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

Sanzo Miyazawa
sanzo.miyazawa@gmail.com

2017-08-26

This supplementary document includes methods, tables and figures provided in the text as part of their full descriptions for reader's convenience.

1. Methods and Materials

1.1. Knowledge of protein folding

A protein folding theory (Shakhnovich and Gutin, 1993a,b; Ramanathan and Shakhnovich, 1994; Pande et al., 1997), which is based on a random energy model (REM), indicates that the equilibrium ensemble of amino acid sequences, $\boldsymbol{\sigma} \equiv (\sigma_1, \cdots, \sigma_L)$ where σ_i is the type of amino acid at site i and L is sequence length, can be well approximated by a canonical ensemble with a Boltzmann factor consisting of the folding free energy, $\Delta G_{ND}(\boldsymbol{\sigma}, T)$ and an effective temperature T_s representing the strength of selection pressure.

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

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

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

where $p^{\mathrm{mut}}(\boldsymbol{\sigma})$ is the probability of a sequence ($\boldsymbol{\sigma}$) randomly occurring in a mutational process and depends only on the amino acid frequencies $\boldsymbol{f}(\boldsymbol{\sigma})$, k_B is the Boltzmann constant, T is a growth temperature, and G_N and G_D are the free energies of the native conformation and denatured state, respectively. Selective temperature T_s quantifies how strong the folding constraints are in protein evolution, and is specific to the protein structure and function. The free energy G_D of the denatured state does not depend on the amino acid order but the amino acid composition, $\boldsymbol{f}(\boldsymbol{\sigma})$, in a sequence (Shakhnovich and Gutin, 1993a,b; Ramanathan and Shakhnovich, 1994; Pande et al., 1997). It is reasonable to assume that mutations independently occur between sites, and therefore the equilibrium frequency

of a sequence in the mutational process is equal to the product of the equilibrium frequencies over sites; $P^{\mathrm{mut}}(\boldsymbol{\sigma}) = \prod_i p^{\mathrm{mut}}(\sigma_i)$, where $p^{\mathrm{mut}}(\sigma_i)$ is the equilibrium frequency of σ_i at site i in the mutational process.

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. That is, the partition function Z for the denatured state is written as follows with the energy density $n(E)$ of conformations that is approximated by a product of a Gaussian probability density and the total number of conformations whose logarithm is proportional to the chain length.

$$Z = \int \exp(\frac{-E}{k_B T})\, n(E) dE \tag{S.4}$$

$$n(E) \approx \exp(\omega L) \mathcal{N}(\bar{E}(\boldsymbol{f}(\boldsymbol{\sigma})), \delta E^2(\boldsymbol{f}(\boldsymbol{\sigma}))) \tag{S.5}$$

where ω is the conformational entropy per residue in the compact denatured state, and $\mathcal{N}(\bar{E}(\boldsymbol{f}(\boldsymbol{\sigma})), \delta E^2(\boldsymbol{f}(\boldsymbol{\sigma})))$ 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. The free energy of the denatured state is approximated as follows.

$$G_D(\boldsymbol{f}(\boldsymbol{\sigma}), T) \approx \bar{E}(\boldsymbol{f}(\boldsymbol{\sigma})) - \frac{\delta E^2(\boldsymbol{f}(\boldsymbol{\sigma}))}{2 k_B T} - k_B T \omega L \tag{S.6}$$

$$= \bar{E}(\boldsymbol{f}(\boldsymbol{\sigma})) - \delta E^2(\boldsymbol{f}(\boldsymbol{\sigma})) \frac{\vartheta(T/T_g)}{k_B T} \tag{S.7}$$

$$\vartheta(T/T_g) \equiv \begin{cases} \frac{1}{2}(1 + \frac{T^2}{T_g^2}) & \text{for } T > T_g \\ \frac{T}{T_g} & \text{for } T \le T_g \end{cases} \tag{S.8}$$

where $\bar{E}$ and δE^2 are estimated as the mean and variance of interaction energies of randomized sequences in the native conformation. T_g is the glass transition temperature of the protein at which entropy becomes zero (Shakhnovich and Gutin, 1993a,b; Ramanathan and Shakhnovich, 1994; Pande et al., 1997).

$$-\frac{\partial G_D}{\partial T}|_{T=T_g} = 0 \tag{S.9}$$

The conformational entropy per residue ω in the compact denatured state can be represented with T_g.

$$\omega L = \frac{\delta E^2}{2(k_B T_g)^2} \tag{S.10}$$

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.

1.2. Probability distribution of homologous sequences with the same native fold in sequence space

The probability distribution $P(\boldsymbol{\sigma})$ of homologous sequences with the same native fold, $\boldsymbol{\sigma} = (\sigma_1, \cdots, \sigma_L)$ where $\sigma_i \in \{\text{amino acids, deletion}\}$, in sequence space with maximum entropy, which satisfies a given amino acid frequency at each site and a given pairwise amino acid frequency at each site pair, is a Boltzmann distribution (Morcos et al., 2011; Marks et al., 2011).

$$P(\boldsymbol{\sigma}) \propto \exp(-\psi_N(\boldsymbol{\sigma})) \tag{S.11}$$

$$\psi_N(\boldsymbol{\sigma}) \equiv -(\sum_{i}^{L}(h_i(\sigma_i) + \sum_{j>i} J_{ij}(\sigma_i, \sigma_j))) \tag{S.12}$$

where h_i and J_{ij} are one-body (compositional) and two-body (covariational) interactions and must satisfy the following constraints.

$$\sum_{\boldsymbol{\sigma}} P(\boldsymbol{\sigma})\, \delta_{\sigma_i a_k} = P_i(a_k) \tag{S.13}$$

$$\sum_{\boldsymbol{\sigma}} P(\boldsymbol{\sigma})\, \delta_{\sigma_i a_k} \delta_{\sigma_j a_l} = P_{ij}(a_k, a_l) \tag{S.14}$$

where $\delta_{\sigma_i a_k}$ is the Kronecker delta, $P_i(a_k)$ is the frequency of amino acid a_k at site i, and $P_{ij}(a_k, a_l)$ is the frequency of amino acid pair, a_k at i and a_l at j; $a_k \in \{\text{amino acids, deletion}\}$. The pairwise interaction matrix J satisfies $J_{ij}(a_k, a_l) = J_{ji}(a_l, a_k)$ and $J_{ii}(a_k, a_l) = 0$. Interactions h_i and J_{ij} can be well estimated from a multiple sequence alignment (MSA) in the mean field approximation (Morcos et al., 2011; Marks et al., 2011), or by maximizing a pseudo-likelihood (Ekeberg et al., 2013, 2014). Because $\psi_N(\boldsymbol{\sigma})$ has been estimated under the constraints on amino acid compositions at all sites, only sequences with a given amino acid composition contribute significantly to the partition function, and other sequences may be ignored.

Hence, from Eqs. (S.2) and (S.11),

$$\psi_N(\boldsymbol{\sigma}) \simeq G_N(\boldsymbol{\sigma})/(k_B T_s) + \text{function of} \boldsymbol{f}(\boldsymbol{\sigma}) \tag{S.15}$$

$$\psi_D(\boldsymbol{f}(\boldsymbol{\sigma}), T) \simeq G_D(\boldsymbol{f}(\boldsymbol{\sigma}), T)/(k_B T_s) + \text{function of} \boldsymbol{f}(\boldsymbol{\sigma}) \tag{S.16}$$

$$\Delta\psi_{ND}(\boldsymbol{\sigma}, T) \simeq \Delta G_{ND}(\boldsymbol{\sigma}, T)/(k_B T_s) \tag{S.17}$$

$$\Delta\psi_{ND}(\boldsymbol{\sigma}, T) \equiv \psi_N(\boldsymbol{\sigma}) - \psi_D(\boldsymbol{f}(\boldsymbol{\sigma}), T) \tag{S.18}$$

$$\psi_D(\boldsymbol{f}(\boldsymbol{\sigma}), T) \approx \bar{\psi}(\boldsymbol{f}(\boldsymbol{\sigma})) - \delta\psi^2(\boldsymbol{f}(\boldsymbol{\sigma}))\vartheta(T/T_g)T_s/T \tag{S.19}$$

$$\omega = (T_s/T_g)^2 \delta\psi^2/(2L) \tag{S.20}$$

where the $\bar{\psi}$ and $\delta\psi^2$ are estimated as the mean and variance of ψ_N over randomized sequences; $\bar{E} \simeq k_B T_s \bar{\psi}$ and $\delta E^2 \simeq (k_B T_s)^2 \delta\psi^2$.

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

Here we assume that the mutational process is a reversible Markov process. That is, 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

$$P^{\text{mut}}(\boldsymbol{\mu}) M_{\boldsymbol{\mu\nu}} = P^{\text{mut}}(\boldsymbol{\nu}) M_{\boldsymbol{\nu\mu}} \tag{S.21}$$

where $P^{\mathrm{mut}}(\boldsymbol{\nu})$ is the equilibrium frequency of sequence $\boldsymbol{\nu}$ in a mutational process, $M_{\boldsymbol{\mu\nu}}$. The mutation rate per population is equal to $2NM_{\boldsymbol{\mu\nu}}$ for a diploid population, where N is the population size. The substitution rate $R_{\boldsymbol{\mu\nu}}$ from $\boldsymbol{\mu}$ to $\boldsymbol{\nu}$ is equal to the product of the mutation rate and the fixation probability with which a single mutant gene becomes to fully occupy the population (Crow and Kimura, 1970).

$$R_{\boldsymbol{\mu\nu}} = 2NM_{\boldsymbol{\mu\nu}}u(s(\boldsymbol{\mu} \to \boldsymbol{\nu})) \tag{S.22}$$

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.

For genic selection (no dominance) or gametic selection in a Wright-Fisher population of diploid, the fixation probability, u, of a single mutant gene, the selective advantage of which is equal to s and the frequency of which in a population is equal to $q_m = 1/(2N)$, was estimated (Crow and Kimura, 1970) as

$$2Nu(s) = 2N\frac{1 - e^{-4N_e s q_m}}{1 - e^{-4N_e s}} \tag{S.23}$$

$$= \frac{u(s)}{u(0)} \qquad \text{with} \qquad q_{\mathrm{m}} = \frac{1}{2N} \tag{S.24}$$

where N_e is effective population size. Eq. (S.23) will be also valid for haploid population if $2N_e$ and $2N$ are replaced by N_e and N, respectively. Also, for Moran population of haploid, $4N_e$ and $2N$ should be replaced by N_e and N, respectively. Fixation probabilities for various selection models, which are compiled from p. 192 and p. 424–427 of Crow and Kimura (1970) and from Moran (1958) and Ewens (1979), are listed in Table S.7. The selective advantage of a mutant sequence $\boldsymbol{\nu}$ to a wildtype $\boldsymbol{\mu}$ is equal to

$$s(\boldsymbol{\mu} \to \boldsymbol{\nu}) = m(\boldsymbol{\nu}) - m(\boldsymbol{\mu}) \tag{S.25}$$

where $m(\boldsymbol{\nu})$ is the Malthusian fitness of a mutant sequence, and $m(\boldsymbol{\mu})$ is for the wildtype.

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

$$P^{\mathrm{eq}}(\boldsymbol{\mu}) = \frac{P^{\mathrm{mut}}(\boldsymbol{\mu})\exp(4N_e m(\boldsymbol{\mu})(1 - q_m))}{\sum_{\boldsymbol{\nu}} P^{\mathrm{mut}}(\boldsymbol{\nu})\exp(4N_e m(\boldsymbol{\nu})(1 - q_m))} \tag{S.26}$$

because both the mutation and fixation processes satisfy the detailed balance conditions, Eq. (S.21) and the following equation, respectively.

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

$$= \frac{\exp(-4N_e m(\boldsymbol{\mu}) q_m) - \exp(-4N_e m(\boldsymbol{\nu}) q_m)}{\exp(-4N_e m(\boldsymbol{\mu})) - \exp(-4N_e m(\boldsymbol{\nu}))} \tag{S.27}$$

$$= \exp(4N_e m(\boldsymbol{\nu})(1 - q_m))\, u(s(\boldsymbol{\nu} \to \boldsymbol{\mu})) \tag{S.28}$$

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

1.4. Relationships between $m(\boldsymbol{\sigma})$, $\psi_N(\boldsymbol{\sigma})$, and $\Delta G_{ND}(\boldsymbol{\sigma})$ of protein sequence

From Eqs. (S.1), (S.11), and (S.26) , 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.

$$P^{\rm eq}(\boldsymbol{\mu}) = \frac{P^{\rm mut}(\boldsymbol{\mu})\exp(4N_e m(\boldsymbol{\mu})(1-q_m))}{\sum_\nu P^{\rm mut}(\boldsymbol{\nu})\exp(4N_e m(\boldsymbol{\nu})(1-q_m)))} \tag{S.29}$$

$$= \frac{P^{\rm mut}(\overline{\boldsymbol{\mu}})\exp(-(\psi_N(\boldsymbol{\mu})-\psi_D(\overline{\boldsymbol{f}(\boldsymbol{\mu})},T)))}{\sum_{\boldsymbol{\nu}} P^{\rm mut}(\overline{\boldsymbol{\nu}})\exp(-(\psi_N(\boldsymbol{\nu})-\psi_D(\overline{\boldsymbol{f}(\boldsymbol{\nu})},T)))} \tag{S.30}$$

$$\simeq \frac{P^{\rm mut}(\boldsymbol{\mu})\exp(-\Delta G_{ND}(\boldsymbol{\mu},T)/(k_BT_s))}{\sum_\nu P^{\rm mut}(\boldsymbol{\nu})\exp(-\Delta G_{ND}(\boldsymbol{\nu},T)/(k_BT_s))} \tag{S.31}$$

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

$$4N_e m(\boldsymbol{\mu})(1-q_m) = -\Delta\psi_{ND}(\boldsymbol{\mu},T) + \text{constant} \tag{S.32}$$

$$\simeq \frac{-\Delta G_{ND}(\boldsymbol{\mu},T)}{k_BT_s} + \text{constant} \tag{S.33}$$

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

$$4N_e s(\boldsymbol{\mu}\to\boldsymbol{\nu})(1-q_m)$$

$$= (4N_e m(\boldsymbol{\nu}) - 4N_e m(\boldsymbol{\mu}))(1-q_m) \tag{S.34}$$

$$= -(\Delta\psi_{ND}(\boldsymbol{\nu},T) - \Delta\psi_{ND}(\boldsymbol{\mu},T)) = -(\psi_N(\boldsymbol{\nu}) - \psi_N(\boldsymbol{\mu})) \tag{S.35}$$

$$\simeq -(\Delta G_{ND}(\boldsymbol{\nu},T) - \Delta G_{ND}(\boldsymbol{\mu},T))/(k_BT_s) = -(G_N(\boldsymbol{\nu}) - G_N(\boldsymbol{\mu}))/(k_BT_s) \tag{S.36}$$

It should be noted here that only sequences for which $f(\boldsymbol{\sigma}) = \overline{f(\boldsymbol{\sigma})}$ contribute significantly to the partition functions in Eq. (S.30), and other sequences may be ignored.

Eq. (S.35) indicates that evolutionary statistical energy ψ should be proportional to effective population size N_e, and therefore it is ideal to estimate one-body (h) and two-body (J) interactions from homologous sequences of species that do not significantly differ in effective population size. Also, Eq. (S.36) indicates that selective temperature T_s is inversely proportional to the effective population size N_e; $T_s \propto 1/N_e$, because free energy is a physical quantity and should not depend on effective population size.

1.5. The ensemble average of folding free energy, $\Delta G_{ND}(\boldsymbol{\sigma},T)$, over sequences

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

$$\langle\Delta G_{ND}(\boldsymbol{\sigma},T)\rangle_{\boldsymbol{\sigma}} \tag{S.37}$$

$$\equiv \Big[\sum_{\boldsymbol{\sigma}} \Delta G_{ND}(\boldsymbol{\sigma},T)P^{\rm mut}(\boldsymbol{\sigma})\exp(-\frac{\Delta G_{ND}(\boldsymbol{\sigma},T)}{k_BT_s})\Big] / \Big[\sum_{\boldsymbol{\sigma}} P^{\rm mut}(\boldsymbol{\sigma})\exp(-\frac{\Delta G_{ND}(\boldsymbol{\sigma},T)}{k_BT_s})\Big] \tag{S.38}$$

$$\approx \Big[\sum_{\boldsymbol{\sigma}\,|\,\boldsymbol{f}(\boldsymbol{\sigma})=\overline{\boldsymbol{f}(\boldsymbol{\sigma}_N)}} G_N(\boldsymbol{\sigma})\exp(-\frac{G_N(\boldsymbol{\sigma})}{k_BT_s})\Big] / \Big[\sum_{\boldsymbol{\sigma}\,|\,\boldsymbol{f}(\boldsymbol{\sigma})=\overline{\boldsymbol{f}(\boldsymbol{\sigma}_N)}} \exp(-\frac{G_N(\boldsymbol{\sigma})}{k_BT_s})\Big] - G_D(\overline{\boldsymbol{f}(\boldsymbol{\sigma}_N)},T) \tag{S.39}$$

$$= \langle G_N(\boldsymbol{\sigma})\rangle_{\boldsymbol{\sigma}} - G_D(\overline{\boldsymbol{f}(\boldsymbol{\sigma}_N)},T) \tag{S.40}$$

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. In Eq. (S.39), the sum over all sequences is approximated by the sum over sequences the amino acid composition of which is the same as that over the natural sequences.

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

$$\langle G_N(\boldsymbol{\sigma})\rangle_{\boldsymbol{\sigma}} \approx \frac{\int E \exp(-E/(k_B T_s))\, n(E)\, dE}{\int \exp(-E/(k_B T_s))\, n(E)\, dE} \tag{S.41}$$

$$= \bar{E}(\overline{\boldsymbol{f}(\boldsymbol{\sigma_N})}) - \delta E^2(\overline{\boldsymbol{f}(\boldsymbol{\sigma_N})})/(k_B T_s) \tag{S.42}$$

$$\langle \psi_N(\boldsymbol{\sigma})\rangle_{\boldsymbol{\sigma}} \equiv [\sum_{\boldsymbol{\sigma}} \psi_{ND}(\boldsymbol{\sigma}) \exp(-\psi_N(\boldsymbol{\sigma}))\,] \,/\, [\sum_{\boldsymbol{\sigma}} \exp(-\psi_N(\boldsymbol{\sigma}))\,] \tag{S.43}$$

$$\approx \bar{\psi}(\overline{\boldsymbol{f}(\boldsymbol{\sigma_N})}) - \delta\psi^2(\overline{\boldsymbol{f}(\boldsymbol{\sigma_N})}) \tag{S.44}$$

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.

$$\overline{\Delta G_{ND}(\boldsymbol{\sigma_N}, T)}/(k_B T_s) = \langle \Delta G_{ND}(\boldsymbol{\sigma}, T)\rangle_{\boldsymbol{\sigma}}/(k_B T_s) \tag{S.45}$$

$$\approx [\,\delta E^2(\overline{\boldsymbol{f}(\boldsymbol{\sigma_N})})\,]\,[\,\vartheta(T/T_g)T_s/T - 1\,]/(k_B T_s)^2 \tag{S.46}$$

$$= \delta\psi^2(\overline{\boldsymbol{f}(\boldsymbol{\sigma_N})})\,[\,\vartheta(T/T_g)T_s/T - 1\,] \tag{S.47}$$

$$= \overline{\Delta G_{ND}(\boldsymbol{\sigma_N}, T_g)}\,/\,(k_B T_s') \tag{S.48}$$

$$T_s' = T_s(T_s/T - 1)/(\vartheta(T/T_g)T_s/T - 1) \tag{S.49}$$

$$\overline{\psi_N(\boldsymbol{\sigma_N})} \equiv \frac{\sum_{\boldsymbol{\sigma}_N} w_{\boldsymbol{\sigma}_N} \psi_N(\boldsymbol{\sigma}_N)}{\sum_{\boldsymbol{\sigma}_N} w_{\boldsymbol{\sigma}_N}} \tag{S.50}$$

$$= \langle \psi_N(\boldsymbol{\sigma})\rangle_{\boldsymbol{\sigma}} \tag{S.51}$$

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. $\Delta G_{ND}(\boldsymbol{\sigma_N}, T_g)$ corresponds to the energy gap (Shakhnovich and Gutin, 1993b) between the native and the glass states, and T_s' will be the selective temperature if $\Delta G_{ND}(\boldsymbol{\sigma_N}, T_g)$ is used for selection instead of $\Delta G_{ND}(\boldsymbol{\sigma_N}, T)$.

