Selection originating from protein stability/foldability: Relationships between protein folding free energy, sequence ensemble, and fitness

Sanzo Miyazawa
sanzo.miyazawa@gmail.com

at Bioinformatics and System Biology, KMUTT, on March 23, 2018

Publication: J. Theor. Biol., 433, 21-38, 2017.

1. Background

- Natural selection maintains protein's stability and foldability over evolutionary timescales. A protein folding theory based on the random energy model (REM) indicates that the equilibrium ensemble of natural protein sequences is well represented by a canonical ensemble characterized by $\exp(-\Delta G_{ND}/k_B T_s)$ or by $\exp(-G_N/k_B T_s)$ if an amino acid composition is kept constant, where $\Delta G_{ND} \equiv G_N - G_D$, G_N and G_D are the native and denatured free energies, and T_s is the effective temperature representing the strength of selection pressure (Shakhnovich et al., 1993).
- It has become clear that the distribution of homologous sequences ($\boldsymbol{\sigma}$) in a protein family can be well approximated by a Boltzmann distribution with $\exp(-\psi_N)$, where the evolutionary statistical energy $\psi_N(\boldsymbol{\sigma}) \equiv -(\sum_i(h_i(\sigma_i) + \sum_{j>i} J_{ij}(\sigma_i, \sigma_j)))$ is represented as the sum of one body (h) (compositional) and pairwise (J) (covariational) interactions over all sites and site pairs (Figliuzzi et al., 2018).
- In population biology, mutation and fixation processes of amino acids in protein evolution are described in terms of fitness (Crow and Kimura, 1970).

These aspects about the distribution of homologous sequences should be unified.

2. Purposes of the present study

A purpose of the present study is to establish relationships between protein foldability/stability, sequence distribution, and protein fitness.

1. We prove that if a mutational process in protein evolution is a reversible Markov process, the equilibrium ensemble of genes will obey a Boltzmann distribution with $\exp(4N_e m(1-1/(2N)))$, where N_e and N are effective and actual population sizes, and m is the Malthusian fitness of a gene. Relationships between $\Delta\psi_{ND}$, ΔG_{ND}, and m are obtained.
2. From the distribution of the change of ψ_N, $\Delta\psi_N$, which results from single amino acid substitutions, we estimate the effective temperature of natural selection (T_s) and then glass transition temperature (T_g) and folding free energy (ΔG_{ND}) of protein on the basis of the REM.
3. Through analyzing the amino acid substitution process in protein evolution, which is characterized by the fitness, $m = -\Delta\psi_{ND}/(4N_e(1-1/(2N)))$, we clarify the relationship between T_s and the amino acid substitution rate, and evaluate the contribution of neutral substitutions under the protein foldability/stability selection.

3-1. Knowledge of protein folding (Shakhnovich et al., 1993; Pande et al., 1997)

A protein folding theory based on a random energy model (REM) indicates:
The probability density of homologous sequence ($\boldsymbol{\sigma}$),

$$P(\boldsymbol{\sigma}) \propto P^{\mathrm{mut}}(\boldsymbol{\sigma}) \exp\left(\frac{-\Delta G_{ND}(\boldsymbol{\sigma}, T)}{k_B T_s}\right) \tag{1}$$

$$\propto \exp\left(\frac{-G_N(\boldsymbol{\sigma})}{k_B T_s}\right) \qquad \text{if } \boldsymbol{f}(\boldsymbol{\sigma}) = \text{constant} \tag{2}$$

$$\Delta G_{ND}(\boldsymbol{\sigma}, T) \equiv G_N(\boldsymbol{\sigma}) - G_D(\boldsymbol{f}(\boldsymbol{\sigma}), T) \tag{3}$$

where

$p^{\mathrm{mut}}(\boldsymbol{\sigma})$ the probability of a sequence ($\boldsymbol{\sigma}$) randomly occurring in a mutational process and depends only on the amino acid frequencies $\boldsymbol{f}(\boldsymbol{\sigma})$

G_N and G_D the free energies of the native conformation and denatured state. The distribution of conformational energies in the denatured state is approximated to be equal to the energy distribution of randomized sequences in the native fold, which is then approximated by a Gaussian distribution.

T and T_s growth temperature and selective temperature representing the strength of selection

3-2. Probability distribution of homologous sequences in sequence space

The probability distribution $P(\boldsymbol{\sigma})$, with maximum entropy, of protein sequences $\boldsymbol{\sigma}(\equiv (\sigma_1, \cdots, \sigma_L))$, which satisfies

$$\begin{aligned} \sum_{\boldsymbol{\sigma}} P(\boldsymbol{\sigma})\, \delta_{\sigma_i a_k} &= P_i(a_k) && (4) \\ \sum_{\boldsymbol{\sigma}} P(\boldsymbol{\sigma})\, \delta_{\sigma_i a_k} \delta_{\sigma_j a_l} &= P_{ij}(a_k, a_l) && (5) \end{aligned}$$

where $\sigma_i, a_k \in$ {amino acids, deletion}, can be represented as

$$\begin{aligned} P(\boldsymbol{\sigma}) &\propto \exp(-\psi_N(\boldsymbol{\sigma})) && (6) \\ \psi_N(\boldsymbol{\sigma}) &\equiv -(\sum_{i}^{L}(h_i(\sigma_i) + \sum_{j>i} J_{ij}(\sigma_i, \sigma_j))) && (7) \end{aligned}$$

where h_i and J_{ij} are one-body (compositional) and two-body (covariational) interactions. Interactions h_i and J_{ij} for homologous sequences **can be estimated** from a multiple sequence alignment (MSA) in the mean field approximation, or by maximizing a pseudo-likelihood, or by MCMC simulation for a Boltzmann machine, and have been shown to well describe the distributions of homologous sequences.

3-3. The equilibrium distribution of sequences in a mutation-fixation process

Assumption: The mutational process is a reversible Markov process; the mutation rate per gene, $M_{\boldsymbol{\mu\nu}}$, from sequence $\boldsymbol{\mu} \equiv (\mu_1, \cdots, \mu_L)$ to $\boldsymbol{\nu}$ satisfies the detailed balance condition,

$$P^{\text{mut}}(\boldsymbol{\mu}) M_{\boldsymbol{\mu\nu}} = P^{\text{mut}}(\boldsymbol{\nu}) M_{\boldsymbol{\nu\mu}} \tag{8}$$

where $P^{\text{mut}}(\boldsymbol{\nu})$ is the equilibrium frequency of sequence $\boldsymbol{\nu}$ in the mutational process, $M_{\boldsymbol{\mu\nu}}$.

The substitution rate $R_{\boldsymbol{\mu\nu}}$ from $\boldsymbol{\mu}$ to $\boldsymbol{\nu}$ for diploid:

$$R_{\boldsymbol{\mu\nu}} = 2N M_{\boldsymbol{\mu\nu}} u(s(\boldsymbol{\mu} \to \boldsymbol{\nu})) \tag{9}$$

$$2Nu(s) = 2N \frac{1 - e^{-4N_e s q_m}}{1 - e^{-4N_e s}} = \frac{u(s)}{u(0)} \qquad \text{with} \quad q_m = \frac{1}{2N} \tag{10}$$

$$s(\boldsymbol{\mu} \to \boldsymbol{\nu}) \equiv m(\boldsymbol{\nu}) - m(\boldsymbol{\mu}) \tag{11}$$

where N and N_e are population size and effective population size, and $u(s(\boldsymbol{\mu} \to \boldsymbol{\nu}))$ is the fixation probability of mutants from $\boldsymbol{\mu}$ to $\boldsymbol{\nu}$ the selective advantage of which is equal to s [Crow and Kimura(1970)]. $m(\boldsymbol{\nu})$ is the Malthusian fitness of a mutant sequence ($\boldsymbol{\nu}$).

