Journal of Theoretical Biology 433 (2017) 21–38

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/jtbi

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

Sanzo Miyazawa

6-5-607 Miyanodai, Sakura, Chiba 285–0857, Japan

ARTICLE INFO

Article history:
Received 27 March 2017
Revised 27 July 2017
Accepted 21 August 2017
Available online 24 August 2017

Keywords:
Folding free energy change
Inverse statistical potential
Boltzmann distribution
Selective temperature
Positive selection

ABSTRACT

Assuming that mutation and fixation processes are reversible Markov processes, we prove that the equilibrium ensemble of sequences obeys a Boltzmann distribution with $\exp(4N_e m(1-1/(2N)))$, where m is Malthusian fitness and N_e and N are effective and actual population sizes. On the other hand, the probability distribution of sequences with maximum entropy that satisfies a given amino acid composition at each site and a given pairwise amino acid frequency at each site pair is a Boltzmann distribution with $\exp(-\psi_N)$, where the evolutionary statistical energy ψ_N is represented as the sum of one body (h) (compositional) and pairwise (J) (covariational) interactions over all sites and site pairs. A protein folding theory based on the random energy model (REM) indicates that the equilibrium ensemble of natural protein sequences is well represented by a canonical ensemble characterized by $\exp(-\Delta G_{ND}/k_B T_s)$ or by $\exp(-G_N/k_B T_s)$ if an amino acid composition is kept constant, where $\Delta G_{ND} \equiv G_N - G_D$, G_N and G_D are the native and denatured free energies, and T_s is the effective temperature representing the strength of selection pressure. Thus, $4N_e m(1-1/(2N))$, $-\Delta\psi_{ND}(\equiv -\psi_N + \psi_D)$, and $-\Delta G_{ND}/k_B T_s$ must be equivalent to each other. With h and J estimated by the DCA program, the changes ($\Delta\psi_N$) of ψ_N due to single nucleotide nonsynonymous substitutions are analyzed. The results indicate that the standard deviation of $\Delta G_N (= k_B T_s \Delta\psi_N)$ is approximately constant irrespective of protein families, and therefore can be used to estimate the relative value of T_s. Glass transition temperature T_g and ΔG_{ND} are estimated from estimated T_s and experimental melting temperature (T_m) for 14 protein domains. The estimates of ΔG_{ND} agree with their experimental values for 5 proteins, and those of T_s and T_g are all within a reasonable range. In addition, approximating the probability density function (PDF) of $\Delta\psi_N$ by a log-normal distribution, PDFs of $\Delta\psi_N$ and K_a/K_s, which is the ratio of nonsynonymous to synonymous substitution rate per site, in all and in fixed mutants are estimated. The equilibrium values of ψ_N, at which the average of $\Delta\psi$ in fixed mutants is equal to zero, well match ψ_N averaged over homologous sequences, confirming that the present methods for a fixation process of mutations and for the equilibrium ensemble of ψ_N give a consistent result with each other. The PDFs of K_a/K_s at equilibrium confirm that T_s negatively correlates with the amino acid substitution rate (the mean of K_a/K_s) of protein. Interestingly, stabilizing mutations are significantly fixed by positive selection, and balance with destabilizing mutations fixed by random drift, although most of them are removed from population. Supporting the nearly neutral theory, neutral selection is not significant even in fixed mutants.

1. Introduction

Natural proteins can fold their sequences into unique structures. Protein's stability and foldability result from natural selection and are not typical characteristics of random polymers (Bryngelson and Wolynes, 1987; Pande et al., 1997; Ramanathan and Shakhnovich, 1994; Shakhnovich and Gutin, 1993a; 1993b). Natural selection maintains protein's stability and foldability over evolutionary timescales. On the basis of the random energy model (REM) for protein folding, it was discussed (Ramanathan and Shakhnovich, 1994; Shakhnovich and Gutin, 1993a; 1993b) that the equilibrium ensemble of natural protein sequences in sequence space is well represented by a canonical ensemble characterized by a Boltzmann factor $\exp(-\Delta G_{ND}(\boldsymbol{\sigma})/k_B T_s)$, where $\Delta G_{ND}(\boldsymbol{\sigma})(\equiv G_N(\boldsymbol{\sigma}) - G_D(\boldsymbol{\sigma}))$ is the folding free energy of sequence $\boldsymbol{\sigma}$, G_N and G_D are the free energies of the native and denatured states, k_B is the Boltzmann constant, and T_s is the effective temperature representing the strength of selection pressure and must satisfy $T_s < T_g < T_m$ for natural proteins to fold into unique native structures; T_g is glass transition temperature and T_m is melting tem-

E-mail address: sanzo.miyazawa@gmail.com

http://dx.doi.org/10.1016/j.jtbi.2017.08.018
0022-5193/

perature. The REM also indicates that the free energy of denatured conformations (G_D) is a function of amino acid frequencies only and does not depend on amino acid order, and therefore the Boltzmann factor will be taken as $\exp(-G_N(\boldsymbol{\sigma})/k_BT_s)$, if amino acid frequencies are kept constant. It was shown by lattice Monte Carlo simulations (Shakhnovich, 1994) that lattice protein sequences selected with this Boltzmann factor were not trapped by competing structures but could fold into unique native structures. Selective temperatures were also estimated (Dokholyan and Shakhnovich, 2001) for actual proteins to yield good correlations of sequence entropy between actual protein families and sequences designed with this type of Boltzmann factor.

On the other hand, the maximum entropy principle insists that the probability distribution of sequences in sequence space, which satisfies constraints on amino acid compositions at all sites and on amino acid pairwise frequencies for all site pairs, is a Boltzmann distribution with the Boltzmann factor $\exp(-\psi_N(\boldsymbol{\sigma}))$, where the total interaction $\psi_N(\boldsymbol{\sigma})$ of a sequence $\boldsymbol{\sigma}$ is represented as the sum of one-body (h) (compositional) and pairwise (J) (covariational) interactions between residues in the sequence; $\psi_N(\boldsymbol{\sigma})$ is called the evolutionary statistical energy by Hopf et al. (2017). The inverse statistical potentials, the one-body (h) and pairwise (J) interactions, that satisfy those constraints for homologous sequences have been estimated (Ekeberg et al., 2014; 2013; Marks et al., 2011; Morcos et al., 2011) as one of inverse Potts problems and successfully employed to predict contacting residue pairs in protein structures (Ekeberg et al., 2014; 2013; Hopf et al., 2012; Marks et al., 2011; Miyazawa, 2013; Morcos et al., 2011; Sułkowska et al., 2012). Morcos et al. (2014) noticed that the ψ_N in the Boltzmann factor is the dimensionless energy corresponding to G_N/k_BT_s, and estimated selective temperatures (T_s) for several protein families by comparing the difference ($\Delta\psi_{ND}$) of ψ between the native and the molten globule states with folding free energies (ΔG_{ND}) estimated with associative-memory, water-mediated, structure, and energy model (AWSEM) (Davtyan et al., 2012).

A purpose of the present study is to establish relationships between protein foldability/stability, sequence distribution, and protein fitness. First, we prove that if mutation and fixation processes in protein evolution are reversible Markov processes, the equilibrium ensemble of genes will obey a Boltzmann distribution with the Boltzmann factor $\exp(4N_em(1-1/(2N)))$, where N_e and N are effective and actual population sizes, and m is the Malthusian fitness of a gene. In other words, correspondences between $-\Delta G_{ND}/k_BT_s$, $-\Delta\psi_{ND}(\equiv\psi_N-\psi_D)$ and $4N_em(1-1/(2N))$ are obtained by equating these three Boltzmann distributions with each other; $\psi_D\simeq G_D/k_BT_s+\text{constant}$.

The second purpose is to analyze the effects ($\Delta\psi_N$) of single amino acid substitutions on the evolutionary statistical energy of a protein, and to estimate from the distribution of $\Delta\psi_N$ the effective temperature of natural selection (T_s) and then glass transition temperature (T_g) and folding free energy (ΔG_{ND}) of protein. We estimate the one-body (h) and pairwise (J) interactions with the DCA program, which is available at "http://dca.rice.edu/portal/dca/home", and then analyze the changes ($\Delta\psi_N$) of the evolutionary statistical energy (ψ_N) of a natural sequence due to single amino acid substitutions caused by single nucleotide changes. The data of $\Delta\psi_N$ due to single nucleotide nonsynonymous substitutions for 14 protein domains show that the standard deviation of $\Delta\psi_N$ over all the substitutions at all sites hardly depends on the evolutionary statistical energy (ψ_N) of each homologous sequence and is nearly constant for each protein family, indicating that the standard deviation of $\Delta G_N\simeq k_BT_s\Delta\psi_N$ is nearly constant irrespective of protein families. From this finding, T_s for each protein family has been estimated in relative to T_s for the PDZ family, which is determined by directly comparing $\Delta\Delta\psi_{ND}(\equiv\Delta(\psi_N-\psi_D)\simeq\Delta\psi_N)$ with the experimental values of folding free energy changes, $\Delta\Delta G_{ND}$, due to single amino acid substitutions. Also T_g and ΔG_{ND} for each protein family are estimated on the basis of the REM from the estimated T_s and an experimental melting temperature T_m. The estimates of T_s and T_g are all within a reasonable range, and those of ΔG_{ND} are well compared with experimental ΔG_{ND} for 5 protein families. The present method for estimating T_s is simpler than the method (Morcos et al., 2014) using AWSEM, and also is useful for the prediction of ΔG_{ND}, because the experimental data of ΔG_{ND} are limited in comparison with T_m, and also experimental conditions such as temperature and pH tend to be different among them. In addition, it has been revealed that $\Delta\psi_N$ averaged over all single nucleotide nonsynonymous substitutions is a linear function of ψ_N/L of each homologous sequence, where L is sequence length; the average of $\Delta\psi_N$ decreases as ψ_N/L increases. This characteristic is required for homologous proteins to stay at the equilibrium state of the native conformational energy $G_N\simeq k_BT_s\psi_N$, and indicates a weak dependency (Miyazawa, 2016; Serohijos et al., 2012) of $\Delta\Delta G_{ND}$ on $\Delta G_{ND}/L$ of protein across protein families.

The third purpose is to study an amino acid substitution process in protein evolution, which is characterized by the fitness, $m=-\Delta\psi_{ND}/(4N_e(1-1/(2N)))$. We employ a monoclonal approximation for mutation and fixation processes of genes, in which protein evolution proceeds with single amino acid substitutions fixed at a time in a population. In this approximation, ψ_N of a protein gene attains the equilibrium, $\psi_N=\psi_N^{eq}$, when the average of $\Delta\psi_N(\simeq\Delta\Delta\psi_{ND})$ over singe nucleotide nonsynonymous mutations fixed in a population is equal to zero. Approximating the distribution of $\Delta\psi_N$ due to singe nucleotide nonsynonymous mutations by a log-normal distribution, their distribution for fixed mutants is numerically calculated and used to calculate the averages of various quantities and also the probability density functions (PDF) of K_a/K_s in all arising mutants and also in fixed mutants only; K_a/K_s is defined as the ratio of nonsynonymous to synonymous substitution rate per site. There is a good agreement between the time average (ψ_N^{eq}) and ensemble average ($\langle\psi_N\rangle_{\boldsymbol{\sigma}}$), which is equal to the sample average, $\overline{\psi_N}$, of ψ_N over homologous sequences, supporting the constancy of the standard deviation of $\Delta\psi_N$ assumed in the monoclonal approximation.

We also study protein evolution at equilibrium, $\psi_N=\psi_N^{eq}$. The common understanding of protein evolution has been that amino acid substitutions observed in homologous proteins are neutral (Kimura, 1968; 1969; Kimura and Ohta, 1971; 1974) or slightly deleterious (Ohta, 1973; 1992), and random drift is a primary force to fix amino acid substitutions in population. The PDFs of K_a/K_s in all arising mutations and in their fixed mutations are examined to see how significant each of positive, neutral, slightly negative,and negative selections is. Interestingly, stabilizing mutations are significantly fixed in population by positive selection, and balance with destabilizing mutations that are also significantly fixed by random drift, although most negative mutations are removed from population. Contrary to the neutral theory (Kimura, 1968; 1969; Kimura and Ohta, 1971; 1974) and supporting the nearly neutral theory (Ohta, 1973; 1992; 2002), the proportion of neutral selection is not large even in fixed mutants. It is also confirmed that the effective temperature (T_s) of selection negatively correlates with the amino acid substitution rate (K_a/K_s) of protein at equilibrium.

2. Methods

2.1. Knowledge of protein folding

A protein folding theory (Pande et al., 1997; Ramanathan and Shakhnovich, 1994; Shakhnovich and Gutin, 1993a; 1993b), which is based on a random energy model (REM), indicates that the equilibrium ensemble of amino acid sequences, $\boldsymbol{\sigma}\equiv(\sigma_1,\ldots,\sigma_L)$ 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.

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

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

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

where $p^{mut}(\boldsymbol{\sigma})$ 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 (Pande et al., 1997; Ramanathan and Shakhnovich, 1994; Shakhnovich and Gutin, 1993a; 1993b). 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; $P^{mut}(\boldsymbol{\sigma}) = \prod_i p^{mut}(\sigma_i)$, where $p^{mut}(\sigma_i)$ 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.

$$Z = \int \exp\left(\frac{-E}{k_B T}\right) n(E) dE \tag{4}$$

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

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.

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

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

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

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 (Pande et al., 1997; Ramanathan and Shakhnovich, 1994; Shakhnovich and Gutin, 1993a; 1993b); $-\partial G_D/\partial T|_{T=T_g} = 0$. The conformational entropy per residue ω in the compact denatured state can be represented with T_g; $\omega L = \delta E^2/(2(k_B T_g)^2)$. Thus, unless $T_g < T_m$, a protein will be trapped at local minima on a rugged free energy landscape before it can fold into a unique native structure.

2.2. 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, $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_L)$ 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 (Marks et al., 2011; Morcos et al., 2011).

$$P(\boldsymbol{\sigma}) \propto \exp(-\psi_N(\boldsymbol{\sigma})) \tag{9}$$

$$\psi_N(\boldsymbol{\sigma}) \equiv -\left(\sum_i^L (h_i(\sigma_i) + \sum_{j>i} J_{ij}(\sigma_i, \sigma_j))\right) \tag{10}$$

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

$$\sum_{\boldsymbol{\sigma}} P(\boldsymbol{\sigma})\, \delta_{\sigma_i a_k} = P_i(a_k) \tag{11}$$

$$\sum_{\boldsymbol{\sigma}} P(\boldsymbol{\sigma})\, \delta_{\sigma_i a_k} \delta_{\sigma_j a_l} = P_{ij}(a_k, a_l) \tag{12}$$

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 (Marks et al., 2011; Morcos et al., 2011), or by maximizing a pseudo-likelihood (Ekeberg et al., 2014; 2013). 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. (2) and (9),

$$\psi_N(\boldsymbol{\sigma}) \simeq G_N(\boldsymbol{\sigma})/(k_B T_s) + \text{function of} \boldsymbol{f}(\boldsymbol{\sigma}) \tag{13}$$

$$\psi_D(\boldsymbol{f}(\boldsymbol{\sigma}), T) \simeq G_D(\boldsymbol{f}(\boldsymbol{\sigma}), T)/(k_B T_s) + \text{function of} \boldsymbol{f}(\boldsymbol{\sigma}) \tag{14}$$

$$\Delta\psi_{ND}(\boldsymbol{\sigma}, T) \simeq \Delta G_{ND}(\boldsymbol{\sigma}, T)/(k_B T_s) \tag{15}$$

$$\Delta\psi_{ND}(\boldsymbol{\sigma}, T) \equiv \psi_N(\boldsymbol{\sigma}) - \psi_D(\boldsymbol{f}(\boldsymbol{\sigma}), T) \tag{16}$$

$$\psi_D(\boldsymbol{f}(\boldsymbol{\sigma}), T) \approx \bar{\psi}(\boldsymbol{f}(\boldsymbol{\sigma})) - \delta\psi^2(\boldsymbol{f}(\boldsymbol{\sigma}))\vartheta(T/T_g) T_s/T \tag{17}$$

$$\omega = (T_s/T_g)^2 \delta\psi^2/(2L) \tag{18}$$

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 $\delta E^2 \simeq (k_B T_s)^2 \delta\psi^2$.

