Evidence of evolutionary selection for co-translational folding

William M. Jacobs and Eugene I. Shakhnovich
Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street, Cambridge, MA, 02138, USA
(Dated: March 28, 2017)

Recent experiments and simulations have demonstrated that proteins can fold on the ribosome, but the importance of co-translational folding for the fitness of an organism remains an open question. Here we report a genome-wide analysis that uncovers evidence of evolutionary selection for co-translational folding. We describe a robust statistical approach to identify conserved loci within genes that are significantly enriched in slowly translated codons. 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 co-translational folding intermediates, which may be smaller than a single domain. We show that the intermediates predicted by a native-centric model of co-translational folding account for the majority of these loci across more than 500 E. 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.

INTRODUCTION

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 co-translationally [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 to denatured and refolded chains, co-translationally 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 optimal co-translational 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 co-translational 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 step-wise manner via structurally distinct intermediates [21–23], and that cooperative folding cannot commence until a minimal number of residues have emerged from the ribosome exit tunnel [24–27]. These general findings suggest that any beneficial pauses during synthesis should occur at specific locations within an amino-acid sequence.

Our analysis of the E. coli proteome reveals that the majority of beneficial translational pauses are associated with stable native-like co-translational folding intermediates, which we predict using a coarse-grained model of the co-translational folding pathway. We find that the relevant folding intermediates may not be complete structural domains, as has often been assumed [28]. Furthermore, the stable native-like structures that are predicted to form during co-translational folding are also likely to differ from the intermediates that are observed when refolding from a denatured ensemble.

To identify pauses in translation, we build our genomic analysis on the observation that synonymous codon usage is extremely non-uniform in highly expressed genes, likely as a result of adaptation for efficient protein production [29]. It has further been established that the subset of codons that appear unusually infrequently in the most highly expressed genes tend to be translated more slowly [30]. Yet to distinguish beneficial pauses from random fluctuations, it is essential that we identify deviations in translation rates that are evolutionarily conserved across species. We therefore describe a statistical method for locating regions within mRNA transcripts where a local enrichment in rare codons is conserved across multiple divergent species.

By comparing putative translational pause sites with a neutral model that accounts for gene-specific codon usage, we can assess the likelihood that the same degree of enrichment and conservation could occur by chance alone. We are thus able to show that, as our confidence in this evolutionary signal increases, the predictive power of our co-translational-folding hypothesis consistently improves as well. Examination of more than 500 proteins suggests that evolutionary selection for optimal co-translational folding is a widespread feature of the E. 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.

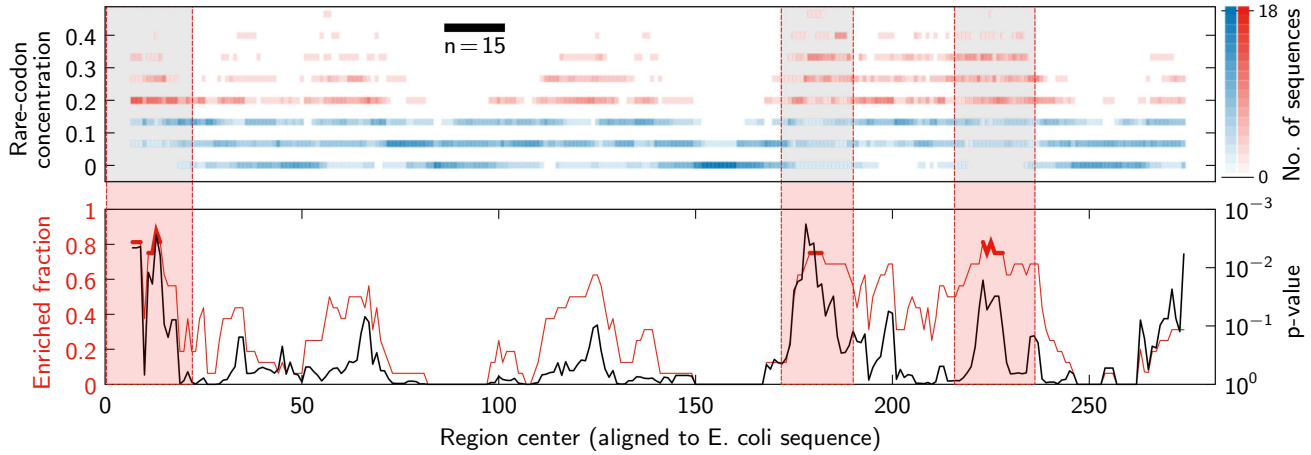

FIG. 1. An example multiple-sequence alignment identifies conserved rare-codon enrichment within the gene *folP*. Above, a histogram shows the number of sequences that have a given local concentration of rare codons at each position in the alignment. The local concentrations of rare codons are determined 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 three rare codons are not enriched and are shown in blue. Below, 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.

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, i.e. codons that are used substantially less often than alternate synonymous codons in highly expressed genes [29]. Despite numerous attempts to predict codon-specific translation rates based on physical factors [31–34], such as tRNA concentrations, translation-speed estimates based on relative-usage metrics [29, 35] remain among the most accurate [30, 36, 37]. Thus, by using the concentration of rare codons 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 these codons.

However, an appropriate neutral model must account for two potential sources of synonymous codon-usage bias at the level of an individual gene. First, we control for the overall rare-codon usage, which is defined as the fraction of rare codons in the entire transcript. Multiple factors have been hypothesized to contribute to the overall degree of codon adaptation in a gene, including evolutionary selection for rapid synthesis, accurate translation and the stability of mRNA transcripts. While the overall rare-codon usage of homologous genes is remarkably similar across genomes (Figure S1), it is nevertheless difficult to say whether this feature is due to selection for an optimal translation rate. By taking a gene's average codon usage into account, we can pick out regions that are locally enriched in rare codons relative to the gene-specific background. Second, we account for synonymous-codon bias due to the amino-acid composition of the protein sequence. We assume that amino-acid sequences are generally under stronger selection pressure and can thus be considered immutable. Yet not all amino acids are coded for by rare codons, and, among the amino acids that are, the probabilities of observing rare codons are typically not equal (Table S1). To account for effects due to the amino-acid sequence, we estimate the rare-codon frequencies for each amino acid type from the average synonymous-codon usage in all genes with a similar average rare-codon usage. We then model neutral rare-codon usage for each individual gene as a Bernoulli process with sequence-dependent probabilities of rare-codon usage (see Methods).

Evaluation of evolutionary conservation

Next, we assess the functional importance of local rare-codon enrichment by looking for conservation across multiple-sequence alignments (Figure 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 usage at each aligned site [38, 39]. As a result, we can identify locally enriched regions that do not align precisely but nevertheless result in translational 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 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 usage, while analyses of single aligned sites tend not to produce statistically significant results.

To be relevant for co-translational 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 require that the local concentration of rare codons deviate from the background distribution by approximately one standard deviation in at least 75% of the sequences in the alignment; Figure S2 shows that our results are robust with respect to this conservation threshold. We then compute 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 of the aligned amino-acid sequences. This second criterion is central to our findings, as we shall 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 in order to establish the evolutionary conservation of translational pauses.

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 (Figure S3a). Although these regions occur throughout the mRNA transcripts, their locations are biased towards both the 5' and 3' ends (Figure 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 evolutionarily 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 (Figure S3b), suggesting that any false positives from our analysis are relatively evenly distributed throughout the transcripts.

Comparison with predicted co-translational folding pathways

To probe the potential consequences of local rare-codon enrichment for protein folding, we next examine 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 in order to be sampled with high probability during co-translational folding. In addition, such intermediates are only likely to form 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 co-translational folding pathway, we predict that stability at equilibrium is necessary for a pause in translation to promote the development of native structure.

Here we apply a coarse-grained model [23] to predict the formation of stable partial structures during nascent-chain elongation. Although this approach does not provide information about potential misfolded structures, which would be necessary for a full kinetic analysis of co-translational folding, it is sufficient to test the hypothesis that translational pausing promotes the assembly of native structure. Importantly, we do not need to assume that domains fold cooperatively or independently, and we can examine potential co-translational folding intermediates at the sub-domain level. We are also not limited to analyzing elements of secondary structure, which fold rapidly and are typically unstable in isolation. In contrast, our coarse-grained model makes predictions regarding the tertiary structure of nascent chains.

To model co-translational folding, we permit a nascent chain of length L to form native contacts among the first L residues of the full protein. We then compute 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 (see Methods). This theory captures the opposing contributions to the free energy from energetically favorable native contacts and the configurational entropy of an unfolded chain. Furthermore, our approach allows native contacts to form in multiple structured regions that are connected by disordered chains. We use a native-centric energy function that emphasizes hydrogen bonds and contacts between larger residues [23], while the thermodynamic stability of the native state is fixed based on the full protein length [42] (see Methods). We show in Figure S4 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 Figure 2, the decreases in the nascent-chain free energy at discrete chain lengths correspond to the stabilization of native-like intermediates; the native contacts contributing to the intermediates are shown in the contact maps below. We also plot the free energy of the nascent-chain configurations in which the terminal residue is constrained to make contact with a pre-existing structured region. At chain lengths where this latter free energy increases, newly synthesized residues cannot bind stably to any nascent native structure. Unsurprisingly, for most proteins, the probability of finding a stable on-pathway intermediate increases as synthesis nears completion (Figure S5).

We are now in a position to test the relationship between translational pausing and the formation of native-like intermediates. Because the ribosome exit tunnel can accommodate between 30 and 60 amino acids, we ex-

