Translation efficiency is determined by both codon bias and folding energy

Tamir Tuller[a,b,1], Yedael Y. Waldman[c], Martin Kupiec[d], and Eytan Ruppin[c,e]

[a]Faculty of Mathematics and Computer Science and [b]Department of Molecular Genetics, Weizmann Institute of Science, Rehovot 76100, Israel; and [c]Blavatnik School of Computer Science, [d]Department of Molecular Microbiology and Biotechnology, and [e]School of Medicine, Tel Aviv University, Ramat Aviv 69978, Israel

Edited by Fred Sherman, University of Rochester School of Medicine and Dentistry, Rochester, NY, and approved December 29, 2009 (received for review August 31, 2009)

Synonymous mutations do not alter the protein produced yet can have a significant effect on protein levels. The mechanisms by which this effect is achieved are controversial; although some previous studies have suggested that codon bias is the most important determinant of translation efficiency, a recent study suggested that mRNA folding at the beginning of genes is the dominant factor via its effect on translation initiation. Using the *Escherichia coli* and *Saccharomyces cerevisiae* transcriptomes, we conducted a genome-scale study aiming at dissecting the determinants of translation efficiency. There is a significant association between codon bias and translation efficiency across all endogenous genes in *E. coli* and *S. cerevisiae* but no association between folding energy and translation efficiency, demonstrating the role of codon bias as an important determinant of translation efficiency. However, folding energy does modulate the strength of association between codon bias and translation efficiency, which is maximized at very weak mRNA folding (i.e., high folding energy) levels. We find a strong correlation between the genomic profiles of ribosomal density and genomic profiles of folding energy across mRNA, suggesting that lower folding energies slow down the ribosomes and decrease translation efficiency. Accordingly, we find that selection forces act near uniformly to decrease the folding energy at the beginning of genes. In summary, these findings testify that in endogenous genes, folding energy affects translation efficiency in a global manner that is not related to the expression levels of individual genes, and thus cannot be detected by correlation with their expression levels.

mRNA folding | protein abundance | synonymous mutations | ribosome density | translation initiation

Synonymous mutations (mutations that alter the coding DNA and RNA sequence without affecting the amino acid sequence of the protein produced) can significantly influence protein abundance via changes in translation efficiency (1–7). Previous studies have suggested two main mechanisms by which protein abundance may be modulated by synonymous mutations: codon bias, denoting the differential usage of synonymous codons depending on the levels of their corresponding tRNAs in the cell (8), and the folding energy of the mRNA transcript, which may influence ribosome binding, and therefore translation initiation (5, 9).

Translation efficiency can be analyzed at two different levels, local and global, where the global level reflects factors that modify the translation efficiency on the transcriptome level but do not change the expression levels of single genes in a causal way (10, 11). A classic example of global mechanisms affecting translation efficiency is the correlation between mRNA levels and codon bias; the usage of efficient codons increases the elongation rate. Assuming constant flux of ribosomes, this would result in fewer ribosomes on mRNA, and thus a better allocation of ribosomes. As a result, the total rate of protein synthesis increases and cell growth is accelerated (11). Genes with higher mRNA levels potentially "consume" more ribosomes, and thus are under stronger selection for global translation efficiency. However, it should be noted that not all global effects are necessarily correlated with expression levels because they may affect translation efficiency in a uniform manner across genes irrespective of their expression levels. As we shall see, these effects play an important role in the following. In difference, factors affecting local translation efficiency are associated with a change in the levels of particular proteins, given their mRNA levels (8). The local translation efficiency of a gene is quantified by the ratio between the protein abundance and the mRNA levels of that gene. The effect of a factor on local translation efficiency can hence be traced by finding a significant correlation between this factor and the ratio between protein abundance and mRNA levels of genes or, equivalently, by finding a significant correlation between the factor and protein abundance when controlling for the mRNA level of the genes.

Recently, Kudla et al. (11) generated a library of 154 genes with different random synonymous mutations encoding the same GFP protein. Studying their influence on its protein levels in *Escherichia coli*, they found that the folding energy of the mRNA segment of the first ~40 nucleotides of the transcript has a significant correlation with the GFP protein abundance, whereas codon bias, measured by the Codon Adaptation Index (12), does not exhibit a significant correlation with protein. Hence, these investigators have suggested that mRNA folding at the beginning of the sequence plays a predominant role in shaping expression levels of individual genes (i.e., local translation efficiency), whereas the previously reported correlations between codon bias and translation efficiency (13, 14) are more likely to arise as a result of selection to increase global translation efficiency across all genes by optimizing ribosome allocation.

Following this work, which focused on a single nonendogenous protein, we examine here the joint role of codon bias and folding energy in determining gene translation efficiency across a whole genome, studying their effects by considering systematically the *E. coli* and *Saccharomyces cerevisiae* transcriptomes. To this end, we employ the tRNA adaptation index (tAI) (15) (*Materials and Methods*,Table S1 and Table S2) as a measure of codon bias; folding energy was calculated using UNAfold software (16) (*Materials and Methods*).

Results

Selection Forces Act To Decrease Folding Energy at the Beginning of Genes. Our first step was to examine whether the mean folding energy of the first 40 nucleotides of each mRNA (of the 4,226 *E. coli* genes) is significantly higher than the mean folding energy of other 40-nt windows. Indeed, we find a significant difference

Author contributions: T.T., M.K., and E.R. designed research; T.T. performed research; T.T. analyzed data; and T.T., Y.Y.W., M.K., and E.R. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. E-mail: tamirtul@post.tau.ac.il.

This article contains supporting information online at www.pnas.org/cgi/content/full/0909910107/DCSupplemental.

between the first window as compared with other windows (e.g., −5 for the interval nucleotides 1–40 vs. −7.95 for nucleotides 41–80, $n = 4{,}226$; Wilcoxon test: $P < 10^{-16}$; similar results were observed for other windows between 41 and 240 nucleotides; all P values were $<10^{-16}$ (Fig. 1*A*), extending the results reported previously (11). In addition, the variance in folding energy is lower in the first window than in all other sliding windows (Fig. 1B), and further analysis of the data of Kudla et al. (11) reveals a significant positive relation between folding energy at the beginning of genes and fitness (measured by the OD of growing cultures; i.e., when there are nonfolding structures at the beginning of the GFP gene, the fitness is higher; see details in *SI Note 1*, Figs. S1 and S2). Similar results were obtained for the *S. cerevisiae* transcriptome (Fig. 1 *C* and *D*); the mean folding energy of the first 40 nucleotides is significantly higher than that of nucleotides 41–80 (−4.3580 vs. −5.1558, $n = 5{,}869$; Wilcoxon test: $P < 10^{-16}$) and significantly higher than the folding energies of all other windows between 41 and 240 nucleotides (all $P <$ 0.003). Interestingly, in both organisms, the mean folding energy of the 41–80-nt interval was lower than the mean folding energy of all other intervals, possibly to minimize the formation of potentially deleterious structures in the region of the ribosome binding site.

To validate further that this finding is not only a result of amino acid bias, we performed an additional test. We compared the folding profiles with those obtained for randomized versions of the genomes of the analyzed organisms, preserving the original codon bias and amino acid composition (*Materials and Methods*). In both *E. coli* and *S. cerevisiae*, we found that for windows more distant from the beginning of the ORF (starting from window index 18 and 10 in *E. coli* and *S. cerevisiae*, respectively; window index denotes the distance in nucleotides from the beginning of the ORF and the beginning of the window; negative window index denotes a window that begins before the beginning of the ORF), these random sequences show higher (significantly higher in most of the windows) folding energy (i.e., weaker folding) than the original profile, thus supporting previous results (17) (Fig. 1). However, when considering the windows that are close to the beginning of the ORF or even partially include the 5′-UTR near the beginning of the ORF [windows whose indexes are between −23 (i.e., they start 23 nucleotides before the first nucleotide in the ORF, and 13 in *E. coli* and windows whose indexes are between −13 and 6 in *S. cerevisiae*], these random sequences show significantly lower folding energy (i.e., stronger mRNA structures) than the original profile (Fig. 1; see *SI Note 2* and Fig. S3 for a similar analysis of the terminal end of ORFs). Taken together, these results support the suggestion that the nonfolding structures at the beginning of ORFs are selected for.