The folding free energy becomes equal to zero at the melting temperature T_m; $\langle \Delta G_{ND}(\boldsymbol{\sigma_N}, T_m)\rangle_{\boldsymbol{\sigma}} = 0$. Thus, the following relationship must be satisfied (Shakhnovich and Gutin, 1993a,b; Ramanathan and Shakhnovich, 1994; Pande et al., 1997).

$$\vartheta(T_m/T_g)\frac{T_s}{T_m} = \frac{T_s}{2T_m}(1 + \frac{T_m^2}{T_g^2}) = 1 \quad \text{with } T_s \le T_g \le T_m \tag{S.52}$$

1.6. Estimation of $\bar{\psi}(\boldsymbol{f}(\boldsymbol{\sigma}))$ and $\delta\psi^2(\boldsymbol{f}(\boldsymbol{\sigma}))$

The mean $\bar{\psi}(\boldsymbol{f}(\boldsymbol{\sigma}))$ and the variance $\delta\psi^2(\boldsymbol{f}(\boldsymbol{\sigma}))$ in the Gaussian approximation for the distribution of conformational energies at the denatured state are estimated as the mean and variance of ψ_N of random sequences in the

native conformation (Pande et al., 1997).

$$\bar{\psi}(\boldsymbol{f}(\boldsymbol{\sigma})) \quad = \quad -\sum_i [\hat{h}_i(::) + \sum_{j>i} \hat{J}_{ij}(::,::)] \tag{S.53}$$

where $\hat{h}_i(::)$ and $\hat{J}_{ij}(::,::)$ are the means of one-body and two-body interactions in random sequences.

$$\hat{h}_i(::) \quad \equiv \quad \sum_k \hat{h}_i(a_k) f_{a_k}(\boldsymbol{\sigma}) \tag{S.54}$$

$$\hat{J}_{ij}(::,::) \quad \equiv \quad \sum_k \sum_l \hat{J}_{ij}(a_k, a_l) f_{a_k}(\boldsymbol{\sigma}) f_{a_l}(\boldsymbol{\sigma}) \tag{S.55}$$

where $f_{a_k}(\boldsymbol{\sigma})$ is the composition of amino acid a_k in the sequence $\boldsymbol{\sigma}$.

$$f_{a_k}(\boldsymbol{\sigma}) \quad = \quad \frac{1}{L} \sum_{i=1}^{L} \delta_{\sigma_i a_k} \tag{S.56}$$

where $\delta_{\sigma_i a_k}$ is the Kronecker delta. The variance, $\delta\psi^2(\boldsymbol{f}(\boldsymbol{\sigma}))$, is

$$\delta\psi^2(\boldsymbol{f}(\boldsymbol{\sigma})) \quad = \quad \sum_k f_{a_k}(\boldsymbol{\sigma}) \sum_i [\, \delta\hat{h}_i(a_k)^2 + \sum_{j \neq i} \{\, 2\delta\hat{h}_i(a_k)\delta\hat{J}_{ij}(a_k, ::) \tag{S.57}$$

$$+ \quad \sum_{m \neq \{i,j\}} \delta\hat{J}_{ij}(a_k, ::)\delta\hat{J}_{im}(a_k, ::) + \frac{1}{2} \sum_l \delta\hat{J}_{ij}(a_k, a_l)^2 f_{a_l}(\boldsymbol{\sigma}) \,\} \,] \tag{S.58}$$

where

$$\delta\hat{h}_i(a_k) \quad \equiv \quad \hat{h}_i(a_k) - \hat{h}_i(::) \tag{S.59}$$

$$\delta\hat{J}_{ij}(a_k, ::) \quad \equiv \quad \hat{J}_{ij}(a_k, ::) - \hat{J}_{ij}(::,::) \tag{S.60}$$

$$\delta\hat{J}_{ij}(a_k, a_l) \quad \equiv \quad \hat{J}_{ij}(a_k, a_l) - \hat{J}_{ij}(::,::) \tag{S.61}$$

1.7. Estimation of one-body (h) and pairwise (J) interactions

The estimates of h and J (Morcos et al., 2011; Marks et al., 2011) are noisy as a result of estimating many interaction parameters from a relatively small number of sequences. Therefore, only pairwise interactions within a certain distance are taken into account; the estimate of J is modified as follows, according to Morcos et al. (Morcos et al., 2014).

$$\hat{J}^{\mathrm{q}}_{ij}(a_k, a_l) \quad = \quad J^{\mathrm{q}}_{ij}(a_k, a_l) H(r_{\mathrm{cutoff}} - r_{ij}) \tag{S.62}$$

where $\hat{J}^{\mathrm{q}}$ is the statistical estimate of J in the mean field approximation in which the amino acid a_q is the reference state, H is the Heaviside step function, and r_{ij} is the distance between the centers of amino acid side chains in

protein structure, and $r_{\rm cutoff}$ is a distance threshold for residue pairwise interactions. Maximum interaction ranges employed for pairwise interactions are $r_{\rm cutoff} \sim 8$ and 15.5 Å , which correspond to the first and second interaction shells between residues, respectively. Here it should be noticed that the total interaction $\psi_N(\boldsymbol{\sigma})$ defined by Eq. (S.12) does not depend on any gauge unless the interaction range for pairwise interactions is limited, but a gauge conversion in which interconversions between h and J occur must not be done before calculating $\hat{J}$, because it may change the estimate of $\psi_N(\boldsymbol{\sigma})$ in the present scheme of Eq. (S.62) in which pairwise interactions are cut off at a certain distance. Thus, a natural gauge must be used before calculating $\hat{J}$.

For example, let us think about the Ising gauge (Ekeberg et al., 2014), in which h^I and J^I can be calculated from h^g and J^g in any gauge through the following conversions.

$$J^{\rm I}_{ij}(a_k, a_l) = J^{\rm g}_{ij}(a_k, a_l) - J^{\rm g}_{ij}(a_k, :) - J^{\rm g}_{ij}(:, a_l) + J^{\rm g}_{ij}(:, :) \tag{S.63}$$

$$h^{\rm I}(a_k) = h^{\rm g}_i(a_k) - h^{\rm g}_i(:) + \sum_{j \neq i} (J^{\rm g}_{ij}(a_k, :) - J^{\rm g}_{ij}(:, :)) \tag{S.64}$$

where

$$h_i(:) \equiv \frac{1}{q} \sum_{k=1}^{q} h_i(a_k) \tag{S.65}$$

$$J_{ij}(:, :) \equiv \frac{1}{q^2} \sum_{k=1}^{q} \sum_{l=1}^{q} J_{ij}(a_k, a_l) \tag{S.66}$$

where q is equal to the total number of amino acid types including deletion, that is, $q = 21$. Thus, the gauge conversion of $\hat{J}$ does not affect the total interaction $\psi_N(\boldsymbol{\sigma})$ but the gauge conversion before calculating $\hat{J}$ may significantly change the total interaction.

In the DCA (Morcos et al., 2011; Marks et al., 2011), the interaction terms are estimated in the mean field approximation as follows.

$$J^{\rm q}_{ij}(a_k, a_l) = -(C^{-1})_{ij}(a_k, a_l) \tag{S.67}$$

$$J^{\rm q}_{ij}(a_q, a_l) = J^{\rm q}_{ij}(a_k, a_q) = J^{\rm q}_{ij}(a_q, a_q) = 0 \tag{S.68}$$

where $i \neq j$ and $1 \leq k, l \leq q - 1$, and the covariance matrix C is defined as

$$C_{ij}(a_k, a_l) \equiv P_{ij}(a_k, a_l) - P_i(a_k) P_j(a_l) \tag{S.69}$$

Here, one (a_q) of the amino acid types including deletion is used as the reference state; $J^{\rm q}$ denotes the J in this gauge, which is called the q gauge here. According to Morcos et al. (Morcos et al., 2011), the probability $P_i(a_k)$ of amino acid a_k at site i and the joint probabilities $P_{ij}(a_k, a_l)$ of amino acids, a_k at site i and a_l at site j, are evaluated by

$$P_i(a_k) = (1 - p_c) f_i(a_k) + p_c \frac{1}{q} \tag{S.70}$$

$$P_{ij}(a_k, a_l) = (1 - p_c) f_{ij}(a_k, a_l) + p_c \frac{1}{q^2} \qquad \text{for} \quad i \neq j \tag{S.71}$$

$$P_{ii}(a_k, a_l) = P_i(a_k) \delta_{a_k a_l} \tag{S.72}$$

where $0 \leq p_c \leq 1$ is the ratio of pseudocount, and $f_i(a_k)$ is the frequency of amino acid a_k at site i and $f_{ij}(a_k, a_l)$ is the frequency of the site pair, a_k at i and a_l at j, in an alignment; $f_{ii}(a_k, a_l)$ is defined as $f_{ii}(a_k, a_l) = f_i(a_k)\delta_{a_k a_l}$.

In the mean field approximation, one body interactions $h_i^q(a_k)$ in the q gauge are estimated by $\hat{h}_i^q(a_k) = \log(P_i(a_k)/P_i(a_q)) - \sum_{j\neq i}\sum_{l\neq q}\hat{J}_{ij}^q(a_k, a_l)P_j(a_l)$. Here, instead the one body interactions $h_i(a_k)$ are estimated in the isolated two-state model (Morcos et al., 2011), that is,

$$P_i(a_k) \propto \exp\left[\, h_{ij}^{\mathrm{q}}(a_k) + J_{ij}^{\mathrm{q}}(a_k, a_l) + h_{ji}^{\mathrm{q}}(a_l)\,\right] \tag{S.73}$$

$$\hat{h}_i^{\mathrm{q}}(a_k) = \frac{1}{L-1}\sum_{j\neq i} h_{ij}^{\mathrm{q}}(a_k) \tag{S.74}$$

These $\hat{h}^q$ and $\hat{J}^q$ in the q gauge are converted to a new gauge, which is called the zero-sum gauge here,

$$\hat{h}_i^{\mathrm{s}}(a_k) = \hat{h}_i^{\mathrm{q}}(a_k) - \hat{h}_i^{\mathrm{q}}(:) \tag{S.75}$$

$$\hat{J}_{ij}^{\mathrm{s}}(a_k, a_l) = \hat{J}_{ij}^{\mathrm{q}}(a_k, a_l) - \hat{J}_{ij}^{\mathrm{q}}(:,:) \tag{S.76}$$

In this gauge, the reference state is the average state over amino acids including deletion, instead of a specific amino acid (a_q) in the q gauge.

1.8. Distribution of $\Delta\Delta\psi_{ND} \simeq \Delta\psi_N$ due to single nucleotide nonsynonymous substitutions

The probability density function (PDF) of $\Delta\Delta\psi_{ND}$, $p(\Delta\Delta\psi_{ND})$, due to single nucleotide nonsynonymous substitutions is approximated by the PDF of $\Delta\psi_N$, $p(\Delta\psi_N)$, because $\Delta\psi_D \simeq 0$ for single amino acid substitutions.

$$\Delta\Delta\psi_{ND} \simeq \Delta\psi_N \tag{S.77}$$

$$p(\Delta\Delta\psi_{ND}) \simeq p(\Delta\psi_N) \tag{S.78}$$

for single nucleotide nonsynonymous substitutions.

For simplicity, a log-normal distribution, $\ln \mathcal{N}(x; \mu, \sigma)$, for which x, μ and σ are defined as follows, is arbitrarily employed here to reproduce observed PDFs of $\Delta\psi_N$, particularly in the domain of $\Delta\psi_N < \overline{\Delta\psi_N}$, although other distributions such as inverse Γ distributions can equally reproduce the observed ones, too.

$$p(\Delta\psi_N) \approx \ln \mathcal{N}(x; \mu, \sigma) \equiv \frac{1}{x}\mathcal{N}(\ln x; \mu, \sigma) \tag{S.79}$$

$$x \equiv \max(\Delta\psi_N - \Delta\psi_N^{\mathrm{o}}, 0) \tag{S.80}$$

$$\exp(\mu + \sigma^2/2) = \overline{\Delta\psi_N} - \Delta\psi_N^{\mathrm{o}} \tag{S.81}$$

$$\exp(2\mu + \sigma^2)(\exp(\sigma^2) - 1) = \overline{(\Delta\psi_N - \overline{\Delta\psi_N})^2)} \tag{S.82}$$

$$\Delta\psi_N^{\mathrm{o}} \equiv \min(\overline{\Delta\psi_N} - n_{\mathrm{shift}}\overline{(\Delta\psi_N - \overline{\Delta\psi_N})^2)}^{1/2}, 0) \tag{S.83}$$

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.

1.9. Probability distributions of selective advantage, fixation rate and K_a/K_s

Now, we can consider the probability distributions of characteristic quantities that describe the evolution of genes. First of all, the probability density function (PDF) of selective advantage s, $p(s)$, of mutant genes can be calculated from the PDF of the change of $\Delta\psi_{ND}$ due to a mutation from $\boldsymbol{\mu}$ to $\boldsymbol{\nu}$, $\Delta\Delta\psi_{ND}(\equiv \Delta\psi_{ND}(\boldsymbol{\nu}, T) - \Delta\psi_{ND}(\boldsymbol{\mu}, T))$. The PDF of $4N_e s$, $p(4N_e s) = p(s)/(4Ne)$, may be more useful than $p(s)$.

$$p(4N_e s) \quad = \quad p(\Delta\Delta\psi_{ND}) \left| \frac{d\Delta\Delta\psi_{ND}}{d4N_e s} \right| = p(\Delta\Delta\psi_{ND})(1 - q_m) \tag{S.84}$$

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. (S.35).

The PDF of fixation probability u can be represented by

$$p(u) \quad = \quad p(4N_e s) \frac{d4N_e s}{du} = p(4N_e s) \frac{(e^{4N_e s} - 1)^2 e^{4N_e s(q_m - 1)}}{q_m(e^{4N_e s} - 1) - (e^{4N_e s q_m} - 1)} \tag{S.85}$$

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

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 synonymous substitutions with s = 0 is

$$\frac{K_a}{K_s} \quad = \quad \frac{u(s)}{u(0)} = \frac{u(s)}{q_m} \tag{S.86}$$

assuming that synonymous substitutions are completely neutral and mutation rates at both types of sites are the same. The PDF of K_a/K_s is

$$p(K_a/K_s) \quad = \quad p(u) \frac{du}{d(K_a/K_s)} = p(u)\, q_m \tag{S.87}$$

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

Now, let us consider fixed mutant genes. The PDF of $\Delta\Delta\psi_{ND}$ in fixed mutants is proportional to that multiplied by the fixation probability.

$$p(\Delta\Delta\psi_{ND,\text{fixed}}) \quad = \quad p(\Delta\Delta\psi_{ND}) \frac{u(s(\Delta\Delta\psi_{ND}))}{\langle u(s(\Delta\Delta\psi_{ND})) \rangle} \tag{S.88}$$

$$\langle u \rangle \quad \equiv \quad \int_{-\infty}^{\infty} u(s) p(\Delta\Delta\psi_{ND}) d\Delta\Delta\psi_{ND} \tag{S.89}$$

Likewise, the PDF of selective advantage in fixed mutants is

$$p(4N_e s_{\text{fixed}}) \quad = \quad p(4N_e s) \frac{u(s)}{\langle u(s) \rangle} \tag{S.90}$$

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

$$p(u_{\text{fixed}}) \quad = \quad p(u) \frac{u}{\langle u \rangle} \tag{S.91}$$

$$p((\frac{K_a}{K_s})_{\text{fixed}}) \quad = \quad p(\frac{K_a}{K_s}) \frac{u}{\langle u \rangle} = p(\frac{K_a}{K_s}) \frac{\frac{K_a}{K_s}}{\langle \frac{K_a}{K_s} \rangle} \tag{S.92}$$

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.

$$\langle \frac{K_a}{K_s} \rangle_{\text{fixed}} = \langle (\frac{K_a}{K_s})^2 \rangle / \langle \frac{K_a}{K_s} \rangle \tag{S.93}$$

1.11. Sequence data

We study the single domains of 8 Pfam (Finn et al., 2016) families and both the single domains and multi-domains from 3 Pfam families. In Table S.1, their Pfam ID for a multiple sequence alignment, and UniProt ID and PDB ID with the starting- and ending-residue positions of the domains are listed. The full alignments for their families at the Pfam are used to estimate one-body interactions h and pairwise interactions J with the DCA program from "http://dca.rice.edu/portal/dca/home" (Marks et al., 2011; Morcos et al., 2011). To estimate the sample ($\overline{\psi_N}$) and ensemble ($\langle \psi_N \rangle_{\boldsymbol{\sigma}}$) averages of the evolutionary statistical energy, M unique sequences with no deletions are used. In order to reduce phylogenetic biases in the set of homologous sequences, we employ a sample weight ($w_{\boldsymbol{\sigma}_N}$) for each sequence, which is equal to the inverse of the number of sequences that are less than 20% different from a given sequence in a given set of homologous sequences. Only representatives of unique sequences with no deletions, which are at least 20% different from each other, are used to calculate the changes of the evolutionary statistical energy ($\Delta\psi_N$) due to single nucleotide nonsynonymous substitutions; the number of the representatives is almost equal to the effective number of sequences (M_{eff}) in Table S.1.

1.12. Estimation of effective temperature T_s for selection

We have examined the changes of ψ_N due to single nucleotide nonsynonymous substitutions over all sites in the homologous sequences of 14 protein families, and have found the following regression equation.

$$\overline{\Delta\psi_N} \approx \alpha_{\psi_N} \frac{\psi_N - \overline{\psi_N}}{L} + \overline{\overline{\Delta\psi_N}} \qquad \text{with } \alpha_{\psi_N} < 0 \tag{S.94}$$

with correlation coefficients, $r_{\psi_N} > 0.9$, where L is sequence length, $\overline{\psi_N}$ denotes the average of ψ_N over all homologous sequences, and $\overline{\Delta\psi_N}$ and $\overline{\overline{\Delta\psi_N}}$ denote the average of $\Delta\psi_N$ over all single nucleotide synonymous substitutions at all sites in a protein sequence and its total average over all homologous sequences in a protein family, respectively. In addition, the following relationship for the standard deviation of $\Delta\psi_N$ has been found.

$$\begin{aligned} \text{Sd}(\Delta\psi_N) &\approx \text{independent of } \psi_N \text{ and} \\ &\quad \text{constant across homologous sequences in every protein family} && \text{(S.95)} \\ &= \text{function of } k_B T_s && \text{(S.96)} \end{aligned}$$

Because

$$\text{Sd}(\Delta G_N) = \text{function that must not explicitly depend on } k_B T_s \text{ but } G_N \tag{S.97}$$

the following important relationship, which can be used to estimate the relative value of T_s, is derived.

$$\begin{aligned} \mathrm{Sd}(\Delta G_N) &\simeq k_B T_s \mathrm{Sd}(\Delta\psi_N) \\ &\approx \text{constant} \end{aligned} \tag{S.98}$$

where $\mathrm{Sd}(\Delta G_N)$ and $\mathrm{Sd}(\Delta\psi_N)$ are the standard deviations of ΔG_N and $\Delta\psi_N$ over all single nucleotide nonsynonymous substitutions at all sites, respectively. These relationships, Eqs. S.94 and S.98, are shown in Figs. S.3 to S.13 , and the regression coefficients (α_{ψ_N}) and correlation coefficients (r_{ψ_N}) are listed in Tables S.2 and S.5.

