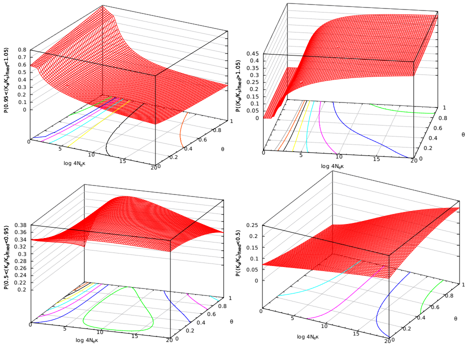
Selection maintaining protein stability at equilibrium



Sanzo Miyazawa
sanzo.miyazawa@gmail.com

December 12, 2015

Background

The common understanding of protein evolution:

- Most amino acid substitutions observed in homologous proteins were selectively neutral and fixed by random drift.
- A proportion of neutral mutations that depends on the strength of structural and functional constraints primarily determines evolutionary rate.

Recently a question has been raised on the common view of protein evolution.

- There are a diversity of protein evolutionary rates among genes.
- Protein evolutionary rate is correlated with gene expression level; highly expressed genes evolve slowly.
- 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 stability.

A generic form of fitness costs due to protein misfolding

- Malthusian fitness for protein dispensability (Drummond et al., 2008):

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

- Malthusian fitness for toxicity of misfolded proteins (Drummond et al., 2008):

$$m_{\text{misfolds}} = -c \sum_i A_i \frac{1 - f_i^{\text{native}}}{f_i^{\text{native}}} \tag{2}$$

- Selection to maintain protein stability (Dasmeh et al., 2014):

$$m = \log f^{\text{native}} \tag{3}$$

The proportion of native conformations, f^{native}, in a two state transition:

$$f^{\text{native}} = \frac{e^{-\beta\Delta G}}{1 + e^{-\beta\Delta G}} \tag{4}$$

where ΔG is the folding free energy of protein.

Because $\exp \beta\Delta G \ll 1$ for typical proteins, all these formulas of Malthusian fitness for misfolded proteins are reduced to

$$m \equiv -\sum_i \kappa_i e^{\beta\Delta G_i} \quad \text{with } \kappa_i \geq 0 \tag{5}$$

The evolution of a single coding gene in a monoclonal approximation

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

$$\begin{aligned} s &\equiv m^{\text{mutant}} - m^{\text{wildtype}} \qquad (6) \\ 4N_e\, s &= 4N_e\, \kappa\, e^{\beta\Delta G}(1 - e^{\beta\Delta\Delta G}) \quad \text{with } \kappa \geq 0 \qquad (7) \end{aligned}$$

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

$$\kappa = cA + \gamma \tag{8}$$

c	$\sim 10^{-4}$	fitness cost per misfolded protein
A	$10 < A < 10^6$	cellular abundance of protein
γ	$0 \leq \gamma \leq 10$	protein indispensability
N_e		effective population size
	$\sim 10^4$ to 10^5	for vertebrates
	$\sim 10^5$ to 10^6	for invertebrates
	$\sim 10^7$ to 10^8	for unicellular eukaryotes
	$> 10^8$	for prokaryotes

Stability changes, $\Delta\Delta G$, due to single amino acid substitutions

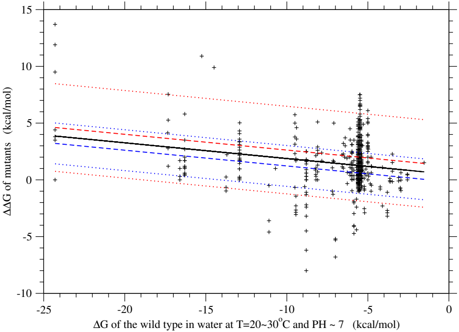
PDF approximated with a weighted sum of two Gaussian functions (Tokuriki et al., 2007):

$$p(\Delta\Delta G) \quad = \quad \theta\mathcal{N}(\mu_s, \sigma_s) + (1-\theta)\mathcal{N}(\mu_c, \sigma_c) \tag{9}$$

$$\text{For surface residues :} \quad \mu_s \quad = \quad -0.14\,\Delta G - 0.17 \quad , \quad \sigma_s = 0.90 \tag{10}$$

$$\text{For core residues :} \quad \mu_c \quad = \quad -0.14\,\Delta G + 1.23 \quad , \quad \sigma_c = 1.93 \tag{11}$$

