Highly Abundant Proteins Are Highly Thermostable

Agusto R. Luzuriaga-Neira[1], Andrew M. Ritchie[2], Bryan L. Payne[1], Oliver Carrillo-Parramon[3], David A. Liberles [2,*], and David Alvarez-Ponce [1,*]

[1]Biology Department, University of Nevada, Reno, USA

[2]Department of Biology and Center for Computational Genetics and Genomics, Temple University, Philadelphia, Pennsylvania, USA

[3]Department of Physics, Universitat Politècnica de Catalunya, Barcelona, Spain

*Corresponding authors: E-mails: dap@unr.edu; daliberles@temple.edu.

Accepted: 08 June 2023

Abstract

Highly abundant proteins tend to evolve slowly (a trend called E-R anticorrelation), and a number of hypotheses have been proposed to explain this phenomenon. The misfolding avoidance hypothesis attributes the E-R anticorrelation to the abundance-dependent toxic effects of protein misfolding. To avoid these toxic effects, protein sequences (particularly those of highly expressed proteins) would be under selection to fold properly. One prediction of the misfolding avoidance hypothesis is that highly abundant proteins should exhibit high thermostability (i.e., a highly negative free energy of folding, ΔG). Thus far, only a handful of analyses have tested for a relationship between protein abundance and thermostability, producing contradictory results. These analyses have been limited by 1) the scarcity of ΔG data, 2) the fact that these data have been obtained by different laboratories and under different experimental conditions, 3) the problems associated with using proteins' melting energy (T_m) as a proxy for ΔG, and 4) the difficulty of controlling for potentially confounding variables. Here, we use computational methods to compare the free energy of folding of pairs of human–mouse orthologous proteins with different expression levels. Even though the effect size is limited, the most highly expressed ortholog is often the one with a more negative ΔG of folding, indicating that highly expressed proteins are often more thermostable.

Key words: protein thermostability, expression levels, misfolding avoidance hypothesis, translational robustness hypothesis, rates of evolution.

Significance

Highly expressed proteins tend to evolve slowly. The misfolding avoidance hypothesis attributes this phenomenon to proteins (particularly highly expressed ones) being under selection to maintain structures that avoid toxic misfolding. One prediction of the misfolding avoidance hypothesis is that highly abundant proteins should be highly thermostable. Are highly abundant proteins indeed highly thermostable? Despite the far-reaching implications of this possibility, attempts to test for the trend using empirical data have thus far provided contradictory and controversial results. By comparing pairs of human–mouse orthologs with different expression levels, we show that the most highly expressed ortholog is often the most thermostable. Our results indicate that highly expressed proteins are indeed, on average, highly thermostable.

Introduction

It is widely appreciated that highly expressed proteins tend to evolve slowly, a trend known as the E-R anticorrelation (Pál et al. 2001). A number of hypotheses have been proposed to try to explain this phenomenon (for review, see Alvarez-Ponce 2014; Zhang and Yang 2015). The

functional maintenance hypothesis suggests that functionally critical proteins are expressed at higher levels and that their function drives their slower evolutionary rate (Cherry 2010a; Gout et al. 2010). The misinteraction avoidance hypothesis suggests that when proteins are at higher concentration in a cell, they will have higher fractional occupancy with an increased number of nonspecific binding partners, leading to selective pressure to avoid sequences on the protein surface or binding interface that would increase the affinity for nonspecific partners (Liberles et al. 2011; Yang et al. 2012). The translational robustness hypothesis argues that translation errors occur at relatively high rates and can result in protein misfolding (whose deleterious effects are abundance dependent), leading to proteins (particularly highly expressed ones) being under selective pressure to maintain structures that can fold despite translation errors (Drummond et al. 2005). The misfolding avoidance hypothesis was proposed as an extension of the translational robustness hypothesis that recognizes that protein misfolding can occur not only as a result of translation errors but also spontaneously in the absence of such errors and posits that proteins (particularly highly expressed ones) are under selective pressures to maintain misfolding-resistant structures throughout evolution (Yang et al. 2010).

The misfolding avoidance hypothesis predicts that highly expressed proteins should have evolved to be highly robust to misfolding (including spontaneous misfolding and misfolding induced by translation errors). One way in which natural selection may have shaped highly expressed proteins to exhibit this robustness is by making them highly thermostable (i.e., with a highly negative free energy of folding, $\Delta G_{folding} = G_{folded} - G_{unfolded}$, hereafter called ΔG). This trend has received some indirect support from analyses based on proxies of protein thermostability, such as protein composition and interatomic contact density (Cherry 2010b; Serohijos et al. 2013), and from simulation studies (Wilke and Drummond 2006; Drummond and Wilke 2008; Yang et al. 2010; Serohijos et al. 2012). However, despite the potentially far-reaching implications of such a trend, attempts to test for it using empirical data have produced contradictory and controversial results. The main reason is that proteins' ΔG is hard to measure experimentally or to estimate computationally, and thus data are scarce, and that using indirect measures of ΔG such as melting temperature (T_m) is potentially problematic (Razban 2019).

Yang et al. (2010) found no correlation between the ΔG of five yeast proteins and their protein abundances. Using data for a small set of *Escherichia coli* proteins (n = 23–28), Plata et al. (2010) found no correlation between mRNA abundance and ΔG or T_m (which they used as a proxy of ΔG). Using T_m values for hundreds of proteins obtained from cell lysates, Leuenberger et al. (2017) found that, in *E. coli* (but not in yeast or human), highly thermostable proteins are more abundant than lowly thermostable ones. Using Leuenberger et al.'s data set, Plata and Vitkup (2018) found a weakly positive correlation between protein abundance and T_m in *E. coli* (largely driven by ribosomal proteins) but, surprisingly, a negative correlation in yeast and human. Razban (2019) noticed that Leuenberger et al.'s proteome-scale measurements of T_m are subjected to noise and that the relationship between T_m and ΔG is not simple and is confounded by protein length, making T_m inappropriate to test whether highly abundant proteins are highly thermostable. After correcting for these effects, he found a positive correlation between protein abundance and thermostability in *E. coli*, yeast, and human. Most recently, using an expanded set of proteins, Usmanova, Plata and Vitkup (2021) found no correlation between protein abundance and ΔG in *E. coli* ($n = 28$) or human ($n = 42$).