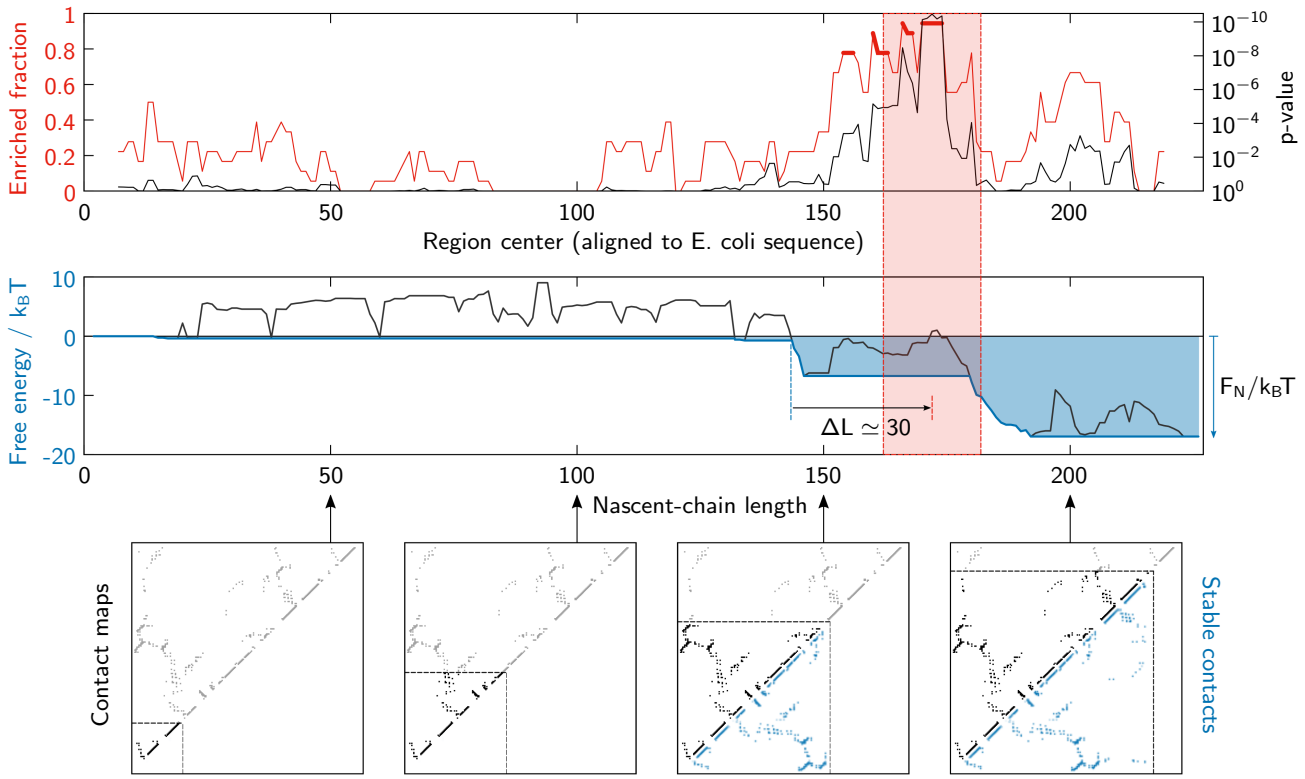

FIG. 2. Predicted co-translational folding intermediates correspond to highly conserved regions of rare-codon enrichment. Above, the fraction of enriched sequences and corresponding p-values for the gene *cmk* are shown as in Figure 1. In the middle panel, the minimum free energy, relative to the unfolded ensemble, of a nascent chain of length L is shown in blue; the stability of the native full-length protein is F_N/k_BT. Native-like intermediates become stable where this minimum-free-energy curve decreases sharply. In contrast, the free energy of nascent-chain configurations in which the Lth residue is constrained to form native interactions is non-monotonic and is shown in black. The lowest p-value for enriched sequences (highlighted region) is approximately $\Delta L = 30$ codons downstream of the first predicted co-translational folding intermediate. Below, contact maps at four nascent-chain lengths show all possible native contacts (black, above the diagonal) as well as the stable contacts (blue, below the diagonal), where the intensity indicates the contact probability in the folded ensemble.

pect that a beneficial pause in synthesis should be separated from a co-translational intermediate by a roughly equivalent distance (see Methods). An example of this correspondence is shown in Figure 2, where a putative translational pause is located approximately 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 co-translational folding intermediates

By applying this analysis to a set of approximately 500 E. coli proteins with known native structures, we find widespread support for our co-translational folding hypothesis. In particular, we find that the co-translational folding intermediates predicted by our coarse-grained model account for a significant proportion of the putative slowly translated regions (Figure 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.

We further tested the sensitivity of our co-translational folding predictions by repeating the above analysis with randomized control sequences. Figure 3 shows that the predicted fraction of putative pause sites in the E. coli genome exceeds a control distribution, in which the total number of pause sites at each p-value threshold is preserved but the locations of the enriched regions are uniformly distributed across the transcripts. Although a significant fraction ($\sim 35\%$) of the fictitious pause sites in the randomized sequences can also be explained by our model, likely due to chance overlaps with predicted

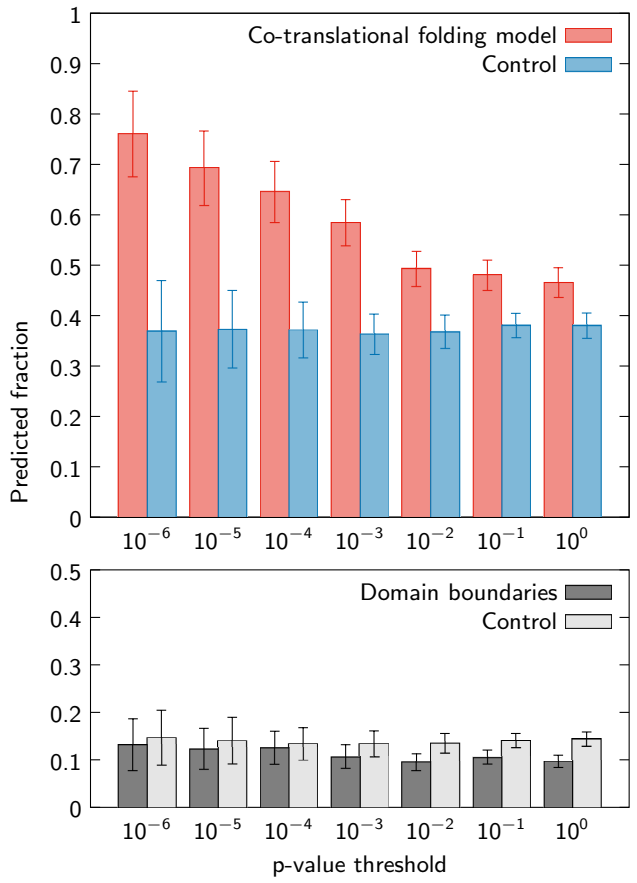

FIG. 3. Above, the fraction of conserved, rare-codon enriched regions that follow a predicted co-translational folding intermediate increases as false positives are systematically eliminated. In contrast, folding intermediates precede a consistently smaller fraction of the uniformly distributed enriched regions in randomized sequence data. Below, analyzing domain boundaries instead of co-translational 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 standard deviation of 100 randomizations, while the error bars on the genomic data are estimated from binomial distributions at each p-value threshold.

intermediates, the difference between the genomic and randomized data increases markedly at the lowest p-value thresholds (one-sided $p < 10^{-7}$ at neutral-model p-value thresholds below 0.01; see Figure S6a). It is important to note that, because we are considering local fluctuations in rare-codon usage, a neutral-model p-value threshold of 0.05 implies approximately one false positive within each 500-amino-acid protein. Consequently, we find that the predictions of our co-translational folding model are not markedly different from the control distributions unless the p-value threshold is below 0.01. Two alternative controls (Figure S7), in which the randomized pause sites are drawn from a non-uniform 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 co-translational 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 (Figure S8). 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 Figure 4 confirm that the association between folding intermediates and translational pause sites is highly significant (one-sided $p < 10^{-7}$ at neutral-model p-value thresholds below 0.01; see Figure S6b) and, importantly, is not related to the overall frequency of predicted co-translational 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 co-translational folding [43]. For these comparisons, we used domain definitions for approximately 800 E. coli proteins from the SCOP database [44]. Surprisingly, we find little evidence of evolutionary selection for translational pausing at domain boundaries. Figure 3 shows that domain boundaries explain a much smaller fraction 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 as expected, and the odds ratios are nearly indistinguishable from the randomized controls (Figure 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 prior to the domain boundary or at a fixed percentage of the domain length (Figure S9). In all cases, the correspondence between the domain boundaries and the conserved, rare-codon enriched loci is significantly weaker than the results of our co-translational folding model. While these findings do not imply that domain boundaries are irrelevant for co-translational 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 co-translational folding in-

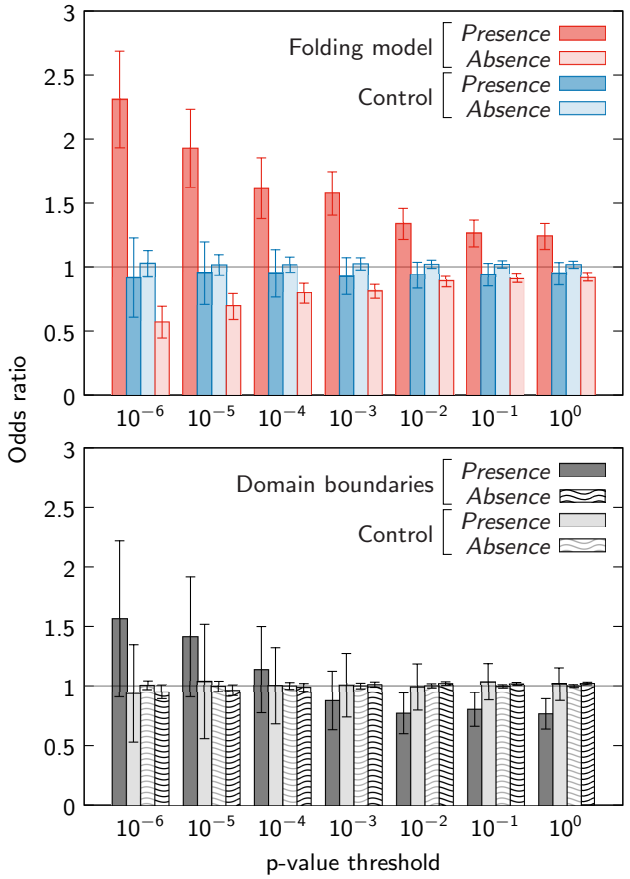

FIG. 4. Conserved regions of rare-codon enrichment are more likely to appear between 20 and 60 codons downstream of a co-translational folding intermediate than elsewhere in a mRNA transcript. Above, the odds ratio of finding an enriched region downstream of a predicted intermediate (presence) or downstream of no predicted intermediate (absence). Unlike the comparisons with randomized control sequences, both ratios deviate significantly from unity and depend on the p-value threshold used. Below, domain boundaries do not exhibit statistically significant associations with conserved pause sites. Error bars are defined as in Figure 3.

termediates. 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 co-translational 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 co-translational folding pathways, as opposed to elements of the static native structure, provides new 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, sub-domain co-translational folding intermediates typically appear when sufficient tertiary contacts are available to compensate for the loss of chain entropy that is required for folding. Of all predicted intermediates, we find that a relatively small, yet highly significant, fraction are associated with conserved translational pauses. The observation that many co-translational folding intermediates are not followed by conserved pauses is also 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. This observation further 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 co-translational folding pathway must be considered when interpreting the effect of a pause in translation. We anticipate that an optimal translation protocol could be predicted with knowledge of the substructure-specific folding and translation rates, as well as their propensities for forming non-native interactions.

The approach that we have taken in this work improves upon earlier studies of rare-codon usage, which have addressed alternative hypotheses regarding translational pausing but yielded mixed results [30, 38, 45–51]. In addition to our focus on the most physically relevant features of co-translational folding, our conclusions are more robust due to our use of a multiple-sequence analysis to detect evolutionary conservation, as well as our formulation of a neutral model that controls for both amino-acid composition and the inherent codon-usage variability across genes. The statistical significance of our results is further increased by the much larger sample size used here.

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 [52]. 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 [53–55]. mRNA-specific effects are thus a likely explanation for the observed 5'-end bias in rare-codon enrichment, where variations in translation speed cannot play a role in co-translational folding. Consequently, we have excluded N-terminal rare codons from our analysis.

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 co-translational protein folding. Further work is needed to understand the relationship between genome-wide codon usage and translation rates and to improve the prediction of co-translational folding intermediates, including those that contain significant amounts of non-native 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.

