Selection originating from protein foldability:  
I. A new method to estimate selection temperature  
Sanzo Miyazawa

Abstract

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 total interaction \( \psi_N \) is represented as the sum of one body and pairwise 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 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 of natural selection. Thus, \( k_B T_s \) was estimated as the ratio of \( \Delta G_{ND} \) to \( \Delta \psi_{ND} \) or \( \Delta \Delta G_{ND} \) to \( \Delta \Delta \psi_{ND} \); \( \Delta \Delta G_{ND} \) is a folding free energy change due to single amino acid substitutions. Here, we examine interaction changes (\( \Delta \psi_N \)) due to single nucleotide nonsynonymous mutations, and have found that the variance of their \( \Delta \psi_N \) over all sites hardly depends on the \( \psi_N \) of each homologous sequence, indicating that the variance of \( \Delta G_N (\equiv k_B T_s \Delta \psi_N) \) is nearly constant irrespective of protein families. As a result, \( T_s \) is estimated from the ratio of the variance of \( \Delta \psi_N \) to that of a reference protein, which is determined by a direct comparison between \( \Delta \Delta \psi_{ND} \) and experimental \( \Delta \Delta G_{ND} \). Based on the REM, glass transition temperature \( T_g \) and \( \Delta G_{ND} \) are estimated from \( T_s \) and experimental melting temperatures (\( T_m \)) for 14 protein domains. The estimates of \( \Delta G_{ND} \) agree well with their experimental values for 5 proteins, and those of \( T_s \) and \( T_g \) and \( \Delta \Delta G_{ND} \) are all within a reasonable range. This method is coarse-grained but much simpler in estimating \( T_s \), \( T_g \) and \( \Delta \Delta G_{ND} \) than previous methods.

Insert Received for publication Date and in final form Date. 
sanzo.miyazawa@gmail.com
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. Natural selection maintains protein’s stability and foldability over evolutionary timescales. On the basis of the random energy model (REM) for protein folding, Pande et al. discussed that the equilibrium ensemble of natural protein sequences is a canonical ensemble characterized by a Boltzmann factor $\exp(-\Delta G_{ND}(\sigma)/k_BT_s)$, where $\Delta G_{ND}(\sigma) = G_N(\sigma) - G_D(\sigma)$ is the folding free energy of sequence $\sigma$, $G_N$ 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 of natural selection 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 temperature; see subsection 4.2 for details. 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 $\exp(-G_N(\sigma)/k_BT_s)$, if amino acid frequencies are kept constant.

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 $\exp(-\psi_N(\sigma)s)$, where $\psi_N(\sigma)s$ is the total interaction of a native sequence $\sigma$ and represented as the sum of one-body ($h$) and pairwise ($J$) interactions between residues in the sequence; see subsection 4.2 for details. The one-body ($h$) and pairwise ($J$) residue-residue interactions that satisfy those constraints for homologous sequences have been estimated as one of Potts problems and successfully used to predict contacting residue pairs in protein structures.

Morcos et al. noticed that the $\psi_N$ in the Boltzmann factor is the dimensionless energy corresponding to $G_N/k_BT_s$, and estimated effective temperatures ($T_s$) of natural selection for several protein families by comparing the difference ($\Delta \psi_{ND}$) of $\psi$ between the native and the molten globule states with folding free energies ($\Delta G_{ND}$) estimated with associative memory, water-mediated, structure, and energy model (AWSEM) for natural proteins to fold into unique native structures; protein families are estimated on the basis of the REM from the estimated $T_s$ and an experimental melting temperature $T_m$. The estimates of $\Delta G_{ND}$ are well compared with experimental $\Delta G_{ND}$ for 5 protein families. The present method for estimating $T_s$ for each protein family has been estimated in relative to $T_g$ for the PDZ family, which is determined by directly comparing $\Delta \Delta \psi_{ND}$ with experimental $\Delta G_{ND}$. Also, $T_s$ and the $\Delta G_{ND}$ 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 $\Delta G_{ND}$ are well compared with experimental $\Delta G_{ND}$ for 5 protein families. The present method for estimating $T_s$ is simpler than the method using AWSEM, and also is useful for the prediction of $\Delta G_{ND}$, because the experimental data of $\Delta G_{ND}$ 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 $\psi_N/L$ of each homologous sequence, where $L$ is a sequence length; the average of $\Delta \psi_N$ decreases as $\psi_N/L$ increases. As shown in the succeeding manuscript, this characteristic is required for homologous proteins to stay at the equilibrium state of the native conformational energy $G_N = k_BT_s \psi_N$, and indicates a weak dependency of $\Delta \Delta G_{ND}$ on $\Delta G_{ND}/L$ of protein across protein families.

1 Materials

1.1 Sequence data

We study the single domains of 8 Pfam families and both the single domains and multi-domains from 3 Pfam families. In Table 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” for 5 protein families. To estimate the sample ($\psi_Ns$) and ensemble ($\langle \psi_N \rangle_s$) averages of the total interaction, $M$ unique sequences with no deletion are used. In order to reduce phylogenetic biases in the set of homologous sequences, we employ a sample weight ($w_{seq}$) 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 deletion, which are at least 20% different from each other, are used to calculate the changes of the total interaction ($\Delta \psi_N$) due to single nucleotide nonsynonymous substitutions; the number of the representatives is almost equal to $M_{eff}$ in Table.
2 Results

2.1 Important parameters in the estimations of one-body and pairwise interactions, \( h \) and \( J \), and of the total interaction, \( \psi_N(\sigma) \)

There are two methods available to estimate one-body (\( h \)) and pairwise (\( J \)) interactions for amino acid order in a protein sequence; \( \psi_N(\sigma) = -\sum_{i} h_i(\sigma_i) + \sum_{i < j} J_{ij}(\sigma_i, \sigma_j) \), where \( \sigma = (\sigma_1, \ldots, \sigma_L) \) and \( \sigma_i \) \( \in \) \{amino acids, deletion\}. See the Supporting Material for details. One of them is a method called the pseudo-likelihood maximization (plmDCA) method (6,7) in which a pseudo-likelihood with a \( L_2 \) regularization term is maximized. Another method is the DCA method (4,5) in which pairwise (\( J \)) interactions are estimated in the mean-filed approximation for a Potts problem.

In the plmDCA, a \( L_2 \) regularization term is equal to \( \lambda_0\|h\| + \lambda_1\|J\| \), which is a weighted sum of the vector and matrix norms of \( h \) and \( J \). The \( \lambda_0 \) and \( \lambda_1 \) are a parameter and the optimum values of \( \|h\| \) and \( \|J\| \) depend strongly on their values. In the case of \( \lambda_0 = \lambda_1 = 0.01 \), which seem to be appropriate for contact predictions (6,7), the values of \( \|h\| \) and \( \|J\| \) are too small to yield reasonable values for an effective temperature \( (T_s) \) of selection. To get an appropriate range of values for \( \|h\| \) and \( \|J\| \), even \( \lambda = 0.0001 \) may be too large. Thus, we use the DCA here rather than the plmDCA, although the plmDCA was very successful in contact predictions. In the case of the DCA method, the ratio of pseudocount \((0 \leq p_c \leq 1) \) defined in Eqs. S51 and S52 is a parameter and controls the values of the ensemble and sample averages of \( \psi_N \) in sequence space, \( \langle \psi_N(\sigma) \rangle_\sigma \) in Eq. A25 and \( \psi_N(\sigma_N) \) in Eq. A28, 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 \( J \) (4,5) include a lot of noise as a result of estimating too many interaction parameters from a relatively small number of sequences, and \( J_{ij} \) may take significant values even for site pairs that are distant from each other in the three dimensional structure of protein. Therefore, according to Morcos et al. (11), the estimate of \( J \) is modified as follows.

\[
J_{ij}(a_k, a_l) = J^0_{ij}(a_k, a_l)H(r_{\text{cutoff}} - r_{ij})
\]

where \( J^0 \) is the statistical estimate of \( J \) in a certain gauge, \( a_k \in \{\text{amino acids, deletion}\} \), \( H \) 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 \( r_{\text{cutoff}} \) is a distance threshold for residue pairwise interactions. If pairwise interactions are cut off at a certain distance, the estimate of the total interaction will depend on the choice of gauge for \( h \) and \( J \); see the Method section in the Supplement for details. The sample and ensemble averages of the total interactions per residue over homologous sequences in the PDZ domain family are plotted against the cutoff distance in Fig. S1 in the Supporting Material. The solid and dotted lines indicate the sample and ensemble averages, respectively, and the plus (black) and cross (blue) marks show those for the simple and Ising gauges, respectively. In the simple gauge, \( \sum_k h^0_i(a_k) = \sum_{k,l} J^0_{ij}(a_k, a_l) = 0 \); see Eqs. S56 and S57. On the other hand, one-body interaction \( (h^0_i) \) in the Ising gauge is generated (7) from any gauge by \( h^0_i(a_k) = h^0_i(a_k) - h^0_i(\cdots) + \sum_{j \neq i}(J^0_{ij}(a_k, \cdot) - J^0_{ij}(\cdots, \cdot)) \); see Eqs. S44 and S45. Those for the simple and Ising gauges agree with each other at a sufficiently large value of the cutoff where all pairwise interactions are included. If pairwise interactions are cut off at a certain distance, however, they will yield very different values for the total interaction. Here the simple gauge is employed, because in the Ising gauge the one-body interactions \( (h^0_i) \) include interactions originated in pairwise interactions \( (J^0_{ij}) \) beyond the cutoff distance; see Eq. S45.

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. 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 total interactions must be equal to the ensemble average, \( \bar{\psi}_N = \langle \psi_N(\sigma) \rangle_\sigma \); see Eqs. A25 and A29. 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 \( r_{\text{cutoff}} \sim 8 \) Å and \( r_{\text{cutoff}} \sim 15.5 \) Å, respectively; see Fig. S1. In Fig. S2, the reflective correlation and regression coefficients between the experimental \( \Delta AG_{\text{PDZ}} \) (13) and \( \bar{\psi}_N \) due to single amino acid substitutions are plotted against the cutoff distance for pairwise interactions in the PDZ domain. The reflective correlation has the maximum at the \( r_{\text{cutoff}} \sim 8 \) Å for \( p_c = 0.205 \) and at \( r_{\text{cutoff}} \sim 15.5 \) Å for \( p_c = 0.33 \), indicating that these cutoff distances are appropriate for those ratios of pseudocount. The ratio of pseudocount and a cutoff distance employed are listed for each protein family in Tables 2 and S5 for \( r_{\text{cutoff}} \sim 8 \) and 15.5 Å, respectively.
2 RESULTS

2.2 Changes of the total interaction, \(\Delta \psi_N\), by single nucleotide nonsynonymous substitutions

The change of the total interaction of a native conformation by a single amino acid substitution from \(\sigma^N_i\) to \(\sigma_i\) at site \(i\) in a native sequence \(\sigma_N\) is defined as

\[
\Delta \psi_N(\sigma^N_{j\neq i}, \sigma_i^N \rightarrow \sigma^N_i) \equiv \psi_N(\sigma^N_{j\neq i}, \sigma_i^N) - \psi_N(\sigma_N)
\]

\(2\)

\[
\Delta \psi_D(\sigma^N_{j\neq i}, \sigma_i^N \rightarrow \sigma_i, T) \equiv \psi_D(\sigma^N_{j\neq i}, \sigma_i^N, T) - \psi_D(\sigma_N, T)
\]

\(3\)

\[
\Delta \psi_D(\sigma^N_{j\neq i}, \sigma_i^N \rightarrow \sigma_i) \equiv \Delta \psi_N(\sigma^N_{j\neq i}, \sigma_i^N \rightarrow \sigma_i) - \Delta \psi_D(\sigma^N_{j\neq i}, \sigma_i^N \rightarrow \sigma_i)
\]

\(4\)

\[
\approx \Delta \psi_N(\sigma^N_{j\neq i}, \sigma_i^N \rightarrow \sigma_i)
\]

\(5\)

where \(L\) is sequence length. This relationship is found in all of the protein families examined here; the correlation and regression coefficients for \(r_{\text{cutoff}} \sim 8\) and 15.5\(\angstrom\) are listed in Tables 2 and S5, 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 total interaction \(\Delta \psi_N\) depends on interaction per residue \(\psi_N/L\) rather than the total interaction \(\psi_N\), because interactions change only for one residue substituted in the sequence. On this line of consideration, the following approximation for the slope is confirmed in Fig. S8.

\[
\alpha_{\psi_N} \approx \frac{\Delta \psi_N(\sigma^N_{j\neq i}, \sigma_i^N \rightarrow \sigma_i)}{L(-\delta \psi^2/L)}
\]

\(6\)

where \(\psi_N\) of the wildtype and \(\psi_N/L\) shown in Fig. 1 and Tables 2 and S5 are 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 selection temperatures \(T_s\) of homologous sequences in a protein family are approximated to be equal.