The equilibrium distribution of sequences in a mutation-fixation process

This Markov process of substitutions, which consists of mutation and fixation processes, in sequence is reversible, and the equilibrium frequency of sequence $\boldsymbol{\mu}$, $P^{\text{eq}}(\boldsymbol{\mu})$, is represented by

$$P^{\text{eq}}(\boldsymbol{\mu}) = \frac{P^{\text{mut}}(\boldsymbol{\mu}) \exp(4N_e m(\boldsymbol{\mu})(1 - q_m))}{\sum_{\boldsymbol{\nu}} P^{\text{mut}}(\boldsymbol{\nu}) \exp(4N_e m(\boldsymbol{\nu})(1 - q_m))} \tag{12}$$

because both the mutation and fixation processes satisfy the detailed balance conditions, Eq. 8 and the following equation, respectively.

$$\begin{aligned} & \exp(4N_e m(\boldsymbol{\mu})(1 - q_m))\, u(s(\boldsymbol{\mu} \to \boldsymbol{\nu})) \\ = \quad & \frac{\exp(-4N_e m(\boldsymbol{\mu}) q_m) - \exp(-4N_e m(\boldsymbol{\nu}) q_m)}{\exp(-4N_e m(\boldsymbol{\mu})) - \exp(-4N_e m(\boldsymbol{\nu}))} \qquad (13) \\ = \quad & \exp(4N_e m(\boldsymbol{\nu})(1 - q_m))\, u(s(\boldsymbol{\nu} \to \boldsymbol{\mu})) \qquad (14) \end{aligned}$$

As a result, the ensemble of homologous sequences in molecular evolution obeys a Boltzmann distribution.

3-4. Relationships between $m(\sigma)$, $\Delta\psi_{ND}(\sigma, T)$, and $\Delta G_{ND}(\sigma, T)$ of protein sequence

From Eqs. 1, 6, and 12 , we can get the following relationships among the Malthusian fitness m, the folding free energy change ΔG_{ND} and $\Delta\psi_{ND}$ of protein sequence.

$$\begin{aligned}
P^{\mathrm{eq}}(\boldsymbol{\mu}) &= \frac{P^{\mathrm{mut}}(\boldsymbol{\mu}) \exp(4N_e m(\boldsymbol{\mu})(1-q_m))}{\sum_{\nu} P^{\mathrm{mut}}(\boldsymbol{\nu}) \exp(4N_e m(\boldsymbol{\nu})(1-q_m)))} && (15) \\
&= \frac{P^{\mathrm{mut}}(\overline{\boldsymbol{\mu}}) \exp(-(\psi_N(\boldsymbol{\mu}) - \psi_D(\overline{\mathbf{f}(\boldsymbol{\mu})}, T)))}{\sum_{\boldsymbol{\nu}} P^{\mathrm{mut}}(\overline{\boldsymbol{\nu}}) \exp(-(\psi_N(\boldsymbol{\nu}) - \psi_D(\overline{\mathbf{f}(\boldsymbol{\nu})}, T)))} && (16) \\
&\simeq \frac{P^{\mathrm{mut}}(\boldsymbol{\mu}) \exp(-\Delta G_{ND}(\boldsymbol{\mu}, T)/(k_B T_s))}{\sum_{\nu} P^{\mathrm{mut}}(\boldsymbol{\nu}) \exp(-\Delta G_{ND}(\boldsymbol{\nu}, T)/(k_B T_s))} && (17)
\end{aligned}$$

where $\overline{\mathbf{f}(\boldsymbol{\sigma})} \equiv \sum_{\boldsymbol{\sigma}} \mathbf{f}(\boldsymbol{\sigma}) P(\boldsymbol{\sigma})$ and $\log P^{\mathrm{mut}}(\overline{\boldsymbol{\sigma}}) \equiv \sum_{\boldsymbol{\sigma}} P(\boldsymbol{\sigma}) \log(\prod_i P^{\mathrm{mut}}(\sigma_i))$. Then, the following relationships are derived for sequences for which $f(\boldsymbol{\mu}) = \overline{f(\boldsymbol{\mu})}$.

$$\begin{aligned}
4N_e m(\boldsymbol{\mu})(1-q_m) &= -\Delta\psi_{ND}(\boldsymbol{\mu}, T) + \text{constant} && (18) \\
&\simeq \frac{-\Delta G_{ND}(\boldsymbol{\mu}, T)}{k_B T_s} + \text{constant} && (19)
\end{aligned}$$

Relationships between $s(\mu \to \nu)$, $\Delta\psi_N(\mu \to \nu)$, and $\Delta G_N(\mu \to \nu)$ of protein sequence

The selective advantage of ν to μ is represented as follows for $f(\mu) = f(\nu) = \overline{f(\sigma)}$.

$$
\begin{aligned}
4N_e s(\mu \to \nu)(1 - q_m) &\equiv (4N_e m(\nu) - 4N_e m(\mu))(1 - q_m) && (20)\\
= \quad & -(\Delta\psi_{ND}(\nu, T) - \Delta\psi_{ND}(\mu, T)) = -(\psi_N(\nu) - \psi_N(\mu)) && (21)\\
\simeq \quad & -(\Delta G_{ND}(\nu, T) - \Delta G_{ND}(\mu, T))/(k_B T_s) = -(G_N(\nu) - G_N(\mu))/(k_B T_s) && (22)
\end{aligned}
$$

It should be noted here that only sequences for which $f(\sigma) = \overline{f(\sigma)}$ contribute significantly to the partition functions in Eq. 16, and other sequences may be ignored.

Eqs. 21 and 22 indicate:

$$
\begin{aligned}
\psi_N &\propto N_e && (23)\\
T_s &\propto 1/N_e, && (24)
\end{aligned}
$$

If $\Delta\Delta G_{ND} \simeq \Delta G_N$ due to single substitutions are known, we can estimate T_s from Eqs. 21 and 22. However, experiments/calculations to estimate the folding energy changes/native free energy changes are not easy. Here **we propose an alternative way to estimate T_s.**

3-5. Denatured state

The distribution of conformational energies in the denatured state (molten globule state), which consists of conformations as compact as the native conformation, is approximated in the random energy model (REM), particularly the independent interaction model (IIM) (Pande et al., 1997) to be equal to the energy distribution of randomized sequences, which is then approximated by a Gaussian distribution, in the native conformation.

$$Z = \int \exp\left(\frac{-E}{k_B T}\right) n(E) dE \quad \text{where} \quad n(E) \approx \exp(\omega L) \mathcal{N}(\bar{E}(\mathbf{f}(\boldsymbol{\sigma})), \delta E^2(\mathbf{f}(\boldsymbol{\sigma}))) \tag{25}$$

where ω is the conformational entropy per residue in the compact denatured state, and $\mathcal{N}(\bar{E}(\mathbf{f}(\boldsymbol{\sigma})), \delta E^2(\mathbf{f}(\boldsymbol{\sigma})))$ is the Gaussian probability density with mean $\bar{E}$ and variance δE^2, which depend only on the amino acid composition of the protein sequence. and are estimated as the mean and variance of interaction energies of randomized sequences in the native conformation.

Free energy of the denatured state

The free energy of the denatured state is approximated as follows.

$$\begin{aligned}
G_D(\mathbf{f}(\boldsymbol{\sigma}), T) &\approx \bar{E}(\mathbf{f}(\boldsymbol{\sigma})) - \frac{\delta E^2(\mathbf{f}(\boldsymbol{\sigma}))}{2k_B T} - k_B T \omega L && (26)\\
&= \bar{E}(\mathbf{f}(\boldsymbol{\sigma})) - \delta E^2(\mathbf{f}(\boldsymbol{\sigma}))\frac{\vartheta(T/T_g)}{k_B T} && (27)\\
\vartheta(T/T_g) &\equiv \begin{cases} \frac{1}{2}(1 + \frac{T^2}{T_g^2}) & \text{for } T > T_g \\ \frac{T}{T_g} & \text{for } T \leq T_g \end{cases} && (28)
\end{aligned}$$

where T_g is the glass transition temperature of the protein at which entropy becomes zero (Shakhnovich and Gutin, 1993); $-\partial G_D/\partial T|_{T=T_g} = 0$. The conformational entropy per residue ω in the compact denatured state can be represented with T_g; $\omega L = \delta E^2/(2(k_B T_g)^2)$. Thus, unless $T_g < T_m$, a protein will be trapped at local minima on a rugged free energy landscape before it can fold into a unique native structure.

