experimental examples of isolated proteins adopting non-native, misfolded tertiary structures along their folding pathways *in vitro* and there are no examples of this occurring co-translationally. More important to note is that even if the co-translational formation of misfolded structures is in reality a common occurrence, the main conclusion of this study will still remain valid. Namely, that our model (Eq. (10)) can predict the influence of codon translation rates on co-translational folding mechanisms involving an arbitrarily large number of states. The reason this conclusion will remain valid, as detailed below, is that as long as the simulation results meet the assumptions of Eq. (10), then the predictive capability of our approach will not depend on whether intermediates are native-like or non-native in nature.

While coarse-grained simulations were used in this study, it is possible to use Eq. (10) with results from all-atom simulations. This is because the assumptions underlying the model are that (1) Markov states can be accurately identified for the system of interest and (2) that sufficient statistics on transitions between states are available to compute accurate interconversion rates. These criteria can be met, in principle, in both all-atom and coarse-grained simulations. The computational cost of all-atom simulations of co-translational folding, however, is currently prohibitive, making it difficult to obtain accurate rates. But as computer power continues to grow, our approach, in combination with the automated Markov state algorithms, should make it possible to efficiently use those results to predict how continuous translation simulations would behave under different translation-rate profiles.

The biological importance and influence of codon translation rates on the proper folding and functioning of nascent proteins are coming to the forefront in a number of fields including molecular and cellular biology,⁸ cancer biology,⁴⁵ personalized medicine,¹⁴ and biotechnology.⁴⁶ What is currently lacking in these fields, however, is a theoretical framework to understand, model, and predict the influence of codon translation rates on these processes. We believe Eq. (10) provides an integral part of that framework as it is a methodology to integrate information, such as folding mechanisms and rates for a specific nascent protein, and makes predictions about the consequences of changing individual codon translation rates for co-translational folding and misfolding. This method enables the efficient study of the co-translational folding of large proteins and opens up the possibility of proteome-wide molecular dynamics studies of co-translational folding.²

ACKNOWLEDGMENTS

E.P.O. thanks Will Noid and Ajeet Sharma for valuable comments on the manuscript, Steven Fillini for providing extra

disk-storage space on Biowulf for the simulations, and Robert Best and David De Sancho for very helpful conversations regarding Markov State modeling. This study utilized the high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health, Bethesda, MD (<http://biowulf.nih.gov>).

