nature COMMUNICATIONS

ARTICLE

Received 14 Nov 2011 | Accepted 17 Apr 2012 | Published 29 May 2012

DOI: 10.1038/ncomms1850

Prediction of variable translation rate effects on cotranslational protein folding

Edward P. O'Brien[1], Michele Vendruscolo[1] & Christopher M. Dobson[1]

The concomitant folding of a protein with its synthesis on the ribosome is influenced by a number of different timescales including the translation rate. Here we present a kinetic formalism to describe cotranslational folding and predict the effects of variable translation rates on this process. Our approach, which utilizes equilibrium data from arrested ribosome nascent chain complexes, provides domain folding probabilities in quantitative agreement with molecular simulations of folding at different translation rates. We show that the effects of single codon mutations in messenger RNA that alter the translation rate can lead to a dramatic increase in the extent of folding under specific conditions. The kinetic formalism that we discuss can describe the cotranslational folding process occurring on a single ribosome molecule as well as for a collection of stochastically translating ribosomes.

[1] Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, UK. Correspondence should be addressed to C.M.D. (email: cmd44@cam.ac.uk).

Ribosome-bound nascent protein chains are particularly vulnerable to misfolding and interacting in aberrant manners with other cellular components[1]. To avoid these potentially dangerous possibilities and facilitate the folding process[2], a variety of quality control mechanisms are associated with translating ribosomes, including those involving molecular chaperones and other ancillary factors[1]. An additional level of control is provided by the opportunity for proteins to fold during synthesis[3–6], thus potentially enhancing folding yields[7] and avoiding misfolded or intermediate species[8,9]. Given its importance, it is not surprising that the cotranslational folding process can be regulated by the modulation of the rates at which successive amino acids are covalently attached to the nascent chain during synthesis. Thus, for example, reduced folding yields have been observed when slow-translating messenger RNA codons are mutated to fast-translating codons[10]. Even single synonymous mutations have been reported to decrease the total enzymatic activity of specific types of proteins, presumably because of cotranslational misfolding[11], leading to disease[12]. Furthermore, slow-translating codons have been observed to appear more frequently at domain boundaries[10], which can result in increased folding yields. All these results indicate that the interplay of the timescales of domain folding and amino-acid addition to the nascent chain is crucial in determining the extent of cotranslational folding (Fig. 1a).

Approaches based on the kinetic modelling of the molecular processes involved in translation have provided profound insights into the diverse functions of the ribosome. For example, the ability of the ribosome to discriminate between cognate and near-cognate transfer RNA has been explained using kinetic equations[13,14]. Here we extend this strategy to the prediction of the extent of nascent chain folding during continuous translation. This approach is based on the use of data on folding kinetics from arrested ribosome nascent chain (RNC) complexes and the time required to add individual residues to the nascent chain, quantities that can be measured using fluorescence or single molecule methods[15,16]. Making predictions based on arrested (that is, equilibrium) RNC data is convenient, because it is experimentally easier to probe such systems as compared with RNCs undergoing continuous, non-equilibrium translation[17]. Our approach is applicable to both single molecule and bulk cotranslational folding occurring during continuous protein synthesis.

Results

The extent of cotranslational folding on a single ribosome. To develop our approach, we first note that, in many instances, the folding of protein domains consisting of less than about 100 residues often occurs in bulk solution without significantly populating any intermediate state, and hence can be described phenomenologically by a two-state model[18] (Fig. 1b). In this scheme, a protein can interconvert between folded (F) and denatured (D) states. In what follows, we will consider this model in the context of a translating protein, allowing us to predict the extent of cotranslational folding at different rates of translation.

Translation introduces the additional timescale, τ_A, of amino-acid addition to the two-state kinetic scheme (Fig. 1c). Because the chemical environment surrounding a protein domain changes, as it is synthesized, the timescales of its folding ($\tau_{F,i}$) and unfolding ($\tau_{D,i}$) are a function of the nascent chain length i, that is, the number

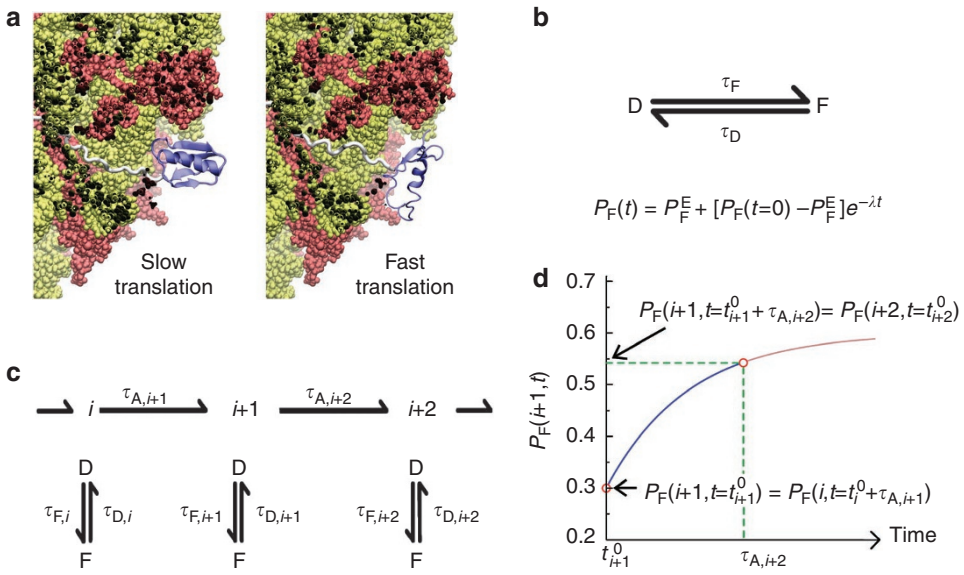

Figure 1 | Kinetic scheme for predicting variable translation rate effects on cotranslational folding. (**a**) Slower translation rates afford a protein domain (blue) more time to fold on the ribosome (left panel) than faster translation rates (right panel). Therefore, a slower translation rate is expected to increase the extent of cotranslational folding[19,31]. Ribosomal protein and RNA molecules are shown in red and yellow, respectively; the nascent chain's polyglycine linker is in white and the protein G domain is in blue. A portion of the ribosome has been removed in the figure to reveal the nascent chain in the exit vestibule. These structures were generated from continuous translation simulations with amino acids incorporated every 60 ms (left panel) and 1.3 ms (right panel). (**b**) A two-state model for protein interconversion between folded (F) and denatured (D) states, with mean folding and unfolding times denoted τ_F and τ_D. The time-dependence of folding, $P_F(t)$, is a function of these two timescales as well as the initial, $P_F(t=0)$, and final, P_F^E, folded fractions ($\lambda = [\tau_F]^{-1} + [\tau_D]^{-1}$). (**c**) A kinetic scheme for cotranslational folding. To make the two-state model applicable to cotranslational folding, we introduce the additional timescale, $\tau_{A,i}$, of amino-acid addition of the ith residue to the C-terminus of the nascent chain. The mean folding and unfolding times, $\tau_{F,i}$ and $\tau_{D,i}$, depend explicitly on the nascent chain length i, and addition of a new amino acid to the nascent chain, which occurs after $\tau_{A,i}$ ms, represents an irreversible reaction. (**d**) Probability that the domain is folded at a time t and corresponding nascent chain length $i+1$ ($P_F(i+1,t)$). For a two-state system, this probability relaxes towards its equilibrium value. The time available for folding at length $i+1$ (that is, the dwell time) is equal to the time $\tau_{A,i+2}$ it takes to incorporate the $i+2$ amino acid (see green dashed line). The final probability of folding at length $i+1$ is equal to the initial probability of folding at nascent chain length $i+2$, that is, $P_F(i+2, t=t^0_{i+2}) = P_F(i+1, t=t^0_{i+2}+\tau_{A,i+2})$.

