G3
Genes | Genomes | Genetics

Accounting for Programmed Ribosomal Frameshifting in the Computation of Codon Usage Bias Indices

Victor Garcia,*,†,1 Stefan Zoller,‡ and Maria Anisimova*,†

*Zurich University of Applied Sciences, Wädenswil, Switzerland, †Swiss Institute of Bioinformatics, Lausanne, Switzerland, and ‡ETH Zurich, Zurich, Switzerland

ORCID ID: 0000-0001-6667-7098 (V.G.)

ABSTRACT Experimental evidence shows that synonymous mutations can have important consequences on genetic fitness. Many organisms display *codon usage bias* (CUB), where synonymous codons that are translated into the same amino acid appear with distinct frequency. Within genomes, CUB is thought to arise from selection for translational efficiency and accuracy, termed the *translational efficiency hypothesis* (TEH). Indeed, CUB indices correlate with protein expression levels, which is widely interpreted as evidence for translational selection. However, these tests neglect -1 programmed ribosomal frameshifting (-1 PRF), an important translational disruption effect found across all organisms of the tree of life. Genes that contain -1 PRF signals should cost more to express than genes without. Thus, CUB indices that do not consider -1 PRF may overestimate genes' true adaptation to translational efficiency and accuracy constraints. Here, we first investigate whether -1 PRF signals do indeed carry such translational cost. We then propose two corrections for CUB indices for genes containing -1 PRF signals. We retest the TEH in *Saccharomyces cerevisiae* under these corrections. We find that the correlation between corrected CUB index and protein expression remains intact for most levels of uniform -1 PRF efficiencies, and tends to increase when these efficiencies decline with protein expression. We conclude that the TEH is strengthened and that -1 PRF events constitute a promising and useful tool to examine the relationships between CUB and selection for translation efficiency and accuracy.

Across an organism's genome, among those codons that decode for the same amino acids (codon families) some are used preferentially over others (Sharp and Li 1986) (reviews in (Hershberg and Petrov 2008; Plotkin and Kudla 2011)). The variation of this *codon usage bias* (CUB) within organisms' genomes is mainly explained by selection for translation efficiency by protein expression demands –the translation

efficiency hypothesis (TEH) (Hershberg and Petrov 2008; Plotkin and Kudla 2011).

The main support for the TEH stems from two separate lines of evidence that concern associations between CUB and tRNA abundances and CUB and protein expression, respectively (Plotkin and Kudla 2011). In the first line, studies revealed that the most frequently used codons –so called *preferred codons* (Hershberg and Petrov 2008)– typically match the most abundant iso-accepting tRNA in both *Escherichia coli* (Ikemura 1985; Yamao *et al.* 1991) and *Saccharomyces cerevisiae* (Ikemura 1985). Similar conclusions were drawn for a number of additional species (Kanaya *et al.* 1999, 2001), but using tRNA gene copy numbers as proxies for tRNA abundances. The assumption here was that tRNA gene copy numbers and tRNA abundances correlate, as is the case in *E. coli* and *yeast* (Ikemura, 1981, 1982). Further studies suggest that employing these more abundant tRNAs in translation provides efficiency (Tuller *et al.* 2010) as well as accuracy gains (at least for *E. coli* (Stoletzki and Eyre-Walker 2007; Varenne *et al.* 1984) and *Drosophila melanogaster* (Akashi 1994)).

Furthermore, exchanging preferred codons in highly expressed genes by non-preferred had strong effects on gene expression (Carlini and Stephan 2003; Carlini 2004). If valid, these two mechanisms together imply that the disproportionately frequent usage of such preferred codons in an mRNA sequence is indicative of high translational efficiency.

The second line of evidence concerns the existence of associations between CUB and protein expression (Ghaemmaghami *et al.* 2003; Jansen *et al.* 2003; Kliman *et al.* 2003; Roth *et al.* 2012; Quax *et al.* 2015). High protein expression demands are assumed to generate a selective pressure for translation efficiency (Hershberg and Petrov 2008; Chamary *et al.* 2006; Duret 2002). Given the mechanisms of the first line of evidence, these efficiency increases could be provided by synonymous mutations that alter codon usage frequencies toward preferred codons (Hershberg and Petrov, 2008, 2009). Taken together with the first line of evidence, these associations suggest that protein expression is also associated with translation efficiency, completing the support of TEH.

The test of TEH by detection of CUB-protein expression association does therefore not rely on a direct comparison between translation efficiency and protein expression. Instead, CUB indices serve as a measure for adaptation due to translational selection and thus as a proxy for translational efficiency (Roth *et al.* 2012). CUB indices are computed from a given mRNA sequence. Thus, the test of the TEH by CUB-protein expression association hinges on the assumption that the CUB index of an analyzed mRNA appropriately reflects its translational efficiency.

The phenomenon of *programmed ribosomal frameshifting* (PRF) (Jacks and Varmus 1985; Craigen *et al.* 1985; Clare *et al.* 1988; Caliskan *et al.* 2015; Dinman, 2006, 2012b, 2012a) indicates that the validity of that assumption may not always be warranted. Programmed ribosomal frameshifting is a process by which ribosomal translation of an mRNA sequence is induced to stall at specific sites, termed *slippery sites*, which leads to a rearrangement of the ribosome on the mRNA sequence (Plant and Dinman 2006; Caliskan *et al.* 2015). Translation then proceeds in a new frame that is shifted relative to the original *open reading frame* (ORF) (Belew and Dinman 2015). This mechanism emerges in all domains of the tree of life (Cobucci-Ponzano *et al.* 2012; Baranov *et al.* 2015; Caliskan *et al.* 2014; Dinman, 2012b). Here, we focus on events that shift ribosomes one nucleotide back in the translational direction, termed *-1 PRF* (Caliskan *et al.* 2015; Advani and Dinman 2016), further restricting ourselves to *yeast*.

The function of -1 PRF differs across organisms. In viruses, -1 PRF is predominantly employed to package more information into available sequence material. By opening a new reading frame, the -1 PRF mechanism allows for dual coding. This function is exemplified in the *gag-pol* gene overlap appearing in many retroviruses (Jacks and Varmus 1985), in particular human immunodeficiency virus (HIV) (Brierley and Dos Ramos 2006). In these viruses, the appropriate proportion of *gag vs. pol* protein expression is regulated by a -1 PRF signal (Maia *et al.* 1996; Dulude *et al.* 2006; Brierley and Dos Ramos 2006). In eukaryotes, -1 PRF is predominantly ($>$99%) used for gene expression regulation (Plant *et al.* 2004; Advani and Dinman 2016). About 10% of genes in yeast are hypothesized to contain functional slippery sites (Belew *et al.* 2008). Ribosomes redirected to -1 shifted frames majorly encounter *premature termination codons* (PTC), that is, stop codons that appear well before the poly-A tails of mRNA sequences (Advani and Dinman 2016). This typically triggers the activation of the non-sense-mediated decay (NMD) pathway (Chang *et al.* 2007). The NMD pathway induces the degradation of both the mRNA, as well as of the partly assembled protein, while leaving the ribosome intact (Advani and Dinman 2016). Thus, -1 PRF acts as an NMD-mediated

destabilizing element of mRNA (Belew *et al.* 2011, 2014; Advani *et al.* 2013). -1 PRF efficiency appears to be controlled by sequence specific elements, such as miRNA (Belew *et al.* 2014), but can be uniformly affected across all PRF-harboring genes by mutations as well as drugs (Advani *et al.* 2013).

The reason why CUB may not appropriately reflect translation efficiency for genes containing -1 PRF signals, is because -1 PRF is likely to conceal costs to translation. When a -1 PRF signal redirects a fraction of the mRNA-translating ribosomes to premature termination (Dinman, 2012b; Advani and Dinman 2016), a translation efficiency cost is incurred that is not reflected in the sequence's CUB index.

In this study, we retest the CUB-protein expression association that lends major support for the TEH while accounting for the -1 PRF phenomenon. We draw from *PRFdb*, a database of predicted -1 PRF signals in the *Saccharomyce cerevisiae* genome (Belew *et al.* 2008). First, we devise a series of hypotheses for how -1 PRF signals are influenced by evolutionary pressures. In accordance with our hypotheses, we find evidence for the existence of a cost of -1 PRF mechanism maintenance. Second, we devise general corrections for any codon usage bias index of mRNA containing -1 PRF signals. We use sequences from PRFdb to compute the new, corrected codon bias indices, and compare them to uncorrected values. We find that the TEH is robust against these corrections and strengthened under biologically plausible assumptions of a -1 PRF dependency on protein expression. We conclude that the -1 PRF signals offer untapped potential to analyze translation efficiency in mRNA.

MATERIALS AND METHODS

Slippery site data

We obtained a dataset of the *Saccharomyces cerevisiae S288C* strain genes with predicted -1 PRF events from the Programmed Ribosomal Frameshifting database (PRFdb) of the University of Maryland (Belew *et al.* 2008). These putatitve -1 PRF signals were identified using first a filter for slippery site identification, and subsequent detection of mRNA pseudoknots by means of the *Nupack* algorithm (Dirks and Pierce 2004). For a slippery site to be found, a confluence of signals in mRNA is required. First, a site must match the pattern (X XXY YYZ), where X is some base, Y is either A or U, and Z is A, C or U. Second, a spacer sequence of a minimum of 8 nucleotides in length needs to exist between the slippery site and the following pseudoknot. Third, a pseudoknot predicted by minimum free energy values of the mRNA secondary structure is expected downstream the slippery site. The dataset also included a full list of gene annotations, accession numbers, the relative position of the slippery sites within the gene in which they were found, as well as the gene mRNA. For each slippery site, we computed were premature stop codons first appear downstream. With this, we computed the frameshifted sequences.