METHODS

Multiple-sequence alignments and neutral model of rare-codon usage

We constructed multiple-sequence 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 (Table S2). For each gene, we calculated the average fraction of rare codons, $\lambda \equiv n_{\rm rare}/N$, where N is the protein length, in all aligned sequences. Assuming a Poisson process with rate constant λ over a region of n consecutive codons, we computed the minimum number of rare codons, $n_{\rm enr}(\lambda)$, that would have to be observed to achieve an approximately one-standard-deviation enrichment. Then, based on the relative codon usage in all genes with similar overall rare-codon usage λ, we estimated the probability of randomly choosing a rare codon for the amino-acid type a, $p_{\rm rare}(a)$ (see Sec. S1 A). To compute the probability of generating an enriched region in a biased reverse translation, $p_{\rm enr}$, we considered all ways in which at least $n_{\rm enr}$ out of n codons could be rare given a fixed amino-acid sequence. Let $a_{s\omega}$ be the amino acid at the aligned position ω of sequence s. We then define

$$p^s_{\rm enr} = \sum_{l=n_{\rm enr}}^{n} \sum_{\Omega}^{\rm perm.} \prod_{i=1}^{l} p^s_{\rm rare}(a_{s\Omega_i}) \prod_{j=l+1}^{n} [1 - p^s_{\rm rare}(a_{s\Omega_j})], \quad (1)$$

where $\{\Omega\}$ is the set of all permutations of the n codon positions in the region of interest, and Ω_i is the position in the alignment at the ith index of the permutation Ω. The probability, within this neutral model, of generating at least $m_{\rm obs}$ enriched sequences out of M aligned sequences is then

$$p_{\rm neu}(m_{\rm obs}, M) = \sum_{m=m_{\rm obs}}^{M} \sum_{\sigma}^{\rm perm.} \prod_{i=1}^{m} p^{\sigma_i}_{\rm enr} \prod_{j=m+1}^{M} (1 - p^{\sigma_j}_{\rm enr}), \quad (2)$$

where $\{\sigma\}$ is the set of all orderings of the M sequences, and σ_i is the sequence at the ith index of the ordering σ. Eq. 2 is the p-value used for all results presented in this study. Note that although we have used a binary classification of rare and common codons in each genome, this p-value accounts for the fact that each amino-acid type has a different probability of being encoded by a rare codon. See Sec. S1 A for further details.

Coarse-grained co-translational folding model

We constructed consensus crystal structures for 516 non-membrane E. coli proteins with 500 residues or fewer by searching the Protein Databank [56] for structures containing sequences with at least 95% amino-acid identity to the E. coli gene. We identified all residue–residue contacts where at least one pair of heavy atoms is less than 4 Å apart in at least 25% of the candidate structures (see Figure S4 for an analysis of this cutoff). Following Ref. 23, we model the configurational free energy of a chain with native contacts among a set of residues, g, as

$$\frac{f(g)}{k_{\rm B}T} = \frac{1}{2}\sum_{u\in g}\sum_{v\in g}\frac{\epsilon_{uv}}{k_{\rm B}T} + \frac{\mu}{k_{\rm B}T}(N_g - 1) - \frac{\Delta S_{\rm l}(g)}{k_{\rm B}}, \quad (3)$$

where u and v are residue indices, ϵ_{uv} is the contact energy between residues u and v, $\mu = 2k_{\rm B}T$ is the entropic cost of ordering a single residue, N_g is the number of ordered residues, and $\Delta S_{\rm l}(g)$ is the total loop entropy of disordered segments. Eq. 3 assumes that the native contacts comprise a single structured region (see Sec. S1 B). The minimum free energy of an L-length nascent chain is then the sum of one or more structured regions (a, b), within which any configuration satisfying $a \le u \le b \,\forall\, u \in g$ is allowed. This free energy can be written as

$$\frac{F(L)}{k_{\rm B}T} = \min_{\boldsymbol{r}} \left\{ -\sum_{(a,b)\in\boldsymbol{r}} \ln \sum_{\{g\}_{ab}} \exp\left[-\frac{f(g)}{k_{\rm B}T}\right] \right\}, \quad (4)$$

where the first sum runs over all structured regions, the second sum includes all configurations consistent with a structured region (a, b), and we choose the lowest free energy of all sets of structured regions $\boldsymbol{r} \equiv \{(a_1, b_1), \ldots, (a_k, b_k) |\, a_1 \ge 0, a_i \ge b_{i-1} \forall i \in [2, k], b_k \le L\}$ with any number of regions $k \ge 0$. The temperature T is chosen such that the dimensionless free-energy of the full-length protein, where $L = N$, satisfies $F(N)/k_{\rm B}T = -0.075N$. In order to evaluate Eq. 4, we employ a mean-field approximation to compute the free-energy of each structured region. Complete details of the model and the mean-field calculations are provided in Sec. S1 B.

Analysis of co-translational folding pathways

Due to the uncertainty in the number of amino acids that are concealed in the ribosome exit tunnel, we consider a rare-codon-enriched region to be associated with a folding intermediate if the enriched region is between 20

and 60 codons downstream from the position at which intermediate first becomes stable (i.e., the co-translational free-energy profile decreases by more than $1k_{\mathrm{B}}T$). In all our comparisons, we ignore enriched regions within the first 80 codons of a transcript, since the conservation of rare codons near the 5' end of the mRNA transcripts is believed to originate from other factors, such as efficient translation initiation. In Figure 3, the 'predicted fraction' is the number of rare-codon-enriched regions that are preceded by a folding intermediate, n_{RI}, divided by the total number of enriched regions, n_{R}. To compute the odds ratios in Figure 4, we first determine the fraction, f_{I}, of all codons in our data set that are 20 to 60 positions downstream of a predicted intermediate. The 'presence' odds ratio is then defined as $(n_{\mathrm{RI}}/n_{\mathrm{R}})/f_{\mathrm{I}}$, and the 'absence' odds ratio is $(1 - n_{\mathrm{RI}}/n_{\mathrm{R}})/(1 - f_{\mathrm{I}})$. To generate the randomized controls shown in Figures 3 and 4, we sampled fictitious rare-codon-enriched regions from uniform distributions over each mRNA transcript, excluding the first 80 codons. The uniform distributions were normalized such that the mean number of randomized enriched regions is equal to the total number of observed enriched regions at each p-value threshold.

ACKNOWLEDGMENTS

We thank Sanchari Bhattacharyya and Michael Manhart for many insightful discussions. This work was supported by NIH grants R01GM068670 and F32GM116231.

[1] A. A. Komar, Trends Biochem. Sci. **34**, 16 (2009).
[2] S. Pechmann, F. Willmund, and J. Frydman, Mol. Cell **49**, 411 (2013).
[3] P. Ciryam, R. I. Morimoto, M. Vendruscolo, C. M. Dobson, and E. P. O'Brien, Proc. Natl. Acad. Sci. **110**, E132 (2013).
[4] W. J. Netzer and F. U. Hartl, Nature **388**, 343 (1997).
[5] J. Frydman, H. Erdjument-Bromage, P. Tempst, and F. U. Hartl, Nat. Struct. Mol. Biol. **6**, 697 (1999).
[6] A. A. Komar, T. Lesnik, and C. Reiss, FEBS Letters **462**, 387 (1999).
[7] S. J. Kim, J. S. Yoon, H. Shishido, Z. Yang, L. A. Rooney, J. M. Barral, and W. R. Skach, Science **348**, 444 (2015).
[8] E. Siller, D. C. DeZwaan, J. F. Anderson, B. C. Freeman, and J. M. Barral, J. Mol. Biol. **396**, 1310 (2010).
[9] K. G. Ugrinov and P. L. Clark, Biophys. J. **98**, 1312 (2010).
[10] D. Agashe, N. C. Martinez-Gomez, D. A. Drummond, and C. J. Marx, Mol. Biol. Evo. **30**, 549 (2013).
[11] P. L. Clark and J. King, J. Biol. Chem. **276**, 25411 (2001).
[12] M. S. Evans, I. M. Sander, and P. L. Clark, J. Mol. Biol. **383**, 683 (2008).
[13] G. Zhang, M. Hubalewska, and Z. Ignatova, Nat. Struct. Mol. Biol. **16**, 274 (2009).
[14] I. M. Sander, J. L. Chaney, and P. L. Clark, J. Am. Chem. Soc. **136**, 858 (2014).
[15] F. Buhr, S. Jha, M. Thommen, J. Mittelstaet, F. Kutz, H. Schwalbe, M. V. Rodnina, and A. A. Komar, Mol. Cell **61**, 341 (2016).
[16] M. Zhou, J. Guo, J. Cha, M. Chae, S. Chen, J. M. Barral, M. S. Sachs, and Y. Liu, Nature **495**, 111 (2013).
[17] E. P. O'Brien, P. Ciryam, M. Vendruscolo, and C. M. Dobson, Acc. Chem. Res. **47**, 1536 (2014).
[18] D. A. Nissley, A. K. Sharma, N. Ahmed, U. A. Friedrich, G. Kramer, B. Bukau, and E. P. O'Brien, Nat. Comm. **7** (2016).
[19] Y. Xu, P. Ma, P. Shah, A. Rokas, Y. Liu, and C. H. Johnson, Nature **495**, 116 (2013).
[20] C. Kimchi-Sarfaty, J. M. Oh, I.-W. Kim, Z. E. Sauna, A. M. Calcagno, S. V. Ambudkar, and M. M. Gottesman, Science **315**, 525 (2007).
[21] F. U. Hartl and M. Hayer-Hartl, Nat. Struct. Mol. Biol. **16**, 574 (2009).
[22] H. Maity, M. Maity, M. M. G. Krishna, L. Mayne, and S. W. Englander, Proc. Natl. Acad. Sci. **102**, 4741 (2005).
[23] W. M. Jacobs and E. I. Shakhnovich, Biophys. J. **111**, 925 (2016).
[24] C. Eichmann, S. Preissler, R. Riek, and E. Deuerling, Proc. Natl. Acad. Sci. **107**, 9111 (2010).
[25] H. Krobath, E. I. Shakhnovich, and P. F. Faísca, J. Chem. Phys. **138**, 06B601_1 (2013).
[26] W. Holtkamp, G. Kokic, M. Jäger, J. Mittelstaet, A. A. Komar, and M. V. Rodnina, Science **350**, 1104 (2015).
[27] E. P. O'Brien, M. Vendruscolo, and C. M. Dobson, Nat. Comm. **3**, 868 (2012).
[28] G. N. Jacobson and P. L. Clark, Curr. Opin. Struct. Biol. **38**, 102 (2016).
[29] P. M. Sharp and W.-H. Li, Nucl. Acids Res. **15**, 1281 (1987).
[30] J. L. Chaney and P. L. Clark, Ann. Rev. Biophys. **44**, 143 (2015).
[31] A. Dana and T. Tuller, Nucl. Acids Res. **42**, 9171 (2014).
[32] G.-W. Li, E. Oh, and J. S. Weissman, Nature **484**, 538 (2012).
[33] V. Pelechano, W. Wei, and L. M. Steinmetz, Cell **161**, 1400 (2015).
[34] D. E. Weinberg, P. Shah, S. W. Eichhorn, J. A. Hussmann, J. B. Plotkin, and D. P. Bartel, Cell Rep. **14**, 1787 (2016).
[35] T. F. Clarke IV and P. L. Clark, PloS One **3**, e3412 (2008).
[36] C.-H. Yu, Y. Dang, Z. Zhou, C. Wu, F. Zhao, M. S. Sachs, and Y. Liu, Mol. Cell **59**, 744 (2015).
[37] P. S. Spencer, E. Siller, J. F. Anderson, and J. M. Barral, J. Mol. Biol. **422**, 328 (2012).
[38] S. Pechmann and J. Frydman, Nat. Struct. Mol. Biol. **20**, 237 (2013).
[39] M. Widmann, M. Clairo, J. Dippon, and J. Pleiss, BMC Genomics **9**, 207 (2008).
[40] T. F. Clarke IV and P. L. Clark, BMC Genomics **11**, 118 (2010).
[41] F. Trovato and E. P. O'Brien, Ann. Rev. Biophys. **45**, 345 (2016).

