Biophysical Journal
Article

Biophysical Society

Slowest-first protein translation scheme: Structural asymmetry and co-translational folding

John M. McBride[1,*] and Tsvi Tlusty[1,2,*]
[1]Center for Soft and Living Matter, Institute for Basic Science, Ulsan, South Korea and [2]Departments of Physics and Chemistry, Ulsan National Institute of Science and Technology, Ulsan, South Korea

ABSTRACT Proteins are translated from the N to the C terminus, raising the basic question of how this innate directionality affects their evolution. To explore this question, we analyze 16,200 structures from the Protein Data Bank (PDB). We find remarkable enrichment of α helices at the C terminus and β strands at the N terminus. Furthermore, this $\alpha - \beta$ asymmetry correlates with sequence length and contact order, both determinants of folding rate, hinting at possible links to co-translational folding (CTF). Hence, we propose the "slowest-first" scheme, whereby protein sequences evolved structural asymmetry to accelerate CTF: the slowest of the cooperatively folding segments are positioned near the N terminus so they have more time to fold during translation. A phenomenological model predicts that CTF can be accelerated by asymmetry in folding rate, up to double the rate, when folding time is commensurate with translation time; analysis of the PDB predicts that structural asymmetry is indeed maximal in this regime. This correspondence is greater in prokaryotes, which generally require faster protein production. Altogether, this indicates that accelerating CTF is a substantial evolutionary force whose interplay with stability and functionality is encoded in secondary structure asymmetry.

SIGNIFICANCE Proteins are inherently asymmetric, as they are translated sequentially from the N to the C terminus. To see how this affects protein evolution, we analyze protein structures, finding corresponding asymmetries in secondary structure: α helices are enriched at the C terminus and β sheets at the N terminus. To explain this significant asymmetry, we propose the "slowest-first" scheme: slow-folding structures are located at the N terminus so they are translated first and thus have more time to fold during translation. The asymmetry peaks when folding time and translation time are similar, where our model predicts that proteins can benefit most from α-β asymmetry. Altogether, this work provides evidence for the evolution of structural asymmetry in proteins to accelerate co-translational folding, and proposes further experimental tests.

INTRODUCTION

All proteins are translated sequentially from the N to the C terminus, and are thus inherently asymmetric (1). One example of such N-to-C asymmetry is signal peptides, which enable translocation across membranes, and are located at the N terminus (2). This raises the general question of whether and how asymmetry in protein production is leveraged to gain evolutionary advantage. Here we examine structural data from the Protein Data Bank (PDB) in search of traces of such adaptation. We analyzed the distribution of secondary structure along the sequence for 16,200 PDB proteins, finding two striking patterns of asymmetry. First, disordered residues are principally located at the ends of sequences, and depleted toward the middle. Second, β strands are enriched by 55% near the N terminus, while α helices are enriched by 22% at the C terminus. These findings agree qualitatively with previous reports (3–10). This $\alpha - \beta$ asymmetry peaks at intermediate values of sequence length and contact order (CO)—which both correlate negatively with folding rate—indicating a possible link between secondary structure asymmetry and folding.

Hence, we further explore the possibility that $\alpha - \beta$ asymmetry may accelerate protein production, and is therefore a signature of evolutionary adaptation. Production of functional proteins from mRNA comprises two concerted processes: directional translation and cooperative folding. The rate of translation is limited by trade-offs between speed, accuracy, and dissipation (11–14). Folding quickly has certain advantages: unfolded proteins lead to aggregation, putting a significant burden on the cell (15–17); faster folding allows quicker responses to environmental changes

Submitted July 12, 2021, and accepted for publication November 17, 2021.
*Correspondence: jmmcbride@protonmail.com or tsvitlusty@gmail.com
Editor: Jianhan Chen.
https://doi.org/10.1016/j.bpj.2021.11.024

(18,19). Moreover, organisms whose fitness depends on fast self-reproduction would benefit from accelerated protein production that can shorten division time (20,21). Proteins may begin folding during translation, although the extent to which this varies across the proteome is unclear (22–33), and depends on the collective dynamics of the folding process. Thus, in principle, faster production times may be achieved if proteins finish folding and translation at around the same time. This co-translational folding (CTF) enables adaptations that increase yield and kinetics of protein production (29–31,33,34). For example, nascent peptides interact with ribosomes and chaperones to reduce aggregation and misfolding (35–39), while translation rates can be tuned to facilitate correct folding (25,40–44). Specifically, we ask if structural asymmetry may have evolved for fast and efficient production via CTF.

We show that the structural asymmetry observed in proteins is consistent with a scheme for accelerating CTF based on the directional nature of translation and the heterogeneity of folding rates along the sequence (45–50); e.g., cooperatively folding protein segments containing β sheets may fold slower compared with those containing α helices (51). In the proposed slowest-first scheme, protein sequences take advantage of this heterogeneity by evolving structural asymmetry: the slowest-folding segments are enriched at the N terminus (5–10,52), so that they are translated first and have more time to fold. This scheme applies to proteins composed of several independently folding segments. Therefore, due to the cooperative nature of the many-body folding process, the scheme is more likely to apply to larger proteins.

A simple model predicts that, under the slowest-first scheme, production rate can be almost doubled when folding time is equivalent to translation time. To examine this hypothesis, we estimate the ratio of folding to translation time of the PDB proteins and compare it with their $\alpha-\beta$ asymmetry, finding that asymmetry peaks when folding time is commensurate with translation time. In this region, proteins are twice as likely to exhibit $\alpha-\beta$ asymmetry that favors the slowest-first scheme. We see more evidence for this scheme in prokaryotic proteins, which is consistent with prokaryotes' greater need for fast protein production due to more frequent cell division. Taken together, these findings suggest that protein sequences have been adapted for accelerated CTF via structural asymmetry.

METHODS

Data

We extracted a set of 16,200 proteins from the PDB (53). We include proteins where the SEQRES records exactly match the corresponding UniProt sequence (not mutated, spliced, or truncated) (54). We include proteins that have been fused or have purification tags, but for these we again only include the part that exactly matches the UniProt sequence. For each unique protein sequence, we only include the most recent structure. We used SIFTS to map PDB and UniProt entries (55). We exclude proteins with predicted signal peptides as little is known about whether such proteins undergo CTF; we used Signal-P5.0 to identify signal peptides (56). Using the above inclusion criteria (and a homology cutoff) we extracted a set of 38,274 domains by matching PDB entries to Pfam domains (57). We assume that domains identified in Pfam via homology are independent folding units. α Helix and β strands are identified through annotations in the PDB; disorder is inferred from residues with missing coordinates. To calculate contact order (CO), we only consider contacts between residues where α carbons are within 10 Å; we confirm that the correlation in Fig. 1 *D* is robust to choice of this cutoff (Fig. S1).

We use the protein folding kinetics database (PFDB) for estimating folding rates (58). For our main results, we only used entries with realistic physical conditions (5<pH<8, and 20°C<T<40°C) and ignored folding rates that had been extrapolated to $T = 25$°C; in total, 122 proteins. We test a second version of the PFDB data set without excluding proteins, and using folding rates that were extrapolated to $T = 25$°C; 141 proteins. We also use the ACPro data set (59) to test the robustness of our conclusions: 125 proteins.

Predicting folding and translation rate

The folding rate, k_{fold} (in units of 1/s), is estimated by a power-law fit as a function of the protein's length:

$$\log_{10} k_{\text{fold}} = A + B\log_{10} L\,, \tag{1}$$

where L is sequence length in residues, and A and B are free parameters. We fit these parameters using data from the PFDB (58) to get 95% confidence intervals (CIs) of $A = 13.8 \pm 2.1$ and $B = -6.1 \pm 1.2$ (with Pearson's correlation coefficient $r = -0.68$, and p<0.005; Fig. S2). We consider that proteins that are observed to fold via multi-state kinetics may be badly approximated by a single folding rate; we thus repeat the analyses with the 89 two-state proteins in the PFDB, finding the area of maximum asymmetry within -1.1<$\log_{10}R$<1.5. The estimate from Eq. 1 is limited for the following reasons: (1) it is extracted from a small set of 122 proteins. (2) it disregards the effects of secondary structure, CO, and other important determinants. (3) The data are from in vitro measurements of post-translational folding. (iv) The data are biased toward small, single-domain proteins. Thus, it is only a rough predictor for the folding rates of individual proteins in the set, as the standard deviation between estimated and empirical folding rates is 1.22. For all these reasons, we use Eq. 1 as an estimator of the average folding rate of sets of proteins of similar length L where the large sampling size of each bin is expected to reduce the errors as $\sim N^{-1/2}$.

