Prediction of variable translation rate effects on cotranslational protein folding

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

Department of Chemistry, University of Cambridge, Cambridge, UK

*cmd44@cam.ac.uk

Supplementary Information

Table S1: Betancourt-Thirumalai values used in calculating the well-depths for the 12-10-6 potential in units of kcal/mol.																				
	cys	**phe**	**leu**	**trp**	**val**	**ile**	**met**	**his**	**tyr**	**ala**	**gly**	**pro**	**asn**	**thr**	**ser**	**arg**	**gln**	**asp**	**lys**	**glu**
cys	2.861																			
phe	1.667	2.095																		
leu	1.623	2.035	2.080																	
trp	1.976	2.035	1.917	1.976																
val	1.637	1.873	2.065	1.800	1.947															
ile	1.593	1.844	2.050	1.844	1.888	1.770														
met	1.608	2.198	1.888	2.272	1.578	1.770	1.711													
his	1.165	1.165	0.738	1.564	0.620	0.605	1.136	1.372												
tyr	1.121	1.608	1.534	1.696	1.283	1.372	1.637	1.195	1.283											
ala	1.268	1.372	1.431	1.475	1.446	1.401	1.224	0.575	1.106	1.180										
gly	1.018	0.723	0.678	1.239	0.826	0.575	0.767	0.546	0.944	0.929	1.180									
pro	1.151	1.165	1.003	1.962	1.003	0.811	1.121	0.959	1.475	0.782	0.900	0.988								
asn	0.472	0.457	0.354	1.018	0.310	0.074	0.413	0.738	0.870	0.531	0.738	0.693	0.944							
thr	0.885	0.885	0.885	0.885	0.885	0.885	0.885	0.885	0.885	0.885	0.885	0.885	0.885	0.885						
ser	0.752	0.738	0.501	0.782	0.516	0.369	0.413	0.664	0.782	0.664	0.738	0.634	0.678	0.885	0.693					
arg	0.413	0.767	0.752	1.490	0.634	0.620	0.634	0.826	1.431	0.487	0.678	0.915	0.856	0.885	0.708	0.693				
gln	0.826	0.944	0.767	1.047	0.634	0.678	0.900	0.560	1.151	0.575	0.590	0.959	0.959	0.885	0.516	1.062	0.678			
asp	0.324	0.177	0.030	0.797	0.089	0.088	0.030	1.210	0.988	0.443	0.634	0.516	1.062	0.885	0.870	1.932	0.708	0.487		
lys	0.369	0.723	0.649	1.298	0.649	0.575	0.560	0.501	1.475	0.590	0.708	0.708	1.092	0.885	0.738	0.147	1.180	1.903	0.324	
glu	0.206	0.383	0.339	1.106	0.280	0.324	0.531	1.047	1.121	0.251	0.177	0.501	0.900	0.885	0.738	1.991	0.738	0.295	2.168	0.221

Table S2: Collision diameter of ribosomal amino acid interaction sites used in modelling inter-molecular non-bonded interactions.

Residue	σ_i (Å)
Gly	2.3
Ala	2.5
Val	2.9
Leu	3.1
Ile	3.1
Met	3.1
Phe	3.2
Pro	2.8
Ser	2.6
Thr	2.8
Cys	2.7
Arg	2.8
Glu	3.0
Tyr	3.2
Trp	3.4
Asp	2.8
Glu	3.0
His	3.0
Lys	3.2
Arg	3.3