There Is a Significant Association Between Codon Bias and Translation Efficiency but Not Between Folding Energy and Translation Efficiency. Because there seems to be a selection for higher folding energy levels at the beginning of *E. coli* and *S. cerevisiae* genes, and following the findings of Kudla et al. (11) regarding the GFP gene, it is pertinent to examine how folding energy at the initial window of the transcript affects the translation efficiency across the whole transcriptome. Surprisingly, in *E. coli*, we do not find a significant correlation between local translation efficiency and the folding energy of the first 40 nucleotides ($r = 0.019$, $P = 0.6971$; $n = 423$; Fig. 2). This observation holds also when conditioning with codon bias [the partial correlation of folding energy and local translation efficiency given codon bias r(Folding Energy, Local Translation Efficiency|Codon Bias) = 0.0219; $P = 0.65$; $n = 423$] or when we examine the correlation between local translation efficiency and the folding energy of other 40-nt windows (we examined all the first 250 windows and performed false discovery rate correction for multiple hypothesis testing). Moreover, no correlation was observed when averaging all the first 250 windows in each gene. In contrast, we do find a significant correlation between local translation efficiency and codon bias [$r = 0.27$ and $P = 1.7 \times 10^{-8}$; $n = 423$; r(Codon Bias, Local Translation Efficiency|Folding Energy) = 0.27, $P = 1.67 \times 10^{-8}$; Fig. 2]. Similar results were obtained across *S. cerevisiae* genes [for tAI: $r = 0.123$ and $P = 1.47 \times 10^{-9}$, r(Codon Bias, Local Translation Efficiency|Folding Energy) = 0.1173 and $P = 1.19 \times 10^{-8}$; for folding energy: $r = 0.0006$ and $P = 0.98$, r(Folding Energy, Translation Efficiency|Codon Bias) = −0.0122 and $P = 0.5553$; $n = 2{,}350$]. Examining the relation with protein abundance levels directly (i.e., a measure of global translation efficiency), we again obtain similar results for both *E. coli* and *S. cerevisiae* (Fig. S4). Finally, the partial correlation between protein abundance and codon bias given the genes' mRNA levels is significant (as opposed to the partial correlation between

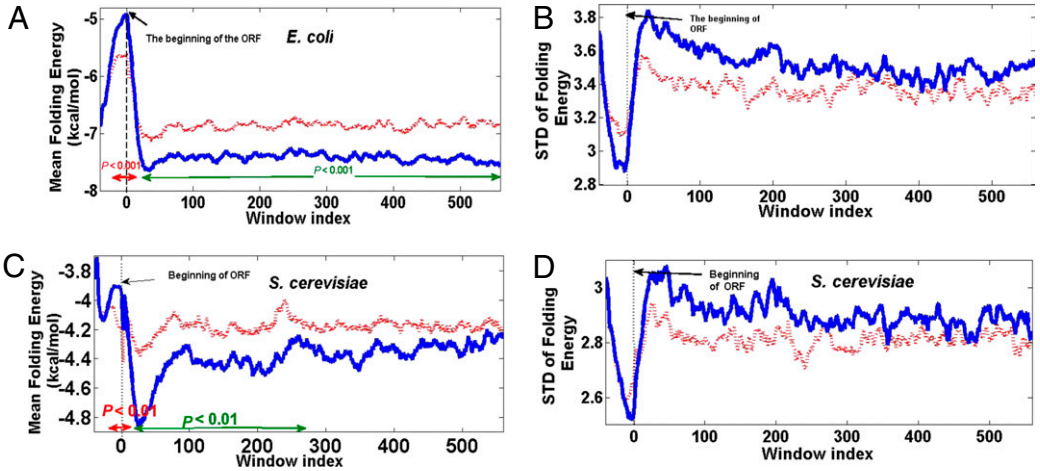

Fig. 1. Endogenous genes in *E. coli* and *S. cerevisiae*. (*A*) Profile of folding energy (mean of sliding window of 40-nt length) across the *E. coli* genome (blue) vs. the profile for a randomized genome (dashed red); the window index denotes the distance (in nucleotides) from the beginning of the ORF to the beginning of the window. The figures also include the 5′-UTR near the beginning of the ORF (negative window indexes). Regions where the folding energy of the real genome is significantly higher (red) or lower (green) than the randomized genome are marked at the bottom of the figure. (*B*) Profile of folding energy STD across the *E. coli* genome (blue) vs. the profile for a randomized genome (dashed red). (*C* and *D*) Similar to *A* and *B* for the *S. cerevisiae* genome.

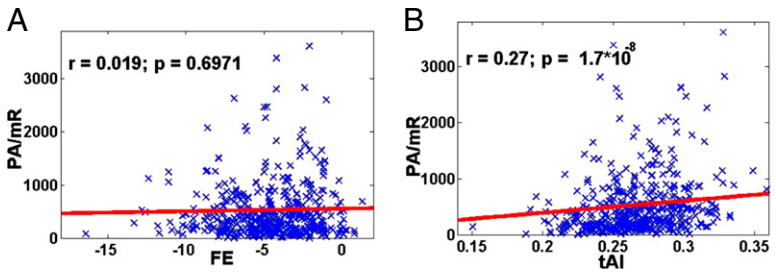

Fig. 2. Endogenous genes in *E. coli*. (*A*) Local translation efficiency (protein abundance/mRNA levels) vs. codon bias (tAI) for all genes. (*B*) Local translation efficiency vs. folding energy of the first 40 nucleotides for all *E. coli* genes.

protein abundance and folding energy given the mRNA levels), further emphasizing the role of codon bias (rather than folding energy) in determining the local translation efficiency of endogenous genes in *E. coli* [r(Protein Abundance, Codon Bias|mRNA Levels) = 0.28, $P = 2.74 \times 10^{-9}$; r(Protein Abundance, Folding Energy|mRNA Levels) = 0.0041, $P = 0.9327$; $n = 423$] and in *S. cerevisiae* [r(Protein Abundance, Codon Bias|mRNA Levels) = 0.38, $P = 8.54 \times 10^{-81}$; r(Protein Abundance, Folding Energy|mRNA Levels) = 0.0095, $P = 0.6458$; $n = 2{,}350$]. These results indicate that the selection for weak mRNA folding at the beginning of genes is global and is not related to the expression level of specific genes, in contrast to codon bias.

Folding Energy Modulates the Relation Between Local Translation Efficiency and Codon Bias. To elucidate the relation between codon bias, folding energy, and local translation efficiency better, we divided all *E. coli* genes into five equal size bins according to their folding energy and measured the correlation between codon bias or folding energy and local translation efficiency in each bin separately. As evident from Fig. 3 *A* and *B*, the codon bias and local translation efficiency correlation is significant in three of the five bins, whereas the folding energy and local translation efficiency correlation is borderline significant only in one window. Specifically, the most significant correlation between codon bias and local translation efficiency is in the bin corresponding to very high folding energy (−1.2 mean folding energy); at these levels, the mRNA folding is very weak and codon bias remains the sole determinant of local translation efficiency. Overall, the relation between codon bias and local translation efficiency as a function of folding energy is not monotonic, as can be seen from the relatively strong correlation in the second bin (−6 mean folding energy). The results for *S. cerevisiae* show a similar trend of more significant codon bias effects but with much lower correlation values that are more evenly distributed among the different folding energy bins (Fig. 3 *C* and *D*).

