JACS
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY
Article
pubs.acs.org/JACS

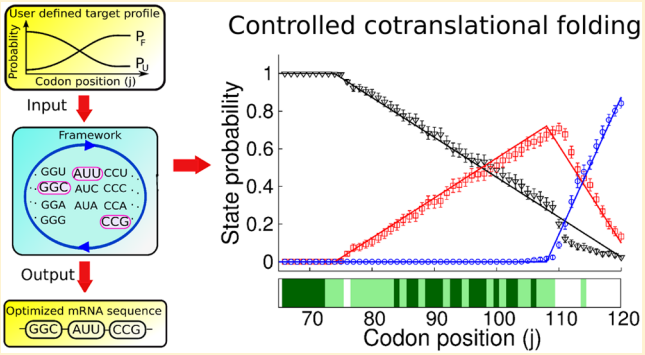
Physical Origins of Codon Positions That Strongly Influence Cotranslational Folding: A Framework for Controlling Nascent-Protein Folding

Ajeet K. Sharma,[†] Bernd Bukau,[‡] and Edward P. O'Brien*[,†]

[†]Department of Chemistry, Pennsylvania State University, University Park, Pennsylvania 16802, United States
[‡]Center for Molecular Biology of the University of Heidelberg (ZMBH), DKFZ-ZMBH Alliance, Im Neuenheimer Feld 282, Heidelberg D-69120, Germany

Supporting Information

ABSTRACT: An emerging paradigm in the field of in vivo protein biophysics is that nascent-protein behavior is a type of nonequilibrium phenomenon, where translation-elongation kinetics can be more important in determining nascent-protein behavior than the thermodynamic properties of the protein. Synonymous codon substitutions, which change the translation rate at select codon positions along a transcript, have been shown to alter cotranslational protein folding, suggesting that evolution may have shaped synonymous codon usage in the genomes of organisms in part to increase the amount of folded and functional nascent protein. Here, we develop a Monte Carlo-master-equation method that allows for the control of nascent-chain folding during translation through the rational design of mRNA sequences to guide the cotranslational folding process. We test this framework using coarse-grained molecular dynamics simulations and find it provides optimal mRNA sequences to control the simulated, cotranslational folding of a protein in a user-prescribed manner. With this approach we discover that some codon positions in a transcript can have a much greater impact on nascent-protein folding than others because they tend to be positions where the nascent chain populates states that are far from equilibrium, as well as being dependent on a complex ratio of time scales. As a consequence, different cotranslational profiles of the same protein can have different critical codon positions and different numbers of synonymous mRNA sequences that encode for them. These findings explain that there is a fundamental connection between the nonequilibrium nature of cotranslational processes, nascent-protein behavior, and synonymous codon usage.



INTRODUCTION

Precise timing is often required for the accuracy and efficiency of the numerous cotranslational processes acting on a nascent protein, which help it to attain its functionality.[1] Therefore, the ability of a nascent-protein molecule to form its native structure and acquire its biological function can be influenced by the rate at which individual codon positions in an mRNA molecule are translated by the ribosome.[2−4] Synthesize a signal sequence too fast and signal recognition particle (SRP) may not be able to bind to it, resulting in a decreased probability of successful cotranslational translocation of the nascent protein through the SEC translocon.[5,6] Change the translation rate at a critical codon position and a protein will switch from cotranslational folding to misfolding, resulting in an increased population of insoluble[7] or soluble, but nonfunctional, protein.[8] For these reasons, evolutionary selection pressures have likely shaped codon usage bias in organisms in part to maximize the efficiency of cotranslational processes by tuning the translation-rate profile along the coding sequence through synonymous codon mutations[9] (Figure 1a).

The physical rules governing why changes in translation rate at select codon positions have a significant effect on nascent-protein folding though changes at other positions have little to no effect are unknown. Hu and co-workers[10] created 342 synonymous mRNA sequence variants of the human anti-IgE antibody and found that the transcripts produced protein of varying solubility and functionality. Some synonymous mutations had no effect on these properties, while others decreased or increased the protein's specific activity by as much as 10-fold. These results support the idea that synonymous mutations at different locations can alter the likelihood of cotranslational folding to varying degrees.

Cells may influence cotranslational behavior by altering the translation-rate profile along an mRNA's coding sequence.[11−15] This suggests that it should be possible to rationally design mRNA sequences using synonymous codons to quantitatively control nascent-protein behavior at all codon positions during

Received: August 3, 2015
Published: December 30, 2015

DOI: 10.1021/jacs.5b08145
J. Am. Chem. Soc. 2016, 138, 1180−1195

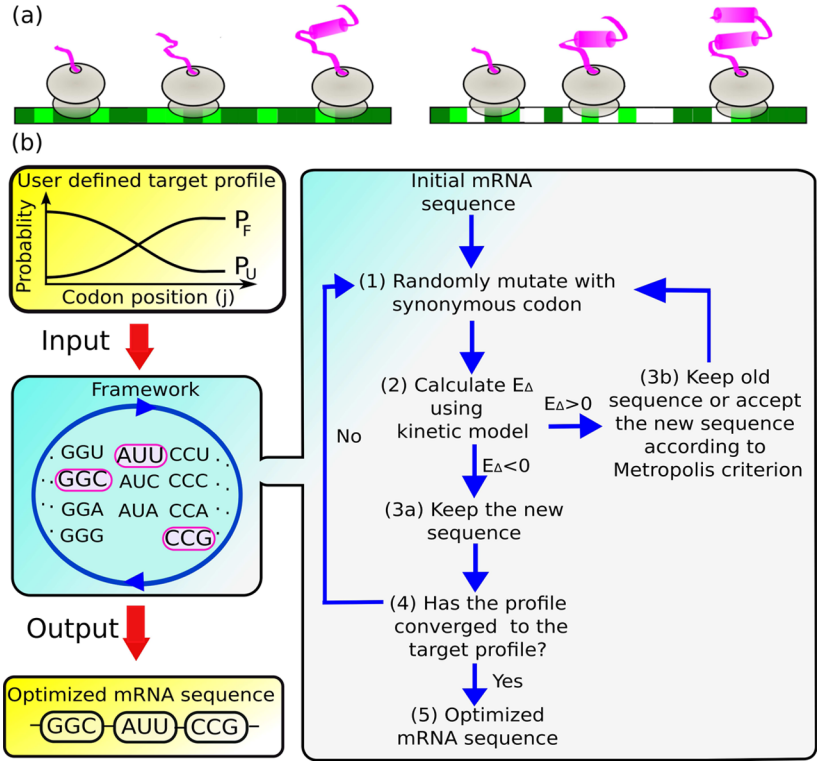


Figure 1. A framework for controlling cotranslational folding through the rational design of mRNA sequences. (a) An illustration of changes in cotranslational folding due to changes in the translation-rate profile induced by synonymous codon substitutions. Dark-green, light-green, and white bars represent the fast-, medium-, and slow-translating codons in the mRNA sequence that encode for the same protein. (b) Flowchart illustrating the steps involved in our Monte Carlo-master-equation-based framework for designing mRNA sequences. A user-defined target profile and rate matrices are provided as an input to our framework, which uses the steps listed in the right to find the mRNA sequence that best reproduce the target profile.

synthesis. Empirical codon optimization strategies for heterologous protein expression often attempt to maximize protein production without regard for the quality of the protein produced in terms of its solubility, folding, and functionality.[16,17] Furthermore, such optimization approaches do not explicitly account for the profound effect that translation-elongation rates can have on nascent-protein behavior.

Here, we focus on the process of cotranslational folding and show that it is possible to rapidly design mRNA sequences to quantitatively control nascent-protein folding at each step during translation elongation given knowledge of the synonymous codon translation rates, which can significantly vary,[18,19] and the rates of interconversion between states of the nascent protein. As a proof of principle, we test the predictions from this framework against coarse-grained molecular dynamics simulations of cotranslational folding in which codons can adopt one of three possible translation rates. We also test our framework in the situation in which each of the 61 sense codons adopt their own unique translation rate. We then explore the rules governing the codon-position-dependent impact that changes in translation rate can have on cotranslational folding.

METHODS

Master Equation for Calculating the Cotranslational Profile of a Protein That Can Populate *N*-States. Implementation of our framework (Figure 1b) requires the accurate prediction of the effect that a change in a codon position's translation rate will have on a protein's cotranslational profile. Therefore, we derived an analytical expression for the steady-state probability that a protein will be in any one of N possible states at each nascent chain length during its synthesis. The probability that a nascent chain of length j is in state l at time t is denoted by $P_l(j,t)$, where $l = \{1, 2, ..., N\}$. The master equation governing the time evolution of $P_l(j,t)$ can be written as

$$\frac{dP_l(j,t)}{dt} = \sum_{i=1,i\neq l}^{N} P_i(j,t)k_{il}(j) - \sum_{i=1,i\neq l}^{N} P_l(j,t)k_{li}(j) + P_l(j-1,t)\omega_l(j) - P_l(j,t)\omega_l(j+1) \quad (1)$$

where $k_{il}(j)$ is the rate at which state i interconverts with state l at codon position j and $\omega_l(j)$ is the rate at which codon $j-1$ is translated when the nascent chain is in state l. The first and second terms of the right-hand side of eq 1 determine the gain and loss in $P_l(j,t)$ arising from the folding kinetics of the protein domain; the third and fourth terms are the gain and loss contributions from translation-elongation kinetics. Equation 1 can also be written as a matrix equation

$$\frac{d\mathbf{P}(j,t)}{dt} = \mathbf{M}(j)\mathbf{P}(j,t) - \mathbf{T}(j)\mathbf{P}(j-1,t) \quad (2)$$

where $\mathbf{P}(j,t)$ is a column vector of the state probabilities

$$\mathbf{P}(j,t) = \begin{bmatrix} P_1(j,t) \\ P_2(j,t) \\ \vdots \\ P_N(j,t) \end{bmatrix}$$

and

DOI: 10.1021/jacs.5b08145
J. Am. Chem. Soc. 2016, 138, 1180−1195

$$\mathbf{M}(j) = \begin{bmatrix} -(\omega_1(j+1) + \sum_{i=2}^{N} k_{1i}(j)) & k_{21}(j) & \cdots & k_{N1}(j) \\ k_{12}(j) & -(\omega_2(j+1) + \sum_{i=1,i\neq 2}^{N} k_{2i}(j)) & \cdots & k_{N2}(j) \\ \vdots & \cdots & \ddots & \vdots \\ k_{1N}(j) & \cdots & \cdots & -(\omega_N(j+1) + \sum_{i=1}^{N-1} k_{Ni}(j)) \end{bmatrix}$$

and

$$\mathbf{T}(j) = \begin{bmatrix} -\omega_1(j) & 0 & \cdots & 0 \\ 0 & -\omega_2(j) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -\omega_N(j) \end{bmatrix}$$

$\mathbf{M}(j)$ and $\mathbf{T}(j)$ are $N \times N$ size transition matrices describing the transitions between states in the nascent protein. Solving eq 2 under steady-state conditions $\left(\text{i.e.,}\frac{d\mathbf{P}(j,t)}{dt} = 0\right)$ results in the recursive relation

$$\mathbf{P}(j) = \mathbf{M}(j)^{-1}\mathbf{T}(j)\mathbf{P}(j-1) \tag{3}$$

That is, the steady-state probabilities ($\mathbf{P}(j)$) at codon j depend on what happened at all earlier codon positions ($\mathbf{P}(j-1)$). The elements of $\mathbf{P}(j)$ are denoted as $P_1(j)$, $P_2(j)$,...,$P_{N-1}(j)$ and $P_N(j)$. Translation, in this model, involves an open system; therefore, the sum of the probabilities of populating different states at codon position j is not equal to one. For this reason, we then normalized these probabilities at each codon position by dividing the term $\sum_{i=1}^{N} P_i(j)$.

Coarse-Grained Model. The coarse-grained simulation model[20−22] of the ribosome-nascent chain complex includes the large ribosomal subunit and the nascent-chain, which, at full length, is composed of a 43-residue polyglycine linker covalently attached to the C terminus of the single-domain MIT (microtubule interacting and trafficking) protein found in humans. The MIT domain is 77 residues in length and adopts an antiparallel three-helix bundle structure in its native state.[23] We simulated this protein's behavior tethered to *E. coli*'s 50S ribosomal subunit during both continuous and arrested translation. In the coarse-grained model, amino acids were represented by a single interaction site at their C_α positions. A $+1e$ charge was assigned to interaction sites representing lysine and arginine residues, and a $-1e$ charge was assigned to the sites representing glutamine and aspartate residues. Other residues were assigned zero charge. Purine and pyrimidine nucleotides in the ribosomal *RNA* were represented by three and four interaction sites,[21] respectively, that represent the ribose ring, the phosphate group, and each conjugated ring in the base. The coarse-grained interaction sites were positioned at the geometric center of each of these groups. The interaction site representing the phosphate group was assigned a charge of $-1e$.

