Avoidance of toxic misfolding does not explain the sequence constraints of highly expressed proteins across organisms

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

[1]Department of Systems Biology, Columbia University, New York, NY, USA. [2]Department of Biomedical Informatics, Columbia University, New York, NY, USA.

Correspondence to DV at dv2121@columbia.edu

Abstract

The avoidance of cytotoxic effects associated with protein misfolding has been proposed as a dominant constraint for the evolution 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 to rigorously test the key predictions of the misfolding avoidance hypothesis. Specifically, that highly expressed proteins are designed to be more stable, and that thermodynamic stability significantly constrains their evolution. Careful re-analyses of the Leuenberger *et al*. data across four different organisms shows no substantial correlation between protein stability and protein abundance. We also find that protein stability does not substantially contribute to sequence constraints of highly abundant proteins. Therefore, the key prediction of the toxic misfolding avoidance hypothesis is not supported by the empirical data.

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 omics data consistently showed that protein evolutionary rates are strongly anticorrelated with their corresponding expression and abundance levels (Pal, et al. 2001; Pal, et al. 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 relatively more misfolded proteins, 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 toxicity avoidance hypothesis is that highly expressed proteins should be more thermodynamically stable than proteins expressed at low levels, and that protein stability should significantly constrain their sequence evolution (Cherry 2010b).

Previously (Plata, et al. 2010), we did not detect any significant correlation between protein expression and thermodynamic stability based on a small set of proteins available in the proTherm (Bava, et al. 2004) database. As a direct empirical test of the hypothesis, we also expressed wild type and destabilized mutant versions of the LacZ protein in *Escherichia coli* and demonstrated that the corresponding fitness cost was not primarily related to misfolding toxicity but to the cost of gratuitous protein production (Plata, et al. 2010). Subsequent experiments in yeast by Geiler-Samerotte *et al*. (Geiler-Samerotte, et al. 2011) and Kafri *et al.* (Kafri, et al. 2016) also revealed that misfolded protein toxicity plays a relatively minor role in explaining the cost behind the E-R anticorrelation.

As the aforementioned results have been obtained using small sets of proteins, additional tests involving large datasets across diverse organisms are essential. Recently, Leuenberger *et al.* (Leuenberger, et al. 2017) measured the thermal stability of hundreds of proteins in two bacteria (*E. coli* and *Thermus thermophilus*) and two eukaryotes (*Saccharomyces cerevisiae* and *Homo sapiens*). The unprecedented size of this dataset, measured directly in the cellular matrix, provides a unique opportunity to empirically test the misfolding toxicity avoidance hypothesis. Based on estimates of protein melting temperatures (T_m) in

E. coli, Leuenberger *et al*. concluded that highly abundant proteins are stable because they are evolutionarily designed to tolerate translational errors (Leuenberger, et al. 2017), supporting the misfolding toxicity avoidance hypothesis. The authors reach their conclusion based on abundance differences between *E. coli* proteins separated into three bins according to their thermal stability (Figure 3I in Leuenberger *et al.*). Notably, analyses of arbitrarily binned data may often hide the magnitude of the effect and lead to misleading conclusions. Therefore, we decided to rigorously investigate the correlation between protein expression and protein stability, and its impact on protein design constraints for all four species in the Leuenberger *et al*. study. We note that despite possible biases and under-sampling of proteins in the study, for the subset of proteins with reported T_m measurements the correlation between sequence constraints and abundance remains strong in all organisms (Table 1). Therefore, these data can be used to investigate the nature of sequence constraints of highly expressed proteins.

