J|A|C|S
ARTICLES

Published on Web 01/04/2011

New Scenarios of Protein Folding Can Occur on the Ribosome

Edward P. O'Brien,[†] John Christodoulou,[‡] Michele Vendruscolo,[†] and Christopher M. Dobson*,[†]

Department of Chemistry, Lensfield Road, University of Cambridge, Cambridge CB2 1EW, United Kingdom, and Institute of Structural and Molecular Biology, University College London and Birkbeck College, University of London, Gower Street, London WC1E 6TB, United Kingdom

Received September 9, 2010; E-mail: cmd44@cam.ac.uk.

Abstract: Identifying and understanding the differences between protein folding in bulk solution and in the cell is a crucial challenge facing biology. Using Langevin dynamics, we have simulated intact ribosomes containing five different nascent chains arrested at different stages of their synthesis such that each nascent chain can fold and unfold at or near the exit tunnel vestibule. We find that the native state is destabilized close to the ribosome surface due to an increase in unfolded state entropy and a decrease in native state entropy; the former arises because the unfolded ensemble tends to behave as an expanded random coil near the ribosome and a semicompact globule in bulk solution. In addition, the unfolded ensemble of the nascent chain adopts a highly anisotropic shape near the ribosome surface and the cooperativity of the folding-unfolding transition is decreased due to the appearance of partially folded structures that are not populated in bulk solution. The results show, in light of these effects, that with increasing nascent chain length folding rates increase in a linear manner and unfolding rates decrease, with larger and topologically more complex folds being the most highly perturbed by the ribosome. Analysis of folding trajectories, initiated by temperature quench, reveals the transition state ensemble is driven toward compaction and greater native-like structure by interactions with the ribosome surface and exit vestibule. Furthermore, the diversity of folding pathways decreases and the probability increases of initiating folding via the N-terminus on the ribosome. We show that all of these findings are equally applicable to the situation in which protein folding occurs during continuous (non-arrested) translation provided that the time scales of folding and unfolding are much faster than the time scale of monomer addition to the growing nascent chain, which results in a quasi-equilibrium process. These substantial ribosome-induced perturbations to almost all aspects of protein folding indicate that folding scenarios that are distinct from those of bulk solution can occur on the ribosome.

Introduction

To what degree does protein folding in the cell resemble protein folding in bulk solution? This question is crucial because most of our quantitative understanding of how proteins self-assemble into their structurally ordered, functional forms comes from studies of the refolding of full length denatured proteins in bulk solution where cellular components are absent.[1−9] By contrast, in the cell protein folding is initiated, at least to some degree, on the ribosome as protein synthesis progresses.[10−12] Differences in folding between these two cases will determine the extent to which our current understanding of protein folding may need to be modified, and whether new scenarios of folding remain to be discovered.

In the process of translation, the ribosome, a 2.5 MDa (in *Escherichia coli*) molecular machine, converts the genetic information encoded in mRNA into a nascent polypeptide chain. As the nascent chain is synthesized, it is threaded through a 80−100 Å long tunnel[13] within the 50S subunit and out into the cellular milieu; the nascent chain can therefore in principle begin to fold while still tethered to tRNA located at the A or P site (Figure 1a), with little or no interactions with its C-terminus which is enclosed within the tunnel or is still to be synthesized. It has been found that a number of relatively large (>100 amino acids (AA)) topologically complex proteins can begin to form tertiary structure while inside the exit tunnel vestibule (the last

[†] University of Cambridge.
[‡] University of London.

(1) Jackson, S. E.; Fersht, A. R. *Biochemistry* **1991**, *30*, 10428–10435.
(2) Go, N. *Adv. Biophys.* **1984**, *18*, 149–164.
(3) Makhatadze, G. I.; Privalov, P. L. *Adv. Protein Chem.* **1995**, *47*, 307–425.
(4) Jackson, S. E. *Folding Des.* **1998**, *3*, R81–R91.
(5) Bryngelson, J. D.; Onuchic, J. N.; Socci, N. D.; Wolynes, P. G. *Proteins: Struc., Funct., Genet.* **1995**, *21*, 167–195.
(6) Dobson, C. M.; Sali, A.; Karplus, M. *Angew. Chem.* **1998**, *37*, 868–893.
(7) Dobson, C. M. *Nature* **2003**, *426*, 884–890.
(8) Dill, K. A.; Ozkan, S. B.; Shell, M. S.; Weikl, T. R. *Ann. Rev. Biophys.* **2008**, *37*, 289–316.
(9) Thirumalai, D.; O'Brien, E. P.; Morrison, G.; Hyeon, C. *Annu. Rev. Biophys.* **2010**, *39*, 159–183.
(10) Gething, M. J.; Sambrook, J. *Nature* **1992**, *335*, 33–45.
(11) Hartl, F. U.; Hayer-Hartl, M. *Science* **2002**, *295*, 1852–1858.
(12) Kramer, G.; Boehringer, D.; Ban, N.; Bakau, B. *Nat. Struc. Mol. Biol.* **2009**, *16*, 589–597.
(13) Voss, N. R.; Gerstein, M.; Steitz, T. A.; Moore, P. B. *J. Mol. Biol.* **2006**, *360*, 893–906.

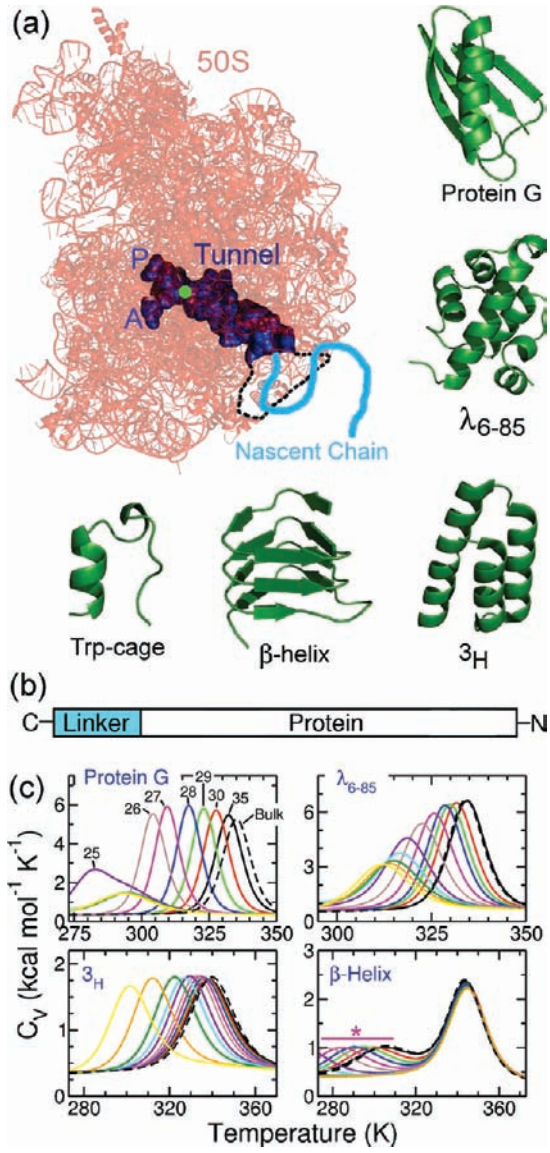

Figure 1. (a) The ribosome nascent chain complex of the 50S subunit, exit tunnel (dark blue),[13] nascent chain (light blue), peptidyl transfer center (green dot), and A and P-sites for tRNA binding are labeled. The exit tunnel vestibule region is outlined by black dashes. The native structure of the five proteins examined in this study are also shown. λ_{6-85} corresponds to monomeric λ -repressor, and 3_H stands for 3-helix bundle. (b) An illustration of the polyglycine linker attached to the C-terminus of the protein of interest. (c) The heat capacity C_V of the nascent chain as a function of temperature at different linker lengths when it is bound to the ribosome. For linker lengths L from 21 AA to 30 AA, the corresponding colors are yellow, orange, dark green, cyan, violet, brown, magenta, blue, light green, and red. Gray and solid black lines correspond to $L = 32$ and 35 AA, while the bulk C_V trace is shown as a dashed black line. This color code for linker lengths is used also in Figure 4a,b. The solid magenta bar and asterisk above the low temperature phase transition for the β-helix emphasize that these unfolding temperatures correspond to that of the C-terminal repeat only (Figure S2).

20 Å of the tunnel, Figure 1a),[14,15] or emerging from the ribosome exit tunnel,[16−18] in a phenomenon referred to as cotranslational folding. By contrast, small globular proteins are likely to fold largely post-translationally, that is, after they have been released from the ribosome and diffused away from it.[19] Since the majority of proteins in organisms are greater than 280 residues in length,[20] some degree of cotranslational folding is very likely to occur for most proteins.

Experimental studies that have demonstrated that cotranslational folding occurs have largely probed only gross structural features of the emerging nascent chain through measurements of properties such as enzymatic activity, cross-linking, proteolytic cleavage, and antibody binding.[14,17,21,22] Thus, little is known about the detailed structural, thermodynamic, and kinetic properties of nascent chains as they emerge from the ribosome exit tunnel and begin to fold. Therefore, higher resolution data from a number of different ribosome nascent chain complexes (RNCs) is needed to begin to quantify in detail the differences between bulk and cotranslational folding. Promising experimental studies on RNCs using NMR,[23−26] FRET,[27] and cryo-EM methods[28,29] suggest that in the near future such high resolution data will begin to become available.

Previously, we examined the ability of a large number of different tertiary structural motifs to sterically fit and fold in the exit tunnel.[15] We demonstrated that tertiary contacts can occur in the exit vestibule for structurally disordered nascent chains, and that a large number of different ordered tertiary structures are sterically allowed to form as well. Interestingly, simple geometric arguments failed to rationalize which structures could fit in the tunnel, emphasizing the need for simulation models that resolve nascent chain behavior at the level of individual residues to fully understand cotranslational folding. Using NMR spectroscopy to probe ribosome-bound nascent chains at the residue level, we recently identified that a partially folded intermediate species is populated in a nascent chain's domain when it sits close to the exit vestibule,[25] suggesting that new scenarios of folding may occur for this protein on the ribosome.

Motivated by these experimental findings, the purpose of the present study is threefold. First, we aim to understand the potential differences between protein folding on the ribosome and in bulk solution using theoretical methods, and thereby generate predictions of the properties of RNCs that are experimentally testable. Our second aim is to identify general

(14) Kosolapov, A.; Deutsch, C. *Nat. Struc. Mol. Biol.* **2009**, *16*, 405–411.
(15) O'Brien, E. P.; Hsu, S. T. D.; Christodoulou, J.; Vendruscolo, M.; Dobson, C. M. *J. Am. Chem. Soc.* **2010**, *132*, 16928–16937.
(16) Frydman, J.; Erdjument-Bromage, H.; Tempst, P.; Hartl, F. U. *Nat. Struc. Biol.* **1999**, *6*, 697–705.
(17) Evans, M. S.; Sander, I. M.; Clark, P. L. *J. Mol. Biol.* **2008**, *383*, 683–692.
(18) Komar, A. A. *Trends Biochem. Sci.* **2009**, *34*, 16–24.
(19) Elcock, A. H. *PLoS Comput. Biol.* **2006**, *2*, e98.
(20) Brocchieri, L.; Karlin, S. *Nucleic Acids Res.* **2005**, *33*, 3390–3400.
(21) Makeyev, E. V.; Kolb, V. A.; Spirin, A. S. *FEBS Lett.* **1996**, *378*, 166–170.
(22) Tsalkova, T.; Odom, O. W.; Kramer, G.; Hardesty, B. *J. Mol. Biol.* **1998**, *278*, 713–723.
(23) Hsu, S. T. D.; Fucini, P.; Cabrita, L. D.; Launay, H.; Dobson, C. M.; Christodoulou, J. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 16516–16521.
(24) Hsu, S. T. D.; Cabrita, L. D.; Fucini, P.; Christodoulou, J.; Dobson, C. M. *J. Am. Chem. Soc.* **2009**, *131*, 8366–8367.
(25) Cabrita, L. D.; Hsu, S. T. D.; Launay, H.; Dobson, C. M.; Christodoulou, J. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 22239–22244.
(26) Cabrita, L. D.; Dobson, C. M.; Christodoulou, J. *Curr. Opin. Struct. Biol.* **2010**, *20*, 33–45.
(27) Woolhead, C. A.; McCormick, P. J.; Johnson, A. E. *Cell* **2004**, *116*, 725–736.
(28) Gilbert, R. J. C.; Fucini, P.; Connell, S.; Fuller, S. D.; Nierhaus, K. H.; Robinson, C. V.; Dobson, C. M.; Stuart, D. I. *Mol. Cell* **2004**, *14*, 57–66.
(29) Birgit, S.; Innis, C. A.; Wilson, D. N.; Gartmann, M.; Armache, J.; Villa, E.; Trabuco, L. G.; Becker, T.; Mielke, T.; Schulten, K.; Steitz, T. A.; Beckmann, R. *Science* **2009**, *326*, 1412–1415.

trends in cotranslational folding behavior by studying a variety of different RNCs. And third, we seek to provide a theoretical framework for interpreting data from current experiments describing the behavior of RNCs.