2.3. 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 $\boldsymbol{\mu} \equiv (\mu_1, \ldots, \mu_L)$ to $\boldsymbol{\nu}$ satisfies the detailed balance condition

$$P^{mut}(\boldsymbol{\mu})M_{\boldsymbol{\mu\nu}} = P^{mut}(\boldsymbol{\nu})M_{\boldsymbol{\nu\mu}} \tag{19}$$

where $P^{mut}(\boldsymbol{\nu})$ 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).

$$R_{\boldsymbol{\mu\nu}} = 2NM_{\boldsymbol{\mu\nu}}u(s(\boldsymbol{\mu} \to \boldsymbol{\nu})) \tag{20}$$

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

$$2Nu(s) = 2N\frac{1 - e^{-4N_e s q_m}}{1 - e^{-4N_e s}} \tag{21}$$

$$= \frac{u(s)}{u(0)} \qquad \text{with} \qquad q_m = \frac{1}{2N} \tag{22}$$

where N_e is effective population size. Eq. (21) 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

$$s(\boldsymbol{\mu} \to \boldsymbol{\nu}) = m(\boldsymbol{\nu}) - m(\boldsymbol{\mu}) \tag{23}$$

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 $\boldsymbol{\mu}$, $P^{eq}(\boldsymbol{\mu})$, in the total process consisting of mutation and fixation processes is represented by

$$P^{eq}(\boldsymbol{\mu}) = \frac{P^{mut}(\boldsymbol{\mu})\exp(4N_e m(\boldsymbol{\mu})(1 - q_m))}{\sum_{\boldsymbol{\nu}} P^{mut}(\boldsymbol{\nu})\exp(4N_e m(\boldsymbol{\nu})(1 - q_m))} \tag{24}$$

because both the mutation and fixation processes satisfy the detailed balance conditions, Eq. (19) and the following equation, respectively.

$$\exp(4N_e m(\boldsymbol{\mu})(1 - q_m))\, u(s(\boldsymbol{\mu} \to \boldsymbol{\nu})) = \frac{\exp(-4N_e m(\boldsymbol{\mu})q_m) - \exp(-4N_e m(\boldsymbol{\nu})q_m)}{\exp(-4N_e m(\boldsymbol{\mu})) - \exp(-4N_e m(\boldsymbol{\nu}))} \tag{25}$$

$$= \exp(4N_e m(\boldsymbol{\nu})(1 - q_m))\, u(s(\boldsymbol{\nu} \to \boldsymbol{\mu})) \tag{26}$$

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

2.4. Relationships between $\mathrm{m}(\boldsymbol{\sigma})$, $\psi_{\mathrm{N}}(\boldsymbol{\sigma})$, and $\Delta\mathrm{G}_{\mathrm{ND}}(\boldsymbol{\sigma})$ of protein sequence

From Eqs. (1), (9), and (24), 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.

$$P^{eq}(\boldsymbol{\mu}) = \frac{P^{mut}(\boldsymbol{\mu})\exp(4N_e m(\boldsymbol{\mu})(1 - q_m))}{\sum_{\nu} P^{mut}(\boldsymbol{\nu})\exp(4N_e m(\boldsymbol{\nu})(1 - q_m)))} \tag{27}$$

$$= \frac{P^{mut}(\overline{\boldsymbol{\mu}})\exp(-(\psi_N(\boldsymbol{\mu}) - \psi_D(\overline{\boldsymbol{f}(\boldsymbol{\mu})}, T)))}{\sum_{\boldsymbol{\nu}} P^{mut}(\overline{\boldsymbol{\nu}})\exp(-(\psi_N(\boldsymbol{\nu}) - \psi_D(\overline{\boldsymbol{f}(\boldsymbol{\nu})}, T)))} \tag{28}$$

$$\simeq \frac{P^{mut}(\boldsymbol{\mu})\exp(-\Delta G_{ND}(\boldsymbol{\mu}, T)/(k_B T_s))}{\sum_{\nu} P^{mut}(\boldsymbol{\nu})\exp(-\Delta G_{ND}(\boldsymbol{\nu}, T)/(k_B T_s))} \tag{29}$$

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

$$4N_e m(\boldsymbol{\mu})(1 - q_m) = -\Delta\psi_{ND}(\boldsymbol{\mu}, T) + \text{constant} \tag{30}$$

$$\simeq \frac{-\Delta G_{ND}(\boldsymbol{\mu}, T)}{k_B T_s} + \text{constant} \tag{31}$$

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

$$4N_e s(\boldsymbol{\mu} \to \boldsymbol{\nu})(1 - q_m) = (4N_e m(\boldsymbol{\nu}) - 4N_e m(\boldsymbol{\mu}))(1 - q_m) \tag{32}$$

$$= -(\Delta\psi_{ND}(\boldsymbol{\nu}, T) - \Delta\psi_{ND}(\boldsymbol{\mu}, T)) = -(\psi_N(\boldsymbol{\nu}) - \psi_N(\boldsymbol{\mu})) \tag{33}$$

$$\simeq -(\Delta G_{ND}(\boldsymbol{\nu}, T) - \Delta G_{ND}(\boldsymbol{\mu}, T))/(k_B T_s) = -(G_N(\boldsymbol{\nu}) - G_N(\boldsymbol{\mu}))/(k_B T_s) \tag{34}$$

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. (28), and other sequences may be ignored.

Eq. (33) 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. (34) 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.

2.5. The ensemble average of folding free energy, $\Delta\mathrm{G}_{\mathrm{ND}}(\boldsymbol{\sigma}, \mathrm{T})$, over sequences

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

$$\langle \Delta G_{ND}(\boldsymbol{\sigma}, T) \rangle_{\sigma} \tag{35}$$

$$\equiv \left[\sum_{\boldsymbol{\sigma}} \Delta G_{ND}(\boldsymbol{\sigma}, T) P^{mut}(\boldsymbol{\sigma}) \exp\left(-\frac{\Delta G_{ND}(\boldsymbol{\sigma}, T)}{k_B T_s}\right)\right] \Big/ \left[\sum_{\boldsymbol{\sigma}} P^{mut}(\boldsymbol{\sigma}) \exp\left(-\frac{\Delta G_{ND}(\boldsymbol{\sigma}, T)}{k_B T_s}\right)\right] \tag{36}$$

$$\approx \left[\sum_{\boldsymbol{\sigma}\,|\,\boldsymbol{f}(\boldsymbol{\sigma}) = \overline{\boldsymbol{f}(\boldsymbol{\sigma}_N)}} G_N(\boldsymbol{\sigma}) \exp\left(-\frac{G_N(\boldsymbol{\sigma})}{k_B T_s}\right)\right] \Big/ \left[\sum_{\boldsymbol{\sigma}\,|\,\boldsymbol{f}(\boldsymbol{\sigma}) = \overline{\boldsymbol{f}(\boldsymbol{\sigma}_N)}} \exp\left(-\frac{G_N(\boldsymbol{\sigma})}{k_B T_s}\right)\right] - G_D(\overline{\boldsymbol{f}(\boldsymbol{\sigma}_N)}, T) \tag{37}$$

$$= \langle G_N(\boldsymbol{\sigma})\rangle_{\boldsymbol{\sigma}} - G_D(\overline{\boldsymbol{f}(\boldsymbol{\sigma}_N)}, T) \tag{38}$$

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. (37), 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).

$$\langle G_N(\boldsymbol{\sigma})\rangle_{\boldsymbol{\sigma}} \approx \frac{\int E \exp(-E/(k_B T_s))\, n(E)\, dE}{\int \exp(-E/(k_B T_s))\, n(E)\, dE} \tag{39}$$

$$= \bar{E}(\overline{\boldsymbol{f}(\boldsymbol{\sigma}_N)}) - \delta E^2(\overline{\boldsymbol{f}(\boldsymbol{\sigma}_N)})/(k_B T_s) \tag{40}$$

$$\langle \psi_N(\boldsymbol{\sigma})\rangle_{\boldsymbol{\sigma}} \equiv \left[\sum_{\boldsymbol{\sigma}} \psi_{ND}(\boldsymbol{\sigma}) \exp(-\psi_N(\boldsymbol{\sigma}))\right] \Big/ \left[\sum_{\boldsymbol{\sigma}} \exp(-\psi_N(\boldsymbol{\sigma}))\right] \tag{41}$$

$$\approx \bar{\psi}(\overline{\boldsymbol{f}(\boldsymbol{\sigma}_N)}) - \delta\psi^2(\overline{\boldsymbol{f}(\boldsymbol{\sigma}_N)}) \tag{42}$$

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.

$$\overline{\Delta G_{ND}(\boldsymbol{\sigma}_N, T)}/(k_B T_s) = \langle \Delta G_{ND}(\boldsymbol{\sigma}, T)\rangle_{\boldsymbol{\sigma}}/(k_B T_s) \tag{43}$$

$$\approx \delta\psi^2(\overline{\boldsymbol{f}(\boldsymbol{\sigma}_N)})\,[\,\vartheta(T/T_g)T_s/T - 1\,] \tag{44}$$

$$\overline{\psi_N(\boldsymbol{\sigma}_N)} \equiv \frac{\sum_{\boldsymbol{\sigma}_N} w_{\boldsymbol{\sigma}_N} \psi_N(\boldsymbol{\sigma}_N)}{\sum_{\boldsymbol{\sigma}_N} w_{\boldsymbol{\sigma}_N}} \tag{45}$$

$$= \langle \psi_N(\boldsymbol{\sigma})\rangle_{\boldsymbol{\sigma}} \tag{46}$$

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

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

$$\vartheta(T_m/T_g)\frac{T_s}{T_m} = \frac{T_s}{2T_m}\left(1 + \frac{T_m^2}{T_g^2}\right) = 1 \quad \text{with } T_s \le T_g \le T_m \tag{47}$$

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

Let us 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 $p(s)$.

$$p(4N_e s) = p(\Delta\Delta\psi_{ND})\,|\frac{d\Delta\Delta\psi_{ND}}{d4N_e s}| = p(\Delta\Delta\psi_{ND})(1 - q_m) \tag{48}$$

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. (33).

The PDF of fixation probability u can be represented by

$$p(u) = p(4N_e s)\frac{d4N_e s}{du} = p(4N_e s)\frac{(e^{4N_e s} - 1)^2 e^{4N_e s(q_m - 1)}}{q_m(e^{4N_e s} - 1) - (e^{4N_e s q_m} - 1)} \tag{49}$$

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

$$\frac{K_a}{K_s} = \frac{u(s)}{u(0)} = \frac{u(s)}{q_m} \tag{50}$$

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

$$p(K_a/K_s) = p(u)\frac{du}{d(K_a/K_s)} = p(u)\, q_m \tag{51}$$

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

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

$$p(\Delta\Delta\psi_{ND,fixed}) = p(\Delta\Delta\psi_{ND})\frac{u(s(\Delta\Delta\psi_{ND}))}{\langle u(s(\Delta\Delta\psi_{ND}))\rangle} \tag{52}$$

$$\langle u\rangle \equiv \int_{-\infty}^{\infty} u(s)p(\Delta\Delta\psi_{ND})d\Delta\Delta\psi_{ND} \tag{53}$$

Likewise, the PDF of selective advantage in fixed mutants is

$$p(4N_e s_{fixed}) = p(4N_e s)\frac{u(s)}{\langle u(s)\rangle} \tag{54}$$

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

$$p(u_{fixed}) = p(u)\frac{u}{\langle u\rangle} \tag{55}$$

$$p((\frac{K_a}{K_s})_{fixed}) = p\left(\frac{K_a}{K_s}\right)\frac{u}{\langle u\rangle} = p\left(\frac{K_a}{K_s}\right)\frac{\frac{K_a}{K_s}}{\langle\frac{K_a}{K_s}\rangle} \tag{56}$$

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

3. Materials

3.1. 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 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

Table 1
Protein families, and structures studied.

Pfam family	UniProt ID	N[a]	N_{eff}[b,c]	M[d]	M_{eff}[c,e]	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_{σ_N}) for a given sequence is equal to the inverse of the number of sequences that are less than 20% different from the given sequence.
[d] The number of unique sequences that include no deletion unless specified. The number in parentheses indicates the maximum number of deletions allowed.
[e] The effective number of unique sequences that include no deletion or at most the specified number of deletions.
[f] The number of residues.
[g] Contacts are calculated in the homodimeric state for these protein.
[h] These proteins consist of two domains, and other ones are single domains.

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 1.

4. Results

First, We describe how one-body and pairwise interactions, h and J, are estimated. Then, the changes of evolutionary statistical energy ($\Delta\psi_N$) due to single nucleotide nonsynonymous changes on natural sequences are analyzed with respect to dependences on the ψ_N of the wildtype sequences. The results indicate that the standard deviation of $\Delta G_N \simeq k_B T_s \Delta\psi_N$ is almost constant over protein families. Hence, the selective temperatures, T_s, of various protein families can be estimated in a relative scale from the standard deviation of $\Delta\psi_N$. The T_s of a reference protein is estimated by comparing the expected values of $\Delta\Delta G_{ND}$ with their experimental values. Folding free energies ΔG_{ND} are estimated from estimated T_s and experimental melting temperature T_m, and compared with their experimental values for 5 protein families. Glass transition temperatures T_g are also estimated from T_s and T_m.

Secondly, based on the distribution of $\Delta\psi_N$, protein evolution is studied. Evolutionary statistical energy (ψ_N) attains the equilibrium when the average of $\Delta\psi_N$ over fixed mutations is equal to zero. The PDF of $\Delta\psi_N$ is approximated by log-normal distributions. The basic relationships are that 1) the standard deviation of $\Delta\psi_N$ is constant specific to a protein family, and 2) the mean of $\Delta\psi_N$ linearly depends on ψ_N. The equilibrium value of ψ_N is shown to agree with the mean of ψ_N over homologous proteins in each protein family. In the present approximation, the standard deviation of $\Delta\psi_N$ and selective temperature T_s at the equilibrium are simple functions of the equilibrium value of mean $\Delta\psi_N$, $\overline{\Delta\psi_N}^{eq}$. Lastly, the probability distribution of K_a/K_s, which is the ratio of nonsynonymous to synonymous substitution rate per site, is analyzed as a function of $\overline{\Delta\psi_N}^{eq}$, in order to examine how significant neutral selection is in the selection maintaining protein stability and foldability. Also, it is confirmed that selective temperature T_s negatively correlates with the mean of K_a/K_s, which represents the evolutionary rate of protein.