3-6. The ensemble average of ψ_N over sequences

The ensemble average of $\Delta G_{ND}(\boldsymbol{\sigma}, T)$ over sequences with Eq. 1 is

$$\begin{aligned}
& \langle \Delta G_{ND}(\boldsymbol{\sigma}, T) \rangle_{\boldsymbol{\sigma}} && (29) \\
\equiv \quad & [\sum_{\boldsymbol{\sigma}} \Delta G_{ND}(\boldsymbol{\sigma}, T) P^{\text{mut}}(\boldsymbol{\sigma}) \exp(-\frac{\Delta G_{ND}(\boldsymbol{\sigma}, T)}{k_B T_s})] \,/\, [\sum_{\boldsymbol{\sigma}} P^{\text{mut}}(\boldsymbol{\sigma}) \exp(-\frac{\Delta G_{ND}(\boldsymbol{\sigma}, T)}{k_B T_s})] && (30) \\
\approx \quad & \langle G_N(\boldsymbol{\sigma}) \rangle_{\boldsymbol{\sigma}} - G_D(\overline{\boldsymbol{f}(\boldsymbol{\sigma}_N)}, T) && (31)
\end{aligned}$$

where $\boldsymbol{\sigma}_N$ denotes a natural sequence, and $\overline{\boldsymbol{f}(\boldsymbol{\sigma_N})}$ denotes the average of amino acid frequencies $\boldsymbol{f}(\boldsymbol{\sigma_N})$ over homologous sequences.

The ensemble averages of G_N and $\psi_N(\boldsymbol{\sigma})$ are estimated in the Gaussian approximation (Pande et al. 1997).

$$\begin{aligned}
\langle G_N(\boldsymbol{\sigma}) \rangle_{\boldsymbol{\sigma}} &\approx \frac{\int E \exp(-E/(k_B T_s))\, n(E)\, dE}{\int \exp(-E/(k_B T_s))\, n(E)\, dE} && (32) \\
&= \bar{E}(\overline{\boldsymbol{f}(\boldsymbol{\sigma_N})}) - \delta E^2(\overline{\boldsymbol{f}(\boldsymbol{\sigma_N})})/(k_B T_s) && (33)
\end{aligned}$$

$$\langle \psi_N(\boldsymbol{\sigma}) \rangle_{\boldsymbol{\sigma}} \approx \bar{\psi}(\overline{\boldsymbol{f}(\boldsymbol{\sigma_N})}) - \delta\psi^2(\overline{\boldsymbol{f}(\boldsymbol{\sigma_N})}) \qquad (34)$$

The sample average of ψ_N over homologous sequences

The ensemble averages of $\Delta G_{ND}(\boldsymbol{\sigma}, T)$ and $\psi_N(\boldsymbol{\sigma})$ over sequences are observable as the sample averages of $\Delta G_{ND}(\boldsymbol{\sigma_N}, T)$ and $\psi_N(\boldsymbol{\sigma_N})$ over homologous sequences fixed in protein evolution, respectively.

$$\begin{aligned}
\overline{\Delta G_{ND}(\boldsymbol{\sigma_N}, T)}/(k_B T_s) &= \langle \Delta G_{ND}(\boldsymbol{\sigma}, T)\rangle_{\boldsymbol{\sigma}}/(k_B T_s) && (35)\\
&\approx \delta\psi^2(\overline{\mathbf{f}(\boldsymbol{\sigma_N})})\,[\,\vartheta(T/T_g)T_s/T - 1\,] && (36)\\
\overline{\psi_N(\boldsymbol{\sigma_N})} &\equiv \frac{\sum_{\boldsymbol{\sigma}_N} w_{\boldsymbol{\sigma}_N}\psi_N(\boldsymbol{\sigma}_N)}{\sum_{\boldsymbol{\sigma}_N} w_{\boldsymbol{\sigma}_N}} && (37)\\
&= \langle \psi_N(\boldsymbol{\sigma})\rangle_{\boldsymbol{\sigma}} && (38)
\end{aligned}$$

where the overline denotes a sample average with a sample weight $w_{\boldsymbol{\sigma}_N}$ for each homologous sequence, which is used to reduce phylogenetic biases in the set of homologous sequences.

The folding free energy becomes equal to zero at the melting temperature T_m; $\langle \Delta G_{ND}(\boldsymbol{\sigma_N}, T_m)\rangle_{\boldsymbol{\sigma}} = 0$.

$$\vartheta(T_m/T_g)\frac{T_s}{T_m} = \frac{T_s}{2T_m}(1 + \frac{T_m^2}{T_g^2}) = 1 \quad \text{with } T_s \leq T_g \leq T_m \qquad (39)$$

4. Results

4-1. Changes of the evolutionary statistical energy, $\Delta\psi_N$, by single nucleotide nonsynonymous substitutions

Fields h_i and couplings J_{ij} were estimated from a MSA of each protein family in the mean field approximation with the DCA program (Marks et. al. 2011).

We calculated the ψ_N of the wildtype and $\Delta\psi_N$ due to all types of single nucleotide nonsynonymous substitutions for all homologous sequences, and their means and variances.

The changes of the evolutionary statistical energy, $\Delta\psi_N$ and $\Delta\psi_D$, due to a single amino acid substitution from σ_i^N to σ_i at site i in a natural sequence $\boldsymbol{\sigma}_N$ are defined as

$$\begin{aligned}
\Delta\psi_N(\sigma^N_{j\neq i}, \sigma^N_i \rightarrow \sigma_i) &\equiv \psi_N(\sigma^N_{j\neq i}, \sigma_i) - \psi_N(\boldsymbol{\sigma}_N) && (40)\\
\Delta\psi_D(\sigma^N_{j\neq i}, \sigma^N_i \rightarrow \sigma_i, T) &\equiv \psi_D(\mathbf{f}(\sigma^N_{j\neq i}, \sigma_i), T) - \psi_D(\mathbf{f}(\boldsymbol{\sigma}_N), T) && (41)\\
\Delta\Delta\psi_{ND}(\sigma^N_{j\neq i}, \sigma^N_i \rightarrow \sigma_i, T) &\equiv \Delta\psi_N(\sigma^N_{j\neq i}, \sigma^N_i \rightarrow \sigma_i) - \Delta\psi_D(\sigma^N_{j\neq i}, \sigma^N_i \rightarrow \sigma_i, T) && (42)\\
&\simeq \Delta\psi_N(\sigma^N_{j\neq i}, \sigma^N_i \rightarrow \sigma_i) \quad \text{because } \mathbf{f}(\boldsymbol{\sigma}_N) \approx \mathbf{f}(\sigma^N_{j\neq i}, \sigma_i) && (43)
\end{aligned}$$

Protein families, and structures studied.

