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

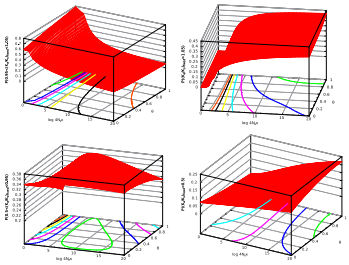
Sanzo Miyazawa

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

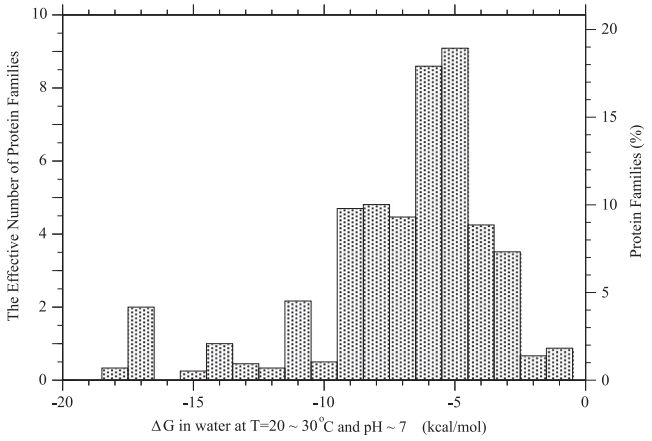


Fig. 1. Distribution of folding free energies of monomeric protein families. Stability data of monomeric proteins for which the item of dG_H2O or dG was obtained in the experimental condition of $6.7 \le \mathrm{pH} \le 7.3$ and $20\ ^\circ\mathrm{C} \le T \le 30\ ^\circ\mathrm{C}$ and their folding-unfolding 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).

may be a function of protein abundance and dispensability; see Methods for details. The distribution of stability change $\Delta\Delta G$ due to single amino acid substitutions is approximated as a weighted sum of two Gaussian functions that was shown (Tokuriki et al., 2007) to well reproduce actual distributions of $\Delta\Delta G$. One of the two Gaussian functions describes substitutions at structurally less-constrained surface sites, and the other at more-constrained core sites of proteins. The proportion of less-constrained surface sites is a parameter (θ).

The fixation probability of a mutant with $\Delta\Delta G$ can be calculated for a duploid population with effective population size N_e (Crow and Kimura, 1970). In the population of genes with such a fitness protein stability is evolutionarily maintained at equilibrium, and equilibrium stability (ΔG_e) negatively correlates with protein abundance/dispensability (κ). The range of ΔG_e is consistent with the observed range of folding free energies shown in Fig. 1.

The probability density functions (PDF) of K_a/K_s, the ratio of nonsynonymous to synonymous substitution rate per site (Miyata and Yasunaga, 1980), at equilibrium and also in the vicinity of equilibrium are numerically examined over a whole domain of the parameters, $0 \le \log 4N_e\kappa \le 20$ and $0 \le \theta \le 1$. The dependences of evolutionary rate on protein abundance/dispensability and on structural constraint are quantitatively described, and it is shown that both factors cannot be ignored on protein evolutionary rate, although protein abundance/indispensability more affect evolutionary rate for less constrained proteins, and structural constraint for less abundant, less essential proteins. Like protein abundance, protein indispensability must correlate with evolutionary rate, but a correlation between them may be hidden by the variation of protein abundance as well as effective population size, and detected only in low-abundant proteins. It has also become clear that nearly neutral selection is predominant only in low-abundant, non-essential proteins with $\log 4N_e\kappa < 2$ or $\Delta G_e > -2.5$ kcal/mol, and in the other proteins positive selection is significant to more stabilize a less-stable wild type. Also, a significant amount of slightly negative mutants are fixed in population by random drift. This view of protein evolution is contrary to the previous understanding. The present model based on a biophysical knowledge of protein stability also indicates that protein stability ($-\beta\Delta G_e$) and the average of K_a/K_s decrease as growth temperature increases.

2. Methods

2.1. Fitness costs due to misfolded proteins

Misfolding can impose costs in three distinct ways (Geiler-Samerotte et al., 2011); loss of function, diversion of protein synthesis resources away from essential proteins, and toxicity of the misfolded molecules. Fitness cost due to functional loss was formulated (Drummond and Wilke, 2008) by taking account of protein dispensability. Assuming that fitness cost of each gene is additive in the Malthusian fitness scale, the total Malthusian fitness of a genome was estimated as

$$m_{\text{dispensability}} \equiv -\sum_i \gamma_i (1 - f_i^{\text{native}}) \tag{1}$$

where $-\gamma_i$ is defined as $-\gamma_i \equiv \log$(deletion-strain growth rate/max growth rate), and f_i^{native} is the fraction of the native conformation for gene i.

Protein folding primarily occurs in the two-state transition, which means that protein conformations are a mixture of completely folded and unfolded conformations (Miyazawa and Jernigan, 1982a, 1982b). Therefore, if the completely folded (native) state is more stable by a free energy difference ΔG than the unfolded (denatured) state, then the native fraction in the conformational ensemble will be equal to

$$f^{\text{native}} = \frac{e^{-\beta\Delta G}}{1 + e^{-\beta\Delta G}} \tag{2}$$

where $\beta = 1/kT$; k is the Boltzmann constant and T is absolute temperature.

Thus, Eq. (1) for the Malthusian fitness of a genome can be transformed as follows in terms of the folding free energy ΔG of the native conformation:

$$m_{\text{dispensability}} = -\sum_i \gamma_i \frac{e^{\beta\Delta G_i}}{e^{\beta\Delta G_i} + 1} \tag{3}$$

Because of $\exp(\beta\Delta G) \ll 1$ in the typical range of folding free energies shown in Fig. 1, the above definition of fitness is approximated by

$$m_{\text{dispensability}} = -\sum_i \gamma_i \left[e^{\beta\Delta G_i} - O(e^{2\beta\Delta G_i}) \right] \tag{4}$$

Drummond and Wilke (2008) took notice of toxicity of misfolded proteins as well as diversion of protein synthesis resources, and formulated the Malthusian fitness (m_{misfolds}) of a genome to be negatively proportional to the total amount of misfolded proteins, which must be produced to obtain the necessary amount of folded proteins (Serohijos et al., 2012):

$$m_{\text{misfolds}} = -c \sum_i A_i \frac{1 - f_i^{\text{native}}}{f_i^{\text{native}}} \tag{5}$$

$$m_{\text{misfolds}} = -c \sum_i A_i e^{\beta\Delta G_i} \tag{6}$$

where c is a positive constant and assumed to be $c = 0.0001$, and A_i is the abundance of protein i.

2.2. Fitness of a linear metabolic pathway

Serohijos and Shakhnovich (2014) examined the evolution of a linear metabolic pathway whose Wrightian fitness was defined as

$$w_{\text{linear pathway}} \equiv w_{\text{flux}} + w_{\text{misfolds}} \tag{7}$$

$$w_{\text{flux}} \equiv \frac{\sum_i \varepsilon_i A_i^{-1}}{\sum_i \varepsilon_i (A_i f_i^{\text{native}})^{-1}} \tag{8}$$

$$w_{\text{misfolds}} \equiv -c \sum_i A_i (1 - f_i^{\text{native}}) \tag{9}$$

where ε_i was defined as enzyme efficiency and assumed to be $\varepsilon_i = 1$. The w_{flux} is a fitness originating from the enzymatic flux of a linear metabolic pathway, and w_{misfolds} represents the effect of toxicity of misfolded proteins, and is the same functional form as Eq. (1), although Eq. (1) is a definition for Malthusian fitness. Then, the Malthusian fitness corresponding to the Wrightian fitness above can be represented as

$$m_{\text{linear\_pathway}} \tag{10}$$

$$\begin{aligned} &= \log \left[\left\{ 1 + \sum_i \frac{\varepsilon_i A_i^{-1}}{\sum_i \varepsilon_i A_i^{-1}} s e^{\beta\Delta G_i} \right\}^{-1} - c \sum_i A_i e^{\beta\Delta G_i} (1 + e^{\beta\Delta G_i})^{-1} \right] \\ &= -\sum_i \left\{ \frac{\varepsilon_i A_i^{-1}}{(\sum_i \varepsilon_i A_i^{-1})} + cA_i \right\} e^{\beta\Delta G_i} \\ &\quad + O\left(\left(\sum_i \left\{ \frac{\varepsilon_i A_i^{-1}}{(\sum_i \varepsilon_i A_i^{-1})} + cA_i \right\} e^{\beta\Delta G_i} \right)^2 \right) \end{aligned} \tag{11}$$

Because $cA_i \le 0.459$ (Serohijos and Shakhnovich, 2014), $\Delta G < -3$ and $\sum_{i=1}^{10} cA_i \exp(\beta\Delta G_i) < 0.03$, the higher order terms can be neglected in this case. However, the fitness costs due to the flux

and misfolded proteins may be formulated to be additive in the Malthusian scale rather than in the Wrightian scale, employing Eq. (6) for the fitness cost due to misfolded proteins;

$$m_{\text{linear\_pathway}} \equiv -\sum_i \left\{ \frac{\varepsilon_i A_i^{-1}}{\sum_i \varepsilon_i A_i^{-1}} + cA_i \right\} e^{\beta \Delta G_i} \tag{12}$$

2.3. Other formulations of protein fitness

Also, the following simple definition for fitness to maintain protein stability was used (Dasmeh et al., 2014):

$$w \propto f_{\text{native}} \tag{13}$$

$$m = -e^{\beta \Delta G} + O(e^{2\beta \Delta G}) + \text{constant} \tag{14}$$

In addition, Eq. (3) for functional loss was employed with $\gamma_i \Rightarrow cA_i$ to represent toxicity of misfolded proteins in (Serohijos et al., 2012, 2013).

2.4. A generic form of protein fitness

Thus, all expressions above for Malthusian fitness of protein can be well approximated by the following expression, because of $\exp(\beta \Delta G) \ll 1$ in the typical range of folding free energies shown in Fig. 1:

$$m \equiv -\sum_i \kappa_i e^{\beta \Delta G_i} \quad \text{with } \kappa_i \geq 0 \tag{15}$$

where κ_i is a parameter. If the fitness costs of functional loss and toxicity due to misfolded proteins are taken into account, κ_i will be defined as

$$\kappa_i = cA_i + \gamma_i \geq 0 \tag{16}$$

assuming their additivity in the Malthusian fitness scale.

