Gene length as a regulator for ribosome recruitment and protein synthesis: theoretical insights

Lucas Dias Fernandes[1,2], **Alessandro de Moura**[2], **and Luca Ciandrini**[3,4,*]

[1]Departamento de Entomologia e Acarologia, Escola Superior de Agricultura ”Luiz de Queiroz” - Universidade de São Paulo, ESALQ - USP, 13418-900, Piracicaba/SP, Brazil
[2]Institute for Complex Systems and Mathematical Biology, University of Aberdeen, Aberdeen, AB24 3UE, UK
[3]DIMNP UMR 5235, Université de Montpellier and CNRS, F-34095, Montpellier, France
[4]Laboratoire Charles Coulomb UMR5221, Université de Montpellier and CNRS, F-34095, Montpellier, France
[*]luca.ciandrini@umontpellier.fr

ABSTRACT

Protein synthesis rates are determined, at the translational level, by properties of the transcript's sequence. The efficiency of an mRNA can be tuned by varying the ribosome binding sites controlling the recruitment of the ribosomes, or the codon usage establishing the speed of protein elongation. In this work we propose transcript length as a further key determinant of translation efficiency. Based on a physical model that considers the kinetics of ribosomes advancing on the mRNA and diffusing in its surrounding, as well as mRNA circularisation and ribosome drop-off, we explain how the transcript length may play a central role in establishing ribosome recruitment and the overall translation rate of an mRNA. We also demonstrate how this process may be involved in shaping the experimental ribosome density-gene length dependence. Finally, we argue that cells could exploit this mechanism to adjust and balance the usage of its ribosomal resources.

Introduction

mRNA translation is, together with transcription, the pillar of the central dogma of molecular biology. In spite of its key role in protein synthesis, the accurate understanding of its dynamical details still remains elusive at the present time, and the sequence determinants of mRNA translation efficiency are not fully understood[1,2]. Initiation of translation regulates the recruitment of ribosomes and it is believed to be modulated by mRNA secondary structures[3,4], while protein elongation is mainly considered to be regulated by tRNA abundances determining the pace of the ribosome[5–7]. The individual steps of translation are thought to be well understood, yet there is no reliable approach quantitatively predicting the overall protein synthesis rates for a given transcript.

A better understanding of the molecular mechanisms of mRNA translation will unravel the physiological determinants of translation efficiency. Besides, this knowledge will be extremely useful in developing applications in synthetic biology and will allow tight control on the average production of a protein and on its expression noise.

The translation efficiency of a transcript is often identified with its experimentally observed polysome state and ribosome density (amount of ribosomes per unit length on the mRNA), meaning that transcripts with high ribosome density are more efficiently translated[8]. Remarkably, many experimental observations show that the ribosome density is related to the length L of the coding sequence (CDS): the longer the mRNA, the smaller the ribosomal density. This indicates the presence of a length-dependent control of translation. As we show in Figure 1, the observation that average ribosomal densities ρ strongly anti-correlate to CDS lengths L appears to be a conserved feature across many organisms, ranging from unicellular systems such as *L. lactis*[9], *S. cerevisiae*[10–12] or *P. falciparum*[13], to more complex organisms such as mouse and human cells[14,15]. The common traits in the density-length dependence suggest that this relationship is dictated by universal mechanisms underlying the translation process.

However, this remark has been strangely overlooked in the literature (with the exception of Guo *et al.*[16]), particularly in the theoretical literature trying to provide models of mRNA translation. A few hypotheses have been proposed to justify the emergence of the length dependence of the ribosome density, which requires a regulation apparatus acting at the initiation[17] or at the elongation stage[12]. These hypotheses have not been examined with the support of a mechanistic model and a mathematical approach. In contrast with previous studies, here we explain qualitatively and quantitatively the relationship between ribosome density and CDS length, making the point that transcript length is a critical determinant of protein synthesis rates.

In Figure 1 we show a log-log plot of measured ribosomal densities as a function of the CDS length for different organisms. The figure suggests a power-law behaviour (L^{-1} is drawn as reference). However, extracting a scaling law from this kind of

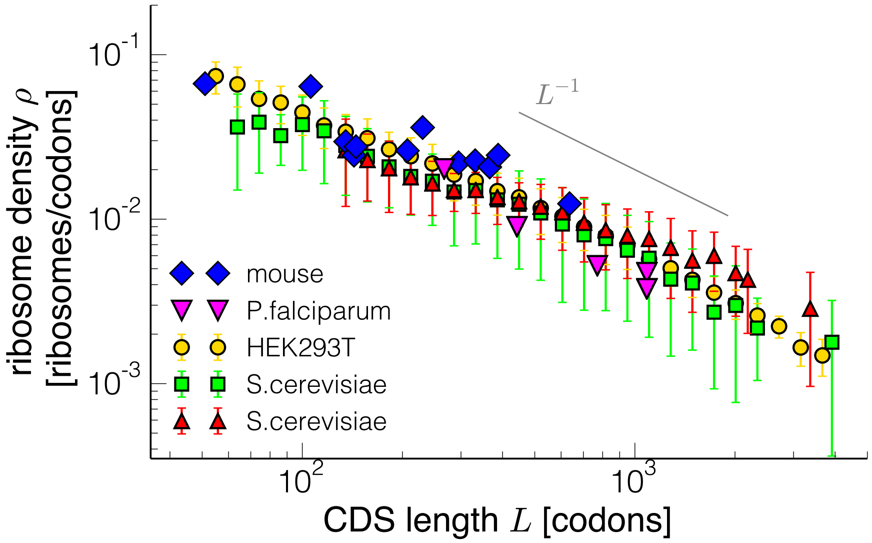

Figure 1. Ribosome density vs CDS length for different datasets. Blue diamonds (mice[14]) and fuchsia down triangles (*P. falciparum*[13]) are individual genes, while yellow circles (HEK293T[15]), green squares (*S. cerevisiae*[10]) and red triangles (*S. cerevisiae*[11]) are length-binned data for the entire genome, with the error bars representing the standard deviation for each bin. The gray line indicates the behaviour of a power law with exponent -1.

data is a phenomenological (and probably a too simplistic) description of this relationship that, moreover, can only be measured for a few decades in L.

Instead, in this paper we propose a mechanistic explanation for the length dependence of translation that is found in experimental data; in the last part of this work we show how this mechanism could be exploited by the cell to adjust and balance its ribosomal resources.

