Protein Stability and Avoidance of Toxic Misfolding Do Not Explain the Sequence Constraints of Highly Expressed Proteins

Germán Plata[1] and Dennis Vitkup*[,1,2]

[1]Department of Systems Biology, Columbia University, New York, NY

[2]Department of Biomedical Informatics, Columbia University, New York, NY

*Corresponding author: E-mail: dv2121@columbia.edu.

Associate editor: Csaba Pal

Abstract

The avoidance of cytotoxic effects associated with protein misfolding has been proposed as a dominant constraint on the sequence evolution and molecular clock of highly expressed proteins. Recently, Leuenberger et al. developed an elegant experimental approach to measure protein thermal stability at the proteome scale. The collected data allow us to rigorously test the predictions of the misfolding avoidance hypothesis that highly expressed proteins have evolved to be more stable, and that maintaining thermodynamic stability significantly constrains their evolution. Notably, reanalysis of the Leuenberger et al. data across four different organisms reveals no substantial correlation between protein stability and protein abundance. Therefore, the key predictions of the misfolding toxicity and related hypotheses are not supported by available empirical data. The data also suggest that, regardless of protein expression, protein stability does not substantially affect the protein molecular clock across organisms.

Key words: **molecular evolution, protein clock, toxic misfolding.**

A fundamental and long-standing question in molecular evolution is what determines protein sequence constraints, or the rate of the protein molecular clock (Zuckerkandl and Pauling 1965; Zhang and Yang 2015). Proteins from the same species accumulate substitutions at rates that span several orders of magnitude, and the causes of such variability have been widely debated (Koonin and Wolf 2010). Analyses of high-throughput genome-scale data consistently showed that protein evolutionary rates are strongly anticorrelated with their corresponding expression and abundance levels (Pal et al. 2001, 2006). This relationship, often referred to as the E–R (Expression–evolutionary Rate) anticorrelation (Zhang and Yang 2015), explains up to a third of the variance in molecular clock rates across proteins (Pal et al. 2006; Drummond and Wilke 2008). Among possible explanations of the E–R anticorrelation is the popular hypothesis that highly expressed proteins evolve slowly to avoid mistranslation-induced (Drummond and Wilke 2008) or spontaneous (Yang et al. 2010) protein misfolding. According to this hypothesis, misfolded proteins are toxic to cells and therefore reduce fitness. As highly abundant proteins have the potential to produce more misfolded proteins compared to proteins with low abundance, their sequences should be under stronger evolutionary constraints to increase protein stability (Drummond and Wilke 2008; Zhang and Yang 2015). Thus, a key prediction of the misfolding avoidance hypothesis is that highly expressed proteins should be more thermodynamically stable than proteins expressed at low levels, and that selection against protein misfolding should significantly constrain their sequence evolution (Cherry 2010; Serohijos et al. 2012, 2013).

Previously (Plata et al. 2010), based on a small set of proteins available in the proTherm database (Bava et al. 2004), we did not detect any significant correlation between protein expression and thermodynamic stability. Furthermore, to empirically test the misfolding hypothesis, we expressed wild type (WT) and destabilized mutant versions of the LacZ protein in *Escherichia coli*. This analysis demonstrated that the corresponding fitness effects were primarily related to the cost of gratuitous protein production and not to misfolding toxicity (Plata et al. 2010). Similar experiments in yeast by Kafri et al. (2016) using WT and destabilized versions of GFP, also showed that misfolded protein toxicity plays a relatively minor role in explaining the fitness cost behind the E–R anticorrelation.