The dependences of the means, μ_c and μ_s, on ΔG are estimated from the regression line of observed values of $\Delta\Delta G$ of mutant proteins on ΔG of the wild-type protein.



Solid: regression; Broken: μ_c, Dotted: $\mu_c \pm \sigma_c$; Broken: μ_s, Dotted: $\mu_s \pm \sigma_s$

PDFs of K_a/K_s in all mutants and in fixed mutants

Instead of pursueing computer simulations of gene populations, we calculate the probability density functions (PDF) of characteristic quantities such as selective advantage, fixation probabilty, and K_a/K_s, and examine the protein evolution of the gene in a monoclonal approximation.

Fixation probability:

$$u(4N_e s) = \frac{1 - e^{-4N_e sq}}{1 - e^{-4N_e s}} \tag{12}$$

where $q = 1/(2N)$ for a mutant gene, and the population size is taken to be $N = 10^6$.

The ratio of nonsynonymous (K_a) to synonymous substitution rate per site (K_s):

$$\frac{K_a}{K_s} = \frac{u(4N_e s)}{u(0)} = \frac{u(4N_e s)}{q} \quad \text{with } q = \frac{1}{2N} \tag{13}$$

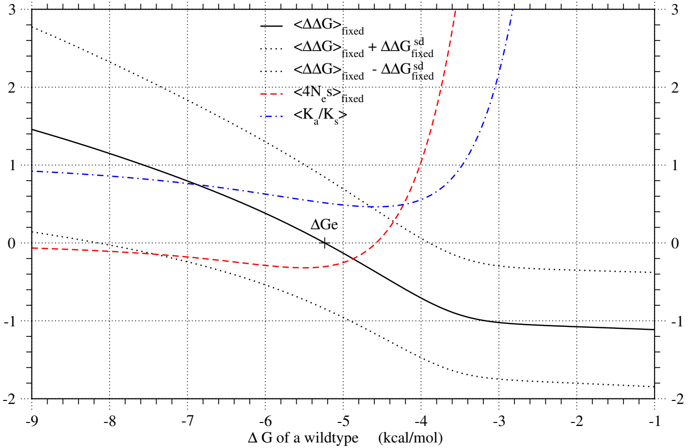
PDF of $\Delta\Delta G$ in fixed mutants:

$$p(\Delta\Delta G_{\text{fixed}}) \equiv p(\Delta\Delta G)\frac{u(4N_e s)}{\langle u \rangle} \tag{14}$$

$$\langle u \rangle \equiv \int_{-\infty}^{\infty} u(4N_e s)p(\Delta\Delta G)d\Delta\Delta G \tag{15}$$

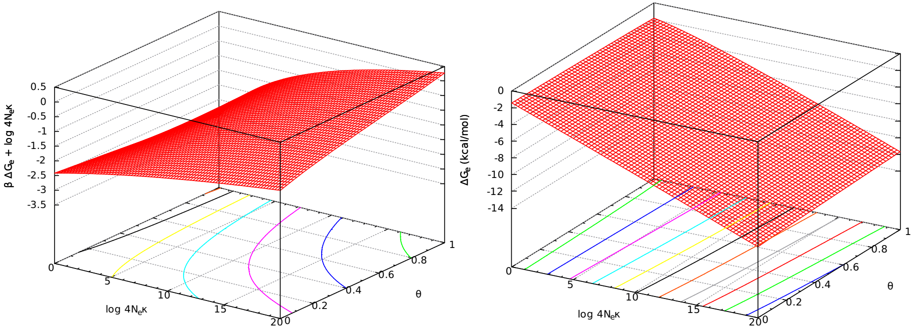
Equilibrium stability, ΔG_e

The average, $\langle\Delta\Delta G\rangle_{\text{fixed}}$, of stability changes over fixed mutants versus protein stability, ΔG, of the wild type.