Results

A stochastic model of translation

We model translating ribosomes as particles moving on a unidimensional discrete track representing the mRNA, as depicted in Figure 2A. In this model particles are injected from one side of the lattice (the 5' end of the mRNA) with a rate α, then advance one site (one codon) with rate p, and are removed at the last site (stop codon) with rate β (Figure 2A). The first step mimics the recruitment of ribosomes (initiation); elongation is given by the dynamics of ribosomes in the bulk; the exit of particles from the last site represents the termination. MacDonald and coworkers introduced this class of model at the end of the 60's precisely in an attempt to mathematically describe the process of mRNA translation[18]. Since then, under the name of *exclusion process*, this model has been extended and thoroughly studied from a theoretical point of view; it became an emblematic framework in out-of-equilibrium physics[19], for which an exact solution is known in the simplest formulation[20].

In the last years, revamped extensions of the exclusion process have appeared in the literature, developed to provide more quantitative models of translation[21–24], and many works are nowadays implicitly based on this framework[25,26]. Here we first look into a variant of the exclusion process that considers particles covering $\ell = 10$ sites of the track[27], as the ribosome footprints cover around 28 nucleotides[12]. Details of the model and simulations can be found in the Materials and Methods section and in the Supplementary Material.

Translation efficiency corresponds to the ribosomal current

From the analytical solution of the exclusion process or by Monte Carlo simulations we can estimate, as a function of the initiation rate α, termination rate β and codon elongation rate p, the two quantities of interest for the translation process: the ribosomal density $\rho(\alpha,\beta,p)$, defined as the average number of ribosomes N divided by the CDS length L, and the ribosomal current $J(\alpha,\beta,p)$, defined as the average number of ribosomes advancing one site in a unit of time. Those quantities can be compared to experimental measurements of polysome profiles and protein production rates. The ribosomal current J corresponds in fact to the protein production rate per mRNA (proteins produced per unit time per mRNA), and we choose to identify

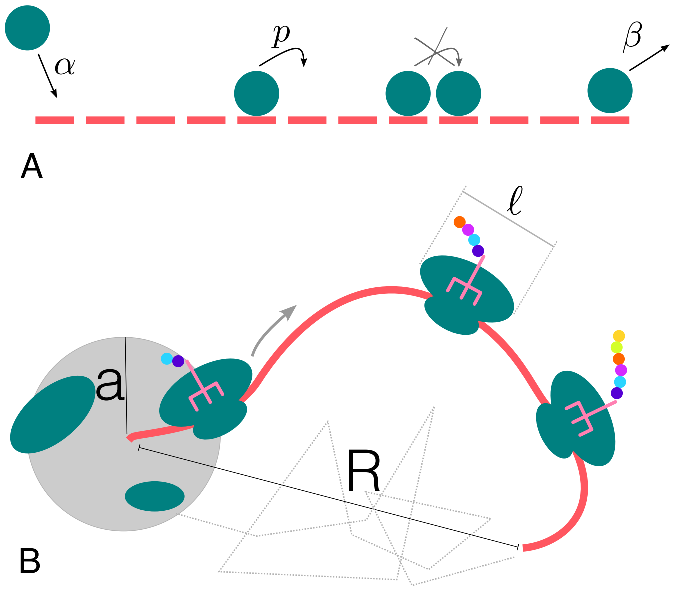

Figure 2. Sketch of the translation model. In the standard exclusion process (**A**) particles can enter the beginning of the lattice with a rate α, move from one site to the next one with rate p (provided that it is not occupied by another particle), and then exit on the last site with rate β. In this study we consider a more refined version of the model (**B**) in which ribosomes cover ℓ sites (codons) and the unidimensional lattice is placed in a three-dimensional environment. R represents the end-to-end distance between the 5'and the 3' region, and a is the radius of the reaction volume for initiation.

it as a better descriptor for the *translation efficiency*. From the analytical solutions one can show that efficiency J and density ρ only depend on $\bar{\alpha} \equiv \alpha/p$ in the regimes analysed in this work (see Supplementary Material). Hence $\rho(\alpha,\beta,p) = \rho(\bar{\alpha})$ and $J(\alpha,\beta,p) = J(\bar{\alpha})$.

In order to make the model more realistic and compare it to experimental datasets, we need to determine the initiation rate α. The estimation of the translation initiation rate has been previously attempted for a few organisms[24,28], and these studies have found a dependence of the initiation rate on the transcript length: the longer the transcript, the weaker the initiation. Here we propose a model that is able to explain this observation by coupling the translation process, in particular translation initiation, to the three-dimensional conformation of the mRNA. We will show how a feedback mechanism controlled by the polysome compaction could induce a length-dependent initiation rate and hence a length-dependent density. Before that, we need to introduce some basic properties of the transcript's spatial conformation.

Transcript end-to-end distance depends on the polysome size

We consider the transcript as a polymer that assumes different spatial conformations and a characteristic 5'-3' end-to-end distance R (see Figure 2B). An undecorated mRNA (without ribosomes on it) can be considered as a polymer with a persistence length $l_p \simeq 1$ nm $\simeq 1$ codon[29], and the average end-to-end distance R can be estimated from basic principles of polymer physics. By assuming an underlying random walk one obtains that the end-to-end distance R depends on the length L of the mRNA as

$$R = \sqrt{2Ll_p}. \tag{1}$$

However, the stiffness produced by the large size ℓ of the ribosomes can drastically change the persistence length of the mRNA. We assume that the persistence length of an mRNA depends on the polysome state via an average between ℓ (a typical ribosome footprint) and l_p (persistence length of an empty mRNA), weighted by the fraction $f = \rho\ell$ of the transcript covered by ribosomes (at the steady-state). After these considerations we write the effective persistence length of the mRNA as

$$\begin{aligned} l_{\text{eff}} =& f\ell + (1-f)l_p \\ =& \ell^2\rho + (1-\rho\ell)l_p, \end{aligned} \tag{2}$$

which is equal to l_p when the mRNA is empty and reaches ℓ when the ribosomal density attains its maximal value $\rho = 1/\ell$. Substituting this value of the effective persistence length into Eq.(1) we obtain the average end-to-end distance of a polysome as a function of the density ρ and the length L:

$$R = \sqrt{2L}\left[\ell^2\rho + (1-\rho\ell)l_p\right]^{1/2}. \tag{3}$$

This way, we have used a coarse-grained model to couple the state of the polysome to its spatial conformation. Intuitively, Eq.(3) means that a translated transcript with many ribosomes on it will be more stretched (so the distance R between 5' and 3' ends will be larger) compared to a situation with a small ribosomal density or an empty mRNA. We have neglected potential formation of secondary structures inside the coding region. Such structures would only slightly decrease the effective length of the sequence (a few codons) and we treat only translationally active transcripts, meaning that moving ribosomes (10-20 codons/s) continuously unfold those structures.

When initiation is limiting, the density ρ of ribosomes is fixed by $\bar{\alpha}$, and via Eq. (3) we are hence able to determine the dependence of the end-to-end distance as a function of the initiation rate and the length of the transcript, $R(\bar{\alpha}, L)$.

