Evolutionary conservation of codon optimality reveals hidden signatures of co-translational folding

Sebastian Pechmann[1,2] and **Judith Frydman**[1,2,3]

[1]Department of Biology, Stanford University, Stanford, CA 94305-5020, USA

[2]BioX Program, Stanford University, Stanford, CA 94305-5020, USA

Abstract

The choice of codons can influence local translation kinetics during protein synthesis. Whether codon preference is linked to co-translational regulation of polypeptide folding remains unclear. Here, we derive a revised translational efficiency scale that incorporates the competition between tRNA supply and demand. Applying this scale to ten closely related yeasts, we uncover the evolutionary conservation of codon optimality in eukaryotes. This analysis reveals universal patterns of conserved optimal and nonoptimal codons, often in clusters, which associate with the secondary structure of the translated polypeptides independent of the levels of expression. Our analysis suggests an evolved function for codon optimality in regulating the rhythm of elongation to facilitate co-translational polypeptide folding, beyond its previously proposed role of adapting to the cost of expression. These findings establish how mRNA sequences are generally under selection to optimize the co-translational folding of corresponding polypeptides.

Introduction

The translational efficiency of individual codons directly modulates the kinetics of protein synthesis[1]. Optimal codons are thought to be translated both faster and more accurately[2, 3]. In turn, nonoptimal codons can slow down protein synthesis. Due to the degeneracy of the genetic code, all amino acids except methionine and tryptophan are encoded by both optimal and nonoptimal codons. The evolutionary forces that shape the codon bias, i.e. the unequal usage of synonymous codons, are the focus of intense study. In particular, the pressure to maintain nonoptimal codons is unclear. One attractive hypothesis is that nonoptimal codons slow translation for biologically relevant functions, such as facilitating co-translational folding by allowing the nascent chain more time to develop native-like structure[4]. Indeed, a link between the mRNA sequences and *in vivo* folding of the encoded proteins has long been suggested[5]. For instance, synonymous substitutions reducing translational efficiency have been found to alter folding[6, 7], and subsequent function[8] of the translated polypeptides. This supports the suggestion that protein synthesis is directly attuned to the co-translational

[3]Corresponding author: jfrydman@stanford.edu.

folding[9–13]. However, no universal correlation across organisms was found between the position of nonoptimal codons and the location of structural units[14] or domain boundaries[15], or between synonymous codon usage and protein secondary structural elements[16]. For *S. cerevisiae*, a codon preference in relation to protein secondary structure could only be found for the amino acids glycine in loops and threonine in helices[15]. Thus, beyond a few individual examples, a global view linking codon optimality to nascent chain folding *in vivo* remains elusive.

The classification into optimal and nonoptimal codons reflects the important role of tRNAs on the rate of protein synthesis. During translation, specific tRNAs recognize the codons in the mRNA and deliver the corresponding amino acids to the ribosome[17] (Fig. 1a). The abundance of tRNAs varies markedly in the cellular pool and is strongly correlated to their gene copy numbers in the genome[18]. Since tRNA gene redundancy is a critical factor explaining the correlation between translational selection and codon bias[19], a scale of codon-specific translational efficiencies has been devised based on the relative abundance of tRNAs as well as selective constraints on codon-tRNA pairings for the "wobble" non-Watson-Crick interactions[19]. Codons that are over-represented in highly expressed genes are also recognized by the most abundant tRNAs, and are thus denoted as optimal[20].

While this "classical" optimality scale has been very useful to derive fundamental insights into the role of the prevalent codon bias, some of its simplifying assumptions may reduce its power. Because the classical view of codon optimality is based on a subset of the most highly expressed genes and not the full genome, it is biased towards optimal codons, as only optimal codons are explicitly defined[20]. In addition, the "classical" scale relies solely on tRNA abundances, without taking into consideration the competition for tRNAs among all translating ribosomes[1, 21]. However, translation rates depend on the balance between supply and demand for charged tRNAs, and a highly used codon may effectively deplete its cognate tRNAs by increasing demand. Indeed, kinetic modeling of translation elongation highlights the key role of this competition in determining translation rate of the corresponding codon[21]. Furthermore, the classical definition of translational efficiency incorporates species-specific tRNA pools and genomic sequences[19, 22], but does not directly include mRNA expression levels, thereby overlooking the effect of divergent gene expression observed even between closely related species[23]. Since mRNA expression levels will affect the demand for tRNAs, its explicit incorporation into a translational efficiency scale may allow for a better comparison between organisms.

To circumvent the above limitations of the classical scale, we developed a new translational efficiency scale incorporating the balance of tRNA supply and demand. The resulting scale reflects the competition for tRNAs in the cell. Importantly, by also including the expression profile of the genome as major characteristic of organism divergence[23], the new scale allows for a better evolutionary comparison of patterns of codon optimality across organisms. We find that codon optimality is evolutionarily conserved across ten closely related yeasts in a site-specific manner that is independent of levels of expression. Importantly, such evolutionary conservation of codon optimality reveals a direct and uniform link between codon optimality in the mRNAs and the secondary structures and folding elements of the encoded translated polypeptides.

Results

A normalized translational efficiency scale

The classical view of translational efficiency (cTE), which is the known tRNA adaptation index[19] (tAI), does not incorporate the cellular dynamics of tRNAs, driven by a trade-off between tRNA supply and demand[1, 21] (Fig. 1b). We hypothesized that a translational efficiency scale that reflects this competition for the cellular pool of tRNAs may better capture the biological forces shaping the codon bias. We thus normalized the cellular tRNA abundances and selective constraints on codon-tRNA interactions as defined in tAI[19] by the codon usage (*see Methods,* Supplementary Fig. 1a). How often a codon is translated in the cell depends on the codon frequencies in the mRNAs, the abundances of the mRNAs that are attached to ribosomes, and the densities of ribosomes on the specific mRNAs[24]. We verified that mRNA abundances alone serve as a sufficient and readily available proxy for the calculation of the codon usage (Supplementary Fig. 1b–d). In this normalized translational efficiency (nTE) scale, codons are considered optimal if the relative availability of cognate tRNAs exceeds their relative usage.

Comparing the nTE to the cTE scale (Fig. 1c–e, Supplementary Fig. 2a, b), the nTE scale has a more shallow plateau-like middle region with two distinct tails of high and low efficiency codons. This stands in contrast to the cTE that scale increases almost linearly (Fig. 1e). The tail of low efficiency codons and plateau-like middle region are unique to the nTE scale, i.e. the ratio of corresponding tRNA availability and codon usage (Supplementary Fig. 2c). Importantly, this analysis suggests that tRNA supply and demand are closely matched for most codons at steady-state expression, likely supporting cost-effective proteome maintenance. Also interesting, the nTE scale contains more optimal codons that encode hydrophobic amino acids, which resonates with observations that optimal codons are associated with structurally sensitive, buried sites[20]. The most efficient new optimal codons in the normalized scale encode glycine, the smallest thus most neutral amino acid, and arginine, cysteine, and proline, which are important gatekeeper residues that can prevent aberrant protein aggregation[25]. The higher fraction of optimal codons encoding gatekeeper and hydrophobic residues revealed by the nTE scale likely reflects the need for increased fidelity during translation of sensitive polypeptide segments critical to correct folding and avoidance of aggregation[20, 26]. In addition, the codons encoding the most abundant amino acid glutamine are the most balanced between supply and demand in the nTE scale. The only non-degenerate amino acids methionine and tryptophan, which are nonoptimal in the cTE scale, are optimal in the nTE scale. Importantly, the nTE scale does not, unlike the cTE scale, correlate with the codon bias; the nTE scale is also not directly correlated with the cTE scale (Fig. 1f). Thus, while the sets of optimal codons before and after normalization largely overlap, the nTE scale provides a new and independent metric of translational efficiency that reflects the cellular competition for tRNAs[1, 21].

