Differential requirements for mRNA folding partially explain why highly expressed proteins evolve slowly

Chungoo Park, Xiaoshu Chen, Jian-Rong Yang, and Jianzhi Zhang[1]

Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI 48109

Edited by Wen-Hsiung Li, University of Chicago, Chicago, IL, and approved January 10, 2013 (received for review October 18, 2012)

The cause of the tremendous among-protein variation in the rate of sequence evolution is a central subject of molecular evolution. Expression level has been identified as a leading determinant of this variation among genes encoded in the same genome, but the underlying mechanisms are not fully understood. We here propose and demonstrate that a requirement for stronger folding of more abundant mRNAs results in slower evolution of more highly expressed genes and proteins. Specifically, we show that: (*i*) the higher the expression level of a gene, the greater the selective pressure for its mRNA to fold; (*ii*) random mutations are more likely to decrease mRNA folding when occurring in highly expressed genes than in lowly expressed genes; and (*iii*) amino acid substitution rate is negatively correlated with mRNA folding strength, with or without the control of expression level. Furthermore, synonymous (d_S) and nonsynonymous (d_N) nucleotide substitution rates are both negatively correlated with mRNA folding strength. However, counterintuitively, d_S and d_N are differentially constrained by selection for mRNA folding, resulting in a significant correlation between mRNA folding strength and d_N/d_S, even when gene expression level is controlled. The direction and magnitude of this correlation is determined primarily by the G+C frequency at third codon positions. Together, these findings explain why highly expressed genes evolve slowly, demonstrate a major role of natural selection at the mRNA level in constraining protein evolution, and reveal a previously unrecognized and unexpected form of nonprotein-level selection that impacts d_N/d_S.

messenger RNA | nonsynonymous substitution rate | synonymous substitution rate

It has been known since the 1960s that different proteins can have drastically different rates of sequence evolution (1). Understanding the cause of this variation has always been a central topic of molecular evolution because it informs us about evolutionary mechanisms (2–8). Although the neutralist's explanation that the rate of protein sequence evolution is determined jointly by mutation rate and functional constraint (4) remains largely valid in principle, the exact meaning and ingredients of "functional constraint" have been elusive. Recent genomic studies, typically examining hundreds to thousands of genes encoded in a genome, identified a number of gene properties that are correlated with the rate of protein sequence evolution (9–21). Among these properties, gene expression level exhibits the strongest correlation at least in bacteria and yeast (10, 11, 14, 18, 20). The cause of the negative correlation between the expression level of a protein and its evolutionary rate, termed E-R anticorrelation, is however enigmatic. For example, one might think that highly expressed proteins evolve slowly simply because they are more important than lowly expressed ones. Nevertheless, the E-R anticorrelation is only slightly weakened by controlling gene importance estimated from the fitness effect of gene deletion (11, 22), suggesting that the correlation between gene importance and expression level is at most a minor contributor to the E-R anticorrelation.

In the last few years, three hypotheses have been proposed to explain the E-R anticorrelation. The protein misfolding avoidance hypothesis posits that natural selection against cytotoxic protein misfolding (23) is stronger for more abundant proteins, which constrains the evolution of these proteins and results in the E-R anticorrelation (18, 24). The protein misinteraction avoidance hypothesis states that natural selection against deleterious protein–protein misinteraction is stronger for more highly expressed proteins, which constrains the surfaces of these proteins and causes the E-R anticorrelation (25). Considering the cost of protein production and the benefit of protein function, the protein function hypothesis (26, 27) asserts that the average adverse fitness effect of a nonsynonymous mutation is greater when occurring in a highly expressed gene than in a lowly expressed gene, generating the E-R anticorrelation. Although the first two hypotheses have received unambiguous empirical support (18, 24, 25), the third remains to be tested. Notwithstanding, it is unclear whether the three hypotheses are sufficient to explain the E-R anticorrelation because, quantitatively, at least the first two hypotheses do not appear to account for the majority of the anticorrelation (25).

While searching for potential additional mechanisms of the E-R anticorrelation, we realize that, in theory, the anticorrelation need not be related to a protein property. The anticorrelation would arise as long as a selective pressure at any level (DNA, mRNA, or protein) is modulated by the mRNA or protein concentration. Prompted by the recent report of a strong positive correlation between mRNA concentration and mRNA folding strength (28), we here propose and demonstrate that mRNA folding strength is under more intense selection for genes of higher expression levels, resulting in stronger evolutionary constraints of more highly expressed proteins, or the E-R anticorrelation. We further show that selection for mRNA folding strength has different impacts on synonymous (d_S) and nonsynonymous (d_N) substitution rates and hence influences d_N/d_S, an index that is commonly interpreted to reflect selection acting on proteins.

Results

Selection for mRNA Folding Intensifies with the Concentration of mRNA. At physiological conditions, an mRNA molecule usually contains

Significance

The expression level of a gene is a leading determinant of its rate of protein sequence evolution, but the underlying mechanisms are unclear. We show that as the mRNA concentration increases, natural selection for mRNA folding intensifies, resulting in larger fractions of mutations deleterious to mRNA folding and lower rates of protein evolution. Counterintuitively, selection for mRNA folding also impacts the nonsynonymous-to-synonymous nucleotide substitution rate ratio, requiring a revision of the current interpretation of this ratio as a measure of protein-level selection. These findings demonstrate a prominent role of selection at the mRNA level in molecular evolution.

Author contributions: C.P. and J.Z. designed research; C.P., X.C., and J.-R.Y. performed research; C.P., X.C., and J.-R.Y. analyzed data; and C.P. and J.Z. 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: jianzhi@umich.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1218066110/-/DCSupplemental.

many intramolecular hydrogen bonds between pairs of nucleotides: three bonds between a G (guanine) and a C (cytosine), two bonds between an A (adenine) and a U (uracil), and two bonds between a G and a U. These hydrogen bonds and base stacking interactions determine the secondary structure of the mRNA molecule, which is typically composed of multiple stems (segments of paired nucleotides) and loops (segments of unpaired nucleotides). The secondary structure of an mRNA molecule can be computationally predicted with moderate accuracies (29, 30). The structures can also be experimentally determined using enzymes that respectively cut single-stranded and double-stranded RNA (31). For example, RNase V1 preferentially cleaves phosphodiester bonds 3′ of double-stranded RNA, whereas S1 nuclease preferentially cleaves 3′ of single-stranded RNA. Recently, Kertesz et al. used this strategy to determine the secondary structures of over 3,000 yeast mRNAs (32). Specifically, by SOLiD (Sequencing by Oligonucleotide Ligation and Detection) sequencing of fragments of RNAs that were respectively treated with RNase V1 and S1 nuclease, they estimated for each nucleotide the ratio between the relative probability that it is paired and the relative probability that it is unpaired. However, because of the difference in enzyme efficiency, these ratios are relative; they are meaningful only when compared among nucleotide sites. We calculated for each gene an mRNA folding score *mF*, which is the arithmetic mean of these ratios across all coding sites in the mRNA that have such information. The higher the *mF* score, the greater the fraction of paired sites in the mRNA and the stronger the overall folding of the mRNA.

Confirming a recent report (28), we observed in yeast a strong positive correlation between the RNA-Seq–based expression level of a gene (33) and its *mF* score (Spearman's rank correlation coefficient $\rho = 0.639$, $P < 10^{-100}$) (Fig. 1*A*). In theory, this correlation may be (*i*) a result of natural selection for stronger folding of more abundant mRNAs or (*ii*) a byproduct of other selections. The latter possibility should be considered, because genes of different expression levels have different amino acid, nucleotide, and synonymous codon frequencies (5, 34–36), all of which may impact the strength of mRNA folding (37).