The coarse-grained force field is the sum of five different energy terms

$$E_{tot} = E_{bond} + E_{angle} + E_{dihedral} + E_{elec} + E_{LJ} \tag{4}$$

The first four energy terms in eq 4 account, respectively, for bond energy, bond angle, dihedral angle, and pairwise electrostatic interactions, where $E_{bond} = \sum_i K_b(r_i - r_0)^2$, $E_{angle} = \sum_i \exp(-\gamma K_\alpha(\theta_i - \theta_\alpha)^2 + \epsilon_\alpha) + \exp(-\gamma K_\beta(\theta_i - \theta_\beta)^2)$,[24] $E_{dihedral} = \sum_{i,j} K_{\psi,ij}(1 + \cos(j\psi_i - \delta_{ij}))$, and $E_{elec} = \sum_{i,j} \frac{q_i q_j}{4\pi\epsilon_0\epsilon_r r_{ij}} \exp\left(-\frac{r_{ij}}{l_D}\right)$ are fully transferable between different protein and *RNA* molecules. K_b is the bond force constant; $(r_i - r_0)$ is the distance of an interaction site from its equilibrium position; θ_α and θ_β are the location of two minima on the bond-angle potential energy surface, and their angle force constants are K_α and K_β, respectively; ϵ_α is used to tune the relative balance between these two bond-angle energy minima; K_ψ is the dihedral force constant; j is the multiplicity of the function; ψ is the dihedral angle; δ_{ij} is the phase shift. In the simulations, we used[22,25] a Debye length l_D = 10 Å and dielectric constant ϵ_r = 78.5.

The last term in eq 4 incorporates structure-dependent van der Waals interactions. We utilized Gō's approach[26] that treats the native interactions as attractive and non-native interactions as repulsive. A modified Lennard-Jones energy term,[27] which accounts for desolvation barriers, was used and is defined as

$$E_{LJ} = \sum_{i,j} \epsilon_{ij}\left[13\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - 18\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{10} + 4\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] \tag{5}$$

The Lennard−Jones well-depth ϵ_{ij} between interaction sites i and j that form native contacts in the MIT protein was set equal to the value of the Bentancourt−Thirumalai statistical potential[28] and scaled by a multiplicative factor to achieve a realistic native-state stability for the MIT protein. All other Lennard−Jones interactions were treated as effectively repulsive by setting their ϵ_{ij} = 0.000132 kcal/mol, as in the Karanicolas−Brooks model.[27] The collision diameter σ_{ij} for the two interaction sites of MIT protein is their distance in the crystal structure divided by $2^{1/6}$. σ_{ij}'s for the intralinker interactions (see Supporting Information) were calculated as defined in ref 22. We used $\sigma_{ij} = (\sigma_i + \sigma_j)/2$ for the interaction between the polyglycine and the MIT domain, the ribosome and the MIT domain, and the ribosome and the linker. The values of σ_i and other parameters are reported in the Supporting Information. In the simulations, the ribosome is held rigid. Therefore, there are no terms in the force field to represent the bonded interactions within ribosome 50S subunit (ribosomal *RNA* and ribosomal protein).

Langevin Dynamics Simulations. We used Charmm[29] version *c*35*b*5 to run Langevin dynamics simulations of the ribosome-nascent chain complex. An integration time step of 0.015 ps, collision frequency of 0.05 ps^{-1}, and system temperature of 310 K were used. For the continuous translation simulations, 1200 independent trajectories were run for the fast and medium mRNA sequences, while 720 trajectories were simulated for all other cases. For the arrested ribosome simulations, we ran 20 independent Langevin dynamics simulations of the arrested ribosome-nascent chain complex at each codon position between 69 and 89 codons (inclusive) and 8 independent trajectories at all other lengths. Different initial velocity distributions were used to initiate each trajectory. The ribosome was held rigid during the simulations by using the "cons fix" command of the constraint module in Charmm. This constraint has no significant effect on the thermodynamics and kinetics of cotranslational protein folding because the ribosome exit tunnel does not exhibit any large-scale fluctuations.[30] System configurations were saved every 50 time steps in the simulations of continuous translation of the slow mRNA sequence, every 150 time steps for the optimized mRNA sequences, and every 500 time steps for all other simulations. The MIT protein cannot populate the intermediate and folded states at codon positions 1−65. The starting structure for all simulations therefore consisted of a ribosome-nascent chain complex containing the 65 N-terminal residues of the protein. We used the procedure described in ref 20 to stochastically add amino acids to the nascent chain. The dwell time of the ribosome at each codon position was inverse of its codon translation rate. Because the Langevin dynamics simulations were performed in the low-friction regime, folding kinetics are faster in comparison to experimental values. Therefore, to keep a reasonable ratio of the time scales of folding and translation we increased the value of the translation rates to 66.4, 664.0, and 6640.0 μs^{-1} for the slow-, medium-, and fast-translating codons, respectively.

Identification of Markov States. In order to identify the three different states MIT can populate during the simulations we used

DOI: 10.1021/jacs.5b08145
J. Am. Chem. Soc. 2016, 138, 1180−1195

Markov state definitions that have been previously published.[20] Specifically, we used two order parameters, the fraction of native contacts between helices 1 and 2 (Q_{1-2}) and between helix 3 and helices 1 and 2 (Q_{12-3}). $Q_{1-2} = \frac{n_{12}}{n_{12}^c}$ and $Q_{12-3} = \frac{n_{12-3}}{n_{12-3}^c}$. n_{12} and n_{12-3} are the number of native contacts formed by helix 1 with 2 and helix 3 with helices 1 and 2 in a given ribosome-nascent chain conformation, while n_{12}^c and n_{12-3}^c are the number of native contacts between these structural elements in the crystal structure. The ribosome simulations involve situations in which the MIT domain might not be fully synthesized. In these cases the calculations of Q_{1-2} and Q_{12-3} are unchanged; n_{12}^c and n_{12-3}^c retain their constant values, and it is only the n_{12} and n_{12-3} terms which will decrease compared to the fully synthesized MIT situation. The free energy surface of the MIT protein as a function of Q_{1-2} and Q_{12-3} displays three basins which correspond to the unfolded, intermediate, and folded states of the MIT protein (Figure S1). A conformation is identified as being folded if $Q_{1-2} > 0.85$ and $Q_{12-3} > 0.85$, as being in the intermediate state if $0.95 > Q_{1-2} > 0.75$ and $Q_{12-3} < 0.05$, and as being unfolded if $Q_{1-2} < 0.05$ and $Q_{12-3} < 0.05$ (Figure S1). The conformations not belonging to any of these states are in the transition region. If a simulation trajectory enters the transition region then the most recently visited state is assigned to these conformations. Studies have shown this type of "core-based" partitioning of the free energy surface can accurately capture the essential kinetics in molecular dynamics simulations.[31,32]

Calculation of Rate Matrices. The master equation calculation requires knowledge of the $\mathbf{M}(j)$ rate matrices. We developed a method to calculate these rates from the time series of different states acquired by the protein during the coarse-grained simulations. Suppose a protein can populate N different states at nascent chain length j. Then assuming transitions between states are Markovian, the first passage time distribution of transitioning out of state i to any other state will decay exponentially as

$$f(i) = \exp(-k(i)t) \tag{6}$$

where

$$k(i) = \sum_{l=1,l\neq i}^{N} k_{il} \tag{7}$$

In eq 7, k_{il} is the rate of transitioning from state i to l. Therefore, $k(i)$ is the sum of the rates leading to all topologically connected states from state i (e.g., Figure 2b). Moreover, if $n(i \rightarrow l)$ is the total number of transitions from state i to l observed during the simulations then

$$\frac{n(i \rightarrow l)}{\sum_{m=1,m\neq i}^{N} n(i \rightarrow m)} = \frac{k_{il}}{\sum_{m=1,m\neq i}^{N} k_{im}} \tag{8}$$

For an N-state protein there are $N - 2$ independent equations of the same form as eq 8. Solving eqs 7 and 8 for k_{il} determines the $N - 1$ transition rates leading to any state from state i.

In order to calculate the rates for MIT we utilized the data from the arrested ribosome simulations. For each of the simulation trajectories we produced the time series of Markov states that were then used to numerically calculate the $f(i)$s and $n(i \rightarrow k)$s. We then performed a least-squares fit of the $f(i)$s to a single-exponential function using Gnuplot 4.2[33] and extracted the $k(i)s$. The transition rates between states were then calculated by solving eqs 7 and 8. These rates define the $\mathbf{M}(j)$s as indicated in eq 2. Note that this method differs from eq 5 utilized in ref 20. In that method, interconversion rates between states were directly calculated from the number of transitions per unit time, whereas in the current approach a numerical fitting to an exponential function is carried out to the dwell-time distribution in state i.

Estimating the Statistical Error of the Steady-State Master Equation Predictions. The rate matrices $\mathbf{M}(j)$ were used in eq 3 to make predictions about the influence of codon translation rates on the cotranslational profile of MIT. To calculate the 95% confidence intervals associated with these predictions we used the following procedure that relies on the construction of rate matrices that preserve the transition probabilities of the original trajectories. First, for each of the X_j independent arrested-ribosome simulations run at codon position j (where X_j = 8 or 20 depending on the codon position), the Markov state time series was constructed. Second, from each of these X_j Markov state time series the probability of transitioning from state i to state k during time interval Δt was calculated as

$$P_{ik}^{[X_j]}(\Delta t) = \frac{n^{[X_j]}(i \rightarrow k|\Delta t)}{\sum_{m=1}^{3} n^{[X_j]}(i \rightarrow m|\Delta t)} \tag{9}$$

where $n^{[X_j]}(i \rightarrow k|\Delta t)$ is the total number of transitions between state i and k during trajectory X_j given that you are sampling the time series of states at a time interval of Δt.[34] The summation in the denominator of eq 9 is over the three different states that the MIT protein can populate. Thus, for the MIT protein, we have 9 different $P_{ik}^{[X_j]}(\Delta t)$ values for each trajectory. Third, these 9 values from trajectory X_j were used to construct 10^4 virtual trajectories through Markov-state space that have the same duration as the original trajectory. The construction of a virtual trajectory is done in the following manner. At time $t = 0$ the system is assumed to be unfolded. At $t = \Delta t$ the probability that the system is now in state i is calculated by selecting a random number between 0 and 1 and determining which region of the number line the random number falls in, where different regions correspond to different transition probabilities $P_{Uk}^{[X_j]}(\Delta t)$ from U to k. For example, if the random number fell in the U → F region the virtual time series for

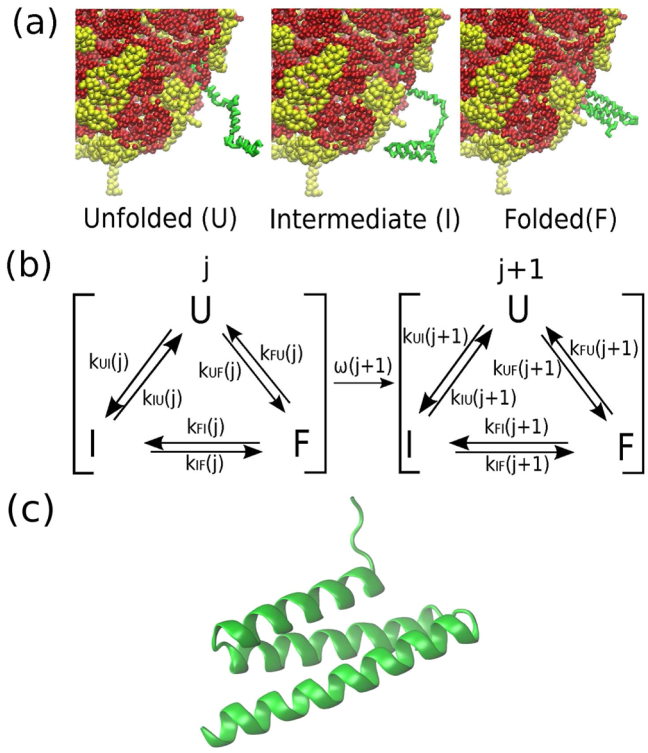


Figure 2. The three-helix bundle MIT protein cotranslationally folds through a 3-state, parallel folding mechanism.[20] (a) Structures from coarse-grained simulations of the ribosome-nascent chain complexes are shown with the MIT protein in the unfolded, intermediate, and folded states. Ribosomal RNA and proteins are shown in red and yellow, respectively, while the MIT protein is shown in green. (b) Parallel cotranslational folding reaction scheme. Codons in the mRNA sequence of MIT protein are sequentially labeled by an integer index j. Incorporation of an amino acid into the nascent chain shifts the ribosome-nascent chain complex to the next codon with rate $\omega(j + 1)$. $k_{UF}(j)$ and $k_{FU}(j)$ are the rates of transition from U to F and F to U, $k_{UI}(j)$ and $k_{IU}(j)$ are the rates of transition from U to I and I to U, and $k_{IF}(j)$ and $k_{FI}(j)$ are the rates of transition from I to F and F to I at codon position j. (c) Folded state of the MIT domain forms a three-helix bundle (PDB ID: 1YXR).

