Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding

Sebastian Pechmann1,2 & Judith Frydman1,2

The choice of codons can influence local translation kinetics during protein synthesis. Whether codon preference is linked to cotranslational 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 yeast species, 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 cotranslational 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 cotranslational folding of corresponding polypeptides.

The translational efficiency of individual codons directly modulates the kinetics of protein synthesis1. Optimal codons are thought to be translated both faster and more accurately2,3. In turn, nonoptimal codons can slow down protein synthesis. Owing 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, that is, 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 cotranslational folding by allowing the nascent chain more time to develop native-like structure4. Indeed, a link between the mRNA sequences and in vivo folding of the encoded proteins has long been suggested5. For instance, synonymous substitutions reducing translational efficiency have been found to alter folding6,7 and subsequent function8 of the translated polypeptides. This supports the suggestion that protein synthesis is directly attuned to cotranslational folding9–13. However, no universal correlation across organisms was found between the position of nonoptimal codons and the location of structural units14 or domain boundaries15, or between synonymous codon usage and protein secondary-structural elements16. For Saccharomyces cerevisiae, a codon preference in relation to protein secondary structure could only be found for the amino acids glycine in loops and threonine in helices15. 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 in determining the rate of protein synthesis. During translation, specific tRNAs recognize the codons in the mRNA and deliver the corresponding amino acids to the ribosome17 (Fig. 1a). The abundance of tRNAs varies markedly in the cellular pool and is strongly correlated to tRNA gene copy numbers in the genome18. Because tRNA gene redundancy is a critical factor in explaining the correlation between translational selection and codon bias19, a scale of codon-specific translational efficiencies has been devised on the basis of the relative abundance of tRNAs as well as the selective constraints on codon-tRNA pairings for the ‘wobble’ non–Watson-Crick interactions19. Codons that are over-represented in highly expressed genes are also recognized by the most abundant tRNAs and are thus denoted as optimal20.

Although this ‘classical’ optimality scale has been very useful in deriving 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 toward optimal codons, as only optimal codons are explicitly defined20. In addition, the classical scale relies solely on tRNA abundances without taking into consideration the competition for tRNAs among all translating ribosomes1,21. 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 the translation rate of the corresponding codon21. Furthermore, the classical definition of translational efficiency incorporates species-specific tRNA pools and genomic sequences19,22 but does not directly include mRNA expression levels, thereby overlooking the effect of divergent gene expression observed even between closely related species23. Because mRNA expression levels will affect the demand for...
tRNAs, their 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. Notably, by also including the expression profile of the genome as a major characteristic of organism divergence, 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 yeast species in a site-specific manner that is independent of levels of expression. Notably, the evolutionary conservation of codon optimality reveals a direct and uniform link between codon optimality in the mRNAs and the secondary structures, and thus the 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 (tAI), does not incorporate the cellular tRNA dynamics driven by a trade-off between tRNA supply and demand. 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 by the codon usage (Online Methods and 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. 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 their relative availability of cognate tRNAs exceeds their relative usage.

Comparison of the nTE to the cTE scale (Fig. 1c-e and Supplementary Fig. 2a,b) shows that 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 scale, which increases almost linearly (Fig. 1e). The tail of low-efficiency codons and the plateau-like middle region are unique to the nTE scale, that is, the ratio of corresponding tRNA availability and codon usage (Supplementary Fig. 2c). Notably, this analysis suggests that tRNA supply and demand are closely matched for most codons at steady-state expression, which probably supports cost-effective proteome maintenance. Also of note, 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. The most efficient new optimal codons in the normalized scale encode glycine, the smallest and thus most neutral amino acid, and arginine, cysteine and proline, which are important gatekeeper residues that can prevent aberrant protein aggregation. The higher fraction of optimal codons encoding gatekeeper and hydrophobic residues revealed by the nTE scale probably reflects the need for increased fidelity during translation of sensitive polypeptide segments critical to correct folding and avoidance of aggregation. 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 nondegenerate amino acids, methionine and tryptophan, which are nonoptimal in the cTE scale, are optimal in the nTE scale. Notably, 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, whereas 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.