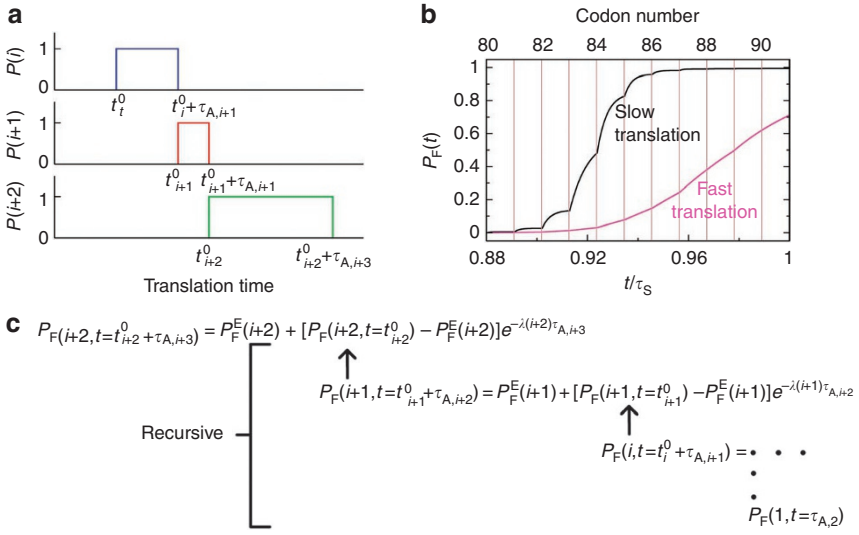

Figure 2 | Aspects of translation that are relevant in the derivation of the kinetic equation discussed in this work. (**a**) Probability of a single ribosome molecule containing a nascent chain length of (from top to bottom) i, $i+1$, and $i+2$ residues as a function of time. Because of the large separation of timescales between the chemical step of peptide bond formation and the ribosome dwell time at a specific codon, this probability is best approximated as a Boxcar function as shown. (**b**) Probability of the protein G domain folding as a function of the synthesis time and codon number (top axis) with new amino acids incorporated every 60 ms (black line) and 1.3 ms (magenta line). The equality of the initial $P_F(i+2,\ t=t^0_{i+2})$ and the final $P_F(i+1,\ t=t^0_{i+1}+\tau_{A,i+2})$ folding probabilities can be clearly seen in this figure. These curves were predicted according to equation (4) based on protein G's folding and unfolding kinetics. To facilitate the comparison of folding at these two different synthesis rates, the time has been divided by their respective total synthesis time $\tau_S=\left(\Sigma_{i=1}^{91}\tau_{A,i+1}\right)$. (**c**) Because of this equality, the folding behaviour at $i+2$ depends recursively on the behaviour at shorter nascent chain lengths. This is illustrated here by the kinetic equations that describe relaxation towards equilibrium (compare with Fig. 1b), with each successive nascent chain length being a function (indicated by the arrows) of the relaxation behaviour at earlier times (that is, at shorter nascent chain lengths) during synthesis. This recursive relation is expressed compactly in equation (4).

of residues comprising the nascent chain at a particular point during its synthesis[4]. The time available for the domain to interconvert between folded and denatured states at length i is equal to $\tau_{A,i+1}$, corresponding to the time it takes to attach the amino-acid $i+1$ to the nascent chain. $\tau_{A,i+1}$ has been shown to be influenced by a number of factors including the identity of the mRNA codon[19], the intracellular concentration of cognate and near-cognate aminoacyl-tRNAs[20], and the presence of secondary structure within the substrate mRNA[21]. For an apparent two-state folding protein, larger $\tau_{A,i+1}$values will increase the probability $P_F(i,t)$ that the domain will fold (that is, achieve its native structure) by affording the domain more time to do so at a nascent chain length i and time t after initiation of synthesis (Fig. 1d).

To derive an equation relating these three timescales ($\tau_{A,i+1},\tau_{F,i},\tau_{D,i}$) we first consider the behaviour of a single ribosome translocating along an mRNA molecule, and the time dependence of its nascent chain length. At a given nascent chain length i, the ribosome will dwell at codon $i+1$ waiting for this codon's cognate tRNA to be selected from the cytosol of the cell. This selection process involves a number of steps and a range of associated molecules such as elongation factor thermo unstable (EF-Tu). As we are concerned specifically with the nascent chain's length dependence as a function of time, we do not need to consider explicitly the details of these other chemical steps, for the reasons that follow. The time it takes to select the cognate tRNA and accommodate it into the A-site of the ribosome structure is stochastic in nature, but, on average, it is estimated in *Escherichia coli* to range from tens to hundreds of milliseconds depending on the identity of the tRNA molecule[20]. Once the A-site and P-site tRNAs are aligned, and receive sufficient thermal energy to pass over the transition state barrier, the chemical step of peptide bond formation, which changes the nascent chain length from i to $i+1$, takes on the order of picoseconds to nanoseconds[22] and is known as the transition path time[23,24]. This six orders-of-magnitude separation in the transition path time and τ_A timescales (picoseconds versus milliseconds) means that, for an individual ribosome molecule, the transition from nascent chain length i to $i+1$ appears instantaneous relative to the time the ribosome spends at either of these chain lengths.

As a consequence, the probability, $P(i)$, that this single ribosome molecule will contain a nascent chain of length i at time t is equal to the boxcar function probability distribution $\prod_{t_i^0,t_i^0+\tau_{A,i+1}}$ which equals 1 in the time interval $[t_i^0,\ t_i^0+\tau_{A,i+1})$ and is zero otherwise (Fig. 2a). t_i^0 is the time at which the ith amino acid is added to the nascent chain after initiation of translation. The change in $P(i)$ with respect to time is

$$\frac{dP(i)}{dt}=\delta(t_i^0)-\delta(t_i^0+\tau_{A,i+1}) \qquad (1)$$

where $\delta(t)$ is the Dirac delta function centred at time t after initiation of translation (Fig. 2a).