To investigate whether the high *mF* values of highly expressed genes are simply a byproduct of their specific protein sequences, amino acid compositions, G+C frequencies, and codon frequencies, which may have been shaped by selection for protein functions, low synthetic costs (34, 35), high translational efficiencies (38), and high translational accuracies (18, 35), we should ideally compare the *mF* of each real mRNA with those of random mRNAs that have the same protein sequence and synonymous codon frequencies as the real mRNA. However, such an analysis is infeasible because the experimentally determined *mF* values are available only for the real mRNAs. We thus turn to computational prediction of mRNA folding (*Materials and Methods*). We first estimated the computationally predicted fraction of paired sites in each real mRNA (mF_p). As expected, mF_p is moderately correlated with the experimentally determined *mF* ($\rho = 0.22$, $P < 10^{-15}$). Furthermore, mF_p is significantly correlated with gene expression level ($\rho = 0.173$, $P < 10^{-15}$), although the correlation is much lower than that between *mF* and expression level ($\rho = 0.639$, $P < 10^{-100}$). For each real mRNA, we generated 100 pseudo-mRNAs by randomly shuffling synonymous codons within the real mRNA; the pseudo-mRNAs all have the same protein sequence, G+C frequencies, and synonymous codon frequencies as the real mRNA. We computationally predicted the secondary structures of these 100 pseudo-mRNAs and estimated their mean fraction of paired sites (mF'_p).

We found that in 59% of the 2,448 yeast genes examined, mF_p exceeds mF'_p, indicating that the folding strengths of the real mRNAs tend to be greater than the expectations based on the protein sequences and synonymous codon frequencies ($P < 10^{-6}$, binomial test), which is consistent with previous reports (39–41). More interestingly, we found that the higher the expression level of a gene, the greater the difference between mF_p and mF'_p ($\rho = 0.06$, $P < 0.003$) (Fig. 1*B*), suggesting that the stronger folding of more abundant mRNAs cannot be fully explained by their specific protein sequences or synonymous codon usages. In other words, highly expressed genes have been under stronger selection for mRNA folding than lowly expressed genes. To gain a rough idea of the estimation error of $mF_p - mF'_p$, for mRNA *i*, we pick mRNA *j* such that mF_p for *j* is similar to mF'_p for *i*, or $mF_p(j) \sim mF'_p(i)$. Thus, $mF_p(i) - mF'_p(i) \sim mF_p(i) - mF_p(j)$. Therefore, the estimation error of $mF_p(i) - mF_p(j)$ approximates that of $mF_p(i) - mF'_p(i)$. We found that the rank correlation between $mF_p(i) - mF_p(j)$ and $mF(i) - mF(j)$ is only 0.095 among 2,448 pairs of *i* and *j*. Assuming that the experimentally determined $mF(i) - mF(j)$ is true, our finding suggests that the computational estimation of $mF_p(i) - mF_p(j)$ or $mF_p(i) - mF'_p(i)$ has a relatively large error. Considering this fact, the true correlation between $mF_p(i) - mF'_p(i)$ and expression level could be substantially greater than the observed value of 0.06. Furthermore, because synonymous codon usage bias increases with gene expression level (5), pseudo-mRNAs tend to be more similar in DNA sequences to their real mRNA as the mRNA concentration gets higher ($P = 0.52$, $P < 10^{-15}$). For example, the mean DNA sequence identity between the 100 pseudo-mRNAs and the real mRNA is usually > 90% for the most highly expressed genes

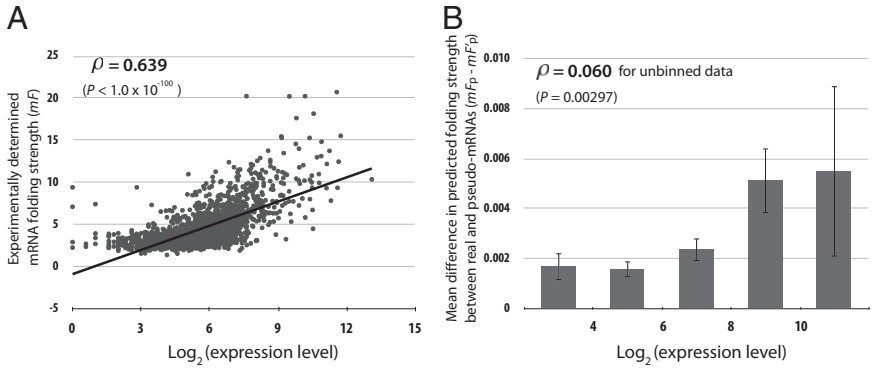

Fig. 1. Selective pressure for mRNA folding intensifies with gene expression level in yeast. (*A*) Experimentally determined mRNA folding strength (*mF*) of a gene increases with its expression level. Each dot represents a gene and the line shows the linear regression. ρ, rank correlation coefficient. (*B*) Mean difference between the predicted folding strengths of real and pseudo-mRNAs increases with gene expression level. Each real mRNA is compared with the average of 100 pseudo-mRNAs that have the same protein sequence and codon frequencies as the real mRNA, and genes with similar expression levels are binned. Error bars show one SE.

but is typically <80% for the least-expressed genes. This difference means that the random shuffling we conducted is less extensive for abundant mRNAs than for rare mRNAs, making $mF_p - mF'_p$ underestimated for the former, compared with the latter. In other words, the correlation in Fig. 1*B* is conservative. Regardless, there is a statistically significant trend that selection for mRNA folding intensifies with the concentration of the mRNA.

Mutational Harm to mRNA Folding Rises with the Folding Strength of mRNA. If the folding strength of an mRNA is selectively maintained at a higher than expected level, a random mutation is more likely to lower its folding strength than enhance it. To verify this prediction, we computationally estimated the folding strength of each yeast mRNA (*Materials and Methods*). Here, folding strength is measured by the free energy of the unfolded mRNA minus that of the folded mRNA, or $-\Delta G_{folding}$: the more positive the $-\Delta G_{folding}$ value, the stronger the folding. We then computationally estimated the folding strengths of all possible mRNAs that are one point mutation away from the real mRNA. As predicted, the greater the folding strength of an mRNA, the higher the fraction of point mutations that reduce the folding strength of the mRNA ($\rho = 0.586$, $P < 10^{-100}$) (Fig. 2*A*). Furthermore, the magnitude of the average mutational harm to mRNA folding ($\Delta\Delta G_{folding}$) also increases with the folding strength of the mRNA ($\rho = 0.407$, $P < 10^{-100}$) (Fig. 2*B*). Because highly expressed genes tend to have stronger mRNA folding, both the fraction of mutations that reduce mRNA folding (Fig. 2*C*) and the magnitude of the average mutational harm (Fig. 2*D*) increase with gene expression, although these trends are relatively weak, likely because of the substantial error associated with the computationally predicted $\Delta\Delta G_{folding}$.

Proteins with Greater mRNA Folding Strengths Evolve More Slowly. Because strong mRNA folding is likely to be selectively maintained (Fig. 1) and because random mutations are more likely to reduce mRNA folding when the folding is stronger (Fig. 2), the fraction of deleterious mutations in a gene is expected to rise with the folding strength of its mRNA. We thus predict that the amino acid substitution rate reduces as the mRNA folding strength of a gene increases. Indeed, among yeast genes, the amino acid substitution rate is negatively correlated with mF ($\rho = -0.429$, $n = 1{,}901$, $P < 10^{-77}$) (Fig. 3*A* and Table 1). This correlation remains significant when we analyze a subset of genes whose expression levels are similar (all within 0.95- to 1.05-times the mean expression level of all genes) ($\rho = -0.260$, $n = 285$, $P < 10^{-5}$) (Fig. 3*B*). Consistent with this finding, the partial rank correlation between the amino acid substitution rate of a gene and its mF, after the control of expression level, is $\rho = -0.227$ ($n = 1901$, $P < 10^{-23}$) (Table 1). Furthermore, within each gene, we calculated the rank correlation between the amino acid substitution rate of a codon and the mean mF of its three nucleotides. This correlation is significantly negative ($z = -5.58$, $P < 10^{-4}$) when combined from all of the genes analyzed (42).

We confirmed the genome-wide negative correlation between the computationally predicted mRNA folding strength (mF_p) of a gene and its amino acid substitution rate in yeast, before and after the control of the mRNA expression level (Table 2). The same correlations are also present in other model organisms examined, including animal, plant, and bacterial species (Table 2). These results suggest that our observations are not limited to yeast.

