In vivo translation rates can substantially delay the cotranslational folding of the *Escherichia coli* cytosolic proteome

Prajwal Ciryam[a,b], Richard I. Morimoto[b], Michele Vendruscolo[a], Christopher M. Dobson[a], and Edward P. O'Brien[a,1]

[a]Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, United Kingdom; and [b]Department of Molecular Biosciences, Northwestern University, Evanston, IL 60208

Edited by Ken A. Dill, Stony Brook University, Stony Brook, NY, and approved November 9, 2012 (received for review August 8, 2012)

A question of fundamental importance concerning protein folding in vivo is whether the kinetics of translation or the thermodynamics of the ribosome nascent chain (RNC) complex is the major determinant of cotranslational folding behavior. This is because translation rates can reduce the probability of cotranslational folding below that associated with arrested ribosomes, whose behavior is determined by the equilibrium thermodynamics of the RNC complex. Here, we combine a chemical kinetic equation with genomic and proteomic data to predict domain folding probabilities as a function of nascent chain length for *Escherichia coli* cytosolic proteins synthesized on both arrested and continuously translating ribosomes. Our results indicate that, at in vivo translation rates, about one-third of the *Escherichia coli* cytosolic proteins exhibit cotranslational folding, with at least one domain in each of these proteins folding into its stable native structure before the full-length protein is released from the ribosome. The majority of these cotranslational folding domains are influenced by translation kinetics which reduces their probability of cotranslational folding and consequently increases the nascent chain length at which they fold into their native structures. For about 20% of all cytosolic proteins this delay in folding can exceed the length of the completely synthesized protein, causing one or more of their domains to switch from co- to posttranslational folding solely as a result of the in vivo translation rates. These kinetic effects arise from the difference in time scales of folding and amino-acid addition, and they represent a source of metastability in *Escherichia coli*'s proteome.

systems biology | synonymous codons | chemical kinetics | chaperone | aggregation

During translation, the ribosome coordinates a series of complex processes, including the unidirectional translocation of an mRNA molecule one codon at a time, the selection for each codon of a complementary tRNA from the cytosol, and the catalysis of the peptide bond. The result is the synthesis of the corresponding nascent chain and its extrusion through a tunnel embedded within the large ribosomal subunit. As individual domains of a multidomain protein emerge from this exit tunnel, they have the opportunity to form secondary and tertiary structures and fold into their native or native-like structures while the rest of the nascent chain is still being synthesized (Fig. 1*A*). Such cotranslational folding has been shown to occur both in vivo and in vitro (1–3), with large multidomain proteins being more likely to exhibit cotranslational folding than small single domain proteins (4).

The time-dependent, irreversible nature of translation means that cotranslational folding occurs out of equilibrium. Consequently, the interplay of time scales arising from different processes involving the ribosome nascent chain (RNC) complex has the potential to modulate the probability that cotranslational folding occurs (5, 6). There are at least three time scales that can have an impact on the cotranslational folding process at nascent chain length i: (*a*) the mean time it takes to add an amino acid to the growing nascent chain (denoted $\tau_{A,i}$) (7), (*b*) the mean time of domain unfolding (denoted $\tau_{U,i}$), and (*c*) the mean time of domain folding (denoted $\tau_{F,i}$). The cotranslational folding of the multidomain SufI protein, for example, was abolished when the translation rate ($1/\tau_{A,i}$) was increased by either replacing slow translating codons located near domain boundaries with fast translating codons or adding excess amounts of aminoacyl-tRNA (8). Moreover, a simulation study that examined the cotranslational folding of protein G (Fig. 1*A*) found that uniformly speeding up translation suppressed cotranslational folding and that placing slow or fast translating codons at specific points along an mRNA molecule could, respectively, increase or decrease the amount of cotranslational folding that occurred (7). Thus, increasing the translation rate can shift the cotranslational folding curve toward longer nascent chain lengths (Fig. 1 *B* and *C*).

Although the cotranslational folding properties of an increasing number of proteins are being characterized experimentally in vitro (9, 10), very little is known about the importance of cotranslational folding in vivo for entire proteomes of specific organisms. Therefore, one fundamental question that arises concerns the percentage of an organism's proteome that folds cotranslationally. A second fundamental question is to what extent proteins that fold cotranslationally do so under a regime governed by the kinetics of translation (which we refer to as kinetic effects) or the thermodynamics of the RNC complex (which we refer to as thermodynamic effects). This issue is important because cotranslational folding populations, pathways, structures, and mechanisms can differ significantly depending on whether kinetic or thermodynamic effects predominate. If kinetic effects occur, the cotranslational folding probability will be less than its thermodynamically determined value at one or more nascent chain lengths and the length at which stable domain folding is achieved (defined as $P_F > 0.5$) will be increased (Fig. 1*B*). Consequently, kinetic effects will tend to delay cotranslational folding as a function of nascent chain length, resulting in an excess population of metastable, ribosome-bound unfolded states.

We address these questions for the cytosolic proteome of *Escherichia coli*. We focus on this organism because it is unicellular, which avoids issues of proteome variation among different cell types. *E. coli* has also been extensively characterized in terms of its genome and proteome, and the concentrations of ribosome and tRNA molecules present at different growth rates have been measured (11), allowing predictive models for $\tau_{A,i}$ to

Author contributions: P.C., R.I.M., M.V., C.M.D., and E.P.O. designed research; P.C. and E.P.O. performed research; P.C. and E.P.O. contributed new reagents/analytic tools; P.C., R.I.M., M.V., C.M.D., and E.P.O. analyzed data; and P.C., R.I.M., M.V., C.M.D., and E.P.O. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. E-mail: eo264@cam.ac.uk.

See Author Summary on page 396 (volume 110, number 2).

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1213624110/-/DCSupplemental.

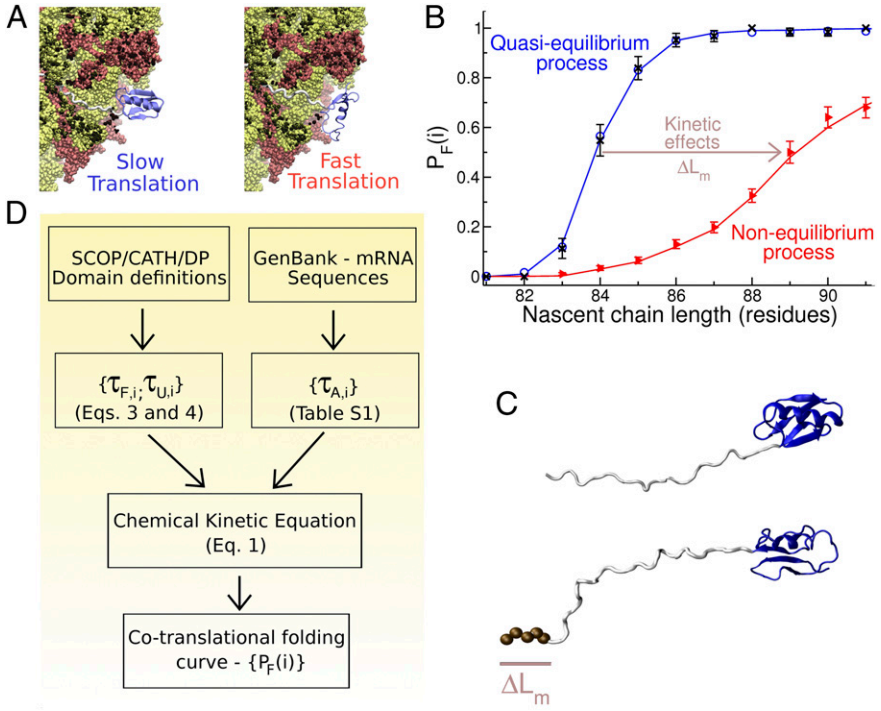

Fig. 1. A systems approach for predicting the cotranslational folding behavior of *E. coli*'s cytosolic proteome. (*A*) 50*S* ribosomal subunit with ribosomal protein and RNA molecules shown, respectively, in red and yellow. The protein G domain (blue) is shown emerging from the exit tunnel, and it is tethered to the P-site tRNA by an unstructured poly-glycine linker (white). (*B*) Cotranslational folding curves (P_F) for a protein G domain synthesized at finite translation rates (red triangles, τ_A = 1.3 ms) and at an infinitely slow translation rate (black X symbols). Data were taken from the coarse-grained molecular dynamics simulations reported by O'Brien et al. (7). The deviation between the $P_F(i)$ curves is characterized by ΔL_m, the separation in nascent chain lengths at which the native state becomes stable, that is when $P_F(i) > 0.5$. The simulation results for τ_A = 60 ms are shown as blue circles; at this translation speed, $\Delta L_m = 0$. (*C*) ΔL_m can be thought of as the additional number of residues (brown spheres) required for a domain to achieve its stable folded structure at finite translation rates (*Lower*) compared with infinitely slow translation (*Upper*). Structures were taken from coarse-grained simulations of protein G reported by O'Brien et al. (7). (*D*) Illustration of the work flow of our systems approach.

be developed (12). We focus on the *E. coli* proteins that are synthesized and remain in the cytosol because this strategy avoids the need to consider additional processes, such as cotranslational protein translocation through translocons (13), that would complicate our approach to modeling and analysis.

Our unique systems approach to these questions combines a kinetic model, which predicts the effect of individual codon translation rates on the cotranslational folding of protein domains (7), with biophysical data on *E. coli* to predict the cotranslational folding curves for protein domains in the *E. coli* cytosol (Fig. 1*D*). In this way, we have been able to calculate cotranslational folding curves at the rate of translation found in *E. coli* (an out-of-equilibrium process), as well as the curves that would arise from an infinitely slow translation process (a quasiequilibrium process; Fig. 1*B*). Our results indicate that a number of proteins exhibit cotranslational folding in *E. coli* and that for many of these proteins, translation kinetics, rather than the thermodynamics of the RNC complex, govern their cotranslational folding behavior. This important physical feature of cotranslational folding has a number of significant implications for cellular biology, biotechnology, and synthetic biology, as we discuss below.

Methods

Creation of a Database Characterizing *E. coli*'s Cytosolic Proteome. The wealth of quantitative data characterizing the K12 MG1655 strain of the Gram-negative bacterium *E. coli* enables us to estimate the cotranslational folding properties of its cytosolic proteome. The data we used in this study were drawn from several sources, which we organized into a database (Dataset S1 and see the *SI Appendix* for dataset formatting information). Full details of the construction of this database are provided in Supplemental Information. Briefly, for the 4,319 unique coding sequences in the *E. coli* transcriptome, we identified those that had an X-ray structure in the Research Collaboratory for Structural Bioinformatics (RCSB) Protein Data Bank (PDB) (14) with a resolution of less than or equal to 3 Å or an NMR entry, and also had domain definitions in the Structural Classification of Proteins (SCOP) (15) or CATH (16) database. This procedure resulted in 802 proteins. For the 489 additional proteins that had an RCSB structure but no SCOP or CATH entry, we used Domain Parser (DP) (17) to identify their domains; the latter algorithm was able to identify domains in 264 of these proteins, and the remaining proteins were not included in the database. A small percentage of SCOP and CATH domains were reported to be made up of multiple segments. Given that the sequence separation between segments is quite large (Fig. S1), we treated these discontiguous segments as separate domains, provided they were at least 50 residues in length. We then used PSORTb 3.0.2 (18) to classify proteins on the basis of their subcellular localization; PSORTb includes 4,148 entries that correspond to 4,129 Uniprot identification numbers, of which 1,898 are cytoplasmic, 818 are of unknown localization, and 54 fall into a category of unknown/multiple compartments. To maximize the coverage of our database, we assumed that proteins of unknown localization were cytoplasmic. After removing redundant structures that mapped onto the same protein domain, the database consisted of 758 unique proteins containing 1,236 domains, which covers 30% of the proteins that are cytosolic in *E. coli*. For each domain entry in our database, we report its corresponding range of residue numbers in the PDB file, its range of codon numbers along the mRNA molecule, its mRNA sequence, its domain structural classification (mostly α, mostly β, or mixed α/β), and its total number of residues. Furthermore, for each protein, we report its corresponding gene expression level in *E. coli* from several experimental sources. These data were used to inform the kinetic model and the analysis that are detailed below.

Although our database contains only the subset of *E. coli* cytosolic proteins for which structural data are available, we find no systematic difference between the length distribution of proteins in our database and that in the full set of cytosolic proteins (Fig. S2), indicating that our database is representative of the full cytosolic proteome, at least in terms of protein size.