Protein expression and mRNA level data

We obtained protein expression as well as mRNA level data from the Supporting Information of (von der Haar 2008). Von der Haar has produced an extensive curated data set that merges data from various sources. For protein expression, the data include the seminal studies of Ghaemmaghami *et al.* (Ghaemmaghami *et al.* 2003), Newman *et al.* (Newman *et al.* 2006), and Lu *et al.* (Lu *et al.* 2007). Additional data stems from 46 further studies, specified in (von der Haar 2008). To ensure comparability, only studies were included in which yeast was grown in rich medium. Transcriptome data for mRNA levels were obtained from (Holstege *et al.* 1998; Arava *et al.* 2003; Holland 2002; Jelinsky and Samson 1999; Roth *et al.* 1998). These data include about

6000 genes. The construction of the curated data set is described in detail in (von der Haar 2008).

Indices for codon usage bias

Many indices have been proposed to assess the degree of codon usage bias present in a gien mRNA sequence. Here, we focus on the codon adaptation index (CAI) (Sharp and Li 1987; Carbone *et al.* 2003). In the following, we give the implemented definition of the CAI. Let L be the length of the mRNA sequence in codons, c is the index of the synonymous codons decoding the same amino acid a, and o_{ac} is the observed count of the synonymous codon c of amino acid a in the sequence. C_a is the index set of all codons within an amino acid a. A is the index set of all amino acids a. We define

$$\omega_{ac} \doteq \frac{o_{ac}}{\sum_{c \in C_a} o_{ac}}, \tag{1}$$

as the relative adaptiveness of codon ac within amino acid a of a given mRNA sequence.

Then, the codon adaptation index is defined as

$$CAI_{\mathrm{mRNA}} \doteq \frac{1}{L} \sum_{a \in A} \sum_{c \in C_a} o_{ac} \ln\left(\omega_{ac}^{\mathrm{ref}}\right), \tag{2}$$

where o_{ac} is measured on the observed on the mRNA of interest, and $\omega_{ac}^{\mathrm{ref}}$ is taken from a reference sequence of highly expressed genes. For the reference sequence, we concatenated all of the mRNA of ribosomal genes (Friberg *et al.* 2004). The ribosomal genes were taken from the ribosomal gene database http://ribosome.med.miyazaki-u.ac.jp/ (Nakao *et al.* 2004).

Data Availability

The code used for analysis of the data are available on Figshare (doi: 10.6084/m9.figshare.5955886; url: https://figshare.com/s/ff89f9e614b129469fd4). Supplemental material available at Figshare: https://doi.org/10.6084/m9.figshare.6970277.

RESULTS

The cost of -1 PRF maintenance

Effective -1 PRF events classically comprise three constitutive elements. The first element is a heptameric sequence –the *slippery site*– (Jacks and Varmus 1985; Plant *et al.* 2004). This site has the structure (X XXY YYZ), where X can be any base (A,C,G,U), Y is either A or U, and lastly, Z is either A, C or U (Plant *et al.* 2004; Plant and Dinman 2006; Ketteler 2012). The second element is a *spacer sequence*, separating the slippery site from the third element (Dinman, 2012b). The spacer sequence typically comprises around 1-12 nucleotides (Ketteler 2012; Dinman, 2012a). The third element is a *pseudoknot*, an mRNA secondary structure that is thought to generate a mechanical tension on the spacer sequence when the slippery site is being translated (Ketteler 2012; Dinman, 2012b, a). To release the tension, the ribosome, while stalling at the slippery site, is pushed one nucleotide back (Dinman, 2012b; Caliskan *et al.* 2015; Ketteler 2012). The frequency with which the ribosome is redirected to the -1 frame is called *-1 PRF efficiency* (Advani *et al.* 2013; Advani and Dinman 2016).

These criteria were used to generate an extensive database of -1 PRF signals, *PRFdb*, of the yeast genome by algorithmic search (see *Materials and Methods*). If PRFdb comprises a sufficiently large set of functional slippery sites, the effects of putative -1 PRF induced costs to translation efficiency should be detectable.
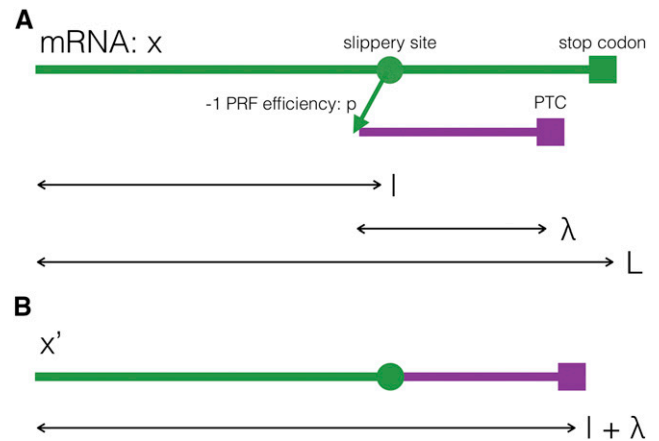

Figure 1 Notation for -1 PRF signals. A) The mRNA sequence in which a -1 PRF signal does not lead to ribosomal frameshifting is termed *x* (green). Arrows indicate the distances in nucleotides. B) The mRNA sequence with -1 PRF is termed *x′*. This is the concatenation of the mRNA *x* until the slippery site (green), and the -1 frame after the slippery site until a premature termination codon is encountered (violet).

To this end, we devised three hypotheses. We derive these hypotheses from the assumption that the maintenance of -1 PRF signals carries an intrinsic cost to the organism. Furthermore, we assume that mutations that alter CUB are accumulated more slowly than mutations that alter -1 PRF efficiency, p. This is because the -1 PRF regulation mechanism is very likely to be affected by only a few mutations, whereas several mutations would be needed to significantly modify CUB. Hence, most short-term changes in protein expression demand in genes with slippery sites are more likely absorbed by -1 PRF.

We first introduce a framework to formulate these hypotheses (see Figure 1). Let L be the length (in nucleotides) of an mRNA sequence x. A single slippery site is located at site l of x, downstream from the start. Ribosomes are redirected to the -1 frame with probability p, the *-1 PRF efficiency*. They continue with the mRNA's translation until encountering a premature termination codon (PTC) at a distance λ post slippery site.

In a first hypothesis, we formed expectations about how the number of slippery sites per gene should vary with the ratio of protein expression to mRNA levels per cell. This ratio has been used as a measure for translation efficiency in past studies (Tuller *et al.* 2010). To formulate the first hypothesis, we conduct a thought experiment. Let us assume that an organism's environmental conditions favor lower expression of a protein from a *-1 PRF*-gene. Then, increasing the average p in that gene will adapt the organism to the new conditions. Conversely, if such environmental changes demanded higher protein expression, genes containing -1 PRF signals may adapt by reducing p. If protein demands exceed the production capacity of a gene with very small p, the -1 PRF mechanism may offer only costs, but no benefits from protein regulation. Then, mutations that remove the heptameric slippery site signature will be selected for. Taken together, these effects imply an asymmetry in the probability of slippery site loss depending on the direction of protein demand changes. Losses of slippery sites are likelier when protein demands increase than when they decrease. -1 PRF signals should thus become rarer with increasing protein expression levels.

The second hypothesis states that the costs of -1 PRF maintenance are reduced as slippery sites approach the 5′ end or start of the mRNA. Every time ribosomes slip through -1 PRF, the energy of translating
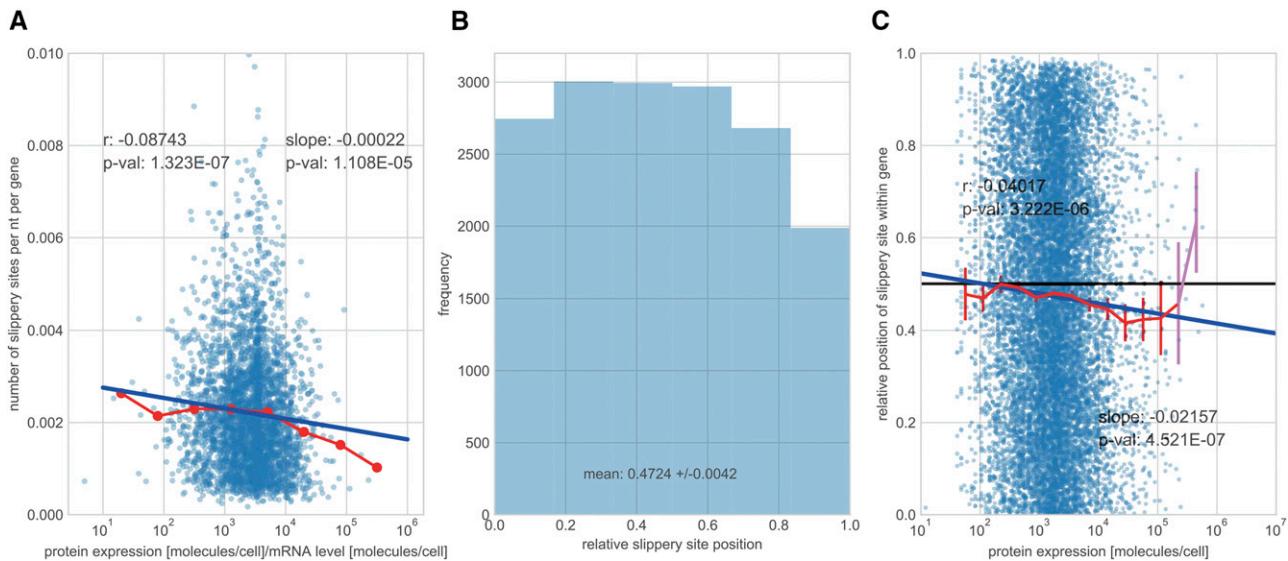
Figure 2 Testing the cost of -1 PRF mechanism maintenance. A) Number of slippery sites –identified by relying on *Nupack* algorithm for pseudoknot detection (Dirks and Pierce 2004)– per gene and per nucleotide across protein expression levels per mRNA level for genes from the von der Haar data set (von der Haar 2008). The red dotted line is the average number of slippery sites per gene. The blue line is a regression line through the data set. The text in the panels gives i) the Spearman correlation coefficient *r* and the p-value for the null-hypothesis that the correlation is zero and ii) the slope of the line and the p-value of the *t*-test for a non-zero slope value. B) The frequency distribution of the within-gene positions of the slippery sites, relative to the length of the gene, l/L. To ensure comparability, only genes from the von der Haar data set are considered. The mean of the distribution, as well as its 95% confidence intervals, are smaller than 0.5 -the expectation in the absence of selective pressure. C) Slippery site positions relative to gene length across protein expression levels in the von der Haar data set. The red line is the average slippery site position (computed across 20 bins of equal width in logarithmic scale) and the widgets are the uncertainty (±1.96 standard error of the mean) around the average estimate. Averages with large uncertainties are in violet. Analogously to A), a regression line with the corresponding correlation coefficient *r* (with p-value) and slope (with p-value) are added.