Next, we note that the experimentally observed timescale of the folding and unfolding process of a protein domain in free solution is typically on the order of milliseconds or more[25], and may be much longer near the ribosome surface[26]. Therefore, the picosecond-to-nanosecond transition-path time of peptide bond formation will also appear as instantaneous relative to the milliseconds or more folding/unfolding timescale. As a consequence, the probability that the nascent chain is in the folded state is equal immediately before (denoted $P_F\ (i,t=t_i^0+\tau_{A,i+1})$) and immediately after (denoted $P_F\ (i+1,\ t=t_{i+1}{}^0)$) the addition of the $i+1$ amino acid (Fig. 1d); that is, the starting point of folding at length $i+1$ is equal to the ending point at length i. Thus, during continuous translation, the extent of folding at a given nascent chain length is a function of the extent of folding at shorter lengths, and, hence, cotranslational folding depends recursively on what has happened at earlier times during the synthesis of the protein (Figs. 2b,c). As translation is a non-equilibrium process, memory effects can become prevalent, and so

it is not surprising that the extent of cotranslational domain folding depends on the states populated at earlier times during synthesis (Fig. 2c).

The specific behaviour of a single ribosome translocating along an mRNA containing N codons is therefore characterized by the series of dwell times at each codon $\{\tau_A\} = (\tau_{A,2}; \tau_{A,3}; ...; \tau_{A,N})$. If we have many independent measurements of domain folding on ribosomes that exhibit the same series of dwell times, then we can treat the domain folding probability as continuous and write down the differential equation defining the domain folding probability with respect to time as

$$\frac{dP_F(i)}{dt} = -\frac{1}{\tau_{D,i}} P_F(i,t) + \frac{1}{\tau_{F,i}}\left[1 - P_F(i,t)\right]. \quad (2)$$

and its solution is

$$P_F(i, t_i^0 \le t < t_i^0 + \tau_{A,i+1}) = P_F^E(i) + \left[P_F(i, t = t_i^0) - P_F^E(i)\right] e^{-\lambda(i)-[t-t_i^0]} \quad (3)$$

We substitute equation (3) into the recursive equations shown in Fig. 2c and rearrange them to find that P_F at arbitrary nascent chain length i and time t is

$$P_F(i, t_i^0 \le t < t_i^0 + \tau_{A,i+1}) = P_F^E(i) + \sum_{j=1}^{i}\left[P_F^E(j-1) - P_F^E(j)\right] e^{-\lambda(i)-[t-t_i^0]\sum_{k=1}^{i-1}\lambda(k)\tau_{A,k+1}}. \quad (4)$$

In equations (3) and (4), $P_F^E(i)$ is the equilibrium probability of folding and equals $\tau_{F,i}^{-1}/\lambda(i)$, with $\lambda(i)$ being the rate of interconversion of the folded and denatured states that equals $[\tau_{F,i}]^{-1} + [\tau_{D,i}]^{-1}$. This is in contrast to the out-of-equilibrium quantity $P_F(i, t = t_i^0)$ in equation (3), which is the folding probability immediately after adding the ith residue to the nascent chain. $\tau_{F,i}$ and $\tau_{D,i}$ are the average times of folding and unfolding at nascent chain length i on an arrested RNC. The placement of the first residue ($i = 1$) in the P-site of the ribosome, corresponding to *fmet*-tRNA in prokaryotes[27], is designated as time point zero, $t_1^0 = 0$ s, and the time at which the ith residue is added is $t_i^0 = \Sigma_{j=1}^{i}\tau_{A,j}$.

Accurate prediction of individual codon translation rate effects. Equation (4) is a function solely of $\tau_{F,i}$, $\tau_{D,i}$, $\tau_{A,i}$ and is a closed form solution to differential equations (1) and (2); therefore it provides an exact solution to the kinetic model shown in Fig. 1c. This equation expresses the probability that a domain is folded at each codon during continuous translation in terms of the equilibrium quantities $P_F^E(i)$ and $\lambda(i)$ that can be measured on arrested ribosomes, and the translation time of each codon ($\tau_{A,i}$), which can be measured by FRET and laser optical tweezer methods[15,16]. To date, however, few such measurements at different nascent chain lengths have been reported. Therefore, to test equation (4) rigorously, we generated an independent data set representing the probability of domain folding at various translation rates using coarse-grained molecular simulations (Supplementary Methods) of the synthesis of protein G on the ribosome from *Thermus thermophilus* (Fig. 1a).

Protein G is a single domain protein whose folded architecture consists of an α-helix located adjacent to a four-stranded β-sheet platform[28]. The coarse-grained model that we use has been shown previously to be consistent with a range of experimental data from arrested RNC complexes[4,29]. As in analogous experiments[17], we attached an unstructured linker to the carboxy terminus of protein G (Fig. 3a) to allow folding and unfolding of this domain to occur near the exit tunnel vestibule, where nascent chain tertiary interactions are sterically permitted[29,30].

We first calculated the equilibrium folding and unfolding kinetics (that is, $\tau_{F,i}$ and $\tau_{D,i}$) of protein G on arrested RNCs containing nascent chain lengths ranging from 81 to 92 AA (Fig. 3b). These timescales can be seen to vary with the nascent chain length, a result attributable to the change in chemical environment around the domain that arises in the simulations from electrostatic and excluded volume interactions between the nascent chain and the ribosome surface. We then simulated the continuous translation of protein G by covalently attaching new glycine residues to the nascent chain's C terminus at the biologically relevant[20] constant time intervals of 60, 10, 5, 2.5, and 1.3 ms, starting from a nascent chain length of 71 AA; at this length, protein G is unfolded on the ribosome as the C-terminal portion of the domain is in the exit tunnel[4]. To obtain statistically significant results, we carried out between 32 and 384 independent protein synthesis simulations at each translation rate.

The effects of translation rate on the extent of protein G folding at each nascent chain length are shown in Fig. 3c, and the corresponding root-mean-squared deviations of the protein G domain from its X-ray structure are shown in Fig. 4. We observe, consistent with previous conjectures[31], that the greater the translation rate the smaller the probability that the domain is folded at a given nascent chain length. Furthermore, at synthesis times close to the average value in *E. coli*, that is, $\tau_A = 50$ ms, we find that continuous translation and arrested RNCs result in the same extent of folding as a function of nascent chain length (Fig. 3c). This result occurs because the folding of protein G, during continuous translation at $\tau_A = 60$ ms, occurs under quasi-equilibrium conditions, where the folding reaction is under thermodynamic control, whereas at $\tau_A = 1.3$ ms cotranslational folding occurs under non-equilibrium conditions, where folding is under kinetic control[4]. It is important to emphasize that domains that fold on timescales of greater than 50 ms are more likely to be under kinetic control at synthesis timescales of $\tau_A \le 50$ ms ($\tau_F = 2$ ms for protein G in free solution[25]), and hence show a deviation between the non-equilibrium and equilibrium folding curves $P_F(i, t)$ and $P_F^E(i)$. In a database of single domain folding timescales[25] under physiologically relevant conditions, a quarter of them have $\tau_F \ge 50$ ms. Thus, at average *E. coli* synthesis rates during exponential growth[32], cotranslational folding of 25% or more of domains in multidomain proteins may be under kinetic control.

Importantly for the purpose of this study, the data in Fig. 3c provide a means to test the accuracy of equation (4). Inserting the arrested RNC folding kinetics from Fig. 3b into equation (4) and setting τ_A to the corresponding value used in the simulations, we find this kinetic formalism accurately and rapidly predicts the extent of cotranslational folding as a function of the translation rate (Fig. 3c). Thus, our approach captures the interplay of translation rate and folding and denaturation timescales and its consequence for the extent of cotranslational folding.