Another way in which natural selection may have made highly expressed proteins robust to misfolding is by altering their sequence in such a way that translation errors have a small impact on protein structure (translation error robustness). If that is the case, we would expect that, for highly expressed proteins, the $\Delta\Delta G$ of translation errors ($\Delta\Delta G_t$ = ΔG of the incorrectly translated protein – ΔG of the correctly translated protein) would be on average less positive than for lowly expressed proteins. However, to our knowledge, this trend has not been tested.

In this work, we computationally test whether highly expressed proteins are highly thermostable and/or highly robust to translation errors by examining a set of pairs of orthologous proteins between human and mouse that are expressed at different levels and have a solved structure in the Protein Data Bank (PDB) (Berman et al. 2000). We show that, in these pairs, the most highly expressed protein is often the most thermostable (the one with a more negative ΔG of folding), indicating that highly expressed proteins are often more thermostable. However, the average $\Delta\Delta G$ of translation errors does not correlate with protein abundance, indicating that highly expressed proteins do not exhibit structures that are more robust to translation errors.

Results

Highly Abundant Proteins Are More Thermostable Than Their Less Abundant Orthologs

We first considered whether highly abundant proteins exhibit high thermostability, as measured from a highly negative free energy of folding (ΔG). Whereas a protein's absolute ΔG is generally unknown and very difficult to estimate computationally (Chen et al. 2008; Bigman and Levy 2018), a number of algorithms allow reliably estimating the difference between the ΔG of two homologous proteins ($\Delta\Delta G$ of mutation), taking as input the structure of one of the proteins and the sequence of the other (Schymkowitz et al. 2005; Delgado et al. 2019). We used this approach

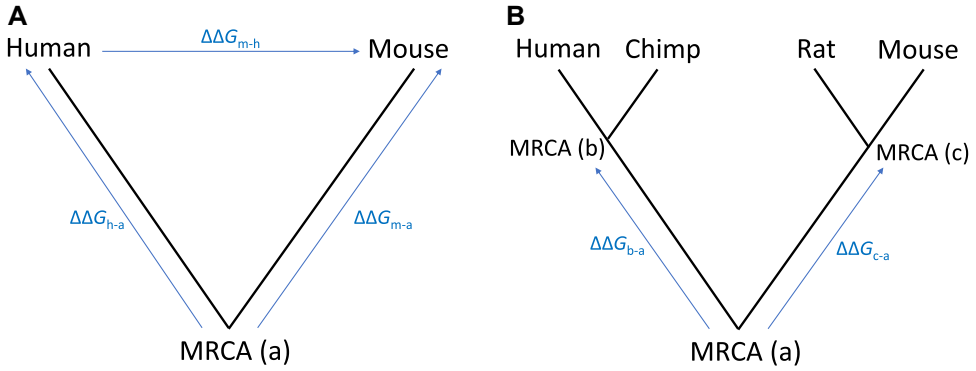

Fig. 1.—Changes in free energy of folding due to mutations ($\Delta\Delta G$ of mutations) used in the study. For each term, the direction of the arrow indicates the initial sequence and the final sequence. Positive $\Delta\Delta G$ values indicate that the final sequence is less stable than the initial sequence. Negative $\Delta\Delta G$ values indicate that the final sequence is more stable than the initial sequence.

to compare the ΔG of pairs of human–mouse orthologous proteins.

We obtained a list of 51 pairs of human–mouse orthologous genes with a protein structure available for at least one of the two orthologs (the available protein structures corresponded to human in 41 cases and to mouse in the other ten cases) and whose gene tree and protein structure met the filtering criteria described in the Materials and Methods section. For each pair of orthologs, we used FoldX 5.0 (Delgado et al. 2019) to estimate $\Delta\Delta G_{m\text{-}h}$: the difference between the free energy of folding of the mouse protein (ΔG_m) and the free energy of folding of the human protein (ΔG_h; fig. 1*A*). These estimations were generated by combining the effects of all individual amino acid differences between human and mouse. For 39 of the pairs, we obtained a positive $\Delta\Delta G_{m\text{-}h}$, indicating that the human protein is more stable than the mouse protein. For the other ten pairs, $\Delta\Delta G_{m\text{-}h}$ was negative, indicating that the mouse protein is more stable than the human one (supplementary table S1, Supplementary Material online). The other two proteins are identical between human and mouse, and thus their $\Delta\Delta G_{m\text{-}h}$ is 0.

For 49 of these proteins, protein abundance data were available for both human and mouse. In 25 cases, abundance was higher in human, whereas in 24 cases, it was higher in mouse. We found a moderate but significantly negative correlation between $\Delta\Delta G_{m\text{-}h}$ and the ratio $R_{m/h}$ = abundance in mouse/abundance in human (Spearman's rank correlation coefficient, $\rho = -0.31$, $n = 49$, $P = 0.022$), indicating that proteins that are more highly abundant in mouse also tend to be more stable in mouse. Removing the two proteins that are identical between human and mouse (at the amino acid residues that are covered by the available protein structures) produced equivalent results ($\rho = -0.36$, $n = 47$, $P = 0.013$; fig. 2). Also equivalent results were obtained when restricting the analyses to the proteins for which a human PDB is available ($\rho = -0.43$, $n = 38$, $P = 0.007$; supplementary fig. S1*A*, Supplementary Material online). In addition, the $\Delta\Delta G$ values of all individual amino acid differences between the human and mouse proteins significantly correlated with $R_{m/h}$ ($\rho = -0.01$, $n = 1302$, $P = 0.005$; supplementary fig. S2, Supplementary Material online).