The selective advantage of a mutant, in which each protein is destabilized by $\Delta\Delta G_i$, to the wild type can be represented by

$$s \equiv m^{\text{mutant}} - m^{\text{wildtype}} = \sum_i s_i \tag{17}$$

$$s_i = \kappa_i e^{\beta \Delta G_i} (1 - e^{\beta \Delta \Delta G_i}) \quad \text{with } \kappa_i \geq 0 \tag{18}$$

3. Results

3.1. Protein stability and fitness

Here, we consider the evolution of a single protein-coding gene in which the selective advantage of mutant proteins in Malthusian parameters is assumed to be

$$s = \kappa e^{\beta \Delta G} (1 - e^{\beta \Delta \Delta G}) \quad \text{with } \kappa \geq 0 \tag{19}$$

and therefore s is upper-bounded by

$$s \leq \kappa e^{\beta \Delta G} \tag{20}$$

where ΔG is the stability of a wild-type protein, $\Delta\Delta G$ is a stability change of a mutant protein, $\beta = 1/kT$; unless specified, $\beta = 1/0.593$ kcal^{-1} mol corresponding to $T = 298^\circ$ K. κ is a parameter whose meaning may depend on the situation; refer to Method for details. If the fitness costs of functional loss and toxicity due to misfolded proteins are both taken into account and assumed to be additive in the Malthusian fitness scale, κ will be defined as

$$\kappa = cA + \gamma \tag{21}$$

where c is fitness cost per misfolded protein (Drummond and Wilke, 2008; Geiler-Samerotte et al., 2011), A is the cellular abundance of the protein (Drummond and Wilke, 2008; Geiler-Samerotte et al., 2011), and γ is indispensability (Drummond and Wilke, 2008) and defined to be $\gamma = -\log($ deletion-strain growth rate/max growth rate). Equation (19) indicates that the selective advantage s is upper-bounded by $\kappa \exp(\beta \Delta G)$. The parameter κ is assumed in the present analysis to take values in the range of $0 \leq \log\ 4N_e\kappa \leq 20$ with effective population size N_e, taking account of the values of the parameters, $c \sim 10^{-4}$ (Drummond and Wilke, 2008), $10 < A < 10^6$ (Ghaemmaghami et al., 2003), $\gamma = 10$ for essential genes (Drummond and Wilke, 2008), and $N_e \sim 10^4$ to 10^5 for vertebrates, $\sim 10^5$ to 10^6 for invertebrates, $\sim 10^7$ to 10^8 for unicellular eukaryotes, and $> 10^8$ for prokaryotes (Lynch and Conery, 2003). The above ranges of the parameters indicate that the effect of protein indispensability (γ) may be hidden by the variation of protein abundance (cA) as well as effective population size (N_e), and may be detected only in low-abundant proteins.

Based on measurements of stability changes due to single amino acid substitutions in proteins, which are collected in the ProTherm database (Kumar et al., 2006), Serohijos et al. (2012) reported that the distribution of $\Delta\Delta G$ is approximately a Gaussian distribution with mean = 1 kcal/mol and standard deviation = 1.7 kcal/mol. In addition, it was shown (Serohijos et al., 2012) that the mean of $\Delta\Delta G$ is negatively proportional to ΔG, and that this dependence of the mean of $\Delta\Delta G$ on ΔG is not large but still important to cause the observed negative correlation between protein abundance and evolutionary rate. On the other hand, (Tokuriki et al., 2007) computationally predicted $\Delta\Delta G$ for all possible single amino acid substitutions in 21 different globular, single domain proteins, and showed that the predicted distributions of $\Delta\Delta G$ were strikingly similar despite a range of protein sizes and folds and largely follow a bi-Gaussian distribution: one of the two Gaussian distributions results from substitutions on protein surfaces and is a narrow distribution with a mildly destabilizing mean $\Delta\Delta G$, whereas the other due to substitutions in protein cores is a wider distribution with a stronger destabilizing mean (Tokuriki et al., 2007).

Here, according to Tokuriki et al. (2007), the distribution of $\Delta\Delta G$ due to single amino acid substitutions is approximated as a bi-Gaussian function with the dependence of mean $\Delta\Delta G$ on ΔG, in order to examine the effects of structural constraint on evolutionary rate. The probability density function (PDF) of $\Delta\Delta G$, $p(\Delta\Delta G)$, for nonsynonymous substitutions is assumed to be

$$p(\Delta\Delta G) = \theta \mathcal{N}(\mu_s, \sigma_s) + (1-\theta)\mathcal{N}(\mu_c, \sigma_c) \tag{22}$$

where $0 \leq \theta \leq 1$, and $\mathcal{N}(\mu, \sigma)$ is a normal distribution with mean μ and standard deviation σ. Since the majority of substitutions appear to be single nucleotide substitutions, the values of the standard deviations (σ_s and σ_c) estimated in Tokuriki et al. (2007) for single nucleotide substitutions are employed here; in kcal/mol units,

$$\mu_s = -0.14\Delta G - 0.17, \quad \sigma_s = 0.90 \tag{23}$$

$$\mu_c = -0.14\Delta G + 1.23, \quad \sigma_c = 1.93 \tag{24}$$

To analyze the dependences of the means, μ_s and μ_c, on ΔG, we plotted the observed values of $\Delta\Delta G$ of single amino acid mutants against ΔG of the wild type, which are collected in the ProTherm database (Kumar et al., 2006); the same analysis was done by Serohijos et al. (2012). Fig. 2 shows a significant dependence of $\Delta\Delta G$ on ΔG; the regression line is $\mu = -0.14\Delta G + 0.49$. The linear slopes of μ_s and μ_c are taken to be equal to the slope (-0.14) of the regression line. The intercepts have been estimated to satisfy the following two conditions:

1. Equations (23) and (24) satisfy $\mu_s(\Delta G_0) = 0.56$ and $\mu_c(\Delta G_0) = 1.96$, which were estimated for single nucleotide

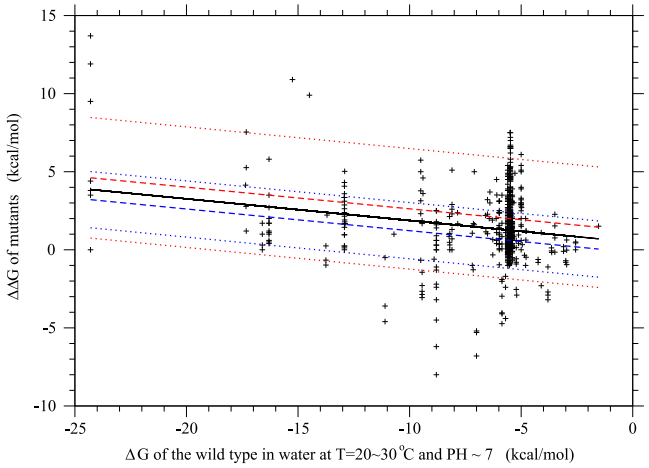


Fig. 2. Dependence of stability changes, $\Delta\Delta G$, due to single amino acid substitutions on the protein stability, ΔG, of the wild type. A solid line shows the regression line, $\Delta\Delta G = -0.14\Delta G + 0.49$; the correlation coefficient and p-value are equal to -0.20 and $< 10^{-7}$, respectively. Broken lines show two means of bi-Gaussian distributions, μ_s in blue and μ_c in red. Blue dotted lines show $\mu_s \pm 2\sigma_s$ and red dotted lines $\mu_c \pm 2\sigma_c$. See Eqs. (22)–(24) for the bi-Gaussian distribution. Stability data of single amino acid mutants for which the items dG_H2O and ddG_H2O or dG and ddG were 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 transitions are two state and reversible are extracted from the ProTherm (Kumar et al., 2006). In the case of dG only thermal transition data are used. In the case in which multiple data exist for the same protein, only one of them is used. The kcal/mol unit is used for $\Delta\Delta G$ and ΔG. A similar distribution was also compiled (Serohijos et al., 2012). (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

substitutions in Tokuriki et al. (2007), at a certain value (ΔG_0) of ΔG.

2. The total mean of the two Gaussian functions agrees with the regression line, $\mu = -0.14\Delta G + 0.49$. The value of θ is taken to be 0.53, which is equal to the average of θ over proteins used in Tokuriki et al. (2007).

A representative value, 7.550, of log $4N_e\kappa$ is determined in such way that the equilibrium value of ΔG is equal to $\Delta G_0 = -5.24$ introduced above; ΔG_e is explicitly defined later. It is interesting that this value $\Delta G_e = -5.24$ kcal/mol agrees with the most probable value of ΔG in the observed distribution of protein stabilities shown in Fig. 1. The fraction θ of less-constrained residues such as most residues on protein surface is correlated with protein length for globular, monomeric proteins; $\theta = 1.27 - 0.33 \cdot \log_{10}(\text{protein length})$ for $50 \le \text{length} \le 330$ (Tokuriki et al., 2007). However, residues taking part in protein-protein interactions may be regarded as core residues rather than surface residues.

The dependence of the PDF, $p(\Delta\Delta G)$, of $\Delta\Delta G$ on θ is shown in Fig. 3. Also, the PDF of selective advantage, $p(4N_es) = -p(\Delta\Delta G)d\Delta\Delta G/\, d4N_es$, is shown in Fig. S.4 to have a peak at a small, positive value of selective advantage, which moves toward more positive values as θ and/or $4N_e\kappa$ increase.

3.2. Equilibrium state of protein stability in protein evolution

The fixation probability u for a mutant gene with selective advantage s and gene frequency q in a duploid system of effective population size N_e was given as a function of $4N_es$ and q by (Crow and Kimura, 1970)

$$u(4N_es) = \frac{1 - e^{-4N_esq}}{1 - e^{-4N_es}} \tag{25}$$

where $q = 1/(2N)$ for a single mutant gene in a population of size N. Population size is taken to be $N = 10^6$. The ratio of the substitution rate per nonsynonymous site (K_a) for nonsynonymous substitutions with selective advantage s to the substitution rate per synonymous site (K_s) for nonsynonymous substitutions with $s = 0$ is

$$\frac{K_a}{K_s} = \frac{u(4N_es)}{u(0)} = \frac{u(4N_es)}{q} \quad \text{with } q = \frac{1}{2N} \tag{26}$$

$$\simeq \frac{4N_es}{1 - e^{-4N_es}} \quad \text{for} \frac{|4N_esq|}{2} \ll 1 \tag{27}$$

assuming that synonymous substitutions are completely neutral and mutation rates at both types of sites are the same. Equations (19) and (25) indicate that $4N_e\kappa$ can be regarded as a single parameter for K_a/K_s. Furthermore, if the dependence of the mean $\Delta\Delta G$ on ΔG could be neglected, $4N_e\kappa\exp(\beta\Delta G)$ could be regarded as a single parameter. In the range of $|4N_esq|/2 \ll 1$, both K_a/K_s and the PDF of K_a/K_s do not depend on $q = 1/(2N)$; see Eq. (27) and (S.15).

The PDF of $\Delta\Delta G$ of fixed mutant genes, $p(\Delta\Delta G_{\text{fixed}})$, is

$$p(\Delta\Delta G_{\text{fixed}}) \equiv p(\Delta\Delta G)\frac{u(4N_es)}{\langle u\rangle} \tag{28}$$

$$\langle u\rangle \equiv \int_{-\infty}^{\infty} u(4N_es)p(\Delta\Delta G)d\Delta\Delta G \tag{29}$$

where $\langle u\rangle$ is the average fixation rate. Fig. 3 shows the PDF of $\Delta\Delta G$ of fixed mutant genes. The PDF of $4N_es$ in fixed mutants is also shown in Fig. S.4; $p(4N_es_{\text{fixed}}) = -p(\Delta\Delta G_{\text{fixed}})d\Delta\Delta G/d4N_es$. Then, the average of $\Delta\Delta G$ in fixed mutant genes can be calculated; $\langle\Delta\Delta G\rangle_{\text{fixed}} \equiv \int_{-\infty}^{\infty} \Delta\Delta G\, p(\Delta\Delta G_{\text{fixed}})\, d\Delta\Delta G$.

Fig. 4 shows the average of the $\Delta\Delta G$ over fixed mutant genes, $\langle\Delta\Delta G\rangle_{\text{fixed}}$, to monotonically decrease with ΔG, indicating that the temporal process of ΔG is stable at $\langle\Delta\Delta G\rangle_{\text{fixed}}(\Delta G_e) = 0$ due to the balance between random drift on destabilizing mutations and positive selection on stabilizing mutations; ΔG_e is the folding free energy at the equilibrium state. If a wild-type protein becomes less stable than the equilibrium, $\Delta G > \Delta G_e$, more stabilizing mutants will fix due to primarily positive selection and secondarily random drift, because stabilizing mutants will increase due to negative shifts of $\Delta\Delta G$ and also the effect of stability change on selective advantage will be more amplified; see Eqs. (23) and (24) for the dependence of $\Delta\Delta G$ on ΔG, and Eq. (19) for the fitness of stability change. As shown in Fig. S.6, the probability of $K_a/K_s > 1.0$, that is, positive selection, significantly increases as ΔG becomes more positive than the equilibrium stability ΔG_e. On the other hand, if a wild-type protein becomes more stable than the equilibrium, $\Delta G < \Delta G_e$, more destabilizing mutants will fix due to random drift, because destabilizing mutants will increase due to positive shifts of $\Delta\Delta G$ and also more destabilizing mutants become nearly neutral due to the less-amplified effect of stability change on selective advantage. As shown later, the PDF of K_a/K_s in the vicinity of equilibrium confirms this mechanism for maintaining protein stability at equilibrium.