4.1. Important parameters in the estimations of one-body and pairwise interactions, h and J, and of the evolutionary statistical energy, $\psi_N(\boldsymbol{\sigma})$

The one-body (h) and pairwise (J) interactions for amino acid order in a protein sequence are estimated here by the DCA method (Marks et al., 2011; Morcos et al., 2011), although there are multiple methods for estimating them (Ekeberg et al., 2014; 2013). In the case of the DCA method, the ratio of pseudocount ($0 \le p_c \le 1$) defined in Eqs. (S.70) and (S.71) is a parameter and controls the values of the ensemble and sample averages of ψ_N in sequence space, $\langle\psi_N(\boldsymbol{\sigma})\rangle_{\boldsymbol{\sigma}}$ in Eq. (42) and $\overline{\psi_N(\boldsymbol{\sigma}_N)}$ in Eq. (45); a weight for observed counts is defined to be equal to $(1 - p_c)$. Sample average means the average over all homologous sequences with a weight for each sequence to reduce phylogenetic biases. An appropriate value must be chosen for the ratio of pseudocount in a reasonable manner.

Another problem is that the estimates of h and J(Marks et al., 2011; Morcos et al., 2011) may be 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. (2014).

$$\hat{J}^q_{ij}(a_k, a_l) = J^q_{ij}(a_k, a_l) H(r_{cutoff} - r_{ij}) \tag{57}$$

where J^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 at sites i and j in a protein structure, and r_{cutoff} is a distance threshold for residue pairwise interactions. The one-body interactions $h_i(a_k)$ are estimated in the isolated two-state model (Morcos et al., 2011) rather than the mean field approximation; see the Method section in the Supplement for details. The zero-sum gauge is employed to represent h and J; $\sum_k \hat{h}^s_i(a_k) = \sum_k \sum_l \hat{J}^s_{ij}(a_k, a_l) = 0$ in the zero-sum gauge.

Candidates for the cutoff distance may be about 8 Å for the first interaction shell and 15–16 Å for the second interaction shell between residues; distance between the centers of side chain atoms is employed for residue distance. Here both the distances are tested for the cutoff distance. Pseudocount in the Bayesian statis-

Table 2
Parameter values for $r_{cutoff} \sim 8$ 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]	$\overline{\mathrm{Sd}(\Delta\psi_N)} \pm$[c] $\mathrm{Sd}(\mathrm{Sd}(\Delta\psi_N))$	r_{ψ_N} for $\overline{\Delta\psi_N}$[d]	α_{ψ_N}	r_{ψ_N} for $\mathrm{Sd}(\Delta\psi_N)$[e]	α_{ψ_N}
HTH_3	53	0.18	7.43	8.22	−0.1997	2.7926	−2.9861	4.2572	5.3503 ± 0.5627	−0.961	−1.5105	−0.598	−0.9888
Nitroreductase	73	0.23	6.38	8.25	−0.1184	2.1597	−2.2788	3.3115	3.6278 ± 0.2804	−0.939	−1.3371	−0.426	−0.3721
SBP_bac_3	218	0.25	9.23	8.10	−0.1000	2.1624	−2.2618	3.2955	3.4496 ± 0.2742	−0.980	−1.5286	−0.841	−0.7876
SBP_bac_3	94	0.37	8.00	7.90	−0.1634	1.2495	−1.4054	1.9291	2.3436 ± 0.1901	−0.959	−1.3938	−0.634	−0.4815
OmpA	95	0.169	8.00	8.20	−0.2457	3.9093	−4.1542	6.5757	7.6916 ± 0.3078	−0.957	−1.5694	−0.410	−0.3804
DnaB	98	0.235	9.65	8.17	−0.2284	3.9976	−4.2291	6.3502	6.1244 ± 0.3245	−0.965	−1.4509	−0.495	−0.4198
LysR_substrate	191	0.235	8.59	7.98	−0.2241	1.4888	−1.7173	2.2784	2.6519 ± 0.1445	−0.964	−1.3347	−0.541	−0.5664
LysR_substrate	103	0.265	8.84	8.25	−0.2244	1.4144	−1.6379	2.2110	2.7371 ± 0.2055	−0.982	−1.4159	−0.727	−0.5307
Methyltransf_5	285	0.13	7.99	7.78	−0.1462	7.2435	−7.3887	12.4689	10.9352 ± 0.3030	−0.981	−1.9140	−0.122	−0.0783
Methyltransf_5	80	0.18	6.78	7.85	−0.1763	5.5162	−5.6896	8.9849	7.6133 ± 0.4382	−0.944	−1.4824	0.125	0.1141
SH3_1	48	0.14	6.42	8.01	−0.1348	3.9109	−4.0434	5.5792	6.1426 ± 0.2935	−0.919	−1.4061	−0.196	−0.1718
ACBP	80	0.22	9.17	8.24	−0.0525	4.6411	−4.7084	7.7612	7.1383 ± 0.2970	−0.972	−1.5884	−0.335	−0.2235
PDZ	81	0.205	9.06	8.16	−0.2398	3.1140	−3.3572	4.7589	4.6605 ± 0.2255	−0.954	−1.5282	−0.369	−0.3042
Copper-bind	125	0.23	9.50	8.27	−0.0940	4.2450	−4.3272	7.2650	6.9283 ± 0.2316	−0.980	−1.8915	−0.282	−0.2352

[a] The average number of contact residues per site within the cutoff distance; the center of side chain is used to represent a residue.
[b] M unique sequences with no deletions are used with a sample weight (w_{σ_N}) for each sequence; w_{σ_N} is equal to the inverse of the number of sequences that are less than 20% different from a given sequence. The M and the effective number M_{eff} of the sequences are listed for each protein family in Table 1.
[c] The averages of $\overline{\Delta\psi_N}$ and $\mathrm{Sd}(\Delta\psi_N)$, which are the mean and the standard deviation of $\Delta\psi_N$ for a sequence, and the standard deviation of $\mathrm{Sd}(\Delta\psi_N)$ 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. (62).
[e] The correlation and regression coefficients of $\mathrm{Sd}(\Delta\psi_N)$ on ψ_N/L.

tics is determined usually as a function of the number of samples (sequences), although the ratio of pseudocount $p_c = 0.5$ was used for all proteins in the contact prediction (Morcos et al., 2011). Here, an appropriate value for the ratio of pseudocount for the certain cutoff distance, either about 8 Å or 15–16 Å, is chosen for each protein family in such a way that the sample average of the evolutionary statistical energies must be equal to the ensemble average, $\overline{\psi_N} = \langle\psi_N\rangle_{\boldsymbol{\sigma}}$; see Eqs. (42) and (46). As shown in Fig. S.1, the value of r_{cutoff}, where $\overline{\psi_N} = \langle\psi_N\rangle_{\boldsymbol{\sigma}}$ is satisfied, monotonously changes as a function of the ratio of pseudocount p_c. The values of p_c, where $\overline{\psi_N} = \langle\psi_N\rangle_{\boldsymbol{\sigma}}$ is satisfied near the specified values of r_{cutoff}, 8 Å and 15.5 Å, are employed for $r_{cutoff} \simeq 8$ Å and 15.5 Å, respectively. In the present multiple sequence alignment for the PDZ domain, with the ratios of pseudocount $p_c = 0.205$ and $p_c = 0.33$, the sample and ensemble averages agree with each other at the cutoff distances $r_{cutoff} \sim 8$ Å and $r_{cutoff} \sim 15.5$ Å, respectively; see Fig. S.1. In Fig. S.2, 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 are plotted against the cutoff distance for pairwise interactions in the PDZ domain. The reflective correlation coefficient has the maximum at the $r_{cutoff} \sim 8$ Å for $p_c = 0.205$ and at $r_{cutoff} \sim 15.5$ Å for $p_c = 0.33$, indicating that these cutoff distances are appropriate for these ratios of pseudocount. The ratio of pseudocount and a cutoff distance employed are listed for each protein family in Table 2 and S.5 for $r_{cutoff} \sim 8$ and 15.5 Å, respectively. The ratios of pseudocount employed here are all smaller than 0.5, which was reported to be appropriate for contact prediction; by using strong regularization, contact prediction is improved but the generative power of the inferred model is degraded (Barton et al., 2016). In the text, only results with $r_{cutoff} \sim 8$ Å are shown. In a supplement, results with $r_{cutoff} \sim 15.5$ Å are provided and discussed in comparison with the results of $r_{cutoff} \sim 8$ Å.

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

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

$$\Delta\psi_N(\sigma_{j\neq i}^N, \sigma_i^N \to \sigma_i) \equiv \psi_N(\sigma_{j\neq i}^N, \sigma_i) - \psi_N(\boldsymbol{\sigma}_N) \tag{58}$$

$$\Delta\psi_D(\sigma_{j\neq i}^N, \sigma_i^N \to \sigma_i, T) \equiv \psi_D(\sigma_{j\neq i}^N, \sigma_i, T) - \psi_D(\boldsymbol{\sigma}_N, T) \tag{59}$$

$$\begin{aligned}\Delta\Delta\psi_{ND}(\sigma_{j\neq i}^N, \sigma_i^N \to \sigma_i) &\equiv \Delta\psi_N(\sigma_{j\neq i}^N, \sigma_i^N \to \sigma_i) \\ &\quad - \Delta\psi_D(\sigma_{j\neq i}^N, \sigma_i^N \to \sigma_i)\end{aligned} \tag{60}$$

$$\begin{aligned}&\simeq \Delta\psi_N(\sigma_{j\neq i}^N, \sigma_i^N \to \sigma_i) \\ &\quad \text{because } \boldsymbol{f}(\boldsymbol{\sigma}_N) \approx \boldsymbol{f}(\sigma_{j\neq i}^N, \sigma_i)\end{aligned} \tag{61}$$

Here, single amino acid substitutions caused by single nucleotide nonsynonymous mutations are taken into account, unless specified. Let us use a single overline to denote the average of the changes of interaction over all types of single nucleotide nonsynonymous mutations at all sites in a specific native sequence, and a double overline to denote their averages over all homologous sequences in a protein family.

We calculated the ψ_N of the wildtype and $\Delta\psi_N$ due to all types of single nucleotide nonsynonymous substitutions for all homologous sequences, and their means and variances. We have examined the dependence of $\overline{\Delta\Delta\psi_{ND}} \simeq \overline{\Delta\psi_N}$ on the ψ_N of each homologous sequence in each protein family. Fig. 1 for the PDZ family and Figs. S.3 to S.13 for all proteins show that $\overline{\Delta\psi_N}$ is negatively proportional to the ψ_N/L of the wildtype, that is,

$$\begin{aligned}&\overline{\Delta\Delta\psi_{ND}(\sigma_{j\neq i}^N, \sigma_i^N \to \sigma_i)} \simeq \overline{\Delta\psi_N(\sigma_{j\neq i}^N, \sigma_i^N \to \sigma_i)} \\ &\approx \alpha_{\psi_N} \frac{\psi_N(\boldsymbol{\sigma}_N) - \overline{\psi_N(\boldsymbol{\sigma}_N)}}{L} + \overline{\overline{\Delta\psi_N(\sigma_{j\neq i}^N, \sigma_i^N \to \sigma_i)}} \\ &\text{with } \alpha_{\psi_N} < 0\end{aligned} \tag{62}$$

where L is sequence length. This relationship is found in all of the protein families examined here; the correlation and regression coefficients for $r_{cutoff} \sim 8$ and 15.5 Å are listed in Table 2 and S.5, respectively. Most of the correlation coefficients are larger than 0.95, and all are greater than 0.9. It is reasonable that the change of the evolutionary statistical energy ($\Delta\psi_N$) depends on interaction per residue (ψ_N/L) rather than the evolutionary statistical energy

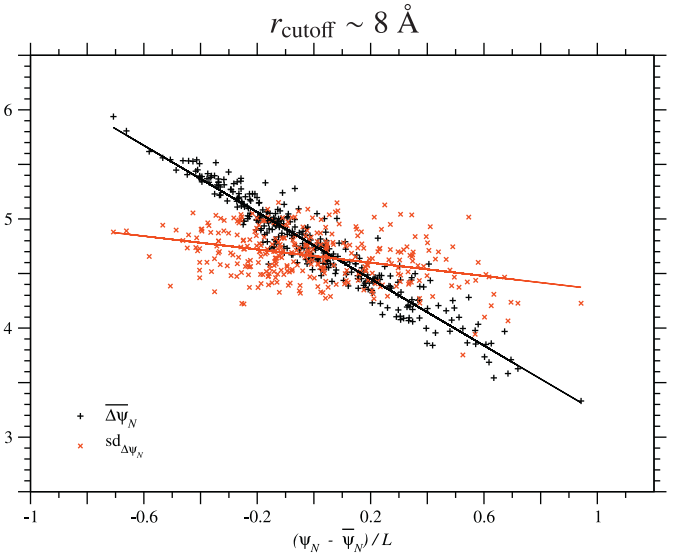

Fig. 1. Correlation between $\Delta\psi_N$ due to single nucleotide nonsynonymous substitutions and ψ_N of homologous sequences in the PDZ domain family. This figure corresponds to the cutoff distance $r_{cutoff} \sim 8$ Å; see Fig. S.12 for $r_{cutoff} \sim 15.5$ Å. Each of the black plus or red cross 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 1. The solid lines show the regression lines for the mean and the standard deviation of $\Delta\psi_N$.

(ψ_N), because interactions change only for one residue substituted in the sequence. Note that the average interactions including a single residue will be equal to $2\psi_N/L$ if all interactions are two-body. The important fact is that the linear dependence of $\Delta\psi_N$ on ψ_N/L shown in Fig. 1 and Table 2 and S.5 is equivalent to the linear dependence of free energy changes caused by single amino acid substitutions on the native conformational energy of the wildtype protein, because the selective temperatures T_S of homologous sequences in a protein family are approximated to be equal.

Is the same type of dependence on ψ_N/L found for the standard deviation of $\Delta\psi_N$ over single nucleotide nonsynonymous substitutions at all sites? Fig. 1, Figs. S.3 to S.13 and Table 2 and S.5 show that the correlation between the standard deviation of $\Delta\psi_N$ and ψ_N of the wildtype is very weak except for Nitroreductase, SBP_bac_3 and LysR_substrate families. Even for these protein families, the standard deviations of $\mathrm{Sd}(\Delta\psi_N)$ are less than 10% of the mean, $\overline{\mathrm{Sd}(\Delta\psi_N)}$; see Table 2 and S.5. Thus, it is indicated that in general the variance/standard deviation of $\Delta\psi_N$ due to single amino acid substitutions is almost constant irrespectively of the ψ_N across homologous sequences. The standard deviations of $\mathrm{Sd}(\Delta\psi_N)$ is relatively large for the HTH_3, because in Fig. S.3 there is a minor sequence group that has a distinguishable value of $\mathrm{Sd}(\Delta\psi_N)$ from the major sequence group.