A conserved translational efficiency "dip" at the start of mRNAs

Evolutionary conservation is a strong indication of functional importance. Thus, any uniform link between translational efficiency patterns in the mRNA sequences and the folding of the encoded polypeptides would expectedly be evolutionarily conserved. We

undertook a systematic analysis of the conservation of codon optimality across ten closely related yeasts (*see Methods*). For this, we computed the nTE scale for all ten yeasts; in all cases we observed the characteristic shape of the nTE scales as observed for *S. cerevisiae*. Previous work using the cTE scale revealed an evolutionarily conserved region of low translational efficiency at the beginning of coding sequences, termed the "ramp", spanning the first ca. 35–50 codons[18]. We thus next tested if the new nTE scale can also detect the "ramp". The average nTE profile of the *S. cerevisiae* genome indeed validated the initial region of low translational efficiency (Fig. 2a), but showed that it is only ca. 10 codons long, i.e. much shorter than found using the cTE scale (Fig. 2b). Importantly, this short "dip" was evolutionarily conserved in all other analyzed yeasts (Supplementary Fig. 3). While the longer "ramp" is not observed in translational efficiency profiles of individual genes using either the cTE or nTE scales (Fig. 2c–e), this characteristic "dip" can be observed in almost all individual translational efficiency profiles, both in highly expressed genes, e.g. hydroxylase *LIA1* or AAA ATPase *CDC48* (Fig. 2c, d), and in lowly expressed RNA polymerase II transcription factor gene *TFB3* (Fig. 2e). Interestingly, the length of the evolutionarily conserved "dip" corresponds to roughly the distance from the peptidyltransferase center to a constriction site within the ribosome[10, 11] (Fig. 2f). The constriction site consists of ribosomal proteins L4 and L17, which are thought to sense nascent chain conformations to signal to the outside of the ribosome[27, 28]. The finding of only a short dip of low translational efficiency also stands in good agreement with the experimental determination of ribosome densities along mRNAs in yeast, which is dominated by systematic ribosome pausing at the very start of the coding sequences[24]. Importantly, this analysis indicates that the combination of the nTE scale and evolutionary analysis can reveal sequence signatures important for translation.

Site-specific evolutionary conservation of codon optimality

A functional link between the positioning of optimal and nonoptimal codons along the mRNA and co-translational folding would predict that codon optimality is conserved in a site-specific manner, that relates to how nascent chains fold. To test for site-specific conservation of codon optimality we constructed sequence alignments across orthologs of ten closely related yeasts, and for each gene computed a conservation score of codon optimality for each position (*see Methods*, Fig. 3a, Supplementary Fig 4a, b). Randomizing the alignments yielded the distributions of random conservation scores, which allowed us to determine the alignment-specific minimal conservation scores found by less than 5% chance (i.e. significance thresholds with $p < 0.05$). Sites in the biological alignment that have higher conservation scores than these significance thresholds are thus considered significantly conserved. If the number of significantly conserved sites in an observed alignment exceeds the corresponding number in the randomized alignment, than these positions in the sequence alignment must be under direct selective pressure for site-specific conservation of codon optimality (Fig. 3b). Because expression alone can explain the fraction of optimal codons in mRNA sequences[29], we devised a randomization procedure that maintains the overall codon composition in each sequence. To avoid a persistent expression bias, all analyses were performed for two curated sets of 404 alignments of high and 302 alignments of low mRNA expression levels respectively (*see Methods,* Supplementary Fig. 4c). Importantly, we also employed independent randomization procedures that take into account the distribution of

optimal and nonoptimal codons for each amino acid in the genetic code. These latter additional analyses demonstrated that the conservation of codon optimality is completely independent of amino acid biases (Supplementary Fig. 5a–g). Furthermore, these results are independent of the 5′ coding regions (Supplementary Fig. 5h). Together, these independent analyses clearly indicate that there is site-specific evolutionary conservation of codon optimality regardless of amino acid bias or expression level.

We found that codon optimality is under selection in almost 80% of the low expression, and over 90% of the high expression genes (Fig. 3c). This is consistent with the observation that highly expressed proteins are generally more conserved[30]. The significance thresholds for optimal and nonoptimal codons in high and low expression genes distribute homogeneously for the nTE scale, and reflect the higher content of optimal codons in highly expressed genes as expected (Fig. 3d). Strikingly, both optimal and nonoptimal codons are found conserved in equal measure in both highly and lowly expressed genes. This challenges the view that mostly optimal codons are selected for in the context of translation. Instead, our findings indicate that codon optimality is not only tuned to expression but rather fulfills an evolutionarily selected function in protein biogenesis[6–8]. Of note, the same analysis using cTE to define codon optimality produces similar results for conservation of optimal codons (Fig. 3e, f). However, the cTE scale reduces the significance of conserved nonoptimal codons (Fig. 3e, f). While optimal codons are important for translational fidelity, nonoptimal codons play a key role in modulating the speed of elongation. For this reason, the nTE scale is better suited to assess the possible link between codon optimality and co-translational folding.

Conserved hidden signatures of co-translational folding

Having determined that optimal and nonoptimal codons are generally conserved raises the question regarding their role in co-translational folding events. We thus mapped the codon conservation profiles from the alignments described above onto the corresponding *S. cerevisiae* protein sequences and structures (e.g. Fig. 4a). Both conserved optimal and nonoptimal codons are generally distributed throughout the mRNA sequences, and often appear in clusters (Fig. 4a). We next tested for statistical associations between conserved optimal and nonoptimal codons and secondary structure elements of the encoded nascent chains (Fig. 4b, c). A control analysis using randomized synonymous codons confirmed the independence of our results from amino acid biases (Supplementary Fig. 6a, b).

We found distinct patterns of codon optimality conservation depending on the secondary structure of the encoded polypeptides, for both high and low expression proteins. Predicted α-helices are enriched in both conserved optimal and conserved nonoptimal codons, independent of expression (Fig 4b). In contrast, β-sheets are enriched in conserved optimal but depleted in conserved nonoptimal codons, in both highly and lowly expressed genes (Fig 4b). Of note, coil regions are always depleted of conserved optimal codons. Conserved nonoptimal codons are weakly enriched in highly expressed genes, and depleted in lowly expressed genes (Fig 4b). Interestingly, α-helices can already form co-translationally even within the ribosomal tunnel[10, 11], while coil regions, comprising loops that fold near the exit of the ribosomal tunnel, have been shown to play key roles in co-translational protein

folding[7]. In contrast, β-sheet containing domains are topologically discontinuous and must await synthesis to begin folding. Furthermore, β-sheets are characterized by their high content of hydrophobic residues, the presence of gatekeepers, and a high aggregation propensity, thus, the general strong enrichment of conserved optimal as well as depletion of conserved nonoptimal codons could primarily serve to reduce the risk of phenotypic missense mutations leading to aggregation.