Pfam family	UniProt ID	N [a]	N_{eff} [bc]	M [d]	M_{eff} [ce]	L [f]	PDB ID
HTH_3	RPC1_BP434/7-59	15315(15917)	11691.21	6286	4893.73	53	1R69-A:6-58
Nitroreductase	Q97IT9_CLOAB/4-76	6008(6084)	4912.96	1057	854.71	73	3E10-A/B:4-76 [g]
SBP_bac_3 [h]	GLNH_ECOLI/27-244	9874(9972)	7374.96	140	99.70	218	1WDN-A:5-222
SBP_bac_3	GLNH_ECOLI/111-204	9712(9898)	7442.85	829	689.64	94	1WDN-A:89-182
OmpA	PAL_ECOLI/73-167	6035(6070)	4920.44	2207	1761.24	95	1OAP-A:52-146
DnaB	DNAB_ECOLI/31-128	1929(1957)	1284.94	1187	697.30	98	1JWE-A:30-127
LysR_substrate [h]	BENM_ACIAD/90-280	25138(25226)	20707.06	85(1)	67.00	191	2F6G-A/B:90-280 [g]
LysR_substrate	BENM_ACIAD/163-265	25032(25164)	21144.74	121(1)	99.27	103	2F6G-A/B:163-265 [g]
Methyltransf_5 [h]	RSMH_THEMA/8-292	1942(1953)	1286.67	578(2)	357.97	285	1N2X-A:8-292
Methyltransf_5	RSMH_THEMA/137-216	1877(1911)	1033.35	975(2)	465.53	80	1N2X-A:137-216
SH3_1	SRC_HUMAN:90-137	9716(16621)	3842.47	1191	458.31	48	1FMK-A:87-134
ACBP	ACBP_BOVIN/3-82	2130(2526)	1039.06	161	70.72	80	2ABD-A:2-81
PDZ	PTN13_MOUSE/1358-1438	13814(23726)	4748.76	1255	339.99	81	1GM1-A:16-96
Copper-bind	AZUR_PSEAE:24-148	1136(1169)	841.56	67(1)	45.23	125	5AZU-B/C:4-128 [g]

[a] The number of unique sequences and the total number of sequences in parentheses; the full alignments in the Pfam[Finn et al.(2016)Finn, Coggill, Eberhardt, Eddy, Mistry, Mitchell, Potter, Punta, Qureshi, Sangrador-Vegas, Salazar, Tate, and Bateman] are used.

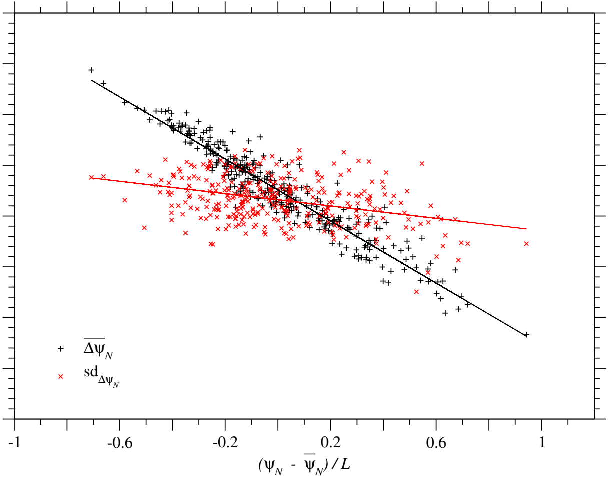
[b] The effective number of sequences.

[c] A sample weight (w_{σ_N}) for a given sequence is equal to the inverse of the number of sequences that are less than 20% different from the given sequence.

[d] The number of unique sequences that include no deletion unless specified. The number in parentheses indicates the maximum number of deletions allowed.

[e] The effective number of unique sequences that include no deletion or at most the specified number of deletions

Sample mean of $\Delta\psi_N$ lineary depends on ψ_N/L, but its standard deviation is almost constant.



Correlation between $\Delta\psi_N$ due to single nucleotide nonsynonymous substitutions and ψ_N of homologous sequences in the PDZ domain family.

Parameter values and sample mean and standard deviation of $\Delta\psi_N$

Pfam family	L	p_c	n_c [a]	r_{cutoff}	$\bar{\psi}/L$ [b]	$\delta\psi^2/L$ [b]	$\overline{\psi_N}/L$ [b]	$\overline{\overline{\Delta\psi_N}}$ [c]	$\overline{\text{Sd}(\Delta\psi_N)}$ ± [c]	r_{ψ_N}	α_{ψ_N}	r_{ψ_N}	
				(Å)					Sd(Sd($\Delta\psi_N$))	for $\overline{\Delta\psi_N}$ [d]		for Sd($\Delta\psi_N$)	
HTH_3	53	0.18	7.43	8.22	−0.1997	2.7926	−2.9861	4.2572	5.3503 ± 0.5627	−0.961	−1.5105	−0.598	−0.9
Nitroreductase	73	0.23	6.38	8.25	−0.1184	2.1597	−2.2788	3.3115	3.6278 ± 0.2804	−0.939	−1.3371	−0.426	−0.3
SBP_bac_3	218	0.25	9.23	8.10	−0.1000	2.1624	−2.2618	3.2955	3.4496 ± 0.2742	−0.980	−1.5286	−0.841	−0.7
SBP_bac_3	94	0.37	8.00	7.90	−0.1634	1.2495	−1.4054	1.9291	2.3436 ± 0.1901	−0.959	−1.3938	−0.634	−0.4
OmpA	95	0.169	8.00	8.20	−0.2457	3.9093	−4.1542	6.5757	7.6916 ± 0.3078	−0.957	−1.5694	−0.410	−0.3
DnaB	98	0.235	9.65	8.17	−0.2284	3.9976	−4.2291	6.3502	6.1244 ± 0.3245	−0.965	−1.4509	−0.495	−0.4
LysR_substrate	191	0.235	8.59	7.98	−0.2241	1.4888	−1.7173	2.2784	2.6519 ± 0.1445	−0.964	−1.3347	−0.541	−0.5
LysR_substrate	103	0.265	8.84	8.25	−0.2244	1.4144	−1.6379	2.2110	2.7371 ± 0.2055	−0.982	−1.4159	−0.727	−0.5
Methyltransf_5	285	0.13	7.99	7.78	−0.1462	7.2435	−7.3887	12.4689	10.9352 ± 0.3030	−0.981	−1.9140	−0.122	−0.0
Methyltransf_5	80	0.18	6.78	7.85	−0.1763	5.5162	−5.6896	8.9849	7.6133 ± 0.4382	−0.944	−1.4824	0.125	0.1
SH3_1	48	0.14	6.42	8.01	−0.1348	3.9109	−4.0434	5.5792	6.1426 ± 0.2935	−0.919	−1.4061	−0.196	−0.1
ACBP	80	0.22	9.17	8.24	−0.0525	4.6411	−4.7084	7.7612	7.1383 ± 0.2970	−0.972	−1.5884	−0.335	−0.2
PDZ	81	0.205	9.06	8.16	−0.2398	3.1140	−3.3572	4.7589	4.6605 ± 0.2255	−0.954	−1.5282	−0.369	−0.3
Copper-bind	125	0.23	9.50	8.27	−0.0940	4.2450	−4.3272	7.2650	6.9283 ± 0.2316	−0.980	−1.8915	−0.282	−0.2

[a] The average number of contact residues per site within the cutoff distance; the center of side chain is used to represent a residue.

[b] M unique sequences with no deletions are used with a sample weight (w_{σ_N}) for each sequence; w_{σ_N} is equal to the inverse of the number of sequences that are less than 20% different from a given sequence. The M and the effective number M_{eff} of the sequences are listed for each protein family in Table **??**.

[c] The averages of $\overline{\Delta\psi_N}$ and Sd($\Delta\psi_N$), which are the mean and the standard deviation of $\Delta\psi_N$ for a sequence, and the standard deviation of Sd($\Delta\psi_N$) over homologous sequences. Representatives of unique sequences with no deletions, which are at least 20% different from each other, are used; the number of the representatives used is almost equal to M_{eff}.

[d] The correlation and regression coefficients of $\overline{\Delta\psi_N}$ on ψ_N/L; see Eq. 44.

[e] The correlation and regression coefficients of Sd($\Delta\psi_N$) on ψ_N/L.