To test the sensitivity of equation (4) to single codon mutations that locally alter the translation rate along an mRNA molecule, we simulated cotranslational folding of protein G when a single 'fast'-translating codon ($\tau_{A,87} = 1.3$ ms) was placed at codon 87 in the context of a 'slow'-translating mRNA sequence ($\tau_A = 10$ ms). Conversely, we also simulated a system in which a single 'slow'-translating codon ($\tau_{A,90} = 10$ ms) was placed at codon 90 in the context of a 'fast'-translating mRNA sequence ($\tau_A = 1.3$ ms). We find that equation (4) accurately predicts the change in the extent of domain folding that results from the change in single synonymous codon mutations (Fig. 5a). This is a crucial demonstration of the utility of this formalism as synonymous mutations have been shown to alter

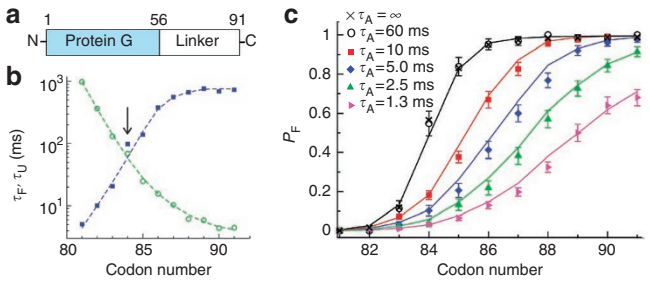

Figure 3 | Probability of cotranslational folding as a function of the translation rate. (**a**) To test the accuracy of the kinetic model (equation (4)), we simulated protein G as a RNC using a well-established coarse-grained model[4,29] (Supplementary Methods). We attached a polyglycine linker (codons 57 to 91) to the C-terminus of protein G (codons 1 to 56) to allow it to fold and unfold near the exit tunnel vestibule. (**b**) Average folding (τ_F, green circles) and unfolding (τ_U, blue squares) times as a function of nascent chain length on an arrested ribosome ($\tau_A = \infty$) calculated from the coarse-grained Langevin dynamics simulations (s.e.m. is shown for 15 replicates). The arrow indicates the nascent chain length at which the domain is folded with close to 50% probability. Lines are to guide the eye and are not based on any model. (**c**) Probability of domain folding during continuous translation as a function of nascent chain length (s.e.m. is shown for 5 replicates). Amino-acid incorporation timescales range from 1 to 60 ms and are constant for a given system. Equilibrium data are shown by black *x* symbols. The results from the coarse-grained simulations are shown as symbols, while the predictions from equation (4) are shown as solid lines and utilize the data from (**b**) as its arguments. The P_F data shown were calculated immediately before addition of the next amino acid to the nascent chain.

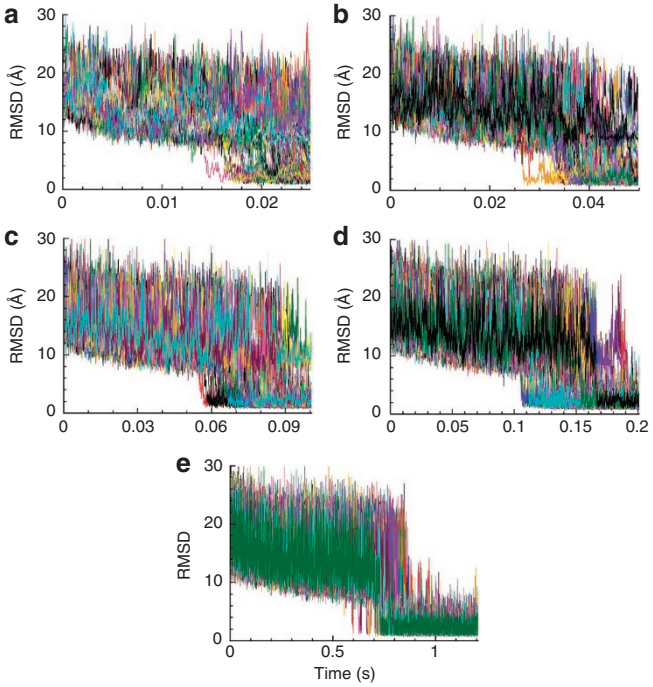

Figure 4 | Root-mean-squared deviation of the protein G domain. Root-mean-squared deviation of the protein G domain from its X-ray structure (PDB code 1GB1) during its continuous translation on the ribosome as a function of the simulation time (mapped onto the experimental timescale—see Methods). Each independent synthesis trajectory is shown as a different colour. Time equal to zero corresponds to a nascent chain length of 72 residues; the final nascent chain length is 91 residues. Panels (**a**) through (**e**) correspond, respectively, to adding a glycine to the C-terminal nascent chain residue every 1.3, 2.5, 5, 10, 60 ms.

folding yields dramatically[19]. These results also demonstrate that the predictions from this kinetic formalism are accurate and sensitive to the effect of variable translation rates at the level of single codons.

While the folding probabilities are shown as a function of nascent chain length in Figs 3c and 5a, equation (4) can also accurately predict these folding curves as a function of the time after the initiation of translation (Fig. 5b).

Application to a collection of translating ribosomes. In the preceding treatment, we considered a single ribosome molecule translocating along an mRNA molecule. Equation (4) therefore represents the average domain folding probability of a nascent chain on a ribosome that translocates with a specific series of dwell times $\{\tau_A\}$. As translocation of a ribosome along mRNA is stochastic, with a distribution of amino-acid addition times at a codon *i*, experiments on different ribosomes can yield different series of dwell times while they translate the same mRNA sequence.

How can we combine the exact result of equation (4), which utilizes a specific series of dwell times, with the stochastic nature of an ensemble of ribosomes, each with their own respective series of dwell times? If the probability density function $P_i(\tau_A)$ of amino-acid addition times at codon *i* is known *a priori*, then for a specific series of *N* dwell times, labelled as set *k* ($\{\tau_A\}_k$), we can calculate the probability p_k of that series occurring by random chance as

$$p_k = \prod_{i=1}^{N} P_i(\tau_A) \qquad (5)$$

Therefore, by inserting the same series of dwell times in both equations (4) and (5) , and multiplying the result as $p_k\ P_F(i=N, t_N^0 + \tau_{A,i+1})$, we obtain the contribution of the $P_F(i)$ folding curve of a single translating ribosome (for example, Fig. 3c) to the folding curve that would result from averaging over a large number of independent, stochastically translating ribosomes.

This result is useful for three reasons. First, equation (5) allows for the calculation of the probability of obtaining a particular single molecule trace (defined by the set of dwell times) in an experiment. Second, it allows for the numerical simulation of an arbitrarily large number of independent, stochastically translating ribosomes and each of their corresponding cotranslational folding probability curves. And finally, with sufficient such simulations, the distribution of folding probability curves and their average can be calculated for an ensemble of stochastically translating ribosomes. Importantly, this approach can be applied to arbitrary $P_i(\tau_A)$ distributions, thus providing it significant versatility.