4.3. Effective temperature $\mathrm{T_s}$ *of selection estimated from the changes of interaction,* $\Delta\psi_\mathrm{N}$*, by single nucleotide nonsynonymous substitutions*

In the previous section, it has been shown that the standard deviation of $\Delta\psi_N$ hardly depends on ψ_N of the wildtype and is nearly constant across homologous sequences in every protein family that has its own characteristic temperature (T_s) for selection pressure, indicating that $\mathrm{Sd}(\Delta\psi_N)$ must be approximated by a function of only k_BT_s. On the other hand, the free energy of the native structure, ΔG_N, must not explicitly depend on k_BT_s, although it may be approximated by a function of G_N. In other words, the following relationships are derived.

$$\begin{aligned}\mathrm{Sd}(\Delta\psi_N(\sigma^N_{j\neq i}, \sigma^N_i \to \sigma_i)) \approx\ & \text{independent of } \psi_N \text{ and} \\ & \text{constant across homologous} \\ & \text{sequences in every protein} \\ & \text{family} \\ =\ & \text{function of } k_BT_s \end{aligned} \tag{63}$$

$$\begin{aligned}\mathrm{Sd}(\Delta G_N(\sigma^N_{j\neq i}, \sigma^N_i \to \sigma_i)) =\ & \text{function that must not explicitly} \\ & \text{depend on } k_BT_s \text{ but } G_N \end{aligned} \tag{64}$$

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

$$\begin{aligned}\mathrm{Sd}(\Delta G_N(\sigma^N_{j\neq i}, \sigma^N_i \to \sigma_i)) &\simeq k_BT_s\,\mathrm{Sd}(\Delta\psi_N(\sigma^N_{j\neq i}, \sigma^N_i \to \sigma_i)) \\ &\approx \text{ constant} \end{aligned} \tag{65}$$

This relationship is consistent with the observation that the standard deviation of $\Delta\Delta G_{ND} (\simeq \Delta G_N)$ is nearly constant irrespectively of protein families (Tokuriki et al., 2007). This relationship allows us to estimate a selective temperature (T_s) for a protein family in a scale relative to that of a reference protein from the ratio of the standard deviation of $\Delta\psi_N$. The PDZ family is employed here as a reference protein, 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 Table 3 and S.6.

$$\begin{aligned}k_B\hat{T}_s = {} & k_B\hat{T}_{s,PDZ} \\ & [\,\overline{\mathrm{Sd}(\Delta\psi_{PDZ}(\sigma^N_{j\neq i}, \sigma^N_i \to \sigma_i))} \,/\, \overline{\mathrm{Sd}(\Delta\psi_N(\sigma^N_{j\neq i}, \sigma^N_i \to \sigma_i))}\,] \end{aligned} \tag{66}$$

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. It will be confirmed in the later section, "the equilibrium value of ψ_N in protein evolution", that the assumption of the constant value specific to each protein family for $\mathrm{Sd}(\Delta\psi_N)$ is appropriate.

4.4. A direct comparison of the changes of interaction, $\Delta\psi_\mathrm{N}(\simeq\Delta\Delta\psi_\mathrm{ND})$*, with the experimental* $\Delta\Delta\mathrm{G_{ND}}$ *due to single amino acid substitutions*

In order to determine the 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 Figs. 2 and S.14. The slope of the least-squares regression line through the origin, which is an estimate of k_BT_s, is equal to $k_B\hat{T}_s = 0.279$ kcal/mol, and the reflective correlation coefficient is equal to 0.93. This estimate 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, which corresponds to 76% of 1.7 kcal/mol (Serohijos et al., 2012) estimated from ProTherm database or 80% of 1.63 kcal/mol (Tokuriki et al., 2007) computationally predicted for single nucleotide mutations by using the FoldX. Using $\overline{\mathrm{Sd}(\Delta\Delta G_{ND})} = 1.30$ estimated from the T_s for PDZ, the absolute values of T_s for other proteins are calculated by Eq. (66) and listed in Table 3; see Table S.6 for $r_{cutoff} \sim 15.5$ Å. The T_s estimated with

Table 3
Thermodynamic quantities estimated with $r_{cutoff} \sim 8$ Å.

Pfam family	r[a]	$k_B\hat{T}_s$[a] (kcal/mol)	$\hat{T}_s$ (°K)	Experimental 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. (18).
[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. (44).

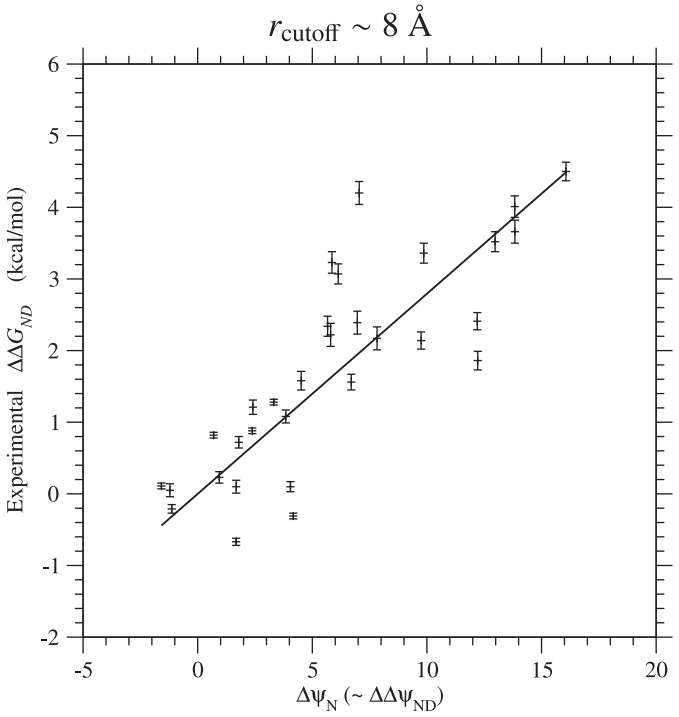

Fig. 2. Regression of the experimental values(Gianni et al., 2007) of folding free energy changes ($\Delta\Delta G_{ND}$) due to single amino acid substitutions on $\Delta\psi_N (\simeq \Delta\Delta\psi_{ND})$ for the same types of substitutions in the PDZ domain. This figure corresponds to the cutoff distance $r_{cutoff} \sim 8$ Å; see Fig. S.14 for $r_{cutoff} \sim 15.5$ Å. The solid line shows the least-squares regression line through the origin with the slope, 0.279 kcal/mol, which is the estimates of k_BT_s. The reflective correlation coefficient is equal to 0.93. The free energies are in kcal/mol units.

$r_{cutoff} \sim 8$ and 15.5 Å are compared with each other in Fig. S.15. 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 $r_{cutoff} = 16$ 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. (2014). The Morcos's estimates of T_s with some exceptions tend to be located between the present estimates with $r_{cutoff} \sim 8$ Å and 15.5 Å which correspond to upper and lower limits for T_s as discussed in the Discussion and the supplement.

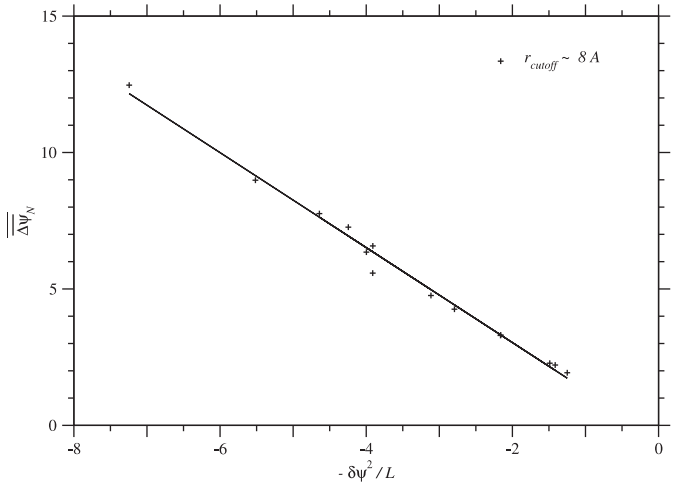

Fig. 3. Dependence of the average of $\overline{\Delta\psi_N}$ due to single nucleotide nonsynonymous substitutions over homologous sequences on $-\delta\psi^2/L$ across protein families. Plus marks indicate the values for each protein family in the case of $r_{cutoff} \sim 8$ Å. The correlation coefficient is equal to 0.995, and the regression line is $\overline{\Delta\psi_N(\sigma^N_{j\neq i}, \sigma^N_i \to \sigma_i)} = -1.74(-\delta\psi^2/L) - 0.445$. See Fig. S.18 for $r_{cutoff} \sim 15.5$ Å.

4.5. Relationship among $\overline{\overline{\Delta\psi_N}}$ of protein families; weak dependency of $\Delta\Delta G_{ND}$ on $\Delta G_{ND}/L$

The weak dependence of $\Delta\Delta G_{ND}$ on ΔG_{ND} was found (Miyazawa, 2016; Serohijos et al., 2012) from the analysis of stability changes due to single amino acid substitutions in proteins, which are collected in the ProTherm database (Kumar et al., 2006). To understand this weak dependence, let us consider the average of $\overline{\Delta\psi_N}$ over homologous sequences in each protein family. The following regression line with $\alpha_{\overline{\psi_N}} = -1.74$ is shown in Fig. 3.

$$\overline{\overline{\Delta\psi_N(\sigma^N_{j\neq i}, \sigma^N_i \to \sigma_i)}} \approx \alpha_{\overline{\psi_N}} \frac{\overline{\psi_N(\boldsymbol{\sigma}_N)} - \bar{\psi}(\overline{\boldsymbol{f}(\boldsymbol{\sigma}_N)})}{L} + \beta_{\overline{\psi_N}} \quad (67)$$

$$= \alpha_{\overline{\psi_N}} \frac{-\delta\psi^2(\overline{\boldsymbol{f}(\boldsymbol{\sigma}_N)})}{L} + \beta_{\overline{\psi_N}} \quad (68)$$

$$\alpha_{\overline{\psi_N}} < 0, \quad \beta_{\overline{\psi_N}} \approx 0 \quad (69)$$

Here, $\overline{\psi_N(\boldsymbol{\sigma}_N)}$ is reduced by $\bar{\psi}$ because the origin of the ψ_N scale is not unique. The correlation between $\overline{\overline{\Delta\psi_N}}$ and $\delta\psi^2/L$ is signif-

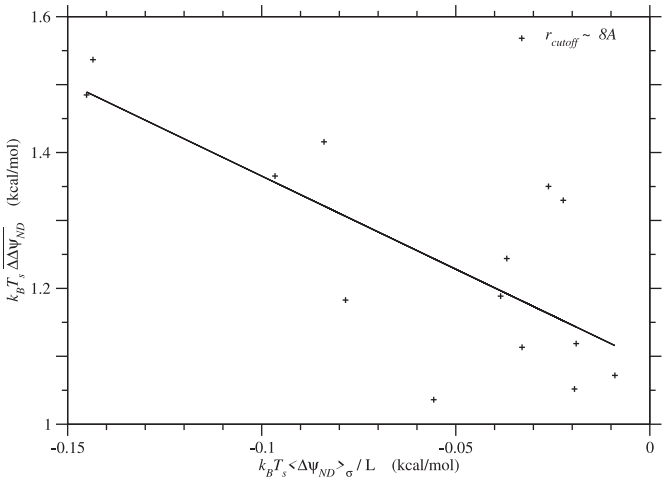

Fig. 4. The sample average of folding free energy change, $\overline{\overline{\Delta\Delta G_{ND}}} \simeq k_B T_s \overline{\overline{\Delta\Delta\psi_{ND}}}$, is plotted against the ensemble average of folding free energy per residue, $\langle \Delta G_{ND} \rangle_{\boldsymbol{\sigma}}/L \simeq k_B T_s \langle \Delta\psi_{ND} \rangle_{\boldsymbol{\sigma}}/L$, for each protein family. The correlation coefficient is $r = -0.75$, and the regression line is $\overline{\overline{\Delta\Delta G_{ND}(\sigma_{j\neq i}^N, \sigma_i^N \to \sigma_i)}} = -2.74 \langle \Delta G_{ND} \rangle_\sigma / L + 1.09$. See Fig. S.19 for $r_{cutoff} \sim 15.5$ Å. The free energies are in kcal/mol units.

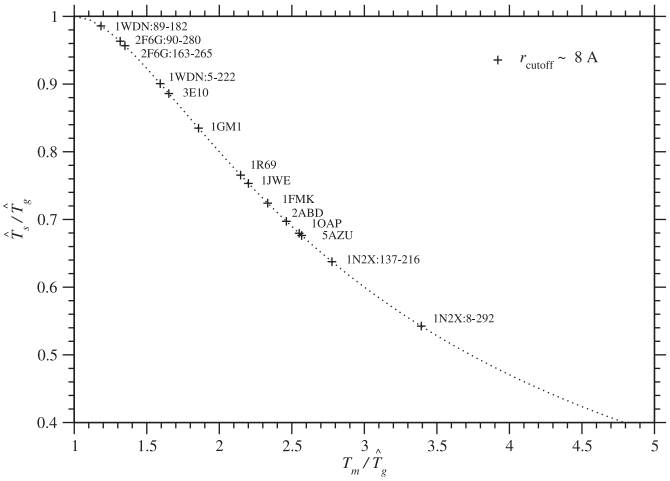

Fig. 5. $\hat{T}_s/\hat{T}_g$ is plotted against $T_m/\hat{T}_g$ for each protein domain. A dotted curve corresponds to Eq. (47), $\hat{T}_s/\hat{T}_g = 2(T_m/\hat{T}_g)/((T_m/\hat{T}_g)^2 + 1)$. Plus marks indicate the values estimated with $r_{cutoff} \sim 8$ Å. See Fig. S.20 for $r_{cutoff} \sim 15.5$ Å. 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.

icant; the correlation coefficient is larger than 0.99. The intercept $\beta_{\overline{\psi_N}}$ should be equal to 0, because if $T_s \to \infty$ then $\delta\psi^2 \to 0$ and $\Delta\psi_N \to 0$. Actually, Fig. 3 shows that $\beta_{\overline{\psi_N}}$ is nearly equal to 0.

