Evidence of evolutionary selection for cotranslational folding

William M. Jacobs\* and Eugene I. Shakhnovich\*,1

\*Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA 02138

Edited by Susan Marqusee, University of California, Berkeley, CA, and approved September 7, 2017 (received for review April 7, 2017)

Recent experiments and simulations have demonstrated that proteins can fold on the ribosome. However, the extent and generality of fitness effects resulting from cotranslational folding remain open questions. Here we report a genome-wide analysis that uncovers evidence of evolutionary selection for cotranslational folding. We describe a robust statistical approach to identify loci within genes that are both significantly enriched in slowly translated codons and evolutionarily conserved. Surprisingly, we find that domain boundaries can explain only a small fraction of these conserved loci. Instead, we propose that regions enriched in slowly translated codons are associated with cotranslational folding intermediates, which may be smaller than a single domain. We show that the intermediates predicted by a native-centric model of cotranslational folding account for the majority of these loci across more than 500 *Escherichia coli* proteins. By making a direct connection to protein folding, this analysis provides strong evidence that many synonymous substitutions have been selected to optimize translation rates at specific locations within genes. More generally, our results indicate that kinetics, and not just thermodynamics, can significantly alter the efficiency of self-assembly in a biological context.


cotranslational folding | synonymous codon usage | protein-folding intermediates | free-energy landscapes

Many proteins can begin folding to their native states before their synthesis is complete (1, 2). As much as one-third of a bacterial proteome is believed to fold cotranslationally (3), with an even higher percentage likely in more slowly translated eukaryotic proteomes. Numerous experiments on both natural and engineered amino acid sequences have shown that folding during synthesis can have profound effects: Compared with denatured and refolded chains, cotranslationally folded proteins may be less prone to misfolding (4–11), aggregation (12), and degradation (13), or they may preferentially adopt alternate stable structures (14–16). Because the timescales for protein synthesis and folding are often similar (17, 18), it is clear that the rate of translation can be used to tune the self-assembly of peptide chains in vivo (19, 20). To this point, however, there exists little evidence that evolution has selected specifically for efficient cotranslational folding kinetics across any substantial fraction of an organism’s proteome.

In this work, we provide evidence that evolutionary selection has tuned protein-translation rates to optimize cotranslational folding pathways. Our approach is motivated by the hypothesis that pauses during protein synthesis may be beneficial for promoting the formation of native structure. By increasing the separation between the timescales for folding and translation, such pauses may promote the assembly of on-pathway intermediates, which, in turn, template the growth of further native structure. Many experimental and computational studies have shown that protein folding naturally proceeds in a stepwise manner via structurally distinct intermediates (21, 22) and that cooperative folding cannot commence until a minimal number of residues have emerged from the ribosome exit tunnel (23–26). These general findings suggest that any beneficial pauses during synthesis should occur at specific locations within an amino acid sequence.

Using a coarse-grained model of cotranslational folding, we find that translational pauses tend to be associated with stable, native-like cotranslational folding intermediates. The relevant folding intermediates are typically not complete structural domains, as has often been assumed (27), and may be distinct from intermediates that are observed when refolding from a denatured ensemble. By comparing putative translational pause sites with a neutral model that accounts for gene-specific codon use, we show that evolutionarily optimized cotranslational folding is a widespread feature of the *Escherichia coli* genome. Our results therefore highlight the extent to which evolution has tuned the self-assembly pathways, and not just the native structures, of complex biomolecules.

Results

Unbiased Identification of Slowly Translated Regions. Our analysis of beneficial pauses in protein synthesis relies on the identification of regions within mRNA transcripts that are enriched in “rare” codons (SI Appendix, Table S1), i.e., codons that are used substantially less often than alternate synonymous codons in highly expressed genes (28). Despite numerous attempts to predict codon-specific translation rates based on physical factors (29–32), such as tRNA concentrations, translation-speed estimates based on relative-use metrics (28, 33) remain among the most accurate (34–36). Thus, using codon rarity as a proxy for translation speed, we can look for pauses in synthesis by identifying regions in a mRNA transcript that are locally enriched in rare codons.