To illustrate these points, consider an amino-acid addition time distribution $P_i(\tau_A)$ that is exponentially distributed and is therefore equal to $\left(1/\left\langle \tau_{A,i+1} \right\rangle\right) e^{-\tau^k_{A,i+1}/\left\langle \tau_{A,i+1} \right\rangle}$, where $\langle \tau_{A,i+1} \rangle$ is the average time required for amino-acid addition to a nascent chain of length *i*. Values of this time have already been estimated for all 48 codons in *E. coli*[20]. $\tau^k_{A,i+1}$ is the time it takes to add the $i+1$ residue to the nascent chain in the *k*th experiment in which a single ribosome translocating along mRNA is monitored. If $N=91$, as in the protein G construct discussed above, and $\langle \tau_{A,i+1} \rangle$ is taken as 60 ms for all codons, then the probability of observing a single ribosome translate a protein in which it dwells at each codon for 20 ms is effectively zero (about 10^{-171}). To simulate the individual folding curves of 1,000 ribosomes stochastically translating this protein G construct; however, we can randomly sample τ_A values from the exponentially distributed $P_{i+1}(\tau_A)$ for each codon (see Methods) and construct

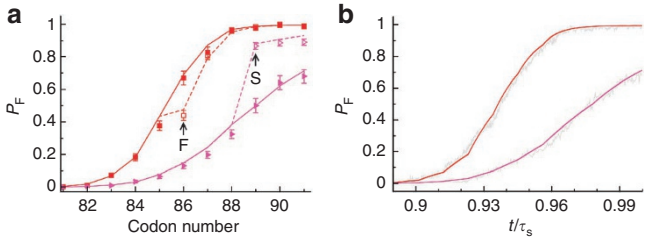

Figure 5 | The effects of synonymous codon mutations on cotranslational folding are accurately predicted by equation (4). (**a**) Comparison of the probability of protein G domain folding during continuous translation with amino acids added every $\tau_A = 10$ ms (solid red squares) and $\tau_A = 1.3$ ms (solid magenta triangles). The s.e.m. is shown for five replicates. A single fast translating codon, with $\tau_{A,87} = 1.3$ ms, was inserted at codon 87 (which shortens the dwell time at position 86 indicated by an arrow and the letter 'F') of the slower translating system ($\tau_A = 10$) and its effect on P_F is shown by the open red squares and the dashed red line. A single slower translating codon, with $\tau_{A,90} = 10$ ms, was inserted at codon 90 (indicated by an arrow and 'S') of the faster translating system ($\tau_A = 1.3$ ms) and its effect on P_F is shown by the open magenta triangles and the dashed magenta line. Coarse-grained simulation results are shown as symbols and predictions from equation (4) are shown as solid and dashed lines. In all systems, the final amino-acid sequence was the same. (**b**) Folding probability of the protein G domain as a function of the translation time for the slow (red line, $\tau_A = 10$) and fast (magenta line, $\tau_A = 1.3$ ms) translating mRNA. For each system, the time has been scaled by their total synthesis time. The predictions from equation (4) are shown as coloured lines and the results from the continuous translation simulations of the coarse-grained model are shown in grey. The difference in the absolute scale τ_S means that there are less data points for the fast translating mRNA system, making the simulation data appear less noisy.

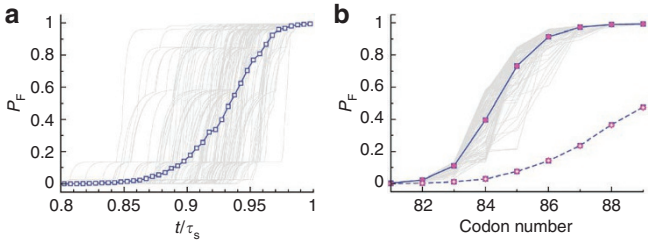

Figure 6 | Behaviour of a collection of stochastically translating ribosomes. (**a**) Cotranslational folding probability of protein G as a function of time for 110 stochastically translating ribosomes (grey lines). The variation in behaviour arises from the different series of dwell times $\{\tau_A\}$ associated with each ribosome. The blue line is the folding probability averaged over 1,000 such stochastically translating ribosome molecules. These data were numerically simulated by assuming $P_i(\tau_A)$ is exponentially distributed with $\langle \tau_{A,i} \rangle = 60$ ms for all i (see Methods). (**b**) Same as (**a**), except that the folding probability is shown per codon number immediately before the addition of the next residue to the nascent chain. Additionally, the ensemble average assuming $\langle \tau_{A,i} \rangle = 1.3$ ms for all i is also shown as a dashed blue line and squares. Predictions from equation (6) are shown as filled and open magenta diamonds for $\langle \tau_{A,i} \rangle = 60$ ms and 1.3 ms, respectively.

1,000 unique dwell time sets $[\{\tau_A\}_k]$. For each τ_A set, we can use equation (4) to calculate the resulting folding curve. Fig. 4 shows these 1,000 folding curves as a function of time (Fig. 6a) and nascent chain length (Fig. 6b). These results show that the kinetic model that we described can be utilized to predict how amino-acid timescales and their underlying distribution affect the extent of cotranslational folding of a protein domain at the resolution of an individual ribosome molecule, or for a large collection of ribosomes.

Exact solution for a collection of ribosomes. When $P_{i+1}(\tau_A)$ is exponentially distributed, it is possible to derive an exact expression relating the average cotranslational folding curve from a collection of stochastically translating ribosomes as a function of nascent chain length (equation (6), Attila Szabo, personal communication). That is, the blue line in Fig. 6b can be predicted without having to resort to the numerical simulations discussed in the previous section, although, by doing so, the information on the underlying distribution of folding curves is lost.

To derive the ensemble averaged folding curve as a function of nascent chain length, denoted $\langle P_F(i) \rangle$, a probabilistic approach can be utilized to analyse the elementary reaction steps in Fig. 1c (see Methods). Under these conditions,

$$\langle P_F(i) \rangle = \sum_{j=1}^{i} \frac{\tau_{F,j}^{-1}}{\tau_{A,j+1}^{-1}} \prod_{k=j}^{i} \frac{\tau_{A,k+1}^{-1}}{(\tau_{A,k+1}^{-1} + \tau_{F,k}^{-1} + \tau_{D,k}^{-1})}, \qquad (6)$$

where the superscript of '-1' indicates the reciprocal of these timescales. To test the accuracy of equation (6), we used it to calculate $\langle P_F(i) \rangle$ for protein G and compared it with results from the numerical simulations described in the previous section. We find excellent agreement between this exact result and the numerical simulations (Fig. 6b). Thus, equation (6) can predict the effect of per codon translation rates on the average cotranslational folding curve that arises from bulk experiments.

Discussion

We have presented two equations (equations (4) and (6)) that predict the extent of cotranslational domain folding based on per codon translation timescales, and the timescales of domain folding and denaturation on arrested RNC complexes at equilibrium (Fig. 3b). We have derived an exact expression for the domain folding probability in the case of a single translating ribosome (equation (4)), and shown how this expression can be utilized to predict the behaviour of a large number of stochastically translating ribosomes. Finally, an exact expression for the cotranslational folding curve was derived for ribosomes translating with exponential dwell times at each codon (equation (6)).