a stretch of length $l + \lambda$ of mRNA is wasted for protein production. Since λ is roughly constant across within-gene sites (see Fig. S1), the cost to translational efficiency incurred from -1 PRF should depend on l. Therefore, diminishing l will minimize cost. Thus, the distribution of slippery site positions relative to the gene length should be skewed toward mRNA starts, that is, small l values.

This effect is expected to be weaker than the effect in the first hypothesis. In the first hypothesis, if protein demands exceed yields generated by a gene and p is small, a fitness cost is incurred from i) the reduction of protein expression, ii) the cost of NMD-mediated mRNA degradation and iii) the cost of unnecessary translation of frameshifted mRNA sequences. Displacing the slippery site would only reduce the latter cost. Thus, we expect the signal for this hypothesis to be weaker than for the first hypothesis.

The third hypothesis states that a skew in the distribution of the location of the slippery site within a gene (relative to that gene's length) should also become more pronounced with higher expression levels. This is because for fixed p, increasing expression levels should mirror an increase in -1 PRF induced translation costs. Thus, the benefits of slippery sites closer to translation-initiation positions should increase with expression levels. As in the second hypothesis, selection for such an effect is likely to be very weak compared to effects in the first hypothesis.

Figure 2 shows the results of testing all three hypotheses. Figure 2A shows that -1 PRF signals in genes become rarer with gene expression, consistent with our first hypothesis. Figure 2B shows that slippery sites positions are more prevalent in the first half of a gene than in the second, consistent with our second hypothesis. The frequency of slippery sites diminishes toward both extremes of mRNA sequences, and more markedly so toward the stop codon of the mRNA. This skew

persists when including all genes from the PRFdb (see Figure S2). Figure 2C shows that the average slippery site position is displaced toward the 5′ end of the mRNA as protein expression levels increase from 10^2 molecules per cell to 10^5 molecules per cell. Above 10^5 molecules per cell, the uncertainty around the averages becomes large due to low sample sizes, and unambiguous deductions become impossible. The decreasing trend in the average l/L with protein expression is confirmed by a linear regression. Lastly, we retested all of the hypotheses with an alternative integrated data set from PaxDB, and obtained the same results (see Figure S3).

Thus, with increasing demand for protein production, the data suggest that production gains will primarily be attained by disposing of the -1 PRF mechanism present in a gene, rather than minimizing the cost of faulty translation. We speculate that, most likely, this occurs by altering the slippery site sequence. The test of the other two hypotheses involving slippery site displacement (or loss of large-l slippery sites) provide further evidence for an intrinsic -1 PRF cost to translation. The selective pressure on this latter adaptive process appears to be weak.

Retesting the translational efficiency hypothesis while accounting for -1 PRF genes

Since -1 PRF signals carry a translational cost, their presence in an mRNA sequence changes the meaning of the associated CUB index. In fact, a slippery site may indicate a mismatch between the CUB index of that sequence and the actual efficiency with which it is translated (see Figure 3).

To elucidate this point, consider an mRNA without -1 PRF signal. In the -1 PRF signal's absence, the CUB index of a translated mRNA sequence appropriately reflects how efficiently ribosomes elongate it in a pre-specified time (see Figure 3A) (Tuller *et al.* 2010; Stoletzki and
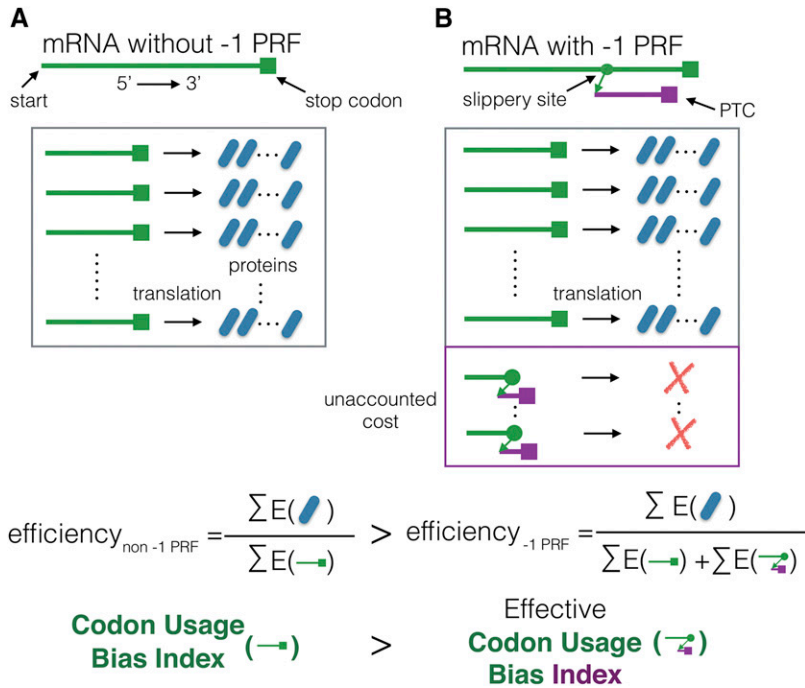
Figure 3 -1 PRF signals bias codon usage bias indices' predictiveness of translation efficiency. A) An mRNA sequence (green line) without a -1 PRF signal. Within a given period of time, the transcribed mRNA in the cell will produce a finite number of proteins. We define the efficiency of the protein synthetisation process as the energy content of the proteins (the nominator) divided by the energy expended by the mRNA transcription and translation processes (the denominator) (Tuller *et al.* 2010). The value of a codon usage bias index should reflect the efficiency of the protein synthesis process under the TEH. B) The same mRNA considering a slippery site (green point). The slippery site redirects ribosomes to a -1 shifted frame, where they encounter premature termination codons (PTC). The same number of proteins are produced as in A). An additional cost, unaccounted for in A), is incurred for protein production due to NMD-mediated protein decay. Thus, the effective translation efficiency when accounting for -1 PRF must be smaller than A). This diminished efficiency should be reflected by a correction in the codon usage bias index value.

Eyre-Walker 2007; Akashi 1994). However, given a functional slippery site in that same mRNA sequence, the interpretation changes (see Figure 3B). Let us assume an average -1 PRF efficiency p. Then, a fraction p of the times in which a ribosome is in the process of translating a specific mRNA, it will be redirected to another frame by -1 PRF. This redirection leads to degradation of the mRNA. Thus, the production of the same number of proteins will require more energy and time in the presence of a -1 PRF signal than it would in its absence. However, traditional CUB indices will give no indication of that process. If accounted for, -1 PRF must therefore lead to a downward correction of the mRNA's translational efficiency and with it, of the associated CUB index.

Since a substantial fraction of genes may contain -1 PRF signals (hypothesized to be $\approx$ 10% of genes in yeast (Belew *et al.* 2008)), neglecting this effect might introduce considerable biases in CUB index values. These biases could affect the associations found between CUB and protein expression (Ghaemmaghami *et al.* 2003; Jansen *et al.* 2003; Kliman *et al.* 2003; Roth *et al.* 2012; Quax *et al.* 2015). To address this issue, in the following we propose two general corrections for CUB indices in the presence of -1 PRF signals, and retest the basis for the TEH stemming from CUB and protein expression correlations.

Codon Usage Bias Index Correction from Translation Efficiency: In a first approach, we derive an estimator for a CUB index correction from basic relationships between CUB indices and translation efficiencies of an mRNA. In the absence of a -1 PRF signal, the CUB on x will be measured by some index function $I(x)$. As noted, if a functional -1 PRF signal is present, $I(x)$ does not account for the translation efficiency loss due to the unproductive translation of the hybrid sequence x' (see Figure 1). We aim to derive a "corrected" index $I_c(x)$ that more appropriately reflects the diminished translation efficiency in the presence of -1 PRF signals.