Calculation of the Cotranslational Folding Curve for a Given Protein Domain. Central to our analysis is the calculation of the cotranslational folding probability of a protein domain as a function of its nascent chain length, referred to as the cotranslational folding curve. We assume the domains in our database are autonomous, cooperative folding units that exist in either a folded or unfolded state with a negligible population of intermediates. For a ribosome that exhibits exponential dwell times at each codon, the ensemble averaged domain folding probability $P_F(i)$ as a function of nascent chain length i is (7)

$$P_F(i) = \sum_{j=1}^{i} \frac{\tau_{F,j}^{-1}}{\tau_{A,j+1}^{-1}} \prod_{k=j}^{i} \frac{\tau_{A,k+1}^{-1}}{\tau_{A,k+1}^{-1} + \tau_{F,k}^{-1} + \tau_{U,k}^{-1}}, \quad [1]$$

where $\tau_{F,i}$, $\tau_{U,i}$, and $\tau_{A,i+1}$ are, respectively, the mean times of domain folding, domain unfolding, and amino acid addition at nascent chain length i. In this equation, $P_F(i)$ is the probability the domain is folded immediately before adding the next amino acid (Fig. 1*B*). The equilibrium cotranslational folding curve $P_F^E(i)$ can be calculated by inserting into Eq. **1** $\tau_{A,i} = \infty$ for all lengths of i, which reduces it to

$$P_F^E(i) = \frac{\tau_{F,i}^{-1}}{\tau_{F,i}^{-1} + \tau_{U,i}^{-1}}. \quad [2]$$

To calculate $\tau_{A,i}$, the mean time of translation of the ith codon in the mRNA sequence, we used the method of Viljoen and colleagues (12), which estimates $\tau_{A,i}$ as a function of the codon identity, and the in vivo concentration of cognate, near-cognate, and noncognate tRNA molecules (*SI Methods*). This method accounts for competitive binding of cognate and noncognate tRNA molecules for a codon and accurately predicts the average translation rate of various proteins measured in vivo. We used the in vivo concentrations of the ribosome and tRNA molecules measured at five different *E. coli* exponential growth rates (11) to calculate $\tau_{A,i}$ for each codon at each growth rate (Tables S1 and S2).

To estimate $\tau_{F,i}$ and $\tau_{U,i}$ for a given domain, we fit an empirical scaling relationship to the cotranslational folding and unfolding kinetics of protein G from previously reported coarse-grained simulations (7) of this RNC complex (Fig. 2*A*). The relationships have the functional forms

$$\tau_{F,i} = \tau_{F,bulk} + a\tau_{F,bulk} e^{-i+l+25} + b\tau_{F,bulk}/i^c, \quad [3]$$

$$\tau_{U,i} = d\tau_{F,bulk} \Big/ \left(1 + e^{-i+l+30}\right), \quad [4]$$

where the fitting parameters a, b, c, and d have values of 404, 3205.5, 1.72, and 0.953, respectively. In Eqs. **3** and **4**, l is the codon number corresponding to the C-terminal residue of the domain, i is the nascent chain length in number of residues, and c is an exponent whose value is equal to 1.72. The simulations on which the fit parameters were determined used an unstructured poly-glycine linker to attach the protein G domain to the peptidyl transferase center (PTC) (7). We do not expect there to be significant changes in the behavior of these equations when different sequence compositions for the linkers are used.

To estimate $\tau_{F,bulk}$ and $\tau_{U,bulk}$, the folding and unfolding times of the domain when it is free in solution, in Eqs. **3** and **4**, we used the De Sancho–Muñoz (DM) model (19), which predicts these two time scales based on the number of residues in the domain, its structural class, as well as the solution temperature (*SI Methods*). The DM model was chosen over others (20) because it provides an estimate of $\tau_{U,bulk}$, which many other models do not provide, and it also accounts for the effect of temperature on these time scales. The latter is important because we will analyze *E. coli*'s proteome behavior near its optimal cellular growth temperature of 37 °C (21) and most quantitative experimental measurements of bulk folding behavior have been carried out near 25 °C (22). The l, $\tau_{F,bulk}$, and $\tau_{U,bulk}$ values for each domain are reported in Dataset S1.

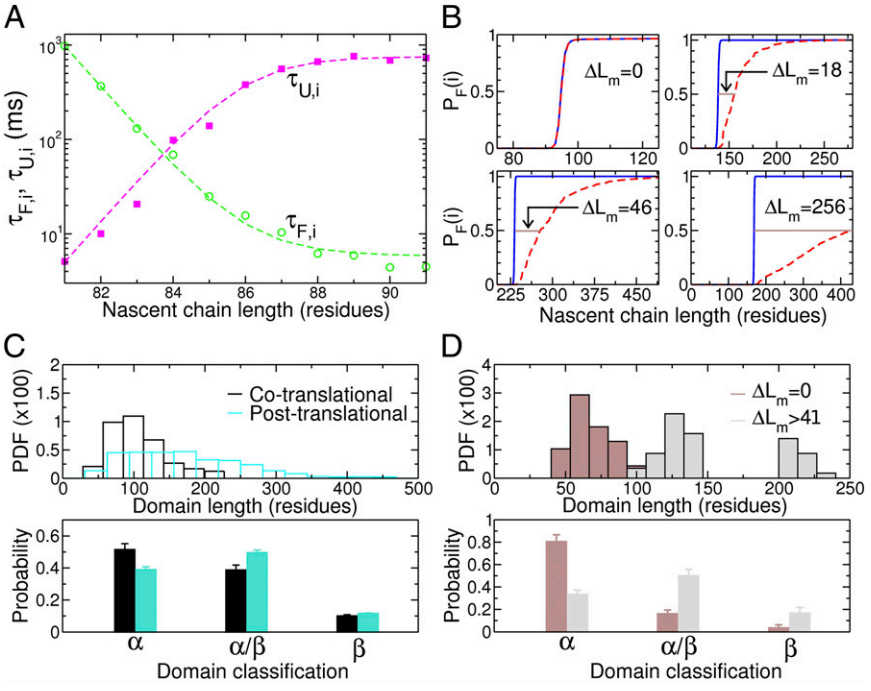

Fig. 2. Cotranslational folding of cytosolic protein domains in *E. coli*. (*A*) Determination of a scaling relationship to model cotranslational domain folding/unfolding kinetics. The mean folding (green) and unfolding (magenta) times of protein G as a function of nascent chain length calculated from coarse-grained simulations (7) were fitted by Eqs. **3** and **4** (dashed lines). (*B*) Examples of cotranslational folding curves calculated for four different protein domains in *E. coli* at in vivo (red, Eq. **1**) and infinitely slow (blue, Eq. **2**) translation rates. The domains correspond, respectively, to (ASNC_ECOLI, domain 1; *Upper Left*), (3MG2_ECOLI, domain 1; *Upper Right*), (ILVC_ECOLI, domain 1; *Lower Left*), and (ENO, domain 1; *Lower Right*) in Dataset S1. (*Upper Left*) Note that the red and blue lines are superimposed. (*C*) Structural characterization of domains that fold cotranslationally and posttranslationally in *E. coli* cells that are dividing every 150 min at 37 °C. (*Upper*) Probability density function (PDF) vs. domain length. (*Lower*) Probability of different domain classifications in terms of mostly α (α), mostly β (β), or mixed α/β secondary structure. (*D*) As in *C*, except the data are from protein domains that fold cotranslationally with $\Delta L_m = 0$ and those that fold with ΔL_m values greater than 41 residues. The noncontiguous distribution for the $\Delta L_m > 41$ distribution arises from the small number of domains used in its construction ($n = 41$ data points).

$\tau_{F,bulk}$ and $\tau_{U,bulk}$ are functions of a domain's size and structural class in the DM model (19). Therefore, in Eqs. **3** and **4**, each domain will exhibit unique $\tau_{F,i}$ and $\tau_{U,i}$ values. The qualitative behavior of these equations is such that as the nascent chain increases in length, the folding time becomes smaller, approaching its bulk value at sufficiently long lengths. Specifically, when more than 28 residues connect the C-terminal residue of a domain to the P-site tRNA on an arrested ribosome, $\tau_{F,i} < \tau_{U,i}$ and the folded state of the domain is thermodynamically more stable than the unfolded state; around 28 residues in linker length, $\tau_{F,i} \approx \tau_{U,i}$ and the domain is at or near its midpoint of stability; below 28 residues, $\tau_{F,i} \gg \tau_{Ui}$ and, consequently, the folded state of the domain is highly unstable, with shorter nascent lengths leading to an exponential increase in the stability of the unfolded state relative to that of the folded state (Fig. 2*A*).

This behavior is consistent with a wide range of experimental results. A study of GFP found that it is capable of forming a stable native fold on the ribosome when between 22 and 31 residues connect it to the peptidyl-tRNA (23). A cross-linking study found that the midpoint of stability of an α-hairpin tertiary structure was between 24 and 30 residues in linker length (24). A single molecule pulling experiment found that the T4 lysozyme protein was folded at a linker length of 41 residues (25); shorter linker lengths, however, were not probed in that study. A similar upper bound was found in NMR (26) and proteolysis experiments (27), as well as by simulation studies (28). Thus, Eqs. **3** and **4** account for structural differences between domains that result in varied kinetics and exhibit realistic ribosome behavior.

Results

About One-Third of the *E. coli* Cytosolic Proteome Exhibits Cotranslational Folding That Primarily Involves Multidomain Proteins. We treat each SCOP-, CATH-, and DP-defined domain in our database as an autonomous cooperative folding unit that folds in a two-state manner with a negligible intermediate state population. For each cytosolic domain in Dataset S1, we calculated its cotranslational folding curve as a function of nascent chain length at the translation rates found in *E. coli* at 37 °C with a doubling time of 150 min (Fig. 2*B*). A protein exhibits cotranslational folding if at least one of its domains folds into its native structure with a probability of greater than 0.5 before synthesis of the full-length protein is complete. Under these conditions, we find that about 37% of *E. coli*'s cytosolic proteome (283 unique proteins in our database out of 758) exhibits cotranslational folding, whereas the remaining proteins fold posttranslationally, with none of the domains in these proteins folding during their synthesis.

Substantial differences are found by comparing the properties of proteins that fold cotranslationally and those that fold posttranslationally. The overwhelming majority (91%) of cotranslationally folding proteins are composed of multiple domains, whereas among posttranslational folders, single and multidomain proteins are almost equally represented (51% vs. 49%). In addition, the structural characteristics of those domains that cotranslationally fold are different from those of posttranslational folding domains. Cotranslationally folding domains are smaller in size on average (106 vs. 177 residues; Fig. 2*C*) and are more likely to be structurally classified as predominantly α-helical (51 ± 4% vs. 39 ± 2% for posttranslationally folding domains; Fig. 2*C*). On the other hand, posttranslationally folding domains tend to consist of a mixed α/β structure (50 ± 2% vs. 39 ± 3% for cotranslational folders; Fig. 2*C*). Thus, the structural characteristics of a domain are a factor in determining whether cotranslational folding occurs.

When the growth rate of *E. coli* increases, the rate of translation also increases for various proteins (29). To probe the effect of growth rate on our results, we used biochemical measurements of tRNA concentrations at 37 °C (11, 30) to calculate $\tau_{A,i}$ at different *E. coli* doubling times (Tables S1 and S2) for use in Eq. **1**. We find that the percentage of the proteome exhibiting cotranslational folding is insensitive to the change in growth rate, staying near 37% at all doubling times (Fig. S3).