However, an appropriate neutral model must account for two potential sources of synonymous codon-use bias at the level of an individual gene. First, we controlled for the overall rare-codon degradation (13), or they may preferentially adopt alternate stable structures. By increasing the separation between the timescales for folding and translation, such pauses may promote the assembly of on-pathway intermediates, which, in turn, template the growth of further native structure. Many experimental and computational studies have shown that protein folding naturally proceeds in a stepwise manner via structurally distinct intermediates (21, 22) and that cooperative folding cannot commence until a minimal number of residues have emerged from the ribosome exit tunnel (23–26). These general findings suggest that any beneficial pauses during synthesis should occur at specific locations within an amino acid sequence.

Using a coarse-grained model of cotranslational folding, we find that translational pauses tend to be associated with stable, native-like cotranslational folding intermediates. The relevant folding intermediates are typically not complete structural domains, as has often been assumed (27), and may be distinct from intermediates that are observed when refolding from a denatured ensemble. By comparing putative translational pause sites with a neutral model that accounts for gene-specific codon use, we show that evolutionarily optimized cotranslational folding is a widespread feature of the *Escherichia coli* genome. Our results therefore highlight the extent to which evolution has tuned the self-assembly pathways, and not just the native structures, of complex biomolecules.

Unbiased Identification of Slowly Translated Regions. Our analysis of beneficial pauses in protein synthesis relies on the identification of regions within mRNA transcripts that are enriched in “rare” codons (SI Appendix, Table S1), i.e., codons that are used substantially less often than alternate synonymous codons in highly expressed genes (28). Despite numerous attempts to predict codon-specific translation rates based on physical factors (29–32), such as tRNA concentrations, translation-speed estimates based on relative-use metrics (28, 33) remain among the most accurate (34–36). Thus, using codon rarity as a proxy for translation speed, we can look for pauses in synthesis by identifying regions in a mRNA transcript that are locally enriched in rare codons.

However, an appropriate neutral model must account for two potential sources of synonymous codon-use bias at the level of an individual gene. First, we controlled for the overall rare-codon

Significance

The mechanisms that affect protein production are crucially important but incompletely understood. In particular, the speed at which a polypeptide chain is synthesized can affect whether it folds to its native state. We propose that the optimal rate of translation depends on the sequence of stable conformations that are sampled by a nascent chain as it is being synthesized. Using optimal and rare synonymous codons as a proxy for translation speed, we find that transient pauses in synthesis are associated with specific intermediate conformations that appear during cotranslational folding and that this behavior is conserved across hundreds of genes and multiple prokaryotic genomes. These results provide striking evidence of evolutionary selection for efficient protein folding in vivo.

Author contributions: W.M.J. and E.I.S. designed research; W.M.J. performed research; W.M.J. contributed new reagents/analytic tools; W.M.J. and E.I.S. analyzed data; and W.M.J. and E.I.S. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

This is an open access article distributed under the PNAS license.

1To whom correspondence should be addressed. Email: shakhnovich@chemistry.harvard.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1705772114/-/DCSupplemental.
use in a gene, which is defined as the fraction of rare codons in the entire transcript (SI Appendix, Fig. S1). Multiple factors have been hypothesized to contribute to the overall degree of codon adaptation of each gene, including evolutionary selection for rapid synthesis, accurate translation, and the stability of mRNA transcripts (37). By taking a gene’s average codon use into account, we instead pick out regions that are locally enriched in rare codons relative to the gene-specific background. Second, we accounted for synonymous-codon bias due to the amino acid composition of the protein sequence. Assuming that amino acid sequences are under stronger selection pressure and can thus be considered immutable, we estimated the average rare-codon frequencies for each amino acid type among all genes with a similar level of rare-codon use. Having controlled for the overall rare-codon use and the amino acid sequence, we modeled neutral codon use as a Bernoulli process with sequence-dependent rare-codon probabilities (SI Appendix, section S1A).