DOI: 10.1021/jacs.5b08145
J. Am. Chem. Soc. 2016, 138, 1180−1195

the first two time points would be recorded as {U,F}; if it fell in the U → U region it would be recorded as {U,U}. At each subsequent time interval this stochastic selection of transitions is repeated until the duration of the trajectory is the same as the original. Fourthly, the third step was repeated for all X_j trajectories. Thus, at this stage of the procedure there are 10^4 sets containing either 8 or 20 different virtual trajectories at codon position j. From each of these sets a rate matrix $\mathbf{M}(j)$ was computed, resulting in 10^4 rate matrices at each codon position. Each rate matrix was then used in eq 3 to independently predict the cotranslational profile of the MIT protein, resulting in 10^4 profiles. The 95% confidence interval of the $P_i(j)$ predicted from eq 3 was calculated as the standard deviation in the 10^4 probabilities of being in state i multiplied by 1.95.[35]

Calculating the Cotranslational Profiles and Their Errors from Coarse-Grained Simulations. From the continuous synthesis simulations, MIT's steady-state cotranslational profiles were calculated as

$$P_i(j) = \frac{\sum_{k=1}^{N_{\mathrm{Traj}}}\left[\sum_{l=1}^{N_{\mathrm{Frames}}(k,j)} \delta_{(i,j,l,k)}/N_{\mathrm{Frames}}(k,j)\right]}{N_{\mathrm{Traj}}} = \frac{\sum_{k=1}^{N_{\mathrm{Traj}}} P_{i,k}(j)}{N_{\mathrm{Traj}}} \quad (10)$$

where $P_i(j)$ is the steady-state probability of being in state i at codon position j and $\delta_{(i,j,l,k)}$ is the Kronecker delta that equals 1 when the system is in state i at codon j in frame l of trajectory k. The summations are over the N_{Traj} independent synthesis trajectories for a given translation-rate profile and the $N_{\mathrm{Frames}}(k,j)$ saved during simulation trajectory k at codon j. In eq 10, $P_{i,k}(j)$ is the probability of being in state i at codon j in trajectory k.

To estimate the statistical errors associated with the $P_i(j)$ values calculated by using eq 10 we used the Bootstrapping method.[36] Specifically, there were either 720 or 1200 $P_{i,k}(j)$ values for a given translation-rate profile. 10^4 bootstrapping cycles were applied to these 720 or 1200 member sets to calculate the 95% confidence interval about $P_i(j)$.

Design of mRNA Sequences That Reproduce the User-Defined Cotranslational Profile. Our framework (Figure 1b) requires a target cotranslational profile as an input, which consists of a list of state probabilities as a function of nascent chain length. In the case of the MIT protein those state probabilities are denoted $P_{\mathrm{U}}^{\mathrm{tar}}(j)$, $P_{\mathrm{I}}^{\mathrm{tar}}(j)$, and $P_{\mathrm{F}}^{\mathrm{tar}}(j)$ for the unfolded, intermediate, and folded state, respectively, at codon position j. The "tar" superscript indicates these are the user-defined target values. The algorithm of our Monte Carlo-master-equation based framework is as follows. We start by supplying an initial mRNA sequence; then a new sequence is generated by synonymously mutating a codon at a randomly selected position (step 1 in Figure 1b). The resulting cotranslational profile (predicted by eq 3) of this new mRNA sequence is compared to the target cotranslational profile on the basis of the energy term

$$E(\mathrm{new}) = \sum_{j=1}^{N_c} |P_{\mathrm{U}}^{\mathrm{tar}}(j) - P_{\mathrm{U}}^{\mathrm{new}}(j)| + |P_{\mathrm{I}}^{\mathrm{tar}}(j) - P_{\mathrm{I}}^{\mathrm{new}}(j)| + |P_{\mathrm{F}}^{\mathrm{tar}}(j) - P_{\mathrm{F}}^{\mathrm{new}}(j)| \quad (11)$$

where $P_{\mathrm{U}}^{\mathrm{new}}(j)$, $P_{\mathrm{I}}^{\mathrm{new}}(j)$, and $P_{\mathrm{F}}^{\mathrm{new}}(j)$ are the steady-state probabilities of the MIT domain populations in unfolded, intermediate, and folded states, respectively, for the new mRNA sequence. However, any other metric which can quantify the deviation between two cotranslational profiles can also be used as an alternative for E(new). Larger E(new) values correspond to greater deviations between these two cotranslational profiles. Therefore, the overall aim of our framework is to identify the translation-rate profile that minimizes the energy term E(new).

The decision to accept or reject this new mRNA sequence (step 2) utilizes the Metropolis criterion by calculating the quantity

$$E_\Delta = E(\mathrm{new}) - E(\mathrm{old}) \quad (12)$$

where E(old) is the energy associated with the old mRNA sequence. If $E_\Delta < 0$ then the new mRNA sequence is accepted and replaces the old sequence (step 3a); if $E_\Delta \geq 0$ the new mRNA sequence is accepted with probability $e^{-E_\Delta/T}$ (step 3b). This process is then repeated 60 million times and yields converged results (Figure 5).

The temperature T used in step 3b effects the probability of accepting the new mRNA sequence when $E_\Delta > 0$. Simulated annealing[37] is a common way to enhance the efficiency of Monte Carlo searches. Therefore, we ran our algorithm using a simulated annealing temperature schedule starting with T = 10 K and ending when $T \leq 2.4 \times 10^{-8}$ K. Every 60 000 Monte Carlo steps the system was quenched to a lower temperature by multiplying the current T by a factor of 0.99.

Exact Solution of a Two-State Cotranslational Folding Model and Its Sensitivity to Synonymous Mutations. Assume a protein domain can only populate one of two possible states during translation, U and F. The rate of transitioning from the unfolded to the folded state at nascent chain length j is denoted $k_{\mathrm{UF}}(j)$, and the reverse transition occurs with rate $k_{\mathrm{FU}}(j)$. Then using eq 3 we find

$$P_{\mathrm{F}}^{i}(j) = \frac{k_{\mathrm{UF}}(j) + \omega^{i}(j+1)P_{\mathrm{F}}^{\mathrm{wt}}(j-1)}{k_{\mathrm{UF}}(j) + k_{\mathrm{FU}}(j) + \omega^{i}(j+1)} \quad (13)$$

where $i \in$ {wt,mut}.

The change in the steady-state probability of the folded state at codon position j upon introducing a synonymous mutation is then (see Supporting Information for full derivation)

$$P_{\mathrm{F}}^{\mathrm{mut}}(j) - P_{\mathrm{F}}^{\mathrm{wt}}(j) = A(P_{\mathrm{F}}^{\mathrm{wt}}(j) - P_{\mathrm{F}}^{\mathrm{eq}}(j)) \quad (14)$$

where

$$A = \frac{\left(\frac{\omega^{\mathrm{mut}}(j+1) - \omega^{\mathrm{wt}}(j+1)}{\omega^{\mathrm{wt}}(j+1)}\right)}{\left(1 + \frac{\omega^{\mathrm{mut}}(j+1)}{k^{\mathrm{eq}}(j)}\right)} \quad (15)$$

In eq 15, $k^{\mathrm{eq}}(j) = k_{\mathrm{UF}}(j) + k_{\mathrm{FU}}(j)$ is the characteristic rate over which a nonequilibrium configuration decays to equilibrium at codon position j. The change in the steady-state probability of the folded state at codon position k downstream of the synonymous mutation site j is (see Supporting Information for full derivation)

$$P_{\mathrm{F}}^{\mathrm{mut}}(k) - P_{\mathrm{F}}^{\mathrm{wt}}(k) = C_{j+1,k}(P_{\mathrm{F}}^{\mathrm{mut}}(j) - P_{\mathrm{F}}^{\mathrm{wt}}(j)) \quad (16)$$

where

$$C_{j+1,k} = \prod_{i=j+1}^{k} \frac{\omega^{\mathrm{wt}}(i+1)}{\omega^{\mathrm{wt}}(i+1) + k^{\mathrm{eq}}(i)} \quad (17)$$

The parameter $C_{j+1,k}$ is always less than or equal to 1. Therefore, the effect of the mutation at codon position j tends to diminish at subsequent codon positions for a two-state system. Inserting eqs 14 and 16 into the expressions for $\chi(j)$ (eq 24) and $\Delta(j)$ (eq 25) yields

$$\chi(j) = |A|B\Delta(j) \quad (18)$$

where

$$B = \sqrt{\frac{1 + \sum_{i=j+1}^{N_c} C_{j+1,i}}{N_c - j + 1}} \quad (19)$$

The parameter B determines how the change in a cotranslational profile upon introducing a synonymous mutation propagates to subsequent codon positions and is bounded by 0 and 1 ($0 \leq B \leq 1$). A smaller B value indicates that the effect of a synonymous mutation will disapear or be significantly reduced after the translation of just a few downstream codon positions, whereas a large B value suggests the cotranslational perturbation is propagated far downstream of the original mutation site.

Estimation of the Codon Translation Rates for *E. coli* Protein Domains. As in ref 38, we used the method of Viljoen and co-workers[39] to estimate the codon translation rates in *E. coli*. The

DOI: 10.1021/jacs.5b08145
J. Am. Chem. Soc. 2016, 138, 1180−1195

method calculates the average time taken to incorporate an amino acid into a nascent chain at 310 K by using the formula

$$\tau(X) = 9.06 + 1.45[10.48C(X) + 0.5R(X)] \quad (20)$$

In eq 20, $\tau(X)$ is the codon translation time in milliseconds for codon X and 9.06 ms is the time of peptide bond formation between two successive amino acids and translocation of the ribosome to the next codon. Coefficients $C(X)$ and $R(X)$ in eq 20 are functions of the specified codon, concentrations of cognate and noncognate *tRNA* molecules, the diffusion constants of *tRNA* molecules, and the temperature of the cytosol. Details of the method and numerical values of codon translation time for *E. coli* under various conditions are given in ref 38. We used the codon translation time for *E. coli* doubling time of 150 min as reported in Table S1 of ref 38 and inverted them to calculate corresponding rates.

Estimation of Folding and Unfolding Rates of *E. coli* Protein Domains at Each Nascent Chain Length. Our framework requires domain's folding ($k_{UF}(j)$) and unfolding ($k_{FU}(j)$) rates at each nascent chain length j. To compute these rates for the *E. coli* domains we used a previously developed[38] model in which

$$k_{UF}(j) = \frac{k_{UF}(\text{bulk})}{1 + ae^{-j+l+25} + \frac{b}{j^c}} \quad (21)$$

and

$$k_{FU}(j) = k_{FU}(\text{bulk})\left(\frac{1 + e^{-j+l+30}}{d}\right) \quad (22)$$

In eqs 21 and 22, $k_{UF}(\text{bulk})$ and $k_{FU}(\text{bulk})$ are the bulk folding and unfolding rates of the *E. coli* domains of interest and l is the number of residues after the most C-terminal residue of the domain. We used previously reported[38] $k_{FU}(\text{bulk})$ and $k_{UF}(\text{bulk})$ values that were obtained using the method developed by de Sancho-Muñoz.[40] The parameters a, b, c, and d are 404, 3205.5, 1.72, and 0.953, respectively.[38]

RESULTS

Framework for Controlling Cotranslational Folding. The cotranslational folding process of a protein domain is characterized by its cotranslational profile, which is the steady-state probability of the domain being in a particular state as a function of the nascent chain length during synthesis. The states that a domain in a multidomain protein can populate during its synthesis include unfolded (U), intermediate (I), folded (F), and misfolded (M) states. Such domain-wise folding can occur before the full-length protein has been synthesized (Figure 2a). Under physiological conditions, the folded state is often the most stable state that a protein can populate. Misfolded states are metastable and contain non-native structure. Intermediates are partly folded and form transiently stable states which can transition to the folded, the unfolded, or the misfolded states (Figure 2b).

Controlling cotranslational folding using synonymous mutations means being able to alter this cotranslational profile to match a user-defined cotranslational profile through the appropriate choice of synonymous codons. These synonymous codons may alter the translation-rate profile along the coding sequence.[41−43] As a consequence, controlling cotranslational folding requires that we rationally alter an mRNA's translation-rate profile. Therefore, any framework to control cotranslational folding must be able to predict how changing the translation rate at a codon position changes a protein's cotranslational profile, and it also must be able to efficiently search the astronomically large synonymous-codon space of a transcripts's coding sequence to find the optimal mRNA sequence that is predicted to reproduce the user-defined cotranslational profile.

Recent studies[20,44,45] have found that a Markov-state-based analysis[34,46,47] is one way to accurately predict the impact that changing codon translation rates have on a protein's cotranslational profile, provided the interconversion rates between states are known. Also, the Metropolis Monte Carlo algorithm[37] is a standard technique in the physical sciences used to search large state spaces for optimal solutions. Therefore, we propose that the Monte Carlo-master-equation-based framework shown in Figure 1b is a way to rationally design mRNA sequences to control cotranslational folding in silico as well as in wet-lab experiments.