We tested whether the predicted folding rates of proteins in the PDB are within certain approximate bounds on realistic folding rates. A lower bound to folding time has been estimated at $\sim L/100$ μs (60), while we take the doubling time of *Escherichia coli*, roughly 20 min, as an approximate upper bound. Of course, many proteins rely on chaperones, so their bare estimated folding time may be longer than the upper bound, while others come from organisms with much longer doubling times. Even so, according to Eq. 1 only 8% of proteins are estimated to have a folding time greater than 20 min, while only 7% of proteins are estimated to fold faster than the lower bound. Given the magnitude of the error in estimating the folding time of individual proteins, Eq. 1 appears to yield estimates that are mostly within the biologically reasonable regime. Furthermore, in estimating the folding rate of large proteins, a common assumption is that they consist of multiple independently folding domains (61)—which considerably reduces the estimated folding time of the slowest proteins—but we neglect to make this assumption.

In principle, we could have used structural/topological measures (such as CO and long-range order (62)) to slightly improve the fit to Eq. 1. However these typically involve numerous methodological choices and additional parameters (63), and the scaling relations are entirely empirical. In contrast, scaling of folding time with length has a robust theoretical background

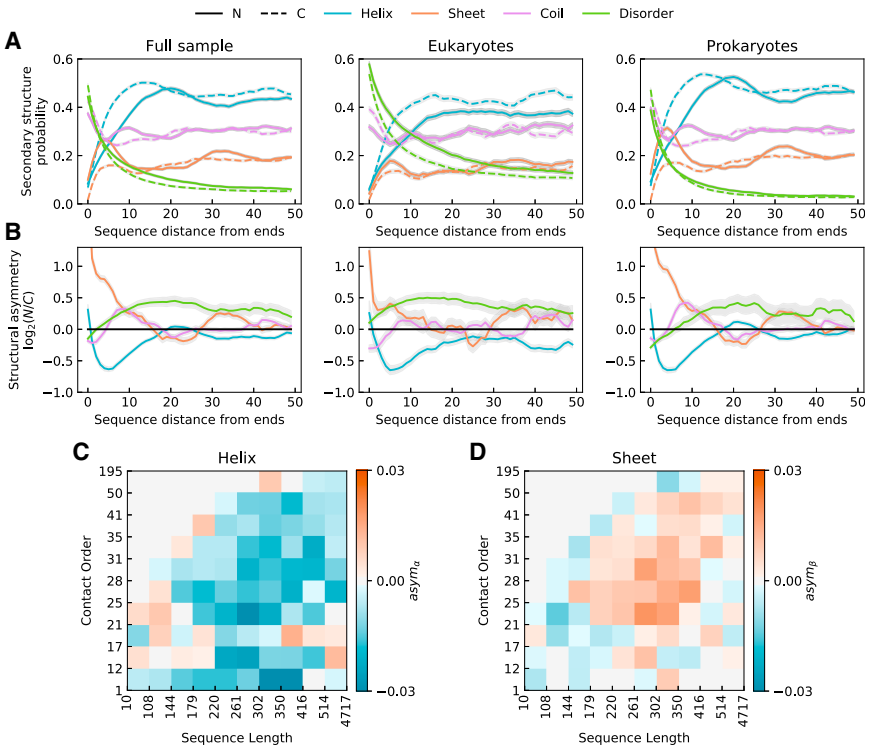

FIGURE 1 (*A*) Distribution of secondary structure along the sequence as a function of distance from the N and C terminus, and (*B*) the structural asymmetry—the ratio of the N and C distributions (in $\log_2$ scale; ± 1 are 2:1 and 1:2 N/C ratios) —for all 16,200 proteins (*left*), 4702 eukaryotic proteins (*middle*), and 10,966 prokaryotic proteins (*right*). Shading indicates bootstrapped 95% CIs in both (*A*) and (*B*). (*C* and *D*) Mean asym_α (*C*) and asym_β (*D*) as a function of sequence length and CO (Eq. 2). The data are split into deciles and the bin edges are indicated on the axes. To reduce noise in the figure due to undersampling, we only color bins where there are 20 or more proteins (full data are shown in Fig. S3).

(64–71); the exact form of the scaling is debated, but a power law is favored slightly (65,72). We do, however, use CO as a measure of topology; CO is the average sequence distance between intra-protein contacts,

$$\text{CO} = \langle |j - i| \rangle, \tag{2}$$

where i and j are pairs of residue indices for each contact (73).

We assume the translation rate, k_{trans}, depends on the organism (host organism for viral proteins), such that k_{trans} is five amino acids per second for eukaryotes and 10 for prokaryotes.

RESULTS

Protein secondary structure is asymmetric

Given the vectorial nature of protein translation, one may expect corresponding asymmetries in protein structure. To probe this, we study a non-redundant set of 16,200 proteins from the PDB (53). We find that these PDB proteins exhibit significant asymmetry in secondary structure (Fig. 1, *A* and *B*), even when counting each SCOP (74) family once (Fig. S4). For example, the first 20 residues at the N terminus are on average 55% more likely to form strands, and the last 20 residues at the C terminus are 22% more likely to form helices (in Fig. 1 *B* this is reported in $\log_2$ scale, but we report percentage here). This asymmetry is stronger for prokaryotic proteins (72%; 20%) than for eukaryotic proteins (20%; 28%). The substantial $\alpha - \beta$ asymmetry points to an evolutionary driving force, which we further investigate.

In both N and C termini, the α helix and β strand distributions have a well-defined shape: they exhibit periodicity in the positioning of these elements along the sequence (Figs 1, *A* and *B*). This periodicity is matched by several $\alpha\beta$-type protein folds where α helices and β strands are arranged in alternating order (Fig. S5). These folds tend to be more abundant in prokaryotic proteins (Table S1); for example, ferredoxin-like folds exhibit high $\alpha - \beta$ asymmetry, significant periodicity at the N terminus, and are about three times more common in prokaryotes.

The distribution of disordered regions exhibits a different pattern of asymmetry: disordered residues are enriched at both ends of proteins compared with the middle (3,4). Eukaryotic proteins are significantly more disordered, where the probability of disorder is well approximated by $\sim D^{-0.5}$, where D is the distance from the end, while in prokaryotic proteins the probability of disorder decays as $\sim D^{-1}$. Proteins also tend to be more disordered at the N termini (3): eukaryotic proteins are 30 % more likely to be disordered within the first 100 residues of the N terminus compared with the C terminus (prokaryotes: 17 %). Although prokaryotic proteins are less disordered than eukaryotic ones, the ratio of the numbers of residues in β strands and α helices is the same.

Structural asymmetry correlates with sequence length and CO

To better understand the $\alpha - \beta$ asymmetry, we examined correlations with sequence length, L, and CO. High CO is likely to result in greater entropy loss between the unfolded and transition state, thus increasing folding time. To

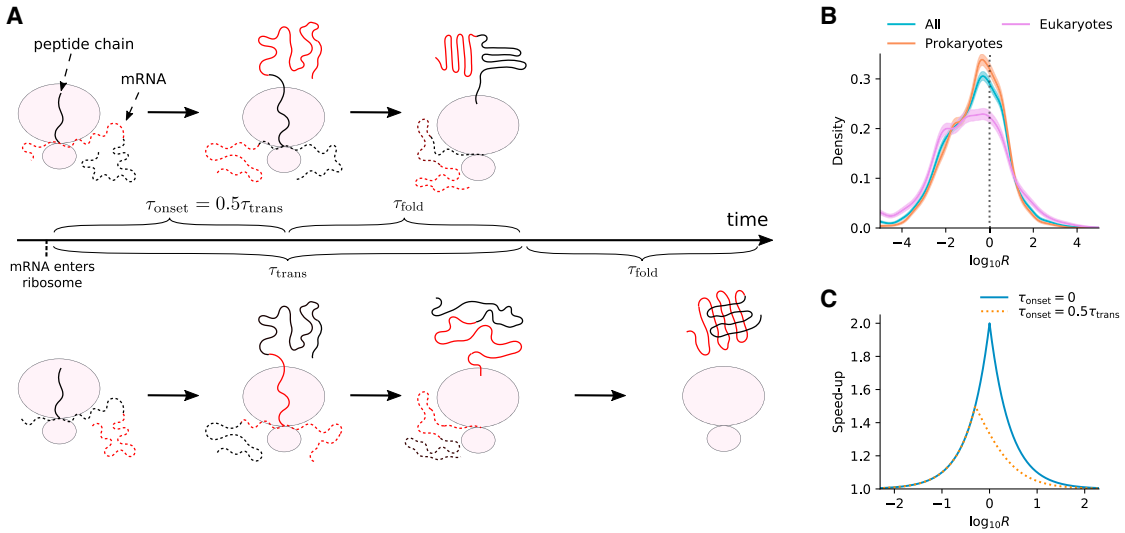

FIGURE 2 CTF and the slowest-first mechanism. (*A*) A section of mRNA (*red*), encoding a protein segment capable of folding cooperatively, is translated starting from the N-terminal side (*left*). The protein segment translocates through the ribosome channel (*middle*), and undergoes folding once the full segment has been translated and is free from steric constraints (*right*). The time from when the segment is about to undergo translation until the onset of folding is labeled τ_{onset}. If this segment is a bottleneck for protein folding, the protein will fold faster when this segment is located at the N terminus (*top*) instead of the C terminus (*bottom*). (*B*) Distribution of $R = \tau_{\text{fold}}/\tau_{\text{trans}}$, the estimated ratio of folding to translation time (Eq. 4), for our entire sample, prokaryotic proteins, and eukaryotic proteins. Solid lines are kernel density estimation fits to histograms; dotted line indicates $R = 1$; shading indicates bootstrapped 95% CIs. (*C*) Theoretical maximum speedup of production rate as a function of R and τ_{onset} (Eq. 5).