Sample mean and standard deviation of $\Delta\psi_N$

- **Sample mean $\overline{\Delta\psi_N}$ is negatively proportional to the ψ_N/L of the wildtype:**

$$\begin{aligned}
\overline{\Delta\Delta\psi_{ND}(\sigma^N_{j\neq i},\sigma^N_i \to \sigma_i)} \simeq \overline{\Delta\psi_N(\sigma^N_{j\neq i},\sigma^N_i \to \sigma_i)} & \\
\approx \quad \alpha_{\psi_N}\frac{\psi_N(\boldsymbol{\sigma}_N)-\overline{\psi_N(\boldsymbol{\sigma}_N)}}{L} + \overline{\overline{\Delta\psi_N(\sigma^N_{j\neq i},\sigma^N_i \to \sigma_i)}} & \qquad \text{with } \alpha_{\psi_N} < 0 \qquad (44)
\end{aligned}$$

where L is sequence length.

- **The standard deviation is almost constant:**

$$\begin{aligned}
\mathrm{Sd}(\Delta\psi_N(\sigma^N_{j\neq i},\sigma^N_i \to \sigma_i)) & \approx & \text{independent of } \psi_N \text{ and} \\
& & \text{constant across homologous sequences in every protein family} \\
& = & \text{function of } k_B T_s \qquad (45)
\end{aligned}$$

Effective temperature T_s of selection is estimated from the changes of interaction, $\Delta\psi_N$, due to single nucleotide nonsynonymous substitutions

$$\begin{aligned}
\mathrm{Sd}(\Delta\psi_N(\sigma^N_{j\neq i}, \sigma^N_i \to \sigma_i)) &\approx \text{independent of } \psi_N \text{ and} \\
& \text{constant across homologous sequences in every protein family} \\
&= \text{function of } k_B T_s \qquad (46) \\
\mathrm{Sd}(\Delta G_N(\sigma^N_{j\neq i}, \sigma^N_i \to \sigma_i)) &= \text{function that must not explicitly depend on } k_B T_s \text{ but } G_N \qquad (47)
\end{aligned}$$

From the equations above, we obtain the important relation that the standard deviation of $\Delta G_N (\simeq k_B T_s \Delta\psi_N)$ does not depend on G_N and is nearly constant irrespective of protein families.

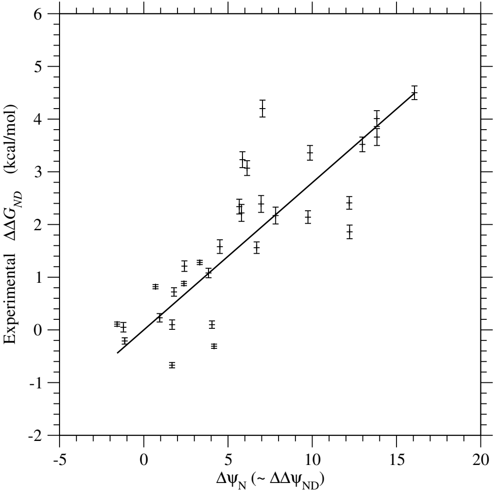
$$\begin{aligned}
\mathrm{Sd}(\Delta G_N(\sigma^N_{j\neq i}, \sigma^N_i \to \sigma_i)) &\simeq k_B T_s\, \mathrm{Sd}(\Delta\psi_N(\sigma^N_{j\neq i}, \sigma^N_i \to \sigma_i)) \\
&\approx \text{constant} \qquad (48)
\end{aligned}$$

PDZ protein is employed as a reference protein to estimate $k_B T_s$ for other proteins.

$$k_B \hat{T}_s = k_B \hat{T}_{s,\,\mathrm{PDZ}} \left[\, \overline{\mathrm{Sd}(\Delta\psi_{\mathrm{PDZ}}(\sigma^N_{j\neq i}, \sigma^N_i \to \sigma_i))} \,/\, \overline{\mathrm{Sd}(\Delta\psi_N(\sigma^N_{j\neq i}, \sigma^N_i \to \sigma_i))} \,\right] \qquad (49)$$

where the overline denotes the average over all homologous sequences.

4-2. A direct comparison of $\Delta\psi_N (\simeq \Delta\Delta\psi_{ND})$ with the experimental $\Delta\Delta G_{ND}$ to estimate $k_B T_s$ for the reference protein, PDZ.



Regression of the experimental values (Gianni et al., 2007) of folding free energy changes ($\Delta\Delta G_{ND}$) due to single amino acid substitutions on $\Delta\psi_N (\simeq \Delta\Delta\psi_{ND})$ for the same types of substitutions in the PDZ domain.

4-3. Thermodynamic quantities estimated with $r_{\text{cutoff}} \sim 8$ Å.

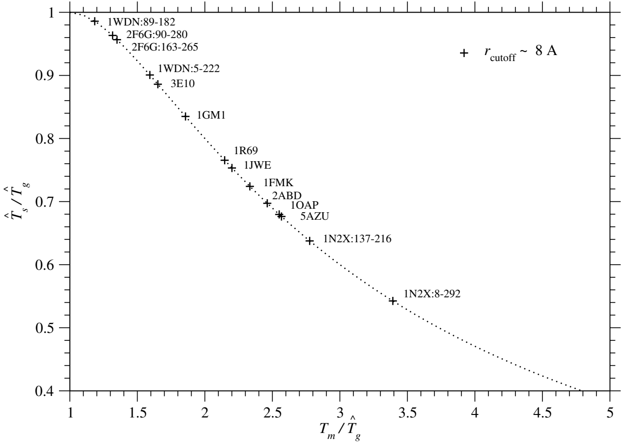
				Experimental				
Pfam family	r [a]	$k_B \hat{T}_s$ [a]	$\hat{T}_s$	T_m	$\hat{T}_g$	$\hat{\omega}$ [b]	T [c]	$\langle \Delta G_{ND} \rangle$ [d]
		(kcal/mol)	(°K)	(°K)	(°K)	(k_B)	(°K)	(kcal/mol)
HTH_3	–	–	122.6	343.7	160.1	0.8182	298	−2.95
Nitroreductase	–	–	180.7	337	204.0	0.8477	298	−2.81
SBP_bac_3	–	–	190.1	336.1	211.0	0.8771	298	−8.03
SBP_bac_3	–	–	279.8	336.1	283.8	0.6072	298	−.85
OmpA	–	–	85.2	320	125.4	0.9027	298	−3.13
DnaB	–	–	107.1	312.8	142.1	1.1341	298	−2.56
LysR_substrate	–	–	247.3	338	256.7	0.6908	298	−3.63
LysR_substrate	–	–	239.6	338	250.4	0.6472	298	−2.00
Methyltransf_5	–	–	60.0	375	110.5	1.0656	298	−41.36
Methyltransf_5	–	–	86.1	375	135.1	1.1214	298	−11.48
SH3_1	0.865	0.1583	106.7	344	147.4	1.0253	295	−3.76
ACBP	0.825	0.1169	91.9	324.4	131.7	1.1281	278	−6.72
PDZ	0.931	0.2794	140.7	312.88	168.5	1.0854	298	−1.81
Copper-bind	0.828	0.1781	94.6	359.3	139.9	0.9709	298	−12.07

[a] Reflective correlation (r) and regression ($k_B \hat{T}_s$) coefficients for least-squares regression lines of experimental $\Delta\Delta G_{ND}$ on $\Delta\psi_N$ through the origin.

[b] Conformational entropy per residue, in k_B units, in the denatured molten-globule state; $\omega = (T_s/T_g)^2 \delta\psi^2/(2L)$

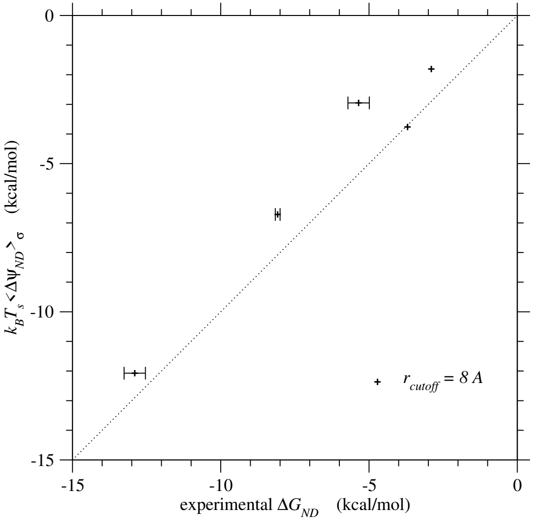
[d] Folding free energy in kcal/mol units; $\langle \Delta G_{ND}(\sigma, T) \rangle_{\sigma}/(k_B T_s) \approx \delta\psi^2 (\overline{\mathbf{f}(\sigma_N)}) \left[\vartheta(T/T_g) T_s/T - 1 \right]$

The values of T_g estimated from the estimated T_s and experimental T_m satisfy the condition for protein folding, $T_s < T_g < T_m$.