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 be expectedly be evolutionarily conserved. We undertook a systematic analysis of the conservation of codon optimality across ten closely related yeasts (Online 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 ~35–50 codons [15]. We thus next tested whether 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 ~10 codons long, that is, much shorter than found by using the cTE scale (Fig. 2b). Notably, this short ‘dip’ was evolutionarily conserved in all other analyzed yeasts (Supplementary Fig. 3). Although the longer ramp is not observed in translational efficiency profiles of individual genes when 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, for example, hydroxylase LIA1 and cell division control protein CDC48 (Fig. 2d), and in the weakly expressed RNA polymerase II transcription factor TFB3 (e). Distance is shown in codons. In contrast, the cTE scale reveals neither the ramp nor the dip in individual profiles. (f) The length of the dip approximately matches 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 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 across the first ~10 codons. (b) The average cTE profile shows an initial region of low translational efficiency that spans ~35–50 codons, previously reported as the ramp. 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 weakly expressed RNA polymerase II transcription factor TFB3 (e). Distance is shown in codons. In contrast, the cTE scale reveals neither the ramp nor the dip in individual profiles. (f) The length of the dip approximately matches 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.

Site-specific evolutionary conservation of codon optimality
A functional link between the positioning of optimal and nonoptimal codons along the mRNA and cotranslational 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 (Online Methods, Fig. 3a and 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 (that is, significance thresholds with P < 0.05). Sites in the biological alignment that have higher conservation scores than these 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, then 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 [25], 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 (Online Methods and Supplementary Fig. 4 c). Notably, 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. 5 h). 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 [26]. 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). Notably, both optimal and nonoptimal codons are found conserved in equal measure in both highly and weakly 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 also fulfills an evolutionarily selected function in protein biogenesis [6–8]. Of note, the same...
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 (gray histogram and black line) and observed (red line) conservation scores. Sites with higher conservation scores than expected by chance, that is, outside the alignment-specific significance thresholds (dashed lines), are considered to be significantly conserved sites. (c) Fraction of analyzed open reading frames (ORFs) that show a higher number of significantly conserved optimal and nonoptimal sites in their respective analyses than would be expected by chance and thus assumed to be under selective pressure. (d) Distributions of significance thresholds for conserved optimal and nonoptimal codons for highly and weakly expressed genes, obtained by 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 weakly expressed genes, obtained by using the cTE scale.

Conserved hidden signatures of cotranslational folding
Having determined that optimal and nonoptimal codons are generally conserved, it is reasonable to ask whether their conservation patterns are related to their involvement in cotranslational folding events. We thus mapped the codon conservation profiles from the alignments described above onto the corresponding *S. cerevisiae* protein sequences and structures (for example, Figure 4a). Both conserved optimal and nonoptimal codons are distributed throughout the mRNA sequences and often appear in clusters (Figure 4a). We next tested for statistical associations between conserved optimal and nonoptimal codons and secondary-structure elements of the encoded nascent chains (Figure 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 (Figure 4b). In contrast, β-sheets are enriched in conserved optimal but depleted in conserved nonoptimal codons in both highly and weakly expressed genes (Figure 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 weakly expressed genes (Fig. 4b).

Notably, α-helices can already form cotranslationally even within the ribosomal tunnel10,11, whereas coil regions, comprising loops that fold near the exit of the ribosomal tunnel, have been shown to have key roles in cotranslational protein folding7. 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 codons as well as depletion of conserved nonoptimal codons could primarily serve to reduce the risk of phenotypic missense mutations leading to aggregation.

Because 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 association is stronger for highly expressed genes (Supplementary Fig. 6c), probably owing to a greater need for translational fidelity in these abundant proteins3,20,26.

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). Notably, PDB structures allow conservation of nonoptimal (red) and optimal (blue) codons, of the exemplary myosin light chain 1 (top, PDB 1M45) and 20S proteasome subunit G (bottom, 1RYP). (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, whereas the transitions into the helix as well as into positions 2 and 3 show a clear preference for nonoptimal codons.