The PDZ family is employed here as a reference protein for T_s, and its T_s is estimated by a direct comparison of $\Delta\psi_N$ and experimental $\Delta\Delta G_{ND}$; the amino acid pair types and site locations of single amino acid substitutions are the most various, and also the correlation between the experimental $\Delta\Delta G_{ND}$ and $\Delta\psi_N$ is the best for the PDZ family in the present set of protein families, SH3_1 (Grantcharova et al., 1998), ACBP (Kragelund et al., 1999), PDZ (Gianni et al., 2005, 2007), and Copper-bind (Wilson and Wittung-Stafshede, 2005); see Tables S.3 and S.6.

$$k_B\hat{T}_s = k_B\hat{T}_{s,\,\mathrm{PDZ}}\,[\,\overline{\mathrm{Sd}(\Delta\psi_{\mathrm{PDZ}})}\,/\,\overline{\mathrm{Sd}(\Delta\psi_N)}\,] \tag{S.99}$$

where the overline denotes the average over all homologous sequences. Here, the averages of standard deviations over all homologous sequences are employed, because T_s for all homologous sequences are approximated to be equal. With estimated T_s and experimental melting temperature T_m, glass transition temperature T_g and folding free energy ΔG_N have been estimated for each protein family on the basis of the REM. The estimates of T_s and T_g are all within a reasonable range, and the estimated values of ΔG_N agree well with their experimental values for 5 protein families, justifying the estimates of T_s.

1.13. Comparison of results between $r_{cutoff} \sim 8$ and 15.5 Å

In order to determine T_s for a reference protein, the experimental values (Gianni et al., 2007) of $\Delta\Delta G_{ND}$ due to single amino acid substitutions in the PDZ domain are plotted against the changes of interaction, $\Delta\psi_N$ for the same types of substitutions in Fig. S.14 for $r_{\mathrm{cutoff}} \sim 8$ and 15.5 Å . The slopes of the least-squares regression lines through the origin, which are estimates of k_BT_s, are equal to $k_B\hat{T}_s = 0.279$ kcal/mol for $r_{\mathrm{cutoff}} \sim 8$ Å and $k_B\hat{T}_s = 0.162$ kcal/mol for $r_{\mathrm{cutoff}} \sim 15.5$ Å , and the reflective correlation coefficients are equal to 0.93 and 0.94, respectively. These estimates of k_BT_s for the PDZ yield $\overline{\mathrm{Sd}(\Delta\Delta G_{ND})} \simeq k_B\hat{T}_s\overline{\mathrm{Sd}(\Delta\psi_N)} = 1.30$ kcal/mol for $r_{\mathrm{cutoff}} \sim 8$ Å , and 1.29 kcal/mol for $r_{\mathrm{cutoff}} \sim 15.5$Å . The reason why $k_B\hat{T}_s\overline{\mathrm{Sd}(\Delta\psi_N)}$ for the PDZ takes similar values for both $r_{\mathrm{cutoff}} \sim 8$ and 15.5Å is that the correlation between the experimental $\Delta\Delta G_{ND}$ and $\Delta\psi_N$ is very good, and the slopes of the regression lines are very close to those of the reflective regression lines through the origin, 0.25 for $r_{\mathrm{cutoff}} \sim 8$ Å , and 0.16 for 15.5Å ; $k_BT_{s,PDZ} = \langle\Delta\Delta G_{ND}\Delta\psi_N\rangle/\langle(\Delta\psi_N)^2\rangle \simeq \langle(\Delta\Delta G_{ND} - \overline{\Delta\Delta G_{ND}})(\Delta\psi_N - \overline{\Delta\psi_N})\rangle/\langle(\Delta\psi_N - \overline{\Delta\psi_N})^2\rangle$, and so $k_B(\hat{T}_s\overline{Sd}(\Delta\psi_N))_{\mathrm{PDZ}} \simeq$ constant for $r_{\mathrm{cutoff}} \sim 8$ and 15.5 Å . In other words, as long as the correlation between the experimental $\Delta\Delta G_{ND}$ and $\Delta\psi_N$ is good, $k_B\hat{T}_s\overline{\mathrm{Sd}}(\Delta\psi_N) \simeq$ constant irrespective of the cutoff value r_{cutoff}, although the estimate of T_s differs depending on $\mathrm{Sd}(\Delta\psi_N)$. This indicates that the correlation between experimental $\Delta\Delta G_{ND}$ and $\Delta\psi_N$ cannot be a good measure for the correctness of estimated $\Delta\psi_N$, although it must be good enough. Other comparisons are needed to judge which estimation of T_s is better.

The estimate of $\overline{\mathrm{Sd}(\Delta\Delta G_{ND})} = 1.30$ or 1.29 kcal/mol corresponds to 76% of 1.7 kcal/mol (Serohijos et al., 2012) estimated from ProTherm database or 79–80% of 1.63 kcal/mol (Tokuriki et al., 2007) computationally

predicted for single nucleotide mutations by using the FoldX. Using $\overline{\mathrm{Sd}(\Delta\Delta G_{ND})} = 1.30$ or 1.29 kcal/mol estimated from T_s for PDZ, the absolute values of T_s for other proteins are calculated by Eq. (S.99) and listed in Tables S.3 and S.6. Fig. S.15 shows that both $r_{\mathrm{cutoff}} \sim 8$ and 15.5Å yield similar values for $\hat{T}_s$ in a scale relative to the $\hat{T}_s$ of the PDZ, because $T_s/T_{s,PDZ} = \overline{\mathrm{Sd}}(\Delta\psi_{N,PDZ})/\overline{\mathrm{Sd}}(\Delta\psi_N)$. In other words, the differences of the absolute values of $\hat{T}_s$ between $r_{\mathrm{cutoff}} \sim 8$ and 15.5Å as shown in Fig. S.15 primarily originate in the difference of $\hat{T}_{s,PDZ}$ for the PDZ.

Larger the standard deviation of $\Delta\psi_N$ is, the smaller the estimate of T_s is. Including the longer range of pairwise interactions tend to increase the variance of $\Delta\psi_N$. The range of interactions should be limited to a realistic value, either the first interaction shell or the second interaction shell. Thus, the estimates of T_s with $r_{\mathrm{cutoff}} \sim 8$ Å and 15.5Å are upper and lower limits, respectively. Morcos et al. (Morcos et al., 2014) estimated T_s by comparing $\Delta\psi_{ND}$ with ΔG_{ND} estimated by the associative-memory, water-mediated, structure, and energy model (AWSEM). They estimated ψ_N with $r_{\mathrm{cutoff}} = 16$ Å and probably $p_c = 0.5$. In Fig. S.16, the present estimates of T_s are compared with those by Morcos et al. (Morcos et al., 2014). The Morcos's estimates of T_s with some exceptions tend to be located between the present estimates with $r_{\mathrm{cutoff}} \sim 8$ Å and 15.5Å , which correspond to the upper and lower limits for T_s.

In Figs. S.18, S.19, S.20, and S.21 , and Figs. S.22, S.23, S.26, S.27, S.28, and S.32 , various results are compared between $r_{\mathrm{cutoff}} \sim 8$ Å and 15.5Å .

References

Armengaud, J., Urbonavicius, J., Fernandez, B., Chaussinand, G., Bujnicki, J. M., Grosjean, H., 2004. n^2-methylation of guanosine at position 10 in tRNA is catalyzed by a THUMP domain-containing, S-adenosylmethionine-dependent methyltransferase, conserved in archaea and eukaryota. J. Biol. Chem. 279, 37142–37152.

Barton, J. P., Leonardis, E. D., Coucke, A., Cocco, S., 2016. ACE: adaptive cluster expansion for maximum entropy graphical model inference. Bioinformatics 32, 3089–3097.

Bryngelson, J. D., Wolynes, P. G., 1987. Spin glasses and the statistical mechanics of protein folding. Proc. Natl. Acad. Sci. USA 84, 7524–7528.

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

D'Auria, S., Scirè, A., Varriale, A., Scognamiglio, V., Staiano, M., Ausili, A., Marabotti, A., MosèRossi, Tanfani, F., 2005. Binding of glutamine to glutamine-binding protein from *Escherichia coli* induces changes in protein structure and increases protein stability. Proteins 58, 80–87.

Davtyan, A., Schafer, N. P., Zheng, W., Clementi, C., Wolynes, P. G., , Papoian, G. A., 2012. AWSEM-MD: Protein structure prediction using coarse-grained physical potentials and bioinformatically based local structure biasing. J. Phys. Chem. B 116, 8494–8503.

Dokholyan, N. V., Shakhnovich, E. I., 2001. Understanding hierarchical protein evolution from first principles. J. Mol. Biol. 312, 289–307.

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

Ekeberg, M., Hartonen, T., Aurell, E., 2014. Fast pseudolikelihood maximization for direct-coupling analysis of protein structure from many homologous amino-acid sequences. J. Comput. Phys. 276, 341–356.

Ekeberg, M., Lövkvist, C., Lan, Y., Weigt, M., Aurell, E., 2013. Improved contact prediction in proteins: Using pseudolikelihoods to infer potts models. Phys. Rev. E 87, 012707–1–16.
URL `http://link.aps.org/doi/10.1103/PhysRevE.87.012707`

Ewens, W. J., 1979. Mathematical Population Genetics. Springer, New York.

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., Potter, S. C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G. A., Tate, J., Bateman, A., 2016. The Pfam protein families database: towards a more sustainable future. Nucl. Acid Res. 44, D279–D285.

Ganguly, T., Das, M., Bandhu, A., Chanda, P. K., Jana, B., Mondal, R., Sau, S., 2009. Physicochemical properties and distinct DNA binding capacity of the repressor of temperate *Staphylococcus aureus* phage ϕ/11. FEBS J. 276, 1975–1985.

Gianni, S., Calosci1, N., Aelen, J. M. A., Vuister, G. W., Brunori, M., Travaglini-Allocatelli, C., 2005. Kinetic folding mechanism of PDZ2 from PTP-BL. Protein Eng. Design Selection 18, 389–395.

Gianni, S., Geierhaas, C. D., Calosci, N., Jemth, P., Vuister, G. W., Travaglini-Allocatelli, C., Vendruscolo, M., Brunori, M., 2007. A PDZ domain recapitulates a unifying mechanism for protein folding. Proc. Natl. Acad. Sci. USA 104, 128–133.

Gillespie, J. H., 1991. The Causes of Molecular Evolution. Oxford Univ. Press, Oxford.

Grantcharova, V. P., Riddle, D. S., Santiago, J. V., Baker, D., 1998. Important role of hydrogen bonds in the structurally polarized transition state for folding of the src SH3 domain. Nature Struct. Biol, 714–720.

Guelorget, A., Roovers, M., Guérineau, V., Barbey, C., Li, X., , Golinelli-Pimpaneau, B., 2010. Insights into the hyperthermostability and unusual region-specificity of archaeal *Pyrococcus abyssi* tRNA m^1A57/58 methyltransferase. Nucl. Acid Res. 38, 6206–6218.

Hopf, T. A., Colwell, L. J., Sheridan, R., Rost, B., Sander, C., Marks, D. S., 2012. Three-dimensional structures of membrane proteins from genomic sequencing. Cell 149, 1607–1621.

Hopf, T. A., Ingraham, J. B., Poelwijk, F. J., Schärfe, C. P. I., Springer, M., Sander, C., Marks, D. S., 2017. Mutation effects predicted from sequence co-variation. Nature Biotech. 35, 128–135.

Jacquin, H., Gilson, A., Shakhnovich, E., Cocco, S., Monasson, R., 2016. Benchmarking inverse statistical approaches for protein structure and design with exactly solvable models. PLoS Comput. Biol. 12, e1004889.

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., 1983. The Neutral Theory of Molecular Evolution. Cambridge Univ. Press, Cambridge.

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.

Knapp, S., Mattson, P. T., Christova, P., Berndt, K. D., Karshikoff, A., Vihinen, M., Smith, C. E., Ladenstein, R., 1998. Thermal unfolding of small proteins with sh3 domain folding pattern. Proteins 31, 309–319.

Kragelund, B. B., Osmark, P., Neergaard, T. B., Schiødt, J., Kristiansen, K., Knudsen, J., Poulsen, F. M., 1999. The formation of a native-like structure containing eight conserved hydrophobic residues is rate limiting in two-state protein folding of ACBP. Nature Struct. Biol 6, 594–601.

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.

Marks, D. S., Colwell, L. J., Sheridan, R., Hopf, T. A., Pagnani, A., Zecchina, R., Sander, C., 12 2011. Protein 3D structure computed from evolutionary sequence variation. PLoS ONE 6 (12), e28766.
URL http://dx.doi.org/10.1371/journal.pone.0028766

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., 01 2013. Prediction of contact residue pairs based on co-substitution between sites in protein structures. PLoS ONE 8 (1), e54252.
URL http://dx.doi.org/10.1371/journal.pone.0054252

Miyazawa, S., 2016. Selection maintaining protein stability at equilibrium. J. Theor. Biol. 391, 21–34.

Moran, P. A. P., 1958. Random processes in genetics. Proc. Cambridge Phil. Soc. 54, 60–71.

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.

Morcos, F., Schafer, N. P., Cheng, R. R., Onuchic, J. N., Wolynes, P. G., 2014. Coevolutionary information, protein folding landscapes, and the thermodynamics of natural selection. Proc. Natl. Acad. Sci. USA 111, 12408–12413.

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.

Ohta, T., 2002. Near-neutrality in evolution of genes and gene regulation. Proc. Natl. Acad. Sci. USA 99, 16134–16137.

Onuchic, J. N., Wolynes, P. G., Lutheyschulten, Z., Socci, N. D., 1995. Toward an outline of the topography of a realistic protein-folding funnel. Proc. Natl. Acad. Sci. USA 92, 3626–3630.

Onwukwe, G. U., Kursula1, P., Koski, M. K., Schmitz, W., Wierenga, R. K., 2014. Human δ^3,δ^2-enoyl-CoA isomerase, type 2: a structural enzymology study on the catalytic role of its ACBP domain and helix-10. FEBS J. 282, 746–768.

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

Pande, V. S., Grosberg, A. Y., Tanaka, T., 2000. Heteropolymer freezing and design: Towards physical models of protein folding. Rev. Mod. Phys. 72, 259–314.

Parsons, L. M., Lin, F., Orban, J., 2006. Peptidoglycan recognition by Pal, an outer membrane lipoprotein. Biochemistry 45, 2122–2128.

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

Rosa, C. L., Milardi, D., Grasso, D., Guzzi, R., Sportelli, L., 1995. Thermodynamics of the thermal unfolding of azurin. J. Phys. Chem. 99, 14864–14870.

Ruiz-Sanz, J., Simoncsits, A., Tőrő, I., Pongor, S., Mateo, P. L., Filimonov, V. V., 1999. A thermodynamic study of the 434-repressor N-terminal domain and of its covalently linked dimers. Eur. J. Biochem 263, 246–253.

Sainsbury, S., Ren, J., Saunders, N. J., Stuarta, D. I., Owens, R. J., 2008. Crystallization and preliminary X-ray analysis of CrgA, a LysR-type transcriptional regulator from pathogenic *Neisseria meningitidis* MC58. Acta Cryst. F64, 797–801.

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

Shakhnovich, E. I., 1994. Proteins with selected sequences fold into unique native conformation. Phys. Rev. Lett. 72, 3907–3911.

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.

Stupák, M., Zŏldák, G., Musatov, A., Sprinzl, M., Sedlák, E., 2006. Unusual effect of salts on the homodimeric structure of NADH oxidase from *Thermus thermophilus* in acidic pH. Biochim. Biophys. Acta 1764, 129–137.

Sułkowska, J. I., Morcos, F., Weigt, M., Hwa, T., Onuchic, J. N., 2012. Genomics-aided structure prediction. Proc. Natl. Acad. Sci. USA 109, 10340–10345.

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.

Torchio, G. M., Ermácora, M. R., Sica, M. P., 2012. Equilibrium unfolding of the PDZ domain of β2-syntrophin. Biophys. J. 102, 2835–2844.

Williams, N. K., Prosselkov, P., Liepinsh, E., Line, I., Sharipo, A., Littler, D. R., Curmi, P. M. G., Otting, G., Dixon, N. E., 2002. *In Vivo* protein cyclization promoted by a circularly permuted *Synechocystis* sp. PCC6803 DnaB mini-intein. J. Biol. Chem. 277, 7790–7798.

Wilson, C. J., Wittung-Stafshede, P., 2005. Snapshots of a dynamic folding nucleus in zinc-substituted *Pseudomonas aeruginosa* azurin. Biochemistry 44, 10054–10062.

Table S.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 (Finn et al., 2016) are used.
[b] The effective number of sequences.
[c] A sample weight ($w_{\boldsymbol{\sigma}_N}$) 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.
[h] These proteins consist of two domains, and other ones are single domains.

Table S.2: **Parameter values for $r_{\mathbf{cutoff}} \sim 8$ Å** employed for each protein family, and the averages of the evolutionary statistical energies ($\overline{\psi_N}$) over all homologous sequences and of the means and the standard deviations of interaction changes ($\overline{\overline{\Delta\psi_N}}$ and $\overline{\mathrm{Sd}(\Delta\psi_N)}$) due to single nucleotide nonsynonymous mutations at all sites over all homologous sequences in each protein family.

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]	$\overline{\mathrm{Sd}(\Delta\psi_N)} \pm$ Sd(Sd($\Delta\psi_N$)) [c]	r_{ψ_N} for $\overline{\Delta\psi_N}$ [d]	α_{ψ_N} for $\overline{\Delta\psi_N}$ [d]	r_{ψ_N} for Sd($\Delta\psi_N$) [e]	α_{ψ_N} for 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 ($w_{\boldsymbol{\sigma}_N}$) for each sequence; $w_{\boldsymbol{\sigma}_N}$ 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 S.1.

[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. (S.94).

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

Table S.3: **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; see Eq. (S.20).
[c] Temperatures are set up for comparison to be equal to the experimental temperatures for ΔG_{ND} or to 298°K if unavailable; see Table S.4 for the experimental data.
[d] Folding free energy in kcal/mol units; see Eq. (S.47).

Table S.4: **Experimental data used.**

Pfam family	experimental values T_m (°K)	T (°K)	ΔG_{ND} (kcal/mol)	ref. for T_m	ref. for ΔG_{ND} and $\Delta\Delta G_{ND}$
HTH_3	343.7	298	-5.33 ± 0.36	(Ganguly et al., 2009)	(Ruiz-Sanz et al., 1999)
Nitroreductase	337.0	-	-	(Stupák et al., 2006)	
SBP_bac_3	336.1	-	-	(D'Auria et al., 2005)	
OmpA	320.0	-	-	(Parsons et al., 2006)	
DnaB	312.8	-	-	(Williams et al., 2002)	
LysR_substrate	338.0	-	-	(Sainsbury et al., 2008)	
Methyltransf_5	375.0	-	-	(Armengaud et al., 2004) (Guelorget et al., 2010)	
SH3_1	344.0	295	-3.70	(Knapp et al., 1998)	(Grantcharova et al., 1998)
ACBP	324.4	278	-8.08 ± 0.08	(Onwukwe et al., 2014)	(Kragelund et al., 1999)
PDZ	312.9	298	-2.90	(Torchio et al., 2012)	(Gianni et al., 2005, 2007)
Copper-bind	359.3	298	-12.90 ± 0.36	(Rosa et al., 1995)	(Wilson and Wittung-Stafshede, 2005)

Table S.5: **Parameter values for $r_{\mathbf{cutoff}} \sim 15.5$ Å** employed for each protein family, and the averages of the evolutionary statistical energies ($\overline{\psi_N}$) over all homologous sequences and of the means and the standard deviations of interaction changes ($\overline{\overline{\Delta\psi_N}}$ and $\overline{\mathrm{Sd}(\Delta\psi_N)}$) due to single nucleotide nonsynonymous mutations at all sites over all homologous sequences in each protein family.