[42] K. Ghosh and K. A. Dill, Proc. Natl. Acad. Sci. **106**, 10649 (2009).
[43] I. J. Purvis, A. J. Bettany, T. C. Santiago, J. R. Coggins, K. Duncan, R. Eason, and A. J. Brown, J. Mol. Biol. **193**, 413 (1987).
[44] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, J. Mol. Biol. **247**, 536 (1995).
[45] C. M. Deane and R. Saunders, Biotech. J. **6**, 641 (2011).
[46] T. A. Thanaraj and P. Argos, Protein Sci. **5**, 1594 (1996).
[47] Y. Lee, T. Zhou, G. G. Tartaglia, M. Vendruscolo, and C. O. Wilke, Proteomics **10**, 4163 (2010).
[48] R. Saunders and C. M. Deane, Nucl. Acids Res. **38**, 6719 (2010).
[49] G. Zhang and Z. Ignatova, PLoS One **4**, e5036 (2009).
[50] M. Chartier, F. Gaudreault, and R. Najmanovich, Bioinformatics **28**, 1438 (2012).
[51] S. Brunak and J. Engelbrecht, Proteins **25**, 237 (1996).
[52] J. B. Plotkin and G. Kudla, Nat. Rev. Genet. **12**, 32 (2011).
[53] G. Kudla, A. W. Murray, D. Tollervey, and J. B. Plotkin, Science **324**, 255 (2009).
[54] D. B. Goodman, G. M. Church, and S. Kosuri, Science **342**, 475 (2013).
[55] S. Bhattacharyya and E. I. Shakhnovich, Private communication (2017).
[56] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, Nucl. Acids Res. **28**, 235 (2000).
[57] M. Wang, M. Weiss, M. Simonovic, G. Haertinger, S. P. Schrimpf, M. O. Hengartner, and C. von Mering, Mol. Cell. Proteomics **11**, 492 (2012).

SUPPLEMENTARY INFORMATION

S1. EXTENDED METHODS

A. Neutral model and statistics of rare-codon enrichment

In this section, we describe a neutral model of gene-specific synonymous codon usage. For each genome, we define a set of rare codons, i.e., the set of codons that are used significantly less frequently than alternative synonymous codons in the most highly expressed genes [29]. Using experimentally determined protein abundances to account for the differing expression levels among genes, we compute the relative synonymous codon frequencies for each amino-acid type,

$$p^s_{\rm use}(c|a) = \frac{\sum_g x_{sg} \sum_{i=1}^{L_{sg}} \mathbf{1}(c_{sgi} = c)}{\sum_g x_{sg} \sum_{i=1}^{L_{sg}} \mathbf{1}(a_{sgi} = a)}, \tag{S1}$$

where c is a codon for an amino acid a, x_{sg} is the protein abundance of gene g in genome s, i is an index that runs over all coding positions up to the protein length L_{sg}, and $\mathbf{1}(\cdot)$ is the indicator function. We then define the set of rare codons as those codons whose protein-abundance-weighted relative usage, $p^s_{\rm use}(c|a)$, is less than 10%. The relative usages of E. coli rare codons, as determined by Eq. S1, are shown in Table S1. We use a composite data set for the protein abundances in E. coli [57] and assume that these abundances are similar for all prokaryotic genomes in our alignment (Table S2). The rare-codon definitions turn out to be very similar, but not identical, for these genomes.

Next, we address the absolute enrichment of rare codons in a local region of a gene, without considering the amino-acid sequence. We begin by calculating the average fraction of rare codons in the M aligned sequences, all of which have at least 50% amino-acid identity with respect to the E. coli gene and differ in length by no more than 20%,

$$\lambda_g \equiv M^{-1} \sum_{s=1}^{M} L_{sg}^{-1} \sum_{i=31}^{L_{sg}} \mathbf{1}[p^{sgi}_{\rm use}(c|a) \le 0.1], \tag{S2}$$

excluding the first 30 codons to avoid the 5'-end bias. Here and below, we write $p^{sgi}_{\rm use}(c|a) \equiv p^s_{\rm use}(c_{sgi}|a_{sgi})$ for brevity. We then define a local enrichment threshold $n_{\rm enr}$ by considering a Poisson process with rare-codon-usage rate λ_g,

$$n^{sg}_{\rm enr} \equiv \min\left(l > 0 \,|\, 1 - p_{\rm Poisson}(\lambda_g, l, n_{sg}) \le 0.15\right), \tag{S3}$$

where $p_{\rm Poisson}(\lambda, l, n)$ is the cumulative distribution function of l events occurring in n trials given a rate λ. The cumulative probability 0.15 corresponds to an approximately one-standard-deviation fluctuation in the local rare-codon usage. Thus, we identify regions of rare-codon enrichment where at least $n^{sg}_{\rm enr}$ out of n_{sg} consecutive codons are rare. In cases where insertions or deletions appear in a multiple-sequence alignment, we define the local regions using the E. coli sequence and recompute $n^{sg}_{\rm enr}$ for the other sequences, which may have a different number of codons n_{sg} in the aligned region, as necessary.

We can now calculate a p-value that accounts for biases in rare-codon usage due to the local amino-acid composition. To do so, we estimate the probability of using a rare codon for each amino-acid type in a given gene by analyzing the relative codon-usage bias of all genes with a similar overall rare-codon usage λ,

$$p^{s,\lambda}_{\rm rare}(a) \equiv \frac{\sum_{\{g\}_\lambda} \sum_{i=31}^{L_{sg}} \mathbf{1}(a_{sgi} = a) \mathbf{1}[p^{sgi}_{\rm use}(c|a) \le 0.1]}{\sum_{\{g\}_\lambda} \sum_{i=31}^{L_{sg}} \mathbf{1}(a_{sgi} = a)}, \tag{S4}$$

where $\{g\}_\lambda$ is the set of genes with similar λ. In practice, we implement Eq. S4 by sorting all genes according to their overall rare-codon usage and splitting them into ten groups, each comprising approximately 400 genes. However, because the average codon usage is not exactly the same for all genes in each of these groups, we then adjust the rare-codon probabilities slightly to match λ_g,

$$p^{sg}_{\rm rare}(a) = \gamma p^{s,\lambda_g}_{\rm rare}(a), \tag{S5}$$

where the scaling factor γ is chosen such that

$$M^{-1} \sum_{s=1}^{M} L_{sg}^{-1} \sum_{i=31}^{L_{sg}} \frac{\gamma p^{s,\lambda_g}_{\rm rare}(a_{sgi})}{1 + (\gamma - 1) p^{s,\lambda_g}_{\rm rare}(a_{sgi})} = \lambda_g. \tag{S6}$$

Eq. S6 is nearly linear in γ while ensuring that all probabilities are positive.

Finally, the neutral-model probability that a local region contains at least $n^{sg}_{\rm enr}$ rare out of n_{sg} codons is

$$p^{sg}_{\rm enr} = \sum_{l=n^{sg}_{\rm enr}}^{n_{sg}} \sum_{\Omega}^{\rm perm.} \prod_{i=1}^{l} p^{sg}_{\rm rare}(a_{s\Omega_i}) \prod_{j=l+1}^{n_{sg}} [1 - p^{sg}_{\rm rare}(a_{s\Omega_j})], \tag{S7}$$

where $\{\Omega\}$ is the set of all permutations of the codon positions within the given region and Ω_i is the position at index i in the permutation Ω. The probability of observing at least $m_{\rm obs}$ enriched sequences is then

$$p^g_{\rm neu}(m_{\rm obs}, M) = \sum_{m=m_{\rm obs}}^{M} \sum_{\sigma}^{\rm perm.} \prod_{i=1}^{m} p^{\sigma_i g}_{\rm enr} \prod_{j=m+1}^{M} (1 - p^{\sigma_j g}_{\rm enr}), \tag{S8}$$

where $\{\sigma\}$ is the set of all orderings of the M sequences, and σ_i is the sequence at the ith index of the ordering σ. We require $m_{\rm obs}/M \ge 0.75$ for an enriched region to be considered conserved (see Figure S2). Eq. S8 is the neutral-model p-value used throughout this work.

B. Quasi-equilibrium co-translational folding model

In this section, we adapt the native-centric coarse-grained model described in Ref. 23 to predict free-energy landscapes for co-translational folding (Figure SM1a). A native contact between a pair of residues is formed when at least one pair of heavy atoms is separated by at most 4 Å in the consensus crystal structure and the residues are at least three positions apart in the sequence. The effective potential energy between residues u and v depends on the number of such heavy-atom contacts, $n_{uv}^{\mathrm{nc}} \geq 1$; whether the pair of residues forms a hydrogen bond, $\mathbf{1}_{uv}^{\mathrm{hb}}$; and whether the contact is part of an α-helix, $\mathbf{1}_{uv}^{\mathrm{helix}}$:

$$\epsilon_{uv} = -\left(\alpha_{\mathrm{helix}}\right)^{\mathbf{1}_{uv}^{\mathrm{helix}}} \left[n_{uv}^{\mathrm{nc}} + \alpha_{\mathrm{hb}} \mathbf{1}_{uv}^{\mathrm{hb}}\right], \tag{S9}$$

with $\alpha_{\mathrm{helix}} = 5/8$ and $\alpha_{\mathrm{hb}} = 16$.

In this discrete model, a microscopic configuration is defined by the set of native contacts that are formed. The free energy of such a configuration is determined by opposing energetic and entropic contributions: native contacts are stabilizing according to Eq. S9, and configurational entropy is gained when native contacts are broken. Furthermore, sets of native contacts can be decomposed into independent structured regions (Figure SM1b). Within each structured region, all possible native contacts are formed among the participating residues; however, different structured regions are linked by chains of disordered residues and can thus move independently. We therefore evaluate the free energy of each structured region separately. The free energy of a configuration g with a single structured region is

$$\frac{f(g)}{k_{\mathrm{B}}T} = \frac{1}{2} \sum_{u,v \in g} \frac{\epsilon_{uv}}{k_{\mathrm{B}}T} + \frac{\mu}{k_{\mathrm{B}}T}(N_g - 1) - \frac{\Delta S_{\mathrm{l}}(g)}{k_{\mathrm{B}}}, \tag{S10}$$

where u and v are residues that form native contacts in the configuration g, N_g is the number of contact-forming residues, k_{B} is the Boltzmann constant, T is the absolute temperature, and $\mu = 2k_{\mathrm{B}}T$ is the entropic penalty associated with native-contact formation [23]. Sequences of disordered residues that begin and end in the same structured region are referred to as loops (Figure SM1b–c). The total loop entropy, $\Delta S_{\mathrm{l}}(g)$, is the sum of $\Delta S_{\mathrm{l}}(l)$ for all loops $\{l\}_g$ that terminate in the structured region,

$$\frac{\Delta S_{\mathrm{l}}(l)}{k_{\mathrm{B}}} \equiv \begin{cases} -\frac{|l|\mu}{k_{\mathrm{B}}T} & \text{if } |l| \leq b, \\ -\frac{b\mu}{k_{\mathrm{B}}T} - \frac{3}{2}\left[\ln \frac{|l|}{b} + \frac{r(l)^2}{b^2|l|}\right] & \text{if } |l| > b, \end{cases} \tag{S11}$$

where $|l|$ is the number of disordered residues in the loop, $r(l)$ is the distance between the loop endpoints, and we assume Gaussian statistics for all loops longer than the Kuhn length $b = 2$ residues.

