Determinants of translation speed are randomly distributed across transcripts resulting in a universal scaling of protein synthesis times

Ajeet K. Sharma, Nabeel Ahmed, and Edward P. O’Brien

Phys. Rev. E **97**, 022409 — Published 20 February 2018

DOI: 10.1103/PhysRevE.97.022409

Determinants of translation speed are randomly distributed across transcripts resulting in a universal scaling of protein synthesis times

Ajeet K. Sharma[1], Nabeel Ahmed[1,2] and Edward P. O'Brien[1,2,*]
[1]Department of Chemistry, Pennsylvania State University, University Park, PA, USA
[2]Bioinformatics and Genomics Program, Pennsylvania State University, University Park, PA, USA

Abstract

Ribosome profiling experiments have found greater than 100-fold variation in ribosome density along mRNA transcripts, indicating that individual codon elongation rates can vary to a similar degree. This wide range of elongation times, coupled with differences in codon usage between transcripts, suggests that the average codon translation-rate per gene can vary widely. Yet, ribosome run-off experiments have found that the average codon translation rate for different groups of transcripts in mouse stem cells is constant at 5.6 AA/s. How these seemingly contradictory results can be reconciled is the focus of this study. Here, we combine knowledge of the molecular factors shown to influence translation speed with genomic information from *E. coli*, *S. cerevisiae* and *H. sapiens* to simulate synthesis of the cytosolic proteins in these organisms. We demonstrate that the almost constant average gene translation rate arises because the molecular determinants of translation speed are distributed nearly randomly amongst most of the transcripts. Consequently, codon translation rates are also randomly distributed and fast-translating segments of a transcript are likely to be offset by equally probable slow-translating segments, resulting in similar average elongation rates for most transcripts. We also show that the codon usage bias does not significantly affect the near random distribution of codon translation rates because only about 10% of the total transcripts in an organism have high codon usage bias while the rest have little to no bias. Analysis of Ribo-Seq data and an *in vivo* fluorescent assay supports these conclusions.

[*] Corresponding email: epo2@psu.edu

Introduction

A set of seemingly contradictory observations concerning codon translation speeds has arisen from ribosome profiling experiments and measurements of synonymous codon usage bias in organisms. Across the coding sequence of individual genes, ribosome density, and hence codon translation rates, have been found to vary up to 100-fold [1–3]. A large number of publications have demonstrated that codons are not randomly distributed between different transcripts [4,5]. Highly expressed genes, for example, are enriched with optimal codons that tend to be translated with a faster rate [6–8]. This evolutionarily shaped bias in codon usage and the large variation in codon translation rates would seem to suggest that the time it takes to synthesize a protein could vary widely from one gene to the next, even for genes of the same length. Consequently, the average codon translation rate should also vary widely between genes. However, a recent ribosome-run-off experiment in mouse stem cells instead found that the average codon translation rate across different sets of transcripts is constant at 5.6 AA/s [9], regardless of the protein's length, expression level, its final location in the cell and whether it had a high or low tRNA Adaptation index. How these observations can be reconciled is not clear. Additionally, the molecular origin of a constant average elongation rate has not been explained, nor whether this occurs in other organisms.

Many factors influence codon translation speeds, including differences in cognate and near-cognate tRNA concentrations [10–13], positively charged nascent-chain residues inside the ribosome exit tunnel [14–16], mRNA secondary structure [17–19], proline residues at either A or P site of the ribosome [20,21] and steric hindrance between ribosomes translating the same mRNA molecule [22,23]. The mechanisms by which these factors influence translation-elongation rates have been studied extensively over the past two decades [13,24,25]. A high concentration of cognate tRNA molecules, for example, can allow ribosomes to quickly translate a codon as the correct tRNA molecule more rapidly encounters the ribosomal A-site through diffusion [10,12]. A high concentration of near-cognate tRNA molecules, however, can slow translation because they compete with the cognate tRNA for binding to the A-site codon [10]. Attractive electrostatic interactions between positively charged nascent-chain residues and the negatively charged ribosome exit tunnel slows translation of downstream codons [14,15]. While secondary structure formed by mRNA downstream of actively translated codon can decrease translation elongation as the ribosome has to first unfold the structured nucleotides before they can be translated [26]. And in the case of proline, the structural constraints imposed by the N-alkyl group at the P or A site of a ribosome can slow down peptide bond formation and reduce the codon's translation rate [20]. Moreover, a dramatic decrease in the translation rate occurs when two proline amino acid residues are translated successively. And steric interactions between ribosomes on a transcript can slow the translation of a codon [27].

The influence of these molecular factors on translation speed have been measured or estimated. The concentration of different tRNA molecules in an *E. coli* cell varies, for example, by more than 20 fold [10,28]. This causes a similar level of variation in the average time it takes a cognate tRNA molecule to bind at the ribosome's A site [5, 21, 22]. Electrostatic interactions between the negatively charged ribosome exit tunnel and positively charged amino acids decreases the translation rate of nascent proteins by transiently arresting the ribosome-nascent chain complexes [14]. Ribosome profiling experiments have quantified this decrease in the codon translation speed and found a 20% increase in the average ribosome density (and translation time) for the five consecutive codons downstream to the codon that encodes a positively charged amino acid [16]. *In vitro* laser optical tweezer experiments have found a 50% decrease in the translation rate of the j^{th} codon in an mRNA transcript, when the $(j+4)^{\text{th}}$ codon is involved in mRNA secondary structure [26]. This effect can also be seen in ribosome profiling data where 10% more ribosome reads are observed due to the presence of mRNA structure [16]. The *in vivo* incorporation of proline is 3-6 times slower, on average, than phenylalanine incorporation due to the peptidyl transfer step that becomes rate limiting when proline is present at the A or P sites [20]. Moreover, the peptidyl transfer reaction between two proline residues occurs with a 25-fold slower rate [10].

Using a simulation model of protein synthesis that accounts for these molecular factors, we provide an explanation for how the incongruous ribosome profiling and synonymous codon-usage bias observations arise from fundamental features of the distribution of codon translation rates and their molecular determinants. Our results show that codon translation speeds and their molecular determinants are near randomly distributed between different mRNA transcripts. As a consequence, the contribution of fast-translating segments of a transcript to translation speed are often offset by equally probable slow translating segments leading to a near constant average codon translation speed across transcripts. This results in a universal linear scaling of protein synthesis time with coding sequence length across organisms. Analysis of ribosome profiling data and *in vivo* synthesis time measurements support these conclusions.

Methods

Simulating protein synthesis. We simulated the protein synthesis process in *E. coli*, *S. cerevisiae* and *H. sapiens*. We chose these species as they represent a diverse set of organisms. From *E. coli,* a simple prokaryotic organism, to *S. cerevisiae,* a eukaryotic organism, and finally *H. sapiens,* a multi-cellular mammalian organism. To simulate the translation of transcripts in these organisms we used an extension of the Totally asymmetric simple exclusion process (TASEP) model known as ℓ-TASEP (Fig. 1) [30,31]. In ℓ-TASEP, a ribosome covers ℓ-successive codon positions along an mRNA sequence, excluding other ribosomes from that section of the transcript. Ribosome profiling experiments have shown that a ribosome covers 10 codons [1], therefore ℓ was set to 10 in the simulations. A ribosome initiates translation in this model with rate α provided the first 5 codons after the start codon are not occupied by another ribosome. It then elongates the nascent chain by moving one codon at a time with a speed determined by various molecular factors. A ribosome cannot move to the next codon in this model if it is already occupied by another ribosome. The ribosome terminates translation and releases the fully synthesized protein with rate β when it encounters the stop codon.

We used the Monte-Carlo method described in Ref. [32] to simulate the time-dependent movement of ribosomes along mRNA molecules in the ℓ-TASEP model. During these simulations, we track the location of all ribosomes translating the transcript at each time point. This allows us to measure the time it takes individual ribosomes to synthesize a protein as the time it takes the ribosome to move from the start codon to the stop codon. For each cytosolic gene in these organisms we simulated a single copy of its transcript and recorded 12,850 protein synthesis events and their corresponding synthesis times. We discarded data from the first 2,850 ribosomes that completed translation to allow steady-state translation conditions to be achieved, *i.e.*, a constant number of actively translating ribosomes per unit time. We then used the remaining 10,000 values to calculate the average synthesis time for that protein.