Supplementary Methods

Coarse grained ribosome-nascent chain model. Protein molecules, including ribosomal protein and the nascent chain, were modeled based on a variant[37] of the Karanicolos-Brooks coarse grained model for proteins[38]. In this model, each residue is represented by one interaction site centered on the C_α backbone atom. Bond distances between neighboring C_α interaction sites along the primary sequence were held fixed at 3.8 Å, and a double-well potential was used for the bond angles allowing triads of C_α interaction sites to adopt angles corresponding to either α or β structure[37]. The dihedral angle potential is transferable and is a function of the amino-acid sequence identity of the four interaction sites defining the given dihedral angle[38]. Electrostatic interactions between groups were modeled using Debye-Huckel theory[37]. A Debye screening length of 10 Å was used, which is close to the value in *E. coli*'s cytoplasm[39], in conjunction with a dielectric constant of 78.5. Lysine and arganine residues were assigned charges of +1*e*, and glutamine and aspartate residues were assigned charges of -1*e*. All other protein interaction sites had 0 charge. Intra-protein van der Waals interactions were modeled using the 12-10-6 potential[38], which captures solvent separated minima effects between interaction sites. Intra-protein native and non-native interactions, as well as the collision diameters between interaction sites, were defined using a standard procedure[38]. Instead of using the Miyazawa-Jernigan values to set the well-depths in the 12-10-6 potential, we used the Betancourt-Thirmulai (BT) values[40], which are in better agreement with measured side chain hydrophobicities. In this potential we set the well depth between the native contact formed by residues *i* and *j* to $\varepsilon_{ij} = \varepsilon_{ij}^{H-bond} + \varepsilon_{ij}^{sb} + \varepsilon_{ij}^{ss}$, where $\varepsilon_{ij}^{H-bond}$ equals 0.75 kcal/mol if a hydrogen bond is present between them in the crystal structure, ε_{ij}^{sb} equals 0.37 kcal/mol if the sidechain (*s*) of residue *i* is in contact with the backbone (*b*) of residue *j*, or vice versa, and ε_{ij}^{ss} is proportional to the BT values if these residue's have side chains that are in contact in the crystal structure. ε_{ij}^{ss} is calculated by subtracting the BT values (Table 2 in reference 4) by 0.6 and the absolute value of the result multiplied by a factor of 1.475 (Table S1), a value chosen to reproduce the experimentally measured stability of protein G in bulk solution. Hydrogen bonds were identified in the crystal structure using the program Stride[41]. We note that in this model four out of the five energy terms in this force field are transferable – corresponding to bond, angle, dihedral and electrostatic energy terms. Only the 12-10-6 potential is non-transferable.

The van der Waals interactions between ribosomal protein and nascent chain were modeled using a short-range repulsive term, $\varepsilon\sigma_{ij}^{12}/r_{ij}^{12}$, where σ_{ij} is the collision diameter between interaction sites *i* and *j* and equals $\sigma_{ij} = (\sigma_i + \sigma_j)/2$, r_{ij} is the distance between them and ε equals 0.000132 kcal/mol. For ribosomal protein, the values of σ_i are proportional to the partial molar volume of free amino acids in solution[42], and their values are listed in Table S2; for nascent chain protein, σ_i were set using a standard procedure[38].

Ribosomal RNA was modeled as before[43] using three interaction sites for A and T nucleotides, and four interaction sites for G and C nucelotides. For each nucleotide interactions sites were centered on the phosphate group, the centroid of the ribose sugar ring, and the centroid of each conjugated ring in the base. The phosphate interaction sites were assigned a charge of -1*e*, while all other RNA interaction sites had zero charge. Debye-Huckel theory, with a 10 Å Debye length, was used to calculate all electrostatic interaction energies between charged protein and RNA groups. No bond, bond angle, or dihedral angle energy terms were defined for the RNA molecules because during the

simulations these interaction sites were held fixed (see below). Van der Waals interactions between RNA and protein were treated with a $\varepsilon\sigma_{ij}^{12}/r_{ij}^{12}$ potential. The phosphate, ribose, and base interaction site's σ_i values equal 4.0 Å each. For the ribosomal protein interactions sites, the σ_j values listed in Table S2 were used, with ε equal to 0.000132 kcal/mol[2].

As done previously[44], the polyglycine linker was modeled using a fully transferable force field potential that models its unstructured conformations.

