Synthesis runs counter to the directional folding pathway of a nascent protein domain

Xiuqi Chen [1], Nandakumar Rajasekaran [1], Kaixian Liu [1,‡] and Christian M. Kaiser [2,3,*]

[1] CMDB Graduate Program, Johns Hopkins University, Baltimore, MD, USA
[2] Department of Biology, Johns Hopkins University, Baltimore, MD, USA
[3] Department of Biophysics, Johns Hopkins University, Baltimore, MD, USA
[‡] present address: Molecular Biology Program, Sloan Kettering Institute, New York, NY
[*] corresponding author (kaiser@jhu.edu)

Abstract

Folding of individual domains in large proteins occurs during translation, helping to avoid otherwise prevalent inter-domain misfolding. How folding intermediates observed *in vitro* for the majority of proteins relate to co-translational folding remains unclear. Combining *in vivo* and single-molecule experiments, we followed the co-translational folding of the G-domain, encompassing the first 293 amino acids of elongation factor G. Surprisingly, the G-domain remains unfolded until it is fully synthesized, without collapsing into molten globule-like states or forming stable intermediates. Upon fully emerging from the ribosome, the G-domain transitions to its stable native structure via folding intermediates. Our results suggest a strictly sequential folding pathway that initiates from the C-terminus and thus runs counter to synthesis. The folding mechanism is likely imposed by the final structure and might have evolved to ensure efficient, timely folding of a highly abundant and essential protein.

Introduction

How proteins fold into the native structures that enable their cellular functions remains a central question in biology. Folding of small proteins or single domains often follows single-exponential kinetics, suggesting a highly cooperative process [1]. However, experimental and computational studies have detected partially folded intermediates for a multitude of proteins, and proteins larger than ~100 amino acids are generally thought to populate folding intermediates before reaching their native structure [2,3].

While the presence of intermediates in most protein folding pathways is universally recognized, their functional roles are less clear. Partially structured states can represent on-pathway intermediates along a multi-step folding pathway, or off-pathway misfolded states that must dissolve before productive folding can be achieved [2,4]. Distinguishing these two scenarios

experimentally is often challenging, although single-molecule approaches have demonstrated their potential to provide this information [5,6].

How intermediates are linked together into a folding pathway remains a subject of debate [7,8]. The "foldon" hypothesis posits that small, cooperative units acquire structure in a strictly sequential order, resulting in a single folding pathway [9]. Considerations based on energy landscape theory propose that a folding protein can reach its native structure through several accessible routes [10], although specific pathways may be energetically favored. Either theory is supported by evidence from both experiments and simulations. It thus seems possible that proteins might fall into distinct categories based on their folding mechanism. Determining the relationship between amino acid sequence, folding pathway and final structure is important for designing sequences that can adopt novel structures and understanding how natural proteins robustly reach their functional conformations.

Folding pathways are not only shaped by protein-specific properties, but also depend strongly on environmental conditions, many of which differ greatly between *in vitro* refolding experiments and folding in living cells [11]. The vast majority of mechanistic folding studies has been carried out with isolated proteins or domains [12]. In the cell, proteins begin to fold while they gradually emerge from the ribosome as it translates the information in the messenger RNA into a polypeptide sequence. Biophysical experiments have shown that interactions of the ribosome with the proximal part of the nascent polypeptide can reduce the stability of native domains [13,14], stabilize secondary structure inside the ribosome exit tunnel [15-17], modulate nascent chain folding kinetics [18] and prevent misfolding [18-20]. Co-translational folding is particularly important for large multi-domain proteins, because it prevents the accumulation of extended unfolded regions during synthesis that have a high propensity for misfolding [21]. Several examples show that the ribosome does not necessarily change the folding pathway [19,22,23], but the connection between vectorial emergence and folding of the nascent chain remains poorly defined.

Here, we have investigated folding of the N-terminal G-domain of *Escherichia coli* elongation factor G (EF-G). Following nascent chain folding in bacterial cells with an arrest peptide-based reporter assay [24], we find that productive folding is not initiated until the full domain has emerged from the ribosome. Force spectroscopy experiments with optical tweezers confirmed the absence of stable structure at shorter chain lengths. These single-molecule measurements also showed that folding of the full domain proceeds through productive folding intermediates both on the ribosome and in isolation. Our studies thus reveal a strictly ordered folding pathway in which the first step requires the extreme C-terminus of the domain. Folding and synthesis of the domain therefore proceed in opposite directions.

Results

Detection of co-translational folding in vivo with a luciferase reporter

To investigate folding of the G-domain from EF-G, we took advantage of the 17 amino acid arrest peptide from the protein SecM [25] (termed *SecM17* here). *SecM17* causes elongation arrest when the N-terminal 16 amino acids have been synthesized [26]. Folding-mediated release of elongation arrest results in translation of the coding sequence downstream of *SecM17* and

production of the full-length encoded protein. The arrest peptide can thus be utilized to detect nascent chain folding that occurs in close proximity to the ribosome [24] (Figure 1). Release is typically quantified by autoradiography after *in vitro* translation [23,27-33]. To monitor folding in living cells, we utilized a genetically encoded luciferase reporter, NanoLuc [34], to detect nascent chain folding (Figure 1).

To validate the *in vivo* folding assay, we generated reporter plasmids in which the G-domain is connected to *SecM17* through 4, 17 or 30 residues of the subsequent domain II, resulting in constructs termed *G+20*, *G+33*, and *G+46* (Figure 2A). The numbers represent the sum of the domain II and *SecM17* residues that separate the G-domain from the peptidyl transferase center at the arrest point. The G-domain (residues 1 to 293 of EF-G) has been shown to stably fold upon emerging from the ribosome exit tunnel [18,35], which sequesters approximately 30 to 40 residues of the nascent chain inside the large subunit of the ribosome.

The *G+33* construct, which places the G-domain close to the tunnel exit, resulted in strong luminescence upon expression in *E. coli* cells (Figure 2B, bar diagram). At this length, the stably folded domain abuts the ribosome, generating a pulling force that destabilizes SecM arrest [24]. Consequently, translation resumes and the NanoLuc reporter is synthesized. As expected, the other two constructs yielded lower levels of reporter expression (Figure 2B, bar diagram). In *G+20*, sequestration of the C-terminal ~15 amino acids in the tunnel destabilizes the G-domain; in *G+46*, unfolded domain II polypeptide separates the folded G-domain from the ribosome, preventing the generation of tension on the nascent chain (Figure 2B, cartoons).

To verify that the luminescence readouts in our experiments reflect increased NanoLuc accumulation, rather than reduced specific activity of NanoLuc caused by fusion to poorly folded polypeptides [36], we visualized arrested and full-length translation products by Western blotting. We observed similar levels of arrested protein for all three constructs, but a significantly higher amount of full-length product for *G+33* compared to the control constructs (Figure 2B, bottom panel). This result indicates that elevated luminescence indeed reports accelerated arrest release.

Colony luminescence identifies arrest releasing candidates

The folding of some small proteins or domains is well described by a two-state model, while proteins of the size of the G-domain usually populate intermediates during folding [2,3]. To determine whether intermediates are formed co-translationally, we generated 75 individual reporter constructs with EF-G inserts ranging in length from 72 to 368 amino acids in steps of four (Figure 3A). When the ribosome stalls at the SecM sequence, the separation of the N-terminal G-domain residue from the peptidyl-transferase center (PTC) of the ribosome, referred to here as length L (Figure 3B), ranges from 88 to 388 aa (16 residues from SecM plus 72 and 368 EF-G residues, respectively).