Parameterization of the TASEP model. To perform the ℓ-TASEP simulations we need to know the translation rate for every codon position in the transcripts as well as initiation and termination rates, α

and β, of each transcript. The sequence of each cytoplasmic gene was downloaded from the NCBI RefSeq database [33,34]. We used the codon translation rates estimated by the Fluitt-Viljoen model [10,35] for the 61 sense codons in *E. coli*, *S. cerevisiae* and *H. sapiens*. The Fluitt-Viljoen model estimates these rates based on the concentrations of the cognate and near-cognate tRNA molecules, ribosomes, and diffusion constant of tRNA molecules.

The Fluitt-Viljoen model ignores the effects of the other molecular factors known to influence codon translation speed. We incorporate the influence of these other factors by imposing an appropriate penalty to the Fluitt-Viljoen codon translation rates (Fig. 1). To account for the effect of downstream mRNA structure on translation speed we identified whether individual codons in a transcript formed secondary structure or not. To do this, we utilized the analysis methods described in the Refs. [36] and [37] and applied them, respectively, to Parallel analysis of RNA structure (PARS) data from *E. coli* [36] and *H. sapiens* [36] and dimethyl sulfate (DMS) sequencing data from *S. cerevisiae* [37]. Using these data in this way allows us to identify the presence of mRNA structure along a transcript. Ribosome profiling experiments have found that there are, on average, approximately 10% more reads three to five codons upstream of structured nucleotides than on average [16]. Therefore, we decreased the translation speed of the j^{th} codon position in a transcript by 10% if the codon at $j+4$ was involved in forming secondary structure in the transcript. Similarly, ribosome profiling experiments have found an approximately 20% increase in the translation time of the next five successive codons after the incorporation of a positively charged amino acid residue [16]. Therefore, we decreased the translation rate by 20% for those codons that are within five positions downstream of a codon encoding lysine, arginine or histidine amino acids.

To account for the translation-rate slowdown caused by one or two prolines we increased the time it takes to form a peptide bond by the experimentally-measured amounts. In the original Fluit-Viljoen model the time required to form a peptide bond between any two amino acids is 9.06 ms. We increased this time by a factor of 1.40 [38] when proline is present at the P site and the A site of the ribosome is occupied by a tRNA that carries a different amino acid. Similarly, when the tRNA that carries proline residue was at the A site and a different amino acid was at the P site then the time required to form a peptide bond between these amino acid residues was increased by a factor of 3.44 [38]. When the proline was present at the both P and A site of the ribosome then the peptide bond formation between these two proline residues were set to 224.7 ms instead of 9.06 ms [38].

The average time a ribosome takes to fully synthesize a protein is also affected by excluded volume interactions between ribosomes [39]. The frequency with which collisions between ribosomes occurs depends on the translation rate parameters, which can vary from transcript to transcript. The ℓ-TASEP model we use accounts for these interactions. And indeed, we find that they contribute to a 5 to 10% slow down of the average synthesis time (Supplemental Results and Fig. S1 [40]).

We use previously reported translation-initiation rates α for each of the cytoplasmic genes in *S. cerevisiae* [29] that were estimated using polysome profiling data [41]. Since individual initiation rates in *E. coli* and *H. sapiens* are not available we estimated them by scaling the initiation rates found in *S. cerevisiae* such that the ratio of the initiation rate to the transcriptome-wide average elongation rate remained constant. To do this, we first calculated the average translation time of each gene by summing the codon translation times (*i.e*, the inverse of the codon translation rate) of each codon position of a gene, and then divided them with the total number of codons in that gene. These average codon translation times were then used to calculate transcriptome-wide average codon translation rate. For example, the transcriptome-wide average codon translation rate in *E. coli* is 8.95 s^{-1}, and in *S.*

cerevisiae, it is 4.05 s^{-1}. Therefore, we re-scaled the translation-initiation rates for *E. coli* genes by multiplying *S. cerevisiae's* set of initiation rates by a factor of 2.21(=8.95/4.05). We then randomly assigned these re-scaled initiation rates to each of the cytoplasmic genes of *E. coli*. Termination is not a rate limiting step of the overall translation process, therefore we set the termination rate to 35 s^{-1} as previously done in Ref. [29].

Proteins simulated. We only simulated the synthesis of proteins targeted to the cytoplasm. The transcript sequences were taken from NCBI RefSeq database [33,34]. The total number of gene sequences in this database that were also experimentally established to produce proteins that are targeted to the cytoplasm are 2,660, 2,709 and 6,488 for *E. coli. S. cerevisiae* and *H. sapiens*, respectively. PARS data is unavailable for a number of *E. coli* transcripts; this decreased the number of *E. coli* transcripts we simulated to 2,618. For *S. cerevisiae*, the lack of translation-initiation rate information for a number of genes meant we only simulated the translation of 2,584 of these transcripts. Thus, we simulated the translation of 2,618, 2,584 and 6,488 cytoplasmic genes in *E. coli*, *S. cerevisiae* and *H. sapiens*. Gene IDs for each of gene we simulated is provided in Supplemental Material [40]**.**

Estimation of codon translation rates from ribosome profiling data. We estimated codon translation rates in *S. cerevisiae* and *E. coli* using ribosome profiling data [12,42]. *In vivo* ribosome profiling data for *S. cerevisiae* were obtained from sample GSM1289257 reported by Weinberg and coworkers [12]. The raw reads were aligned to the transcriptome and quantified by their 5′ end mapped position by following the procedure used in Ref. 16. Gene annotations for *S. cerevisiae* were obtained from Saccharomyces Genome Database (http://www.yeastgenome.org/), version R64-2-1 on 15 January 2016. All reads in the coding sequence were assigned to nucleotide positions according to the center-weighting method [2]. According to this method, 11 nucleotides from both 5' and 3' ends of a read are excluded and the read is distributed equally among the remaining center-most nucleotide positions. To avoid experimental sampling errors in our analysis, we selected only the subset of high coverage genes with non-zero reads at every codon position. In *S. cerevisiae*, 1,255 genes meet this criterion in our dataset. For each gene, the reads at each codon position were divided by the average number of reads per codon in the transcript, which is widely used in the field as a proxy for normalized codon translation rates [12,24]. These normalized ribosome densities from all genes were aggregated and constitute a transcriptome-wide distribution of codon speeds.

We followed the same procedure to estimate the codon translation rates in *E.coli* from ribosome profiling data of sample GSM1572273 [42]. For assignment of reads, we used an offset of 12 nucleotides from the 3′ end as suggested by Woolstenhulme *et al* [42]. The remaining steps of analysis were the same as described for *S. cerevisiae*. In *E. coli*, 202 genes meet the criterion of non-zero reads at every codon position.

Defining highly-similar distributions of codon translation rates using the Jensen-Shannon metric. Using the Jensen-Shannon (JS) divergence metric [43] we defined a threshold to characterize the distribution of codon translation rates in a transcript as being highly similar to the transcriptome-wide distribution of codon translation rates. To determine this threshold we first calculated the distribution of codon translation rates for each of the cytoplasmic genes in *E. coli*, *S. cerevisiae* and *H. sapiens* as well as the transcriptome-wide distribution of codon translation rates using a bin size of 3 amino acids per second. This bin size gives a non-zero probability density at all the points between the minimum and maximum codon translation rates in these organisms. Next, we calculated the values of the JS metric, comparing the per-transcript distribution against the transcriptome-wide distribution of codon translation rates. We then visually inspected the similarity of these two distributions for randomly selected transcripts at different JS values. We find that if the JS value is equal to or greater than 0.995 then the individual transcript distribution and the transcriptome-wide distribution look highly similar to each other (Figs. S2, S3 and S4 [40]).

Results

Protein synthesis times scale linearly with coding-sequence length. To understand how variations in the translation rates of individual codons affect the time it takes a ribosome to synthesize a cytoplasmic protein we calculated the average synthesis time of cytoplasmic proteins in *E. coli*, *S. cerevisiae* and *H. sapiens*. To do this we identified those genes that result in proteins targeted to the cytoplasm, and then simulated the translation of their corresponding mRNA molecules. We used the ℓ-TASEP model to simulate this process.