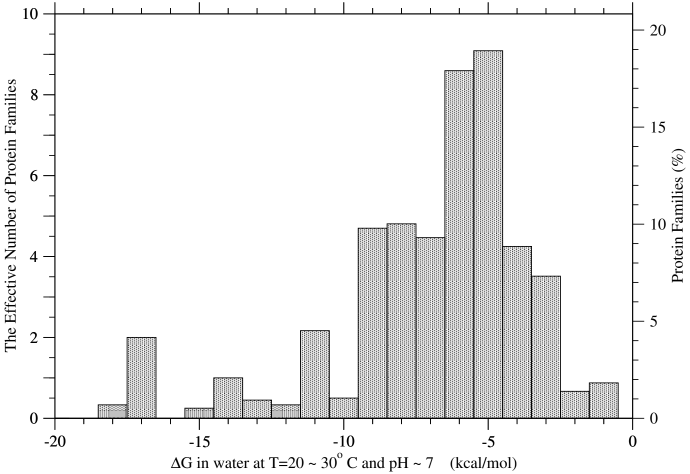
ΔG_e is the stable equilibrium point for ΔG, where $\langle\Delta\Delta G\rangle_{\text{fixed}} = 0$.

Dependence of equilibrium stability, ΔG_e, on parameters, $4N_e\kappa$ and θ.



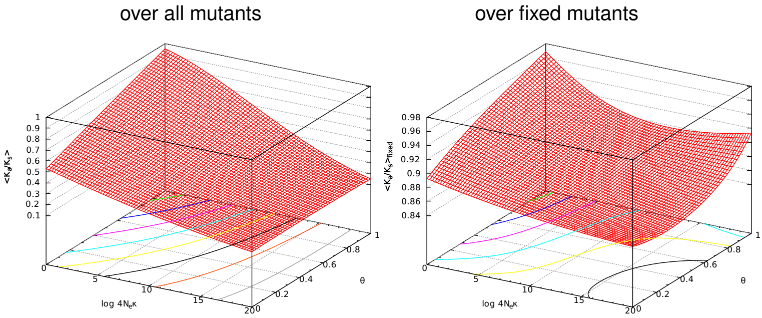
- The value of $\beta\Delta G_e + \log 4N_e\kappa$ is the upper bound of $\log 4N_e s$, and would be constant if the mean of $\Delta\Delta G$ in all arising mutants did not depend on ΔG.
- ΔG_e decreases as $\log 4N_e\kappa$, effective population size or protein abundance/indispensability, increases.

Distribution of folding free energies of monomeric protein families



- The observed range of ΔG shown above is consistent with that range, -2 to -12.5 kcal/mol, expected from the present model.

The average of K_a/K_s at equilibrium of protein stability, $\Delta G = \Delta G_e$



over all mutants

over fixed mutants

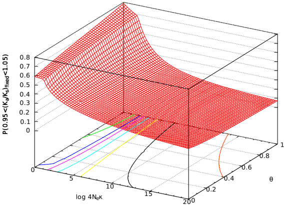
- Protein abundance/indispensability and effective population size, $4N_e\kappa$, more decrease evolutionary rate for less-constrained proteins.
- Structural constraint, $1-\theta$, more decreases evolutionary rate for less-abundant, less-essential proteins.
- $\langle K_a/K_s \rangle < 1$ over a whole range of the parameters.

Probability of each selection category

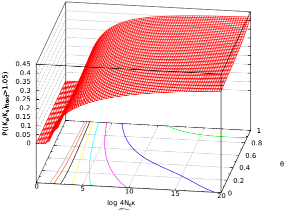
in fixed mutants at equilibrium of protein stability, $\Delta G = \Delta G_e$.