We transformed *E. coli* cells with a mixture of all plasmids at similar concentrations and grew colonies on inducing agar. When we applied the luciferase substrate to the plate, a fraction of colonies exhibited strong luminescence (Figure 3C). Sequencing revealed that constructs yielding highly luminescent colonies all encoded candidate proteins with lengths ranging from 308 to 328 amino acids (Figure 3D, cyan dots). It thus appears that colony luminescence reliably identifies arrest-releasing constructs in the region where full folding of the G-domain is expected. Non-luminescent colonies, selected from a small area of the plate that also contained highly luminescent colonies (Figure 3C, blowup), all had candidate lengths outside this region (Figure 3D, grey dots).

This result suggests that, surprisingly, only full G-domain folding constitutes a major folding waypoint during EF-G synthesis.

Full in vivo folding profile of the G-domain

The on-plate assay provides a convenient format for analyzing a pool of candidates in a simple experiment. However, colony luminescence only provides a binary readout, and subtle differences in signal are not resolved. To compare arrest release rates more quantitatively, we assayed the 75 EF-G truncation constructs individually by growing them separately in liquid cultures. After normalizing by cell density, culture luminescence intensity shows a clear peak with the maximum at L = 332 (Figure 4). Strong arrest release is therefore detected when the C-terminal G-domain residue (aa 293) is separated from the PTC by 39 aa, a value similar to that observed for other relatively large domains [31].

Notably, a signal increase to approximately 25% of the maximum intensity coincides with completion of G-domain synthesis (L = 309). At this length, offset from the maximum of the peak by 23 amino acids (Figure 4, vertical bar), all of the G-domain except for the C-terminal helix has been extruded from the ribosome. The slightly elevated arrest release rates in this length range could suggest the formation of meta-stable structures outside the ribosome [29,32], or formation of secondary structure inside the exit tunnel [16,31,37]. Remarkably, all chain lengths shorter than 308 amino acids exhibit only basal luciferase activity (below 20% of the peak value; grey box in Figure 4), suggesting that the G-domain does not form stably folded intermediates. Given that the G-domain is large, this finding is unexpected.

Stably structured states are not formed until synthesis is complete

Several folding scenarios can explain why most of the G-domain truncations do not exhibit luminescence above the baseline level. If partially folded structures are destabilized by the ribosome, they may become stably folded only after an intervening "spacer" has been synthesized, preventing detection of these intermediates in the *in vivo* reporter assay. Alternatively, the G-domain may only begin to form stable structures once the complete domain has been synthesized and is mostly exposed to the outside of the ribosome.

To distinguish between these scenarios, we probed the structure of nascent EF-G polypeptides directly using single-molecule force spectroscopy with optical tweezers (Figure 5). We generated terminally stalled ribosome-nascent chain complexes by *in vitro* translation of non-stop messenger RNAs [18]. Tag sequences at the N-terminus of the nascent chain and on protein L17 in the large ribosome subunit allowed us to tether these complexes for mechanical manipulation with optical tweezers (Figure 5A). Mechanical force acts as a denaturant that destabilizes folded structures. In "force ramp" experiments, a continuously increasing tension is applied to the tethered molecule. In the resulting force-extension curves, unfolding of nascent chain structure results in "rips", sudden increases in molecular extension as the polypeptide transitions from a compact folded to an extended unfolded state. The method is thus suitable to detect tertiary structure in individual protein molecules.

In ribosome nascent chain complexes with a length of L = 328 (termed 328_{RNC} here), unfolding of the native G-domain results in a characteristic transition (Figure 5B), as observed previously [18,35]. As expected, a very short nascent chain (44_{RNC}) that exposes only a few EF-G residues outside the exit tunnel does not exhibit any transitions in these experiments. Surprisingly, however, a

substantially longer nascent chains (252_{RNC} and 316_{RNC}) do not exhibit defined unfolding transitions, either (Figure 5B). Occasionally, we detect heterogeneous transitions at these chain lengths events (Supplementary Figure 1). Their distribution is distinct from that expected for unfolding of well-defined states and might be reminiscent of misfolded states that have been observed with other proteins [19,38]. Regardless of what these heterogeneous transitions represent, the results from our single-molecule experiments indicate that the nascent G-domain does not form stable folding intermediates or collapsed states, even when almost all of its sequence has emerged from the ribosome.

Multi-state folding of the full domain follows a strict order

To follow folding of the ribosome-bound G-domain, we carried out optical tweezers experiments with 328_{RNC} in "force clamp". In these experiments, the force is held at a constant value while changes in molecular extension are recorded. After fully unfolding the G-domain, we jumped the force to 3.5 pN to initiate refolding. The molecule transitions repeatedly between the unfolded and partially folded states (Figure 6A), reflecting the complexity in folding that is expected for a large domain. Transitions cease upon forming a fully structured stable state (Figure 6A, open arrowhead). The partially structured states exhibit several distinct extensions that may represent productive on-pathway folding intermediates or misfolded off-pathway species. Folding to the native state occurs from a partially folded structure (Figure 6A, grey arrowhead), demonstrating the presence of at least one on-pathway intermediate.

A complex pattern of folding intermediates is also observed in force clamp experiments with the isolated G-domain in the absence of the ribosome (Figure 6B). Previous work [18] showed that the overall folding rate of 328_{RNC} is reduced compared to that of the isolated G-domain, implying interactions of the nascent chain with the ribosome prior to folding. The ribosome may therefore affect the stability of folding intermediates or transitions between them. Nevertheless, the intermediates exhibit similar extensions in the isolated G-domain and the ribosome-bound nascent chain, suggesting similar folding pathways in both scenarios. As observed for 328_{RNC}, the final folding step to the stably structured domain initiates from a partially structured state (Figure 6B, grey arrowhead).

The on-pathway folding intermediate is approximately 10 nm more extended than the natively folded G-domain (Figure 6 A and B, grey vs open arrowheads). At the refolding force of 3.5 pN, this length corresponds to ~90 amino acids of unfolded polypeptide, as calculated using a worm-like chain model [39] with a persistence length of 0.65 nm and a contour length of 0.36 nm per amino acid. The intermediate would therefore be composed of ~200 structured G-domain residues. Transitions consistent with unfolding of such a structure are observed in some of force ramp recordings obtained with 328_{RNC} (Supplementary Figure 1, dashed grey line). The absence of similar transitions in shorter nascent chains (252_{RNC} and 316_{RNC}, Supplementary Figure 1) suggests the C-terminal part of the G-domain, encompassing the last alpha-helix (Figure 6C), is required for the formation of the obligatory on-pathway intermediate. Taken together, our measurements are consistent with a strictly ordered folding pathway that begins at the extreme C-terminus of the G-domain.

Discussion

We have defined the folding waypoints and intermediates of the nascent G-domain from elongation factor G using a combination of *in vivo* (Figure 2 – 4) and single-molecule measurements (Figures 5 and 6). Notably, no stable folding intermediates are detected *in vivo* or *in vitro* until the complete domain of 293 amino acids is synthesized. The small src SH3 domain (64 amino acids) was previously found to fold only upon reaching the exit of the ribosome tunnel [22]. However, polypeptides composed of more than 100 amino acids are commonly assumed to populate folding intermediates [2,3]. It is thus surprising that the N-terminal ~250 amino acids of the G-domain do not appear to acquire stable structures.