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 to low cooperativity that depend on an initial nucleation step31. Furthermore, this is the main folding event that has been found to occur inside the ribosome exit tunnel11,32,33; indeed, helix formation has been experimentally observed to occur very early in the ribosomal tunnel, between the peptidyltransferase center and the constriction site28. Notably, 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 positions 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, independently 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 inside the ribosome33, with the strongest compaction into secondary structure observed proximal to the peptidyltransferase center33. Helix formation inside the ribosomal tunnel appears strongly sequence dependent but cannot be explained by sequence hydrophobicity or helical propensity alone33. The specialized environment within the tunnel is proposed to create a rugged solvation landscape that may slow down folding34. 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

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 1M45) and 20S proteasome subunit G (bottom, 1RYP). (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, whereas the transitions into the helix as well as into positions 2 and 3 show a clear preference for nonoptimal codons.
correspond to a distinct codon bias and have been suggested to have a role in the cotranslational folding of the encoded polypeptides. However, a coherent and uniform link has so far remained elusive and may be difficult to detect, in part probably owing 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 nTE incorporates the biologically relevant competition for tRNAs among all ribosomes, which is known to influence the kinetics of translation elongation\(^1,2\). As a result, nTE allows for a better comparison between organisms. This is important, as a general and systematic link between codon optimality and cotranslational 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, which further confirms 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. Notably, 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\(^1,2,8,33\). 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 uncovering. Because 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 cotranslational folding outside the ribosome. For instance, exemplary experimental work has demonstrated the importance of nonoptimal codons in loops for the successful cotranslational folding of protein domains\(^2\).

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 those 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 cotranslational folding.

Although mRNA secondary structure may add an additional layer of translational regulation\(^13\), 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 a helicase, unraveling the translated mRNAs\(^36\).

One key aspect of the new nTE scale is that tRNA supply and demand are very balanced for most codons at steady state. Almost all amino acids are encoded by equal numbers of optimal and nonoptimal codons in nTE. As a result, the strong and consistent link between conserved codon optimality and protein secondary structure that we observed 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 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 tempting to speculate that tRNA recycling, dynamics and modifications\(^27,38\) may further influence the rhythm of translation elongation, both at steady state and in response to cellular stresses such as amino acid starvation\(^29\).

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.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

We thank the Frydman lab for helpful discussions. We gratefully acknowledge support from an European Molecular Biology Organization Long-Term Fellowship (ALTF 1334-2010) to S.P. and US National Institutes of Health grants GM56433 and AI91575 to J.F.

AUTHOR CONTRIBUTIONS

S.P performed all analyses; S.P. and J.F. designed research, interpreted the data and wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/doifinder/10.1038/nsmb.2466. Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Gingold, H. & Pilpel, Y. Determinants of translation efficiency and accuracy. Mol. Syst. Biol. 7, 481 (2011).
2. Akashi, H. Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy. Genetics 136, 927–935 (1994).
3. Drummond, D.A. & Wilke, C.O. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell 134, 341–352 (2008).
4. Cabrita, L.D., Dobson, C.M. & Christodoulou, J. Protein folding on the ribosome. Curr. Opin. Struct. Biol. 20, 33–45 (2010).
5. Thanaraj, T.A. & Argos, P. Ribosome-mediated translational pause and protein domain organization. Protein Sci. 5, 1594–1612 (1996).
ONLINE METHODS

Data sources. Genomic sequences and ortholog assignments for S. cerevisiae, Candida glabrata, Debaryomyces hansenii, Kluyveromyces lactis, Saccharomyces bayanus, Saccharomyces kluyveri, Saccharomyces mikatae, Saccharomyces paradoxus, Schizosaccharomyces pombe and Yarrowia lipolytica were retrieved from the Broad Institute (http://www.broadinstitute.org/regev/orthogroups)41, and alignments of the genetic sequences between orthologs were computed with ClustalW42 using the corresponding amino acid sequences. tRNA counts were retrieved from the tRNA database (http://grnmdb.ucsc.edu/) if available or were otherwise predicted with tRNAscan-SE from the genomic sequences43. mRNA expression levels for S. cerevisiae were obtained from ref. 43, for S. pombe from ref. 44 and for all other yeasts from ref. 45. 1,574 alignments contained more than seven 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 s.d. across orthologs. Removing alignments with more than 30% gaps yielded a set of 404 alignments of highly expressed genes and 302 alignments of weakly expressed genes. The mapping of PDB structures to S. cerevisiae genes was obtained from ref. 46, yielding a curated set of 357 alignments.