To find an expression for $I_c(x)$, we begin with the assumption that there exists a monotonous mapping F that maps the translation efficiency $\eta(x)$ of a sequence x (in the absence of -1 PRF) to a codon usage bias index I: $I(x) = F(\eta(x))$. F is a monotonously increasing function,

where increases in translation efficiency are reflected by increases in the codon usage bias index. Research by Tuller *et al.* further suggests that F is concave (Tuller *et al.* 2010), such that codon usage bias index values saturate with increasing translation efficiency values. We define translation efficiency $\eta(x)$ classically as an input-output energy ratio. More specifically, $\eta(x)$ is the ratio of the total per-protein energy $E_p(x)$ contained in the $n(x)$ synthesized proteins from x to the energy, $E_i(x)$, used for producing those proteins in a sufficiently long time frame Δt :

$$\eta(x) = \frac{n(x)E_p(x)}{E_i(x)}. \tag{3}$$

Here, the energy going into the synthesization machinery, $E_i(x)$, does not contain the expenditure for -1 PRF's regulatory use. Thus, for a sequence x that additionally carries a functional slippery site, we have

$$\eta(x) \rightarrow \eta_c(x) \equiv \frac{n_c(x)E_{p,c}(x)}{E_{i,c}(x)} = \frac{n(x)E_p(x)}{E_i(x) + E_i(x')}. \tag{4}$$

Here, the same number of proteins were produced as when -1 PRF was not considered, that is, $n_c(x) = n(x)$. Since the final proteins are structurally equivalent to those in the absence of -1 PRF, we also have $E_{p,c}(x) \equiv E_p(x)$. However, more energy was expended to produce these proteins, and therefore $E_{i,c}(x) = E_i(x) + E_i(x') > E_i(x)$. Their synthesization requires at least $E_i(x)$. A part of the expended energy on translation is not implemented in the proteins, but in the translation of x' and the NMD pathway activation, $E_i(x')$. Thus, it follows that $\eta_c(x) < \eta(x)$.

The corrected usage bias statistic $I_c(x)$ is defined as

$$I_c(x) \doteq F(\eta_c(x)) = F\left(\frac{n(x)E_p(x)}{E_i(x) + E_i(x')}\right). \tag{5}$$

Since the form of F as well as the values of $E_p(x)$, $E_i(x)$ and $E_i(x')$ are either difficult to measure or unknown, we aim to to compute $I_c(x)$ indirectly from $I(x)$. To this end, we separate the translation efficiency $\eta_c(x)$ into two components:

$$\eta_c(x) = \eta(x) + Q(x, x'). \qquad (6)$$

Solving for $Q(x, x')$ gives

$$Q(x, x') = -\frac{n(x)E_p(x)}{(E_i(x) + E_i(x'))}\frac{E_i(x')}{E_i(x)} = -\eta_c(x)R(x, x'), \qquad (7)$$

where $R(x, x') \equiv E_i(x')/E_i(x)$. With this, we have

$$\eta_c(x) = \frac{\eta(x)}{1 + R(x, x')}. \qquad (8)$$

Further analysis of the ratio $R(x, x')$ is complicated by the inability to directly measure $E_i(x)$ and $E_i(x')$. To address this issue, we pursue an approach where the ratio is approximated by information about the relative one-elongation energies spent translating x and x'. More specifically, we assume that each time x is translated, an energy input of $T(x) \doteq E_i(x)/n(x)$ is expended per time unit Δt. The total number of times translation is initiated on the mRNA sequence, $N(x)$, is $N(x) = n(x)/(1 - p)$. This is because $n(x)$ corresponds to the fraction $1 - p$ of times that an mRNA-elongating ribosome remains in frame. Analogously, the number of times translation is interrupted by a -1 PRF event is $n(x') \equiv N(x)p = n(x)p/1 - p$. Hence, each time the sequence x' is translated, an energy expenditure of $T(x') \doteq E_i(x')/n(x')$ is ensued. Note that $n(x')$ does not correspond to a protein number. The ratio between these two is approximated by:

$$\frac{T(x')}{T(x)} = \frac{E_i(x')}{E_i(x)}\left(\frac{1 - p}{p}\right) \approx \frac{l + \lambda + \pi}{L}, \qquad (9)$$

which is independent of the number of synthesized proteins $n(x)$. Here, we have assumed that ratio of the per-run translation energies expended for the non-frameshifting to the frameshifting scenarios corresponds roughly to the ratio of the lengths of the translated sequences x' to x, respectively. However, there is an extra cost to each frameshifted mRNA, x' stemming from NMD-mediated degradation. This cost is accounted for by π, which is an unknown fixed cost associated with each -1 PRF event. π is measured in units of equivalents of translation cost per nucleotide.

With (9), we can approximate the $R(x, x')$:

$$R(x, x') \doteq \frac{E_i(x')}{E_i(x)} \approx \frac{l + \lambda + \pi}{L}\left(\frac{p}{1 - p}\right). \qquad (10)$$

With this, we are ready to address the last approximation required to find an analytical expression for $I_c(x)$. Since F is concave, it follows that $F(a\eta) \geq aF(\eta)$ for $a \in \mathbb{R}$. Thus, setting $a = (1 + R(x, x'))^{-1}$, and using definition (4), we find a lower bound for $I_c(x)$:

$$I_c(x) \doteq F(\eta_c(x)) = F\left(\left(1 + R(x, x')\right)^{-1}\eta(x)\right)$$
$$\geq \left(1 + R(x, x')\right)^{-1}F(\eta(x))$$
$$= \hat{I}_c(x) \doteq \left(1 + \frac{l + \lambda + \pi}{L}\left(\frac{p}{1-p}\right)\right)^{-1}I(x).$$
$$(11)$$

The approximation for $R(x, x')$ results in reasonable and useful properties of $\eta_c(x)$. More specifically, as $p \to 1$ we have that $\eta_c(x) \to 0$. This entails that no mRNA is translated into proteins, as expected. The same follows if $\pi \to \infty$. The extraction of a from within the

brackets leads to the undesirable behavior that as $p \to 1$ or $\pi \to \infty$, we have that $\hat{I}_c(x) \to 0$. Instead, $\hat{I}_c(x)$ should approximate some minimal value $I_{c,\min}(x) = F(0)$.

A major drawback of $\hat{I}_c(x)$ is its reliance on p and π, which are both unknown. Typically, p is assumed to lie between $1 - 10\%$, but may reach up to 70% (for example the *EST2* gene in yeast, (Advani *et al.* 2013)). In the following, and if not stated otherwise, we assume that p is constant across all genes. It is unclear how large π should be. We have found no evidence that the value of π is dependent on any characteristic of mRNA, such as its length, or codon composition (Chang *et al.* 2007; Dinman, 2012b). NMD is an evolutionarily conserved surveillance pathway (Chang *et al.* 2007). Its activation may thus differ in energy expenditure between organisms (indeed, in *yeast*, NMD does not require an exon-junction complex, unlike most other eukaryotes). However, we have found no indication that the energy required to fully complete the NMD process will substantially vary from mRNA to mRNA within an organism. Thus, we assume that π is uniform across genes, although we do not know how large it is relative to other energy inputs. Again, if not stated otherwise, we assume that $\pi = 0$ in order to give a conservative estimate of the effects of -1 PRF on the TEH.

Codon Usage Bias Index Correction from Averaging: In a second approach, we define a correction for a -1 PRF-aware CUB index using a balancing principle. To this end, we add the CUB index $I(x)$ in successful, non-NMD mediated translations of x and the index $I(x')$ of untranslated x', while weighing both with their respective probability of occurrence:

$$I_a(x) \doteq (1 - p)I(x) + pI(x'). \qquad (12)$$

The corrected index $I_a(x)$ is thus the expected value of I when considering that x is translated a fraction $(1 - p)$ of the time, and x' is translated a fraction p of the time. $I(x')$ therefore acts as a penalization function. I_a may underestimate the extra energy costs to -1 PRF incurred from the activation of the NMD-mediated degradation processes, since these are not comprised in $I(x')$. Similarly to $\hat{I}_c(x)$, the major drawback of the estimator $I_a(x)$ is its dependence on p. As before, we thus evaluate $I_a(x)$ for different, plausible values of p to assess -1 PRF's effect on the TEH. However, $I_a(x)$ has the advantage that it is also well defined for $p = 1$.

Reexamining the TEH with corrected CUB indices: Both corrections $\hat{I}_c$ and I_a represent a change in the value of a codon usage bias index given a -1 PRF efficacy. The models for correcting I presented here do not relate these measures to protein expression levels, P. They do therefore not predict protein expression. Instead, they allow us to reexamine the association between protein expression and a corrected codon usage bias index that in part underpins the TEH.

To examine how these corrections affect associations between CUB and protein expression, we used the widely employed *codon adaptation index* (CAI) (Carbone *et al.* 2003) as an example for a CUB index I. We then computed both corrected $\hat{I}_c$ and I_a for different values of p and for all genes in the von der Haar data set (von der Haar 2008) that contain -1 PRF signals. For $\hat{I}_c$ we also assumed different values of π.