Initiation can be enhanced by a length-dependent mechanism

We will consider that the magnitude of the initiation rate α is determined by the concentration c of free ribosomal subunits via a first rate equation: $\alpha = \alpha_0 c$, with α_0 being the initiation rate constant depending, for instance, on the affinity between the ribosome and the 5' UTR binding site on the mRNA. The concentration c is the local concentration of subunits in the reaction volume of radius a around the ribosome binding site at the 5' end of the transcript; we introduce c_∞ as the homogeneous concentration of free subunits far from the transcript. The local concentration c in the reaction volume is affected by the ribosomes terminating translation at the 3' end of a transcript, then diffusing into the reaction volume and contributing to the abundance of free ribosomal subunits in this volume. The contribution c_R to the local concentration c due to this feedback mechanism depends on the end-to-end distance R previously calculated in Eq. (3). It can be shown for different organisms, considering typical values of transcript numbers and cytoplasm volume, that the average end-to-end distance is smaller than the average separation between transcripts. This corroborates the intrinsic assumption that the translating processes of distinct mRNAs do not interfere with each other. Thus each individual mRNA can be thought of as a sink-source system for ribosomes (the ribosome binding site representing the sink and the ribosome termination site representing the source) immersed in an environment with a constant background ribosomal concentration c_∞. Hence, the local ribosome concentration around the ribosome binding site can be written as $c = c_\infty + c_R$.

By considering ribosomes as particles performing free diffusion when they are not bound to the mRNA (see the Supplementary Material) we can obtain a mathematical expression of the initiation rate as a function of the system's parameters. Regarding translation as a steady state process with ribosomal density and current values given by ρ and J respectively, we find the initiation rate $\bar{\alpha}$ (see Supplementary Material for a complete derivation):

$$\bar{\alpha} = \bar{\alpha}_0(c_\infty + c_R) = \bar{\alpha}_\infty + \lambda \frac{J(\bar{\alpha})}{R(\bar{\alpha}, L)}, \tag{4}$$

where we have emphasised the dependence of the ribosomal current J on $\bar{\alpha}$. Similarly, the end-to-end distance R also depends on $\bar{\alpha}$ and L as shown in Eq. (3).

We highlight that the parameter $\bar{\alpha}_0$ and thus $\bar{\alpha}_\infty$ and λ depend on the binding between the ribosome and the mRNA, which is supposed to be mainly regulated by mRNA secondary structures. We will consider the parameters of the model to be independent on the transcript length. This is justified by the a weak correlation (Pearson $r = -0.01$, p-value 0.5)[30] between free energies of secondary structures in the 5'UTRs and the transcript length L (see also Figure S5 in the Supplementary Material). The parameter $\bar{\alpha}_\infty$ is adimensional and $\alpha_\infty = p\bar{\alpha}_\infty$ represents the initiation rate without the feedback mechanism, or equivalently when the end-to-end distance R is large enough to make this mechanism negligible. The parameter λ characterises the strength of the feedback. It has the dimensions of a length and it corresponds to the typical separation between 5' and 3' below which the feedback mechanism becomes relevant. We measure this parameters in units of codon length, which roughly corresponds to 1 nm.

Consequently, the current of ribosomes leaving the end of a transcript increases the concentration of ribosomal subunits around their binding region; through modulation of the mRNA stiffness due to the ribosome load, this feedback leads to initiation rates that are strongly length-dependent. Equation (4) is an implicit equation that can be numerically solved to obtain the initiation rate, and thus the density $\rho(\bar{\alpha})$ and the current as a function of L. We also developed a simulation scheme that allows us to fix the initiation rate via a self-consistent method.

Although Eq. (4) considers the ribosome as a single diffusing particle, we can explicitly consider the diffusion of the two ribosomal subunits. This would generate a dependence on R^2 instead of R in Eq. (4). However, the qualitative behaviour of our results does not significantly change and for the sake of simplicity we decided to present the outcomes of the theory described by Eq. (4). We include the analysis of this more refined model in the Supplementary Material (see also Figure S8 and S11).

Experimental density-length dependence emerges from initiation enhanced effects

We then compare the outcome of the model to experimental measurements of ribosome densities. The result of this analysis is shown in Figure 3: the predicted ribosome density is quantitatively comparable to the experimental quantification, and our mechanistic model based on basic physical principles is able to capture the length dependence of the ribosome load. Our theory could therefore explain the observed length dependence of the translational properties.

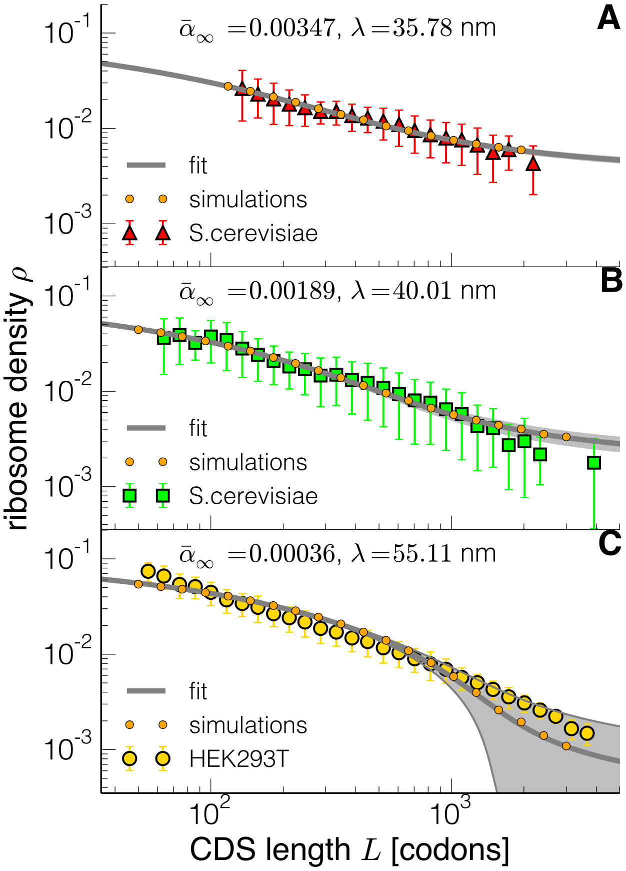

Figure 3. Comparison between theory and experimental ribosome densities in yeast (**A-B**)[10, 11] and human embryonic kidney cells (**C**)[15]. The symbols and datasets correspond to the ones of Figure (1). The grey lines represent the best fit of the model (the parameter values are written in each panel), while the shadow areas correspond to the regions spanned within the margin of error of the estimated $\bar{\alpha}_\infty$. Orange circles are the outcome of stochastic simulations used to test the numerical solution of the equation using the parameters obtained from the best fit.