Consistent with previous results [18,35], the G-domain forms a stable structure after fully emerging from the ribosome (L = 328), which manifests as a large peak in the SecM reporter assay (Figure 4). Non-random signal fluctuations at shorter chain lengths (48 < L < 320) might represent secondary structure formations, especially α-helices, which have previously been shown to accelerate SecM arrest release [29,31,40]. The identity of nascent chain sequences inside the ribosome exit tunnel have been described to also affect arrest release [29,31,32] and might thus account for the observed signal. Regardless of their origin, the amplitudes of these fluctuations are small compared to the main peak, suggesting the absence of stably folded structures until the G-domain is fully synthesized.

Single-molecule optical tweezers experiments confirm that folding begins only after synthesis is complete. We do not detect well-defined stable structures in nascent chains at lengths up to L = 316 (Figure 5). Small, heterogeneous transitions that are occasionally observed for the incomplete G-domain (Supplementary Figure 1) suggest misfolded states that are likely suppressed *in vivo* by molecular chaperones. Optical tweezers measurements readily detect partially folded (e.g. in alpha-synuclein [38]) or collapsed, molten globule-like structures (e.g. in ribonuclease H [41], T4 lysozyme [19] and apomyoglobin [42]). The contact order [43] of the G-domain is moderately low (relative contact order: 0.07), and its overall hydrophobicity [44] is similar to that of other proteins (Supplementary Figure 2). These factors therefore do not appear to account for the observed lack of structure formation or collapse during synthesis. Its sequence properties might help to keep the domain extended and prevent (premature) collapse into kinetically trapped states [45]. Ribosome interactions, previously shown to destabilize native [13,46] and non-native [19,35] nascent chain structure, might further contribute to keeping the nascent G-domain unfolded before its synthesis is complete.

Interestingly, our measurements suggest a relatively sharp transition in the propensity to form compact structures as the nascent chain is elongated. At L = 316, the nascent chain has properties of an intrinsically disordered protein, whereas the addition of just 12 amino acids to L = 328 results in the formation of collapsed or partially structured states and subsequent folding to the native structure. Theoretical studies concluded that the formation of compact states is an evolved property of natural proteins [47]. The G-domain may be an attractive model to investigate how this property is related to protein sequence and structure.

Our studies provide an example of folding occurring in the direction opposite to that of synthesis and thus contrasts with previous findings of gradual compaction and folding concomitant with protein elongation [48]. Once the full G-domain has emerged from the ribosome, folding occurs in several steps that appear similar on the ribosome and in isolation (Figure 6A). The ribosome therefore does not seem to change the folding pathway. The late onset of folding and the detection

of defined intermediates suggests that folding proceeds along a sequential pathway. This scenario is consistent with the "foldon hypothesis", in which well-defined states are formed in a prescribed order [49].

The strict sequentiality of G-domain folding might be dictated by the final structure. The region around the C-terminal helix is completely engulfed in the folded structure (Figure 6C and D). Perhaps folding has to occur with the helix serving as a central nucleus around which structure is formed subsequently, rather than by inserting the helix into preformed intermediates. Interestingly, part of the "enclosure" around the C-terminal helix is formed by the G' domain, an insertion present in some, but not all G-domain containing elongation factors [50] (Supplementary Figure 3). In future studies with homologous G-domains (such as those shown in Supplementary Figure 3), it will be interesting to examine whether lack of the G' insertion allows intermediate formation during synthesis and relaxes the strict order of folding that we observe here for EF-G. Nascent G-domains appear as attractive models for investigating how folding pathways co-evolve with structures that enable crucial cellular functions.

EF-G is a highly abundant protein (top 1% in the *E. coli* proteome [51]) that fulfills an essential cellular function. The efficiency of EF-G synthesis and folding may have been under evolutionary pressure. The coding sequence contains very few rare codons (Supplementary Figure 4), suggesting swift translation [52]. Collapsed states and intermediates can kinetically trap folding proteins in non-native states [4]. Structure acquisition through a strictly ordered sequential pathway upon completion of synthesis might have evolved as a mechanism to ensure timely folding of EF-G.

Materials and methods

Bacterial strains, plasmids and reagents

In vivo folding experiments were carried out in *E.coli* strain Lemo21(DE3) (New England Biolabs, NEB, C2528J). All plasmids used in this study are based on a backbone with a pUC origin, Lac-operator-controlled T7 promoter and Ampicillin resistance gene [35]. SecM and NanoLuc sequences were synthesized (Integrated DNA Technologies) and inserted into the open reading frame with Gibson Assembly Master Mix (NEB E2611). All commercially available enzymes were purchased from NEB unless stated otherwise. PCR reactions were carried out with Phusion high-fidelity DNA polymerase (Thermo Scientific, F530S). Chemicals were purchased from Sigma-Aldrich unless stated otherwise. Streptavidin-HRP for immunoblot was from SouthernBiotech™ (#7100-05).

Construction of EF-G truncation library

All EF-G fragments were PCR-amplified with a NheI tailed universal forward primer and reverse primers at designated positions along the EF-G open reading frame. The backbone with SecM-NanoLuc sequence was amplified with SpeI tailed forward primer at the AviTag and reverse primer at the SecM. After PCR clean-up, backbone and insert were mixed at a 1 to 3 ratio (40 fmol

in total) in CutSmart® Buffer (NEB) supplemented with 3mM ATP and 10mM DTT. The reaction was provided with 4U of each NheI, SpeI, T4PNK and T4 DNA ligase at 10µl final volume. After 2hr at 37°C and 16°C overnight, the product was transformed and verified by colony PCR with Taq DNA polymerase. Plasmids DNA from colonies showing the correct insert sizes were isolated, and their correct sequences were verified by Sanger sequencing.

On-plate NanoLuc assay

Cells were transformed with indicated plasmid per manufacturer instructions and spread on a LB agar supplemented with ampicillin, chloramphenicol, 500 µM L-rhamnose and 500 µM IPTG on a 9 cm diameter round plate. Colonies were grown at 37°C for 12 to 16 hours. We reconstituted 500µl Nano-Glo® Live Cell Reagent (Promega N2011) for each plate and sprayed evenly onto the plate with an airbrush (Neo Iwata HP-CN N4500). Images were taken in a shaded home-made imaging box equipped with a Canon Rebel T3 camera. Camera settings were at Neutral from 5 to 15 seconds exposure and ISO800 for optimal contrast. Colonies that showed high luminescence and were well separated were selected for colony PCR and Sanger sequencing.

Liquid culture NanoLuc assay

Cells transformed with individual plasmids were spread on LB agar with antibiotics and allowed to form colonies overnight. LB supplemented with ampicillin and chloramphenicol was inoculated with a single colony to grow an overnight culture. Overnight cultures were diluted into fresh LB containing to OD_{600}=0.01 and incubated in a 37°C shaker at 220rpm. Cell densities were monitored with a plate reader (ThermoMax Microplate Reader, Molecular Devices). At OD_{600}=0.2-0.4, cultures were induced with 500 µM L-rhamnose and 500 µM IPTG for 1 hour. Cell densities were measured and 100µl cultures were put onto a white round-bottom 96-well plate for luminescence measurement. All NanoLuc assays in this study were carried out with Nano-Glo® Live Cell Assay from Promega (N2011) according to manufacturer instructions. Luminescence was measured on a GloMax® Navigator Microplate Luminometer (Promega). Signals were linearly normalized to OD_{600}=0.4.

