20431P

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

Sanzo Miyazawa
sanzo.miyazawa@gmail.com

at the 58th BPSJP annual meeting in Gunma on September 16-18, 2020

Slides: https://www.sanzo.org/tmp/BPSJP20431P.pdf
Publication-1: J. Theor. Biol. 433, 21-38, 2017 (arXiv:1612.09379)
Publication-2: IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2020 (arXiv:1909.05006)

1. Background

- The probability distribution ($P(\boldsymbol{\sigma})$) of homologous sequences ($\boldsymbol{\sigma}$) in a protein family can be well approximated by a Boltzmann distribution (Figliuzzi et al., 2018):

$$P(\boldsymbol{\sigma}) \quad \propto \quad \exp(-\psi_N(\boldsymbol{\sigma})) \ , \ \psi_N(\boldsymbol{\sigma}) \equiv -(\sum_{i}^{L}(h_i(\sigma_i) + \sum_{j>i} J_{ij}(\sigma_i, \sigma_j))) \tag{1}$$

  where h_i and J_{ij} are one-body at site i and two-body interactions between sites i and j; in this study, h_i and J_{ij} were estimated from a MSA of each protein family in the mean field approximation with the DCA program (Marks et. al. 2011).

- A protein folding theory based on the random energy model (REM) indicates:

$$P(\boldsymbol{\sigma}) \quad \propto \quad P^{\mathrm{mut}}(\boldsymbol{\sigma}) \exp(\frac{-\Delta G_{ND}(\boldsymbol{\sigma}, T)}{k_B T_s}) \quad \propto \exp(\frac{-G_N(\boldsymbol{\sigma})}{k_B T_s}) \quad \text{if } \boldsymbol{f}(\boldsymbol{\sigma}) = \text{constant} \tag{2}$$

  where $\Delta G_{ND} \equiv G_N - G_D$, G_N and G_D are the native and denatured free energies, T_s is the effective temperature representing the strength of selection pressure, and $P^{\mathrm{mut}}(\boldsymbol{\sigma})$ is the probability of sequence $\boldsymbol{\sigma}$ in the mutational process (Shakhnovich et al., 1993).

- 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

We 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:

$$P(\sigma) \quad \propto \quad P^{\mathrm{mut}}(\sigma) \exp(4N_e m(1 - 1/(2N))) \tag{3}$$

   where N_e and N are effective and actual population sizes, and m is the Malthusian fitness of a gene.

2. Relationships between $\Delta\psi_{ND}$, ΔG_{ND}, and m are obtained from Eqs. 1, 2, and 3 .
3. 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.
4. 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. The equilibrium distribution of sequences in a mutation-fixation process

Assumption: The mutational process is a reversible Markov process;
$P^{\mathrm{mut}}(\boldsymbol{\mu})M_{\boldsymbol{\mu\nu}} = P^{\mathrm{mut}}(\boldsymbol{\nu})M_{\boldsymbol{\nu\mu}}$, where $M_{\boldsymbol{\mu\nu}}$ is the mutation rate per gene from sequence $\boldsymbol{\mu}$ to $\boldsymbol{\nu}$.

A Markov process with the substitution rate $R_{\boldsymbol{\mu\nu}}$ from $\boldsymbol{\mu}$ to $\boldsymbol{\nu}$ for diploid is reversible.

$$R_{\boldsymbol{\mu\nu}} \equiv 2NM_{\boldsymbol{\mu\nu}}u(s(\boldsymbol{\mu} \to \boldsymbol{\nu})) \tag{4}$$

$$2Nu(s) = 2N\frac{1-e^{-4N_e s q_m}}{1-e^{-4N_e s}} = \frac{u(s)}{u(0)} \quad \text{with} \quad q_m = \frac{1}{2N} \tag{5}$$

$$s(\boldsymbol{\mu} \to \boldsymbol{\nu}) \equiv m(\boldsymbol{\nu}) - m(\boldsymbol{\mu}) \tag{6}$$