Since hydrophobicity is linked to both protein folding and aggregation, we also tested for associations between conserved codon optimality and hydrophobicity. As expected, conserved optimal codons are enriched in hydrophobic regions, and conserved nonoptimal codons are depleted. This observation is stronger for highly expressed genes (Supplementary Fig. 6c), likely owing to a greater need for translational fidelity in these abundant proteins[3, 20, 26].

We next considered only sites that appear in clusters. This analysis provided a more stringent test for our results that validated and increased the significance of all the above associations between codon optimality and secondary structure propensity (Fig. 4c). In particular, the enrichment of conserved nonoptimal codons appearing in clusters is much stronger in α-helices at both high and low expression levels, and in coil regions of highly expressed genes.

To validate these observations with confirmed secondary structures, we analyzed 357 experimentally determined protein structures from the protein data bank (PDB). Importantly, PDB structures allow to distinguish between more defined structural elements such as hydrogen-bonded turns that often connect the more regular elements of secondary structure, α-helix and β-strand, and are important for folding into the final structure. This analysis indicated that conserved nonoptimal codons are strongly enriched in turns (Fig. 5a, Supplementary Fig. 6d, e). The enrichment of conserved nonoptimal codons only in α-helices and in turns is remarkable, since these secondary structure elements can form co-translationally[10, 11]. These findings are not detected when using the classical definition of codon optimality (Fig. 5b). All the above findings for conserved optimal codons are weaker using cTE and all significant associations to conserved nonoptimal codons are lost (Fig. 5b). This may explain why the link between codon optimality and co-translational folding has been hard to detect. Individual examples support our findings. For instance, the all-helical Myosin light chain 1 shows alternating patterns of conserved optimal and nonoptimal codons in all its helices (Fig. 5c). The 20S proteasome subunit G is characterized by a clear conservation of optimal codons in its structural core of β-sheets, complemented by conservation of both optimal and nonoptimal codons in its helices and turns (Fig. 5c).

What stands out from this analysis is the enrichment of both conserved optimal and nonoptimal codons in α-helices. The formation of α-helices is one of the elementary steps in protein folding, characterized by complex kinetics due low cooperativity that depend on an initial nucleation step[31]. Furthermore, this is the main folding event that has been found to occur inside the ribosome exit tunnel[11, 32, 33]; indeed helix formation has been experimentally observed to occur very early in the ribosomal tunnel, between the peptidyltransferase center and the constriction site[28]. Strikingly, for the α-helices in the

experimental protein structures of *S. cerevisiae*, we found a distinct alternating pattern of codon optimality. Specifically, we observed a preference for a nonoptimal codon at the transition into the helix, followed by an enrichment of optimal codons at position 1 and 4, interspersed with a strong preference for nonoptimal codons at positions 2 and 3 (Fig. 5d). Moreover, the significance of this profile only extends across the first full helix turn, independent of helix length (Fig. 5d). This suggests that codon optimality may be evolutionarily selected to tune the translation and folding rates of helices early in their entry into the ribosomal tunnel. The ribosomal tunnel has been found to possess distinct folding zones that may facilitate helix formation deep inside the ribosome[33], with the strongest compaction into secondary structure observed proximal to the peptidyltransferase center[33]. Helix formation inside the ribosomal tunnel appears strongly sequence dependent, but cannot be explained by sequence hydrophobicity or helical propensity alone[33]. The specialized environment within the tunnel is proposed to create a rugged solvation landscape that may slow down folding[34]. We speculate that the evolutionarily conserved patterns of codon optimality in helices could facilitate specific interactions with the exit tunnel wall, and may even assist helix nucleation inside the exit tunnel.

Discussion

We propose a new translational efficiency scale that incorporates the cellular competition for tRNAs into the definition of codon optimality. Codon specific translational efficiencies have long been found to correspond to a distinct codon bias, and have been suggested to play a role in the co-translational folding of the encoded polypeptides. However, a coherent and uniform link has so far remained elusive, and may be difficult to detect, in part likely due to generally weak selection on synonymous substitutions and complex but robust polypeptide folding patterns. Our analysis provides conceptual advances on two fundamental aspects of this problem. First, our normalized translational efficiency scale (nTE) incorporates the biologically relevant competition for tRNAs among all ribosomes, which is known to influence the kinetics of translation elongation[1, 21]. As a result, nTE allows for a better comparison between organisms. This is important, as a general and systematic link between codon optimality and co-translational folding, would have to be evolutionarily conserved. Second, we uncover a uniform relationship between the evolutionary conservation of codon optimality and the folding patterns of the nascent polypeptides, further confirming our overall approach.

Our new nTE scale reveals an evolutionarily conserved relationship between preferences in optimal and nonoptimal codons in the mRNA sequences and the secondary structures of the corresponding translated polypeptides. In light of the established correlation between the levels of expression and the fraction of optimal codons, it is even more remarkable that the site-specific evolutionary conservation of codon optimality is observed independent of expression levels. This suggests a functional role in setting a rhythm of translation elongation that correlates with the folding elements in the nascent polypeptides. Strikingly, we found that conserved nonoptimal codons are only enriched in α-helices and hydrogen-bonded turns. Helices comprise the structural elements that have been observed to fold already deep inside the ribosomal exit tunnel[11, 28, 32, 33]. The sensing of helical conformations at the constriction site near the peptidyltransferase center has been shown to

critically influence ribosome conformation and signaling[28], and further physiological roles and detailed mechanisms await to be uncovered. Since hydrogen-bonded turns and loops connect more defined folding elements within the emerging polypeptide, their enrichment in nonoptimal codons may reflect their role in coordinating co-translational folding outside the ribosome. For instance, exemplary experimental work has demonstrated the importance of nonoptimal codons in loops for the successful co-translational folding of protein domains[7].

Our results point to a complex trade-off in the selection of optimal and nonoptimal codons to balance the need to allow time for successful protein folding while avoiding aberrant aggregation (Fig. 6). We find optimal codons predominantly at sites where translational fidelity is important to prevent aggregation, namely in β-sheets and for gatekeeper residues. This complements the found preference for optimal codons at structurally sensitive and aggregation-prone sites[20, 26]. Nonoptimal codons, often in clusters, can slow translation elongation and thus coordinate co-translational folding.

While mRNA secondary structure may add an additional layer of translational regulation[35], we found that its evolutionary conservation is much lower (Supplementary Fig. 6f–h). This suggests a weaker role in orchestrating the timing of elongation, consistent with the fact that the ribosome itself acts as helicase, unraveling the translated mRNAs[36].

One interesting aspect of the new nTE scale is that tRNA supply and demand are very balanced for most codons at steady state. Remarkably, almost all amino acids are encoded by equal numbers of optimal and non-optimal codons in nTE. As a result, the strong and consistent link between conserved codon optimality and protein secondary structure we observe is independent of amino acid biases and must thus be the direct result of site-specific selection on codon optimality. Evolutionary selection appears to exploit and amplify very subtle effects, as highlighted by the remarkable fact that we consistently find clusters of conserved optimal and nonoptimal codons. Our definition of nTE assumes steady state conditions, and no limitation in amino acid supply. It is interesting to speculate that tRNA recycling, dynamics and modifications[37, 38] may further influence the rhythm of translation elongation, both at steady state and in response to cellular stresses such as amino acid starvation[39].