Finally, the regression of $\Delta\Delta G_{ND}$ on ΔG_{ND} would be derived if T_g, T_s, and T were constant.

$$\overline{\overline{\Delta\Delta G_{ND}(\sigma_{j\neq i}^N, \sigma_i^N \to \sigma_i)}}$$

$$\approx -\alpha_{\overline{\psi_N}} k_B T_s \frac{\delta\psi^2(\overline{\boldsymbol{f}(\boldsymbol{\sigma}_N)})}{L} + k_B T_s \beta_{\overline{\psi_N}} \tag{70}$$

$$= \alpha_{\Delta G_{ND}} k_B T_s \frac{\delta\psi^2(\overline{\boldsymbol{f}(\boldsymbol{\sigma}_N)})}{L} \left(\vartheta(T/T_g) \frac{T_s}{T} - 1 \right) + \beta_{\Delta G_{ND}} \tag{71}$$

$$= \alpha_{\Delta G_{ND}} \frac{\langle \Delta G_{ND}(\boldsymbol{\sigma}_N, T) \rangle}{L} + \beta_{\Delta G_{ND}} \tag{72}$$

In general, T_s and T_g are different among protein families, so that the correlation between $\overline{\overline{\Delta\Delta G_{ND}}}$ and $\langle \Delta G_{ND} \rangle / L$ cannot be strong. In Fig. 4, $\overline{\overline{\Delta\Delta G_{ND}}}$ for the present proteins are plotted against $\langle \Delta G_{ND} \rangle / L$. However, it should be noted that the correlation is not expected for $\overline{\overline{\Delta\Delta G_{ND}}}$ and $\langle \Delta G_{ND} \rangle$ but for $\overline{\overline{\Delta\Delta G_{ND}}}$ and $\langle \Delta G_{ND} \rangle / L$.

4.6. Estimation of T_g, ω, and $\langle \Delta G_{ND}(\boldsymbol{\sigma}) \rangle_{\boldsymbol{\sigma}}$ from T_s and T_m

To estimate glass transition temperature T_g, the conformational entropy per residue ω in the compact denatured state, and the ensemble average of folding free energy in sequence space $\langle \Delta G_{ND} \rangle_{\boldsymbol{\sigma}}$, melting temperature T_m must be known for each protein; see Eqs. (47), (18), and (44) for T_g, ω and $\langle \Delta G_{ND} \rangle_{\boldsymbol{\sigma}}$, respectively. The experimental value of T_m (Armengaud et al., 2004; D'Auria et al., 2005; Ganguly et al., 2009; Guelorget et al., 2010; Knapp et al., 1998; Onwukwe et al., 2014; Parsons et al., 2006; Rosa et al., 1995; Sainsbury et al., 2008; Stupák et al., 2006; Torchio et al., 2012; Williams et al., 2002) employed for each protein is listed in Tables 3 and S.6. For comparison, temperature T is set up to be equal to the experimental temperature for ΔG_{ND} or to 298°K if unknown.

An estimate of glass transition temperature, $\hat{T}_g$, has been calculated with $\hat{T}_s$ and T_m by Eq. (47), and is listed in Tables 3 and S.6 for each protein. In Fig. 5, $\hat{T}_s/\hat{T}_g$ is plotted against $T_m/\hat{T}_g$ for each protein family. Unless $T_g < T_m$, a protein will be trapped at local minima on a rugged free energy landscape before it folds into a unique native structure. Protein foldability increases as T_m/T_g increases. A condition, $\Delta G_{ND} = 0$ at $T = T_m$, for the first order transition requires that Eq. (47), which is indicated by a dotted curve in Fig. 5, must be satisfied. As a result, T_s/T_g must be lowered to increase T_m/T_g; in other words, proteins must be selected at lower T_s. The present estimates of T_s and T_g would be within a reasonable range (Morcos et al., 2014; Onuchic et al., 1995; Pande et al., 2000) of values required for protein foldability.

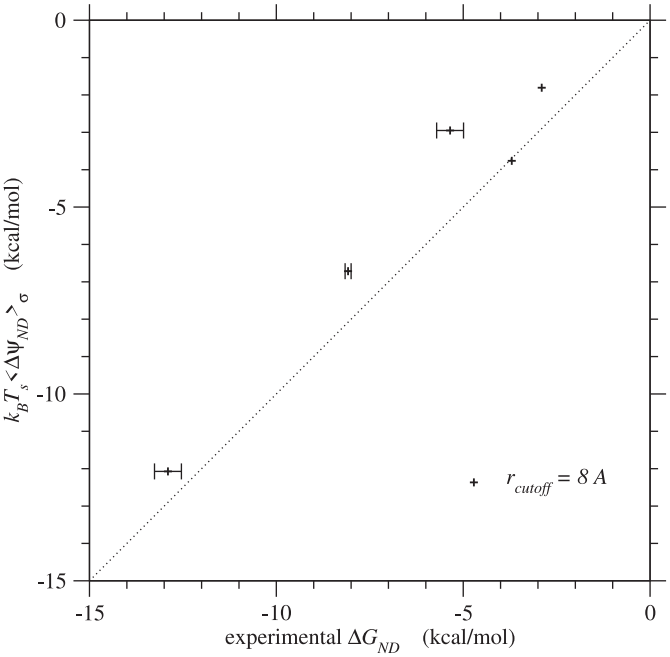

Fig. 6. Folding free energies, $\langle \Delta G_{ND} \rangle_{\boldsymbol{\sigma}} \simeq k_B T_s \langle \Delta\psi_{ND} \rangle_{\boldsymbol{\sigma}}$, predicted by the present method are plotted against their experimental values, $\Delta G_{ND}(\boldsymbol{\sigma}_N)$. Plus marks indicate the values estimated with $r_{cutoff} \sim 8$ Å. See Fig. S.21 for $r_{cutoff} \sim 15.5$ Å. The free energies are in kcal/mol units.

In Tables 3 and S.6, the ensemble average of $\Delta G_{ND}(\boldsymbol{\sigma})$ over sequences calculated by Eq. 44, and the conformational entropy per residue ω in the compact denatured state by Eq. (18) are also listed for each protein. Fig. 6 shows the comparison of their

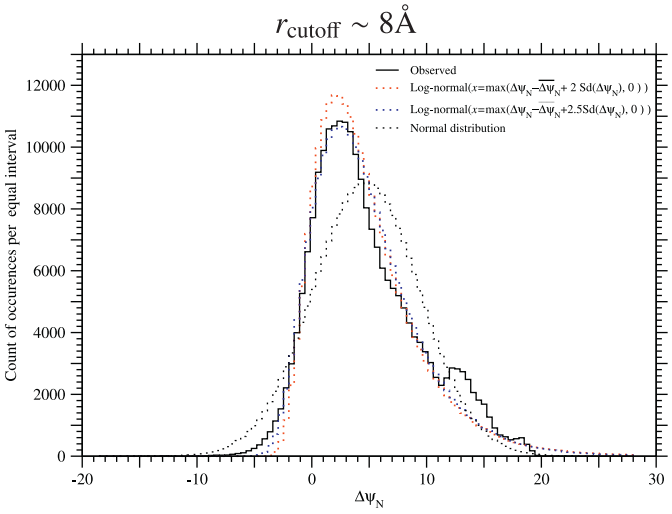

Fig. 7. The observed frequency distribution and the fitted distributions of $\Delta\psi_N$ in the PDZ protein family. A black solid line indicates the observed frequency distribution of $\Delta\psi_N$ per equal interval in homologous sequences of the PDZ protein family, and red dotted and blue dotted lines indicate the total frequencies of log-normal distributions with $n_{shift} = 2$ or 2.5 and parameters estimated with the mean and variance of the observed distribution for each protein; see Eqs. 74) to (78). 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 1. See Fig. S.22 for $r_{cutoff} \sim 15.5$ Å.

ensemble averages, $\langle \Delta G_{ND}(\boldsymbol{\sigma})\rangle_{\boldsymbol{\sigma}}$, and the experimental values of $\Delta G_{ND}(\boldsymbol{\sigma}_N)$ (Gianni et al., 2005; 2007; Grantcharova et al., 1998; Kragelund et al., 1999; Ruiz-Sanz et al., 1999; Wilson and Wittung-Stafshede, 2005) listed in Table S.4. The correlation in the case of $r_{cutoff} \sim 8$ Å is quite good, indicating that the constancy approximation (Eq. (65)) for the variance of ΔG_N is appropriate. The conformational entropy per residue in the compact denatured state, $\hat{\omega}$ in Eq. (18), estimated from the condition for the first order transition falls into the range of $0.60–1.13k_B$ for $r_{cutoff} \sim 8$ Å, which agrees well with the range estimated by Morcos et al. (2014).

4.7. The equilibrium value of evolutionary statistical energy ψ_N in the mutation–fixation process of amino acid substitutions

Let us consider the fixation process of amino acid substitutions in a monoclonal approximation, in which protein evolution is assumed to proceed with single amino acid substitutions fixed at a time in a population. In this approximation, $\Delta\psi_{ND}$ and ψ_N are at equilibrium and the ensemble of protein sequences attains to the equilibrium state, when the average of $\Delta\Delta\psi_{ND} \simeq \Delta\psi_N$ over singe nucleotide nonsynonymous mutations fixed in a population is equal to zero; an amino acid composition is assumed to be constant in protein evolution.

$$\langle \Delta\Delta\psi_{ND}\rangle_{fixed} \simeq \langle \Delta\psi_N\rangle_{fixed} = 0 \Longleftrightarrow \Delta\psi_{ND} \text{ and } \psi_N \text{ are at equilibrium.} \quad (73)$$

The average of $\Delta\psi_N$ over fixed mutations, $\langle \Delta\psi_N\rangle_{fixed}$, is calculated numerically with the probability density function (PDF) of $\Delta\Delta\psi_{ND}(\simeq \Delta\psi_N)$ for single nucleotide nonsynonymous mutations; see Eqs. (52) and (53). $N = 10^6$ is employed.

The PDF of $\Delta\Delta G_{ND}$ were approximated with a normal distribution(Serohijos et al., 2012) or a bi-normal distribution (Tokuriki et al., 2007). Figs. 7, S.22, and S.23, however, show that a single normal distribution with the observed mean and standard deviation cannot well reproduce the observed distribution of $\Delta\psi_N$ due to single nucleotide nonsynonymous mutations. For simplicity, a log-normal distribution, $\ln\mathcal{N}(x;\mu,\sigma)$, for which x, μ and σ defined as follows, is arbitrarily used here to better reproduce observed distributions 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 well reproduce the observed ones, too.

$$p(\Delta\psi_N) \approx \ln\mathcal{N}(x;\mu,\sigma) \equiv \frac{1}{x}\mathcal{N}(\ln x;\mu,\sigma) \quad (74)$$

$$x \equiv \max(\Delta\psi_N - \Delta\psi_N^0, 0) \quad (75)$$

$$\exp(\mu + \sigma^2/2) = \overline{\Delta\psi_N} - \Delta\psi_N^0 \quad (76)$$

$$\exp(2\mu + \sigma^2)(\exp(\sigma^2) - 1) = \overline{(\Delta\psi_N - \overline{\Delta\psi_N})^2)} \quad (77)$$

$$\Delta\psi_N^0 \equiv \min(\overline{\Delta\psi_N} - n_{\rm shift}\overline{(\Delta\psi_N - \overline{\Delta\psi_N})^2)}^{1/2}, 0) \quad (78)$$

where $\Delta\psi_N^0$ is the origin for the log-normal distribution and the shifting factor $n_{\rm shift}$ is taken to be equal to 2, unless specified. It is shown in Figs. 7, S.22, and S.23 that log-normal distributions can better reproduce the observed distribution of $\Delta\psi_N$ due to single nucleotide nonsynonymous mutations except in the tails. Disagreements between the log-normal and observed distributions in the domain of $\Delta\psi_N > \overline{\Delta\psi_N}$ do not much affect the PDF of $\overline{\Delta\psi_N}$ in fixed mutants, because fixation probabilities for $\Delta\psi_N(> \overline{\Delta\psi_N})$ are too low.

The average of $\Delta\psi_N$ over fixed mutants is uniquely determined by the distribution of $\Delta\Delta\psi_N(\simeq \Delta\psi_N)$, which is approximated here by a log-normal distribution estimated from the mean and variance of $\Delta\psi_N$; it depends also on q_m, which is assumed to be constant, through fixation probability, because $2N_e s \simeq -\Delta\psi_N/(1 - q_m)$. In other words, $\langle \Delta\psi_N\rangle_{fixed}$ is uniquely determined by the mean and variance of $\Delta\psi_N$. Therefore, under the equilibrium condition $\langle \Delta\psi_N\rangle_{fixed} = 0$, only one of the mean and variance can be freely specified, and the other is uniquely determined. We employ $\overline{\Delta\psi_N}$ or ψ_N as a parameter, because $\overline{\Delta\psi_N}$ depends on ψ_N, and only one of them can be specified. We define $\overline{\Delta\psi_N}^{eq}$ as $\overline{\Delta\psi_N}$ at which $\langle \Delta\psi_N\rangle_{fixed} = 0$.

Suppose that the regression equation, Eq. (62), of $\Delta\psi_N$ on ψ_N is exact, and the standard deviation of $\Delta\psi_N$ is constant irrespective of ψ_N; the slope (α_{ψ_N}), $\overline{\overline{\Delta\psi_N}}$, $\overline{\mathrm{Sd}(\Delta\psi_N)}$, and $\overline{\psi_N}$ that are estimated with $r_{cutoff} \sim 8$ Å for the PDZ and listed in Table 2 are employed here. In Fig. 8, the average of $\Delta\psi_N$ over single nucleotide nonsynonymous substitutions fixed in a population, $\langle \Delta\psi_N\rangle_{fixed}$, is plotted against ψ_N/L of a wildtype for the PDZ protein family. This figure shows that $\langle \Delta\psi_N\rangle_{fixed}$ changes its value from positive to negative as ψ_N increases, that is, the value of ψ_N at which $\langle \Delta\psi_N\rangle_{fixed} = 0$, ψ_N^{eq}, is the stable equilibrium value for ψ_N. In order for protein to have such a stable equilibrium value for folding free energy ($\Delta G_{ND} \simeq k_B T_s \Delta\psi_{ND}$), the regression coefficient of $\overline{\Delta\psi_N}$ on ψ_N must be more negative than that of the standard deviation, $\mathrm{Sd}(\Delta\psi_N)$, because otherwise stabilizing mutations increase as ψ_N decreases. This condition is, of course, satisfied for all protein families studied here, because the mean of $\Delta\psi_N$ over all substitutions at all sites is negatively proportional to ψ_N of a wildtype, but its standard deviation is nearly constant irrespective of ψ_N across homologous sequences; see Tables 2 and S.5.