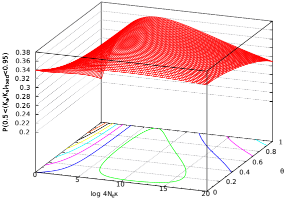
- Nearly neutral selection is predominant only for low-abundant, non-essential proteins.
- Positive selection is significant for the other proteins.

nearly neutral selection, $P(0.95 < (K_a/K_s)_{fixed} < 1.05)$

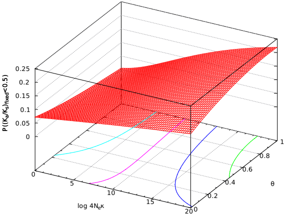


positive selection, $P(1.05 < (K_a/K_s)_{fixed})$





slightly negative selection, $P(0.5 < (K_a/K_s)_{fixed} < 0.95)$



negative selection, $P((K_a/K_s)_{fixed} < 0.5)$

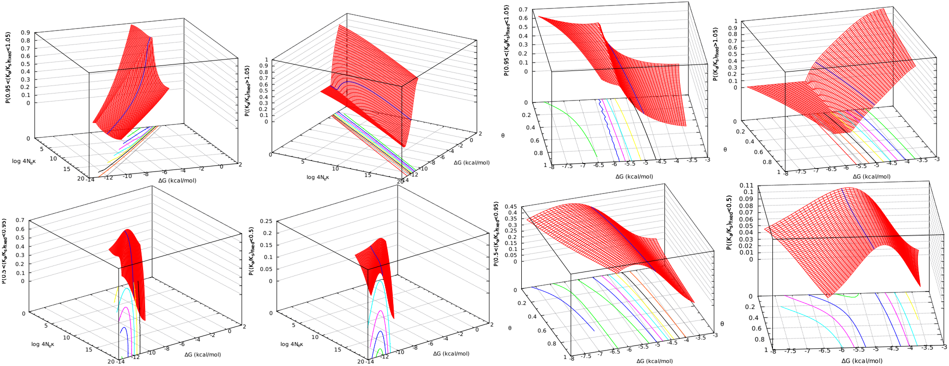
- Slightly negative selection is always significant.

Dependence of probability of each selection on ΔG and $4N_e\kappa$ or θ

in fixed mutants; shown within $2 \cdot \Delta\Delta G^{\text{sd}}_{\text{fixed}}$ around $\Delta G = \Delta G_e$ indicated by a blue line.

- Positive selection is predominant in $\Delta G > \Delta G_e$.
- Nearly neutral and slightly negative selections are predominant in $\Delta G < \Delta G_e$.

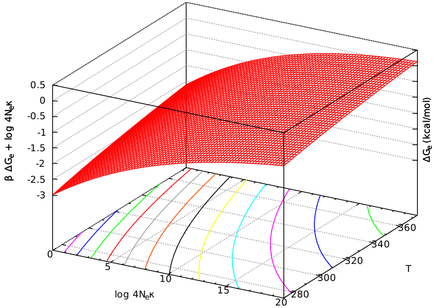
nearly neutral positive selection nearly neutral positive selection

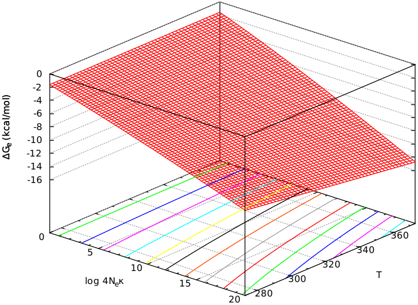


slightly negative negative selection slightly negative negative selection

Dependence of equilibrium stability, ΔG_e, on growth temperature T

- Protein stability ($-\Delta G_e/kT$) is predicted to decrease as growth temperature increases.





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.
- 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 mutants. 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 $(1-\theta)$ 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.
- The present model indicates that protein stability $(-\beta\Delta G_e)$ and $\langle K_a/K_s \rangle$ decrease as growth temperature increases.