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]	$\overline{\mathrm{Sd}(\Delta\psi_N)} \pm$ Sd(Sd($\Delta\psi_N$)) [c]	r_{ψ_N} for $\overline{\Delta\psi_N}$ [d]	α_{ψ_N} for $\overline{\Delta\psi_N}$ [d]	r_{ψ_N} for Sd($\Delta\psi_N$) [e]	α_{ψ_N} for Sd($\Delta\psi_N$) [e]
HTH_3	53	0.245	32.90	15.67	−0.2548	4.0057	−4.2642	6.8512	6.9544 ± 0.5309	−0.955	−1.5717	−0.519	−0.5727
Nitroreductase	73	0.315	28.71	15.75	−0.1476	3.7093	−3.8565	6.3226	5.6267 ± 0.5440	−0.953	−1.5765	−0.694	−0.6640
SBP_bac_3	218	0.35	55.48	15.90	−0.0669	3.4004	−3.4674	5.7978	4.8666 ± 0.4517	−0.971	−1.6708	−0.821	−0.8874
SBP_bac_3	94	0.455	42.81	15.45	−0.1628	2.3208	−2.4831	4.0963	3.7760 ± 0.3970	−0.968	−1.6628	−0.770	−0.6408
OmpA	95	0.235	35.58	15.69	−0.2552	5.8175	−6.0757	10.4102	11.8829 ± 0.4108	−0.948	−1.6212	−0.354	−0.3599
DnaB	98	0.35	46.65	15.57	−0.2351	6.1890	−6.4167	10.7294	8.0204 ± 0.3493	−0.894	−1.5176	−0.311	−0.3037
LysR_substrate	191	0.335	52.30	15.58	−0.2826	2.5962	−2.8789	4.4194	4.1701 ± 0.1782	−0.963	−1.6196	−0.613	−0.4726
LysR_substrate	103	0.37	44.33	15.60	−0.2816	2.4438	−2.7239	4.1276	4.2029 ± 0.3674	−0.984	−1.5436	−0.769	−0.5462
Methyltransf_5	285	0.175	53.52	15.53	−0.1687	12.8982	−13.0658	23.6376	18.7982 ± 0.4701	−0.952	−1.9804	−0.171	−0.1630
Methyltransf_5	80	0.24	37.02	15.11	−0.1632	9.9944	−10.1576	17.5749	13.9124 ± 0.4756	−0.862	−1.6406	−0.290	−0.2822
SH3_1	48	0.165	28.46	15.76	−0.1350	7.6161	−7.7523	11.9725	13.3845 ± 0.4719	−0.896	−1.5944	−0.255	−0.2420
ACBP	80	0.28	36.27	15.34	−0.0235	7.4707	−7.4947	13.1892	9.7188 ± 0.4242	−0.911	−1.7087	0.085	0.0861
PDZ	81	0.33	40.82	15.77	−0.3022	5.2295	−5.5313	8.6909	7.9383 ± 0.2930	−0.966	−1.7215	−0.316	−0.2328
Copper-bind	125	0.295	45.22	15.32	−0.0999	8.5521	−8.6592	15.5941	9.6566 ± 0.3556	−0.951	−1.7441	−0.175	−0.1981

[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 without deletions are used with a sample weight ($w_{\boldsymbol{\sigma}_N}$) for each sequence; $w_{\boldsymbol{\sigma}_N}$ 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 S.1.

[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 without 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. (S.94).

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

Table S.6: **Thermodynamic quantities estimated with $r_{\mathbf{cutoff}} \sim 15.5$ Å.**

Pfam family	r [a]	$k_B\hat{T}_s$ [a] (kcal/mol)	Experimental $\hat{T}_s$ (°K)	Experimental T_m (°K)	Experimental $\hat{T}_g$ (°K)	$\hat{\omega}$ [b] (k_B)	T [c] (°K)	$\langle \Delta G_{ND} \rangle$ [d] (kcal/mol)
HTH_3	–	–	93.1	343.7	136.0	0.9378	298	−3.70
Nitroreductase	–	–	115.0	337	152.9	1.0501	298	−4.56
SBP_bac_3	–	–	133.0	336.1	166.9	1.0794	298	−12.85
SBP_bac_3	–	–	171.4	336.1	196.6	0.8818	298	−3.85
OmpA	–	–	54.5	320	97.6	0.9060	298	−3.38
DnaB	–	–	80.7	312.8	120.4	1.3908	298	−3.38
LysR_substrate	–	–	155.2	338	184.5	0.9185	298	−9.22
LysR_substrate	–	–	154.0	338	183.6	0.8598	298	−4.68
Methyltransf_5	–	–	34.4	375	82.3	1.1299	298	−46.26
Methyltransf_5	–	–	46.5	375	96.4	1.1630	298	−13.04
SH3_1	0.836	0.0821	48.4	344	94.6	0.9954	295	−4.24
ACBP	0.823	0.0689	66.6	324.4	109.7	1.3763	278	−8.79
PDZ	0.944	0.1619	81.5	312.88	121.1	1.1852	298	−2.39
Copper-bind	0.888	0.1015	67.0	359.3	115.2	1.4466	298	−19.28

[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; see Eq. (S.20).
[c] Temperatures are set up for comparison to be equal to the experimental temperatures for ΔG_{ND} or to 298°K if unavailable; see Table S.4 for the experimental data.
[d] Folding free energy in kcal/mol units; see Eq. (S.47).

Table S.7: **Fixation probabilities of a single mutant in various models.**

A) For Wright-Fisher population; compiled from p. 192 and pp. 424–427 of Crow and Kimura (1970).

Fitness/Selection [a]	h [a]	$M_{\delta x}$ [b]	$V_{\delta x}$ [c]	u [de]	q_m [f]
No dominance	$1/2$	$sx(1-x)$	$x(1-x)/(2N_e)$	$(1-e^{-4N_e sq_m})/(1-e^{-4N_e s})$	$1/(2N)$
Dominance favored	1	$2sx(1-x)^2$	$x(1-x)/(2N_e)$	[e]	$1/(2N)$
Recessive favored	0	$2sx^2(1-x)$	$x(1-x)/(2N_e)$	[e]	$1/(2N)$
Gametic selection		$sx(1-x)$	$x(1-x)/(2N_e)$	$(1-e^{-4N_e sq_m})/(1-e^{-4N_e s})$	$1/(2N)$
Haploid		$sx(1-x)$	$x(1-x)/N_e$	$(1-e^{-2N_e sq_m})/(1-e^{-2N_e s})$	$1/N$

B) For Moran population (Moran, 1958; Ewens, 1979)

Fitness/Selection [a]	$M_{\delta x}$	$V_{\delta x}$ [c]	u [de]	q_m [f]
Haploid	$sx(1-x)/N_e$	$2x(1-x)/N_e^2$	$(1-e^{-N_e sq_m})/(1-e^{-N_e s})$	$1/N$

[a] For zygotic selection, $2s$ and $2sh$ are the selective advantages of mutant homogeneous and heterogeneous zygotes, respectively. For others, s is the selective advantage of mutant gene.
[b] Mean in the rate of the change of gene frequency per generation; $M_{\delta x} = 2sx(1-x)(h+(1-2h)x)$ for zygotic selection.
[c] Variance in the rate of the change of gene frequency per generation.
[d] Fixation probability.
[e] $u(q_m) = F(q_m)/F(1)$ where $F(q_m) = \int_0^{q_m} G(x)dx$ and $G(x) = \exp(-\int 2M_{\delta x}/V_{\delta x}dx)$.
[f] Frequency of a single mutant gene.

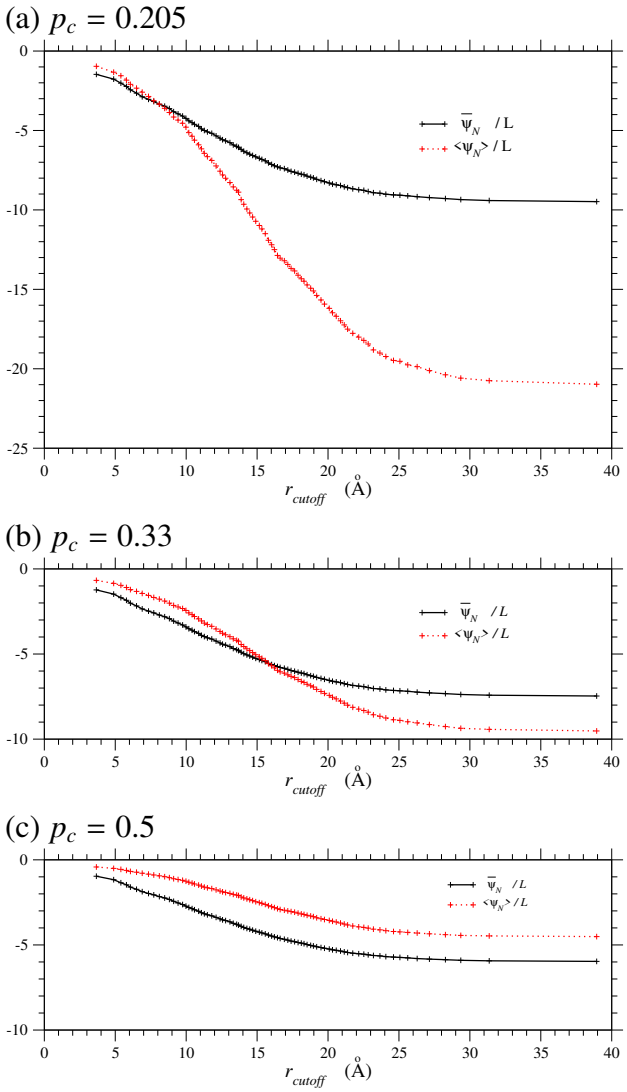

Figure S.1: **Dependences of the sample ($\overline{\psi_N}/L$) and ensemble ($\langle\psi_N\rangle_{\boldsymbol{\sigma}}/L$) averages of evolutionary statistical energy per residue on the cutoff distance for pairwise interactions in the PDZ domain.** The ratios of pseudocount p_c = 0.205 and 0.33 are employed here for the cutoff distance r_{cutoff} ~ 8 and 15.5 Å, respectively. The black solid and red dotted lines indicate the sample and ensemble averages, respectively.

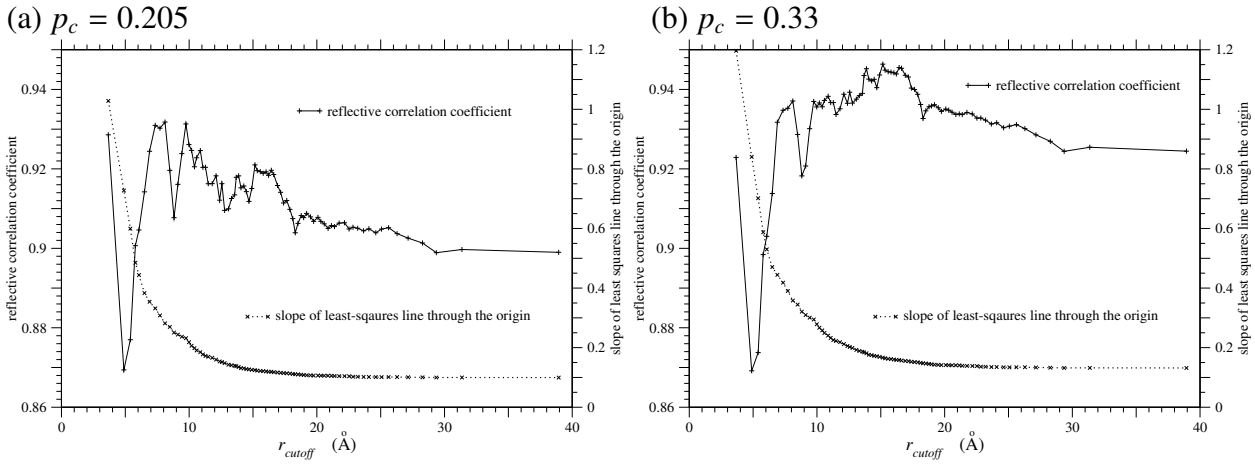

Figure S.2: **Dependences of the reflective correlation and regression coefficients between the experimental $\Delta\Delta G_{ND}$ (Gianni et al., 2007) and $\Delta\psi_N$ due to single amino acid substitutions on the cutoff distance for pairwise interactions in the PDZ domain.** The left and right figures are for the ratios of pseudocount, p_c = 0.205 and 0.33, respectively. The solid and dotted lines show the reflective correlation and regression coefficients for the least-squares regression line through the origin, respectively. The sample ($\overline{\psi_N}/L$) and ensemble ($\langle\psi_N\rangle_{\boldsymbol{\sigma}}/L$) averages of evolutionary statistical energy agree with each other at the cutoff distance $r_{\rm cutoff} \sim 8$ Å for p_c = 0.205 and $r_{\rm cutoff} \sim 15.5$ Å for p_c = 0.33, where the reflective correlation coefficients attain to the maximum.

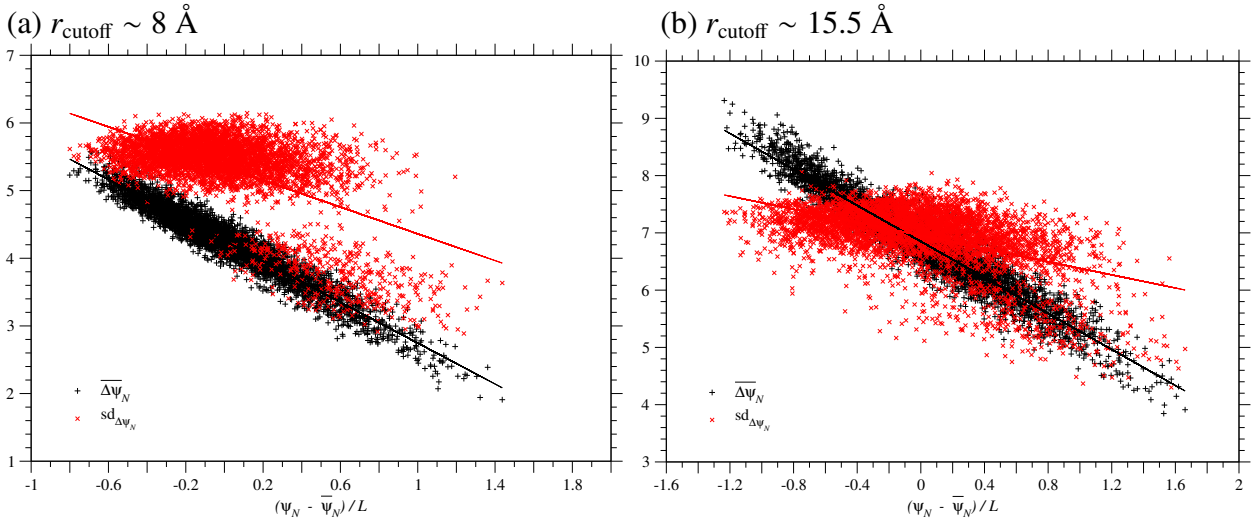

Figure S.3: **Correlation between $\Delta\psi_N$ due to single nucleotide nonsynonymous substitutions and ψ_N of homologous sequences in the HTH_3 family of the domain, 1R69-A:6-58.** The left and right figures correspond to the cutoff distance $r_{\mathrm{cutoff}} \sim 8$ and 15.5 Å, respectively. Each of the black plus or red cross marks corresponds to the mean or the standard deviation of $\Delta\psi_N$ due to all types of single nucleotide nonsynonymous substitutions over all sites in each of the homologous sequences. Representatives of unique sequences, which are at least 20% different from each other, are employed; the number of the representatives is almost equal to M_{eff} in Table S.1. The solid lines show the regression lines for the mean and the standard deviation of $\Delta\psi_N$.

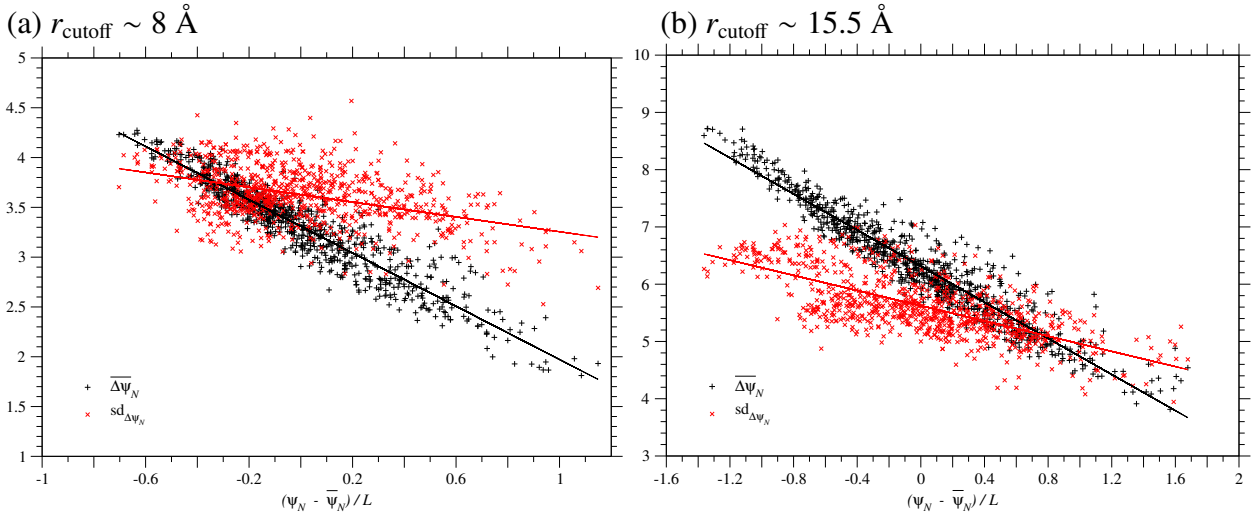

Figure S.4: **Correlation between $\Delta\psi_N$ due to single nucleotide nonsynonymous substitutions and ψ_N of homologous sequences in the Nitroreductase family of the domain, 3E10-A/B:4-76.** The left and right figures correspond to the cutoff distance $r_{\mathrm{cutoff}} \sim 8$ and 15.5 Å, respectively. Each of the black plus or red cross marks corresponds to the mean or the standard deviation of $\Delta\psi_N$ due to all types of single nucleotide nonsynonymous substitutions over all sites in each of the homologous sequences. Representatives of unique sequences, which are at least 20% different from each other, are employed; the number of the representatives is almost equal to M_{eff} in Table S.1. The solid lines show the regression lines for the mean and the standard deviation of $\Delta\psi_N$.

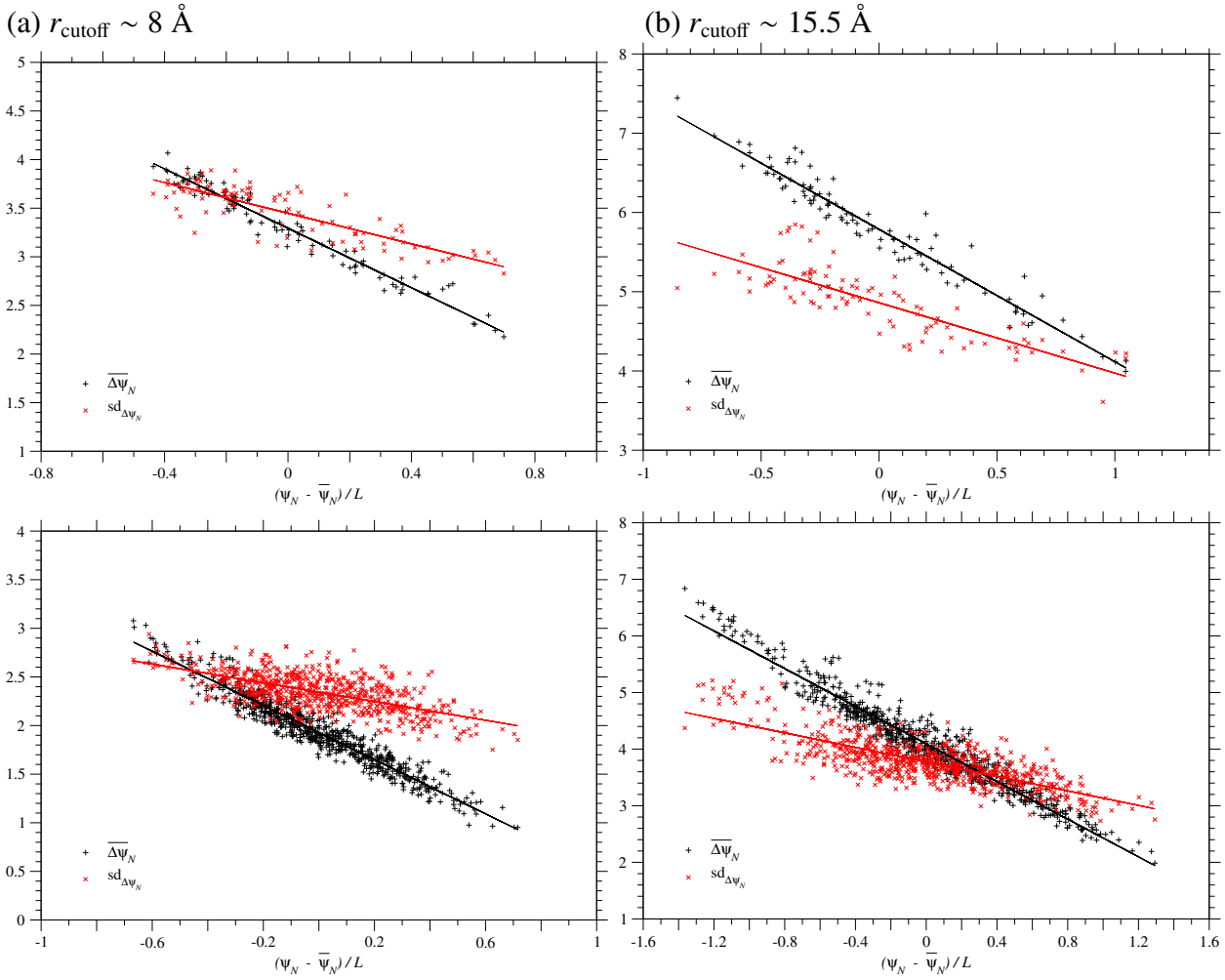

Figure S.5: **Correlation between $\Delta\psi_N$ due to single nucleotide nonsynonymous substitutions and ψ_N of homologous sequences in the SBP_bac_3 family of the domains, 1WDN-A:5-222 (upper) and 1WDN-A:89-182 (lower).** The left and right figures correspond to the cutoff distance r_{cutoff} ~ 8 and 15.5 Å, respectively. Each of the black plus or red cross marks corresponds to the mean or the standard deviation of $\Delta\psi_N$ due to all types of single nucleotide nonsynonymous substitutions over all sites in each of the homologous sequences. Representatives of unique sequences, which are at least 20% different from each other, are employed; the number of the representatives is almost equal to M_{eff} in Table S.1. The solid lines show the regression lines for the mean and the standard deviation of $\Delta\psi_N$.