quantify secondary structure asymmetry, we calculate the magnitude of asymmetry normalized by length,

$$\begin{aligned} \text{asym}_\alpha &= (N_\alpha - C_\alpha)/L\,, \\ \text{asym}_\beta &= (N_\beta - C_\beta)/L\,, \end{aligned} \tag{3}$$

where N_α (N_β) and C_α (C_β) are the number of residues in a α helices (β strands) in the N and C halves of a protein sequence.

We find that $\alpha - \beta$ asymmetry is a non-monotonic function of both L and CO (Fig. 1, *C* and *D*; this is to some extent expected since L and CO are correlated, $r = 0.65$). In particular, there is a region of intermediate length ($179 - 416$) and intermediate CO ($21 - 35$) where structural asymmetry is most apparent, and extends as far as 100 residues from the ends (Fig. S6). The fact that both quantities correlate negatively with folding rate (L, $r = -0.68$; CO, $r = -0.64$; Fig. S2) (58,69,73,75), taken together with proteins' inherent asymmetry due to vectorial translation, leads us to suspect that the origins of this $\alpha - \beta$ asymmetry may be related to CTF.

CTF appears to be widespread

During protein production, the ribosome advances along the mRNA from the 5′ to the 3′ end, producing the corresponding protein from the N to the C terminus (Fig. 2 *A*). Proteins may fold in stages during translation (23–27,76–81), but the extent to which this happens in still unclear (32,36,77,82–86). In principle, one way to maximize the rate of production and to minimize aggregation is by making proteins fold faster than they are translated, or at a similar rate. We can obtain a rough approximation of how often this occurs by estimating folding rates and translation rates of proteins. We estimate the folding rate k_{fold} using a power-law scaling with length (not CO) fitted to data from the PFDB (58) (see section, "methods"). We assume an average translation rate k_{trans} that depends on the organism. Thus we can estimate the ratio R of folding time τ_{fold} to translation time τ_{trans}:

$$R = \frac{\tau_{\text{fold}}}{\tau_{\text{trans}}} = \frac{1/k_{\text{fold}}}{L/k_{\text{trans}}}\,. \tag{4}$$

The estimated R distribution exhibits a peak in the region of commensurate times $R \approx 1$ (Fig. 2 *B*). For the 68% of proteins (CI 53%–88%; Fig. S7) that lie in the region $R \leq 1$, folding may be quicker than translation, suggesting that CTF is common. In comparison, a more rigorous method estimated that, in 37% of proteins in *E. coli*, at least one domain will fully fold before translation finishes (87). Examining prokaryotic proteins and eukaryotic proteins separately reveals a sharper peak in the R distribution for prokaryotic proteins in the region of commensurate folding and translation times, $1/10 < R < 10$. Notably, a greater fraction of prokaryotic proteins (56%) are in this regime compared with eukaryotic proteins (41%).

Folding rate asymmetry can speed up CTF

Fig. 2 *A* shows two scenarios where the folding rate τ_{fold} is determined by a rate-limiting fold (RLF; red segment)

(50,86,88,89). This bottleneck represents an independent, cooperatively folding segment (i.e., a "foldon" (48)) that folds slowly. If the bottleneck is located at the N terminus (top in Fig. 2 *A*), then the production time is minimal, $\tau_{\min} = \max(\tau_{\text{fold}} + \tau_{\text{onset}}, \tau_{\text{trans}})$, where τ_{onset} is the time it takes for the segment to be translated, pass through the $\sim$ 10 -nm-long ribosome tunnel, and be free of steric constraints (29,75,76,90). In the other extreme, if the RLF includes the C terminus (bottom in Fig. 2 *A*), production time is maximized (91), $\tau_{\max} = \tau_{\text{fold}} + \tau_{\text{trans}}$. In this case, the last element can escape the ribosome quickly after being translated since it is not delayed by downstream translation (92). Thus, production rate can be accelerated by a factor:

$$\text{speedup} = \frac{\tau_{\max}}{\tau_{\min}} . \tag{5}$$

In the limit $\tau_{\text{onset}} \ll \tau_{\text{trans}}$, one finds from Eqs. 4 and 5 that the speedup as a function of $R = \tau_{\text{fold}}/\tau_{\text{trans}}$ (Fig. 2 *D*) is

$$\text{speedup} = 1 + e^{-|\ln R|} . \tag{6}$$

A maximal, 2-fold speedup is achieved when translation time equals folding time, $R = 1$, and taking $\tau_{\text{onset}} > 0$ shifts this maximum toward $R<1$. Since the maximal speedup occurs when the RLF is part of the N terminus, and when $\tau_{\text{onset}} \sim 0$, this effect should be more pronounced at the ends; we note that asymmetry is greatest close to the termini (Fig. S6).

In practice, τ_{onset} depends on the size of the cooperatively folding protein segment. It has been suggested that such a foldon can have as few as 20 residues (48,93), but, in general, cooperatively folding units appear to be larger than this; proteins that are observed to fold via two-state kinetics can vary greatly in size (94–96). Thus, this speedup may only be applicable to larger proteins that fold via multi-state kinetics; this is compatible with the prediction that a speedup is substantial only when $\tau_{\text{trans}} \approx \tau_{\text{fold}}$.

Structural asymmetry is maximum for commensurate folding and translation times

The speedup curve (Fig. 2 *C*) implies that proteins can benefit the most from structural asymmetry when $R = \tau_{\text{fold}}/\tau_{\text{trans}} \approx 1$. Hence, we estimate the magnitude of $\alpha - \beta$ asymmetry as a function of R and plot the distributions in Fig. 3 *A*. We note that R is not estimated very precisely, and thus it is useful to average out the errors by examining the proteins in large bins (deciles). At intermediate R, the means of the distributions shift away from zero, indicative of strong bias.

To capture the magnitude of these shifts we calculate the N-terminal enrichment, E, defined as the fraction of proteins with positive asymmetry (i.e., enriched at the N terminus) minus the fraction of proteins with negative asymmetry (enriched at the C terminus), for both helices and strands:

$$\begin{aligned} E_\alpha &= P(\text{asym}_\alpha > 0) - P(\text{asym}_\alpha < 0), \\ E_\beta &= P(\text{asym}_\beta > 0) - P(\text{asym}_\beta < 0). \end{aligned} \tag{7}$$

Fig. 3 *B* shows that, in the R deciles with maximum asymmetry, proteins in the PDB are 2.0 times as likely to be enriched in β strands in the N terminus, while α helices are 1.9 times more likely to be enriched in the C-terminal half. This maximum is found when $-0.6 \le \log_{10} R \le 0.1$ (the 95% CIs for the k_{fold} estimate give $-1.1 \le \log_{10} R \le 0.7$; Fig. S8). This region of maximal asymmetry overlaps with the region of maximal speedup (Fig. 2 *D*, Eq. 6), suggesting that