Mechanism for the E-R Anticorrelation. The findings in the above three sections suggest a mechanistic model for the E-R anticorrelation. This model, hereby termed the mRNA folding-strength model, makes two claims that have been demonstrated above. First, natural selection for mRNA folding is stronger for more highly expressed genes, resulting in a positive correlation between gene expression level and mRNA folding strength (Fig. 1). Second, a random mutation in a gene is more likely to weaken the mRNA folding as the folding gets stronger (Fig. 2); conse-

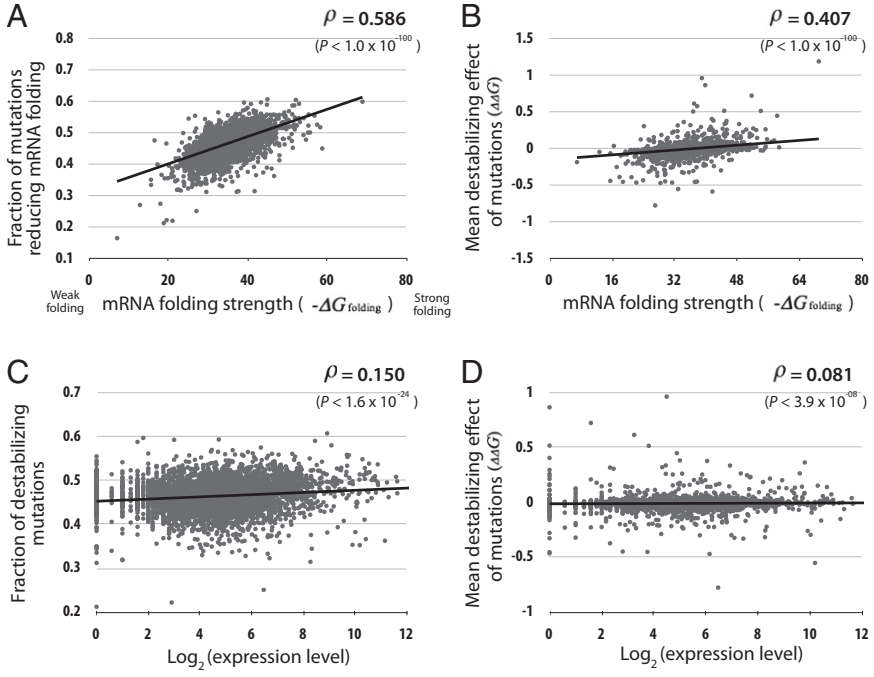

Fig. 2. Random point mutations are more harmful to mRNA folding when occurring in genes with stronger mRNA folding. (*A*) The fraction of mutations that reduce mRNA folding increases with the folding strength of the original mRNA. (*B*) The mean destabilizing effect of mutations increases with the folding strength of the original mRNA. (*C*) The fraction of mutations that reduce mRNA folding increases with the expression level of the gene. (*D*) The mean destabilizing effect of mutations increases with the expression level of the gene. In all panels, each dot represents a yeast gene and the line shows the linear regression. ρ, rank correlation coefficient.

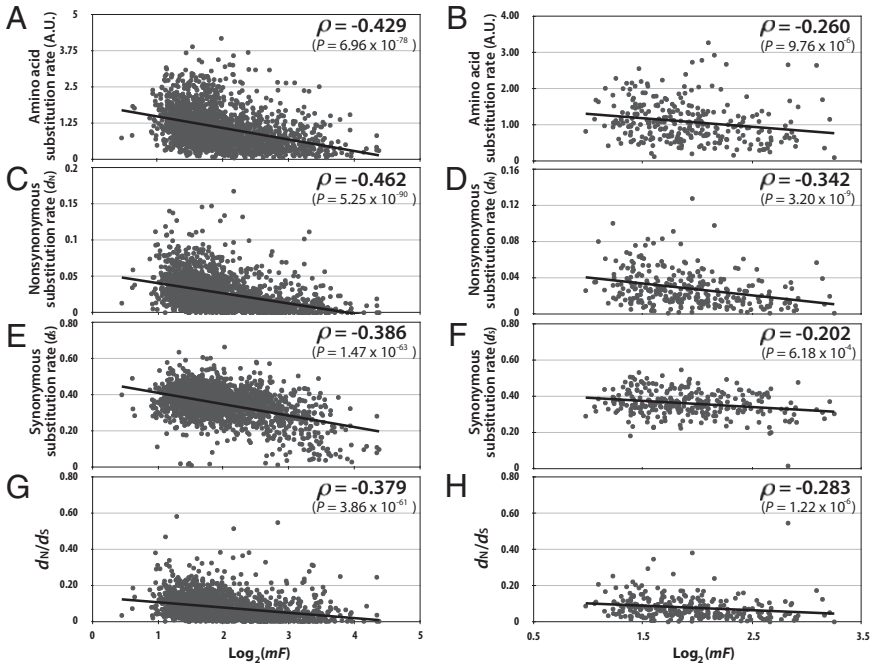

Fig. 3. Correlations between the yeast mRNA folding strength (*mF*) of a gene and its various rates of sequence evolution. (*A*) Correlation between *mF* and amino acid substitution rate. (*B*) Correlation between *mF* and amino acid substitution rate for the subset of genes whose expression levels are between 0.95- and 1.05-times the mean expression of all genes. (*C*) Correlation between *mF* and d_N. (*D*) Correlation between *mF* and d_N for the aforementioned subset of genes. (*E*) Correlation between *mF* and d_S. (*F*) Correlation between *mF* and d_S for the aforementioned subset of genes. (*G*) Correlation between *mF* and d_N/d_S. (*H*) Correlation between *mF* and d_N/d_S for the aforementioned subset of genes. In all panels, each dot represents a gene and the line shows the linear regression. ρ, rank correlation coefficient.

quently, amino acid substitutions are slower as the mRNA folding strength increases (Fig. 3). These two claims together lead to an E-R anticorrelation. If our model is correct, the E-R anticorrelation should be weakened when the mRNA folding strength is controlled. Indeed, this control causes the E-R anticorrelation to decrease from 0.423 ($P < 10^{-75}$) to 0.214 ($P < 10^{-20}$) (Table 1), a statistically significant drop ($P < 10^{-12}$). Nevertheless, the decreased E-R anticorrelation remains statistically significant, suggesting that selection for mRNA folding is not the sole cause of the E-R anticorrelation, which is consistent with the existence of other contributors such as selections against protein misfolding and misinteraction (18, 24, 25).

mRNA Folding Strength also Impacts Synonymous (d_S) and Nonsynoymous (d_N) Substitution Rates, and d_N/d_S. Just as the mRNA folding strength of a gene negatively impacts its amino acid substitution rate, the folding strength should also negatively impact the synonymous (d_S) and nonsynoymous (d_N) substitution rates. Indeed, we observed significant negative correlations between *mF* and both d_S and d_N among yeast genes (Figs. 3 *C* and *E*, and Table 1). The result on d_S supports an earlier report of selection acting on synonymous mutations because of their impacts on mRNA folding (40). These correlations remain significant after the control of gene expression level by either limiting the examined genes to those with similar expressions (Figs. 3 *D* and *F*) or using partial correlations (Table 1). Furthermore, *mF* also nega-

Table 1. Correlations between evolution rates and mRNA folding strength or expression level among yeast genes

	Rank correlations with various evolutionary rates			
Evolutionary rates considered	Correlation	*P* value	Partial correlation	*P* value
Amino acid substitution rate				
mRNA folding (*mF*)	−0.4286	6.96×10^{-78}	−0.2274*	2.56×10^{-24}
mRNA expression level	−0.4227	8.26×10^{-76}	−0.2143†	1.20×10^{-21}
Nonsynonymous rate (d_N)				
mRNA folding (*mF*)	−0.4615	5.25×10^{-90}	−0.2792*	8.71×10^{-37}
mRNA expression level	−0.4168	9.56×10^{-74}	−0.1787†	2.48×10^{-15}
Synonymous rate (d_S)				
mRNA folding (*mF*)	−0.3861	1.47×10^{-63}	−0.2595*	1.17×10^{-31}
mRNA expression level	−0.3070	7.95×10^{-41}	−0.0850†	2.00×10^{-4}
d_N/d_S				
mRNA folding (*mF*)	−0.3785	3.86×10^{-61}	−0.2133*	1.87×10^{-21}
mRNA expression level	−0.3521	3.85×10^{-53}	−0.1549†	8.39×10^{-12}

*After controlling mRNA expression level.
†After controlling *mF*.