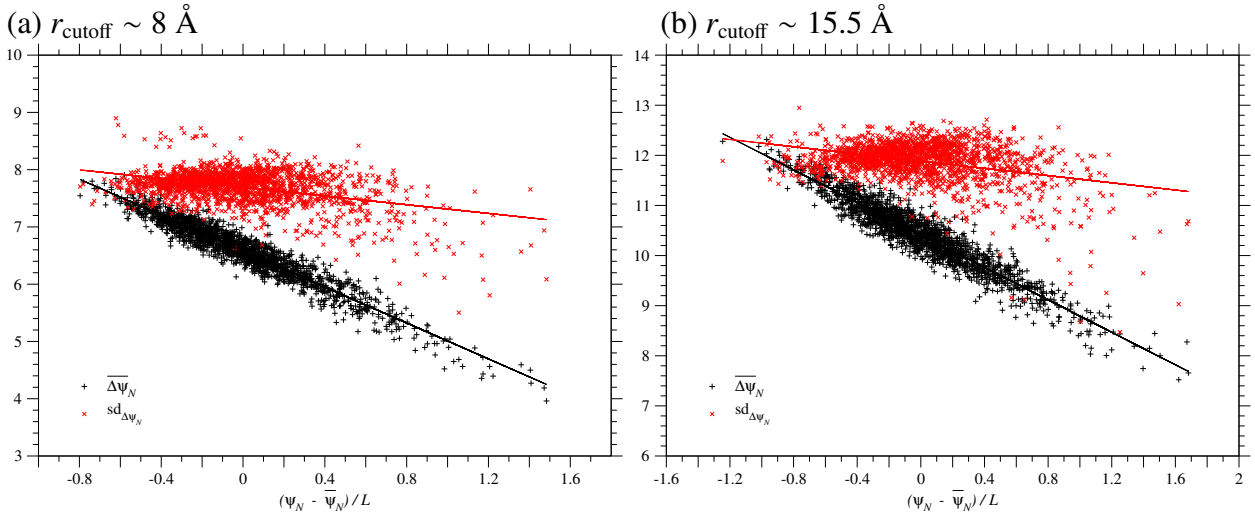

Figure S.6: **Correlation between $\Delta\psi_N$ due to single nucleotide nonsynonymous substitutions and ψ_N of homologous sequences in the OmpA family of the domains, 1OAP-A:52-146.** The left and right figures correspond to the cutoff distance $r_{\mathrm{cutoff}} \sim 8$ and 15.5 Å, respectively. Each of the black plus or red cross marks corresponds to the mean or the standard deviation of $\Delta\psi_N$ due to all types of single nucleotide nonsynonymous substitutions over all sites in each of the homologous sequences. Representatives of unique sequences, which are at least 20% different from each other, are employed; the number of the representatives is almost equal to M_{eff} in Table S.1. The solid lines show the regression lines for the mean and the standard deviation of $\Delta\psi_N$.

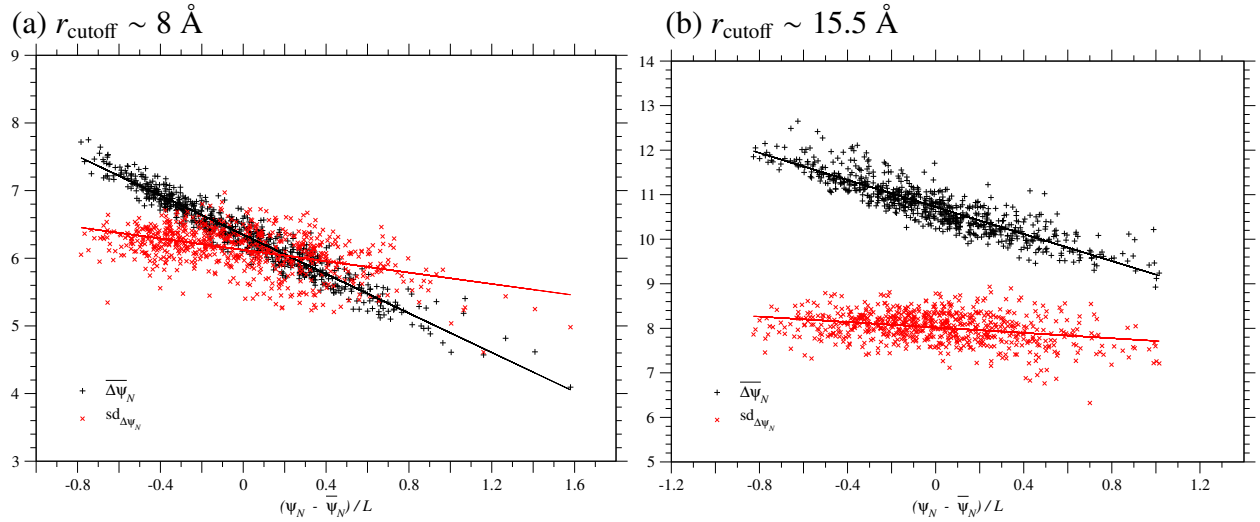

Figure S.7: **Correlation between $\Delta\psi_N$ due to single nucleotide nonsynonymous substitutions and ψ_N of homologous sequences in the DnaB family of the domains, 1JWE-A:30-127.** The left and right figures correspond to the cutoff distance r_{cutoff} ~ 8 and 15.5 Å, respectively. Each of the black plus or red cross marks corresponds to the mean or the standard deviation of $\Delta\psi_N$ due to all types of single nucleotide nonsynonymous substitutions over all sites in each of the homologous sequences. Representatives of unique sequences, which are at least 20% different from each other, are employed; the number of the representatives is almost equal to M_{eff} in Table S.1. The solid lines show the regression lines for the mean and the standard deviation of $\Delta\psi_N$.

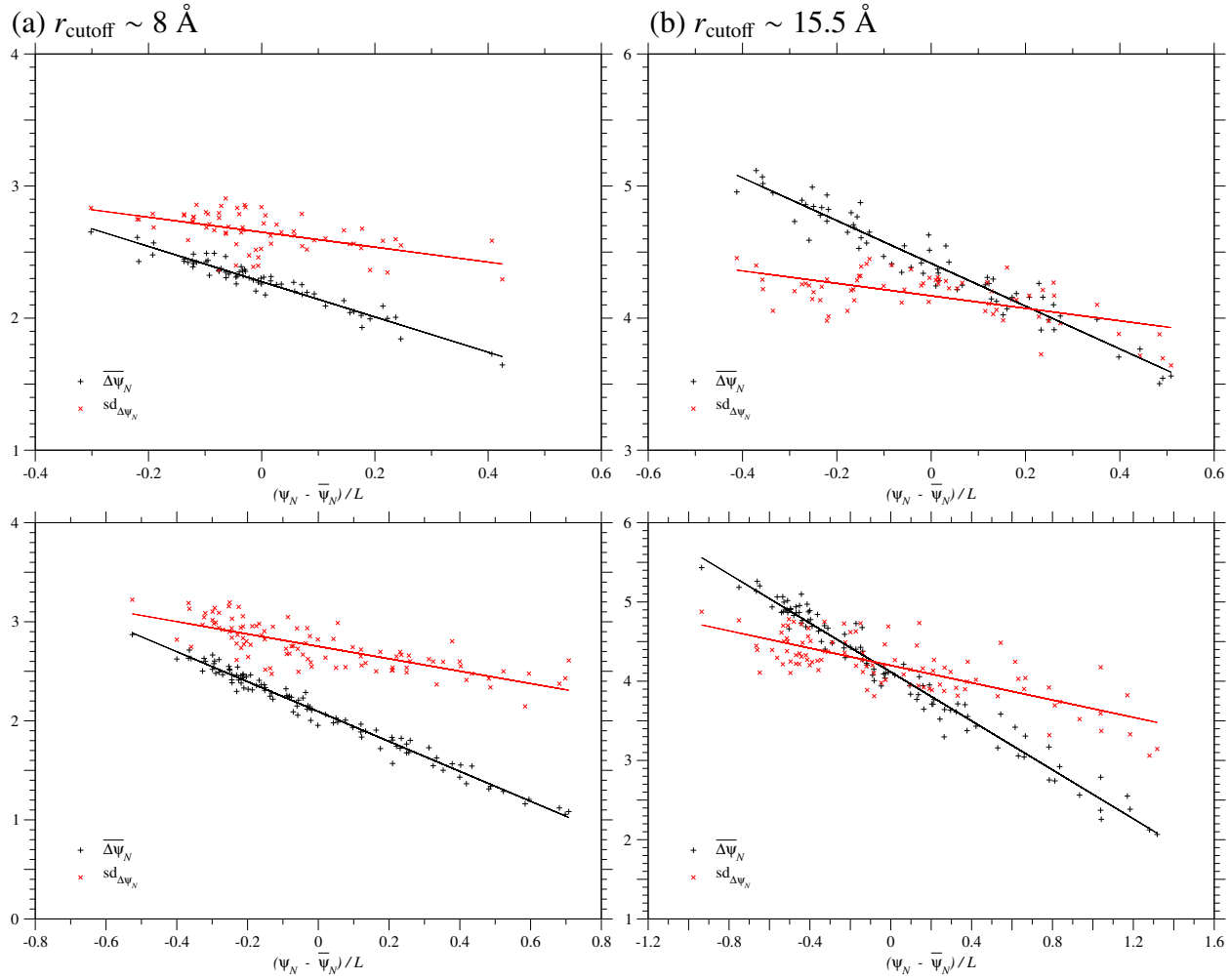

Figure S.8: **Correlation between $\Delta\psi_N$ due to single nucleotide nonsynonymous substitutions and ψ_N of homologous sequences in the LysR_substrate family of the domains, 2F6G-A:90-280 (above) and 2F6G-A:163-265 (below).** The left and right figures correspond to the cutoff distance r_{cutoff} ~ 8 and 15.5 Å, respectively. Each of the black plus or red cross marks corresponds to the mean or the standard deviation of $\Delta\psi_N$ due to all types of single nucleotide nonsynonymous substitutions over all sites in each of the homologous sequences. Representatives of unique sequences, which are at least 20% different from each other, are employed; the number of the representatives is almost equal to M_{eff} in Table S.1. The solid lines show the regression lines for the mean and the standard deviation of $\Delta\psi_N$.

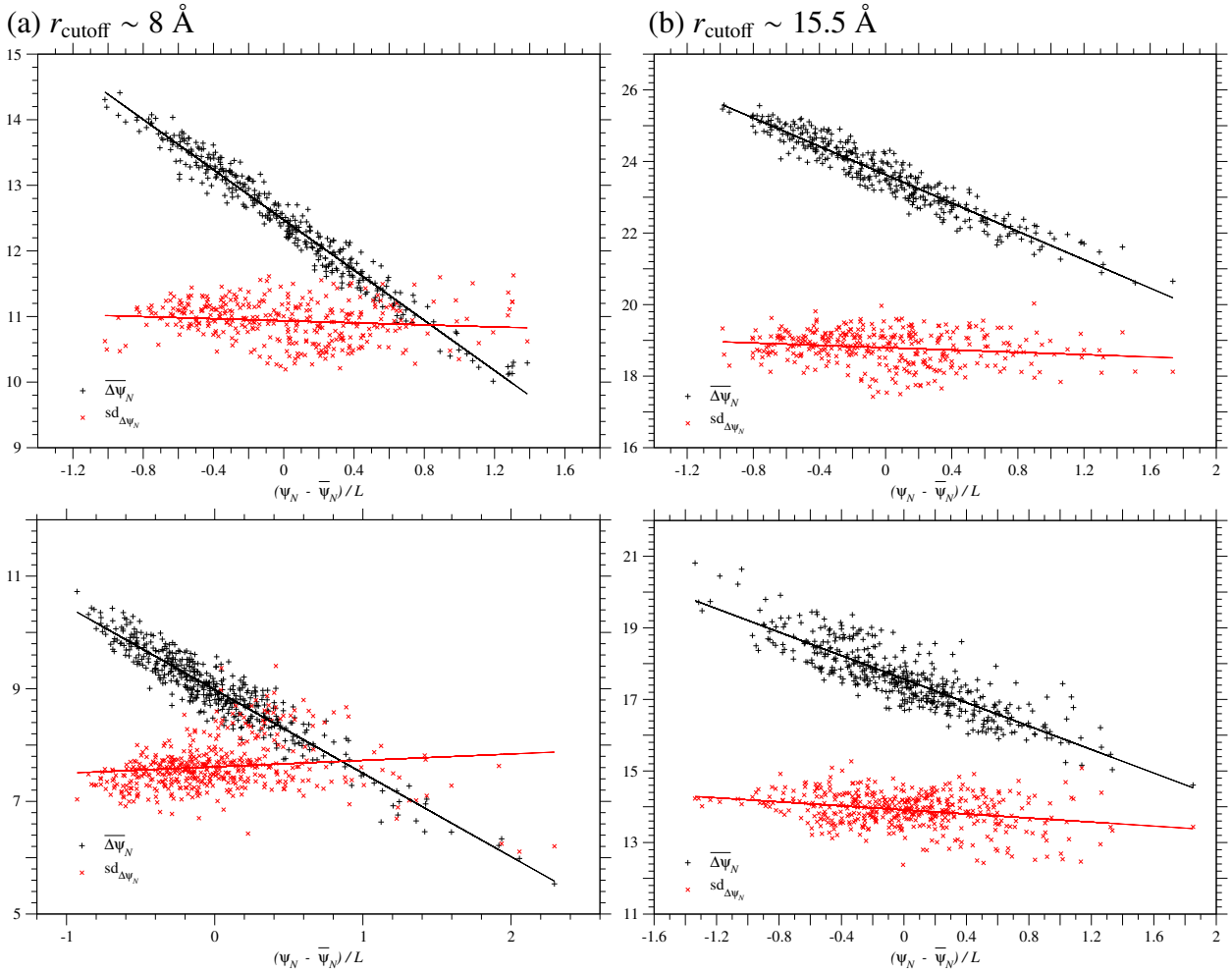

Figure S.9: **Correlation between $\Delta\psi_N$ due to single nucleotide nonsynonymous substitutions and ψ_N of homologous sequences in the Methyltransf_5 family of the domains, 1N2X-A:8-292 (above) and 1N2X-A:137-216 (below).** The left and right figures correspond to the cutoff distance r_{cutoff} ~ 8 and 15.5 Å, respectively. Each of the black plus or red cross marks corresponds to the mean or the standard deviation of $\Delta\psi_N$ due to all types of single nucleotide nonsynonymous substitutions over all sites in each of the homologous sequences. Representatives of unique sequences, which are at least 20% different from each other, are employed; the number of the representatives is almost equal to M_{eff} in Table S.1. The solid lines show the regression lines for the mean and the standard deviation of $\Delta\psi_N$.

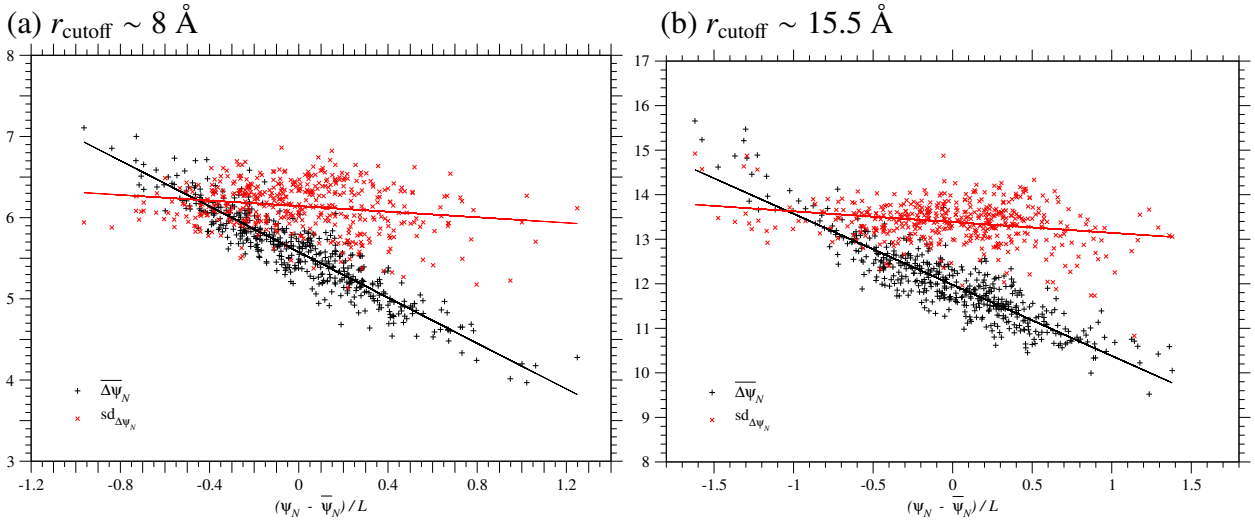

Figure S.10: **Correlation between $\Delta\psi_N$ due to single nucleotide nonsynonymous substitutions and ψ_N of homologous sequences in the SH3_1 family of the domain, 1FMK-A:87-134.** The left and right figures correspond to the cutoff distance r_{cutoff} ~ 8 and 15.5 Å, respectively. Each of the black plus or red cross marks corresponds to the mean or the standard deviation of $\Delta\psi_N$ due to all types of single nucleotide nonsynonymous substitutions over all sites in each of the homologous sequences. Representatives of unique sequences, which are at least 20% different from each other, are employed; the number of the representatives is almost equal to M_{eff} in Table S.1. The solid lines show the regression lines for the mean and the standard deviation of $\Delta\psi_N$.