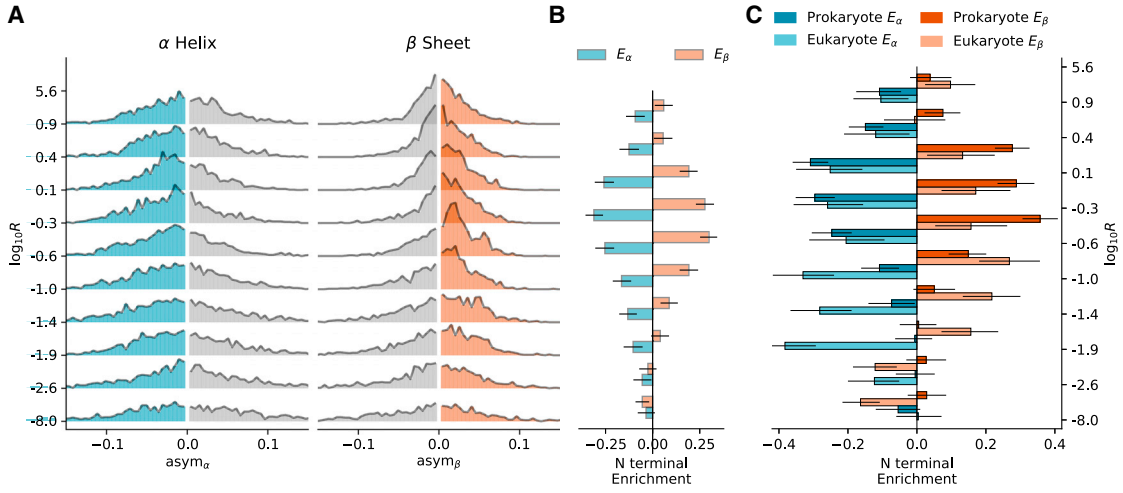

FIGURE 3 (*A*) The $\alpha - \beta$ asymmetry distributions as a function of R, the folding/translation time ratio (Eq. 4). Proteins are divided into deciles according to R; bin edges are shown on the y axis. (*B*) N-terminal enrichment—the degree to which strands/helices are enriched in the N over the C terminus (Eq. 7) —is shown for the deciles given in (*B*). (*C*) N-terminal enrichment as a function of R for 4702 eukaryotic proteins and 10,966 prokaryotic proteins. Proteins are divided into bins according to R; bin edges, shown on the y axis, are the same as in (*A*) and (*B*). Whiskers indicate bootstrapped 95% CIs.

asymmetry evolves because it enhances CTF. To help put these figures in context, in the pentile where proteins have maximum asymmetry ($-0.6 \leq \log_{10} R \leq 0.1$), 11% of proteins have $\text{asym}_\beta \geq 0.05$ (which, for the mean length, $L = 303$, and an average β strand length of five residues, corresponds to about three extra β strands near the N terminus), while half as many have $\text{asym}_\beta \leq -0.05$ (the opposite case) (Fig. S9).

Prokaryotes exhibit greater asymmetry than eukaryotes

We looked at $\alpha - \beta$ asymmetry for prokaryotic and eukaryotic proteins separately, finding that, when asymmetry is maximum, prokaryotes exhibit more asymmetry than eukaryotes; strands are 36% more likely to be enriched at the N terminus in prokaryotes compared with eukaryotes (Fig. 3 *C*). Typically, prokaryotic cells divide more frequently than eukaryotic cells (21) and thus have a greater need for fast production of functional proteins. The analysis is therefore consistent with the slowest-first scheme, which implies that the stronger pressure on prokaryotes should lead to greater asymmetry.

Multi-domain proteins are optimized for CTF via distinct mechanisms

Multi-domain proteins can be potentially adapted at two levels: within domains and between domains (97) (Fig. 4 *A*). To test this, we isolated individual domains in the PDB (using Pfam) (57), and calculated CO and $\alpha - \beta$ asymmetry for each domain as in Fig. 3. While intra-domain optimization of secondary structure clearly occurs within single-domain proteins, it is much weaker within multi-domain proteins (Fig. 4, *B* and *C*). Inter-domain optimization entails ordering the slowest-folding domains at the N terminus, for which we find no significant bias (Fig. S10). Instead, we find that, as the number of domains increases, the size and CO of individual domains decreases (Fig. 4, *D* and *E*). Domains in multi-domain proteins also contain about 20% less β sheets than single-domain proteins. Thus CTF is maintained in multi-domain proteins mostly by using faster-folding domains throughout.

DISCUSSION

We examined the hypothesis that proteins are selected for CTF to hasten protein production and reduce aggregation/misfolding, but this may not be equally true for all proteins. As an example, we showed that in prokaryotes, which have a greater burden of cell growth, proteins tend to have more asymmetry than in eukaryotes. Topological constraints may preclude early folding of the N terminus if the C terminus is also part of the folding nucleus (86). More generally, CTF may be hindered in some proteins by interactions with

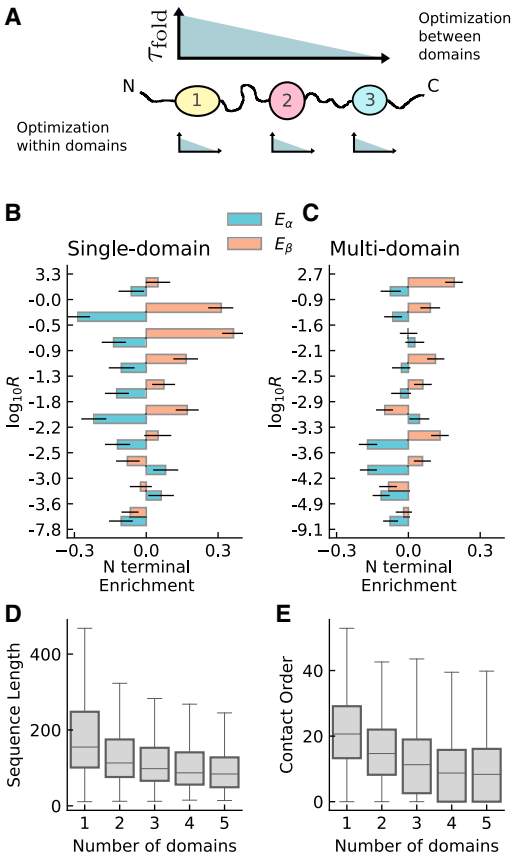

FIGURE 4 (*A*) Multi-domain proteins can be optimized via asymmetry between domains and/or within domains. (*B* and *C*) N-terminal enrichment within domains as a function of *R* for single-domain proteins (*B*: 14,442 domains) and multi-domain proteins (*C*: 23,832 domains). Domains are split into deciles based on *R*, and the bin edges are shown on the y axis; whiskers indicate bootstrapped 95% CIs. (*D* and *E*) Domain sequence length and CO distributions for proteins with different numbers of domains.

the ribosome (77). Long-lived proteins (98) may derive little benefit from an increase in production speed. On the other hand, proteins produced in large quantities need to fold quickly as aggregation can increase non-linearly with concentration (99). Some of these predictions can be tested when sufficient data for protein lifespan (100), expression levels (101), and structure become available. While we showed that $\alpha - \beta$ asymmetry is apparent in a broad set of proteins, further analysis of an extended data set may be able to detect the sub-classes of proteins that will benefit most from $\alpha - \beta$ asymmetry.

Our model proposes an absolute maximum of a 2-fold acceleration in protein production, which, at first glance, may seem insignificant compared with the range of protein folding times (about nine orders of magnitude). However, we show that this acceleration is only possible when folding time is similar to translation time (i.e., on the order of seconds to minutes). We expect that, even for a single protein, decreasing production time by seconds to minutes can increase fitness, and, when this is applied over a large group

of proteins, this should be quite substantial. In addition, locating an RLF at the N-terminal side rather than the other end will reduce the rate of aggregation, reducing both waste and harmful amyloid formation.

To experimentally test the slowest-first mechanism, we suggest studying CTF of multiple proteins with $R \approx 1$, which differ in asym_α and asym_β. In particular, we propose to use proteins whose sequences are related by "circular permutation," while having identical structures (85,102–104). Circular permutants with opposite structural asymmetry, as the example in Fig. 5, should fold at significantly different rates. Additional experimental control of R is possible via synonymous codon mutations (105) or in vitro expression systems (28). Thus, one can test whether asymmetry in secondary structure can lead to acceleration of CTF, and how this depends on R. The extensively used experimental technique of studying folding on stalled ribosomes also allows an indirect test of our hypothesis (36,81).

CTF for multi-domain proteins is more complex

Multi-domain proteins exhibit less asymmetry than single-domain proteins. Due to interactions between domains (39,42,76,90,107), optimization via asymmetry may not be feasible; instead, a safe strategy is to fold each domain before translating subsequent domains. To explain the lack of intra-domain $\alpha - \beta$ asymmetry (Fig. 4 *C*), we propose a simple mechanical argument. When a β sheet forms, the protein chain contracts. This results in a pulling force on both the ribosome (108,109) and on any upstream domains. This extra resistance to β sheet formation may preclude the early formation of β strands at the N-terminal side of a domain. If this is true, then the domain in position 1 should still exhibit $\alpha - \beta$ asymmetry; we currently lack sufficient statistical power to conclusively test this (Fig. S11). Further tests could look at CTF of a β-rich domain in the presence or absence of an upstream domain (110,111).

Disorder is enriched at both sequence ends

The N and C termini principally share a notable tendency for disorder near the end, which suggests that they are affected by the same physical "end effect." The amino acid at the end is linked to the chain by only one peptide bond, leaving it more configurational freedom than an amino acid in the center of the protein, which is constrained by two bonds. This entropic contribution to the free energy of the loose ends, of order k_BT, can induce disorder in marginally stable structures.