In these simulations, we record the time it takes for individual ribosomes to traverse from the start codon to the stop codon of each transcript. We start our *in silico* measurements after the translation-system achieves steady-state conditions. This results in a time-independent distribution of ribosome density, when averaged over sufficiently long observation times. The average synthesis time for a given protein was then computed from the 10,000 recorded times of individual ribosomes completing the translation process. The average synthesis time of cytosolic proteins are plotted as a function of the number of codons in the corresponding mRNA for *E. coli*, *S. cerevisiae* and *H. sapiens*, respectively, in Figs. 2(A), 2(B) and 2(C). We find that the average synthesis time of a protein is strongly correlated with its coding sequence length in these three different organisms (Pearson $R^2 \geq 0.95, p < 10^{-200}$).

We next investigated whether such a high correlation between the synthesis time of proteins and gene length is an artifact of a few data points at the largest gene lengths. To do this we removed genes in the top 1-percentile for length and recalculated R^2. We find the correlation remain greater than 0.94 (*p*-value <10^{-200}) for all three organisms. Thus, the strong correlation we observe is not due to sparse sampling at long transcript lengths.

The average elongation rate is similar between transcripts. This scaling relationship suggests that different genes in an organism are translated with a highly similar average codon translation rate. To test this hypothesis we calculated the distribution of the average codon translation rate per transcript. We find that the coefficient of variation (*i.e.*, the standard deviation of the distribution divided by its average value) is 0.11, 0.27 and 0.07, respectively, in *E. coli*, *S. cerevisiae* and *H. sapiens* (Fig. S5 [40]). This means that the average codon translation rate from one gene to the next, within an organism, does not exhibit much variation.

A prediction from this result is that the average synthesis time of a transcript can be accurately estimated by multiplying the transcriptome-wide-averaged codon translation speed by the number of codons in the coding sequence. To test this, we first calculated the synthesis time of proteins by multiplying the slope of the best-fit line in Figs. 2(A), 2(B) and 2(C) (*i.e.*, the global average codon translation speed) by the number of codons in the corresponding mRNA transcript and then compared it with the synthesis time of the same protein obtained from the simulations. The error in these predictions is characterized by the absolute percent error $\frac{|\tau_{predict}-\tau_{sim}|}{\tau_{sim}}100\%$, where τ_{sim} is the synthesis time of a protein calculated from the TASEP simulations and $\tau_{predict}$ is the synthesis time of a protein obtained by using the best fit translation time. The median percent error is less than 5.8, 9.1 and 4.7% in *E. coli, S. cerevisiae* and *H. sapiens*, respectively (Figs. 2D-2F). Thus, the global translation speed in an organism can be used to accurately predict the synthesis time of a cytoplasmic protein.

The scaling relationship is robust to changes in the molecular factors. Our simulation model accounts for the effects of tRNA concentration and four other factors that are known to influence codon translation rates (Fig. 1). It is likely that there are yet to be discovered molecular factors influencing translation speed. Therefore, it is important to understand how robust this scaling relationship is to changes in the number of molecular factors influencing translation speed. If the conclusions we have drawn are insensitive to such changes then our conclusions are more likely to be accurate. To test for this we removed some of the factors in our model and examined how the results change. To do this, we eliminated from our model the influence of mRNA structure, prolines, and charged residues in the exit tunnel on translation speed. This left us with codon translation rates that depend only on differences in tRNA concentration and ribosome traffic. We used these values to simulate translation of the genes and computed the average synthesis times of the proteins. We again find a very strong correlation (Pearson $R^2 > 0.96,\ p < 10^{-200}$) between the protein synthesis times and coding sequence length for the organisms (Fig. S6 [40]). Thus, increasing the number of molecular determinants that affect the translation speed (Fig. 2) does not cause a significant change in the correlation between the protein synthesis time and gene length. This suggests that the scaling relationship we observed will likely remain even if new factors are discovered in the future.

A narrow distribution of codon translation rates alone is not sufficient to explain the scaling relationship. Small variation in the translation rates of individual codons across the transcriptome could give rise to a nearly constant average translation speed in different genes, thereby resulting in a linear scaling of protein synthesis time with transcript length (Fig. 2). However, if such a small variation is sufficient to cause the scaling relationship in Fig. 2 then this relationship should remain robust to any biased redistribution of codon translation rates among the transcripts in our dataset. We tested this by creating artificially biased translation-rate profiles for each of the genes. To do this we first rank ordered the transcriptome-wide codon translation rates from fastest to slowest. Next, we randomly selected a gene and sequentially assigned the first N_c fastest codon translation rates to each codon position of that gene, where N_c is the number of codons in the transcript. We then removed those N_c codon translation rates from our rank-ordered list and repeated the same procedure until we created artificial translation rate profiles for each of the cytoplasmic genes in that organism.

We performed protein synthesis simulations for each of these genes and calculated their average synthesis times. The artificial assignment of codon translation rates drastically decreased the R^2-correlation coefficient between the protein synthesis time and the coding-sequence length to 0.29, 0.19 and 0.68 for *E. coli*, *S. cerevisiae* and *H. sapiens* (Fig. S7 [40]), respectively. If the limited variation in the codon translation rates was sufficient to cause the linear scaling relation in Fig. 2 then we would have observed a very strong R^2 correlation between protein synthesis time and gene length for these biased translation-rate profiles. This result indicates that how codon translation rates are distributed between different transcripts plays a central role in the scaling relationship observed in Fig. 2.

The Law of the Large Numbers explains the mathematical origin of the scaling relationship. The narrow distribution of the average codon translation rate per gene (Fig. S5 [40]) suggests that the Law of Large Numbers is at play during the translation of transcripts. According to the Strong Law of Large Numbers [44] the mean of a sample of n observations, μ_n, converges to the average value μ of the population from which those observations were drawn as n approaches infinity. To map this statement onto the process of translation let the "mean of the sample" be the average codon translation rate $\overline{t_n}$ of a transcript (*i.e.*, $\mu_n = \overline{t_n}$); let the "sample" be the coding sequence of the transcript; let n be the number of codons in the coding sequence, *i.e.* N_c; and let an "observation" be the translation rate of a single codon position in the coding sequence. The sample mean (*i.e.*, the average codon translation rate of a transcript) is then, by definition,

$$\overline{t_{N_C}} = \frac{t_1+t_2+t_3+\cdots+t_i+\cdots+t_{N_C}}{N_c},$$

[1]

where t_i is the average translation time of the i^{th} codon position in the transcript. The Law of Large Numbers predicts that for sufficiently large genes $\overline{t_{N_C}} = \mu$. Substituting this relationship into eq. 1, and carrying out algebraic rearrangement, yields

$$N_c\mu = t_1 + t_2 + t_3 + \cdots + t_{N_C} = t_s.$$

[2]

The right-hand-side of Eq. 2 equals the average synthesis time of the transcript, denoted by t_S. Therefore, the Law of Large Numbers predicts that $t_S = N_c\mu$. That is, if you double the length of a coding sequence (*i.e.*, double N_c), you will double the time it takes to synthesize the protein. Thus, the Law of Large Numbers predicts the linear scaling relationship we have observed in Figs. 2 and S6 [40].

To test if the Law of Large numbers is relevant to translation we examined its prediction that the average codon translation time per gene should converge towards the transcriptome-wide average codon translation speed as the coding sequence gets longer. We find that for all three organisms $\overline{t_{N_C}}$ converges towards μ as the transcript length increases (Fig. 3). Thus, the Law of Large Numbers provides a mathematical explanation for the origin of the scaling of synthesis time with transcript length. This also suggests that the estimation of the synthesis time of a protein using the transcriptome-wide average codon translation rate (Figs. 2D-DF) tends to be more accurate for longer transcripts.

A near-random distribution of codon translation rates is the proximal cause of the scaling relationship with a strong correlation. The fact that some of the features of translation in our model follows the Law of Large numbers is not a satisfying explanation as it does not provide a physical explanation for the scaling relationship's origin. However, an additional assumption of the Law of Large numbers – that the composition of codon translation rates in a coding sequence are randomly sampled from the population-wide distribution [45] – would provide such an explanation. This assumption predicts that the codon translation rates in a coding sequence are randomly sampled from the transcriptome-wide distribution of codon translation rates. As a consequence, the distribution of codon translation rates in a coding sequence should be statistically indistinguishable from the distribution of codon translation rates across all transcripts. We tested this prediction by calculating the percentage of transcripts for which this was the case using the two-sample Kolmogorov-Smirnov (KS) test [46]. We find that in *E. coli, S. cerevisiae* and *H. sapiens*, respectively, 77%, 55% and 35% of genes contain distributions of codon translation rates that are indistinguishable from the transcriptome-wide distribution.