Protein secondary structures were predicted with PSIPRED47. RNA secondary structures were predicted with the Vienna package48. For alignments with assigned PDB structure, secondary structure and relative accessible surface area were extracted with the DSSP program49. Sequence hydrophobicity was computed by using the Kyte and Doolittle scale.

Translational efficiency and codon optimality. The classical translational efficiency cTE, for each codon is the published tRNA adaptation index (tAI) as computed with codonR19. It estimates the tRNA availability for each codon i from a weighted sum of the gene copy numbers tGCNi of the matching tRNA isoacceptors j under a selective constraint sj on the efficiency of the codon-anticodon coupling, incorporating Crick’s wobble rules19.

\[
W_j = \frac{\sum_{i=1}^{N_j} (1-s_j) tG C N_i}{W_{\text{max}}}
\]

\[
cTE = W_j / W_{\text{max}}
\]

The selective constraint on codon-anticodon interactions sj is 0 for cognate tRNAs and small for wobble interactions19. 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 on the basis of the codon-anticodon interaction19. The division of the individual translational efficiencies Wj by the maximum translational efficiency Wmax linearly rescales all translational efficiencies so that the maximum value is 1.

The codon usage cuj was defined as a 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 over all ORFs (Supplementary Fig. 1a). For codon i, this is the sum of the counts ci of the codon i in gene j, weighted by the transcript abundance ai of gene j, considering all genes in the genome g. For comparability, the codon usage is also rescaled to have a maximum value of 1.

\[
U_j = \sum_{g=1}^{G} a_i c_{ij}
\]

\[
cu_j = U_j / U_{\text{max}}
\]

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

\[
n\text{TE}^j = c\text{TE} / c\text{uj}
\]

\[
n\text{TE}_i = n\text{TE}^j / n\text{TE}_\text{max}
\]

Codons i with cTEi ≥ cuj are considered optimal and are otherwise considered nonoptimal. For comparison, we used the set of classical optimal codons reported in ref. 20. They are those found significantly enriched in the highest-expressed genes by a chi-squared test20. Average translational efficiency profiles were computed as described in ref. 18.

Randomization procedure and significant sites. We tested for evolutionary conservation of codon optimality in ten closely related yeasts. The conservation score S of optimal or nonoptimal codons at any given position i is defined as 

\[
S_i = \frac{n_i}{N_i}
\]

where ni is the number of optimal or nonoptimal codons at position i and Ni is the total number of aligned codons at position i. We only considered alignments with at least seven orthologous sequences, and a minimum of five codons have to be aligned for a conservation score to be computed at that position. Each alignment of orthologs was randomized 1,000 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 Figs. 5 and 6). From the distribution of the conservation scores in the randomized alignments, we extracted the alignment-specific minimal conservation scores that are observed at less than 5% chance as the significance threshold. We considered sites with higher conservation scores than the significance thresholds as significantly conserved. If more optimal and nonoptimal codons from their respective analyses are significantly conserved in the biological alignment than in the randomized alignment, the site-specific evolutionary conservation of optimal and nonoptimal codons 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 footprint18. 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 ref. 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 s.d. for each position18.

Statistical testing. All statistical testing of associations was performed by using the Cochran-Mantel-Haenszel test in the statistical computing environment R (http://www.r-project.org/) and as described in ref. 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 Supplementary Tables 1–8.

Data availability. All data sets of this study are available at http://www.stanford.edu/group/frydman/codons/.