Single-molecule force spectroscopy with RNCs

Stalled ribosome-bound nascent chain complexes (RNCs) were generated as described previously [18,35], dissolved in HKMβ buffer (20 mM HEPES, 100 mM KCl, 5 mM $MgCl_2$, 5 mM β-mercaptoethanol), and stored in small aliquots after flash-freezing. Single molecule experiments were carried out using a custom home-built instrument [53]. The experiments were carried out in HKMβ buffer. For force ramp experiments, the trap was moved at a constant speed of 100 nm/s to apply continuously increasing forces on the nascent chain in the range from 2 to 50 pN. The force was ramped down at the same loading rate, and the molecule was held at 2 pN for 10 s before being pulled again. Data were collected at a sampling frequency of 1000 Hz and averaged to 33 Hz for plotting. The extension changes were determined using custom MATLAB scripts as described in detail previously [18]. For constant force experiments, the molecule was first subjected to force ramp cycles to ensure it exhibited the characteristic unfolding transitions. After fully unfolding the molecule at 30 pN, the force was reduced to 3.5 pN to initiate refolding. The change in molecular extension was recorded at a sampling frequency of 1000 Hz.

References

1. Sosnick, T.R. & Barrick, D. The folding of single domain proteins--have we reached a consensus? *Current opinion in structural biology* **21**, 12-24 (2011).
2. Brockwell, D.J. & Radford, S.E. Intermediates: ubiquitous species on folding energy landscapes? *Curr Opin Struct Biol* **17**, 30-7 (2007).
3. Malhotra, P. & Udgaonkar, J.B. How cooperative are protein folding and unfolding transitions? *Protein Sci* **25**, 1924-1941 (2016).
4. Sosnick, T.R., Mayne, L., Hiller, R. & Englander, S.W. The barriers in protein folding. *Nat Struct Biol* **1**, 149-56 (1994).
5. Stigler, J., Ziegler, F., Gieseke, A., Gebhardt, J.C. & Rief, M. The complex folding network of single calmodulin molecules. *Science* **334**, 512-6 (2011).
6. Bustamante, C., Alexander, L., Maciuba, K. & Kaiser, C.M. Single Molecule Studies of Protein Folding with Optical Tweezers. *Annu Rev Biochem* **in press**(2020).
7. Englander, S.W. & Mayne, L. The case for defined protein folding pathways. *Proc Natl Acad Sci U S A* **114**, 8253-8258 (2017).
8. Eaton, W.A. & Wolynes, P.G. Theory, simulations, and experiments show that proteins fold by multiple pathways. *Proc Natl Acad Sci U S A* **114**, E9759-E9760 (2017).
9. Englander, S.W., Mayne, L., Kan, Z.Y. & Hu, W. Protein Folding-How and Why: By Hydrogen Exchange, Fragment Separation, and Mass Spectrometry. *Annu Rev Biophys* **45**, 135-52 (2016).
10. Onuchic, J.N. & Wolynes, P.G. Theory of protein folding. *Curr Opin Struct Biol* **14**, 70-5 (2004).
11. Gruebele, M., Dave, K. & Sukenik, S. Globular Protein Folding In Vitro and In Vivo. *Annu Rev Biophys* **45**, 233-51 (2016).
12. Braselmann, E., Chaney, J.L. & Clark, P.L. Folding the proteome. *Trends in biochemical sciences* **38**, 337-44 (2013).
13. Cabrita, L.D. et al. A structural ensemble of a ribosome-nascent chain complex during cotranslational protein folding. *Nat Struct Mol Biol* **23**, 278-85 (2016).
14. Samelson, A.J., Jensen, M.K., Soto, R.A., Cate, J.H. & Marqusee, S. Quantitative determination of ribosome nascent chain stability. *Proc Natl Acad Sci U S A* **113**, 13402-13407 (2016).
15. Lu, J. & Deutsch, C. Folding zones inside the ribosomal exit tunnel. *Nat Struct Mol Biol* **12**, 1123-9 (2005).
16. Bhushan, S. et al. alpha-Helical nascent polypeptide chains visualized within distinct regions of the ribosomal exit tunnel. *Nat Struct Mol Biol* **17**, 313-7 (2010).
17. Holtkamp, W. et al. Cotranslational protein folding on the ribosome monitored in real time. *Science* **350**, 1104-7 (2015).
18. Liu, K., Maciuba, K. & Kaiser, C.M. The Ribosome Cooperates with a Chaperone to Guide Multi-domain Protein Folding. *Mol Cell* **74**, 310-319 e7 (2019).
19. Kaiser, C.M., Goldman, D.H., Chodera, J.D., Tinoco, I., Jr. & Bustamante, C. The ribosome modulates nascent protein folding. *Science* **334**, 1723-7 (2011).
20. Alexander, L.M., Goldman, D.H., Wee, L.M. & Bustamante, C. Non-equilibrium dynamics of a nascent polypeptide during translation suppress its misfolding. *Nat Commun* **10**, 2709 (2019).

21. Kaiser, C.M. & Liu, K. Folding up and Moving on—Nascent Protein Folding on the Ribosome. *Journal of Molecular Biology* **430**, 4580-4591 (2018).
22. Guinn, E.J., Tian, P., Shin, M., Best, R.B. & Marqusee, S. A small single-domain protein folds through the same pathway on and off the ribosome. *Proc Natl Acad Sci U S A* **115**, 12206-12211 (2018).
23. Tian, P. et al. Folding pathway of an Ig domain is conserved on and off the ribosome. *Proc Natl Acad Sci U S A* **115**, E11284-E11293 (2018).
24. Goldman, D.H. et al. Ribosome. Mechanical force releases nascent chain-mediated ribosome arrest in vitro and in vivo. *Science* **348**, 457-60 (2015).
25. Nakatogawa, H. & Ito, K. The ribosomal exit tunnel functions as a discriminating gate. *Cell* **108**, 629-36 (2002).
26. Muto, H., Nakatogawa, H. & Ito, K. Genetically encoded but nonpolypeptide prolyl-tRNA functions in the A site for SecM-mediated ribosomal stall. *Mol Cell* **22**, 545-52 (2006).
27. Nilsson, O.B. et al. Cotranslational Protein Folding inside the Ribosome Exit Tunnel. *Cell reports* **12**, 1533-40 (2015).
28. Nilsson, O.B. et al. Cotranslational folding of spectrin domains via partially structured states. *Nat Struct Mol Biol* **24**, 221-225 (2017).
29. Notari, L., Martinez-Carranza, M., Farias-Rico, J.A., Stenmark, P. & von Heijne, G. Cotranslational Folding of a Pentarepeat beta-Helix Protein. *J Mol Biol* **430**, 5196-5206 (2018).
30. Kudva, R. et al. The shape of the bacterial ribosome exit tunnel affects cotranslational protein folding. *Elife* **7**(2018).
31. Farias-Rico, J.A., Ruud Selin, F., Myronidi, I., Fruhauf, M. & von Heijne, G. Effects of protein size, thermodynamic stability, and net charge on cotranslational folding on the ribosome. *Proc Natl Acad Sci U S A* **115**, E9280-E9287 (2018).
32. Kemp, G., Kudva, R., de la Rosa, A. & von Heijne, G. Force-Profile Analysis of the Cotranslational Folding of HemK and Filamin Domains: Comparison of Biochemical and Biophysical Folding Assays. *J Mol Biol* **431**, 1308-1314 (2019).
33. Marino, J., von Heijne, G. & Beckmann, R. Small protein domains fold inside the ribosome exit tunnel. *FEBS Lett* **590**, 655-60 (2016).
34. Hall, M.P. et al. Engineered luciferase reporter from a deep sea shrimp utilizing a novel imidazopyrazinone substrate. *ACS Chem Biol* **7**, 1848-57 (2012).
35. Liu, K., Rehfus, J.E., Mattson, E. & Kaiser, C.M. The ribosome destabilizes native and non-native structures in a nascent multidomain protein. *Protein Sci* **26**, 1439-1451 (2017).
36. Waldo, G.S., Standish, B.M., Berendzen, J. & Terwilliger, T.C. Rapid protein-folding assay using green fluorescent protein. *Nat Biotechnol* **17**, 691-5 (1999).
37. Ziv, G., Haran, G. & Thirumalai, D. Ribosome exit tunnel can entropically stabilize alpha-helices. *Proc Natl Acad Sci U S A* **102**, 18956-61 (2005).
38. Neupane, K., Solanki, A., Sosova, I., Belov, M. & Woodside, M.T. Diverse metastable structures formed by small oligomers of alpha-synuclein probed by force spectroscopy. *PLoS One* **9**, e86495 (2014).
39. Bustamante, C., Marko, J.F., Siggia, E.D. & Smith, S. Entropic elasticity of lambda-phage DNA. *Science* **265**, 1599-600 (1994).
40. Marsden, A.P. et al. Investigating the Effect of Chain Connectivity on the Folding of a Beta-Sheet Protein On and Off the Ribosome. *J Mol Biol* **430**, 5207-5216 (2018).