It is the same type of dependence on \(\psi_N/L\) found for the standard deviation of \(\Delta \psi_N\) over single nucleotide nonsynonymous substitutions at all sites? Fig. 1 and Tables 2 and S5 show that the correlation between the standard deviation of \(\Delta \psi_N\) and \(\psi_N\) of the wildtype is very weak except for Nitroreductase, SBP_bac,3 and LysR,substrate families. Even for these protein families, the regression coefficients are less negative than those for \(\Delta \psi_N\). 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 \(\psi_N\) across homologous sequences.

2.3 Effective temperature \(T_s\) of selection estimated from the changes of interaction, \(\Delta \psi_N\), by single nucleotide nonsynonymous substitutions

The variance of \(\Delta \psi_N\) must be approximated by a function of \(k_BT_s\), because it does not depend on \(\psi_N\) of the wildtype and is nearly constant across homologous sequences in every protein family that has its own characteristic temperature \(T_s\) for natural selection. On the other hand, the free energy of the native structure, \(\Delta G_N\), may be approximated by a function of \(G_N\) but must not explicitly depend on \(k_BT_s\). In other words, the following relationships are derived.

\[
\text{Var}(\Delta \psi_N(\sigma^N_{j\neq i}, \sigma_i^N \rightarrow \sigma_i)) \approx \text{constant across homologous sequences in every protein family}
\]

\(8\)

\[
= \text{function of } k_BT_s
\]

\(9\)

\[
\text{Var}(\Delta G_N(\sigma^N_{j\neq i}, \sigma_i^N \rightarrow \sigma_i)) = \text{function that must not explicitly depend on } k_BT_s \text{ but } G_N
\]

\(10\)
From the equations above, we obtain the important relation that the variance of $\Delta G_N (= k_BT_s \Delta \psi_N)$ does not depend on $G_N$ and is nearly constant irrespective of protein families.

$$\text{Var}(\Delta G_N (\sigma_{j\neq i}^N, \sigma_i^N \to \sigma_i)) = (k_BT_s)^2 \text{Var}(\Delta \psi_N (\sigma_{j\neq i}^N, \sigma_i^N \to \sigma_i)) \approx \text{constant}$$

This relationship is consistent with the observation that the variance of $\Delta \Delta G_N (= \Delta \psi_N)$ is nearly constant irrespectively of protein families $\Delta \psi_N (\neq G_N)$ [15]. This relationship allows us to estimate a selection 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 $\Delta \Delta G_N$: because the experimental data of $\Delta \Delta G_N$ are available for many types of single amino acid substitutions in the PDZ domain.

$$k_BT_s = k_BT_{s,\text{PDZ}} \left( \frac{\text{Var}(\Delta \psi_{\text{PDZ}} (\sigma_{j\neq i}^N, \sigma_i^N \to \sigma_i)) / \text{Var}(\Delta \psi_N (\sigma_{j\neq i}^N, \sigma_i^N \to \sigma_i))}{1/2} \right)$$

where the overline denotes the average over all homologous sequences. Here, the averages of variances over all homologous sequences are employed, because $T_s$ for all homologous sequences are approximated to be equal.

### 2.4 A direct Comparison of the changes of interaction, $\Delta \psi_N (= \Delta \Delta \psi_N)$, with the experimental $\Delta \Delta G_N$ due to single amino acid substitutions

The effective temperature ($T_s$) of selection for the PDZ family has been estimated by directly comparing $\Delta \psi_N (= \Delta \Delta \psi_N)$ with experimental $\Delta \Delta G_N$ [13] for single amino acid substitutions. In Fig. 2, the experimental values [15] of $\Delta \Delta G_N$ 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. The slopes of the least-squares regression lines through the origin, which are estimates of $k_BT_s$, are equal to $k_BT_s = 0.279$ kcal/mol for $r_{\text{cutoff}} \sim 8 \text{ Å}$ and $k_BT_s = 0.162$ kcal/mol for $r_{\text{cutoff}} \sim 15.5 \text{ Å}$, and the reflective correlation coefficients are equal to 0.93 and 0.94, respectively.

These estimates of $k_BT_s$, 0.279 and 0.162, for the PDZ yield $(\text{Var}(\Delta \psi_{\text{PDZ}}))^{1/2} = 1.28 \text{ kcal/mol}$, which corresponds to 75% of 1.7 kcal/mol [17] estimated from ProTherm database or 79% of 1.63 kcal/mol [16] computationally predicted for single nucleotide mutations by using the FoldX. Using $(\text{Var}(\Delta \Delta G_N))^{1/2} = 1.28 \text{ kcal/mol}$ estimated from $T_s$ for PDZ, the absolute values of $T_s$ for other proteins are calculated by Eq. 12 and listed in Table 3 and S6. Fig. 3 shows that $r_{\text{cutoff}} \sim 15.5 \text{ Å}$ yields slightly larger values in a scale relative to the $T_s$ of the PDZ but lower absolute values of $T_s$ for all the proteins than $r_{\text{cutoff}} \sim 8 \text{ Å}$.

Morcos et al. [11] estimated $T_s$ by comparing $\Delta \psi_N$ with $G_N$ estimated by the associative memory, water-mediated, structure, and energy model (AWSM). They estimated $\psi_N$ with $r_{\text{cutoff}} = 16 \text{ Å}$ and probably $p_c = 0.5$; the gauge employed is unknown. In Fig. S7 the present estimates of $T_s$ are compared with those by Morcos et al. [11].

### 2.5 Expected dependency of folding free energy change ($\Delta \Delta G_N$) due to single nucleotide nonsynonymous substitutions on folding free energy per residue ($\Delta G_N/L$)

Let us consider the average of $\Delta \psi_N$ over homologous sequences in each protein family. The following regression line is shown in Fig. 4.

$$\overline{\Delta \psi_N (\sigma_{j\neq i}^N, \sigma_i^N \to \sigma_i)} \approx \frac{\overline{\psi_N (\sigma_N) - \tilde{f} (\sigma_N)}}{L} + \beta_{\overline{\psi_N}}$$

(13)

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

(14)

Here, $\psi_N (\sigma_N)$ is reduced by $\tilde{f}$ because the origin of the $\psi_N$ scale is not unique. The correlation between $\overline{\Delta \psi_N}$ and $\delta \tilde{f} / L$ is significant; the correlation coefficient is larger than 0.99. It should be noted here that the intercept $\beta_{\overline{\psi_N}}$ should be equal to 0, because if $T_s \to \infty$ then $\delta \tilde{f} \to 0$ and $\Delta \psi_N \to 0$. Actually, Fig. 4 shows that $\beta_{\overline{\psi_N}}$ is nearly equal to 0.

Biophysical Journal 00(00) 1–S-24
Finally, the dependence of $\Delta \Delta G_{ND}$ on $G_{ND}$ can be predicted.

$$\Delta \Delta G_{ND}(\sigma^N_{p_i},\sigma^N_{\psi_i}) \approx -\alpha_{\omega}k_BT_s \frac{\delta \omega^2(f(\sigma_N))}{L} + \beta_{\omega}$$

In general, $T_s$ and $T_m$ are different among protein families, so that the correlation between $\Delta \Delta G_{ND}$ and $\langle G_{ND} \rangle / L$ cannot be strong. In Fig. 5, $\Delta \Delta G_{ND}$ for the present proteins are plotted against $\langle G_{ND} \rangle / L$. However, it should be noted that the correlation is not expected for $\Delta \Delta G_{ND}$ and $\langle G_{ND} \rangle$ but for $\Delta \Delta G_{ND}$ and $\langle G_{ND} \rangle / L$.

2.6 Estimation of $T_s$, $\omega$, and $\langle G_{ND}(\sigma) \rangle$ from $T_s$ and $T_m$

To estimate glass transition temperature, $T_g$, the conformational entropy per residue $\omega$ in the compact denatured state, and the ensemble average of folding free energy in sequence space $\langle G_{ND} \rangle$, melting temperature $T_m$ must be known for each protein; see Eqs. A30 and A27 for $T_s$, $\omega$ and $\langle G_{ND} \rangle$, respectively. The experimental value of $T_m$ employed for each protein is listed in Table 3 and S6. For comparison, temperature $T$ is set up to be equal to the experimental temperature for $G_{ND}$ or to 298 K if unknown.

An estimate of glass transition temperature, $T_g$, has been calculated from $T_s$ and $T_m$ by Eq. A30, and is listed in Table 3 and S6 for each protein. In Fig. 6, $T_g/T_s$ is plotted against $T_m/T_s$ 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 $T_m/T_s$ increases. A condition, $\Delta \Delta G_{ND} = 0$ at $T = T_m$, for the first order transition requires that Eq. A30 which is verified for $\Delta \Delta G_{ND}$, $T_s$, and $\langle G_{ND} \rangle$ is satisfied. As a result, $T_g/T_s$ must be lowered to increase $T_m/T_s$; 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 of values required for protein foldability.

In Table 3 and S6 the ensemble average of $G_{ND}(\sigma)$ over sequences calculated by Eq. A27 and the conformational entropy per residue $\omega$ in the compact denatured state by Eq. A18 are also listed for each protein. Fig. 7 shows the distribution of $\omega$, $\omega$ estimated from the condition for the first order transition falls into the ranges of 0.6–1.1 $k_B$ for $r_{cutoff} = 8$ Å and 0.8–1.4 $k_B$ for $r_{cutoff} = 15.5$ Å. This range for $r_{cutoff} = 8$ Å agrees well with the range estimated by Morcos et al. (11).

3 Discussion

We have analyzed the interaction changes (\Delta \psi_N) due to single nucleotide nonsynonymous substitutions. As studied in the succeeding manuscript (13), the regression coefficient of their mean (\bar{\Delta \psi_N}) on \psi_N must be more negative than that of their standard deviation (Sd(\Delta \psi_N)), otherwise the folding free energy, $G_{ND} = k_BT_s \Delta \psi_{ND}$, of protein could not have a stable equilibrium value; actually Tables 2 and 3 show that their mean over all the substitutions at all sites is negatively proportional to \psi_N of a wildtype, but their standard deviation is nearly constant irrespective of \psi_N across homologous sequences.

On the basis of the random energy model (REM) (3), the effective temperatures ($T_e$), natural selection, glass transition temperatures ($T_g$), and folding free energies ($G_{ND}$) for 14 protein domains are estimated in the empirical approximation that the variance of $\Delta \psi_N$ is constant across homologous sequences with different $\psi_N$, so that their estimates may be coarse-grained, however, this method is easier and faster than the method (1) using the AWSEM (12). Experimental data for $G_{ND}$ are 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 $G_{ND}$.

We have employed the cutoff distances for pairwise interactions, $r_{cutoff} = 8$ and 15.5 Å, which correspond to the first and second interaction shells between residues, respectively. Also the ratio of pseudocount $p_j$, was chosen to yield the sample and ensemble averages are equal, i.e., $\bar{\psi}_N = \psi_N$. Morcos et al. (11) employed probably $p_j = 0.5$, which was successfully employed in contact prediction, and chose $r_{cutoff} = 16$ Å on the basis of the correlation between $\Delta \psi_{ND}$ and $\Delta \psi_N = \Delta \psi_N$ due to single amino acid substitutions for the PDZ protein family. In the present method, $p_j = 0.5$ cannot be used, because
Fig. S1 shows that for $p_c = 0.5$, $\psi_N$ coincides with $\langle \psi_N \rangle_a$ at $r_{\text{cutoff}} \approx 8\text{Å}$ in the Ising gauge but is more negative than $\langle \psi_N \rangle_a$ at any cutoff distance in the simple gauge. The condition of $\psi_N = \langle \psi_N \rangle_a$ is required to estimate folding free energies $\Delta G_{ND}$ with Eq. A27 on the basis of the REM.

