Highly Abundant Proteins Favor More Stable 3D Structures in Yeast

Adrian W. R. Serohijos, S. Y. Ryan Lee, and Eugene I. Shakhnovich*
Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts

ABSTRACT To understand the variation of protein sequences in nature, we need to reckon with evolutionary constraints that are biophysical, cellular, and ecological. Here, we show that under the global selection against protein misfolding, there exists a scaling among protein folding stability, protein cellular abundance, and effective population size. The specific scaling implies that the several-orders-of-magnitude range of protein abundances in the cell should leave imprints on extant protein structures, a prediction that is supported by our structural analysis of the yeast proteome.

Received for publication 26 September 2012 and in final form 29 November 2012.

*Correspondence: shakhnovich@chemistry.harvard.edu

In molecular biophysics, the view that properties of proteins can be determined from first principles of physics and chemistry is almost a canon law. Advances in molecular dynamics, protein folding, ab initio structure prediction, and design of novel protein folds and function all support this view. Notwithstanding these developments, to what extent can physics and chemistry account for the diversity of biophysical and biochemical properties of proteins in nature?

From comparative genomics, one emerging constraint in the evolution of the coding regions of the genome is global selection against the cytotoxic effects of protein misfolding (1). Misfolded proteins are detrimental to the cell because they can form aggregates that can be toxic (2). The apparent universality of this constraint is manifested in the consistent observation that highly expressed proteins evolve more slowly across all forms of life—from bacteria to nematodes, mammals, and humans (1). Apart from explaining the universal correlation between abundance and the rate of evolution, a major prediction of the misfolding hypothesis is that more abundant proteins will evolve toward greater stability (1,3,4). One can show that this prediction arises from the interplay of population dynamics and protein biophysics.

Assuming monoclonality, the rate of protein evolution (ratio of nonsynonymous and synonymous substitutions) can be expressed as (5,6)

$$\omega(s) = N_e \frac{1-\exp(-2s)}{1-\exp(-2N_e s)} \tag{1}$$

where N_e is the effective population size and s is the change in fitness due to the substitution (the selection coefficient). In a recent work (4), we showed that under the selection against protein misfolding, and assuming a two-state folding process, s is explicitly expressed as

$$s = -cA\left[\frac{1}{1+\exp(-\beta(\Delta G+\Delta\Delta G))} - \frac{1}{1+\exp(-\beta\Delta G)}\right] \tag{2}$$

where A is the cellular abundance of a protein, ΔG is the folding stability, c is the fitness cost per misfolded protein (measured in yeast to be ~32/(total cellular protein concentration (7)), and $\beta = 1/k_B T$. From Eq. 2, the rate is a function of premutation gene properties (abundance and ΔG) and the change in stability due to the arising mutation ($\Delta\Delta G$).

Integrating over all possible mutational effects $p(\Delta\Delta G)$, the molecular clock surface is

$$\int_{-\infty}^{+\infty} p(\Delta\Delta G)\omega(s)d(\Delta\Delta G) \tag{3}$$

The distribution $p(\Delta\Delta G)$ is approximately a Gaussian with mean $\Delta\Delta G_{mean}$ (1 kcal/mol) and standard deviation $\Delta\Delta G_{sd}$ (1.7 kcal/mol). Estimates for both parameters are derived from empirical measurements of folding stability changes due to single point mutations (ProTherm database (8)). This integral (Eq. 3) defines the molecular clock surface shown in Fig. 1. Because fixation of a mutation changes ΔG, the evolution of a gene is essentially a walk on the molecular clock surface, and this walk is slowest in the neighborhood of the gully (Fig. 1, *red line*). Consequently, on evolutionary timescales, genes tend to cluster in the gully of the surface under mutation-selection balance (4). Indeed, evolutionary simulations from various groups have predicted this correlation between abundance and stability (1,3,4).

The surface defined by Eq. 3 has a minimum at

$$A = \left(\frac{1}{\beta}\frac{\Delta\Delta G_{sd}}{\Delta\Delta G_{mean}^2}\right)\frac{1}{(N_e-1)c}\frac{(1+e^{-\beta\Delta G})^2}{e^{-\beta\Delta G}} \tag{4}$$

Editor: Bertrand Garcia-Moreno.

http://dx.doi.org/10.1016/j.bpj.2012.11.3838

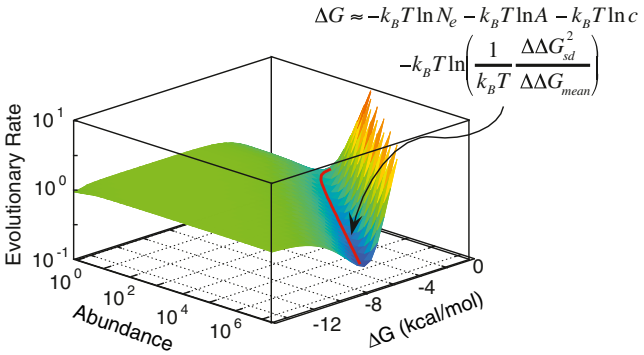

FIGURE 1 Rate of evolution of a protein as a function of its cellular abundance and folding stability. Rate is defined as dN/dS (the rate of nonsynonymous substitutions per nonsynonymous sites). The rate is slowest in the gully (*red line*), which defines the average relationship among the folding stability, abundance, and population size (see Serohijos et al. (4) and Supporting Material).