In summary, we found uniform and evolutionarily conserved signatures in the mRNA sequences that link to folding patterns of the encoded polypeptides. The ribosome emerges in this analysis as a very active folding environment and the choice of the coding sequence emerges as finely tuned to the action of the ribosome. Our findings present a promising avenue to increase our understanding of *in vivo* protein folding, still a fundamental and poorly understood problem in biology.

Online Methods

Data sources

Genomic sequences and ortholog assignments for *S. cerevisiae*, *C. glabrata*, *D. hansenii*, *K. lactis*, *S. bayanus*, *S. kluyveri*, S. mikatae, *S. paradoxus*, *S. pombe*, and *Y. lipolytica* were retrieved from the Broad Institute (http://www.broadinstitute.org/regev/orthogroups)[40], and

alignments of the genetic sequences between orthologs were computed with ClustalW[41] *via* the corresponding amino acid sequences. tRNA counts were retrieved from the tRNA database (http://www.gtrnadb.ucsc.edu) if available, otherwise predicted with tRNAscan-SE from the genomic sequences[42]. mRNA expression levels for *S. cerevisiae* were obtained from[43], for S. pombe from[44], and for all other yeasts from[45]. 1574 alignments contained more than 7 sequences, our requirement to obtain meaningful conservation scores. To remove any intrinsic expression bias, we normalized the expression levels across the yeasts by quantile normalization, and selected 500 alignments each of high and low expression and with the lowest inner expression divergence measured by the standard deviation across orthologs. Removing alignments with more than 30% gaps yielded a set of 404 alignments of highly, and 302 alignments of lowly expressed genes. The mapping of PDB structures to *S. cerevisiae* genes was obtained from[46], yielding a curated set of 357 alignments.

Protein secondary structures were predicted with PSIPRED[47]. RNA secondary structures were predicted with the Vienna package[48]. For alignments with assigned PDB structure, secondary structure and relative accessible surface area were extracted with the DSSP program[49]. Sequence hydrophobicity was computed using the Kyte&Doolittle scale.

Translational efficiency and codon optimality

The classical translational effiency cTE_i for each codon is the published tRNA adaptation index (tAI) as computed with the codonR program[19]. It estimates the tRNA availability for each codon i from a weighted sum of the gene copy numbers $tGCN_{ij}$ of the matching tRNA isoacceptors j under a selective constraint s_{ij} on the efficiency of the codon–anticodon coupling, incorporating Crick's wobble rules[19]: n_i

$$W_i = \sum_{j=1}^{n_i} (1-s_{ij})tGCN_{ij}$$
$$cTE_i = W_i/W_{\max}$$

The selective constraint on codon–anticodon interactions s_{ij} is 0 for cognate tRNAs, and small for wobble interactions[19]. The overall efficiency of a codon is thus given by the sum of the contributions of the recognizing individual tRNAs under consideration of specific selective constraints based on the codon–anticodon interaction[19]. The division of the individual translational efficiencies W_i by the maximum translational efficiency W_{max} linearly rescales all translational efficiencies so that the maximum value is 1.

The codon usage cu_i was defined as relative estimate of how often each codon is translated. It is derived from the number of occurrences of each codon in an ORF, weighted by the corresponding transcript abundance, and summed up over all ORFs (Supplementary Fig. 1a). For codon i this is the sum of the counts c_{ij} of the codon i in gene j, weighted by the transcript abundance a_j of gene j, considering all genes in the genome g. For comparability, the codon usage is also rescale to have a maximum value of 1.

$$U_i = \sum_{j=1}^{g} a_j c_{ij}$$
$$cu_i = U_i / U_{\max}$$

In this work, the normalized translational efficiency nTE_i is subsequently defined as the ratio of tRNA availability cTE_i (supply), which is based on cellular tRNA abundance and selective constraints for wobble interactions, and codon usage cu_i (demand), linearly rescaled to have a maximum value of 1.

$$nTE'_i = cTE_i / cu_i$$
$$nTE_i = nTE'_i / nTE'_{\max}$$

Codons i with $cTE_i \quad cu_i$ are considered optimal, and nonoptimal otherwise. For comparison, we used the set of classical optimal codons reported in[20]. They are those found significantly enriched in the highest expressed genes by a Chi-square test[20]. Average translational efficiency profiles were computed as described in reference 18.

Randomization procedure & significant sites

We tested for evolutionary conservation of codon optimality in 10 closely related yeasts. The conservation score S of optimal or nonoptimal codons at any given position i is defined as $S_i = n_i / N_i$, where n_i is the number of optimal or nonoptimal codons at position i, and N_i the total number of aligned codons at position i respectively. We only considered alignments with at least 7 orthologous sequences, and a minimum of 5 codons has to be aligned for a conservation score to be computed at that position. Each alignment of orthologs was randomized 1000 times by individually shuffling each sequence to keep its original composition and maintain the individual codon bias. We employed additional randomization schemes to verify independence of amino acid biases (Supplementary Fig. 5,6). From the distribution of the conservation scores in the randomized alignments, we extracted alignment-specific the minimal conservation scores that are observed at less than 5% chance as significance thresholds. We considered sites with higher conservation scores than the significance thresholds as significantly conserved. If more optimal and nonoptimal codons respectively are significantly conserved in the biological alignment than in the randomized alignment, the site-specific evolutionary conservation of optimal and nonoptimal codons respectively is the result of selective pressure.

Individual and average translational efficiency profiles

Individual translational efficiency profiles were computed with the nTE and cTE scales, and smoothed with a sliding window of size 15, the size of the immediate ribosome footprint[18]. Average translational efficiency profiles were computed by aligning all genes at the start codon, and subsequently calculating the average translational efficiency for each position as described in[18]. Thus, the average profiles only reflect those fluctuations of the individual profiles that are present in all sequences. We randomly reshuffled all sequences to calculate the mean and standard deviation for each position[18].

Statistical testing

All statistical testing of associations was performed using the Cochran–Mantel–Haenszel test in the statistical computing environment R (http://www.r-project.org), and as described in reference 20. The enrichment of optimal or nonoptimal codons at specific helix positions in *S. cerevisiae* sequences was tested with Fisher's exact test and corrected for multiple-testing with the Benjamini & Hochberg (1995) method. All test statistics and definitions of optimal and nonoptimal codons are listed in the Supplementary Information (Supplementary Tables 1–8).