- ¹G. M. Whitesides and B. Grzybowski, "Self-assembly at all scales," *Science* **295**, 2418–2421 (2002).
- ²P. Ciryam, R. I. Morimoto, M. Vendruscolo, C. M. Dobson, and E. P. O'Brien, "In vivo translation rates can substantially delay the cotranslational folding of the *Escherichia coli* cytosolic proteome," *Proc. Natl. Acad. Sci. U. S. A.* **110**, E132–E140 (2013).
- ³E. P. O'Brien, M. Vendruscolo, and C. M. Dobson, "Prediction of variable translation rate effects on cotranslational protein folding," *Nat. Commun.* **3**, 868 (2012).
- ⁴A. A. Komar, T. Lesnik, and C. Reiss, "Synonymous codon substitutions affect ribosome traffic and protein folding during in vitro translation," *FEBS Lett.* **462**, 387–391 (1999).
- ⁵D. A. Nissley and E. P. O'Brien, "Timing is everything: Unifying codon translation rates and nascent proteome behavior," *J. Am. Chem. Soc.* **136**, 17892–17898 (2014).
- ⁶A. A. Komar, "A pause for thought along the co-translational folding pathway," *Trends Biochem. Sci.* **34**, 16–24 (2009).
- ⁷G. Zhang, M. Hubalewska, and Z. Ignatova, "Transient ribosomal attenuation coordinates protein synthesis and co-translational folding," *Nat. Struct. Mol. Biol.* **16**, 274–280 (2009).
- ⁸P. S. Spencer, E. Siller, J. F. Anderson, and J. M. Barral, "Silent substitutions predictably alter translation elongation rates and protein folding efficiencies," *J. Mol. Biol.* **422**, 328–335 (2012).
- ⁹E. Siller, D. C. DeZwaan, J. F. Anderson, B. C. Freeman, and J. M. Barral, "Slowing bacterial translation speed enhances eukaryotic protein folding efficiency," *J. Mol. Biol.* **396**, 1310–1318 (2010).
- ¹⁰T. F. Clarke IV and P. L. Clark, "Rare codons cluster," *PLoS One* **3**, e3412 (2008).
- ¹¹R. Saunders and C. M. Deane, "Synonymous codon usage influences the local protein structure observed," *Nucleic Acids Res.* **38**, 6719–6728 (2010).
- ¹²T. A. Thanaraj and P. Argos, "Ribosome-mediated translational pause and protein domain organization," *Protein Sci.* **5**, 1594–1612 (1996).
- ¹³S. Pechmann and J. Frydman, "Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding," *Nat. Struct. Mol. Biol.* **20**, 237–243 (2013).
- ¹⁴C. Kimchi-Sarfaty *et al.*, "A 'silent' polymorphism in the MDR1 gene changes substrate specificity," *Science* **315**, 525–528 (2007).
- ¹⁵M. Zhou *et al.*, "Non-optimal codon usage affects expression, structure and function of clock protein FRQ," *Nature* **495**, 111–115 (2013).
- ¹⁶P. Cortazzo *et al.*, "Silent mutations affect in vivo protein folding in *Escherichia coli*," *Biochem. Biophys. Res. Commun.* **293**, 537–541 (2002).
- ¹⁷J. Ninio, "Alternative to the steady-state method: Derivation of reaction rates from first-passage times and pathway probabilities," *Proc. Natl. Acad. Sci. U. S. A.* **84**, 663–667 (1987).
- ¹⁸E. P. O'Brien, M. Vendruscolo, and C. M. Dobson, "Kinetic modelling indicates that fast-translating codons can coordinate cotranslational protein folding by avoiding misfolded intermediates," *Nat. Commun.* **5**, 2988 (2014).
- ¹⁹E. P. O'Brien, J. Christodoulou, M. Vendruscolo, and C. M. Dobson, "Trigger factor slows co-translational folding through kinetic trapping while sterically protecting the nascent chain from aberrant cytosolic interactions," *J. Am. Chem. Soc.* **134**, 10920–10932 (2012).
- ²⁰E. P. O'Brien, J. Christodoulou, M. Vendruscolo, and C. M. Dobson, "New scenarios of protein folding can occur on the ribosome," *J. Am. Chem. Soc.* **133**, 513–526 (2011).
- ²¹S. E. Radford, C. M. Dobson, and P. A. Evans, "The folding of hen lysozyme involves partially structured intermediates and multiple pathways," *Nature* **358**, 302–307 (1992).
- ²²P. L. Clark and J. King, "A newly synthesized, ribosome-bound polypeptide chain adopts conformations dissimilar from early in vitro refolding intermediates," *J. Biol. Chem.* **276**, 25411–25420 (2001).
- ²³N. V. Buchete and G. Hummer, "Coarse master equations for peptide folding dynamics," *J. Phys. Chem. B* **112**, 6057–6069 (2008).