Thermostability Changes in the Human versus Mouse Lineages

For 47 of the 51 pairs of human–mouse orthologs, we were able to identify orthologs in a number of outgroup species, which we used to infer the protein sequences of the most recent common ancestor (MRCA) of human and mouse (node "a" in fig. 1). For each protein, ten possible ancestral proteins were sampled from the posterior distribution. This approach has been shown to correct for sampling biases and to produce ancestral sequences with more realistic amino acid compositions and biochemical properties (Williams et al. 2006). We then used the inferred ancestral sequences and FoldX 5.0 to estimate the difference in the free energy of folding between the MRCA and either the human or mouse sequences. We computed $\Delta\Delta G_{h\text{-}a}$ as the difference between the free energy of folding of the human protein (ΔG_h) and the free energy of folding of the ancestral protein (ΔG_a), taking the average across the ten inferred ancestral sequences (fig. 1*A*). Similarly, $\Delta\Delta G_{m\text{-}a}$ was computed as the difference between the free energy of folding of the mouse protein (ΔG_m) and ΔG_a, averaged across the ten inferred ancestral sequences (fig. 1*A*). The values of $\Delta\Delta G_{m\text{-}a}$ tended to be higher than those of $\Delta\Delta G_{h\text{-}a}$ (median $\Delta\Delta G_{h\text{-}a}$: −0.935, median $\Delta\Delta G_{m\text{-}a}$: 2.494, Mann–Whitney *U* test, $P = 0.0001$; supplementary fig. S3*A*, Supplementary Material online); however, the trend was inverted when only proteins with an available mouse structure were included in the analysis (median $\Delta\Delta G_{h\text{-}a}$: −4.658, median $\Delta\Delta G_{m\text{-}a}$: −3.162, Mann–Whitney *U* test, $P = 0.003$; supplementary fig. S3*C*, Supplementary Material online). Similar results were obtained when analyzing

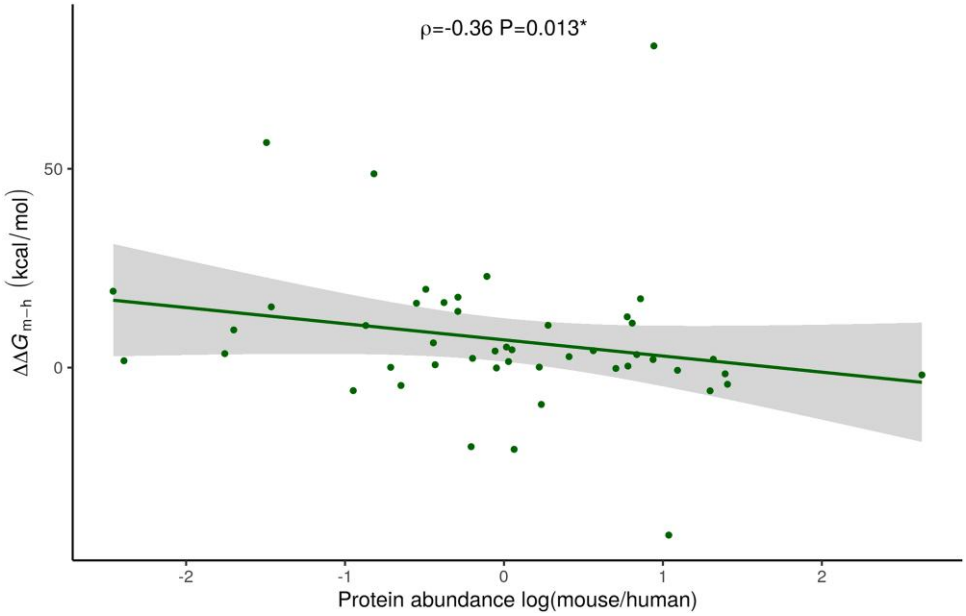

Fig. 2.—Correlation between the difference in the free energy of folding of human and mouse proteins and the difference in the abundance of human and mouse. Each dot corresponds to a pair of human–mouse orthologs ($n = 47$). The shaded area represents the 95% confidence interval. Spearman's rank correlation test significance level: *, $P < 0.05$.

the $\Delta\Delta G$ values of individual amino acid changes that accumulated along the human and mouse branches (supplementary figs. S4 and S5, Supplementary Material online). This suggests the possibility of a bias introduced by computational methods for structure determination, but as indicated, it has been controlled for by separating the analysis by the organism that the solved structure came from.

By using protein abundances from a number of outgroup species, we were able to infer the protein abundance in the MRCA of human and mouse for 38 of the 47 pairs of human–mouse orthologs. For each protein, we used these ancestral protein abundances to estimate the ratio of abundance change within each branch as the ratios $R_{h/a}$ = abundance in human/abundance in the MRCA and $R_{m/a}$ = abundance in mouse/abundance in the MRCA. The correlation between $\Delta\Delta G_{h\text{-}a}$ and $R_{h/a}$ ($\rho = -0.01$ $n = 38$, $P = 0.995$; fig. 3*A*) was not significant. However, the correlation between $\Delta\Delta G_{m\text{-}a}$ and $R_{m/a}$ ($\rho = -0.33$, $n = 38$, $P = 0.004$; fig. 3*C*) was significant, indicating that increases in protein abundance along the rodent branch resulted in increases in protein thermostability. Within each branch, no significant differences were observed between the $\Delta\Delta G$ values of individual amino acid changes that occurred on proteins whose expression increased versus on proteins whose expression decreased (Mann–Whitney *U* test, human branch: $P = 0.328$, mouse branch: $P = 0.06$; supplementary fig. S4, Supplementary Material online).

A fraction of the observed amino acid differences between the human and mouse proteins represents polymorphisms rather than fixed mutations. Polymorphic nonsynonymous mutations are more likely than fixed nonsynonymous substitutions to be destabilizing (Saunders and Baker 2002; Hunt et al. 2014; Baugh et al. 2016; Ancien et al. 2018), which may be biasing our results. To discard this possibility, we added chimpanzee and rat to our alignments and inferred the sequences of the MRCA of human and chimpanzee (node "b" in fig. 1*B*) and the MRCA of mouse and rat (node "c" in fig. 1*B*). For each node, ten ancestral sequences were sampled from the posterior distribution. Presumably, the inferred ancestral sequences are virtually free from polymorphic mutations, since the MRCA of human and chimpanzee existed 4–6 million years ago (Chimpanzee Sequencing and Analysis Consortium 2005) and the MRCA of mouse and rat existed 12–24 million years ago (Gibbs et al. 2004). We then used FoldX 5.0 to estimate $\Delta\Delta G_{b\text{-}a}$ (the change in the free energy of folding from the MRCA of human and mouse to the MRCA of human and chimpanzee) and $\Delta\Delta G_{c\text{-}a}$ (the change in the free energy of folding from the MRCA of human and mouse to the MRCA of mouse and rat; fig. 1*B*). The average $\Delta\Delta G_{b\text{-}a}$ was 0.131 kcal/mol and the average $\Delta\Delta G_{c\text{-}a}$ was 4.019 kcal/mol. In 33 out of the 47 groups of orthologs, $\Delta\Delta G_{c\text{-}a}$ was higher than $\Delta\Delta G_{b\text{-}a}$ (binomial test, $P = 0.007$; paired Wilcoxon signed-rank test, $n = 47$, $P = 0.002$), suggesting again that proteins overall accumulated more destabilizing changes along the rodent branch than along the primate branch. Whereas $\Delta\Delta G_{b\text{-}a}$ did not significantly correlate with $R_{h/a}$ ($\rho = -0.01$, $n = 38$, $P = 0.696$; fig. 3*B*), $\Delta\Delta G_{c\text{-}a}$ did negatively correlate with $R_{m/a}$ ($\rho = -0.33$, $n = 38$, $P = 0.041$; fig. 3*D*), confirming that increases in protein abundance along the rodent branch resulted in