$\hat{T}_s/\hat{T}_g$ **is plotted against** $T_m/\hat{T}_g$ **for each protein domain.** A dotted curve corresponds to the condition of $\langle \Delta G_{ND}(\boldsymbol{\sigma_N}, T_m)\rangle_{\boldsymbol{\sigma}} = 0$, $\hat{T}_s/\hat{T}_g = 2(T_m/\hat{T}_g)/((T_m/\hat{T}_g)^2 + 1)$.

The values of $\langle \Delta G_{ND}(\sigma, T)\rangle_{\sigma}$ estimated from the estimated T_s and experimental T_m almost agree with their experimental values.



Folding free energies, $\langle \Delta G_{ND}\rangle_{\sigma} \simeq k_B T_s \langle \Delta\psi_{ND}\rangle_{\sigma}$, predicted by the present method are plotted against their experimental values, $\Delta G_{ND}(\sigma_N)$.

4-4. The equilibrium state of evolutionary statistical energy ψ_N in the mutation–fixation process of amino acid substitutions

- Monoclonal approximation for the fixation process of amino acid substitutions:
  Protein evolution is assumed to proceed with a single amino acid substitution fixed at a time in a population.

  Equilibrium condition for $\Delta\psi_{ND}$ and ψ_N:

$$\langle\Delta\Delta\psi_{ND}\rangle_{\text{fixed}} \simeq \langle\Delta\psi_N\rangle_{\text{fixed}} = 0 \quad\Longleftrightarrow\quad \Delta\psi_{ND} \text{ and } \psi_N \text{ are at equilibrium.} \tag{50}$$

  The PDF of $\Delta\Delta\psi_{ND}$ in fixed mutants is proportional to that multiplied by the fixation probability.

$$p(\Delta\Delta\psi_{ND,\text{fixed}}) = p(\Delta\Delta\psi_{ND})\frac{u(s(\Delta\Delta\psi_{ND}))}{\langle u(s(\Delta\Delta\psi_{ND}))\rangle} \tag{51}$$

$$\langle u(s(\Delta\Delta\psi_{ND}))\rangle \equiv \int_{-\infty}^{\infty} u(s(\Delta\Delta\psi_{ND}))p(\Delta\Delta\psi_{ND})d\Delta\Delta\psi_{ND} \tag{52}$$

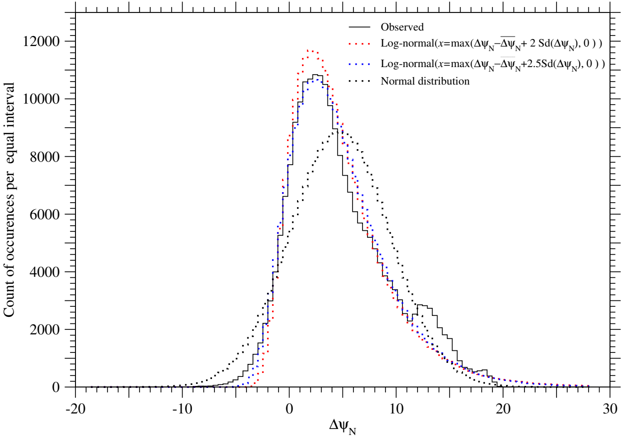
  The selective advantage of $\boldsymbol{\mu}$ to $\boldsymbol{w}$ildtype is represented as follows for $f(\boldsymbol{w}) = f(\boldsymbol{\mu}) = \overline{f(\boldsymbol{\sigma})}$.

$$4N_e s(\boldsymbol{w} \to \boldsymbol{\mu})(1 - q_m) \equiv (4N_e m(\boldsymbol{\mu}) - 4N_e m(\boldsymbol{w}))(1 - q_m) \tag{53}$$

$$= -(\Delta\psi_{ND}(\boldsymbol{\mu}, T) - \Delta\psi_{ND}(\boldsymbol{w}, T)) = -(\psi_N(\boldsymbol{\mu}) - \psi_N(\boldsymbol{w})) \tag{54}$$

$$\simeq -(\Delta G_{ND}(\boldsymbol{\mu}, T) - \Delta G_{ND}(\boldsymbol{w}, T))/(k_B T_s) = -(G_N(\boldsymbol{\mu}) - G_N(\boldsymbol{w}))/(k_B T_s) \tag{55}$$

4-5. The observed frequency distribution of $\Delta\psi_N$ can be well approximated by a log-normal distribution.



The observed frequency distribution and the fitted distributions of $\Delta\psi_N$ in the PDZ protein family. Only representatives of unique sequences with no deletions, which are at least 20% different from each other, are employed; the total count is equal to 222,466 over 335 homologous sequences.

Approximation of the observed frequency distribution of $\Delta\psi_N$ with a log-normal

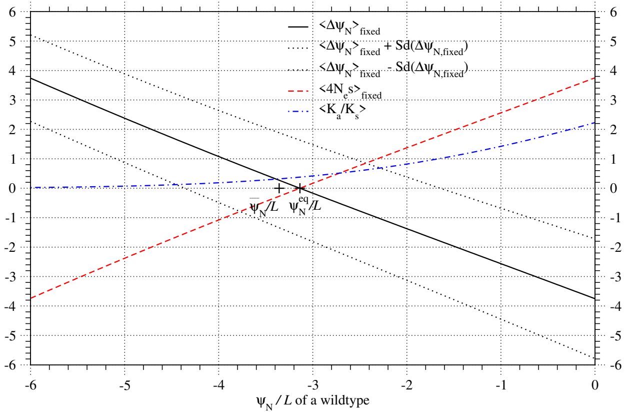
The equilibrium condition is calculated by approximating the probability density of $\Delta\psi_N$ with a log-normal distribution.

$$\begin{aligned}
p(\Delta\psi_N) &\approx \ln\mathcal{N}(x;\mu,\sigma) \equiv \frac{1}{x}\mathcal{N}(\ln x;\mu,\sigma) && (56)\\
x &\equiv \max(\Delta\psi_N - \Delta\psi_N^{\mathrm{o}}, 0) && (57)\\
\exp(\mu + \sigma^2/2) &= \overline{\Delta\psi_N} - \Delta\psi_N^{\mathrm{o}} && (58)\\
\exp(2\mu + \sigma^2)(\exp(\sigma^2) - 1) &= \overline{(\Delta\psi_N - \overline{\Delta\psi_N})^2)} && (59)\\
\Delta\psi_N^{\mathrm{o}} &\equiv \min(\overline{\Delta\psi_N} - n_{\mathrm{shift}}\overline{(\Delta\psi_N - \overline{\Delta\psi_N})^2)}^{1/2}, 0) && (60)
\end{aligned}$$

where $\Delta\psi_N^{\mathrm{o}}$ is the origin for the log-normal distribution and the shifting factor n_{shift} is taken to be equal to 2, unless specified. The parameters for a log-normal distribution are determined by employing the regression line of $\overline{\Delta\psi_N}$ on ψ_N and $\overline{Sd(\Delta\psi_N)}$ observed in respective protein families.