$$\exp(4N_e m(\boldsymbol{\mu})(1-q_m))\, u(s(\boldsymbol{\mu} \to \boldsymbol{\nu})) = \exp(4N_e m(\boldsymbol{\nu})(1-q_m))\, u(s(\boldsymbol{\nu} \to \boldsymbol{\mu})) \tag{7}$$

where $u(s(\boldsymbol{\mu} \to \boldsymbol{\nu}))$ 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). Thus, the equilibrium distribution is

$$P(\boldsymbol{\sigma}) \propto P^{\mathrm{mut}}(\boldsymbol{\sigma}) \exp(4N_e m(1-1/(2N))) \tag{8}$$

3-2. Relationships between $m(\boldsymbol{\sigma})$, $\Delta\psi_{ND}(\boldsymbol{\sigma}, T)$, and $\Delta G_{ND}(\boldsymbol{\sigma}, T)$ of protein sequence

From Eqs. 1, 2, and 3 , 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.

$$
\begin{aligned}
P^{\mathrm{eq}}(\boldsymbol{\sigma}) &\propto P^{\mathrm{mut}}(\boldsymbol{\sigma}) \exp(4N_e m(\boldsymbol{\sigma})(1-q_m)) && (9)\\
&\propto P^{\mathrm{mut}}(\overline{\boldsymbol{\sigma}}) \exp(-(\psi_N(\boldsymbol{\sigma}) - \psi_D(\overline{\boldsymbol{f}(\boldsymbol{\sigma})}, T))) && (10)\\
&\propto \simeq P^{\mathrm{mut}}(\boldsymbol{\sigma}) \exp(-\Delta G_{ND}(\boldsymbol{\sigma}, T)/(k_B T_s)) && (11)
\end{aligned}
$$

where $\overline{\boldsymbol{f}(\boldsymbol{\sigma})} \equiv \sum_{\boldsymbol{\sigma}} \boldsymbol{f}(\boldsymbol{\sigma}) P(\boldsymbol{\sigma})$ 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 $f(\boldsymbol{\sigma}) = \overline{f(\boldsymbol{\sigma})}$.

$$
\begin{aligned}
4N_e m(\boldsymbol{\sigma})(1-q_m) &= -\Delta\psi_{ND}(\boldsymbol{\sigma}, T) + \text{constant} && (12)\\
&\simeq \frac{-\Delta G_{ND}(\boldsymbol{\sigma}, T)}{k_B T_s} + \text{constant} && (13)\\
4N_e s(\boldsymbol{\mu} \to \boldsymbol{\nu})(1-q_m) &= -(\Delta\psi_{ND}(\boldsymbol{\nu}, T) - \Delta\psi_{ND}(\boldsymbol{\mu}, T)) = -(\psi_N(\boldsymbol{\nu}) - \psi_N(\boldsymbol{\mu})) && (14)\\
\psi_N(\boldsymbol{\sigma}) &\simeq G_N(\boldsymbol{\sigma})/(k_B T_s) + \text{ function of } \boldsymbol{f}(\boldsymbol{\sigma}) && (15)\\
\psi_D(\boldsymbol{f}(\boldsymbol{\sigma}), T) &\simeq G_D(\boldsymbol{f}(\boldsymbol{\sigma}), T)/(k_B T_s) + \text{ function of } \boldsymbol{f}(\boldsymbol{\sigma}) && (16)
\end{aligned}
$$

- The distribution of conformational energies in the denatured state (molten globule state) is approximated in the random energy model (REM) (Shakhnovich and Gutin, 1993; 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.