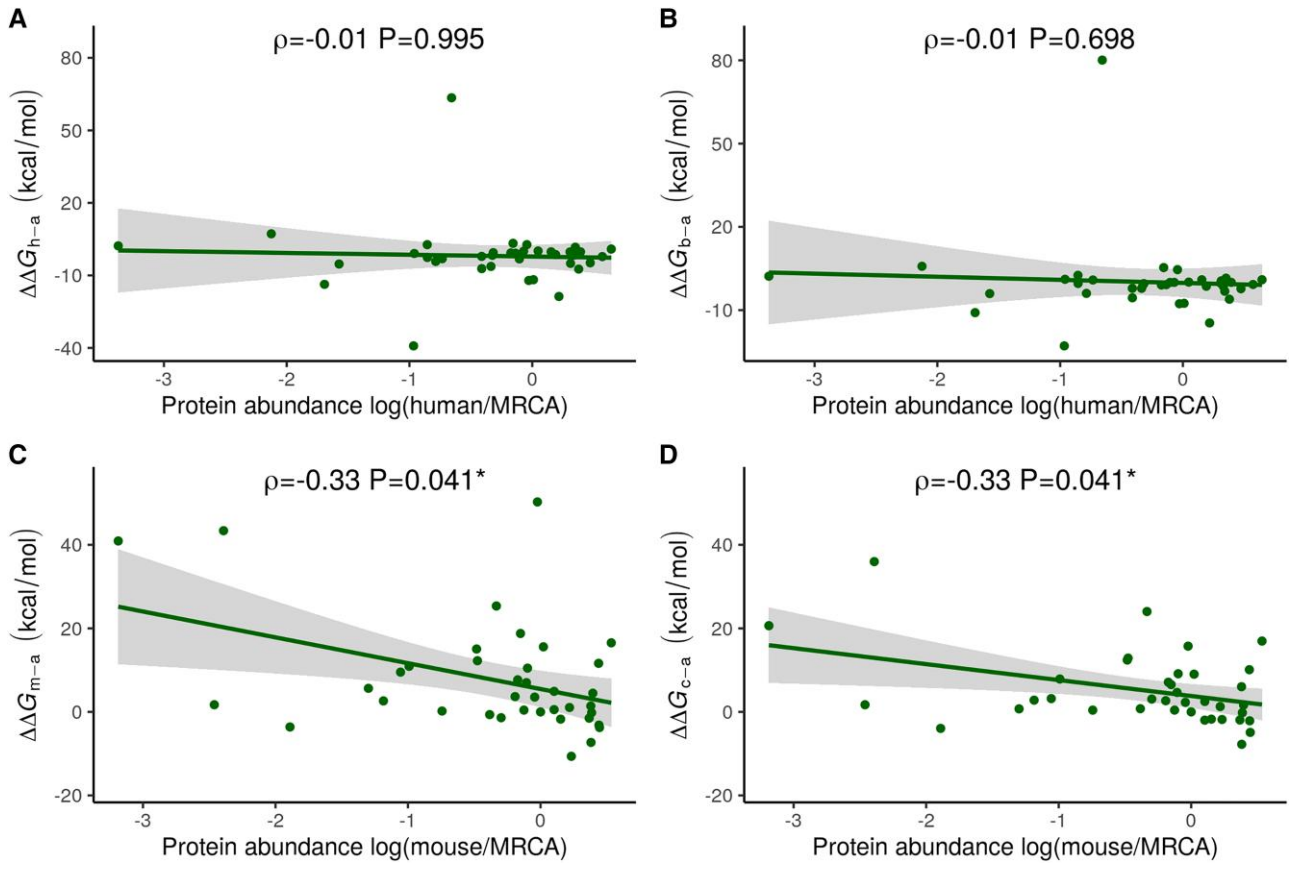

Fig. 3.—Correlation between changes in free energy of folding ($\Delta\Delta G$) and changes in protein abundance at different branches of the primate and rodent phylogeny. Each dot corresponds to a human or mouse protein. The shaded areas represent the 95% confidence intervals. Spearman's rank correlation test significance level: *, $P < 0.05$.

increases in protein thermostability. Of note, more mutations accumulated along the mouse branch than along the human one (supplementary table S1, Supplementary Material online), consistent with the faster substitution rates of rodents (Wu and Li 1985), which probably results in a higher statistical power to detect correlations in this lineage.

Highly Abundant Proteins Are Not Highly Robust to Translation Errors

We then considered whether highly abundant proteins exhibit structures that are highly robust to translation errors, as measured from the average and median change in free energy resulting from all possible translation errors (average $\Delta\Delta G$ of translation error or $\Delta\Delta G_t$). For each of the 51 protein structures in our data set, we used FoldX 5.0 to estimate the $\Delta\Delta G_t$ resulting from each possible translation error (n = length of the protein × 19 for each protein) and obtained the average, the median, and the fraction of destabilizing translation errors (those with $\Delta\Delta G_t > 1$). Neither of the three metrics correlated with protein abundances (average $\Delta\Delta G_t$: $\rho = -0.14$, $n = 51$, $P = 0.323$, fig. 4*A*; median $\Delta\Delta G_t$: $\rho = -0.15$, $n = 51$, $P = 0.286$, fig. 4*B*; fraction of destabilizing translation errors: $\rho = -0.03$, $n = 51$, $P = 0.821$, fig. 4*C*), indicating that highly abundant proteins are not more robust to translation errors than lowly abundant ones.