The utility of each of these equations depends on the questions that one is interested in addressing and the type of experiment (bulk versus single molecule) that is being carried out. In analysing and predicting cotranslational folding behaviour on individual ribosomes, equation (4) is perhaps the most relevant. The application, via numerical methods, of equation (4) to a collection of stochastically translating ribosomes is of direct consequence to both single molecule and bulk experiments as this approach offers the ability to calculate the individual ribosome folding curves as well as the ensemble average. This numerical approach can handle arbitrary distributions of amino-acid addition timescales and is thus not limited to the exponential dwell time distributions. Bulk experiments, where the average cotranslational folding curve as a function of nascent chain length may be measured from a collection of ribosomes, can be predicted using equation (6). Thus, these equations are applicable under a wide range of conditions.

Laser optical tweezers have recently[26] been used to measure the folding rate under tension of T4-lysozyme arrested on the ribosome at two different linker lengths. While the unfolding rate at zero force was not estimated at either length, these experiments clearly demonstrate that it is possible to measure $\tau_{F,i}$ and $\tau_{D,i}$ experimentally, which are key inputs in our approach. We expect that, as more studies measuring these rates are carried out on this and other proteins, such data, when combined with our approach,

will be useful in predicting what happens during continuous translation.

A number of additional translation-associated processes were not explicitly considered in the reaction scheme (Fig. 1c). For example, the competitive (and reversible) binding of near- and non-cognate tRNAs for a codon can slow down the rate of amino-acid addition by cognate tRNA molecules[20]. Furthermore, chaperones such as trigger factor directly interact with nascent chains during their synthesis, and can slow the rate of cotranslational folding of at least some proteins[33]. These processes do not diminish the utility of our approach, because each of these additional processes can effectively be accounted for by incorporating them into the timescales of amino-acid addition (in the case of competitive binding) and into the rates of folding and unfolding (in the case of trigger factor). The mathematical dependence of τ_A on near-cognate and non-cognate tRNA concentrations and their competitive binding rates has been worked out previously[20]. Thus, combining that model with equation (4) provides a means to model the effect of competitive tRNA binding on cotranslational folding. Similarly, when quantitative experimental measurements become available for the effect of trigger factor on the rates of domain folding and unfolding, they can be incorporated implicitly into effective timescales in these equations.

The kinetic models that we have described here are based on domains that fold cotranslationally in a two-state manner. This property applies to a variety of small proteins (typically ≤100 residues) and enables an analytical solution to be obtained for the kinetics of cotranslational folding. For protein domains larger than those examined here, which may populate intermediate states on the ribosome, the kinetic scheme in Fig. 1c can be modified to account for such additional states as they are experimentally identified. Although the additional complexity of such kinetic schemes may make it difficult to find an analytical solution, they could always be solved by numerical methods.

The approach we have proposed here has many potential applications in the areas of *in vivo* protein folding, biotechnology, and synthetic biology. For example, when coupled with models of translation rates that account for codon usage and tRNA concentrations[34], this formalism provides a means to predict cotranslational folding behaviour of entire proteomes under varying cellular conditions and aid in the design and of synthetic transcriptomes that optimize the extent of cotranslational folding. Equations (4) and (6) also provide a way for experimentalists to map directly the results from more easily studied arrested RNCs to the realistic situation of continuous translation. Thus, such kinetic modelling of the ribosome, when combined with a variety of different experimental data, provides new research avenues and the potential for novel insights in a number of different areas.

Methods

An exact solution for a collection of translating ribosomes. An equation can be derived relating the domain folding probability immediately before the addition of the next amino acid to the nascent chain for a ribosome that dwells with an exponential waiting time distribution at each codon (equation (6), Attila Szabo, personal communication). This probability, denoted $P_F^{final}(i)$, is equal to the probability of taking the pathway in Fig. 1c $F_i \rightarrow F_{i+1}$ and can be calculated as

$$P_F^{final}(i) = P_F^{final}(i-1)\cdot P(F_i \rightarrow F_{i+1}\#D_{i+1}) + P_D^{final}(i-1)\cdot P(D_i \rightarrow F_{i+1}\#D_{i+1}) \quad (7)$$

where $P_F^{final}(i-1)$ and $P_D^{final}(i-1)$ are, respectively, the probabilities that, when the nascent chain changes from length $i-1$ to i, the domain was either in the folded or denatured state. $P(F_i \rightarrow F_{i+1}\#D_{i+1})$ is the probability, that, beginning in the folded state at length i, the RNC complex will reach the folded state at length $i+1$ before reaching the denatured state at length $i+1$. Likewise, $P(D_i \rightarrow F_{i+1}\#D_{i+1})$ is the probability that, beginning in the denatured state at length i, the system will reach the folded state at length $i+1$ before reaching the denatured state at length $i+1$. Because there are only two states in our reaction scheme (Fig. 1c), we have that $P_D^{final}(i-1) = 1 - P_F^{final}(i-1)$. Substituting this into equation (7), we have

$$P_F^{final}(i) = P_F^{final}(i-1)\left[P(F_i \rightarrow F_{i+1}\#D_{i+1}) - P(D_i \rightarrow F_{i+1}\#D_{i+1})\right] + P(D_i \rightarrow F_{i+1}\#D_{i+1}) \quad (8)$$

and we see that $P_F^{final}(i)$ is a recursive relationship.

Using the probabilistic method[35] for calculating pathway probabilities in reaction schemes, $P(F_i \rightarrow F_{i+1}\#D_{i+1})$ and $P(D_i \rightarrow F_{i+1}\#D_{i+1})$ can be easily shown to equal, respectively, $1-\tau_{D,k}^{-1}/(\tau_{A,k+1}^{-1}+\tau_{F,k}^{-1}+\tau_{D,k}^{-1})$ and $\tau_{F,k}^{-1}/(\tau_{A,k+1}^{-1}+\tau_{F,k}^{-1}+\tau_{D,k}^{-1})$. Inserting these terms into equation (8), and using the boundary condition that at $i=1$ $P_F^{final}(0)=0$ and $P_D^{final}(0)=1$ (that is, for a nascent chain comprising one residue, the domain is denatured), this recursive relationship when solved equals equation (6).

Numerical simulation of a collection of translating ribosomes. To simulate the stochastic nature of translation, it is necessary to consider the randomly distributed dwell times $\{\tau_A\}$ that the ribosome exhibits during its translation of an mRNA molecule. As the underlying $P_i(\tau_A)$ distribution of amino-acid addition times at codon i has not yet been experimentally determined, here we assume it to be exponentially distributed with $P_i(\tau_A) = (1/\langle\tau_{A,i+1}\rangle)e^{(-\tau_{A,i+1}^k/\langle\tau_{A,i+1}\rangle)}$. For each ribosome, we constructed its $\{\tau_A\}$ by randomly sampling from this distribution, using an inverse transform sampling in which $\tau_{A,i+1}^k = \langle\tau_{A,i+1}\rangle\cdot\ln[R]$, where R is a random number selected from a uniform distribution in the range of (0,1). For each ribosome, this procedure results in 91 dwell times, representing a ribosome stochastically translating the protein G construct (Fig. 3a). This procedure was repeated 1,000 times, each yielding a unique $\{\tau_A\}$, which represents the behaviour of 1,000 different synthesis events of this protein. Each $\{\tau_A\}$ was then inserted into equation (4) to yield their corresponding cotranslational folding curve (Fig. 6).