As shown in Fig. S2, the estimates of $T_s$ depend on the cutoff distance. The estimates of $T_s$ are determined in relation to $T_s$ of the PDZ, which has been estimated to be equal to the reflective regression coefficient of experimental $\Delta G_{ND}$ on $\Delta \psi_N (\approx \Delta \Delta \psi_N)$, although the variance of $\Delta \psi_N$ is, the smaller the estimate of $T_s$ is. Including the longer range of pairwise interactions increases the variance of $\Delta \psi_N$. Correlation between $\Delta \psi_N$ and $\Delta G_{ND}$ is not a good measure to judge which cutoff distance is better. $T_s$ 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. It is not certain at this time which value better suits the present analysis, but consistencies between various quantities, particularly between $\langle \Delta G_{ND} \rangle_a$ and experimental $\Delta G_{ND}$, may indicate that $r_{\text{cutoff}} \sim 8\text{Å}$ slightly better suits the present method than $r_{\text{cutoff}} \sim 15.5\text{Å}$; see Figs. S3 and S4.

4 Appendix

4.1 Knowledge of protein folding

A protein folding theory (3), which is based on a random energy model (REM), indicates that the equilibrium ensemble of amino acid sequences, $\sigma = (\sigma_1, \cdots, \sigma_L)$ where $\sigma_i$ is the type of amino acid at site $i$ and $L$ is sequence length, should be a canonical ensemble with a Boltzmann factor consisting of the folding free energy, $\Delta G_{ND}(\sigma, T)$ and an effective temperature $T_s$ of natural selection.

$$P(\sigma) \propto p_{\text{mut}}(\sigma) \exp(-\Delta G_{ND}(\sigma, T) / k_BT_s)$$  \text{(A1)}$$

$$\Delta G_{ND}(\sigma, T) \equiv G_N(\sigma) - G_D(f(\sigma), T)$$  \text{(A2)}$$

$$\text{where } p_{\text{mut}}(f(\sigma)) \text{ is the probability of a sequence } (\sigma) \text{ randomly occurring in a mutational process and depends only on the amino acid frequencies } f(\sigma), k_B \text{ is the Boltzmann constant, } T \text{ is a growth temperature, and } G_N \text{ and } G_D \text{ are the free energies of the native conformation and denatured state, respectively. Selection temperature } T_s \text{ quantifies how strong the folding constraints have been during evolution (11), and is specific to the protein structure and function. The free energy } G_D \text{ of the denatured state does not depend on the amino acid order but the amino acid composition, } f(\sigma), \text{ in a sequence. It is reasonable to assume that mutations independently occur between sites, and therefore the equilibrium frequency of a sequence in the mutational process is equal to the product of the equilibrium frequencies over sites; } P_{\text{mut}}(\sigma) = \prod_i p_{\text{mut}}(\sigma_i), \text{ where } p_{\text{mut}}(\sigma_i) \text{ is the equilibrium frequency of } \sigma_i \text{ at site } i \text{ in the mutational process.}$$

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) (3), to be equal to the energy distribution of randomized sequences, which is then approximated by a Gaussian distribution, in the native conformation. That is, a 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.

$$Z = \int \exp(-E/k_BT_s) n(E)dE$$  \text{(A4)}$$

$$\bar{n}(E) = \exp(\omega L) N(\bar{E}(f(\sigma)), \delta E^2(f(\sigma)))$$  \text{(A5)}$$

where $\omega$ is the conformational entropy per residue in the compact denatured state, and $N(\bar{E}(f(\sigma)), \delta E^2(f(\sigma)))$ is the Gaussian probability density with mean $\bar{E}$ and variance $\delta E^2$, which depend only on the amino acid composition of the protein sequence. The free energy of the denatured state is approximated as follows.

$$G_D(f(\sigma), T) \approx \bar{E}(f(\sigma)) - \frac{\delta E^2(f(\sigma))}{2k_BT} - k_BT\omega L$$  \text{(A6)}$$

$$\bar{E}(f(\sigma)) = \bar{E}(f(\sigma)) - \delta E^2(f(\sigma)) \frac{T/T_g}{k_BT}$$  \text{(A7)}$$

$$\bar{T}(T/T_g) \equiv \begin{cases} \frac{1}{T} + \frac{T}{T_g} & \text{for } T > T_g \\ \frac{T}{T_g} & \text{for } T \leq T_g \end{cases}$$  \text{(A8)}$$
where $\bar{E}$ and $\delta E^2$ 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 \cite{3}; $-\partial \bar{G}_D/\partial T|_{T=T_g} = 0$. The conformational entropy per residue $\omega$ in the compact denatured state can be represented with $T_g$; $\omega L = \delta E^2/(2k_BT_g^2)$. 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.

4.2 Probability distribution of homologous sequences in sequence space

The probability distribution $P(\sigma)$ of sequences, $\sigma \equiv (\sigma_1, \ldots, \sigma_J)$ where $\sigma_i \in \{$amino acids, deletion$\}$, with maximum entropy in sequence space that satisfies a given amino acid frequency at each site and a given pairwise amino acid frequency at each conformational entropy per residue $\bar{T}$.

$$P(\sigma) \propto \exp(-\psi_N(\sigma))$$ (A9)

where $h_i$ and $J_{ij}$ are one-body and two-body interactions and must satisfy the following constraints.

$$\sum_{\sigma, \sigma_i = a_k} P(\sigma) = P_i(a_k)$$ (A11)

$$\sum_{\sigma, \sigma_i = a_k, \sigma_j = a_l} P(\sigma) = P_{ij}(a_k, a_l)$$ (A12)

where $P_i(a_k)$ 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 $j$; $a_k \in \{$amino acids, deletion$\}$. The pairwise interaction matrix $J$ satisfies $J_{ij}(a_k, a_l) = J_{ji}(a_l, a_k)$ and $J_{ij}(a_k, a_l) = 0$. Interactions $h_i$ and $J_{ij}$ can be well estimated from a multiple sequence alignment (MSA) in the mean field approximation \cite{4,5}, or by maximizing a pseudo-likelihood \cite{6,7}. Here we must notice that $\psi_N(\sigma)$ has been estimated under the constraints on amino acid compositions at all sites, and therefore the amino acid composition of a whole sequence must be constant across sequences.

From Eqs. \[A2\] and [A9],

$$G_N(\sigma) = k_BT_s[\psi_N(\sigma) + \text{function of } f(\sigma)]$$ (A13)

$$G_D(f(\sigma), T) = k_BT_s[\psi_D(\sigma) + \text{function of } f(\sigma)]$$ (A14)

$$\Delta G_{ND}(\sigma, T) = k_BT_s\Delta \psi_{ND}(\sigma)$$ (A15)

$$\Delta \psi_{ND}(\sigma) \equiv \psi_N(\sigma) - \psi_D(\sigma)$$ (A16)

$$\omega = \left(T_s/T_g\right)^2\delta \psi^2/(2L)$$ (A17)

where $\bar{E} = k_BT_s\bar{\psi}$ and $\delta E^2 = (k_BT_s)^2\delta \psi^2$. The mean $\bar{\psi}$ and variance $\Delta \psi^2$ are estimated as the mean and variance of $\psi_N$ over randomized sequences in the native structure \cite{3}.

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

The ensemble average of $\Delta G_{ND}(\sigma, T)$ over sequences with Eq. \[A1\] is

$$\langle \Delta G_{ND}(\sigma, T) \rangle_\sigma$$ (A19)

$$\equiv \left[ \sum_\sigma \Delta G_{ND}(\sigma, T)P^{\text{mut}}(\sigma)\exp(-\frac{\Delta G_{ND}(\sigma, T)}{k_BT_s}) \right] / \left[ \sum_\sigma P^{\text{mut}}(\sigma)\exp(-\frac{\Delta G_{ND}(\sigma, T)}{k_BT_s}) \right]$$ (A20)

$$\approx \left[ \sum_{\sigma | f(\sigma) = f_N} G_N(\sigma)\exp(-\frac{G_N(\sigma)}{k_BT_s}) \right] / \left[ \sum_{\sigma | f(\sigma) = f_N} \exp(-\frac{G_N(\sigma)}{k_BT_s}) \right] - G_D(f_N(T))$$ (A21)

$$= \langle G_N(\sigma) \rangle_N - G_D(f(N), T)$$ (A22)
where $\sigma_N$ denotes a native sequence, and $f(\sigma_N)$ denotes the average of $f(\sigma_N)$ over homologous sequences. In Eq. [A21], the sum over all sequences is approximated by the sum over sequences the amino acid composition of which is the same as that of the native sequences. The ensemble average of $G_N$ is also estimated in the Gaussian approximation [3].

$$\langle G_N(\sigma) \rangle_{\sigma} = \frac{\int E \exp(-E/(k_B T_s)) n(E) dE}{\int \exp(-E/(k_B T_s)) n(E) dE} \quad (A23)$$

$$\langle \psi_N(\sigma) \rangle_{\sigma} = \frac{\int \psi f(\sigma_N) - \delta E^2 f(\sigma_N)/(k_B T_s) dE}{\int \exp(-E/(k_B T_s)) n(E) dE} \quad (A24)$$

$$\langle \psi_N(\sigma) \rangle_{\sigma} = \frac{\int \psi f(\sigma_N) - \delta \psi^2 f(\sigma_N) dE}{\int \exp(-E/(k_B T_s)) n(E) dE} \quad (A25)$$

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

$$\Delta G_{ND}(\sigma_N, T) = \langle \Delta G_{ND}(\sigma, T) \rangle_{\sigma} \quad (A26)$$

$$\psi_N(\sigma_N) = \frac{\sum_{\sigma_s} w_{\sigma_s} \psi_N(\sigma_N)}{\sum_{\sigma_s} w_{\sigma_s}} \quad (A27)$$

where the overline denotes sample average with a sample weight $w_{\sigma_s}$ 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 $T_m$; $\langle \Delta G_{ND}(\sigma_N, T_m) \rangle_{\sigma} = 0$. Thus, the following relationship must be satisfied [3].

$$\vartheta(T_m/T_s) T_s = T_s \left(1 + \frac{T_s^2}{T_g^2} \right) = 1 \quad \text{with } T_s \leq T_g \leq T_m \quad (A30)$$

5 Supporting Citation

References [20–36] appear in the Supporting Material.

SUPPLEMENTARY MATERIAL

An online supplement to this article can be found by visiting BJ Online at http://www.biophysj.org.

References

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

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

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

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

5. Marks, D. S., L. J. Colwell, R. Sheridan, T. A. Hopf, A. Pagnani, R. Zecchina, and C. Sander, 2011. Protein 3D Structure Computed from Evolutionary Sequence Variation. PLoS ONE 6:e28766. http://dx.doi.org/10.1371/journal.pone.0028766

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

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

8. Miyazawa, S., 2013. Prediction of Contact Residue Pairs Based on Co-Substitution between Sites in Protein Structures. PLoS ONE 8:e54252. http://dx.doi.org/10.1371/journal.pone.0054252

9. Sulikowska, J. I., F. Morcos, M. Weigt, T. Hwa, and J. N. Onuchic, 2012. Genomics-aided structure prediction. Proc. Natl. Acad. Sci. USA 109:10340–10345.
10. Hopf, T. A., L. J. Colwell, R. Sheridan, B. Rost, C. Sander, and D. S. Marks, 2012. Three-Dimensional Structures of Membrane Proteins from Genomic Sequencing. Cell 149:1607–1621.

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

12. Davtyan, A., N. P. Schaefer, W. Zheng, C. Clementi, P. G. Wolynes, and G. A. Papoian, 2012. AWSEM-MD: Protein Structure Prediction Using Coarse-Grained Physical Potentials and Bioinformatically Based Local Structure Biasing. J. Phys. Chem. B 116:8494–8503.

13. Miyazawa, S., 2016. Selection originating from protein foldability: Folding free energy, sequence ensemble, and fitness.

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

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

16. Tokuriki, N., F. Stricher, J. Schymkowitz, L. Serrano, and D. S. Tawfik, 2007. The Stability Effects of Protein Mutations Appear to be Universally Distributed. J. Mol. Biol. 369:1318–1332.

17. Serohijos, A., Z. Rimas, and E. Shakhnovich, 2012. Protein Biophysics Explains Why Highly Abundant Proteins Evolve Slowly. Cell Reports 2:249 – 256. [http://dx.doi.org/10.1016/j.celrep.2012.06.022]

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

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

20. Ganguly, T., M. Das, A. Bandhu, P. K. Chanda, B. Jana, R. Mondal, and S. Sau, 2009. Physicochemical properties and distinct DNA binding capacity of the repressor of temperate Staphylococcus aureus phage φ11. FEBS J. 276:1975–1985.

21. Wilson, C. J., and P. Wittung-Stafshede, 2005. Snapshots of a Dynamic Folding Nucleus in Zinc-Substituted Pseudomonas aeruginosa Azurin. Biochemistry 44:10054–10062.