41. Cecconi, C., Shank, E.A., Bustamante, C. & Marqusee, S. Direct observation of the three-state folding of a single protein molecule. *Science* **309**, 2057-60 (2005).
42. Elms, P.J., Chodera, J.D., Bustamante, C. & Marqusee, S. The molten globule state is unusually deformable under mechanical force. *Proc Natl Acad Sci U S A* **109**, 3796-801 (2012).
43. Plaxco, K.W., Simons, K.T. & Baker, D. Contact order, transition state placement and the refolding rates of single domain proteins. *J Mol Biol* **277**, 985-94 (1998).
44. Roseman, M.A. Hydrophilicity of polar amino acid side-chains is markedly reduced by flanking peptide bonds. *J Mol Biol* **200**, 513-22 (1988).
45. Clark, P.L., Plaxco, K.W. & Sosnick, T.R. Water as a Good Solvent for Unfolded Proteins: Folding and Collapse are Fundamentally Different. *J Mol Biol* (2020).
46. Waudby, C.A. et al. Systematic mapping of free energy landscapes of a growing filamin domain during biosynthesis. *Proc Natl Acad Sci U S A* **115**, 9744-9749 (2018).
47. Thirumalai, D., Samanta, H.S., Maity, H. & Reddy, G. Universal Nature of Collapsibility in the Context of Protein Folding and Evolution. *Trends Biochem Sci* **44**, 675-687 (2019).
48. Wruck, F., Katranidis, A., Nierhaus, K.H., Buldt, G. & Hegner, M. Translation and folding of single proteins in real time. *Proc Natl Acad Sci U S A* **114**, E4399-E4407 (2017).
49. Englander, S.W. & Mayne, L. The nature of protein folding pathways. *Proc Natl Acad Sci U S A* **111**, 15873-80 (2014).
50. Wittinghofer, A. & Vetter, I.R. Structure-function relationships of the G domain, a canonical switch motif. *Annu Rev Biochem* **80**, 943-71 (2011).
51. Schmidt, A. et al. The quantitative and condition-dependent Escherichia coli proteome. *Nat Biotechnol* **34**, 104-10 (2016).
52. Clarke, T.F.t. & Clark, P.L. Rare codons cluster. *PLoS One* **3**, e3412 (2008).
53. Smith, S.B., Cui, Y. & Bustamante, C. Optical-trap force transducer that operates by direct measurement of light momentum. *Methods Enzymol* **361**, 134-62 (2003).
54. Pettersen, E.F. et al. UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem* **25**, 1605-12 (2004).

Figures

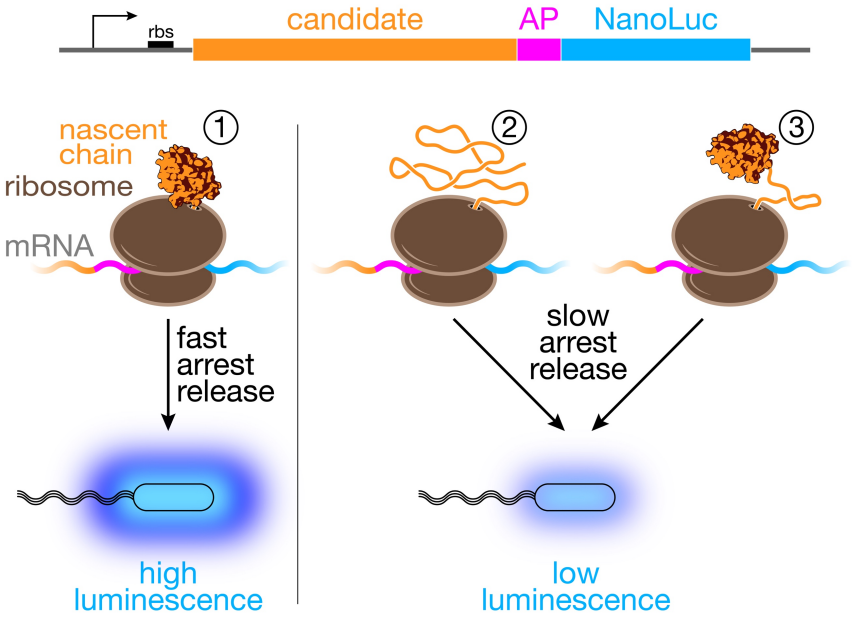


Figure 1. NanoLuc reporter assay for monitoring nascent chain folding *in vivo*. Plasmids for expression in *E. coli* encode candidate proteins fused to the SecM arrest peptide (AP) and the NanoLuc reporter (diagram on top). Upon expression in host cells, nascent chain synthesis is arrested at the end of the AP. When the candidate is folded and in close proximity to the ribosome at the arrest point (①), force on the nascent polypeptide inside the ribosome exit tunnel accelerates arrest release. Accumulation of the full-length fusion protein that includes NanoLuc results in high luminescence. When the nascent chain is not stably folded (②) or separated from the ribosome by an unfolded segment (③), arrest release occurs more slowly, and cells exhibit low luminescence.