The Majority of Cotranslational Folding Events in the *E. coli* Cytosol Exhibit Kinetic Effects That Can Substantially Delay Folding. Two cotranslational folding curves of a domain can be calculated from the chemical kinetic model described in *Methods*: the folding curve at the translation rates found in *E. coli* (Eq. **1**), where translation kinetics have the potential to decrease the probability of cotranslational domain folding (Figs. 1*B* and 2*B*), and the folding curve generated at an infinitely slow translation rate (Eq. **2**), where the thermodynamics of the RNC complex determine the cotranslationally folded population. The difference in the midpoint nascent chain length between these two curves, ΔL_m, provides a measure of the deviation from quasiequilibrium behavior exhibited by cotranslational folding in *E. coli* and an estimate of the number of nascent chain lengths at which kinetic effects are exhibited. Another way to interpret ΔL_m is that it represents the additional number of residues necessary to allow a domain to reach its native state due to the finite translation rates in *E. coli* (Fig. 1*C*). When $\Delta L_m = 0$, the folding curves superimpose (or are at least highly similar) and thermodynamic properties govern the extent of cotranslational folding at each nascent chain length. When $\Delta L_m > 0$, the curves deviate from one another and kinetic

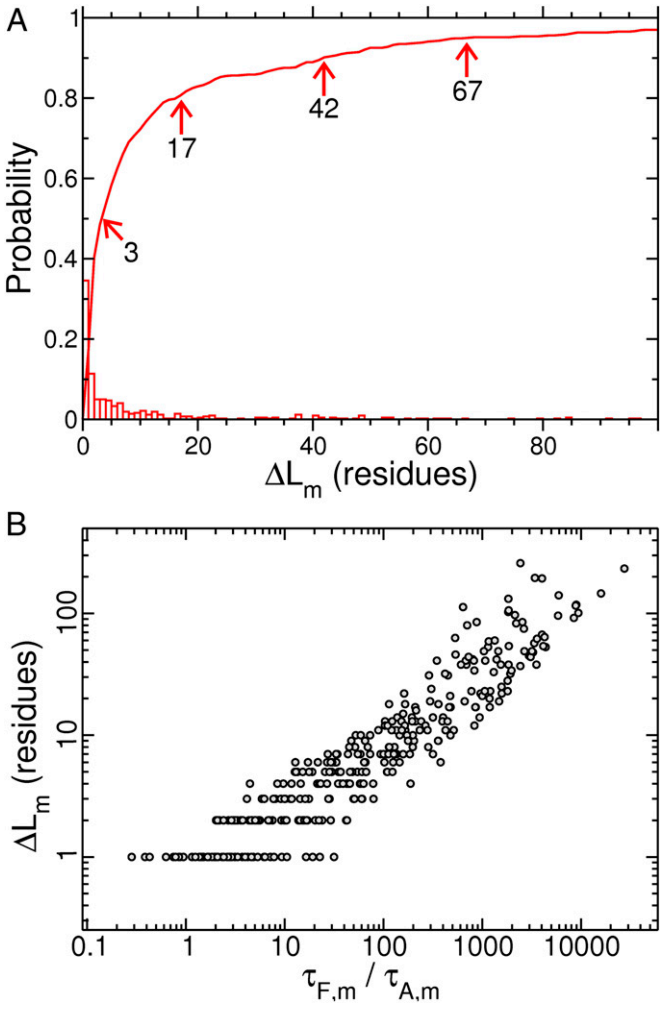

Fig. 3. Extent to which kinetic effects are exhibited during in vivo cotranslational folding and its correlation with the separation in time scales. (*A*) Probability distribution of ΔL_m values for domains that exhibit cotranslational folding in *E. coli* doubling every 150 min at 37 °C (n = 422). The cumulative distribution function (CDF) is shown as a solid red line. The arrows and numbers indicate (from left to right) the ΔL_m values at which the CDF equals 0.5, 0.8, 0.9, and 0.95, respectively. (*B*) ΔL_m for cotranslationally folding proteins as a function of the ratio of $\tau_{F,m}$ to $\tau_{A,m}$ at nascent chain length m at which the midpoint of domain folding stability occurs at an infinitely slow translation rate.

effects determine the folded domain population at one or more nascent chain lengths (Fig. 2*B*).

Of the 1,236 cytosolic domains in our database, 602 cotranslationally fold at infinitely slow translation rates and the remaining domains fold posttranslationally, as they have thermodynamically stable folded structures in free solution (Dataset S1). The reason the latter domains do not fold cotranslationally is that they correspond to either single domain proteins with no unstructured C terminus or to the C-terminal domain in multidomain proteins, which means they must be released from the ribosome to emerge fully from the exit tunnel and have the opportunity to fold. At the faster translation rates found in *E. coli* at 37 °C, 180 of the domains that fold cotranslationally on arrested ribosomes switch to posttranslational folding in vivo (Dataset S2). The folding of these domains takes longer than the synthesis of the complete protein, and their thermodynamically stable folded state is never reached during translation (Fig. S4). These 180 domains arise from 163 unique cytosolic proteins. Therefore, for 22% of the cytosolic proteome (163 of 758 unique proteins in our database), the finite rate of in vivo translation switches one or more of their domains from co- to posttranslational folding.

For the domains that switch from co- to posttranslational folding, ΔL_m is greater than the full length of the protein and an exact ΔL_m value cannot be calculated, but for the 422 domains that do exhibit cotranslational folding at in vivo translation rates, we plot the probability distribution of their ΔL_m values (Fig. 3*A*). The range of ΔL_m values spans from 0 to 259 residues, and the majority of cotranslationally folding domains exhibit delays of $\Delta L_m \geq 3$ residues. This finding indicates that for most cotranslationally folding domains, kinetic effects are exhibited at three or more nascent chain lengths. For a number of proteins, the delay in cotranslational folding can be much larger: 20% of them exhibit delays of $\Delta L_m \geq 17$ residues, 10% exhibit delays of $\Delta L_m \geq 42$ residues, and 5% exhibit delays of greater than 67 residues (Fig. 3*A*). Examples of cotranslational folding curves exhibiting various ΔL_m values are shown in Fig. 2*B*, and highly similar results are found at faster *E. coli* growth rates (Fig. 4*A*). Thus, the finite rate of synthesis in the *E. coli* cytosol can significantly reduce the cotranslationally folded population at a given nascent chain length and can substantially increase the nascent chain length at which a domain folds, switching the domain from co- to posttranslational folding in many cases.

Characterizing the structures of those domains whose cotranslational folding curve is governed solely by thermodynamics (i.e., $\Delta L_m = 0$ residues) vs. those exhibiting very significant kinetic effects (which we define as domains with $\Delta L_m \geq 42$ residues), we find the average lengths of the domains are, respectively, 69 and 155 residues (Fig. 2*D*). Domains that are classified as predominantly α-helical are significantly overrepresented in the thermodynamically governed group (80 ± 6% vs. 33 ± 4% in the kinetically governed group), whereas mixed α/β structure domains are more common in domains that exhibit kinetic effects (50 ± 6% vs. 16 ± 3% in the thermodynamically governed group). These results indicate that the cotranslational folding of large β-strand–rich domains is more likely to be governed by translation kinetics than that of small α-helical domains.

We then asked whether gene expression levels in *E. coli* alter these results. To test this possibility, we weighted the data in Fig. 3*A* using gene expression levels estimated from mRNA copy numbers (29). We used mRNA copy number data because they provide substantially better coverage of the proteins in our database than do protein abundance data. Although there is a weak correspondence between gene expression and protein abundance in single cells, there is a correlation ($r = 0.77$) between these quantities on a population basis (31). We find the distribution of ΔL_m is largely unchanged when gene expression levels are accounted for appropriately (Fig. 4*B*). Another possibility we

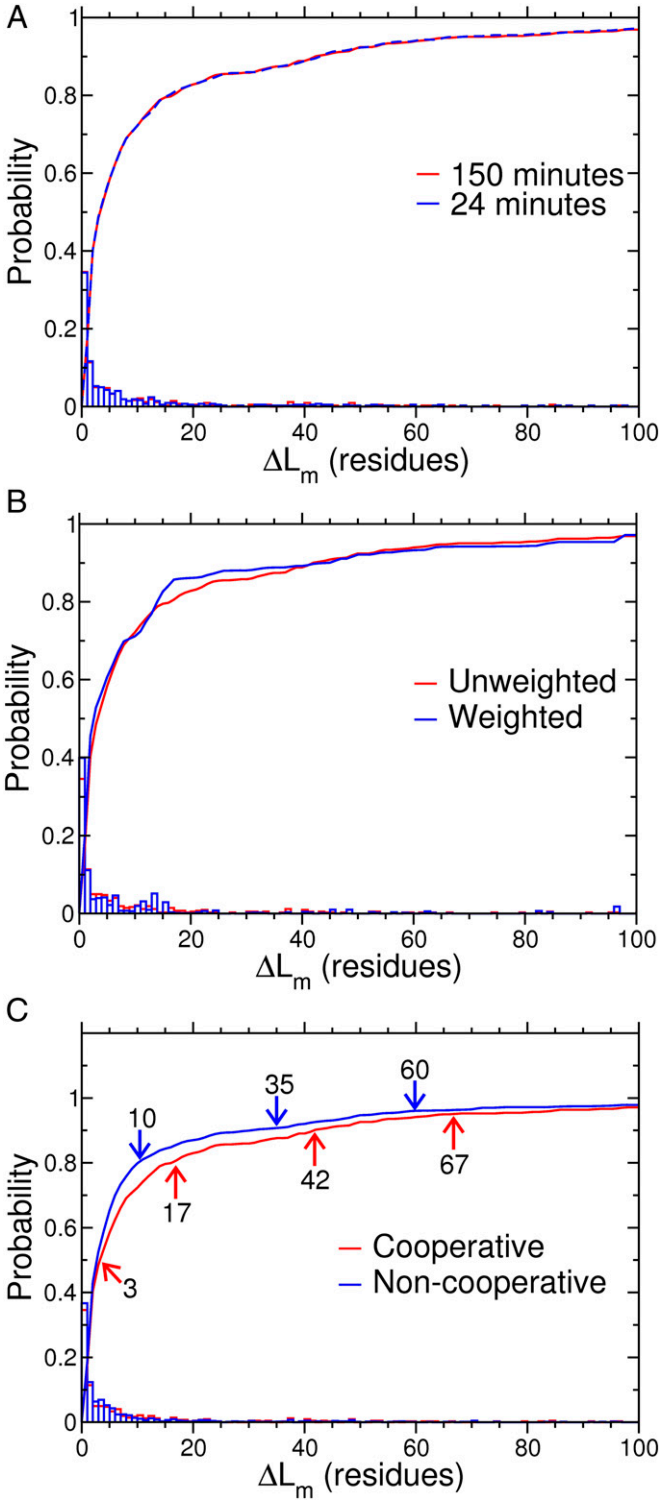

Fig. 4. Probability distribution of ΔL_m values (histograms) and cumulative distributions (solid lines) at two different *E. coli* growth rates listed in the legend as the doubling time (*A*), weighted by the protein expression level data denoted by BLT_WT in Dataset S1 (*B*), and for proteins that exhibit cooperative and noncooperative domain folding (*C*). In *C*, the ΔL_m values are shown for probabilities corresponding to 0.8, 0.9, and 0.95 for the noncooperative dataset, whereas the cooperative dataset includes an additional ΔL_m value reported at a probability of 0.5.

explored was that ΔL_m correlates with the expression level, but we observe no such correlation in our data (Fig. S5). These results therefore demonstrate that the extent of kinetic and thermodynamic effects is uncorrelated with gene expression levels.

Kinetic Effects Are Exhibited Even When Noncooperative Subdomain Folding Occurs. In the analysis above, we assumed the SCOP, CATH, and DP domain definitions represent autonomous, cooperative folding units. Under this assumption, these domains are only capable of forming a stable, folded tertiary structure once all the residues comprising the domain have emerged from the exit tunnel and are sterically permitted to come into contact, which can include tertiary structure formation in the last 2 nm of the exit tunnel (24, 32). Recent experimental results, however, call this assumption into question because the CATH-defined NBD1 domain of human CFTR protein was found to exhibit stable subdomain folding during synthesis on the ribosome (33). Thus, noncooperative folding can occur on the ribosome, with a stable native tertiary structure forming before synthesis of the entire domain is complete. This phenomenon has the potential to alter the conclusions of this study because subdomains, being smaller, will, on average, fold faster than full-size domains (34). This will change the interplay of the time scales as expressed in Eqs. **1** and **2**.

To estimate the robustness of our conclusion of widespread kinetic effects against the potential for subdomain folding, we divided each domain in Dataset S1 that was greater than 150 residues in length into two halves and defined the resulting subdomains as autonomous folding units. This procedure represents an extreme limit in which subdomain folding is widespread among proteins in the cytosolic proteome. As a consequence, the estimated folding and unfolding times on the ribosome are altered (Eqs. **3** and **4**). With this new set of domain definitions, we find that the distribution of ΔL_m is shifted slightly toward smaller values (Fig. 4*C*). The overall trend, however, is highly similar to that in Fig. 3*A*, and our conclusion that the majority of proteins in the cytosolic proteome exhibit kinetic effects is unaltered by the occurrence of subdomain folding during translation.

Separation of Folding and Amino Acid Addition Time Scales Determines the Extent to Which Kinetic Effects Are Exhibited. A key thermodynamic concept is that the ratio of time scales during an irreversible process determines whether it is a quasiequilibrium process (in which thermodynamics dominate) or a nonequilibrium process (in which kinetics dominate; Fig. 1*B*). Therefore, we hypothesized that the deviation of a given cotranslational folding curve in *E. coli* from quasiequilibrium (the arrested ribosome case) should be a function of the ratio of the mean folding time to the time of amino acid addition at its midpoint nascent chain length on an arrested ribosome. We plotted ΔL_m as a function of this ratio and observe a trend in which ΔL_m increases with increasing $\frac{\tau_{F,m}}{\tau_{A,m}}$ (Fig. 3*B*). When $\frac{\tau_{F,m}}{\tau_{A,m}} \approx 1$, we observe ΔL_m is close to 0, and at $\frac{\tau_{F,m}}{\tau_{A,m}} > 1$, we observe $\Delta L_m \gg 0$, with increased $\frac{\tau_{F,m}}{\tau_{A,m}}$ values correlating with increased ΔL_m values. Thus, the ratio of these two time scales is a primary determinant of the extent to which kinetic effects are exhibited during cotranslational folding.