As the aforementioned results have been obtained using small sets of proteins, additional tests involving large data sets across diverse organisms are essential. Recently, Leuenberger et al. (2017) measured the thermal stability of thousands of proteins across two bacteria (*E. coli* and *Thermus thermophilus*) and two eukaryotes (*Saccharomyces cerevisiae* and *Homo sapiens*). The unprecedented size of this data set, measured directly in the cellular matrix, makes it possible to empirically test the misfolding toxicity hypothesis at the proteome scale. Using protein melting temperatures (T_m) from *E. coli*, Leuenberger et al. (2017) concluded that highly abundant proteins are stable because they are evolutionarily designed to tolerate translational errors, supporting the misfolding toxicity avoidance hypothesis. The authors reached their conclusion based on different abundances of *E. coli* proteins separated into three bins according to their thermal stability (figure 3I in Leuenberger et al.), but did not perform similar analyses for the remaining three species. Notably, analyses of

Letter

Table 1. Correlation between T_m, Gene and Protein Expression, and Evolutionary Rate[a].

Species	Protein Abundance versus Ka	Gene Expression versus Ka	T_m versus Protein Abundance	T_m versus Gene Expression	T_m versus Ka
Escherichia coli	−0.38** (−0.38**)	−0.40** (−0.40**)	0.08*	−0.02	0.02
Saccharomyces cerevisiae	−0.47** (−0.47**)	−0.45** (−0.45**)	−0.16**	−0.06	0.05
Homo sapiens	−0.16** (−0.17**)	−0.17** (−0.17**)	−0.19**	−0.14**	0.02
Thermus thermophilus	NA	−0.35** (−0.35**)	NA	0.04	−0.04

NOTE.—Values in parentheses show the partial Spearman correlation between abundance/expression and Ka after controlling for T_m.
[a]Only proteins with measured T_m were considered, ribosomal proteins were excluded (see supplementary table S1, Supplementary Material online, for results including ribosomal proteins).
P values for Spearman's rank correlation are indicated as *<0.05 and **$<5\times10^{-3}$.

arbitrarily binned data often obscure the effect size and thus may lead to misleading conclusions. Therefore, we decided to investigate the correlation between protein abundance and stability, and its impact on evolutionary sequence constraints using unbinned data from all four species analyzed by Leuenberger et al.

We note that despite possible biases and uneven sampling of proteins in different organisms, the correlation of sequence constraints, commonly quantified as the rate of nonsynonymous substitutions per site (Ka), with protein abundance (table 1, second column) and gene expression (table 1, third column) remains strong for the subset of proteins with reported T_m measurements. Therefore, these data can be used to investigate the nature of sequence constraints in all organisms analyzed by Leuenberger et al. Moreover, although proteins with similar T_m may have different folding stabilities at physiological temperatures (Becktel and Schellman 1987), using data from the ProTherm database we found a significant correlation between proteins' T_m and their unfolding Gibbs free energies (Spearman's $r=0.64$, $P<10^{-20}$, Pearson's $r=0.75$, $P<10^{-20}$, supplementary fig. S1, Supplementary Material online). Consequently, reported protein melting temperatures do reflect, at least on an average, protein stabilities at physiological temperatures.

Using protein stabilities and abundances from Leuenberger et al., we first confirmed a weak but significant positive correlation between T_m and protein abundance in *E. coli* (Spearman's $r=0.16$, $P=6\times10^{-6}$; Pearson's $r=0.2$, $P=7\times10^{-8}$). Surprisingly, for the other two organisms with protein abundance data (yeast and human) we found significant negative correlations with T_m (Spearman's $r=-0.11$ and -0.19, respectively, both $P<0.005$, Pearson's $r=-0.09$ and -0.13, both $P<0.02$), contrary to the prediction that abundant proteins should be more stable. Moreover, because ribosomal proteins are highly abundant and generally enriched among stable proteins, it is possible that the weak correlation of T_m and protein abundance is primarily driven in *E. coli* by the properties of ribosomal proteins. Indeed, excluding 46 ribosomal proteins (out of 730 considered proteins) substantially decreased both the magnitude and the significance of the correlation in *E. coli* (fig. 1*a*; Spearman's $r=0.08$, $P=0.03$, Pearson's $r=0.09$, $P=0.02$), whereas for yeast and human data, we still observed small negative correlations (fig. 1*b* and *c*, and table 1, fourth column). We next calculated, after removing ribosomal proteins, the correlation between T_m and mRNA expression in all four species (fig. 1*d*–*g*, and table 1, fifth column). Similar to protein abundances, and contrary to the expectation of the misfolding avoidance hypothesis, the correlations were either nonsignificant or negative. Furthermore, when T_m was calculated considering data from all peptides associated with each protein, rather than only peptides assigned to the least stable protein domain (the approach used by Leuenberger et al. [2017]), we again observed only a weak positive correlation between T_m and protein abundance in *E. coli* (Spearman's $r=0.07$, $P=0.05$, Pearson's $r=0.09$, $P=0.01$), but not in any other organism.