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

23. Stupak, M., G. Zoldák, A. Musatov, M. Sprinzl, and E. Sedláček, 2006. Unusual e

24. Grantcharova, V. P., D. S. Riddle, J. V. Santiago, and D. Baker, 1998. Important role of hydrogen bonds in the structurally polarized transition state for folding of the src homology 3 domain. Nature Struct. 5:44–50.

25. Parsons, L. M., F. Lin, and J. Orban, 2006. Peptidoglycan Recognition by Pal, an Outer Membrane Lipoprotein. Biochemistry 45:2122–2128.

26. Williams, N. K., P. Prosselkov, E. Liepins, I. Line, A. Sharipo, D. R. Littler, P. M. G. Curmi, G. Otting, and N. E. Dixon, 2004. In Vivo Protein Cyclization Promoted by a Circularly Permuted Synechocystis sp. PCC6803 DnaB Mini-intein. J. Biol. Chem. 279:16800–16806.

27. Sainsbury, S., J. Ren, N. Saunders, D. I. Stuarta, and R. J. Owens, 2008. Crystallization and preliminary X-ray analysis of CrgA, a LysR-type transcriptional regulator from pathogenic Neisseria meningitidis MC38. Acta Cryst. F64:797–801.

28. Armengaud, J., J. Urbonavicius, B. Fernandez, G. Chauvissand, J. M. Bujnicki, and H. Grosjean, 2004. N²-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.BiolChem 279:37142–37152.

29. Guelorget, A., M. Roovers, V. Guérineau, C. Barbey, X. Li, and B. Golinelli-Pimpaneau, 2010. Insights into the hyperthermostability and unusual region-specificity of archaeal Pseudomonas aeruginosa. J. Biol. Chem. 285:246–253.

30. Knapp, S., P. T. Mattson, P. Christova, K. D. Berndt, A. Karshiko, M. Vihinen, C. E. Smith, and R. Ladenstein, 1998. Thermal Unfolding of Small Proteins With SH3 Domain Folding Pattern. Proteins 31:309–319.

31. Grantharova, V. P., D. S. Riddle, J. V. Santiago, and D. Baker, 1998. Important role of hydrogen bonds in the structurally polarized transition state for folding of the src homology 3 domain. Nature Struct. 5:44–50.

32. Onwukwe, G. U., P. Kurszul, M. K. Koski, W. Schmitz, and R. R. Wierenga, 2014. Human Δ1,Δ2-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.

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

34. Torchio, G. M., M. R. Ermacora, and M. P. Sica, 2012. Equilibrium Unfolding of the PDZ Domain of β2-Syntrophin. Biophys. J. 102:2835–2844.

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

36. Rosa, C. L., D. Milardi, D. Grasso, R. Guzzi, and L. Sportelli, 1995. Thermodynamics of the Thermal Unfolding of Azurin. J. Phys. Chem. 99:14864–14870.
Table 1:  **Protein families, and structures studied.**

| Pfam family | UniProt ID      | $N^a$ | $N_{\text{eff}}^{bc}$ | $M^a$ | $M_{\text{eff}}^{ce}$ | $L^f$ | PDB ID     |
|------------|-----------------|-------|------------------------|-------|------------------------|-------|------------|
| HTH3       | RPC1_BP434/7-59 | 15315 | 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|
| SBP_bac3   | GLNH_ECOLI/27-244 | 9874(9972) | 7374.96         | 140   | 99.70                  | 218   | 1WDN-A:5-222|
| SBP_bac3   | 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 | BENM_ACIAD/90-280 | 25138(25226) | 20707.06       | 85(1) | 67.00                  | 191   | 2F6G-A/B:90-280|
| LysR_substrate | BENM_ACIAD/163-265 | 25032(25164) | 21144.74       | 121(1) | 99.27                 | 103   | 2F6G-A/B:163-265|
| Methyltransf5 | RSMH_THEMA/8-292 | 1942(1953) | 1286.67        | 578(2) | 357.97                | 285   | 1N2X-A:8-292|
| Methyltransf5 | 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-A/B/C/D:4-128|

$^a$ The number of unique sequences and the total number of sequences in parentheses; the full alignments in the Pfam (14) are used.  
$^b$ The effective number of sequences.  
$^c$ A sample weight ($w_{\sigma_N}$) for a given sequence is equal to the inverse of the number of sequences that are less than 20% different from the given sequence.  
$^d$ The number of unique sequences that include no deletion unless specified. The number in parentheses indicates the maximum number of deletions allowed.  
$^e$ The effective number of unique sequences that include no deletion or at most the specified number of deletions.  
$^f$ The number of residues.  
$^g$ These proteins consist of two domains, and other ones are single domains.
Table 2:  **Parameter values for** \( r_{\text{cutoff}} \approx 8 \, \text{Å} \) **employed for each protein family,** and the averages of the total interactions \( (\bar{\psi}_N) \) over all homologous sequences and of the means and the variances of interaction changes (\( \bar{\Delta\psi}_N \) and \( \text{Var}(\Delta\psi_N) \)) due to single nucleotide nonsynonymous mutations at all sites over all homologous sequences in each protein family.

| Pfam family   | \( L \) | \( p_c \) | \( n_s \) | \( r_{\text{cutoff}} \) (Å) | \( \bar{\psi}_N/L \) | \( \delta\psi^2/L \) | \( \bar{\Delta\psi}_N/c \) | \( \text{Var}(\Delta\psi_N)^c \) | \( r_{\psi N} \) for \( \bar{\Delta\psi}_N \) | \( \alpha_{\psi N} \) for \( \text{Var}(\Delta\psi_N)^{1/2} \) |
|---------------|------|------|-------|------------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| HTH\_3        | 53   | 0.175| 7.434 | 8.22             | -0.2039         | 2.8837          | -3.0712         | 4.0412          | 30.8500         | -0.94           | -1.3671         | -0.23           | -0.1741         |
| Nitroreductase | 73   | 0.24 | 6.411 | 8.39             | -0.1153         | 2.0825          | -2.1975         | 2.9168          | 11.0879         | -0.94           | -1.3915         | -0.59           | -0.5270         |
| SBP\_bac\_3   | 218  | 0.25 | 9.230 | 8.10             | -0.1000         | 2.1624          | -2.2618         | 3.1534          | 12.2140         | -0.98           | -1.5267         | -0.84           | -0.8111         |
| SBP\_bac\_3   | 94   | 0.37 | 8.000 | 7.90             | -0.1634         | 1.2495          | -1.4054         | 1.8075          | 5.6101          | -0.96           | -1.4191         | -0.65           | -0.5171         |
| OmpA          | 95   | 0.169| 8.000 | 8.20             | -0.2457         | 3.9093          | -4.1542         | 6.0657          | 55.3105         | -0.96           | -1.5882         | -0.47           | -0.4471         |
| DnaB          | 98   | 0.235| 9.653 | 8.17             | -0.2284         | 3.9976          | -4.2291         | 6.1006          | 38.5736         | -0.96           | -1.4654         | -0.52           | -0.4385         |
| LysR\_substrate| 191  | 0.23 | 7.843 | 7.87             | -0.2248         | 1.4450          | -1.6712         | 1.9999          | 6.7385          | -0.97           | -1.3646         | -0.52           | -0.5814         |
| LysR\_substrate| 103  | 0.265| 8.835 | 8.25             | -0.2240         | 1.4132          | -1.6372         | 2.0309          | 7.4763          | -0.99           | -1.4629         | -0.78           | -0.6311         |
| Methyltransfer\_S | 285  | 0.13 | 7.993 | 7.78             | -0.1462         | 7.2435          | -7.3887         | 12.2738         | 127.1400        | -0.99           | -1.7945         | -0.08           | -0.0461         |
| Methyltransfer\_S | 80   | 0.18 | 6.775 | 7.85             | -0.1763         | 5.5162          | -5.6896         | 8.7386          | 61.5293         | -0.96           | -1.5601         | 0.06            | 0.0612          |
| SH3\_l        | 48   | 0.14 | 6.417 | 8.01             | -0.1348         | 3.9109          | -4.0434         | 5.4273          | 39.8445         | -0.94           | -1.4457         | -0.22           | -0.1995         |
| ACBP          | 80   | 0.22 | 9.175 | 8.24             | -0.0525         | 4.6411          | -4.7084         | 7.8272          | 55.1762         | -0.97           | -1.6189         | -0.30           | -0.1986         |
| PDZ           | 81   | 0.205| 9.061 | 8.16             | -0.2398         | 3.1140          | -3.3572         | 4.3897          | 21.0789         | -0.96           | -1.5014         | -0.43           | -0.3293         |
| Copper-bind   | 125  | 0.23 | 9.232 | 8.26             | -0.0838         | 4.1946          | -4.2657         | 7.2514          | 51.7793         | -0.98           | -1.9254         | -0.20           | -0.1475         |

\( ^a \) The average number of contact residues per site within the cutoff distance; the center of side chain is used to represent a residue.

\( ^b \) \( M \) unique sequences without deletions are used with a sample weight \( (n_{\text{w}}) \) for each sequence; \( n_{\text{w}} \) 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 \( M_{\text{eff}} \) of the sequences are listed for each protein family in Table 4.

\( ^c \) Representatives of unique sequences without deletions, which are at least 20\% different from each other, are used; the number of the representatives used is almost equal to \( M_{\text{eff}} \). For HTH\_3, 4431 sequences are used; abnormal sequences are removed.

\( ^d \) The correlation and regression coefficients of \( \bar{\Delta\psi}_N \) on \( \bar{\psi}_N/L \); see Eq. 6.

\( ^e \) The correlation and regression coefficients of \( \text{Var}(\Delta\psi_N) \) on \( \bar{\psi}_N/L \).
Table 3: Thermodynamic quantities estimated with $r_{\text{cutoff}} \sim 8 \text{ Å}$.

| Pfam family | $r^a$ | $k_B \tilde{T}_s$ $^a$ (kcal/mol) | $\tilde{T}_s$ (°K) | $T_m$ (°K) | $\hat{T}_x$ (°K) | $T^b$ (°K) | $\omega^c$ ($k_B$) | $\langle \Delta G_{\text{ND}} \rangle^d$ (kcal/mol) |
|-------------|-------|----------------------------------|-------------------|----------|-----------------|----------|---------------|-------------------|
| HTH_3       | -     | -                               | 116.3             | 343.7    | 155.1           | 298      | 0.8107        | -2.98             |
| Nitroreductase | -     | -                               | 194.0             | 337.0    | 214.2           | 298      | 0.8537        | -2.62             |
| SBP_bac_3   | -     | -                               | 184.8             | 336.1    | 207.0           | 298      | 0.8622        | -8.14             |
| SBP_bac_3   | -     | -                               | 272.7             | 336.1    | 277.7           | 298      | 0.6025        | -0.99             |
| OmpA        | -     | -                               | 86.9              | 320.0    | 126.8           | 298      | 0.9171        | -3.16             |
| DnaB        | -     | -                               | 104.0             | 312.8    | 139.7           | 298      | 1.1082        | -2.52             |
| LysR_substrate | -   | -                               | 248.8             | 338.0    | 258.0           | 298      | 0.6722        | -3.46             |
| LysR_substrate | -   | -                               | 236.2             | 338.0    | 247.7           | 298      | 0.6425        | -2.05             |
| Methyltransf_5 | -   | -                               | 57.3              | 375.0    | 107.8           | 298      | 1.0220        | -39.91            |
| Methyltransf_5 | -   | -                               | 82.3              | 375.0    | 131.7           | 298      | 1.0783        | -11.14            |
| SH3_1       | 0.86  | 0.1583                          | 102.3             | 344.0    | 143.8           | 295      | 0.9904        | -3.68             |
| ACBP        | 0.82  | 0.1169                          | 87.0              | 324.4    | 127.6           | 278      | 1.0774        | -6.51             |
| PDZ         | 0.93  | 0.2794                          | 140.7             | 312.9    | 168.5           | 298      | 1.0854        | -1.81             |
| Copper-bind | 0.83  | 0.1781                          | 89.8              | 359.3    | 135.8           | 298      | 0.9171        | -11.55            |

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

$^b$ Temperatures are set up for comparison to be equal to the experimental temperatures for $\Delta G_{\text{ND}}$ or to 298°K if unavailable; see Table S4 for the experimental data.

$^c$ Conformational entropy per residue, in $k_B$ units, in the denatured molten-globule state; see Eq. A18.