Synonymous Codon Mutations Can Significantly Decrease Kinetic Effects for a Small Percentage of Cotranslational Folding Domains. Synonymous codon mutations alter the mRNA sequence but leave the translated protein sequence unaltered. Synonymous codons can also change the translation rate, which can itself directly affect the extent of cotranslational folding (7, 8, 12). Thus, altering the mRNA sequence such that the slowest translating synonymous mutation is present at each codon has the potential to minimize kinetic effects by decreasing the $\frac{\tau_{F,i}}{\tau_{A,i+1}}$ ratio.

To examine the extent to which kinetic effects on cotranslational folding can be reduced by synonymous codons, we replaced each codon in each mRNA sequence *in silico* with its slowest translating synonymous codon. We then recalculated the cotranslational folding curves for each domain and calculated the change in ΔL_m (denoted $\Delta\Delta L_m$) between the WT and mutant mRNA transcripts. The majority of domains exhibit a decrease in ΔL_m of no more than 2 residues due to the synonymous codon-induced slowdown in translation (Fig. 5), but larger decreases are observed for a minority of proteins, with 5% of the cytosolic domains exhibiting a decrease in $\Delta\Delta L_m$ of greater than 43 residues. Thus, for the vast majority of cytosolic proteins, synonymous mutations have little effect on the separation of folding curves between quasiequilibrium and finite translation rates; for a minority, however, synonymous mutations can significantly reduce the influence of translation rate on cotranslational folding, and thereby increase the folded population at shorter nascent chain lengths.

Robustness of the Results Against Changes in Cotranslational Folding Time Scale Estimates. The cotranslational folding time of T4-lysozyme on an arrested ribosome has recently been measured at two different nascent chain lengths using laser optical tweezers (LOT) (25). It was found that at a linker length of 41 residues, the value of τ_F, under a constant tension of 4 pN, is approximately two orders of magnitude larger than in bulk solution, whereas at a linker length of 60 residues, τ_F is within one order of magnitude of its bulk value. This trend toward the bulk folding time at longer nascent chain lengths is consistent with our model (Fig. 2*A*). However, at these experimental linker lengths, $\tau_{F,i}$ is close to its bulk value in our model. We therefore tested whether or not our conclusions are altered by introducing the longer range effect observed in the LOT experiments by reparameterizing the scaling relationship in Eq. **3** to reproduce the trend seen in the experiments. In this model, $\tau_{F,i}$ is two orders and one order of magnitude larger than its bulk folding time at linker lengths of 41 and 61 residues, respectively (Fig. S6*A*). Using this alternative scaling relationship, we recalculated the cotranslational folding properties of the cytosolic proteins in Dataset S1.

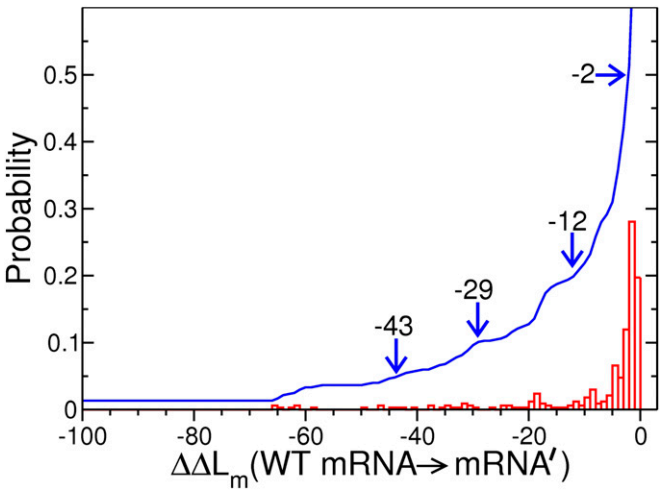

Fig. 5. Slow translating synonymous codon mutations and their affect on the deviation of cotranslational folding from quasiequilibrium. The probability distribution of the change in ΔL_m values for cotranslational folding domains on converting each codon in the WT mRNA transcripts to its corresponding slowest translating synonymous codon is shown. Cotranslational folders that have $\Delta L_m = 0$ for the WT mRNA are not included in this analysis because their $\Delta\Delta L_m$ could never be anything other than 0. The CDF is shown in blue. The arrows and numbers indicate (from left to right) the ΔL_m values at which the CDF equals 0.05, 0.10, 0.20, and 0.5, respectively.

We find our results are largely unaltered by this change in $\tau_{F,i}$ behavior. In this model, 37% of the cytosolic proteome is predicted to fold cotranslationally at in vivo translation rates; in addition, 22% of the cytosolic proteins contain at least one domain that switches from co- to posttranslational folding due to in vivo translation rates. Kinetic effects are still observed in a majority of the proteome (Fig. S6*B*), and the cotranslational folding of α-helical domains remains much more likely to be governed by RNC thermodynamics (Fig. S6*C*). Thus, our conclusions are robust to changes and uncertainties in the ribosome folding and unfolding time scales.

Discussion

The systems approach presented in this paper (Fig. 1*D*) offers several important insights into proteome-wide protein folding in the cellular context. With it, we estimate that about one-third of the cytosolic *E. coli* proteome exhibits cotranslational folding (Fig. S3), and we observe that these domains tend to be small and α-helical, whereas posttranslational folding domains tend to be larger and contain a higher β-strand content (Fig. 2*C*). Further, we find that in *E. coli*, a majority of cotranslationally folding domains exhibit a deviation from their folding curve on an arrested ribosome at one or more nascent chain lengths (Figs. 2B and 3A). This deviation is a manifestation of the nonequilibrium cotranslational folding process that these domains experience during continuous translation. It demonstrates that in vivo translation kinetics, rather than RNC thermodynamics, can determine the folded population of protein domains at one or more nascent chain lengths. As a consequence, the nascent chain length at which these domains reach their stable folded state in *E. coli* can be much greater than that expected under quasiequilibrium (i.e., arrested ribosome) conditions. At the nascent chain lengths that intervene between the midpoints of the quasiequilibrium and nonequilibrium folding curves (Fig. 2*B*), the folded state is thermodynamically stable but is not populated to the extent that would be expected based on native state thermodynamic stability alone, leading to an excess population of the unfolded state. Thus, the finite rates of translation in *E. coli* reduce cotranslational folding, delay it as a function of nascent chain length, and enrich the population of unfolded nascent chains for a considerable fraction of the cytosolic proteome.

For example, for the domain analyzed in Fig. 2*B* (*Lower Right*), the midpoint of stability is shifted by 256 residues as a result of the finite translation rate in *E. coli*. At a nascent chain length of 250 residues, the folding probability of this domain is reduced from ≈1.0 on an arrested ribosome to 0.18 on a ribosome translating at in vivo rates, and, conversely, its unfolded state probability is increased from 0 to 0.82. Thus, the predominantly populated unfolded state at this nascent chain length is, in fact, metastable. Remarkably, we estimate that 22% of cytosolic proteins contain domains that would cotranslationally fold if translation were slow enough but do not do so in *E. coli* under exponential growth conditions in rich medium (Fig. S4 and Dataset S2). The reason is that the rate of translation in *E. coli* is fast enough to produce a ΔL_m value for these domains that is greater than that of the full-length protein. Consequently, these domains switch from co- to posttranslational folding, demonstrating that kinetic effects can substantially delay cotranslational folding in vivo.

There are few in vivo experimental data to compare with our predictions of the proteins that fold cotranslationally (Dataset S2). One *E. coli* protein for which convincing evidence exists is the five-domain β-galactosidase protein (3), where the concomitant appearance of the full-length protein and enzymatic activity suggests cotranslational folding is occurring. Our model predicts that this protein exhibits cotranslational folding at in vivo translation rates, with domains 2 and 4 folding before the full-length protein is released from the ribosome and domains 1, 3, and 5 folding posttranslationally. This prediction is consistent with the experimental conclusion that β-galactosidase folds cotranslationally because the current experimental data do not eliminate the possibility that some β-galactosidase domains may not fold cotranslationally.

The separation in time scales of domain folding relative to amino acid addition determines whether cotranslational folding is a quasiequilibrium or a nonequilibrium process and, consequently, whether thermodynamics or kinetics govern the cotranslationally folded populations. This provides an explanation for the correlation between ΔL_m (a measure of the deviation from quasiequilibrium) and the ratio of the τ_F-to-τ_A time scales (Fig. 3*B*). At the molecular level, when $\frac{\tau_{F,m}}{\tau_{A,m}} \leq 1$, the domain has sufficient time to equilibrate and reach its global free energy minimum at this nascent chain length. Thus, thermodynamic stability governs the cotranslationally folded population of this domain, and the process is quasiequilibrium in nature. When $\frac{\tau_{F,m}}{\tau_{A,m}} \gg 1$, a domain does not have sufficient time to equilibrate at each nascent chain length because there is a smaller time window for the domain to relax to its global free energy minimum. Under these conditions, kinetic effects suppress the cotranslationally folded population below its thermodynamic value, leading to an enrichment of metastable, ribosome-bound unfolded states. Thus, it is the separation of these time scales that determines the extent to which thermodynamics or kinetics govern cotranslational domain folding.

We have also found that large β-strand–rich domains are more likely to exhibit kinetic effects than small α-helical domains (Fig. 2*D*). It has been established previously that domain size is an important factor determining the magnitude of $\tau_{F,bulk}$ (34); that is, larger proteins tend to take longer to fold because they have a larger number of residues to incorporate into an ordered state. An additional factor that influences $\tau_{F,bulk}$ is the topological complexity of the native structure, because more complex structures with larger sequence separation between native contacts tend to fold more slowly than domains enriched in local contacts (20). Thus, small α-helical domains will typically have smaller $\frac{\tau_{F,m}}{\tau_{A,m}}$ ratios than large β-strand–rich domains, which means the latter are more likely to exhibit kinetic effects during cotranslational folding than the former. We note that we do not see any codon bias for the different classes of domains, as indicated by a similar average translation rate for the codons following the different domain classes (Fig. S7). This suggests domain topology and domain size, rather than codon bias, is the predominant factor influencing the ratio $\frac{\tau_{F,m}}{\tau_{A,m}}$.

These findings also suggest one reason why it is often difficult to express eukaryotic proteins in *E. coli* (35). Eukaryotic proteins tend to be larger on average than prokaryotic proteins (36); thus, they will normally have larger $\tau_{F,bulk}$ values (19, 34). Furthermore, the translation rate in *E. coli* is up to sevenfold faster than in eukaryotes [10–22 amino acids per second vs. 3–6 amino acids per second (29)], indicating that $\tau_{A,m}$ is smaller in *E. coli*. Thus, for a given eukaryotic protein, its $\frac{\tau_{F,m}}{\tau_{A,m}}$ ratio will be larger when it is expressed in *E. coli* than in a eukaryotic cell. This larger separation of time scales suggests that eukaryotic proteins are more likely to exhibit suppressed cotranslational folding in *E. coli*, which can lead to an increase in the probability that they misfold, aggregate, or are degraded, all of which will decrease their soluble folding yield. These findings also suggest that slowing down the translation rate in *E. coli* will increase successful heterologous expression, a prediction that is consistent with results from experiments in which τ_A was modulated using streptomycin-sensitive ribosomes (37).

Trigger factor (TF) and DnaK are two cotranslationally acting *E. coli* chaperones that preferentially bind to the nascent chains of larger proteins compared with smaller ones (38, 39). In *E. coli* cells in which DnaK and TF were deleted, it was found that larger proteins were much more likely to aggregate than smaller proteins

(38–40). Pulse-chase radio labeling of nascent chains in *E. coli* cells containing nonfunctional chaperone GroEL has demonstrated that such aggregation involves only newly synthesized proteins and not preexisting, folded proteins (41). Together, these data indicate that the nascent chains of large proteins are more likely to interact with DnaK and TF, and to be aggregation-prone. Our results suggest this bias for larger proteins may arise from the increased likelihood that larger domains will remain unfolded on the ribosome and fold posttranslationally. Such posttranslational folding increases the affinity of these chaperones for the nascent chain (42) and affords them more time to interact with the unfolded nascent chain.

A key challenge in systems biology studies is accurately estimating the values for a large number of kinetic parameters. Here, we estimated the various codon translation times using experimental information on in vivo tRNA concentrations (*Methods*). An alternative approach could be to use ribosome profiling data (43), whose measured distribution of ribosome positions on a transcript is a function of $\tau_{A,i}$.

A common experimental procedure to obtain high-resolution cotranslational folding data is to use arrested ribosomes (25, 44). Our results suggest that the folding mechanism in this situation can sometimes differ significantly from that during continuous translation. Transition state ensemble (TSE) properties of the folding domain are influenced by the relative thermodynamic stability of the folded and unfolded states, with the TSE resembling more closely whichever of these two states is less stable [i.e., the Hammond effect (45)]. Because kinetic effects delay folding, and hence cause the unfolded state to be less stable than the folded state once folding does occur (Fig. 2*B*), it follows that the TSE will shift toward the unfolded state in nascent chains undergoing continuous translation compared with the TSE on an arrested ribosome. Thus, care must be taken when extrapolating protein folding results on arrested ribosomes to the situation on actively translating ribosomes.