A precise computation of $\hat{I}_c$ and I_a would require information about the value of p across genes. Since no such information is available as of now, we follow two approaches. First, we treat all -1 PRF sites equally, assuming an equal p value for all, independent of the protein expression levels P of the genes they are located in. Second, we assume that p is associated with P, and allow p to slowly decrease with the P of the gene.
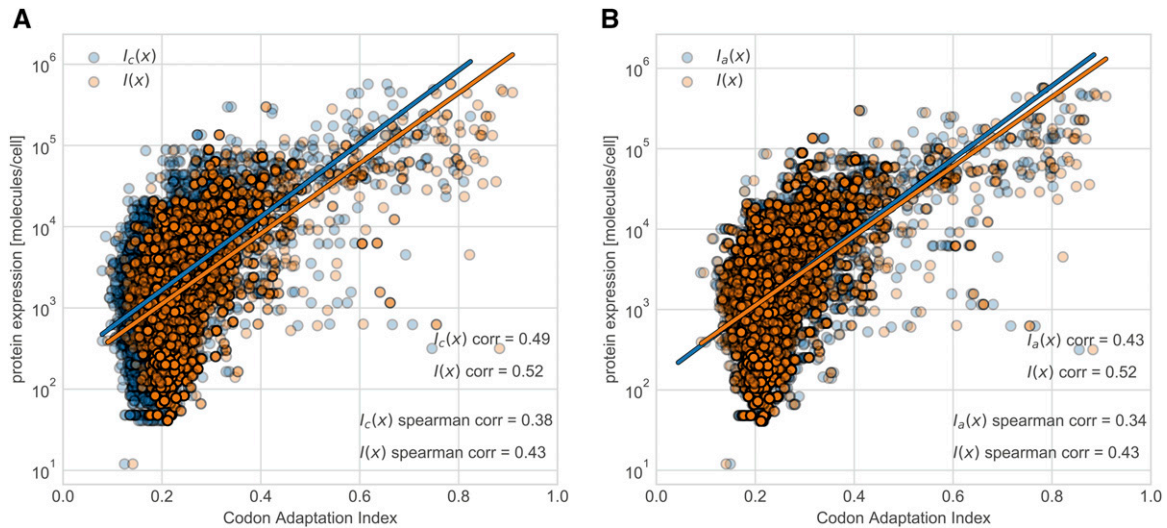
Figure 4 Corrected and original codon adaptation indices *vs.* protein expression levels for genes containing algorithmically identified -1 PRF signals and with known expression levels (von der Haar 2008), assuming $p = 0.3$ and $\pi = 0$. A) The original codon adaptation index values of mRNA sequences x of genes from the (von der Haar 2008) data set are shown as orange, semi-transparent points. The shown genes all contain algorithmically identified -1 PRF signals. The shape of the relation between CAI and protein expression is similar to that shown in (Ghaemmaghami *et al.* 2003) for all genes with known expression levels. Blue, semi-transparent points denote the values of -1 PRF corrected CAI values, $\hat{I}_c$, *vs.* same protein expression levels. Regressions between the $\log_{10}$-protein expression levels and both corrected an original CAI values are shown as orange and blue lines, respectively. B) is analogous to A) using the corrected CUB index I_a.

Figure 4 shows that correcting for the presence of -1 PRF signals with a -1 PRF efficiency of up to $p = 0.3$ does not substantially affect the CUB index to protein expression relationship for both CUB index corrections, if the extra cost of mRNA degradation by NMD is neglected ($\pi = 0$). Figure 4A) shows that all CAI values for all genes are diminished when using the correction $\hat{I}_c$. A substantial correlation between protein expression and corrected CAI index remains. Accounting for the frameshifted sequences x' also diminishes codon adaptation measures I_a considerably compared to uncorrected CUB index values, as shown in Figure 4B). However, unlike with $\hat{I}_c$ this reduction does not correspond to a uniform negative offset. Instead, the effect of accounting for -1 PRF is to both shift and broaden the distribution of $I_a(x)$ values relative to $I(x)$ (see Figure S4).

The two corrections $\hat{I}_c$ and I_a show different sensitivities to the -1 PRF efficiency. In particular, the sensitivity of $\hat{I}_c$ to -1 PRF efficiency is mediated by the value of π. Figure 5A shows that the correlations between both $\hat{I}_c$ and I_a to protein expression levels decline for increasing values of p and different values of π. As p increases, and for low costs of NMD-mediated mRNA decay, the correlation of $\hat{I}_c$ to protein expression levels declines to very low values (~ 0.2 at $p = 0.9$ for $\pi = 0, 10, 100$). If the energetic cost of -1 PRF induced mRNA degradation, π, becomes larger ($\pi = 1000, 10000$), the correlation declines very rapidly with p, and almost vanishes when p reaches ~ 0.5. For I_a estimates, correlations' dependency on p is moderate, never falling below 0.4 at biologically implausible values of p of unity.

These results corroborate the support for the translation efficiency hypothesis. For uniform -1 PRF efficiencies and for $\hat{I}_c$-based estimates, correlations of corrected CAI to protein expression levels decline with p. They only become sufficiently diminished to challenge the TEH when both the values of π and p are very high. Indeed, according to our mathematical framework, $\pi = 1000, 10000$ corresponds to the cost of either translating a large gene, or tenfold

that cost. Except for $\pi = 10000$, all correlation values between corrected CUB indices and protein expression levels remain at around $0.4 - 0.5$ for biologically relevant -1 PRF efficiencies of $p = 1 - 10\%$. For I_a estimates, correlation seems generally robust to changes in p.

Varying -1 PRF efficiency with protein expression levels: Since the data in Figure 2 strongly suggest a cost to -1 PRF maintenance, we also explored how a -1 PRF efficiency decline with protein expression levels consistent with such cost would affect the correlations in Figure 5A. Unlike a uniform p across protein levels, corrections to $\hat{I}_c$ are expected to be larger at low expression levels than at high levels. The -1 PRF decline is assumed as follows: $p(P) = p_b/\log_{10}(P)$, where we call p_b the *baseline -1 PRF efficiency*.

For $\hat{I}_c$, this additional assumption leads to a measure:

$$\hat{I}_c(x) = \left(1 + \frac{l + \lambda + \pi}{L} \left(\frac{p_b}{\log_{10}(P) - p_b} \right) \right)^{-1} I(x). \qquad (13)$$

Analogously, I_a becomes

$$I_a(x) = I(x) - \left(p_b - \log_{10}(P) \right) \left(I(x) - I(x') \right). \qquad (14)$$

Figure 5B shows that under this assumption, correlations between protein expression and $\hat{I}_c$ rise with larger baseline -1 PRF efficiencies for all except the largest values of π, while again I_a remains unaffected. We observe analogous results using the Spearman rank correlation (see Figure S5).

When -1 PRF efficiencies decline with protein expression –in accordance with a cost to -1 PRF– most correlations between $\hat{I}_c$-based estimates and protein expression rise with -1 PRF baselines. Except for $\pi = 10000$, all correlation values increase with -1 PRF baselines. This also adds support for the TEH, suggesting that
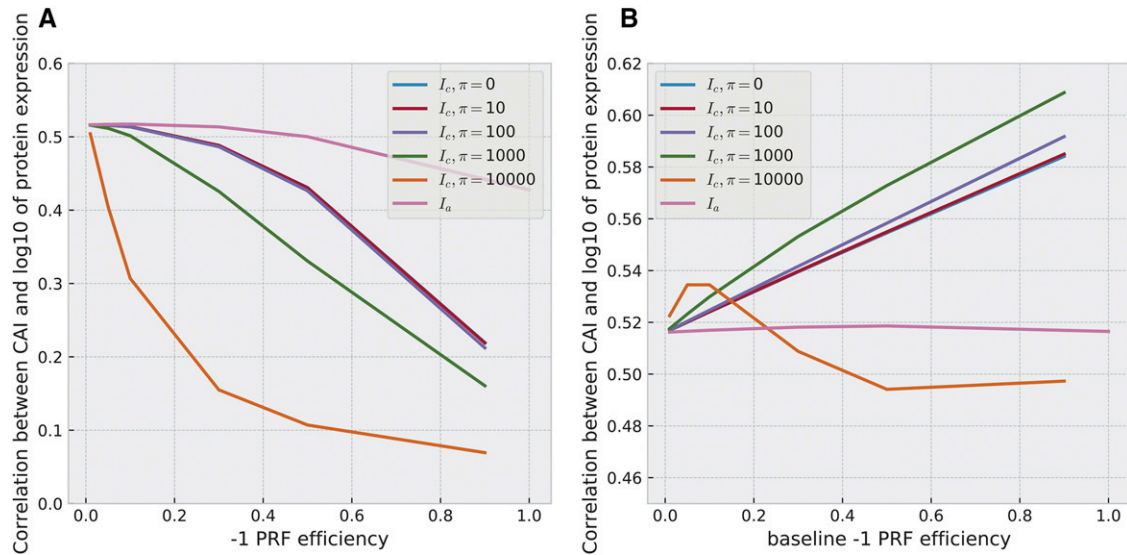
Figure 5 Correlation values between corrections of CAI to -1 PRF presence ($\hat{I}_c$, I_a) to protein expression levels P (von der Haar 2008) for different -1 PRF efficiencies, p. A) For $\hat{I}_c$, we considered different values of π, the cost of degrading an mRNA after -1 PRF, namely $\pi = 0, 10, 100, 1000, 10000$. For $\hat{I}_c$, correlation values were computed at -1 PRF efficiencies $p = 0.01, 0.05, 0.1, 0.3, 0.5, 0.9$. I_a was computed at the same levels except for $p = 0.9$, where instead $p = 1$ was used. B) Analogous to A), but with p dependent on protein expression. For each gene, the -1 PRF efficiency is computed by $p(P) = p_b/\log_{10}(P)$, where for p_b, the *baseline -1 PRF efficiency*, we chose the same values as for p in A).

-1 PRF conceals stronger associations than measured with uncorrected codon usage bias indices.