In this design algorithm, a user-defined cotranslational profile is supplied as an input. Then an initial starting mRNA sequence is randomly mutated with a single synonymous codon substitution to create a new sequence (step 1, Figure 1b). The cotranslational profiles of the old and new sequences are predicted using a master equation (eq 3), and the deviations of these profiles from the user-defined cotranslational profile are calculated (eq 11). These deviations are then used to compute E_Δ (eq 12, step 2), which is employed in the Metropolis criterion to either accept or reject the newly mutated mRNA sequence (step 3a or step 3b). This process is iterated until an mRNA sequence is found that results in the best agreement with the target cotranslational profile.

Here, we test this framework in silico by designing optimal mRNA sequences (i.e., translation-rate profiles) that accurately control the cotranslational profile generated from coarse-grained molecular dynamics simulations of protein synthesis. To build this framework, we must first establish that the master equation approach, used in step 2 (Figure 1b), is accurate enough to predict the impact of changing codon translation rates at specific codon positions. If an accurate master equation cannot be created the framework will not work.

We start by solving the steady-state master equation for the probabilities that a nascent chain can populate various states as a function of the codon position in the corresponding coding sequence. The result is (eq 3) the recursive relation

$$\mathbf{P}(j) = \mathbf{M}(j)^{-1}\mathbf{T}(j)\mathbf{P}(j-1) \quad (3)$$

where $\mathbf{P}(j)$ is a column vector containing the steady-state probabilities for the N states (e.g., U, I, and F) that the nascent chain can potentially populate during its synthesis. $\mathbf{M}(j)$ and $\mathbf{T}(j)$ in eq 3 are $N \times N$ matrices. The elements of $\mathbf{M}(j)$ are functions of the rates at which the nascent chain interconverts between different states at nascent chain length j and the elements of $\mathbf{T}(j)$ are a function of the codon translation rate at j. Detailed expressions of $\mathbf{M}(j)$ and $\mathbf{T}(j)$ are given in eq 2. Thus, provided the rate matrix $\mathbf{M}(j)$ is known, eq 3 can be used to predict how the cotranslational profile (i.e., $\mathbf{P}(j)$ versus j) of a protein changes due to changes in individual codon translation rates.

To test the accuracy of predictions from the master equation (eq 3), we used previously generated[20] coarse-grained simulation data of the synthesis of the single-domain MIT protein. MIT is a 77-residue protein that forms a 3-helix bundle in its folded state and can populate unfolded, intermediate, and folded states (Figure 2c). Specifically, arrested ribosome simulations of this ribosome-nascent chain construct in which the MIT protein was C-terminally fused to a 43-amino-acid-long unstructured linker (to mimic multidomain folding) were analyzed, and the rate matrix $\mathbf{M}(j)$ was computed at various

DOI: 10.1021/jacs.5b08145
J. Am. Chem. Soc. 2016, 138, 1180−1195

nascent chain lengths (see Methods). These $\mathbf{M}(j)$ values were then used in eq 3 to predict how uniformly changing the translation rate at all codon positions altered the cotranslational profile of this protein. These predictions were then tested against explicit, continuous synthesis simulations at those same global translation rates. We find that the master equation approach accurately predicts (Figure 3, solid lines) the

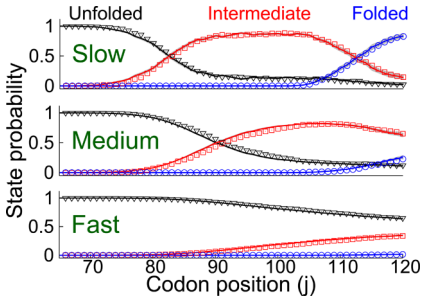


Figure 3. Master equation accurately predicts the effect of changing codon translation rates on the cotranslational profile of the MIT protein. The steady-state probabilities of the MIT domain being in the unfolded, intermediate, and folded states are plotted as a function of the nascent chain length in black, red, and blue, respectively. The probabilities arising from mRNA sequences consisting of only slow-, medium-, or fast-translating codons are shown in the top, middle, and bottom panels, respectively. Solid lines are the numerical predictions made by the master equation approach, and the gray area around these curves covers the 95% confidence interval of the numerical predictions. (Note, the error is only slightly greater than the thickness of the lines.) Error bars for the coarse-grained simulation data are smaller than their symbols. The lowest R^2 values among unfolded, intermediate, and folded state probability curves for the fast-, medium-, and slow-translating mRNA sequences are, respectively, 0.956, 0.994, and 0.995. For each curve, all p values are less than 0.0001 (Table S1).

cotranslational folding behavior from the coarse-grained simulations (Figure 3, discrete data points). (Note, no folding events can occur before the 69th codon position because approximately 30 residues are needed to span the ribosome exit tunnel, and populating the intermediate state requires an additional 40 residues to be outside the tunnel.) Thus, it is possible to predict the effect of changing codon translation rates on a protein's cotranslational profile provided the $\mathbf{M}(j)$s are known.

Having established that the master equation approach yields accurate predictions for the MIT protein, we next tested whether our framework could design mRNA sequences to control its cotranslational folding. To do this for the MIT protein we defined six different cotranslational profiles that we wanted our framework to reproduce (Figure 4, solid lines). Some of these profiles are quite irregular in shape, including step function and linear-ramp behaviors. These six profiles were chosen solely because they exhibit a wide range of behaviors. In all organisms, there are on average three synonymous codons per amino acid (i.e., 61 sense codons for 20 naturally occurring amino acids). Therefore, in this proof-of-principle test, we assumed that three synonymous codons exist for each amino acid—a fast-, medium-, and slow-translating codon. Thus, for MIT's mRNA, which is 120 codons in length, there are 3^{120} ($\sim 10^{57}$) different mRNA sequences that can encode its primary structure. Using these six different profiles as user-defined inputs to our framework yields six different optimized mRNA sequences (Figure 4 and Table S2). As an explicit check that these optimized sequences actually control folding in the user-defined manner we ran Langevin dynamics simulations of continuous synthesis of the MIT protein for these six optimized translation-rate profiles and calculated the resulting cotranslational profiles from these simulations. We find excellent agreement between the first five user-defined profiles and the profiles generated from the coarse-grained simulations of protein synthesis (Figure 4). The last profile (Figure 4f) shows poor agreement for reasons we explain later. Thus, our framework can rapidly and accurately design mRNA sequences to quantitatively control cotranslational folding during synthesis.

Number of Degenerate mRNA Sequences Depends on the Cotranslational Profile. With this in silico-validated method in hand, we can now address a range of important biological questions. Fundamental questions that have not been addressed in the literature include how many different synonymous mRNA sequences can give rise to a particular cotranslational profile and whether this degeneracy depends on the cotranslational profile. These are important questions because they are relevant to the evolutionary processes shaping synonymous-codon usage in organisms.

We address these questions by calculating how many different mRNA sequences give rise to each of the five cotranslational profiles shown in Figure 4a–e. To calculate this quantity we ran our framework (Figure 1b) 32 independent times for each profile and recorded each unique mRNA sequence that reproduced the optimized target profile within a threshold of $E(MC^k) \leq 0.075$, where $E(MC^k)$ is defined as

$$E(MC^k) = \sum_{j=1}^{N_c} |P_{\mathrm{U}}^{\mathrm{opt}}(j) - P_{\mathrm{U}}^{\mathrm{MC},k}(j)| + |P_{\mathrm{I}}^{\mathrm{opt}}(j) - P_{\mathrm{I}}^{\mathrm{MC},k}(j)| + |P_{\mathrm{F}}^{\mathrm{opt}}(j) - P_{\mathrm{F}}^{\mathrm{MC},k}(j)| \quad (23)$$

In eq 23, $P_{\mathrm{U}}^{\mathrm{opt}}(j)$, $P_{\mathrm{I}}^{\mathrm{opt}}(j)$, and $P_{\mathrm{F}}^{\mathrm{opt}}(j)$ are the steady-state probabilities of the MIT domain being in states U, I, and F, respectively, that arise from the optimal mRNA sequences shown in Figure 4. $P_{\mathrm{U}}^{\mathrm{MC},k}(j)$, $P_{\mathrm{I}}^{\mathrm{MC},k}(j)$, and $P_{\mathrm{F}}^{\mathrm{MC},k}(j)$ are the steady-state probabilities arising from the kth random mRNA sequence generated by our framework. N_c in eq 23 is the number of codons in the coding sequence of the mRNA. Thus, by using the 0.075 threshold, we can distinguish mRNA sequences that give rise to the same cotranslational profile (i.e., "degenerate" sequences) or significantly different profiles (i.e., "nondegenerate" sequences). An example of the similarity between an optimized and a degenerate profile is shown in Figure S2. Since only the unfolded state of MIT can be populated at codon positions 1–69 we ignore any degeneracy arising from this region and instead focus on the region between codon positions 70 and 120, as it is in this region that changing translation rates can alter the state that is populated by the nascent chain.

We find that the number of degenerate mRNA sequences depends on the cotranslational profile and ranges from 84 to 13 359 for the five profiles we tested (Figure 5 and Table S4). For example, the profile in Figure 4e is reproduced by only 84 mRNA sequences, while the profile in Figure 4a is reproduced by 13 359 sequences. Thus, this degeneracy spans an astonishing 3 orders of magnitude for the same protein and depends on the details of the cotranslational profile. These results have the biological implication that highly degenerate

DOI: 10.1021/jacs.5b08145
J. Am. Chem. Soc. 2016, 138, 1180–1195

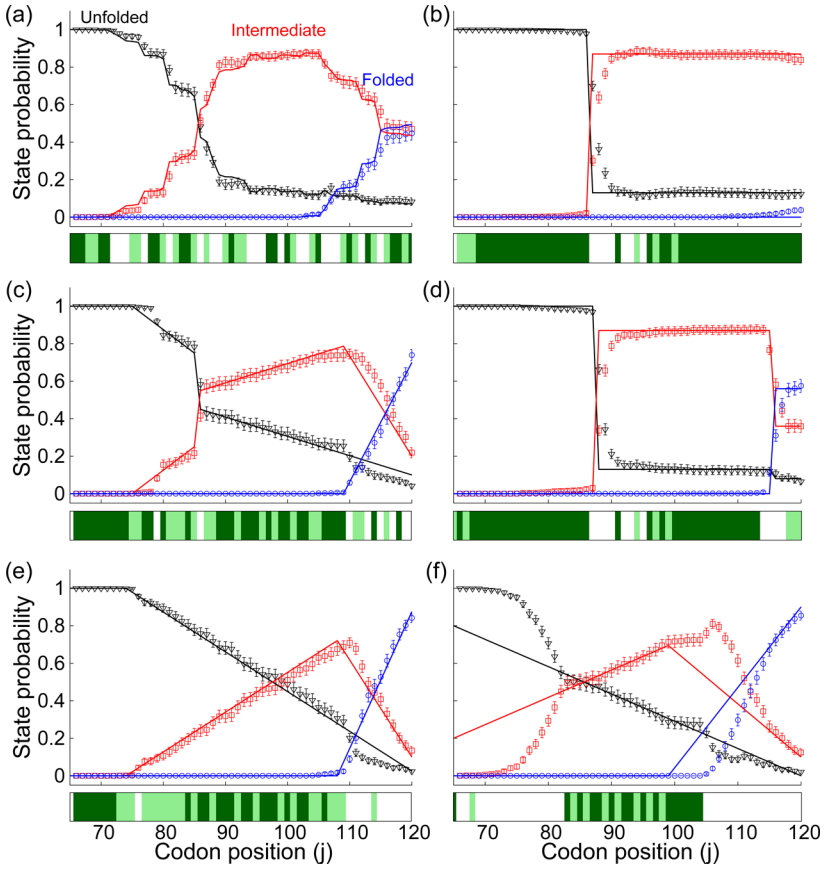


Figure 4. Monte Carlo-master-equation-based framework successfully designs mRNA sequences that reproduce user-defined cotranslational profiles (a–f). Probabilities of populating the unfolded, intermediate, and folded states of the MIT domain are plotted against the nascent chain length in black, red, and blue, respectively. User-defined target cotranslational profiles are plotted as solid lines, while the discrete data points were obtained from the coarse-grained simulations of the continuous translation process from the optimized mRNA sequences generated by our framework (Table S2). Simulation data are plotted with error bars representing the 95% confidence interval. At the bottom of each panel, the translation-rate profile of the optimized mRNA sequence is shown; dark-green, green, and white bars represent fast-, medium-, and slow-translating codons, respectively. The lowest R^2 values among the unfolded, intermediate, and folded state probability curves in a, b, c, d, e, and f are, respectively, 0.993, 0.962, 0.979, 0.947, 0.981, and 0.653. For all curves, the p values are less than 0.0001 (Table S3).

