Journal of Theoretical Biology 391 (2016) 21–34

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Selection maintaining protein stability at equilibrium

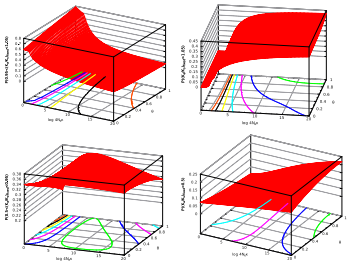
Sanzo Miyazawa

6-5-607 Miyanodai, Sakura, Chiba 285-0857, Japan

HIGHLIGHTS

- Protein stability is kept at equilibrium by random drift and positive selection.
- Neutral selection is predominant only for low-abundant, non-essential proteins.
- Protein abundance more decreases evolutionary rate for less-constrained proteins.
- Structural constraint more decreases evolutionary rate for less-abundant, less-essential proteins.
- Protein stability ($-\Delta G_e/kT$) and $\langle K_a/K_s\rangle$ are predicted to decrease as growth temperature increases.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 31 July 2015
Received in revised form
29 November 2015
Accepted 1 December 2015
Available online 8 December 2015

Keywords:
Neutral theory
Positive selection
Evolutionary rate
Structural constraints
Protein abundance

ABSTRACT

The common understanding of protein evolution has been that neutral mutations are fixed by random drift, and a proportion of neutral mutations depending on the strength of structural and functional constraints primarily determines evolutionary rate. Recently it was indicated that fitness costs due to misfolded proteins are a determinant of evolutionary rate and selection originating in protein stability is a driving force of protein evolution. Here we examine protein evolution under the selection maintaining protein stability.

Protein fitness is a generic form of fitness costs due to misfolded proteins; $s=\kappa\exp(\Delta G/kT)(1-\exp(\Delta\Delta G/kT))$, where s and $\Delta\Delta G$ are selective advantage and stability change of a mutant protein, ΔG is the folding free energy of the wildtype protein, and κ is a parameter representing protein abundance and indispensability. The distribution of $\Delta\Delta G$ is approximated to be a bi-Gaussian distribution, which represents structurally slightly- or highly-constrained sites. Also, the mean of the distribution is negatively proportional to ΔG.

The evolution of this gene has an equilibrium point (ΔG_e) of protein stability, the range of which is consistent with observed values in the ProTherm database. The probability distribution of K_a/K_s, the ratio of nonsynonymous to synonymous substitution rate per site, over fixed mutants in the vicinity of the equilibrium shows that nearly neutral selection is predominant only in low-abundant, non-essential proteins of $\Delta G_e > -2.5$ kcal/mol. In the other proteins, positive selection on stabilizing mutations is significant to maintain protein stability at equilibrium as well as random drift on slightly negative mutations, although the average $\langle K_a/K_s\rangle$ is less than 1. Slow evolutionary rates can be caused by both high protein abundance/indispensability and large effective population size, which produces positive shifts of $\Delta\Delta G$ through decreasing ΔG_e, and strong structural constraints, which directly make $\Delta\Delta G$ more positive. Protein abundance/indispensability more affect evolutionary rate for less constrained proteins,

E-mail address: sanzo.miyazawa@gmail.com

http://dx.doi.org/10.1016/j.jtbi.2015.12.001

and structural constraint for less abundant, less essential proteins. The effect of protein indispensability on evolutionary rate may be hidden by the variation of protein abundance and detected only in low-abundant proteins. Also, protein stability ($-\Delta G_e/kT$) and $\langle K_a/K_s \rangle$ are predicted to decrease as growth temperature increases.

1. Introduction

The common understanding of protein evolution has been that amino acid substitutions observed in homologous proteins are selectively neutral (Kimura, 1968, 1969; Kimura and Ohta, 1971, 1974) or slightly deleterious (Ohta, 1973, 1992), and random drift is a primary force to fix amino acid substitutions in population. The rate of protein evolution has been understood to be determined primarily by the proportion of neutral mutations, which may be measured by the ratio of nonsynonymous to synonymous substitution rate per site (K_a/K_s) (Miyata and Yasunaga, 1980) and determined by functional density (Zuckerkandl, 1976) weighted by the relative variability at specific-function sites of a protein (Go and Miyazawa, 1980). Since then, these theories have been widely accepted, however, recently a question has been raised on whether the diversity of protein evolutionary rate among genes can be explained only by the proportion and the variability of specific-function sites, and molecular and population-genetic constraints on protein evolutionary rate have been explored.

Recent works have revealed that protein evolutionary rate is correlated with gene expression level; highly expressed genes evolve slowly, accounting for as much as 34% of rate variation in yeast (Pál et al., 2001). Of course, there are many reports that support a principle of lower evolution rate for stronger functional density. Broadly expressed proteins in many tissues tend to evolve slower than tissue-specific ones (Kuma et al., 1995; Duret and Mouchiro, 2000). The connectivity of well-conserved proteins in a network is shown (Fraser et al., 2002) to be negatively correlated with their rate of evolution, because a greater proportion of the protein sequence is directly involved in its function. A fitness cost due to protein–protein misinteraction affects the evolutionary rate of surface residues (Yang et al., 2012). Protein dispensability in yeast is correlated with the rate of evolution (Hirsh and Fraser, 2001, 2003), although there is a report insisting on no correlation between them (Pál et al., 2003). Other reports indicate that the correlation between gene dispensability and evolutionary rate, although low, is significant (Zhang and He, 2005; Wall et al., 2005; Jordan et al., 2002).