Role of Folding Energy in Determining Global Translation Efficiency Can Be Explained by Examining the Association Between Folding Energy and Ribosomal Density. The recent findings of Ingolia et al. (18) reporting genome-wide measurements of ribosome densities at a resolution of single nucleotides for *S. cerevisiae* in two conditions [growing on yeast peptone dextrose (YPD) and in starvation] may help to shed light on the findings reported in the previous section. These data have enabled us to compare the relation between the genomic profile of folding energy and the genomic profile of ribosome density. A plot of the spatial genomic ribosome density [based on the data of Ingolia et al. (18)] and the spatial mean genomic folding energy (measured in sliding windows of 40 nucleotides, as before) appears in Fig. 4. The correlation between the profile of ribosome density in YPD and the profile of folding energy is −0.63 ($P = 2.4 \times 10^{-8}$; $n = 66$); the correlation between the ribosome density in starvation and the folding energy is −0.51 ($P = 1.1 \times 10^{-5}$; $n = 66$). These

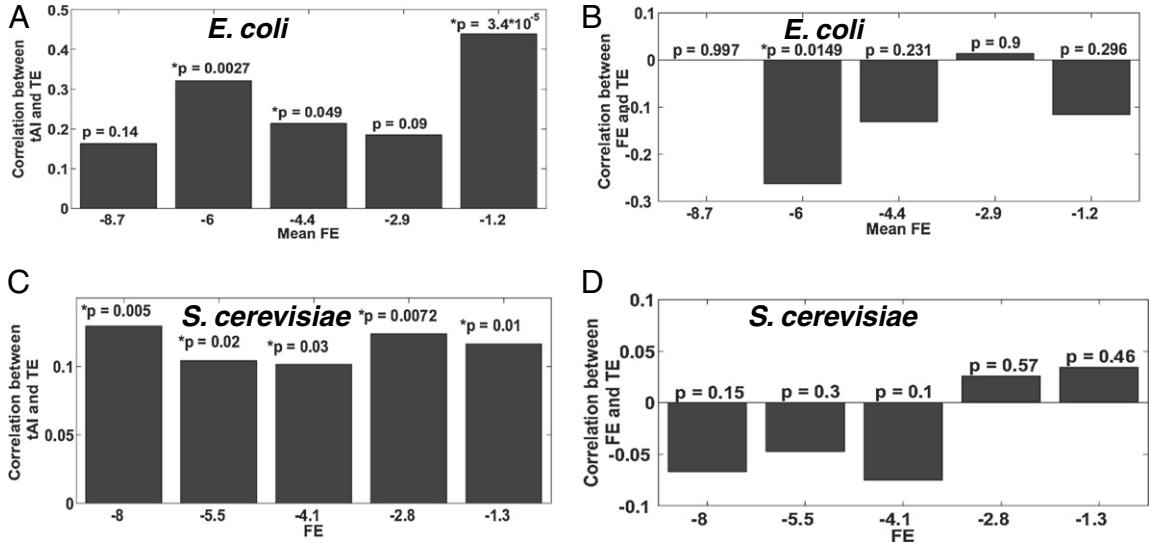

Fig. 3. *E. coli* and *S. cerevisiae*. (A) Correlation between codon bias and local translation efficiency (*y* axis) for five equal-sized bins according to folding energy values (*x* axis). (B) Correlation between folding energy and translation efficiency (*y* axis) for five equal-sized bins according to folding energy values (*x* axis). (*C* and *D*) Same correlations (but with much lower magnitudes) are detected for *S. cerevisiae*.

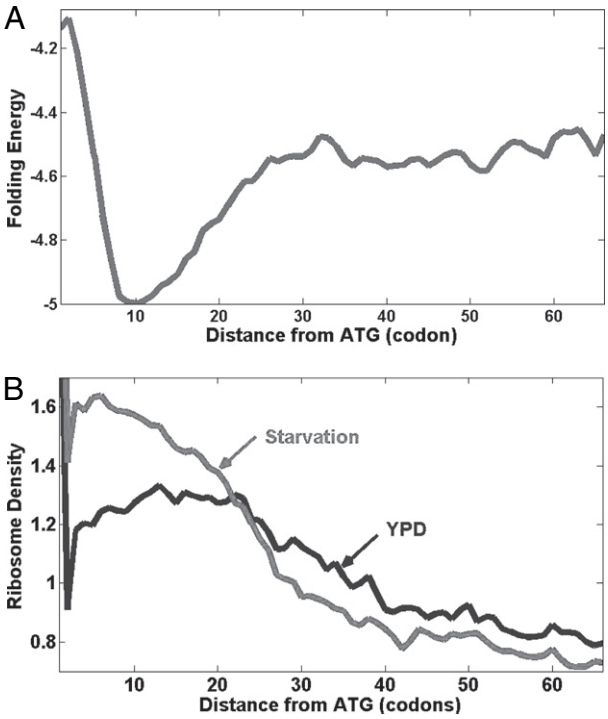

Fig. 4. Profile of folding energy (*A*) explains the profiles of ribosome density in starvation and YPD (*B*).

inverse relations indicate that lower folding energies (which correspond to more elaborate mRNA structures) slow down the velocity of ribosomal movement on mRNA, because under the assumption of a constant flux of ribosomes, the density of ribosomes is higher for lower ribosome velocity. This result suggests that folding energy influences the rate of translation elongation (and not only translation initiation). Thus, it further demonstrates how folding energy plays a part in determining global translation efficiency.

Codon Bias Better Explains Translation Efficiency and Protein Abundance Changes Across Species than Folding Energy. Finally, we studied the influence of folding energy and codon bias on protein abundance and translation efficiency from an evolutionary standpoint. If folding energy or codon bias is a central determinant of translation efficiency (local and global) in endogenous genes, one would expect evolutionary forces to act to shape their levels according to the desired level of translation efficiency. To this end, we ranked the folding energy, codon bias, protein abundance, and local translation efficiency of endogenous genes in each of the two yeast species, *S. cerevisiae* and *S. pombe* (for which genome-wide protein abundance and mRNA data are available). Next, we measured the correlation between the change in the folding energy rank of ortholog genes between the two species and the corresponding change in their protein abundance and local translation efficiency ranks, finding it to be nonsignificant or of borderline significance (for protein abundance: $r = 0.0079$, $P = 0.8204$; for local translation efficiency: $r = 0.076$, $P = 0.032$; based on 873 gene pairs). However, a similar analysis, when performed for delineating the effects of codon bias, reveals a significant correlation between the tAI and protein abundance and local translation efficiency changes across these species (for protein abundance: $r = 0.2257$, $P = 1.5 \times 10^{-11}$; for translation efficiency: $r = 0.115$, $P = 0.001$; $n = 873$; see Dataset S1 for rankings of the orthologs and additional information). Thus, also from an evolutionary viewpoint, codon bias better explains protein abundance and local translation efficiency changes than folding energy. These results again show that the selection for weak mRNA folding at the beginning of genes is global and is not related to changes across evolution of the genes' expression levels. On the other hand, codon bias does change across evolution in accordance with the changes occurring in gene expression levels.

Discussion and Conclusions

In the current study, we analyze the role of codon bias (in terms of coadaptation of the tRNA pool, the tAI measure) and folding energy in translational processes on a genome scale. We find that there is a global selection for nonfolding structures at the beginning of *E. coli* and *S. cerevisiae* genes (compared with the other parts of the coding sequences). This selection probably acts to allow faster binding of ribosomes to the transcript so as to initiate translation. In addition, in *S. cerevisiae*, the genomic spatial distribution of folding energy can explain the global spatial distribution of ribosomes reported (18). Thus, folding energy affects not only translation initiation but elongation speed.

When comparing between codon bias and folding energy as determinants of translation efficiency, we find the former to be more correlative with gene expression. In the case of local translation efficiency, we observe a correlation between codon bias and protein-to-mRNA level ratio, whereas a similar analysis for folding energy reveals no correlation. On a more refined level, however, when grouping the genes according to their folding energy levels, the strength of association between codon bias and local translation efficiency is dependent on the levels of folding energy. Finally, from an evolutionary standpoint, we again find that codon bias better correlates with changes across yeast species in protein abundance and protein-to-mRNA ratios than folding energy.

Our results suggest that there is selection for structures with weak folding at the beginning of genes; this selection, however, is global and not related to protein abundance or mRNA levels of genes; hence, it cannot be detected by the conventional measure of correlation with gene expression. Under the constraints of the

global selection for weak folding observed at the beginning of genes in *E. coli* and *S. cerevisiae* transcriptomes, codon bias, rather than folding energy, is the rate-limiting factor in the translation process of individual genes.

These results seem to contradict those reported recently by Kudla et al. (11) regarding the fact that there is no correlation between protein abundance and codon bias in the artificial GFP gene. To explain these differences, we compared the data of Kudla et al. (11) with endogenous genes in *E. coli*. First, we found that the folding energy values in the artificial GFP gene are significantly lower than those of endogenous genes [mean: −8.1 vs. −5, respectively; 0.95 standard deviations (STDs) from the mean folding energy of endogenous genes; P value $= 3.5 \times 10^{-27}$, Wilcoxon rank sum test; $n1 = 148, n2 = 4{,}226$]. Second, we found that the partial correlation between codon bias and protein abundance given folding energy is significant [r(Local Translation Efficiency, Codon Bias|Folding Energy) $= 0.17, P = 0.04; n = 148$]. Finally, a detailed analysis of the correlations between codon bias, folding energy, and protein abundance across five different folding energy bins of the GFP data of Kudla et al. (11) reveals that, indeed, significant correlations between codon bias and protein abundance and nonsignificant correlations between folding energy and protein abundance can be detected in bins having folding levels in the range detected for endogenous genes (more details and further analysis are provided in *SI Note 3*); in addition, *SI Note 4*, Fig. S5 and Table S3 include specific examples from the literature in which synonymous changes (rather than folding energy) affect translation efficiency.