We next tested for the percentage of transcripts that have a near-random distribution of codon translation speeds, as the average translation speed of these transcripts would be highly similar to the transcriptome-wide average and contribute to the scaling relationship. To do this we used the Jensen-Shannon divergence metric, which measures the similarity between two distributions, and defined a threshold for characterizing two distributions as highly similar (see Methods). We find that the distribution of codon translation rates in 81%, 91% and 94% of transcripts are highly similar to the transcriptome-wide distributions in *E. coli*, *S. cerevisiae* and *H sapiens*, respectively. Thus, in the vast majority of transcripts, codon translation rates are randomly, or nearly randomly, distributed between the different transcripts.

This finding explains the physical origin of the scaling relationship, and the near constant average codon translation rates between different genes in our simulations and observed in experiments [9]. Since codon translation rates are randomly distributed between transcripts, then it follows that for every fast-translating segment of a transcript there is just as likely to be a slow-translating segment elsewhere along the same transcript. In this way, the average codon translation rate across a transcript never deviates far from the transcriptome-wide average translation rate. Consequently, an increase in the coding sequence length results in a directly proportional increase in the average synthesis time of the transcript.

A near-random distribution of molecular determinants is the ultimate cause. The observation that the codon translation rates are randomly or near-randomly distributed across most transcripts suggests that the molecular factors determining codon translation rates should also be randomly distributed. If they are randomly distributed then each of these molecular factors must scale linearly with the length of coding sequence. We observe this is indeed the case for positively charged residues, proline resides, codons that take part in mRNA structure, and the number of times a particular tRNA molecule is called for by a transcript (Fig. 4). For example, for a lysine-carrying tRNA molecule, we find a correlation between the number of codon positions at which it is called and the transcript length ($\mathrm{R}^2 > 0.69$, Fig. 4(D)). More generally, for the 46 unique tRNA molecules in *E. coli,* the majority have an R^2 greater than 0.6 (Fig. 4(E)).

Next, we examined how similar the distributions of molecular factors are to a true random distribution. We did this by constructing artificial coding sequences whose number and length equals that found in the organisms, but across which we randomly assigned the molecular factors found in the organism's transcriptome. For example in *E. coli* there are 35,775 prolines in the 2,618 transcripts in our dataset. We randomly distributed these prolines across the 2,618 artificial transcripts, thereby maintaining the fraction of prolines found in nature. From these artificial transcripts we computed the distribution of each molecular factor present per 100 codons and compared it to the actual distribution found in *E. coli*. For codons that encode positively charged residues and prolines residues we find that the observed distributions and random distributions are similar (Figs. S8(A) and S8(B) [40]) in *E. coli*. And the distributions of codons that are decoded by the top-third, middle-third or bottom-third percentile of cognate-tRNA concentrations [10–13] are very similar to the random distribution (Figs. S8(D), S8(E) and S8(F) [40]). There is poor agreement, however, between the observed fraction of codons forming structure and the random distribution (Fig. S8(C) [40]). Similar results are found in *S. cerevisiae* (Figs. S9 and S10 [40]) and *H. sapiens* (Figs. S11 and S12 [40]). These results suggest that for most transcripts, the molecular determinants of codon translation rates are randomly or near-randomly distributed between transcripts. As a consequence, regardless of the coding-sequence length of the transcript, the overall effect of these molecular determinants across the transcriptome remains almost the same, resulting in a linear scaling of protein synthesis time with length of the coding sequence (Fig. 2).

Ribosome profiling experiments are consistent with near-randomly distributed codon translation rates. Ribosome profiling (Ribo-Seq) data can test our prediction that codon translation rates are randomly distributed between transcripts. Ribo-Seq is a next-generation sequencing technique that provides a measure of the number and location of actively translating ribosomes across the transcriptome at codon resolution. The signal from Ribo-Seq can be converted into an estimate of normalized individual codon translation rates across a coding sequence, as detailed in the Methods Section. Therefore, we tested whether the distribution of normalized codon translation rates in high-coverage transcripts was statistically different from the distribution of codon translation rates from all the transcripts in our dataset. We find that 78% of the 199 high-coverage transcripts in *E. coli*, and 80% of the 1,255 high-coverage transcripts in *S. cerevisiae*, are statistically indistinguishable from their respective global distribution of codon translation rates (two-sample Kolmogorov-Smirnov test, Fig. 5). This indicates that despite any potential codon usage bias that may be present, codon-translation rates are randomly distributed across transcripts in these two organisms, resulting in an average gene translation speed which is very similar to the transcriptome-wide average translation speed. This is experimental support for one of the key conclusions of this study.

***In vivo* fluorescent measurements detect a scaling relationship.** While ribosome-runoff experiments have measured average codon translation rates, they did not measure the synthesis time of individual proteins from start to finish as they never reach a time point at which all the ribosomes have completed translation. A method for measuring synthesis times was published recently [47]. In that method, fluorescently-labelled antibodies bind to peptide epitopes that are genetically encoded in the gene of interest. By detecting the time-dependent appearance of fluorescent foci arising from the antibodies binding to nascent proteins during synthesis, the time it takes to synthesize the protein of interest can be measured. The synthesis time of three different proteins in *U2OS* cells has been measured in this way. Plotting these times versus coding sequence length results in a linear scaling between them (Fig. S13 [40]). While it would be ideal if many more proteins were measured in this way, these results are consistent with the main conclusion of this study that protein synthesis times are directly proportional to coding sequence length.

Codon usage bias does not affect the linear nature of the scaling relationship. Our results would seem to contradict the observation that codon usage is non-random across genes [48–50]. In fact, only a small percentage of genes in organisms exhibit large codon usage biases. Plotting the distribution of Codon Bias Index (CBI) [51,52] values for the genes in *E. coli* and *S. cerevisiae* (Figs. S14(A) and S14(B) [40]) reveals that many genes have a CBI value near zero (*i.e.*, random, or near random codon usage), however long tails toward higher CBI values are also present. To quantify the percentage of genes that have a large codon usage bias we utilized the observation that highly expressed genes tend to be enriched in optimal codons [53,54]. Taking the average CBI for these genes (Figs. S14(C) and S14(D) [40]), and subtracting off the standard deviation defines a threshold CBI value above which we classify a gene as having a large codon usage bias. This threshold is 0.38 and 0.49 in *E. coli* and *S. cerevisiae*, respectively. Only 11% and 7% genes in *E. coli* and *S. cerevisiae* are above this threshold. Thus around 90% of genes in these organisms exhibit small to no codon usage bias, and therefore the global scaling relationship (Fig. 2) is not significantly affected by this subpopulation of genes with high codon usage bias.

Since highly expressed genes tend to have higher codon usage bias [53,54] we examined whether the scaling relationship persists for these genes. We find that indeed it does (Fig. S15 [40]), albeit with slightly faster average translation speeds. This means that the average codon translation rate is similar between highly expressed genes and suggests that the molecular factors are randomly distributed within those highly-expressed genes. The distribution from which these molecular factors are randomly sampled however, is slightly different from their transcriptome-wide counterpart, thus giving rise to a different average elongation rate in Fig. S15 [40].

Discussion

Despite the large variation in codon translation rates that can occur from one codon position to the next, and the codon usage bias that is present in the genomes of organisms, the average codon translation rate between most transcripts are highly similar. This was seen in previously reported Ribo-Seq experiments of mouse stem cells, as well as in our simulation results of the translation of cytosolic proteins in *E. coli*, *S. cerevisiae* and *H. sapiens*. Our results indicate that the molecular origin of this observation is that the determinants of translation speed (tRNA concentration, ribosome traffic, proline-sequence motifs, charged residues in the exit tunnel, and mRNA structure) are near-randomly distributed between genes (Figs. 4, S8-S12 [40]). As a consequence, codon translation rates are also near-randomly distributed between transcripts, meaning that the number of fast-translating codons will be nearly equal to the number of slow-translating codons in most transcripts. Thus, the average codon translation rate per gene follows the Law of Large Numbers – converging towards the transcriptome-wide average as transcript coding sequence length gets longer (Fig. 3) – and resulting in a narrow distribution centered on the transcriptome-wide average (Fig. S5 [40]).