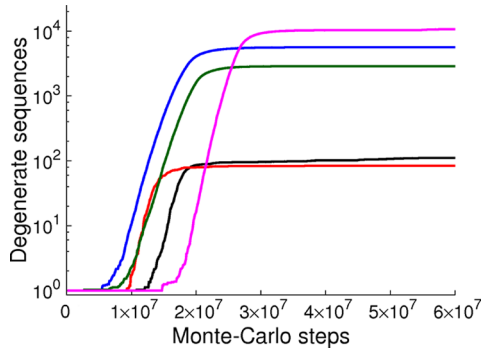


Figure 5. The number of degenerate mRNA sequences that encode the same cotranslational profile is a function of the cotranslational profile. The number of unique mRNA sequences (i.e., degenerate sequences) that give rise to the same cotranslational profile as the optimized mRNA sequence as a function of the Monte Carlo step number in a run of our framework. The number of degenerate sequences shown at each step is averaged over all 32 independent runs. The number of degenerate sequences identified for the cotranslational profiles shown in Figure 4a, 4b, 4c, 4d, and 4e are indicated by the lines colored in magenta, blue, black, green, and red, respectively.

cotranslational profiles are more likely to be robust to random synonymous mutations.

To test whether our search for degenerate sequences was exhaustive, we plotted the number of unique, degenerate sequences that were found as a function of the Monte Carlo step in our framework. For all profiles, converged behavior (i.e., a plateau region) is observed (Figure 5), consistent with an exhaustive search. Further support for an exhaustive search is that of the 32 independent runs of our framework; a majority were able to find all of the degenerate sequences within a single run (Table S4). Thus, our calculations are very likely to have identified all degenerate sequences for each of the profiles we tested.

Critical Codon Positions Depend on the Cotranslational Profile. Our master equation model (eq 3) allows us to rapidly identify the codon positions where a synonymous codon substitution will have a significant effect on the cotranslational profile. We carried out an in silico synonymous-codon scanning experiment in which single-point mutations were made at each and every codon position in the optimized mRNA sequences (Figure 4), and the effect on MIT's cotranslational profile was predicted using eq 3 for each mutation separately. The effect of a mutation at codon position j on the cotranslational profile was measured as the root-squared deviation between the mutated and the optimized profile

DOI: 10.1021/jacs.5b08145
J. Am. Chem. Soc. 2016, 138, 1180–1195

$$\chi(j) = \sqrt{\sum_{k=j}^{N_c} ((P_U^{org}(k) - P_U^{mut}(k))^2 + (P_I^{org}(k) - P_I^{mut}(k))^2 + (P_F^{org}(k) - P_F^{mut}(k))^2)} \quad (24)$$

In eq 24, $P_U^{org}(k)$, $P_I^{org}(k)$, and $P_F^{org}(k)$ are the steady-state probabilities of the MIT domain being in states U, I, and F, respectively, for the original mRNA sequence (Figure 4) and $P_U^{mut}(k)$, $P_I^{mut}(k)$, and $P_F^{mut}(k)$ are the steady-state probabilities after the synonymous mutation was introduced. (Note, any initial mRNA sequence can be used in eq 24; use of an optimized sequence as the starting sequence is not required.) In this in silico experiment, we have two choices for the synonymous substitution at each codon position since we assume there are only three synonymous codons per amino acid. For example, if the original codon at a particular position is a fast-translating codon, we can substitute either a medium or a slow codon in its place. We carry out both possible types of single-point substitutions at each codon position and calculate the sensitivity, i.e., $\chi(j)$, for each case.

Critical codon positions are identified by large $\chi_{max}(j)$ (i.e., the maximum sensitivity between the two possible synonymous substitutions for a codon position) values. We find that the positions of critical codons change depending on the cotranslational profile (Figure 6). For example, in Figure 6d,

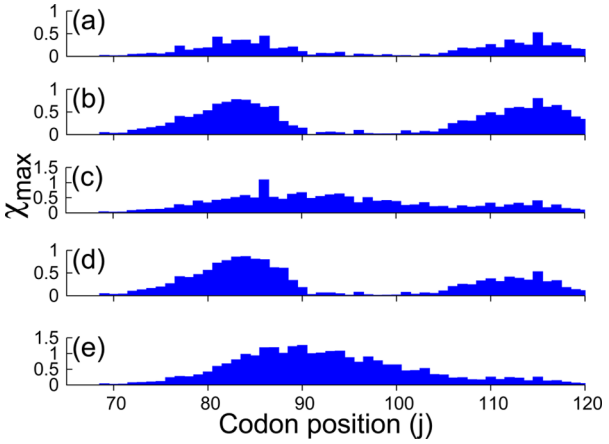


Figure 6. Sensitivity profiles for the optimized mRNA sequences. $\chi_{max}(j)$ as a function of the codon position for the different MIT mRNA sequences. Panels a, b, c, d, and e display the sensitivity profiles arising from the optimized mRNA sequences shown, respectively, in Figure 4a, 4b, 4c, 4d, and 4e.

the sensitivity distribution is bimodal, whereas in Figure 6e, the distribution is unimodal. Between these two distributions the most sensitive codon positions switch from codon positions 84 and 115 in Figure 6d to position 90 in Figure 6e. Thus, for the same protein, a synonymous codon substitution at a particular position may or may not have a significant impact on folding depending on the original cotranslational profile. Later in the paper we discuss the physical origins of these critical codon positions.

Degeneracy Correlates with Sensitivity to Single-Point Synonymous Codon Substitutions. We hypothesized that the number of mRNA sequences that give rise to the same cotranslational profile (i.e., the profile's degeneracy) should be related to the sensitivity of the cotranslational profile to single-point synonymous substitutions. To test this hypothesis, we examined if there was a correlation between the number of degenerate sequences (Table S4) and the number of insensitive mutations. An insensitive mutation is one in which a synonymous substitution into the optimized mRNA sequence causes no change in the cotranslational profile, with a threshold of $E(MC^k) \leq 0.075$ (eq 23). We only have five points to plot; therefore, to increase the statistical power of this test we created three additional cotranslational profiles for MIT that we had our framework create optimized sequences for and for which we calculated their degeneracy and number of insensitive mutations. These three profiles (Figure S3) are dissimilar to the other five profiles. We find an R^2 Pearson correlation coefficient of 0.74 between a profile's degeneracy and the number of insensitive mutations (Figure 7, p value = 0.0053). This

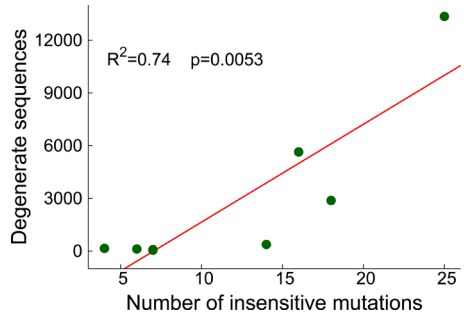


Figure 7. The number of degenerate mRNA sequences per cotranslational profile shown in Figures 4a–e and Figure S3 as a function of the number of single-point-synonymous mutations that have no effect on the cotranslational profile (i.e., insensitive mutations). The eight (two of them are overlapping) different data points in this plot arise from the eight optimized mRNA sequences of the MIT protein generated by our framework (Figures 4a–e and S3.).

supports our hypothesis that a relationship exists between a cotranslational profile's mRNA sequence degeneracy and that profile's robustness to single-point synonymous mutations. This result has the experimental implication that knowledge of the impact that a small number of synonymous mutations have on a protein's cotranslational folding can provide information on the much larger sequence space of transcripts that give rise to the same cotranslational behavior.

We tested the robustness of the correlation in Figure 7 by changing the threshold of $E(MC^k)$ (eq 23) used to identify degenerate sequences to first 0.0825 and then 0.0900 and also by changing the functional form of $E(MC^k)$ to a root-mean-square form (eq S17) with a $E(MC^k)$ threshold of 0.0150. We find the resulting R^2 correlations range from 0.52 to 0.93 for these three cases (Figure S4) with statistically significant p values for two of them. In one case the p value is just above the significance level threshold of 0.05 (p value = 0.0657, Figure S4(c)). This suggests the correlation is fairly robust, although testing this correlation for other proteins in future research is important in establishing whether this is a general phenomenon.

Sensitive Codon Positions Tend To Be Far from Equilibrium. A fundamental question in molecular biology is why synonymous codon substitutions at some codon positions

DOI: 10.1021/jacs.5b08145
J. Am. Chem. Soc. 2016, 138, 1180–1195

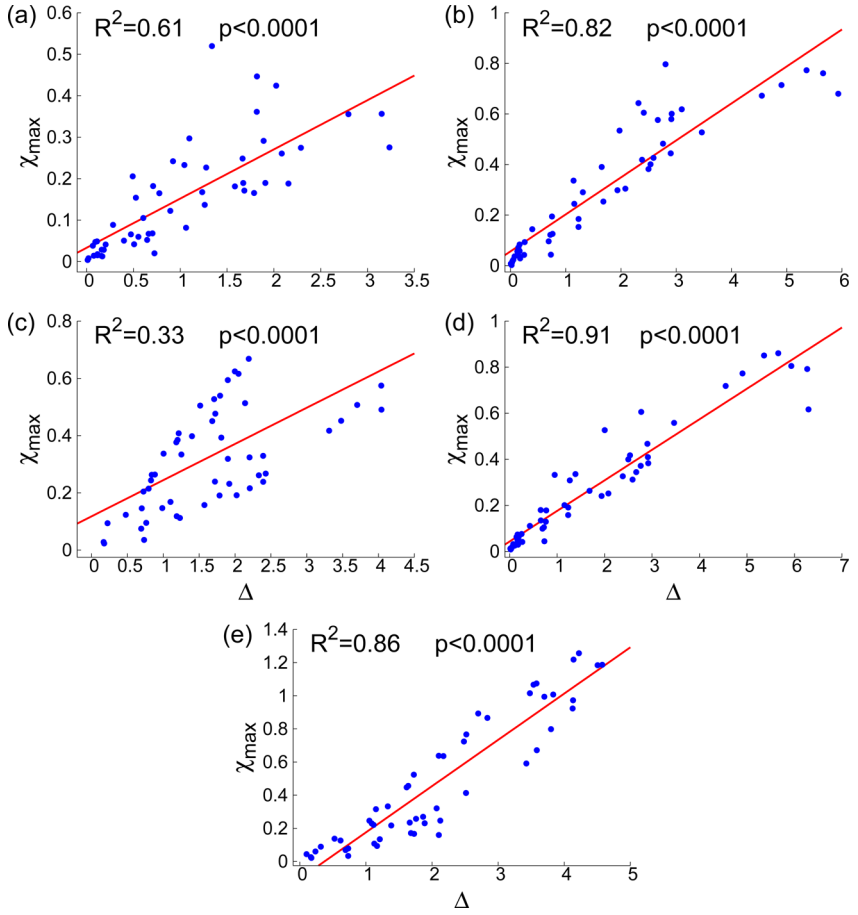


Figure 8. Scatter plots of χ_{max} versus $\Delta(j)$ for the various cotranslational profiles of the MIT protein shown in Figure 4a–e. Panels a–e show data for the five optimized mRNA sequences shown in Figure 4a–e, respectively. These plots exhibit a linear correlation indicating that the impact that a single synonymous point mutation has on a cotranslational profile is a function of the deviation of the original cotranslational profile from equilibrium.

along a coding sequence have a big effect on cotranslational behavior but mutations at other positions do not. For example, for the profile in Figure 4d, why is codon position 85 highly sensitive to synonymous mutations while at position 97 such mutations have no effect on the folding process (Figure 6d)?

Our recent Perspective article[1] emphasized that because translation-elongation kinetics can have such a large impact on nascent-protein behavior it must be the case that for many proteins their cotranslational folding occurs under non-equilibrium (i.e., "kinetically controlled") conditions rather than quasi-equilibrium (i.e., "thermodynamically controlled") conditions. Therefore, we hypothesized that codon positions that result in significant cotranslational profile changes upon a synonymous mutation are those that are furthest from equilibrium. This hypothesis predicts that there should exist a correlation between $\chi(j)$ and a measure of the distance from the equilibrium process of cotranslational folding.

To test this hypothesis, we first define $\Delta(j)$ as

$$\Delta(j) = \sqrt{(N_c - j + 1)((P_U^{org}(j) - P_U^{eq}(j))^2 + (P_I^{org}(j) - P_I^{eq}(j))^2 + (P_F^{org}(j) - P_F^{eq}(j))^2)} \quad (25)$$

where the equilibrium state probabilities, $P^{eq}_{ie\{U,I,F\}}$,are calculated under conditions in which translation is arrested ($\omega(j) = 0$ at all j). In eq 25, we are multiplying the difference between the equilibrium and the original cotranslational profiles at codon j where a synonymous mutation could be potentially made by the number of codon positions downstream of $j - 1$ which have the potential to be impacted by the mutation at j.

We find moderate to strong correlations between $\chi_{max}(j)$ and $\Delta(j)$ for the five cotranslational profiles (Figure 8) of MIT protein. For three out of the five profiles, 82% or more of the variance in codon position sensitivity to synonymous mutations can be predicted by the deviation of the original profile from equilibrium. For the other two profiles, 61% and 33% of χ_{max}'s variance can be explained by Δ. This means that in the majority of cases for the MIT protein, simply comparing the optimized and arrested-ribosome cotranslational profiles provides a reasonable predictor of which codon positions are likely to strongly influence cotranslational folding upon a synonymous mutation.