To achieve these aims, we have carried out coarse grained molecular simulations of five different RNCs in which translation of the nascent chain is arrested at 10 or more different sequence positions. Arrested constructs have been used in a number of experimental studies to provide 'snapshots' of the properties of a nascent chain as it is synthesized.[25,27,30] In our approach, we studied a folding-competent globular protein attached to the ribosome via a folding-incompetent polyglycine linker in order to explore the impact the exit tunnel vestibule and exterior surface of the ribosome have on folding. The linker length L was varied from 22 to 35 AA, which places the C-terminal residue of the globular protein either at the exit vestibule (for $L \approx 24$ AA[29]) or fully outside the tunnel (for $L = 35$ AA). The objective of the simulations was to allow reversible folding and unfolding of the protein to occur on the ribosome and enable a rigorous analysis of the thermodynamic, kinetic, and structural properties of the RNCs at equilibrium.

The results of this analysis are that the presence of the ribosome exit tunnel vestibule and surface alters significantly the properties of the nascent chain and the manner in which it folds, most prominently for nascent chains that fold into large or topologically complex native structures. The unfolded ensemble of a protein attached to the ribosome becomes expanded and anisotropic in shape, resembling a scalene ellipsoid. The native state is destabilized by the ribosome and never exceeds its stability in bulk solution; the driving force for this destabilization results from the fact that near the ribosome surface the unfolded state entropy is constant or increases while the native state entropy decreases. Furthermore, as the linker length increases unfolding rates decrease, and folding rates increase in a linear manner toward their bulk values. The cooperativity of the folding transition also decreases on the ribosome due to the appearance of partially folded N-terminal intermediates that are not populated in bulk solution, a finding that supports the hypothesis that sequential folding is promoted by the ribosome. Moreover, the transition state ensemble for some proteins becomes more compact and native-like indicating that the details of the folding process are altered by attachment to the ribosome; analysis of the distribution of folding routes after a temperature quench reveals that the ribosome decreases their diversity and increases the probability that folding will start from the N-terminus. These ribosome-induced changes in almost all aspects of protein folding indicate that the scenarios of folding on the ribosome can differ from those in bulk solution. We discuss the experimental implications of these findings.

Results and Discussion

Bulk Solution Properties of the Nascent Chains. To identify differences between protein folding in bulk solution and on the ribosome for the five proteins shown in Figure 1a, it is necessary first to characterize fundamental aspects of their behavior in the former. The five proteins we examined (Table 1) span a range of secondary structure content from all β (β-helix), mixed α/β (protein G), all α (λ_{6-85} and 3_H), to a simple helix-turn motif (Trp-cage); moreover, both globular and repeat (β-helix) protein architectures are represented in this set of proteins.

Table 1. Bulk Solution Properties of the Proteins Studied in this Work at 310 K (unless Otherwise Indicated)

	protein[a]					
property	protein G	λ_{6-85}	3_H	β -helix	Trp-cage	max RSE[b]
C.O.[c]	9.5	7.4	5.1	8.5	3.7	–
N_P[d]	56	80	54	57	20	–
T_m (K)	335.1	334.1	339.9	343.5	341.5	0.1%
γ[e]	1.05	1.12	1.21	1.53	1.50	0.5%
ΔG_{FU}[f]	−4.8	−2.7	−2.4	−0.96	−3.20	0.8%
$R_{g,F}$ (Å)[g]	11.2	12.7	12.5	13.6	7.94	0.1%
$R_{g,U}$ (Å)	18.4	14.2	14.8	16.8	10.3	0.5%
$\sigma^2_{R_{g,U}}$ (Å^2)	13.5	11.5	7.8	5.8	1.40	0.3%
a_S	211	54.4	119	193	65.3	0.8%
a_C	40.0	26.2	25.0	15.3	13.5	1.0%
k_F (ns^{-1})	0.15	0.53	1.8	1.2	78	3.0%
k_U (ns^{-1})	6.7×10^{-5}	7.0×10^{-3}	0.14	0.53	12	6.0%

[a] Note that in the bulk simulations all of the proteins have a disordered polyglycine linker of 35 AA attached to their C-terminus. [b] The maximum value, out of the set of five proteins, of the relative standard error (RSE) for the calculated property A. The RSE is calculated as $\sigma_A/([N]^{1/2}\langle A\rangle)\cdot 100\%$, where σ_A is the standard deviation in A, N is the number of independent measurements of A (typically $N = 5$), and $\langle A\rangle$ is the average value of A. [c] The absolute contact order of the protein calculated as $N^{-1}\sum^N \Delta S_{ij}$, where N is the total number of side chain − side chain contacts in the native state structure and ΔS_{ij} is the sequence separation between residues i and j that are in contact in the native state.[70] [d] Number of residues in the protein of interest. [e] $\gamma = \Delta H_{vH}/\Delta H_{Cal}$. [f] In units of kcal/mol. [g] $R_{g,F}$ and $R_{g,U}$ are the radius of gyration of the folded and unfolded states respectively. $\sigma^2_{R_{g,U}}$ is the variance in the $P(R_{g,U})$ distribution.

Proteins larger than 80 residues were not studied here because we often found it difficult with current computer resources to achieve sufficient equilibrium sampling on the ribosome.

Table 1 lists the calculated properties of each of these proteins in bulk solution including their unfolding temperature T_m (identified as the maximum in the heat capacity (C_v) versus temperature trace, Figure 1c) and native stability ΔG_{FU}, the potential energy ($\Delta E_{FU} = E_F - E_U$) and entropy ($\Delta S_{FU} = S_F - S_U$) differences between the folded (**F**) and unfolded (**U**) states, and a number of metrics quantifying the size (radius of gyration R_g), shape (as measured by the asphericity a_S and acylindricity a_C), and the variance in the size distribution ($\sigma^2_{R_{g,U}}$) of the unfolded ensemble. Single exponential folding (k_F) and unfolding (k_U) kinetics were observed upon temperature quench in bulk solution (Figure S1) and the rate constants are listed in Table 1. We define a given protein structure to be folded if it has a root mean squared deviation from the C_α atoms of the crystal structure of less than 4 Å; otherwise, it is defined as unfolded. We note that protein G's calculated ΔG_{FU} and γ values are in reasonable agreement with its experimentally measured values,[31] while λ_{6-85}'s ΔG_{FU} is less stable than its experimentally measured value by about 1.5 kcal/mol.[32] We find that such discrepancies do not alter the qualitative trends of nascent chain behavior with synthesis length. This is evidenced in part by the fact that the three largest proteins studied here exhibit the same qualitative behavior on the ribosome and that those trends are unchanging with changes in ribosome-nascent chain interaction strength (see below).

In bulk solution, the heat capacity trace of the β-helix is the only one from the set of proteins to exhibit two maxima (Figure

(30) Lu, J. L.; Deutsch, C. *Nat. Struc. Mol. Biol.* **2005**, *12*, 1123–1129.

(31) Thoms, S.; Max, K. E. A.; Wunderlich, W.; Jacso, T.; Lilie, H.; Reif, B.; Heinemann, U.; Schmid, F. X.; Franz, X. *J. Mol. Biol.* **2009**, *391*, 918–932.

(32) Sancho, D. D.; Doshi, U.; Munoz, V. *J. Am. Chem. Soc.* **2009**, *131*, 2074–2075.

1c). To identify the structural origin of these two apparent phase transitions, we determined the unfolding temperature per residue (see Methods). Figure S2 clearly demonstrates that the C_V maxima at $T \approx 305$ K corresponds to the unfolding of the C-terminal repeat of this protein, and the maxima at $T \approx 344$ K corresponds to the unfolding of the rest of the structure; this finding is consistent with experimental observations of end fraying of other repeat proteins[33,34] indicating that this explanation is reasonable. We will see below that as a result of this fraying the folding behavior of the β-helix is not significantly altered on the ribosome relative to that in bulk solution, in contrast to the findings for some of the other proteins in this work.

Coarse-Grained Simulations of the Ribosome Nascent Chain Complex Yield Robust Results. To model the RNC, we represented the *E. coli* 50S subunit of the ribosome using two interaction sites per amino acid and three (for C, T) or four (for A, G) interaction sites per RNA base, a process that yielded more than 17 000 interaction sites in the coarse grained model (Figure S3).[15] During the simulations, we held the ribosome interaction sites immobile, while the nascent chain, tethered by its C-terminus near the peptidyl transferase center (PTC, Figure 1a), was free to fluctuate between conformations. Keeping the ribosome rigid allowed simulations to be carried out on much longer time scales, yielding more precise results than would be the case if ribosome dynamics were included. However, we tested the impact of ribosome structural fluctuations on the properties of the RNC containing a Trp-cage nascent chain, and found that simulations in which the ribosome was held rigid yielded essentially identical thermodynamic and structural results as simulations of a flexible ribosome model ($R^2 \geq 0.97$, Figure S4a and S4b).

In this coarse grained model, interactions between the ribosome and a nascent chain are treated as short ranged and repulsive (i.e., the ribosome has only steric interactions with the nascent chain and is therefore inert). We tested the impact of introducing nonspecific attractive interactions between the nascent chain and ribosome on the results of simulations of the protein G-RNC by adding an attractive Lennard-Jones' term between the two components (see Methods). We observed that the trends in the thermodynamic and structural properties of the nascent chain with changes in L were the same in both the inert and attractive ribosome simulations ($R^2 \geq 0.78$, Figures S4c and S4d); thus, in these simulations, the trends with linker length are independent of such interactions.

The results presented in the following sections are based largely on simulations of arrested RNCs, although, in nature, effectively continuous translation occurs. We tested these two situations and found that they yield equivalent results for the protein G-RNC (Figures S4e and S4f). As we discuss in detail below, this equivalence arises when folding of a protein domain is fast compared to its synthesis, in which case a protein experiences a quasi-equilibrium process during continuous translation. Thus, for the proteins we study here, understanding folding on an arrested RNC is directly relevant to the process of folding during continuous translation. These results (Figure S4) therefore show that the findings presented below are robust to alterations in the model and simulation parameters and justify the use of a coarse grained model representing a rigid, inert and arrested ribosome to compare cotranslational folding and that taking place in bulk solution.