It was claimed (Serohijos et al., 2012, 2013) that the equilibrium point would correspond to the minimum of the average fixation probability. However, in Fig. 4 for log $4N_e\kappa = 7.550$ and $\theta = 0.53$, the average $\langle s\rangle_{\text{fixed}}$ of selective advantage in fixed mutants has a minimum at $\Delta G = -5.50$ kcal/mol and changes its sign at $\Delta G = -4.58$ kcal/mol, where the average $\langle K_a/K_s\rangle = \langle u\rangle/q$ has a minimum and which is more positive than the equilibrium stability $\Delta G_e = -5.24$ kcal/mol. In other words, Fig. 4 and Fig. S.16 show that the values of ΔG at $\langle\Delta\Delta G\rangle_{\text{fixed}} = 0$ and at the minimum of $\langle K_a/K_s\rangle$ may be close but differ from each other, and indicate that the value of ΔG corresponding to the minimum of $\langle K_a/K_s\rangle$ is not a good approximation for the equilibrium stability, because $\langle K_a/K_s\rangle$ gently changes in the vicinity of the equilibrium stability as shown in Fig. S.16.

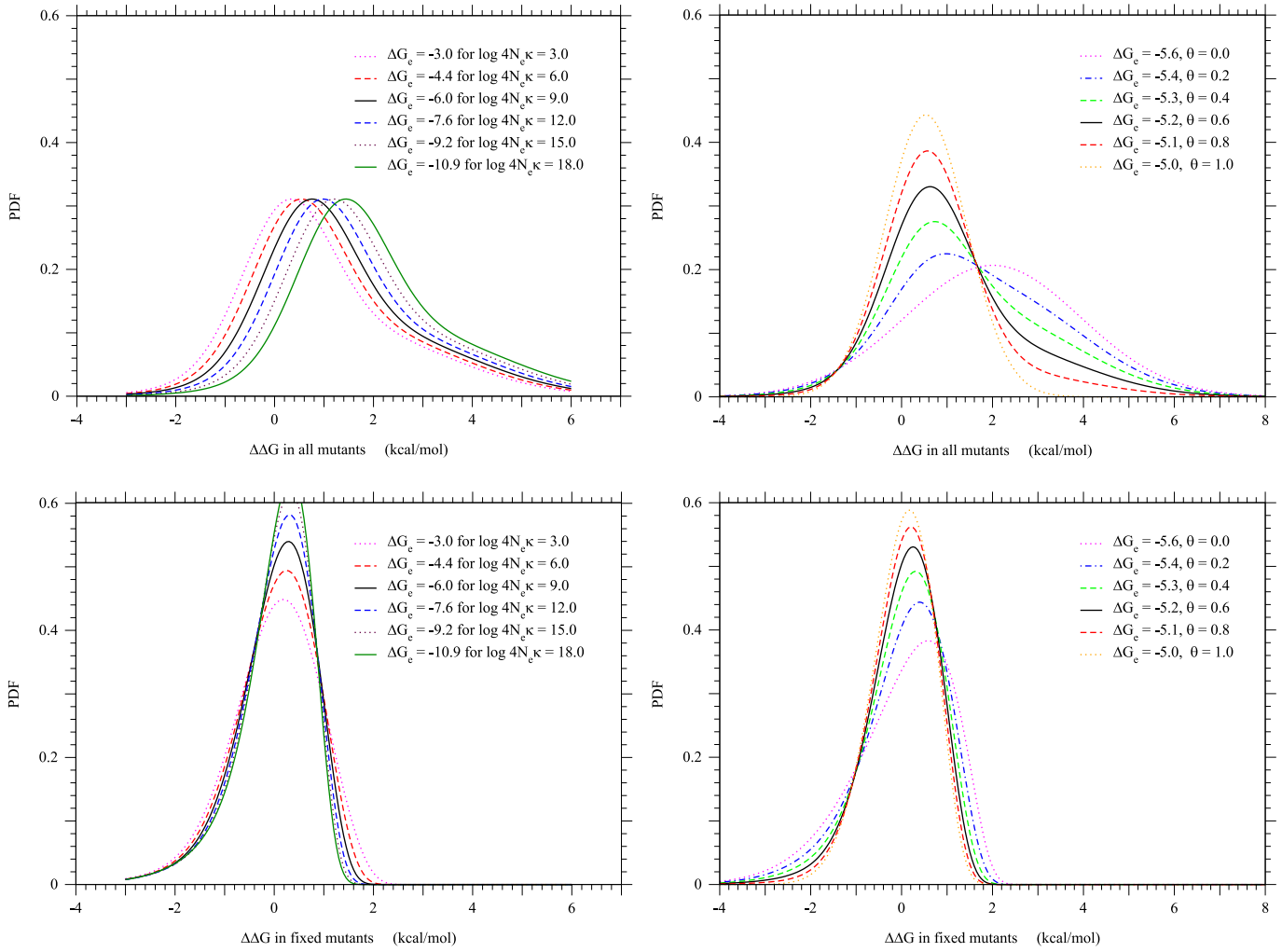


Fig. 3. PDFs of stability changes, $\Delta\Delta G$, due to single amino acid substitutions in all mutants and in fixed mutants at equilibrium of protein stability, $\Delta G = \Delta G_e$. The PDF of $\Delta\Delta G$ due to single amino acid substitutions in all arising mutants is assumed to be bi-Gaussian; see Eq. (22). Unless specified, log $4N_e\kappa = 7.55$ and $\theta = 0.53$ are employed. The kcal/mol unit is used for $\Delta\Delta G$ and ΔG_e.

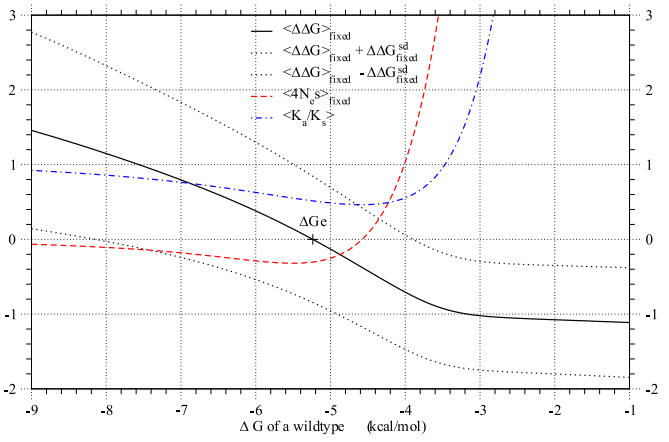


Fig. 4. The average, $\langle\Delta\Delta G\rangle_{\text{fixed}}$, of stability changes over fixed mutants versus protein stability, ΔG, of the wild type. ΔG_e, where $\langle\Delta\Delta G\rangle = 0$, is the stable equilibrium value of folding free energy, ΔG, in protein evolution. The averages of $\Delta\Delta G$, $4N_e s$, and K_a/Ks over fixed mutants are plotted against protein stability, ΔG, of the wild type by solid, broken, and dash-dot lines, respectively. Thick dotted lines show the values of $\langle\Delta\Delta G\rangle_{\text{fixed}} \pm \Delta\Delta G^{\text{sd}}_{\text{fixed}}$, where $\Delta\Delta G^{\text{sd}}_{\text{fixed}}$ is the standard deviation of $\Delta\Delta G$ over fixed mutants. log $4N_e\kappa = 7.55$ and $\theta = 0.53$ are employed. The kcal/mol unit is used for $\Delta\Delta G$ and ΔG.

3.3. Equilibrium stability, ΔG_e

The equilibrium value, ΔG_e, of ΔG that satisfies $\langle\Delta\Delta G_{\text{fixed}}\rangle = 0$ in fixed mutants depends directly on θ and indirectly on $4N_e\kappa$ through fixation probability; see Eqs. (19) and (25). As shown in Fig. 5, ΔG_e depends weakly on θ. On the other hand, ΔG_e depends more strongly on and is almost negatively proportional to log $4N_e\kappa$, as also shown in real proteins (Serohijos et al., 2013). If the dependence of the means, μ_s and μ_c in Eqs. (23) and (24), of $\Delta\Delta G$ in all mutants on ΔG could be neglected, $4N_e\kappa \exp(\beta\Delta G)$ could be regarded as a single parameter, and so $4N_e\kappa \exp(\beta\Delta G_e)$ would be constant, irrespective of $4N_e\kappa$. Thus, the dependence of log $4N_e\kappa \exp(\beta\Delta G_e)$ on log $4N_e\kappa$ shown in Fig. 5 is caused solely by the linear dependence of the means μ_s and μ_c of $\Delta\Delta G$ on ΔG (Serohijos et al., 2012). It is interesting to know that as log $4N_e\kappa$ varies from 0 to 20, ΔG_e changes from -1.5 to -12.5 kcal/mol, the range of which is consistent with experimental values of protein folding free energies shown in Fig. 1.

3.4. K_a/K_s at equilibrium, $\Delta G = \Delta G_e$

Equations (23) and (24) indicate that the distribution of $\Delta\Delta G$ shifts toward the positive direction as ΔG becomes more negative.

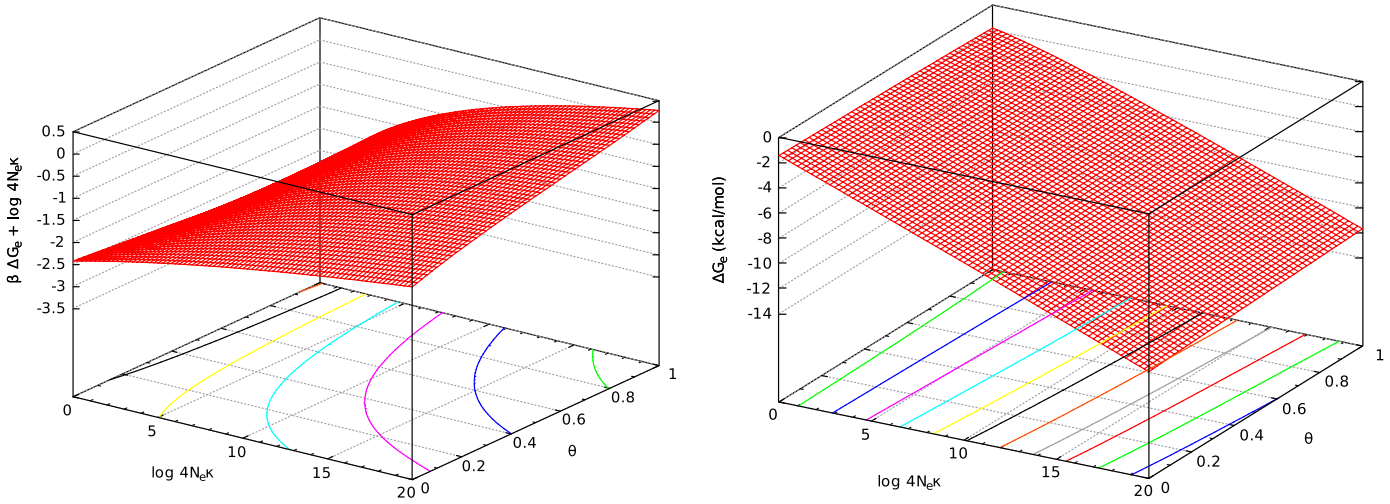


Fig. 5. Dependence of equilibrium stability, ΔG_e, on parameters, $4N_e\kappa$ and θ. ΔG_e is the equilibrium value of folding free energy,ΔG, in protein evolution. The value of $\beta\Delta G_e + \log 4N_e\kappa$ is the upper bound of $\log 4N_e s$, and would be constant if the mean of $\Delta\Delta G$ in all arising mutants did not depend on ΔG; see Eq. (19). The kcal/mol unit is used for ΔG_e.