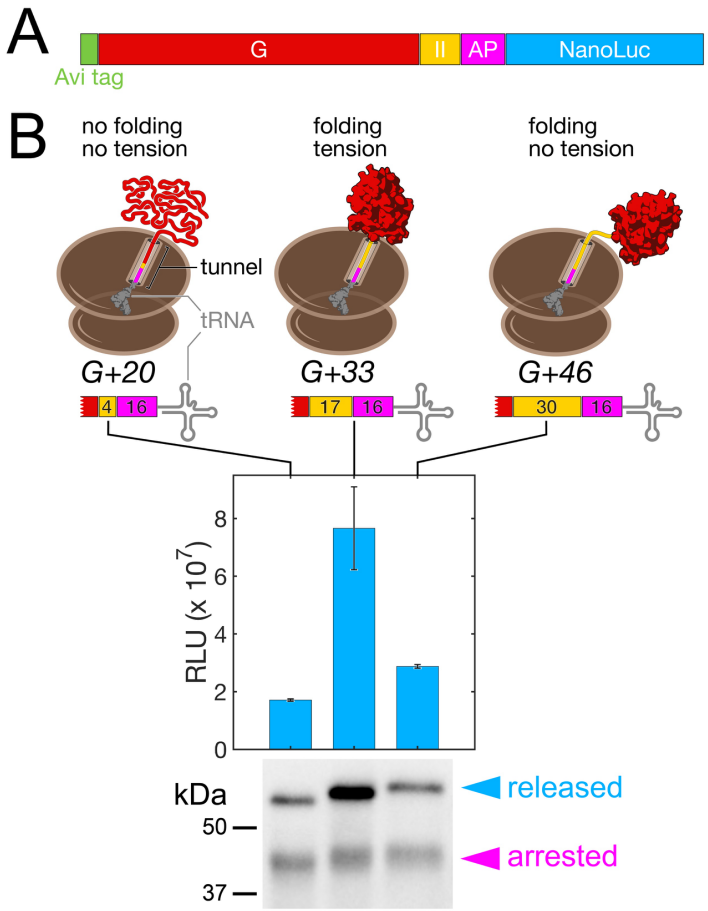


Figure 2. The G-domain folds upon emerging from the ribosome during EF-G synthesis. A. Construct design for measuring folding of the G-domain (G) in live cells. The G-domain is extended by a variable number of domain II (II) residues in the folding sensor fusion construct (compare to Figure 1A). **B.** Arrest release measurements for three distinct lengths of EF-G nascent chains. Luminescence measurements (bar diagram) show that arrest release is high in the *G+33* construct, compared to shorter (*G+20*) or longer (*G+46*) nascent chains. Cartoons (top) illustrate the positioning of the G-domain (red), the domain II segment (yellow) and the arrest peptide (magenta) relative to the tRNA in the ribosome P-site. Western blot analysis (bottom panel) shows similar amounts of arrested nascent chains (magenta arrowhead), but clear differences in the accumulation of the full-length fusion protein (blue arrowhead), confirming that the luminescence signal is due to accelerated arrest release. Error bars represent the standard deviation from 5 independent experiments. RLU: relative light units.

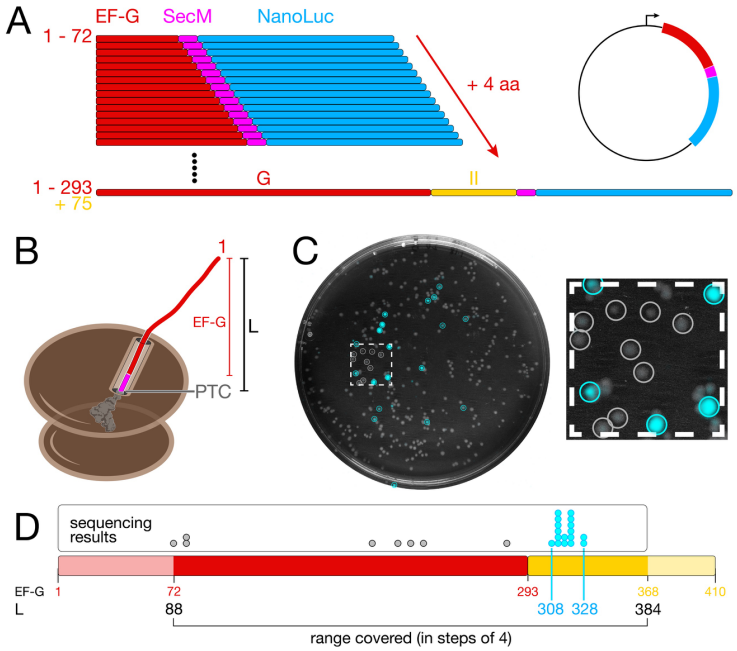


Figure 3. On-plate screening identifies nascent chain folding in a pool of constructs. A. Construction of a plasmid pool containing increasingly longer EF-G candidate polypeptides. The shortest construct encodes 72 N-terminal amino acids of the G-domain (red), the longest construct encodes the full G-domain plus 75 amino acids from domain II (yellow). **B.** Cartoon illustrating the nomenclature for the plasmid pool. The length L represents the number of amino acids from the peptidyl-transferase center (PTC) of the ribosome to residue 1 of the EF-G candidate. **C.** Merged brightfield image of the agar plate (grayscale) and colony luminescence (cyan). The inset shows a magnification of the boxed region. Colonies selected for sequencing are circled. **D.** Colony sequencing results. Each dot represents the sequencing result from an individual colony, mapped onto the position within the EF-G candidate sequence (bar diagram). All constructs yielding high luminescence (cyan dots) cluster in the region $308 \leq L \leq 328$. All non-luminescent samples have shorter chain lengths (grey dots). Red and yellow numbers indicate residue positions in the EF-G amino acid sequence.

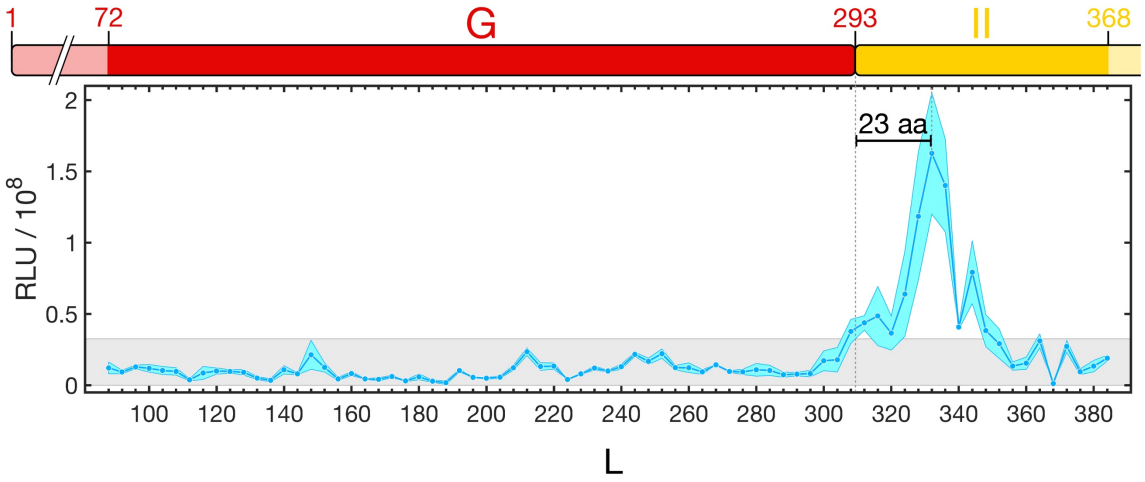


Figure 4. Folding profile of EF-G nascent chains. The plot shows the luminescence readings from 75 individual cultures. The blue line indicates the mean from three independent measurements, with the standard deviation indicated by the cyan-colored area. A strong peak centered around L = 332, where the G-domain has been extruded from the ribosome. Shorter and longer constructs outside the main peak region show only background luminescence (grey box indicates 20% of maximum signal), suggesting the absence of folding intermediates before synthesis of the entire domain is complete. The EF-G domain diagram on top is drawn for reference to indicate synthesis progress. Sequence positions are offset from L by 16 amino acids, the length of the SecM arrest peptide.