We fit the expression of $\rho(\bar{\alpha})$ (continuous lines in Figure 3) and obtain the two parameters $\bar{\alpha}_\infty$ and λ for three available datasets (two yeast datasets[10, 11] and a human embryonic kidney cells dataset[15]), then check the accuracy of the solution with the stochastic simulation scheme developed as explained in the Material and Methods section. Taking the standard errors of the parameter estimation there is no significant deviation between the data and the fitted model(details can be found in the Materials and Methods). However, for the last set, there is a stronger dependence of the solution on the parameter $\bar{\alpha}_\infty$: shaded regions in Figure 3 represent the solution considering the standard error of this parameter. This could explain the slight deviation between theory and experimental data for large mRNAs. We also emphasise that the parameter $\bar{\alpha}_\infty$, depending on the global availability of ribosomes, is supposed to be the most affected by experimental variations (for instance by growth rate dependence or cell cycle stage). The estimation of $\bar{\alpha}_\infty$ is more accurate when we take into account the diffusion of the two subunits (see Supplementary Material).

Our simulations also allow us to extract the amount of ribosomes bound to a transcript, from which we can extract the monosome:polysome ratio, which is also subject to length-effects (see Supplementary Material). By increasing the CDS length we observe a reduction of the monosome:polysome ratio following a power-law like behaviour. A recent work[31] has identified, by merging polysome and ribosome profiling techniques, the amount of active monosomes, i.e. mRNAs with only one ribosome elongating the protein. Consistently with our findings, the monosome:polysome ratio also shows signs of a marked anti-correlation with the mRNA length.

Ribosome drop-off cannot alone be responsible for the density-length dependence

In this section we study the consequences of ribosome drop-off[32,33], so far neglected in our model, on the observed density-length dependence. In order to do that, we performed simulations including ribosome drop-off at a rate $\delta = 10^{-3}$ s^{-1}. This is justified by the estimated drop-off rates of the order of 10^{-4}/codon[32,33], and by the codon elongation rates we considered in this paper of the order of 10 codons s^{-1}. Figure 4 shows that the simulations of the process with drop-off (full lines) do not largely differ from the model without drop-off (dashed lines), and the deviations starts to become relevant for large sequences (order 10^4 codons). For such lengths the extended model is actually reproducing the experimental behaviour even better than the model without drop-off.

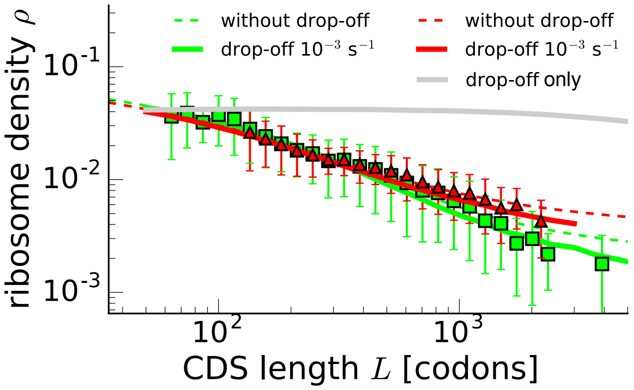

Figure 4. Model with ribosome drop-off. Green lines correspond to the Arava dataset[10], while red ones correspond to the Mackay dataset[11] Symbols of experimental points in yeast correspond to the ones of Figure 1, while dashed lines represent the solutions of the model described in the previous section with the same parameters used in panels A and B of Figure 3. The continuous lines are the outcome of simulations of our model including ribosome drop-off at a rate of 10^{-3} s^{-1}, and the grey line shows the outcome of the simulations with drop-off only.

To further exclude the possibility that the length dependence rises from ribosome prematurely leaving the transcript, we simulated ribosome drop-off occurring during the translation process without the feedback mechanism that we propose (basically, we simulated a standard TASEP with particle detachment rate δ). This is represented by the grey line in Figure 4. We were not able to obtain any length dependence for biologically relevant values of the drop-off rate (see Suppelementary Figure S9), and we can therefore conclude that the behaviour observed in Figure 1 cannot originate by ribosomal drop-off alone.

mRNA circularisation does not change the phenomenology of the model

The 5' and 3' end of eukaryotic mRNAs interact with each other via protein-protein interaction, for instance between the Poly(A)-binding protein PAPB and the initiation factor eIF4F bound at the 5' cap; this coupling is believed to induce the formation of transcripts with circular structures[34]. However, depending on the energies at play, the transcript dynamically switches between an *open*, linear state, and a *closed*, circularised state (see Figure 5A). When in the closed state, the end-to-end distance R will be only of a few nanometers (the order of magnitude of the two molecular partners supposedly involved in this interactions). We will denote as d the distance between 5' and 3' in the closed state. The mRNA is found in its circularised state with a given probability P_c, and in an open state with probability $P_o = 1 - P_c$, with a difference in free energies between the two states $\Delta G = G_c - G_o$.

To find how the average end-to-end distance is affected by this interaction we weight its value in the closed and open state with their respective probabilities:

$$R_{\text{circ}} = P_o R + P_c d \,. \tag{5}$$

If $P_o \approx 1$ then our feedback model well approximates the mRNA translation process. While in prokaryotes we can likely assume $P_o = 1$, this is probably an oversimplification for eukaryotes. In order to consider transcript circularisation we have now to compute how P_o depends on the CDS length L and on the interaction energy ε (in $k_B T$ units) between the two ends. The intuitive explanation we used before to determine differences in local concentrations of ribosomes close to the 5' end can be reproduced here to compute P_c. For a fixed ε and a short mRNA, we expect to find circularised transcripts with a probability P_c close to one; in contrast, very large transcripts should be hardly found in the close state. This length dependence will also contribute to the ribosome density behaviour observed in experiments (Fig. 1). We computed P_c as a function of L, ℓ_{eff} and ε, which turns out to be:

$$P_c = \frac{1}{1+e^{\beta\Delta G}} = \frac{1}{1+\left[\left(\frac{l_{\text{eff}}L}{d^2}\right)^{\frac{3}{2}}\sqrt{\frac{4\pi}{3}}-1\right]e^{\left(\frac{2\pi^2 l_{\text{eff}}}{L}+\varepsilon\right)}}. \tag{6}$$

The details of the calculation of P_c can be found in the Supplementary Material. Here ε is considered as a fitting parameter. By inserting Eq.(6) in Eq.(5) and computing the end-to-end distance to be plugged in Eq. (4) we eventually find, now as a function of $\bar{\alpha}_\infty$, λ and ε, how the initiation rate is affected by the concentration increase of ribosomes in the 5' reaction volume with also considering transcript circularisation. We have then fitted $\rho(\bar{\alpha})$ to the dataset we have previously used, and the outcome is shown in Figure 5B (fitting values and other datasets can be found in the Supplementary Material). We did not find a significative difference from the best fit of the simpler model previously introduced (dashed line in Figure 5B). In Figure 5C we show how the probabilities of finding a circularised or open mRNA depend on the transcript length, together with the end-to-end distance R. The results agree with our intuitive explanation.