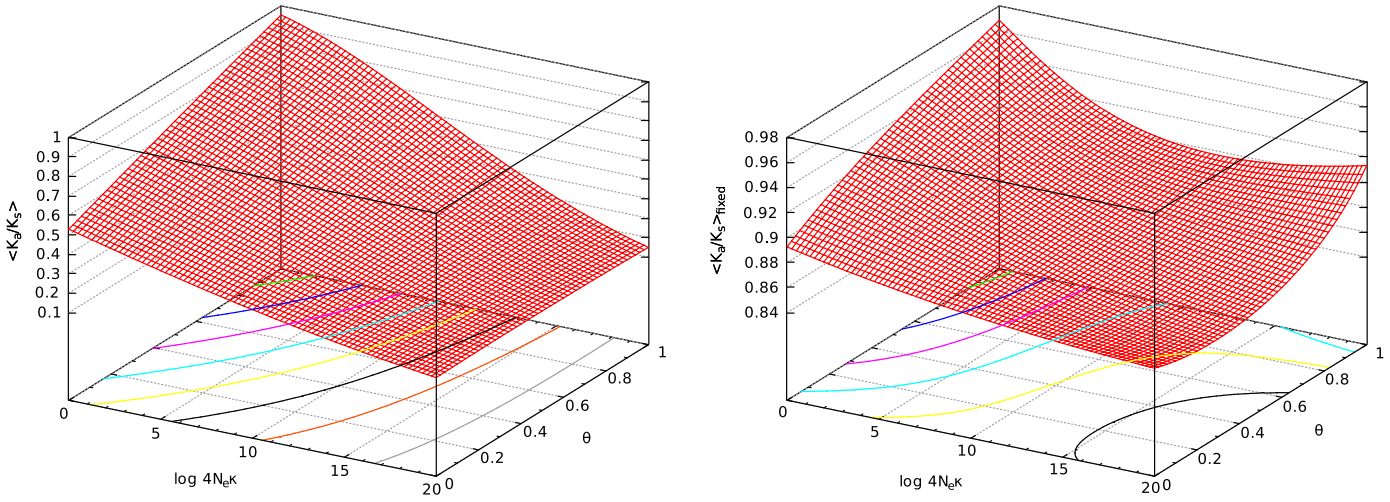


Fig. 6. The average of K_a/K_s over all mutants or over fixed mutants only at equilibrium of protein stability, $\Delta G = \Delta G_e$.

Hence, increasing $4N_e\kappa$ that makes ΔG_e more negative results in positive shifts of the distribution of $\Delta\Delta G$, which increase destabilizing mutations. In addition, as indicated by Eq. (19), the upper bound of $4N_e s$, $4N_e\kappa\, ex\, p(\beta\Delta G_e)$, scales the effect of $\Delta\Delta G$ on protein fitness. The larger $4N_e\kappa \exp(\beta\Delta G_e)$ is, the larger the effect of $\Delta\Delta G$ on selective advantage becomes. Thus, the increase of $4N_e\kappa\exp(\beta\Delta G_e)$ caused by the increase of κ and/or N_e increases both destabilizing mutations and their fitness costs, and results in slow evolutionary rates for proteins with large κ and/or N_e. In other words, highly expressed and indispensable genes, and genes with a large effective population size must evolve slowly. On the other hand, the decrease of θ, that is, the increase of highly constrained residues directly shifts the average of $\Delta\Delta G$ in all arising mutants toward the positive direction, and causes slow evolutionary rates.

The average of K_a/K_s over all mutants, which can be observed as the ratio of average nonsynonymous substitution rate per nonsynonymous site to average synonymous substitution rate per synonymous site, and also that over fixed mutants only are shown in Fig. 6. At any value of θ, $\langle K_a/K_s\rangle$ decreases as $\log 4N_e\kappa$ increases, explaining the observed relationship that highly expressed and indispensable genes evolve slowly (Drummond et al., 2005; Drummond and Wilke, 2008; Serohijos et al., 2012). Likewise, at any value of $\log 4N_e\kappa$, $\langle K_a/K_s\rangle$ increases as θ increases. In other words, the more structurally constrained a protein is, the more slowly it evolves. The effect of protein abundance/indispensability on evolutionary rate is more remarkable for less constrained proteins and the effect of structural constraint is more remarkable for less abundant, less essential proteins.

The average of K_a/K_s over all mutants, $\langle K_a/K_s\rangle$, is less than 1.0 on the whole domain shown in the figure, indicating that the average of K_a/K_s over a long time interval and over many sites should not show any positive selection. Even the average of K_a/K_s over fixed mutants is less than 1.0, and falls into a narrow range of 0.97–0.85, which is much narrower than a range of 0.96–0.15 for that over all mutants; the average of K_a/K_s over fixed mutants is equal to $\langle(K_a/K_s)^2\rangle/\langle K_a/K_s\rangle$, and as a matter of course must be equal to or larger than the averages of K_a/K_s over all mutants. However, the average of K_a/K_s over a short time interval and over a small number of sites may exhibit values larger than one. In Fig. 7, the PDFs of K_a/K_s for all mutants and also for fixed mutants only are shown; $p(K_a/K_s) = p(4N_e s)d(4N_e s)/d(K_a/K_s)$. A significant fraction of fixed mutants fix with $K_a/K_s > 1$.

Arbitrarily, the value of K_a/K_s is categorized into four classes; negative, slightly negative, nearly neutral, and positive selection categories whose K_a/K_s are within the range of $K_a/K_s \le 0.5, 0.5 < K_a/K_s \le 0.95, 0.95 < K_a/K_s \le 1.05$, and $1.05 < K_a/K_s$,

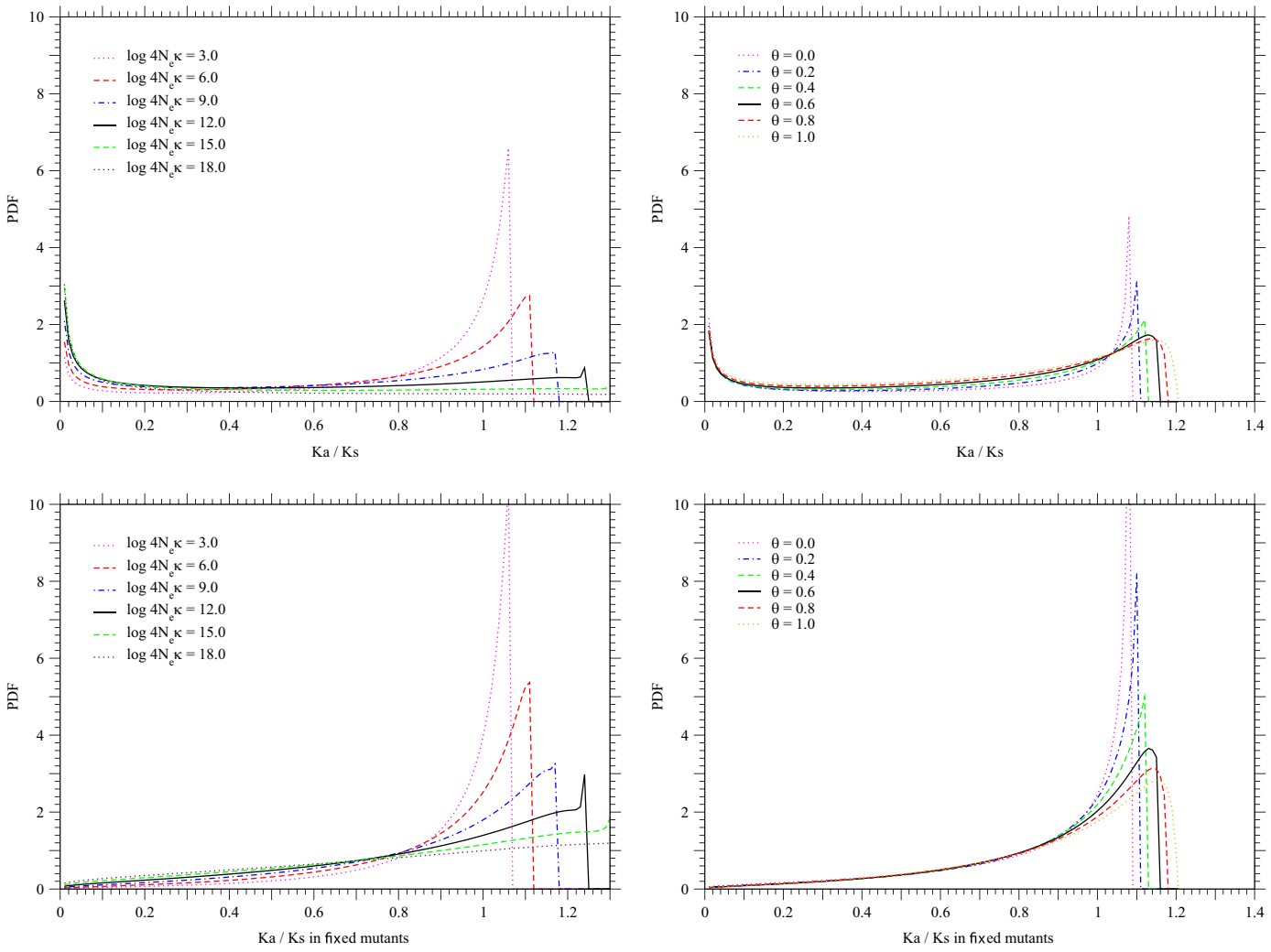


Fig. 7. PDFs of K_a/K_s in all mutants and in fixed mutants only at equilibrium of protein stability, $\Delta G = \Delta G_e$. Unless specified, log $4N_e\kappa = 7.55$ and $\theta = 0.53$ are employed.

respectively. Then, the probabilities of each selection category in all mutants and in fixed mutants are calculated and shown in Fig. S.10 and Fig. 8, respectively. At the largest abundance (log $4N_e\kappa = 20$) most arising mutations are negative mutations whose K_a/K_s are less than 0.5. This is reasonable, because at this condition the wild-type protein is very stable with low equilibrium values ΔG_e as shown in Fig. 5, and therefore most mutations destabilize the wild-type and tend to be removed from population. Most fixed mutants are positive mutants or slightly negative mutants fixed by random drift. Nearly neutral mutants are less than 3% of all mutants, and less than 15% of fixed mutants.

On the other hand, at the other extreme of log $4N_e\kappa < 2$, there are no mutations of the positive selection category, this is because the upper bound of K_a/K_s, which corresponds to the upper bound ($4N_e\kappa \exp(\beta\Delta G)$) of $4N_es$, at the equilibrium stability ΔG_e becomes less than 1.05 that is the lower bound for the positive selection category; see Eq. (19). The significant amount of mutations become nearly neutral. As θ changes from 1 to 0, that is, structural constraints increase, the proportion of nearly neutral mutations changes from 0.75(0.56) to 0.31(0.22), and instead negative mutations increase and most of them are removed from population. Thus, the selection mechanism for fixing stabilizing mutants in little expressed, non-essential genes (log $4N_e\kappa < 2$) is not positive selection but nearly neutral selection, that is, random drift.