According to the ProTherm database (8), most proteins have stabilities < -3 kcal/mol. In this regime, the above expression takes a simpler form:

$$A \approx \left(\frac{1}{\beta}\frac{\Delta\Delta G_{sd}}{\Delta\Delta G_{mean}^2}\right)\frac{1}{N_e c}e^{-\beta\Delta G} \quad (5)$$

or

$$\Delta G \approx -k_B T \ln N_e - k_B T \ln A - k_B T \ln c - k_B T \ln\left(\frac{1}{k_B T}\frac{\Delta\Delta G_{sd}^2}{\Delta\Delta G_{mean}}\right) \quad (6)$$

which defines a peculiar scaling relationship among the average stability of proteins in a proteome (ΔG), their cellular abundance A, and the organism's effective population size N_e. All of the variables on the right-hand side of Eq. 6 have been measured or estimated empirically, allowing one to assign the relative contribution of population size and abundance to the evolution of protein folding stability (Table 1). Indeed, the variation of protein folding stability in nature could be largely due to protein abundance and population size (Table 1).

Considering that protein cellular abundances span 10–10^6 copies per cell (as shown in yeast (9)), with an energetic equivalence of ~7 kcal/mol in protein stability (Table 1), we reasoned that abundance should systematically manifest in the structural properties of proteins across a genome. To date, the strongest empirical support for the interdependence of abundance and stability is the observation that highly abundant, slowly evolving proteins and proteins from thermophilic bacteria share a similar amino acid composition (10). To demonstrate this prediction more unambiguously, we extracted all of the yeast proteins from the Protein Data Bank, partitioned them into domains as defined by SCOP (11), and then mapped their experimentally measured abundance (9). Also, we excluded domains with gaps in the structure. This procedure yielded 302 domains on which we performed a structural analysis (Fig. 2 and Table S1 in the Supporting Material). Using the modeling tool Eris (12), we calculated the hydrogen-bonding energy and van der Waals interaction energy (two major contributors to the folding free energy) within each domain (13). Residues in more abundant proteins form more extensive hydrogen bonds between their side chains and backbones (r = −0.29***) and among their side chains (r = −0.30***). Abundance likewise correlates with increasing van der Waals interaction (r = −0.30***).

We note, however, that the manifestation of the scaling (Eq. 6) is strong on protein structural properties that directly influence stability (Fig. 2, *A–C*), but could be less manifested in indirect indictators of stability, such as protein length (14). For example, in the 302 proteins we analyzed,

TABLE 1 Energetic equivalence of constraints imposed by evolutionary variables

Variable[a]	Observed/estimated values in nature	Energetic equivalence (kcal/mol)
A	10–10^6 (9)	−1 to −8
N_e	10^4–10^8 (17)	−5 to −11

[a]We used the scaling in Eq. 6 to assign the relative strength of constraints imposed by the evolutionary variables on the evolution of folding stability (ΔG). Abundance has been measured in yeast. Effective population sizes have been estimated across all kingdoms of life (10^4 in mammals and 10^8 in prokaryotes). The calculation assumes monoclonality ($\mu << 1$), and the effect of the mutation rate μ on the scaling remains an open question. $k_B T = 0.593$ kcal/mol.

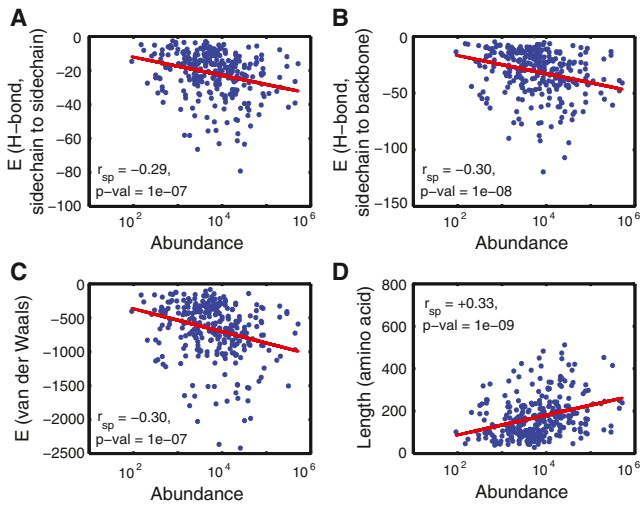

FIGURE 2 (*A–C*) Correlation between abundance and structural properties (hydrogen-bond content and strength of van der Waals interaction) of protein domains in yeast. (*D*) Stability is an extensive property, and thus abundance correlates with domain length (14). Indicated are the values of the Spearman rank correlation. See also Table S1.

the more-abundant domains were generally longer ($r = 0.33^{***}$). When we expanded the set to include domains (15) that do not have empirically determined structures (Fig. S1), we found no correlation between domain length and abundance, because length is a coarse descriptor of stablity. The general observation that more abundant genes tend to be shorter ($r = -0.19^{***}$) reflects the fact that they have fewer domains ($r = -0.12^{***}$; Fig. S1).