$$G_D(\boldsymbol{f}(\boldsymbol{\sigma}), T) \approx \bar{E}(\boldsymbol{f}(\boldsymbol{\sigma})) - \frac{\delta E^2(\boldsymbol{f}(\boldsymbol{\sigma}))}{2k_B T} - k_B T \omega L = \bar{E}(\boldsymbol{f}(\boldsymbol{\sigma})) - \delta E^2(\boldsymbol{f}(\boldsymbol{\sigma})) \frac{\vartheta(T/T_g)}{k_B T} \quad (17)$$

$$\vartheta(T/T_g) \equiv \begin{cases} (1 + T^2/T_g^2)/2 & \text{for } T > T_g \\ T/T_g & \text{for } T \le T_g \end{cases} \quad (18)$$

  where ω is the conformational entropy per residue in the compact denatured state, and 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 ensemble average of $\Delta G_{ND}(\boldsymbol{\sigma}, T)$ over sequences with Eq. 2 is

$$\langle \Delta G_{ND}(\boldsymbol{\sigma}, T) \rangle_{\boldsymbol{\sigma}} \approx \langle G_N(\boldsymbol{\sigma}) \rangle_{\boldsymbol{\sigma}} - G_D(\overline{\boldsymbol{f}(\boldsymbol{\sigma}_N)}, T) \quad (19)$$

  where $\boldsymbol{\sigma}_N$ denotes a natural sequence.
- $\langle G_N(\boldsymbol{\sigma}) \rangle_{\boldsymbol{\sigma}}$ is estimated in the Gaussian approximation (Pande et al. 1997).

$$\langle G_N(\boldsymbol{\sigma}) \rangle_{\boldsymbol{\sigma}} \approx \bar{E}(\overline{\boldsymbol{f}(\boldsymbol{\sigma_N})}) - \delta E^2(\overline{\boldsymbol{f}(\boldsymbol{\sigma_N})})/(k_B T_s) \quad (20)$$

4-1. 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 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.

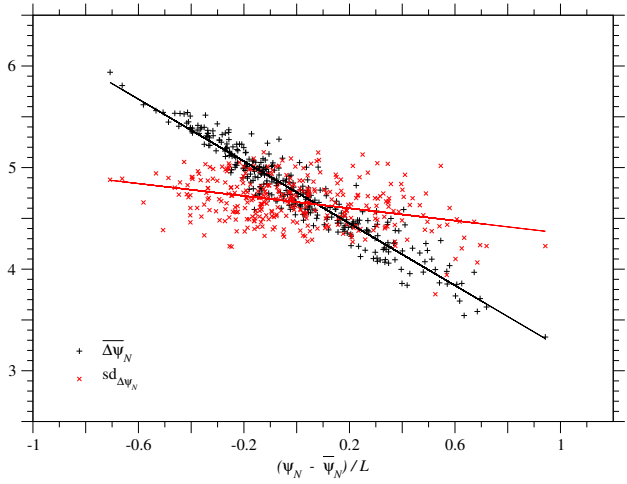
[d] The number of unique sequences that include no deletion unless specified. The number in parentheses indicates the maximum number of deletions allowed.

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

[f] The number of residues.

[g] Contacts are calculated in the homodimeric state for these protein

4-2. Changes of the evolutionary energy, $\Delta\psi_N$, due to single nucleotide nonsynonymous substitutions: The 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 the sample mean and standard deviation of $\Delta\psi_N$

Pfam family	L	p_c	n_c [a]	r_{cutoff} (Å)	$\bar{\psi}/L$ [b]	$\delta\psi^2/L$ [b]	$\overline{\psi_N}/L$ [b]	$\overline{\overline{\Delta\psi_N}}$ [c]	$\overline{\text{Sd}(\Delta\psi_N)} \pm$ [c] $\text{Sd}(\text{Sd}(\Delta\psi_N))$	r_{ψ_N} for $\overline{\Delta\psi_N}$ [d]	α_{ψ_N} for $\overline{\Delta\psi_N}$ [d]	r_{ψ_N} for $\text{Sd}(\Delta\psi_N)$ [e]	α_{ψ_N} for $\text{Sd}(\Delta\psi_N)$ [e]
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	−0.9888
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	−0.3721
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	−0.7876
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	−0.4815
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	−0.3804
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	−0.4198
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	−0.5664
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	−0.5307
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	−0.0783
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	0.1141
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	−0.1718
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	−0.2235
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	−0.3042
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	−0.2352

[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 7.