$^d$ Folding free energy in kcal/mol units; see Eq. A27.
Figure 1: Correlation between $\Delta \psi_N$ due to single nucleotide nonsynonymous substitutions and $\psi_N$ of homologous sequences in the PDZ domain family. The left and right figures correspond to the cutoff distance $r_{\text{cutoff}} \sim 8$ and 15.5 Å respectively. Each of the black plus or red cross marks corresponds to the mean or the standard deviation of $\Delta \psi_N$ due to all types of single nucleotide nonsynonymous substitutions over all sites in each of the homologous sequences of the PDZ domain family. Only 335 representatives of unique sequences with no deletion, which are at least 20% different from each other, are employed; the number of the representatives is almost equal to $M_{\text{eff}}$ in Table 1. The solid lines show the regression lines for the mean and the standard deviation of $\Delta \psi_N$. 
Figure 2: Regression of the experimental values (15) of folding free energy changes ($\Delta \Delta G_{ND}$) due to single amino acid substitutions on $\Delta \psi_N (\approx \Delta \Delta \psi_{ND})$ for the same types of substitutions in the PDZ domain. The left and right figures correspond to the cutoff distance $r_{\text{cutoff}} \sim 8$ Å and 15.5 Å respectively. The solid lines show the least-squares regression lines through the origin with the slopes, 0.279 kcal/mol for $r_{\text{cutoff}} \sim 8$ Å and 0.162 kcal/mol for $r_{\text{cutoff}} \sim 15.5$ Å, which are the estimates of $k_B T_s$. The reflective correlation coefficients for them are equal to 0.93 and 0.94, respectively. The free energies are in kcal/mol units.
Figure 3: Comparison of selection temperatures ($T_s$) estimated with different cutoff distances by the present method. The abscissa and ordinate correspond to the cases of $r_{	ext{cutoff}} \sim 8$ and 15.5Å respectively. The $T_s$ is in °K units.
Figure 4: Dependence of the average of $\Delta\psi_N$ due to single nucleotide nonsynonymous substitutions over homologous sequences on $-\delta\psi^2/L$ across protein families. Plus and open circle marks indicate the values for each protein family in the cases of $r_{\text{cutoff}} \sim 8$ and 15.5 Å, respectively. In the case of the cutoff distance 8 Å, the correlation coefficient is equal to 0.99, and the regression line is $\Delta\psi_N(\sigma_N^j, \sigma_i^N \rightarrow \sigma_i) = -1.75(-\delta\psi^2/L) - 0.67$. In the case of $r_{\text{cutoff}} \sim 16$ Å, the correlation coefficient is equal to 0.995, and the regression line is $\Delta\psi_N(\sigma_N^j, \sigma_i^N \rightarrow \sigma_i) = -1.82(-\delta\psi^2/L) - 0.81$. 
Figure 5: The sample average of folding free energy change, $\overline{\Delta G_{ND}} \equiv k_B T_s \Delta \overline{\Delta \psi_{ND}}$, is plotted against the ensemble average of folding free energy per residue, $\langle \Delta G_{ND} \rangle_{\sigma} / L = k_B T_s \langle \Delta \psi_{ND} \rangle_{\sigma} / L$, for each protein family. In the case of the cutoff distance $8 \text{ Å}$, the correlation coefficient is $r = -0.74$, and the regression line is $\Delta \Delta G_{ND}(\sigma_{j,i}^N, \sigma_i^N \rightarrow \sigma_i) = -2.90 \langle \Delta G_{ND} \rangle_{\sigma} / L + 1.00$. In the case of $r_{\text{cutoff}} \sim 16 \text{ Å}$, the correlation coefficient is $r = -0.60$, and the regression line is $\Delta \Delta G_{ND}(\sigma_{j,i}^N, \sigma_i^N \rightarrow \sigma_i) = -3.23 \langle \Delta G_{ND} \rangle_{\sigma} / L + 1.14$. The free energies are in kcal/mol units.
Figure 6: $\hat{T}_s/\hat{T}_g$ is plotted against $T_m/\hat{T}_g$ for each protein domain. A dotted curve corresponds to Eq. A30, $\hat{T}_s/\hat{T}_g = 2(T_m/\hat{T}_g)/(T_m/\hat{T}_g)^2 + 1)$. Plus and open circle marks indicate the values estimated with $r_{\text{cutoff}} \sim 8$ and 15.5 Å, respectively. The $T_s$ and $T_g$ must satisfy $T_s < T_g < T_m$ for proteins to be able to fold into unique native structures.
Figure 7: Folding free energies, $\langle \Delta G_{ND} \rangle_\sigma$, predicted by the present method are plotted against their experimental values, $\Delta G_{ND}(\sigma_N)$. The free energies are in kcal/mol units.
Supplementary material
for
"Selection originating from protein foldability:
I. A new method to estimate selection temperature"

Sanzo Miyazawa
sanzo.miyazawa@gmail.com

1 Methods

1.1 Knowledge of protein folding

A protein folding theory \(3\), which is based on a random energy model (REM), indicates that the equilibrium ensemble of amino acid sequences, \(\sigma = (\sigma_1, \cdots, \sigma_L)\) where \(\sigma_i\) is the type of amino acid at site \(i\) and \(L\) is sequence length, should be a canonical ensemble with a Boltzmann factor consisting of the folding free energy, \(\Delta G_{ND}(\sigma, T)\) and an effective temperature \(T_s\) of natural selection.

\[
P(\sigma) \propto p_{\text{mut}}(\sigma) \exp(-\Delta G_{ND}(\sigma, T)/k_B T_s) \quad (S1)
\]

where \(p_{\text{mut}}(f(\sigma))\) is the probability of a sequence (\(\sigma\)) randomly occurring in a mutational process and depends only on the amino acid frequencies \(f(\sigma)\), \(k_B\) 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. Selection temperature \(T_s\) quantifies how strong the folding constraints have been during evolution \(11\), 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, \(f(\sigma)\), in a sequence. 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, \(P_{\text{mut}}(\sigma) = \prod_i p_{\text{mut}}(\sigma_i)\), where \(p_{\text{mut}}(\sigma_i)\) is the probability density with mean \(\bar{f}(\sigma)\) and variance \(\delta E^2(\sigma)\), which depend only on the amino acid composition of the sequence.

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) \(3\), to be equal to the energy distribution of randomized sequences, which is then approximated by a Gaussian distribution, in the native conformation. That is, a 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.

\[
Z = \int \exp(-E/k_BT) n(E) dE \quad (S4)
\]

\[
n(E) = \exp(\omega L) N(\bar{E}(f(\sigma)), \delta E^2(\sigma)) \quad (S5)
\]

where \(\omega\) is the conformational entropy per residue in the compact denatured state, and \(N(\bar{E}(f(\sigma)), \delta E^2(\sigma))\) is the Gaussian probability density with mean \(\bar{E}\) and variance \(\delta E^2\), which depend only on the amino acid composition of the protein sequence. The free energy of the denatured state is approximated as follows.

\[
G_D(f(\sigma), T) \approx \bar{E}(f(\sigma)) - \frac{\delta E^2(\sigma)}{2k_B T} - k_B T \omega L \quad (S6)
\]

\[
= \bar{E}(f(\sigma)) - \delta E^2(\sigma) \frac{\theta(T/T_g)}{k_B T} \quad (S7)
\]

\[
\theta(T/T_g) \equiv \begin{cases} \frac{1}{2} + \frac{1}{2} \frac{T}{T_g} & \text{for } T > T_g \\ \frac{T}{T_g} & \text{for } T \leq T_g \end{cases} \quad (S8)
\]
where $E$ and $\delta E^2$ 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 (3).

$$\left. -\frac{\partial G_D}{\partial T}\right|_{T=T_g} = 0 \quad (S9)$$

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

$$\omega L = \frac{\delta E^2}{2(k_B T_g)^2} \quad (S10)$$

Thus, unless $T_g < T_m$, a protein will be trapped at local minima on a rugged free energy landscape before it can fold into a unique native structure.

### 1.2 Probability distribution of homologous sequences in sequence space

The probability distribution $P(\sigma)$ of sequences, $\sigma \equiv (\sigma_1, \ldots, \sigma_L)$ where $\sigma_i \in \{\text{amino acids, deletion}\}$, with maximum entropy in sequence space that satisfies a given amino acid frequency at each site and a given pairwise amino acid frequency at each site pair is a Boltzmann distribution (4, 5).

$$P(\sigma) \propto \exp(-\psi_N(\sigma)) \quad (S11)$$

$$\psi_N(\sigma) \equiv -(\sum_i h_i(\sigma_i) + \sum_{ij} J_{ij}(\sigma_i, \sigma_j)) \quad (S12)$$

where $h_i$ and $J_{ij}$ are one-body and two-body interactions and must satisfy the following constraints.

$$\sum_{\sigma, \sigma_i = a_k} P(\sigma) = P_i(a_k) \quad (S13)$$

$$\sum_{\sigma, \sigma_i = a_k, \sigma_j = a_l} P(\sigma) = P_{ij}(a_k, a_l) \quad (S14)$$

where $P_i(a_k)$ 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 $J_{ij}(a_k, a_l) = 0$. Interactions $h_i$ and $J_{ij}$ can be well estimated from a multiple sequence alignment (MSA) in the mean field approximation (4, 5), or by maximizing a pseudo-likelihood (6, 7). Here we must notice that $\psi_N(\sigma)$ has been estimated under the constraints on amino acid compositions at all sites, and therefore the amino acid composition of a whole sequence must be constant across sequences.

From Eqs. (S2) and (S11)

$$G_N(\sigma) = k_B T_s \left[ \psi_N(\sigma) + \text{function of f(\sigma)} \right] \quad (S15)$$

$$G_D(f(\sigma), T) = k_B T_s \left[ \psi_D(\sigma) + \text{function of f(\sigma)} \right] \quad (S16)$$

$$\Delta G_{ND}(\sigma, T) = k_B T_s \Delta \psi_{ND}(\sigma) \quad (S17)$$

$$\Delta \psi_{ND}(\sigma) \equiv \psi_N(\sigma) - \psi_D(\sigma) \quad (S18)$$

$$\psi_D(\sigma) \equiv \hat{\psi}(f(\sigma)) - \delta \hat{\psi}^2(f(\sigma)) \partial(T/T_s)T_s/T \quad (S19)$$

$$\omega = (T_s/T)^2 \delta \hat{\psi}^2/(2L) \quad (S20)$$

where $\hat{E} = k_B T_s \hat{\psi}$ and $\delta E^2 = (k_B T_s)^2 \delta \hat{\psi}^2$. The mean $\hat{\psi}$ and variance $\Delta \hat{\psi}^2$ are estimated as the mean and variance of $\psi_N$ over randomized sequences in the native structure (3).
1.3 The ensemble average of folding free energy, $\Delta G_{ND}(\sigma, T)$, over sequences

The ensemble average of $\Delta G_{ND}(\sigma, T)$ over sequences with Eq. [S1] is

$$
\langle \Delta G_{ND}(\sigma, T) \rangle_{\sigma} = \frac{1}{\sum_{\sigma} \Delta G_{ND}(\sigma) P_{\text{mut}}(\sigma) \exp(-\Delta G_{ND}(\sigma, T)/k_B T_s)} / \left[ \sum_{\sigma} P_{\text{mut}}(\sigma) \exp(-\Delta G_{ND}(\sigma, T)/k_B T_s) \right] (S21)
$$

$$
\approx \left[ \sum_{\sigma} \Delta G_{N}(\sigma) \exp(-\Delta G_{N}^{\text{mut}}(\sigma))/k_B T_s \right] / \left[ \sum_{\sigma} \exp(-\Delta G_{N}(\sigma)/k_B T_s) \right] - G_D(\bar{f}(\sigma_N), T) (S22)
$$

$$
= \langle G_N(\sigma) \rangle_{\sigma} - G_D(\bar{f}(\sigma_N), T) (S23)
$$

where $\sigma_N$ denotes a native sequence, and $\bar{f}(\sigma_N)$ denotes the average of $f(\sigma_N)$ over homologous sequences. In Eq. [S23], the sum over all sequences is approximated by the sum over sequences the amino acid composition of which is the same as that of the native sequences. The ensemble average of $G_N$ is also estimated in the Gaussian approximation (3).

$$
\langle G_N(\sigma) \rangle_{\sigma} = \int E \exp(-E/(k_B T_s)) n(E) dE / \int \exp(-E/(k_B T_s)) n(E) dE (S25)
$$

$$
= \hat{E}(\bar{f}(\sigma_N)) - \delta E^2(\bar{f}(\sigma_N))/k_B T_s (S26)
$$

$$
\langle \psi_N(\sigma) \rangle_{\sigma} = \frac{\sum_{\sigma_s} w_{\sigma_s} \psi_N(\sigma_N)}{\sum_{\sigma_s} w_{\sigma_s}} (S27)
$$