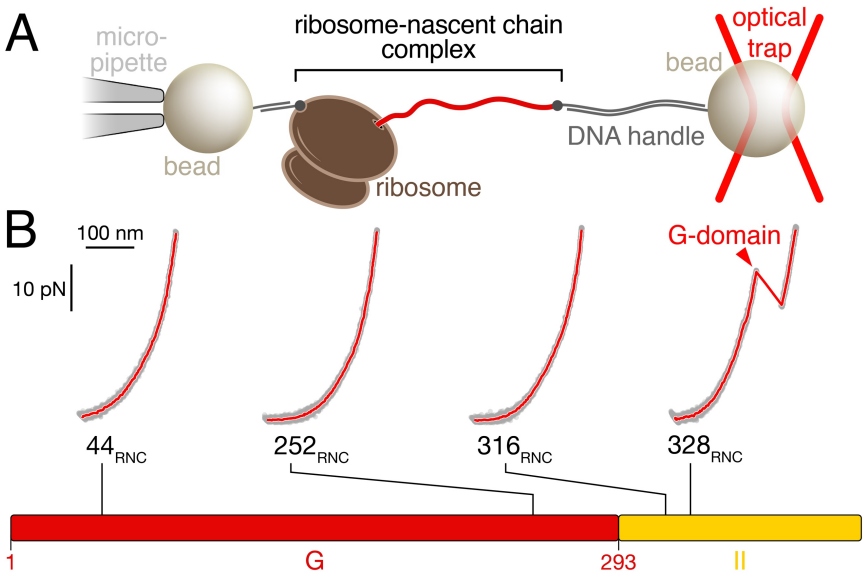


Figure 5. Nascent chain structure probed with single-molecule force spectroscopy. A. Cartoon schematic (not to scale) of the experimental setup. A single ribosome-nascent chain complex (RNC) is tethered between two polystyrene beads by means of DNA handles. Moving the bead held in the optical trap enables application of force to nascent EF-G polypeptides. **B.** Representative force-extension curves of RNCs with four distinct lengths (grey dots: raw data, 1000 Hz; red lines: filtered data, 30 Hz). When the G-domain is fully outside the ribosome (328_{RNC}), the characteristic unfolding transition of the G-domain is observed (red arrowhead). Shorter chain lengths (44_{RNC}, 252_{RNC}, 316_{RNC}) do not show defined unfolding transitions. The bar diagram at the bottom indicates the positions of the chain lengths in the EF-G primary structure.

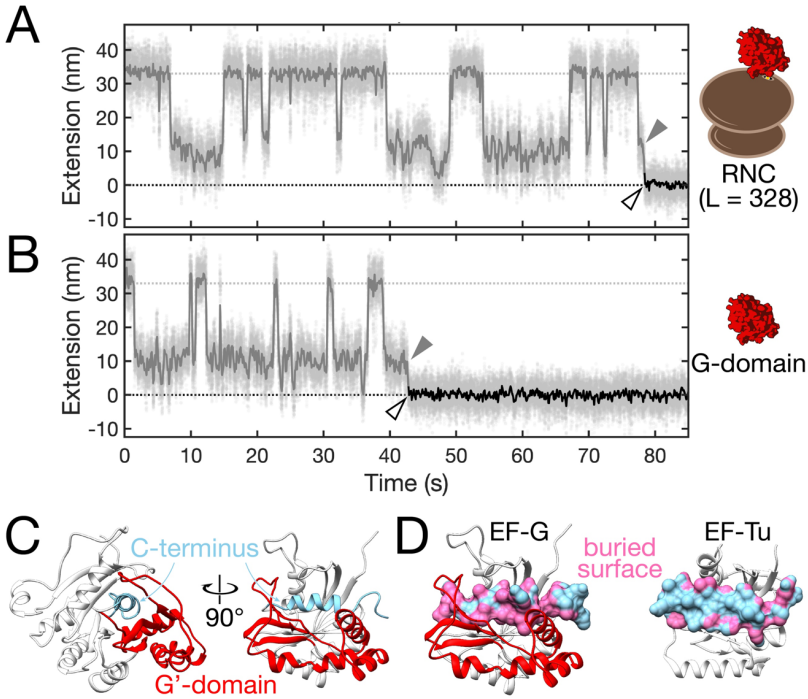


Figure 6. Refolding of the G-domain on and off the ribosome. A. Refolding of the ribosome-bound G-domain. Starting with the unfolded G-domain at t = 0, refolding against a constant force of 3.5 pN is followed by recording changes in molecular extension. The domain visits several partially folded states with extensions between the unfolded (grey dotted line) and folded (black dotted line) states. At least one state represents an on-pathway folding intermediate (grey arrowhead) that is visited immediately before the domain reaches the folded state (open arrowhead). Grey dots: raw data (1000 Hz), solid lines: filtered data (30 Hz). **B.** Same as A. for the isolated G-domain. The overall trajectory resembles that of the ribosome-bound G-domain. **C.** Cartoon representation of the G-domain. The G' insertion (red) and the remaining part of the G-domain (white) enclose the C-terminal region (light blue) containing the last α-helix. **D.** G-domains from EF-G (left) and EF-Tu (right) in the same orientation as the right panel in C., with the C-terminal region shown in surface representation. The surface area of this region that is buried in the native structure is shown in pink. Whereas the C-terminus is mostly buried in EF-G, the face closest to the viewer is largely accessible in EF-Tu. Structural representations were generated with Chimera [54] using pdb coordinates 4v9p (EF-G) and 6eze (EF-Tu).

Supplementary Information

Synthesis runs counter to the directional folding pathway of a nascent protein domain

Xiuqi Chen [1], Nandakumar Rajasekaran [1], Kaixian Liu [1,‡] and Christian M. Kaiser [2,3,*]

[1] CMDB Graduate Program, Johns Hopkins University, Baltimore, MD, USA

[2] Department of Biology, Johns Hopkins University, Baltimore, MD, USA

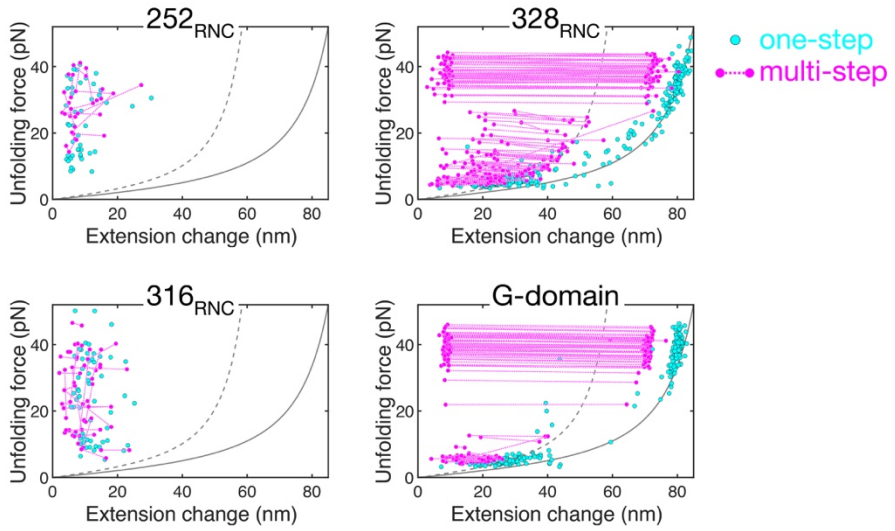
[3] Department of Biophysics, Johns Hopkins University, Baltimore, MD, USA

[‡] present address: Molecular Biology Program, Sloan Kettering Institute, New York, NY