Since disordered regions do not need time to fold, placing them toward the C terminus gives the other residues more time to fold. Yet, we find a similar, slightly stronger, tendency for disorder near the N terminus (green curves in Fig. 1 *B*), particularly in eukaryotes. This may result from other determinants of protein evolution; e.g., disordered regions tend to interact with some ribosome-associating chaperones (112,113). If disorder at the N terminus is related to chaperones, we expect that asymmetry will be higher for slow-folding proteins as they are more prone to aggregation. We find that bias for disorder at the N terminus is strongest for slow-folding proteins (high R, L, and CO; Fig. S12), but only for prokaryotes, not eukaryotes. Given the absence of a correlation between R, L, and CO and disorder asymmetry in eukaryotic proteins, the question of why eukaryotic proteins are more disordered at the N terminus remains open.

The generality of these findings is potentially hindered by the fact that the PDB consists of a limited set of proteins whose structures are easier to determine. To avoid such biases, we look at statistics of disordered residues in structures predicted by the alphaFold algorithm (114) of the entire human and *E. coli* proteomes (Fig. S13). We find that both the end-middle asymmetry, and the N-C asymmetry is evident even when taking into account entire proteomes.

How strong is the link between theory and experiment?

In this study, we show correlations that support the slowest-first hypothesis, and this exploratory analysis alone is not sufficient to confirm or deny the hypothesis. In our estimation, the strongest link between theory and experiment is that the former predicts asymmetry at $R \sim 1$, and the latter shows maximal $\alpha - \beta$ asymmetry $R \sim 1$. It is important to note that the model directly predicts asymmetry in the positioning of RLFs, from which we infer the observed $\alpha - \beta$ asymmetry. This inference relies on the assumption that RLFs are more likely to be enriched in β strands than α helices.

We first explore why we judge this assumption to be most probable, compared with the opposite assumption, or with the null hypothesis that there is no enrichment of β strands and depletion of α helices in RLFs. On a basic level, interactions in α helices are inherently local, whereas β strands can be either local (e.g., hairpins) or long ranged. Consequentially, on average, β strands take longer to form contacts, and suffer a greater entropic penalty to stability (115). There is ample evidence linking topology to folding rate; structures with more long-range contacts tend to fold slower (73,104,116–121). Experimental support for this topological effect is found in designed peptides that fold close

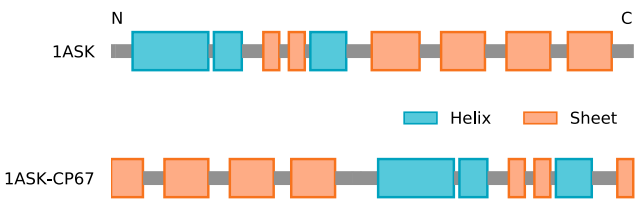

FIGURE 5 Secondary structure for nuclear transport factor 2 H66A mutant (PDB: 1ASK (106)) and a circular permutant, 1ASK-CP67, which may fold faster during translation.

to the speed limit. Helices have been found to fold at submicrosecond timescales, whereas the fastest-folding β hairpins are an order of magnitude slower than helices (122). This speed gap appears to widen when comparing whole proteins instead of peptides: we find that all-α proteins in the PFDB tend to fold about two orders of magnitude faster than all-β proteins, even when controlling for sequence length (Fig. S14). To further test the link between $\alpha - \beta$ asymmetry and RLFs, we run coarse-grained simulations of CTF of three structurally asymmetric proteins while varying R, for both the original sequence and for the reverse sequence. We find that these proteins fold faster when β strands are translated first, in the relevant region of $R \sim 1$ (Fig. S15). The above evidence from theory, experiment, and simulation strongly suggests that RLFs will tend to be enriched in β strands.

We can further test the hypothesis by examining the experimental data on other structural features that should be enriched in RLFs. First we consider the distribution of long-range contacts along the sequence, by calculating the average sequence distance between hydrogen bonds that are exclusively within the N-terminal half of the protein versus the bonds in the C-terminal half. This metric, calculated independently of the secondary structure, shows a maximum in the region $-0.6 \leq \log_{10} R \leq 0.1$, mirroring the trends found for $\alpha - \beta$ asymmetry (Fig. S16). Another independent observation consistent with the theory is in knotted proteins: there is a significant bias toward the N terminus being entangled rather than the C terminus (52,123). Another feature that we might expect from RLFs is that the folding nucleus may be less accessible to solvent. We thus look at asymmetry in surface accessibility. β Strands at the N terminus are less likely to be exposed to solvent than β strands at the C terminus; this bias is stronger for prokaryotic proteins (first 20 residues: 41 %) compared with eukaryotic proteins (13 %) (Fig. S17). Since solvent-exposed β strands are less likely to form part of a folding nucleus (124), this suggests that β strands at the N terminus are more likely to nucleate folding compared with those at the C terminus. These findings cannot prove or disprove the hypothesis on their own, as this work presents correlations rather than mechanistic causation. Nevertheless, taken together, the findings provide strong, consistent support for the hypothesis. The most direct test of the hypothesis would be whether the folding nuclei are found at the N terminus, specifically when $R \sim 1$. Thoroughly evaluating this prediction will require many more experimental studies on folding trajectories, or high-throughput studies (125,126), and the development of computational tools to predict the location of RLFs (127,128).

Alternatives explanations for asymmetry

We can think of only few examples of inherent asymmetry in proteins. The ends have a slight chemical difference (carbon versus nitrogen), but one atom seems insignificant. Amino acids are chiral, but we cannot say how this could lead to asymmetry in secondary structure. The only remaining driving force for asymmetry, in our view, is the unidirectional translation process.

One theory exists that predicts asymmetry in α helices: Marenduzzo et al. show that, in a model of a growing, self-interacting string, the string forms helices at the growing (C) end (7), and they suggest that this bias affected protein evolution. However, this fails to address why there are more β sheets at the N terminus, nor does it explain why asymmetry depends on protein length.

We could also consider that β strands are typically less open to solvent, and hence we also see asymmetry in solvent accessibility (Fig. S17). Thus, it is possible that β strand asymmetry is a by-product of a driving force for having the N-terminal side less exposed to solvent. The only explanation we can think of for such a driving force is to have the folding nucleus located at the N terminus, which is effectively covered by the slowest-first hypothesis since a folding nucleus can be considered an RLF.

We discussed reduction in aggregation as one of the potential benefits of faster protein production, but it is also possible that negative selection for aggregation propensity is the main driving force for asymmetry. One view is that β strands are more aggregation prone; it may be more efficient to prioritize their translation, as they are partially shielded in the ribosome vestibule and can interact with chaperones recruited by the ribosome. The reason β strands are more aggregation prone is their high hydrophobicity, so a more direct test of this alternative hypothesis is to look for asymmetry in hydrophobicity. We find that N-terminal halves are indeed more hydrophobic on average than C-terminal halves. However, there is no correlation with R (Fig. S18); neither do we have any grounds to expect that negative selection for avoiding aggregation should peak at $R \sim 1$. This indicates that there may indeed be a driving force to translate the aggregation-prone parts first, but this fails to explain why $\alpha - \beta$ asymmetry depends on length.

Finally, there remains the possibility that the observed asymmetry is simply a by-product of the diversification of ancestral proteins through gene duplication and divergence; if some asymmetry was present in early proteins, this could have become locked-in. We tested this by running the analyses on a subset of proteins where we only count each fold once, and showed that the asymmetry is still apparent (Fig. S4). It is not possible, however, to test the correlation between asymmetry and R, since this subset of proteins is too small (1526) to test for higher-order correlations. Nonetheless, even if we could show that the observed correlation between asymmetry and R is due to a number of popular folds, we could propose two contrasting explanations for this: (1) the folds are prevalent due to some other reasons, and they tend to show $\alpha - \beta$ asymmetry by chance; (2) the folds are popular due to the functional benefits of

$\alpha - \beta$ asymmetry. Our intention here is not to lend credence to either side but rather to highlight the difficulty in evaluating this particular hypothesis.

We need more data on folding rates

The data used to fit Eq. 1 are sparse (122 proteins); biased toward small, single-domain proteins; and typically obtained from in vitro refolding experiments (58). To test whether our conclusions are robust to sampling, we estimate CIs using bootstrapping with sample sizes equal to the original sample size, and half that amount; we perform this test on both the reduced version of the PFDB data set used in the main figures and on a second version of the PFDB data set (see section, "methods"; Fig. S8). In addition, we calculate the main results using a different protein folding data set, ACPro (59), which partially overlaps with PFDB but includes larger proteins (Fig. S19). In all of the above analyses, the point of maximum asymmetry is found to be $1/100<R<100$, which corresponds to the region where CTF speedup is possible. However, to have better clarity on this issue, we would welcome more experimental data on the following questions: how do large proteins fold, and how long do they take to fold? When, and how, is CTF different to in vitro refolding? How do large proteins fold on the ribosome?