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

$$
\Delta G_{ND}(\sigma_N, T) = \langle \Delta G_{ND}(\sigma, T) \rangle_{\sigma} (S28)
$$

$$
= \delta E^2(\bar{f}(\sigma_N)) \left( \frac{1}{T_s} T_s/T - 1 \right) (S29)
$$

$$
\psi_N(\sigma_N) = \frac{\sum_{\sigma_s} w_{\sigma_s} \psi_N(\sigma_N)}{\sum_{\sigma_s} w_{\sigma_s}} (S30)
$$

where the overline denotes sample average with a sample weight $w_{\sigma_s}$ 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 $T_m$; $\langle \Delta G_{ND}(\sigma_N, T_m) \rangle_{\sigma} = 0$. Thus, the following relationship must be satisfied (3).

$$
\vartheta(T_m/T_g) \frac{T_s}{T_m} = \frac{T_s}{2T_m} (1 + \frac{T_s^2}{T_g}) = 1 \quad \text{with } T_s \leq T_g \leq T_m (S33)
$$

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

The mean $\bar{\psi}(f(\sigma))$ and the variance $\delta \psi^2(f(\sigma))$ in the Gaussian approximation for the distribution of conformational energies at the denatured state are estimated as the mean and variance of $\psi_N$ of random sequences in the native conformation (3).

$$
\bar{\psi}(f(\sigma)) = -\sum_i [h_i(::) + \sum_{i \neq j} J_{ij}(::::)] (S44)
$$

where $h_i(::)$ and $J_{ij}(::::)$ are the means of one-body and two-body interactions in random sequences.

$$
h_i(::) = \sum_k h_i(a_k) f_{a_k}(\sigma) (S35)
$$

$$
J_{ij}(::::) = \sum_k \sum_l J_{ij}(a_k, a_l) f_{a_k}(\sigma) f_{a_l}(\sigma) (S36)
$$
where \( f_{\alpha}(\sigma) \) is the composition of amino acid \( a_k \) in the sequence \( \sigma \).

\[
f_{\alpha}(\sigma) = \frac{1}{L} \sum_{i=1}^{L} \delta_{\sigma_{ai}} \tag{S37}
\]

where \( \delta_{\sigma_{ai}} \) is the Kronecker delta. The variance, \( \delta \phi^2(f(\sigma)) \), is

\[
\delta \phi^2(f(\sigma)) = \sum_k f_{\alpha}(\sigma) \sum_i \left[ \delta h_i(a_k) + \sum_{j \neq i} (2 \delta h_i(a_k) \delta J_{ij}(a_k, ::) + \sum_{m \neq i, j} \delta J_{ij}(a_k, ::) \delta J_{im}(a_k, ::) + \frac{1}{2} \sum_j \delta J_{ij}(a_k, a_l) f_{\alpha}(\sigma) \right] \tag{S38}
\]

where

\[
\begin{align*}
\delta h_i(a_k) & \equiv h_i(a_k) - h_i(:, :) \\ 
\delta J_{ij}(a_k, ::) & \equiv J_{ij}(a_k, :) - J_{ij}(::, :) \\ 
\delta J_{ij}(a_k, a_l) & \equiv J_{ij}(a_k, a_l) - J_{ij}(::, :) 
\end{align*} \tag{S40-S42}
\]

### 1.5 Estimation of one-body (\( h \)) and pairwise (\( J \)) interactions

The estimates of \( h \) and \( J \) are noisy, and \( J_{ij} \) may take significant values even for site pairs that are distantly located in the three dimensional structure of protein. Therefore, according to Morcos et al. (11), the estimate of \( J \) is modified as follows.

\[
J_{ij}(a_k, a_l) = J_{ij}^0(a_k, a_l) H(r_{\text{cutoff}} - r_{ij}) \tag{S43}
\]

where \( J^0 \) is the statistical estimate of \( J \) in a certain gauge, \( H \) is the Heaviside step function, and \( r_{ij} \) is the distance between the centers of amino acid side chains in protein structure, and \( r_{\text{cutoff}} \) is a distance threshold for residue pairwise interactions. Maximum interaction ranges employed for pairwise interactions are \( r_{\text{cutoff}} \sim 8 \) and \( 15.5 \) Å, which correspond to the first and second interaction shells between residues, respectively. Here it should be noticed that the total interaction \( \psi_N(\sigma) \) defined by Eq. 12 does not depend on any gauge unless the interaction range for pairwise interactions is limited, but a gauge conversion in which interconversions between \( h \) and \( J \) occur may change the estimate of \( \psi_N(\sigma) \) in the present scheme of Eq. 43 in which pairwise interactions are cut off at a certain distance. Thus, a natural gauge must be used before calculating \( J \).

For example, let us think about the Ising gauge (7), in which \( h' \) and \( J' \) can be calculated from any gauge through the following conversions.

\[
\begin{align*}
J_{ij}^0(a_k, a_l) &= J_{ij}^0(a_k, a_l) - J_{ij}^0(a_k, :) - J_{ij}^0(:, a_l) + J_{ij}^0(:, :) \\
h^0(a_k) &= h^0(a_k) - h^0( :) + \sum_{j \neq i} (J_{ij}^0(a_k, :) - J_{ij}^0(:, :)) 
\end{align*} \tag{S44-S45}
\]

where

\[
\begin{align*}
h( :) &= \frac{1}{q} \sum_{k=1}^{q} h_k(a_k) \\
J_{ij}(::, :) &= \frac{1}{q^2} \sum_{k=1}^{q} \sum_{l=1}^{q} J_{ij}(a_k, a_l)
\end{align*} \tag{S46-S47}
\]

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

In the DCA (4,5), the interaction terms are estimated in the mean field approximation as follows.

\[
\begin{align*}
J_{ij}^0(a_k, a_l) &= -(C^{-1})_{ij}(a_k, a_l) \\
J_{ij}^0(a_q, a_l) &= J_{ij}^0(a_k, a_q) = J_{ij}^0(a_q, a_q) = 0
\end{align*} \tag{S48-S49}
\]
where \( i \neq j \) and \( 1 \leq k, l \leq q - 1 \), and the covariance matrix \( C \) is defined as

\[
C_{ij}(a_k, a_l) = P_{ij}(a_k, a_l) - P_i(a_k)P_j(a_l)
\]  

(S50)

Here, one \((a_q)\) of the amino acid types including deletion is used as the reference state; \( J^q \) denotes the \( J \) in this gauge. According to Morcos et al. (4), the probability \( P_i(a_k) \) of amino acid \( a_k \) at site \( i \) and the joint probabilities \( P_{ij}(a_k, a_l) \) of amino acids, \( a_k \) at site \( i \) and \( a_l \) at site \( j \), are evaluated by

\[
P_i(a_k) = (1 - p_c)f_i(a_k) + p_c \frac{1}{q}
\]  

(S51)

\[
P_{ij}(a_k, a_l) = (1 - p_c)f_{ij}(a_k, a_l) + p_c \frac{1}{q^2} \quad \text{for} \quad i \neq j
\]  

(S52)

\[
P_{ii}(a_k, a_l) = P_i(a_k)\delta_{a_k a_l}
\]  

(S53)

where \( 0 \leq p_c \leq 1 \) is the ratio of pseudocount, and \( f_i(a_k) \) is the frequency of amino acid \( a_k \) at site \( i \) and \( f_{ij}(a_k, a_l) \) is the frequency of the site pair, \( a_k \) at \( i \) and \( a_l \) at \( j \), in an alignment; \( f_{ii}(a_k, a_l) \) is defined as \( f_{ii}(a_k, a_l) = f_i(a_k)\delta_{a_k a_l} \). Then, according to Morcos et al. (4), the one body interactions \( h_i(a_k) \) are estimated in the isolated two-state model, that is,

\[
P_i(a_k) \propto \exp \left[ h^q_i(a_k) + J^q_{ij}(a_k, a_l) + h^q_{ji}(a_l) \right]
\]  

(S54)

\[
h^q_i(a_k) = \frac{1}{L-1} \sum_{j \neq i} h^q_{ij}(a_k)
\]  

(S55)

These \( h^q \) and \( J^q \) in the \( q \) gauge are converted to a new gauge, which is called a simple gauge here,

\[
h^s_i(a_k) = h^q_i(a_k) - h^q_{i:} \]  

(S56)

\[
J^s_{ij}(a_k, a_l) = J^q_{ij}(a_k, a_l) - J^q_{i:}(a_l) \]  

(S57)

In this gauge, the reference state is the average state over amino acids including deletion, instead of a specific amino acid \((a_q)\) in the \( q \) gauge. The estimate of \( J \) in this gauge is used to calculate \( \hat{J} \) in Eq. S43.
2 Materials

2.1 Sequence data

We study the single domains of 8 Pfam families and both the single domains and multi-domains from 3 Pfam families. In Table S1, 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” (4, 5). To estimate the sample \( \langle \psi_N \rangle \) and ensemble \( \langle \langle \psi_N \rangle \rangle \) averages of the total interaction, \( M \) unique sequences with no deletion are used. In order to reduce phylogenetic biases in the set of homologous sequences, we employ a sample weight \( w_{\sigma N} \) 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 deletion, which are at least 20% different from each other, are used to calculate the changes of the total interaction \( \Delta \psi_N \) due to single nucleotide nonsynonymous substitutions; the number of the representatives is almost equal to \( M \) in Table S1.

3 Empirical rules found in the analysis of \( \Delta \psi_N \)

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

\[
\Delta \psi_N \approx \alpha_{\psi N} \psi_N - \bar{\psi}_N + \Delta \bar{\psi}_N \quad \text{with } \alpha_{\psi N} < 0 \quad (S58)
\]

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

\[
\text{Var}(\Delta \psi_N) \approx \text{constant across homologous sequences in every protein family} \quad (S59)
\]

Because

\[
\text{Var}(\Delta G_N) = \text{function that must not explicitly depend on } k_B T_s \text{ but } G_N \quad (S61)
\]

the following important relationship has been derived and used to estimate the relative value of \( T_s \) against \( T_{s,PDZ} \) of the PDZ family.

\[
\text{Var}(\Delta G_N) = (k_B T_s)^2 \text{Var}(\Delta \psi_N) \approx \text{constant} \quad (S62)
\]

where \( \text{Var}(\Delta G_N) \) and \( \text{Var}(\Delta \psi_N) \) are the variances of \( \Delta G_N \) and \( \Delta \psi_N \) over all single nucleotide nonsynonymous substitutions at all sites, respectively. These relationships, Eqs. S58 and S62, are also shown in Figs. S3 and S4, and the regression coefficients \( (\alpha_{\psi N}) \) and correlation coefficients \( (r_{\psi N}) \) are listed in Tables S2 and S5. With estimated \( T_s \) and experimental melting temperature \( T_m \), glass transition temperature \( T_g \) and folding free energy \( \Delta G_N \) have been estimated for each protein family on the basis of the REM. The estimates of \( T_s \) and \( T_g \) are all within a reasonable range, and the estimated values of \( \Delta G_N \) agree well with their experimental values for 5 protein families, justifying the estimates of \( T_s \).
Table S1: Protein families, and structures studied.

| Pfam family | UniProt ID | N<sup>a</sup> | N<sub>eff</sub><sup>b</sup> | M<sup>c</sup> | M<sub>eff</sub><sup>c</sup> | L<sup>d</sup> | PDB ID  |
|-------------|------------|----------------|----------------|---------|----------------|---------|---------|
| HTH<sub>3</sub> | 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 |
| SBP_bac<sub>3</sub> | GLNH_ECOLI/27-244 | 9874(9972) | 7374.96 | 140 | 99.70 | 218 | 1WDN-A:5-222 |
| SBP_bac<sub>3</sub> | GLNH_ECOLI/111-204 | 9712(9898) | 7442.85 | 829 | 689.64 | 94 | 1WDN-A:89-182 |
| OmpA | PDL_ECOLI/73-167 | 6035(6070) | 4920.44 | 2207 | 1761.24 | 95 | 1OAP-A:52-146 |
| DnaB | DNAP_ECOLI/31-128 | 1929(1957) | 1284.94 | 1187 | 697.30 | 98 | 1JWE-A:30-127 |
| LysR<sub>substrate</sub> | BENM_ACIAD/90-280 | 25138(25226) | 20707.06 | 85(1) | 67.00 | 191 | 2F6G-A/B:90-280 |
| LysR<sub>substrate</sub> | BENM_ACIAD/163-265 | 25032(25164) | 21144.74 | 121(1) | 99.27 | 103 | 2F6G-A/B:163-265 |
| Methyltransf<sub>5</sub> | RSMH_THEMA/8-292 | 1942(1953) | 1286.67 | 578(2) | 357.97 | 285 | 1N2X-A:8-292 |
| Methyltransf<sub>5</sub> | RSMH_THEMA/137-216 | 1877(1911) | 1033.35 | 975(2) | 465.53 | 80 | 1N2X-A:137-216 |
| SH3<sub>1</sub> | 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_PSEA:24-148 | 1136(1169) | 841.56 | 67(1) | 45.23 | 125 | 5AZU-A/B/C/D:4-128 |

<sup>a</sup> The number of unique sequences and the total number of sequences in parentheses; the full alignments in the Pfam (<sup>14</sup>) are used.