The probability of each selection category in fixed mutants depends strongly on $4N_e\kappa$, but much less on θ. Current common understanding is that amino acid substitutions in protein evolution are either neutral (Kimura, 1968) or lethal, at most slightly deleterious (Ohta, 1973) or lethal, unless functional selection operates and functional changes occur. On the contrary, nearly neutral fixations are predominant only in proteins with log $4N_e\kappa < 2$ or $\Delta G_e > -2.5$ kcal/mol, and positive selection is significant in the other proteins. On the other hand, slightly negative selection is always significant. An interesting result is that the effects of structural constraint on K_a/K_s are the most remarkable in proteins with log $4N_e\kappa < 2$ or instead $\Delta G_e > -2.5$ kcal/mol in which nearly neutral fixations are predominant.

3.5. K_a/K_s in the vicinity of equilibrium

In Fig. 4, the $\langle\Delta\Delta G\rangle_{\text{fixed}} \pm$ standard deviation of $\Delta\Delta G$ of fixed mutants are also drawn. The standard deviation of $\Delta\Delta G$ of fixed mutants is equal to 0.84 kcal/mol at the equilibrium, ΔG_e, indicating that protein stability ΔG fluctuates more or less within $\Delta G_e \pm 0.84$ kcal/mol instantaneously. Such a deviation from the equilibrium must be canceled by compensatory substitutions that consecutively occur, otherwise the protein stability would far depart from its equilibrium point.

In Figs. 9 and 10 and Figs. S.12 and S.14, the probabilities of each selection category in fixed mutants and in all arising mutants are shown as a function of ΔG and $4N_e\kappa$ or θ, respectively. The range of ΔG around ΔG_e shown by a blue line on the surface grid is within

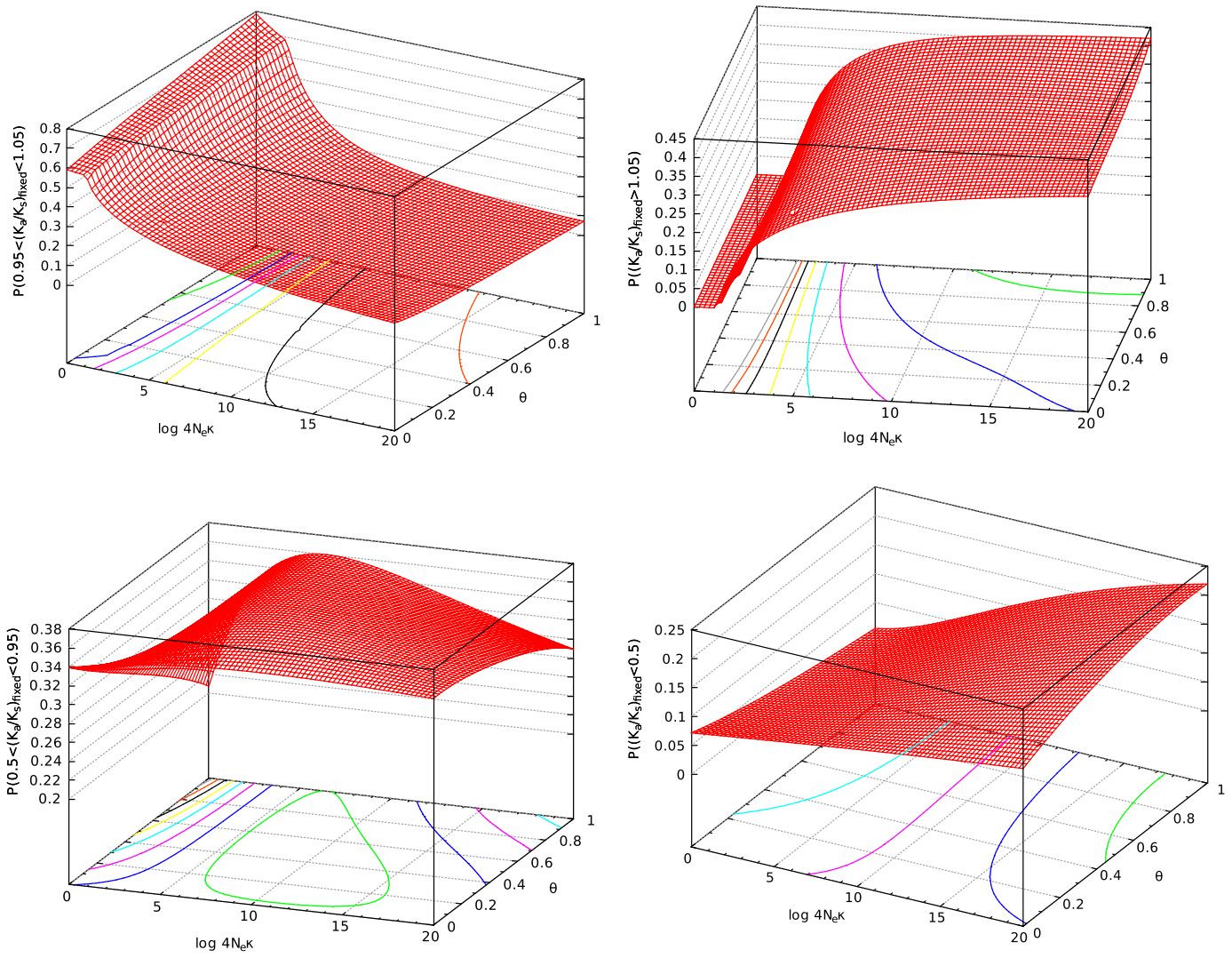


Fig. 8. Probability of each selection category in fixed mutants at equilibrium of protein stability, $\Delta G = \Delta G_e$. Arbitrarily, the value of K_a/K_s is categorized into four classes; negative, slightly negative, nearly neutral, and positive selection categories in which K_a/K_s is within the ranges of $K_a/K_s \le 0.5, 0.5 < K_a/K_s \le 0.95, 0.95 < K_a/K_s \le 1.05$, and $1.05 < K_a/K_s$, respectively.

two times of the standard deviation of $\Delta\Delta G$ over fixed mutants at ΔG_e.

As indicated by Eqs. (23) and (24), it is shown in Figs. S.12 and S14 that stabilizing mutations increase due to negative shifts of $\Delta\Delta G$ as the wild type becomes less stable than the equilibrium, $\Delta G > \Delta G_e$, and that destabilizing mutations increase due to positive shifts of $\Delta\Delta G$ as $\Delta G < \Delta G_e$. In addition, as indicated by Eq. (19), it is shown in Figs. 9 and 10 that positive selection on stabilizing mutants is more amplified as $\Delta G > \Delta G_e$, and that more destabilizing mutations become nearly neutral due to the less-amplified effect of stability change on selective advantage as $\Delta G < \Delta G_e$. This is a mechanism of maintaining protein stability at equilibrium.

3.6. Lower bound of K_a/K_s for adaptive substitutions on protein function

The observed value of $K_a/K_s > 1$ is often used to indicate functional selection. The averages of K_a/K_s over all mutants and even over fixed mutants are less than 1 as shown in Fig. 6. Therefore, the average of K_a/K_s over long time or many sites does not indicate positive selection. However, the probability of $K_a/K_s > 1$ is not negligible as shown in Figs. 7 and 8. Then, a question is how large K_a/K_s due to selection on protein stability can be.

The distribution of K_a/K_s significantly changes with ΔG, as shown in Fig. S.6 and Fig. 11. It may be appropriate to see the average of K_a/K_s, $\langle K_a/K_s\rangle_{\text{fixed}}$, in mutants fixed at $\Delta G > \Delta G_e$, because the upper bound of K_a/K_s becomes larger for $\Delta G > \Delta G_e$ than at the equilibrium, and also positive mutants must fix to improve the protein stability of the wild type. Fig. 11 shows that $\langle K_a/K_s\rangle_{\text{fixed}}$ can be very large for proteins with low equilibrium stabilities (large $4N_e\kappa$ and small θ), although $\langle K_a/K_s\rangle_{\text{fixed}} \sim 1$ in $\Delta G < \Delta G_e$ in which nearly neutral and slightly negative selections are predominant 1.7(1.2) at $\Delta G_e + \Delta\Delta G^{\text{sd}}_{\text{fixed}}$ and 6.1(5.6) at $\Delta G_e + 2\cdot\Delta\Delta G^{\text{sd}}_{\text{fixed}}$ for $\log 4N_e\kappa = 20$ ($\theta = 0.0$), where $\Delta\Delta G^{\text{sd}}_{\text{fixed}}$ means the standard deviation of $\Delta\Delta G$ in fixed mutants at ΔG_e. The 85% of fixed mutants have $\Delta\Delta G$ within the standard deviation. Therefore, a lower bound for adaptive substitutions may be taken to be about 1.7, which is almost equal to the upper bound of K_a/K_s at the equilibrium for $\log 4N_e\kappa = 20$ and $\theta = 1$; see Fig. S.17. However, as shown in Fig. S.17, the more genes are expressed and/or the stronger structural constraints are, the larger the upper bound of K_a/K_s at the equilibrium is. Judging of adaptive