DISCUSSION

In this study, we have investigated the hypothesis that -1 PRF presupposes translational costs to an organism, while at the same time generating benefits associated with protein expression regulation. We explored whether such a cost might be identified more directly in data and used both PRFdb (Belew *et al.* 2008) and the von der Haar data set (von der Haar 2008) to address this question. We devised three hypotheses for likely signals of such cost in -1 PRF carrying genes. We could not find contradictory evidence for any of those hypotheses in the data. In a second step, we explored whether these costs, if not accounted for, bias important measures of CUB. We devised two new general approaches to correct CUB indices for the presence of -1 PRF signals. We then tested whether the concealment of such costs may unduly influence, falsify or strengthen, one classical test of the translational efficiency hypothesis: the association of CUB indices with protein expression levels. Under the assumption of uniform -1 PRF efficiencies, the energetic costs related with NMD activation would need to be implausibly high to warrant this conclusion. We find that on the contrary, assuming that -1 PRF efficiency decreases with protein expression levels -as suggested by the existence of a cost to -1 PRF-, the TEH is strengthened. Thus, taken together, our results suggests that high organismal demands for specific proteins are reflected in CUB-mediated translation efficiency gains.

Our study comes with a series of caveats. A first caveat of this study lies in that the slippery sites are inferred (utilizing the *Nupack* algorithm (Dirks and Pierce 2004) for mRNA pseudoknot detection), but not always experimentally confirmed (Advani *et al.* 2013). Previous studies suggest that inferred slippery sites are likely to be functional. In (Advani *et al.* 2013) it is reported that in previous work, nine out of nine high confidence -1 PRF sites detected by methods used in the PRFdb in *S. cerevisae* were confirmed to be functional *in vivo*. Crucially, whether a -1 PRF site is regarded to be *functional* depends on whether

it exceeds a predetermined *p*-threshold. For example, in (Advani *et al.* 2013), a search of -1 PRF slippery sites identified 10 candidate genes in *EST2*, three in *EST1*, 2 in *STN*, and 1 in *CDC* in mRNA involved in telomerase. Out of these, and employing a cutoff of 1% for -1 PRF efficiency, seven carry functional slippery sites (*EST2*: 3, *EST1*: 2, *STN*: 1, *CDC*: 1) (see (Advani *et al.* 2013), Table 1). Had a cutoff of 0% -1 PRF efficiency been employed, all slippery sites except one would be functional. This limited sample provides confidence that the methods described by (Belew *et al.* 2008) appropriately capture biological mechanisms.

Even with such uncertainty, it is unlikely that the presence of non-functional -1 PRF sites in the analyzed data would affect the claim of an intrinsic -1 PRF maintenance cost (Figure 2). This claim would only be biased if the probability of a candidate site to be functional were affected by either protein expression levels of the gene within which it resides or alternatively, the relative position in the gene of the putative slippery site. Note that these effects should arise from mechanisms that are independent of the ones studied here. Again, we are unaware of any such mechanisms acting in yeast, except for the requirement of a spacer sequence. A spacer sequence of a minimum length of say, 8 nt, will prohibit slippery sites to be located within 8 nt of the mRNA end. This restriction will slightly bias the *a priori* position of candidate slippery sites. However, as average mRNA by far exceed this length, such a restriction cannot explain the effects documented here.

Another caveat of our study was the lack of estimates for *p* for the whole data set. Due to this restriction, we could not explore how -1 PRF efficiency *p* relates to CAI or protein expression. As assumed in Figure 5B, we expect that -1 PRF maintenance costs should on average translate into an inverse relationship between *p* and protein expression levels, because higher protein expression demands by the organism should be countered with reductions of *p*, minimizing the loss of mRNA to NMD-induced degradation. However, it is also possible that there exist cases in which -1 PRF has important functions in highly expressed genes. Hence, although there are fewer slippery sites for highly expressed genes and they are costly in terms of translation efficiencies, these sites could have large -1 PRF efficiencies. We have

not found evidence in the literature to support this notion, and it would concomitantly contradict the evidence here presented. Additional testing of our hypotheses on the costs -1 PRF maintenance would be greatly helped if information on how -1 PRF efficiencies vary with protein expression was available.

While our analysis of the behavior of $\hat{I}_c$ gives us an indication how -1 PRF costs to translation efficiency could bias CUB indices, these insights rely on the assumptions made during $\hat{I}_c$'s derivation. Importantly, $\hat{I}_c$ is a lower bound to I_c, which means that downward corrections of I_c are potentially exaggerated. Thus, the claim that corrected CUB indices will leave the basis for the TEH unaffected is resistant to such a bias. Another key assumption is that CUB indices of an mRNA sequence should increase monotonically with translation efficiency. In practice, this association is surely not perfect, due to multiple additional influences on CUB from unrelated biological processes. For these processes to systematically bias our results they should i) dominate over translational efficiency effects or ii) result in directional effects when combined. Despite these possible shortcomings, the assumption reflects key properties with which many CUB indices are designed, namely to mirror translational efficiency. For example, the CAI uses ribosomal mRNA as a reference to compute preferred codons because it is highly plausible that they are translationally efficient.

Moreover, our results suggest that utilizing expected values for corrected CUB index definition, such as in the case of I_a, is suboptimal. Analysis of the distribution of I_a levels with $p = 1$ of genes in the von der Haar data set show that corrected codon adaptation can increase. This behavior contradicts the rationale behind introducing such corrections in the first place, as, surprisingly, $I(x') > I(x)$ can occur.

To test the hypotheses about the TEH derived here, we have only utilized genes with identified slippery sites, and not all genes. Testing on all genes would be indicated if the TEH was challenged. However, if the correlation coefficient measured within a subset in which a correction has been applied is not substantially diminished relative to uncorrected values, the same correction will not affect the whole set either. Therefore testing it on the whole set is not necessary.

The results indicating intrinsic -1 PRF maintenance cost and TEH support are in mutual agreement. -1 PRF maintenance induces a reduction of the size of l with increasing protein expression levels (Figure 2C). At fixed p, as assumed in our analysis, significant reductions in l would diminish the penalization to the original, un-corrected CUB. This could only be compensated by increases of p or π with protein expression levels, contrary to intuition and available evidence (Advani *et al.* 2013). In fact, our analysis shows that a more plausible p dependency to protein expression leads to a strengthening of the basic CUB index to expression level association.

These results have to be interpreted in the context of current codon usage bias research. The mechanism of *translational selection* (TEH) remains the main explanation put forward for selection based origins of codon bias. This explanation presupposes that silent mutations affect fitness via translation processes. These fitness increases originate from translation efficiency and accuracy gains, that in turn are hypothesized to stem from translation initiation and elongation processes. In fact, mRNA elongation rates do indeed appear to correlate positively with preferred codon frequencies (Gustafsson *et al.* 2004), although better evidence would be desirable. However, for initiation processes –which are assumed to contribute the bulk of these fitness increases–, the effects of CUB on initiation rate increase remain subject of debate (Quax *et al.* 2015). In fact, Tuller and Zur have analyzed the effect of the structure of the 5′ end of an mRNA on translation initiation and elongation rates and found various regulatory signals that affect these rates in different ways (Tuller and Zur 2015).

Indeed, the induced folding at the 5′ end of the ORF appears to affect translation efficiency.

Overall, the evidence from codon usage bias statistics and its associations to tRNA abundance and protein expression offers a compelling narrative for the TEH. However, current research efforts aiming to identify the exact mechanisms that give rise to these associations must account for conflating selective forces. Indeed, the mechanisms laid out in the TEH may not be the only way in which selection shapes codon usage frequencies. For example, besides the abundance of tRNAs, other factors have been discovered to crucially affect elongation rates and hence, to be possible targets selection (Chamary *et al.* 2006; Resch *et al.* 2007, 2009; Hunt *et al.* 2009). More precisely, specific synonymous changes can influence mRNA splicing, mRNA secondary structure, protein stability as well as protein folding (Komar 2009; Chamary *et al.* 2006; Carlini and Genut 2006; Parmley *et al.* 2006; Tsai *et al.* 2008; Komar 2007). Synonymous changes may also alter the *secondary structure of mRNA* and thus affect the rate of translation –as established *in vitro* (Ivanov *et al.* 1997; Parmley and Hurst 2007; Tuller and Zur 2015) and subsequently *in vivo* (Kimchi-Sarfaty *et al.* 2007). A comprehensive review on how protein expression is fine tuned by codon usage bias is given by Quax (Quax *et al.* 2015). Further, in a very recent study, a synonymous difference between mammalian cytoskeletal β- and γ-actin proteins was found to affect co-translational processing, ubiquitination, and co-translational degradation, leading to differential stability properties of the corresponding protein products (Zhang *et al.* 2010).

Within genomes, the biologial role of a gene influences that gene's CUB in various additional ways. Some codons are more abundant in genes depending on their function, displaying distinct codon bias patterns. Supek has reviewed how gene function modifies codon preferences (Supek 2016). Selective pressure to maintain (or alter) gene function is superimposed to what is expected from translational efficiency and accuracy optimization. Genes where altered CUB patterns have been found are involved in diverse functions: amino acid starvation responses, cyclical protein expression, tissue specific expression, cellular differentiation, stress responses, and carcinogenesis. While genes can differ in function, there are also differences in function within a gene's sequence that also affect local CUB. For example, Tuller and Zur have surveyed the multiple roles of the 5′ end of coding sequences in gene expression regulation (Tuller and Zur 2015). They hypothesize that due to multitude of regulatory signals found in that region, selection pressures regarding codon utilization are likely different than in other regions. Unlike cross-gene function, such effects are local, which should not affect our analysis. How effects from gene function may affect our results, crucially depends on their frequency and direction. The literature reviewed here does not appear to warrant the assumption that all the function-dependent selective pressures will align to influence CUB in the same way, leading to systematic bias. The phenomenon of -1 PRF is thus only one of many ways in which CUB may be influenced.