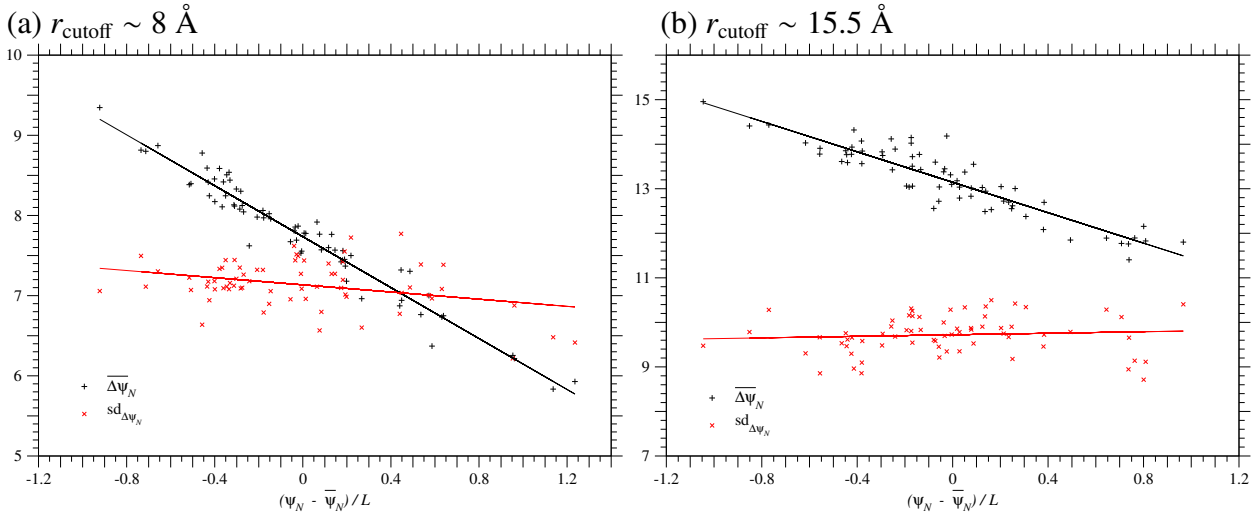

Figure S.11: **Correlation between $\Delta\psi_N$ due to single nucleotide nonsynonymous substitutions and ψ_N of homologous sequences in the ACBP family of the domain, 2ABD-A:2-81.** The left and right figures correspond to the cutoff distance r_{cutoff} ~ 8 and 15.5 Å, respectively. Each of the black plus or red cross marks corresponds to the mean or the standard deviation of $\Delta\psi_N$ due to all types of single nucleotide nonsynonymous substitutions over all sites in each of the homologous sequences. Representatives of unique sequences, which are at least 20% different from each other, are employed; the number of the representatives is almost equal to M_{eff} in Table S.1. The solid lines show the regression lines for the mean and the standard deviation of $\Delta\psi_N$.

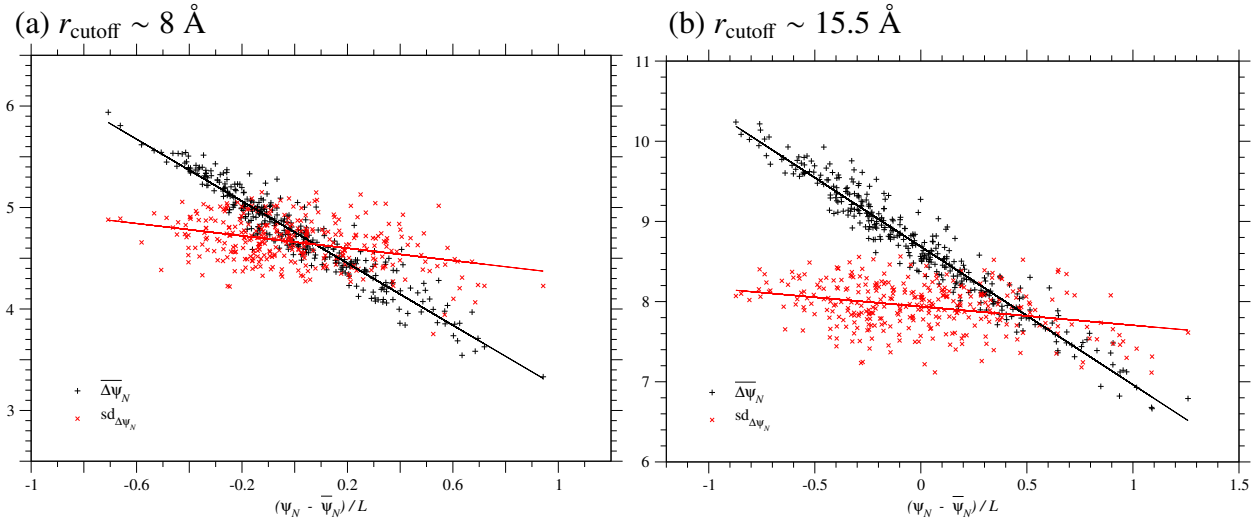

Figure S.12: **Correlation between $\Delta\psi_N$ due to single nucleotide nonsynonymous substitutions and ψ_N of homologous sequences in the PDZ domain family.** The left and right figures correspond to the cutoff distance r_{cutoff} ~ 8 and 15.5 Å, respectively. Each of the black plus or red cross marks corresponds to the mean or the standard deviation of $\Delta\psi_N$ due to all types of single nucleotide nonsynonymous substitutions over all sites in each of the homologous sequences of the PDZ domain family. Only 335 representatives of unique sequences with no deletions, which are at least 20% different from each other, are employed; the number of the representatives is almost equal to M_{eff} in Table S.1. The solid lines show the regression lines for the mean and the standard deviation of $\Delta\psi_N$.

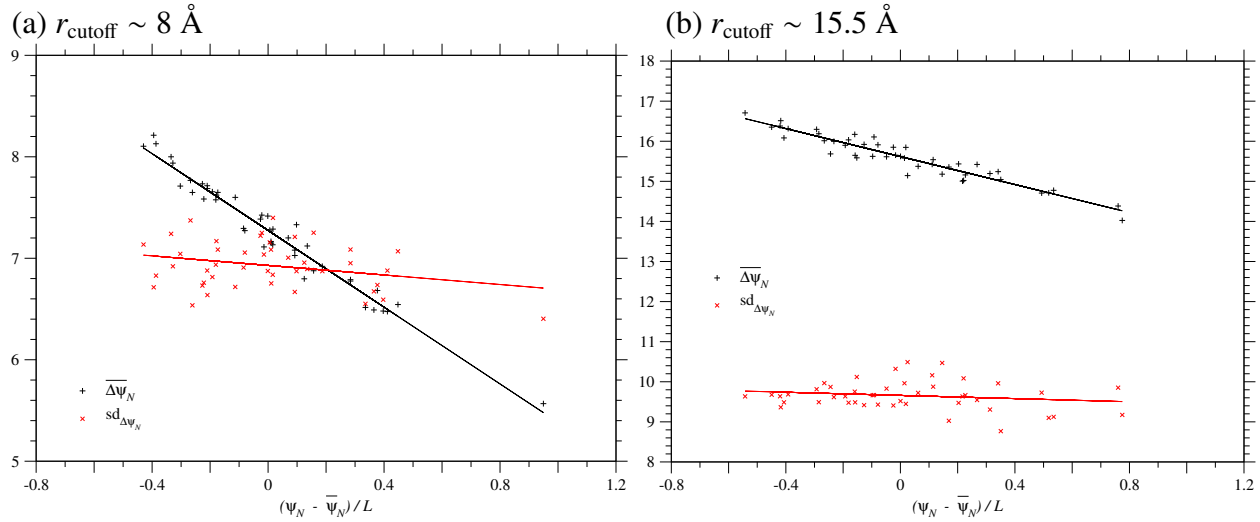

Figure S.13: **Correlation between $\Delta\psi_N$ due to single nucleotide nonsynonymous substitutions and ψ_N of homologous sequences in the Copper-bind family of the domain, 5AZU-B/D:4-128.** The left and right figures correspond to the cutoff distance $r_{\mathrm{cutoff}} \sim 8$ and 15.5 Å, respectively. Each of the black plus or red cross marks corresponds to the mean or the standard deviation of $\Delta\psi_N$ due to all types of single nucleotide nonsynonymous substitutions over all sites in each of the homologous sequences. Representatives of unique sequences, which are at least 20% different from each other, are employed; the number of the representatives is almost equal to M_{eff} in Table S.1. The solid lines show the regression lines for the mean and the standard deviation of $\Delta\psi_N$.

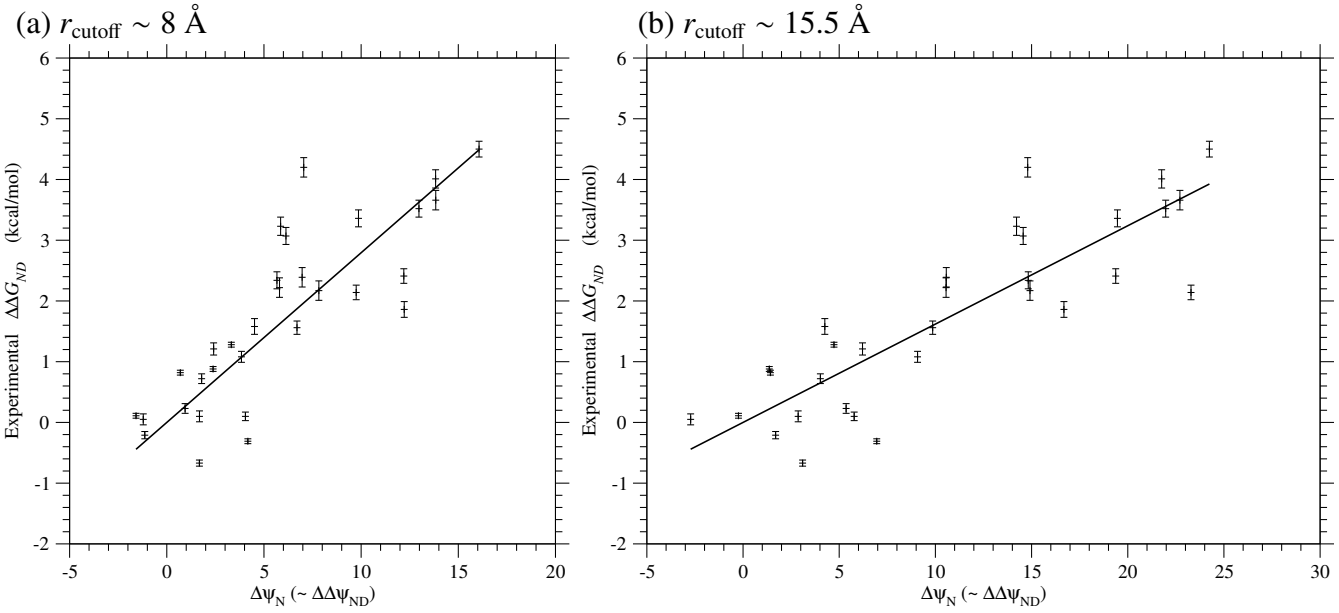

Figure S.14: **Regression of the experimental values (Gianni et al., 2007) of folding free energy changes ($\Delta\Delta G_{ND}$) due to single amino acid substitutions on $\Delta\psi_N(\simeq \Delta\Delta\psi_{ND})$ for the same types of substitutions in the PDZ domain.** The left and right figures correspond to the cutoff distance r_{cutoff} ~ 8 and 15.5 Å, respectively. The solid lines show the least-squares regression lines through the origin with the slopes, 0.279 kcal/mol for r_{cutoff} ~ 8 Å and 0.162 kcal/mol for r_{cutoff} ~ 15.5 Å, which are the estimates of k_BT_s. The reflective correlation coefficients for them are equal to 0.93 and 0.94, respectively. The free energies are in kcal/mol units.

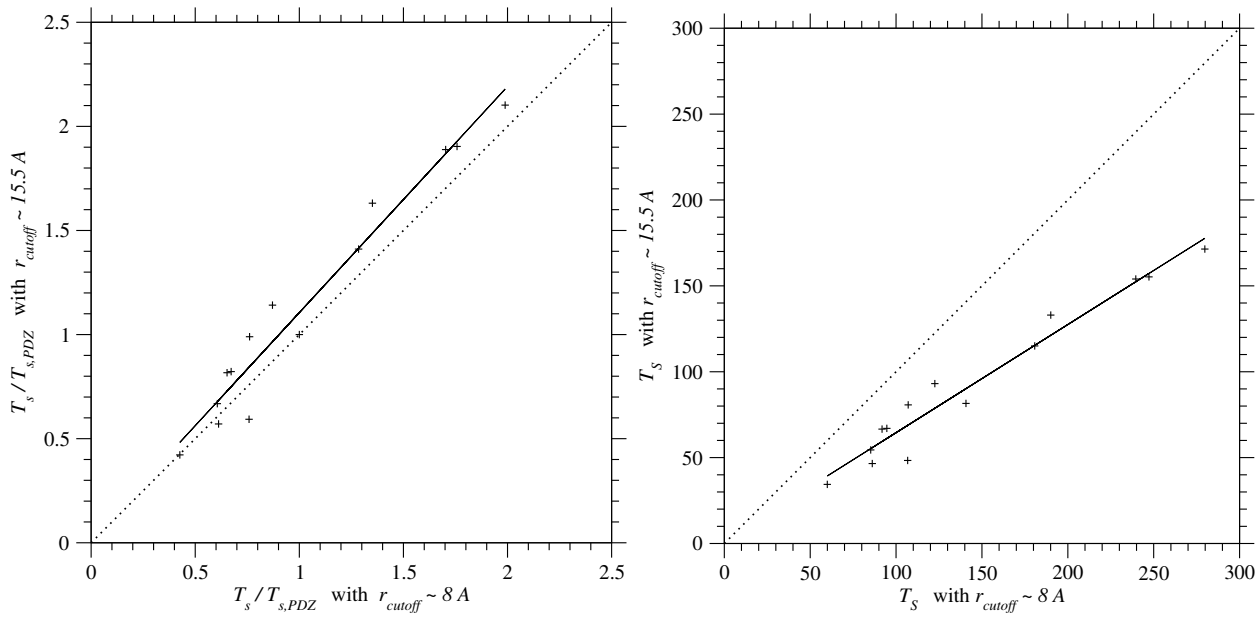

Figure S.15: **Comparison of selective temperatures (T_s) estimated with different cutoff distances by the present method.** The abscissa and ordinate correspond to the cases of $r_{\mathrm{cutoff}} \sim 8$ and 15.5 Å, respectively. The T_s is in °K units. The solid lines show the regression lines, $(T_s/T_{s,PDZ})_{15.5A} = 1.09(T_s/T_{s,PDZ})_{8A} + 0.02$ and $(T_s)_{15.5A} = 0.630(T_s)_{8A} + 1.57$. The correlation coefficients are equal to 0.98 for both.

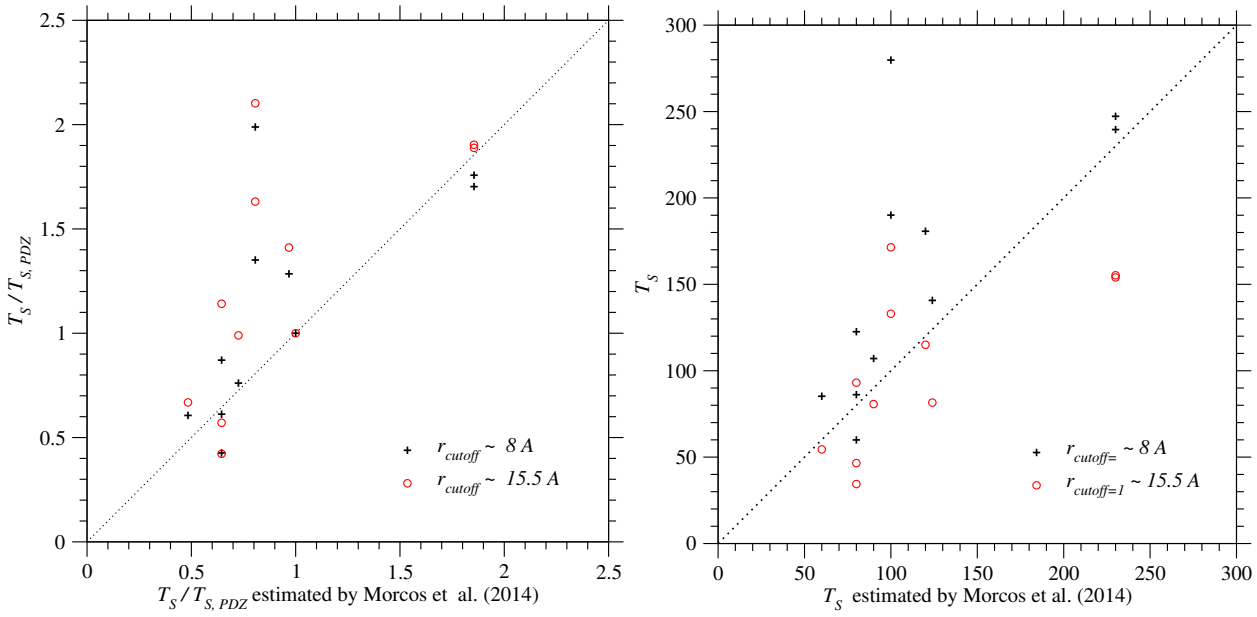

Figure S.16: **Selective temperatures (T_s) estimated by the present method are plotted against those estimated by Morcos et al. (Morcos et al., 2014); their estimated values of T_s tend to fall between the upper (r_{cutoff} ~ 8) and lower (r_{cutoff} ~ 15.5 Å) estimates of T_s. Plus and open circle marks correspond to the cases of r_{cutoff} ~ 8 and 15.5 Å, respectively.**

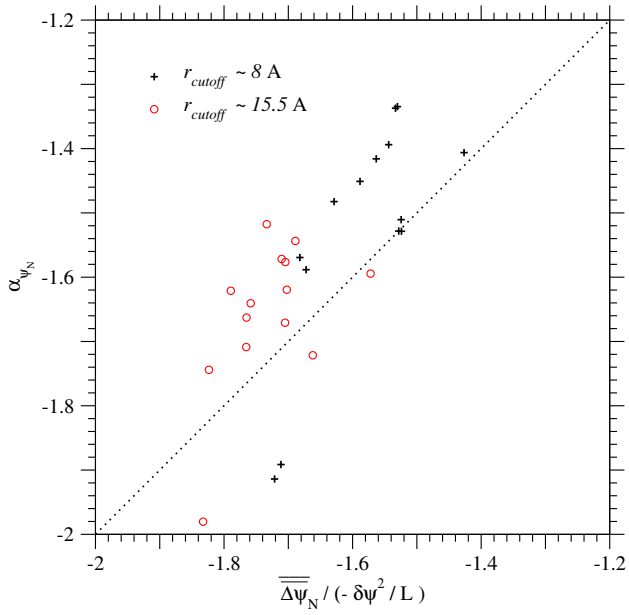

Figure S.17: **Comparison of α_{ψ_N}, which is the regression coefficient of $\overline{\Delta\psi_N}$ on ψ_N/L, with $\overline{\overline{\Delta\psi_N}}/(-\delta\psi^2/L)$ for each protein family.** Plus and open circle marks correspond to the cases of $r_{\mathrm{cutoff}} \sim 8$ and 15.5 Å, respectively.

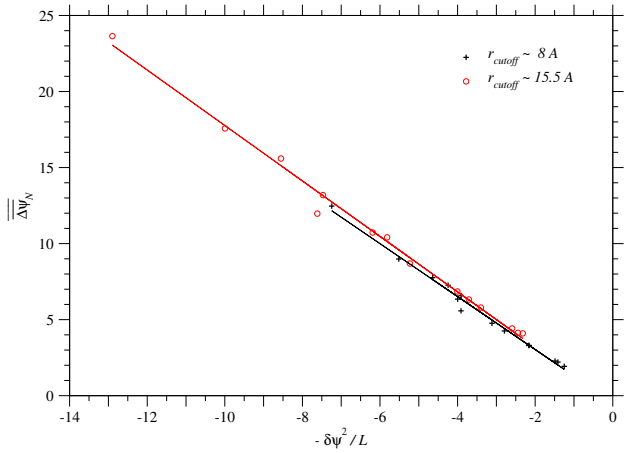

Figure S.18: **Dependence of the average of $\overline{\Delta\psi_N}$ due to single nucleotide nonsynonymous substitutions over homologous sequences on $-\delta\psi^2/L$ across protein families.** Plus and open circle marks indicate the values for each protein family in the cases of $r_{\text{cutoff}} \sim 8$ and 15.5 Å, respectively. In the case of the cutoff distance 8 Å, the correlation coefficient is equal to 0.995, and the regression line is $\overline{\overline{\Delta\psi_N(\sigma^N_{j\neq i}, \sigma^N_i \to \sigma_i)}} = -1.74(-\delta\psi^2/L) - 0.445$. In the case of $r_{\text{cutoff}} \sim 15.5$ Å, the correlation coefficient is equal to 0.996, and the regression line is $\overline{\overline{\Delta\psi_N(\sigma^N_{j\neq i}, \sigma^N_i \to \sigma_i)}} = -1.82(-\delta\psi^2/L) - 0.466$.