PDBs 2WDN[45] for the ribosome, and 1GB1[46] for protein G were used as starting coordinates for this coarse graining procedure. This ribosome structure was chosen over others because most of the heavy atoms and the position of the C_α atom of the first P-site nascent chain residue were crystolographically resolved. Furthermore, at the start of this project, it had the largest number of ribosomal protein molecules resolved in the 50S subunit.

Simulation details of the coarse-grained ribosome model. Due to the large spatial separation between folding events that can occur at the exit tunnel vestibule[44] and the 30S subunit (>8 nm) only the 50S subunit of the ribosome was explicitly represented in these simulations. Simulations were carried out in CHARMM[47] (version c35b5) using Langevin dynamics with a 15 fs integration time step and a damping coefficient of 0.05 s^{-1}. This damping coefficient was found to maximize the frequency of folding/unfolding transitions in free solution per unit of simulation time. Using the Constraints module in CHARMM, ribosome interaction sites were held rigid during the simulations. This extended the maximum simulation time scale that could be obtained by orders-of-magnitude. We've previously shown that holding the ribosome rigid has negligible effect on tertiary structure formation within the exit vestibule[43] and therefore we don't expect this aspect of the simulation to affect the results of this study. Using the MMFP module, we modeled the covalent attachment between the C-terminus of the nascent chain and the P-site tRNA by harmonically constraining the C-terminal interaction site of the nascent chain to its crystal position (in PDB 2WDN) with a force constant of 50 kcal/(mol $Å^2$). Non-bonded interactions were truncated at 20 Å, with a shift function applied starting at 18 Å. We tested the effect of this cut-off on the calculated thermodynamic properties of protein G in free solution by calculating its melting temperature using a cut-off of 80 Å. We find a negligible difference using these two different cut-offs. This indicates that using a shorter cut-off does not alter the conclusions of this study. The Shake algorithm[48] was used to keep all bond lengths fixed at their equilibrium value during the simulations.

Stalled, equilibrium RNC simulations. To calculate the equilibrium properties of nascent protein G folding we studied RNCs consisting of nascent chain lengths ranging from 81 to 91 residues, corresponding to poly-glycine linker lengths of 25 to 35 residues (Fig. 2A). Initial configurations at each nascent chain were constructed by unfolding the nascent chain near the exit tunnel opening using high temperature, and then applying a series of harmonic restraints to its C-terminus to pull it into the exit tunnel until the C-terminus finally reached the PTC. To evaluate protein G-RNCs' thermodynamic properties we used Replica Exchange simulations (RES)[49] consisting of 16 different temperature replicas that ranged between 200 to 450 K. For each replica, temperature swaps with neighboring temperature windows were attempted every 1300 integration time steps, with system coordinates of the last frame saved for later analysis. A total of $1.3x10^5$ swap attempts were made. The first $4x10^4$ swap attempts were discarded during the analysis phase to allow for system equilibration. The percentage of successful swap attempts ranged from 10 to 80% across replicas,

and each trajectory was observed to make multiple transitions between the highest and lowest temperature replicas at each nascent chain length.

To calculate the folding kinetics of protein G on the stalled ribosome we carried out first passage time simulations in which a given trajectory *j* is started in the unfolded state and the simulation time it takes to first reach the folded state, $\tau_{F,j}$, recorded. The average of many of these first passage times is equal to the inverse of the folding rate constant, k_F. At each nascent chain length we prepared the system in an unfolded conformation by first equilibrating it for 5x10^6 integration time steps at 365 K, which is above protein G's melting temperature, and then quenching the system temperature instantaneously (defined as time zero) to 310 K. Once 70% of the native contacts (see below) were made within the protein G domain the simulation was stopped and the simulation time recorded as the first passage time $\tau_{F,j}$. Anywhere from 64 to 384 independent folding trajectories were carried out at each nascent chain length in this manner, with the smaller the τ_A value the more trajectories that were run. Protein G is an apparent two state folding protein[50], therefore, the average unfolding time, τ_D was calculated as $\tau_D = \tau_F \mathrm{Exp}(-\beta \Delta G_{ND})$, with τ_F and ΔG_{ND} having been determined from the simulations described above.