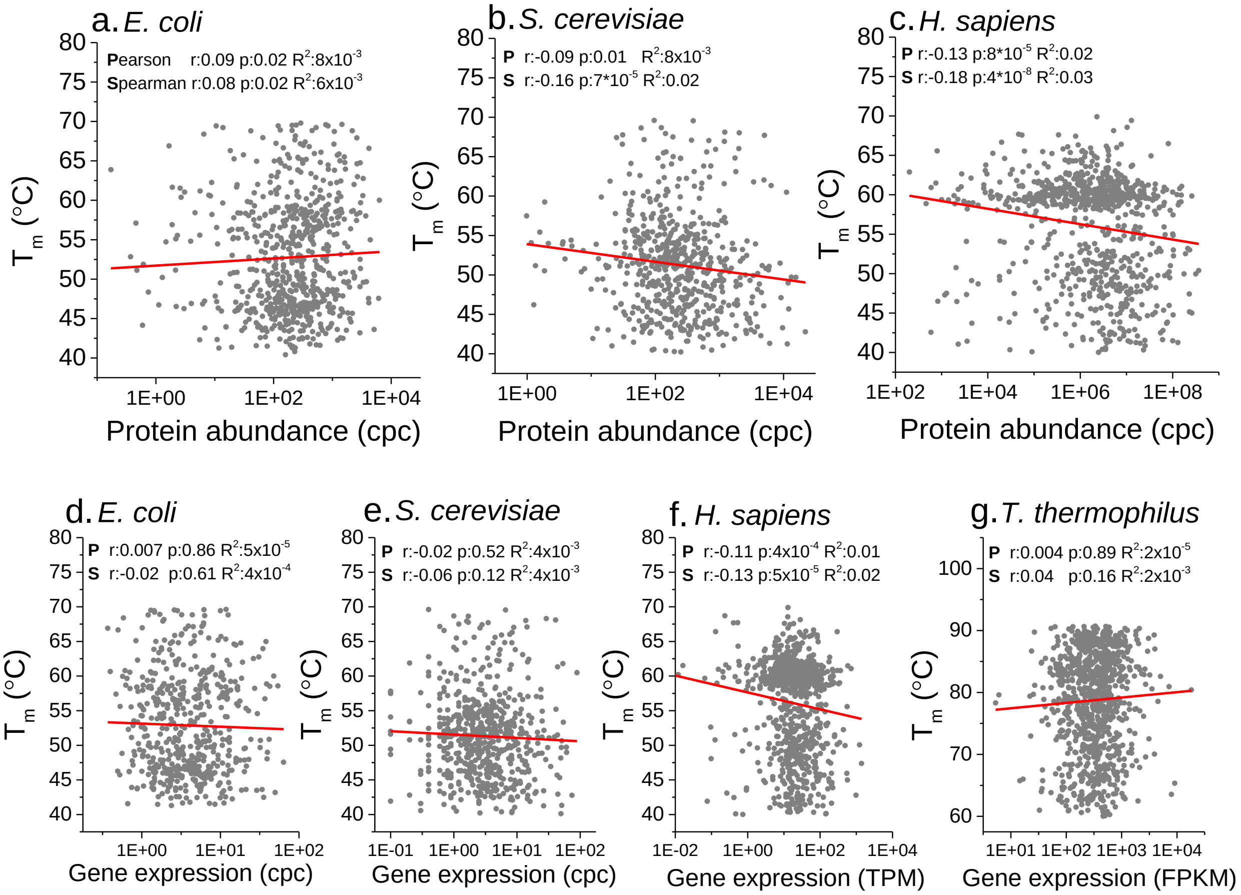

Figure 1. Protein melting temperature (T_m) calculated by Leuenberger *et al*. as a function of protein abundance for three species (**a.** *E. coli*, **b.** *S. cerevisiae* and **c.** *H. sapiens*). Proteins annotated as ribosomal are excluded from analysis. **d-g.** T_m as a function of mRNA expression for four species. The red lines

represent linear fits for the log-transformed protein abundance and mRNA data; correlation coefficients, p-values, and corresponding R^2 are shown for Pearson (**P**) and Spearman (**R**) in each panel. cpc: counts per cell; TPM: Transcripts Per Kilobase Million; FPKM: Fragments Per Kilobase Million.

First, using protein-level stabilities and abundances from Leuenberger *et al.*, we confirmed a weak but significant positive correlation between T_m and protein counts in *E. coli* (Spearman's r: 0.16, $p=6x10^{-6}$; Pearson's r: 0.2, $p=7x10^{-8}$, or ~4% of the variance explained). 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, p<0.005), contrary to the prediction that abundant proteins should be more stable. Moreover, because ribosomal proteins are highly abundant and generally enriched among thermostable proteins, it is possible that the weak correlation of T_m and protein abundance in *E. coli* primarily reflects the properties of ribosomal proteins, rather than a general effect of protein abundance. Indeed, excluding 46 ribosomal proteins (out of 730 proteins considered) significantly decreased both the magnitude and significance of the correlation in *E. coli* (Figure 1a; Pearson's r:0.08, p=0.03, or less than 1% of the variance explained), whereas for yeast and human data, we still observed negative correlations (Figure 1b,c, Table 1).

Table 1. Correlation between T_m, gene and protein expression, and evolutionary rate [a]

Species	Protein abundance vs. Ka[b,c]	Gene expression vs. Ka	T_m vs. protein abundance	T_m vs. Gene expression	T_m vs. Ka
E. coli	$-0.38^{**}(-0.38^{**})$	$-0.40^{**}(-0.40^{**})$	0.08^{*}	-0.02	0.02
S. cerevisiae	$-0.47^{**}(-0.47^{**})$	$-0.45^{**}(-0.45^{**})$	-0.16^{**}	-0.06	0.05
H. sapiens	$-0.14^{**}(-0.15^{**})$	$-0.19^{**}(-0.19^{**})$	-0.18^{**}	-0.13^{**}	-0.03
T. thermophilus	N.A.	$-0.35^{**}(-0.35^{**})$	N.A.	0.04	-0.04