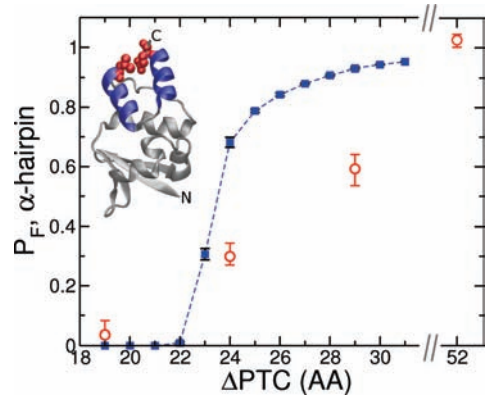

Figure 2. Comparison of the calculated (filled blue squares) and experimentally measured (open red circles) probability that the α-hairpin be folded (P_F, α-hairpin) as a function of ΔPTC for the Kv1.2 protein at 297 K. ΔPTC corresponds to the number of nascent chain residues separating the PTC from the C-terminal residue of the α-hairpin. The crystal structure of Kv1.2 is shown with the α-hairpin colored blue. Experimental data were taken from cross-linking measurements (Figure 3 in ref 14) between cysteines that were engineered into the two residue positions shown in a space-filling representation on the Kv1.2 structure. To obtain P_F, α-hairpin from the experiments, we rescaled the data by the cross-linking probability values at ΔPTC = 3 and 52 AA. In the simulations of Kv1.2, α -hairpin structures were considered folded when the backbone C_α interaction sites were within 2.5 Å root mean square deviation from the crystal structure. The dashed blue line is to guide the eye.

Simulation Results Are Consistent with Experimental Data. To validate our RNC model against experimental data, we simulated the Kv1.2 protein (Figure 2, inset) arrested on the ribosome at different nascent chain lengths. Kv1.2 is 99 residues in length and consists of both α-helical and β-strand content. The Kv1.2 protein was investigated because intranascent chain cross-linking data has been reported for the sequence homologous Kv1.3 protein.[14] A crystal structure exists for Kv1.2[35] but not for Kv1.3; however, the 96% sequence identity between the two forms of this protein indicates that they are highly likely to share the same native state structure and yield similar experimental and simulation results.

Experimentally, it has been found that a native α-hairpin tertiary structure located at the C-terminus of Kv1.3 can form on the ribosome with as few as 19 AA separating the C-terminal residue of the α-hairpin from the PTC, and that the midpoint of folding stability of this α-hairpin occurs with approximately 27 AA separating the α-hairpin's C-terminal residue from the PTC.[14] We calculated both of these quantities from our simulations, identifying the shortest linker length at which this α-hairpin was found to be in a folded conformation, and the midpoint linker length as the point at which it is folded approximately half of the time. At 310 K, we find the α-hairpin can first fold with 21 AA separating it from the PTC, and its midpoint of stability occurs at ≈24 AA (Figure 2). These findings are in good agreement with the experimental values and strongly supports the validity of our model and its predictions.

Native State Stability and Unfolding Temperatures Decrease near the Ribosome Surface. To examine the effect of the ribosome on the **F** ⇌ **U** transition of the nascent chain while it is arrested on the ribosome, we covalently linked a folding

(33) Cortajarena, A. L.; Mochrie, S. G. J.; Regan, L. *J. Mol. Biol.* **2008**, *379*, 617–626.
(34) Zweifel, M. E.; Leahy, D. J.; Hughson, F. M.; Barrick, D. *Protein Sci.* **2003**, *12*, 2622–2632.
(35) Minor, D. L.; Lin, Y. F.; Mobley, B. C.; Avelar, A.; Jan, Y. N.; Jan, L. Y.; Berger, J. M. *Cell* **2000**, *102*, 657–670.

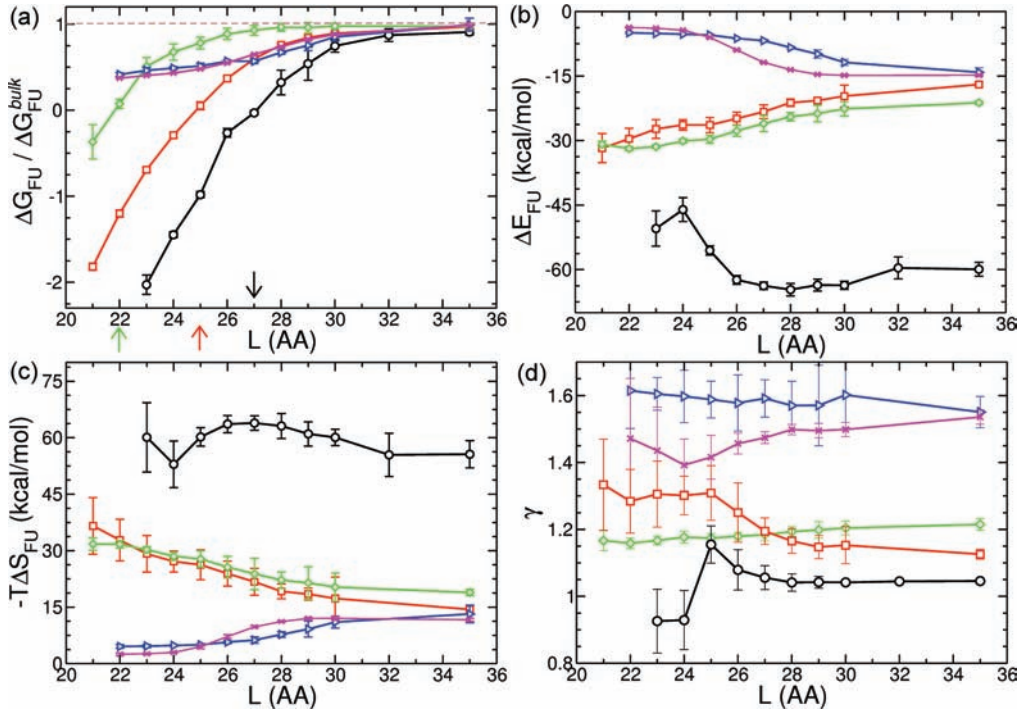

Figure 3. Thermodynamic properties of the five proteins in Figure 1a at different linker lengths. Data for protein G, λ_{6-85}, 3_H, β-helix, and Trp-cage are shown, respectively, as black circles, red squares, green diamonds, blue triangles, and magenta × marks. Brown dashed lines indicate when the thermodynamic property under consideration equals its value in bulk solution listed in Table 1. (a) The native state stability $\Delta G_{FU}(= -k_BT \ln(P_F/1 - P_F)$, where k_B is Boltzmann's constant, T is the temperature, and P_F is the probability of a protein being folded) divided by its bulk value ΔG_{FU}^{bulk} at 310 K. The L_m values (i.e., the L value at which $\Delta G_{FU} \approx 0$) for protein G, λ_{6-85}, and 3_H are indicated by the black, red, and green arrows, respectively. (b) The potential energy difference ΔE_{FU} between the folded and unfolded ensembles. (c) The entropy difference, $-T\Delta S_{FU}$ between the folded and unfolded ensembles. (d) The ratio γ of the van 't Hoff energy to the calorimetric energy at the unfolding temperature $T_m(L)$ (Figure S5).

incompetent polyglycine linker sequence to the C-terminus of the folding competent protein of interest (Figure 1b). The use of such linker sequences is common in experimental studies of RNCs[25,30] in part because it allows for the study of the behavior of a full length protein sequence at the exit tunnel vestibule and indeed far outside the tunnel (Figure 1a).

We find that as L increases, and the center-of-mass of the extruded portion of the nascent chain moves further away from the ribosome surface, the location of the maximum in C_V increases for all the proteins examined (Figure 1c), although for the β-helix only the low temperature phase transition (i.e., unfolding of the C-terminal end repeat) exhibits this behavior; the high temperature phase transition (i.e., unfolding of the N-terminal repeat of the β-helix, Figure S2b) is largely unchanged. Extracting the unfolding temperatures (T_m) from these traces reveals that they tend toward their bulk values (i.e., $T_m/T_m^{bulk} \rightarrow 1$) as L increases (Figure S5). The largest relative change in T_m occurs for the Trp-cage protein, which exhibits a 23% decrease relative to bulk at $L = 22$ AA, while the T_m of the β-helix (which we associate with unfolding of the N-terminal repeats) is largely unchanged for the L values examined here.

In a similar manner, the native state of the proteins is destabilized near the ribosome surface. As L increases, ΔG_{FU} approaches but never exceeds the bulk stability (Figure 3a). This finding contrasts with the theoretical prediction that the ribosome, when treated as an inert wall, will enhance native state stability due to excluded volume (crowding) interactions (see below).[36] From Figure 3a, the linker lengths corresponding to the midpoints of stability, that is, the L values (denoted L_m) at which $\Delta G_{FU}(L) \approx 0$, can be extracted. The native states of the proteins 3_H, λ_{6-85}, and protein G become thermodynamically stable when $L \geq 22$, 25, and 27 AA, respectively; the L_m values for nascent chains containing the β-helix and Trp-cage are, however, smaller than the shortest linker lengths examined here.

The reason that the L_m of the Trp-cage is so low can be attributed to the fact that its small size ($R_{g,F} = 7.9$ Å, Table 1) allows it to fit in to the ribosome exit vestibule in its folded conformation (Figure S6). The low L_m of the β-helix does not, however, result from its small size (its $R_{g,F} = 13.6$ Å), but because of the unstructured nature of its C-terminus. The 23 residues closest to the C-terminus of the β-helix are predominately unstructured at 310 K (Figure S2), which means the effective linker length at $L = 22$ AA is more like 45 AA (=22 + 23 AA). Thus, the folded N-terminal portion of the β-helix is not in as close proximity to the ribosome surface as might be expected from the L value alone and is therefore minimally affected by the presence of the ribosome.

To understand the origin of this native state destabilization by the ribosome, we dissect the free energy into its potential energy, ΔE_{FU}, and entropic, ΔS_{FU}, components as a function of L. Figure 3b,c reveals that, for λ_{6-85}, 3_H, and protein G (for $L > 26$), the destabilization is entropic in origin, that is, there is a greater entropic penalty for folding near the ribosome surface than in bulk solution. This fact is evident from the surprising observation that as L decreases the potential energy difference between **F** and **U** actually becomes more not less favorable toward **F** for the largest proteins (Figure 3b) and that the entropic penalty for folding, as quantified by $-T\Delta S_{FU}$, becomes larger when the protein is located closer to the ribosome surface (Figure 3c).

(36) Zhou, H. X.; Dill, K. A. *Biochemistry* **2001**, *40*, 11289–11293.

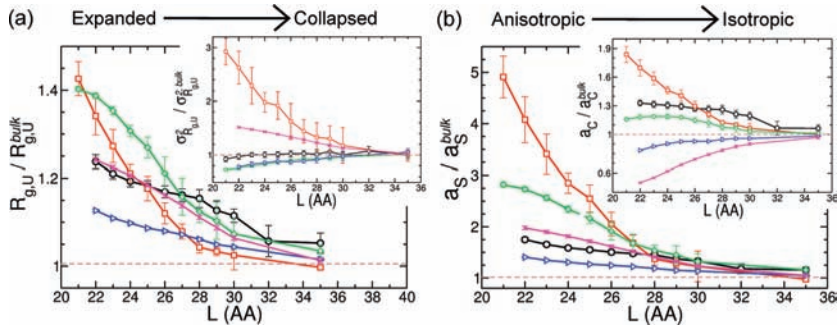

Figure 4. The behavior of the unfolded ensemble of the nascent chain when arrested on the ribosome at various linker lengths at 310 K. Figure 3's color code for the proteins is used here. (a) The radius of gyration of the unfolded ensemble $R_{g,U}$ scaled by its bulk solution value. The inset shows the variance in the unfolded ensemble $R_{g,U}$, denoted $\sigma^2_{R_{g,U}}(\equiv \langle R^2_{g,U}\rangle - \langle R_{g,U}\rangle^2)$. (b) The asphericity $a_S(\equiv \lambda_2^2 - 0.5(\lambda_0^2 + \lambda_1^2))$ and acylindricity $a_C(\equiv \lambda_0^2 - \lambda_1^2)$ of the unfolded ensemble.

There is an Entropic Penalty for Folding on the Ribosome Originating in Part from Structural Changes in the Unfolded State. From the perspective of simple excluded volume arguments, it is a surprising and counterintuitive result that the ribosome would increase the entropic cost of folding. In particular, excluded volume arguments would posit that near an inert surface, as which to a first approximation the ribosome can be treated, the equilibrium between **U** and **F** ensembles will be shifted toward compact native-like conformations because unfolded conformations, which are larger in size, will be disfavored due to steric conflict with the wall. Indeed, a theoretical study has concluded that the native state is *stabilized* near an inert wall due to entropic destabilization of the *unfolded* state.[36] In our simulations we find the opposite result, the native state is *destabilized* due to an increase in the entropic cost of folding into the *native* state.