We now introduce a mean-field approximation to calculate the free energy of an ensemble of microscopic configurations. We assume that all configurations in the ensemble contain a single structured region with an essentially equivalent set of loops (i.e., the loops grow or shrink by adding or removing adjacent residues, but otherwise the structured region remains unchanged; see Figure SM1c). Instead of considering the set of discrete configurations $\{g\}$, we now characterize the ensemble by the average ordering of each residue in this set of microscopic configurations, $\{\rho_u\}$. Ignoring correlations in the residues beyond nearest-neighbor native contacts, we can write the mean-field free energy of this ensemble of microscopic configurations as the sum of the average energetic and

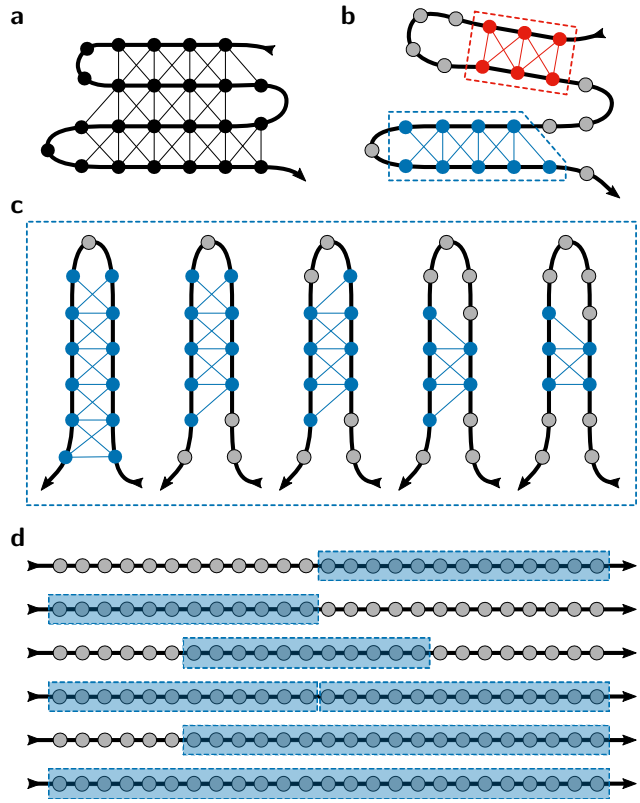

FIG. SM1. Schematic of the coarse-grained model (adapted from Ref. 23). (a) Native contacts (thin black lines) are drawn between residues (circles) that are connected by the polymer backbone (thick black line). (b) A partially folded configuration may contain one or more structured regions. In this example configuration, native contacts are formed within two such regions (red and blue sets of residues), while the disordered gray residues do not form any native contacts; each structured region is associated with a single loop of either four disordered residues (red region) or one disordered residue (blue region). (c) An example ensemble of similar configurations. The configurations differ by the number of residues in the structured region as well as the length, but not the position, of the single disordered loop. (d) Example sets of non-overlapping structured regions (dashed blue boxes). For each structured region, we minimize the mean-field free energy (Eq. S13); the total free energy is the sum of the minimized mean-field free energies of all structured regions.

entropic contributions,

$$\frac{F_{\rm mf}(\{\rho_u\})}{k_{\rm B}T} = \frac{U_{\rm mf}}{k_{\rm B}T} - \frac{S_{\rm mf}}{k_{\rm B}} \tag{S12}$$

$$= -\frac{1}{2}\sum_{u,v}\frac{\rho_u\tilde{\epsilon}_{uv}\rho_v}{k_{\rm B}T} - \sum_u \ln q_u \tag{S13}$$

$$-\frac{\langle \Delta S_{\rm l}\rangle_{\{\rho_u\}}}{k_{\rm B}},$$

where q_u is the single-residue mean-field partition function

$$q_u = 1 + \exp\left(-\tilde{\mu} - \sum_{v\in g}\tilde{\epsilon}_{uv}\rho_v\right). \tag{S14}$$

Native contacts are correlated on a length scale comparable to the Kuhn length due to the finite size of the residues and the chain connectivity [23]. We have excluded native contacts between residues that are fewer than three positions apart in the sequence when constructing the energy matrix $\{\epsilon_{uv}\}$, because such contacts are likely to be present in both the denatured and folded ensembles. Nevertheless, we need to account for local geometric correlations in native contact formation. To impose correlations between residues that are very close in the sequence, we add nearest and next-nearest-neighbor couplings to $\{\epsilon_{uv}\}$ and $\{\mu_u\}$,

$$\tilde{\epsilon}_{uv} = \epsilon_{uv} - \mathbf{1}_r(v{+}1)[\delta_{u,v+1}\lambda_1 - \mathbf{1}_r(v{+}2)\delta_{u,v+2}\lambda_2], \tag{S15}$$

$$\tilde{\mu}_u = \mu + \mathbf{1}_r(v{+}1)\left[\lambda_1 + \mathbf{1}_r(v{+}2)\lambda_2\right], \tag{S16}$$

where δ is the Kronecker delta, $\mathbf{1}_r(v)$ indicates whether the residue v can form native contacts in the structured region, and we assume that $u < v$ and $\mathbf{1}_r(u) = 1$ for notational simplicity. We have chosen coupling constants $\lambda_1 = 0.8$ and $\lambda_2 = 0.4$ to obtain comparable results to the Monte Carlo simulations reported in Ref. 23. Choosing $\lambda_1, \lambda_2 > 0$ tends to increase the average ordering at the free-energy minimum for all residues that form multiple contacts in the structured region.

To determine $F_{\rm mf}$ and $\{\rho_u\}$, we first minimize the mean-field free energy without considering the loop term, $\Delta S_{\rm l}$. Eq. S13 can be minimized numerically using a standard non-linear, multi-dimensional minimization algorithm. The probability of forming a native contact between residues u and v in the mean-field model is then $p_{uv}^{\rm contact} = \rho_u \mathbf{1}(\epsilon_{uv} < 0)\rho_v$; example contact maps showing $\{p_{uv}^{\rm contact}\}$ at the free-energy minimum can be found in Figure 2. While it is not guaranteed that Eq. S13 will have a single minimum, in practice we find that, excluding the trivial solution, this is almost always the case. Such behavior is expected from our previous study [23] of the model presented in Eq. S10.

Next, we compute the ensemble-averaged loop entropy, $\langle \Delta S_{\rm l}\rangle_{\{\rho_u\}}$, perturbatively by assuming that this contribution to Eq. S13 does not alter $\{\rho_u\}$ at the free-energy minimum. To allow disordered loops to grow or shrink by adding or removing adjacent residues, we calculate the

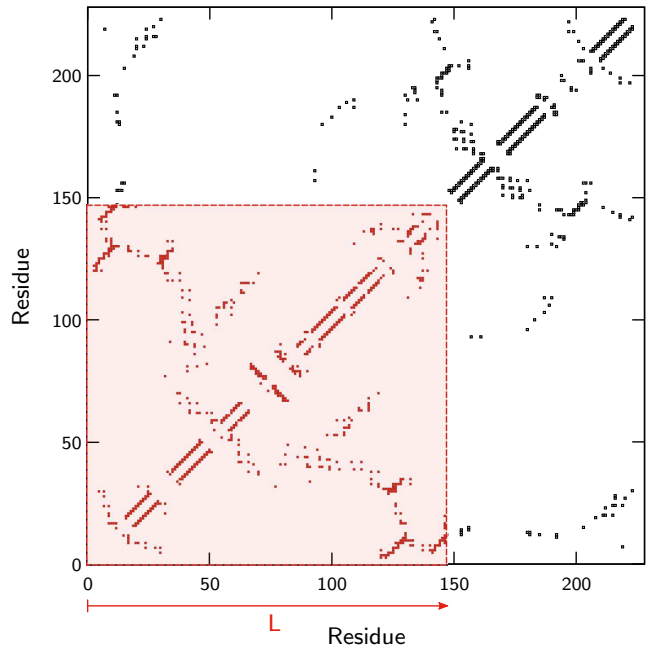

FIG. SM2. An example consensus contact map for the E. coli gene *cmk*. The highlighted region shows the native contacts that can be formed by a nascent chain of length L.

Boltzmann-weighted average loop entropy for all possible pairs of terminal residues $\{(u,v)\}_l$ for each loop l,

$$\langle \Delta S_{\rm l}(l)\rangle_{\{\rho_u\}} = \frac{\sum_{\{(u,v)\}_l}\rho_u p_{\rm l}([u,v])\rho_v e^{\frac{\Delta S_{\rm l}([u,v])}{k_{\rm B}}} S_{\rm l}([u,v])}{\sum_{\{(u,v)\}_l}\rho_u p_{\rm l}([u,v])\rho_v e^{\frac{\Delta S_{\rm l}([u,v])}{k_{\rm B}}}}, \tag{S17}$$

where $p_{\rm l}([u,v]) \equiv \prod_{u>t>v}(1-\rho_t)$, and $\Delta S_{\rm l}([u,v])$ implies Eq. S11 computed with the loop endpoints (u,v).

Finally, to compute the free energy of a nascent chain of length L, we consider native contacts among the first L residues (Figure SM2). At each chain length, $F(L)$ reports the minimum free energy relative to a random coil of length L. Any combination of non-overlapping structured regions using the first L residues is permitted (Figure SM1d). The set of allowed sets of structured regions can be written as $\boldsymbol{r} \equiv \{(a_1,b_1),\ldots,(a_k,b_k)|\, a_1 \geq 0, a_i \geq b_{i-1}\forall i \in [2,k], b_k \leq L\}$, where the pair (a_i,b_i) indicates the first and last residues that can participate in the structured region with index i, and the number of structured regions is $k \geq 0$. We minimize the mean-field free energy by finding the lowest-free-energy set of structured regions,

$$\frac{F(L)}{k_{\rm B}T} = \min_{\boldsymbol{r}}\left\{\sum_{(a,b)\in\boldsymbol{r}}\frac{F_{\rm mf}(\{\rho_u\}_{ab})}{k_{\rm B}T}\right\}, \tag{S18}$$

where $\{\rho_u\}_{ab}$ implies that $\rho_u = 0$ for all $u < a$ and $u > b$. The free energy defined in Eq. S18 is used for identifying co-translational folding intermediates as described in the main text.