CONCLUSION

To summarize, in the proposed the slowest-first mechanism, CTF can be accelerated by positioning the slowest-folding parts of a protein near the N terminus so that they have more time to fold. A survey of the PDB shows that the estimated acceleration correlates with asymmetry in secondary structure. In particular, the rate of production can be almost doubled when translation time is similar to folding time, and indeed these proteins exhibit the maximal asymmetry in secondary structure distribution. This correlation is robust to extensive analyses, and the hypothesis is contrasted with several alternative hypotheses, but further experiments are needed to definitively test the theory. Altogether, there appears to be substantial evolutionary selection, manifested in sequence asymmetry, for proteins that can fold during translation.

DATA AVAILABILITY

The non-redundant sets of proteins and domains, and data used in figures, are available in Zenodo, at http://doi.org/10.5281/zenodo.5612430. Code for analysis, simulations, and figures are available on GitHub, at https://github.com/jomimc/FoldAsymCode.

SUPPORTING MATERIAL

Supporting material can be found online at https://doi.org/10.1016/j.bpj.2021.11.024.

AUTHOR CONTRIBUTIONS

J.M. and T.T. designed research. J.M. performed research. J.M. analyzed data. J.M. and T.T. wrote the paper.

ACKNOWLEDGMENTS

We acknowledge Albert J. Libchaber for stimulating discussions and comments on the manuscript.

This work was supported by the Institute for Basic Science, project code IBS-R020-D1.

REFERENCES

1. Salas, M., M. A. Smith, and S. Ochoa. 1965. Direction of reading of the genetic message. *J. Biol. Chem.* 3995:3988–3995.
2. von Heijne, G. 1990. The signal peptide. *J. Membr. Biol.* 115:195–201.
3. Lobanov, M. Y., E. I. Furletova, and O. V. Galzitskaya. 2010. Library of disordered patterns in 3d protein structures. *PLoS Comput. Biol.* 6:1–10.
4. Lobanov, M. Y., I. V. Likhachev, and O. V. Galzitskaya. 2020. Disordered residues and patterns in the protein data bank. *Molecules.* 25:1522.
5. Thornton, J. M., and B. L. Chakauya. 1982. Conformation of terminal regions in proteins. *Nature.* 298:296–297.
6. Bhattacharyya, R., D. Pal, and P. Chakrabarti. 2002. Secondary structures at polypeptide-chain termini and their features. *Acta Crystallogr. D Biol. Crystallogr.* 58:1793–1802.
7. Marenduzzo, D., T. X. Hoang, and A. Maritan. 2005. Form of growing strings. *Phys. Rev. Lett.* 95:098103.
8. Krishna, M. M. G., and S. W. Englander. 2005. The N-terminal to C-terminal motif in protein folding and function. *Proc. Natl. Acad. Sci. U S A.* 102:1053–1058.
9. Laio, A., and C. Micheletti. 2006. Are structural biases at protein termini a signature of vectorial folding? *Proteins.* 62:17–23.
10. Saunders, R., M. Mann, and C. M. Deane. 2011. Signatures of co-translational folding. *Biotechnol. J.* 6:742–751.
11. Hopfield, J. J. 1974. Kinetic proofreading: a new mechanism for reducing errors in biosynthetic processes requiring high specificity. *Proc. Natl. Acad. Sci. U S A.* 71:4135–4139.
12. Ninio, J. 1975. Kinetic amplification of enzyme discrimination. *Biochimie.* 57:587–595.
13. Drummond, D. A., and C. O. Wilke. 2009. The evolutionary consequences of erroneous protein synthesis. *Nat. Rev. Genet.* 10:715–724.
14. Piñeros, W. D., and T. Tlusty. 2020. Kinetic proofreading and the limits of thermodynamic uncertainty. *Phys. Rev. E.* 101:022415.
15. Dobson, C. M. 1999. Protein misfolding, evolution and disease. *Trends Biochem. Sci.* 24:329–332.
16. López-Otín, C., M. A. Blasco, and G. Kroemer. 2013. The hallmarks of aging. *Cell.* 153:1194–1217.
17. Santra, M., K. A. Dill, and A. M. R. de Graff. 2019. Proteostasis collapse is a driver of cell aging and death. *Proc. Natl. Acad. Sci. U S A.* 116:22173–22178.
18. Spriggs, K. A., M. Bushell, and A. E. Willis. 2010. Translational regulation of gene expression during conditions of cell stress. *Mol. Cell.* 40:228–237.
19. de Nadal, E., G. Ammerer, and F. Posas. 2011. Controlling gene expression in response to stress. *Nat. Rev. Genet.* 12:833–845.
20. Dill, K. A., K. Ghosh, and J. D. Schmit. 2011. Physical limits of cells and proteomes. *Proc. Natl. Acad. Sci. U S A.* 108:17876–17882.

21. Zubkov, M. V. 2014. Faster growth of the major prokaryotic versus eukaryotic CO_2 fixers in the oligotrophic ocean. *Nat. Commun.* 5:3776.

22. Alexandrov, N. 1993. Structural argument for N-terminal initiation of protein folding. *Protein Sci.* 2:1989–1991.

23. Evans, M. S., I. M. Sander, and P. L. Clark. 2008. Cotranslational folding promotes beta-helix formation and avoids aggregation in vivo. *J. Mol. Biol.* 383:683–692.

24. Holtkamp, W., G. Kokic, and M. V. Rodnina. 2015. Cotranslational protein folding on the ribosome monitored in real time. *Science.* 350:1104–1107.

25. Kim, S. J., J. S. Yoon, and W. R. Skach. 2015. Translational tuning optimizes nascent protein folding in cells. *Science.* 348:444–448.

26. Wruck, F., A. Katranidis, and M. Hegner. 2017. Translation and folding of single proteins in real time. *Proc. Natl. Acad. Sci. U S A.* 114:E4399–E4407.

27. Nilsson, O. B., A. A. Nickson, and J. Clarke. 2017. Cotranslational folding of spectrin domains via partially structured states. *Nat. Struct. Mol. Biol.* 24:221–225.

28. Samelson, A. J., E. Bolin, and S. Marqusee. 2018. Kinetic and structural comparison of a protein's cotranslational folding and refolding pathways. *Sci. Adv.* 4:eaas9098.

29. Liutkute, M., E. Samatova, and M. V. Rodnina. 2020. Cotranslational folding of proteins on the ribosome. *Biomolecules.* 10:97.

30. Zhang, G., and Z. Ignatova. 2011. Folding at the birth of the nascent chain: coordinating translation with Co-translational folding. *Curr. Opin. Struct. Biol.* 21:25–31.

31. Trovato, F., and E. P. O'Brien. 2016. Insights into cotranslational nascent protein behavior from computer simulations. *Annu. Rev. Biophys.* 45:345–369.

32. Cabrita, L. D., A. M. E. Cassaignau, and J. Christodoulou. 2016. A structural ensemble of a ribosome–nascent chain complex during cotranslational protein folding. *Nat. Struct. Mol. Biol.* 23:278–285.

33. Waudby, C. A., C. M. Dobson, and J. J. Christodoulou. 2019. Nature and regulation of protein folding on the ribosome. *Trends Biochem. Sci.* 44:914–926.

34. Kramer, G., A. Shiber, and B. Bukau. 2019. Mechanisms of cotranslational maturation of newly synthesized proteins. *Annu. Rev. Biochem.* 88:337–364.

35. Kaiser, C. M., H. C. Chang, and J. M. Barral. 2006. Real-time observation of trigger factor function on translating ribosomes. *Nature.* 444:455–460.

36. Kaiser, C. M., D. H. Goldman, and C. Bustamante. 2011. The ribosome modulates nascent protein folding. *Science.* 334:1723–1727.

37. O'Brien, E. P., J. Christodoulou, and C. M. Dobson. 2012. Trigger factor slows co-translational folding through kinetic trapping while sterically protecting the nascent chain from aberrant cytosolic interactions. *J. Am. Chem. Soc.* 134:10920–10932.

38. O'Brien, E. P., M. Vendruscolo, and C. M. Dobson. 2014. Kinetic modelling indicates that fast-translating codons can coordinate cotranslational protein folding by avoiding misfolded intermediates. *Nat. Commun.* 5:2988.

39. Liu, K., K. Maciuba, and C. M. Kaiser. 2019. The ribosome cooperates with a chaperone to guide multi-domain protein folding. *Mol. Cell.* 74:310–319.e7.