Evaluation of Evolutionary Conservation. Next, we assessed the functional importance of local rare-codon enrichment by looking for conservation of rare-codon use across multiple-sequence alignments (Fig. 1). We extended the neutral model described above to 18 sufficiently diverged prokaryotic genomes, with rare-codon definitions and gene-specific rare-codon probabilities computed for each genome independently. Here our approach differs from conventional conservation analyses, because we are interested in the enrichment of rare codons within contiguous 15-codon segments of a transcript, as opposed to the codon use at each aligned site (38, 39). By examining conservation of rare-codon enrichment, we can identify local regions that do not align precisely but nevertheless result in transcriptional pauses at similar places within the protein sequence. This approach also allows for a meaningful comparison of the local rare-codon enrichment in sequence alignments that contain insertions and deletions. Our choice of a 15-codon enrichment region is comparable to choosing the length of a typical element of protein secondary structure, and we verified that regions with widths of 10 and 20 codons yield similar results. In contrast, larger enrichment regions defined on the basis of complete domains rarely differ significantly from the background rare-codon use, while analyses of single aligned sites tend not to produce statistically significant results.

To be relevant for cotranslational folding, putative slowly translated regions must meet two criteria: a high degree of conservation of slowly translated codons and a low probability of such an occurrence in the neutral model. For a region to be considered both enriched and conserved, we required that the local concentration of rare codons deviate from the background distribution by approximately 1 SD in at least 75% of the sequences in the alignment; SI Appendix, Fig. S2 shows that our results are robust with respect to this conservation threshold. We then computed an associated $P$ value that reports the probability, within the neutral model, of randomly generating at least the observed number of enriched regions from reverse translations, i.e., by sampling synonymous sequences using the aligned amino acid sequences and a probabilistic model of the codon use for each amino acid type (SI Appendix, section S1A). This second criterion is central to our findings, as we discuss below. We emphasize that these criteria are distinct: Depending on the amino acid identities, it is possible to observe low $P$ values without significant rare-codon enrichment relative to the background, and vice versa. Consequently, both criteria must be satisfied to constitute evidence for evolutionary selection.

Our analysis reveals numerous rare-codon enrichment loci in the E. coli genome that are inconsistent with the neutral model and are thus likely to be a result of evolutionary selection (SI Appendix, Fig. S3A). Although these regions occur throughout the mRNA transcripts, their locations are biased toward both the 5′ and 3′ ends (SI Appendix, Fig. S3B). While these trends have been noted previously (40), our analysis confirms that the increased probability of rare-codon enrichment at the 3′ end is evolutionary conserved and is not a consequence of the amino acid sequences. Furthermore, we find that these biases become more pronounced as we lower the $P$-value threshold used for comparison with the neutral model (SI Appendix, Fig. S3B), suggesting that any false positives from our analysis are relatively evenly distributed throughout the transcripts. We also analyzed the codon-level similarity among the genomes in our alignments and verified that these results reflect conservation of rarity as opposed to conservation of specific codons (SI Appendix, Fig. S4).

Comparison with Predicted Cotranslational Folding Pathways. To probe the potential consequences of local rare-codon enrichment for protein folding, we next examined the formation of native-like intermediates during protein synthesis. A large body of simulation evidence (41) has shown that intermediates must be stable at equilibrium to be sampled with high probability during cotranslational folding and are likely to form only when the folding rate is fast relative to the protein elongation rate. Therefore, while an intermediate’s equilibrium free energy does not completely determine whether it will appear on a cotranslational folding pathway, we assume that stability at equilibrium is necessary for a pause in translation to promote the development of native structure.

**Fig. 1.** An example multiple-sequence alignment identifies conserved rare-codon enrichment within the gene folP. (Upper) A histogram shows the number of sequences that have a given local concentration of rare codons at each position in the alignment. The local concentrations of rare codons are determined within 15-codon regions. Based on the average occurrence of rare codons in folP, local rare-codon concentrations of at least 3/15 are considered to be enriched and are colored red in the histogram, while 15-codon regions with fewer than 3 rare codons are not enriched and are shown in blue. (Lower) The fraction of sequences that are enriched at each position in the alignment (red) is shown along with the corresponding neutral-model $P$ value (black), as explained in the main text. Conserved regions, where at least 75% of the sequences are enriched, are highlighted.
Here we applied a coarse-grained model (22) to predict the formation of stable partial structures during nascent-chain elongation. Importantly, this model captures the tertiary structure of nascent chains and does not assume that domains fold cooperatively or independently. To model cotranslational folding, a nascent chain of length $L$ is allowed to form native contacts among the first $L$ residues of the full protein. We then computed the minimum free energy of a nascent chain, relative to an unfolded ensemble, using a mean-field theory based on the protein’s native structure (SI Appendix, section S1B). This approach captures the opposing contributions to the free energy from energetically favorable native contacts and the configurational entropy of an unfolded chain. We used a native-centric energy function that emphasizes hydrogen bonds and contacts between larger residues (22), while the thermodynamic stability of the native state is fixed based on the full protein length (42). We show in SI Appendix, Fig. S5 that tuning the native-state stability does not significantly affect the results of our analysis.