Similar Results Are Found Even at Physiologically Relevant Rates. All previous results were based on applying our sequence design method to the MIT domain simulated using a coarse-grained representation and low-friction Langevin dynamics—two modeling techniques known to artificially accelerate the rates of protein folding by orders of magnitude.[48] For this reason, the codon translation rates were also

DOI: 10.1021/jacs.5b08145
J. Am. Chem. Soc. 2016, 138, 1180–1195

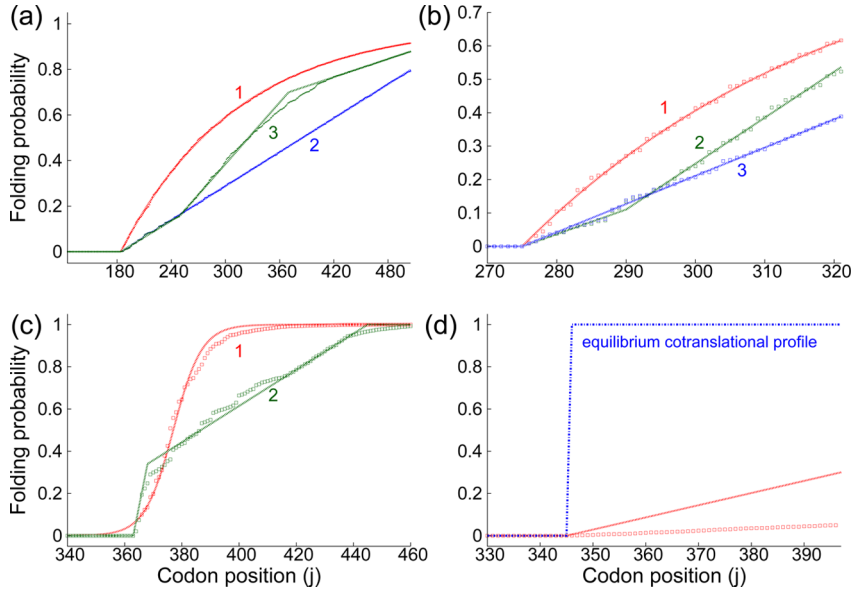


Figure 9. Monte Carlo-master-equation-based framework successfully optimizes mRNA sequences that reproduce user-defined cotranslational profiles for *E. coli* proteins (a–d). Folding probabilities of the (a) first domain of SYK1, (b) third domain of TRXB, (c) fourth domain of FDHF, and (d) second domain of AAT proteins are plotted as a function of nascent chain length during synthesis. Different cotranslational profiles of the same protein are plotted in a different color and labeled by an integer index 1, 2, or 3. User-defined target cotranslational profiles are plotted with dashed lines, whereas optimized cotranslational profiles are plotted with discrete data points. In (d), the dashed blue line is the equilibrium folding curve of AAT protein.

accelerated in those synthesis simulations (see Methods). This raises the question of whether the conclusions drawn from MIT will hold for proteins that are modeled at more realistic, physiologically relevant rates.

To address this question we applied our method to domains from four different proteins from *E. coli* (SYK1, domain 1; TRXB, domain 3; FDHF, domain 4; AAT, domain 2) that were previously predicted[38] to cotranslationally fold. We used estimated codon translation rates for *E. coli* growing at 310 K (see Methods) and doubling every 150 min. In these estimates the 64 codons translate with rates that range from 2.5 to 54.0 AA/s. We used a previously published phenomenological model[39] that predicts a domain's $k_{FU}(j)$ and $k_{UF}(j)$ values (eqs 21 and 22) on the ribosome based, in part, on the domain's size and structural class. These models provided us with all the rates necessary to use our design framework. Thus, we did not need to use molecular simulations to obtain the interconversion rate matrix $\mathbf{M(j)}$ as we did for MIT.

We find that for three out of the four of *E. coli* protein domains our method can find mRNA sequences that accurately reproduce user-defined profiles (Figure 9). The algorithm did not work well for one of the domains (Figure 9d); we discuss this failure in depth later. Thus, as with MIT, our algorithm works even at physiologically relevant rates of folding, unfolding, and codon translation.

We also examined whether the biologically relevant findings from MIT also held for these proteins. We find that indeed they do: critical codon positions can shift depending on the cotranslational profile (Figure S5) as seen by plotting χ_{max} vs j for the two cotranslational profiles of FDFH protein (Figure 9c). Mild to moderate correlations between χ_{max} and Δ are observed for these proteins (Figure S6), indicating that the more sensitive a codon position is (i.e., the larger χ_{max}) the more likely it is to be located at positions where the cotranslational profile is further from the equilibrium profile. Also, tentatively, we observe that, as with the MIT domain, the number of unique mRNA sequences that yield the same cotranslational profile for TRXB protein depends on the original cotranslational profile and likely spans almost an order of magnitude or more (Table S5). We emphasize that the last observation is tentative because we know that, despite our best efforts, our estimates of the number of degenerate mRNA sequences are not converged for all of the cotranslational profiles associated with the TRXB protein. Unlike MIT, only for a minority of the 32 independent runs of our framework do we find the same unique sequences (data not shown). This means the values reported in Table S5 are lower bounds of the true number of degenerate mRNA sequences. Because the numbers of degenerate sequences are not exact, we cannot test whether, like for MIT, the degeneracy correlates with the sensitivity to single-point synonymous mutations for these proteins. Thus, where the data permits rigorous testing, all of the conclusions drawn from these *E. coli* proteins are consistent with those from MIT.

Ratio of the Time Scales Also Contributes to the Sensitivity of a Codon Position to Synonymous Mutations. Not all of the variance in Figures 8 and S6 is explained by Δ (i.e., $R^2 \neq 1$). Thus, other factors must also contribute to the magnitude of the effect that a synonymous mutation at a particular codon position has on a cotranslational profile. To gain insight into these additional factors, we analytically solved a chemical kinetic model describing cotranslational-domain folding involving only two states, U and F, for an exact relationship between $\chi(j)$ and $\Delta(j)$. (An analytic solution for a three-state folder is not possible to our knowledge.) We find that the relationship between $\chi(j)$ and $\Delta(j)$ (derivation provided in the Methods section) in this situation is

$$\chi(j) = |A|B\Delta(j) \tag{18}$$

where the A and B are defined by eqs 15 and 19, respectively. Thus, while χ is a function of Δ, it is also a function of A and B,

DOI: 10.1021/jacs.5b08145
J. Am. Chem. Soc. 2016, 138, 1180–1195

which are terms involving ratios of rates associated with translation-elongation kinetics and the kinetics of domain folding. If eq 18 is accurate, it should be the case that plotting χ vs $|A|\Delta B$ for the *E. coli* domains (which are two-state folders) should yield a perfect one-to-one correspondence (i.e., $R^2 = 1$), and indeed, this is what we find (Figure S7). Thus, the sensitivity of a cotranslational profile to synonymous mutation is a function of Δ and a ratio of time scales.

To estimate the relative contributions of the terms A and B to χ we calculated the correlation of χ with A, and χ with B, for the *E. coli* proteins. We find χ has little to no correlation with B for all cases tested ($R^2 \leq 0.27$, data not shown). We find that χ correlates with A very strongly in some cases (Figures S8(b), S8(e), and S8(h)) and weakly in others (Figures S8(c) and S8(f)). We also note that the quantities A and Δ appear to be anticorrelated: the TRXB protein shows a very good correlation with $|A|$ (R^2 values vary from 0.59 to 0.72, Figure S8), whereas the same protein exhibits a weak correlation between χ_{max} and Δ ($R^2 \leq 0.16$, Figure S6); FDHF protein shows a good correlation of χ_{max} with Δ ($R^2 = 0.58$, Figure S6), but its χ_{max} does not correlate well with $|A|$ ($R^2 \leq 0.08$, Figure S8). To test this, we calculated the correlation of χ with $|A|\Delta$ and we find $R^2 \geq 0.99$ in all cases Figure S9. These results indicate A and Δ are more important than B in determining χ_{max}. Thus, the terms A and Δ appear to be the greatest determinants of critical codon positions.

Situations in which This Framework Yields Poor Agreement with User-Defined Profiles. We reported two examples where our framework designs mRNA sequences that result in cotranslational profiles that are in poor agreement with the user-defined profiles (Figures 4f and 9d). These provide case studies that illustrate the limitations of our approach. The simplest failure to understand is that shown in Figure 9d for a domain in the *E. coli* protein AAT. At equilibrium (i.e., infinitely slow translation), this domain cotranslationally folds (Figure 9d, dashed blue line). However, when we design a user-defined target cotranslational profile no mRNA sequence can be found that recapitulates it. This appears to be a conundrum because Δ is large in this case at essentially all codon positions beyond 345. However, the folding and unfolding rates for this domain are much smaller than the individual codon translation rates, meaning that it never has time to fold at the elongation rates found in *E. coli*. More precisely, within the analytic expression relating χ and Δ (eq 18), this very slow domain folding drives $A \rightarrow 0$, meaning that no matter how large Δ is, the ratio of time scales ensures we cannot alter the domain's folding behavior with *E. coli*'s range of available codon translation rates. Specifically, the denominator in the expression of A (eq 15) contains the term $\frac{\omega^{mut}(j+1)}{k^{eq}(j)}$; with such a small $k_{UF}(j)$ and $k_{FU}(j)$ for this domain, $k^{eq}(j)$ tends toward zero and the term $\frac{\omega^{mut}(j+1)}{k^{eq}(j)}$ achieves a very large value, driving A toward zero.

The other example is the user-defined profile shown in Figure 4f (solid lines), which cannot be accurately reproduced at codon positions 65−80 and positions 100−110. This failure can be understood by noting that the equilibrium cotranslational profile of MIT is accurately represented by the quasi-equilibrium profile shown in Figure 3 (top), and thus, at equilibrium the intermediate state in not significantly populated between codons 65 and 80 and the folded state is not significantly populated between codon 100 and 110. The target profile in Figure 4f requires these states to be significantly populated in these respective regions. Thus, our framework yields poor agrement here because there is no thermodynamic driving force to populate intermediate or folded states in these regions. Expressed more technically, the maximum extent to which a state can be populated at codon position j is always less than or equal to the maximum equilibrium probability that the state can be populated between codon positions 1 and j, that is, $P_i(j) \leq \max\{P_i^{eq}(1),...,P_i^{eq}(j)\}$, where $P_i(j)$ and $P_i^{eq}(j)$ are, respectively, the probability of being in state i at codon position j during continuous synthesis and at equilibrium (i.e., arrested translation). Thus, despite the fact that our framework exploits the nonequilibrium nature of translation to control the cotranslational folding process, thermodynamic properties of the ribosome-nascent chain complex set hard bounds on the range of possible nascent chain behaviors that an mRNA sequence can encode.

DISCUSSION

Synonymous codon-usage bias in the genomes of organisms appears to have been evolutionarily selected for in part to alter the translation-rate profiles across coding sequences to influence the process of cotranslational protein folding.[9,49,50] Fundamental biological questions related to this phenomenon remain unanswered, including why there is a differential impact on cotranslational folding depending on the location at which a synonymous mutation is introduced along a transcript. In the present study, we developed a computational framework (Figure 1b) that exploits the nonequilibrium nature of translation to design mRNA sequences that can control cotranslational folding in a user-defined manner. We addressed biological questions on the physical origins of critical codon positions and examined the extent to which a cotranslational profile can be encoded by different synonymous mRNA sequences.

Our results show that knowledge of the underlying rates of codon translation, folding, and unfolding of a domain during synthesis can be used as inputs to our framework to rationally design mRNA sequences to manipulate nascent-chain behavior. Specifically, the desired cotranslational behavior is first defined as the probability of a protein segment being in a particular state at different nascent chain lengths during synthesis, which we refer to as the target cotranslational profile. The framework then uses Metropolis Monte Carlo to search the astronomically large mRNA sequence space that encodes the protein, which arises from the various combinations of synonymous codons, to find the optimal mRNA sequence that most closely reproduces the user-defined cotranslational profile as predicted by a master equation model. Varying the mRNA sequence through synonymous codon mutations alters the rate at which the ribosome moves along it. Provided the rate matrix $\mathbf{M(j)}$ and codon translation rates are known and regardless of the molecular origin of these rates,[51−53] our method is applicable in principle to all kingdoms of life. Thus, our framework finds the optimal translation-rate profile that guides a protein's cotranslational protein folding in a user-defined manner. In this way, translation-elongation dynamics are used in our framework to control cotranslational folding.

With this framework we demonstrated that a wide range of cotranslational folding behaviors can be encoded in an mRNA sequence constructed by using different combinations of synonymous codons. For the MIT domain, which can populate an on-pathway, native-like intermediate, we encoded cotranslational behavior in the state probabilities that displayed step-

DOI: 10.1021/jacs.5b08145
J. Am. Chem. Soc. 2016, 138, 1180−1195

function (Figure 4b and 4d), linear-ramp (Figure 4e), and a combination of step-function and linear-ramp (Figure 4c) changes during synthesis. Reproduction of these profiles using the optimized mRNA sequences in independent coarse-grained simulations (Figure 4) demonstrates the precision our framework can achieve in controlling the cotranslational folding process.