Thus, the differences between the findings of this global analysis and those of Kudla et al. (11) suggest that repeating the experiment of Kudla et al. (11) with a protein encoded by mRNA with higher levels of folding energy (weaker folding at the beginning) is likely to demonstrate a much stronger relation between codon bias and protein abundance than reported, as we find for both *E. coli* and *S. cerevisiae* transcriptomes. More generally, repeating the experiment of Kudla et al. (11) with different genes is likely to demonstrate different levels of correlation between translation efficiency and folding energy or codon bias. Interestingly, a recent (small-scale) study by Welch et al. (19) did not find a correlation between translation efficiency and folding energy in two endogenous *E. coli* genes, but in the same token, did also not find such a correlation with codon bias (though it did find a strong correlation between synonymous codon changes and protein levels). This probably indicates that there are still many open issues that need to be further studied to elucidate the determinants of translation efficiency.

Materials and Methods

Protein Abundance and mRNA Levels. Protein abundance values and mRNA measurements of *E. coli* were taken from the work of Lu et al. (20); protein abundance values and mRNA levels of *S. cerevisiae* were taken from the work of Newman et al. (21) and Wang et al. (22), respectively; and protein abundance and mRNA values of *S. pombe* vs. *S. cerevisiae* were taken from the work of Schmidt et al. (23). We analyzed organisms whose large-scale gene expression of protein abundance and mRNA levels are available. Other recent data on protein abundance either include relatively small number of measurements (24) or do not include corresponding measurements of mRNA levels (25).

Profiles of Ribosome Density. Profiles of ribosome density at a resolution of single nucleotides in *S. cerevisiae* were downloaded from the work of Ingolia et al. (18).

Coding Sequences. Coding sequences of the fungi were taken from the work of Man and Pilpel (26), and the coding sequences of *E. coli* were downloaded from the National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/ftp/) in August 2008.

5′-UTR Sequences. Forty nucleotides of 5′-UTR sequences near the beginning of the ORF of *S. cerevisiae* and *E. coli* and 40 nucleotides of 3′-UTR sequences near the end of the ORF of *S. cerevisiae* and *E. coli* were taken from the NCBI (http://www.ncbi.nlm.nih.gov/ftp/) in November 2009.

Position in *E. coli* Operons. Data about the order of *E. coli* genes in operons were downloaded from the work of Gama-Castro et al. (27). The folding profile of groups of genes that are present in the end of operons, monocistronic genes, was compared with the folding profile of genes in the beginning or middle of an operon.

tAI. The tAI was computed following the work of dos Reis et al. (15), which defined this measure. This measure gauges the availability of tRNAs for each codon. Because codon–anticodon coupling is not unique as a result of wobble interactions, several anticodons can recognize the same codon, with different efficiency weights [see the article by dos Reis et al. (15) for all the relations between codons and anticodons].

Let n_i be the number of tRNA isoacceptors recognizing codon i. Let *tCGNij* be the copy number of the jth tRNA that recognizes the ith codon, and let S_{ij} be the selective constraint on the efficiency of the codon–anticodon coupling. We define the absolute adaptiveness, W_i, for each codon i as follows:

$$Wi = \sum_{j=1}^{ni} (1 - Sij) tCGNij$$

From W_i, we obtain w_i, which is the relative adaptiveness value of codon i by normalizing the values of W_i (dividing them by the maximal of all 61 W_is).

The final tAI of a gene, g, is the geometric mean of all its codons

$$tAIg = \left(\prod_{k=1}^{lg} wikg \right)^{1/lg},$$

where i_{kg} is the codon defined by the kth triplet on gene g and lg is the length of the gene (excluding stop codons).

For tAI calculation, tRNA copy numbers of the two fungi were downloaded from the work of Man and Pilpel (26). tRNA copy numbers of *E. coli* were downloaded in November 2008 from the Genomic tRNA Database (http://lowelab.ucsc.edu/GtRNAdb/) (28); tRNA copy numbers of all organisms analyzed in this study appear in Table S1.

The S_{ij} values can be organized in a vector (S-vector) as described by dos Reis et al. (15); each component in this vector is related to one wobble nucleoside–nucleoside pairing (e.g., I:U, G:U, G:C, I:C, U:A, I:A). The w_i values for all codons (except stop codons) of all organisms analyzed in this study appear in Table S2.

Ortholog Mapping. For comparing orthologs of *S. pombe* and. *S. cerevisiae*, we used the ortholog mapping technique of Lu et al. (20).

Computing Folding Energy and Profiles of Folding Energy. Folding energy was calculated by UNAfold software (16) for windows of 40 nucleotides along the genes' sequences. Let FE_i denote the folding energy of a window of 40-nt length, starting from the ith nucleotide of the gene.

The local profile of a gene was defined as the vector of the folding energy, *FE*, values assigned to the sliding windows of 40-nt length of the gene codons

$$Local\_FE_{Gene_i} = (FE_1, FE_2, \ldots, FE_n)$$

For a particular species, all the genes in the genome were lined up once according to their start codon and once according to their stop codon. The profiles (start and end) of mean *FE* were calculated as

$$\begin{aligned} \overline{Local\_FE_{start}} &= (\overline{FE_2}, \overline{FE_3}, \overline{FE_4}, \ldots) \overline{Local\_FE_{end}} \\ &= (\overline{FE_n}, \overline{FE_{n-1}}, \overline{FE_{n-2}}, \ldots), \end{aligned}$$

where

$$\overline{FE_i} = \sum_{Genes_i} FE_i / |Genes_i|$$

and $Genes_i$ is the number of genes with at least $i + 1$ 40-nt windows.

Let $STD(v)$ denote the STD of a vector v of real numbers; the profiles (head and tail) of the STD of *FE* were calculated as follows:

$$Local\_STD\_FE_{start} = (STD(FE_1), STD(FE_2), STD(FE_3), \ldots)$$
$$Local\_STD\_FE_{end} = (STD(FE_n), STD(FE_{n-1}), STD(FE_{n-2}), \ldots),$$

where

$STD(FE_i)$ is the STD for the vector that includes the *FE* of the *i*th window of all the genes with at least *I* + *1* 40-nt sliding windows

Randomized Profiles of Folding Energy. To show that the profiles of folding energy (weaker folding energy at the beginning of ORF) are selected for, we compared the genomic profile of folding energy with a profile of folding energy observed for a randomization of the genome. The genome was randomized in the following way. Each codon was replaced by a random codon, according to the distribution (frequency) of codons coding the same amino acid in the genome of the organism. Thus, the randomized genomes maintained both the amino acid content of each coding sequence and the codon frequencies of the original genome. We compared the mean of 10 randomized profiles with the original profile.

Correlations and *P* Values. All the correlations reported are the nonparametric Spearman correlation; *P* values were computed by the nonparametric Wilcoxon test. In the case of the comparison of the mean of the randomized profile energy with the original profile, we performed a Kolmogorov–Smirnov test (Wilcoxon test yields similar results) for each window index to compare the values of the folding energy of genes in the original genome with the mean folding energy of genes in the randomized genomes.

ACKNOWLEDGMENTS. We thank Prof. Plotkin for providing us with the protein abundance measurements for different synonymous mutations of the GFP protein and for helpful discussions. T.T. is a Koshland Scholar at Weizmann Institute of Science. Y.Y.W. was supported in part by a fellowship from the Edmond J. Safra Bioinformatics program at Tel-Aviv University. M.K. was supported by grants from the Israel Science Foundation and the United States–Israel Binational Fund. E.R. was supported by grants from the Israel Science Foundation and the European Union Pathogenomics consortium.