[c] The averages of $\overline{\Delta\psi_N}$ and $\text{Sd}(\Delta\psi_N)$, which are the mean and the standard deviation of $\Delta\psi_N$ due to single nucleotide nonsynonymous mutations for a sequence, and the standard deviation of $\text{Sd}(\Delta\psi_N)$ 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.

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

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

$$\begin{aligned}
\mathrm{Sd}(\Delta\psi_N(\sigma^N_{j\neq i}, \sigma^N_i \to \sigma_i)) &\approx \text{independent of } \psi_N \text{ and} \\
&\quad \text{constant across homologous sequences in every protein family} \\
&= \text{function of } k_B T_s \qquad (21) \\
\mathrm{Sd}(\Delta G_N(\sigma^N_{j\neq i}, \sigma^N_i \to \sigma_i)) &= \text{function that must not explicitly depend on } k_B T_s \text{ but } G_N \qquad (22)
\end{aligned}$$

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.

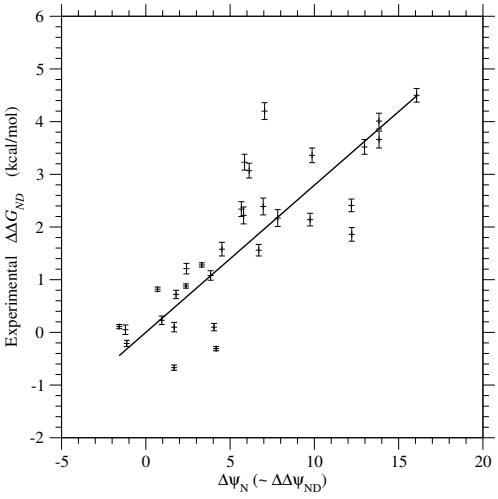
$$\begin{aligned}
\mathrm{Sd}(\Delta G_N(\sigma^N_{j\neq i}, \sigma^N_i \to \sigma_i)) &\simeq k_B T_s \, \mathrm{Sd}(\Delta\psi_N(\sigma^N_{j\neq i}, \sigma^N_i \to \sigma_i)) \\
&\approx \text{constant} \qquad (23)
\end{aligned}$$

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

$$k_B \hat{T}_s = k_B \hat{T}_{s,\,\mathrm{PDZ}} \left[\overline{\mathrm{Sd}(\Delta\psi_{\mathrm{PDZ}}(\sigma^N_{j\neq i}, \sigma^N_i \to \sigma_i))} \,/\, \overline{\mathrm{Sd}(\Delta\psi_N(\sigma^N_{j\neq i}, \sigma^N_i \to \sigma_i))} \right] \qquad (24)$$

where the overline denotes the average over all homologous sequences.

4-3. 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.

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

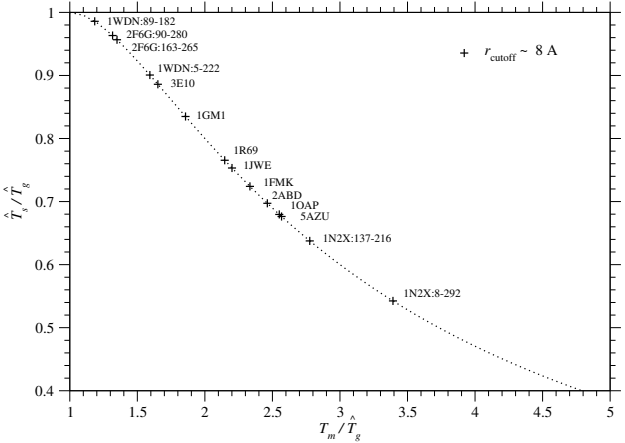
Pfam family	r [a]	$k_B \hat{T}_s$ [a] (kcal/mol)	Experimental $\hat{T}_s$ (°K)	T_m (°K)	$\hat{T}_g$ (°K)	$\hat{\omega}$ [b] (k_B)	T [c] (°K)	$\langle \Delta G_{ND} \rangle$ [d] (kcal/mol)
HTH_3	–	–	122.6	343.7	160.1	0.8182	298	−2.95
Nitroreductase	–	–	180.7	337	204.0	0.8477	298	−2.81
SBP_bac_3	–	–	190.1	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; $\omega = (T_s/T_g)^2 \delta\psi^2/(2L)$