Our calculations predict that, in general, native structure forms discontinuously during nascent-chain elongation. In the example shown in Fig. 2, Lower, decreases in the nascent-chain free energy occur at distinct chain lengths. These sudden drops correspond to the appearance of stable intermediates with native-like tertiary structure. In contrast, at chain lengths corresponding to the intervening plateaus, the nascent-chain free energy remains constant because the newly synthesized residues cannot form sufficient stabilizing contacts with any existing tertiary structure. Unsurprisingly, the probability of finding a stable on-pathway intermediate increases as synthesis nears completion (SI Appendix, Fig. S6).

We are now in a position to test the relationship between translocation pausing and the formation of native-like intermediates. The ribosome exit tunnel is widely believed to conceal between 30 aa and 40 aa (35, 43), although a greater number may be accommodated in partially helical conformations (44). In addition, some tertiary structure formation may commence within the exit-tunnel vestibule (45). A beneficial pause in synthesis should therefore be separated from a cotranslational intermediate by a distance that is roughly equivalent to the exit-tunnel length (Materials and Methods). An example of this correspondence is shown in Fig. 2, where a putative translational pause is located ~30 residues downstream of the formation of a predicted intermediate. However, we emphasize that, according to the present hypothesis, the formation of an intermediate is necessary but not sufficient to expect that a translational pause would be beneficial. For example, intermediates that fold quickly relative to the average translation rate or appear less than the exit-tunnel distance from the end of the protein are unlikely to be accompanied by a productive pause.

**Conserved, Enriched Regions Associate with Predicted Cotranslational Folding Intermediates.** By applying this analysis to a set of ~500 E. coli proteins with known native structures, we find widespread support for our cotranslational folding hypothesis. In particular, we find that the cotranslational folding intermediates predicted by our coarse-grained model account for a significant proportion (>50%) of the putative slowly translated regions (Fig. 3). Most importantly, we find that the fraction of rare-codon–enriched regions that can be explained by our model increases consistently as we reduce the $P$-value threshold for establishing evolutionary conservation. In other words, the predictive power of our model improves as false positives related to the random clustering of rare codons are preferentially eliminated. This trend is also robust with respect to variations in the precise definition of codon rarity (SI Appendix, Fig. S7).

We further tested the sensitivity of our cotranslational folding predictions by repeating the above analysis with randomized control sequences, which preserve the total number of pause sites at each $P$-value threshold but uniformly distribute their locations across the transcripts (Fig. 3 and Materials and Methods). Although a significant fraction (~35%) of the fictitious pause sites in the randomized sequences can also be explained by our model, likely due to chance overlaps with predicted intermediates, the difference between the genomic and randomized data increases markedly at lower $P$-value thresholds (one-sided $P < 10^{-7}$ at neutral-model $P$-value thresholds below 0.01; SI Appendix, Fig. S8A). Two alternative controls (SI Appendix, Fig. S9), in which the randomized pause sites are drawn from a nonuniform distribution with a 3′-end bias or obtained directly from reverse translations, verify that our results are not solely a consequence of the 3′-end rare-codon bias in the mRNA transcripts or the amino acid sequences of the proteins.