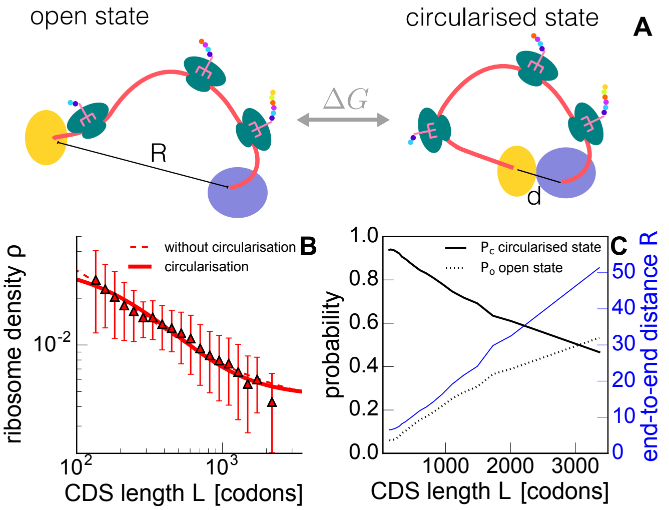

Figure 5. Model with mRNA circularisation. Sketches of the two possible mRNA conformations, open and circularised, whose transitions depend on the free energy gap ΔG (**A**). In blue and yellow we have represented the interacting proteins (e.g. PABP and eIF4F) bound at the 3' and 5' ends; the black line is the end-to-end distance that is equal to d in the circularised state. We have fixed $d = 5$ nm in our calculation. Ribosome density computed taking into account mRNA circularisation (continuous line) is then compared to experimental data (triangles, cfr. symbols used in Figure 1) and the previous model neglecting circularisation (dashed line) (**B**). The fitted parameters are $\bar{\alpha}_\infty = (4.7 \pm 0.6)\ 10^{-3}\ \text{s}^{-1}$, $\lambda = 7.0 \pm 0.6$ nm and $\varepsilon = -8.3 \pm 0.4$ (k_BT). End-to-end distance (blue curve, right axis) and calculated probabilities P_c and $P_o = 1 - P_c$ as a function of the CDS length L (**C**).

Length-dependent competition for resources

In this section we discuss some potential applications of our model related to bacterial growth laws[35]. Specifically, we will show that cells can adjust, based on a length-discriminatory mechanism, their relative expression of genes at different ribosome concentrations. This result from our model leads us to predict a new regulation mechanism for gene expression.

We observe that changes in ribosome densities (or in translation efficiency) induced by changes in the ribosomal pool are conditional on the mRNA length (Figure 6A). Short transcripts are less affected by the amount of available ribosomes compared to long ones, suggesting a possible mechanism to regulate the relative protein production rate at different growth rates based on transcript length only. As a proof of principle, in Figure 6B we plot the relative expression $\eta(L_1, L_2) = J_{L_1}/J_{L_2}$ between the

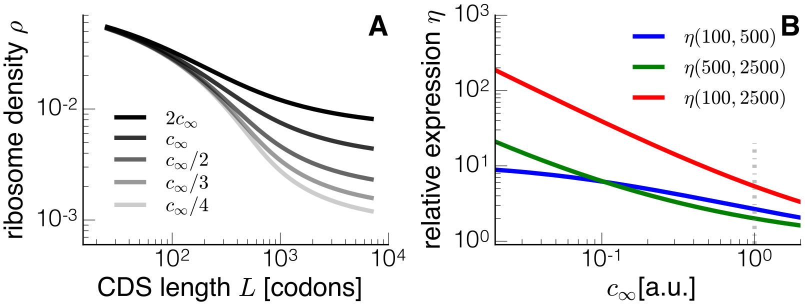

Figure 6. Effects of competition for resources (ribosomes) on the protein production rate. The ribosome density depends on the overall ribosome concentration c_∞. We show the ribosome density as a function of the transcript length for different concentrations of available ribosomes c_∞ (**A**). The curve denoted with c_∞ in the legend is built starting from the same parameters of Figure 3A. We change c_∞ as described in the legend for the other curves. Short transcripts are less affected by changes in c_∞, as we also show in (**B**), where we plot the relative expression of transcripts η (defined in the text) as a function of c_∞. We used transcript with three different lengths (here $L = 100, 500$ and 2500). According to these results, ribosomal proteins that are short should be relatively more expressed under high ribosome competition regimes compared to other types of proteins.

translation efficiencies of transcripts with lengths L_1 and L_2. When ribosomes are limiting, short transcripts are relatively more translated than long ones.

This behaviour can be intuitively explained. The main contribution to the initiation of long mRNAs is the concentration of free ribosomes c_∞, meaning that long mRNAs are largely influenced by changes in the ribosomal pool. Short transcripts instead can more efficiently take advantage of the length dependent contribution c_R to initiation.

This constitutes a mechanism for regulating the relative expression between short and long genes. Ribosomal proteins are composed only of a few dozens of amino-acids and, as a consequence, our theory predicts that at least qualitatively ribosomal proteins should be proportionately more efficiently translated than longer proteins under strong ribosome competition regimes, i.e. low growth rates. This difference should decrease when ribosomal resources get less tight, and in the limit of infinite ribosomal resources, η should tend to 1 since the length dependence becomes less and less relevant. Our model suggests a translational mechanism to relatively over express short proteins (e.g. ribosomal proteins) at the cost of longer ones.

Discussion

Many aspects of translation are still puzzling researchers. Theoretical approaches propose designing principles for modulating the translation efficiency at the level of initiation[4] and of elongation[24], mainly based on the role of RNA secondary structures, codon bias or amino acid properties. However, when tested on synthetic constructs, the hypotheses underlying the theoretical models are often contradicted[8], so that the identification of transcript-dependent determinants of translation efficiency is still debated in the literature. In this work we have identified and studied another factor modulating the translation efficiency: the length of the transcript.

We connected the emergence of the ribosome density-mRNA length dependence shown in Figure 1 to a mechanistic model built on basic physical principles and on the properties of the translation process. Our theory then establishes a link between densities and mRNA lengths and explains experimental data with an excellent agreement without invoking an evolutionary selection of genes based on their length. As a matter of fact, a selection process towards short efficient genes to improve cellular fitness could also be conjectured[36,37]. Without direct experimental observation we cannot rule out this hypothesis, although this would not explain the same behaviour observed for different organisms (experimental data in Figure 1 seems to collapse to a unique universal curve). Moreover, the poor correlation between free energy of secondary structures at the 5' end of an mRNA and CDS length is a signature that binding sites are not significantly weaker for long genes, as it would assume an evolutionary argument. Instead our theory predicts, as an outcome, that short genes have larger initiation rates compared to long ones.