Analysis of coarse-grained simulations. Mapping simulation timescales to experimental timescales. Low viscosity Langevin dynamics, as used in the coarse-grained simulations (Supplementary Methods), accelerate molecular dynamics while leaving the thermodynamic properties of the system unaltered. To map these accelerated kinetics to the experimentally relevant high-viscosity situation in aqueous media, we multiply the simulation time by the constant $\tau_{F,E}^{bulk}/\tau_{F,S}^{bulk} = 6.7\times10^5$, the ratio of experimentally measured to calculated folding times. $\tau_{F,E}^{bulk}$ is the experimentally (E) measured folding time of protein G in bulk solution and equals 2.4 ms (ref. 25), whereas $\tau_{F,S}^{bulk}$ is the average folding time from these simulations in the absence of the ribosome and equals 3.6 ns. This constant represents a linear scaling between the simulation time and the experimental time. Thus, in these simulations, when a new glycine residue is inserted into the growing chain during continuous translation every 90 ns ($=6\times10^6$ integration time steps) of simulation time, this interval corresponds to an experimental time of $\tau_A = 60$ ms. Likewise, τ_F, τ_D and the other τ_A values reported in the main text are the results of multiplying their simulation times by this constant.

The probability of domain folding at equilibrium, P_F^E, was calculated from the Replica exchange simulations (Supplementary Methods). A given simulation conformation of protein G was considered to be folded, if its fraction of native contacts was greater than 50%, and otherwise was considered unfolded. The folded/unfolded time series for each replica was constructed using this definition, and the time series from replicas at different temperatures combined in the WHAM equations[36] to calculate P_F^E. The stability of the folded state of protein G with respect to its denatured state, ΔG_{ND}, is equal to $-k_B T\ln[P_F^E/1-P_F^E]$, where k_B is Boltzmann's constant and T is the temperature.

The mean folding time τ_F of protein G equals the average of the set of first passage times $\{\tau_{F,i}\}$ determined from temperature quench simulations at various nascent chain lengths (Supplementary Methods). τ_D is calculated as $\tau_D = \tau_F\mathrm{Exp}(-\Delta G_{ND}/k_B T)$, where k_B is Boltzmann's constant and T is the simulation temperature.

The probability of domain folding during continuous synthesis simulations, $P_F(i)$, was calculated as $(1/N)\Sigma_{i=1}^N\Theta(Q_{BB,i}-0.50)$, where the summation is over the N-independent trajectories simulated for the given system, $\theta(Q_{BB,i}-0.50)$ is the Heaviside step function that equals 1 if more than half of the native backbone contacts $Q_{BB,i}$ in the structure of protein G are made in the last frame of the simulation at nascent chain length i and 0 otherwise. $Q_{BB} = (1/C)\ \Sigma_{j=1}^{S-4}\Sigma_{k=j+4}^{S}\Theta(r_{jk}^F - r_{jk})$, where C is the number of native backbone contacts within the crystal structure, S (=56) is the number of interaction sites in protein G, and r_{jk}^F and r_{jk} are, respectively, the spatial distances between interaction sites j and k in the crystal structure and the simulation structure. In this analysis, a native contact is identified in the crystal structure if any heavy atoms between residues j and k are within 4.5 Å of each other.

The standard error about the mean of τ_F was calculated by breaking the 152 independent folding trajectories into 15 sets of 10 or 11 $\tau_{F,i}$ values each, calculating the average value of each set and then calculating the standard deviation of the 15 averages divided by $\sqrt{15}$. To calculate the s.e.m. of ΔG_{ND}, the replica exchange sim-

ulation time-series data were broken into 5 independent sets, with approximately 20,000 points in each replica in each set. We then calculated ΔG_{ND} using each data set in the WHAM equations, and calculated the s.e.m., using these five ΔG_{ND} values. τ_D's s.e.m. was calculated using standard propagation of error equations.