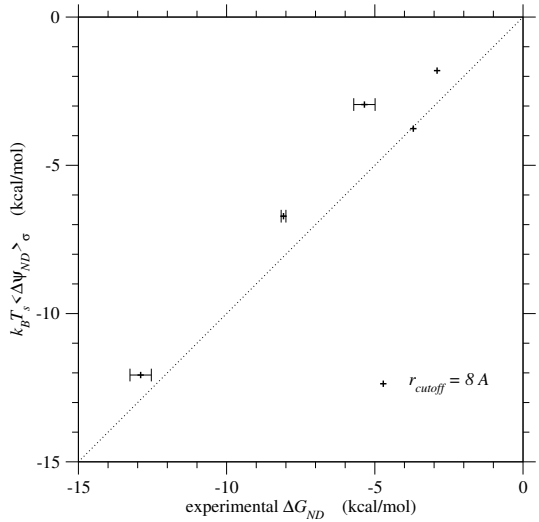
[d] Folding free energy in kcal/mol units; $\langle \Delta G_{ND}(\boldsymbol{\sigma}, T) \rangle_{\boldsymbol{\sigma}} / (k_B T_s) \approx \delta\psi^2 (\overline{\mathbf{f}(\boldsymbol{\sigma}_N)}) \, [\vartheta(T/T_g) T_s/T - 1]$

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



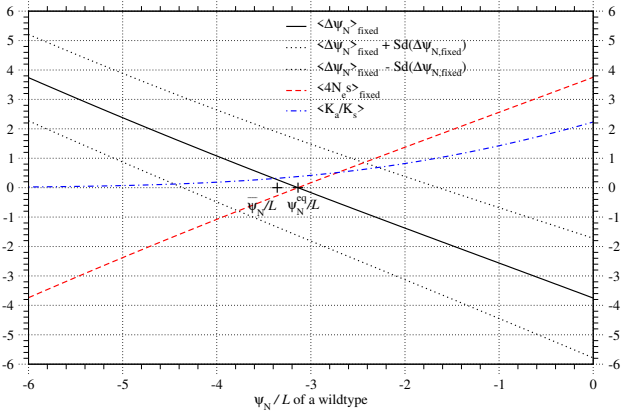
$\hat{T}_s/\hat{T}_g$ **is plotted against** $T_m/\hat{T}_g$ **for each protein domain.** A dotted curve corresponds to the condition of $\langle \Delta G_{ND}(\boldsymbol{\sigma_N}, T_m) \rangle_{\boldsymbol{\sigma}} = 0$, $\hat{T}_s/\hat{T}_g = 2(T_m/\hat{T}_g)/((T_m/\hat{T}_g)^2 + 1)$.

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



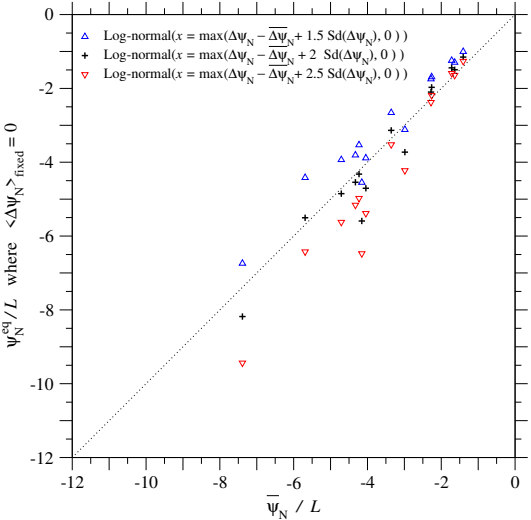
Folding free energies, $\langle \Delta G_{ND} \rangle_{\boldsymbol{\sigma}} \simeq k_B T_s \langle \Delta \psi_{ND} \rangle_{\boldsymbol{\sigma}}$, predicted by the present method are plotted against their experimental values, $\Delta G_{ND}(\boldsymbol{\sigma_N})$.