Three hypotheses may explain this surprising result. With decreasing L: (i) the unfolded state entropy $S_U(L)$ is constant and the native state entropy $S_F(L)$ decreases; (ii) $S_F(L)$ is constant and $S_U(L)$ increases; or (iii) $S_U(L)$ and $S_F(L)$ both decrease, with $S_F(L)$ decreasing faster than $S_U(L)$. We tested these hypotheses by calculating $TS_U(L)$ and $TS_F(L)$ as a function of L from our simulations and plot the results in Figure S7. Remarkably, the unfolded state entropy for protein G, and 3_H stay relatively constant with changes in L, while for λ_{6-85} it increases significantly close to the ribosome (Figure S7a). The native state entropy tends to decrease (Figure S7b) for these three proteins at least over some range of L values. Thus, we find that a combination of hypotheses (i) and (ii) explains the ribosome-induced entropic destabilization of the native state.

There is still a question though in structural terms as to what can explain this constant or increased unfolded state entropy on the ribosome when we would normally expect the nascent chain to lose configurational entropy near an inert surface due to steric clashes with the ribosome. The answer would appear to be that the unfolded state of the nascent chain undergoes a structural transition from an expanded to a collapsed set of conformations as L increases from 22 to 35 AA, a phenomenon that is not included in the polymer theory treatment of this situation.[36] This coil to globule-like transition in the unfolded ensemble is revealed by the fact that the characteristic size of **U**, as reported by its average radius of gyration $R_{g,U}$, is much larger close to the ribosome than it is when further away (Figure 4a). More important and relevant, however, is that polymers behaving as expanded random coils with relatively high potential energies have greater configurational entropy than they do when behaving as collapsed globules that populate low potential energy states. This result is indicated by the fact that for λ_{6-85} the variance in the probability distribution along $R_{g,U}$ ($\sigma^2_{R_{g,U}}$) is larger when it is close to the ribosome than when it is further away (Figure 4a, inset), which indicates that a greater diversity of unfolded state conformations is populated close to the ribosome surface. For protein G and 3_H, however, $\sigma^2_{R_{g,U}}$ is largely independent of L, indicating the diversity of unfolded state conformations is unchanged close to the ribosome (Figure 5).

If this transition from expanded to collapsed structures in the unfolded ensemble accounts for the increased unfolded state entropy of λ_{6-85} observed as L is reduced from 29 to 21 AA in Figure S7a, then the *apparent* unfolded state entropy, computed from the probability distribution of $R_{g,U}$ ($P(R_{g,U})$), should correlate with the *true* unfolded state entropy calculated thermodynamically. The physical basis for this hypothesis results from the fact that the fields of Thermodynamics and Statistical Mechanics offer two different but equivalent ways in which to calculate the entropy of **U**. If the average potential energy (E_U) and free energy (G_U) are known, then Thermodynamics tells us $S_U(L) = (E_U(L) - G_U(L))/T$. In Statistical Mechanics, if every unfolded conformation and its corresponding potential energy (E_i) is known, then $S_U(L) = -k_B\sum_{i\in U} P_i \ln(P_i)$, where the sum is over each conformation i in the unfolded ensemble ($i \in$ U) and the probability of observing that conformation is given by the Boltzmann distribution at temperature T, $P_i = e^{-E_i/(k_BT)}$. To test the hypothesis stated above, instead of calculating $S_U(L)$ using P_i, we need to calculate $S_U(L)$ using $P(R_{g,U})$, that is $S_U^{app}(L) = -k_B\sum_{R_{g,min}}^{R_{g,max}} P(R_{g,U}) \ln(P(R_{g,U}))$. In rewriting the equation in this way, we are now calculating an *apparent* entropy (denoted S_U^{app}) because we are no longer using the microstate probabilities P_i but instead we are using the macrostate probabilities ($P(R_{g,U})$), for which a large number of unfolded conformations correspond to the same approximate $R_{g,U}$ value. A practical consequence of this is that $S_U^{app}(L) \neq S_U(L)$, although the trends in $S_U^{app}(L)$ versus L are still physically meaningful.

Calculating $S_U^{app}(L)$, and $S_U(L)$ from the thermodynamic definition, we find there is a near perfect correlation between these two quantities for λ_{6-85}-RNC ($R^2 = 0.99$, Figure S8a). This correlation tells us that it is the coil to globule-like transition in the unfolded ensemble that explains why $S_U(L)$ increases close to the ribosome for λ_{6-85}. On the other hand, the shape of $P(R_{g,U})$ for the protein G-RNC at $L = 25$ and 35 AA is largely unchanged (Figure S8b), which means that its apparent entropy does not change significantly with L. Thus, although the unfolded state of protein G expands in size near the ribosome surface (Figure 4a), its unfolded state entropy stays relatively

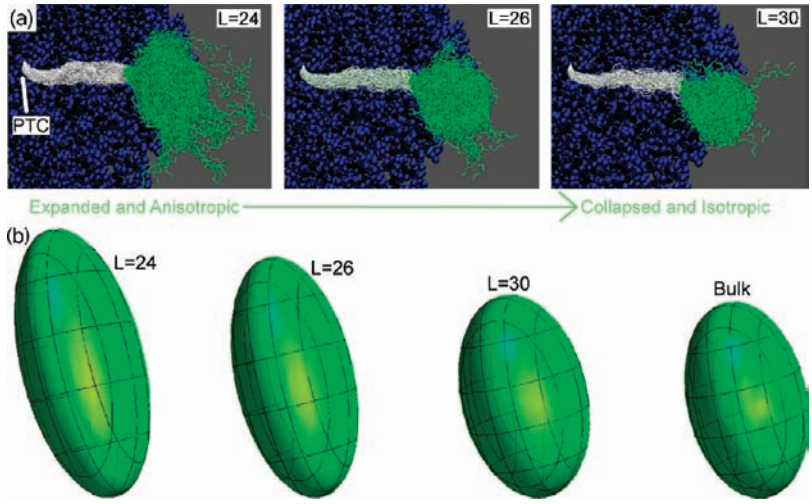

Figure 5. Visualization of the unfolded ensemble of the nascent chain when arrested on the ribosome at various linker lengths at 310 K. (a) 100−500 randomly selected simulation structures of unfolded λ_{6-85} on the ribosome with linker lengths, from left to right, of 24, 26, and 30 AA, respectively. A cut-away of the ribosome is provided to view the nascent chain within the exit tunnel. The linker residues are colored white while the λ_{6-85}'s residues are colored green. (b) Ellipsoid representations of the unfolded structures of λ_{6-85} whose radii correspond to the gyration tensor eigenvalues λ_0, λ_1, and λ_2. From left to right the ellipsoids were calculated for λ_{6-85} unfolded structures at $L = 24$, 26, 30 AA and bulk, respectively.

constant as the loss of highly compact conformations is offset by a near equal gain in highly expanded conformations.

Taken together, these data demonstrate that the native state is destabilized relative to the unfolded state near the ribosome surface, which reduces the unfolding temperature of the protein molecule. For large or more topologically complex proteins, this destabilization is primarily caused by an increased entropic penalty for folding. The increased entropic penalty arises because the native state tends to lose entropy near the ribosome surface and the unfolded state entropy stays either constant or increases with decreasing L; this increased entropy in **U** is related to an expanded to collapsed transition in the unfolded state of the nascent chain. Such native state destabilization, having either entropic and energetic origins, has been observed for proteins tethered to flat, inert surfaces.[37] The essential elements that the current polymer theory of folding near an inert wall lacks, but these simulations capture, are that the entropy of the native state can change and that an aqueous solution such as the cytoplasmic environment may be a poor solvent for the unfolded state. This means the unfolded state should be modeled as a polymer in a poor solvent[38] instead of a polymer in a theta solvent (i.e., a Gaussian chain).

The Cooperativity of Folding Is Decreased on the Ribosome Due to the Appearance of Partially Folded N-Terminal Structures. To examine whether the cooperativity of the **F** ⇌ **U** transition is altered by the ribosome, we plot the ratio γ ($\equiv \Delta H_{vH}/\Delta H_{Cal}$) of the van 't Hoff enthalpy (ΔH_{vH}) divided by the calorimetric enthalpy (ΔH_{Cal}) as a function of L (Figure 3d). As interpreted in bulk protein folding studies, a value of $\gamma = 1.0$ suggests that the **F** ⇌ **U** transition is two-state in nature with negligibly populated intermediate states.[1] If values of γ are greater than 1.0, the folding transition is said to be less cooperative and less two-state as a result of a greater population of intermediate states between **F** and **U**.

Figure 3d indicates that for the smallest proteins studied here (Trp-cage and 3_H), and the β-helix, γ is largely unchanged from its bulk value (the error bars suggest their is no meaningful change in $\gamma(L)$ relative to bulk), and therefore, the ribosome does not significantly perturb the cooperativity of the folding of these proteins. For protein G and λ_{6-85}, which are among the largest and most topologically complex proteins of the proteins examined in this study (Table 1), γ exhibits an increase with decreasing L. For λ_{6-85}, γ increases from 1.12 in bulk to 1.35 at $L = 21$ AA, and for protein G, γ increases from 1.05 in bulk to 1.16 at $L = 24$ AA. We note that the γ values for both λ_{6-85} and protein G deviate significantly from their values in bulk only at L values around or below their L_m (Figure 3d). For example, the L_m values of protein G and λ_{6-85} are, respectively, 27 and 25 AA (Figure 3a), yet their γ values deviate by less than 10% from the bulk values until $L = 25$ AA for protein G and $L = 26$ AA for λ_{6-85}. This finding indicates that intermediate states during cotranslational folding are more populated around $L \approx L_m$ than at $L \gg L_m$, or in bulk solution, and is consistent with the experimental observation of a partially folded intermediate structure of an Ig-like domain when $L = 21$ AA.[25]

To test this hypothesis that intermediates may be populated on the ribosome, and to gain insight into the structural origin of the behavior of γ, we plot the free energy profile $F(Q)$ versus the fraction of native contacts Q formed by the nascent chains at various linker lengths (Figure 6). Q is a probe of the extent of native structure formation in the nascent chain (Methods); a value of zero means no native contacts are formed, while a value of unity means all native contacts are present. Figure 6a clearly shows that at $L = 24$ AA, protein G populates a metastable intermediate state between **U** and **F**, which manifests itself as a local minimum in $F(Q)$ located between $0.25 < Q < 0.5$; this intermediate state has more native structure than **U**, as indicated by its location along the Q-axis, and becomes unstable at $L \geq 29$ AA. Likewise, λ_{6-85} at $L \leq 25$ AA exhibits a local free energy minimum at intermediate values of Q ($0.4 < Q < 0.6$) that becomes unstable for $L \geq 25$ AA (Figure 6a), indicating that a partially folded ensemble is populated. For 3_H, β-helix, and Trp-cage, metastable intermediates are not observed in the free-energy profile at any value of L (Figures 6b and S9). Thus, the decrease in folding cooperativity on the ribosome is due to

(37) Friedel, M.; Baumketner, A.; Shea, J. E. *J. Chem. Phys.* **2007**, *126*, 095101.
(38) Wilkins, D. K.; Grimshaw, S. B.; Receveur, V.; Dobson, C. M.; Jones, J. A.; Smith, L. *J. Biochem.* **1999**, *38*, 16424–16431.