Because not all translation errors are equally likely, we repeated our analyses giving more weight to more frequent translation errors. For each protein, we randomly sampled 100,000 translation errors. The likelihood of sampling each translation error was proportional to the probability of occurrence of the error (see Materials and Methods section). Again, neither of the three metrics correlated with protein abundances (average $\Delta\Delta G_t$: $\rho = -0.19$, $n = 51$, $P = 0.171$, supplementary fig. S6*A*, Supplementary Material online; median $\Delta\Delta G_t$: $\rho = -0.06$, $n = 51$, $P = .681$, supplementary fig. S6*B*, Supplementary Material online; fraction of destabilizing translation errors: $\rho = -0.05$, $n = 51$, $P = 0.741$; supplementary fig. S6*C*, Supplementary Material online).

Discussion

There exist at least two nonmutually exclusive evolutionary paths by which proteins may have become highly robust to

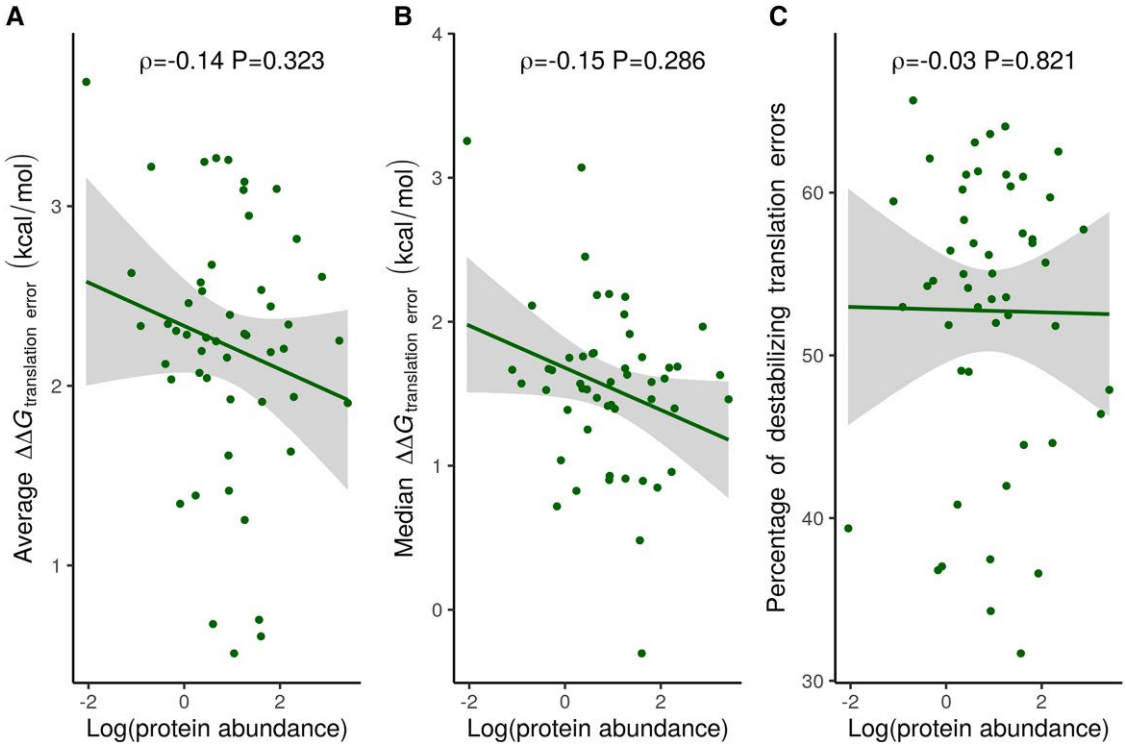

Fig. 4.—Correlation between the differences in the free energy of folding due to all possible translation errors ($\Delta\Delta G_t$) and protein abundance. Each dot corresponds to one human or mouse protein structure ($n = 51$). Average and median $\Delta\Delta G_t$ values were estimated by substituting every amino acid position with the other 19 standard amino acids. The percent of destabilizing translation errors was computed as the fraction of translation errors with $\Delta\Delta G_t > 1$. The shaded areas represent the 95% confidence intervals.

misfolding. First, they may have evolved to have a highly negative free energy of folding (ΔG); second, they may have evolved to have structures that can properly fold despite some translation errors (leading to a low average $\Delta\Delta G$ of translation error). By comparing pairs of human–mouse orthologous proteins, we show a negative correlation between the mouse/human abundance ratio and the difference between the thermostabilities of the mouse and human proteins (fig. 2 and supplementary fig. S2, Supplementary Material online). These results indicate that the most highly expressed ortholog tends to be the most stable. Consistent results were obtained when analyzing mutations that accumulated along the branch separating the MRCA of human and mouse (node "a" in fig. 1) and mouse (fig. 3*C*), and when analyzing mutations that accumulated along the branch separating the MRCA of human and mouse and the MRCA of mouse and rat (node "c" in fig. 1) (fig. 3*D*). One would expect that the strength of selection for an expression effect would be amplified by organismal effective population size. The continuous nature of the relationship in figure 2 does not show a strong effect for this, in that there is no significant difference for expression ratios above and below 1. Because this slope is the same, this suggests that the strength of the effect is not significantly different between higher effective population size mice and lower effective population size humans. Furthermore, highly abundant proteins do not exhibit a lower average or median $\Delta\Delta G$ of translation error (fig. 4 and supplementary fig. S6, Supplementary Material online). Thus, our results lend support to the first, but not the second evolutionary path. It is possible, however, that future analyses using larger data sets lend support to the second evolutionary path too. The second effect is expected to be a weaker effect, as it is a specific secondary effect. Selection depends upon enough specific translational errors to occur at individual sites rather than acting on the $\Delta G_{folding}$ for the whole protein. This has a similarity to other secondary quality control mechanisms in evolution, where local secondary selection has been observed to act only in larger effective population size species (Xiong et al. 2017).

Previous analyses of the relationship between protein abundances and ΔG have relied on ΔG values generated in different laboratories and under different experimental conditions, deposited in the ProTherm database (Nikam et al. 2021). In contrast, our analyses rely on $\Delta\Delta G$ values that have been inferred using the same method across all orthologous pairs.