4-5. Evolutionary energy ψ_N in the mutation–fixation process of amino acid substitutions has a stable equilibrium value, because $\langle \Delta\psi_N \rangle_{\text{fixed}}$ is a decreasing function of ψ_N/L with $-2 <$ slope < 0; $\langle \Delta\Delta\psi_{ND} \rangle_{\text{fixed}} \simeq \langle \Delta\psi_N \rangle_{\text{fixed}} = 0$ at equilibrium.



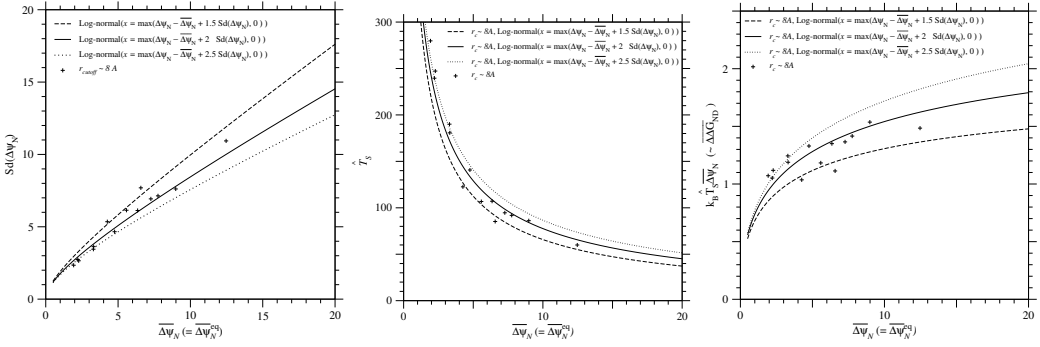
The average of $\Delta\psi_N (\simeq \Delta\Delta\psi_{ND})$ over fixed single nucleotide nonsynonymous mutations versus ψ_N/L of a wildtype for the PDZ protein family **by approximating $p(\Delta\psi_N)$ with a log-normal distribution;**

4-6. The equilibrium value (ψ_N^{eq}) of ψ_N almost agrees with the sample average ($\overline{\psi_N}$) of ψ_N over all homologous sequences.



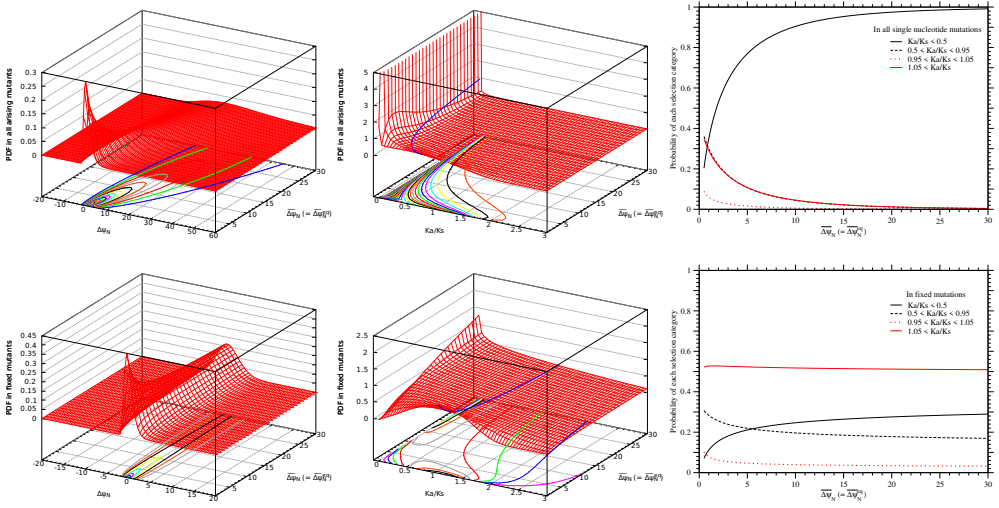
The distribution of $\Delta\psi_N$ due to single nucleotide nonsynonymous mutations **is approximated by a log-normal distribution.**