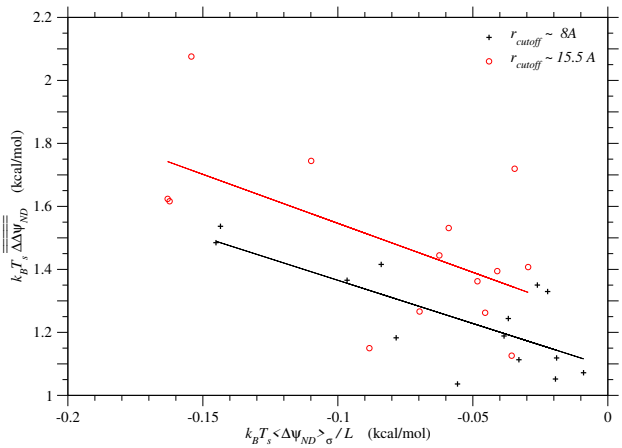

Figure S.19: **The sample average of folding free energy change, $\overline{\overline{\Delta\Delta G_{ND}}} \simeq k_B T_s \overline{\overline{\Delta\Delta\psi_{ND}}}$, is plotted against the ensemble average of folding free energy per residue, $\langle \Delta G_{ND} \rangle_{\boldsymbol{\sigma}} / L \simeq k_B T_s \langle \Delta\psi_{ND} \rangle_{\boldsymbol{\sigma}} / L$, for each protein family.** In the case of the cutoff distance 8 Å, the correlation coefficient is $r = -0.75$, and the regression line is $\overline{\overline{\Delta\Delta G_{ND}(\sigma^N_{j\neq i}, \sigma^N_i \to \sigma_i)}} = -2.74 \langle \Delta G_{ND} \rangle_{\boldsymbol{\sigma}} / L + 1.09$. In the case of $r_{\text{cutoff}} \sim 15.5$ Å, the correlation coefficient is $r = -0.59$, and the regression line is $\overline{\overline{\Delta\Delta G_{ND}(\sigma^N_{j\neq i}, \sigma^N_i \to \sigma_i)}} = -3.11 \langle \Delta G_{ND} \rangle_{\boldsymbol{\sigma}} / L + 1.24$. The free energies are in kcal/mol units.

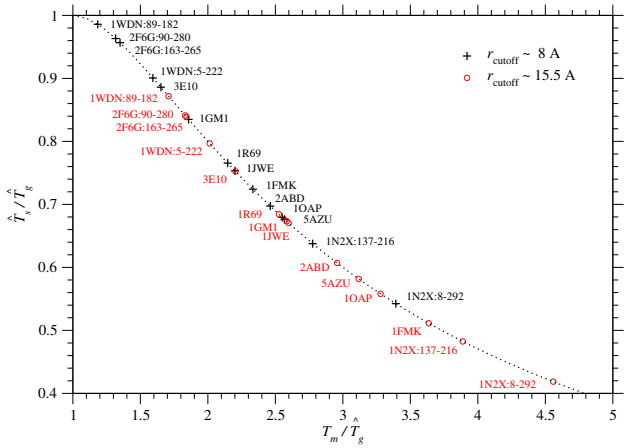

Figure S.20: $\hat{T}_s/\hat{T}_g$ **is plotted against** $T_m/\hat{T}_g$ **for each protein domain.** A dotted curve corresponds to Eq. (S.52), $\hat{T}_s/\hat{T}_g = 2(T_m/\hat{T}_g)/((T_m/\hat{T}_g)^2 + 1)$. Plus and open circle marks indicate the values estimated with $r_{\text{cutoff}} \sim 8$ and 15.5 Å, respectively. The effective temperature T_s for selection and glass transition temperature T_g must satisfy $T_s < T_g < T_m$ for proteins to be able to fold into unique native structures.

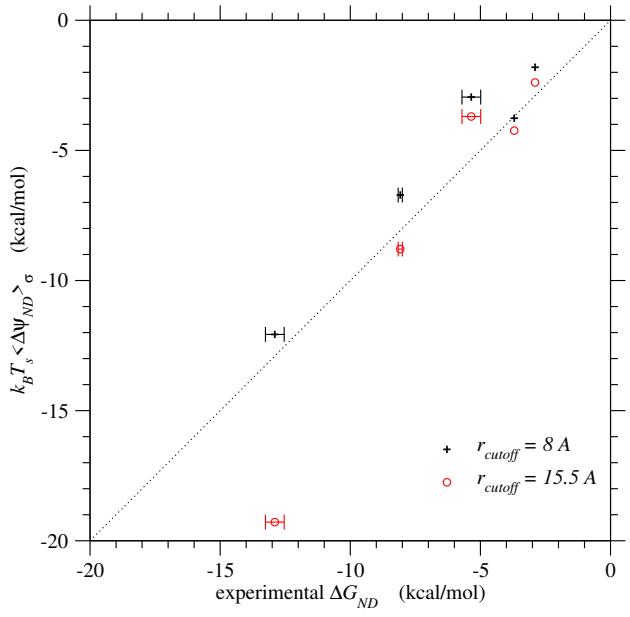

Figure S.21: **Folding free energies, $\langle \Delta G_{ND} \rangle_{\boldsymbol{\sigma}} \simeq k_B T_s \langle \Delta \psi_{ND} \rangle_{\boldsymbol{\sigma}}$, predicted by the present method are plotted against their experimental values, $\Delta G_{ND}(\boldsymbol{\sigma}_N)$.** Plus and open circle marks indicate the values estimated with $r_{\text{cutoff}} \sim 8$ and 15.5 Å, respectively. The free energies are in kcal/mol units.

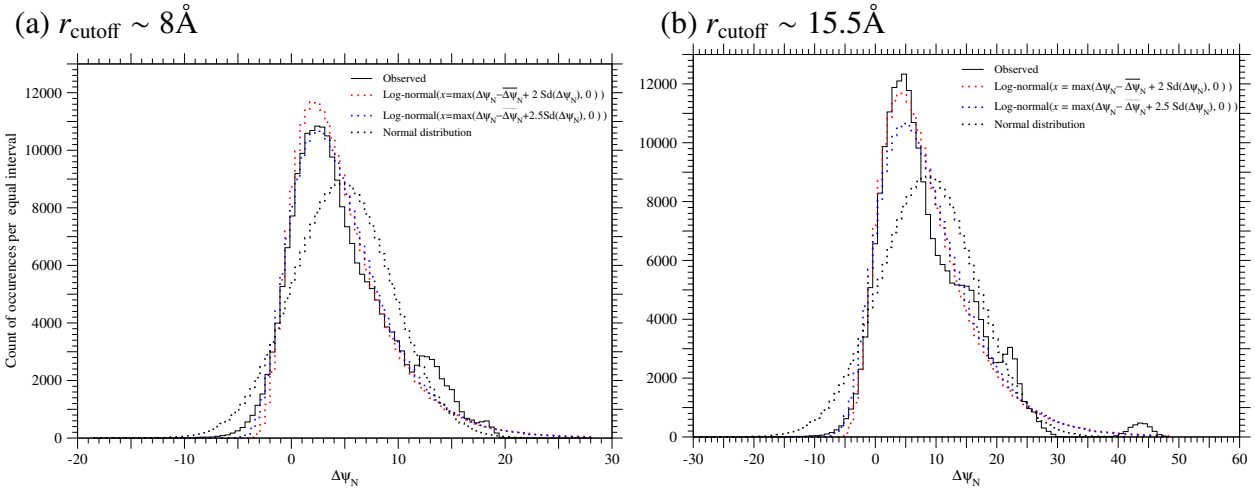

Figure S.22: **The observed frequency distribution and the fitted distributions of $\Delta\psi_N$ in the PDZ protein family.** A black solid line indicates the observed frequency distribution of $\Delta\psi_N$ per equal interval in homologous sequences of the PDZ protein family, and red dotted and blue dotted lines indicate the total frequencies of log-normal distributions with $n_{\mathrm{shift}} = 2$ or 2.5 and parameters estimated with the mean and variance of the observed distribution for each protein; see Eqs. (S.79) to (S.83). A black dotted line indicates the total frequencies of normal distributions the mean and variance of which are equal to those of the observed distribution for each protein. 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, which is almost equal to M_{eff} in Table S.1.

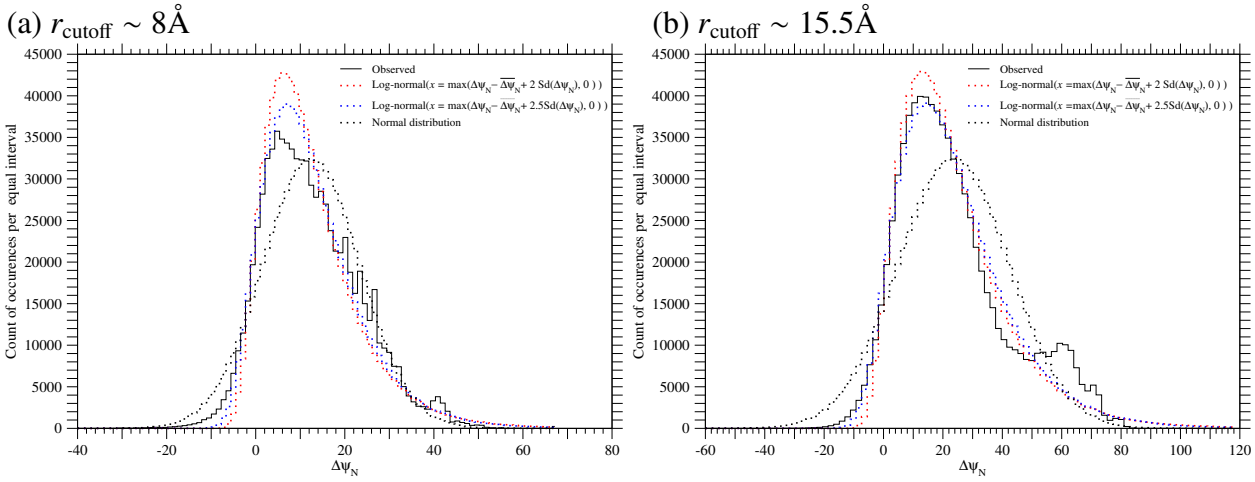

Figure S.23: **The observed frequency distribution and the fitted distribution of $\Delta\psi_N$ in the Methyltransf_5 family of the domain, 1N2X-A:8-292.** A black solid line indicates the observed frequency distribution of $\Delta\psi_N$ per equal interval in homologous sequences of the Methyltransf_5 protein family, and red dotted and blue dotted lines indicate the total frequencies of log-normal distributions with $n_{\mathrm{shift}} = 2$ or 2.5 and parameters estimated with the mean and variance of the observed distribution for each protein; see Eqs. (S.79) to (S.83). A black dotted line indicates the total frequencies of normal distributions the mean and variance of which are equal to those of the observed distribution for each protein. Only representatives of unique sequences, which are at least 20% different from each other, are employed; the total count is equal to 814549 over 354 homologous sequences, which is almost equal to M_{eff} in Table S.1.

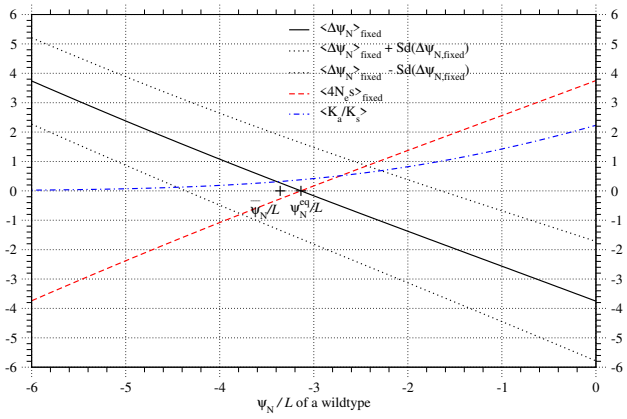

Figure S.24: **The average of $\Delta\psi_N(\simeq \Delta\Delta\psi_{ND})$ over fixed single nucleotide nonsynonymous mutations versus ψ_N/L of a wildtype for the PDZ protein family.** The averages of $\Delta\psi_N(\simeq \Delta\Delta\psi_{ND})$ and $4N_e s$ over the fixed mutants, and the average of $K_a/K_s(\equiv u(s)/u(0))$ over all the mutants are plotted against ψ_N/L of a wildtype by solid, broken, and dash-dot lines, respectively; $q_m = 1/(2\times10^6)$ is assumed. Dotted lines show the values of $\langle\Delta\psi_N\rangle_{\rm fixed} \pm {\rm sd}$, where the sd is the standard deviation of $\Delta\psi_N$ over fixed mutants. Fixation probability has been calculated with $\Delta\Delta\psi_{ND} \simeq \Delta\psi_N$; see Eqs. (S.23) and (S.35). Here the empirical relationships of Eqs. (S.94) and (S.96) are assumed; that is, the mean of $\Delta\psi_N$ linearly decreases as ψ_N increases, but the standard deviation of $\Delta\psi_N$ is constant irrespective of ψ_N. The slope (α_{ψ_N}) and intercept ($-\alpha_{\psi_N}\overline{\psi_N}/L + \overline{\overline{\Delta\psi_N}}$) and the average of ${\rm Sd}(\Delta\psi_N)$ over homologous sequences that are estimated with $r_{\rm cutoff} \sim 8$Å for the PDZ and listed in Table S.2 are employed here. The distribution of $\Delta\psi_N$ due to single nucleotide nonsynonymous mutations is approximated by a log-normal distribution with $n_{\rm shift} = 2.0$; see Eqs. (S.79) to (S.83). The $\psi_N^{\rm eq}$, where $\langle\Delta\Delta\psi_{ND}\rangle_{\rm fixed} \simeq \langle\Delta\psi_N\rangle_{\rm fixed} = 0$, is the stable equilibrium value of ψ_N in the protein evolution of the PDZ protein family. The $\psi_N^{\rm eq}$ is close to the average of ψ_N over homologous sequences ($\overline{\psi_N}$), indicating that the present approximations for $\psi_N^{\rm eq}$ and for $\overline{\psi_N} = \langle\psi_N\rangle_{\boldsymbol{\sigma}}$ are consistent to each other.

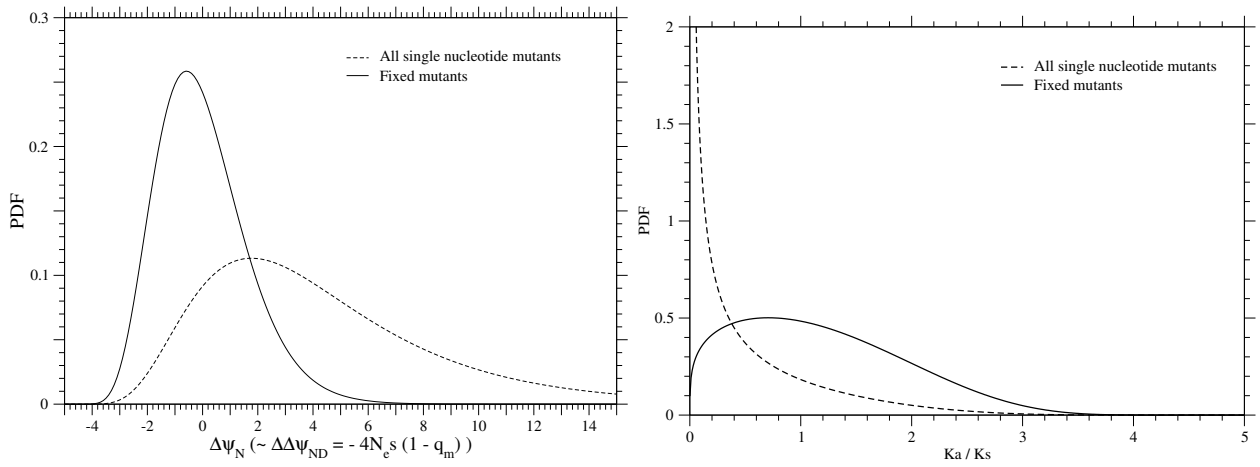

Figure S.25: **PDFs of $\Delta\psi_N(\simeq \Delta\Delta\psi_{ND} = -4N_e s(1-q_m))$ and of K_a/K_s for all single nucleotide nonsynonymous mutants and for their fixed mutants at equilibrium ($\langle\Delta\psi_N\rangle_{\mathbf{fixed}} = 0$) for the PDZ protein family.** K_a/K_s is defined as the ratio of nonsynonymous to synonymous substitution rate per site, $u(s)/u(0)$; see Eq. (S.86). Fixation probability has been calculated with $\Delta\Delta\psi_{ND} \simeq \Delta\psi_N$; see Eqs. (S.23) and (S.35). The equilibrium value $\psi_N^{\rm eq}$, where $\langle\Delta\Delta\psi_{ND}\rangle_{\rm fixed} \simeq \langle\Delta\psi_N\rangle_{\rm fixed} = 0$, is calculated by using the linear dependency of $\overline{\Delta\psi_N}$ on ψ_N (Eq. (S.94)) and estimated values with $r_{\rm cutoff} \sim 8$Å for the PDZ in Tables S.2. The standard deviation of $\Delta\psi_N$ is approximated to be constant and equal to $\overline{\rm Sd(\Delta\psi_N)}$; see Eq. (S.96). The distribution of $\Delta\psi_N$ due to single nucleotide nonsynonymous mutations is approximated by a log-normal distribution with $n_{\rm shift} = 2.0$; see Eqs. (S.79) to (S.83).

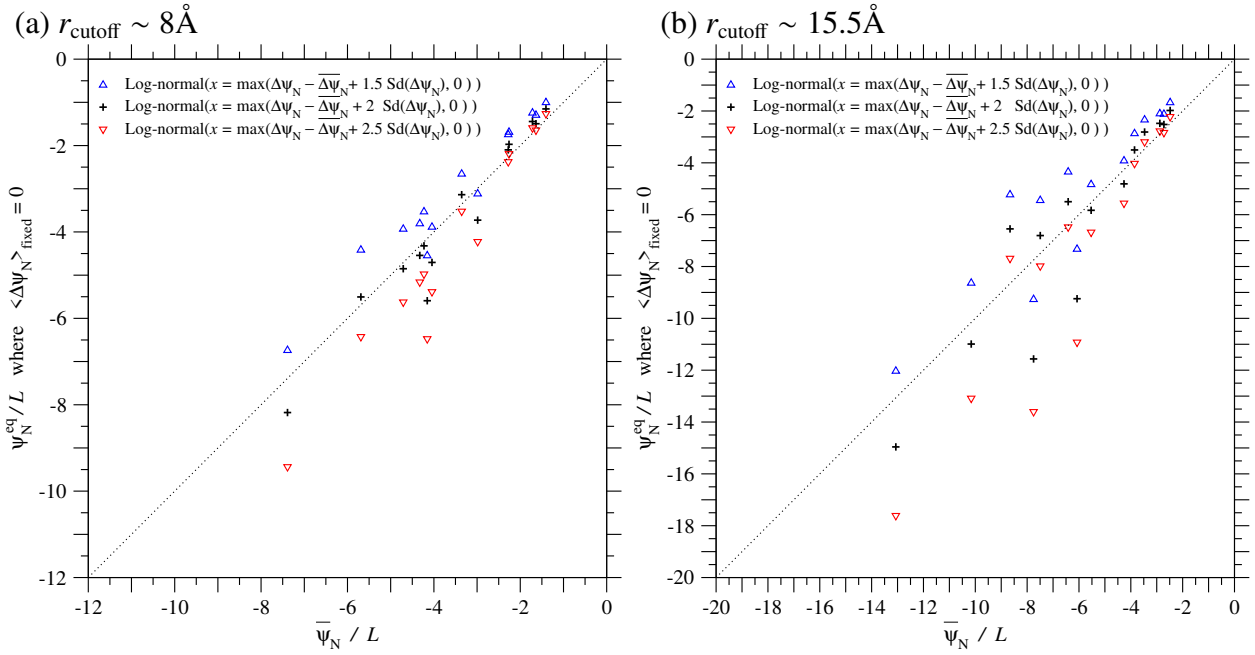

Figure S.26: **The equilibrium value of ψ_N/L, where $\langle\Delta\psi_N\rangle_{\mathbf{fixed}} = 0$, is plotted against the average of ψ_N/L over homologous sequences for each protein family.** The cutoff distances, (a) $r_{\mathrm{cutoff}} = 8$Å and (b) $r_{\mathrm{cutoff}} = 15.5$Å, are employed to estimate ψ_N of each protein family. The equilibrium values ψ_N^{eq}, where $\langle\Delta\psi_N\rangle_{\mathrm{fixed}} = 0$, are calculated by using the linear dependency of $\overline{\Delta\psi_N}$ on ψ_N (Eq. (S.94)) and estimated values with $r_{\mathrm{cutoff}} \sim 8$ or 15.5Å in Tables S.2 or S.5. The standard deviation of $\Delta\psi_N$ is approximated to be constant and equal to $\mathrm{Sd}(\Delta\psi_N)$; see Eq. (S.96). Plus, upper triangle, and lower triangle marks indicate the cases of log-normal distributions with $n_{\mathrm{shift}} = 1.5, 2.0$, and 2.5 employed to approximate the distribution of $\Delta\psi_N$, respectively; see Eqs. (S.79) to (S.83).