<sup>b</sup> The effective number of sequences.

<sup>c</sup> A sample weight (w<sub>σ</sub>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.

<sup>d</sup> The number of unique sequences that include no deletion unless specified. The number in parentheses indicates the maximum number of deletions allowed.

<sup>e</sup> The effective number of unique sequences that include no deletion or at most the specified number of deletions.

<sup>f</sup> The number of residues.

<sup>g</sup> These proteins consist of two domains, and other ones are single domains.
Table S2: Parameter values for $r_{cutoff} \sim 8 \text{ Å}$ employed for each protein family, and the averages of the total interactions $(\Delta \psi_N)$ over all homologous sequences and of the means and variances of interaction changes ($\Delta \psi_N$ and $\text{Var}(\Delta \psi_N)$) due to single nucleotide nonsynonymous mutations at all sites over all homologous sequences in each protein family.

| Pfam family  | $L$  | $p_e$ | $n_e$ | $r_{cutoff}$ (Å) | $\bar{\psi}/L$ | $\Delta \psi^2 / L$ | $\bar{\psi} / L$ | $\Delta \psi_N$ | $\text{Var}(\Delta \psi_N)$ | $r_{\phi_N}$ | $\alpha_{\phi_N}$ | $r_{\phi_N}$ for $\Delta \psi_N$ | $\alpha_{\phi_N}$ for $\text{Var}(\Delta \psi_N)^{1/2}$ |
|--------------|------|------|------|------------------|----------------|-------------------|----------------|----------------|----------------|----------------|----------------|--------------------------|--------------------------|
| HTH3         | 53   | 0.175| 7.434| 8.22             | -0.2039        | 2.8837            | -3.0712        | 4.0412         | 30.8500        | -0.94          | -1.3671        | -0.23                     | -0.1741                   |
| Nitroreductase| 73   | 0.24 | 6.411| 8.39             | -0.1153        | 2.0825            | -2.1975        | 2.9168         | 11.0879        | -0.94          | -1.3915        | -0.59                     | -0.5270                   |
| SBP_bac3     | 218  | 0.25 | 9.230| 8.10             | -0.1000        | 2.1624            | -2.2618        | 3.1534         | 12.2140        | -0.98          | -1.5267        | -0.84                     | -0.8111                   |
| SBP_bac3     | 94   | 0.37 | 8.000| 7.90             | -0.1634        | 1.2495            | -1.4054        | 1.8075         | 5.6101         | -0.96          | -1.4191        | -0.65                     | -0.5171                   |
| OmpA         | 95   | 0.169| 8.000| 8.20             | -0.2457        | 3.9093            | -4.1542        | 6.0657         | 55.3105        | -0.96          | -1.5882        | -0.47                     | -0.4471                   |
| DnaB         | 98   | 0.235| 9.653| 8.17             | -0.2294        | 3.9976            | -4.2291        | 6.1006         | 38.5736        | -0.96          | -1.4654        | -0.52                     | -0.4385                   |
| LysR_substrate| 191  | 0.23 | 7.843| 7.87             | -0.2248        | 1.4450            | -1.6712        | 1.9999         | 6.7385         | -0.97          | -1.3646        | -0.52                     | -0.5814                   |
| LysR_substrate| 103  | 0.265| 8.835| 8.25             | -0.2240        | 1.4132            | -1.6372        | 2.0309         | 7.4763         | -0.99          | -1.4629        | -0.78                     | -0.6311                   |
| Methyltransfer_5 | 285  | 0.13 | 7.993| 7.78             | -0.1462        | 7.2435            | -7.3887        | 12.2738        | 127.1400       | -0.99          | -1.7945        | -0.08                     | -0.0461                   |
| Methyltransfer_5 | 80   | 0.18 | 6.775| 7.85             | -0.1763        | 5.5162            | -5.6896        | 8.7386         | 61.5293        | -0.96          | -1.5601        | 0.06                      | 0.0612                    |
| SH3_1        | 48   | 0.14 | 6.417| 8.01             | -0.1348        | 3.9109            | -4.0434        | 5.4273         | 39.8445        | -0.94          | -1.4457        | -0.22                     | -0.1995                   |
| ACBP         | 80   | 0.22 | 9.175| 8.24             | -0.0525        | 4.6411            | -4.7084        | 7.8272         | 55.1762        | -0.97          | -1.6189        | -0.30                     | -0.1986                   |
| PDZ          | 81   | 0.205| 9.061| 8.16             | -0.2398        | 3.1140            | -3.3572        | 4.3897         | 21.0789        | -0.96          | -1.5014        | -0.43                     | -0.3293                   |
| Copper-bind  | 125  | 0.23 | 9.232| 8.26             | -0.0838        | 4.1946            | -4.2657        | 7.2514         | 51.7793        | -0.98          | -1.9254        | -0.20                     | -0.1475                   |

\(^a\) The average number of contact residues per site within the cutoff distance; the center of side chain is used to represent a residue.

\(^b\) $M$ unique sequences without deletions are used with a sample weight $(w_{seq})$ for each sequence; $n_{seq}$ 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 S1.

\(^c\) Representatives of unique sequences without deletions, which are at least 20% different from each other, are used; the number of the representatives used is almost equal to $M_{eff}$. For HTH3, 4431 sequences are used; abnormal sequences are removed.

\(^d\) The correlation and regression coefficients of $\Delta \psi_N$ on $\bar{\psi} / L$; see Eq. S58.

\(^e\) The correlation and regression coefficients of ($\text{Var}(\Delta \psi_N)$)$^{1/2}$ on $\bar{\psi} / L$. 
Table S3: **Thermodynamic quantities estimated with $r_{\text{cutoff}} \sim 8\ \text{Å}$.**

| Pfam family       | $r^a$ | $k_B\hat{T}_x^a$ (kcal/mol) | $\hat{T}_x$ (°K) | $T_m$ (°K) | $T^b$ (°K) | $\hat{\omega}^c$ ($k_B$) | $(\Delta G_{ND})^d$ (kcal/mol) |
|-------------------|-------|-----------------------------|------------------|------------|------------|-------------------------|---------------------------------|
| HTH\_3            | -     | -                           | 116.3            | 343.7      | 155.1      | 298                     | 0.8107                          | -2.98                           |
| Nitroreductase     | -     | -                           | 194.0            | 337.0      | 214.2      | 298                     | 0.8537                          | -2.62                           |
| SBP\_bac\_3       | -     | -                           | 184.8            | 336.1      | 207.0      | 298                     | 0.8622                          | -8.14                           |
| SBP\_bac\_3       | -     | -                           | 272.7            | 336.1      | 277.7      | 298                     | 0.6025                          | -0.99                           |
| OmpA               | -     | -                           | 86.9             | 320.0      | 126.8      | 298                     | 0.9171                          | -3.16                           |
| DnaB               | -     | -                           | 104.0            | 312.8      | 139.7      | 298                     | 1.1082                          | -2.52                           |
| LysR\_substrate    | -     | -                           | 248.8            | 338.0      | 258.0      | 298                     | 0.6722                          | -3.46                           |
| LysR\_substrate    | -     | -                           | 236.2            | 338.0      | 247.7      | 298                     | 0.6425                          | -2.05                           |
| Methyltransfer\_5 | -     | -                           | 57.3             | 375.0      | 107.8      | 298                     | 1.0220                          | -39.91                          |
| Methyltransfer\_5 | -     | -                           | 82.3             | 375.0      | 131.7      | 298                     | 1.0783                          | -11.14                          |
| SH3\_1            | 0.86  | 0.1583                      | 102.3            | 344.0      | 143.8      | 295                     | 0.9904                          | -3.68                           |
| ACBP               | 0.82  | 0.1169                      | 87.0             | 324.4      | 127.6      | 278                     | 1.0774                          | -6.51                           |
| PDZ                | 0.93  | 0.2794                      | 140.7            | 312.9      | 168.5      | 298                     | 1.0854                          | -1.81                           |
| Copper-bind        | 0.83  | 0.1781                      | 89.8             | 359.3      | 135.8      | 298                     | 0.9171                          | -11.55                          |

- $^a$ Reflective correlation ($r$) and regression ($k_B\hat{T}_x$) coefficients for least-squares regression lines of experimental $\Delta \Delta G_{ND}$ on $\Delta \psi_N$ through the origin.
- $^b$ Temperatures are set up for comparison to be equal to the experimental temperatures for $\Delta G_{ND}$ or to 298°K if unavailable; see Table S4 for the experimental data.
- $^c$ Conformational entropy per residue, in $k_B$ units, in the denatured molten-globule state; see Eq S20.
- $^d$ Folding free energy in kcal/mol units; see Eq S30.
Table S4: Experimental data used.

| Pfam family          | experimental values | ref. for $T_m$ | ref. for $\Delta G_{ND}$ and $\Delta \Delta G_{ND}$ |
|----------------------|---------------------|----------------|-----------------------------------------------------|
|                      | $T_m$ (°K) | $T$ (°K) | $\Delta G_{ND}$ (kcal/mol) |                                           |
| HTH_3                | 343.7     | 298      | $-5.33 \pm 0.36$ | (20) (22)                                      |
| Nitroreductase       | 337.0     | -        | -                  | (23)                                           |
| SBP_bac_3            | 336.1     | -        | -                  | (24)                                           |
| OmpA                 | 320.0     | -        | -                  | (25)                                           |
| DnaB                 | 312.8     | -        | -                  | (26)                                           |
| LysR_substrate       | 338.0     | -        | -                  | (27)                                           |
| Methyltransfer_5     | 375.0     | -        | -                  | (28, 29)                                       |
| SH3_1                | 344.0     | 295      | $-3.70$            | (30)                                           |
| ACBP                 | 324.4     | 278      | $-8.08 \pm 0.08$   | (32)                                           |
| PDZ                  | 312.9     | 298      | $-2.90$            | (34)                                           |
| Copper-bind          | 359.3     | 298      | $-12.90 \pm 0.36$  | (36)                                           |

S-10
### Table S5: Parameter values for \( \tau_{\text{cutoff}} \sim 15.5 \text{ Å} \) employed for each protein family, and the averages of the total interactions (\( \psi_N \)) over all homologous sequences and of the means and the variances of interaction changes (\( \overline{\Delta \psi}_N \) and \( \overline{\text{Var}(\Delta \psi_N)} \)) due to single nucleotide nonsynonymous mutations at all sites over all homologous sequences in each protein family.