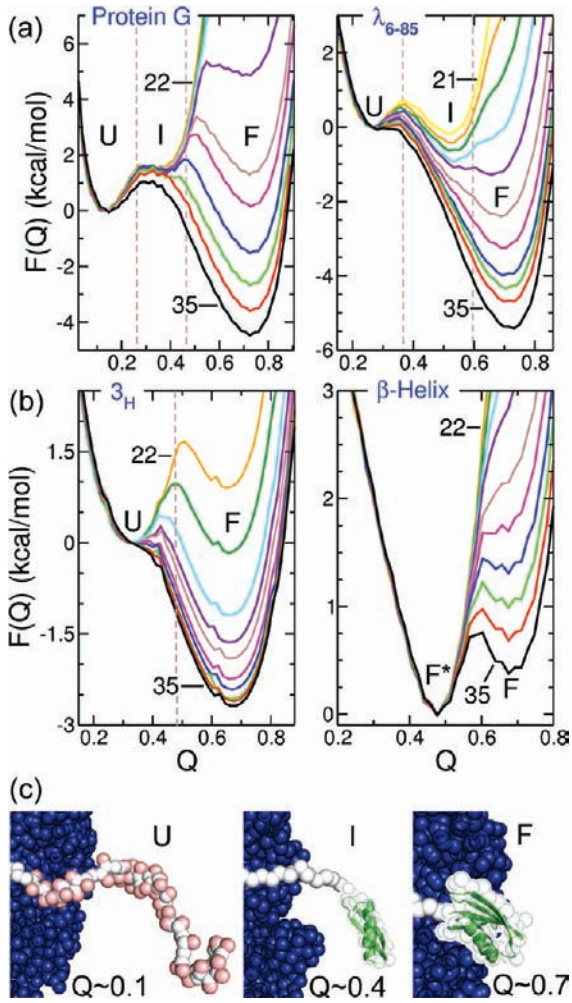

Figure 6. The free energy $F(Q) = -k_BT \ln(P(Q)) + C$ versus the fraction of native contacts Q for (a) protein G and λ_{6-85}, and (b) 3_H and β-helix, at different L values on the ribosome and 310 K. $P(Q)$ is the probability of observing a value of Q during the simulation, and C is an arbitrary offset to make comparison of $F(Q)$ easier at various L values. The color code for linker lengths in Figure 1c is used here. The nominal boundaries between the fully unfolded (**U**), partially folded intermediate (**I**), and folded (**F**) regions in $F(Q)$ are indicated by dashed brown lines for protein G and λ_{6-85}. Note for the β-helix that two folded states are identified, F* and **F**, that correspond, respectively, to the folded N-terminus and frayed C-terminus (Figure S2), and to the fully folded structure. (c) Representative simulation structures of **U**, **I**, and **F** for protein G at $L = 26$ AA. The ribosome is shown in blue, and the nascent chain in white. For clarity, the nascent chain side chain interaction sites are not shown, except for **U** where they are shown as pink spheres. Note that for **I**, the N-terminal β-hairpin forms a partially folded structure with the α-helix consistent with the data in Figure S10. The crystal structures of the folded portions of the nascent chain are superimposed (green) on the simulated structures.

the appearance of partially folded metastable states that are not populated in bulk solution.

For protein G and λ_{6-85} we can identify these intermediate, partially folded structures by calculating Q_i, the fraction of native contacts per tertiary substructure i (see Methods). Plotting Q_i versus Q in Figure S10 for protein G and λ_{6-85} at $L = 26$ and 24 AA, respectively, we find that the low values of Q in Figure 6a ($Q \approx 0.15$ to 0.25 for protein G and $Q \approx 0.20$ to 0.35 for λ_{6-85}) correspond to largely unfolded structures containing little to no residual structure (Figure 6c). Intermediate values of Q for protein G correspond to partially folded structures in which the N-terminal β-hairpin and α-helix adopt their native conformation but the C-terminal β-hairpin is unstructured and partially sequestered in the exit tunnel vestibule (Figures 6c and S10a). For λ_{6-85}, the intermediate Q-values represent a more diverse set of partially folded conformations in which different combinations of four out of five native helices located near the N-terminus form folded tertiary structure (Figure S10b). The Q values between 0.7 and 0.8 correspond to the fully folded nascent chains (Figures 6c and S10). We note that **F** typically has an average Q value below 1 due to thermally induced structural fluctuations within the native state ensemble.

When, therefore, small proteins with low topological complexity and low folding cooperativity are being synthesized, their cooperativity and any intermediates that they may populate on the ribosome are largely unchanged relative to the situation in bulk solution. However, when larger proteins with greater topological complexity and more cooperative bulk folding transitions are being synthesized by the ribosome, they can populate metastable partially folded states near the exit vestibule that are not populated in bulk solution. This finding indicates that the ribosome may promote folding pathways that fold via the N-terminus first; we show below this is indeed the case for protein G. This finding suggests that cotranslational folding pathways and bulk pathways can differ significantly for large complex proteins, such as multidomain proteins where domains can independently fold while still tethered to the ribosome.[19]

The metastable intermediates that we observe for these proteins indicate a 'weak' sequential cotranslational folding mechanism, in that once folded, these N-terminal intermediates do not necessarily stay folded. Instead, due to their metastable nature they rapidly fluctuate between unstructured and partially folded states, which can act as folding nuclei as the remainder of the nascent chain is synthesized.[39] It has been shown previously that the stability of a folded protein structure is, on average, directly proportional to the number of amino acids comprising it.[32,40] It is therefore possible that, for proteins larger than the ones studied here, partially folded N-terminal intermediates are likely to be more stable and hence a sequential cotranslational folding mechanism may become predominate for such RNCs. In addition, it is possible that proteins that have long-range contacts in their native fold may form intermediates that are richer in local contacts and non-native structure as emphasized in a lattice model study,[41] and suggested in an *in vitro* model system.[42]

The Unfolded Ensemble Expands and Becomes Highly Anisotropic when Bound to the Ribosome. We next examined whether the unfolded ensemble of the nascent chain is altered on the ribosome by calculating its characteristic size and the variance of its size distribution as quantified by its average radius of gyration $R_{g,U}$ and its variance in $R_{g,U}$, denoted $\sigma^2_{R_{g,U}}$. Figure 4a shows that the unfolded ensembles of all the proteins examined here exhibit a monotonic increase in the magnitude of their $R_{g,U}$ values as L decreases (Figure 4a), that is, when the protein is closer to the ribosome surface. Furthermore, the breadths of their $R_{g,U}$ values also increase or stay relatively constant (Figure 4a, inset). Remarkably, at the smallest L values studied here, the unfolded ensemble can become 40% larger

(39) Wang, P. Y.; Klimov, D. K. *Proteins: Struct., Funct., Bioinf.* **2008**, *70*, 925–937.
(40) Kouza, M.; Li, M. S.; O'Brien, E. P.; Hu, C. K.; Thirumalai, D. *J. Phys. Chem. A* **2006**, *110*, 671–676.
(41) Morrissey, M. P.; Ahmed, Z.; Shakhnovich, E. I. *Polymer* **2004**, *45*, 557–571.
(42) Chow, C. C.; Chow, C.; Raghunathan, V.; Huppert, T. J.; Kimball, E. B.; Cavagnero, S. *Biochemistry* **2003**, *42*, 7090–7099.

than its size in bulk solution and the variance in its size distribution can increase by more than a factor of 2 relative to the bulk value (Figure 4a, inset). Thus, the unfolded ensemble is highly expanded and can have a greater diversity of sizes when it is bound near to the ribosome surface than in bulk solution, and undergoes a collapse transition as L increases.

A more expanded unfolded ensemble is likely to expose a higher proportion of hydrophobic groups to solvent than a less expanded ensemble. To test this hypothesis, we calculated the solvent accessible surface areas of hydrophobic groups in the unfolded ensemble of protein G, λ_{6-85}, and 3_H. Figure S11 shows that in the simulation of λ_{6-85}-RNC there is a large increase in hydrophobic surface area relative to its bulk unfolded ensemble. Protein G and 3_H, on the other hand, show only slightly greater exposure of hydrophobic groups. Thus, for some nascent chains, the unfolded ensemble can acquire a much greater hydrophobic character on the ribosome, a phenomenon that could make the nascent chain even more aggregation prone under these conditions than in bulk solution.[43]

To examine whether the characteristic shape of the unfolded ensemble is altered by the ribosome, we have plotted the shape parameters of asphericity ($a_S = \lambda_2^2 - 0.5(\lambda_0^2 + \lambda_1^2)$) and acylindricity ($a_C = \lambda_1^2 - \lambda_0^2$) in Figure 4b. These quantities measure how spherical ($a_S = 0$ corresponds to a perfectly spherical distribution) and cylindrical ($a_C = 0$ corresponds to a perfectly cylindrical distribution) are the unfolded ensemble; a_S and a_C, which are functions of the $R_{g,U}$ tensor eigenvalues (denoted λ_0, λ_1, and λ_2), can be determined experimentally from NMR or SAXS data.[44,45] Figure 4b indicates that the unfolded ensembles of all the proteins studied here change shape significantly when near the ribosome surface. The unfolded ensembles become less spherical as L decreases (i.e., $a_S/a_S^{bulk} \neq 1$) and for the protein G, λ_{6-85}, and 3_H nascent chains, their unfolded ensembles also become less cylindrical (Figure 4b, inset). This means that the unfolded ensembles of these proteins become highly anisotropic, with large inequalities between the tensor eigenvalues ($\lambda_0 \neq \lambda_1 \neq \lambda_2$). The average shape of these unfolded ensembles near the ribosome surface is therefore best represented as a scalene ellipsoid. To make these shape parameters more intuitive, unfolded simulation structures of λ_{6-85} at various L values are shown in Figure 5a, with their corresponding ellipsoids (with axes equal to the tensor eigenvalues) given in Figure 5b. These two figures dramatically illustrate the size and shape changes that the unfolded ensemble undergoes during its biosynthesis on the ribosome.

For the β-helix, only a modest decrease in the unfolded state value of a_C occurs, while a_C decreases to 0.6 of its bulk value for the Trp-cage, indicating that its unfolded ensemble becomes more cylindrical near the ribosome (Figure 4b). This result is due to the very short length of this polypeptide (20 AA), which means it becomes highly extended relative to its total contour length, when it is unfolded near the ribosome surface and hence becomes more cylindrical. On the basis of these results, we conclude the unfolded ensemble of a protein can be drastically different when bound to the ribosome from that in bulk solution; its size, shape, and distribution of conformations can be highly perturbed.

The increase in unfolded ensemble size near the ribosome is consistent with polymer theory predictions that when a polymer is tethered to an inert wall its end-to-end distance increases.[46] These changes in denatured state properties affect the relative stabilities of the native states, as discussed earlier, and can lead to greater exposure of hydrophobic surface area which tend to make the nascent chain more aggregation prone. From a biological perspective, this feature of cotranslational folding is particularly dangerous in the context of polysomes, in which a string of ribosomes simultaneously translates the same mRNA strand, thereby increasing the local concentration of the nascent chain. This suggests that one of the biological roles for the chaperone Trigger Factor, which binds to RNCs and is thought to form a cradle over the emerging nascent chain,[47] is to prevent aberrant nascent chain protein aggregation by sequestering hydrophobic groups of the nascent chain from the cytosol and potentially via compaction of the unfolded ensemble through confinement effects.[48] In addition, nature appears to organize polysomes to minimize contact formation between nascent chains on neighboring ribosomes,[49] an effect that will also serve to inhibit their aggregation.

Folding Rates Decrease and Unfolding Rates Increase near the Ribosome Surface. We next examined the effect of the ribosome on the folding and unfolding kinetics of the arrested nascent chains by performing *in silico* temperature quench experiments and constant temperature simulations, respectively (see Methods). We note that it was necessary to calculate folding kinetics from temperature quench simulations because, at 310 K, **F** is overwhelmingly populated at equilibrium and therefore to few unfolding to folding transitions are observed to obtain sufficient statistics. On the ribosome, k_F and k_U of the Trp-cage, β-helix, and 3_H are only slightly different from their bulk values (Figure 7a,b), with the β-helix displaying a modest increase in k_U below $L = 28$ AA, and Trp-cage and 3_H showing slight decreases in k_F. On the other hand, protein G and λ_{6-85} (which are larger and topologically more complex) exhibit relatively slower folding rates and faster unfolding rates as L is decreased. Indeed, for protein G, λ_{6-85}, and 3_H, there appears to be a correlation between the contact order of the proteins, a measure of the topological complexity of their native folds (Table 1), and their change in k_F with L (Figure 7b); the larger the contact order, the greater the change in k_F relative to bulk. Trp-cage and the β-helix are not globular proteins, and therefore may not follow this potential trend.