Experimental results in fact suggest that translation initiation is also dependent on mRNA length[10,17]. Since the ribosome recruitment rate must depend on the local concentration of ribosomal subunits around the 5'UTR, we conjecture that the local concentration is modulated by the CDS length via a feedback mechanism coupling protein synthesis and initiation rates. We can roughly name this process "recycling", as subunits terminating translation will contribute to the increase of the local concentration c_R and thus be more easily re-used as sketched in Figure 2. This coarse-grained physical model has allowed us to reproduce the scaling behaviour of experimental ribosome densities (Figure 3) and it constitutes, to our understating, the first quantitative explanation of how translational features are affected by the transcript length. Previous works have studied the effect of particle recycling[38–41] but, with the exception of Chou (2003)[38], they do not explicitly compute how the recycling term is regulated by the end-to-end distance R.

The compaction of the transcript (here characterised by the end-to-end distance), also depends on its polysome state. Intuitively, an mRNA with many translating ribosomes will be more stretched than an empty mRNA. We captured this feature by introducing a ribosome density dependence on the end-to-end distance through Eq. (3). We emphasise that this is the simplest choice for coupling elongation properties and the three-dimensional conformation of the mRNA, and one could introduce more complicated relationships linking end-to-end distance, ribosome density and elongation rate; here we wanted to show that, as a proof of principle, by a feedback mechanism enhancing initiation we can reproduce experimental data very well.

We have also studied how the results change by explicitly considering the diffusion of the two ribosomal subunits, and we found no significant change (see Supplementary Material). To make the model more realistic we considered two further extensions of the model. We considered (i) ribosome drop-off and (ii) mRNA circularisation. The former brought an improvement in the comparison between data and theory for large transcripts only (see Figure 4), while the inclusion of the latter did not lead to a particular phenomenological change of the model's outcomes (Figure 5). This suggests that ribosome recycling, as considered in the basic model, is the fundamental element originating the length-dependence translation.

We have then speculated on how the length could be exploited to create differences in the relative expression of genes at different growth rates, here used as measure of the free ribosomes abundances. When resources are constrained, i.e. when the amount of free ribosomes c_∞ is small, competition for resources might become relevant[42–44] and our theory predicts that the length-dependent term of the initiation rate dominates the process. In other words, when ribosomes are strongly limiting, ribosome recruitment is mainly due to recycled ribosomes, meaning that short transcripts can better capitalise the resources. This mechanisms could also be a way to translationally favour the production of ribosomal proteins (which are short) in a scenario of deficiency of ribosomes. In order to formulate this hypothesis, we neglect known mechanisms responsible for ribosome biogenesis, a complex process that is beyond the scope of our work. Our conjecture has then to be interpreted in the perspective of ribosomal concentrations fixed by a certain amount of ribosome production (established, for instance, by the richness of the growth medium): for a given concentration of ribosomes we make strong predictions on how the ribosomal pool should be partitioned among the different transcripts with just a length-discrimination mechanism.

The model could be further extended to consider translation of bacterial operons: in this case, in fact, one transcript is composed of different sinks (ribosome binding sites) and sources (stop codons) of ribosomes, while here we have discussed the case of a transcript translating a single gene (with one ribosome binding site and one stop codons). Having an operon translating different genes will increase the complexity of the feedback term, and it could in principle create counter-intuitive phenomenologies.

Our findings are compared to experimental ribosomal densities, and our framework can quantitatively reproduce the measurements. We have used our model to estimate the protein production rates of synonymous genes, and the method was successful (see Figure S12). In this work we have used our model to predict the ribosome-length dependence, i.e. we have emphasised the dependence of Eq. (4) on L, but our theory predicts that there is a feedback between elongation (codon usage) and initiation, that we have exploited in Figure S12. To further test the model, it will be necessary to make fusions of a reporter gene with peptides of variable lengths, and then measure ribosome density or translation efficiency with the aim of experimentally reproducing Figure 3 in a controlled manner. However, one should pay attention to the changes in mRNA degradation, translation elongation and initiation induced by the added nucleotides coding the fused peptides. Figure 6B also constitutes a good way to test our hypotheses. One of our predictions is the relative change of expression of short/long transcripts when changing the cellular growth rate. This could be obtained by concurrently expressing two reporter genes of different lengths, and measuring their relative expression at different growth rates obtained by changing growth medium or by different antibiotics.

Methods

The exclusion process

We base our model on the exclusion process, which is also introduced in the section Results and in Figure 2A. More accurately, this model is known in the literature as TASEP: Totally Asymmetric Simple Exclusion Process, for which nowadays there exists a plethora of extensions applied in many different fields, from vehicular traffic to intracellular transport[19]. Each site of the track in Figure 2A corresponds to a codon, and the particles can advance from site to site -provided that the next site is not occupied by another particle- mimicking the elongation process. We give a thorough description of the exclusion process and the known results in the Supplementary Material.

To simulate the dynamics of the exclusion process we used a kinetic Gillespie-like Monte Carlo as used in Ciandrini *et al.*[24].

Fitting and numerical solutions

We substitute the expression for $\bar{\alpha}$, Eq. (4), in the equation for the density $\rho(\bar{\alpha})$ in the low density phase of the ℓ-TASEP (see Supplementary Material). The current J is given by the $J(\rho)$ correction in the ℓ-TASEP and the end-to-end distance is R found in Eq.(3). Thus, we obtain an implicit equation $\rho = \rho(\bar{\alpha}_\infty, \lambda, L)$ that can be numerically solved for each set of variables $\{\bar{\alpha}_\infty, \lambda, L\}$ and used to fit the experimental data $\rho_{\mathrm{exp}}(L)$ to obtain the parameters $\bar{\alpha}_\infty$ and λ for each dataset, and their standard errors (see Supplementary Material).

We have used built-in functions of Mathematica[45] to obtain numerical solutions for the density and currents and to fit the three datasets used in this study.