- ²⁴C. Schütte, F. Noé, J. Lu, M. Sarich, and E. Vanden-Eijnden, "Markov state models based on milestone," *J. Chem. Phys.* **134**, 204105 (2011).
- ²⁵F. Noé, C. Schütte, E. Vanden-Eijnden, L. Reich, and T. R. Weikl, "Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations," *Proc. Natl. Acad. Sci. U. S. A.* **106**, 19011–19016 (2009).
- ²⁶J. H. Prinz *et al.*, "Markov models of molecular kinetics: Generation and validation," *J. Chem. Phys.* **134**, 174105 (2011).
- ²⁷J. N. Onuchic and P. G. Wolynes, "Theory of protein folding," *Curr. Opin. Struct. Biol.* **14**, 70–75 (2004).
- ²⁸D. K. Klimov and D. Thirumalai, "Viscosity dependence of the folding rates of proteins," *Phys. Rev. Lett.* **79**, 317–320 (1997).
- ²⁹Y. Ueda, H. Taketomi, and N. Gō, "Studies on protein folding, unfolding, and fluctuations by computer simulation. II. A. Three-dimensional lattice model of lysozyme," *Biopolymers* **17**, 1531–1548 (1978).
- ³⁰E. P. O'Brien, G. Ziv, G. Haran, B. R. Brooks, and D. Thirumalai, "Effects of denaturants and osmolytes on proteins are accurately predicted by the molecular transfer model," *Proc. Natl. Acad. Sci. U. S. A.* **105**, 13403–13408 (2008).
- ³¹S. Piana, K. Lindorff-Larsen, and D. E. Shaw, "Atomic-level description of ubiquitin folding," *Proc. Natl. Acad. Sci. U. S. A.* **110**, 5915–5920 (2013).
- ³²R. B. Best, G. Hummer, and W. A. Eaton, "Native contacts determine protein folding mechanisms in atomistic simulations," *Proc. Natl. Acad. Sci. U. S. A.* **110**, 17874–17879 (2013).
- ³³D. De Sancho and V. Muñoz, "Integrated prediction of protein folding and unfolding rates from only size and structural class," *Phys. Chem. Chem. Phys.* **13**, 17030–17043 (2011).
- ³⁴B. R. Brooks *et al.*, "CHARMM: The biomolecular simulation program," *J. Comput. Chem.* **30**, 1545–1614 (2009).
- ³⁵W. Feller, "An introduction to probability theory and its applications," *Technometrics* **2**, 509 (1968).
- ³⁶P. G. Wolynes, J. N. Onuchic, and D. Thirumalai, "Navigating the folding routes," *Science* **267**, 1619–1620 (1995).
- ³⁷V. A. Voelz, G. R. Bowman, K. Beauchamp, and V. S. Pande, "Molecular simulation of *ab initio* protein folding for a millisecond folder NTL9(1-39)," *J. Am. Chem. Soc.* **132**, 1526–1528 (2010).
- ³⁸Z. K. Majumdar, R. Hickerson, H. F. Noller, and R. M. Clegg, "Measurements of internal distance changes of the 30 S ribosome using FRET with multiple donor-acceptor pairs: Quantitative spectroscopic methods," *J. Mol. Biol.* **351**, 1123–1145 (2005).
- ³⁹S. T. D. Hsu *et al.*, "Structure and dynamics of a ribosome-bound nascent chain by NMR spectroscopy," *Proc. Natl. Acad. Sci. U. S. A.* **104**, 16516–16521 (2007).
- ⁴⁰G. W. Li, E. Oh, and J. S. Weissman, "The anti-Shine–Dalgarno sequence drives translational pausing and codon choice in bacteria," *Nature* **484**, 538–541 (2012).
- ⁴¹A. Tsai *et al.*, "Heterogeneous pathways and timing of factor departure during translation initiation," *Nature* **487**, 390–393 (2012).
- ⁴²J. D. Chodera, N. Singhal, V. S. Pande, K. A. Dill, and W. C. Swope, "Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics," *J. Chem. Phys.* **126**, 155101 (2007).
- ⁴³M. Senne, B. Trendelkamp-Schroer, A. S. J. S. Mey, C. Schütte, and F. Noé, "EMMA: A software package for Markov model building and analysis," *J. Chem. Theory Comput.* **8**, 2223–2238 (2012).
- ⁴⁴K. A. Beauchamp *et al.*, "MSMBuilder2: Modeling conformational dynamics on the picosecond to millisecond scale," *J. Chem. Theory Comput.* **7**, 3412–3419 (2011).
- ⁴⁵J. J. Gartner *et al.*, "Whole-genome sequencing identifies a recurrent functional synonymous mutation in melanoma," *Proc. Natl. Acad. Sci. U. S. A.* **110**, 13481–13486 (2013).
- ⁴⁶E. Angov, C. J. Hillier, R. L. Kincaid, and J. A. Lyon, "Heterologous protein expression is enhanced by harmonizing the codon usage frequencies of the target gene with those of the expression host," *PLoS One* **3**, e2189 (2008).
- ⁴⁷See supplementary material at <http://dx.doi.org/10.1063/1.4916914> for implementation of Eq. (10) in Matlab code and a Charmm script for simulating continuous translation of a nascent protein.