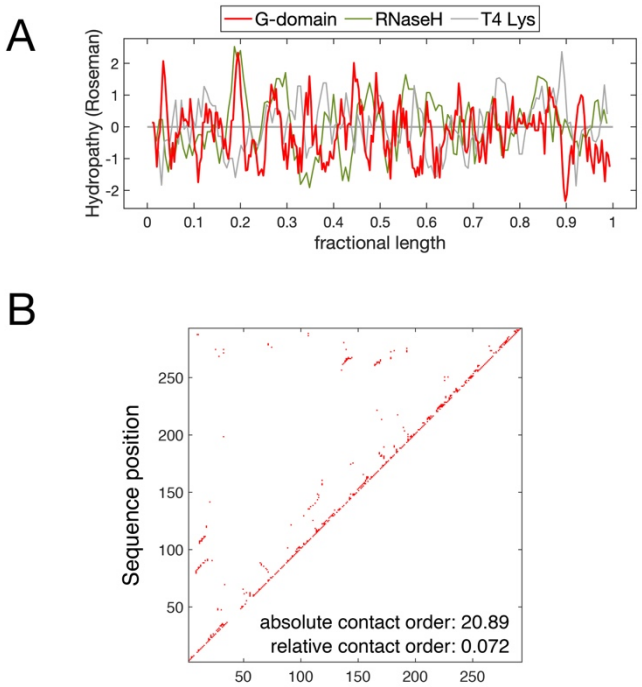
[*] corresponding author (kaiser@jhu.edu)

Supplementary Information includes Supplementary Figures 1 to 4

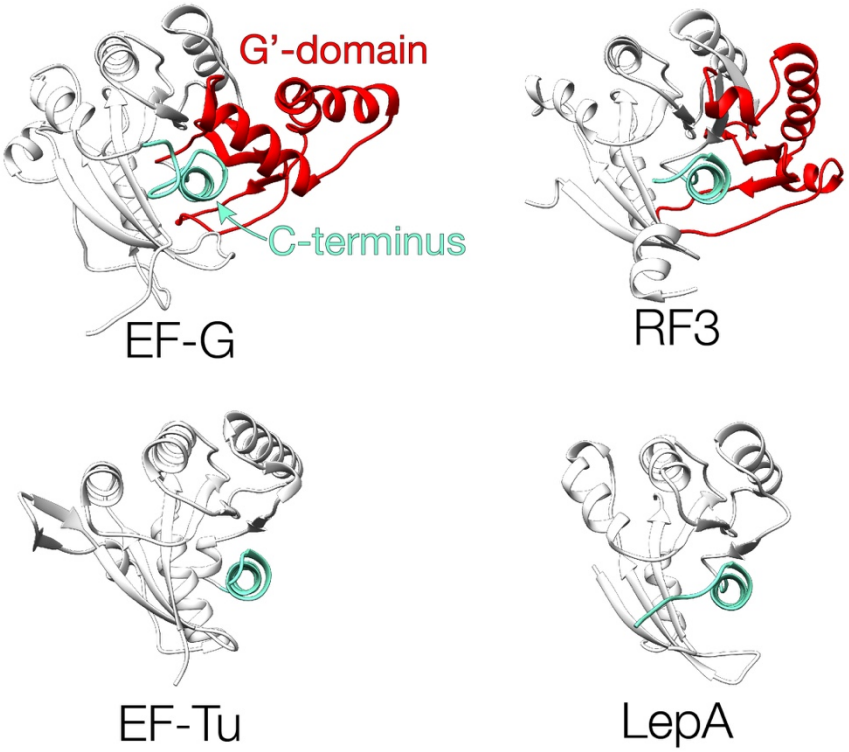
Supplementary Figures



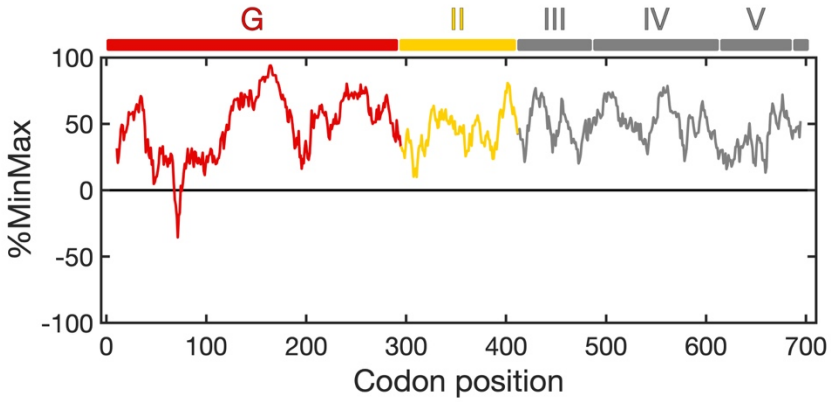
Supplementary Figure 1. Unfolding transitions from force ramp experiments. The scatter plots show the transitions observed in force-ramp measurements of nascent EF-G polypeptides (*252RNC*, *316RNC*, 328*RNC*) and the isolated G-domain. Every circle represents an unfolding event, characterized by the force at which it occurs and the associated change in molecular extension. Single unfolding events are shown in cyan. Serial transitions from the same force-extension curve are represented as magenta circles, connected by dotted lines. *328RNC* and the G-domain show defined unfolding transitions in the range between ~30 pN to ~45 pN, characteristic of G-domain unfolding [18,35]. *328RNC* also shows transitions resulting from the incompletely folded G-domain. In contrast, transitions recorded with *252RNC* and *316RNC* are short, heterogeneous and spread over a wide force range. Almost all *328RNC* and G-domain traces contained measurable transitions. Transitions were observed infrequently for the incomplete nascent chains (*252RNC*: 60 traces with transitions, 213 traces total; *316RNC*: 85 traces with transitions, 363 traces total). WLC-models with contour lengths of L_C = 103.7 nm (full G-domain unfolding) and L_C = 73.1 nm (intermediate lacking the C-terminal 90 amino acids; see main text for details) are indicated by solid and dashed grey lines, respectively. The data shown for *328RNC* and the G-domain have previously been published [35].



Supplementary Figure 2. Hydrophobicity and contact order of the G-domain. A. Hydrophobicity for the EF-G G-domain (red), calculated using the Roseman scale [44], averaged with a window size of 5 amino acids. Fractional length is the ratio of sequence position and total sequence length. The hydrophobicity profiles for ribonuclease H (olive) and T4 lysozyme (grey) are plotted for comparison, showing similar hydrophobicity. **B.** Contact map of the EF-G G-domain. Red squares indicate positions where residue pairs are in contact (atom distances within 6 Å). The absolute and relative contact order [43] are indicated in the plot.



Supplementary Figure 3. Examples of G-domain structures. Cartoon representations of the G-domains from 4 elongation factors (EF-G, pdb: 4v9p; RF3, pdb: 2h5e; EF-Tu, pdb: 6eze; LepA: pdb: 3cb4). The C-terminal regions (aquamarine) in EF-G and RF3 are mostly buried due to the presence of the G' insertion (red) in these proteins. The same region is more surface accessible in EF-Tu and LepA, which lack the insertion. Cartoons were generated using Chimera [54].



Supplementary Figure 4. Codon usage of the EF-G coding sequence. Codon usage over the complete EF-G coding sequence was calculated using a web interface (www.codons.org) based on the %MinMax algorithm [52] with a window size of 18. Domain boundaries in the EF-G protein are indicated on top. Extensive clusters of rare codons are absent, except for a short region around codon position 70.