The conjecture that highly expressed proteins are stable because they are designed to tolerate translational errors (Leuenberger et al. 2017) can be directly tested by analyzing the effect of protein stability on the correlation between protein abundance and sequence constraints. Such an analysis demonstrates that the significant negative correlation between protein abundance and evolutionary constraints (Ka), with or without ribosomal proteins, remains essentially unchanged after controlling for protein stability in all analyzed organisms (the correlations in parentheses in the second and third columns in table 1 and supplementary table S1, Supplementary Material online).

Interestingly, the Leuenberger et al. data also suggest that protein stability, irrespective of protein abundance or mRNA expression, does not substantially affect the protein molecular clock. In none of the four species the correlation between T_m and Ka is either strong or significant (table 1, last column and fig. 2). There is also no significant correlation between protein stability and the clock rate when only single domain proteins are considered (supplementary fig. S2, Supplementary Material online). These results indicate that, beyond the avoidance of misfolding toxicity, any theory requiring the optimization of protein stability as a dominant constraint of the protein molecular clock is not consistent with the empirical data.

Overall, our analyses demonstrate that there is no substantial correlation between protein stability and protein abundance (at most 1–4% of the variance explained). In two of the analyzed organisms, the correlation between stability and

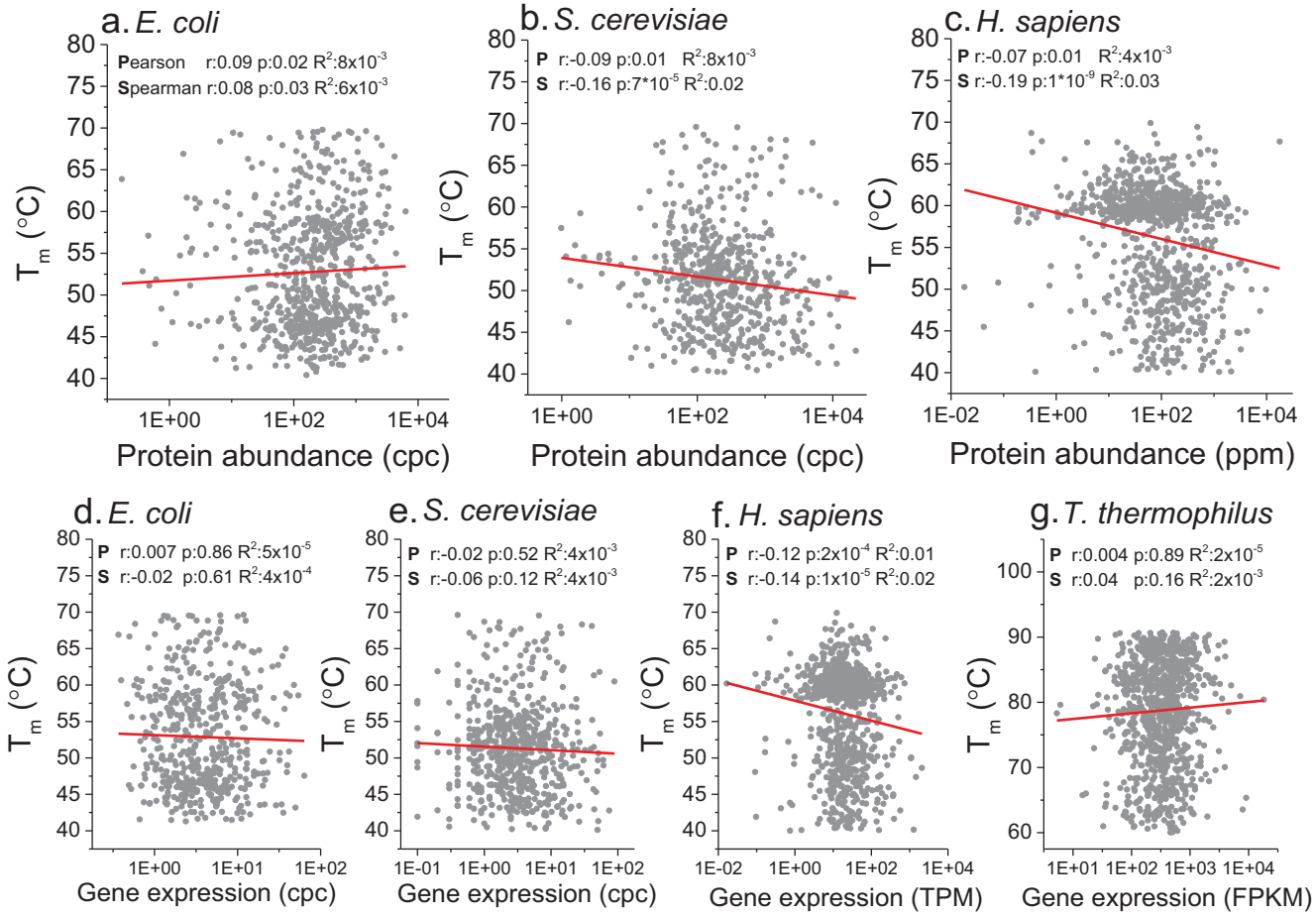