S2. SUPPLEMENTARY TABLES AND FIGURES

Amino acid	Codon	Rel. usage	Max. rel. usage
G	GGA	0.0387	0.4474 (GGC)
	GGG	0.0811	
I	ATA	0.0161	0.6062 (ATC)
L	CTA	0.0269	0.6785 (CTG)
	CTC	0.0880	
	CTT	0.0853	
	TTA	0.0495	
	TTG	0.0718	
P	CCC	0.0842	0.6344 (CCG)
R	AGA	0.0097	0.5283 (CGT)
	AGG	0.0043	
	CGA	0.0170	
	CGG	0.0260	
S	AGT	0.0539	0.3198 (TCT)
	TCA	0.0547	
	TCG	0.0657	
T	ACA	0.0840	0.5224 (ACC)

TABLE S1. Rare codons in the E. coli genome. A codon is considered to be rare if the abundance-weighted relative usage, as defined in Eq. S1, is less than 10%. For comparison, the highest relative-usage codon for each amino acid is shown in the rightmost column.

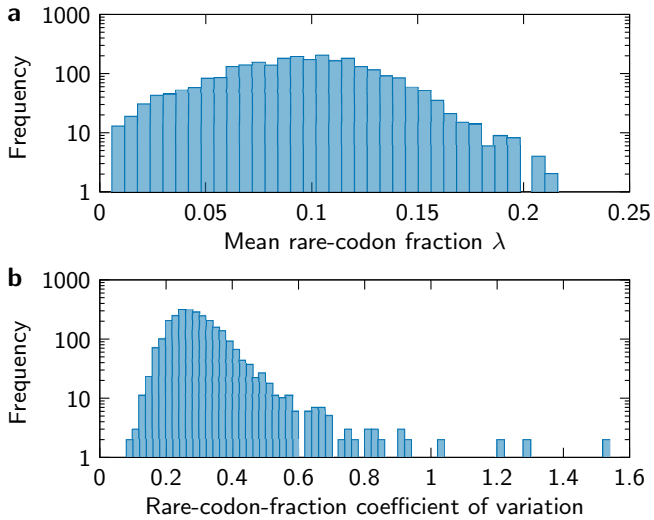

FIG. S1. The overall rare-codon usage in the E. coli genome. (a) E. coli genes exhibit a relatively broad range of average rare-codon usages, λ. (b) However, for each gene, the rare-codon usage is remarkably similar across prokaryotic genomes. The similarity among genomes is quantified here by the coefficient of variation (CV), defined as the ratio of the standard deviation to the mean rare-codon fraction across genomes. The high-CV outliers are overwhelmingly associated with the lowest λ genes.

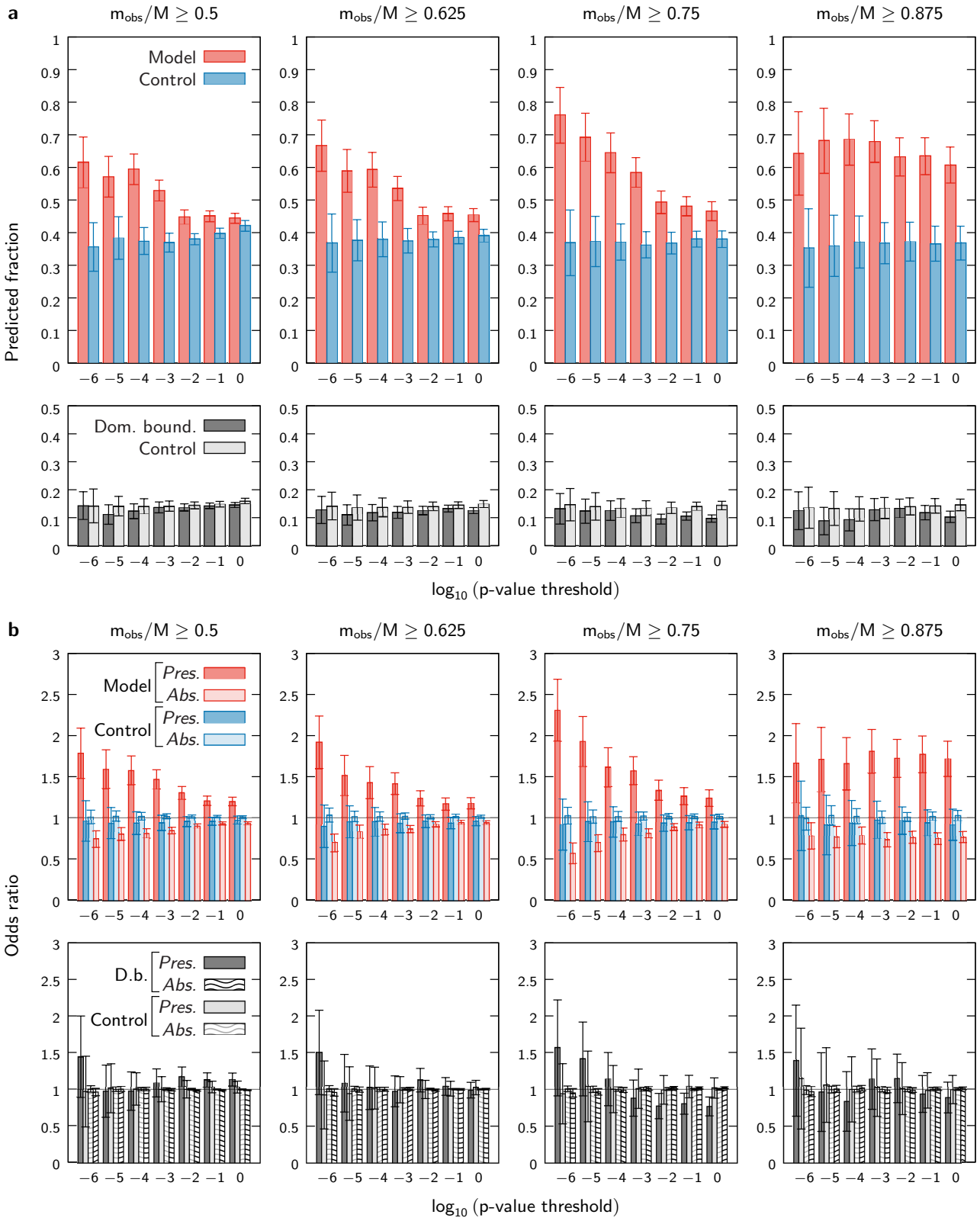

FIG. S2. The accuracy of the co-translational folding model predictions does not depend strongly on the conservation cutoff used in determining putative pause sites. In each column, we require that the fraction of the aligned sequences exhibiting a rare-codon enriched region, $m_{\rm obs}/M$, equal or exceed a prescribed cutoff (see Sec. S1 A); throughout this work, we have used a cutoff of 75%. Panels in (a) are analogous to Figure 3, while panels in (b) are analogous to Figure 4. For the 87.5% cutoff, almost all conserved regions are associated with very low p-values, and thus the folding model outperforms the control at all p-value thresholds; however, the scarcity of enriched regions at this strict conservation cutoff tends to reduce the statistical significance.

Genome ID	Codon identity E. coli	Codon identity Average	Amino-acid identity E. coli	Amino-acid identity Average	Species
NC_000913.3	1.000	0.390	1.000	0.682	Escherichia coli K-12 MG1655
NC_007606.1	0.775	0.376	0.830	0.654	Shigella dysenteriae
NZ_CP007557.1	0.473	0.382	0.803	0.687	Citrobacter freundii
NC_003198.1	0.448	0.397	0.767	0.681	Salmonella enterica
NC_014121.1	0.420	0.380	0.739	0.675	Enterobacter cloacae
NC_015663.1	0.407	0.384	0.726	0.675	Enterobacter aerogenes
NZ_CP007215.1	0.406	0.375	0.722	0.673	Enterobacter sacchari
NC_016845.1	0.405	0.384	0.724	0.671	Klebsiella pneumoniae
NC_009778.1	0.384	0.372	0.694	0.659	Cronobacter sakazakii
NZ_CP009450.1	0.382	0.369	0.702	0.658	Pluralibacter gergoviae
NZ_CP009451.1	0.373	0.359	0.692	0.655	Cedecea neteri
NC_017910.1	0.349	0.342	0.658	0.631	Shimwellia blattae
NZ_CP009454.1	0.306	0.318	0.591	0.598	Pantoea rwandensis
NC_016818.1	0.304	0.310	0.590	0.592	Rahnella aquatilis
NC_014500.1	0.301	0.313	0.582	0.586	Dickeya dadantii
NC_008800.1	0.284	0.292	0.594	0.604	Yersinia enterocolitica
NC_003143.1	0.273	0.283	0.578	0.592	Yersinia pestis
NC_022546.1	0.268	0.289	0.502	0.520	Plautia stali

TABLE S2. The prokaryotic genomes used in the multiple-sequence alignments. The middle columns show the average codon and amino-acid identities of all genes relative to both the E. coli genome and the other species in the alignment. When computing codon and amino-acid identities, we have excluded genes whose lengths differ from the E. coli homolog by more than 20%, and we have ignored all insertions and deletions.

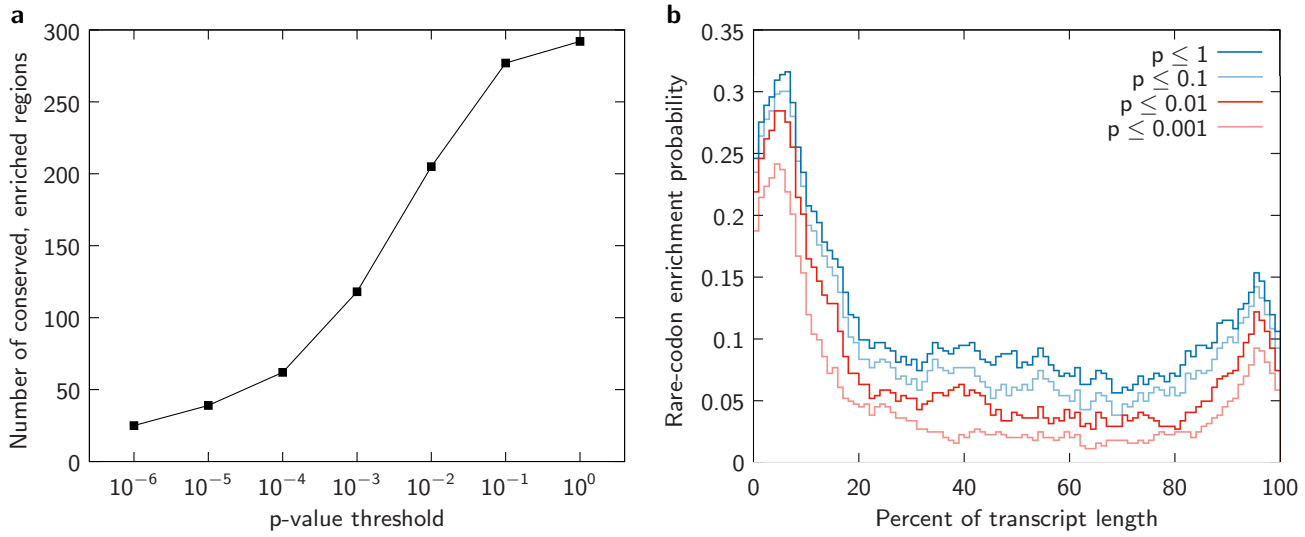

FIG. S3. The distribution of conserved, rare-codon enriched regions in the 516-protein data set used in this study. As discussed in the main text, we include only those regions that are conserved across at least 75% of the aligned sequences. (a) Excluding enriched regions within 80 codons of the N-terminus, the number of conserved, enriched regions at each p-value threshold (Eq. S8). (b) The distribution of enriched regions within mRNA transcripts at various p-value thresholds. The locations within the transcripts are normalized by the lengths of the transcripts. The apparent biases in the locations of the enriched regions towards both the 5' and 3' ends of the transcripts become more significant as the p-value threshold is lowered. In contrast, random reverse translations of the E. coli proteome using the gene-specific rare-codon probabilities (Eqs. S4–S6) yield uniform distributions at all p-value thresholds. The bias towards rare-codon enrichment at the 3' end is therefore not a consequence of the amino-acid composition near the C-terminus.