40. Zhou, M., T. Wang, and Y. Liu. 2015. Nonoptimal codon usage influences protein structure in intrinsically disordered regions. *Mol. Microbiol.* 97:974–987.

41. Jacobs, W. M., and E. I. Shakhnovich. 2017. Evidence of evolutionary selection for cotranslational folding. *Proc. Natl. Acad. Sci. U S A.* 114:11434–11439.

42. Bitran, A., W. M. Jacobs, and E. Shakhnovich. 2020. Cotranslational folding allows misfolding-prone proteins to circumvent deep kinetic traps. *Proc. Natl. Acad. Sci. U S A.* 117:1485–1495.

43. Walsh, I. M., M. A. Bowman, and P. L. Clark. 2020. Synonymous codon substitutions perturb cotranslational protein folding in vivo and impair cell fitness. *Proc. Natl. Acad. Sci. U S A.* 117:3528–3534.

44. Cope, A. L., and M. A. Gilchrist. 2021. Quantifying shifts in natural selection on codon usage between protein regions: a population genetics approach. *bioRxiv* https://doi.org/10.1101/2020.12.18.423529.

45. Fersht, A. R. 1997. Nucleation mechanisms in protein folding. *Curr. Opin. Struct. Biol.* 7:3–9.

46. Ensign, D. L., P. M. Kasson, and V. S. Pande. 2007. Heterogeneity even at the speed limit of folding: large-scale molecular dynamics study of a fast-folding variant of the villin headpiece. *J. Mol. Biol.* 374:806–816.

47. Lindorff-Larsen, K., S. Piana, and D. E. Shaw. 2011. How fast-folding proteins fold. *Science.* 334:517–520.

48. Englander, S. W., and L. Mayne. 2014. The nature of protein folding pathways. *Proc. Natl. Acad. Sci. U S A.* 111:15873–15880.

49. Malhotra, P., and J. B. Udgaonkar. 2016. How cooperative are protein folding and unfolding transitions? *Protein Sci.* 25:1924–1941.

50. Jacobs, W. M., and E. I. Shakhnovich. 2016. Structure-based prediction of protein-folding transition paths. *Biophys. J.* 111:925–936.

51. Morrissey, M. P., Z. Ahmed, and E. I. Shakhnovich. 2004. The role of cotranslation in protein folding: a lattice model study. *Polymer.* 45:557–571.

52. Baiesi, M., E. Orlandini, and A. Trovato. 2019. Sequence and structural patterns detected in entangled proteins reveal the importance of co-translational folding. *Sci. Rep.* 9:8426.

53. Berman, H. M., J. Westbrook, and P. E. Bourne. 2000. The Protein Data Bank. *Nucleic Acids Res.* 28:235–242.

54. The Uniprot Consortium. 2018. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* 47:D506–D515.

55. Dana, J. M., A. Gutmanas, and S. Velankar. 2018. SIFTS: updated structure integration with function, taxonomy and sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins. *Nucleic Acids Res.* 47:D482–D489.

56. Armenteros, J. J. A., K. D. Tsirigos, and H. Nielsen. 2019. SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat. Biotechnol.* 37:420–423.

57. El-Gebali, S., J. Mistry, and R. D. Finn. 2019. The Pfam protein families database in 2019. *Nucleic Acids Res.* 47:D427–D432.

58. Manavalan, B., K. Kuwajima, and J. Lee. 2019. PFDB: a standardized protein folding database with temperature correction. *Sci. Rep.* 9:1588.

59. Wagaman, A. S., A. Coburn, and S. S. Jaswal. 2014. A comprehensive database of verified experimental data on protein folding kinetics. *Protein Sci.* 23:1808–1812.

60. Kubelka, J., J. Hofrichter, and W. A. Eaton. 2004. The protein folding 'speed limit. *Curr. Opin. Struct. Biol.* 14:76–88.

61. Rollins, G. C., and K. A. Dill. 2014. General mechanism of two-state protein folding kinetics. *J. Am. Chem. Soc.* 136:11420–11427.

62. Song, J., K. Takemoto, and T. Akutsu. 2010. Prediction of protein folding rates from structural topology and complex network properties. *IPSJ Trans. Bioinform.* 3:40–53.

63. Wagaman, A. S., and S. S. Jaswal. 2014. Capturing protein folding-relevant topology via absolute contact order variants. *J. Theor. Comput. Chem.* 13:1450005.

64. Thirumalai, D. 1995. From minimal models to real proteins: time scales for protein folding kinetics. *J. Phys.* 5:1457–1467.

65. Gutin, A. M., V. I. Abkevich, and E. I. Shakhnovich. 1996. Chain length scaling of protein folding time. *Phys. Rev. Lett.* 77:5433–5436.

66. Cieplak, M., T. X. Hoang, and M. S. Li. 1999. Scaling of folding properties in simple models of proteins. *Phys. Rev. Lett.* 83:1684–1687.

67. Koga, N., and S. Takada. 2001. Roles of native topology and chain-length scaling in protein folding: a simulation study with a GŌ-like model. *J. Mol. Biol.* 313:171–180.

68. Li, M. S., D. K. Klimov, and D. Thirumalai. 2002. Dependence of folding rates on protein length. *J. Phys. Chem. B.* 106:8302–8305.

69. Naganathan, A. N., and V. Muñoz. 2005. Scaling of folding times with protein size. *J. Am. Chem. Soc.* 127:480–481.

70. Lane, T. J., and V. S. Pande. 2012. A simple model predicts experimental folding rates and a hub-like topology. *J. Phys. Chem. B.* 116:6764–6774.

71. Garbuzynskiy, S. O., D. N. Ivankov, and A. V. Finkelstein. 2013. Golden triangle for folding rates of globular proteins. *Proc. Natl. Acad. Sci. U S A.* 110:147–150.

72. Lane, T. J., and V. S. Pande. 2013. Inferring the rate-length law of protein folding. *PLoS One.* 8:1–5.

73. Ivankov, D. N., S. O. Garbuzynskiy, and A. V. Finkelstein. 2003. Contact order revisited: influence of protein size on the folding rate. *Protein Sci.* 12:2057–2062.

74. Andreevna, A., D. Howorth, C. Chothia, E. Kulesha, and A. Murzin. 2013. SCOP2 prototype: a new approach to protein structure mining. *Nucleic Acids Res.* 42:D310–D314. https://doi.org/10.1093/nar/gku841.

75. Zhao, V., W. M. Jacobs, and E. I. Shakhnovich. 2020. Effect of protein structure on evolution of cotranslational folding. *Biophys. J.* 119:1123–1134.

76. Notari, L., M. Martínez-Carranza, and G. von Heijne. 2018. Cotranslational folding of a pentarepeat beta-helix protein. *J. Mol. Biol.* 430:5196–5206.

77. Kemp, G., R. Kudva, and G. von Heijne. 2019. Force-profile analysis of the cotranslational folding of HemK and filamin domains: comparison of biochemical and biophysical folding assays. *J. Mol. Biol.* 431:1308–1314.

78. Mercier, E., and M. V. Rodnina. 2018. Co-translational folding trajectory of the HemK helical domain. *Biochemistry.* 57:3460–3464.

79. Liutkute, M., M. Maiti, and M. V. Rodnina. 2020. Gradual compaction of the nascent peptide during cotranslational folding on the ribosome. *eLife.* 9:e60895.

80. Elfageih, R., A. Karyolaimos, and R. Kudva. 2020. Cotranslational folding of alkaline phosphatase in the periplasm of *Escherichia coli*. *Protein Sci.* 29:2028–2037.

81. Khushoo, A., Z. Yang, and W. R. Skach. 2011. Ligand-driven vectorial folding of ribosome-bound human CFTR Nbd1. *Mol. Cell.* 41:682–692.

82. Waudby, C. A., T. Wlodarski, and J. Christodoulou. 2018. Systematic mapping of free energy landscapes of a growing filamin domain during biosynthesis. *Proc. Natl. Acad. Sci. U S A.* 115:9744–9749.

83. Eichmann, C., S. Preissler, and E. Deuerling. 2010. Cotranslational structure acquisition of nascent polypeptides monitored by NMR spectroscopy. *Proc. Natl. Acad. Sci. U S A.* 107:9111–9116.

84. Guinn, E. J., P. Tian, and S. Marqusee. 2018. A small single-domain protein folds through the same pathway on and off the ribosome. *Proc. Natl. Acad. Sci. U S A.* 115:12206–12211.

85. Marsden, A. P., J. J. Hollins, and J. Clarke. 2018. Investigating the effect of chain connectivity on the folding of a beta-sheet protein on and off the ribosome. *J. Mol. Biol.* 430:5207–5216.