1. Zuckerkandl E, Pauling L (1965) Molecules as documents of evolutionary history. *J Theor Biol* 8:357–366.
2. Ikemura T (1985) Codon usage and tRNA content in unicellular and multicellular organisms. *Mol Biol Evol* 2:13–34.
3. Ikemura T (1981) Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes. *J Mol Biol* 146: 1–21.
4. Parmley JL, Hurst LD (2007) How do synonymous mutations affect fitness? *BioEssays* 29:515–519.
5. Nackley AG, et al. (2006) Human catechol-O-methyltransferase haplotypes modulate protein expression by altering mRNA secondary structure. *Science* 314:1930–1933.
6. Boycheva S, Chkodrov G, Ivanov I (2003) Codon pairs in the genome of Escherichia coli. *Bioinformatics* 19:987–998.
7. Coleman JR, et al. (2008) Virus attenuation by genome-scale changes in codon pair bias. *Science* 320:1784–1787.
8. Gustafsson C, Govindarajan S, Minshull J (2004) Codon bias and heterologous protein expression. *Trends Biotechnol* 22:346–353.
9. Hall MN, Gabay J, Débarbouillé M, Schwartz M (1982) A role for mRNA secondary structure in the control of translation initiation. *Nature* 295:616–618.
10. Andersson SG, Kurland CG (1990) Codon preferences in free-living microorganisms. *Microbiol Rev* 54:198–210.
11. Kudla G, Murray AW, Tollervey D, Plotkin JB (2009) Coding-sequence determinants of gene expression in Escherichia coli. *Science* 324:255–258.
12. Sharp PM, Li WH (1987) The Codon Adaptation Index—A measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* 15: 1281–1295.
13. Ghaemmaghami S, et al. (2003) Global analysis of protein expression in yeast. *Nature* 425:737–741.
14. Tuller T, Kupiec M, Ruppin E (2007) Determinants of protein abundance and translation efficiency in S. cerevisiae. *PLoS Comput Biol* 3:2510–2519.
15. dos Reis M, Savva R, Wernisch L (2004) Solving the riddle of codon usage preferences: A test for translational selection. *Nucleic Acids Res* 32:5036–5044.
16. Markham NR, Zuker M (2005) DINAMelt web server for nucleic acid melting prediction. *Nucleic Acids Res* 33 (Web Server issue):W577–W581.
17. Katz L, Burge CB (2003) Widespread selection for local RNA secondary structure in coding regions of bacterial genes. *Genome Res* 13:2042–2051.
18. Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* 324:218–223.
19. Welch M, et al. (2009) Design parameters to control synthetic gene expression in Escherichia coli. *PLoS One* 4:1–10.
20. Lu P, Vogel C, Wang R, Yao X, Marcotte EM (2007) Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat Biotechnol* 25:117–124.
21. Newman JR, et al. (2006) Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. *Nature* 441:840–846.
22. Wang Y, et al. (2002) Precision and functional specificity in mRNA decay. *Proc Natl Acad Sci USA* 99:5860–5865.
23. Schmidt MW, Houseman A, Ivanov AR, Wolf DA (2007) Comparative proteomic and transcriptomic profiling of the fission yeast Schizosaccharomyces pombe. *Mol Syst Biol* 3:1–12.
24. Picotti P, Bodenmiller B, Mueller LN, Domon B, Aebersold R (2009) Full dynamic range proteome analysis of S. cerevisiae by targeted proteomics. *Cell* 138:795–806.
25. Malmström J, et al. (2009) Proteome-wide cellular protein concentrations of the human pathogen Leptospira interrogans. *Nature* 460:762–765.
26. Man O, Pilpel Y (2007) Differential translation efficiency of orthologous genes is involved in phenotypic divergence of yeast species. *Nat Genet* 39:415–421.
27. Gama-Castro S, et al. (2008) RegulonDB (version 6.0): Gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic Acids Res* 36 (Database issue): D120–D124.
28. Lowe TM, Eddy SR (1997) tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* 25:955–964.

Supporting Information

Tuller et al. 10.1073/pnas.0909910107

SI Note 1: Folding Energy vs. Fitness

The reported results (Fig. 1, Figs. S1 and S2, and *SI Note 2*) suggest that very low folding energy at the beginning or end of genes should affect the fitness of the analyzed organism. Indeed, when we returned to the data of Kudla et al. (1) and divided the genes into two groups [genes whose folding energy at the beginning was below −10 (hardly seen in *E. coli* endogenous genes) and genes with folding energy above −10], we found the fitness (measured by OD) of the second group was significantly higher than that of the first group (mean OD = 0.71 vs. mean OD = 0.76, $n = 148$; Wilcoxon test, $P = 0.01$). Similarly, when controlling for the codon bias, the correlation between OD and folding energy is significant [r(Folding Energy, OD|Codon Bias) = 0.17, $P = 0.03$]. The data of Kudla et al. (1) do not include genes with very low folding energy at the ends. Thus, we could not perform a similar analysis for the folding energy at the end of genes.

SI Note 2: Folding Energy at the End of Genes

Interestingly, we find that in addition to the first mRNA window, the folding energy at the terminal windows was significantly higher than that of adjacent windows, suggesting that constraints may apply to translation termination (e.g., −6.5 for the last 40 nucleotides vs. −7.4 for the 80–40 nucleotides before the end, $n = 4{,}226$; Wilcoxon test, $P < 10^{-16}$; Fig. S1). This pattern was similar for genes that are located at the end of an operon and for genes that are not located at the end of an operon (Fig. S3).

Similar results were obtained for the *S. cerevisiae* transcriptome [the folding energy at the terminal windows was significantly higher than that of adjacent windows (e.g., −4.285 for the last 40 nucleotides vs. −4.79 for the 80–40 nucleotides before the end, $n = 5{,}869$; Wilcoxon test: $P < 10^{-16}$), and the mean folding energy and variance (in absolute values) are lower at the terminal of the genes] (Fig. S2). However, when we compared the folding energy at the terminal end of the original profile with the mean folding energy of a randomized profile, we did not find a significant difference in *E. coli* and *S. cerevisiae* (in the case of *E. coli*, in all the window indexes that are more than 8 nucleotides from the end of the ORF, the folding energy of the original genome was significantly lower than the folding energy of the randomized genomes; in the case of *S. cerevisiae*, 180 of the window indexes exhibited significantly lower folding energy than the randomized genomes).

SI Note S3: Folding Energy vs. Codon Bias in Artificial Genes

The results reported here describe the relations between codon bias, folding energy, protein abundance, and translation efficiency that have been shaped by selection forces. In the case of artificial genes (1), there may be a significant correlation between local translation efficiency and folding energy rather than between local translation efficiency and codon bias. This discrepancy can be explained using the findings reported in this paper.

In the case of the artificial genes (1), we find that the folding energy values are significantly lower than those of endogenous genes (mean: −8.1 vs. −5, respectively; 0.95 STDs from the mean folding energy of endogenous genes; P value $= 3.5 \times 10^{-27}$, Wilcoxon rank sum test; $n1 = 148$, $n2 = 4{,}226$; Fig. S5*A*). Specifically, the bin with lower folding energy (where the codon bias and local translation efficiency correlation was the strongest) is absent from the set, whereas it includes a "tail" with relatively low folding energy absent from endogenous genes (folding energy less than −10). The dataset of Kudla et al. (1) hence does not represent accurately the folding energy of endogenous genes, and its comparatively low folding energy values (implying tighter folding) may mask the effects of other factors such as codon bias acting in a natural environment. Remarkably, we find that although codon bias does not significantly correlate with protein abundance in this dataset ($r = 0.03$, $P = 0.7$; $n = 148$), the partial correlation between codon bias and protein abundance given folding energy is significant [r(Local Translation Efficiency, Codon Bias|;Folding Energy) = 0.17, $P = 0.04$; $n = 148$]. We should also note that in the GFP dataset, the partial correlation between folding energy and protein abundance given codon bias is significant [r(Local Translation Efficiency, Folding Energy|Codon Bias) = 0.66, P value $= 4.32 \times 10^{-20}$; $n = 148$], suggesting that in the case of the data of Kudla et al. (1), both folding energy and codon bias have a significant effect on the protein abundance of the GFP gene. A detailed analysis of the correlations between codon bias, folding energy, and protein abundance across five different folding energy bins of the GFP data of Kudla et al. (1) reveals that, indeed, significant correlations between codon bias and protein abundance and nonsignificant correlations between folding energy and protein abundance can be detected in bins having folding levels in the range detected for endogenous genes (Fig. S5 *B–F*). Thus, even for the dataset of Kudla et al. (1), the codon bias does influence translation efficiency (*SI Note 4* and Table S3). As opposed to codon bias, the strongest folding energy and protein abundance correlation is observed for a relatively low folding energy bin ($r = 0.58$, $P = 0.0013$; $n = 29$; Fig. S5*D*), where the mRNA folding is very strong. However, such a folding energy range is rarely seen in endogenous genes (Fig. S5*A*).