References

1. Hartl, F. U. & Hayer-Hartl, M. Converging concepts of protein folding *in vitro* and *in vivo*. *Nat. Struct. Mol. Biol.* **16,** 574–581 (2009).
2. Thirumalai, D., O'Brien, E. P., Morrison, G. & Hyeon, C. Theoretical perspectives on protein folding. *Annu. Rev. Biophys.* **39,** 159–183 (2010).
3. Nicola, A. V., Chen, W. & Helenius, A. Co-translational folding of an alphavirus capsid protein in the cytosol of living cells. *Nat. Cell Biol.* **1,** 341–345 (1999).
4. O'Brien, E. P., Christodoulou, J., Vendruscolo, M. & Dobson, C. M. New scenarios of protein folding can occur on the ribosome. *J. Am. Chem. Soc.* **133,** 513–526 (2011).
5. Elcock, A. H. Molecular simulations of cotranslational protein folding: fragment stabilities, folding cooperativity, and trapping in the ribosome. *PLOS Comput. Biol.* **2,** 824–841 (2006).
6. Fedyukina, D. V. & Cavagnero, S. Protein folding at the exit tunnel. *Ann. Rev. Biophys.* **40,** 337–359 (2011).
7. Ugrinov, K. G. & Clark, P. L. Cotranslational folding increases GFP folding yield. *Biophys. J.* **98,** 1312–1320 (2010).
8. Clark, P. L. & King, J. A newly synthesized, ribosome-bound polypeptide chain adopts conformations dissimilar from early *in vitro* refolding intermediates. *J. Biol. Chem.* **276,** 25411–25420 (2001).
9. Netzer, W. J. & Hartl, F. U. Recombination of protein domains facilitated by co-translational folding in eukaryotes. *Nature* **388,** 343–9 (1997).
10. Komar, A. A., Lesnik, T. & Reiss, C. Synonymous codon substitutions affect ribosome traffic and protein folding during *in vitro* translation. *FEBS Lett.* **462,** 387–391 (1999).
11. Tsai, C. J. *et al.* Synonymous mutations and ribosome stalling can lead to altered folding pathways and distinct minima. *J. Mol. Biol.* **383,** 281–291 (2008).
12. Kimchi-Sarfaty, C. *et al.* A 'silent' polymorphism in the MDR1 gene changes substrate specificity. *Science* **315,** 525–528 (2007).
13. Hopfield, J. J. Kinetic proofreading - new mechanism for reducing errors in biosynthetic processes requiring high specificity. *Proc. Natl Acad. Sci. USA* **71,** 4135–4139 (1974).
14. Ninio, J. Kinetic Amplification of enzyme discrimination. *Biochimie* **57,** 587–595 (1975).
15. Khushoo, A., Yang, Z., Johnson, A. E. & Skach, W. R. Ligand-driven vectorial folding of ribosome-bound human CFTR NBD1. *Mol. Cell* **41,** 682–692 (2011).
16. Uemura, S. *et al.* Real-time tRNA transit on single translating ribosomes at codon resolution. *Nature* **464,** 1012–U73 (2010).
17. Hsu, S. T. D. *et al.* Structure and dynamics of a ribosome-bound nascent chain by NMR spectroscopy. *Proc. Natl Acad. Sci. USA* **104,** 16516–16521 (2007).
18. Jackson, S. E. & Fersht, A. R. Folding of chymotrypsin inhibitor-2.1. Evidence for a 2-state transition. *Biochemistry* **30,** 10428–10435 (1991).
19. Zhang, G., Hubalewska, M. & Ignatova, Z. Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. *Nat. Struct. Mol. Biol.* **16,** 274–280 (2009).
20. Fluitt, A., Pienaar, E. & Vijoen, H. Ribosome kinetics and aa-tRNA competition determine rate and fidelity of peptide synthesis. *Comput. Biol. Chem.* **31,** 335–346 (2007).
21. Qu, X. *et al.* The ribosome uses two active mechanisms to unwind messenger RNA during translation. *Nature* **475,** 118–21 (2011).
22. Schwartz, S. D. & Schramm, V. L. Enzymatic transition states and dynamic motion in barrier crossing. *Nat. Chem. Biol.* **5,** 552–559 (2009).
23. Dellago, C., Bolhuis, P. G., Csajka, F. S. & Chandler, D. Transition path sampling and the calculation of rate constants. *J. Chem. Phys.* **108,** 1964–1977 (1998).
24. Chung, H. S., McHale, K., Louis, J. M. & Eaton, W. A. Single-molecule fluorescence experiments determine protein folding transition path times. *Science* **335,** 981–4 (2012).
25. De Sancho, D., Doshi, U. & Munoz, V. Protein folding rates and stability: how much is there beyond size? *J. Am. Chem. Soc.* **131,** 2074–2075 (2009).
26. Kaiser, C. M., Goldman, D. H., Chodera, J. D., Tinoco, I. Jr & Bustamante, C. The ribosome modulates nascent protein folding. *Science* **334,** 1723–7 (2011).
27. Bingel-Erlenmeyer, R. *et al.* A peptide deformylase-ribosome complex reveals mechanism of nascent chain processing. *Nature* **452,** 108–111 (2008).
28. Gronenborn, A. M. *et al.* A novel, highly stable fold of the immunoglobulin binding domain of streptococcal protein-G. *Science* **253,** 657–661 (1991).
29. O'Brien, E. P., Hsu, S. T. D., Christodoulou, J., Vendruscolo, M. & Dobson, C. M. Transient tertiary structure formation within the ribosome exit port. *J. Am. Chem. Soc.* **132,** 16928–16937 (2010).
30. Kosolapov, A. & Deutsch, C. Tertiary interactions within the ribosomal exit tunnel. *Nat. Struct. Mol. Biol.* **16,** 405–411 (2009).
31. Purvis, I. J. *et al.* The efficiency of folding of some proteins is increased by controlled rates of translation invivo - a hypothesis. *J. Mol. Biol.* **193,** 413–417 (1987).
32. Young, R. & Bremer, H. Polypeptide-chain-elongation rate in escherichia-coli B-R as a function of growth-rate. *Biochem. J.* **160,** 185–194 (1976).
33. Agashe, V. R. *et al.* Function of trigger factor and DnaK in multidomain protein folding: Increase in yield at the expense of folding speed. *Cell* **117,** 199–209 (2004).
34. Czech, A., Fedyunin, I., Zhang, G. & Ignatova, Z. Silent mutations in sight: co-variations in tRNA abundance as a key to unravel consequences of silent mutations. *Mol. Biosyst.* **6,** 1767–1772 (2010).
35. Ninio, J. Alternative to the steady-state method: derivation of reaction rates from first-passage times and pathway probabilities. *Proc. Natl Acad. Sci. USA* **84,** 663–7 (1987).
36. Kumar, S., Bouzida, D., Swendsen, R. H., Kollman, P. A. & Rosenberg, J. M. The weighted histogram analysis method for free-energy calculations on biomolecules.1. the method. *J. Comput. Chem.* **13,** 1011–1021 (1992).

Acknowledgements

We thank Attila Szabo for stimulating discussions on chemical kinetic modelling and for proposing and deriving equation (6); Sophie Jackson for a careful reading of the manuscript; Robert Best for providing CHARMM source code for the double-well dihedral potential and Debye-Huckel electrostatic calculations; Changbong Hyeon for useful suggestions on modelling electrostatic interactions in coarse-grained models; and John Christodoulou for illuminating discussions about cotranslational folding. This work was supported by an NSF postdoctoral grant (EPO), BBSRC and the Wellcome Trust (MV and CMD), and the EPSRC (EPO, MV and CMD). This study utilized the high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health, Bethesda, Maryland. (http://biowulf.nih.gov).

Author contributions

E.P.O., M.V., and C.M.D. designed the research. E.P.O. carried out the research and analysed the data. E.P.O., M.V., and C.M.D. interpreted the data and wrote the manuscript.

Additional information

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: O'Brien, E. P. *et al.* Prediction of variable translation rate effects on cotranslational protein folding. *Nat. Commun.* 3:868 doi: 10.1038/ncomms1850 (2012).

nature COMMUNICATIONS

DOI: 10.1038/ncomms2559

Erratum: Prediction of variable translation rate effects on cotranslational protein folding

Edward P. O'Brien, Michele Vendruscolo & Christopher M. Dobson

Nature Communications 3:868 doi:10.1038/ncomms1850 (2012); Published 29 May 2012; Updated 26 February 2013

This article contains typographical errors in equations (3) and (4) that were introduced during the production process. These errors do not affect the analysis or results presented in the paper. Equations (3) and (4) should read as follows.

$$P_{\mathrm{F}}(i,t_i^0 \le t < t_i^0 + \tau_{\mathrm{A},i+1}) = P_{\mathrm{F}}^{\mathrm{E}}(i) + [P_{\mathrm{F}}(i, t = t_i^0) - P_{\mathrm{F}}^{\mathrm{E}}(i)]e^{-\lambda(i)\cdot[t - t_i^0]} \tag{3}$$

$$P_{\mathrm{F}}(i,t_i^0 \le t < t_i^0 + \tau_{\mathrm{A,i}+1}) = P_{\mathrm{F}}^{\mathrm{E}}(i) + \sum_{j=1}^{i} [P_{\mathrm{F}}^{\mathrm{E}}(j-1) - P_{\mathrm{F}}^{\mathrm{E}}(j)]e^{-\lambda(i)\cdot[t - t_i^0]\sum_{k=j}^{i-1}\lambda(k)\tau_{\mathrm{A},k+1}} \tag{4}$$

In addition, the mathematical term in the sentence immediately following equation (5) also contains a typographical error, and should read: $p_k P_{\mathrm{F}}(i = N, t_N^0 + \tau_{\mathrm{A},N+1})$.