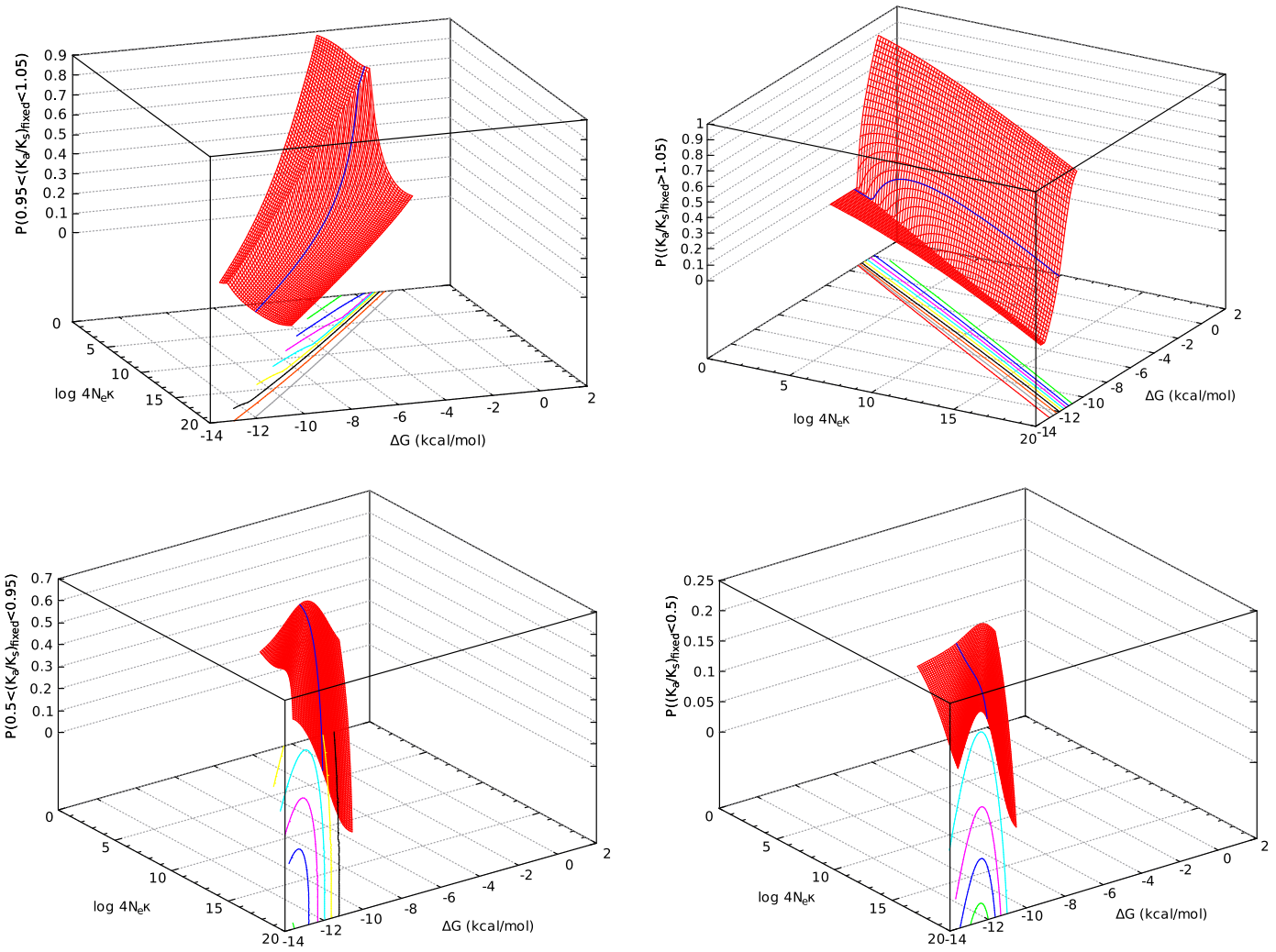


Fig. 9. Dependence of the probability of each selection category in fixed mutants on $4N_e\kappa$ and ΔG. A blue line on the surface grid shows $\Delta G = \Delta G_e$, which is the equilibrium value of ΔG in protein evolution. The range of ΔG shown in the figures is $|\Delta G - \Delta G_e| < 2 \cdot \Delta\Delta G^{sd}_{fixed}$, where $\Delta\Delta G^{sd}_{fixed}$ is the standard deviation of $\Delta\Delta G$ over fixed mutants at $\Delta G = \Delta G_e$. Arbitrarily, the value of K_a/K_s is categorized into four classes; negative, slightly negative, nearly neutral, and positive selection categories in which K_a/K_s is within the ranges of $K_a/K_s \le 0.5, 0.5 < K_a/K_s \le 0.95, 0.95 < K_a/K_s \le 1.05$, and $1.05 < K_a/K_s$, respectively. $\theta = 0.53$ is employed. The kcal/mol unit is used for ΔG. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

changes may need not only $K_a/K_s > 1$ but also other supporting evidences; such that substitutions are localized at specific sites.

3.7. Dependences of protein stability ($\beta\Delta G_e$) and evolutionary rate ($\langle K_a/K_s \rangle$) on growth temperature

It is natural that the folding free energies, ΔG_e, of proteins in organisms growing at higher temperatures must be lower than those at the normal temperature, in order to attain the same stabilities and fitnesses as in the normal temperature. Equations (2) and (15) indicate that the same stability and fitness will be attained if $\beta\Delta G_e$ is constant. It means that it is sufficient for ΔG_e at the 100 °C to decrease $373/298 = 1.25$ times of that at the normal temperature (25 °C). Is this a figure expected for folding free energies of thermophilic proteins at high growth temperature? It is not enough data of ΔG at high temperature in the ProTherm (Kumar et al., 2006) to answer this question; $\Delta G(T = 75\ ^\circ\text{C}) = 10.76$ kcal/mol for oxidized and 4.3 for reduced CuA domain of cytochrome oxidase from Thermus thermophilus (Wittung-Stafshede et al., 1998) and $\Delta G(T = 60\ ^\circ\text{C}) = 13.01$ kcal/mol for pyrrolidone carboxyl peptidase from Pyrococcus (Ogasahara et al., 1998). The present model indicates that $\beta\Delta G_e$ slightly increases as growth temperature increases.

In Fig. 12 and Fig. S.18, $\beta\Delta G_e + \log\ 4N_e\kappa$ is shown as a function of absolute temperature T and $\log\ 4N_e\kappa$ or θ, assuming that the distribution of $\Delta\Delta G$ and its dependency on ΔG do not depend on T, that is, Eqs. (22)–(24). At fixed values of $\log\ 4N_e\kappa$ and θ, $\beta\Delta G_e + \log\ 4N_e\kappa$ increases as T increases, meaning that protein stability, $-\beta\Delta G$, decreases as growth temperature increases. This tendency is slightly larger at smaller values of $\log\ 4N_e\kappa$, that is, for less abundant proteins.

The effects of growth temperature on K_a/K_s are shown in Fig. S.19. The present model predicts that $\langle K_a/K_s \rangle$ decreases as growth temperature increases unless any other parameter does not change.

4. Discussion

Recently, fitness costs due to misfolded proteins have been widely noticed, particularly neurological disorder linked to misfolded protein toxicity (Bucciantini et al., 2002). Fitness costs that originate in functional loss (Geiler-Samerotte et al., 2011) and in diversion of protein synthesis and aggregation of proteins have been evaluated (Drummond and Wilke, 2008) to be related to the proportion of misfolded proteins. Also, previous studies indicate

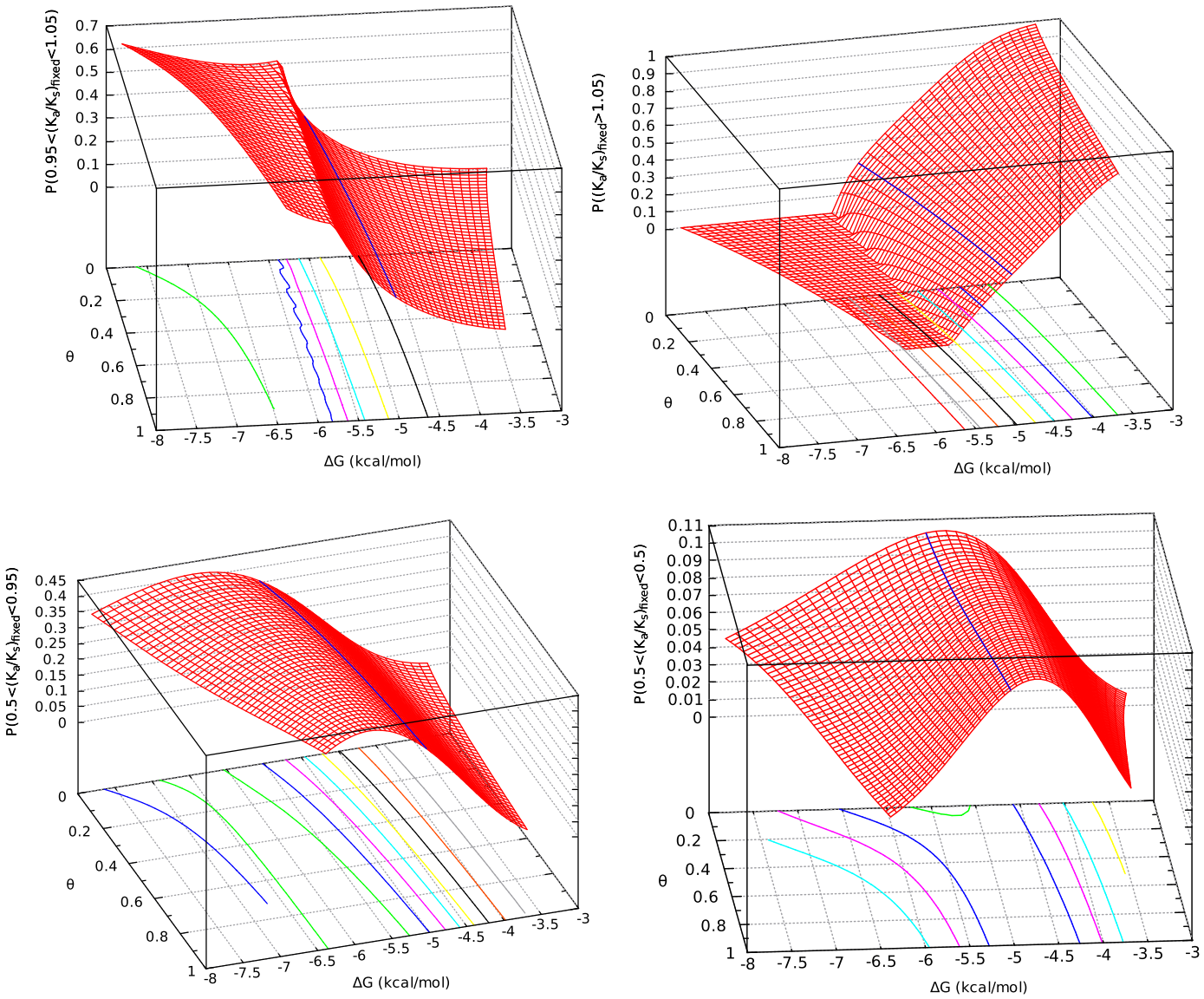


Fig. 10. Dependence of the probability of each selection category in fixed mutants on θ and ΔG. A blue line on the surface grid shows $\Delta G = \Delta G_e$, which is the equilibrium value of ΔG in protein evolution. The range of ΔG shown in the figures is $|\Delta G - \Delta G_e| < 2 \cdot \Delta\Delta G^{sd}_{fixed}$, where $\Delta\Delta G^{sd}_{fixed}$ is the standard deviation of $\Delta\Delta G$ over fixed mutants at $\Delta G = \Delta G_e$. Arbitrarily, the value of K_a/K_s is categorized into four classes; negative, slightly negative, nearly neutral, and positive selection categories in which K_a/K_s is within the ranges of $K_a/K_s \le 0.5, 0.5 < K_a/K_s \le 0.95, 0.95 < K_a/K_s \le 1.05$, and $1.05 < K_a/K_s$, respectively. $\log 4N_e\kappa = 7.55$ is employed. The kcal/mol unit is used for ΔG. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

that factors that relate protein stability to protein fitness are protein abundance, protein indispensability, and structural constraints of protein. Current knowledge of protein folding can provide an exact formulation for the proportion of misfolded proteins as a function of folding free energy, and reasonable predictions (Schymkowitz et al., 2005; Yin et al., 2007; Tokuriki et al., 2007) of stability changes due to single amino acid substitutions in protein native structures. Thus, on the basis of knowledge of protein biophysics it became possible to study the effects of amino acid substitutions on protein stability and then the evolution of protein (Drummond and Wilke, 2008; Serohijos et al., 2012; Serohijos and Shakhnovich, 2014; Echave et al., 2015; Faure and Koonin, 2015).

Here, the effects of protein abundance and indispensability (κ) and of structural constraint (θ) on protein evolutionary rate (K_a/K_s) have been examined in detail. Both the effects are represented with different functional forms. Structural constraints affect the distribution of stability change $\Delta\Delta G$ due to mutations. On the other hand, protein abundance/indispensability affects the effectiveness of stability change on protein fitness as well as the distribution of $\Delta\Delta G$.