We find that for all of these RNCs the change in k_F with L is linear (Figure 7b). To understand the origin of this behavior, we applied Transition State Theory to examine whether changes in the barrier height to folding, due to changes in entropy or enthalpy, can lead to this linear dependence. If we assume that Transition State Theory is accurate in the context of an RNC and that the prefactor is unchanging with L, then $k_F(L)/k_F^{bulk} = e^{-\Delta\Delta G^{TS}(L)/k_B T}$, where $\Delta\Delta G^{TS}(L) = \Delta G^{TS}(L) - \Delta G^{TS,bulk}$, and ΔG^{TS} is the difference in free energy between the transition state and unfolded state. Using the thermodynamic relation that $\Delta\Delta G^{TS}(L) = \Delta\Delta E^{TS}(L) - T\Delta\Delta S^{TS}(L)$, we have that $k_F(L)/k_F^{bulk}$

(43) Calamai, M.; Taddei, N.; Stefani, M.; Ramponi, G.; Chiti, F. *Biochemistry* **2003**, *42*, 15078–15083.
(44) Ryabov, Y.; Suh, J. Y.; Grishaev, A.; Clore, G. M.; Schwieters, C. D. *J. Am. Chem. Soc.* **2009**, *131*, 9522–9531.
(45) Svergun, D. I.; Koch, M. H. J. *Curr. Opin. Struc. Biol.* **2002**, *12*, 654–660.
(46) Sanchez, I. C. *Physics of Polymer Surfaces and Interfaces*; Butterworth-Heinemann: London, 1992.
(47) Merz, F.; Boehringer, D.; Schaffizel, C.; Preissler, S.; Hoffmann, A.; Maier, T.; Rutkowska, A.; Lozza, J.; Ban, N.; Bukau, B.; Deuerling, E. *EMBO J.* **2008**, *27*, 1622–1632.
(48) Klimov, D. K.; Newfield, D.; Thirumalai, D. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 8019–8024.
(49) Brandt, F.; Etchells, S. A.; Ortiz, J. O.; Elcock, A. H.; Hartl, F. U.; Baumeister, W. *Cell* **2009**, *136*, 261–271.

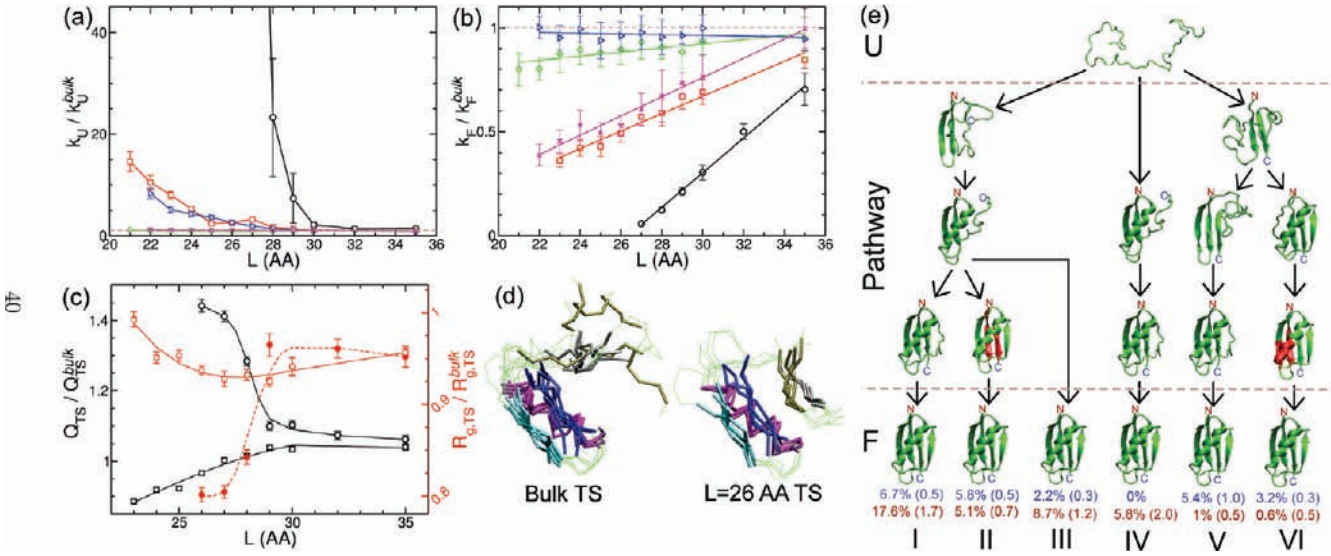

Figure 7. The unfolding rate k_U (a) and folding rate k_F (b) divided by their corresponding values in bulk solution at 310 K. Figure 3's color code for the proteins is used here. The lines are simply to guide the eye. (c) Average properties of the transition state ensemble at various L values. Circles are for protein G and squares are for λ_{6-85}; red symbols refer to the $R_{g,TS}/R_{g,TS}^{bulk}$ axis (on the right) and black symbols refer to the fraction of native contacts in the transition state ensemble (Q_{TS}/Q_{TS}^{bulk} axis on the left). The lines are again just to guide the eye and are based on a polynomial fit. (d) Transition state ensemble structures of protein G in bulk solution and on the ribosome with $L = 26$ AA. (e) Productive folding pathways, characterized by their acquisition of native tertiary structure, of protein G upon temperature quench from 360 to 310 K. The percentage of trajectories that folded via a given pathway are listed for folding in bulk solution (blue) and on the ribosome (red) with $L = 27$ AA at the bottom of the figure. The numbers in parentheses are the standard error about the mean. In pathways II and VI, some structures have a portion colored red to indicate these portions have not yet folded into their native structure (i.e., they have less than half of their native contacts present). In pathway II, the red colored structure indicates that the C and N-terminal β-strands have not yet come together in their folded structure while the rest of the structure has formed. In pathway V, the red colored structure indicates that the α-helix and N-terminal β-hairpin have not yet docked into their folded structure.

$= e^{-\Delta\Delta E^{TS}(L)/k_BT} \cdot e^{T\Delta\Delta S^{TS}(L)/k_BT}$. In the extreme case that there is little change in potential energy with L (i.e., $\Delta\Delta E^{TS}(L) \ll k_BT$), then $k_F(L)/k_F^{bulk} \approx e^{T\Delta\Delta S^{TS}(L)/k_BT}$ and entropy changes are the primary origin of the linear behavior in $k_F(L)/k_F^{bulk}$. Conversely, if there is little change in entropy with L (i.e., $-T\Delta\Delta S^{TS}(L) \ll k_BT$), then $k_F(L)/k_F^{bulk} \approx e^{T\Delta\Delta S^{TS}(L)/k_BT}$ and enthalpy is the primary determinant. To examine whether entropy or potential energy changes dominate the kinetic behavior, we calculated $-T\Delta\Delta S^{TS}(L)$ and $\Delta\Delta E^{TS}(L)$ from the equilibrium simulations, which are independent and distinct from the set of temperature quench simulations that were used to calculate $k_F(L)/k_F^{bulk}$, and incorporated the results into the respective approximations to predict $k_F(L)/k_F^{bulk}$.

The results show that for protein G both the enthalpic (Figure S12a) and entropic (Figure S12b) approximations yield predicted $k_F(L)/k_F^{bulk}$ behavior that are linear with respect to L and correlate with the simulated $k_F(L)/k_F^{bulk}$ observed in Figure 7b. This indicates that for this protein a combination of both potential energy and entropy changes in the folding free-energy barrier contribute to the linear dependence of $k_F(L)/k_F^{bulk}$. On the other hand, for λ_{6-85} and 3_H, the predicted folding kinetics based on $\Delta\Delta E^{TS}(L)$ do correlate with the observed kinetics (albeit weakly in the case of 3_H), but do not correlate based on changes in $-T\Delta\Delta S^{TS}(L)$ alone, suggesting that enthalpy changes in the folding barrier give rise to the linear dependence in $k_F(L)/k_F^{bulk}$ for these two proteins.

These results therefore show that large or topologically more complex globular proteins near the ribosome surface exhibit larger changes in k_U and k_F at a given L value relative to bulk solution. k_F varies linearly with L, the origin of which can be due, depending on the protein, to either enthalpy changes alone, or a combination of enthalpy and entropy changes in the folding free energy barrier.

The Transition State Ensemble Is More Compact and Native-Like on the Ribosome. From the temperature quench simulations used to calculate k_F, we can identify the nascent chain structures that comprise the transition state ensemble at different L values (see Methods). This identification is possible because it has been shown for the class of simulation model used here that the order parameter Q accurately identifies the location where the probability of relaxing into the native basin before reaching the unfolded basin (often referred to as the *p-fold* criterion[50]) is 50%, which corresponds to the **TS**.[50] Characterizing the size and structure of these ensembles (Figure 7c), we observe that for protein G the ribosome makes the **TS** more compact (the quantity $R_{g,TS}/R_{g,TS}^{bulk}$ decreases by 20% relative to bulk solution) and significantly increases the amount of native-structure (Q_{TS}/Q_{TS}^{bulk} increases by up to 40%).

Examining the probability of the formation of individual tertiary substructures within the **TS** (Figure S13) reveals that all such elements become more ordered as a result of the presence of the ribosome. This finding is reflected in the **TS** structures shown in Figure 7d of protein G, where the C-terminal hairpin can be seen clearly to be more structured on the ribosome compared to when it is in bulk solution. These structural changes with L explain the increased **TS** compaction and structure formation observed in Figure 7c. Surprisingly, when we couple this finding with the fact that the midpoint linker length of protein G is $L_m = 27$ AA (Figure 3a), we conclude that it is likely that half the nascent chains of this RNC will fold through this highly perturbed transition state during continuous translation (under the quasi-equilibrium assumption, that we demonstrate below is a valid assumption for this construct). On the other hand, the **TS** of λ_{6-85} at most L values is only slightly

(50) Cho, S. S.; Levy, Y.; Wolynes, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 586–591.

more compact than in bulk and does not exhibit significant changes in native structure content except when sitting close ($L \leq 24$ AA) to the ribosome surface where a 10% decrease is observed relative to bulk. Thus, while compaction of the **TS** is a common feature on the ribosome, the impact on structure formation can differ between proteins.

The Ribosome Decreases the Diversity of Folding Pathways and Promotes the Initiation of Folding via the N-Terminus. The **TS** conformations (Figure 7d), while indicating whether a given tertiary structure is or is not formed, do not tell us the routes a protein takes while transiting the hyperdimensional free energy surface between **U** and **F**.[5] To determine the impact of the ribosome on these folding routes, and their respective probabilities, we analyzed the temperature quench simulations of the protein G-RNC and λ_{6-85}-RNC. For each folding trajectory, we determined the order in which different native tertiary substructures are acquired during the **U** to **F** transition. We are thus coarse graining the folding pathway from the time series of $3N$ Cartesian coordinates, where N is the number of interaction sites in the system, that uniquely defines each folding trajectory, and projecting it on to a collective variable, which is a function of the number of native tertiary substructures and their type. In addition, because folding is a stochastic process, involving many recrossings between states of greater and lesser tertiary structure, we only keep track of those transitions that increase the total fraction of native structure during the simulations (see Methods for full details). This approach reduces the complexity of representing productive folding[51] and is analogous to keeping track of only the forward flux in network models of folding.[52]

When we perform this analysis for protein G in bulk solution, we find that there is a wide diversity of observed folding behavior; a minimum of 22 unique routes are necessary to account for 50% of the folding trajectories and 141 routes are needed to account for all of them (Figure S14). The four most probable pathways in bulk solution account for 21% of the folding trajectories and involve two branches (Figure 7e), one in which the N-terminal β-hairpin forms first and the other in which the N and C β-strands form first. For the protein G-RNC with $L = 27$ AA, however, the diversity of folding pathways decreases; this is revealed by the fact that only 10 pathways are necessary to account for 50% of the folding trajectories and 101 for all of them (Figure S14). Interestingly, the four most probable ways in which folding occurs for a nascent chain attached to the ribosome (Figure 7e) account for 34% of the folding trajectories (a net increase of 13% relative to bulk) and involve an increase in the initiation of folding via the N-terminal β-hairpin (e.g., pathway I in Figure 7e). This increase comes, however, at the expense of folding via initial structure formation in the N and C-terminal β-strands (pathways V and VI in Figure 7e). The latter in fact show decreased probabilities for the protein attached to the ribosome. The probability that the N-terminal β-hairpin folds before the C-terminal β-hairpin is 44% in bulk solution and 71% on the ribosome, while the probability that the C-terminal β-hairpin forms before the N-terminal β-hairpin is 19% in bulk solution and 15% on the ribosome.