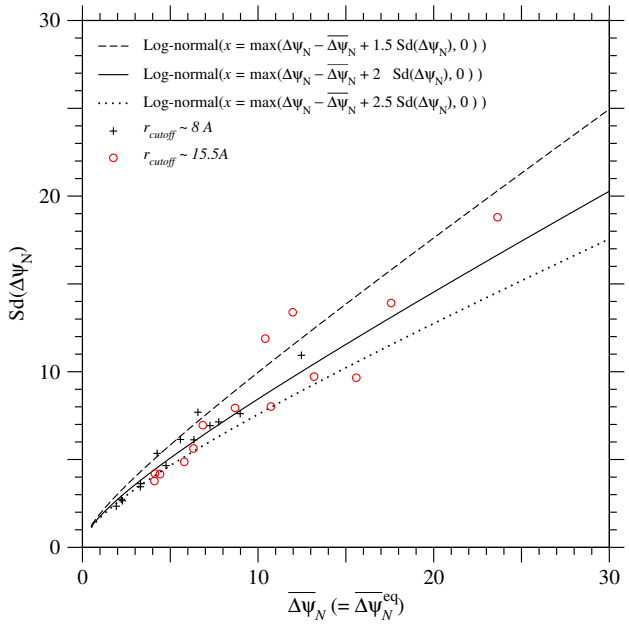

Figure S.27: **Relationship between the mean and the standard deviation of $\Delta\psi_N$ due to single nucleotide nonsynonymous mutations at equilibrium,** $\langle\Delta\psi_N\rangle_{\mathbf{fixed}} = 0$ The standard deviation of $\Delta\psi_N$ that satisfies $\langle\Delta\psi_N\rangle_{\rm fixed} = 0$ is plotted against its mean, $\overline{\Delta\psi_N}$. Broken, solid, and dotted lines indicate the cases of log-normal distributions with $n_{\rm shift} = 1.5, 2.0$ and 2.5 employed to approximate the distribution of $\Delta\psi_N$, respectively; see Eqs. (S.79) to (S.83). Plus and open circle marks indicate the averages, $\overline{\overline{\Delta\psi_N}}$ and $\overline{{\rm Sd}(\Delta\psi_N)}$, over homologous sequences in each protein family for $r_{\rm cutoff} \sim 8$ and 15.5Å, respectively; see Tables S.2 and S.5.

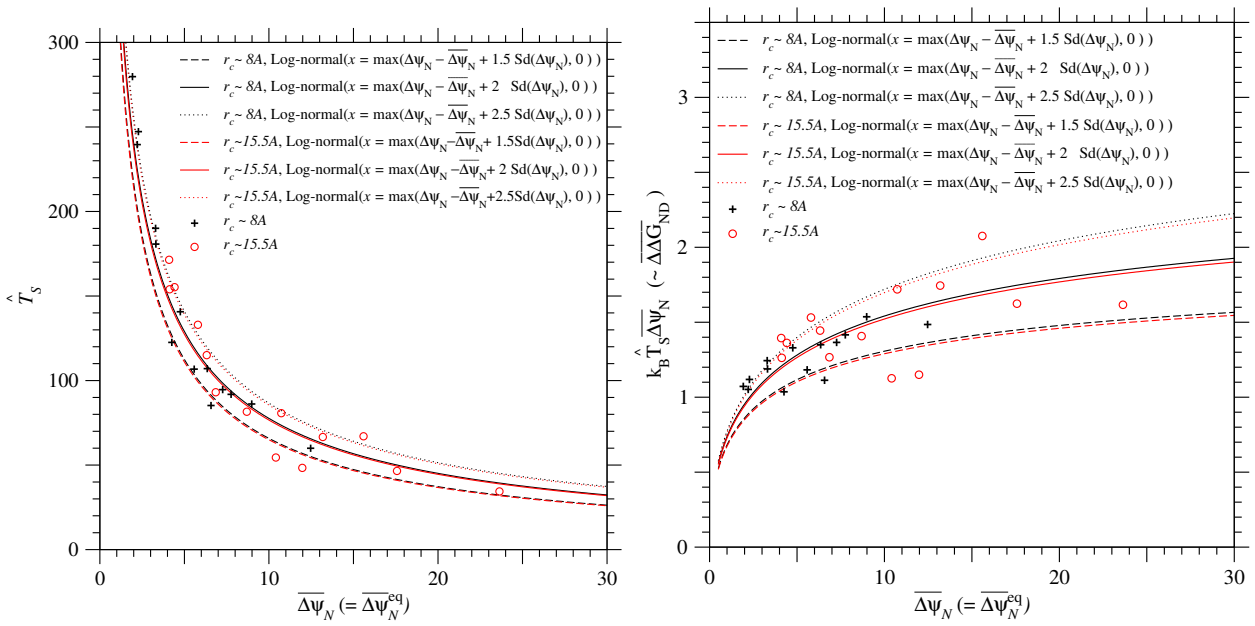

Figure S.28: **Relationships between $\hat{T}_s$ and $\overline{\Delta\psi_N}$ and between $k_B\hat{T}_s\overline{\Delta\psi_N}(\simeq \overline{\Delta\Delta G_{ND}})$ and $\overline{\Delta\psi_N}$ at equilibrium, $\langle\Delta\psi_N\rangle_{\mathbf{fixed}} = 0$.** The estimate $\hat{T}_s(= (\hat{T}_s\overline{Sd}(\Delta\psi_N))_{PDZ}/Sd(\Delta\psi_N))$ of effective temperature for selection and the estimate of mean folding free energy change, $k_B\hat{T}_s\overline{\Delta\psi_N}(= k_B(\hat{T}_s\overline{Sd}(\Delta\psi_N))_{PDZ}/Sd(\Delta\psi_N) \cdot \overline{\Delta\psi_N} \simeq \overline{\Delta\Delta G_{ND}})$, are plotted against $\overline{\Delta\psi_N}$ under the condition of $\langle\Delta\psi_N\rangle_{\text{fixed}} = 0$. The T_s is estimated in relative to the T_s of the PDZ family in the approximation that the standard deviation of ΔG_N due to single nucleotide nonsynonymous mutations is constant irrespective of protein families; see Eq. (S.98). Broken, solid, and dotted lines indicate the cases of log-normal distributions with $n_{\text{shift}} = 1.5, 2.0$ and 2.5 employed to approximate the distribution of $\Delta\psi_N$, respectively; see Eqs. (S.79) to (S.83). Plus and open circle marks indicate those estimates against the average of $\overline{\Delta\psi_N}$ over homologous sequences for each protein family with $r_{\text{cutoff}} \sim 8$ and 15.5Å, respectively; see Tables S.2 and S.5. The curves for $r_{\text{cutoff}} \sim 8$ and 15.5Å almost overlap with each other, because the estimates of $(\hat{T}_s\overline{Sd}(\Delta\psi_N))_{PDZ}$ for the PDZ with $r_{\text{cutoff}} \sim 8$ and 15.5Å are almost equal to each other.

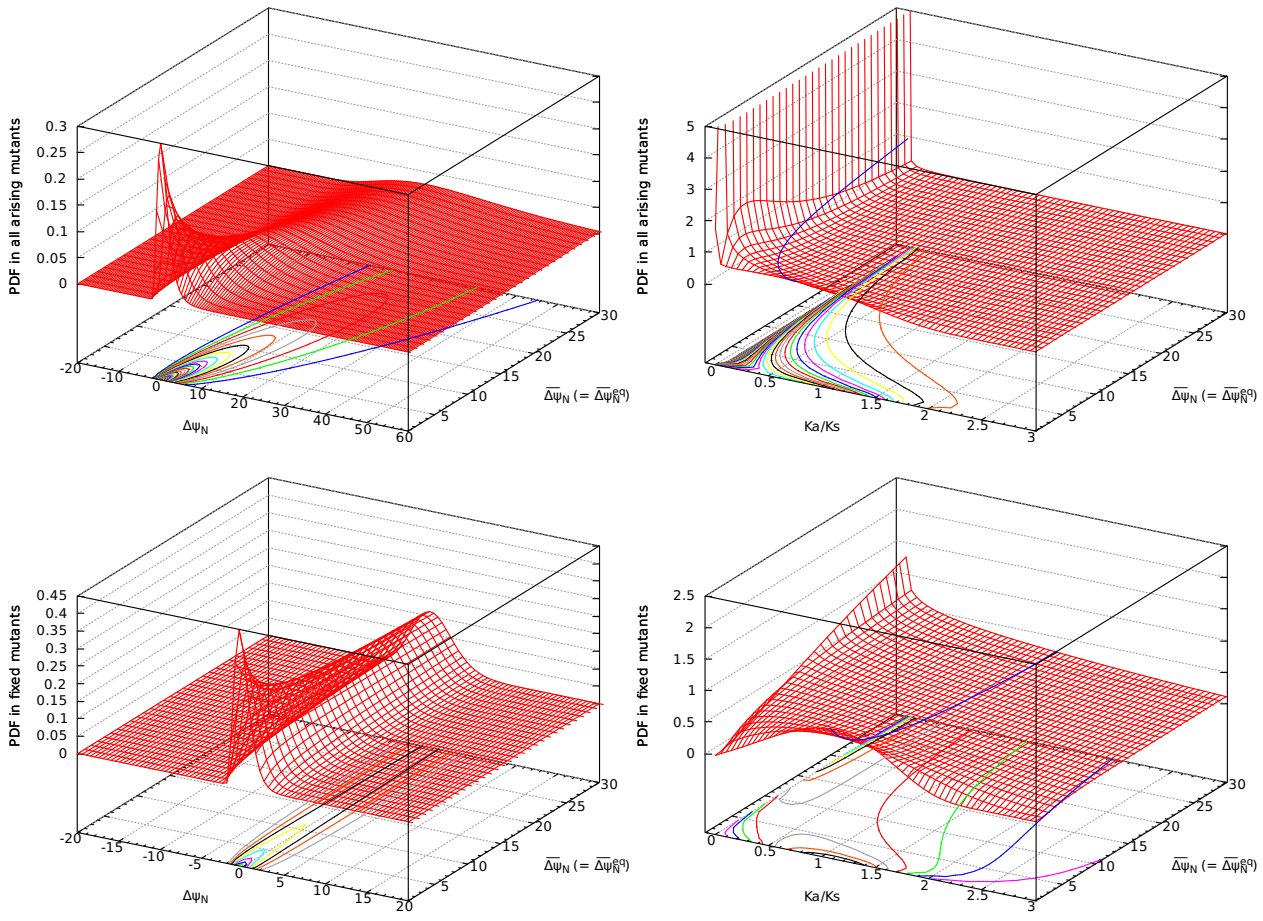

Figure S.29: **PDFs of $\Delta\psi_N$ (left) and K_a/K_s (right) in all singe nucleotide nonsynonymous mutants (upper) and in their fixed mutants (lower) as a function of $\Delta\psi_N$ at equilibrium, $\langle\Delta\psi_N\rangle_{\mathbf{fixed}} = 0$.** Fixation probability has been calculated with $\Delta\Delta\psi_{ND} \simeq \Delta\psi_N$; see Eqs. (S.23) and (S.35). The distribution of $\Delta\psi_N$ due to single nucleotide nonsynonymous mutations is approximated by a log-normal distribution with $n_{\rm shift} = 2.0$; see Eqs. (S.79) to (S.83). The standard deviation of $\Delta\psi_N$ is determined to satisfy $\langle\Delta\psi_N\rangle_{\rm fixed} = 0$ at $\overline{\Delta\psi_N} = \overline{\Delta\psi_N}^{\rm eq}$.

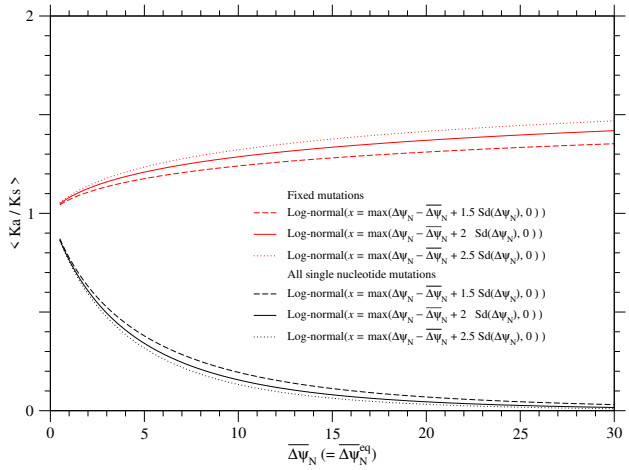

Figure S.30: **The averages of K_a/Ks over all single nucleotide nonsynonymous mutations and over their fixed mutations as a function of $\overline{\Delta\psi_N}$ at equilibrium, $\langle\Delta\psi_N\rangle_{\mathbf{fixed}} = 0$.** Black and red lines indicate $\langle K_a/K_s\rangle$ and $\langle K_a/K_s\rangle_{\rm fixed}$, respectively. Fixation probability has been calculated with $\Delta\Delta\psi_{ND} \simeq \Delta\psi_N$; see Eqs. (S.23) and (S.35). Broken, solid, and dotted lines indicate the cases of log-normal distributions with $n_{\rm shift} = 1.5, 2.0$ and 2.5 employed to approximate the distribution of $\Delta\psi_N$, respectively; see Eqs. (S.79) to (S.83). The standard deviation of $\Delta\psi_N$ is determined to satisfy $\langle\Delta\psi_N\rangle_{\rm fixed} = 0$ at $\overline{\Delta\psi_N} = \overline{\Delta\psi_N}^{\rm eq}$.

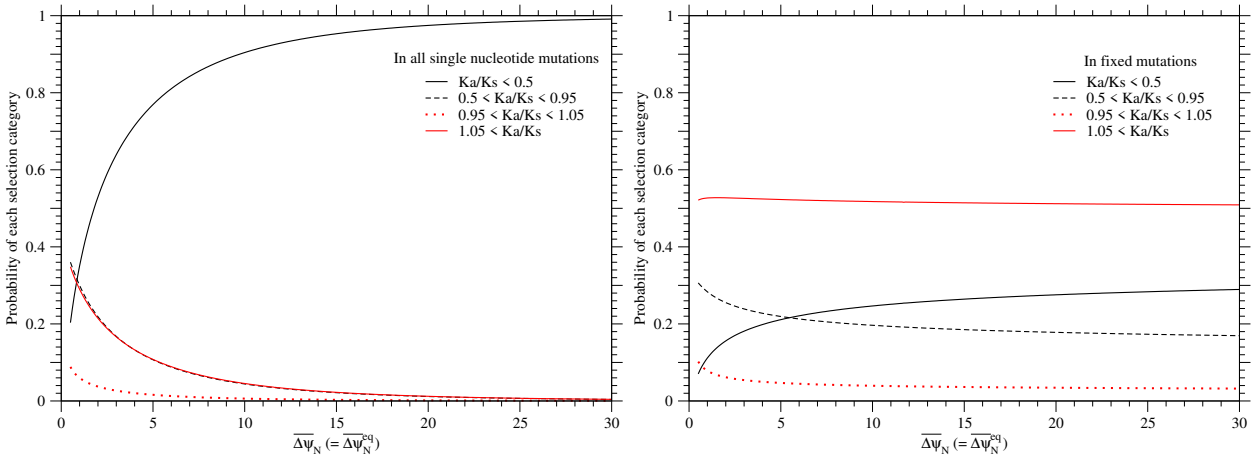

Figure S.31: **The probabilities of each selection category in all single nucleotide nonsynonymous mutations and in their fixed mutations as a function of $\overline{\Delta\psi_N}$ at equilibrium, $\langle\Delta\psi_N\rangle_{\mathbf{fixed}} = 0$.** The left and right figures are for single nucleotide nonsynonymous mutations and for their fixed mutations, respectively. Red solid, red dotted, black broken, and black solid lines indicate positive, neutral, slightly negative and negative selection categories, respectively; the values of K_a/K_s are divided arbitrarily into four categories, $K_a/K_s > 1.05$, $1.05 > K_a/K_s > 0.95$, $0.95 > K_a/K_s > 0.5$, and $0.5 > K_a/K_s$, which correspond to their selection categories, respectively. Fixation probability has been calculated with $\Delta\Delta\psi_{ND} \simeq \Delta\psi_N$; see Eqs. (S.23) and (S.35). The distribution of $\Delta\psi_N$ due to single nucleotide nonsynonymous mutations is approximated by a log-normal distribution with $n_{\rm shift} = 2.0$; see Eqs. (S.79) to (S.83). The standard deviation of $\Delta\psi_N$ is determined to satisfy $\langle\Delta\psi_N\rangle_{\rm fixed} = 0$ at $\overline{\Delta\psi_N}(= \overline{\Delta\psi_N}^{\rm eq})$.

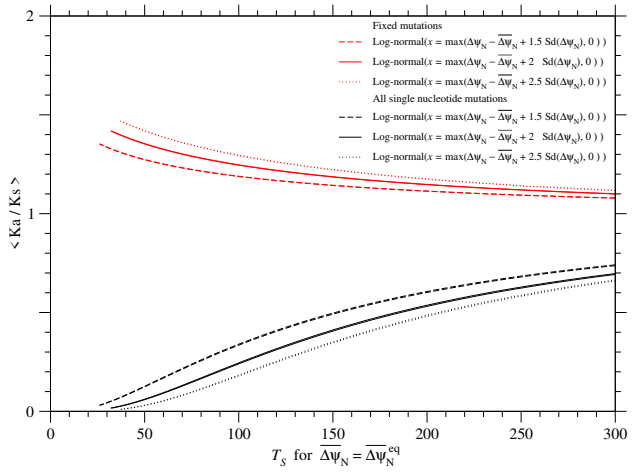

Figure S.32: **The averages of K_a/Ks over all single nucleotide nonsynonymous mutations and over their fixed mutations as a function of the effective temperature of selection, $T_s(= (T_s\overline{Sd}(\Delta\psi_N))_{PDZ}/Sd(\Delta\psi_N))$, at equilibrium, $\langle\Delta\psi_N\rangle_{\mathbf{fixed}} = 0$.** Black and red lines indicate $\langle K_a/K_s\rangle$ and $\langle K_a/K_s\rangle_{\rm fixed}$, respectively. Fixation probability has been calculated with $\Delta\Delta\psi_{ND} \simeq \Delta\psi_N$; see Eqs. (S.23) and (S.35). The distribution of $\Delta\psi_N$ due to single nucleotide nonsynonymous mutations is approximated by a log-normal distribution with $n_{\rm shift} = 2.0$; see Eqs. (S.79) to (S.83). The standard deviation of $\Delta\psi_N$ is determined to satisfy $\langle\Delta\psi_N\rangle_{\rm fixed} = 0$ at $\overline{\Delta\psi_N}(= \overline{\Delta\psi}_N^{\rm eq})$. The T_s is estimated in the scale relative to the T_s of the PDZ family in the approximation that the standard deviation of ΔG_N due to single nucleotide nonsynonymous mutations is constant irrespective of protein families; see Eq. (S.98). Broken, solid, and dotted lines indicate the cases of log-normal distributions with $n_{\rm shift} = 1.5, 2.0$ and 2.5 employed to approximate the distribution of $\Delta\psi_N$, respectively; see Eqs. (S.79) to (S.83). The curves for $r_{\rm cutoff} \sim 8$ and 15.5Å almost overlap with each other, because the estimates of $(\hat{T}_s\overline{Sd}(\Delta\psi_N))_{PDZ}$ for the PDZ with $r_{\rm cutoff} \sim 8$ and 15.5Å are almost equal to each other.