There is experimental support for two key results from this study. Through an analysis of Ribo-Seq data from *E. coli* and *S. cerevisiae* we have found that the distribution of ribosome density across individual transcripts is highly similar to the transcriptome-wide distribution of ribosome densities (Fig. 5) – consistent with a near random distribution of codon translation speeds between transcripts. And *in vivo* measurements of protein synthesis times observe linear scaling with coding sequence length (Fig. S13 [40]). We also note that the analysis of the distribution of molecular determinants of translation speed (Figs. 4, S8-S12 [40]) comes solely from the experimentally determined genome sequences and tRNA concentrations. Thus, there is strong experimental support for the molecular origins and consequences we have identified. A potential point of confusion for readers is what we mean when we say there is a random distribution of factors, or codon translation speeds, across transcripts. If we were to artificially design a translation rate profile we would have to carry out two essential steps. First, we would need to pick out codon translation rates for each codon position in the mRNA sequence. Second, we would need to order those codon translation rates from the 5′ to 3′ end of the coding sequence. These two steps are independent, meaning both steps could be random, both could be non-random, or one could be random and the other not. For example, we could randomly select codon translation rates from a genome-wide distribution of the codon translation rates of the organism of interest, but we could then non-randomly order them within the coding sequence. In this paper, when we say factors (or speed) are randomly distributed across transcripts we are referring to the first step in this process. Thus, this study does not comment on whether codon translation rates or their molecular determinants are randomly distributed *within* a given coding sequence. Indeed, that question is not germane as the Law of Large numbers only requires that the first step be random, not the second step. What we have clearly established in this study is that the composition (not the ordering) of those speeds or factors is near-randomly distributed across transcripts.

A number of studies have found that translation elongation kinetics can influence the structure and function of some proteins [52,55,56]. When a small number of synonymous mutations are made at critical codon positions along a transcript, which changes the elongation rate but not the nascent protein's primary structure, there can be dramatic changes in co-translational folding that results in more misfolding and decreased specific activity [52,56]. Indeed, there is evidence evolution may have encoded patterns of translation-elongation rate information in mRNA molecules to coordinate co-translational processes to optimize the efficiency of protein maturation [57]. Our results are not in contradiction with such findings as these observations indicate it is the non-random ordering of codon translation speeds within a coding sequence that is critical, not whether the composition is randomly sampled from the underlying transcriptome-wide distribution.

While many of the molecular factors we examined were near-randomly distributed, there are clear examples of non-random distributions. For example, the distribution of the number of codons involved in mRNA structure in all three organisms is significantly different from their random distribution (Figs S8(C), S10(C) and S12(C) [40]). Structure formation involves interactions between nucleobases that are distantly separated along the mRNA sequence, leading to long range correlations that could contribute to this non-random distribution. Additionally, G-C rich portions of mRNA, which form more stable structure, could contribute as well [58]. We also find a non-random distribution of codons that call for tRNAs with high, medium and low abundance in *H. sapiens* (Figs. S12(D), S12(E) and S12(F) [40]). However, despite tRNA concentration being the most important determinant of codon translation rate in our model, we find a very strong correlation between the protein synthesis time and coding sequence length in *H. sapiens* (Fig. 2(C)). The reason for this is that the variation in the concentration of tRNA molecules in *H. sapiens* is significantly lower as compared to the *E. coli* and *S. cerevisiae*. The coefficient of variation for tRNA concentration in *E. coli*, *S. cerevisiae* and *H. sapiens* are 0.81, 0.71 and 0.19, respectively [10,59]. Therefore, the non-random distribution of codons corresponding to the tRNAs with high, medium and low abundances does not significantly impact the average codon translation time of a transcript. For the same reason, we find the strongest correlation between the protein synthesis times and coding sequence length in biased translation rate profiles of *H. sapiens* (Fig. S7 [40]). For a few tRNAs, we also see a low R^2 correlation between coding sequence length and the number of times those tRNAs are called in a coding sequence (Figs. 4(E), S9(E) and S11(E) [40]). We find that the codons corresponding to the tRNAs with the five lowest R^2 correlations are on average used ten times less frequently than the rest of the other codons in *E. coli* and three times less frequently used in *S. cerevisiae* and *H. sapiens*. Therefore, despite their low R^2 values they do not have a significant effect on the scaling relationship.

Consistent with a near random distribution, we find that the variance in average codon translation time per transcript decreases with increasing transcript length (Fig. 3). However, evolutionary biases in codon usage are also likely to contribute to the variation in average translation speed of shorter transcripts away from the global mean value. For example, highly expressed genes tend to be shorter in length (average length of 233 versus 306 codons in *E. coli*, and 219 versus 533 codons in *S. cerevisiae*), and have stronger codon usage bias [7,8]. Additionally, shorter mRNA transcripts often have higher translation-initiation rates [60,61] suggesting there is a greater fitness benefit for biasing codon usage in such transcripts to efficiently use the pool of cellular ribosomes. Indeed, we find that highly expressed proteins have a faster translation speed in our model, but still maintain a linear scaling (Fig. S15 [40]), indicating that short, highly expressed proteins also have randomly distributed molecular factors, but that are being randomly drawn from a particular subset of the molecular factor/translation speed space. This is consistent with different evolutionary pressures shaping mRNA sequence evolution of highly and lowly expressed transcripts [62,63]. For these reasons, the results of this study also do not contradict the finding that codon usage bias plays a functional role in regulating gene expression [5,64,65] by, for example, ensuring the efficient use of ribosomes [66,67].

In principle, any non-random distribution of codon translation rates that does not depend upon transcript length could also give rise to a linear scaling of protein synthesis time with transcript length. For example, in this study, we created artificial transcripts that had biased translation rate profiles and found the synthesis time with transcript length could be described using a linear function. However, the important point to note is that these biased profiles resulted in a weak correlation, with R^2 as low as 0.19. Thus, the very strong correlations we observe in the naturally occurring transcripts (R^2>0.95) coupled with the near-random distribution of many molecular factors, is evidence that randomness, on average, plays a greater role than any non-random contributions to synthesis time.

The slope of the scaling relationship we have observed (Fig. 2) could be affected by changes in cellular condition. For example, an increase in the cellular concentration of guanosine triphosphate (GTP) molecule, which are hydrolyzed during translation [22,23], would uniformly increase the translation rate of each codon. This increase in codon translation rates would not affect their random distribution between different transcripts. Therefore, the scaling relationship would likely remain robust to such changes in cellular conditions. However, increased codon translation rates would tend to decrease the slope of the best fit line between protein synthesis time and the coding sequence length.

In our simulation model we accounted for the influence of five different molecular factors known to influence codon translation speeds. With rapid advances in experimental techniques that probe translation [47,68,69], and growing interest of the impact of translation kinetics on nascent protein behavior, it will undoubtedly be the case that additional factors that influence translation speed will be identified. We have demonstrated that the scaling relationship we have observed is robust to changes in which factors are modeled and which are not. Thus, even if new factors are identified in the future, it is likely that the conclusions of this paper will remain unchanged.

Acknowledgements

This work was supported by Human Frontier Science Program research grant RGP0038/201 and National Science Foundation CAREER grant MCB-1553291.

References

[1] N. T. Ingolia, S. Ghaemmaghami, J. R. S. Newman, and J. S. Weissman, Science **324**, 218 (2009).

[2] E. Oh, A. H. Becker, A. Sandikci, D. Huber, R. Chaba, F. Gloge, R. J. Nichols, A. Typas, C. A. Gross, G. Kramer, J. S. Weissman, and B. Bukau, Cell **147**, 1295 (2011).

[3] T. Tuller and H. Zur, Nucleic Acids Res. **43**, 13 (2014).

[4] Y. Prat, M. Fromer, N. Linial, and M. Linial, BMC Evol. Biol. **9**, 285 (2009).