The particular appeal of the -1 PRF phenomenon lies in its potential to elucidate many of these processes. Because slippery sites are precisely localized and -1 PRF efficiencies measurable, -1 PRF signals constitute natural experiments to translation efficiency and accuracy theories. This is because the separation of the mRNA by a slippery site should create differential translational costs across that mRNA. This translational cost gradient should be reflected in codon usage bias differences. -1 PRF based approaches to analyzing codon usage bias behavior, like the one presented in this study, may thus offer novel tools to better understand the means by which translation efficiency gains are realized in nature.

LITERATURE CITED

Advani, V. M., A. T. Belew, and J. D. Dinman, 2013 Yeast telomere maintenance is globally controlled by programmed ribosomal frameshifting and the nonsense-mediated mrna decay pathway. Translation (Austin) 1: e24418. https://doi.org/10.4161/trla.24418

Advani, V. M., and J. D. Dinman, 2016 Reprogramming the genetic code: The emerging role of ribosomal frameshifting in regulating cellular gene expression. BioEssays 38: 21–26. https://doi.org/10.1002/bies.201500131

Akashi, H., 1994 Synonymous codon usage in drosophila melanogaster: natural selection and translational accuracy. Genetics 136: 927–935.

Arava, Y., Y. Wang, J. D. Storey, C. L. Liu, P. O. Brown et al., 2003 Genome-wide analysis of mrna translation profiles in saccharomyces cerevisiae. Proc. Natl. Acad. Sci. USA 100: 3889–3894. https://doi.org/10.1073/pnas.0635171100

Baranov, P. V., J. F. Atkins, and M. M. Yordanova, 2015 Augmented genetic decoding: global, local and temporal alterations of decoding processes and codon meaning. Nat. Rev. Genet. 16: 517–529. https://doi.org/10.1038/nrg3963

Belew, A. T., V. M. Advani, and J. D. Dinman, 2011 Endogenous ribosomal frameshift signals operate as mrna destabilizing elements through at least two molecular pathways in yeast. Nucleic Acids Res. 39: 2799–2808. https://doi.org/10.1093/nar/gkq1220

Belew, A. T., and J. D. Dinman, 2015 Cell cycle control (and more) by programmed -1 ribosomal frameshifting: implications for disease and therapeutics. Cell Cycle 14: 172–178. https://doi.org/10.4161/15384101.2014.989123

Belew, A. T., N. L. Hepler, J. L. Jacobs, and J. D. Dinman, 2008 Prfdb: a database of computationally predicted eukaryotic programmed -1 ribosomal frameshift signals. BMC Genomics 9: 339. https://doi.org/10.1186/1471-2164-9-339

Belew, A. T., A. Meskauskas, S. Musalgaonkar, V. M. Advani, S. O. Sulima et al., 2014 Ribosomal frameshifting in the ccr5 mrna is regulated by mirnas and the nmd pathway. Nature 512: 265–269. https://doi.org/10.1038/nature13429

Brierley, I., and F. J. Dos Ramos, 2006 Programmed ribosomal frameshifting in hiv-1 and the sars-cov. Virus Res. 119: 29–42. https://doi.org/10.1016/j.virusres.2005.10.008

Caliskan, N., V. I. Katunin, R. Belardinelli, F. Peske, and M. V. Rodnina, 2014 Programmed -1 frameshifting by kinetic partitioning during impeded translocation. Cell 157: 1619–1631. https://doi.org/10.1016/j.cell.2014.04.041

Caliskan, N., F. Peske, and M. V. Rodnina, 2015 Changed in translation: mrna recoding by -1 programmed ribosomal frameshifting. Trends Biochem. Sci. 40: 265–274. https://doi.org/10.1016/j.tibs.2015.03.006

Carbone, A., A. Zinovyev, and F. Képès, 2003 Codon adaptation index as a measure of dominating codon bias. Bioinformatics 19: 2005–2015. https://doi.org/10.1093/bioinformatics/btg272

Carlini, D. B., 2004 Experimental reduction of codon bias in the drosophila alcohol dehydrogenase gene results in decreased ethanol tolerance of adult flies. J. Evol. Biol. 17: 779–785. https://doi.org/10.1111/j.1420-9101.2004.00725.x

Carlini, D. B., and J. E. Genut, 2006 Synonymous snps provide evidence for selective constraint on human exonic splicing enhancers. J. Mol. Evol. 62: 89–98. https://doi.org/10.1007/s00239-005-0055-x

Carlini, D. B., and W. Stephan, 2003 In vivo introduction of unpreferred synonymous codons into the drosophila adh gene results in reduced levels of adh protein. Genetics 163: 239–243.

Chamary, J. V., J. L. Parmley, and L. D. Hurst, 2006 Hearing silence: non-neutral evolution at synonymous sites in mammals. Nat. Rev. Genet. 7: 98–108. https://doi.org/10.1038/nrg1770

Chang, Y.-F., J. S. Imam, and M. F. Wilkinson, 2007 The nonsense-mediated decay rna surveillance pathway. Annu. Rev. Biochem. 76: 51–74. https://doi.org/10.1146/annurev.biochem.76.050106.093909

Clare, J. J., M. Belcourt, and P. J. Farabaugh, 1988 Efficient translational frameshifting occurs within a conserved sequence of the overlap between the two genes of a yeast ty1 transposon. Proc. Natl. Acad. Sci. USA 85: 6816–6820. https://doi.org/10.1073/pnas.85.18.6816

Cobucci-Ponzano, B., M. Rossi, and M. Moracci, 2012 Translational recoding in archaea. Extremophiles 16: 793–803. https://doi.org/10.1007/s00792-012-0482-8

Craigen, W. J., R. G. Cook, W. P. Tate, and C. T. Caskey, 1985 Bacterial peptide chain release factors: conserved primary structure and possible frameshift regulation of release factor 2. Proc. Natl. Acad. Sci. USA 82: 3616–3620. https://doi.org/10.1073/pnas.82.11.3616

Dinman, J. D., 2006 Programmed ribosomal frameshifting goes beyond viruses: Organisms from all three kingdoms use frameshifting to regulate gene expression, perhaps signaling a paradigm shift. Microbe Wash. D.C. 1: 521–527.

Dinman, J. D., 2012a Control of gene expression by translational recoding. Adv. Protein Chem. Struct. Biol. 86: 129–149. https://doi.org/10.1016/B978-0-12-386497-0.00004-9

Dinman, J. D., 2012b Mechanisms and implications of programmed translational frameshifting. Wiley Interdiscip. Rev. RNA 3: 661–673. https://doi.org/10.1002/wrna.1126

Dirks, R. M., and N. A. Pierce, 2004 An algorithm for computing nucleic acid base-pairing probabilities including pseudoknots. J. Comput. Chem. 25: 1295–1304. https://doi.org/10.1002/jcc.20057

Dulude, D., Y. A. Berchiche, K. Gendron, L. Brakier-Gingras, and N. Heveker, 2006 Decreasing the frameshift efficiency translates into an equivalent reduction of the replication of the human immunodeficiency virus type 1. Virology 345: 127–136. https://doi.org/10.1016/j.virol.2005.08.048

Duret, L., 2002 Evolution of synonymous codon usage in metazoans. Curr. Opin. Genet. Dev. 12: 640–649. https://doi.org/10.1016/S0959-437X(02)00353-2

Friberg, M., P. von Rohr, and G. Gonnet, 2004 Limitations of codon adaptation index and other coding dna-based features for prediction of protein expression in saccharomyces cerevisiae. Yeast 21: 1083–1093. https://doi.org/10.1002/yea.1150

Ghaemmaghami, S., W.-K. Huh, K. Bower, R. W. Howson, A. Belle et al., 2003 Global analysis of protein expression in yeast. Nature 425: 737–741. https://doi.org/10.1038/nature02046

Gustafsson, C., S. Govindarajan, and J. Minshull, 2004 Codon bias and heterologous protein expression. Trends Biotechnol. 22: 346–353. https://doi.org/10.1016/j.tibtech.2004.04.006

Hershberg, R., and D. A. Petrov, 2008 Selection on codon bias. Annu. Rev. Genet. 42: 287–299. https://doi.org/10.1146/annurev.genet.42.110807.091442

Hershberg, R., and D. A. Petrov, 2009 General rules for optimal codon choice. PLoS Genet. 5: e1000556. https://doi.org/10.1371/journal.pgen.1000556

Holland, M. J., 2002 Transcript abundance in yeast varies over six orders of magnitude. J. Biol. Chem. 277: 14363–14366. https://doi.org/10.1074/jbc.C200101200

Holstege, F. C., E. G. Jennings, J. J. Wyrick, T. I. Lee, C. J. Hengartner et al., 1998 Dissecting the regulatory circuitry of a eukaryotic genome. Cell 95: 717–728. https://doi.org/10.1016/S0092-8674(00)81641-4

Hunt, R., Z. E. Sauna, S. V. Ambudkar, M. M. Gottesman, and C. Kimchi-Sarfaty, 2009 Silent (synonymous) snps: should we care about them? Methods Mol. Biol. 578: 23–39. https://doi.org/10.1007/978-1-60327-411-1_2

Ikemura, T., 1981 Correlation between the abundance of escherichia coli transfer rnas and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the e. coli translational system. J. Mol. Biol. 151: 389–409. https://doi.org/10.1016/0022-2836(81)90003-6

Ikemura, T., 1982 Correlation between the abundance of yeast transfer rnas and the occurrence of the respective codons in protein genes.

differences in synonymous codon choice patterns of yeast and escherichia coli with reference to the abundance of isoaccepting transfer rnas. J. Mol. Biol. 158: 573–597. https://doi.org/10.1016/0022-2836(82)90250-9

Ikemura, T., 1985   Codon usage and trna content in unicellular and multicellular organisms. Mol. Biol. Evol. 2: 13–34.