86. Tian, P., A. Steward, and R. B. Best. 2018. Folding pathway of an Ig domain is conserved on and off the ribosome. *Proc. Natl. Acad. Sci. U S A.* 115:E11284–E11293.

87. Ciryam, P., R. I. Morimoto, and E. P. O'Brien. 2013. In vivo translation rates can substantially delay the cotranslational folding of the *Escherichia coli* cytosolic proteome. *Proc. Natl. Acad. Sci. U S A.* 110:E132–E140.

88. Friel, C. T., A. P. Capaldi, and S. E. Radford. 2003. Structural analysis of the rate-limiting transition states in the folding of Im7 and Im9: similarities and differences in the folding of homologous proteins. *J. Mol. Biol.* 326:293–305.

89. Hanazono, Y., K. Takeda, and K. Miki. 2016. Structural studies of the N-terminal fragments of the WW domain: insights into Co-translational folding of a beta-sheet protein. *Sci. Rep.* 6:34654.

90. Liu, K., X. Chen, and C. M. Kaiser. 2019. Energetic dependencies dictate folding mechanism in a complex protein. *Proc. Natl. Acad. Sci. U S A.* 116:25641–25648.

91. Chen, X., N. Rajasekaran, and C. M. Kaiser. 2020. Synthesis runs counter to directional folding of a nascent protein domain. *Nat. Commun.* 11:5096.

92. Nissley, D. A., Q. V. Vu, and E. P. O'Brien. 2020. Electrostatic interactions govern extreme nascent protein ejection times from ribosomes and can delay ribosome recycling. *J. Am. Chem. Soc.* 142:6103–6110.

93. Hu, W., Z. Y. Kan, and S. W. Englander. 2016. Cytochrome C folds through foldon-dependent native-like intermediates in an ordered pathway. *Proc. Natl. Acad. Sci. U S A.* 113:3809–3814.

94. de Prat Gay, G., J. Ruiz-Sanz, and A. R. Fersht. 1995. Conformational pathway of the polypeptide chain of chymotrypsin inhibitor-2 growing from its N terminus in vitro. Parallels with the protein folding pathway. *J. Mol. Biol.* 254:968–979.

95. Fersht, A. 1999. Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding. Macmillan.

96. Jones, K., and P. Wittung-Stafshede. 2003. The largest protein observed to fold by two-state kinetic mechanism does not obey contact-order correlation. *J. Am. Chem. Soc.* 125:9606–9607.

97. Jacob, E., R. Unger, and A. Horovitz. 2013. N-terminal domains in two-domain proteins are biased to be shorter and predicted to fold faster than their C-terminal counterparts. *Cell Rep.* 3:1051–1056. https://doi.org/10.1016/j.celrep.2013.03.032.

98. Toyama, B. H., and M. W. Hetzer. 2013. Protein homeostasis: live long, won't prosper. *Nat. Rev. Mol. Cell Biol.* 14:55–61.

99. Wang, W. 2005. Protein aggregation and its inhibition in biopharmaceutics. *Int. J. Pharm.* 289:1–30.

100. Fornasiero, E. F., S. Mandad, and S. O. Rizzoli. 2018. Precisely measured protein lifetimes in the mouse brain reveal differences across tissues and subcellular fractions. *Nat. Commun.* 9:4230.

101. Yun, H., J. W. Lee, and S. Y. Lee. 2007. EcoProDB: the Escherichia coli protein database. *Bioinformatics.* 23:2501–2503.

102. Miller, E. J., K. F. Fischer, and S. Marqusee. 2002. Experimental evaluation of topological parameters determining protein-folding rates. *Proc. Natl. Acad. Sci. U S A.* 99:10359–10363.

103. Lo, W. C., C. C. Lee, and P. C. Lyu. 2008. Cpdb: a database of circular permutation in proteins. *Nucleic Acids Res.* 37:D328–D332.

104. Kemplen, K. R., D. De Sancho, and J. Clarke. 2015. The response of Greek key proteins to changes in connectivity depends on the nature of their secondary structure. *J. Mol. Biol.* 427:2159–2165.

105. Komar, A. A., T. Lesnik, and C. Reiss. 1999. Synonymous codon substitutions affect ribosome traffic and protein folding during in vitro translation. *FEBS Lett.* 462:387–391.

106. Clarkson, W. D., A. H. Corbett, and M. Stewart. 1997. Nuclear protein import is decreased by engineered mutants of nuclear transport factor 2 (Ntf2) that do not bind gdp-ran11edited by I. B. Holland. *J. Mol. Biol.* 272:716–730.

107. Han, J. H., S. Batey, and J. Clarke. 2007. The folding and evolution of multidomain proteins. *Nat. Rev. Mol. Cell Biol.* 8:319–330.

108. Goldman, D. H., C. M. Kaiser, and C. Bustamante. 2015. Mechanical force releases nascent chain–mediated ribosome arrest in vitro and in vivo. *Science.* 348:457–460.

109. Leininger, S. E., F. Trovato, and E. P. O'Brien. 2019. Domain topology, stability, and translation speed determine mechanical force generation on the ribosome. *Proc. Natl. Acad. Sci. U S A.* 116:5523–5532.

110. Batey, S., L. G. Randles, and J. Clarke. 2005. Cooperative folding in a multi-domain protein. *J. Mol. Biol.* 349:1045–1059.

111. Kemp, G., O. B. Nilsson, and G. von Heijne. 2020. Cotranslational folding cooperativity of contiguous domains of alpha-spectrin. *Proc. Natl. Acad. Sci. U S A.* 117:14119–14126.

112. Alamo, M., D. J. Hogan, and J. Frydman. 2011. Defining the specificity of cotranslationally acting chaperones by systematic analysis

of mRNAs associated with ribosome-nascent chain complexes. *PLoS Biol.* 9:1–23.

113. Willmund, F., M. del Alamo, and J. Frydman. 2013. The cotranslational function of ribosome-associated Hsp70 in eukaryotic protein homeostasis. *Cell.* 152:196–209.

114. Tunyasuvunakool, K., J. Adler, and D. Hassabis. 2021. Highly accurate protein structure prediction for the human proteome. *Nature.* 596:590–596.

115. Stanger, H. E., F. A. Syud, and S. H. Gellman. 2001. Length-dependent stability and strand length limits in antiparallel beta-sheet secondary structure. *Proc. Natl. Acad. Sci. U S A.* 98:12015–12020.

116. Chiti, F., N. Taddei, and C. M. Dobson. 1999. Mutational analysis of acylphosphatase suggests the importance of topology and contact order in protein folding. *Nat. Struct. Biol.* 6:1005–1009.

117. Lindberg, M. O., J. Tångrot, and M. Oliveberg. 2001. Folding of circular permutants with decreased contact order: general trend balanced by protein stability. *J. Mol. Biol.* 314:891–900.

118. Woodson, S. A. 2002. Folding mechanisms of group I ribozymes: role of stability and contact order. *Biochem. Soc. Trans.* 30:1166–1169.

119. Kaya, H., and H. S. Chan. 2003. Contact order dependent protein folding rates: kinetic consequences of a cooperative interplay between favorable nonlocal interactions and local conformational preferences. *Proteins.* 52:524–533.

120. Steward, A., G. S. McDowell, and J. Clarke. 2009. Topology is the principal determinant in the folding of a complex all-alpha Greek key death domain from human FADD. *J. Mol. Biol.* 389:425–437.

121. Bandyopadhyay, B., T. Mondal, and A. Horovitz. 2019. Contact order is a determinant for the dependence of GFP folding on the chaperonin GroEL. *Biophys. J.* 116:42–48.

122. Eaton, W. A., V. Muñoz, and J. Hofrichter. 2000. Fast kinetics and mechanisms in protein folding. *Annu. Rev. Biophys. Biomol. Struct.* 29:327–359.

123. Norcross, T. S., and T. O. Yeates. 2006. A framework for describing topological frustration in models of protein folding. *J. Mol. Biol.* 362:605–621.

124. Nölting, B., and D. A. Agard. 2008. How general is the nucleation-condensation mechanism? *Proteins.* 73:754–764.

125. Magliery, T. J., J. J. Lavinder, and B. J. Sullivan. 2011. Protein stability by number: high-throughput and statistical approaches to one of protein science's most difficult problems. *Curr. Opin. Chem. Biol.* 15:443–451.

126. Fowler, D. M., and S. Fields. 2014. Deep mutational scanning: a new style of protein science. *Nat. Methods.* 11:801–807.

127. Matsuoka, M., and T. Kikuchi. 2014. Sequence analysis on the information of folding initiation segments in ferredoxin-like fold proteins. *BMC Struct. Biol.* 14:15.

128. Sacquin-Mora, S. 2015. Fold and flexibility: what can proteins' mechanical properties tell us about their folding nucleus? *J. R. Soc. Interface.* 12:20150876.