[5] T. E. F. Quax, N. J. Claassens, D. Söll, and J. van der Oost, Mol. Cell **59**, 149 (2015).

[6] S. Ghaemmaghami, W.-K. Huh, K. Bower, R. W. Howson, A. Belle, N. Dephoure, E. K. O'Shea, and J. S. Weissman, Nature **425**, 737 (2003).

[7] Z. Zhou, Y. Dang, M. Zhou, L. Li, C. Yu, J. Fu, S. Chen, and Y. Liu, Proc. Natl. Acad. Sci. **113**, E6117 (2016).

[8] Y. Hiraoka, K. Kawamata, T. Haraguchi, and Y. Chikashige, Genes to Cells **14**, 499 (2009).

[9] N. T. Ingolia, L. F. Lareau, and J. S. Weissman, Cell **147**, 789 (2011).

[10] A. Fluitt, E. Pienaar, and H. Viljoen, Comput. Biol. Chem. **31**, 335 (2007).

[11] A. Dana and T. Tuller, Nucleic Acid Res. **42**, 9171 (2014).

[12] D. E. Weinberg, P. Shah, S. W. Eichhorn, J. A. Hussmann, J. B. Plotkin, and D. P. Bartel, Cell Rep. **14**, 1787 (2016).

[13] M. V. Rodnina, Protein Sci. **25**, 1390 (2016).

[14] J. Lu and C. Deutsch, J. Mol. Biol. **384**, 73 (2008).

[15] R. Sabi and T. Tuller, BMC Genomics **16 Suppl 1**, S5 (2015).

[16] C. Charneski and L. Hurst, PLoS Biol. **11**, e1001508 (2013).

[17] J. Wen, L. Lancaster, C. Hodges, A. Zeri, S. H. Yoshimura, H. F. Noller, C. Bustamante, and I. T. Jr, Nature **452**, 598 (2008).

[18] C. Chen, H. Zhang, S. L. Broitman, M. Reiche, I. Farrell, B. S. Cooperman, and Y. E. Goldman, Nat. Struct. Mol. Biol. **20**, 582 (2013).

[19] Y. Mao, H. Liu, Y. Liu, and S. Tao, Nucleic Acids Res. **42**, 4813 (2014).

[20] M. Y. Pavlov, R. E. Watts, Z. Tan, V. W. Cornish, M. Ehrenberg, and A. C. Forster, Proc. Natl. Acad. Sci. U. S. A. **106**, 50 (2009).

[21] C. G. Artieri and H. B. Fraser, Genome Res. **24**, 2011 (2014).

[22] A. K. Sharma and D. Chowdhury, Phys. Biol. **8**, 26005 (2011).

[23] D. Chowdhury, Phys. Rep. **529**, 1 (2013).

[24] C. Pop, S. Rouskin, N. T. Ingolia, L. Han, E. M. Phizicky, J. S. Weissman, and D. Koller, Mol. Syst. Biol. **10**, 770 (2014).

[25] T. E. Gorochowski, Z. Ignatova, R. A. L. Bovenberg, and J. A. Roubos, Nucleic Acids Res. **43**, 3022 (2015).

[26] X. Qu, J.-D. Wen, L. Lancaster, H. F. Noller, C. Bustamante, and I. Tinoco, Nature **475**, 118 (2011).

[27] N. Mitarai and S. Pedersen, Phys. Biol. **10**, 56011 (2013).

[28] H. Dong, L. Nilsson, and C. G. Kurland, J. Mol. Biol. **260**, 649 (1996).

[29] L. Ciandrini, I. Stansfield, and M. C. Romano, PLoS Comput. Biol. **9**, e1002866 (2013).

[30] L. B. Shaw, R. K. P. Zia, and K. H. Lee, Phys. Rev. E **68**, 021910 (2003).

[31] A. B. Kolomeisky, G. M. Schütz, E. B. Kolomeisky, and J. P. Straley, J. Phys. A: Math. Gen. **31**, 6911 (1999).

[32] T. Chou, K. Mallick, and R. K. P. Zia, Reports Prog. Phys. **74**, 116601 (2011).

[33] T. Tatusova, M. Dicuccio, A. Badretdin, V. Chetvernin, E. P. Nawrocki, L. Zaslavsky, A. Lomsadze, K. D. Pruitt, M. Borodovsky, and J. Ostell, Nucleic Acids Res. **44**, 6614 (2016).

[34] J. R. Brister, D. Ako-Adjei, Y. Bao, and O. Blinkova, Nucleic Acids Res. **43**, D571 (2015).

[35] M. Siwiak and P. Zielenkiewicz, PLoS Comput. Biol. **6**, e1000865 (2010).

[36] C. Del Campo, A. Bartholomäus, I. Fedyunin, and Z. Ignatova, PLoS Genet. **11**, e1005613 (2015).

[37] S. Rouskin, M. Zubradt, S. Washietl, M. Kellis, and J. S. Weissman, Nature **505**, 701 (2014).

[38] L. K. Doerfel, I. Wohlgemuth, C. Kothe, F. Peske, H. Urlaub, and M. V. Rodnina, Science **339**, 85 (2013).

[39] P. Greulich and A. Schadschneider, in *Traffic Granul. Flow '07*, Springer, Berlin, Heidelberg (2009).

[40] See Supplemental Material at [URL] for Supplemental Results and Figs. S1-S16.

[41] V. L. MacKay, X. Li, M. R. Flory, E. Turcott, G. L. Law, K. A. Serikawa, X. L. Xu, H. Lee, D. R. Goodlett, R. Aebersold, L. P. Zhao, and D. R. Morris, Mol. Cell. Proteomics **3**, 478 (2004).

[42] C. J. Woolstenhulme, N. R. Guydosh, R. Green, and A. R. Buskirk, Cell Rep. **11**, 13 (2015).

[43] C. D. Manning and H. Schütze, *Foundations of Natural Language Processing,* The MIT press (2000).

[44] R. Nelson, *Probability, Stochastic Processes, and Queueing Theory,* Springer New York (1995).

[45] I. I. Gorban, The *Statistical Probability Phenomenon,* Springer International Publishing, Cham (2017)

[46] T. T. Soong, Fundamentals of Probability and Statistics for Engineers, Willey (20014).

[47] T. Morisaki, K. Lyon, K. F. DeLuca, J. G. DeLuca, B. P. English, Z. Zhang, L. D. Lavis, J. B. Grimm, S. Viswanathan, L. L. Looger, T. Lionnet, and T. J. Stasevich, Science **352**, 1425 (2016).

[48] P. M. Sharp and W. H. Li, Nucleic Acids Res. **15**, 1281 (1987).

[49] D. Lal, M. Verma, S. K. Behura, and R. Lal, Res. Microbiol. **167**, 669 (2016).

[50] G. Brandis and D. Hughes, PLOS Genet. **12**, e1005926 (2016).

[51] J. L. Bennetzen and B. D. Hall, J. Biol. Chem. **257**, 3026 (1982).

[52] M. Zhou, J. Guo, J. Cha, M. Chae, S. Chen, J. M. Barral, M. S. Sachs, and Y. Liu, Nature **495**, 111 (2013).

[53] M. Bulmer, Genetics **129**, 897 (1991).

[54] P. Shah and M. a Gilchrist, Proc. Natl. Acad. Sci. U. S. A. **108**, 10231 (2011).

[55] D. A. Nissley, A. K. Sharma, N. Ahmed, U. Friedrich, G. Kramer, B. Bukau, and E. P. O'Brien, Nat. Commun. **7**, 10341 (2015).

[56] D. A. Nissley and E. P. O'brien, J. Am. Chem. Soc. **136**, 17892 (2014).

[57] S. Pechmann and J. Frydman, Nat. Struct. Mol. Biol. **20**, 237 (2013).

[58] G. Faure, A. Y. Ogurtsov, S. A. Shabalina, and E. V Koonin, Nucleic Acids Res. **44**, 10898 (2016).

[59] M. Siwiak and P. Zielenkiewicz, PLoS One **8**, e73943 (2013).

[60] L. D. Fernandes, A. P. S. de Moura, and L. Ciandrini, Sci. Rep. **7**, 17409 (2017).