The widespread presence of kinetic effects in the cotranslational folding of the proteome suggests that translation-related time scales can be manipulated for synthetic biology, biotechnology, and cellular engineering purposes to boost (or suppress) cotranslational folding of individual proteins. This possibility has been demonstrated in experiments in vitro and in molecular simulations in which insertion of slow-translating codons at particular points along an mRNA molecule altered the folded population (7, 8). Although synonymous mutations alone may have a significant impact on the cotranslational folding curves of some proteins, we have found that they do not significantly reduce kinetic effects for the majority of the proteome (Fig. 5). Therefore, to alter these individual $\tau_{A,i}$ values to an extent that they affect more of the proteome, higher order processes will need to be considered. These manipulations could include the alteration of mRNA secondary structure (46), the variation in the concentration of charged tRNA molecules in vivo (47), or the introduction of anti–Shine–Dalgarno sequences (48, 49).

In addition to informing strategies for synthetic manipulation of the protein-coding sequences, our results may have implications for understanding sequence changes during evolution. Translational fidelity and its importance to protein structural stability has been proposed as an explanation for the strong relationship between mRNA expression levels and evolutionary rates (50, 51). Mistranslation events, in which an incorrect amino acid is incorporated into the nascent chain, could affect the rates of protein folding and unfolding, and thereby alter the likelihood of co- and posttranslational folding events. The fitness consequences of this could therefore link the cotranslational folding process to the evolution of protein and mRNA sequences.

In conclusion, our approach offers a theoretical framework that can be applied to analyze other cells and tissues in both prokaryotes and eukaryotes, and it has the potential to be used to redesign entire transcriptomes rationally.

ACKNOWLEDGMENTS. P.C. thanks the US-UK Fulbright Commission and St. John's College, University of Cambridge, for postgraduate funding. E.P.O. thanks David De Sancho for valuable conversations regarding domain definitions and implementation of the DM model (19). We thank an anonymous reviewer for helpful comments. This work was supported by a National Science Foundation postdoctoral grant (to E.P.O.) and the Engineering and Physical Sciences Research Council (E.P.O., M.V., and C.M.D.).

1. Nicola AV, Chen W, Helenius A (1999) Co-translational folding of an alphavirus capsid protein in the cytosol of living cells. *Nat Cell Biol* 1(6):341–345.
2. Netzer WJ, Hartl FU (1997) Recombination of protein domains facilitated by co-translational folding in eukaryotes. *Nature* 388(6640):343–349.
3. Agashe VR, et al. (2004) Function of trigger factor and DnaK in multidomain protein folding: Increase in yield at the expense of folding speed. *Cell* 117(2):199–209.
4. Elcock AH (2006) Molecular simulations of cotranslational protein folding: Fragment stabilities, folding cooperativity, and trapping in the ribosome. *PLOS Comput Biol* 2(7):e98.
5. Komar AA, Lesnik T, Reiss C (1999) Synonymous codon substitutions affect ribosome traffic and protein folding during in vitro translation. *FEBS Lett* 462(3):387–391.
6. Clarke TF, Clark PL (2010) Increased incidence of rare codon clusters at 5′ and 3′ gene termini: Implications for function. *BMC Genomics* 11:118.
7. O'Brien EP, Vendruscolo M, Dobson CM (2012) Prediction of variable translation rate effects on cotranslational protein folding. *Nat Commun* 3:868.
8. Zhang G, Hubalewska M, Ignatova Z (2009) Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. *Nat Struct Mol Biol* 16(3):274–280.
9. Clark PL, King J (2001) A newly synthesized, ribosome-bound polypeptide chain adopts conformations dissimilar from early in vitro refolding intermediates. *J Biol Chem* 276(27): 25411–25420.
10. Hartl FU, Hayer-Hartl M (2009) Converging concepts of protein folding in vitro and in vivo. *Nat Struct Mol Biol* 16(6):574–581.
11. Dong H, Nilsson L, Kurland CG (1996) Co-variation of tRNA abundance and codon usage in Escherichia coli at different growth rates. *J Mol Biol* 260(5):649–663.
12. Fluitt A, Pienaar E, Viljoen H (2007) Ribosome kinetics and aa-tRNA competition determine rate and fidelity of peptide synthesis. *Comput Biol Chem* 31(5-6):335–346.
13. Sijbrandi R, et al. (2003) Signal recognition particle (SRP)-mediated targeting and Sec-dependent translocation of an extracellular Escherichia coli protein. *J Biol Chem* 278(7): 4654–4659.
14. Berman HM, et al. (2000) The Protein Data Bank. *Nucleic Acids Res* 28(1):235–242.
15. Andreeva A, et al. (2008) Data growth and its impact on the SCOP database: New developments. *Nucleic Acids Res* 36(Database issue):D419–D425.
16. Cuff AL, et al. (2011) Extending CATH: Increasing coverage of the protein structure universe and linking structure with function. *Nucleic Acids Res* 39(Database issue):D420–D426.
17. Xu Y, Xu D, Gabow HN (2000) Protein domain decomposition using a graph-theoretic approach. *Bioinformatics* 16(12):1091–1104.
18. Yu NY, et al. (2010) PSORTb 3.0: Improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics* 26(13):1608–1615.
19. De Sancho D, Muñoz V (2011) Integrated prediction of protein folding and unfolding rates from only size and structural class. *Phys Chem Chem Phys* 13(38):17030–17043.
20. Ivankov DN, et al. (2003) Contact order revisited: Influence of protein size on the folding rate. *Protein Sci* 12(9):2057–2062.
21. Bronikowski AM, Bennett AF, Lenski RE (2001) Evolutionary adaptation to temperature. VIII. Effects of temperature on growth rate in natural isolates of Escherichia coli and Salmonella enterica from different thermal environments. *Evolution* 55(1):33–40.
22. De Sancho D, Doshi U, Muñoz V (2009) Protein folding rates and stability: How much is there beyond size? *J Am Chem Soc* 131(6):2074–2075.
23. Kelkar DA, Khushoo A, Yang ZY, Skach WR (2012) Kinetic analysis of ribosome-bound fluorescent proteins reveals an early, stable, cotranslational folding intermediate. *J Biol Chem* 287(4):2568–2578.
24. Kosolapov A, Deutsch C (2009) Tertiary interactions within the ribosomal exit tunnel. *Nat Struct Mol Biol* 16(4):405–411.
25. Kaiser CM, Goldman DH, Chodera JD, Tinoco I, Jr., Bustamante C (2011) The ribosome modulates nascent protein folding. *Science* 334(6063):1723–1727.
26. Eichmann C, Preissler S, Riek R, Deuerling E (2010) Cotranslational structure acquisition of nascent polypeptides monitored by NMR spectroscopy. *Proc Natl Acad Sci USA* 107(20):9111–9116.
27. Hoffmann A, et al. (2012) Concerted action of the ribosome and the associated chaperone trigger factor confines nascent polypeptide folding. *Mol Cell* 48(1):63–74.
28. O'Brien EP, Christodoulou J, Vendruscolo M, Dobson CM (2011) New scenarios of protein folding can occur on the ribosome. *J Am Chem Soc* 133(3):513–526.
29. Liang ST, Xu YC, Dennis P, Bremer H (2000) mRNA composition and control of bacterial gene expression. *J Bacteriol* 182(11):3037–3044.
30. Robinson AC, Donachie WD (1987) Cell division: Parameter values and the process. *Escherichia coli and Salmonella typhimurium: Cellular and Molecular Biology*, eds Ingraham J, et al. (American Society for Microbiology, Washington, DC), Vol 2, pp 1578–1593.

31. Taniguchi Y, et al. (2010) Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science* 329(5991):533–538.
32. O'Brien EP, Hsu STD, Christodoulou J, Vendruscolo M, Dobson CM (2010) Transient tertiary structure formation within the ribosome exit port. *J Am Chem Soc* 132(47): 16928–16937.
33. Khushoo A, Yang Z, Johnson AE, Skach WR (2011) Ligand-driven vectorial folding of ribosome-bound human CFTR NBD1. *Mol Cell* 41(6):682–692.
34. Thirumalai D (1995) From minimal models to real proteins—Time scales for protein-folding kinetics. *J Phys I France* 5(11):1457–1467.
35. Baneyx F, Mujacic M (2004) Recombinant protein folding and misfolding in Escherichia coli. *Nat Biotechnol* 22(11):1399–1408.
36. Brocchieri L, Karlin S (2005) Protein length in eukaryotic and prokaryotic proteomes. *Nucleic Acids Res* 33(10):3390–3400.
37. Siller E, DeZwaan DC, Anderson JF, Freeman BC, Barral JM (2010) Slowing bacterial translation speed enhances eukaryotic protein folding efficiency. *J Mol Biol* 396(5): 1310–1318.
38. Deuerling E, et al. (2003) Trigger Factor and DnaK possess overlapping substrate pools and binding specificities. *Mol Microbiol* 47(5):1317–1328.
39. Calloni G, et al. (2012) DnaK functions as a central hub in the E. coli chaperone network. *Cell Rep* 1(3):251–264.
40. Mogk A, et al. (1999) Identification of thermolabile Escherichia coli proteins: Prevention and reversion of aggregation by DnaK and ClpB. *EMBO J* 18(24):6934–6949.
41. Chapman E, et al. (2006) Global aggregation of newly translated proteins in an Escherichia coli strain deficient of the chaperonin GroEL. *Proc Natl Acad Sci USA* 103(43):15800–15805.
42. Rutkowska A, et al. (2008) Dynamics of trigger factor interaction with translating ribosomes. *J Biol Chem* 283(7):4124–4132.
43. Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* 324(5924):218–223.
44. Hsu STD, et al. (2007) Structure and dynamics of a ribosome-bound nascent chain by NMR spectroscopy. *Proc Natl Acad Sci USA* 104(42):16516–16521.
45. Taskent H, Cho JH, Raleigh DP (2008) Temperature-dependent Hammond behavior in a protein-folding reaction: Analysis of transition-state movement and ground-state effects. *J Mol Biol* 378(3):699–706.
46. Qu X, et al. (2011) The ribosome uses two active mechanisms to unwind messenger RNA during translation. *Nature* 475(7354):118–121.
47. Zhang G, et al. (2010) Global and local depletion of ternary complex limits translational elongation. *Nucleic Acids Res* 38(14):4778–4787.
48. Li GW, Oh E, Weissman JS (2012) The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature* 484(7395):538–541.
49. Takahashi S, Tsuji K, Ueda T, Okahata Y (2012) Traveling Time of a translating ribosome along messenger RNA monitored directly on a quartz crystal microbalance. *J Am Chem Soc* 134(15):6793–6800.
50. Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH (2005) Why highly expressed proteins evolve slowly. *Proc Natl Acad Sci USA* 102(40):14338–14343.
51. Drummond DA, Wilke CO (2008) Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* 134(2):341–352.

This database is a collection of sequence, expression, and PDB fidelity information for *E. coli* strain K12 MG1655 protein domains.

Each entry of the database has two hierarchical levels. The top level is the protein level, denoted by ##Unibegin, which marks the beginning of an entry and refers to one entire protein (denoted by its Uniprot ID).

The second level is the domain level, denoted by #Dombegin, which marks the beginning of information pertinent to a particular domain of the protein. Each protein has one or more domains.

The protein level of the entry contains the following information:

ND: the number of domains in the protein.
mRNA: the mRNA sequence for the corresponding gene, obtained from GenBank.
TAA: the translated protein sequence for the mRNA.
SCL: the predicted subcellular localization, obtained from PSORTb.
RPA_M: the relative protein abundance obtained from the Marcotte dataset.
RPA_F: the relative protein abundance obtained from the Frischman dataset.
RPA_X: the relative protein abundance obtained from the Xie dataset.
RNE_C: the relative mRNA expression obtained from the Church dataset.
RNE_XM: the relative mean mRNA expression obtained from the Xie dataset.
RNE_XL: the relative lifetime mRNA expression obtained from the Xie dataset.
SEC: value can be 'yes' or 'no' obtained from SignalP. If 'yes' then the protein is predicted to be secreted via the Sec pathway.
TAT: value can be 'yes' or 'no' obtained from TatP (based on D-score). If 'yes' then the protein is predicted to be secreted via the Tat pathway.
BLT_WT: the relative mRNA expression in wild-type conditions obtained from the Blattner dataset.
BLT-HS: the relative mRNA expression in the heat shock condition obtained from the Blattner dataset.

The domain level of the entry contains the following information on domains within the protein. Each domain is labelled Di, where *i* is the number of the domain in the protein.