Protein thermostabilities are affected by a number of factors. Thus, a proper test of the relationship between protein abundance and thermostability should include appropriate controls for these factors. One factor affecting protein thermostabilities is protein length, given that long proteins are on average more capable of establishing stabilizing intramolecular interactions and have larger hydrophobic cores (Chan and Dill 1991; Kumar et al. 2000; Ghosh and Dill 2009). Other factors are the number of protein–protein interactions, whether proteins are part of protein complexes, and whether they are folded by chaperones, because intermolecular interactions can contribute to stabilizing proteins (Gershenson and Gierasch 2011; Chi and Liberles 2016; Leuenberger et al. 2017). Nonetheless, our analysis is based on comparison of pairs of human–mouse orthologous proteins, most likely with identical or very similar length and interaction patterns, which is expected to remove any potential effect of these factors. In addition, we discarded from our analyses those proteins whose structures had been solved in complex with other proteins and/or cofactors (other than metal ions).

Unexpectedly, we found that, on average, amino acid substitutions that accumulated along the mouse branch were more destabilizing than those that accumulated along the human branch (supplementary fig. S3, Supplementary Material online). Primates exhibit a lower effective population size (N_e) than rodents (Ohta 1993; Chimpanzee Sequencing and Analysis Consortium 2005), and theory predicts that more destabilizing changes will accumulate in populations with low N_e (Goldstein 2013). Our observations may be partially biased by the fact that, for most of the pairs of human–mouse orthologs used in our study, the human protein structure ($n = 41$) rather than the mouse one ($n = 10$) is available. This may have resulted in inferred ancestral protein structures often resembling more the human structure than the mouse one. In support of this possibility, when we restricted our analyses to the orthologous pairs for which the mouse protein structure is available, the trend inverted: Mutations that accumulated along the human branch were more destabilizing than those that accumulated along the mouse branch (supplementary figs. S3C, S4, and S5, Supplementary Material online). On the other hand, the tendency for mouse proteins to be less thermostable can be observed in figure 2, which is based on analyses that do not rely on ancestral protein sequence or structure reconstruction. The figure shows that $\Delta\Delta G_{m\text{-}h}$ is positive in 37 and negative in ten cases and that the regression line exhibits a positive *y*-intercept—that is, the linear model predicts proteins with equal abundance in human and mouse to be less stable in mouse than in human. These results suggest that mouse proteins may indeed be less thermostable than human ones. In any case, we do not expect this potential bias to affect the main conclusions of our work, because 1) our key results (the negative correlation between $\Delta\Delta G_{m\text{-}h}$ and $R_{m/h}$ shown in fig. 2) do not rely on ancestral protein structure reconstruction and 2) they remain significant after removing the pairs of orthologs for which the mouse protein structure is available (correlation between $\Delta\Delta G_{m\text{-}h}$ and $R_{m/h}$: $\rho = -0.43$, $n = 38$, $P = 0.007$, supplementary fig. S1*A*, Supplementary Material online; correlation between $\Delta\Delta G_{c\text{-}a}$ and $R_{m/a}$: $\rho = -0.46$, $n = 38$, $P = 0.008$).

Our observation that highly expressed proteins are highly thermostable (figs. 2 and 3 and supplementary fig. S2, Supplementary Material online) is one of the predictions of the translational robustness hypothesis (Drummond et al. 2005) and its extension, the misfolding avoidance hypothesis (Yang et al. 2010). However, we would like to clarify that our results, on their own, are insufficient to fully demonstrate that these hypotheses are correct. It is possible that higher thermostability results from reducing surface hydrophobic content, a prediction of the misinteraction avoidance hypothesis. Demonstrating whether these hypotheses are correct, or favoring them over alternative hypotheses, is beyond the scope of the current work. Along these lines, we note that the strength of the correlations that we observe is rather weak. Nonetheless, they have potentially far-reaching implications for our understanding protein structure and evolution.

Materials and Methods

Human–Mouse Pairs of Orthologs Selection

Curated reconciled gene family trees including human or mouse members were obtained from The Adaptive Evolution Database (TAED; Liberles et al. 2001; Hermansen et al. 2017). Gene trees containing exactly one member each in human and mouse (ignoring subspecies) were identified and collected into a list. Integrated whole-organism protein abundance data for each identified gene were collected from the human (9,606) and mouse (10,090) data sets in the protein abundance database PaxDB v4.0 (Wang et al. 2015). Protein structures for each gene family were identified using the information in TAED and retrieved from the PDB (Berman et al. 2000). All ortholog pairs with an available PDB structure with a human or mouse source were collected and subjected to quality control. Gene or protein family names containing the terms "LOW QUALITY," "partial," or "fragment," were removed, as were pairs from gene families with fewer than five known members. PDB structures were individually scrutinized and rejected if they contained bound substrates or inhibitors, or cofactors other than metal ions, or were crystallized in complex with other proteins. Structures with stability-affecting mutations, membrane proteins, and structures of individual domains or motifs were also excluded. This procedure afforded 51 mouse–human ortholog pairs as the final data set.

Multiple Sequence Alignment and Ancestral Sequence Inference

For each protein, we obtained the amino acid sequences of human, mouse, chimpanzee, rat, and four outgroups (dog, cat, cow, and African elephant, where available) from the UniProt database (Apweiler et al. 2004). We then aligned the sequences using the MEGA X software (Kumar et al. 2018) and used the maximum likelihood algorithm implemented in the FastML v3.11 program (Ashkenazy et al. 2012) to infer the sequences of the MRCA of human and mouse (node "a" in fig. 1), the MRCA of human and chimpanzee (node "b"), and the MRCA of mouse and rat (node "c"). Using the marginal reconstruction estimates, ten sequences were sampled from the posterior distribution for each ancestral node. These sampled sequences were used in further $\Delta\Delta G$ calculations (averaging $\Delta\Delta G$ values across the ten sequences). For ancestral sequence reconstruction, we employed the JTT + G substitution model and a phylogenetic tree retrieved from the TimeTree platform v5.0 (Kumar et al. 2022).

Ancestral Protein Abundance Inference

For each protein, we retrieved whole-organism protein abundances for human, mouse, and four outgroups (horse, dog, cow, and pig, where available) from the PaxDB database v4 (Wang et al. 2015). We then used the anc.ML function from the phytools R package v1.2 (Revell 2012) to estimate the abundance for the MRCA of human and mouse. The anc.ML method uses a maximum likelihood framework to infer ancestral continuous traits. For each protein, we estimated the ancestral protein abundance under the Brownian motion (Felsenstein 1973; Schluter et al. 1997) and Ornstein–Uhlenbeck (Felsenstein 1988) models, and we used the Akaike information criterion (AIC) to select the best model for our estimations. In all cases, the Brownian motion model was the one exhibiting the best fit. The phylogenetic tree used in this analysis was retrieved from the TimeTree platform v5.0 (Kumar et al. 2022). Inferences were only carried out when protein abundance data were available for at least two of the outgroup species.