Table 2. Genome-wide correlation between the amino acid substitution rate of a gene and its computationally predicted mRNA folding strength (mF_p)

Species	Correlation	*P* value	Partial correlation*	*P* value
Escherichia coli	−0.1736	6.38×10^{-12}	−0.1822	3.36×10^{-13}
Saccharomyces cerevisiae	−0.2919	2.31×10^{-44}	−0.2376	1.21×10^{-31}
Drosophila melanogaster	−0.0808	4.72×10^{-05}	−0.0484	1.50×10^{-02}
Arabidopsis thaliana	−0.1509	1.26×10^{-05}	−0.1270	2.30×10^{-04}
Mus musculus	−0.1261	1.36×10^{-31}	−0.1221	3.96×10^{-30}

*After controlling the mRNA expression level.

tively correlates with d_N/d_S (Fig. 3*G* and Table 1), and the correlation holds even after the control of the expression level (Fig. 3*H* and Table 1). This observation suggests that, as the mRNA folding strength rises, nonsynonymous mutations become more likely than synonymous mutations to harm mRNA folding, which is counterintuitive, because the differential fitness effects of synonymous and nonsynonymous mutations should manifest at the protein level rather than the mRNA level.

To identify the cause of the negative correlation between *mF* and d_N/d_S, we examined whether synonymous and nonsynonymous mutations have different effects on mRNA folding. Instead of examining all possible point mutations in a gene as a group (Fig. 2), we separated them into synonymous and nonsynonymous. We found that the average harm of a nonsynonymous mutation to mRNA folding ($\Delta\Delta G_N$) increases with the folding strength ($-\Delta G_{folding}$) of the mRNA ($\rho = 0.426$, $P < 10^{-100}$) (Fig. 4*A*). A similar pattern was observed for synonymous mutations ($\Delta\Delta G_S$) ($\rho = 0.256$, $P < 10^{-83}$) (Fig. 4*B*). As inferred in the previous section, the average harm of a nonsynonymous mutation minus that of a synonymous mutation ($\Delta\Delta G_N - \Delta\Delta G_S$) increases with the mRNA folding strength ($\rho = 0.205$, $P < 10^{-53}$) (Fig. 4*C*).

To understand the underlying cause of the above surprising observation, we examined A+U and G+C nucleotides separately. Because on average a G or C forms a stronger pair than an A or U, mutations from A/T to G/C in a gene tend to enhance mRNA folding and hence are referred to as strengthening (st) mutations. In contrast, mutations from G/C to A/T are called weakening (wk) mutations. Other mutations are referred to as equivalent (eq) mutations. We calculated the mean harm of each type of mutations to mRNA folding in each gene. The mean harm of strengthening mutations ($\Delta\Delta G_{st}$) decreases with the rise of the mRNA folding strength ($\rho = -0.220$, $P < 10^{-60}$) (Fig. 5*A*). This finding is understandable because, as mRNA folding gets stronger its G+C frequency tends to increase. Consequently, an A/T to G/C mutation is less likely to destroy a pairing but more likely to add a pairing. For the same reason, the mean harm of weakening mutations ($\Delta\Delta G_{wk}$) increases with the mRNA folding strength ($\rho = 0.424$, $P < 10^{-100}$) (Fig. 5*B*). The mean harm of equivalent mutations ($\Delta\Delta G_{eq}$) also increases with the mRNA folding strength ($\rho = 0.399$, $P < 10^{-100}$) (Fig. 5*C*). This result is probably because, in a G/C-rich mRNA, which tends to have strong folding, an equivalent mutation tends to impair a G:C pair, whereas in an A/U-rich mRNA, which tends to have weak folding, an equivalent mutation tends to impair an A:U pair, and impairing a G:C pair decreases folding strength more than impairing an A:U pair. We examined the composition of the three types of mutations among all possible synonymous and nonsynonymous mutations in yeast genes, respectively. Interestingly, compared with synonymous mutations, nonsynonymous mutations are enriched with equivalent mutations and are deprived of strengthening mutations (Fig. 5*D*). Thus, as the mRNA folding strength rises, nonsynonymous mutations become more deleterious, compared with synonymous mutations, which results in a negative correlation between *mF* and d_N/d_S.

One wonders why strengthening mutations are relatively enriched among synonymous mutations, whereas equivalent mutations are relatively enriched among nonsynonymous mutations. The answer lies in the genetic code table and the synonymous codon usage in yeast. Synonymous mutations occur primarily at third codon positions. Among the synonymous codons of each amino acid, the

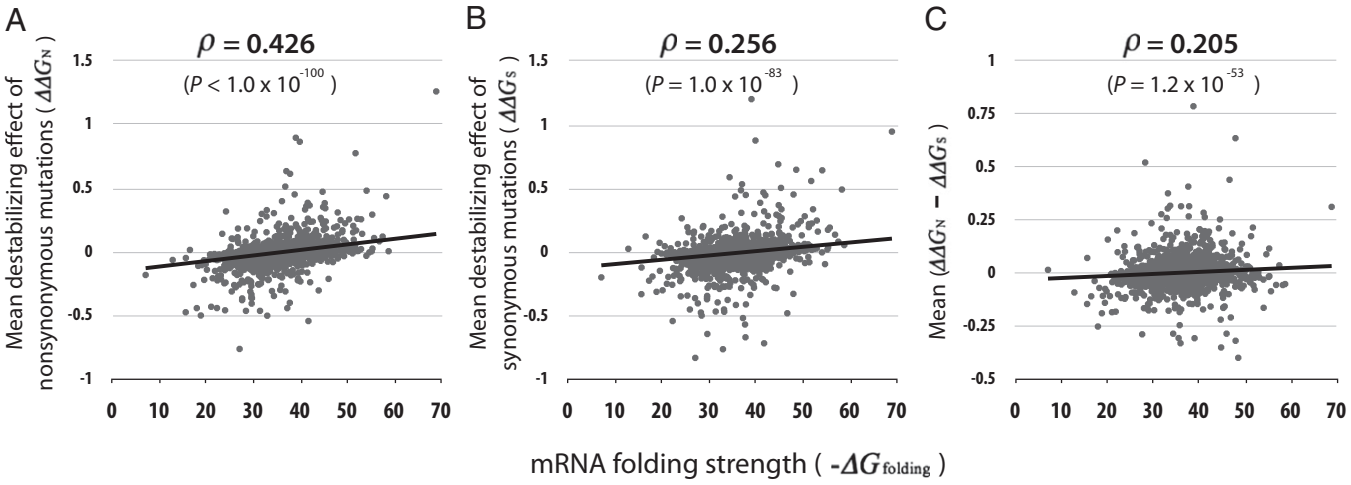

Fig. 4. Destabilizing effects of nonsynonymous and synonymous mutations on mRNA folding. (*A*) Mean destabilizing effect of nonsynonymous mutations increases with the folding strength of the original mRNA. (*B*) Mean destabilizing effect of synonymous mutations increases with the folding strength of the original mRNA. (*C*) Mean difference in destabilizing effect between nonsynonymous and synonymous mutations increases with the folding strength of the original mRNA. In all panels, each dot represents a yeast gene and the line shows the linear regression. ρ, rank correlation coefficient.