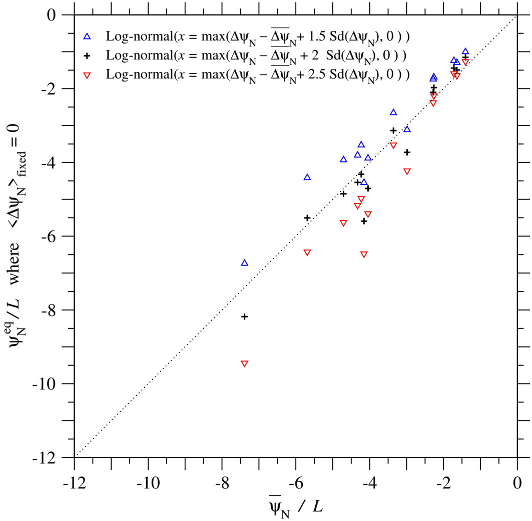
$$\begin{aligned}
\overline{\Delta\psi_N} &\approx \alpha_{\psi_N}\frac{\psi_N(\boldsymbol{\sigma}_N) - \overline{\psi_N(\boldsymbol{\sigma}_N)}}{L} + \overline{\overline{\Delta\psi_N(\sigma^N_{j\neq i}, \sigma^N_i \to \sigma_i)}} \quad \text{with } \alpha_{\psi_N} < 0 && (61)\\
\overline{(\Delta\psi_N - \overline{\Delta\psi_N})^2}^{1/2} &\approx \overline{\overline{(\Delta\psi_N(\sigma^N_{j\neq i}, \sigma^N_i \to \sigma_i) - \overline{\Delta\psi_N(\sigma^N_{j\neq i}, \sigma^N_i \to \sigma_i)})^2}}^{1/2} && (62)
\end{aligned}$$

4-6. Evolutionary statistical energy ψ_N in the mutation–fixation process of amino acid substitutions has a stable equilibrium value, because $\overline{\Delta\psi_N}$ and therefore $\langle\Delta\psi_N\rangle_{\text{fixed}}$ are decreasing functions of ψ_N/L.



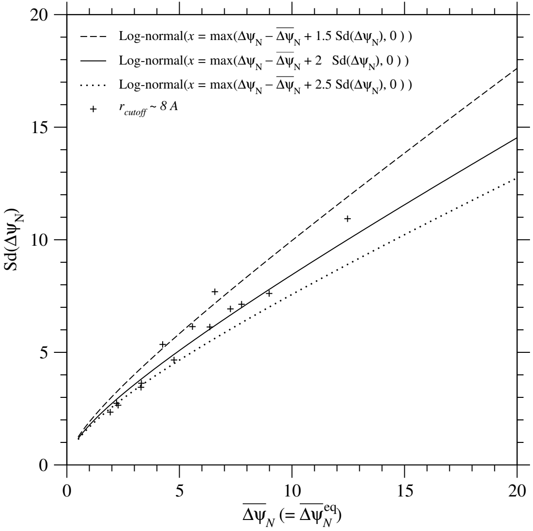
The average of $\Delta\psi_N (\simeq \Delta\Delta\psi_{ND})$ over fixed single nucleotide nonsynonymous mutations versus ψ_N/L of a wildtype for the PDZ protein family.

4-7. The equilibrium value of ψ_N almost agrees with the sample average of ψ_N over all homologous sequences.



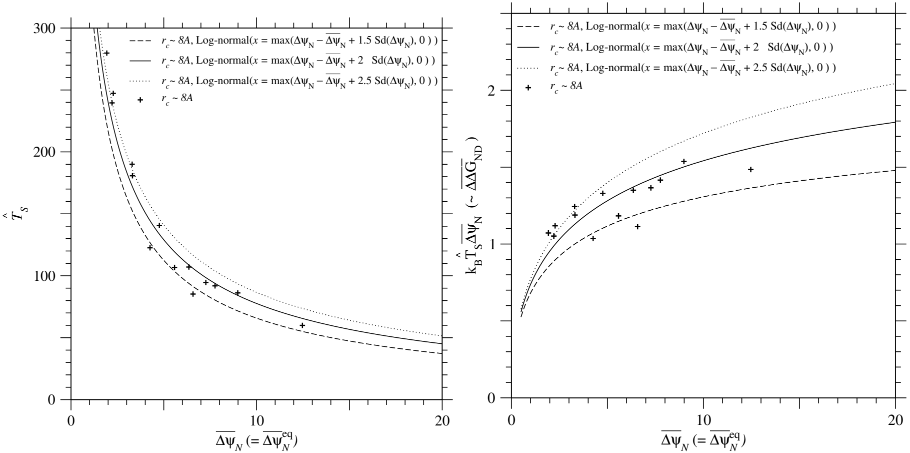
The equilibrium value of ψ_N/L, where $\langle\Delta\psi_N\rangle_{\text{fixed}} = 0$, is plotted against the average of ψ_N/L over homologous sequences for each protein family.

4-8. The relationship between the standard deviation and the mean of $\Delta\psi_N$ at the equilibrium state of ψ_N



Relationship between the mean and the standard deviation of $\Delta\psi_N$ due to single nucleotide nonsynonymous mutations at equilibrium, $\langle\Delta\psi_N\rangle_{\textbf{fixed}} = 0$

The relationship between T_s and the mean of $\Delta\psi_N$ at the equilibrium state of ψ_N



Relationships between $\hat{T}_s$ and $\overline{\Delta\psi_N}$ and between $k_B\hat{T}_s\overline{\Delta\psi_N}(\simeq \overline{\Delta\Delta G_{ND}})$ and $\overline{\Delta\psi_N}$ at equilibrium, $\langle\Delta\psi_N\rangle_{\text{fixed}} = 0$.

4-9. Probability distributions of selective advantage, fixation rate, and K_a/K_s

Let us consider the PDFs of characteristic quantities that describe the evolution of genes.

- The PDF of $4N_e s$:

$$p(4N_e s) \quad = \quad p(\Delta\Delta\psi_{ND})\,|\frac{d\Delta\Delta\psi_{ND}}{d4N_e s}| = p(\Delta\Delta\psi_{ND})(1-q_m) \tag{63}$$

  where $\Delta\Delta\psi_{ND}$ must be regarded as a function of $4N_e s$, that is, $\Delta\Delta\psi_{ND} = -4N_e s(1-q_m)$; see Eq. 21.

- The PDF of fixation probability u:

$$p(u) \quad = \quad p(4N_e s)\frac{d4N_e s}{du} = p(4N_e s)\frac{(e^{4N_e s}-1)^2 e^{4N_e s(q_m-1)}}{q_m(e^{4N_e s}-1)-(e^{4N_e s q_m}-1)} \tag{64}$$

  where $4N_e s$ must be regarded as a function of u.

- The ratio of the nonsynonymous substitution rate per nonsynonymous site (K_a) with selective advantage s to the substitution rate per synonymous site (K_s) with s = 0

$$\frac{K_a}{K_s} \quad = \quad \frac{u(s)}{u(0)} = \frac{u(s)}{q_m} \quad , \quad p(K_a/K_s) = p(u)\frac{du}{d(K_a/K_s)} = p(u)\,q_m \tag{65}$$

  assuming that synonymous substitutions are completely neutral and mutation rates at both types of sites are the same.