To estimate the rates at which MIT interconverts between states we used coarse-grained molecular dynamics simulations. Such simulations have the well-characterized effect of dramatically speeding up folding rates,[48] and therefore, the codon translation rates were increased in the continuous-translation simulations to a rate that is much faster than occurs in vivo. To test if our framework offered a similar level of control over cotranslational folding when the values of these rates were realistic, we utilized estimated *E. coli* codon translation rates[39] that range from 2.4 to 54.0 AA/s and estimated[38] domain folding and unfolding rates that agree with experimental values.[38,40] We applied our framework to domains in four *E. coli* proteins and found that in most cases the framework could find mRNA sequences that quantitatively reproduced the user-defined cotranslational profiles (Figure 9), demonstrating that the framework works equally well at physiologically relevant rates.

Compared to other strategies for influencing nascent-protein behavior, our framework is unique in a number of ways. Conventional techniques, referred to as codon optimization methods,[17,54] frequently focus on designing mRNA sequences that maximize the amount of protein that is produced by an mRNA molecule. These methods often utilize various characteristics of codons as surrogates for codon translation rates, such as whether they are rare or common in the organism's genome or have low or high cognate *tRNA* abundances. For example, a sequence design technique[54] for heterologous protein expression is to use the most common synonymous codon at each position in an mRNAs coding sequence based on the assumption that common codons are translated quickly and more accurately.[55] What sets our approach apart from these other methods is that our framework explicitly accounts for the influence of translation dynamics on cotranslational protein behavior. Thus, our framework can design proteins that avoid misfolding and thereby increase the amount of functional protein in heterologous expression. Unlike other approaches, our method can predict its own success or failure. For example, for the user-defined profile in Figure 4f (solid lines), our framework predicted (data not shown) that the best an mRNA sequence could do is reproduce the profile between codon positions 80 and 105 and fail elsewhere, which is what we observed when we ran the explicit coarse-grained simulations of protein synthesis.

A challenge in applying our framework is knowing the states a nascent protein populates during translation, the interconversion rates between those states, and the individual codon translation rates. High-throughput and single-molecule experiments, however, have made it possible to measure some of these quantities. For example, ribosome profiling data are being used to estimate in vivo codon translation rates.[63−66] A number of techniques[67,68] can provide a measure of cotranslational folding curves. For example, FactSeq[67] can evaluate equilibrium cotranslational profiles. With advances in fluorescence-based techniques[69,70,72] it should be possible to study the time-dependent folding kinetics of a nascent protein on an arrested ribosome. As described by Johnson,[70] donor (D) and acceptor (A) dyes can be attached to two chemically modified nascent chain residues to measure changes in FRET efficiency[71] and thereby probe folding and unfolding rates. Performing these experiments on a series of truncated mRNA sequences of different lengths would generate the time series of states at those codon positions, which could then be used in our master-equation approach. In addition, a recent publication from the Bustamante[73] group has measured the folding and unfolding rates of a domain stalled on the ribosome. Information obtained from such experiments could potentially be used to develop a phenomenological model to predict the domain folding and unfolding rates near the ribosome surface. Even when such data are not available, there are theoretical models that predict an organism's codon translation rates[74,75] and a domain's bulk folding and unfolding rates.[40,76,77] Thus, our framework can be utilized for various proteins using measured or estimated rates currently available in the literature.

This framework can be extended to design mRNA sequences to control other cotranslational nascent-protein behaviors. In eukaryotes and prokaryotes, there are at least 11[80] different cotranslational processes[3,15,56−58] that can act on a nascent chain during its synthesis. Each one of these processes can be represented as a different state in the cotranslational reaction network (Figure 2b), resulting in a modification to the rate matrix $\mathbf{M}(j)$ used in the master equation (eq 3). For example, to incorporate the effects of the cotranslationally acting *E. coli* chaperone trigger factor on two-state cotranslational folding would require an additional state to be added to the reaction scheme. This new state would account for the binding of trigger factor to the unfolded state of the domain. Trigger factor has been shown to slow down the cotranslational folding of one protein.[22] Therefore, folding rates of the protein would have to be decreased for this new state by an appropriate amount. Besides this, no changes are required to our framework (Figure 1b) to account for these additional processes. Extending the model in this way would open up a large range of possibilities, such as making it possible to design mRNA sequences that minimize the chances of premature nascent-protein degradation[59] and maximize the efficiency of the cotranslational translocation of secretory proteins into the endoplasmic reticulum.[60]

Seven synonymous variants of the EgFABP1 gene were found to exhibit widely different behavior in *E. coli*:[61] one out of the seven variants produced a large fraction of insoluble, aggregated protein, suggesting cotranslational folding was altered by this specific set of mutations. On the other hand, the other six variants changed the fraction of insoluble protein very little relative to wild-type. Our results provide an explanation for this codon-position-dependent impact on cotranslational folding. In this study, we found that in silico the MIT protein as well as the *E. coli* proteins recapitulate the experimental observation that mutations at some codon positions can have a bigger impact on cotranslational folding than others, as χ_{max} was found to be profile dependent (Figures 6 and S5). This allowed us to explore within these models the factors determining such critical-codon positions. We found that at a given codon position j, the further the nascent chain state probabilities were from their equilibrium values the greater the impact a single synonymous codon substitution could have at that position (Figures 8 and S6). That is, the farther a cotranslational profile is from equilibrium, the more likely it is that synonymous codon substitutions will significantly affect that profile and the more likely it will be

DOI: 10.1021/jacs.5b08145
J. Am. Chem. Soc. 2016, 138, 1180−1195

that critical positions are those positions that exhibit the greatest deviation from equilibrium. This is supported by the moderate to strong correlation of Δ, a measure of the deviation from equilibrium, with χ_{max}, a measure of the impact of synonymous mutations on the cotranslational profile (Figures 8 and S6).

The deviation from equilibrium is not the only determinant of critical codon positions because $R^2 \neq 1$ in Figures 8 and S6. To gain insight into what these other factors are we derived an exact expression (eq 18) for the relationship between χ and Δ by solving a chemical reaction scheme representing the cotranslational folding of domains that fold in a two-state manner. We found that two additional terms, *A* and *B*, appear in eq 18. These terms are functions of ratios of various time scales associated with translation and folding. For example, *A* is proportional to the relative change in a codon position's translation rate upon mutation as well as the relaxation rate of the nascent-protein dynamics. These additional factors account for the unexplained variance in Figures 8 and S6, with *A* and Δ as the two major determinants of the position-dependent impact synonymous mutations have on cotranslational folding (Figure S9); *B* exhibits little correlation with χ. When *A* strongly correlates with χ_{max}, Δ exhibits a weak correlation and vice versa (see, for example, proteins TRXB and FDHF in Figures S6 and S8). This suggests that for some domains the deviation of the cotranslational profile from equilibrium is the primary determinant of critical codon positions, and for others a ratio of time scales as represented by *A* is more important.

For the MIT protein and *E. coli* protein TRXB we found that the critical-codon positions change depending on the cotranslational profile (Figures 6 and S5). This phenomenon is consistent with studies of the Multidrug Resistance 1 (MDR1) protein.[62] Single-point synonymous mutations introduced at nucleotide position 1236 (C > T) or 3435 (C > T) in the wild-type MDR1 gene did not change the drug-transport activity of the mutant transcripts, suggesting they did not perturb the protein's structure.[62] However, MDR1's functionality was altered when both of these mutations were present simultaneously in the transcript. In light of our results, this experimental result suggests that the codon position containing nucleotide 3435 is not a critical codon position for the wild-type cotranslational profile but becomes a critical codon position for the cotranslational profile encoded by the 1236 (C > T) mutant transcript and vice versa.

We observed that the cotranslational profile influences the number of degenerate mRNA sequences that give rise to it, and that this degeneracy correlates with the sensitivity of the cotranslational profile to single-point mutations. For example, the number of mRNA sequences that give rise to the same cotranslational profile can span 3 orders of magnitude (Table S4) depending on the starting cotranslational profile of the MIT domain. (The results for the TRXB protein were inconclusive on this point because the degeneracy did not converge in the calculations.) We found that the mRNA-sequence degeneracy of these cotranslational profiles correlated with the number of codon positions that upon synonymous mutation had no impact on the cotranslational profile (Figure 7). The correlation was statistically significant in two out of three robustness tests we carried out (Figure S4), suggesting that the validity of this observation is tentative and needs to be established for other proteins as well. (Again, the degeneracy for the *E. coli* proteins did not converge, meaning we could not test this correlation for them.) This observation suggests experimentalists may be able to estimate the relative degeneracy of a cotranslational profile from a single-point synonymous mutation scanning experiment. For example, assume there are two orthologous proteins of the same length. Our observation suggests that a synonymous-codon scanning experiment can be run, and the protein that can withstand the largest number of single-point codon mutations without changing its cotranslational-folding behavior will have more degenerate mRNA sequences. This correlation is not a tautology because the number of insensitive mutations arising from single-point substitutions across an mRNA sequence might not capture the combinatorial complexity, nor the additive and subtractive effects that multiple, simultaneous mutations could have on a cotranslational profile. Indeed, the aforementioned MDR1 gene is an example in which there is a synergistic effect due to a double mutation.

The sensitivity of far from equilibrium cotranslational profiles to synonymous codon mutations has implications for mRNA sequence evolution and codon usage in organisms (Figure 10).

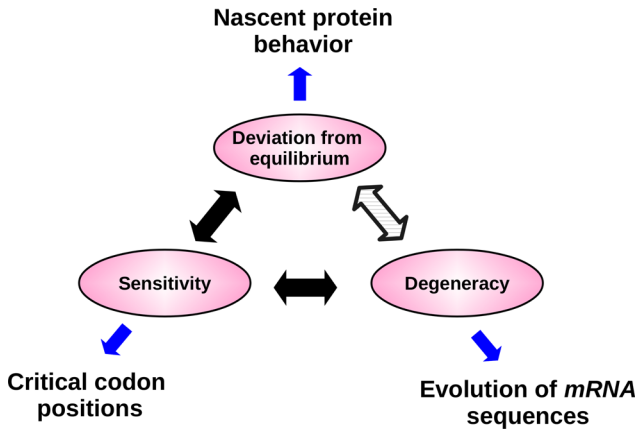


Figure 10. The connections made in the study between different phenomena related to the cotranslational folding of a protein and their biological implications. Black arrows indicate the establishment of a direct connection between phenomena, while the shaded arrow indicates an indirect link. Blue arrows indicate the biological phenomena that are directly impacted by these physical properties of translation-rate and cotranslational profiles.

Consider the following thought experiment: Two different proteins in an organism are both crucial to a cell's viability. One of these proteins, however, has a highly degenerate cotranslational profile, meaning its cotranslational behavior is robust to synonymous codon mutations. The other protein has a low-degeneracy cotranslational profile, and hence, its nascent behavior can be perturbed by just a few synonymous mutations. Our results suggest that if cotranslational processing is important to the maturation of these proteins then the latter protein will exhibit fewer synonymous codon mutations across a species population than the former. This suggests that the nonequilibrium nature of cotranslational processes can contribute to shaping the codon usage across the genomes of organisms.

In summary, this study has provided a number of biological insights by identifying phenomenological and physical rules governing which positions along an mRNA molecule will have a significant impact on cotranslational behavior due to a synonymous mutation. Our results show that a cotranslational profile's deviation from equilibrium, its sensitivity to single-

DOI: 10.1021/jacs.5b08145
J. Am. Chem. Soc. 2016, 138, 1180−1195

point mutations, and its mRNA-sequence degeneracy are interrelated quantities and that each of these factors has direct implications for nascent-protein behavior, critical codon positions, and mRNA sequence evolution (Figure 10).

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/jacs.5b08145.

Additional information about the CHARMM simulation parameters, iso-contour plot of the free energy surface of the MIT protein at 320 K, an example of the similarity between the optimal and a degenerate profile, three additional cotranslational profiles of the MIT construct that our framework designed optimized mRNA sequences for, robustness analysis of the correlation between the number of degenerate sequences and the number of insensitive mutations, sensitivity profiles for the optimized mRNA sequences of the FDHF protein, correlation between the sensitivity of a codon position to synonymous mutations (χ_{max}) and Δ for three *E. coli* proteins, correlation between χ_{max} and $|A|B\Delta$ for three *E. coli* proteins, correlation between χ_{max} and $|A|$ for three *E. coli* proteins, correlation between χ_{max} and $|A|\Delta$ for three *E. coli* proteins, statistics of the correlation between the simulation data and the master equation predictions in Figure 3, nine optimized MIT mRNA sequences shown in Figures 4 and S3, statistics of the correlation for each curve in Figure 4, number of degenerate sequences for the eight optimized mRNA sequences in Figures 4a–e and S3, lower bound of the number of degenerate sequences for the three optimized mRNA sequences of TRXB protein in Figure 9b, and an exact derivation of χ vs Δ relation (PDF)

CHARMM topology files (TXT)

CHARMM topology files (TXT)

■ AUTHOR INFORMATION

Corresponding Author

*epo2@psu.edu

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank William Noid, Philip Bevilacqua, Güenter Kramer, and the O'Brien Lab members, especially Dan Nissley and Fabio Trovato, for a critical reading of the manuscript. We also thank Robert Best and William Swope for a useful discussion on Markov state modeling. This work was supported in part by an HFSP Research Grant and Penn State start up funds.