Continuous, non-equilibrium RNC simulations. Continuous translation simulations were started from a nascent chain length of 71 residues at which protein G is always unfolded[44] and has a 15 residue glycine linker attached to its C-terminus. New glycine residues were covalently attached to the nascent chain every 6x10^6, 1x10^6, 5x10^5, 2.5x10^5, or 1.25x10^5 integration time steps, corresponding, respectively, to experimental τ_A times of 60, 10, 5, 2.5, and 1.3 ms (see Methods). The process of residue addition consisted of several steps. First, the simulation at nascent chain length *i* was stopped and a new residue was introduced to the system 3.8 Å from the C-terminus of the nascent chain (towards the location of the 30S subunit). The 'link' command in CHARMM was issued to covalently attach the new glycine to the C-terminus. Second, the harmonic constraint on the C-terminal interaction site was released and applied to the new residue, with a target position located at the C_α atom position in crystal structure 2WDM[45]. Third, all other interaction sites, except for the newly added glycine, were held fixed and the system was minimized using the Adopted Basis Newton-Raphson method[50]. Finally, the constraints holding the nascent chain fixed were released, and Langevin dynamics performed until the next glycine arrived, whose arrival time is set by $\tau_{A,i+2}$. The new glycine was added to the nascent chain using the same procedure.

Supplementary References

37. Best, R.B., Chen, Y.G. & Hummer, G. Slow protein conformational dynamics from multiple experimental structures: The helix/sheet transition of arc repressor. *Structure* **13**, 1755-1763 (2005).
38. Karanicolas, J. & Brooks, C.L. The origins of asymmetry in the folding transition states of protein L and protein G. *Protein Sci* **11**, 2351-2361 (2002).
39. Spitzer, J.J. & Poolman, B. Electrochemical structure of the crowded cytoplasm. *Trends Biochem Sci* **30**, 536-541 (2005).
40. Betancourt, M.R. & Thirumalai, D. Pair potentials for protein folding: Choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. *Protein Sci* **8**, 361-369 (1999).

41. Frishman, D. & Argos, P. Knowledge-based protein secondary structure assignment. *Proteins* **23**, 566-579 (1995).
42. O'Brien, E.P., Ziv, G., Haran, G., Brooks, B.R. & Thirumalai, D. Effects of denaturants and osmolytes on proteins are accurately predicted by the molecular transfer model. *P Natl Acad Sci USA* **105**, 13403-13408 (2008).
43. 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).
44. 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).
45. Voorhees, R.M., Weixlbaumer, A., Loakes, D., Kelley, A.C. & Ramakrishnan, V. Insights into substrate stabilization from snapshots of the peptidyl transferase center of the intact 70S ribosome. *Nat Struct Mol Biol* **16**, 528-533 (2009).
46. Gronenborn, A.M. et al. A Novel, Highly Stable Fold of the Immunoglobulin Binding Domain of Streptococcal Protein-G. *Science* **253**, 657-661 (1991).
47. Brooks, B.R. et al. CHARMM: The Biomolecular Simulation Program. *J Comput Chem* **30**, 1545-1614 (2009).
48. Vangunsteren, W.F. & Berendsen, H.J.C. Algorithms for Macromolecular Dynamics and Constraint Dynamics. *Mol Phys* **34**, 1311-1327 (1977).
49. Sugita, Y. & Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chem Phys Lett* **314**, 141-151 (1999).
50. Kim, D.E., Fisher, C. & Baker, D. A breakdown of symmetry in the folding transition state of protein L. *J Mol Biol* **298**, 971-984 (2000).