SI Note S4: Various Examples in Which Codon Bias Can Explain Changes in Protein Abundance in *E. coli* Genes

As additional support to the role of codon bias (rather than mRNA folding energy) in translation efficiency, we report here 13 examples of *E. coli*. genes whose sequence was changed; as a result, their protein abundance was also changed (Table S3). The examples are taken from the literature (2–4). In all these examples, the folding energy of the first 40 nucleotides did not change or the change in general folding energy was not in the expected direction (i.e., the protein abundance increased, but the folding energy decreased or did not change, or vice versa), whereas the change in codon bias was in the expected direction.

Importantly, in these cases (2, 4), increasing the tRNA levels of rare codons also led to an increase in protein levels, an additional control that further implies a direct connection between protein abundance and codon bias via tRNA/codon coadaptation.

Another interesting result is the analysis of Sørensen and Pedersen (5), who measured the translation rate of two glutamate codons: GAA and GAG. They found them to have a 3-fold difference in translation rate (21.6 and 6.4 codons per second, respectively). Remarkably, the w_i of these codons, which is based on the tRNA pool and affinity of codon–anticodon coupling and is the basis for the tAI calculation, captures the ratio of the translation rate between the two codons. Calculating w_i values for *E. coli*, we found that the ratio between the w_i of GAA and GAG is 3.125 (0.5/0.16) as compared with the ratio of 3.34 reported in the experiments (21.4/6.4). This result suggests that there is a direct relation between the adaptation of a codon to the tRNA pool and the time it takes to translate it, further supporting the connection between codon bias and local translation efficiency.

1. Kudla G, Murray AW, Tollervey D, Plotkin JB (2009) Coding-sequence determinants of gene expression in Escherichia coli. *Science* 324:255–258.
2. Rosenberg AH, Goldman E, Dunn JJ, Studier FW, Zubay G (1993) Effects of consecutive AGG codons on translation in Escherichia coli, demonstrated with a versatile codon test system. *J Bacteriol* 175:716–722.
3. Gonzalez de Valdivia EI, Isaksson LA (2004) A codon window in mRNA downstream of the initiation codon where NGG codons give strongly reduced gene expression in Escherichia coli. *Nucleic Acids Res* 32:5198–5205.
4. Burgess-Brown NA, et al. (2008) Codon optimization can improve expression of human genes in Escherichia coli: A multi-gene study. *Protein Expression Purif* 59:94–102.
5. Sørensen MA, Pedersen S (1991) Absolute in vivo translation rates of individual codons in Escherichia coli. The two glutamic acid codons GAA and GAG are translated with a threefold difference in rate. *J Mol Biol* 222:265–280.

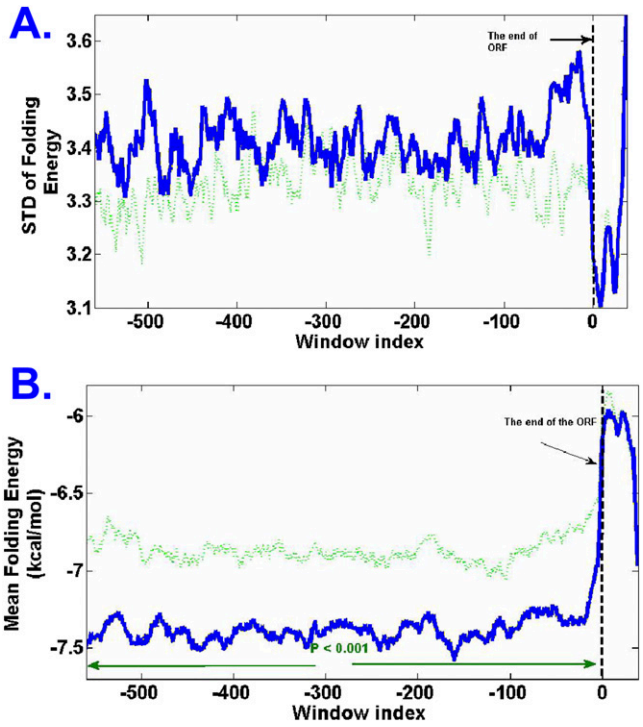

Fig. S1. Folding energy [mean (*A*) and STD (*B*)] at the end of genes in *E. coli*. The randomized profile (green) and original profile (blue) are shown.

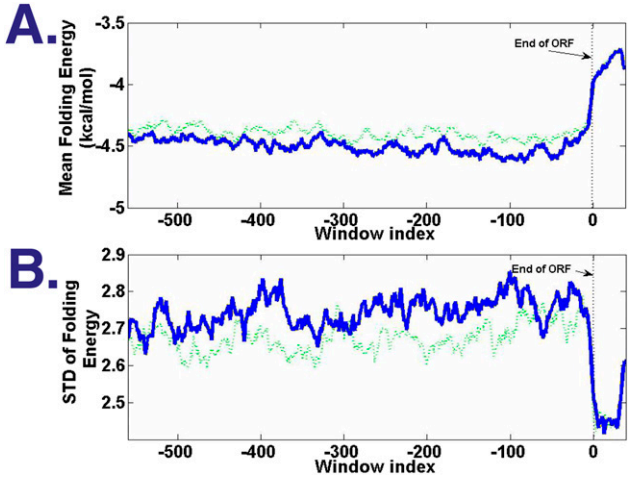

Fig. S2. Folding energy [mean (*A*) and STD (*B*)] at the end of genes in *S. cerevisiae*. The randomized profile (green) and original profile (blue) are shown.

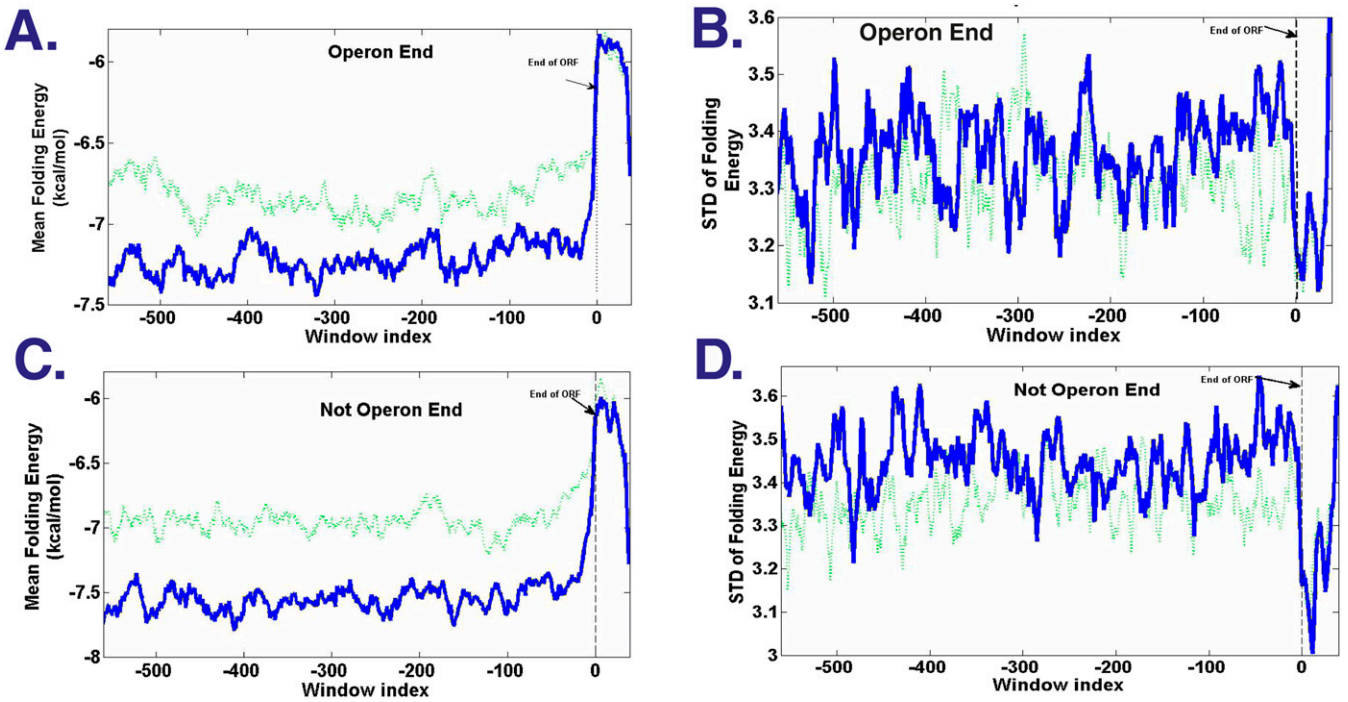

Fig. S3. Folding energy [mean (*A* and *C*) and STD (*B* and *D*)] at the end of genes in *E. coli* for genes that are at the end of an operon (*A* and *B*) and for genes that are not at the end of an operon (*C* and *D*).