It was proposed (Drummond et al., 2005; Drummond and Wilke, 2008; Geiler-Samerotte et al., 2011) that low substitution rates of highly expressed genes could be explained by fitness costs due to functional loss and toxicity (Stoebel et al., 2008; Geiler-Samerotte et al., 2011) of misfolded proteins. Misfolding reduces the concentration of functional proteins, and wastes cellular time and energy on production of useless proteins. Also misfolded proteins form insoluble aggregates (Geiler-Samerotte et al., 2011). Fitness cost due to misfolded proteins is larger for highly expressed genes than for less expressed ones.

Fitness cost due to misfolded proteins was formulated (Drummond and Wilke, 2008; Geiler-Samerotte et al., 2011) to be related to the proportion of misfolded proteins. Knowledge of protein folding indicates that protein folding primarily occurs in two-state transition (Miyazawa and Jernigan, 1982a, 1982b), which means that the ensemble of protein conformations are a mixture of completely folded and unfolded conformations. Free energy (ΔG) of protein stability, which is equal to the free energy of the denatured state subtracted from that of the native state, and stability change ($\Delta\Delta G$) due to amino acid substitutions are collected in the ProTherm database (Kumar et al., 2006), although the data are not sufficient. Prediction methods, however, for $\Delta\Delta G$ are improved enough to reproduce real distributions of $\Delta\Delta G$ (Schymkowitz et al., 2005; Yin et al., 2007). Therefore, on the biophysical basis, the distribution of fitness can be estimated and protein evolution can be studied. Shakhnovich group studied protein evolution on the basis of knowledge of protein folding (Serohijos and Shakhnovich, 2014; Dasmeh et al., 2014) and showed (Serohijos et al., 2012) that the negative correlation between protein abundance and K_a/K_s was caused by the distribution of $\Delta\Delta G$ that negatively correlates with the ΔG of a wild type. Also, it was shown (Serohijos et al., 2013) that highly abundant proteins had to be more stable than low abundant ones. Relationship between evolutionary rate and protein stability is studied from various points of view (Echave et al., 2015; Faure and Koonin, 2015).

Here we study relationship between evolutionary rate and selection on protein stability in a monoclonal approximation. A fitness assumed here for a protein is a generic form to which all formulations (Drummond and Wilke, 2008; Geiler-Samerotte et al., 2011; Serohijos et al., 2012, 2013; Serohijos and Shakhnovich, 2014; Dasmeh et al., 2014) previously employed for protein fitness are reduced in the condition of $\exp(\beta\Delta G) \ll 1$, which is satisfied in the typical range of folding free energies shown in Fig. 1; $\beta = 1/(kT)$, k is the Boltzmann constant and T is absolute temperature. The generic form of Malthusian fitness of a protein-coding gene is $m \equiv -\kappa \exp(\beta\Delta G)$, where κ is a parameter, which

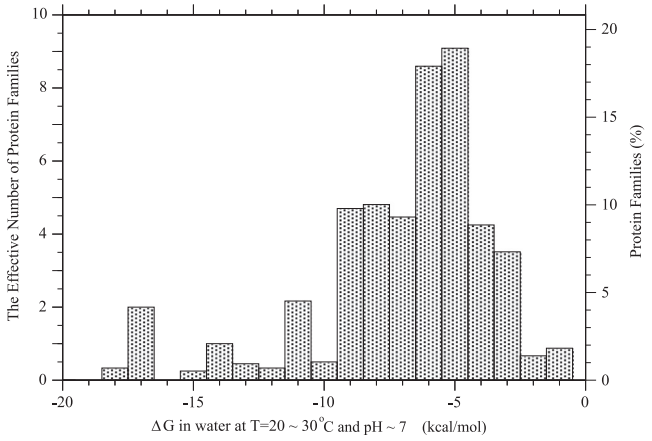


Fig. 1. Distribution of folding free energies of monomeric protein families. Stability data of monomeric proteins for which the item of dG_H2O or dG was obtained in the experimental condition of $6.7 \le \mathrm{pH} \le 7.3$ and $20\ ^\circ\mathrm{C} \le T \le 30\ ^\circ\mathrm{C}$ and their folding-unfolding transition is two state and reversible are extracted from the ProTherm (Kumar et al., 2006); in the case of dG only thermal transition data are used. Thermophilic proteins, and proteins observed with salts or additives are also removed. An equal sampling weight is assigned to each species of homologous protein, and the total sampling weight of each protein family is normalized to one. In the case in which multiple data exist for the same species of protein, its sampling weight is divided to each of the data. However, proteins whose stabilities are known may be samples biased from the protein universe. The value, $\Delta G_e = -5.24$, kcal/mol of equilibrium stability at the representative parameter values, $\log 4N_e\kappa = 7.55$ and $\theta = 0.53$, agrees with the most probable value of ΔG in the distribution above. Also, the range of ΔG shown above is consistent with that range, -2 to -12.5 kcal/mol, expected from the present model. The kcal/mol unit is used for ΔG. A similar distribution was also compiled (Zeldovich et al., 2007).