The equilibrium value of ψ_N for each protein domain is calculated with the estimated values of α_{ψ_N}, $\overline{\psi_N}$, $\overline{\overline{\Delta\psi_N}}$, and $\overline{Sd(\Delta\psi_N)}$ listed in Tables 2 and S.5; it should be noticed here that $\overline{Sd(\Delta\psi_N)}$

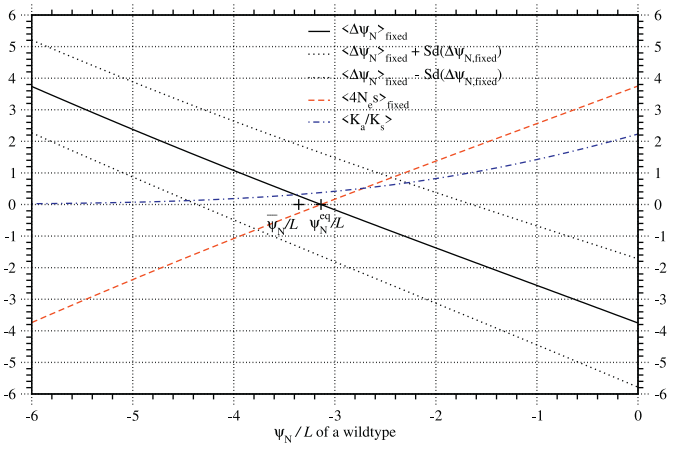

Fig. 8. The average of $\Delta\psi_N(\simeq\Delta\Delta\psi_{ND})$ over fixed single nucleotide nonsynonymous mutations versus ψ_N/L of a wildtype for the PDZ protein family. The averages of $\Delta\psi_N(\simeq\Delta\Delta\psi_{ND})$ and $4N_es$ over the fixed mutants, and the average of $K_a/K_s(\equiv u(s)/u(0))$ 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 $\langle\Delta\psi_N\rangle_{fixed}\pm$ 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. (21) and (33). Here the empirical relationships of Eqs. (62) and (63) 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 ($-\alpha_{\psi_N}\overline{\psi_N}/L+\overline{\Delta\psi_N}$) and the average of $\mathrm{Sd}(\Delta\psi_N)$ over homologous sequences that are estimated with $r_{cutoff}\sim8$Å for the PDZ and listed in Table 2 are employed here. The distribution of $\Delta\psi_N$ due to single nucleotide nonsynonymous mutations is approximated by a log-normal distribution with $n_{shift}=2.0$; see Eqs. (74) to (78). The ψ_N^{eq}, where $\langle\Delta\Delta\psi_{ND}\rangle_{fixed}\simeq\langle\Delta\psi_N\rangle_{fixed}=0$, is the stable equilibrium value of ψ_N in the protein evolution of the PDZ protein family. The ψ_N^{eq} is close to the average of ψ_N over homologous sequences ($\overline{\psi_N}$), indicating that the present approximations for ψ_N^{eq} and for $\overline{\psi_N}=\langle\psi_N\rangle_\sigma$ are consistent to each other.

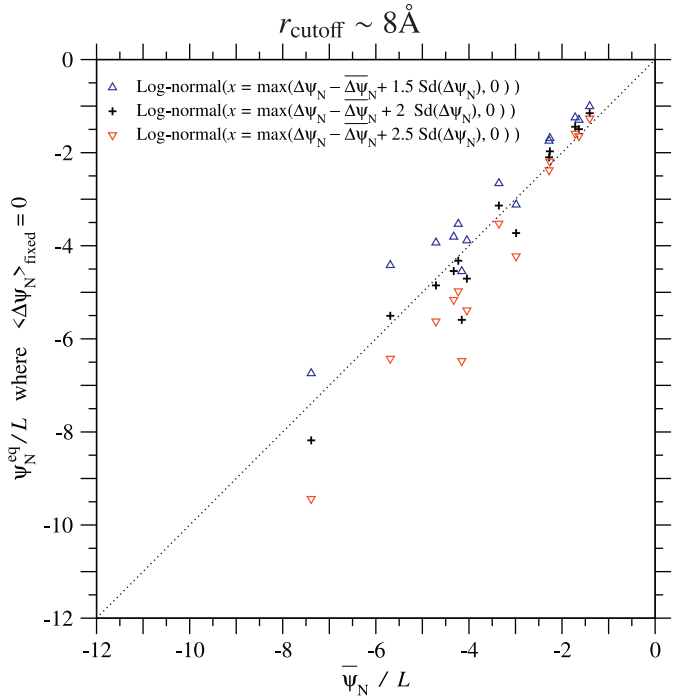

Fig. 9. The equilibrium value of ψ_N/L, where $\langle\Delta\psi_N\rangle_{fixed}=0$, is plotted against the average of ψ_N/L over homologous sequences for each protein family. The cutoff distance $r_{cutoff}=8$Å is employed to estimate ψ_N of each protein family; see Fig. S.26 for $r_{cutoff}=15.5$Å. The equilibrium values ψ_N^{eq}, where $\langle\Delta\psi_N\rangle_{fixed}=0$, are calculated by using the linear dependency of $\overline{\Delta\psi_N}$ on ψ_N (Eq. (62)) and estimated values with $r_{cutoff}\sim8$ in Table 2. The standard deviation of $\Delta\psi_N$ is approximated to be constant and equal to $\overline{\mathrm{Sd}(\Delta\psi_N)}$; see Eq. (63). Plus, upper triangle, and lower triangle marks indicate the cases of log-normal distributions with $n_{shift}=1.5, 2.0$, and 2.5 employed to approximate the distribution of $\Delta\psi_N$, respectively; see Eqs. (74) to (78).

is assumed to be constant. In Figs. 9 and S.26, the equilibrium values of ψ_N/L estimated with $n_{shift}=1.5, 2$, and 2.5 in the monoclonal approximation are plotted against the average of ψ_N/L over homologous sequences for each protein family. The agreement between the time average (ψ_N^{eq}) and ensemble average ($\langle\psi_N\rangle_\sigma(=\overline{\psi_N})$) is better for $r_{cutoff}\sim8$ Å than for $r_{cutoff}\sim15.5$ Å and is not bad in the case of $r_{cutoff}\sim8$ Å, indicating that the present methods for the fixation process of amino acid substitutions and for the equilibrium ensemble of ψ_N give a consistent result with each other, and also that it is a good approximation to assume the standard deviation of $\Delta\psi_N$ not to depend on ψ_N in each protein family.

4.8. Relationships between $\overline{\Delta\psi_N}(=\overline{\Delta\psi}_N^{eq})$ and the standard deviation of $\Delta\psi_N$, $\hat{T}_s$, and $\Delta\Delta\hat{G}_{ND}$ at equilibrium

In the present model, the equilibrium values, ψ_N^{eq} and the corresponding $\overline{\Delta\psi}_N^{eq}$, are functions of the mean and standard deviation of $\Delta\psi_N$ only, because the distribution of $\Delta\Delta\psi_{ND}(\simeq\Delta\psi_N)$ is approximately estimated with the mean and standard deviation of $\Delta\psi_N$. On the other hand, ψ_N^{eq} and $\overline{\Delta\psi}_N^{eq}$ should be equal to $\overline{\psi_N}=\langle\psi_N\rangle$ and $\overline{\Delta\psi_N}$, respectively; the time average and ensemble average should be consistent. Actually ψ_N^{eq} almost agrees with $\overline{\psi_N}$ as shown in Fig. 9. Therefore the standard deviation of $\Delta\psi_N$ is uniquely determined from its mean as long as ψ_N and $\overline{\Delta\psi_N}$ are at equilibrium; conversely the equilibrium value of $\overline{\Delta\psi_N}$ is determined by $\mathrm{Sd}(\Delta\psi_N)$. In Fig. 10, the standard deviation of $\Delta\psi_N$ is plotted against $\overline{\Delta\psi_N}(=\overline{\Delta\psi}_N^{eq})$. Likewise the estimate of effective temperature of selection, $\hat{T}_s(=(\hat{T}_s\overline{Sd}(\Delta\psi_N))_{PDZ}/Sd(\Delta\psi_N))$, and that of folding free energy change, $\Delta\Delta\hat{G}_{ND}(=k_B(\hat{T}_s\overline{Sd}(\Delta\psi_N))_{PDZ}/Sd(\Delta\psi_N)\cdot\overline{\Delta\psi_N})$, are plot-

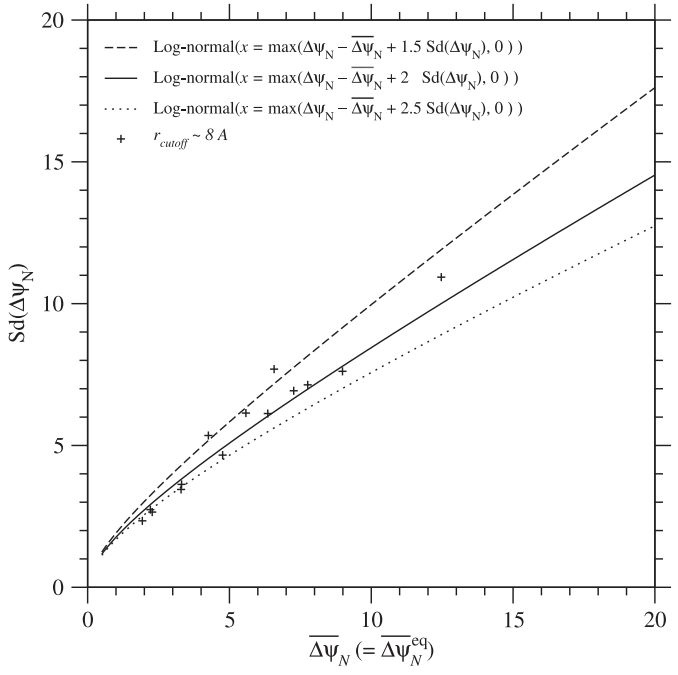

Fig. 10. Relationship between the mean and the standard deviation of $\Delta\psi_N$ due to single nucleotide nonsynonymous mutations at equilibrium, $\langle\Delta\psi_N\rangle_{fixed}=0$. The standard deviation of $\Delta\psi_N$ that satisfies $\langle\Delta\psi_N\rangle_{fixed}=0$ is plotted against its mean, $\overline{\Delta\psi_N}$. Broken, solid, and dotted lines indicate the cases of log-normal distributions with $n_{shift}=1.5, 2.0$ and 2.5 employed to approximate the distribution of $\Delta\psi_N$, respectively; see Eqs. (74) to (78). Plus marks indicate the averages, $\overline{\Delta\psi_N}$ and $\overline{\mathrm{Sd}(\Delta\psi_N)}$, over homologous sequences in each protein family for $r_{cutoff}\sim8$ Å, which are listed in Table 2. See Fig. S.27 for $r_{cutoff}\sim15.5$ Å.

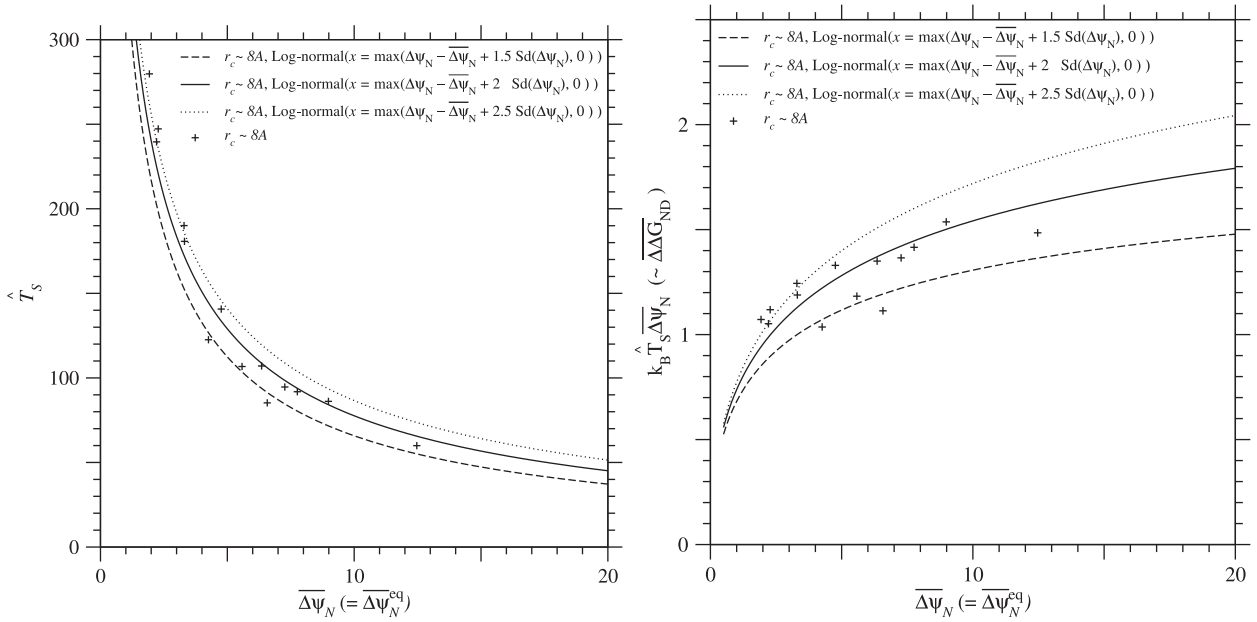

Fig. 11. Relationships between $\hat{T}_s$ and $\overline{\Delta\psi_N}$ and between $k_B\hat{T}_s\overline{\Delta\psi_N}(\simeq \overline{\Delta\Delta G_{ND}})$ and $\overline{\Delta\psi_N}$ at equilibrium, $\langle\Delta\psi_N\rangle_{fixed} = 0$. The estimate $\hat{T}_s(= (\hat{T}_s\overline{Sd}(\Delta\psi_N))_{PDZ}/Sd(\Delta\psi_N))$ of effective temperature for selection and the estimate of mean folding free energy change, $k_B\hat{T}_s\overline{\Delta\psi_N}(\simeq \overline{\Delta\Delta G_{ND}})$, are plotted against $\overline{\Delta\psi_N}$ under the condition of $\langle\Delta\psi_N\rangle_{fixed} = 0$. 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. (65). Broken, solid, and dotted lines indicate the cases of log-normal distributions with $n_{shift} = 1.5, 2.0$ and 2.5 employed to approximate the distribution of $\Delta\psi_N$, respectively; see Eqs. (74) to (78). Plus marks indicate those estimates against the average of $\overline{\Delta\psi_N}$ over homologous sequences for each protein family with $r_{cutoff} \sim 8$Å, which are listed in Tables 2 and 3. See Fig. S.28 for $r_{cutoff} \sim 15.5$Å.

ted as a function of $\overline{\Delta\psi_N}(= \overline{\Delta\psi}_N^{eq})$ in Fig. 11. These figures show that the averages, $\overline{\overline{\Delta\psi_N}}$ and $\overline{\mathrm{Sd}(\psi_N)}$, over homologous sequences scatter along the expected curves.

4.9. Protein evolution at equilibrium, $\langle\Delta\psi_N\rangle_{fixed} = 0$

The common understanding of protein evolution has been that amino acid substitutions observed in homologous proteins are neutral (Kimura, 1968; 1969; Kimura and Ohta, 1971; 1974) or slightly deleterious (Ohta, 1973; 1992), and random drift is a primary force to fix amino acid substitutions in population. In order to see how significant neutral/slightly deleterious substitutions are in protein evolution, the PDFs of K_a/K_s in all single nucleotide nonsynonymous mutations and in their fixed mutations are calculated; K_a/K_s is the ratio of nonsynonymous to synonymous substitution rate per site (Miyata and Yasunaga, 1980) and defined here as $K_a/K_s \equiv u(s)/u(0)$, where $u(s)$ is a fixation probability for selective advantage s; see Eq. (50).