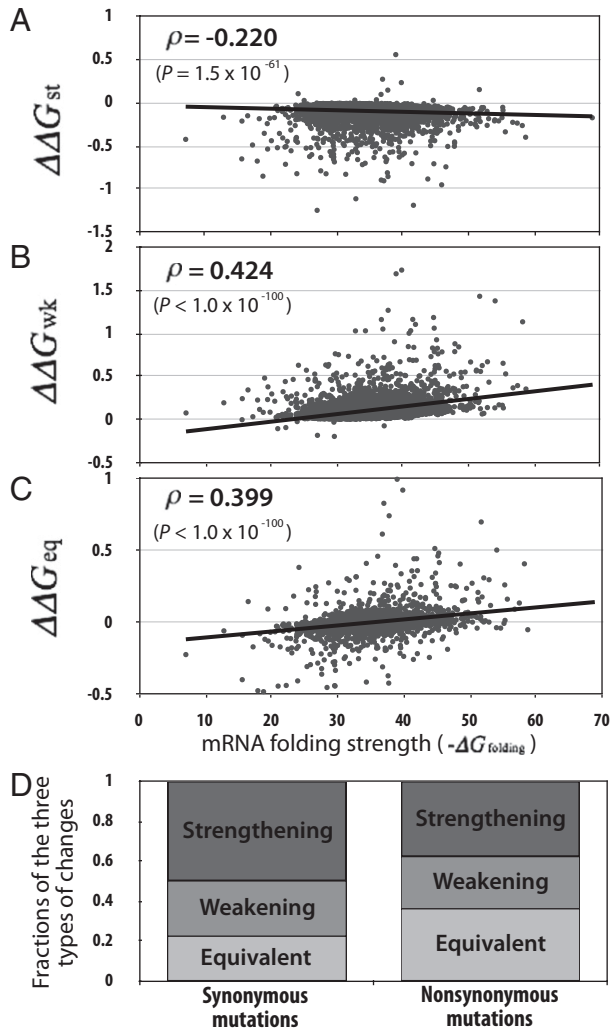

Fig. 5. Destabilizing effects of strengthening (st), weakening (wk), and equivalent (eq) mutations. (*A*) Mean destabilizing effect of strengthening mutations decreases with the rise of the folding strength of the original mRNA. (*B*) Mean destabilizing effect of weakening mutations increases with the folding strength of the original mRNA. (*C*) Mean destabilizing effect of equivalent mutations increases with the folding strength of the original mRNA. (*D*) Fractions of strengthening, weakening, and equivalent mutations among synonymous and nonsynonymous mutations. In *A*–*C*, each dot represents a yeast gene, the line shows the linear regression, and ρ is the rank correlation coefficient.

preferred one always ends with an A or T in yeast (http://cbio.ufs.ac.za/lectures/yeast_codon_table.txt). At these third codon positions with A or T, transitions (from A to G; from T to C) would be strengthening, but transversions (from A to C or T; from T to G or A) would be 50% strengthening and 50% equivalent. Because the genetic code table is structured such that transitions at third codon positions are more likely than transversions to be synonymous, synonymous mutations are relatively enriched with the strengthening type, but nonsynonymous mutations are relatively enriched with the equivalent type, when third codon positions are considered. At the first and second codon positions, the G+C frequency in yeast is ∼40%. If the mutation rates among the four nucleotides are equal, 40% strengthening, 26.7% weakening, and 33.3% equivalent mutations are expected at these positions, and the vast majority of them are nonsynonymous. When all three positions are considered together, synonymous mutations are still enriched with the strengthening type, whereas the nonsynonymous mutations are enriched with the equivalent type. Note that the above calculation is meant to offer an intuitive explanation of a counterintuitive phenomenon. It is not a precise calculation, because of the omission of synonymous mutations at first codon positions and other complications arising from the structure of the genetic code table and amino acid and codon usage biases. The result from the exact calculation is shown in Fig. 5*D*, where the actual gene sequences from yeast are used.

The above findings suggest that the relative fractions of strengthening, weakening, and equivalent mutations among synonymous and nonsynonymous mutations may vary among species, depending on the G+C frequencies at various codon positions. Consequently, the relationship between mF and d_N/d_S may vary among species. Indeed, analyzing nine pairs of bacterial genomes with varying G+C frequencies reveals a positive correlation between the G+C frequency at third codon positions and the correlation between mF_p and d_N/d_S (Fig. 6*A*). In species with high G+C frequencies, mF_p and d_N/d_S are positively correlated, whereas in other species they are negatively correlated, as in yeast. A similar result is found when the gene expression level is controlled (Fig. 6*B*). Use of genomic G+C frequencies yielded similar results (Fig. S1).

Discussion

In this study, we proposed and demonstrated that natural selection for stronger folding of more abundant mRNAs constrains the evolution of highly expressed genes, resulting in a negative correlation between the expression level of a gene and its rate of protein sequence evolution. However, what is the benefit of strong folding of abundant mRNAs? Recent studies indicated that strong folding does not increase the half-life of the mRNA (28) and does not enhance the translational efficiency of the mRNA (28, 43–45). In fact, low mRNA folding at the beginning of an mRNA (5′ untranslated region and the first ∼30 coding nucleotides) appears to enhance the translational initiation rate (43–45). Zur and Tuller (28) proposed that mRNA aggregation (i.e., pairing between two mRNA molecules) becomes problematic as the mRNA concentration rises and that strong mRNA folding may prevent such aggregation. For the following two reasons, this hypothesis is unlikely to be correct. First, although the aggregation of mRNA molecules is expected to decrease translational efficiency, mRNA folding has a similar effect (46) at least qualitatively. If mRNA folding is energetically more stable than mRNA aggregation, folding would probably hamper translation more than aggregation. Second, using a biophysical model of the competition between mRNA folding and aggregation, we attempted to identify yeast genes for which mRNA aggregation is energetically favored over folding (*Materials and Methods*). However, only one gene was found (Fig. 7). Furthermore, the significant negative correlation between gene expression level and the energetic preference for aggregation (Fig. 7) suggests a reduced rather than an increased risk of mRNA aggregation for more abundant mRNAs. Thus, avoiding mRNA aggregation cannot explain the selective pressure for stronger folding of more abundant mRNAs. Because mRNA folding likely affects one or more aspects of translation, such as elongation speed and cotranslational protein folding (46–48), which are under differential selective pressures for genes of different expression levels (18, 38), it is likely that the selection for mRNA folding is related to the translational process. However, exactly which aspect and property of translation is selected for that has led to enhanced folding of abundant mRNAs awaits further investigation.

The identification of yet another mechanism of the E-R anticorrelation strongly supports the view that the anticorrelation has multiple distinct causes (25). Different from the three mechanisms previously proposed (protein misfolding avoidance, protein misinteraction avoidance, and protein function), the mechanism identified here relates directly to a property of mRNA rather than protein, although it is possible that this mRNA property

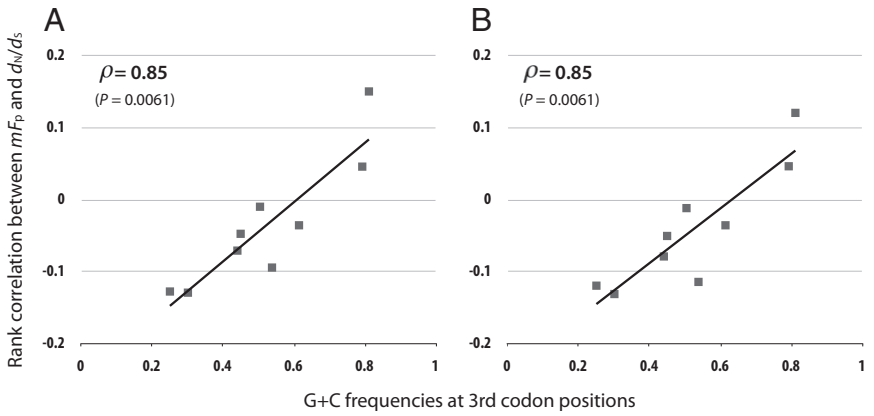

Fig. 6. The rank correlation between mRNA folding strength (mF_p) and d_N/d_S depends on the G+C frequency at third codon positions. (*A*) The rank correlation between mF_p and d_N/d_S increases with the G+C frequency at third codon positions. (*B*) The rank correlation between mF_p and d_N/d_S after the control of expression level increases with the G+C frequency at third codon positions. In both panels, each dot represents a pair of closely related bacterial genomes, the line shows the linear regression, and ρ is the rank correlation coefficient.