Data availability

All datasets of this study are available at http://www.stanford.edu/group/frydman/codons.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

References

1. Gingold H, Pilpel Y. Determinants of translation efficiency and accuracy. Mol Syst Biol. 2011; 7:481. [PubMed: 21487400]

2. Akashi H. Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy. Genetics. 1994; 136:927–935. [PubMed: 8005445]

3. Drummond DA, Wilke CO. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell. 2008; 134:341–352. [PubMed: 18662548]

4. Cabrita LD, Dobson CM, Christodoulou J. Protein folding on the ribosome. Curr Opin Struct Biol. 2010; 20:33–45. [PubMed: 20149635]

5. Thanaraj TA, Argos P. Ribosome-mediated translational pause and protein domain organization. Protein Sci. 1996; 5:1594–612. [PubMed: 8844849]

6. Zhang F, Saha S, Shabalina SA, Kashina A. Differential arginylation of actin isoforms is regulated by coding sequence-dependent degradation. Science. 2010; 329:1534–1537. [PubMed: 20847274]

7. Zhang G, Hubalewska M, Ignatova Z. Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. Nat Struct Mol Biol. 2009; 16:274–80. [PubMed: 19198590]

8. Kimchi-Sarfaty C, et al. A "silent" polymorphism in the MDR1 gene changes substrate specificity. Science. 2007; 315:525–528. [PubMed: 17185560]

9. Plotkin JB, Kudla G. Synonymous but not the same: the causes and consequences of codon bias. Nat Rev Genet. 2011; 12:32–42. [PubMed: 21102527]

10. Kramer G, Boehringer D, Ban N, Bukau B. The ribosome as a platform for co-translational processing, folding and targeting of newly synthesized proteins. Nat Struct Mol Biol. 2009; 16:589–597. [PubMed: 19491936]

11. Wilson DN, Beckmann R. The ribosomal tunnel as a functional environment for nascent polypeptide folding and translational stalling. Curr Opin Struct Biol. 2011; 21:274–282. [PubMed: 21316217]

12. Komar AA. A pause for thought along the co-translational folding pathway. Trends Biochem Sci. 2009; 34:16–24. [PubMed: 18996013]

13. Warnecke T, Hurst LD. GroEL dependency affects codon usage--support for a critical role of misfolding in gene evolution. Mol Syst Biol. 2010; 6:340. [PubMed: 20087338]

14. Brunak S, Engelbrecht J. Protein structure and the sequential structure of mRNA: alpha-helix and beta-sheet signals at the nucleotide level. Proteins. 1996; 25:237–52. [PubMed: 8811739]

15. Saunders R, Deane CM. Synonymous codon usage influences the local protein structure observed. Nucleic Acids Res. 2010; 38:6719–28. [PubMed: 20530529]

16. Gupta SK, Majumdar S, Bhattacharya TK, Ghosh TC. Studies on the relationships between the synonymous codon usage and protein secondary structural units. Biochem Biophys Res Commun. 2000; 269:692–6.

17. Petrov A, et al. Dynamics of the translational machinery. Curr Opin Struct Biol. 2011; 21:137–45. [PubMed: 21256733]

18. Tuller T, et al. An evolutionarily conserved mechanism for controlling the efficiency of protein translation. Cell. 2010; 141:344–354. [PubMed: 20403328]

19. dos Reis M, Savva R, Wernisch L. Solving the riddle of codon usage preferences: a test for translational selection. Nucl Acids Res. 2004; 32:5036–5044. [PubMed: 15448185]

20. Zhou T, Weems M, Wilke CO. Translationally optimal codons associate with structurally sensitive sites in proteins. Mol Biol Evol. 2009; 26:1571–1580. [PubMed: 19349643]

21. Zhang G, Fedyunin I, Miekley O, Valleriani A, Moura A, Ignatova Z. Global and local depletion of ternary complex limits translation elongation. Nucl Acid Res. 2010; 38:4778–4787.

22. Man O, Pilpel Y. Differential translation efficiency of orthologous genes is involved in phenotypic divergence of yeast species. Nat Genet. 2007; 39:415–21. [PubMed: 17277776]

23. Fraser HB, Moses AM, Schadt EE. Evidence for widespread adaptive evolution of gene expression in budding yeast. Proc Natl Acad Sci. 2010; 107:2977–82. [PubMed: 20133628]

24. Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science. 2009; 324:218–223. [PubMed: 19213877]

25. Reumers J, Maurer-Stroh S, Schymkowitz J, Rousseau F. Protein sequences encode safeguards against aggregation. Hum Mutat. 2009; 30:431–437. [PubMed: 19156839]

26. Lee Y, Zhou T, Tartaglia GG, Vendruscolo M, Wilke CO. Translationally optimal codons associate with aggregation-prone sites in proteins. Proteomics. 2010; 10:4163–71. [PubMed: 21046618]

27. Berndt U, Oellerer S, Zhang Y, Johnson AE, Rospert S. A signal-anchor sequence stimulates signal recognition particle binding to ribosomes from inside the exit tunnel. Proc Natl Acad Sci. 2009; 106:1398–1403. [PubMed: 19164516]

28. Lin PJ, Jongsma CG, Pool MR, Johnson AE. Polytopic membrane protein folding at L17 in the ribosome tunnel initiates cyclical changes at the translocon. J Cell Biol. 2011; 195:55–70. [PubMed: 21949410]

29. Shah P, Gilchrist MA. Explaining complex codon usage patterns with selection for translational efficiency, mutation bias, and genetic drift. Proc Natl Acad Sci. 2011; 108:10231–10236. [PubMed: 21646514]

30. Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. Why highly expressed proteins evolve slowly. Proc Natl Acad Sci. 2005; 102:14338–43. [PubMed: 16176987]

31. De Sancho D, Best RB. What is the time scale for α-helix nucleation? J Am Chem Soc. 2011; 133:6809–6816. [PubMed: 21480610]

32. O'Brien EP, Hsu ST, Christodoulou J, Vendruscolo M, Dobson CM. Transient Tertiary Structure Formation within the Ribosome Exit Port. J Am Chem Soc. 2010; 132:16928–16937. [PubMed: 21062068]

33. Lu J, Deutsch C. Folding zones inside the ribosomal exit tunnel. Nat Struct Mol Biol. 2005; 12:1123–1129. [PubMed: 16299515]

34. Lucent D, Snow CD, Aitken CE, Pande VS. Non-Bulk-Like Solvent Behavior in the Ribosome Exit Tunnel. PLoS Comput Biol. 2010; 6:e1000963. [PubMed: 20975935]

35. Qu X, et al. The ribosome uses two active mechanisms to unwind messenger RNA during translation. Nature. 2011; 475:118–21. [PubMed: 21734708]

36. Takyar S, Hickerson RP, Noller HF. mRNA helicase activity of the ribosome. Cell. 2005; 20:49–58. [PubMed: 15652481]

37. Cannarozzi G, Schraudolph NN, Faty M, von Rohr P, Friberg MT, Roth AC, Gonnet P, Gonnet G, Barral Y. A role for codon order in translation dynamics. Cell. 2010; 141:355–67. [PubMed: 20403329]

38. Alexandrov A, et al. Rapid tRNA decay can result from lack of nonessential modifications. Mol Cell. 2006; 21:87–96. [PubMed: 16387656]

39. Elf J, Nilsson D, Tenson T, Ehrenberg M. Selective charging of tRNA isoacceptors explains patterns of codon usage. Science. 2003; 300:1718–22. [PubMed: 12805541]

40. Wapinski I, Pfeffer A, Friedman N, Regev A. Natural history and evolutionary principles of gene duplication in fungi. Nature. 2007; 449:54–61. [PubMed: 17805289]

41. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research. 1994; 22:4673–4680. [PubMed: 7984417]

42. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Research. 1997; 25:955–964. [PubMed: 9023104]

43. Holstege FC, et al. Dissecting the regulatory circuitry of a eukaryotic genome. Cell. 1998; 95:717–728. [PubMed: 9845373]

44. Chen D, et al. Global transcriptional responses of fission yeast to environmental stress. Mol Biol Cell. 2003; 14:214–229. [PubMed: 12529438]

45. Tsankov AM, Thompson DA, Socha A, Regev A, Rando OJ. The role of nucleosome positioning in the evolution of gene regulation. PLoS Biol. 2010; 8:e1000414. [PubMed: 20625544]

46. Tóth-Petróczy A, Tawfik DS. Slow protein evolutionary rates are dictated by surface-core association. Proc Natl Acad Sci. 2011; 108:11151–6. [PubMed: 21690394]

47. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol. 1999; 292:195–202. [PubMed: 10493868]

48. Hofacker IL, Priwitzer B, Stadler PF. Prediction of locally stable RNA secondary structures for genome-wide surveys. Bioinformatics. 2004; 20:186–190. [PubMed: 14734309]

49. Joosten RP, et al. A series of PDB related databases for everyday needs. Nucleic Acids Research. 2011; 39:D411–419. [PubMed: 21071423]
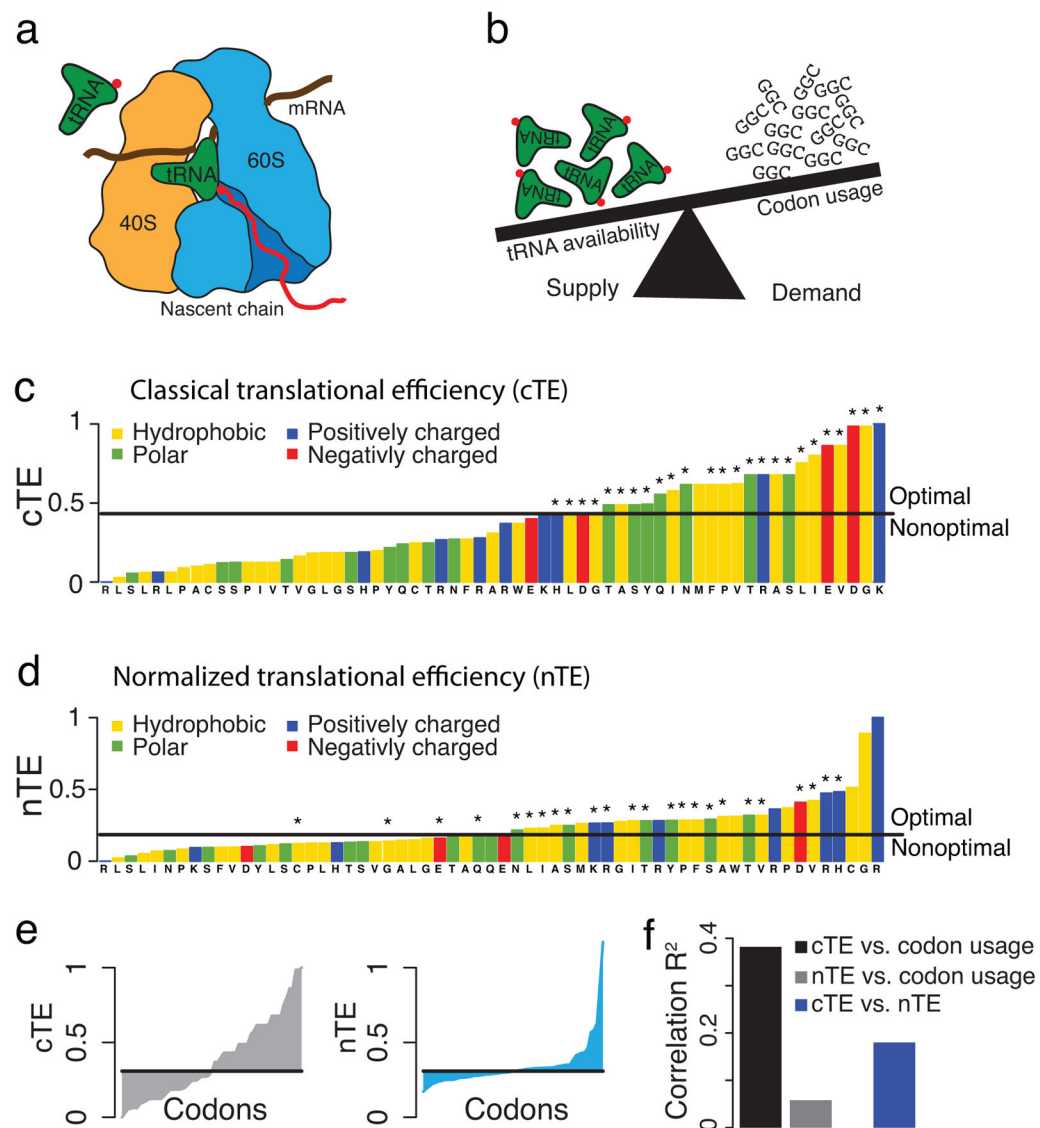
Figure 1.
A normalized translational efficiency scale that balances tRNA supply and demand. (**a**) Charged tRNAs recognize the codons in the mRNA and deliver the corresponding amino acid for elongation of the nascent chain. (**b**) The competition for tRNAs depends on both their availability in the cellular tRNA pool (supply), and the usage of the corresponding codons in the cell (demand). (**c**) The classical translational efficiency (cTE) scale is the established tRNA adaptation index. Codons that are over-represented in the most highly expressed genes are marked as optimal, and nonoptimal otherwise. *(**d**) The normalized translational efficiency (nTE) scale is defined by dividing the cTE by the codon usage, incorporating the competition between tRNA supply and demand. Codons are optimal if their relative tRNA availability exceeds the relative codon usage. The set of optimal codons from the classical scale is indicated with *. (**e**) Comparison of the cTE and nTE scales. (**f**) Correlations between the cTE and nTE scales and the codon usage.
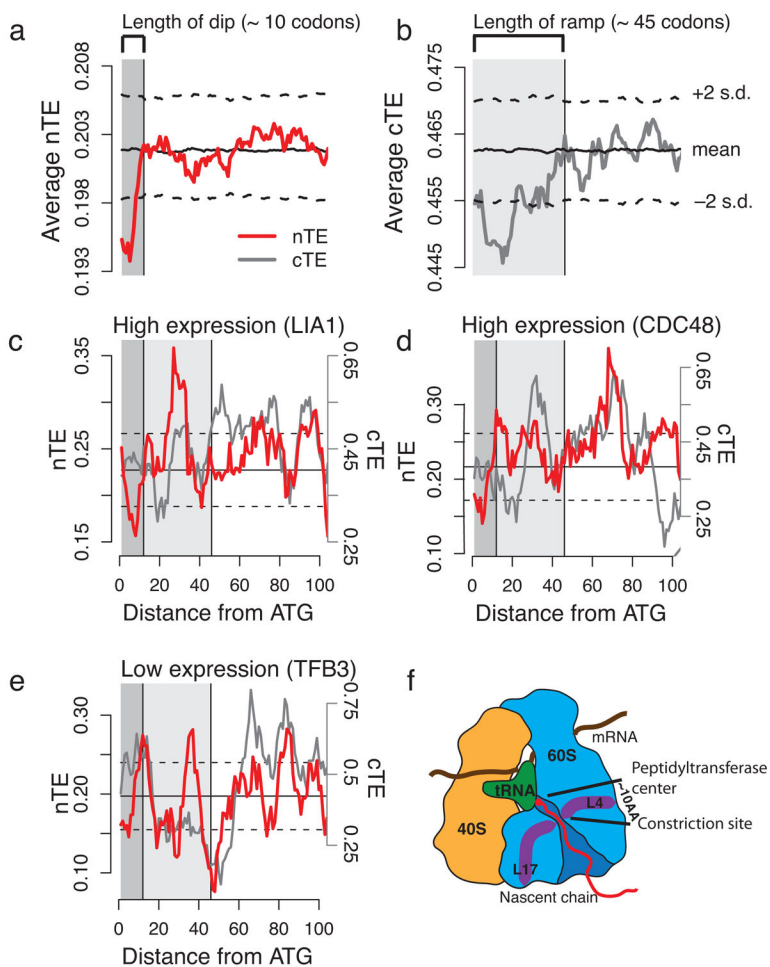
Figure 2.
A conserved short "dip" of low translational efficiency at the beginning of mRNAs. (**a**) The average nTE profile of the *S. cerevisiae* genome exhibits a very short "dip" of low translational efficiency at the beginning of the coding regions that spans approx. across the first 10 codons. (**b**) The average cTE profile shows an initial region of low translational efficiency that spans approx. across 35–50 codons, previously reported as the "ramp"[18]. (**c–e**) The nTE scale reveals short "dips" in individual genes such as the highly expressed hydroxylase LIA1 (**c**) and cell division control protein CDC48 (**d**), and the lowly expressed RNA polymerase II transcription factor TFB3 (**e**). In contrast, the cTE scale does neither reveal the "ramp" nor the "dip" in individual profiles. (**f**) The length of the "dip" matches approx. the distance from the peptidyltransferase center of the ribosome to a constriction site within the exit tunnel, where ribosomal proteins L4 and L17 sense nascent chains.