■ REFERENCES

(1) Nissley, D. A.; O'Brien, E. P. *J. Am. Chem. Soc.* **2014**, *136*, 17892−17898.
(2) Komar, A. A. *Trends Biochem. Sci.* **2009**, *34*, 16−24.
(3) Kramer, G.; Boehringer, D.; Ban, N.; Bukau, B. *Nat. Struct. Mol. Biol.* **2009**, *16*, 589−597.
(4) Sauna, Z. E.; Kimchi-Sarfaty, C. *Nat. Rev. Genet.* **2011**, *12*, 683−691.
(5) Pechmann, S.; Chartron, J. W.; Frydman, J. *Nat. Struct. Mol. Biol.* **2014**, *21*, 1100−1105.
(6) Zhang, D.; Shan, S. *J. Biol. Chem.* **2012**, *287*, 7652−7660.
(7) Spencer, P. S.; Siller, E.; Anderson, J. F.; Barral, J. M. *J. Mol. Biol.* **2012**, *422*, 328−335.
(8) Komar, A. A.; Lesnik, T.; Reiss, C. *FEBS Lett.* **1999**, *462*, 387−391.
(9) Pechmann, S.; Frydman, J. *Nat. Struct. Mol. Biol.* **2013**, *20*, 237−243.
(10) Hu, S.; Wang, M.; Cai, G.; He, M. *J. Biol. Chem.* **2013**, *288*, 30855−30861.
(11) Siller, E.; DeZwaan, D. C.; Anderson, J. F.; Freeman, B. C.; Barral, J. M. *J. Mol. Biol.* **2010**, *396*, 1310−1318.
(12) Zhang, G.; Hubalewska, M.; Ignatova, Z. *Nat. Struct. Mol. Biol.* **2009**, *16*, 274−280.
(13) Sander, I. M.; Chaney, J. L.; Clark, P. L. *J. Am. Chem. Soc.* **2014**, *136*, 858−861.
(14) Goder, V.; Spiess, M. *EMBO J.* **2003**, *22*, 3645−3653.
(15) Ruiz-Canada, C.; Kelleher, D. J.; Gilmore, R. *Cell* **2009**, *136*, 272−283.
(16) Angov, E. *Biotechnol. J.* **2011**, *6*, 650−659.
(17) Angov, E.; Hillier, C. J.; Kincaid, R. L.; Lyon, J. A. *PLoS One* **2008**, *3*, e2189.
(18) Doerfel, L. K.; Wohlgemuth, I.; Kothe, C.; Peske, F.; Urlaub, H.; Rodnina, M. V. *Science* **2013**, *339*, 85−88.
(19) Martens, A. T.; Taylor, J.; Hilser, V. J. *Nucleic Acids Res.* **2015**, *43*, 3680−3687.
(20) Caniparoli, L.; O'Brien, E. P. *J. Chem. Phys.* **2015**, *142*, 145102.
(21) O'Brien, E. P.; Christodoulou, J.; Vendruscolo, M.; Dobson, C. M. *J. Am. Chem. Soc.* **2011**, *133*, 513−526.
(22) O'Brien, E. P.; Christodoulou, J.; Vendruscolo, M.; Dobson, C. M. *J. Am. Chem. Soc.* **2012**, *134*, 10920−10932.
(23) Takasu, H.; Jee, J. G.; Ohno, A.; Goda, N.; Fujiwara, K.; Tochio, H.; Shirakawa, M.; Hiroaki, H. *Biochem. Biophys. Res. Commun.* **2005**, *334*, 460−465.
(24) Best, R. B.; Chen, Y. G.; Hummer, G. *Structure* **2005**, *13*, 1755−1763.
(25) Spitzer, J. J.; Poolman, B. *Trends Biochem. Sci.* **2005**, *30*, 536−541.
(26) Ueda, Y.; Taketomi, H.; Gō, N. A., II. *Biopolymers* **1978**, *17*, 1531−1548.
(27) Karanicolas, J.; Brooks, C. L. *Protein Sci.* **2002**, *11*, 2351−2361.
(28) Betancourt, M. R.; Thirumalai, D. *Protein Sci.* **1999**, *8*, 361−369.
(29) Brooks, B. R.; Brooks, C. L., III; Mackerell, A. D., Jr.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; C. Boresch, S.; Calfisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M. *J. Comput. Chem.* **2009**, *30*, 1545−1614.
(30) O'Brien, E. P.; Hsu, S. T. D.; Christodoulou, J.; Vendruscolo, M.; Dobson, C. M. *J. Am. Chem. Soc.* **2010**, *132*, 16928−16937.
(31) Schütte, E.; Noé, F.; Lu, J.; Sarich, M.; Vanden-Eijnden, E. *J. Chem. Phys.* **2011**, *134*, 204105.
(32) Buchete, N. V.; Hummer, G. J. *J. Phys. Chem. B* **2008**, *112*, 6057−6069.
(33) http://www.gnuplot.info/docs_4.2/gnuplot.html.
(34) Prinz, J.-H.; Hao, W.; Marco, S.; Keller, B.; Senne, M.; Held, M.; Chodera, J. D.; Schütte, C.; Noé, F. *J. Chem. Phys.* **2011**, *134*, 174105.
(35) Smithson, M. *Confidence interval*; Sage Publications, Inc.: Thousand Oaks, CA, 2003.
(36) Efron, B. *annals of statistics.* **1979**, *7*, 1−26.
(37) Leach, A. R. *Molecular modelling: principles and applications*, 2nd ed.; Pearson Prentice Hall: Essex, England, 2001.
(38) Ciryam, P.; Morimoto, R. I.; Vendruscolo, M.; Dobson, C. M.; O'Brien, E. P. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, E132−E140.
(39) Fluitt, A.; Pienaar, E.; Viljoen, H. *Comput. Biol. Chem.* **2007**, *31*, 335−346.
(40) De Sancho, D.; Muñoz, V. *Phys. Chem. Chem. Phys.* **2011**, *13*, 17030−17043.

DOI: 10.1021/jacs.5b08145
J. Am. Chem. Soc. 2016, 138, 1180−1195

(41) Sørensen, M. A.; Kurland, C. G.; Pedersen, S. *J. Mol. Biol.* **1989**, *207*, 365−377.
(42) Letzring, D. P.; Dean, K. M.; Grayhack, E. J. *RNA* **2010**, *16*, 2516−2528.
(43) Pedersen, S. *EMBO J.* **1984**, *3*, 2895−2898.
(44) O'Brien, E. P.; Vendruscolo, M.; Dobson, C. M. *Nat. Commun.* **2012**, *3*, 868.
(45) O'Brien, E. P.; Vendruscolo, M.; Dobson, C. M. *Nat. Commun.* **2014**, *5*, 2988.
(46) Pande, V. S.; Beauchamp, K.; Bowman, G. R. *Methods* **2010**, *52*, 99−105.
(47) Chodera, J. D.; Singhal, N.; Pande, V. S.; Dill, K. A.; Swope, W. C. *J. Chem. Phys.* **2007**, *126*, 155101.
(48) Klimov, D. K.; Thirumalai, D. *Phys. Rev. Lett.* **1997**, *79*, 317−320.
(49) Deane, C. M.; Saunders, R. *Biotechnol. J.* **2011**, *6*, 641−649.
(50) Thanaraj, T. A.; Argos, P. *Protein Sci.* **1996**, *5*, 1973−1983.
(51) Li, G.; Oh, E.; Weissman, J. S. *Nature* **2012**, *484*, 538−541.
(52) Kudla, G.; Murray, A. W.; Tollervey, D.; Plotkin, J. B. *Science* **2009**, *324*, 255−258.
(53) Doerfel, L. K.; Wohlgemuth, I.; Kothe, C.; Peske, F.; Urlaub, H.; Rodnina, M. V. *Science* **2013**, *339*, 85−88.
(54) Gustafsson, C.; Govindarajan, S.; Minshull, J. *Trends Biotechnol.* **2004**, *22*, 346−353.
(55) Drummond, D. A.; Wilke, C. O. *Cell* **2008**, *134*, 341−352.
(56) Mogk, A.; Huber, D.; Bukau, B. *Cold Spring Harbor Perspect. Biol.* **2011**, *3*, a004366.
(57) Bulleid, N. J.; Freedman, R. B. *EMBO J.* **1990**, *9*, 3527−3532.
(58) Oh, W. H.; Wu, C.; Kim, S. J.; Facchinetti, V.; Julien, L.; Finlan, M.; Roux, P. P.; Su, B.; Jacinto, E. *EMBO J.* **2010**, *29*, 3939−3951.
(59) Duttler, S.; Pechmann, S.; Frydman, J. *Mol. Cell* **2013**, *50*, 379−393.
(60) Zhang, X.; Shan, S. O. *Annu. Rev. Biophys.* **2014**, *43*, 381−408.
(61) Cortazzo, P.; Cervenansky, C.; Marín, M.; Reiss, C.; Ehrlich, R.; Deana, A. *Biochem. Biophys. Res. Commun.* **2002**, *293*, 537.
(62) Kimchi-Sarfaty, C.; Oh, J. M.; Kim, I.; Sauna, Z. E.; Calcagno, A. M.; Ambudkar, S. V.; Gottesman, M. M. *Science* **2007**, *315*, 525.
(63) Oh, E.; Becker, A. H.; Sandikci, A.; Huber, D.; Chaba, R.; Gloge, F.; Nichols, R. J.; Typas, A.; Gross, C. A.; Kramer, G.; Weissman, J. S.; Bukau, B. *Cell* **2011**, *147*, 1295−1308.
(64) Pop, C.; Rouskin, S.; Ingolia, N. T.; Han, L.; Phizicky, E. M.; Weissman, J. S.; Koller, D. *Mol. Syst. Biol.* **2014**, *10*, 770.
(65) Ingolia, N. T.; Ghaemmaghami, S.; Newman, J. R. S.; Weissman, J. S. *Science* **2009**, *324*, 218−223.
(66) Dana, A.; Tuller, T. *G3: Genes, Genomes, Genet.* **2015**, *5*, 73−80.
(67) Han, Y.; David, A.; Liu, B.; Magadán, J. G.; Bennink, J. R.; Yewdell, J. W.; Qian, S. B. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, 12467−12472.
(68) Nicola, N. V.; Chen, W.; Helenius, A. *Nat. Cell Biol.* **1999**, *1*, 341−345.
(69) Khushoo, A.; Yang, Z.; Johnson, A. E.; Skach, W. R. *Mol. Cell* **2011**, *41*, 682−692.
(70) Johnson, A. E. *FEBS Lett.* **2005**, *579*, 916−920.
(71) Woolhead, C.; McCormick, P. J.; Johnson, A. E. *Cell* **2004**, *116*, 725−736.
(72) Kim, S. J.; Yoon, J. S.; Shishido, H.; Yang, Z.; Rooney, L. A.; Barral, J. M.; Skach, W. R. *Science* **2015**, *348*, 444−448.
(73) Goldman, D. H.; Kaiser, C. M.; Milin, A.; Righini, M.; Tinoco, I., Jr.; Bustamante, C. *Science* **2015**, *348*, 457−460.
(74) Zhang, G.; Ignatova, Z. *PLoS One* **2009**, *4*, e5036.
(75) Huang, T.; Wan, S.; Xu, Z.; Zheng, Y.; Feng, K.; Li, H.; Kong, X.; Cai, Y.; et al. *PLoS One* **2011**, *6*, e16036.
(76) Muñoz, V.; Eaton, W. A. *Proc. Natl. Acad. Sci. U. S. A.* **1999**, *96*, 11311−11316.
(77) Gromiha, M. M.; Selvaraj, S.; Thangakani, A. M. *J. Chem. Inf. Model.* **2006**, *46*, 1503−1508.
(78) Kaiser, C. M.; Goldman, D. H.; Chodera, J. D.; Tinoco, I., Jr.; Bustamante, C. *Science* **2011**, *334*, 1723−1727.
(79) Zhou, M.; Guo, J.; Cha, J.; Chae, M.; Chen, S.; Barral, J. M.; Sachs, M. S.; Liu, Y. *Nature* **2013**, *495*, 111.
(80) In prokaryotes and eukaryotes: misfolding, glycosylation, acetylation, SEC-translocation, and SRP binding. In prokaryotes: deformylation, demethionation, and the binding of trigger factor. In eukaryotes: ubiquitination, phosphorylation, and NAC binding.

DOI: 10.1021/jacs.5b08145
J. Am. Chem. Soc. 2016, 138, 1180−1195