Estimation of $\Delta\Delta G$ of Amino Acid Substitutions

We used the FoldX 5.0 program (Delgado et al. 2019) to infer differences in ΔG between pairs of human and mouse orthologous proteins ($\Delta\Delta G_{m\text{-}h}$). The program takes as input a protein's three-dimensional structure (which we obtained from the PDB; Berman et al. 2000) and a set of amino acid changes (which we inferred from human–mouse protein sequence alignments). The software then uses an algorithm based on an empirical force field to infer the effect of each amino acid change on protein stability, as well as the combined effect of all changes. Throughout this manuscript, we use combined $\Delta\Delta G$ values for each protein, unless noted otherwise (supplementary figs. S2, S4, and S5, Supplementary Material online). Our analyses involved two steps: We first optimized the PDB files using the RepairPDB command, and we then ran the BuildModel function, completing five cycles for each amino acid change.

In 41 cases, we used the human protein structure as input to infer $\Delta\Delta G_{m\text{-}h}$ directly. In ten cases, we used the mouse structure as input, and then we inverted the sign of the resulting $\Delta\Delta G$ estimates to obtain $\Delta\Delta G_{m\text{-}h}$. Amino acid changes (or combinations of amino acid changes) with negative $\Delta\Delta G$ values are expected to be stabilizing, whereas those with positive $\Delta\Delta G$ values are expected to be destabilizing. Thus, positive (negative) $\Delta\Delta G_{m\text{-}h}$ values indicate that the mouse protein is less (more) stable than the human protein.

For each group of orthologs, we inferred the structure of the protein of the MRCA of human and mouse (node "a" in fig. 1) from the available PDB structure and the ten sampled ancestral sequences using the BuildModel function. For each sampled ancestral sequence, a separate structural model was built. We then used the BuildModel function again to infer ten $\Delta\Delta G_{h\text{-}a}$, ten $\Delta\Delta G_{m\text{-}a}$, ten $\Delta\Delta G_{b\text{-}a}$, and ten $\Delta\Delta G_{c\text{-}a}$ values for each group of orthologs, completing three cycles for each amino acid change. Each of the ten $\Delta\Delta G_{h\text{-}a}$ and $\Delta\Delta G_{m\text{-}a}$ values was inferred by picking one of the ten structural models for node "a" and the sequence of the relevant species (human or mouse, respectively); the ten values were then averaged, and the resulting average was used in all our analyses (fig. 3 and supplementary fig. S3 and table S1, Supplementary Material online). Each of the ten $\Delta\Delta G_{b\text{-}a}$ and $\Delta\Delta G_{c\text{-}a}$ values was inferred by picking one of the ten structural models for node "a" and one of the ten ancestral sequences sampled for the relevant internal node (nodes "b" or "c," respectively); the ten values were then averaged, and the resulting average was used in all our analyses (fig. 3 and supplementary fig. S3 and table S1, Supplementary Material online).

Estimation of $\Delta\Delta G$ of Translation Errors

We used FoldX 5.0 (Delgado et al. 2019) to infer the effect of each possible translation error on protein stability ($\Delta\Delta G$ of translation error or $\Delta\Delta G_t$). Using repaired PDB files, each amino acid was "mutated" to all other 19 standard amino acids by conducting a single run of the PositionScan function.

Then, for each protein, we randomly sampled 100,000 translation errors. In each drawing, the probability of picking a given translation error was proportional to the likelihood of the translation error occurring, taking into account that 1) unpreferred codons are approximately five times more likely to undergo translation errors (Drummond and Wilke 2008) and 2) the translation error table inferred by Yang et al. (2010). Coding sequences

(CDSs) were obtained from the Ensembl genome database (Hubbard et al. 2002), and a list of human preferred codons was obtained from the Codon Statistics Database (Subramanian et al. 2022).

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

This work was supported by grants MCB 1818288 and MCB 1817413 from the National Science Foundation.

Data Availability

All data used in this work are publicly available, as described in the Materials and Methods section. Accession information is shown in supplementary table S2, Supplementary Material online.

Literature Cited

Alvarez-Ponce D. 2014. Why proteins evolve at different rates: the determinants of proteins' rates of evolution. In: Fares MA, editor. Natural selection: methods and applications. London: CRC Press (Taylor & Francis). p. 126–178.

Ancien F, Pucci F, Godfroid M, Rooman M. 2018. Prediction and interpretation of deleterious coding variants in terms of protein structural stability. Sci Rep. 8:4480.

Apweiler R, et al. 2004. Uniprot: the universal protein knowledgebase. Nucleic Acids Res. 32:D115–D119.

Ashkenazy H, et al. 2012. FastML: a web server for probabilistic reconstruction of ancestral sequences. Nucleic Acids Res. 40: W580–W584.

Baugh EH, et al. 2016. Robust classification of protein variation using structural modelling and large-scale data integration. Nucleic Acids Res. 44:2501–2513.

Berman HM, et al. 2000. The Protein Data Bank. Nucleic Acids Res. 28: 235–242.

Bigman LS, Levy Y. 2018. Stability effects of protein mutations: the role of long-range contacts. J Phys Chem B. 122:11450–11459.

Chan HS, Dill KA. 1991. Polymer principles in protein structure and stability. Annu Rev Biophys Biophys Chem. 20:447–490.

Chen Y, et al. 2008. Protein folding: then and now. Arch Biochem Biophys. 469:4–19.

Cherry JL. 2010a. Expression level, evolutionary rate, and the cost of expression. Genome Biol Evol. 2:757–769.

Cherry JL. 2010b. Highly expressed and slowly evolving proteins share compositional properties with thermophilic proteins. Mol Biol Evol. 27:735–741.

Chi PB, Liberles DA. 2016. Selection on protein structure, interaction, and sequence. Protein Sci. 25:1168–1178.

Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437:69–87.