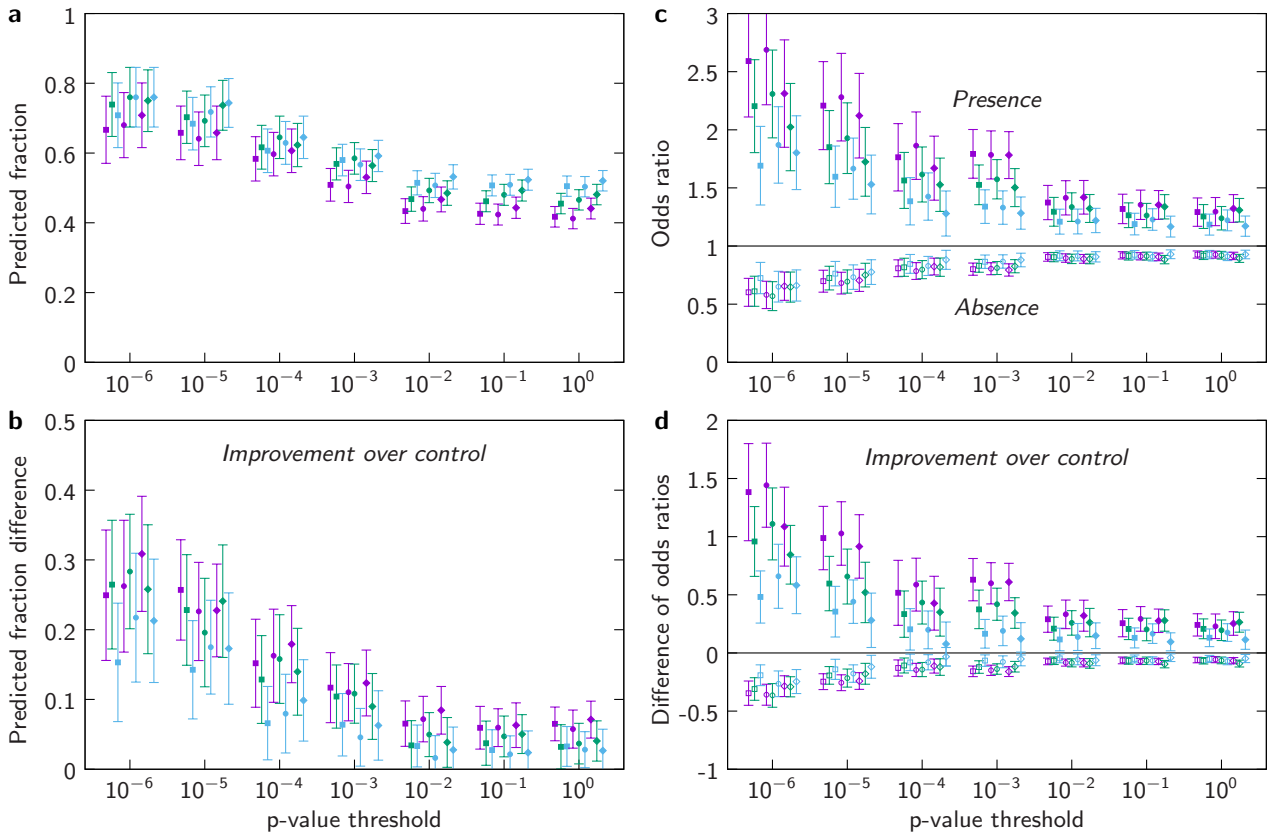

FIG. S4. The co-translational folding predictions are relatively insensitive to the model parameters. Here we vary both the consensus crystal structure definition (squares: residue–residue contacts appear in at least one crystal structure; circles: residue–residue contacts appear in at least 25% of the crystal structures; diamonds: residue–residue contacts appear in at least 50% of the crystal structures) and the native-state stabilities (purple: $F(N)/N = -0.05k_{\mathrm{B}}T$; green: $F(N)/N = -0.075k_{\mathrm{B}}T$; blue: $F(N)/N = -0.1k_{\mathrm{B}}T$, where N is the protein length). Panels (a) and (c) are analogous to Figures 3 and 4, respectively, while panels (b) and (d) show the differences between the model results and the mean of the randomized control distribution for each set of model parameters.

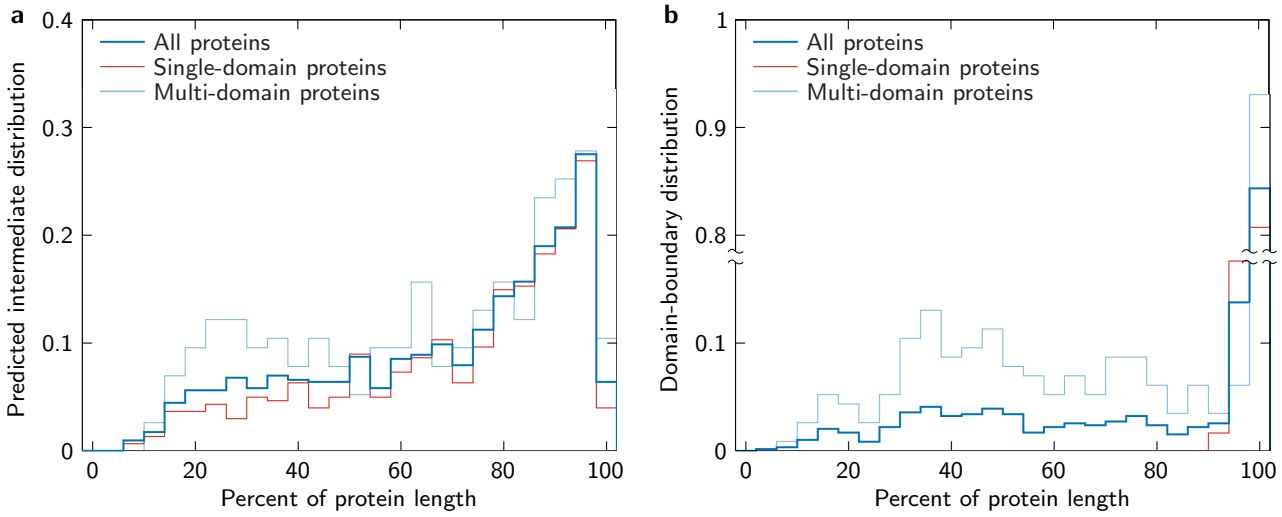

FIG. S5. The distributions of (a) predicted co-translational folding intermediates and (b) domain boundaries for the data set used in this study. The locations within the protein sequences are normalized by the protein length. Note that some domain boundaries in single-domain proteins do not coincide with the full protein length due to small discrepancies between the crystal-structure sequences and the complete sequences extracted from the E. coli genome.

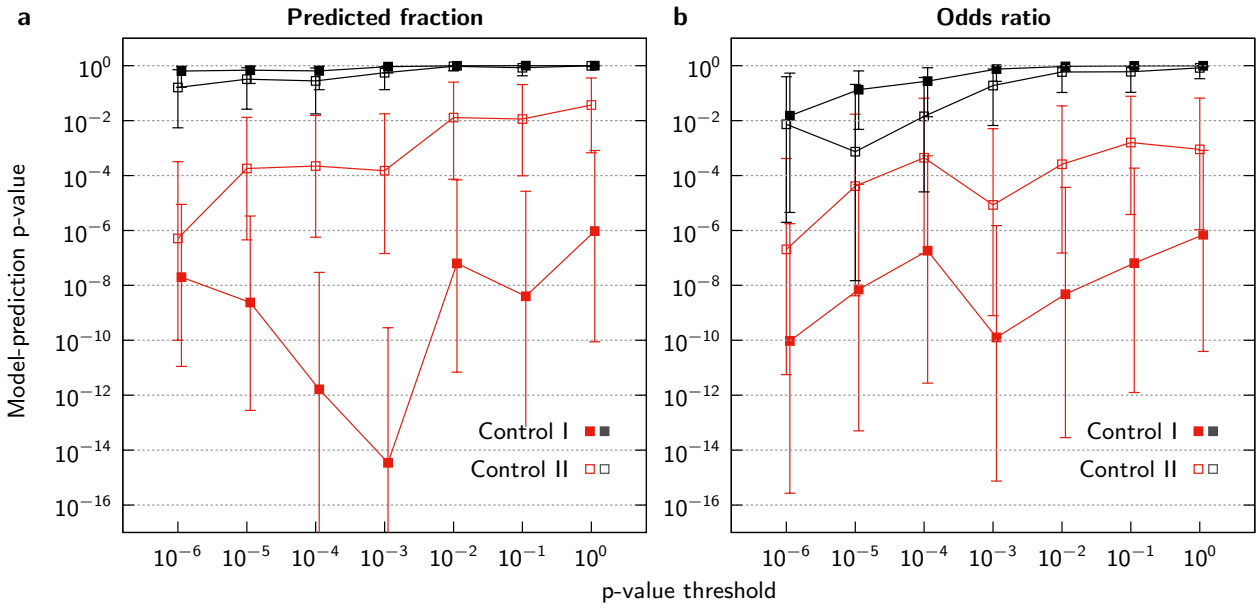

FIG. S6. The probability of generating the observed correspondence between co-translational folding intermediates and putative pause sites from the randomized control data sets. The 'model-prediction p-values' are calculated from the control distributions (Figures 3, 4 and S7) that were estimated from 100 independent randomizations, and should not be confused with the neutral-model p-value threshold used to identify putative pause sites. Control I (uniform distribution of random pause sites) refers to the control discussed in the main text, while Control II (3'-end-biased pause sites) is defined in Figure S7. Model-prediction p-values are shown in red for the co-translational folding model and in gray for the domain-boundary hypothesis. To account for the fact that we have performed calculations using a 516-protein subset of the E. coli proteome, we computed 95% confidence intervals (shown as error bars) for the model-prediction p-values, assuming that our subset is a representative sample of the E. coli proteome. The overall statistical significance of our results is influenced by the number of putative pause sites, which decreases as the p-value threshold is lowered. Consequently, the model-prediction p-values are not monotonic functions of the neutral-model p-value threshold.