These data show that the ribosome affects the distribution of folding routes in two ways: it decreases their diversity and biases them toward initiating folding via the N-terminus. We emphasize that this effect is *not* a consequence of the vectorial emergence of the nascent chain from the exit tunnel because the ribosome is arrested, and at $L = 27$ AA protein G is already located at the tunnel vestibule and is capable of making all of its native interactions between its N and C-terminal residues. Rather, it is a consequence both of steric interactions between the nascent chain and ribosome surface and the fact that the nascent chain is tethered to ribosome-bound tRNA. While it is natural to suggest that the reason their is a decreased diversity in folding pathways is because their is a decreased conformational space available to the protein close to the ribosomal surface, this conclusion is misleading because, as discussed above, the total entropy of the unfolded state is relatively constant for protein G on the ribosome (Figure S7a). A more accurate assessment is that new areas of configurational space become accessible to a nascent chain attached to the ribosome and other areas become restricted (Figure S8b) with the loss of highly compact conformations being offset by a gain in more expanded conformations. It is this alteration in accessible configurational space which likely contributes to changes in the distribution of folding pathways. Indeed, folding from the N-terminus of a protein has been independently observed in simulations of a two domain serine protease on the ribosome,[19] suggesting this phenomenon may be general.

Folding can be Equivalent during Arrested and Continuous Translation of Nascent Chains. Cotranslational folding of an arrested nascent chain occurs under equilibrium conditions, yet continuous translation means cotranslational folding occurs under nonequilibrium conditions; the possibility therefore exists that these different conditions may cause folding properties to differ significantly. Because many experimental studies have been reported on arrested nascent chain behavior, it is crucial to determine under what circumstances these two situations will yield equivalent results. We hypothesize that these conditions will yield approximately equivalent results when, at $L = L_m$, the folding time $\tau_F \ll \tau_S$, which means that folding and unfolding occur many times with respect to the time scale of monomer addition to the C-terminus of the nascent chain τ_S. This situation is therefore a quasi-equilibrium one, with a large separation in the characteristic time scales of monomer addition and folding. The result is that on the slowest time scales additional degrees of freedom (i.e., amino acids) are being added to the nascent chain during translation, but on much faster time scales the nascent chain is sampling all of its accessible configurations sufficiently quickly that it is effectively equilibrated at each nascent chain length, with the nascent chain properties exhibiting an equilibrium (Boltzmann) distribution at each nascent chain length. Quasi-equilibrium also implies the domain can reach its global free-energy minimum state at each nascent chain length. From experiments, we know that $\tau_S \approx 0.05$ s in *E. coli*[53] and $\tau_F = 2$ ms for protein G[54] in bulk solution; hence, $\tau_F/\tau_S = 0.04 \ll 1$ and quasi-equilibrium conditions should therefore be achieved during protein G's continuous translation.

We have tested this hypothesis by carrying out simulations of the continuous biosynthesis of the protein G-linker nascent chain on the ribosome (see Methods) and compared its properties with those calculated when protein G is arrested at different nascent chain lengths on the ribosome. In the continuous

(51) Noe, F.; Schutte, C.; Vanden-Eijnden, E.; Reich, L.; Weikl, T. R. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 19011–19016.
(52) Berezhkovskii, A.; Hummer, G.; Szabo, A. *J. Chem. Phys.* **2009**, *130*, 205102.
(53) Bremer, H.; Dennis, P. P. *Escherichia coli and Salmonella typhimurium: Cellular and Molecular Biology*, 2nd ed.; ASM Press: London, 1996.
(54) McCallister, E. L.; Alm, E.; Baker, D. *Nat. Struc. Biol.* **2000**, *7*, 669–673.

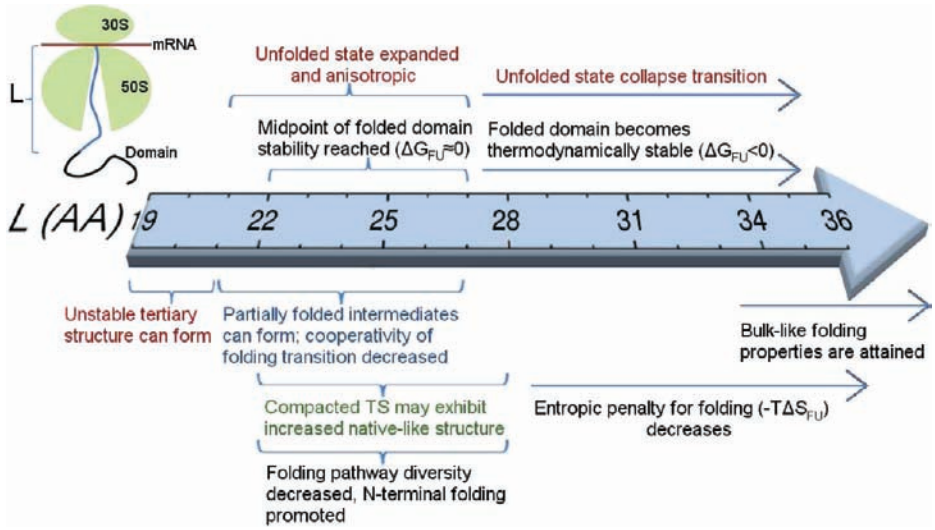

Figure 8. A schematic summary of the perspective offered by these simulations on cotranslational folding. The arrow, from left to right, indicates the length L of the folding-incompetent linker as it is increased, while the properties of the folding domain at each L value are indicated above and below this arrow. A cartoon of the ribosome is shown in the upper left and indicates the linker (blue) and folding domain (black). Note that this summary is based on a linker that adopts an extended conformation; linker sequences that adopt compact conformations will shift these results to higher L values. Furthermore, it is reasonable to expect that increasing the strength of attractive ribosome-nascent chain interactions will also shift these results to higher L values.

translation simulations, we approximate the experimentally determined ratio of time scales $\tau_F/\tau_S = 0.04$ by incorporating a new residue into the nascent chain after every 100 ns of simulation (protein G's simulated τ_F =6.7 ns, Table 1). In terms of both nascent chain size (Figure S2e) and native structure content (Figure S2f), we find excellent agreement between arrested and continuous translation. This result tells us that for fast folding protein domains, at least, the properties of the nascent chains are essentially equivalent during arrested and continuous translation. A similar result was found in lattice model simulations of cotranslational folding.[39] Such a situation will not, however, occur for proteins in which folding takes place on a time scale comparable to that of as monomer addition ($\tau_F \approx \tau_S$) as quasi-equilibrium conditions will be violated and nonequilibrium conditions will dominate. In this case, the nascent chain conformations will not be populated as a function of their inherent stabilities, folding pathways the domain takes will be dependent on the previous conformational states it populated, and the domain may reach only kinetically accessible states instead of its global free minimum. We estimate that for proteins greater than 85 AA in length this quasi-equilibrium condition is likely to be violated and therefore arrested and continuous studies of such RNCs may yield different results (O'Brien, unpublished results).

Conclusions

The results of this simulation study indicate that for large, topologically complex proteins essentially all aspects of cotranslational folding can differ from that of folding in bulk solution. The origin of these differences is the effect of the ribosome vestibule and outer surface on the nascent chain with an additional contribution from the vectorial nature of nascent chain synthesis. On the ribosome, the diversity of folding routes decreases and those involving folding of the N-terminus before the C-terminus are promoted. In addition, the denatured state expands and becomes anisotropic in nature and can take on a greater hydrophobic character. By contrast, the transition state becomes more compact and can acquire greater native-like structure, and the native state is destabilized due to a decrease in native state entropy and an increase in denatured state entropy. Furthermore, partially folded intermediates are populated, which decreases the cooperativity of the folding transition, and folding rates vary in a linear manner with linker length. Of particular relevance to experiments involving arrested synthesis, we demonstrated that continuous translation and arrested translation yield equivalent results when the quasi-equilibrium condition $\tau_F \ll \tau_S$ is satisfied.

These findings lead us to the following perspective on cotranslational folding that is summarized in Figure 8. Consider a protein consisting of a series of domains that fold independently of one another. As this protein is being synthesized and the first N-terminal domain enters the exit vestibule of the ribosome tunnel, the native state of the domain is thermodynamically unstable, being populated less than 1% of the time. This unfolded domain takes on distinct properties relative to that in bulk solution, in particular it expands and has the anisotropic shape of a scalene ellipsoid. Furthermore, its hydrophobic character may increase, which might aid in chaperone recruitment. If this domain is over 80 AA in size, then partially folded tertiary structures, probably involving N-terminal residues, are likely to be populated and may be thermodynamically stable even at this length of the nascent chain.

When another three to five residues are incorporated in to the C-terminus of the nascent chain, even small globular domains (<80 AA) populate N-terminally folded intermediates, although they are typically metastable, being populated between 10 and 50% of the time. Concomitantly, the unfolded state decreases in size by up to 20% and loses some of its anisotropic shape. At this nascent chain length, the domain is close to its midpoint of stability, resulting in appreciable populations of unfolded, partially folded and folded species.[25] These partially folded N-terminal species, which are not populated in bulk solution, decrease the cooperativity of the folding transition and shift the **TS** toward more compact structures than are found in bulk.

On the addition of a further one or two amino acids to the nascent chain, the equilibrium of the domain shifts in favor of

F and its native state becomes marginally stable ($-2 \leq \Delta G_{FU} < 0$ kcal/mol) and the unfolded state collapses to within 15% of its size in bulk solution. The partially folded N-terminal species are still populated within the distribution of nascent chain structures but to a reduced extent. Structurally, the folded state is located either partially inside the vestibule or fully outside the tunnel at this chain length. The rate of unfolding is now slower than the rate of folding, and the folding rate varies linearly as new amino acids are incorporated. For those cases in which a nascent chain folds at this chain length, it will be more likely to be initiated from the N-terminus than from the C-terminus, despite the fact that N- and C-terminal interactions within the domain are sterically allowed by the ribosome.

Allowing incorporation of another five amino acids, the center of mass of the domain moves further away on average from the ribosome, which decreases the effect of the ribosome on its folding behavior. The properties of the domain, although still tethered to the ribosome by its yet-to-be fully synthesized neighboring domain, are similar to their bulk solution properties. At this nascent chain length, folding occurs in a similar fashion as in bulk solution. Thus, a kinetic partitioning must occur during continuous translation of multidomain proteins, as some fraction of the domains will fold in a bulk solution-like manner if they fold at longer nascent chain lengths, while others will behave very differently if they fold at shorter nascent chain lengths. This conclusion is consistent with results of folding kinetics from lattice simulations of cotranslational folding.[39]

There are a number of caveats to this perspective on cotranslational folding. First, most single-domain proteins are likely to fold post-translationally because they require almost all of their native interactions to be thermodynamically stable,[55] and they only fully emerge from the exit tunnel after they have been covalently released from tRNA. Thus, folding of such single-domain proteins are unlikely to be perturbed significantly by the presence of the ribosome unless they have a C-terminal linker or extension similar to that used in this study or bind strongly to the outer ribosome surface. This conclusion is supported by results from simulations of CI2 on the ribosome.[19] Second, time scales are crucial in delineating cotranslational folding scenarios during the nonequilibrium process of translation. Nascent chains that fold on much longer time scales than monomer incorporation ($\tau_F \gg \tau_S$) will not be perturbed as much by attachment to the ribosome compared to those that fold with $\tau_F \approx \tau_S$ or $\tau_F \ll \tau_S$. In this last regime ($\tau_F \ll \tau_S$), quasi-equilibrium is achieved and our simulation results for fast folding proteins are directly relevant to continuous translation. Finally, in the simulations described in this paper, we did not account for the plethora of auxiliary proteins that are known to interact with the emerging nascent chain;[12] these chaperone-nascent chain interactions may alter cotranslational folding further and may be crucial in avoiding misfolding and aggregation.