Figure 3.
Site-specific evolutionary conservation of codon optimality. (**a**) Optimal (O) and nonoptimal
(N) codons are projected onto an exemplary sequence alignment of the *S. cerevisiae* gene
RIB5, and shown together with positional conservation scores. Significantly conserved
optimal codons are indicated in blue, and nonoptimal codons in red. (**b**) Comparison of the
distributions of random (grey histogram and black line) and observed (red line) conservation
scores. Sites with higher conservation scores than expected by chance, i.e. outside the
alignment-specific significance thresholds (dotted lines), are considered significantly
conserved sites. (**c**) Fraction of analyzed ORFs that show a higher number of significantly
conserved optimal and nonoptimal sites than would be expected by chance, thus can be
assumed under selective pressure. (**d**) Distributions of significance thresholds for conserved
optimal and nonoptimal codons for highly and lowly expressed genes, obtained using the
nTE scale. (**e**) Fraction of ORFs under selective pressure computed with classical codon

optimality. If perfect site-specific conservation of optimal codons can already be observed by chance, selection for the site-specific conservation of optimal codons cannot be assumed, and these cases are thus indicated as 'not significant'. (**f**) Distributions of significance thresholds for conserved optimal and nonoptimal codons for highly and lowly expressed genes, obtained using the cTE scale.
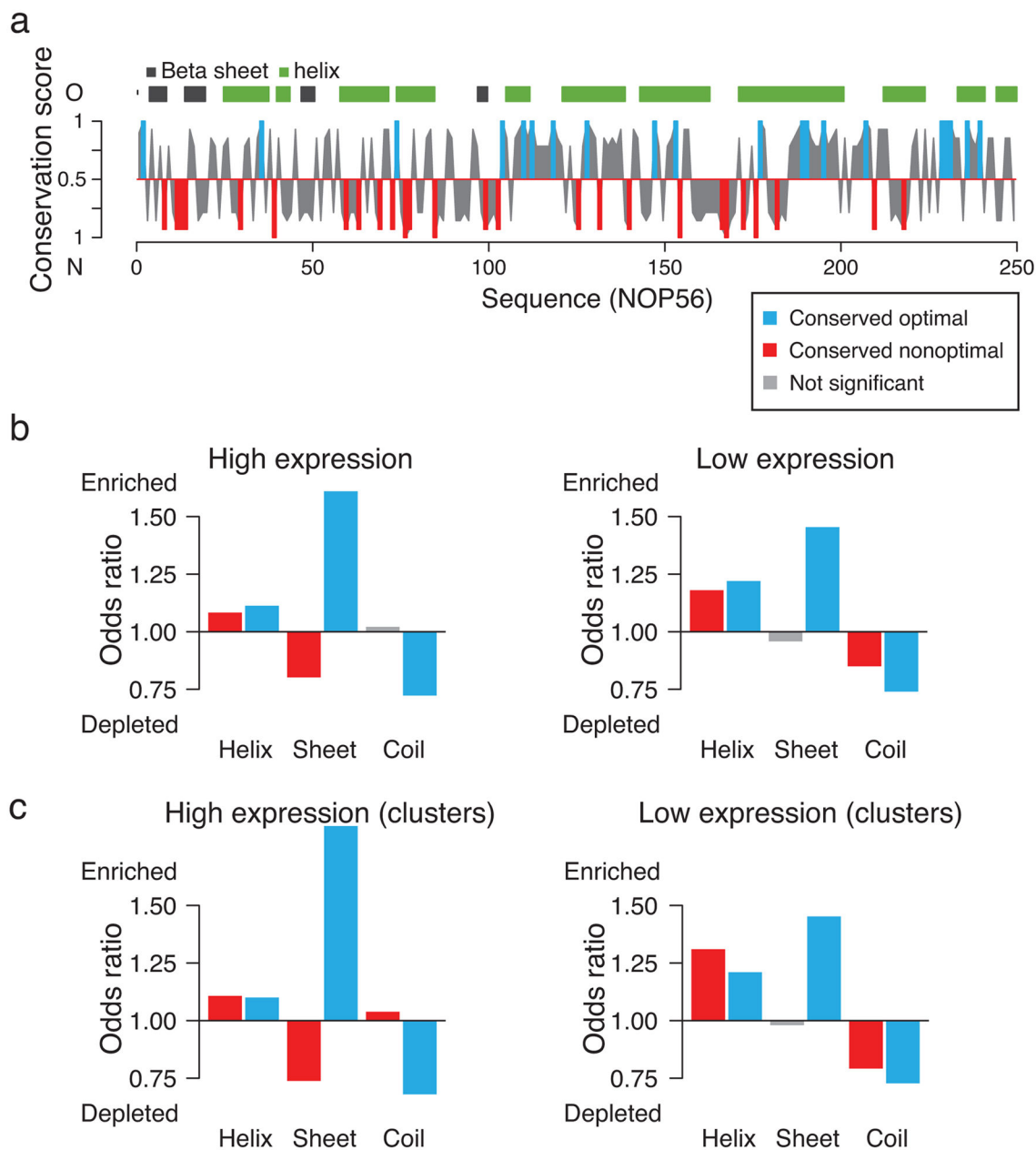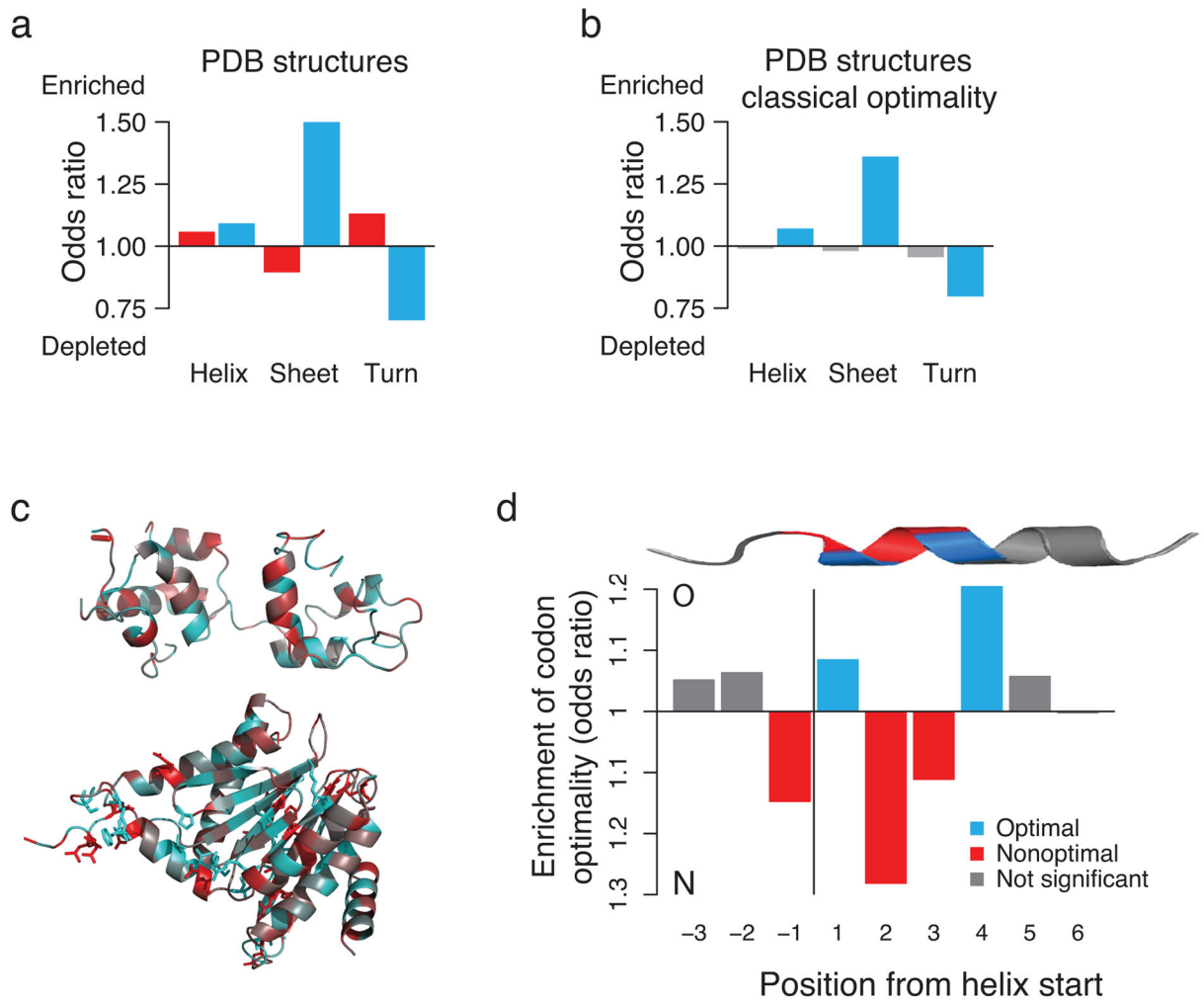
Figure 4.