Di_PDBC: PDB ID and chain, listed in the format ID-Chain.
Di_DB: database ('SCOP', 'CATH', or 'DP') from which the domain information was obtained.
Di_DRC: codon range for the domain, with position numbering based on the translated mRNA sequence. Format is Start/End.
Di_DRO: residue range of domain as originally reported in the database (SCOP or CATH) from which it was derived. Format is PDB chain ID:Start/End.
Di_AA: amino acid sequence of the domain, based on its codon range in the translated mRNA sequence.
Di_NR: the number of residues comprising the translated domain.
Di_MR: the number of missing residues in the PDB domain sequence with respect to the translated protein sequence.
Di_MM: the number of mismatched residues in the PDB domain sequence when BLASTED against the translated protein sequence.
Di_IR: the number of inserted residues in the PDB domain sequence with respect to the translated protein sequence.

Di_MC: value can be 'n' or '*'. If '*', then BLAST search was used to align segment yielded multiple alignments, of which the first was used for further analysis.

Di_NO: the number of residues taking part in ordered structure (*ie*., residues comprising either α-helical or β-strand structures in the PDB domain).

Di_NA: the number of residues taking part in α-helical structure.

Di_NB: the number of residues taking part in β-strand structure.

Di_FA: the fraction of ordered residues that are taking part in α-helical structure.

Di_FB: the fraction of ordered residues that are taking part in β-strand structure.

Di_CLASS: domain classification that can have values of '*alpha'*, '*alpha-beta*' or '*beta*.' Domains that have **Di_FA** values of greater than 0.70 are classified as predominately alpha-helical domains ('*alpha'*); domains with **Di_FB** values of greater than 0.70 are classified as predominately β-strand domains ('*beta*'); otherwise, the domains are '*alpha-beta*'.

Di_KF: the bulk folding rate of the domain when free in solution computed from the De Sancho-Munoz model, in units of s^{-1}.

Di_KU: the bulk unfolding rate of the domain when free in solution computed from the De Sancho-Munoz model, in units of s^{-1}.

Di_DG: the bulk stability of the folded domain relative to the unfolded state at 310 K calculated as –RT*ln*[Di_KF / Di_KU], in units of *kcal/mol*.

Supporting Information

Ciryam et al. 10.1073/pnas.1213624110

SI Methods

Construction of the *Escherichia coli* Database. The *Escherichia coli* database that was constructed in this study is shown in Dataset S1.

Collecting Escherichia coli K12 MG1655 Protein Data Bank files. All Protein Data Bank (PDB) files containing proteins from the *Escherichia coli* species and whose structures were determined using solution-state or solid-state NMR or X-ray diffraction, with a resolution ≤3.0 Å, were downloaded from www.rcsb.org. This yielded 11,036 PDB chains from 4,352 PDB files. For these PDBs, we identified their corresponding Uniprot identification numbers (IDs) when available from the RCSB. If there were multiple Uniprot IDs associated with a given PDB, we assumed these were fusion proteins and set them aside to be handled separately. Only fusion proteins constituted exclusively of *E. coli* proteins were considered, and these were divided into different entries. If the Uniprot ID was not available directly from the RCSB, we converted other available IDs (PDB, GenBank, or Norine) to a Uniprot ID. If a PDB mapped to multiple Uniprot IDs in which only one corresponded to *E. coli*, we used the Uniprot ID corresponding to *E. coli*. If the PDB mapped to multiple *E. coli* Uniprot IDs, we determined to which chain each ID corresponded. For those PDBs for which we were not able to obtain a Uniprot ID directly or by conversion using the summary file for the PDB search, we attempted to extract IDs from the PDB directly by checking the DBREF field in the PDB header.

We restricted this study to the *E. coli* K12 strain MG1655 (which corresponds to Uniprot IDs ending in _ECOLI and from the Uniprot Complete Proteome Set). Our database (Dataset S1) initially included a mix of K12 MG1655 proteins, non-K12 MG1655 *E. coli* proteins, and non-*E. coli* proteins. We deleted non-*E. coli* proteins from our database. For non-K12 MG1655 proteins, we used the BLASTp function of BLAST 2.2.25 (default settings) to identify homologous K12 MG1655 proteins. We included in our database (with the K12 MG1655 nomenclature) those non-K12 MG1655 proteins that had ≥98% sequence identity with a protein in K12 MG1655. If more than one K12 MG1655 protein was thus identified, we selected the one with the greatest sequence identity. After these conversions, we had 8,942 PDB chains (corresponding to 3,696 PDBs) remaining, representing 1,215 unique Uniprot IDs.

In constructing our final database of domains (see below), we determined that 15 Uniprot IDs in our database (CRCA_ECOLI, GST_ECOLI, GUDH_ECOLI, OXAA_ECOLI, RP5M_ECOLI, SUFI_ECOLI, TESC_ECOLI, THD1_ECOLI, YBDB_ECOLI, YEBR_ECOLI, YHIQ_ECOLI, YJGF_ECOLI, YJJX_ECOLI, FRUR_ECOLI, and MVIM_ECOLI) were deprecated. They had been replaced in Uniprot by, respectively, PAGP_ECOLI, GSTA_ECOLI, GUDD_ECOLI, YIDC_ECOLI, HPF_ECOLI, FTSP_ECOLI, FADM_ECOLI, ILVA_ECOLI, ENTH_ECOLI, MSRC_ECOLI, RSMJ_ECOLI, RIDA_ECOLI, NCPP_ECOLI, CRA_ECOLI, and YCEM_ECOLI. The translated mRNA sequences in our database corresponded exactly to the sequences for these updated Uniprot IDs, with the exception of RSMJ_ECOLI, CRA_ECOLI, and YCEM_ECOLI. In these cases, the protein sequence listed in Uniprot was identical for the old and updated IDs, but this sequence differed from the translated mRNA of the corresponding ID of Blattner and colleagues (1) at the first residue (in all three cases, the first position was listed as methionine in Uniprot but as valine when the mRNA sequence in the GenBank was translated *in silico*). Nevertheless, we replaced all 15 deprecated terms with their updated terms, using the GenBank-translated sequences for further analysis.

Extracting sequences from PDB files. We replaced modified or nonstandard amino acids with their standard equivalents as available. Two common modifications are methylated lysine and selenomethionine. In addition, we scanned the lines starting with MODRES in each PDB for information on modified residues. When these were different from previously encountered modifications, we added them to our conversion dictionary for use on future PDBs in our database. We ignored covalent linkers that do not resemble amino acids. We discarded chain information for very short sequences that clearly did not correspond to an *E. coli* protein.

Collecting DNA, mRNA, and translated protein information. *E. coli* cDNA sequences were downloaded from GenBank. Each cDNA was associated with a particular ID of Blattner and colleagues (1), and we included in our set those cDNAs whose Blattner and colleagues' ID had a corresponding Uniprot ID. These were transcribed to their corresponding mRNA sequences, which were then translated into protein sequences, halting at stop codons. In some cases, there was a frame problem in which the total length of the cDNA, as well as its corresponding mRNA, was not divisible by 3. In the 21 instances in which this occurred the Uniprot sequence was used for the translated sequence.

Identifying domains. We used the Structural Classification of Proteins (SCOP) version 1.75 (2) and CATH version 2.0 (3) databases to identify domains within our proteins of interest. Because of the potential for double-counting domains if we combined the two databases, we assembled an integrated list of domains using a hierarchical approach. For each PDB file in our set, we first identified the SCOP and CATH domains associated with it. Within each PDB, if a given chain had associated domains in CATH, we included these in the database. If it did not, we checked whether that chain had domains in the SCOP database. If so, we included these. For the 2,549 PDB chains for which there were not SCOP or CATH entries, we used pDomains to identify domains using the Domain Parser (DP) software. Because pDomains contain only a subset of PDBs, we were able to identify domains for 1,195 of these PDB chains.

In some cases, DP domains did not map well onto the numberings in the PDB files themselves. We conducted a series of transformations to correct for this. If the first residue in the DP domain had a lower numbering than the first numbered residue in the PDB chain, we shifted all the residues in the domain, such that the first residue in the DP domain and the PDB chain matched. We performed a similar calculation for the last residue of the PDB domain. First, we considered single domain proteins. We determined whether any segment of a given domain in DP had a starting residue with a lower numerical position than the starting residue in the PDB chain or an ending residue with a higher numerical position than the ending residue in the PDB chain. We adjusted the domain positions to account for the difference between the first domain position and the first chain position and between the last domain position and the last chain position. For multiple domain proteins, we conducted a similar procedure, but one that shifted all domains on the basis of those that lay outside the range of the PDB chain itself.

SCOP and CATH define domains based on evolutionary and homology relationships. Therefore, the domains they define will not always represent the autonomous folding units that are the relevant definition of a domain in this study. For example, 90% of the domains in our proteins of interest are composed of one contiguous segment along the primary structure; 10% consist of two or more segments that are noncontiguous along the primary

structure. The segments in the latter domains are separated by 126 residues on average (Fig. S1). This large separation means that it is reasonable to treat these two segments as autonomous folding units on the ribosome because they will have the opportunity to fold sequentially on the ribosome without influence from the other segment. Therefore, for the 10% of domains that consisted of multiple segments, we treated each segment as a separate domain in our database. Naturally occurring protein segments of less than 50 residues in length rarely are capable of folding autonomously (i.e., in isolation). Therefore, we removed from our database any domain that was less than 50 residues in length. This procedure resulted in the 1,236 cytosolic domains in our database.

Assigning domain residue and codon numbering. To analyze cotranslational folding, we needed to be able to compare codons on the mRNA with residues in the protein sequences. However, PDB sequences often include only fragments of the original protein or leave off beginning and ending residues. Therefore, we aligned codon and residue numbering using BLAST (4).

We used BLAST to analyze each segment of each domain in our database in comparison to the full-length translated sequence (as derived from the cDNA sequence). This enabled us to adjust the codon number on the basis of missing, mismatched, and inserted residues in the PDB sequence.

When a residue was missing from the PDB sequence of the domain, we added one to the residue number of the downstream residues (to reflect the offset). When a residue was inserted in the PDB sequence that did not exist in the original sequence, we skipped over numbering it and continued with numbering at the first noninserted residue. Mismatched residues were numbered normally.

In some cases, the BLAST analysis returned multiple alignments of a segment against the full-length sequence. In these cases, we only considered the first (highest bit score) alignment. In all cases, BLAST yielded an alignment.

Quantifying the structural fidelity and uniqueness of a domain. As a measure of the fidelity of the PDBs used in our analysis, we kept track of the number of missing, inserted, and mismatched residues for each domain. In the RCSB, multiple PDBs often refer to the same protein. Therefore, there is a great potential for duplicate domains in the final database. We sought to guard against this by using the following method to identify duplicate domains.

For the set of domains corresponding to a given Uniprot ID, we assumed domains were different if they possessed a different number of segments. For those domains with the same number of segments, we used the following test.

Consider two domains of protein A, denoted A_m and A_n. A_m and A_n each have i segments. Let x be an arbitrary tolerance threshold. $A_{m,i}$ is the ith segment of domain m of protein A. If $A_{m,i}$ and $A_{n,i}$ started more than x residues apart for one or more values of i, we considered the two domains to be different. Alternatively, if $A_{m,i}$ and $A_{n,i}$ ended more than x residues apart for one or more values of i, we considered the two domains to be different.

Otherwise, we considered the domains to be the same, and we used the following procedure to select the most accurate domain structure to include in our database. First, if one of the domains was derived from DP and the other was derived from SCOP or CATH, we deleted the DP domain. If this did not delete one of the domains, we compared the total number of missing residues. If the total numbers of missing residues in the two domains were different, we selected the domain with the fewest missing residues. If these values were the same, we compared the total number of inserted residues. If the total numbers of inserted residues in the two domains were different, we selected the domain with the fewest inserted residues. If these values were the same, we compared the total number of mismatched residues. If the total numbers of mismatched residues in the two domains were different, we selected the domain with the fewest mismatched residues. If these values were the same, we compared the total length of the domains. If the total lengths of the domains were different, we selected the domain with the greatest length. If these were the same, we compared the type of experiment from which the PDB data were obtained. If these were different, we selected the domain obtained by X-ray diffraction. If these were the same, we compared the resolution of the two structures. If these were different, we selected the domain with the highest resolution. If these were the same, we compared the deposition date of the structures. If these were different, we selected the domain of the structure most recently deposited. If these values were the same, we arbitrarily selected the first domain in our list.

For sets of domains considered to be the same that included more than two entries, we iterated this process, comparing pairs of domains in order of their listing, selecting one of each pair as we proceeded. We ended the procedure when only one domain remained.

We tried a variety of tolerance values, using $x = 30$ for the final database. Another possibility that we considered was that domains overlapped with each other. In cases in which this occurred, we used only domains from a single PDB chain, selecting this chain using the criteria listed above.