First let us see the distributions of $\Delta\psi_N$ at equilibrium, $\langle\Delta\psi_N\rangle_{fixed} = 0$. Fig. 12 shows the PDFs of $\Delta\psi_N$ in all single nucleotide nonsynonymous mutations and in their fixed mutations as a function of $\overline{\Delta\psi_N}(= \overline{\Delta\psi}_N^{eq})$, respectively. Because $4N_es(1-q_m) = -\Delta\Delta\psi_{ND} \simeq -\Delta\psi_N$, the PDFs of $\Delta\psi_N$ can be regarded as the PDFs of $-4N_es(1-q_m)$. At equilibrium, the distribution of $\Delta\psi_N$ in all single nucleotide nonsynonymous mutants becomes wider as the mean of $\Delta\psi_N$ increases, however, that in fixed mutants remains to be narrow with a peak near zero.

The PDFs of K_a/K_s in all single nucleotide nonsynonymous mutations and in their fixed mutations are shown in Fig. 12. The blue line on the landscape of the PDF shows the averages of K_a/K_s. The averages of K_a/K_s in all single nucleotide nonsynonymous mutations and in their fixed mutations are also shown in Fig. 13. The average of K_a/K_s in all the arising mutants is less than 1 and decreases as $\overline{\Delta\psi_N} \simeq \overline{\Delta\psi}_N^{eq}$ increases, indicating that negative mutants significantly occur and increase as $\overline{\Delta\psi_N}$ increases. On the other hand, $\langle K_a/K_s\rangle_{fixed}$ in fixed mutants is larger than 1 and increases as $\Delta\psi_N^{eq}$ increases, indicating that positive mutants fix significantly in population and increase as equilibrium folding free energy change increases, that is, equilibrium protein stability decreases. To see each contribution of positive, neutral, slightly negative and negative selections, the value of K_a/K_s is 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 $0.5 > K_a/K_s$ for their selection categories, respectively. The probabilities of each selection category in all single nucleotide nonsynonymous mutations and in their fixed mutations are shown in Fig. 14. The almost 50% of fixed mutations are stabilizing mutations fixed by positive selection ($1.05 < K_a/K_s$), and another 50% are destabilizing mutations fixed by random drift. They are balanced with each other, and the stability of protein is maintained. Contrary to the neutral theory (Kimura, 1968; 1969; Kimura and Ohta, 1971; 1974), the proportion of neutral selection is not large even in fixed mutations, and slightly negative mutations are significantly fixed. Neutral mutations fixed with $0.95 < K_a/K_s < 1.05$ are only less than 10%, and slightly negative mutations fixed with $0.5 < K_a/K_s < 0.95$ and negative mutations fixed with $K_a/K_s < 0.5$ are both from 10 to 30%. The nearly neutral theory (Ohta, 1973; 1992; 2002) insists that most fixed mutations satisfy $|N_es| \leq 2$. This condition corresponds to $0.003 \leq K_a/K_s(= u(s)/u(0)) \leq 8$; see Eqs. (21) and (50). The PDF of K_a/K_s shown in Fig. 14 indicates that this condition is satisfied, supporting the nearly neutral theory.

4.10. Relationship between T_s and K_a/K_s

The effective temperature (T_s) of protein for selection, which is defined in Eq. (1), represents the strength of selection originating from protein stability and foldability. Thus, it must be related with the evolutionary rate (amino acid substitution rate) of protein. As the effective temperature of selection (T_s) decreases, the mean change of evolutionary statistical energy ($\overline{\Delta\psi}_N^{eq}$) due to single amino acid substitutions increases; see Fig. 11. Therefore,

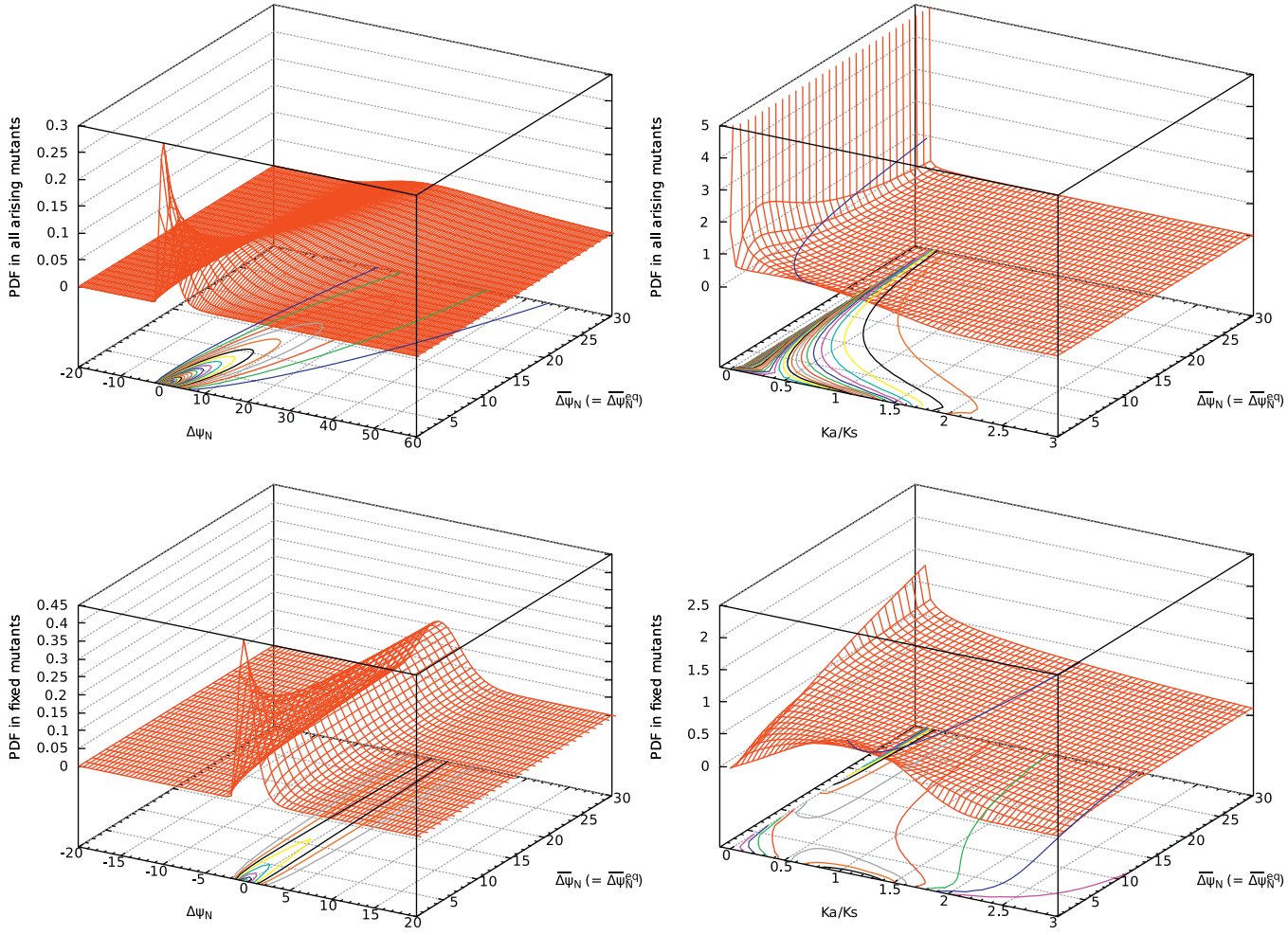

Fig. 12. PDFs of $\Delta\psi_N$ (left) and K_a/K_s (right) in all singe nucleotide nonsynonymous mutants (upper) and in their fixed mutants (lower) as a function of $\overline{\Delta\psi_N}$ at equilibrium, $\langle\Delta\psi_N\rangle_{fixed}=0$. Fixation probability has been calculated with $\Delta\Delta\psi_{ND}\simeq\Delta\psi_N$; see Eqs. (21) and (33). The distribution of $\Delta\psi_N$ due to single nucleotide nonsynonymous mutations is approximated by a log-normal distribution with $n_{shift}=2.0$; see Eqs. (74) to (78). The standard deviation of $\Delta\psi_N$ is determined to satisfy $\langle\Delta\psi_N\rangle_{fixed}=0$ at $\overline{\Delta\psi_N}=\overline{\Delta\psi_N}^{eq}$.

destabilizing mutations increase, and an amino acid substitution rate is expected to decrease. Fig. 13 shows that the average of K_a/K_s decreases as $\overline{\Delta\psi_N}^{eq}$ increases. The direct relationship between substitution rate and $T_s(=(T_s\overline{Sd}(\Delta\psi_N))_{PDZ}/Sd(\Delta\psi_N))$ is shown in Fig. 15; the average of K_a/K_s decreases as T_s increases. In the selection maintaining protein foldability/stability, the effective temperature of selection is directly reflected in the average amino acid substitution rate.

5. Discussion

A main purpose of the present study is to formulate protein fitness originating from protein foldability and stability. From a phenomenological viewpoint, Drummond and Wilke (2008) took notice of toxicity of misfolded proteins as well as diversion of protein synthesis resources, and formulated a Malthusian fitness of a genome to be negatively proportional to the total amount of misfolded proteins, which must be produced to obtain the necessary amount of folded proteins (Serohijos et al., 2012). They also formulated a Malthusian fitness based on protein dispensability to be negatively proportional to the ratio of unfolded proteins. These formulas of protein fitness can be well approximated by a generic form, $m=-\kappa\exp(\Delta G_{ND}/(k_BT))$, where T is growth temperature, and $\kappa(\geq 0)$ is a parameter that depends on protein disability and cellular abundance of protein (Miyazawa, 2016).

In the comparison of this generic formula of protein fitness with the present one, it may be interpreted that $4N_e(1-q_m)\kappa/T\sim 1/T_s$, if $|\Delta G_{ND}/(k_BT)|\ll 1$, however, the growth temperature T and folding free energy do not always satisfy this condition. These two types of selection should be considered to be the different types of selection, although both are related with protein stability (ΔG_{ND}). Selective advantage of mutant is not upper-bounded in the present scheme of a Malthusian fitness but in the case of $m=-\kappa\exp(\Delta G_{ND}/(k_BT))$. As a result, PDFs of K_a/K_s in all arising mutations and in fixed mutations have very different shapes between these two formulas of fitness (Miyazawa, 2016). Selection modeled here is one that yields the distribution of homologous sequences in protein evolution. In other word, the present formula for protein fitness models natural selection maintaining protein's stability, foldability, and function over the evolutionary time scale, which is much longer than the time scale for the selection originating from toxicity of misfolded proteins.

The present formulas for protein fitness, Eqs. (31) and (30), have been derived on the basis of a protein folding theory, particularly the random energy model, and the maximum entropy principle for the distribution of homologous sequences with the same fold in sequence space, respectively. The former indicates that the equilibrium ensemble of sequences can be well approximated by a canonical ensemble with the Boltzmann factor $\exp(-\Delta G_{ND}/k_BT_s)$, and

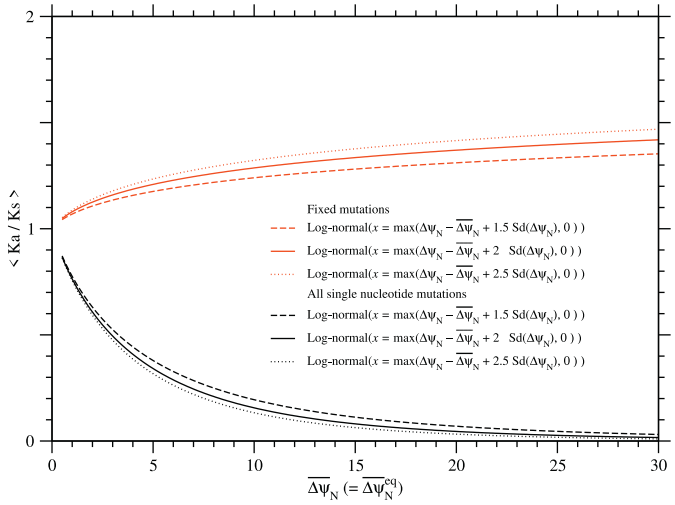

Fig. 13. The averages of K_a/Ks over all single nucleotide nonsynonymous mutations and over their fixed mutations as a function of $\overline{\Delta\psi_N}$ at equilibrium, $\langle\Delta\psi_N\rangle_{fixed} = 0$. Black and red lines indicate $\langle K_a/K_s\rangle$ and $\langle K_a/K_s\rangle_{fixed}$, respectively. Fixation probability has been calculated with $\Delta\Delta\psi_{ND} \simeq \Delta\psi_N$; see Eqs. (21) and (33). Broken, solid, and dotted lines indicate the cases of log-normal distributions with $n_{shift} = 1.5, 2.0$ and 2.5 employed to approximate the distribution of $\Delta\psi_N$, respectively; see Eqs. (74) to (78). The standard deviation of $\Delta\psi_N$ is determined to satisfy $\langle\Delta\psi_N\rangle_{fixed} = 0$ at $\overline{\Delta\psi_N} = \overline{\Delta\psi_N}^{eq}$.

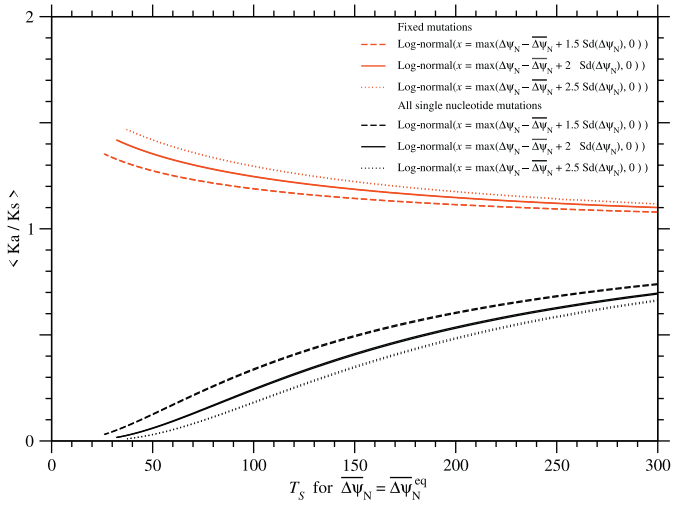

Fig. 15. The averages of K_a/Ks over all single nucleotide nonsynonymous mutations and over their fixed mutations as a function of the effective temperature of selection, $T_s (= (T_s\overline{Sd}(\Delta\psi_N))_{PDZ}/Sd(\Delta\psi_N))$, at equilibrium, $\langle\Delta\psi_N\rangle_{fixed} = 0$. Black and red lines indicate $\langle K_a/K_s\rangle$ and $\langle K_a/K_s\rangle_{fixed}$, respectively. Fixation probability has been calculated with $\Delta\Delta\psi_{ND} \simeq \Delta\psi_N$; see Eqs. (21) and (33). The distribution of $\Delta\psi_N$ due to single nucleotide nonsynonymous mutations is approximated by a log-normal distribution with $n_{shift} = 2.0$; see Eqs. (74) to (78). The standard deviation of $\Delta\psi_N$ is determined to satisfy $\langle\Delta\psi_N\rangle_{fixed} = 0$ at $\overline{\Delta\psi_N} (= \overline{\Delta\psi_N}^{eq})$. 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. (65). Broken, solid, and dotted lines indicate the cases of log-normal distributions with $n_{shift} = 1.5, 2.0$ and 2.5 employed to approximate the distribution of $\Delta\psi_N$, respectively; see Eqs. (74) to (78). The curves for $r_{cutoff} \sim 8$ and 15.5 Å almost overlap with each other, because the estimates of $(\hat{T}_s\overline{Sd}(\Delta\psi_N))_{PDZ}$ for the PDZ with $r_{cutoff} \sim 8$ and 15.5 Å are almost equal to each other.