Next, we performed inverse tests to assess whether cotranslational folding intermediates are preferentially associated with putative translational pauses. However, because the formation of an intermediate is not in itself a sufficient condition for a translational pause to be beneficial, we find that the overall frequency of such associations is small relative to the number of predicted intermediates (SI Appendix, Fig. S10). We therefore computed the odds ratio of finding conserved, rare-codon–enriched regions just downstream of a predicted intermediate, as opposed to elsewhere in a mRNA transcript. The results shown in Fig. 4 confirm that the association between folding intermediates and translational pause sites is highly significant (one-sided $P < 10^{-7}$ at
Fig. 3. (Upper) The fraction of conserved, rare-codon–enriched regions that follow a predicted cotranslational folding intermediate increases as false positives are systematically eliminated. In contrast, folding intermediates precede a consistently smaller fraction of the uniformly distributed enriched regions in randomized sequences. (Lower) Analyzing domain boundaries instead of folding intermediates similarly exhibits no dependence on the P-value threshold and accounts for a much lower percentage of the observed rare-codon enrichment loci. The error bars on the control distributions indicate the SD of 100 randomizations, while the error bars on the genomic data are estimated from binomial distributions at each P-value threshold.

neutral-model P-value thresholds below 0.01; SI Appendix, Fig. S8B) and, importantly, is not related to the overall frequency of predicted cotranslational intermediates. Here again, the predictive power of our model shows a strong dependence on the P-value threshold used for screening putative pause sites. In contrast, tests with randomized control sequences do not deviate from an odds ratio of unity.

We also applied our analysis to structural domain boundaries, which have previously been suggested to play a role in coordinating cotranslational folding (46). Nevertheless, in agreement with more recent works (27), we find little evidence of selection for translational pausing at domain boundaries. For these comparisons, we used domain definitions for approximately 800 E. coli proteins from the Structural Classification of Proteins (SCOP) database (47). Fig. 3 shows that domain boundaries explain a much smaller fraction (≤12%) of the putative pause sites than our folding model. Furthermore, the predictive power of the domain-boundary hypothesis does not vary with the P-value threshold, and the odds ratios are nearly indistinguishable from the randomized controls (Fig. 4). These conclusions also hold for various related hypotheses: Instead of assuming that a domain must be completely synthesized before folding, we tested models where native structure begins to form either at a fixed number of residues before the domain boundary or at a fixed percentage of the domain length (SI Appendix, Fig. S11). In all cases, the correspondence between the domain boundaries and the conserved, rare-codon–enriched loci is significantly weaker than the results of our cotranslational folding model. While these findings do not imply that domain boundaries are irrelevant for cotranslational folding, we can conclude that the domain-boundary hypothesis is insufficient to explain the vast majority of conserved, slowly translated regions.

Discussion

By integrating a multiple-sequence analysis of synonymous codon conservation with protein-folding theory, we have shown that highly conserved rare-codon clusters preferentially associate with predicted cotranslational folding intermediates. The putative pause sites in the E. coli genome that are both evolutionarily conserved and unaccounted for by the neutral model systematically appear downstream of predicted cotranslational folding intermediates at distances that are similar to the length of the ribosome exit tunnel. Our large-scale study therefore supports...
the hypothesis that beneficial pauses during protein synthesis follow key steps in the assembly of native structure. Comparisons with randomized control sequences confirm that our observations are highly significant.

This analysis of cotranslational folding pathways, as opposed to elements of the static native structure, provides insights into the interplay between translation and the self-assembly kinetics of nascent proteins. The stabilization of a partial structure often occurs well before a native domain is completely synthesized, especially in cases where the domain comprises more than 200 residues. In particular, subdomain cotranslational folding intermediates typically appear when sufficient tertiary contacts are available to compensate for the loss of chain entropy that is required for folding. Overall, a relatively small fraction ($\leq 15\%$) of all predicted intermediates are followed by conserved translational pauses (SI Appendix, Fig. S10), but the association between folding intermediates and conserved pauses is highly significant (Fig. 4). This observation is consistent with our hypothesis, since the effect of a pause depends on the relative timescales for translation and folding, as well as potential interference due to non-native interactions. In addition, this observation explains why pause sites are not preferentially associated with domain boundaries: Although fully synthesized domains may be stable on the ribosome, the prior formation of a partial-chain intermediate is likely to affect the subsequent folding rates for other parts of the protein. As a result, the entire cotranslational folding pathway must be considered when interpreting the effect of a pause in translation. We anticipate that an optimal translation protocol (40) occurs well before a native domain is completely synthesized such that the expected number of fictitious enriched regions is equal to the randomized control sequences from which the control distributions are available to compensate for the loss of chain entropy that is required for folding.