Density and current via a self-consistent simulation scheme

Equation (4) allows us to obtain, via simulation, the values for ρ for different values of L, taking into account the feedback mechanism coupling protein synthesis and initiation rate and finite size effects (the later intrinsic to the numerical simulations). We obtain this with the following self-consistent method:

(i) We initialise the system with an arbitrary value of $\alpha = \alpha^{(0)}$, let the system evolve until the steady-state is reached and then evaluate the current $J^{(0)}$ and the density $\rho^{(0)}$;

(ii) Compute R as in equation (3) and update $\alpha = \alpha^{(1)}$ according to equation (4) with $J^{(0)}$ and $\rho^{(0)}$ computed in (i);

(iii) Repeat the previous points for several iterations until $|\alpha^{(i)} - \alpha^{(i-1)}|/\alpha^{(i)} < 0.01$ (in general less than 10 iterations are needed to make the algorithm converge);

(iv) The final value of α is then used to obtain the final densities and currents.

With this iteration process, for a given choice of the parameters α_∞ and λ, we can obtain the steady-state density and current, which vary with the length L of the transcript, due to the joint contribution of recycling and finite-size effects.

This self-consistent method allows us to simulate the system without explicitly considering, thanks to Equation (4), the diffusion of particles when they are not bound to the lattice and the dynamics of the mRNA.

Choice of datasets

We restricted our analysis to measures made by sucrose gradient methods. We are aware that a more recent technique like ribosome profiling[12] would provide ribosome densities with codon resolution, but this method does not provide an *absolute* ribosome density (see definition below). The length-correlation has been shown to hold in ribosome profiling experiments[12]. Instead of assuming arbitrary normalisation of ribosome footprints to match our theory, we analysed absolute ribosome densities that are available in the literature. For the Mackay *et al.* dataset[11] we used the *reliable* subset of data.

Definitions

We define the ribosome density to be the number N of translating ribosomes divided by the length L (expressed in number of codons) of the CDS. We embrace this definition instead of alternative ones (ribosomes per 100 or 1000 nucleotides) for practical reasons. Thus, the density $\rho \equiv N/L$ is expressed in ribosomes per codons, and it can be thought of as the probability of a codon being covered by the the A-site of the ribosome (i.e., a codon being translated). For steric reasons, this density is bound by $1/\ell$, where ℓ is the length of the ribosome footprint (in codons). For instance, $\ell \sim 10$ in *S. cerevisiae*.

References

1. Gingold, H. & Pilpel, Y. Determinants of Translation Efficiency and Accuracy. *Molecular Systems Biology* **7**, 481 (2011). DOI 10.1038/msb.2011.14.

2. Quax, T. E. F., Claassens, N. J., Söll, D. & van der Oost, J. Codon Bias as a Means to Fine-Tune Gene Expression. *Molecular Cell* **59**, 149–161 (2015). DOI 10.1016/j.molcel.2015.05.035.

3. Kudla, G., Murray, A. W., Tollervey, D. & Plotkin, J. B. Coding-Sequence Determinants of Gene Expression in Escherichia coli. *Science* **324**, 255–258 (2009). DOI 10.1126/science.1170160.

4. Salis, H. M., Mirsky, E. A. & Voigt, C. A. Automated design of synthetic ribosome binding sites to control protein expression. *Nat Biotech* **27**, 946–950 (2009). DOI 10.1038/nbt.1568.

5. Plotkin, J. B. & Kudla, G. Synonymous but not the same: the causes and consequences of codon bias. *Nat Rev Genet* **12**, 32–42 (2011). DOI 10.1038/nrg2899.

6. Kemp, A. J. *et al.* A yeast tRNA mutant that causes pseudohyphal growth exhibits reduced rates of CAG codon translation. *Molecular Microbiology* **87**, 284–300 (2013). DOI 10.1111/mmi.12096.

7. Gorgoni, B., Ciandrini, L., McFarland, M. R., Romano, M. C. & Stansfield, I. Identification of the mRNA targets of tRNA-specific regulation using genome-wide simulation of translation. *Nucl. Acids Res.* gkw630 (2016). DOI 10.1093/nar/gkw630.

8. Li, G.-W. How do bacteria tune translation efficiency? *Current Opinion in Microbiology* **24**, 66–71 (2015). DOI 10.1016/j.mib.2015.01.001.

9. Picard, F. *et al.* Bacterial translational regulations: high diversity between all mRNAs and major role in gene expression. *BMC Genomics* **13**, 528 (2012). DOI 10.1186/1471-2164-13-528.

10. Arava, Y. *et al.* Genome-wide analysis of mRNA translation profiles in Saccharomyces cerevisiae. *PNAS* **100**, 3889–3894 (2003). DOI 10.1073/pnas.0635171100.

11. MacKay, V. L. *et al.* Gene expression analyzed by high-resolution state array analysis and quantitative proteomics: response of yeast to mating pheromone. *Mol. Cell Proteomics* **3**, 478–489 (2004). DOI 10.1074/mcp.M300129-MCP200.

12. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S. & Weissman, J. S. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324**, 218–223 (2009). DOI 10.1126/science.1168978.

13. Lacsina, J. R., LaMonte, G., Nicchitta, C. V. & Chi, J.-T. Polysome profiling of the malaria parasite Plasmodium falciparum. *Mol. Biochem. Parasitol.* **179**, 42–46 (2011). DOI 10.1016/j.molbiopara.2011.05.003.

14. Cataldo, L., Mastrangelo, M. A. & Kleene, K. C. A quantitative sucrose gradient analysis of the translational activity of 18 mRNA species in testes from adult mice. *Mol. Hum. Reprod.* **5**, 206–213 (1999).

15. Hendrickson, D. G. *et al.* Concordant Regulation of Translation and mRNA Abundance for Hundreds of Targets of a Human microRNA. *PLOS Biol* **7**, e1000238 (2009). DOI 10.1371/journal.pbio.1000238.

16. Guo, J., Lian, X., Zhong, J., Wang, T. & Zhang, G. Length-dependent translation initiation benefits the functional proteome of human cells. *Mol Biosyst* **11**, 370–378 (2015). DOI 10.1039/c4mb00462k.

17. Arava, Y. Compaction of polyribosomal mRNA. *RNA Biol* **6**, 399–401 (2009).

18. MacDonald, C. T., Gibbs, J. H. & Pipkin, A. C. Kinetics of biopolymerization on nucleic acid templates. *Biopolymers* **6**, 1–5 (1968). DOI 10.1002/bip.1968.360060102.

19. Chou, T., Mallick, K. & Zia, R. K. P. Non-equilibrium statistical mechanics: from a paradigmatic model to biological transport. *Rep. Prog. Phys.* **74**, 116601 (2011). DOI 10.1088/0034-4885/74/11/116601.

20. Blythe, R. A. & Evans, M. R. Nonequilibrium steady states of matrix-product form: a solver's guide. *J. Phys. A: Math. Theor.* **40**, R333 (2007). DOI 10.1088/1751-8113/40/46/R01.

21. Mitarai, N., Sneppen, K. & Pedersen, S. Ribosome Collisions and Translation Efficiency: Optimization by Codon Usage and mRNA Destabilization. *Journal of Molecular Biology* **382**, 236–245 (2008). DOI 10.1016/j.jmb.2008.06.068.