Probability distributions of $\Delta\Delta\psi_{ND}$, $4N_e s$, u, and K_a/K_s in fixed mutant genes

The PDF of $\Delta\Delta\psi_{ND}$ in fixed mutants is proportional to that multiplied by the fixation probability.

$$p(\Delta\Delta\psi_{ND,\text{fixed}}) = p(\Delta\Delta\psi_{ND})\frac{u(s(\Delta\Delta\psi_{ND}))}{\langle u(s(\Delta\Delta\psi_{ND}))\rangle} \tag{66}$$

$$\langle u(s(\Delta\Delta\psi_{ND}))\rangle \equiv \int_{-\infty}^{\infty} u(s(\Delta\Delta\psi_{ND}))p(\Delta\Delta\psi_{ND})d\Delta\Delta\psi_{ND} \tag{67}$$

Likewise, the PDF of selective advantage in fixed mutants is

$$p(4N_e s_{\text{fixed}}) = p(4N_e s)\frac{u(s)}{\langle u(s)\rangle} \tag{68}$$

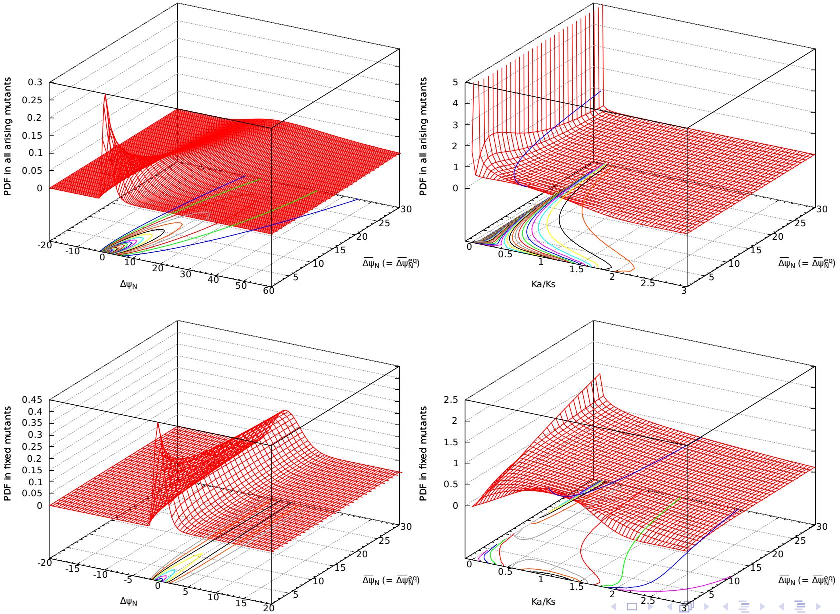
and those of the u and K_a/K_s in fixed mutants are

$$p(u_{\text{fixed}}) = p(u)\frac{u}{\langle u\rangle} \tag{69}$$

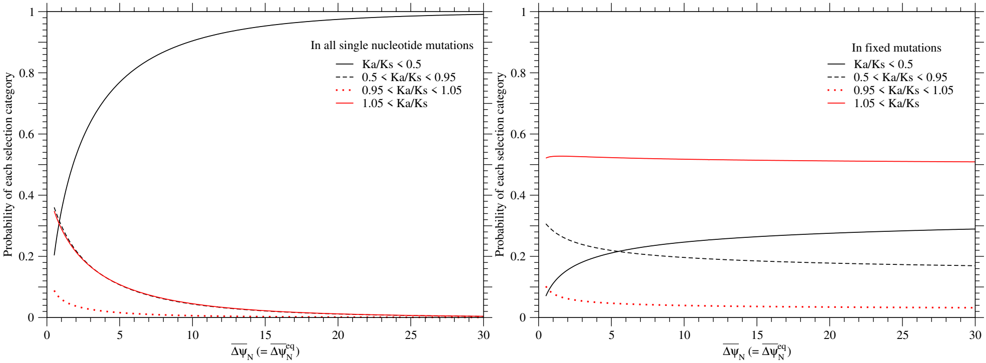
$$p((\frac{K_a}{K_s})_{\text{fixed}}) = p(\frac{K_a}{K_s})\frac{u}{\langle u\rangle} = p(\frac{K_a}{K_s})\frac{\frac{K_a}{K_s}}{\langle\frac{K_a}{K_s}\rangle} \tag{70}$$

The average of K_a/K_s in fixed mutants is equal to the ratio of the second moment to the first moment of K_a/K_s in all arising mutants; $\langle K_a/K_s\rangle_{\text{fixed}} = \langle (K_a/K_s)^2\rangle/\langle K_a/K_s\rangle$.

4-10. PDFs of $\Delta\psi_N$ and K_a/K_s as a function of $\overline{\Delta\psi_N}$

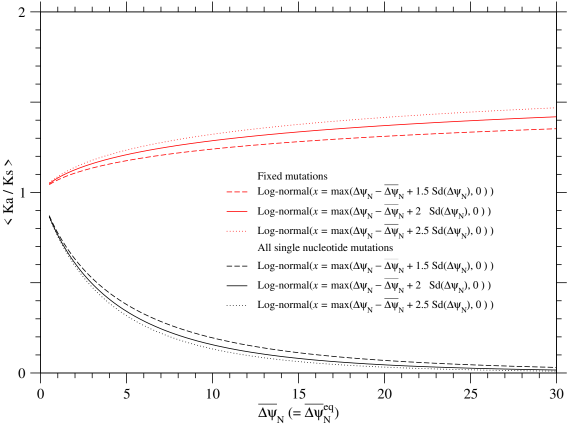


4-11. The probability of neutral ($0.95 < K_a/K_s < 1.05$) selection category is insignificant in fixed mutations.



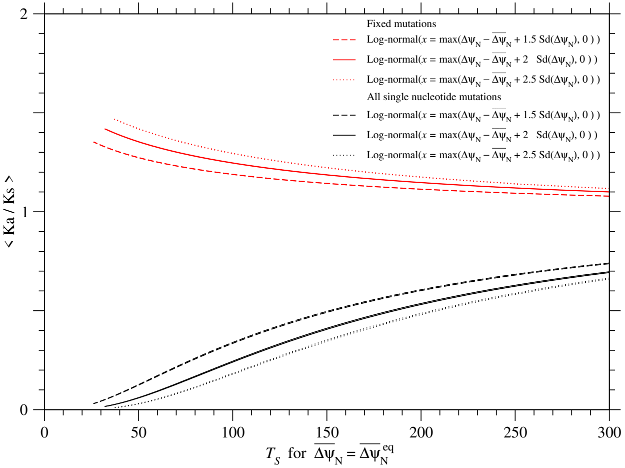
The probabilities of each selection category in all single nucleotide nonsynonymous mutations and in their fixed mutations as a function of $\overline{\Delta\psi_N}$ at equilibrium, $\langle\Delta\psi_N\rangle_{\text{fixed}} = 0$.

$\langle K_a/K_s\rangle < 1$ in arising mutations, although $\langle K_a/K_s\rangle_{\text{fixed}} > 1$ in fixed mutations.



The averages of K_a/Ks over all single nucleotide nonsynonymous mutations and over their fixed mutations as a function of $\overline{\Delta\psi_N}$ at equilibrium, $\langle\Delta\psi_N\rangle_{\text{fixed}} = 0$.

4-12. $\langle K_a/K_s \rangle$ as a function of T_s at the equilibrium state of ψ_N



The averages of K_a/Ks over all single nucleotide nonsynonymous mutations and over their fixed mutations as a function of the effective temperature of selection, $T_s (= (T_s\overline{Sd}(\Delta\psi_N))_{PDZ}/Sd(\Delta\psi_N))$, at equilibrium, $\langle\Delta\psi_N\rangle_{\textbf{fixed}} = 0$.

5. Summary

- A Boltzmann distribution with protein fitness is derived under the assumption that amino acid substitutions are at equilibrium in a reversible Markov process.
- Relationships are obtained for folding free energy, folding statistical energy and fitness.
- Selective temperature, and then, glass transition temperature and folding free energy are estimated for 14 protein domains with the estimated T_s and experimental T_m. Their estimated values fall in a reasonable range.
- The equilibrium value of ψ_N at $\langle \Delta\psi_N \rangle_{\text{fixed}} = 0$ well agrees with the mean of ψ_N over all the homologous sequences in each protein family, indicating the consistency of the present theory.
- Selective temperature is directly related to substitution rate (K_a / K_s).
- Protein stability and foldability are kept in a balance of positive selection and random drift.
- Positive and negative mutations are significantly fixed in stability/foldability selection, supporting the nearly neutral theory rather than the neutral theory for protein evolution.