Delgado J, Radusky LG, Cianferoni D, Serrano L. 2019. FoldX 5.0: working with RNA, small molecules and a new graphical interface. Bioinformatics 35:4168–4169.

Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. 2005. Why highly expressed proteins evolve slowly. Proc Natl Acad Sci U S A. 102:14338–14343.

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

Felsenstein J. 1973. Maximum-likelihood estimation of evolutionary trees from continuous characters. Am J Hum Genet. 25(5): 471–492.

Felsenstein J. 1988. Phylogenies and quantitative characters. Ann Rev Ecol Syst. 19(1):445–471.

Gershenson A, Gierasch LM. 2011. Protein folding in the cell: challenges and progress. Curr Opin Struct Biol. 21:32–41.

Ghosh K, Dill KA. 2009. Computing protein stabilities from their chain lengths. Proc Natl Acad Sci U S A. 106:10649–10654.

Gibbs RA, et al. 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature 428: 493–521.

Goldstein RA. 2013. Population size dependence of fitness effect distribution and substitution rate probed by biophysical model of protein thermostability. Genome Biol Evol. 5:1584–1593.

Gout JF, Kahn D, Duret L, Consortium PP-G. 2010. The relationship among gene expression, the evolution of gene dosage, and the rate of protein evolution. PLoS Genet. 6:e1000944.

Hermansen RA, et al. 2017. The Adaptive Evolution Database (TAED): a new release of a database of phylogenetically indexed gene families from chordates. J Mol Evol. 85:46–56.

Hubbard T, et al. 2002. The Ensembl genome database project. Nucleic Acids Res. 30:38–41.

Hunt RC, Simhadri VL, Iandoli M, Sauna ZE, Kimchi-Sarfaty C. 2014. Exposing synonymous mutations. Trends Genet. 30: 308–321.

Kumar S, et al. 2022. TimeTree 5: an expanded resource for species divergence times. Mol Biol Evol. 39(8):msac174.

Kumar S, Stecher G, Li M, Knyaz C, Tamura K. 2018. MEGA X: molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol. 35:1547–1549.

Kumar S, Tsai CJ, Nussinov R. 2000. Factors enhancing protein thermostability. Protein Eng. 13:179–191.

Leuenberger P, et al. 2017. Cell-wide analysis of protein thermal unfolding reveals determinants of thermostability. Science 355(6327):eaai7825.

Liberles DA, Schreiber DR, Govindarajan S, Chamberlin SG, Benner SA. 2001. The Adaptive Evolution Database (TAED). Genome Biol. 2: RESEARCH0028.

Liberles DA, Tisdell MD, Grahnen JA. 2011. Binding constraints on the evolution of enzymes and signalling proteins: the important role of negative pleiotropy. Proc Biol Sci. 278:1930–1935.

Nikam R, Kulandaisamy A, Harini K, Sharma D, Gromiha MM. 2021. ProThermDB: thermodynamic database for proteins and mutants revisited after 15 years. Nucleic Acids Res. 49:D420–D424.

Ohta T. 1993. Amino acid substitution at the *Adh* locus of *Drosophila* is facilitated by small population size. Proc Natl Acad Sci U S A. 90:4548–4551.

Pál C, Papp B, Hurst LD. 2001. Highly expressed genes in yeast evolve slowly. Genetics 158:927–931.

Plata G, Gottesman ME, Vitkup D. 2010. The rate of the molecular clock and the cost of gratuitous protein synthesis. Genome Biol. 11:R98.

Plata G, Vitkup D. 2018. Protein stability and avoidance of toxic misfolding do not explain the sequence constraints of highly expressed proteins. Mol Biol Evol. 35:700–703.

Razban RM. 2019. Protein melting temperature cannot fully assess whether protein folding free energy underlies the universal abundance-evolutionary rate correlation seen in proteins. Mol Biol Evol. 36:1955–1963.

Revell LJ. 2012. phytools: an R package for phylogenetic comparative biology (and other things). Methods Ecol Evol. 3(2):217–223.
Saunders CT, Baker D. 2002. Evaluation of structural and evolutionary contributions to deleterious mutation prediction. J Mol Biol. 322: 891–901.
Schluter D, Price T, Mooers AØ, Ludwig D. 1997. Likelihood of ancestor states in adaptive radiation. Evolution 51(6):1699–1711.
Schymkowitz J, et al. 2005. The FoldX web server: an online force field. Nucleic Acids Res. 33:W382–W388.
Serohijos AW, Lee SY, Shakhnovich EI. 2013. Highly abundant proteins favor more stable 3D structures in yeast. Biophys J. 104:L1–L3.
Serohijos AW, Rimas Z, Shakhnovich EI. 2012. Protein biophysics explains why highly abundant proteins evolve slowly. Cell Rep. 2: 249–256.
Subramanian K, Payne B, Feyertag F, Alvarez-Ponce D. 2022. The codon statistics database: a database of codon usage bias. Mol Biol Evol. 39(8):msac157.
Usmanova DR, Plata G, Vitkup D. 2021. The relationship between the misfolding avoidance hypothesis and protein evolutionary rates in the light of empirical evidence. Genome Biol Evol. 13: evab006.
Wang M, Herrmann CJ, Simonovic M, Szklarczyk D, von Mering C. 2015. Version 4.0 of PaxDb: protein abundance data, integrated across model organisms, tissues, and cell-lines. Proteomics 15: 3163–3168.
Wilke CO, Drummond DA. 2006. Population genetics of translational robustness. Genetics 173:473–481.
Williams PD, Pollock DD, Blackburne BP, Goldstein RA. 2006. Assessing the accuracy of ancestral protein reconstruction methods. PLoS Comput Biol. 2(6):e69.
Wu CI, Li WH. 1985. Evidence for higher rates of nucleotide substitution in rodents than in man. Proc Natl Acad Sci U S A. 82: 1741–1745.
Xiong K, McEntee JP, Porfirio DJ, Masel J. 2017. Drift barriers to quality control when genes are expressed at different levels. Genetics 205: 397–407.
Yang JR, Liao BY, Zhuang SM, Zhang J. 2012. Protein misinteraction avoidance causes highly expressed proteins to evolve slowly. Proc Natl Acad Sci U S A. 109:E831–E840.
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.
Zhang J, Yang JR. 2015. Determinants of the rate of protein sequence evolution. Nat Rev Genet. 16:409–420.

Associate editor: Mario dos Reis