[a] Only proteins with measured T_m were considered for the correlations, ribosomal proteins were excluded

[b] P-values for Spearman's rank correlation are indicated as * <0.05 and $^{**}<5x10^{-3}$

[c] Values in parentheses show the partial correlation between abundance/expression and Ka after controlling for T_m

Because protein sequence constraints –commonly quantified as the rate of non-synonymous substitutions per site, Ka– also tend to strongly correlate with gene expression levels, we next calculated the correlation of T_m and mRNA expression for all four species (Figure 1. d-g, Table 1). In all cases, the correlation was either non-significant or negative, i.e. directly opposite to the prediction of the misfolding avoidance hypothesis. Moreover, if the avoidance of toxic misfolded proteins is a major determinant of sequence constraints, we would expect a negative correlation between T_m and Ka. In contrast, our analysis showed that in none of the four species such a correlation is either strong or significant (Table 1).

A stated conclusion of Leuenberger *et al*. is that highly expressed proteins are stable because they are designed to tolerate translational errors (Leuenberger, et al. 2017). This conclusion can be directly tested by analyzing the effect of protein stability on the relationship between protein abundance and sequence constraints. Importantly, for the hypothesis to be valid, it is not enough to demonstrate a positive correlation between protein abundance and stability, but one must also show a significant effect on the sequence constraints of highly expressed proteins. Contrary to this expectation, we found 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 organisms (Table 1, first two columns, in parenthesis). We note that even if there were a substantial contribution of protein stability to sequence constraints, there are (Chimpanzee 2005) multiple other reasons, unrelated to mistranslation misfolding toxicity, for abundant proteins to be more stable. For example, functional cost-benefit tradeoffs (Cherry 2010a; Gout, et al. 2010).

Overall, our analyses demonstrate that there is no substantial correlation between protein stability and protein abundance (1-4% of the variance explained). In two of the analyzed organisms the correlation between stability and abundance is actually 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. Most importantly, there is no detectable effect of protein stability on the relationships between protein abundance and evolutionary sequence constraints. Therefore, the analysis of the extensive dataset generated by Leuenberger *et al*., similar to previous studies (Plata, et al. 2010; Kafri, et al. 2016), suggests that either mistranslation-induced or spontaneous misfolding toxicity is unlikely to substantially affect protein sequence constraints and the molecular clock rate of highly expressed proteins.

Data sources

T_m data and protein abundances for *E. coli* and yeast were obtained from supplementary Table 3 in the Leuenberger *et al*. study (Leuenberger, et al. 2017). Average human protein abundances were obtained for NCI60 cell lines (Gholami, et al. 2013). *E. coli, T. thermophilus* and *S. cerevisiae* expression data were obtained from Lu *et al*. (Lu, et al. 2007), Swarts *et al.* (Swarts, et al. 2015) and Holstege *et al.*

(Holstege, et al. 1998), respectively. Human expression data were averaged across the main 9 tissues in the Melé *et al*. (Mele, et al. 2015) study. Human Ka values were obtained from the study by the Chimpanzee S & A Consortium (Chimpanzee 2005). Ka values for *E. coli*, *S. cerevisiae* and *T. thermophilus* were calculated with the PAML package (Yang 1997) relative to *Salmonella enterica*, *Saccharomyces bayanus*, and *Thermophilus aquaticus* orthologs, respectively.

References

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:D120-121.

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.

Chimpanzee SAC. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437:69-87.

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

Geiler-Samerotte KA, Dion MF, Budnik BA, Wang SM, Hartl DL, Drummond DA. 2011. Misfolded proteins impose a dosage-dependent fitness cost and trigger a cytosolic unfolded protein response in yeast. Proc Natl Acad Sci U S A 108:680-685.

Gholami AM, Hahne H, Wu Z, Auer FJ, Meng C, Wilhelm M, Kuster B. 2013. Global proteome analysis of the NCI-60 cell line panel. Cell Rep 4:609-620.

Gout JF, Kahn D, Duret L, Paramecium Post-Genomics C. 2010. The relationship among gene expression, the evolution of gene dosage, and the rate of protein evolution. PLoS Genet 6:e1000944.

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 95:717-728.

Kafri M, Metzl-Raz E, Jona G, Barkai N. 2016. The Cost of Protein Production. Cell Rep 14:22-31.

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

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.

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

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 348:660-665.

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

Pal C, Papp B, Lercher MJ. 2006. An integrated view of protein evolution. Nat Rev Genet 7:337-348.

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

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:e0124880.

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. 13:555-556.

Zhang J, Yang JR. 2015. Determinants of the rate of protein sequence evolution. Nat Rev Genet 16:409-420.

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.