As was recently pointed out (16), population size constrains the cellular distribution of folding stabilities such that organisms with small effective population sizes (e.g., endosymbiotic parasites that undergo episodic bottlenecking) will evolve less thermodynamically stable proteins, simply because deleterious mutations will fix at a higher probability in smaller population sizes. On the contrary, organisms with higher population sizes, which experience stronger purifying selection, are predicted to evolve more stable proteins. Additionally, assuming that all other things are equal, vertebrates (with effective population sizes of 10^4–10^5 (17)) are predicted by Eq. 6 to evolve proteins that are on average 6 kcal/mol less stable than proteins in prokaryotes (whose population sizes are $\geq 10^8$ (17); Table 1). Systematically proving this prediction is the subject of future work. Nonetheless, protein structures of viruses, which undergo episodic bottlenecking (and hence have a low effective population size), already show low van der Waals and hydrogen-bond contact densities (18).

Stability is the most universal and well understood biophysical property of proteins, and successes in protein folding and engineering are testaments to how much we understand stability from first principles. However, in nature, protein evolution must reckon with the stochastic processes of mutation and purifying selection, making the effective population size a crucial variable (16). Protein evolution likewise needs to reckon with emerging constraints in cell biology, such as the selection against protein misfolding (1,19), where abundance scales with the selective pressure felt by an evolving gene.

SUPPORTING MATERIAL

Supplementary Table S1 and Fig. S1 are available at http://www.biophysj.org/biophysj/supplemental/S0006-3495(12)05148-X.

ACKNOWLEDGMENTS

We thank N. Dokholyan for the use of ERIS, and Z. Rimas for discussions.

This work was supported by the National Institutes of Health. S.Y.R. Lee received funding from the Harvard College Research Program.

REFERENCES and FOOTNOTES

1. Drummond, D. A., and C. O. Wilke. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell.* 134:341–352.
2. Bucciantini, M., E. Giannoni, …, M. Stefani. 2002. Inherent toxicity of aggregates implies a common mechanism for protein misfolding diseases. *Nature.* 416:507–511.
3. Yang, J. R., S. M. Zhuang, and J. Zhang. 2010. Impact of translational error-induced and error-free misfolding on the rate of protein evolution. *Mol. Syst. Biol.* 6:421.
4. Serohijos, A. W., Z. Rimas, and E. I. Shakhnovich. 2012. Protein biophysics explains why highly abundant proteins evolve slowly. *Cell Rep.* 2:249–256.
5. Yang, Z., and R. Nielsen. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* 17:32–43.
6. Kryazhimskiy, S., and J. B. Plotkin. 2008. The population genetics of dN/dS. *PLoS Genet.* 4:e1000304.
7. Geiler-Samerotte, K. A., M. F. Dion, …, D. A. Drummond. 2011. Misfolded proteins impose a dosage-dependent fitness cost and trigger a cytosolic unfolded protein response in yeast. *Proc. Natl. Acad. Sci. USA.* 108:680–685.
8. Kumar, M. D., K. A. Bava, …, A. Sarai. 2006. ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. *Nucleic Acids Res.* 34(Database issue):D204–D206.
9. Ghaemmaghami, S., W. K. Huh, …, J. S. Weissman. 2003. Global analysis of protein expression in yeast. *Nature.* 425:737–741.
10. Cherry, J. L. 2010. Highly expressed and slowly evolving proteins share compositional properties with thermophilic proteins. *Mol. Biol. Evol.* 27:735–741.
11. Murzin, A. G., S. E. Brenner, …, C. Chothia. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247:536–540.
12. Yin, S., F. Ding, and N. V. Dokholyan. 2007. Eris: an automated estimator of protein stability. *Nat. Methods.* 4:466–467.
13. Kortemme, T., A. V. Morozov, and D. Baker. 2003. An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. *J. Mol. Biol.* 326:1239–1259.
14. Ghosh, K., and K. Dill. 2010. Cellular proteomes have broad distributions of protein stability. *Biophys. J.* 99:3996–4002.
15. Malmström, L., M. Riffle, …, D. Baker. 2007. Superfamily assignments for the yeast proteome through integration of structure prediction with the gene ontology. *PLoS Biol.* 5:e76.
16. Wylie, C. S., and E. I. Shakhnovich. 2011. A biophysical protein folding model accounts for most mutational fitness effects in viruses. *Proc. Natl. Acad. Sci. USA.* 108:9916–9921.
17. Lynch, M., and J. S. Conery. 2003. The origins of genome complexity. *Science.* 302:1401–1404.
18. Tokuriki, N., C. J. Oldfield, …, D. S. Tawfik. 2009. Do viral proteins possess unique biophysical features? *Trends Biochem. Sci.* 34:53–59.
19. Chen, Y., and N. V. Dokholyan. 2008. Natural selection against protein aggregation on self-interacting and essential proteins in yeast, fly, and worm. *Mol. Biol. Evol.* 25:1530–1533.