The results we have discussed in this paper make several specific predictions that are amenable to experimental validation. For example, we predict broadening of the size distribution of the unfolded ensemble that could be tested by single molecule FRET methods, where a broader FRET distribution is expected for unfolded proteins attached to the ribosome. The predicted change in shape of the unfolded ensemble could in principle be tested using SAXS or NMR. The partially folded intermediate conformations that are predicted to be populated on the ribosome could in principle be trapped using flash freezing or cross-linking and subsequently probed via cryo-EM. The predicted changes in native state stability could be measured using force unfolding techniques such as Atomic Force Microscopy or Laser Optical Tweezers.

Overall, therefore, these findings suggest that an understanding of protein folding in the cell will include new scenarios of protein folding that have not been observed in bulk solution, within the framework of the generic principles of protein folding. In particular, knowledge of the impact of the ribosome, and of other cellular components such as molecular chaperones,[19,41,56,57] promises to provide insights into the manner in which folding is promoted and misfolding suppressed to enable cells to function effectively.

Methods

Proteins Examined. We examined the five globular and repeat proteins shown in Figure 1a that range in length from 20 amino acids (Trp-cage) up to 80 amino acids (λ_{6-85}). The crystal structures used in the modeling described below for protein G, λ_{6-85}, 3_H, β-helix, and Trp-cage correspond, respectively, to PDB identification codes 1gb1,[58] 1lmb,[59] 1oks,[60] 1m8n,[61] and 1l2y.[62] The β-helix was excised from its corresponding PDB file corresponding to residues 46−102 in chain A. For 3_H, selenomethionine residues were treated as methionine in constructing the coarse grained model.

Validation against the cross-linking experimental data was carried out by simulating the Kv1.2 protein whose PDB identification code is 1qdv.[35]

Protein and 50S Ribosome Model. The coarse-grained ribosome nascent chain model used in this study has been described in detail elsewhere.[15] Briefly, the nascent chain, whether bound to the ribosome or in bulk, is modeled using the C_α side chain model (C_α-SCM),[63,64] which represents a protein molecule by using two interaction sites per amino acid, one at the C_α carbon backbone position and the other at the center of mass of the side chain. The C_α-SCM is a Go model, that is, some of its force field parameters are based on the crystal structure of the protein under study. The C_α-SCM energy function accounts for the sequence composition of the protein, in terms of variable side chain−side chain interaction strengths and the sizes of the side chain interaction sites, and backbone hydrogen bonding, among other terms.[64] The polyglycine linker is modeled using a modified form of the C_α-SCM with a transferable, sequence dependent, backbone torsion potential.[65] The linker is in an extended strand conformation with an equilibrium C_α bond angle of 123° that was parametrized using PDB statistics (see details in ref 15).

We simulate only the large 50S subunit of the ribosome and do not include the 30S subunit because the nascent chain in these simulations never comes closer than 30 Å to the location of the 30S subunit. Furthermore, at 310 K in a medium with a dielectric of 80 the Bjerrum length is around 7 Å, meaning that electrostatic

(55) Neira, J. L.; Fersht, A. R. *J. Mol. Biol.* **1999**, *285*, 1309–1333.
(56) Kaiser, C. M.; Chang, H. C.; Agashe, V. R.; Lakshmipathy, S. K.; Etchells, S. A.; Hayer-Hartl, M.; Hartl, F. U.; Barral, J. M. *Nature* **2006**, *444*, 455–460.
(57) McGuffee, S. R.; Elcock, A. H. *PLoS Comp. Biol.* **2010**, *6*, e1000694.
(58) Gronenborn, A. M.; Filpula, D. R.; Essig, N. Z.; Achari, A.; Whitlow, M.; Wingfield, P. T.; Clore, G. M. *Science* **1991**, *253*, 657–661.
(59) Beamer, L. J.; Pabo, C. O. *J. Mol. Biol.* **1992**, *227*, 177–196.
(60) Johansson, K.; Bourhis, J. M.; Campanacci, V.; Cambillau, C.; Canard, B.; Longhi, S. *J. Biol. Chem.* **2003**, *278*, 44567–44573.
(61) Leinala, E. K.; Davies, P. L.; Doucet, D.; Tyshenko, M. G.; Walker, V. K.; Jia, Z. *J. Biol. Chem.* **2002**, *277*, 33349–33352.
(62) Neidigh, J. W.; Fesinmeyer, R. M.; Andersen, N. H. *Nat. Struct. Biol.* **2002**, *9*, 425–430.
(63) Klimov, D. K.; Thirumalai, D. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 2544–2548.
(64) O'Brien, E. P.; Ziv, G.; Haran, G.; Brooks, B. R.; Thirumalai, D. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *36*, 13403–13408.
(65) Karanicolas, J.; Brooks, C. L. *Protein Sci.* **2002**, *11*, 2351–2361.

interactions between the nascent chain and the 30S subunit are less than the thermal energy and are therefore negligible.

We use the atomic coordinates from an *E. coli* 50S crystal structure with PDB identification 1PNY to generate a coarse grained 50S subunit containing approximately 17,000 interaction sites (Figure S3). Ribosomal RNA is modeled using the 3,4-Interaction Site Model[15] and ribosomal protein using the C_α-SCM. tRNA was not modeled in the simulations as its effect is unlikely to be significant in the present study given its 60 Å distance from where tertiary folding can begin along the ribosome tunnel.[14,15] The location of the covalent bond between the C-terminus of the nascent chain and the tRNA is precisely known (Figure 1a),[66] and therefore, we modeled this spatial constraint as a harmonic restraint applied to the C-terminal C_α-atom of the nascent chain.

Simulation Details. To calculate the thermodynamic properties of the nascent chain in bulk solution or on the ribosome, we use replica exchange (REX) simulations that enhance conformational sampling and thereby increase the precision of the calculated quantities.[67] A REX run consists of running independent trajectories, conventionally referred to as replicas, in the canonical (NVT) ensemble at different temperatures. Periodically during the simulations, the swapping of system coordinates between replicas at different temperatures is attempted and is accepted if $\chi \leq 1$ and with a probability proportional to $e^{-\chi}$ if $\chi > 1$, where $\chi = (E_1 - E_2)/(k_B(T_2 - T_1))$ and the temperature and total potential energy of replica i are denoted T_i and E_i, respectively.

In this study, one REX simulation was performed for each ribosome nascent chain construct and for the protein alone in bulk solution. In the bulk simulations, the proteins contained a 35 AA polyglycine linker attached to their C-terminus. Eight temperature windows were used with minimum and maximum temperatures between 160 and 460 K and the range of temperatures in any single REX run did not exceed 200 K. A total of 90 000 swap attempts between neighboring temperature windows were carried out, with the first 15 000 discarded to allow for equilibration. The probability of successfully swapping coordinates between temperatures was found to be between 0.10 and 0.50.

The program CHARMM[68] was used to simulate the time evolution of each replica using Langevin dynamics with a damping coefficient of 0.05 ps^{-1} and an integration time step of 10 fs in the NVT ensemble. Each replica was simulated for 13 ps before swapping was attempted. An in-house Perl script was used to perform REX. In the simulations, the ribosome was held fixed using the Constraint module in CHARMM, and the C-terminus of the nascent chain was tethered near the PTC using a harmonic restraint with a spring constant of 2.5 kcal/(mol $Å^2$).

To demonstrate that holding the ribosome interaction sites fixed during the simulations does not change our results, we performed REX in which the ribosome interaction sites fluctuate in a harmonic potential with a variance of 1 $Å^2$. This was done using the 'constraints harmonic' command in CHARMM and the Block module, with which we turned off intraribosome interactions. In this REX simulation, we modified the acceptance criteria for temperature swaps such that instead of using the total potential energy of the system (E_i), we used E_i minus the sum of the harmonic restraint energies of the ribosome interaction sites. This modified acceptance criteria significantly enhanced sampling because it allowed larger spacings between temperature windows. This approach should approximately preserve the underlying canonical distribution provided the ribosome interaction sites fluctuate on much faster time scales than the characteristic time scale of nascent chain rearrangement, yielding a quasi-equilibrium condition.

To calculate k_F and k_U in bulk solution and on the ribosome, we performed, respectively, temperature quench simulations starting from the unfolded ensemble at 360 K → 310 K, and constant temperature simulations starting from the folded state at 310 K. Langevin dynamics was used with a 10 fs integration time step and damping coefficient of 0.05 ps^{-1}. The folding simulations were equilibrated at 360 K for 150 000 integration time steps and then production runs started at the final temperature of 310 K. For the constant temperature simulations, an rmsd restraint was applied to the protein during equilibration to keep it folded, and released upon the start of the production run. Anywhere from 300 to 600 independent trajectories were simulated to calculate each k_F and k_U value. Each trajectory was stopped when it reached a predefined fraction of native contacts Q_{Bound} (see Table S1 and below) and its simulation time τ recorded for use in first passage time analysis.

For protein G and λ_{6-85}, the time scales for unfolding of RNCs with $L > 28$ AA at 310 K are much longer than the time scales that can reasonably be simulated. To estimate k_U at 310 K, we assumed that these proteins behave as apparent two-state systems and therefore $k_U = k_F e^{\Delta G_{FU}/(k_B T)}$, where k_F and ΔG_{FU} are precisely known from the temperature quench and replica exchange simulations, respectively. This two-state assumption is reasonable considering protein G and λ_{6-85} have γ values close to 1 (Table 1).

Analysis. We used the WHAM equations[69] to calculate thermodynamic quantities from the REX simulations. Definitions of the quantities calculated in this way can be found in the Supporting Information. We used first passage time analysis to calculate the rate $k_{A\to B} = \langle \tau_{A\to B} \rangle^{-1}$, where $\langle \tau_{A\to B} \rangle$ is the average time it takes a trajectory to reach state B starting from state A. State B is defined by an absorbing boundary placed at Q_{Bound} (Table S1). $\langle \tau_{A\to B} \rangle = N^{-1} \sum_{i=1}^{N} \tau_i$, where N is the number of folding or unfolding trajectories simulated upon temperature quench or constant temperature simulations and is between 300 and 600, and τ_i is the first passage time recorded for the i^{th} trajectory. Details of the analysis of folding pathways and their probabilities can be found in the Supporting Information.

Acknowledgment. We thank Sophie Jackson for a careful reading of the manuscript, and are grateful for financial support for this work through an NSF Postdoctoral Fellowship to E.P.O, a Human Frontier Young Investigators Award (RGY67/2007) and BBSRC grant (9015651/1) to J.C., and a Program Grant from the Wellcome Trust to C.M.D.

Supporting Information Available: Details on simulation methods, protocols, and analysis, as well as additional discussion on the validation of the coarse-grained ribosome nascent chain model. This information is available free of charge via the Internet at http://pubs.acs.org/.

JA107863Z

(66) Voorhees, R. M.; Weixlbaumer, A.; Loakes, D.; Kelley, A. C.; Ramakrishnan, V. *Nat. Struct. Mol. Biol.* **2009**, *16*, 528–533.
(67) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314*, 141–151.
(68) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187–217.
(69) Kumar, S.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A.; Rosenberg, J. M. *J. Comput. Chem.* **1992**, *13*, 1011–1021.
(70) Plaxco, K. W.; Simons, K. T.; Baker, D. *J. Mol. Biol.* **1998**, *227*, 985–994.