ultimately impacts some protein properties. It is important to note that controlling mRNA folding strength weakens the E-R anticorrelation from 0.423 to 0.214, about 10-times the effect of controlling several proxies of protein misfolding and protein misinteraction (25), suggesting that selection for strong mRNA folding contributes to the E-R anticorrelation more than the other two mechanisms. An alternative interpretation is that the proxies for the other two mechanisms are less accurate than the experimentally determined *mF* score for mRNA folding strength, resulting in smaller consequences to the E-R anticorrelation when removed. We found that controlling the computationally predicted mF_p also significantly ($P = 0.031$) reduces the E-R anticorrelation, from 0.423 to 0.372, despite the moderate correlation between mF_p and *mF*. The magnitude of this reduction also exceeds those caused by controlling proxies for the other two mechanisms (25). Thus, it is likely that the newly discovered mechanism is a primary contributor to the E-R anticorrelation.

It is notable that in yeast the anticorrelation between mRNA folding strength and protein evolutionary rate (0.4286) (Table 1) is slightly but not significantly stronger than the E-R anticorrelation (0.4227) (Table 1), suggesting that the overall impact of mRNA folding on protein evolutionary rate is at least as great as that of expression level. Because mRNA folding strength and expression level are highly correlated, one could control one factor when evaluating the other. The partial anticorrelation between mRNA folding strength and protein evolutionary rate (0.2274) (Table 1) is again slightly but not significantly stronger than the partial E-R anticorrelation (0.2143) (Table 1), suggesting that the net impacts of mRNA folding strength and expression level on protein evolutionary rate are comparable, after the exclusion of their overlapping impact. That a major part of the functional constraint of protein evolution arises from mRNA folding is a strong testament of the centrality of RNAs in molecular biology (49).

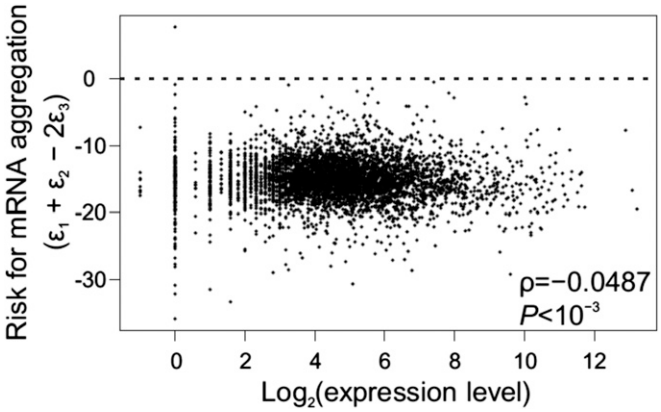

Fig. 7. Risk for mRNA aggregation is negligible for all but one yeast gene. Each dot represents a yeast gene. Intermolecular aggregation is energetically favored over intramolecular folding above the dotted line ($\varepsilon_1 + \varepsilon_2 = 2\varepsilon_3$), and folding is energetically favored over aggregation below the dotted line. ρ, rank correlation coefficient.

In addition to lowering the amino acid substitution rate, we predicted and demonstrated that selection for strong mRNA folding reduces the d_S and d_N of a gene. What is unexpected, however, is that mRNA folding strength also impacts the d_N/d_S ratio, which is widely used to measure selective pressures at the protein level (5, 7). Furthermore, the impact of the mRNA folding strength on d_N/d_S is dependent on the G+C frequency of the genome or third codon positions. For example, in yeast, strong mRNA folding negatively impacts d_N/d_S. As shown in Fig. 3*G*, all yeast genes analyzed have $d_N/d_S < 1$. In such cases, the common interpretation is that, the lower the d_N/d_S ratio, the stronger the selection against protein sequence changes. Our finding, however, indicates that a low d_N/d_S ratio could also be caused by a selective maintenance of strong mRNA folding that may be unrelated to protein function. In other words, d_N/d_S may overestimate the protein-level purifying selection. In species with high G+C frequencies, however, strong mRNA folding would increase d_N/d_S (Fig. 6) and lead to an underestimation of protein-level purifying selection. This result would be particularly problematic when the observed $d_N/d_S > 1$, because such a ratio is commonly interpreted as the action of positive selection for protein sequence changes (5, 7), but apparently it could also be a byproduct of the selective maintenance of mRNA folding in species with high G+C frequencies. Thus, cautions are required in interpreting d_N/d_S ratios. More importantly, the expected d_N/d_S ratio could be different among genes or genomes of different G+C frequencies even when they are under the same extent of protein-level selection. Hence, one may not unambiguously infer the relative strengths of natural selection on two proteins (or two proteomes) by comparing their d_N/d_S ratios (or mean d_N/d_S ratios). More studies are needed to evaluate the extent to which such inferences are affected and to correct potential biases.

We note that most of the correlations reported in this study are strong when experimentally determined mRNA folding strengths are used, but weak when computationally predicted folding strengths are used. This contrast is not unexpected, given

the limited accuracy of computational prediction of mRNA folding, evident from the moderate correlation between *mF* and mF_p (0.22). Our analyses illustrate the importance of experimental determination of mRNA folding and call for further improvement of the computational prediction.

In conclusion, selection for strong mRNA folding constrains the evolution of both gene and protein sequences and has variable effects on d_N/d_S depending on the G+C frequency of the species considered. Future research should aim to understand the benefits of mRNA folding, especially for highly expressed genes, and to devise methods that distinguish the natural selection at the protein level from that at the mRNA level.

Materials and Methods

Yeast mRNA Folding Strength. We used the same definition of folding score *mF* as in ref. 28. That is, the ratio between the probability that a nucleotide is paired and the probability that it is unpaired was calculated using the yeast experimental data obtained from ref. 32. The ratios were then averaged across all nucleotides of an mRNA to get the *mF* score of the mRNA.

The secondary structures of mRNAs were also computationally predicted using RNAfold (50) found in the ViennaRNA package (www.tbi.univie.ac.at/~ivo/RNA/), under the folding temperature of 30 °C. The prediction was based on a thermodynamic model with the assumption that the secondary structure of an mRNA is determined by the minimum free energy (51). Because long-range interactions within an mRNA molecule are rare (52, 53) and because predicting long-range interactions is computationally demanding (54), our secondary structure prediction considered windows of 150 nucleotides (55) with a step size of 10 nucleotides. Using window sizes of 50 nucleotides and 100 nucleotides yielded similar results. At each nucleotide, the probability that it is paired was estimated by the number of sliding windows in which it is paired divided by the number of sliding windows that include the nucleotide. The probabilities were averaged among all sites of an mRNA to yield the folding score mF_p of the mRNA.

To test whether the secondary structure of an mRNA is more stable than expected by chance, we generated 100 pseudo-mRNA sequences from each real mRNA by randomly shuffling all synonymous codons in the real mRNA. Thus, each pseudo-mRNA has the same nucleotide frequencies, codon frequencies, and amino acid sequence as the real mRNA. The mean of the mF_p values of the 100 pseudo-mRNA sequences is designated mF'_p and is compared with mF_p of the real mRNA. Our analysis is superior to that in ref. 28, where pseudo-mRNAs and the real mRNA have the same amino acid sequence but different codon frequencies. Although an earlier study controlled dinucleotide frequencies is generating pseudo-mRNA sequences (41), we did not do so because this control was later deemed inappropriate (40) for the reason that dinucleotide frequencies may be altered by natural selection for mRNA folding.

In examining the impact of random mutations on the folding stability of an mRNA, we first used RNAfold (50) to computationally estimate the folding strength ($-\Delta G_{folding}$) of a real mRNA, which is the free energy of the unfolded mRNA minus that of the folded mRNA. We then estimated the harm of a mutation to mRNA folding by $\Delta\Delta G$, which is $(-\Delta G_{folding})_{wild\text{-}type}$ minus $(-\Delta G_{folding})_{mutant}$.