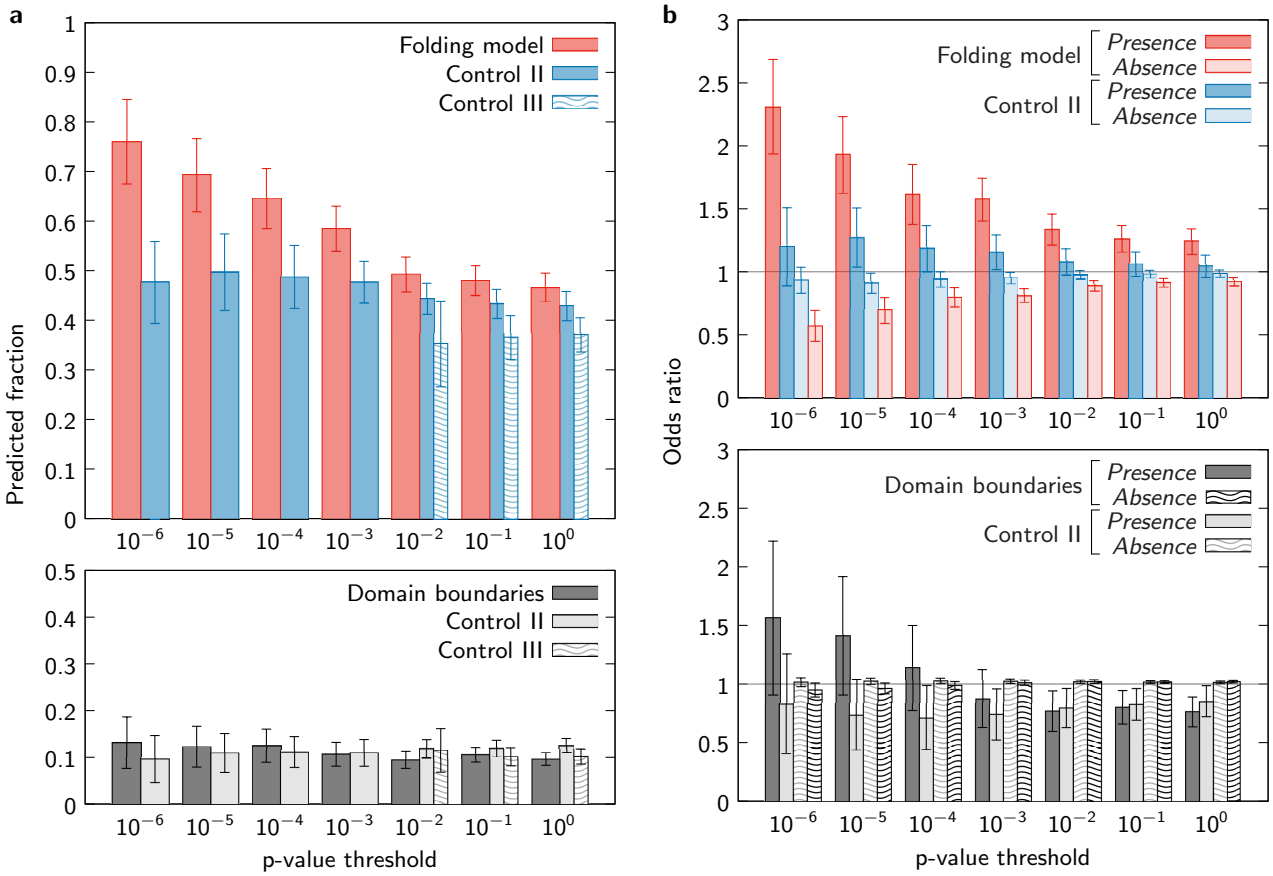

FIG. S7. The performance of the co-translational folding model and the domain-boundary hypothesis relative to two alternative controls. Panels (a) and (b) show the fraction of correctly predicted pause sites and the associated odds ratios as in Figures 3 and 4, respectively. For Control II, randomized pause sites were generated according to the observed distribution within mRNA transcripts (Figure S3b) at each p-value threshold, resulting in a bias towards the 3' end. Comparison with this control shows that the predictions of our model are not simply a result of this bias. For Control III, reverse translations of the E. coli proteome were carried out using the gene-specific rare-codon probabilities defined in Eqs. S4–S6, and then all calculations were repeated as described in Sec. S1 A. Comparison with this control shows that our results are not a consequence of a hidden amino-acid-sequence bias. Because the reverse translations were generated from the neutral model, we do not obtain a sufficient number of fictitious pause sites to compare with the results of our folding model at p-value thresholds less than 10^{-2}. As in Figures 3 and 4, the error bars on the controls report the standard deviation of the results obtained from 100 independent randomizations.

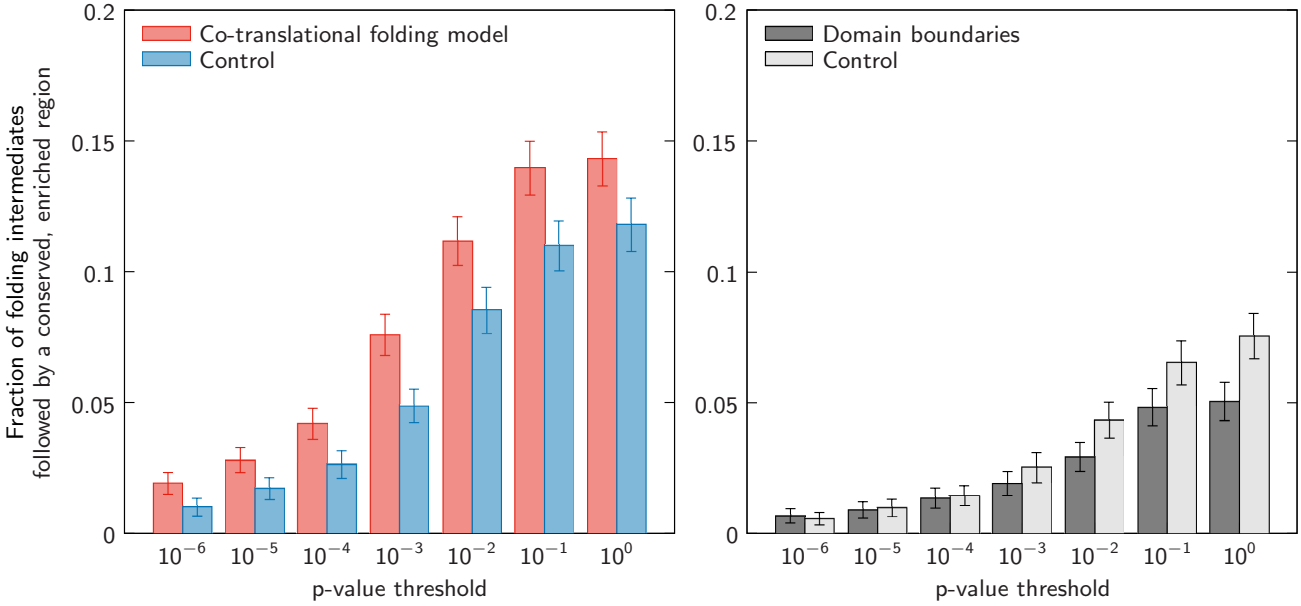

FIG. S8. The fraction of predicted intermediates that are followed by conserved, rare-codon enriched regions. On the left, the predictions of the co-translational folding model, and on the right, comparison with the domain-boundary hypothesis. The randomized control distributions and error bars are defined as in Figure 3. Note that, for both the folding model and the domain-boundary hypothesis, the results are strongly affected by the total number of putative conserved sites, which increases with the p-value threshold (Figure S3a). However, the statistical significance of the folding-model predictions is maximized at a p-value threshold of 10^{-3}, which provides an optimal balance of false positives and false negatives in the identification of putative pause sites.

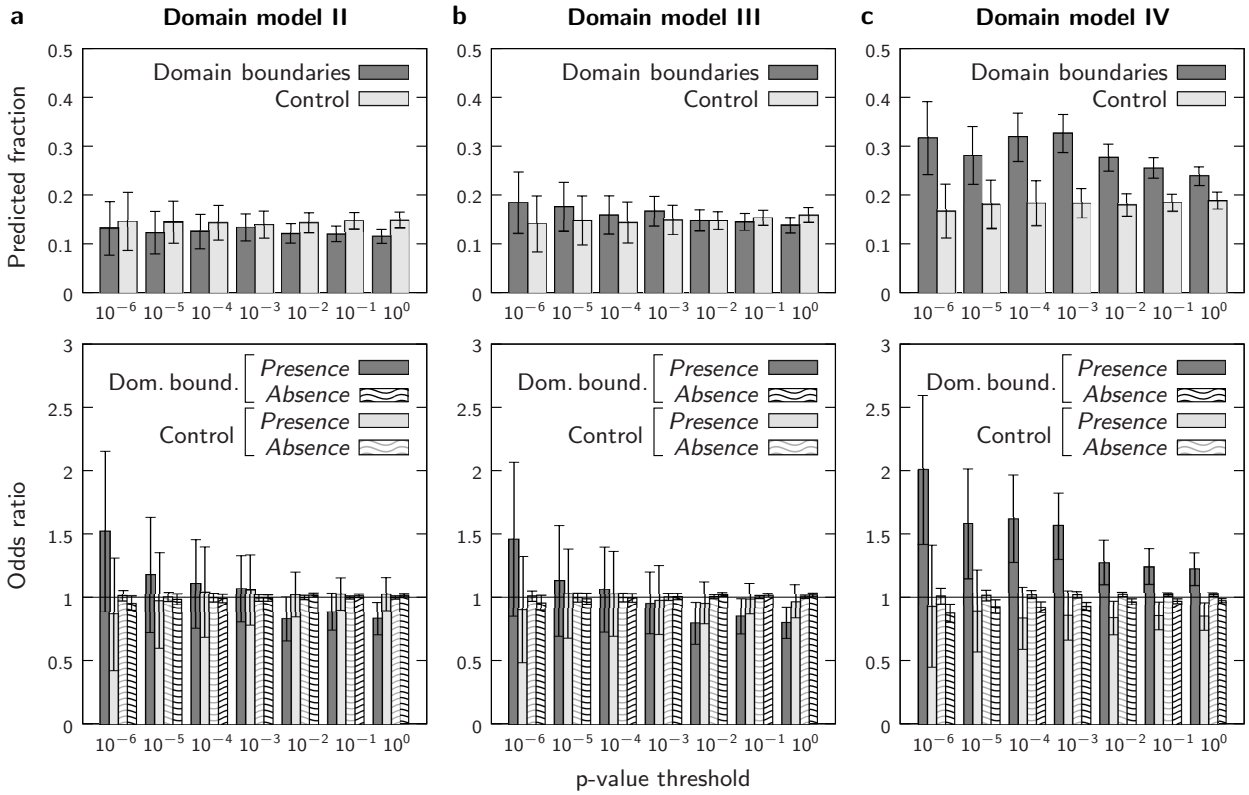

FIG. S9. Alternate domain-boundary models yield similar results. In each of these three models, we assume that the native structure forms prior to the C-terminus of the domain: model II, at 10 residues before the domain boundary; model III, at 95% of the domain length; and model IV, at 90% of the domain length. Only model IV, which assumes that native structure forms well before the domain boundary for all domains, makes statistically significant predictions; nevertheless, model IV accounts for less than 35% of the putative pause sites at all p-value thresholds. All panels are analogous to Figure 3 (top row) or Figure 4 (bottom row).