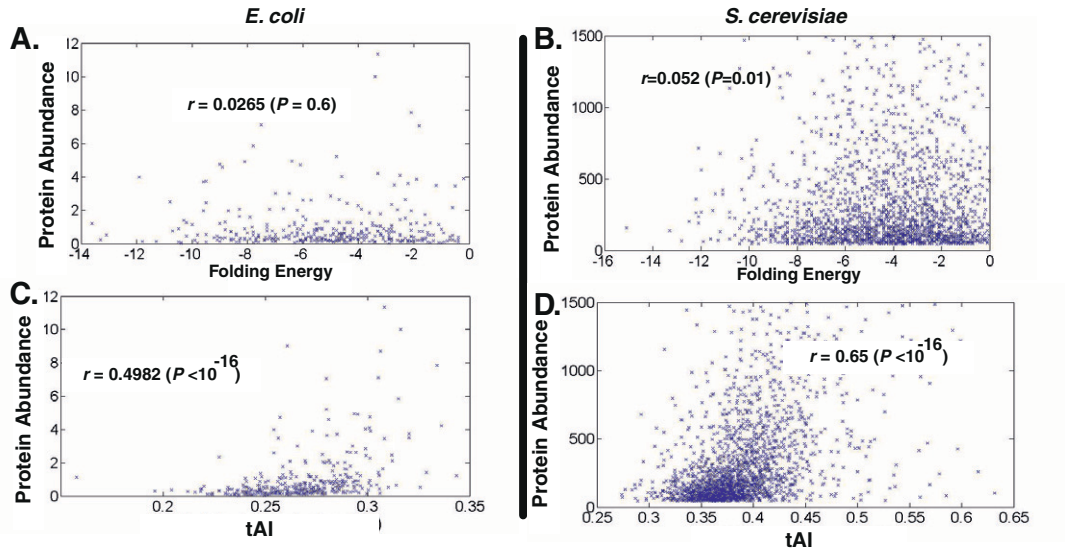

Fig. S4. Codon bias (tAI) and folding energy vs. protein abundance: Codon bias significantly correlates with protein abundance ($r = 0.4982$, $P < 10^{-16}$ and $r = 0.65$, $P < 10^{-16}$ for *E. coli* and *S. cerevisiae*, respectively), whereas we find nonsignificant or very low correlation between folding energy and protein abundance ($r = -0.0265$, $P = 0.6$ and $r = 0.0521$, $P = 0.01$ for *E. coli* and *S. cerevisiae*, respectively). (*A, C*) *E. coli* folding energy vs. protein abundance and codon bias vs. protein abundance. (*B, D*) *S. cerevisiae* folding energy vs. protein abundance and codon bias vs. protein abundance.

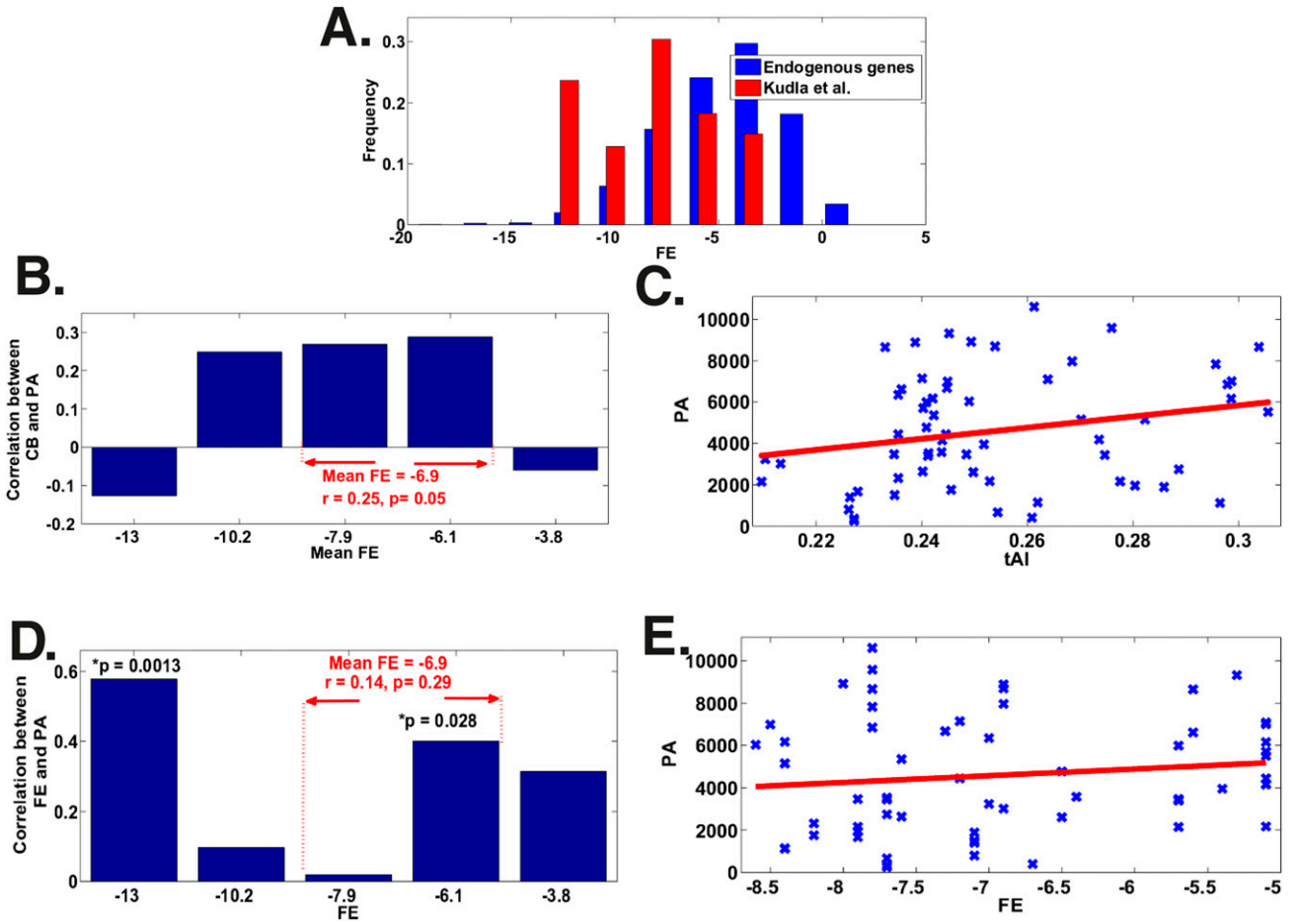

Fig. S5. Analysis of the data of Kudla et al. [Kudla G, Murray AW, Tollervey D, Plotkin JB (2009) Coding-sequence determinants of gene expression in Escherichia coli. *Science* 324:255–258]. (*A*) Folding energy distribution at the beginning of endogenous genes and in the dataset of Kudla et al. (*B*) Codon bias (tAI) and protein abundance correlation (*y* axis) for five equal-sized bins according to folding energy values (*x* axis). When folding energy is in the midrange, the tAI/protein abundance correlation is significant (three middle bins: $r = 0.3$, $P = 0.0041$, $n = 90$; two middle bins marked in the figure: $r = 0.25$, $P = 0.05$, $n = 60$). (*C*) Scatter plot of protein abundance vs. tAI for the two marked intermediate bins. (*D*) Folding energy and protein abundance correlation (*y* axis) for five equal-sized bins according to folding energy values in the beginning of the sequence (*x* axis; the same five equal-sized bins as in *B*). When the folding energy is very low, the correlation between protein abundance and folding energy is most significant [$r = 0.58$, $n = 29$ ($P = 0.0013$; the last bin)]. (*E*) Scatter plot of protein abundance vs. folding energy for the two marked intermediate bins (−6.9 mean folding energy).