The common understanding of protein evolution has been that amino acid substitutions found 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. However, there is a selection maintaining protein stability at equilibrium (Drummond and Wilke, 2008; Serohijos et al., 2012; Serohijos and Shakhnovich, 2014). From the present analysis of the PDF of K_a/K_s, it has become clear how the equilibrium of stability is maintained; see Figs. 9 and 10. In less-stable proteins of $\Delta G > \Delta G_e$, more stabilizing mutations fix due to positive selection, because negative shifts of $\Delta\Delta G$ increase stabilizing mutants and also more amplify the effect of stability change on selective advantage; see Eqs. (23), (24) and (19). In more-stable proteins of $\Delta G < \Delta G_e$, more destabilizing mutants are fixed by random drift, because positive shifts of $\Delta\Delta G$ increase destabilizing mutants and also make more destabilizing mutants become nearly neutral with

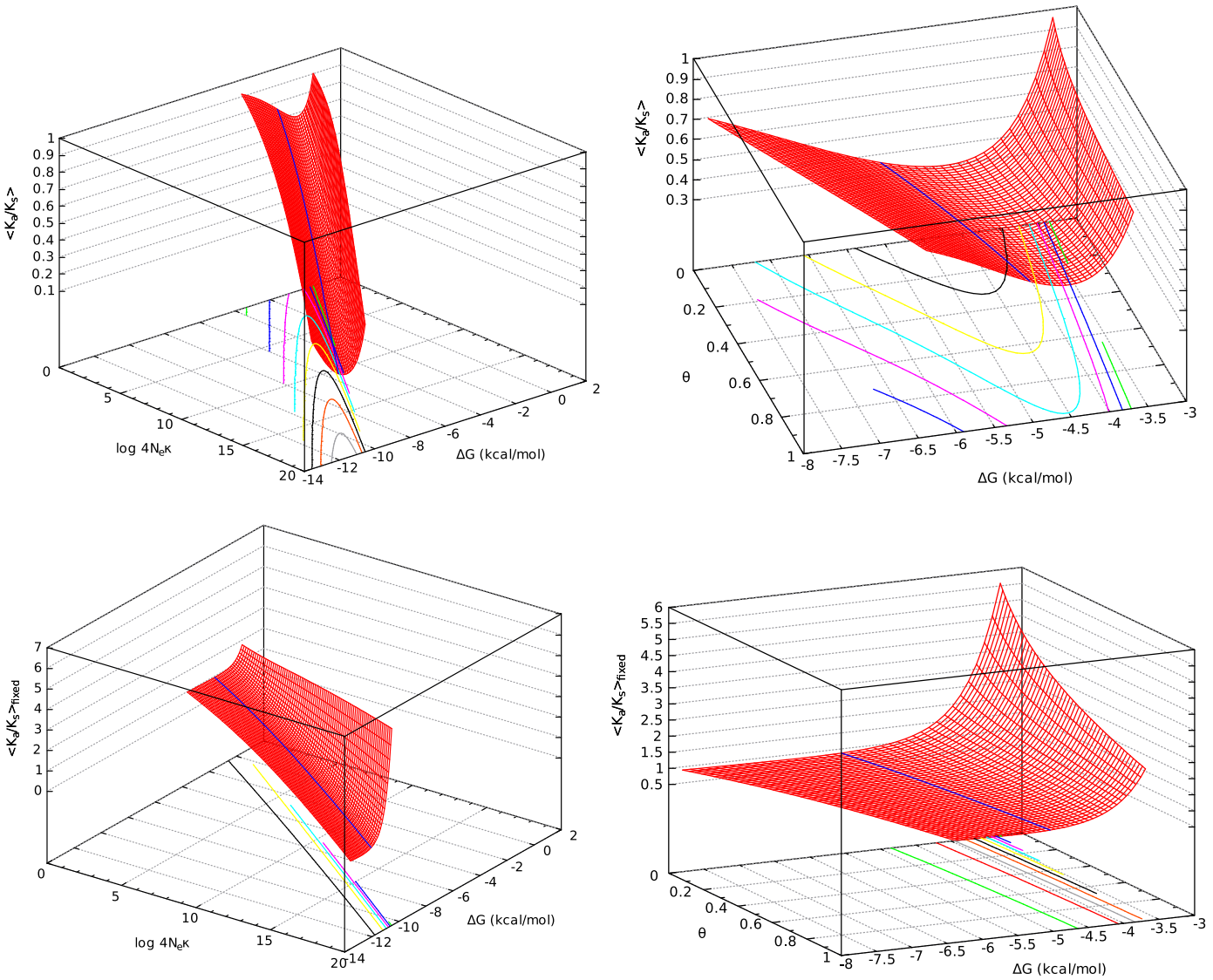


Fig. 11. Dependence of the average of K_a/K_s over all mutants or over fixed mutants only on protein stability, ΔG, of the wild type. A blue line on the surface grid shows $\Delta G = \Delta G_e$, which is the equilibrium value of ΔG in protein evolution. The range of ΔG shown in the figures is $|\Delta G - \Delta G_e| < 2 \cdot \Delta\Delta G^{sd}_{fixed}$, where $\Delta\Delta G^{sd}_{fixed}$ is the standard deviation of $\Delta\Delta G$ over fixed mutants at $\Delta G = \Delta G_e$. Unless specified, log $4N_e\kappa = 7.55$ and $\theta = 0.53$ are employed. The kcal/mol unit is used for ΔG. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

the less-amplified effect of stability change on selective advantage. It has been revealed that contrary to the neutral theory nearly neutral selection is predominant only in low-abundant, non-essential proteins with log $4N_e\kappa < 2$ or with low equilibrium stability ($\Delta G_e > -2.5$ kcal/mol); see Fig. 8.

The average $\langle K_a/K_s \rangle$ and even $\langle K_a/K_s \rangle_{fixed}$ at equilibrium stability $\Delta G = \Delta G_e$ are less than one over the whole parameter range; see Fig. 6. Hence, as far as selection is on protein stability, the average of K_a/K_s over a long time interval and over many sites will be expected to be less than one, if all synonymous mutations are neutral (Spielman and Wilke, 2015). However, because the probability of $K_a/K_s > 1$ is significant, branches with $K_a/K_s > 1$ in phylogenetic trees may be observed, as observed in a population dynamics simulation (Serohijos and Shakhnovich, 2014), even though synonymous mutations are neutral and no adaptive selection operates on protein function. According to the present estimate, a lower bound of K_a/K_s to indicate adaptive substitutions must be at least as large as 1.7.

Protein equilibrium stability (ΔG_e) has been clearly described here as a function of $4N_e\kappa$ and θ. The more expressed a gene is (the larger $4N_e\kappa$ is), the stabler the wild-type protein at equilibrium is (the more negative ΔG_e becomes); see Fig. 5. The decrease of ΔG_e shifts the distribution of $\Delta\Delta G$ toward the positive direction, generating more highly destabilizing mutants; see Eqs. (23) and (24). In addition, as $4N_e\kappa$ increases, the net effect, $4N_e\kappa \exp(\beta\Delta G_e)$, increases and more amplifies the effects of stability changes ($\Delta\Delta G$) on selective advantage (s); see Fig. 5 and Eq. (19). As a result, highly expressed and indispensable genes, and genes with a large effective population size evolve slowly; see Fig. 6. However, if the distribution of $\Delta\Delta G$ did not depend on ΔG, $4N_e\kappa \exp(\beta\Delta G_e)$ would be constant, and K_a/K_s would not depend on $4N_e\kappa$, that is, protein abundance/indispensability and effective population size.

On the other hand, structural constraints on protein affect protein evolutionary rate by changing the distribution of $\Delta\Delta G$ due to amino acid substitutions. As shown in Fig. 6, at any value of log $4N_e\kappa$, $\langle K_a/K_s \rangle$ decreases as θ decreases. In other word, the more a protein is structurally constrained, the more slowly it evolves, as claimed by Zuckerkandl (1976). Fig. 6 shows that the effect of protein abundance/indispensability on evolutionary rate is more remarkable for less constrained proteins, and the effect of structural constraint is more remarkable for less abundant, less essential proteins.

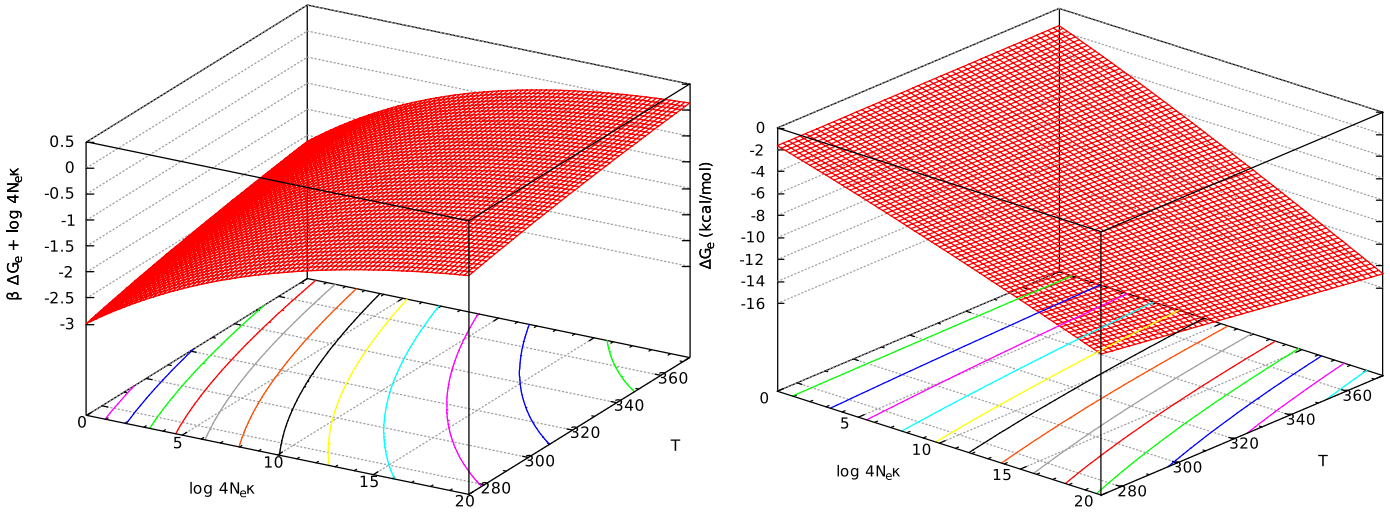


Fig. 12. Dependence of equilibrium stability, ΔG_e, on parameters, $4N_e\kappa$ and T. ΔG_e is the equilibrium value of folding free energy, ΔG, in protein evolution. T is absolute temperature; $\beta = 1/kT$, where k is the Boltzmann constant. Equations (22)–(24) are assumed for the distribution of $\Delta\Delta G$ and its dependency on ΔG; they are assumed to be independent of T. $\theta = 0.53$ is employed. The value of $\beta\Delta G_e + \log 4N_e\kappa$ is the upper bound of $\log 4N_e s$, and would not depend on $\log 4N_e\kappa$ if the mean of $\Delta\Delta G$ in all arising mutants did not depend on ΔG; see Eq. (19). The kcal/mol unit is used for ΔG_e.

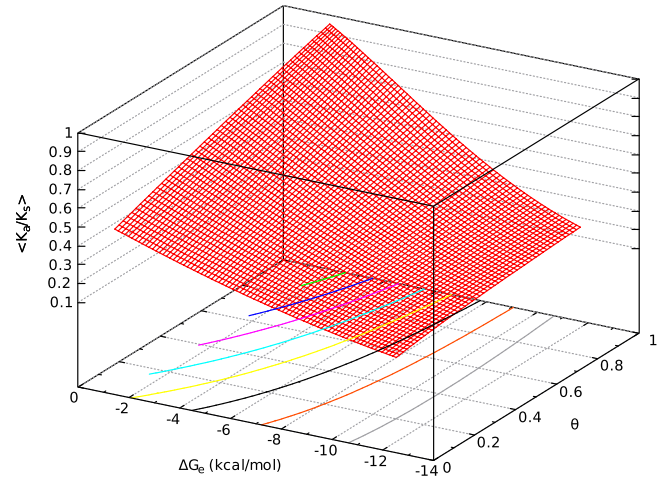


Fig. 13. The average of K_a/K_s over all mutants as a function of ΔG_e and θ.