Sample size used here.

While this paper was under review, we became aware of a contemporaneous study (57) that identifies conserved rare-codon clusters via a complementary statistical analysis. The authors also observe extensive rare-codon conservation across mRNA transcripts and similarly find no evidence of enrichment near domain boundaries.

Synonymous substitutions can also affect protein synthesis through mechanisms that are unrelated to protein folding, most notably via changes to mRNA secondary structure and stability (37). However, many experimental studies have shown that these effects originate predominantly from substitutions near the 5′-end of the mRNA transcripts and typically modulate the total protein production as opposed to the protein quality (58, 59). Such mRNA-specific effects are thus a likely explanation for the observed 5′-end bias in rare-codon enrichment, where variations in translation speed are unlikely to play a role in cotranslational folding. Consequently, we have excluded N-terminal rare codons from our analysis. In addition to rare-codon use, various studies have proposed that additional factors, such as interactions between the nascent chain and the ribosome exit tunnel (35) or the presence of internal Shine–Dalgarno motifs (30), can affect translation rates. It is likely that a more complete picture of sequence-dependent translation kinetics will enable further refinements to the cotranslational folding model presented here.

In conclusion, our study highlights the importance of optimal kinetic pathways for efficient biomolecular self-assembly. Although a protein’s amino acid sequence entirely determines its thermodynamically stable structure, it is becoming increasingly clear that synonymous mutations are not always silent. Our analysis provides strong evidence that evolutionary selection has tuned local translation rates to improve the efficiency of cotranslational protein folding. Further work is needed to understand the relationship between genome-wide codon use and translation rates and to improve the prediction of cotranslational folding intermediates, including those that contain significant amounts of nonnative structure. Nevertheless, our results indicate that folding kinetics play a role in evolutionary selection and suggest that similar relationships may exist for other biological self-assembly phenomena, such as the assembly of macromolecular complexes.

### Materials and Methods

We constructed alignments based on the amino acid sequences of homologous genes from 18 prokaryotic species with between 50% and 85% average amino acid sequence identity to E. coli (SI Appendix, Table S2). We then computed P values associated with rare-codon–enriched regions, assuming biased reverse translations and a gene-specific model for the probability of each amino acid type being encoded by a rare codon. Consensus crystal structures were constructed for 511 nonmembrane E. coli proteins with 500 residues or fewer, using Protein Data Bank (60) entries containing complete structures for sequences with at least 95% amino acid identity to the E. coli gene. SCOP domain assignments were obtained from ref. 47 for all proteins with at most 500 residues. Due to the uncertainty in the number of amino acids that are concealed in the ribosome exit tunnel and the potential for steric interactions between folding intermediates and the ribosome, we consider a rare-codon–enriched region to be associated with a folding intermediate if the enriched region is anywhere between 20 and 60 codons downstream from the position at which an intermediate first becomes stable, ignoring enriched regions within the first 80 codons of a transcript. An intermediate is identified whenever the cotranslational free-energy profile decreases by more than 1 kT relative to the previous free-energy plateau; for example, see the pattern of alternating plateaus and precipitous free-energy decreases in Fig. 2, Lower. To generate the randomized control sequences from which the control distributions in Figs. 3 and 4 were calculated, we sampled locations for fictitious rare-codon–enriched regions from a uniform distribution over each mRNA transcript, excluding the first 80 codons. This uniform distribution was normalized such that the expected number of fictitious enriched regions is equal to the total number of observed enriched regions at each P-value threshold. See SI Appendix, section S1 for complete details of all methodologies. Essential data are provided in Dataset S1. All code necessary to reproduce these results is available at https://faculty.chemistry.harvard.edu/shakhnovich/software.

### ACKNOWLEDGMENTS

We thank Sanchari Bhattacharyya and Michael Manhart for many insightful discussions. This work was supported by National Institutes of Health Grants R01GM124044 and F32GM116231.

Weinberg DE, et al. (2016) Improved ribosome-footprint and mRNA measurements