Determination of subcellular localization using PSORTb. We used PSORTb version 3.0 (5) with default settings for Gram-negative bacteria to predict subcellular localizations for each protein in our database.

Determination of proteins in the Sec and Tat pathways using SignalP and TatP. We used SignalP version 4.0 (6) with default settings for Gram-negative bacteria to identify proteins predicted to be secreted via the Sec pathway. We used TatP version 1.0 (7) with default settings to identify proteins predicted to be secreted via the Tat pathway. There are several scores produced by TatP, but the D-score is used by TatP for differentiation of secretory vs. nonsecretory proteins. We used this score for our classification.

Appending mRNA and protein abundance data. We include in our database literature data on mRNA expression and protein abundance from a variety of sources, as described below. In all cases, identifiers for data were first converted to Uniprot IDs. Further analysis was restricted to those data for which there was an unambiguous Uniprot ID (cases in which a single ID pointed to multiple Uniprot IDs were eliminated). Each restricted dataset was normalized by dividing its data points by the sum of the values in the set.

Marcotte dataset. For the dataset of Marcotte and colleagues (8), protein abundance data were reported using absolute protein expression methodology, which provides an estimate of protein abundance in molecules per cell. The original labels were IDs of Blattner and colleagues (1).

Frishman dataset. For the dataset of Frishman et al. (9), protein abundance data were reported as an estimate of the copy number per cell based on the emPAI score. The original labels were Uniprot IDs.

Church dataset. For the dataset of Church and colleagues (10), mRNA expression data were reported as an estimate of copy number per cell based on microarray experiments. The original labels were IDs of Blattner and colleagues (1).

Xie dataset. We include three entries from the dataset of Xie and colleagues (11): mean RNA expression, lifetime RNA expression, and protein abundance. RNA data were reported from RNA sequencing, whereas protein data were based on fluorescence measurements of fluorescently tagged proteins. The original labels were IDs of Blattner and colleagues (1).

Blattner dataset. For the dataset of Blattner and colleagues (1), mRNA expression data were reported for WT and heat shock conditions as an estimate of copy number per cell based on microarray experiments. The original labels were IDs of Blattner and colleagues (1).

Calculation of Codon Translation Times at Different *E. coli* Growth Rates. We use the method of Viljoen and colleagues (12) to calculate the mean time it takes to add an amino acid to the growing nascent chain at nascent chain length i, $\tau_{A,i}$, at 37 °C. In this approach $\tau_{A,i} = 9.06 + 1.45[10.48C(i) + 0.5R(i)]$ in milliseconds, where 9.06 is the time it takes for the chemical step of peptide bond formation and translocation of the A-site tRNA to the P-site and the term in brackets accounts for the time it takes for a cognate tRNA to bind to codon i $[10.48C(i)]$ and the delay due to kinetic competition of noncognate tRNA binding $[0.5R(i)]$. Full details of this model can be found in the study by Viljoen and colleagues (12). Briefly, the parameters $C(i)$ and $R(i)$ are a function of the codon identity, the density of cognate and noncognate tRNA molecules in the cell, the diffusion constants of tRNA, and the solution temperature. The cognate and noncognate tRNA identities were taken from table 1 in ref. 12. The number of tRNA and release factor (RF) molecules in *E. coli* were calculated using the cellular concentrations reported in Table 5 of Ref. 13 multiplied by *E. coli's* volume, V (Table S2) (13). tRNA molecules are typically fully charged in *E. coli* cells under nonstarvation conditions (14). Therefore, we assumed the numbers in Table S2 correspond to the number of charged tRNA molecules. The volume of *E. coli* in units of cubic micrometers was calculated from the empirical relationship $V = 0.4 \cdot 2^{dph} 10^{-18}$, where *dph* is the number of *E. coli* doublings per hour. The diffusion constants of the different tRNA molecules were taken from table 1 in ref. 12; however, for RF1 and RF2, we used a diffusion coefficient of $0.257 \cdot 10^{-11}$ m^2/s because we believe this is more realistic than the value of $0.3947 \cdot 10^{-11}$ m^2/s originally used. This is based on the fact that RF1 and RF2 are of a similar size, shape, and mass as the tRNA molecules. This change leads to the differences between the calculated $\tau_{A,i}$ values in Table S1 and table 5 in ref. 12.

Calculation of Domain Folding and Unfolding Kinetics in Bulk Solution. We used the de Sancho–Muñoz (DM) model (15) to estimate $\tau_{F,bulk}$ and $\tau_{U,bulk}$ at 310 K. The DM model uses experimentally informed enthalpy (table 3 in ref. 15), entropy (equations 1 and 2 in ref. 15), and heat capacity (equation 6 in ref. 15) per residue estimates to predict the transition state barrier height on folding and unfolding, denoted, respectively, as $\Delta_{U\to F}$ and $\Delta_{U\to F}$. These barrier values are then inserted into transition state theory to predict $\tau_{F,bulk}$ as being equal to $k_0^{-1} e^{\Delta_{U\to F}/RT}$ and $\tau_{U,bulk}$ as being equal to $k_0^{-1} e^{\Delta_{F\to U}/RT}$, where R is the universal gas constant and T is the solution temperature. In our analysis, we used the eight parameter values as originally reported for the DM model. However, we used a solution temperature of 310 K as opposed to the original value of 298 K. Given that the DM model uses temperature-dependent thermodynamic equations for the enthalpy, entropy, and heat capacity, we believe this is a reasonable approximation.

Structural Classification of Domains. The DM model uses the structural classification of a domain in its parameter selection. Domains were classified in Dataset S1 based on their secondary structural content using the program Stride (16). Stride's algorithm uses backbone dihedral angles as well as hydrogen bonding patterns to identify helical and β-strand structures. A domain was classified as either mostly α-helical or mostly β-strand if more than 70% of the residues comprising the ordered secondary structure were either α-helical or β-strand. A domain was considered mixed α/β if both the α-helical and β-strand content was greater than 30% of the ordered structure.

1. Allen TE, et al. (2003) Genome-scale analysis of the uses of the Escherichia coli genome: Model-driven analysis of heterogeneous data sets. *J Bacteriol* 185(21):6392–6399.
2. Andreeva A, et al. (2008) Data growth and its impact on the SCOP database: New developments. *Nucleic Acids Res* 36(Database issue):D419–D425.
3. Cuff AL, et al. (2011) Extending CATH: Increasing coverage of the protein structure universe and linking structure with function. *Nucleic Acids Res* 39(Database issue): D420–D426.
4. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215(3):403–410.
5. Yu NY, et al. (2010) PSORTb 3.0: Improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics* 26(13):1608–1615.
6. Petersen TN, Brunak S, von Heijne G, Nielsen H (2011) SignalP 4.0: Discriminating signal peptides from transmembrane regions. *Nat Methods* 8(10):785–786.
7. Bendtsen JD, Nielsen H, Widdick D, Palmer T, Brunak S (2005) Prediction of twin-arginine signal peptides. *BMC Bioinformatics* 6:167.
8. Lu P, Vogel C, Wang R, Yao X, Marcotte EM (2007) Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat Biotechnol* 25(1):117–124.
9. Ishihama Y, et al. (2008) Protein abundance profiling of the Escherichia coli cytosol. *BMC Genomics* 9:102.
10. Selinger DW, et al. (2000) RNA expression analysis using a 30 base pair resolution Escherichia coli genome array. *Nat Biotechnol* 18(12):1262–1268.
11. Taniguchi Y, et al. (2010) Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science* 329(5991):533–538.
12. Fluitt A, Pienaar E, Viljoen H (2007) Ribosome kinetics and aa-tRNA competition determine rate and fidelity of peptide synthesis. *Comput Biol Chem* 31(5-6): 335–346.
13. Dong H, Nilsson L, Kurland CG (1996) Co-variation of tRNA abundance and codon usage in Escherichia coli at different growth rates. *J Mol Biol* 260(5):649–663.
14. Dittmar KA, Sørensen MA, Elf J, Ehrenberg M, Pan T (2005) Selective charging of tRNA isoacceptors induced by amino-acid starvation. *EMBO Rep* 6(2):151–157.
15. De Sancho D, Muñoz V (2011) Integrated prediction of protein folding and unfolding rates from only size and structural class. *Phys Chem Chem Phys* 13(38): 17030–17043.
16. Heinig M, Frishman D (2004) STRIDE: A web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic Acids Res* 32(Web Server issue): W500-2.

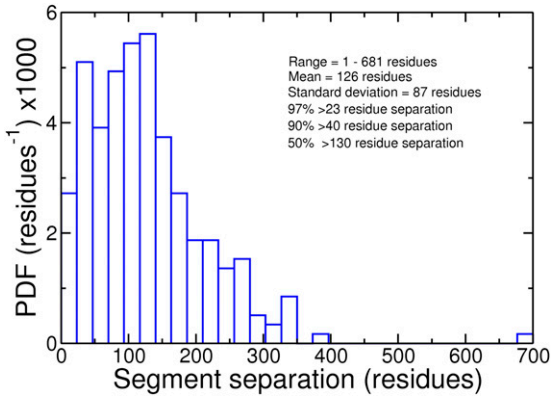

Fig. S1. Probability density function (PDF) distribution of the number of residues separating segments of the primary structure that comprises a single domain according to SCOP and CATH. This distribution was calculated from only those SCOP and CATH domains that contained more than one segment in our dataset of proteins of interest.

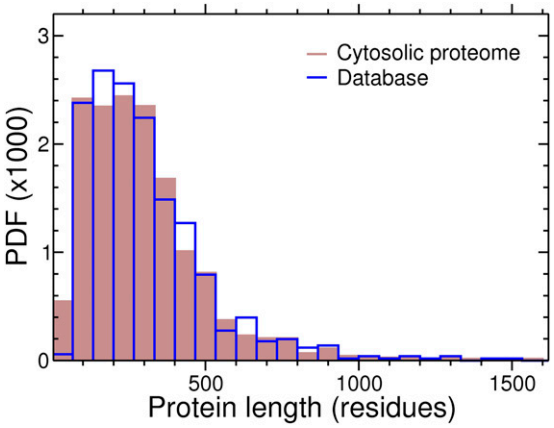

Fig. S2. Probability density function (PDF) of protein lengths of the cytoplasmic proteome in *E. coli* (brown) and the cytosolic proteins in our database (blue).

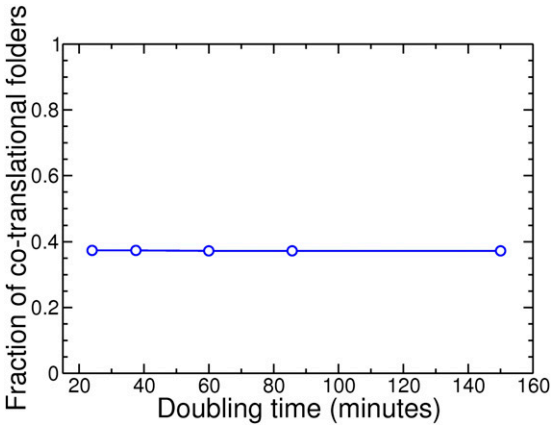

Fig. S3. Fraction of the cytosolic proteome that exhibits cotranslational folding as a function of *E. coli*'s growth rate at 37 °C. The line is to guide the eye and is not based on any model.

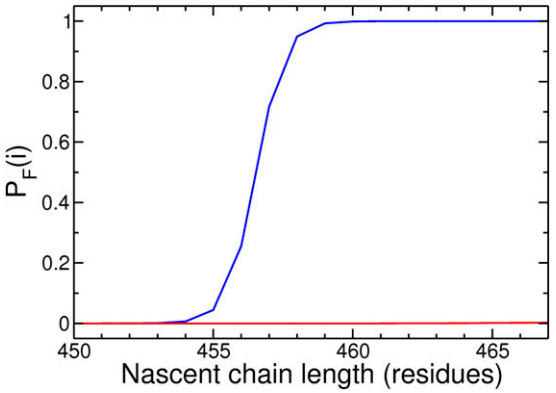

Fig. S4. Cotranslational folding curves of domain 2 from protein 6PDG at an infinitely slow translation (blue) and the finite translation rates in *E. coli* at 37 °C and a doubling time of 150 min (red). Note that this domain exhibits cotranslational folding at extremely slow translation rates but not at the translation rates in *E. coli*.