Fig. 1. Protein melting temperature (T_m) calculated by Leuenberger et al. as a function of protein abundance in three species ([*a*] *Escherichia coli*, [*b*] *Saccharomyces cerevisiae*, and [*c*] *Homo sapiens*). T_m as a function of mRNA expression in four species ([*d*] *E. coli*, [*e*] *S. cerevisiae*, [*f*] *H. sapiens*, and [*g*] *Thermus thermophilus*). The solid lines represent linear fits to the log-transformed protein abundance and mRNA expression data; correlation coefficients, corresponding *P* values, and R^2 are shown for Pearson's (P) and Spearman's (S) correlations in each panel. cpc, counts per cell; ppm, parts per million; TPM, Transcripts Per Kilobase Million; FPKM, Fragments Per Kilobase Million. Proteins annotated as ribosomal were excluded from the analysis.

abundance is weak and opposite to the main prediction of the misfolding avoidance hypothesis. The weak correlation observed in *E. coli* is primarily driven by the properties of ribosomal proteins. There are also no detectable effects of protein stability on the relationships between protein abundance and evolutionary sequence constraints. Therefore, the analysis of the extensive data set recently generated by Leuenberger et al., similar to previous studies (Plata et al. 2010; Kafri et al. 2016), suggests that neither mistranslation-induced nor spontaneous misfolding toxicity is likely to substantially affect protein sequence constraints and the rate of the protein molecular clock.

Given no significant correlation between T_m and Ka, it is likely that common biophysical mechanisms for protein stabilization, such as the burial of several additional hydrophobic residues (Dill and Bromberg 2011), may not significantly increase the evolutionary constraints on hundreds of other sites in a protein. Therefore, it will be important to further investigate how effects associated with the costs of protein production, protein cellular abundance and functional optimization, contribute to evolutionary sequence constraints and the protein molecular clock (Cherry 2010; Plata et al. 2010).