Ivanov, I. G., A. A. Saraffova, and M. G. Abouhaidar, 1997   Unusual effect of clusters of rare arginine (agg) codons on the expression of human interferon alpha 1 gene in escherichia coli. Int. J. Biochem. Cell Biol. 29: 659–666. https://doi.org/10.1016/S1357-2725(96)00161-6

Jacks, T., and H. E. Varmus, 1985   Expression of the rous sarcoma virus pol gene by ribosomal frameshifting. Science 230: 1237–1242. https://doi.org/10.1126/science.2416054

Jansen, R., H. J. Bussemaker, and M. Gerstein, 2003   Revisiting the codon adaptation index from a whole-genome perspective: analyzing the relationship between gene expression and codon occurrence in yeast using a variety of models. Nucleic Acids Res. 31: 2242–2251. https://doi.org/10.1093/nar/gkg306

Jelinsky, S. A., and L. D. Samson, 1999   Global response of saccharomyces cerevisiae to an alkylating agent. Proc. Natl. Acad. Sci. USA 96: 1486–1491. https://doi.org/10.1073/pnas.96.4.1486

Kanaya, S., Y. Yamada, M. Kinouchi, Y. Kudo, and T. Ikemura, 2001   Codon usage and trna genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with cg-dinucleotide usage as assessed by multivariate analysis. J. Mol. Evol. 53: 290–298. https://doi.org/10.1007/s002390010219

Kanaya, S., Y. Yamada, Y. Kudo, and T. Ikemura, 1999   Studies of codon usage and trna genes of 18 unicellular organisms and quantification of bacillus subtilis trnas: gene expression level and species-specific diversity of codon usage based on multivariate analysis. Gene 238: 143–155. https://doi.org/10.1016/S0378-1119(99)00225-5

Ketteler, R., 2012   On programmed ribosomal frameshifting: the alternative proteomes. Front. Genet. 3: 242. https://doi.org/10.3389/fgene.2012.00242

Kimchi-Safarty, C., J. M. Oh, I.-W. Kim, Z. E. Sauna, A. M. Calcagno et al., 2007   A "silent" polymorphism in the mdr1 gene changes substrate specificity. Science 315: 525–528. https://doi.org/10.1126/science.1135308

Kliman, R. M., N. Irving, and M. Santiago, 2003   Selection conflicts, gene expression, and codon usage trends in yeast. J. Mol. Evol. 57: 98–109. https://doi.org/10.1007/s00239-003-2459-9

Komar, A. A., 2007   Silent snps: impact on gene function and phenotype. Pharmacogenomics 8: 1075–1080. https://doi.org/10.2217/14622416.8.8.1075

Komar, A. A., 2009   A pause for thought along the co-translational folding pathway. Trends Biochem. Sci. 34: 16–24. https://doi.org/10.1016/j.tibs.2008.10.002

Lu, P., C. Vogel, R. Wang, X. Yao, and E. M. Marcotte, 2007   Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. Nat. Biotechnol. 25: 117–124. https://doi.org/10.1038/nbt1270

Maia, I. G., K. Séron, A. L. Haenni, and F. Bernardi, 1996   Gene expression from viral rna genomes. Plant Mol. Biol. 32: 367–391. https://doi.org/10.1007/BF00039391

Nakao, A., M. Yoshihama, and N. Kenmochi, 2004   Rpg: the ribosomal protein gene database. Nucleic Acids Res. 32: D168–D170. https://doi.org/10.1093/nar/gkh004

Newman, J. R. S., S. Ghaemmaghami, J. Ihmels, D. K. Breslow, M. Noble et al., 2006   Single-cell proteomic analysis of s. cerevisiae reveals the architecture of biological noise. Nature 441: 840–846. https://doi.org/10.1038/nature04785

Parmley, J. L., J. V. Chamary, and L. D. Hurst, 2006   Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. Mol. Biol. Evol. 23: 301–309. https://doi.org/10.1093/molbev/msj035

Parmley, J. L., and L. D. Hurst, 2007   How do synonymous mutations affect fitness? BioEssays 29: 515–519. https://doi.org/10.1002/bies.20592

Plant, E. P., and J. D. Dinman, 2006   Comparative study of the effects of heptameric slippery site composition on -1 frameshifting among different eukaryotic systems. RNA 12: 666–673. https://doi.org/10.1261/rna.2225206

Plant, E. P., P. Wang, J. L. Jacobs, and J. D. Dinman, 2004   A programmed -1 ribosomal frameshift signal can function as a cis-acting mrna destabilizing element. Nucleic Acids Res. 32: 784–790. https://doi.org/10.1093/nar/gkh256

Plotkin, J. B., and G. Kudla, 2011   Synonymous but not the same: the causes and consequences of codon bias. Nat. Rev. Genet. 12: 32–42. https://doi.org/10.1038/nrg2899

Quax, T. E. F., N. J. Claassens, D. Söll, and J. van der Oost, 2015   Codon bias as a means to fine-tune gene expression. Mol. Cell 59: 149–161. https://doi.org/10.1016/j.molcel.2015.05.035

Resch, A. M., L. Carmel, L. Mariño-Ramírez, A. Y. Ogurtsov, S. A. Shabalina et al., 2007   Widespread positive selection in synonymous sites of mammalian genes. Mol. Biol. Evol. 24: 1821–1831. https://doi.org/10.1093/molbev/msm100

Resch, A. M., A. Y. Ogurtsov, I. B. Rogozin, S. A. Shabalina, and E. V. Koonin, 2009   Evolution of alternative and constitutive regions of mammalian 5′utrs. BMC Genomics 10: 162. https://doi.org/10.1186/1471-2164-10-162

Roth, A., M. Anisimova, and G. M. Cannarozzi, 2012, pp. 189–217 in Measuring codon usage bias. Codon evolution: mechanisms and models, Oxford University Press Inc, New York. https://doi.org/10.1093/acprof:osobl/9780199601165.003.0013

Roth, F. P., J. D. Hughes, P. W. Estep, and G. M. Church, 1998   Finding dna regulatory motifs within unaligned noncoding sequences clustered by whole-genome mrna quantitation. Nat. Biotechnol. 16: 939–945. https://doi.org/10.1038/nbt1098-939

Sharp, P. M., and W. H. Li, 1986   An evolutionary perspective on synonymous codon usage in unicellular organisms. J. Mol. Evol. 24: 28–38. https://doi.org/10.1007/BF02099948

Sharp, P. M., and W. H. Li, 1987   The codon adaptation index–a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. 15: 1281–1295. https://doi.org/10.1093/nar/15.3.1281

Stoletzki, N., and A. Eyre-Walker, 2007   Synonymous codon usage in escherichia coli: selection for translational accuracy. Mol. Biol. Evol. 24: 374–381. https://doi.org/10.1093/molbev/msl166

Supek, F., 2016   The code of silence: Widespread associations between synonymous codon biases and gene function. J. Mol. Evol. 82: 65–73. https://doi.org/10.1007/s00239-015-9714-8

Tsai, C.-J., Z. E. Sauna, C. Kimchi-Safarty, S. V. Ambudkar, M. M. Gottesman et al., 2008   Synonymous mutations and ribosome stalling can lead to altered folding pathways and distinct minima. J. Mol. Biol. 383: 281–291. https://doi.org/10.1016/j.jmb.2008.08.012

Tuller, T., Y. Y. Waldman, M. Kupiec, and E. Ruppin, 2010   Translation efficiency is determined by both codon bias and folding energy. Proc. Natl. Acad. Sci. USA 107: 3645–3650. https://doi.org/10.1073/pnas.0909910107

Tuller, T., and H. Zur, 2015   Multiple roles of the coding sequence 5′ end in gene expression regulation. Nucleic Acids Res. 43: 13–28. https://doi.org/10.1093/nar/gku1313

Varenne, S., J. Buc, R. Lloubes, and C. Lazdunski, 1984   Translation is a non-uniform process. effect of trna availability on the rate of elongation of nascent polypeptide chains. J. Mol. Biol. 180: 549–576. https://doi.org/10.1016/0022-2836(84)90027-5

von der Haar, T., 2008   A quantitative estimation of the global translational activity in logarithmically growing yeast cells. BMC Syst. Biol. 2: 87. https://doi.org/10.1186/1752-0509-2-87

Yamao, F., Y. Andachi, A. Muto, T. Ikemura, and S. Osawa, 1991   Levels of trnas in bacterial cells as affected by amino acid usage in proteins. Nucleic Acids Res. 19: 6119–6122. https://doi.org/10.1093/nar/19.22.6119

Zhang, F., S. Saha, S. A. Shabalina, and A. Kashina, 2010   Differential arginylation of actin isoforms is regulated by coding sequence-dependent degradation. Science 329: 1534–1537. https://doi.org/10.1126/science.1191701

Communicating editor: R. Kulathinal