In all of our analyses, mRNA folding was examined for the entire protein-coding region of an mRNA. Because the folding strength of mRNA near the start codon is lower in highly expressed genes than lowly expressed ones (44, 45), we recalculated *mF* after excluding the first 45 nucleotides of the coding region. Our results were unaltered by using the new *mF* values (Fig. S2 and Table S1).

Gene Expression Levels. Yeast mRNA concentrations were obtained from Illumina-based RNA sequencing measurements in the log growth phase under the rich medium YPD (33). We similarly used the RNA sequencing data from *Escherichia coli* in the log growth phase in Luria Broth (LB) (56), *Drosophila melanogaster* adults of mixed ages (57), *Mus musculus* testis (58), and *Arabidopsis thaliana* seeds under the accession number of ME00360 (59).

Evolutionary Rates. The amino acid substitution rates of *Saccharomyces cerevisiae* proteins were estimated using the alignments of orthologs from six species (*S. cerevisiae*, *Saccharomyces paradoxus*, *Saccharomyces mikatae*, *Saccharomyces bayanus*, *Candida glabrata*, and *Saccharomyces castellii*) and were obtained from ref. 25.

To estimate the synonymous (d_S) and nonsynonymous (d_N) substitution rates, we first obtained the orthologous gene sequences from *S. cerevisiae* and *S. paradoxus* using the annotations in ftp://ftp.sanger.ac.uk/pub/users/dmc/yeast/latest/ (60). The protein-coding nucleotide sequences were aligned using MUSCLE (61) with the default option, and d_S and d_N were estimated using PAML (62).

We also analyzed protein evolutionary rates by comparing four additional model organisms with their respective close relatives: *Escherichia coli–Salmonella enterica*, *Drosophila melanogaster–Drosophila simulans*, *Arabidopsis thaliana–Arabidopsis lyrata*, and *Mus musculus–Rattus norvegicus*. Their orthologous gene sequences were obtained from the Integrated Microbial Genomes system (http://img.jgi.doe.gov/cgi-bin/w/main.cgi), FlyBase (http://flybase.org/), Phytozome (www.phytozome.net/), and ENSEMBL (http://useast.ensembl.org/index.html), respectively. The amino acid sequences were aligned using MUSCLE (61) with the default option and the amino acid substitution rate was estimated using PAML (62).

Competition Between mRNA Folding and Aggregation. We adapted the model in ref. 63 to access the competition between mRNA intramolecular folding and intermolecular aggregation. In this model, sequence heterogeneity is neglected so that the pairing energy between nucleotide *i* and *j* is denoted by

$$\varepsilon_{i,j} = \begin{cases} \varepsilon_1 \text{if } i,j \in \text{RNA}-1 \\ \varepsilon_2 \text{if } i,j \in \text{RNA}-2 \\ \varepsilon_3 \text{ if } i \in \text{RNA}-1 \text{ and } j \in \text{RNA}-2, \text{or vice versa,} \end{cases} \quad [1]$$

where RNA-1 and RNA-2 are the two RNA molecules considered in the model. The pairing gets stronger as ε becomes more negative. We dissected both mRNA sequences into 150-nucleotide nonoverlapping segments; the remaining fragments at the 3′ end of the mRNAs that are shorter than 150 nucleotides were ignored. We calculated ε_1 and ε_2 by averaging the per-nucleotide minimum folding energy of each segment from the corresponding mRNAs by RNAfold. To calculate ε_3, we averaged the per-nucleotide duplex folding (i.e., only intermolecular base pairs are allowed to form) energies between every possible intermolecular segment pairs.

According to ref. 63, the dual RNA system has a phase transition at $\varepsilon_1+\varepsilon_2 = 2\varepsilon_3$. That is, when $\varepsilon_1+\varepsilon_2 < 2\varepsilon_3$, mRNA folding dominates the system; otherwise, aggregation dominates. To investigate how the competition between aggregation and folding depends on the expression level of a gene, we first picked a focal gene and estimated its ε_1. We then randomly picked 100 mRNA molecules from the transcriptome based on the relative concentrations of all mRNAs and calculated the expected values of ε_2 and ε_3.

Prokaryotic Data Analysis. Publicly available microarray gene expression data and genome sequences of nine pairs of bacterial genomes were obtained from a previous study (64). We chose closely related genome pairs such that their d_S can be accurately estimated and the expression levels measured in one genome can approximate those of the other. The expression data were originally retrieved from the Stanford Microarray Database (*Bacillus subtilis 168*, ID: 66211; *Helocobacter pylori*, ID: 16576; *Mycobacterium tuberculosis*, ID: 14047; *Salmonella typhimurium*, ID: 23956; and *Vibrio cholerae*: ID 66211) and the National Center for Biotechnology Information Gene Expression Omnibus (*Dehalococcoides ethenogenes*, GSE 10185; *Geobacter sulfurreducens*, GSE 22511; *Listeria monocytogenes*, GSE 16336; and *Streptococcus agalactiae A909*, GSE 21564). The genomic data, including protein and DNA sequences for all protein-coding genes and sets of orthologs, were downloaded from the Integrated Microbial Genomes system (http://img.jgi.doe.gov/cgi-bin/w/main.cgi) (65).

Orthologous genes from nine species pairs were analyzed separately: *Bacillus subtilis* 168 and *B. subtilis spizizenii*, *Helocobacter pylori* and *Helocobacter acinonychis*, *Mycobacterium tuberculosis* and *Mycobacterium marinum*, *Salmonella typhimurium* and *Salmonella bongori*, *Vibrio cholerae* and *Vibrio mimicus*, *Dehalococcoides ethenogenes* and *Dehalococcoides* sp. CBDB1, *Geobacter sulfurreducens* and *Geobacter metallireducens*, *Listeria monocytogenes* and *Listeria innocua*, and *Streptococcus agalactiae* A909 and *Streptococcus agalactiae* NEM316. The DNA sequence alignments were obtained using TRANALIGN in EMBOSS (http://helixweb.nih.gov/emboss/html/tranalign.html) after the protein sequence alignments made using MUSCLE (61). We then calculated d_S and d_N using PAML (62). The folding scores (mF_p) for prokaryotic mRNAs were predicted by RNAfold under the same parameters used for yeast mRNAs.

ACKNOWLEDGMENTS. We thank Wei-Chin Ho, Wenfeng Qian, Jinrui Xu, and two anonymous reviewers for valuable comments. This work was supported in part by a research grant from the National Institutes of Health (to J.Z.).