22. Brackley, C. A., Romano, M. C. & Thiel, M. The Dynamics of Supply and Demand in mRNA Translation. *PLOS Comput Biol* **7**, e1002203 (2011). DOI 10.1371/journal.pcbi.1002203.

23. Zia, R. K. P., Dong, J. J. & Schmittmann, B. Modeling translation in protein synthesis with TASEP: A tutorial and recent developments. *J Stat Phys* **144**, 405 (2011). DOI 10.1007/s10955-011-0183-1.

24. Ciandrini, L., Stansfield, I. & Romano, M. C. Ribosome traffic on mRNAs maps to gene ontology: genome-wide quantification of translation initiation rates and polysome size regulation. *PLoS Comput. Biol.* **9**, e1002866 (2013). DOI 10.1371/journal.pcbi.1002866.

25. Reuveni, S., Meilijson, I., Kupiec, M., Ruppin, E. & Tuller, T. Genome-Scale Analysis of Translation Elongation with a Ribosome Flow Model. *PLOS Comput Biol* **7**, e1002127 (2011). DOI 10.1371/journal.pcbi.1002127.

26. Tian, C. *et al.* Rapid Curtailing of the Stringent Response by Toxin-Antitoxin Encoded mRNases. *J. Bacteriol.* JB.00062–16 (2016). DOI 10.1128/JB.00062-16.

27. Shaw, L. B., Zia, R. K. P. & Lee, K. H. Totally asymmetric exclusion process with extended objects: A model for protein synthesis. *Phys. Rev. E* **68**, 021910 (2003). DOI 10.1103/PhysRevE.68.021910.

28. Siwiak, M. & Zielenkiewicz, P. A Comprehensive, Quantitative, and Genome-Wide Model of Translation. *PLOS Comput Biol* **6**, e1000865 (2010). DOI 10.1371/journal.pcbi.1000865.

29. Vanzi, F., Takagi, Y., Shuman, H., Cooperman, B. S. & Goldman, Y. E. Mechanical Studies of Single Ribosome/mRNA Complexes. *Biophysical Journal* **89**, 1909–1919 (2005). DOI 10.1529/biophysj.104.056283.

30. Ringnér, M. & Krogh, M. Folding free energies of 5'-utrs impact post-transcriptional regulation on a genomic scale in yeast. *PLOS Computational Biology* **1**, 1–8 (2005). DOI 10.1371/journal.pcbi.0010072.

31. Heyer, E. E. & Moore, M. J. Redefining the Translational Status of 80s Monosomes. *Cell* **164**, 757–769 (2016). DOI 10.1016/j.cell.2016.01.003.

32. Sin, C., Chiarugi, D. & Valleriani, A. Quantitative assessment of ribosome drop-off in E. coli. *Nucl. Acids Res.* **44**, 2528–2537 (2016). DOI 10.1093/nar/gkw137.

33. Bonnin, P., Kern, N., Young, N. T., Stansfield, I. & Romano, M. C. Novel mrna-specific effects of ribosome drop-off on translation rate and polysome profile. *PLOS Computational Biology* **13**, 1–38 (2017). DOI 10.1371/journal.pcbi.1005555.

34. Wells, S. E., Hillner, P. E., Vale, R. D. & Sachs, A. B. Circularization of mrna by eukaryotic translation initiation factors. *Molecular Cell* **2**, 135 – 140 (1998). DOI https://doi.org/10.1016/S1097-2765(00)80122-7.

35. Scott, M., Gunderson, C. W., Mateescu, E. M., Zhang, Z. & Hwa, T. Interdependence of Cell Growth and Gene Expression: Origins and Consequences. *Science* **330**, 1099–1102 (2010). DOI 10.1126/science.1192588.

36. Castillo-Davis, C. I., Mekhedov, S. L., Hartl, D. L., Koonin, E. V. & Kondrashov, F. A. Selection for short introns in highly expressed genes. *Nat Genet* **31**, 415–418 (2002).

37. Eisenberg, E. & Levanon, E. Y. Human housekeeping genes are compact. *Trends in Genetics* **19**, 362–365 (2003). DOI 10.1016/S0168-9525(03)00140-9.

38. Chou, T. Ribosome Recycling, Diffusion, and mRNA Loop Formation in Translational Regulation. *Biophysical Journal* **85**, 755–773 (2003). DOI 10.1016/S0006-3495(03)74518-4.

39. Sharma, A. K. & Chowdhury, D. Stochastic theory of protein synthesis and polysome: Ribosome profile on a single mRNA transcript. *Journal of Theoretical Biology* **289**, 36–46 (2011). DOI 10.1016/j.jtbi.2011.08.023.

40. Margaliot, M. & Tuller, T. Ribosome flow model with positive feedback. *Journal of the Royal Society Interface* **10**, 20130267 (2013). DOI 10.1098/rsif.2013.0267.

41. Marshall, E., Stansfield, I. & Romano, M. C. Ribosome recycling induces optimal translation rate at low ribosomal availability. *J R Soc Interface* **11**, 20140589 (2014). DOI 10.1098/rsif.2014.0589.

42. Greulich, P., Ciandrini, L., Allen, R. J. & Romano, M. C. Mixed population of competing totally asymmetric simple exclusion processes with a shared reservoir of particles. *Phys. Rev. E* **85**, 011142 (2012). DOI 10.1103/PhysRevE.85.011142.

43. Shah, P., Ding, Y., Niemczyk, M., Kudla, G. & Plotkin, J. B. Rate-limiting steps in yeast protein translation. *Cell* **153**, 1589–1601 (2013). DOI 10.1016/j.cell.2013.05.049.

44. Raveh, A., Margaliot, M., Sontag, E. D. & Tuller, T. A model for competition for ribosomes in the cell. *Journal of the Royal Society Interface* **13**, 20151062 (2016). DOI 10.1098/rsif.2015.1062.

45. Mathematica. Wolfram research, inc. *Version 10.2* (2015).

Acknowledgements

The authors would like to acknowledge the funding provided by the European Union Seventh Framework Programme [FP7/2007-2013] (NICHE; grant agreement 289384) (LDF). AM was partially funded by the UK Biotechnology and Biological Research Council (BBSRC), through grant BB/N015711/1. LC would like to acknowledge Maria Carmen Romano, Jean Hausser, Marco Cosentino Lagomarsino, Jean-Charles Walter and Norbert Kern for early discussions on this work.

Author contributions statement

L.C. and A.M. conceived and discussed the theory, L.C. and L.D.F. performed the simulations and analysed the data, L.C., L.D.F. and A.M. wrote the manuscript text and supplementary information. All authors reviewed the manuscript.