Materials and Methods

T_m data, and protein abundances for *E. coli* and yeast, as well as the number of domains per protein, were obtained from supplementary table 3 in the Leuenberger et al. study (2017). Human protein abundances were obtained from the whole organism integrated data set in the PaxDB v.4 database (Wang et al. 2012). *Escherichia coli*, *T. thermophilus*, and *S. cerevisiae* expression data were obtained from Lu et al. (2007), Swarts et al. (2015) and Holstege et al. (1998), respectively. Human expression data were averaged across the main nine tissues in the Mele et al. (2015)'s study. Ka values for *E. coli*, *S. cerevisiae*, *H. sapiens*, and *T. thermophilus* were calculated with the PAML package (Yang 1997) relative to *Salmonella enterica*, *Saccharomyces bayanus*, *Macaca mulatta*, and *Thermophilus aquaticus* orthologs, respectively. Orthologs were identified as bidirectional best hits (BBHs) using protein BLAST (Altschul et al. 1997); we only considered for the analysis BBHs for which at least 70% of the length of

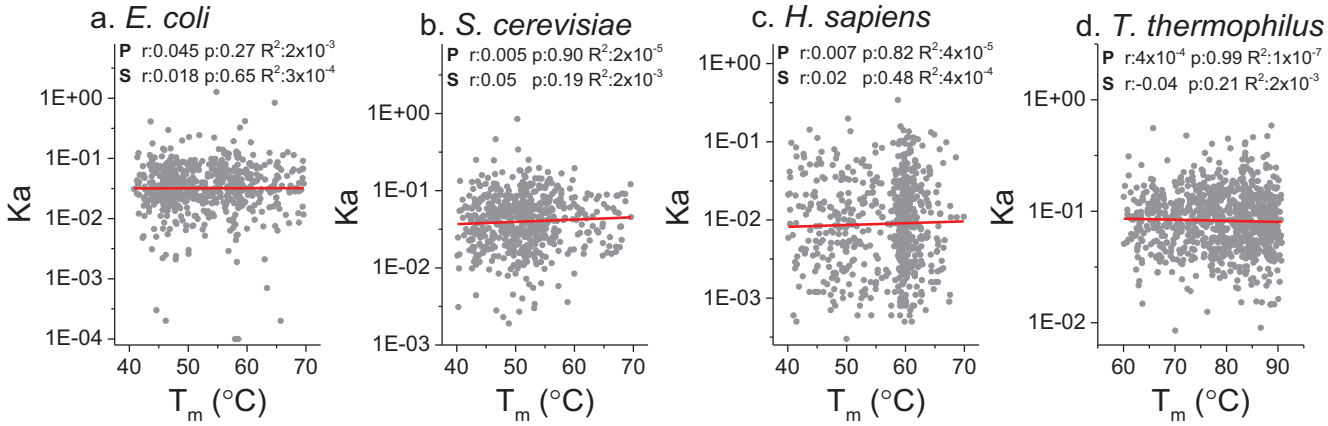

Fig. 2. The rate of nonsynonymous substitutions per nonsynonymous site (Ka) as a function of protein melting temperature (T_m) calculated by Leuenberger et al. (2017). Results are shown for four different organisms ([*a*] *Escherichia coli*, [*b*] *Saccharomyces cerevisiae*, [*c*] *Homo sapiens*, and [*d*] *Thermus thermophilus*). Ribosomal proteins were excluded from the analysis. The solid lines represent linear fits of the log-transformed Ka data; correlation coefficients, corresponding *P* values, and R^2 are shown for Pearson (P) and Spearman (S) correlations, in each panel.

the shortest protein was aligned. Unfolding free energy and melting temperature data used in supplementary figure S1, Supplementary Material online, were obtained from the ProTherm (Feb. 2013) database (Bava et al. 2004). The Ribosomal Protein Gene Database was used to identify ribosomal proteins (Nakao et al. 2004).

Acknowledgments

We thank Eugene Koonin and Dinara Usmanova for helpful discussions. This work was supported in part by the National Institute of General Medical Sciences grant GM079759 to DV.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25(17):3389–3402.