Conserved codon optimality associates with signatures of co-translational folding. (**a**) Exemplary conservation profile of codon optimality for the gene NOP56, calculated using the nTE scale, together with the predicted protein secondary structure. Conserved optimal codons are shown in blue, conserved nonoptimal codons in red, and not significantly conserved sites in grey. (**b**) Associations between conserved codon optimality and predicted secondary structure for highly (left) and lowly (right) expressed genes. (**c**) As in (**b**), but for conserved optimal and nonoptimal codons that appear in clusters.

Figure 5.
Conserved codon optimality maps onto known protein structures. (**a**) Associations between clusters of conserved codon optimality and secondary structures in experimentally determined protein structures. (**b**) Associations between codon optimality as defined in cTE, and secondary structures in experimental PDB structures. All associations for nonoptimal codons are lost. (**c**) Secondary structure representations, colored according to the conservation of nonoptimal (red) and optimal (blue) codons, of the exemplary Myosin light chain 1 (top, PDB structure: 1M45_A) and 20S proteasome subunit (bottom, 1RYP_G). (**d**) A distinct and positional pattern of optimal and nonoptimal codons characterizes α-helices in *S. cerevisiae*. Positions 1 and 4 are strongly enriched in optimal codons, while the transition into the helix as well as positions 2 and 3 show a clear preference for nonoptimal codons
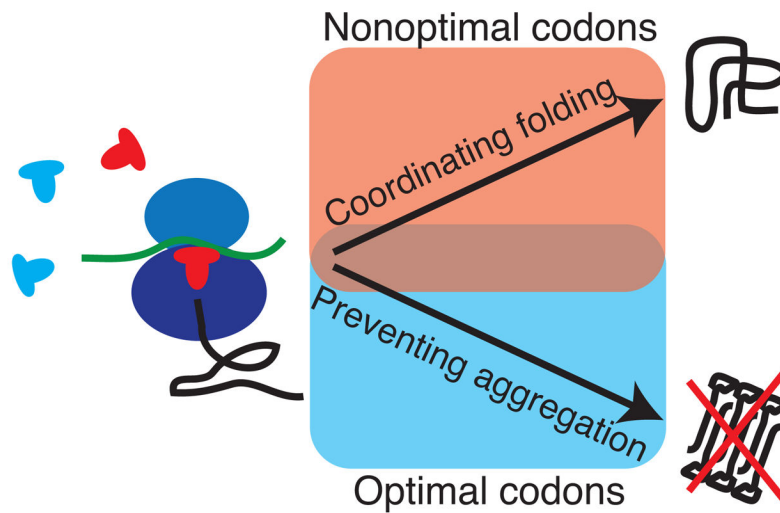
Figure 6.
Optimal and nonoptimal codons in protein synthesis and co-translational folding. Optimal codons are translated faster and more accurately, while nonoptimal codons can induce translational pausing. Evolutionarily conserved optimal codons are found predominantly at sites where translational accuracy is important to prevent aggregation. Evolutionarily conserved nonoptimal codons are found preferentially in secondary structure elements that can fold co-translationally, even already inside the ribosomal exit tunnel.