the latter insists that the probability distribution of homologous sequences, which satisfies a given amino acid composition at each site and a given pairwise amino acid frequency at each site pair, can be represented as a Boltzmann distribution with $\exp(-\psi_N)$, in which the evolutionary statistical energy (ψ_N) is represented as the sum of one-body (compositional) and pairwise (covariational) interactions between sites. On the other hand, assuming mutation and fixation processes to be reversible Markov processes leads us to a formulation that the equilibrium ensemble of sequences also obeys a Boltzmann distribution with $\exp(4N_e m(1-q_m))$. As a result, we obtain the correspondences between folding free energy ($-\Delta G_{ND}/k_B T_s$), and $-\Delta\psi_{ND}$ and protein fitness ($4N_e m(1-q_m)$): the equality between the latter two variables (Eq. (33)), which indicates that $\Delta\psi_N$ is proportional to fitness (s), and the approximate equality between the former two variables (Eq. (34)) since a canonical ensemble with $\Delta G_{ND}/(k_B T_s)$ is an approximate for the sequence ensemble under natural selection. A discrepancy between evolutionary statistical energies J_{ij} and actual interaction energies was pointed out for non-contacting residue pairs in Monte Carlo simulations of lattice proteins (Jacquin et al., 2016). Also, the ra-

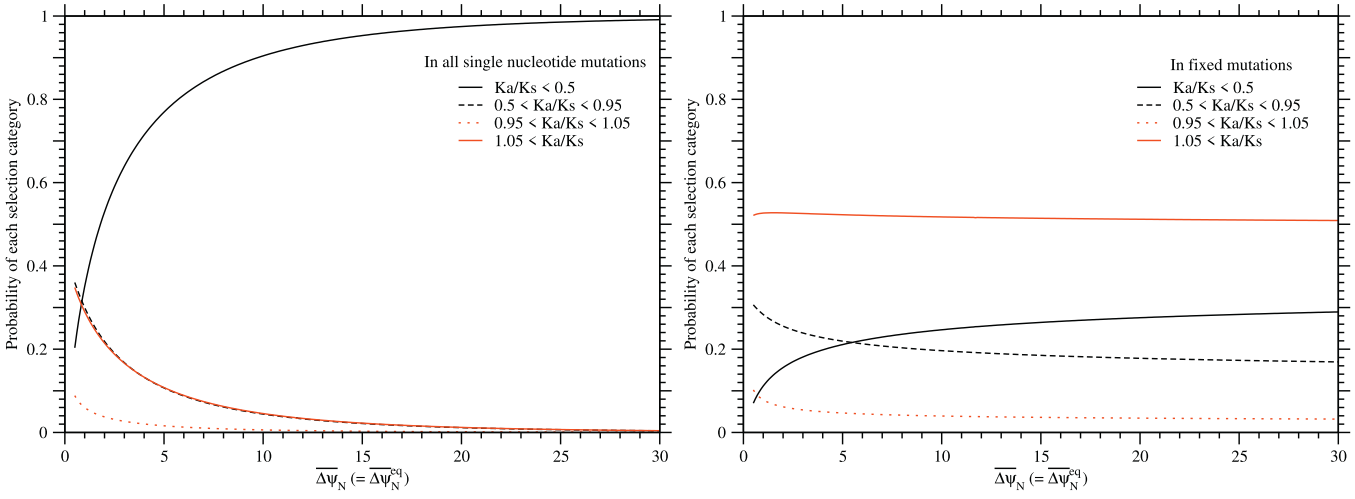

Fig. 14. The probabilities of each selection category in all single nucleotide nonsynonymous mutations and in their fixed mutations as a function of $\overline{\Delta\psi_N}$ at equilibrium, $\langle\Delta\psi_N\rangle_{fixed} = 0$. The left and right figures are for single nucleotide nonsynonymous mutations 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. (21) and (33). The distribution of $\Delta\psi_N$ due to single nucleotide nonsynonymous mutations is approximated by a log-normal distribution with $n_{shift} = 2.0$; see Eqs. (74) to (78). The standard deviation of $\Delta\psi_N$ is determined to satisfy $\langle\Delta\psi_N\rangle_{fixed} = 0$ at $\overline{\Delta\psi_N} (= \overline{\Delta\psi_N}^{eq})$.

tio of $-J_{ij}(a_k, a_l)$ to the corresponding actual contact energy was shown to differ among contact site pairs. On the other hand, Hopf et al. (2017) successfully predicted mutation effects with evolutionary statistical energy and showed that the change of evolutionary statistical energy ($\Delta\psi_N$) due to amino acid substitutions can capture experimental fitness landscapes and identify deleterious human variants.

In the analysis of the interaction changes ($\Delta\psi_N$) due to single nucleotide nonsynonymous substitutions, we have employed the cutoff distances for pairwise interactions, $r_{cutoff} \sim 8$ and 15.5 Å, which correspond to the first and second interaction shells between residues, respectively. Both the cutoff distances yield similar values for $T_s/T_{s,PDZ} = \overline{Sd}(\Delta\psi_{N,PDZ})/\overline{Sd}(\Delta\psi_N)$; see Fig. S.15. Thus, the differences in the estimation of T_s between these two cutoff distances principally originate in the estimation of T_s for the reference protein, PDZ. The absolute value of $k_B\hat{T}_{s,PDZ}$ for the PDZ has been estimated to be equal to the slope of the reflective regression line of $\Delta\Delta G_{ND}$ on $\Delta\psi_N$. Therefore, as long as the correlation between $\Delta\Delta G_{ND}$ and $\Delta\psi_N$ is good enough as shown in Figs. 2 and S.14, $k_B(\hat{T}_s\overline{\mathrm{Sd}(\psi_N)})_{PDZ}$ takes a similar value irrespective of r_{cutoff}, and the estimate $\hat{T}_{s,PDZ}$ differs depending on $\overline{\mathrm{Sd}(\psi_{N,PDZ})}$. Thus, $\Delta\psi_N$ must correlate with experimental $\Delta\Delta G_{ND}$, but on the basis of the correlation coefficient one cannot determine which estimation of $\Delta\psi_N$ is better. Larger the standard deviation of $\Delta\psi_N$ is, the smaller the estimate of T_s from a direct comparison between $\Delta\Delta G_{ND}$ and $\Delta\psi_N$ is. Including the longer range of pairwise interactions tend to increase the variance of $\Delta\psi_N$. The range of interactions must be limited to a realistic value, either the first interaction shell or the second interaction shell. Thus, the estimates of T_s with $r_{cutoff} \sim 8$ Å and 15.5Å would be upper and lower limits, respectively. Unfortunately T_s is not directly observable. Comparison of the estimates of folding free energies with their experimental values may be appropriate to judge which value is more appropriate for the cutoff distance, although the number of experimental data is limited. Actual values of T_s may be closer to the estimates with $r_{cutoff} \sim 8$Å, because contact predictions based on the estimate of pairwise interactions J succeed for close contacts within the first interaction shell. Also, the estimation of ΔG_{ND} and the correlation between ψ_N^{eq} and $\overline{\psi_N}$ are slightly better with $r_{cutoff} \sim 8$ Å than 15.5 Å; see Figs. 6, 9, S.21, and S.26.

On the basis of the random energy model(REM) (Pande et al., 1997; Shakhnovich and Gutin, 1993a; 1993b), glass transition temperatures (T_g) and folding free energies (ΔG_{ND}) for 14 protein domains are estimated under the condition of $\overline{\psi_N} = \langle\psi_N\rangle_{\boldsymbol{\sigma}}$. The first order transition for protein folding is assumed to estimate the folding free energies by Eq. (44). Selective temperature, T_s, is estimated in the empirical approximation that the standard deviation of $\Delta\psi_N$ is constant across homologous sequences with different ψ_N, so that their estimates may be more coarse-grained, however, this method is easier and faster than the method (Morcos et al., 2014) using the AWSEM (Davtyan et al., 2012). Experimental data for ΔG_{ND} are very limited, and also experimental conditions such as temperature and pH tend to be different among them. A prediction method for folding free energy would be useful in such a situation, although the present method requires the knowledge of melting temperature (T_m) besides sequence data, however, experimental data of T_m are more available than for ΔG_{ND}.

For proteins to have a stable equilibrium value of ψ_N in protein evolution, the regression coefficient of mean interaction change ($\overline{\Delta\psi_N}$) on ψ_N must be more negative than that of their standard deviation ($\mathrm{Sd}(\Delta\psi_N)$), otherwise stabilizing mutations increase as ψ_N decreases. Actually Tables 2 and S.5 show that their mean over all the substitutions at all sites is negatively proportional to ψ_N of a wildtype, but their standard deviation is nearly constant irrespective of ψ_N across homologous sequences. The equilibrium value ψ_N^{eq}, where the average of $\Delta\psi_N$ over fixed mutants is equal to zero, is calculated with the approximation of the distribution of $\Delta\psi_N$ by a log-normal distribution and the empirical rules of Eqs. (62) and (63). In the monoclonal approximation, it has been confirmed that the time average (ψ_N^{eq}) and ensemble average ($\overline{\psi_N} = \langle\psi_N\rangle_{\boldsymbol{\sigma}}$) of evolutionary statistical energy (ψ_N) almost agree with each other. Therefore, this result also supports these approximations and empirical rules, particularly Eq. (63), that is, the constancy of the standard deviation of $\Delta\psi_N$ across homologous sequences. In the log-normal distribution approximation, $\overline{\Delta\psi_N}^{eq}$, $\mathrm{Sd}(\Delta\psi_N)^{eq}$, $\hat{T}_s$, and $\Delta\Delta\hat{G}_{ND}$ can be determined as a function of any one of them. Here they have been shown as a function of $\overline{\Delta\psi_N}^{eq}$.

We have also studied the evolution of protein at equilibrium, at which the ensemble of homologous sequences obeys a Boltzmann distribution with $\exp(-\psi_N)(\simeq \exp(-\Delta\psi_{ND}))$, and the ensemble averages of evolutionary statistical energy ($\psi_N \simeq G_N/(k_BT_s)$) and its change due to a mutation ($\Delta\psi_N \simeq \Delta\Delta\psi_{ND} \simeq \Delta\Delta G_{ND}/(k_BT_s)$) agree with their steady values; $\langle\psi_N\rangle_{\boldsymbol{\sigma}} = \overline{\psi_N} = \psi_N^{eq}$ and $\langle\overline{\Delta\psi_N}\rangle_{\boldsymbol{\sigma}} = \overline{\overline{\Delta\psi_N}} = \overline{\Delta\psi_N}^{eq}$. The PDFs of $\Delta\psi_N$ and K_a/K_s in all the mutants and in their fixed mutants have been estimated. It is confirmed that the effective temperature (T_s) of selection negatively correlates with the amino acid substitution rate (K_a/K_s) of protein.

New alleles can become fixed owing to random drift or to positive selection of substantially advantageous mutations (Gillespie, 1991; Kimura, 1983; Ohta, 2002). The present study indicates that the stability of protein is maintained in such a way that stabilizing mutations are significantly fixed by positive selection, and balance with destabilizing mutations fixed by random drift. As shown in Fig. 14, the almost 50% of fixed mutations are stabilizing mutations fixed by positive selection ($1.05 < K_a/K_s$), and another 50% are destabilizing mutations fixed by random drift. An interesting fact is that contrary to the neutral theory (Kimura, 1968; 1969; Kimura and Ohta, 1971; 1974), the proportion of neutral selection is not large even in fixed mutants. In the selection to maintain protein stability/foldability, neutral mutations fixed with $0.95 < K_a/K_s < 1.05$ are only less than 10%, and slightly negative mutations fixed with $0.5 < K_a/K_s < 0.95$ and negative mutations fixed with $K_a/K_s < 0.5$ are both from 10 to 30%. As a result, at equilibrium the average of K_a/K_s in all the mutants is less than 1, but that in their fixed mutants is larger than 1. The PDF of K_a/K_s shown in Fig. 14 supports the nearly neutral theory (Ohta, 1973; 1992; 2002), which insists that most fixed mutations satisfy $|N_e s| \leq 2$ corresponding to $0.003 \leq K_a/K_s (= u(s)/u(0)) \leq 8$. It should be noted that these conclusions based on the PDFs of $\Delta\psi_N$ and K_s/K_s require only an equilibrium condition of $\overline{\Delta\psi_N} = \overline{\Delta\psi_N}^{eq}$, but does not require the approximation of constancy for the variance of $\Delta\psi_N$ across homologous sequences, which is used only to estimate T_s and ψ_N^{eq} and other relations based on T_s.

In the present study, we have analyzed the mutation-fixation process in equilibrium. The equilibrium state will vary if an environmental condition varies. The evolutionary statistical energy ψ_N and the inverse of selective temperature $1/T_s$ are linearly proportional to the effective population size N_e, as indicated by Eq. (33). Thus, the equilibrium values, ψ_N^{eq}, $\overline{\Delta\psi_N}^{eq}$ and $\mathrm{Sd}(\Delta\psi_N)^{eq}$, are all linearly proportional to the effective population size N_e. On the other hand, $\mathrm{Sd}(\Delta\psi_N)^{eq}$ is not linearly proportional to $\overline{\Delta\psi_N}^{eq}$ but downward-concave, as shown in Fig. 10. As a result, as N_e decreases, $k_BT_s\,\overline{\Delta\psi_N}^{eq} \simeq k_BT_s\,\overline{\Delta\Delta\psi_{ND}}^{eq} (\simeq \overline{\Delta\Delta G_{ND}}^{eq})$ decreases. In other words, the equilibrium value of the mean folding free energy change becomes less positive and therefore that of folding free energy $\overline{\Delta G_{ND}}^{eq} \simeq k_BT_s\,\overline{\Delta\psi_{ND}}^{eq}$ is expected to be higher (less stable) for a smaller number of effective population size N_e; see Eq. (72).

Supplementary document

File 1 — Supplementary methods, tables, and figures

A PDF file in which the details of the methods are described and additional tables and figures are provided; methods, tables, and figures provided in the text are also included as part of their full descriptions for reader's convenience.

Funding This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Acknowledgement

I would like to thank a reviewer for his excellent comments and suggestions that have helped me improve the paper considerably.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.jtbi.2017.08.018

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. doi: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. 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. Des. Select. 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. Nat. 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. Nat. 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. Nat. 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., 2011. Protein 3D structure computed from evolutionary sequence variation. PLoS ONE 6 (12). E28766. doi: 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., 2013. Prediction of contact residue pairs based on co-substitution between sites in protein structures. PLoS ONE 8 (1). E54252. doi: 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., Törö, 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 Rep. 2 (2), 249–256. doi: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., Zöldák, 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.