In the result, the average of K_a/K_s over all arising mutants decreases roughly by 0.4–0.8 as $\log 4N_e\kappa$ increases from 0 to 20; see Fig. 6. On the other hand, it decreases by 0.1–0.4 as the proportion of the residues of the surface type, θ, decreases from 1 to 0. For monomeric, globular proteins, the proportion of protein surface may range from 0.7 to 0.45. Thus, in typical globular proteins, protein abundance/indispensability may cause larger differences of evolutionary rate between proteins than structural constraint. However, proteins that interact with other molecules on protein surface effectively reduce residues of the protein-surface type (Franzosa and Xia, 2009). Both protein abundance/indispensability and structural constraint must be taken into account for protein evolutionary rate.

Protein abundance and indispensability both affect evolutionary rate similarly through protein fitness. It was shown in real proteins that protein abundance correlates with evolutionary rate (Pál et al., 2001). The present model of protein fitness Eq. (19) also indicates that protein indispensability must correlate with evolutionary rate (Hirsh and Fraser, 2001, 2003), but a correlation between them may be hidden by the variation of protein abundance and detected only in low-abundant proteins (Pál et al., 2003); see Eq. (21). In addition, effective population size must affect ΔG_e and $\langle K_a/K_s\rangle$ together with κ as $4N_e\kappa$.

In the present model, protein equilibrium stability (ΔG_e) and evolutionary rate ($\langle K_a/K_s\rangle$) are predictable from θ and $4N_e\kappa$. The proportion of the surface type of residues may be estimated as those whose surface accessibility values (ASA) are less than 0.25 (Tokuriki et al., 2007), but experimental measurements of protein abundance, indispensability, and effective population size to determine $4N_e\kappa$ may be relatively hard. Instead the experimental value of protein stability may be employed as equilibrium stability to predict evolutionary rate and others, although it is not an independent variable. Fig. 13 shows evolutionary rate as a function of ΔG_e and θ. Needless to say, mutational effects on $\Delta\Delta G$, such as θ and the distribution of $\Delta\Delta G$, must be well estimated for various categories of proteins (Faure and Koonin, 2015) to obtain successful predictions. Also, accurate estimations of ΔG for various proteins are needed to examine the present predictions. It is interesting to examine if protein stability ($-\beta\Delta G$) and the average of K_a/K_s decrease as growth temperature increases.

5. Conclusions

- The range, -2 to -12.5 kcal/mol, of equilibrium values, ΔG_e, of protein stability calculated with the present fitness model is consistent with the distribution of experimental values shown in Fig. 1.
- Contrary to the neutral theory, nearly neutral selection is predominant only in low-abundant, non-essential proteins of $\log 4N_e\kappa < 2$ or $\Delta G_e > -2.5$ 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. However, $\langle K_a/K_s\rangle$ and even $\langle K_a/K_s\rangle_{\text{fixed}}$ at $\Delta G = \Delta G_e$ are less than 1.
- Protein abundance/indispensability (κ) and effective population size (N_e) more affect evolutionary rate for less constrained proteins, and structural constraint (θ) for less abundant, less essential proteins.
- Protein indispensability must negatively correlate with evolutionary rate like protein abundance, but the correlation between

them may be hidden by the variation of protein abundance and detected only in low-abundant proteins.
- Evolutionary rates of proteins may be predicted from equilibrium stability (ΔG_e) and structural constraints (PDF of $\Delta\Delta G$) of the protein.
- The present model indicates that protein stability ($-\beta\Delta G_e$) and $\langle K_a/K_s\rangle$ decrease as growth temperature increases.

Appendix A. Supplementary data

Supplementary data associated with this paper can be found in the online version at http://dx.doi.org/10.1016/j.jtbi.2015.12.001.

References

Bucciantini, M., Giannoni, E., Chiti, F., Baroni, F., Formigli, L., Zurdo, J., Taddei, N., Ramponi, G., Dobson, C.M., Stefani, M., 2002. Inherent toxicity of aggregates implies a common mechanism for protein misfolding diseases. Nature 416, 507–511.

Crow, J.F., Kimura, M., 1970. An Introduction to Population Genetics Theory. Harper & Row publishers, New York.

Dasmeh, P., Serohijos, A.W., Kepp, K.P., Shakhnovich, E.I., 2014. The influence of selection for protein stability on dN/dS estimations. Genome Biol. Evol. 6, 2956–2967.

Drummond, D.A., Bloom, J.D., Adami, C., Wilke, C.O., Arnold, F.H., 2005. Why highly expressed proteins evolve slowly. Proc. Natl. Acad. Sci. USA 102 (40), 14338–14343.

Drummond, D.A., Wilke, C.O., 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell 134 (2), 341–352. http://dx.doi.org/10.1016/j.cell.2008.05.042.

Duret, L., Mouchiro, D., 2000. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. Mol. Biol. Evol. 17, 68–85.

Echave, J., Jackson, E.L., Wilke, C.O., 2015. Relationship between protein thermodynamic constraints and variation of evolutionary rates among sites. Phys. Biol. 12 (2), 025002 ⟨http://stacks.iop.org/1478-3975/12/i=2/a=025002⟩.

Faure, G., Koonin, E.V., 2015. Universal distribution of mutational effects on protein stability, uncoupling of protein robustness from sequence evolution and distinct evolutionary modes of prokaryotic and eukaryotic proteins. Phys. Biol. 12 (3), 035001 ⟨http://stacks.iop.org/1478-3975/12/i=3/a=035001⟩.

Franzosa, E.A., Xia, Y., 2009. Structural determinants of protein evolution are context-sensitive at the residue level. Mol. Biol. Evol. 26, 2387–2395.

Fraser, H.B., Hirsh, A.E., Steinmetz, L.M., Scharfe, C., Feldman, M.W., 2002. Evolutionary rate in the protein interaction network. Science 296, 750–752.

Geiler-Samerotte, K.A., Dion, M.F., Budnik, B.A., Wang, S.M., Hartl, D.L., Drummond, D.A., 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.

Ghaemmaghami, S., Huh, W.-K., Bower, K., Howson, R.W., Belle, A., Dephoure, N., ÓShea, E.K., Weissman, J.S., 2003. Global analysis of protein expression in yeast. Nature 425, 737–741.

Go, M., Miyazawa, S., 1980. Relationship between mutability, polarity and exteriority of amino acid residues in protein evolution. Int. J. Pept. Protein Res. 15, 211–224.

Hirsh, A.E., Fraser, H.B., 2001. Protein dispensability and rate of evolution. Nature 411, 1047–1049.

Hirsh, A.E., Fraser, H.B., 2003. Genomic function (communication arising): rate of evolution and gene dispensability. Nature 421, 497–498.

Jordan, I.K., Rogozin, I.B., Wolf, Y.I., Koonin, E.V., 2002. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. Genome Res. 12, 962–968.

Kimura, M., 1968. Evolutionary rate at the molecular level. Nature 217, 624–626.

Kimura, M., 1969. The rate of molecular evolution considered from the standpoint of population genetics. Proc. Natl. Acad. Sci. USA 63, 1181–1188.

Kimura, M., Ohta, T., 1971. Protein polymorphism as a phase of molecular evolution. Nature 229, 467–469.

Kimura, M., Ohta, T., 1974. On some principles governing molecular evolution. Proc. Natl. Acad. Sci. USA 71, 2848–2852.

Kuma, K., Iwabe, N., Miyata, T., 1995. Functional constraints against variations on molecules from the tissue level: slowly evolving brain-specific genes demonstrated by protein kinase and immunoglobulin supergene families. Mol. Biol. Evol. 12, 123–130.

Kumar, M., Bava, K., Gromiha, M., Prabakaran, P., Kitajima, K., Uedaira, H., Sarai, A., 2006. ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. Nucl. Acid Res. 34, D204–D206.

Lynch, M., Conery, J.S., 2003. The origins of genome complexity. Science 302, 1401–1404.

Miyata, T., Yasunaga, T., 1980. Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its applications. J. Mol. Evol. 16, 23–36.

Miyazawa, S., Jernigan, R.L., 1982a. Equilibrium folding and unfolding pathways for a model protein. Biopolymers 21, 1333–1363.

Miyazawa, S., Jernigan, R.L., 1982b. Most probable intermediates in protein folding-unfolding with a non-interacting globule-coil model. Biochemistry 21, 5203–5213.

Ogasahara, K., Nakamura, M., Nakura, S., Tsunasawa, S., Kato, I., Yoshimoto, T., Yutani, K., 1998. The unusually slow unfolding rate causes the high stability of pyrrolidone carboxyl peptidase from a hyperthermophile, Pyrococcus furiosus: equilibrium and kinetic studies of guanidine hydrochloride-induced unfolding and refolding. Biochemistry 37, 17537–17544.

Ohta, T., 1973. Slightly deleterious mutant substitutions in evolution. Nature 246, 96–98.

Ohta, T., 1992. The nearly neutral theory of molecular evolution. Annu. Rev. Ecol. Syst. 23, 263–286.

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

Pál, C., Papp, B., Hurst, L.D., 2003. Genomic function (communication arising): rate of evolution and gene dispensability. Nature 421, 496–497.

Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., Serrano, L., 2005. The FoldX web server: an online force field. Nucl. Acid Res. 33, W382–W388.

Serohijos, A., Rimas, Z., Shakhnovich, E., 2012. Protein biophysics explains why highly abundant proteins evolve slowly. Cell Rep. 2 (2), 249–256 ⟨http://dx.doi.org/10.1016/j.celrep.2012.06.022⟩.

Serohijos, A., Ryan Lee, S.Y., Shakhnovich, E., 2013. Highly abundant proteins favor more stable 3D structures in yeast. Biophys. J. 104 (3), L1–L3. http://dx.doi.org/10.1016/j.bpj.2012.11.3838.

Serohijos, A.W., Shakhnovich, E.I., 2014. Contribution of selection for protein folding stability in shaping the patterns of polymorphisms in coding regions. Mol. Biol. Evol. 31, 165–176.

Spielman, S.J., Wilke, C.O., 2015. The relationship between dN/dS and scaled selection coefficients. Mol. Biol. Evol. 32, 1097–1108.

Stoebel, D.M., Dean, A.M., Dykhuizen, D.E., 2008. The cost of expression of Escherichia coli lac operon proteins ss in the process, not in the product. Genetics 178, 1653–1660.

Tokuriki, N., Stricher, F., Schymkowitz, J., Serrano, L., Tawfik, D.S., 2007. The stability effects of protein mutations appear to be universally distributed. J. Mol. Biol. 369, 1318–1332.

Wall, DP., Hirsh, AE., Fraser, HB., Kumm, J., Giaever, G., Eisen, MB., Feldman, MW., 2005. Functional genomic analysis of the rates of protein evolution. Proc. Natl. Acad. Sci. USA 102, 5483–5488.

Wittung-Stafshede, P., Malmstrom, B.G., Sanders, D., Fee, J.A., Winkler, J.R., Gray, H. B., 1998. Effect of redox state on the folding free energy of a thermostable electron-transfer metalloprotein: the CuA domain of cytochrome oxidase from thermus thermophilus. Biochemistry 37, 3172–3177.

Yang, JR., Liao, BY., Zhuang, SM., Zhang, J., 2012. Protein misinteraction avoidance causes highly expressed proteins to evolve slowly. Proc. Natl. Acad. Sci. USA 109, E831–E840.

Yin, S., Ding, F., Dokholyan, N.V., 2007. Eris: an automated estimator of protein stability. Nat. Methods 4, 466–467.

Zeldovich, K.B., Chen, P., Shakhnovich, E.I., 2007. Protein stability imposes limits on organism complexity and speed of molecular evolution. Proc. Natl. Acad. Sci. USA 104, 16152–16157.

Zhang, J., He, X., 2005. Significant impact of protein dispensability on the instantaneous rate of protein evolution. Mol. Biol. Evol. 22, 1147–1155.

Zuckerkandl, E., 1976. Evolutionary processes and evolutionary noise at the molecular level. J. Mol. Evol. 7, 167–183.