Bava KA, Gromiha MM, Uedaira H, Kitajima K, Sarai A. 2004. ProTherm, version 4.0: thermodynamic database for proteins and mutants. *Nucleic Acids Res.* 32(Database issue):D120–D121.

Becktel WJ, Schellman JA. 1987. Protein stability curves. *Biopolymers* 26(11):1859–1877.

Cherry JL. 2010. Highly expressed and slowly evolving proteins share compositional properties with thermophilic proteins. *Mol Biol Evol.* 27(3):735–741.

Dill KA, Bromberg S. 2011. Molecular driving forces: statistical thermodynamics in biology, chemistry, physics, and nanoscience. London; New York: Garland Science.

Drummond DA, Wilke CO. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* 134(2):341–352.

Holstege FC, Jennings EG, Wyrick JJ, Lee TI, Hengartner CJ, Green MR, Golub TR, Lander ES, Young RA. 1998. Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* 95(5):717–728.

Kafri M, Metzl-Raz E, Jona G, Barkai N. 2016. The cost of protein production. *Cell Rep.* 14(1):22–31.

Koonin EV, Wolf YI. 2010. Constraints and plasticity in genome and molecular-phenome evolution. *Nat Rev Genet.* 11(7):487–498.

Leuenberger P, Ganscha S, Kahraman A, Cappelletti V, Boersema PJ, von Mering C, Claassen M, Picotti P. 2017. Cell-wide analysis of protein thermal unfolding reveals determinants of thermostability. *Science* 355(6327):eaai7825.

Lu P, Vogel C, Wang R, Yao X, Marcotte EM. 2007. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat Biotechnol.* 25(1):117–124.

Mele M, Ferreira PG, Reverter F, DeLuca DS, Monlong J, Sammeth M, Young TR, Goldmann JM, Pervouchine DD, Sullivan TJ, et al. 2015. Human genomics. The human transcriptome across tissues and individuals. *Science* 348(6235):660–665.

Nakao A, Yoshihama M, Kenmochi N. 2004. RPG: the Ribosomal Protein Gene database. *Nucleic Acids Res.* 32(Database issue):D168–D170.

Pal C, Papp B, Hurst LD. 2001. Highly expressed genes in yeast evolve slowly. *Genetics* 158(2):927–931.

Pal C, Papp B, Lercher MJ. 2006. An integrated view of protein evolution. *Nat Rev Genet.* 7(5):337–348.

Plata G, Gottesman ME, Vitkup D. 2010. The rate of the molecular clock and the cost of gratuitous protein synthesis. *Genome Biol.* 11(9):R98.

Serohijos AW, Lee SY, Shakhnovich EI. 2013. Highly abundant proteins favor more stable 3D structures in yeast. *Biophys J.* 104(3):L1–L3.

Serohijos AW, Rimas Z, Shakhnovich EI. 2012. Protein biophysics explains why highly abundant proteins evolve slowly. *Cell Rep.* 2(2):249–256.

Swarts DC, Koehorst JJ, Westra ER, Schaap PJ, van der Oost J. 2015. Effects of argonaute on gene expression in *Thermus thermophilus*. *PLoS One* 10(4):e0124880.

Wang M, Weiss M, Simonovic M, Haertinger G, Schrimpf SP, Hengartner MO, von Mering C. 2012. PaxDb, a database of protein abundance averages across all three domains of life. *Mol Cell Proteomics* 11(8):492–500.

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.

Yang ZH. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci.* 13(5):555–556.

Zhang J, Yang JR. 2015. Determinants of the rate of protein sequence evolution. *Nat Rev Genet.* 16(7):409–420.

Zuckerkandl E, Pauling L. 1965. Evolutionary divergence and convergence in proteins. In: Bryson V, Vogel H, editors. Evolving genes and proteins. New York: Academic Press. p. 97–166.