1. Zukerkandl E, Pauling L (1965) Evolutionary divergence and convergence in proteins. *Evolving Genes and Proteins*, eds Bryson V, Vogel HJ (Academic, New York), pp 97–166.
2. Kimura M (1968) Evolutionary rate at the molecular level. *Nature* 217(5129):624–626.
3. Kimura M, Ota T (1974) On some principles governing molecular evolution. *Proc Natl Acad Sci USA* 71(7):2848–2852.
4. Kimura M (1983) *The Neutral Theory of Molecular Evolution* (Cambridge Univ Press, Cambridge).
5. Li W (1997) *Molecular Evolution* (Sinauer, Sunderland, MA).
6. Nei M (1987) *Molecular Evolutionary Genetics* (Columbia Univ Press, NY).
7. Nei M, Kumar S (2000) *Molecular Evolution and Phylogenetics* (Oxford Univ Press, NY).
8. Lobkovsky AE, Wolf YI, Koonin EV (2010) Universal distribution of protein evolution rates as a consequence of protein folding physics. *Proc Natl Acad Sci USA* 107(7): 2983–2988.
9. Hurst LD, Smith NG (1999) Do essential genes evolve slowly? *Curr Biol* 9(14):747–750.
10. Pál C, Papp B, Hurst LD (2001) Highly expressed genes in yeast evolve slowly. *Genetics* 158(2):927–931.
11. Zhang J, He X (2005) Significant impact of protein dispensability on the instantaneous rate of protein evolution. *Mol Biol Evol* 22(4):1147–1155.
12. Liao BY, Scott NM, Zhang J (2006) Impacts of gene essentiality, expression pattern, and gene compactness on the evolutionary rate of mammalian proteins. *Mol Biol Evol* 23(11):2072–2080.
13. Zhang L, Li WH (2004) Mammalian housekeeping genes evolve more slowly than tissue-specific genes. *Mol Biol Evol* 21(2):236–239.
14. Drummond DA, Raval A, Wilke CO (2006) A single determinant dominates the rate of yeast protein evolution. *Mol Biol Evol* 23(2):327–337.
15. Hirsh AE, Fraser HB (2001) Protein dispensability and rate of evolution. *Nature* 411(6841):1046–1049.
16. Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW (2002) Evolutionary rate in the protein interaction network. *Science* 296(5568):750–752.
17. Wolf YI, Carmel L, Koonin EV (2006) Unifying measures of gene function and evolution. *Proc Biol Sci* 273(1593):1507–1515.
18. Drummond DA, Wilke CO (2008) Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* 134(2):341–352.
19. Subramanian S, Kumar S (2004) Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. *Genetics* 168(1):373–381.
20. Rocha EP, Danchin A (2004) An analysis of determinants of amino acids substitution rates in bacterial proteins. *Mol Biol Evol* 21(1):108–116.
21. Liao BY, Weng MP, Zhang J (2010) Impact of extracellularity on the evolutionary rate of mammalian proteins. *Genome Biol Evol* 2:39–43.
22. Wang Z, Zhang J (2009) Why is the correlation between gene importance and gene evolutionary rate so weak? *PLoS Genet* 5(1):e1000329.
23. Geiler-Samerotte KA, et al. (2011) Misfolded proteins impose a dosage-dependent fitness cost and trigger a cytosolic unfolded protein response in yeast. *Proc Natl Acad Sci USA* 108(2):680–685.
24. Yang JR, Zhuang SM, Zhang J (2010) Impact of translational error-induced and error-free misfolding on the rate of protein evolution. *Mol Syst Biol* 6:421.
25. Yang JR, Liao BY, Zhuang SM, Zhang J (2012) Protein misinteraction avoidance causes highly expressed proteins to evolve slowly. *Proc Natl Acad Sci USA* 109(14):E831–E840.
26. Gout JF, Kahn D, Duret L; Paramecium Post-Genomics Consortium (2010) The relationship among gene expression, the evolution of gene dosage, and the rate of protein evolution. *PLoS Genet* 6(5):e1000944.
27. Cherry JL (2010) Expression level, evolutionary rate, and the cost of expression. *Genome Biol Evol* 2:757–769.
28. Zur H, Tuller T (2012) Strong association between mRNA folding strength and protein abundance in *S. cerevisiae*. *EMBO Rep* 13(3):272–277.
29. Mathews DH, Sabina J, Zuker M, Turner DH (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol* 288(5):911–940.
30. Zuker M (2000) Calculating nucleic acid secondary structure. *Curr Opin Struct Biol* 10(3):303–310.
31. Felden B (2007) RNA structure: Experimental analysis. *Curr Opin Microbiol* 10(3): 286–291.
32. Kertesz M, et al. (2010) Genome-wide measurement of RNA secondary structure in yeast. *Nature* 467(7311):103–107.
33. Nagalakshmi U, et al. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 320(5881):1344–1349.
34. Akashi H, Gojobori T (2002) Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis*. *Proc Natl Acad Sci USA* 99(6): 3695–3700.
35. Akashi H (2003) Translational selection and yeast proteome evolution. *Genetics* 164(4):1291–1303.
36. Kudla G, Lipinski L, Caffin F, Helwak A, Zylicz M (2006) High guanine and cytosine content increases mRNA levels in mammalian cells. *PLoS Biol* 4(6):e180.
37. Seffens W, Digby D (1999) mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences. *Nucleic Acids Res* 27(7):1578–1584.
38. Qian W, Yang JR, Pearson NM, Maclean C, Zhang J (2012) Balanced codon usage optimizes eukaryotic translational efficiency. *PLoS Genet* 8(3):e1002603.
39. Stoletzki N (2008) Conflicting selection pressures on synonymous codon use in yeast suggest selection on mRNA secondary structures. *BMC Evol Biol* 8:224.
40. Chamary JV, Hurst LD (2005) Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. *Genome Biol* 6(9):R75.
41. Katz L, Burge CB (2003) Widespread selection for local RNA secondary structure in coding regions of bacterial genes. *Genome Res* 13(9):2042–2051.
42. Han C-P (1989) Combining tests for correlation coefficients. *Am Stat* 43(4):211–215.
43. Kudla G, Murray AW, Tollervey D, Plotkin JB (2009) Coding-sequence determinants of gene expression in *Escherichia coli*. *Science* 324(5924):255–258.
44. Gu W, Zhou T, Wilke CO (2010) A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes. *PLOS Comput Biol* 6(2): e1000664.
45. Tuller T, Waldman YY, Kupiec M, Ruppin E (2010) Translation efficiency is determined by both codon bias and folding energy. *Proc Natl Acad Sci USA* 107(8):3645–3650.
46. Qu X, et al. (2011) The ribosome uses two active mechanisms to unwind messenger RNA during translation. *Nature* 475(7354):118–121.
47. Wen JD, et al. (2008) Following translation by single ribosomes one codon at a time. *Nature* 452(7187):598–603.
48. Watts JM, et al. (2009) Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature* 460(7256):711–716.
49. Darnell JE (2011) *RNA: Life's Indispensable Molecule* (Cold Spring Harbor Lab Press, Cold Spring Harbor, N.Y.), pp xiv, 416 pp.
50. Hofacker IL, et al. (1994) Fast folding and comparison of RNA secondary structures. *Monatsh Chem* 125(2):167–188.
51. Zuker M, Stiegler P (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res* 9(1):133–148.
52. Doktycz MJ, Larimer FW, Pastrnak M, Stevens A (1998) Comparative analyses of the secondary structures of synthetic and intracellular yeast MFA2 mRNAs. *Proc Natl Acad Sci USA* 95(25):14614–14621.
53. Parsch J, Stephan W, Tanda S (1998) Long-range base pairing in *Drosophila* and human mRNA sequences. *Mol Biol Evol* 15(7):820–826.
54. Doshi KJ, Cannone JJ, Cobaugh CW, Gutell RR (2004) Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction. *BMC Bioinformatics* 5:105.
55. Lange SJ, et al. (2012) Global or local? Predicting secondary structure and accessibility in mRNAs. *Nucleic Acids Res* 40(12):5215–5226.
56. Giannoukos G, et al. (2012) Efficient and robust RNA-seq process for cultured bacteria and complex community transcriptomes. *Genome Biol* 13(3):R23.
57. Graveley BR, et al. (2011) The developmental transcriptome of *Drosophila melanogaster*. *Nature* 471(7339):473–479.
58. Brawand D, et al. (2011) The evolution of gene expression levels in mammalian organs. *Nature* 478(7369):343–348.
59. Lamesch P, et al. (2012) The *Arabidopsis* Information Resource (TAIR): Improved gene annotation and new tools. *Nucleic Acids Res* 40(Database issue):D1202–D1210.
60. Liti G, et al. (2009) Population genomics of domestic and wild yeasts. *Nature* 458(7236):337–341.
61. Edgar RC (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32(5):1792–1797.
62. Yang Z (2007) PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 24(8):1586–1591.
63. Guttal V, Bundschuh R (2006) Model for folding and aggregation in RNA secondary structures. *Phys Rev Lett* 96(1):018105.
64. Park C, Zhang J (2012) High expression hampers horizontal gene transfer. *Genome Biol Evol* 4(4):523–532.
65. Markowitz VM, et al. (2010) The integrated microbial genomes system: An expanding comparative analysis resource. *Nucleic Acids Res* 38(Database issue):D382–D390.