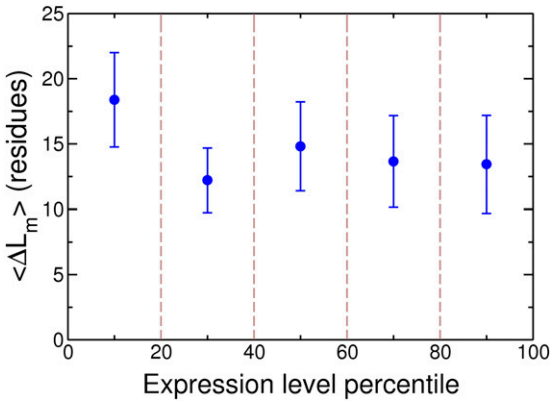

Fig. S5. Average ΔL_m value as a function of gene expression level as reported in the BLT_WT field in Dataset S1. The average ΔL_m value, and its standard error, were calculated from the set of cotranslational folding proteins that fell within a given quintile of expression level.

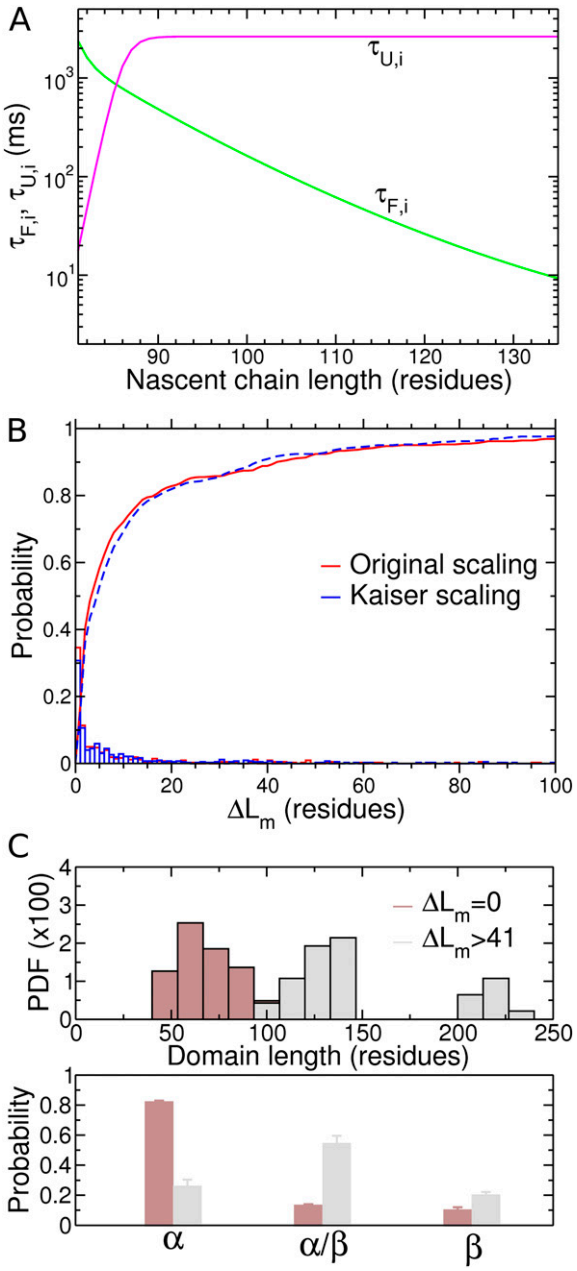

Fig. S6. Results using a different scaling relationship for $\tau_{F,i}$ (Eq. 3) that reproduces the trends observed in the laser optical tweezers (LOT) experiments of Kaiser et al. (1). (*A*) Behavior of $\tau_{F,i}$ for protein G using Eq. 3 with parameter values for *a*, *b*, and *c* of 402.371, 5.21917×10^{22}, and 10.4368, respectively. Note that we use the experimentally determined $\tau_{U,bulk}$ here (2). (*B*) Cumulative distribution function of ΔL_m using the original parameters (red) and the parameters (blue) of Kaiser et al. (1). (*C*) Structural characterization of domains that fold cotranslationally with $\Delta L_m = 0$ and those that fold with ΔL_m values greater than 41 residues in *E. coli* cells that are dividing every 150 min at 37 °C. (*Upper*) Probability density function (PDF) vs. domain length. (*Lower*) Probability of different domain classifications in terms of mostly α (α), mostly β (β), or mixed α/β secondary structure.

1. Kaiser CM, Goldman DH, Chodera JD, Tinoco I, Jr., Bustamante C (2011) The ribosome modulates nascent protein folding. *Science* 334(6063):1723–1727.
2. De Sancho D, Doshi U, Muñoz V (2009) Protein folding rates and stability: How much is there beyond size? *J Am Chem Soc* 131(6):2074–2075.

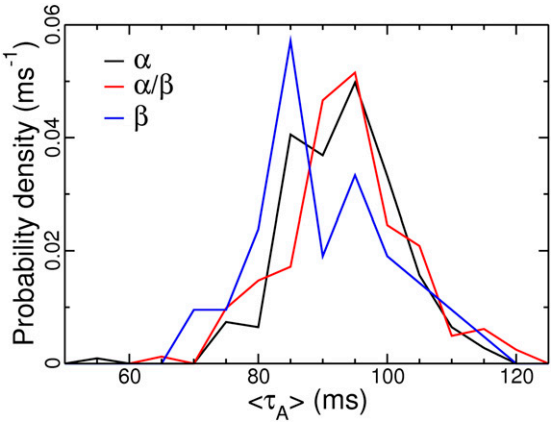

Fig. S7. Probability density distribution of the average translation rate of codons after the C terminus of a cotranslationally folding domain that is predominantly α-helical (black), β-strand (blue), or mixed α/β (red). The Mann–Whitney *U* test, corrected for multiple hypotheses using the Holm–Bonferroni method, indicates the median translation rates of these distributions are the same within statistical error.

Table S1. τ_A per codon (milliseconds) at different *E. coli* growth rates at 37 °C

		E. coli doubling time, min				
Codon	Label	150	86	60	37.5	24
UUU	1	100.7	96.8	89.5	100.9	99.0
UUC	2	161.7	152.0	138.5	164.2	156.4
UUG	3	36.3	37.3	36.9	37.6	42.1
UUA	4	111.0	127.3	117.4	163.1	168.0
UCU	5	35.8	31.9	33.0	32.9	34.3
UCC	6	194.7	194.7	187.6	191.2	184.5
UCG	7	88.4	70.5	77.6	79.1	81.5
UCA	8	75.5	60.8	66.7	63.1	65.8
UGU	9	51.4	55.3	56.3	53.4	59.5
UGC	10	86.2	92.2	93.8	90.6	97.0
UGG	11	125.3	140.2	126.5	133.2	117.3
UGA	12	28.2	25.2	28.9	38.4	39.9
UAU	13	33.8	36.6	37.6	33.0	38.1
UAC	14	57.9	62.1	65.4	55.6	63.4
UAG	15	114.6	98.2	106.7	143.7	184.9
UAA	16	24.9	22.7	25.3	33.5	35.0
CUU	17	214.7	192.1	179.7	207.1	178.5
CUC	18	162.7	151.2	141.7	157.0	136.1
CUG	19	27.8	28.1	27.4	26.9	28.9
CUA	20	232.5	212.6	227.2	227.8	254.5
CCU	21	110.9	104.2	111.8	103.5	116.3
CCC	22	58.4	59.0	61.1	53.3	61.9
CCG	23	108.7	126.2	101.5	142.8	156.0
CCA	24	173.4	179.6	171.5	177.4	194.2
CGU	25	21.1	23.1	21.8	19.9	20.4
CGC	26	28.1	31.2	30.1	27.2	28.3
CGG	27	339.2	269.5	529.6	453.3	483.5
CGA	28	27.2	30.1	29.7	25.3	26.1
CAU	29	233.2	223.3	204.2	218.4	190.3
CAC	30	163.5	166.9	153.6	153.7	138.0
CAG	31	188.5	182.3	179.1	149.5	134.8
CAA	32	128.2	124.6	103.9	149.2	125.0
GUU	33	18.5	19.1	20.1	17.9	18.3
GUC	34	177.3	172.0	156.8	170.3	175.2
GUG	35	30.5	32.5	37.0	30.0	30.8
GUA	36	60.4	65.2	77.1	61.6	66.4
GCU	37	27.0	26.0	24.1	24.8	23.9
GCC	38	355.3	360.4	363.3	336.0	362.3
GCG	39	31.7	30.4	28.1	29.3	27.7
GCA	40	69.7	64.4	56.6	64.7	61.9
GGU	41	27.2	26.8	26.3	28.5	25.7
GGC	42	41.4	40.6	39.8	44.1	40.9
GGG	43	61.9	63.5	63.9	56.5	65.6
GGA	44	289.0	291.3	289.7	278.7	321.1
GAU	45	58.3	57.7	59.4	57.3	53.5
GAC	46	95.9	96.3	100.7	92.5	86.8
GAG	47	29.9	30.6	30.3	28.8	28.0
GAA	48	51.6	52.6	52.0	50.7	48.4
AUU	49	74.4	76.2	75.8	69.4	66.4
AUC	50	105.9	106.4	104.7	95.2	88.8
AUG	51	210.9	202.2	198.9	186.2	213.8
AUA	52	105.2	106.2	104.6	95.3	90.7
ACU	53	36.5	36.5	37.0	35.3	32.7
ACC	54	121.4	118.5	119.4	129.2	134.6
ACG	55	102.7	102.1	107.9	98.6	92.2
ACA	56	136.2	133.7	139.7	114.8	96.9
AGU	57	57.9	61.5	63.5	68.4	71.3
AGC	58	100.2	107.2	112.5	124.4	137.0
AGG	59	407.9	345.1	369.4	309.7	416.0
AGA	60	145.1	181.6	171.4	178.2	189.7
AAU	61	74.8	79.6	76.1	68.6	66.6
AAC	62	127.2	133.6	127.7	116.6	114.6
AAG	63	83.8	83.3	85.0	93.1	94.7
AAA	64	57.5	54.3	54.1	59.8	59.1

Table S2. Number of molecules per cell at different *E. coli* growth rates at 37 °C

		E. coli doubling time, min				
tRNA	Label	150	86	60	37.5	24
Ala1	1	3,257	4,590	7,110	12,793	28,574
Ala2	2	619	829	1,178	2,329	4,864
Arg2	3	4,767	5,689	7,858	17,357	34,843
Arg3	4	638	1,021	733	1,650	3,134
Arg4	5	870	919	1,335	2,380	4,796
Arg5	6	390	614	814	1,796	2,997
Asn	7	1,198	1,510	2,199	4,454	9,933
Asp1	8	2,402	3,181	4,258	8,791	21,066
Cys	9	1,592	1,909	2,644	5,140	9,633
Gln1	10	766	1,064	1,835	2,314	5,968
Gln2	11	883	1,205	1,754	3,702	8,543
Glu2	12	4,729	6,096	8,450	17,613	39,993
Gly1	13	1,072	1,404	1,957	3,998	7,549
Gly2	14	1,072	1,404	1,957	3,998	7,549
Gly3	15	4,373	5,951	8,470	14,487	34,011
His	16	642	856	1,330	2,446	5,968
Ile1	17	1,741	2,318	3,352	6,907	16,856
Ile2	18	1,741	2,318	3,352	6,907	16,856
Leu1	19	4,484	5,834	8,475	15,568	30,250
Leu2	20	944	1,357	2,043	3,446	8,080
Leu3	21	667	974	1,324	2,329	4,319
Leu4	22	1,919	2,477	3,524	7,054	12,672
Leu5	23	1,134	1,357	2,058	2,665	5,150
Lys	24	1,932	2,660	3,717	6,374	14,212
Metf1	25	1,214	1,886	3,039	4,622	13,926
Metf2	26	718	892	1,193	2,468	5,137
Metm	27	708	1,013	1,471	2,993	6,036
Phe	28	1,039	1,408	2,169	3,424	6,963
Pro1	29	902	954	1,775	2,008	3,638
Pro2	30	721	982	1,142	2,928	5,109
Pro3	31	581	739	1,122	1,862	3,488
Sec	32	219	336	485	766	1,417
Ser1	33	1,300	2,175	2,766	5,096	10,029
Ser2	34	346	406	591	1,000	1,975
Ser3	35	1,411	1,717	2,290	3,943	7,726
Ser5	36	766	1,017	1,451	2,687	5,491
Thr1	37	101	160	273	408	912
Thr2	38	543	782	1,067	1,949	4,251
Thr3	39	1,099	1,459	1,957	3,548	7,549
Thr4	40	918	1,240	1,643	3,643	9,388
Trp	41	947	1,087	1,694	3,030	6,840
Tyr1	42	772	943	1,365	3,366	5,709
Tyr2	43	1,265	1,510	1,896	3,811	6,867
Val1	44	3,852	4,723	5,598	13,867	27,784
Val2A	45	632	782	1,203	1,971	3,801
Val2B	46	635	935	1,335	2,636	6,022
RF1	47	1,200	1,800	2,300	3,250	4,900
RF2	48	6,000	9,345	10,590	12,760	24,900

Other Supporting Information Files

SI Appendix (PDF)
Dataset S1 (TXT)
Dataset S2 (XLSX)