4-7. Relationships between $\overline{\Delta\psi_N}$ and $Sd(\Delta\psi_N)$, $\hat{T}_s$, and $k_B\hat{T}_s\overline{\Delta\psi_N}$ at the equilibrium state of ψ_N



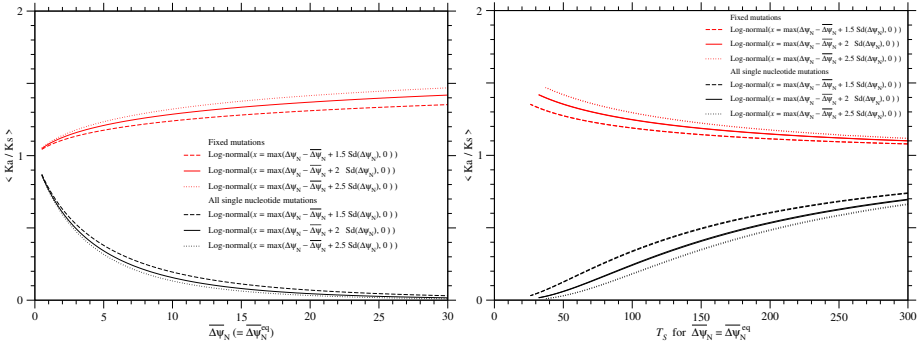
$\Delta\psi_N$ is the change of ψ_N due to single nonsynonymous nucleotide mutations.

4-8. The probability of neutral ($0.95 < K_a/K_s < 1.05$) selection category is insignificant in fixed mutations.



K_a/K_s: the ratio of the substitution rate per nonsynonymous site (K_a) to the substitution rate per synonymous site (K_s).

4-9. $\langle K_a/K_s \rangle$ 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 $\overline{\Delta\psi_N}\,(=\overline{\Delta\psi_N^{\mathbf{eq}}})$ or the effective temperature of selection, $T_s\,(=(T_s\overline{Sd}(\Delta\psi_N))_{PDZ}/Sd(\Delta\psi_N))$, at equilibrium, $\langle\Delta\psi_N\rangle_{\mathbf{fixed}}=0$.

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 ($\langle K_a/K_s \rangle$).
- 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.

References

Crow, J.F., Kimura, M., 1970.
An Introduction to population genetics theory.
Harper & Row publishers, New York.

Marks, D.S., Colwell, L.J., Sheridan, R., Hopf, T.A., Pagnani, A., Zecchina, R., Sander, C., 2011.
Protein 3D structure computed from evolutionary sequence variation.
PLoS ONE 6, e28766.
URL: http://dx.doi.org/10.1371/journal.pone.0028766, doi:10.1371/journal.pone.0028766.

Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D.S., Sander, C., Zecchina, R., Onuchic, J.N., Hwa, T., Weigt, M., 2011.
Direct-coupling analysis of residue coevolution captures native contacts across many protein families.
Proc. Natl. Acad. Sci. USA 108, E1293–E1301.
doi:10.1073/pnas.1111471108.

Pande, V.S., Grosberg, A.Y., Tanaka, T., 1997.
Statistical mechanics of simple models of protein folding and design.
Biophys. J. 73, 3192–3210.

Ramanathan, S., Shakhnovich, E., 1994.
Statistical mechanics of proteins with evolutionary selected sequences.
Phys. Rev. E 50, 1303–1312.

Shakhnovich, E.I., Gutin, A.M., 1993a.
Engineering of stable and fast-folding sequences of model proteins.
Proc. Natl. Acad. Sci. USA 90, 7195–7199.

Shakhnovich, E.I., Gutin, A.M., 1993b.
A new approach to the design of stable proteins.
Protein Eng. 6, 793–800.