Table S1. tRNA copy numbers of the analyzed organisms

Codon	*S. cerevisiae*	*S. pombe*	*E. coli*
TTT	0	0	0
TTC	10	5	2
TTA	7	2	1
TTG	10	4	1
TCT	11	7	0
TCC	0	0	2
TCA	3	2	1
TCG	1	1	1
TAT	0	0	0
TAC	8	4	3
TAA	0	0	0
TAG	0	0	0
TGT	0	0	0
TGC	4	3	1
TGA	0	0	1
TGG	6	3	1
CTT	0	5	0
CTC	1	0	1
CTA	3	1	1
CTG	0	1	4
CCT	2	6	0
CCC	0	0	1
CCA	10	2	1
CCG	0	1	1
CAT	0	0	0
CAC	7	4	1
CAA	9	4	2
CAG	1	2	2
CGT	6	8	4
CGC	0	0	0
CGA	0	1	0
CGG	1	1	1
ATT	13	8	0
ATC	0	0	3
ATA	2	1	0
ATG	10	7	8
ACT	11	7	0
ACC	0	0	2
ACA	4	2	1
ACG	1	1	2
AAT	0	0	0
AAC	10	6	4
AAA	7	3	6
AAG	14	9	0
AGT	0	0	0
AGC	2	3	1
AGA	11	2	1
AGG	1	1	1
GTT	14	9	0
GTC	0	0	2
GTA	2	2	5
GTG	2	1	0
GCT	11	9	0
GCC	0	0	2
GCA	5	2	3
GCG	0	1	0
GAT	0	0	0
GAC	16	8	3
GAA	14	4	4
GAG	2	6	0
GGT	0	0	0
GGC	16	8	4
GGA	3	3	1
GGG	2	1	1

Copy numbers are shown according to the corresponding (perfectly matched) codons.

Table S2. w_i values (for tAI calculation) for codons of the analyzed organisms

Codon	*S. cerevisiae*	*S. pombe*	*E. coli*
TTT	0.27032	0.220382	0.10975
TTC	0.615764	0.502008	0.25
TTA	0.431034	0.200803	0.125
TTG	0.753695	0.465863	0.165
TCT	0.67734	0.702811	0.10975
TCC	0.487685	0.506024	0.25
TCA	0.184797	0.200873	0.125
TCG	0.12069	0.164659	0.165
TAT	0.216256	0.176305	0.164625
TAC	0.492611	0.401606	0.375
TAA	NaN	NaN	NaN
TAG	NaN	NaN	NaN
TGT	0.108128	0.132229	0.054875
TGC	0.246305	0.301205	0.125
TGA	NaN	NaN	NaN
TGG	0.369458	0.301205	0.165
CTT	0.027032	0.502008	0.054875
CTC	0.061576	0.361446	0.125
CTA	0.184729	0.100452	0.125
CTG	0.059113	0.13253	0.54
CCT	0.123153	0.60241	0.054875
CCC	0.08867	0.433735	0.125
CCA	0.615776	0.200863	0.125
CCG	0.197044	0.164659	0.165
CAT	0.189224	0.176305	0.054875
CAC	0.431034	0.401606	0.125
CAA	0.554187	0.401606	0.25
CAG	0.238916	0.329317	0.33
CGT	0.369458	0.803213	0.5
CGC	0.26601	0.578313	0.36
CGA	0.000037	0.100482	0.00005
CGG	0.061576	0.13253	0.125
ATT	0.800493	0.803213	0.164625
ATC	0.576355	0.578313	0.375
ATA	0.123233	0.100482	0.163204
ATG	0.615764	0.702811	1
ACT	0.67734	0.702811	0.10975
ACC	0.487685	0.506024	0.25
ACA	0.246373	0.200873	0.125
ACG	0.140394	0.164659	0.29
AAT	0.27032	0.264458	0.2195
AAC	0.615764	0.60241	0.5
AAA	0.431034	0.301205	0.75
AAG	1	1	0.24
AGT	0.054064	0.132229	0.054875
AGC	0.123153	0.301205	0.125
AGA	0.67734	0.200803	0.125
AGG	0.278325	0.164659	0.165
GTT	0.862069	0.903614	0.10975
GTC	0.62069	0.650602	0.25
GTA	0.123239	0.200894	0.625
GTG	0.162562	0.164659	0.2
GCT	0.67734	0.903614	0.10975
GCC	0.487685	0.650602	0.25
GCA	0.30795	0.200894	0.375
GCG	0.098522	0.164659	0.12
GAT	0.432512	0.35261	0.164625
GAC	0.985222	0.803213	0.375
GAA	0.862069	0.401606	0.5
GAG	0.399015	0.730924	0.16
GGT	0.432512	0.35261	0.2195
GGC	0.985222	0.803213	0.5
GGA	0.184729	0.301205	0.125
GGG	0.182266	0.196787	0.165

NaN values appear for stop codons (excluded from this analysis).

Table S3. Relation between codon bias or folding energy and protein abundance in experiments of synonymous manipulation of codons

Example no.	Example description	Protein abundance increased (+)/decreased (–)	Folding energy change (whole gene)	tAI change (whole gene)
1	Rosenberg, et al. [1]T7 gene 9; add five AGG codons after codon 13	–	0 (54 nucleotides)	–0.0274/0.0484 = –0.5661 (54 nucleotides)
2	Burgess-Brown, et al. [2]Synthetic vs. native DHRS1	+	–21.59/35 = –0.6169	0.0438/0.0236 = 1.8559
3	Burgess-Brown, et al. [2]Gene synthetic vs. native DHRS4	+	–17.88/35 = –0.5109	0.0597/0.0236 = 2.5297
4	Burgess-Brown, et al. [2]Synthetic vs. native GMDS	+	–13.59/35 = –0.3883	0.0723/0.0236 = 3.0636
5	Burgess-Brown, et al. [2]Synthetic vs. native HADH2	+	–10.7/35 = –0.3057	0.0531/0.0236 = 2.2500
6	Burgess-Brown, et al. [2]Synthetic vs. native HPGD.	+	–28.8/35 = –0.8229	0.0647/0.0236 = 2.7415
7	Burgess-Brown, et al. [2]Synthetic vs. native HSD17B2	+	–27.4/35 = –0.7829	0.0680/0.0236 = 2.8814
8	Burgess-Brown, et al. [2] Synthetic vs. native HSD17B4	+	–68.95/35 = –1.97	0.0859/0.0236 = 3.6398
9	Burgess-Brown, et al. [2]Synthetic vs. native; MAT2B	+	–22.14/35 = –0.6326	0.0633/0.0236 = 2.6822
10	Burgess-Brown, et al. [2]Synthetic vs. native RDH5	+	–16.69/35 = –0.4769	0.0430/0.0236 = 1.8220
11	Burgess-Brown, et al. [2]Synthetic vs. native RETSDR2	+	–38.12/35 = –1.0891	0.0684/0.0236 = 2.8983
12	Burgess-Brown, et al. [2]Synthetic vs. native RETSDR4	+	–41.74/35 = –1.1926	0.0994/0.0236 = 4.2119
13	Burgess-Brown, et al. [2]Synthetic vs. native TGDS	+	–55.87/35 = –1.5963	0.0823/0.0236 = 3.4873

The change in codon bias and folding energy are normalized by the STD of these values on *E. coli* sequences of the same length. For all the cases that appear in this table, the first 40 nucleotides, and therefore their tAI and folding energy values, were not changed. Thus, we considered the effects on the entire sequence.

1. Rosenberg AH, Goldman E, Dunn JJ, Studier FW, Zubay G (1993) Effects of consecutive AGG codons on translation in Escherichia coli, demonstrated with a versatile codon test system. *J Bacteriol* 175:716–722.
2. Burgess-Brown NA, et al. (2008) Codon optimization can improve expression of human genes in Escherichia coli: A multi-gene study. *Protein Expression Purif* 59:94–102.

Other Supporting Information Files

Dataset S1 (XLS)