[61] Y. Arava, Y. Wang, J. D. Storey, C. L. Liu, P. O. Brown and D. Herschlag, Proc. Natl. Acad. Sci. **100**, 3889 (2003).

[62] D. A. Drummond, J. D. Bloom, C. Adami, C. O. Wilke, and F. H. Arnold, Proc. Natl. Acad. Sci. **102**, 14338 (2005).

[63] C. Pál, B. Papp, and L. D. Hurst, Genetics **158**, 927 (2001).

[64] T. Tuller, A. Carmi, K. Vestsigian, S. Navon, Y. Dorfan, J. Zaborske, T. Pan, O. Dahan, I. Furman, and Y. Pilpel, Cell **141**, 344 (2010).

[65] C. H. Yu, Y. Dang, Z. Zhou, C. Wu, F. Zhao, M. S. Sachs, and Y. Liu, Mol. Cell **59**, 744 (2015).

[66] G. E. Andersson and C. G. Kurland, Mol. Biol. Evol. **8**, 530 (1991).

[67] S. Klumpp, J. Dong, and T. Hwa, PLoS One **7**, e48542 (2012).

[68] B. Wu, C. Eliscovich, Y. J. Yoon, and R. H. Singer, Science **352**, 1430 (2016).

[69] X. Yan, T. A. Hoek, R. D. Vale, and M. E. Tanenbaum, Cell **165**, 976 (2016).

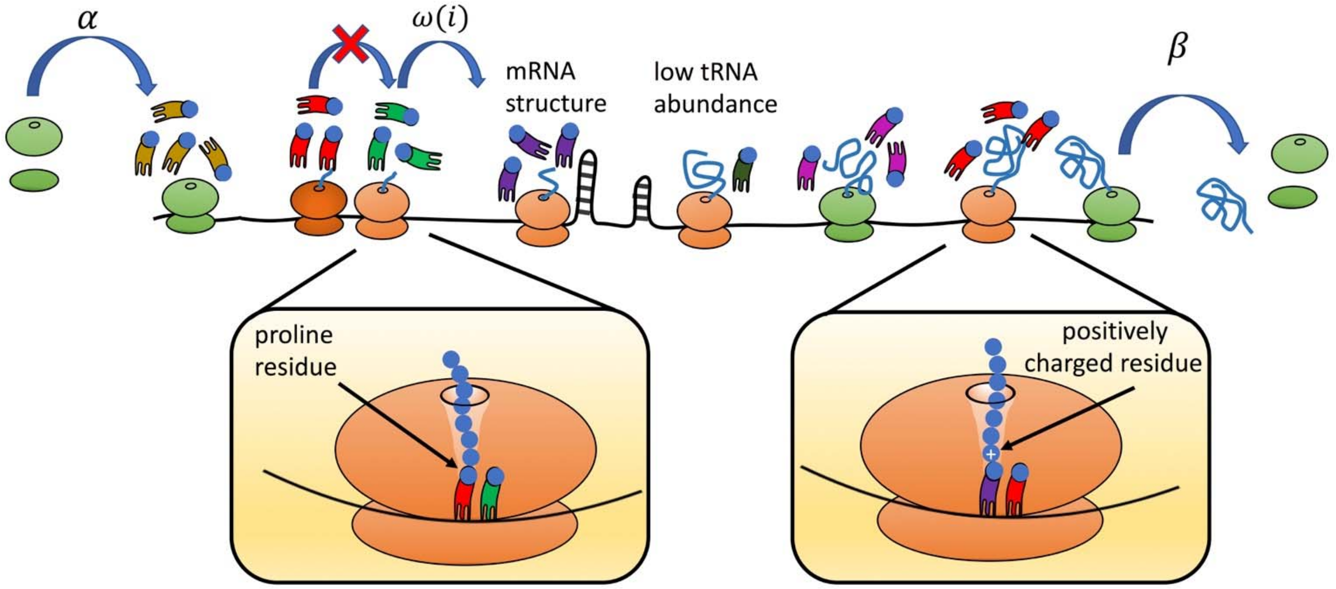

Figure 1: **Illustration of the ℓ-TASEP model and its parameterization.** A ribosome initiates translation with rate α when the first five codon positions after the start codon are not blocked by another ribosome. The ribosome translates the j^{th} codon position with rate $\omega(j)$ when no downstream ribosome occupies $(j+10)^{th}$ codon position and terminates the translation process with rate β. tRNA abundance, mRNA structure, proline residues at ribosome A and P site and positively charged residues affect the translation rate of a codon and are accounted for in the model. Ribosomes in green and light-red color are translating optimal and non-optimal codons, respectively, whereas the ribosome colored in dark-red is sterically blocked by a downstream ribosome.

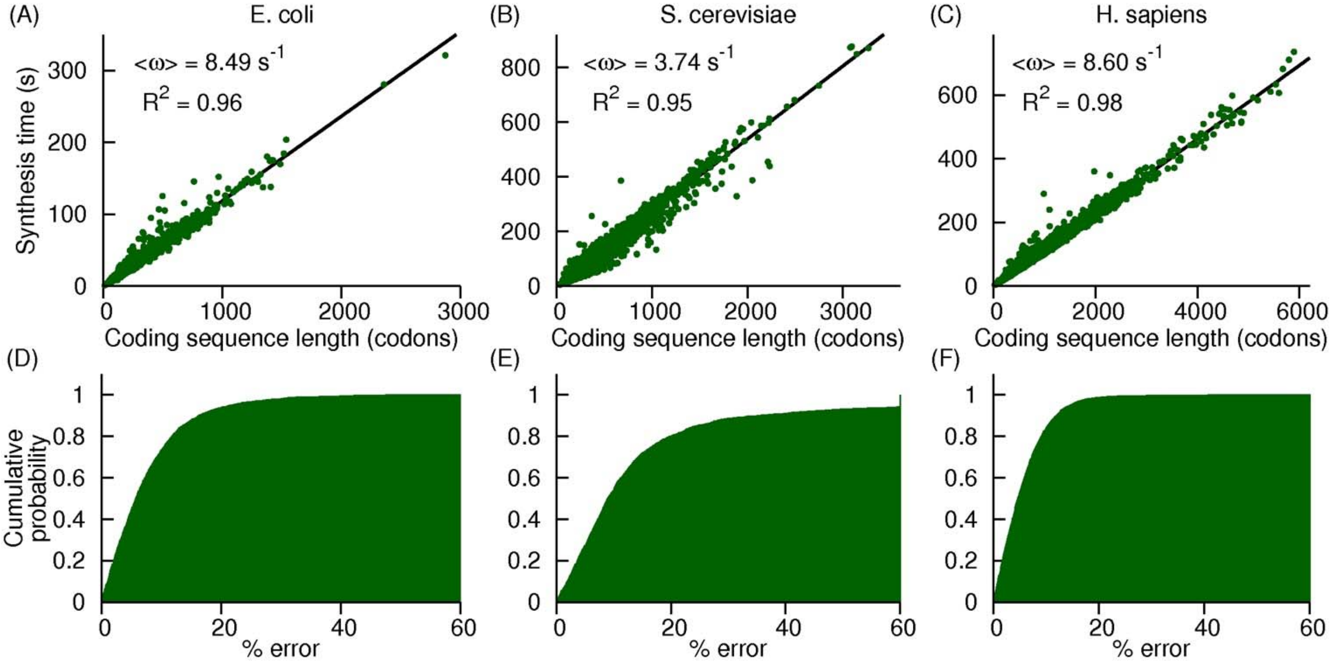

Figure 2: **Synthesis time of proteins scale linearly with coding sequence length.** Average synthesis times of proteins as a function of the length of the corresponding coding sequence are plotted, respectively, for *E. coli*, *S. cerevisiae* and *H. sapiens* in (**A**), (**B**) and (**C**). The cumulative probability distribution of the percent error in the predicted synthesis times is plotted in (**D**), (**E**) and (**F**) for *E. coli, S. cerevisiae* and *H. sapiens*, respectively.

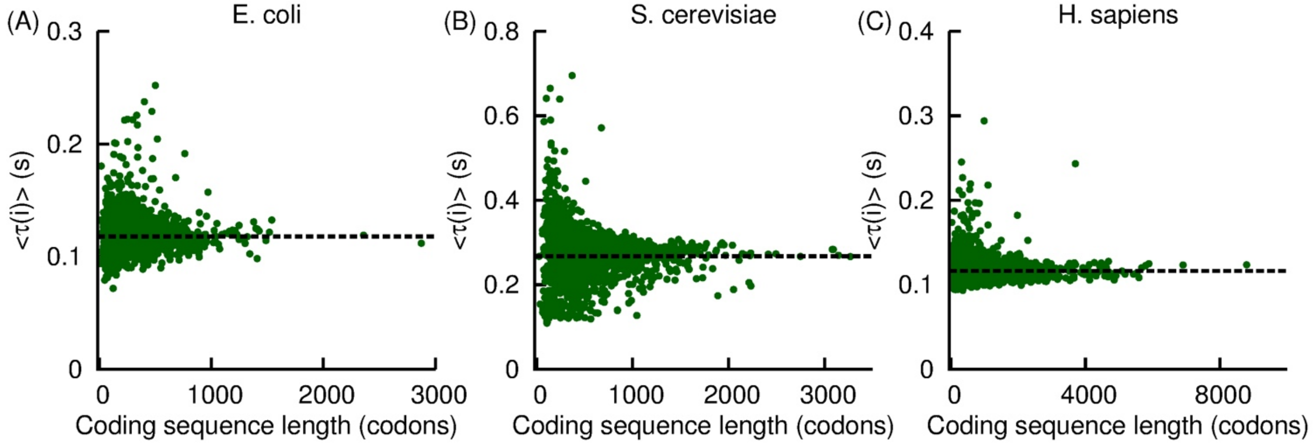

Figure 3: Average translation time of a gene follows the Law of Large Numbers. The average codon translation time of a transcript versus gene length for *E. coli, S. cerevisiae* and *H. sapiens* in (**A**), (**B**) and (**C**), respectively. Results are from Monte-Carlo Simulations of the translation process. The dotted lines represent the transcriptome-wide average codon translation time in these organisms.

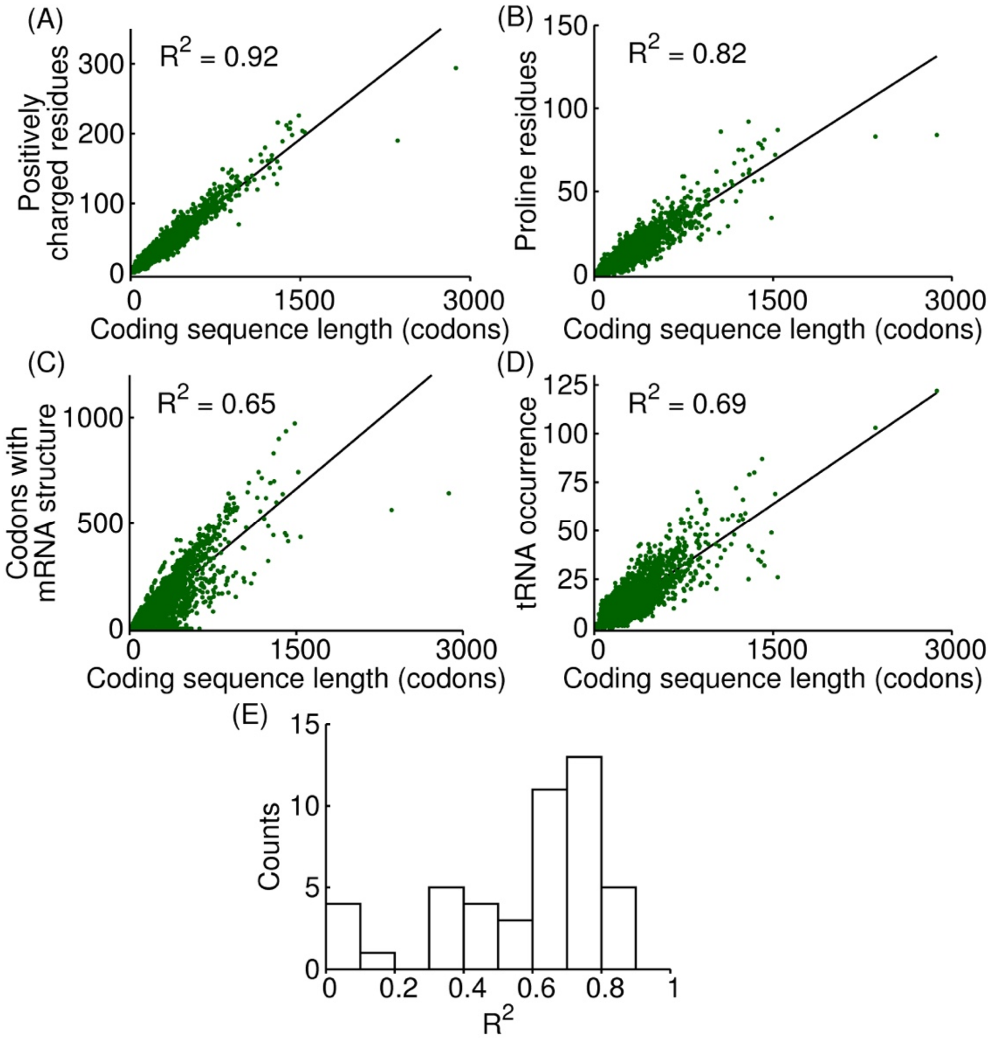

Figure 4: The molecular determinants of codon translation speeds scale linearly with coding sequence length in *E. coli*. The number of codon positions that encode positively charged residues, proline residues, and those codon positions that take part in mRNA structure are plotted against the coding sequence length in (**A**), (**B**) and (**C**), respectively. The occurrence of codon positions in a transcript that pair with a tRNA whose anti-codon is UUU and carries a lysine amino acid is plotted against the transcript length (**D**). A histogram of the correlation coefficient R^2 between the transcript length and the number of codon positions in a transcript that pair with 46 unique tRNA molecules in *E. coli* are plotted in (**E**).

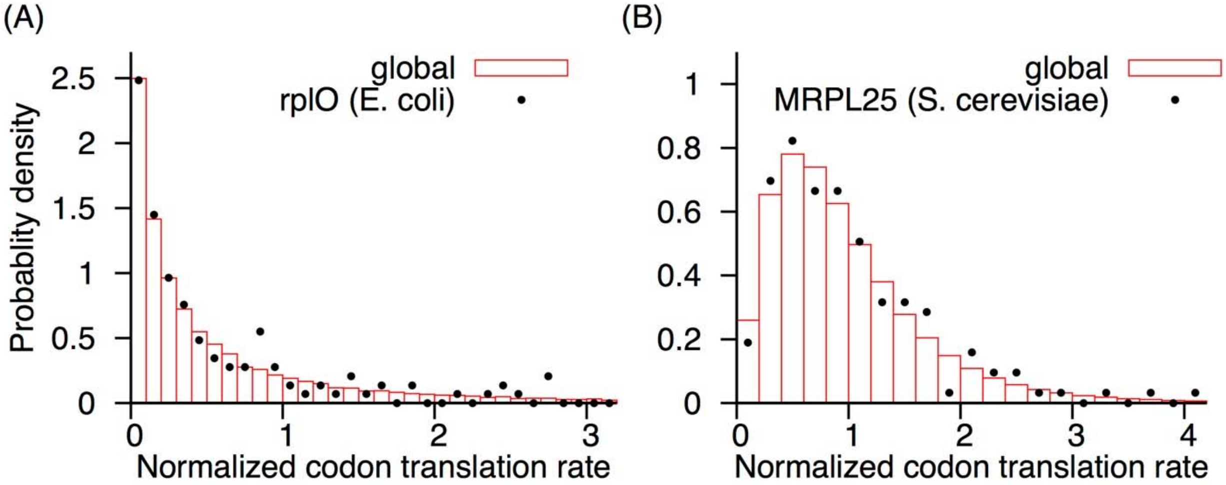

Figure 5: **Comparison between the distribution of normalized codon translation rates in an individual gene with the global distribution of codon translation rates.** (**A**) The distribution of codon translation rate for rplO *E. coli* gene and global distribution of codon translation rates in *E. coli* were plotted in black data points and red bars, respectively. (**B**) Same as (**A**) except for gene MRPL25 in *S. cerevisiae*.