| Pfam family | \( L \) | \( p_x \) | \( n_x \) | \( \tau_{\text{cutoff}} \) (Å) | \( \overline{\psi}/L \) | \( \overline{\Delta \psi}/L \) | \( \overline{\text{Var}(\Delta \psi_N)} \) | \( r_{\bar{\psi}} \) | \( \sigma_{r_{\bar{\psi}}} \) | \( r_{\bar{\psi}} \) | \( \sigma_{r_{\bar{\psi}}} \) |
|-------------|-------|---------|-------|------------------|-----------------|-----------------|-----------------|----------------|----------------|----------------|----------------|
| HTH\textsubscript{3}  | 53    | 0.24    | 32.30 | 15.46            | -0.2591         | 4.0640          | 6.3128          | 48.5075        | -0.93           | -1.4723         | -0.44           | -0.2813         |
| Nitroreductase | 73    | 0.32    | 25.51 | 15.62            | -0.1452         | 3.3359          | -3.4815         | 5.2466          | -0.94           | -1.5730         | -0.68           | -0.5961         |
| SBP\textsubscript{bac}\textsubscript{3} | 218   | 0.35    | 55.48 | 15.90            | -0.0669         | 3.4004          | -3.4674         | 5.5888          | 24.5240         | -0.97           | -1.6605         | -0.84           | -0.9034         |
| SBP\textsubscript{bac}\textsubscript{3} | 94    | 0.45    | 42.81 | 15.45            | -0.1628         | 3.2308          | -2.4831         | 3.8780          | 14.6360         | -0.97           | -1.6840         | -0.77           | -0.6742         |
| OmpA         | 95    | 0.235   | 35.58 | 15.69            | -0.2552         | 5.8175          | -6.0757         | 9.5326          | 125.0540        | -0.95           | -1.6945         | -0.44           | -0.4231         |
| DnaB         | 98    | 0.35    | 46.65 | 15.57            | -0.2351         | 6.1890          | -6.4167         | 10.3965         | 67.5994         | -0.90           | -1.5787         | -0.33           | -0.3279         |
| LysR\textsubscript{substrate} | 191   | 0.33    | 44.16 | 15.13            | -0.2668         | 2.3641          | -2.6309         | 3.7028          | 14.7899         | -0.96           | -1.5508         | -0.53           | -0.4839         |
| LysR\textsubscript{substrate} | 103   | 0.37    | 44.06 | 15.60            | -0.2834         | 2.4271          | -2.7111         | 3.8248          | 17.9478         | -0.99           | -1.5685         | -0.84           | -0.6683         |
| Methyltransf\textsubscript{S} | 285   | 0.175   | 53.52 | 15.53            | -0.1687         | 12.8982         | -13.0658        | 23.2012         | 369.5520        | -0.98           | -1.7992         | -0.14           | -0.1249         |
| Methyltransf\textsubscript{S} | 80    | 0.24    | 37.02 | 15.11            | -0.1632         | 9.9944          | -10.1576        | 16.9985         | 194.4250        | -0.91           | -1.6482         | -0.38           | -0.3845         |
| SH3\textsubscript{1}    | 48    | 0.165   | 28.46 | 15.76            | -0.1350         | 7.6161          | -7.7523         | 11.5955         | 187.9570        | -0.92           | -1.6770         | -0.29           | -0.2693         |
| ACBP         | 80    | 0.28    | 36.27 | 15.34            | -0.0235         | 7.4707          | -7.4947         | 13.1960         | 102.0870        | -0.92           | -1.6896         | 0.12            | 0.1203          |
| PDZ          | 81    | 0.33    | 40.82 | 15.77            | -0.3022         | 5.2295          | -5.5313         | 8.0540          | 62.2153         | -0.97           | -1.7009         | -0.35           | -0.2408         |
| Copper-bind  | 125   | 0.295   | 44.34 | 15.43            | -0.0934         | 8.5012          | -8.5928         | 15.4936         | 102.3600        | -0.97           | -1.7939         | -0.24           | -0.2361         |

\(^a\) The average number of contact residues per site within the cutoff distance; the center of side chain is used to represent a residue.

\(^b\) \( M \) unique sequences without deletions are used with a sample weight \( n_{\omega_x} \) for each sequence; \( n_{\omega_x} \) 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 \( M_{\text{eff}} \) of the sequences are listed for each protein family in Table S1.

\(^c\) Representatives of unique sequences without deletions, which are at least 20% different from each other, are used; the number of the representatives used is almost equal to \( M_{\text{eff}} \). For HTH\textsubscript{3}, 4461 sequences are used; abnormal sequences are removed.

\(^d\) The correlation and regression coefficients of \( \overline{\Delta \psi}_N \) on \( \psi_N/L \); see Eq S58.

\(^e\) The correlation and regression coefficients of \( \text{Var}(\Delta \psi_N) \) on \( \psi_N/L \).
Table S6: Thermodynamic quantities estimated with $r_{\text{cutoff}} \sim 15.5$ Å.

| Pfam family     | $r^a$ | $k_B \hat{T}_r$ | $T_r$ | $T_m$ | $\hat{T}_m$ | $T^b$ | $\hat{T}_m$ | $\langle \Delta \tilde{G}_{ND} \rangle^d$ (kcal/mol) |
|-----------------|-------|-----------------|-------|-------|-------------|-------|-------------|--------------------------------------------------|
| HTH_3           | -     | -               | 92.3  | 343.7 | 135.4       | 298   | 0.9452      | -3.73                                            |
| Nitroreductase  | -     | -               | 131.3 | 337.0 | 165.8       | 298   | 1.0466      | -4.30                                            |
| SBP_bac_3       | -     | -               | 129.9 | 336.1 | 164.5       | 298   | 1.0601      | -12.76                                           |
| SBP_bac_3       | -     | -               | 168.1 | 336.1 | 194.1       | 298   | 0.8705      | -3.86                                            |
| OmpA            | -     | -               | 57.5  | 320.0 | 100.5       | 298   | 0.9516      | -3.53                                            |
| DnaB            | -     | 78.2            |       | 312.8 | 118.2       | 298   | 1.3542      | -3.32                                            |
| LysR_substrate  | -     | 167.2           |       | 338.0 | 193.8       | 298   | 0.8803      | -8.38                                            |
| LysR_substrate  | -     | 151.8           |       | 338.0 | 118.1       | 298   | 0.8453      | -4.64                                            |
| Methyltransf_5  | -     | 33.5            |       | 375.0 | 81.0        | 298   | 1.0994      | -45.10                                           |
| Methyltransf_5  | -     | 46.1            |       | 375.0 | 96.0        | 298   | 1.1537      | -12.95                                           |
| SH3_1           | 0.84  | 0.0821          | 46.9  | 344.0 | 93.1        | 295   | 0.9678      | -4.13                                            |
| ACBP            | 0.82  | 0.0689          | 63.7  | 324.4 | 107.0       | 278   | 1.3220      | -8.51                                            |
| PDZ             | 0.94  | 0.1619          | 81.5  | 312.9 | 121.1       | 298   | 1.1852      | -2.39                                            |
| Copper-bind     | 0.89  | 0.0997          | 63.6  | 359.3 | 111.9       | 298   | 1.3710      | -18.42                                           |

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

$b$ Temperatures are set up for comparison to be equal to the experimental temperatures for $\Delta G_{ND}$ or to 298°K if unavailable; see Table S4 for the experimental data.

d Conformational entropy per residue, in $k_B$ units, in the denatured molten-globule state; see Eq. S20.

d Folding free energy in kcal/mol units; see Eq. S30.
Figure S1: Dependences of the sample ($\bar{\psi}_N/L$) and ensemble ($\langle\psi_N\rangle_{\sigma}/L$) averages of the total interaction per residue on the cutoff distance for pairwise interactions in the PDZ domain. The ratios of pseudocount $p_c = 0.205$ and 0.33 are employed here for the cutoff distance $r_{\text{cutoff}} \sim 8$ and 15.5Å respectively. The solid and dotted lines indicate the sample and ensemble averages, respectively, and black plus and blue cross marks show those for the simple and Ising gauges, respectively.
Figure S2: Dependences of the reflective correlation and regression coefficients between the experimental $\Delta\Delta G_{ND}$ (15) and $\Delta\psi_N$ due to single amino acid substitutions on the cutoff distance for pairwise interactions in the PDZ domain. The left and right figures are for the ratios of pseudocount, $p_c = 0.205$ and 0.33, respectively. The solid and dotted lines show the reflective correlation and regression coefficients for the least-squares regression line through the origin, respectively. The sample ($\overline{\psi_N}/L$) and ensemble ($\langle\psi_N\rangle_r/L$) averages of the total interaction agree with each other at the cutoff distance $r_{\text{cutoff}} \sim 8$ Å for $p_c = 0.205$ and $r_{\text{cutoff}} \sim 15.5$ Å for $p_c = 0.33$, where the reflective correlation coefficients attain to the maximum.
Figure S3: Correlation between $\Delta \psi_N$ due to single nucleotide nonsynonymous substitutions and $\psi_N$ of homologous sequences in the PDZ domain family. The left and right figures correspond to the cutoff distance $r_{\text{cutoff}} \sim 8$ and 15.5 Å respectively. Each of the black plus or red cross marks corresponds to the mean or the standard deviation of $\Delta \psi_N$ due to all types of single nucleotide nonsynonymous substitutions over all sites in each of the homologous sequences of the PDZ domain family. Only 335 representatives of unique sequences with no deletion, which are at least 20% different from each other, are employed; the number of the representatives is almost equal to $M_{\text{eff}}$ in Table S1. The solid lines show the regression lines for the mean and the standard deviation of $\Delta \psi_N$. 
Figure S4: Correlation between $\Delta \psi_N$ due to single nucleotide nonsynonymous substitutions and $\psi_N$ of homologous sequences in the SBP$_{bac\_3}$ family of the domain, 1WDN-A:5-222. The left and right figures correspond to the cutoff distance $r_{\text{cutoff}} \sim 8$ and 15.5 Å respectively. Each of the black plus or red cross marks corresponds to the mean or the standard deviation of $\Delta \psi_N$ due to all types of single nucleotide nonsynonymous substitutions over all sites in each of the homologous sequences in the SBP$_{bac\_3}$ family of the domain, 1WDN-A:5-222. Only 100 representatives of unique sequences with no deletion, which are at least 20% different from each other, are employed; the number of the representatives is almost equal to $M_{\text{eff}}$ in Table S1. The solid lines show the regression lines for the mean and the standard deviation of $\Delta \psi_N$. 

(a) $r_{\text{cutoff}} \sim 8$ Å

(b) $r_{\text{cutoff}} \sim 15.5$ Å
Figure S5: Regression of the experimental values (15) of folding free energy changes ($\Delta \Delta G_{ND}$) due to single amino acid substitutions on $\Delta \psi_N (\approx \Delta \Delta \psi_{ND})$ for the same types of substitutions in the PDZ domain. The left and right figures correspond to the cutoff distance $r_{\text{cutoff}} \sim 8$ Å and 15.5 Å respectively. The solid lines show the least-squares regression lines through the origin with the slopes, 0.279 kcal/mol for $r_{\text{cutoff}} \sim 8$ Å and 0.162 kcal/mol for $r_{\text{cutoff}} \sim 15.5$ Å, which are the estimates of $k_B T_s$. The reflective correlation coefficients for them are equal to 0.93 and 0.94, respectively. The free energies are in kcal/mol units.
Figure S6: Comparison of selection temperatures ($T_s$) estimated with different cutoff distances by the present method. The abscissa and ordinate correspond to the cases of $r_{\text{cutoff}} \sim 8$ and 15.5 Å respectively. The $T_s$ is in °K units.
Figure S7: Selection temperatures ($T_s$) estimated by the present method are plotted against those estimated by Morcos et al. [11]. Plus and open circle marks correspond to the cases of $r_{\text{cutoff}} \sim 8$ Å and $r_{\text{cutoff}} \sim 15.5$ Å, respectively.
Figure S8: Comparison of \( \alpha_{\psi_N} \), which is the regression coefficient of \( \Delta \psi_N \) on \( \psi_N/L \), with \( \Delta \psi_N/(-\delta \psi^2/L) \) for each protein family. Plus and open circle marks correspond to the cases of \( r_{\text{cutoff}} \sim 8 \) and 15.5Å, respectively.
Figure S9: Dependence of the average of $\Delta \psi_N$ due to single nucleotide nonsynonymous substitutions over homologous sequences on $-\delta \psi^2/L$ across protein families. Plus and open circle marks indicate the values for each protein family in the cases of $r_{\text{cutoff}} \sim 8$ and 15.5 Å, respectively. In the case of the cutoff distance 8 Å, the correlation coefficient is equal to 0.99, and the regression line is $\Delta \psi_N(\sigma^N_{jN}, \sigma^N_i \rightarrow \sigma_i) = -1.75(-\delta \psi^2/L) + 0.67$. In the case of $r_{\text{cutoff}} \sim 16$ Å, the correlation coefficient is equal to 0.995, and the regression line is $\Delta \psi_N(\sigma^N_{jN}, \sigma^N_i \rightarrow \sigma_i) = -1.82(-\delta \psi^2/L) + 0.81$. 
Figure S10: The sample average of folding free energy change, $\overline{\Delta \Delta G_{ND}} = k_B T_s \overline{\Delta \Delta \psi_{ND}}$, is plotted against the ensemble average of folding free energy per residue, $\langle \Delta G_{ND}\rangle_{\sigma}/L = k_B T_s \langle \Delta \psi_{ND}\rangle_{\sigma}/L$, for each protein family. In the case of the cutoff distance 8 Å, the correlation coefficient is $r = -0.74$, and the regression line is $\Delta \Delta G_{ND}(\sigma^N_{j \alpha}, \sigma^N_{i} \rightarrow \sigma_i) = -2.90(\Delta G_{ND})_{\sigma}/L + 1.00$. In the case of $r_{\text{cutoff}} \sim 16$ Å, the correlation coefficient is $r = -0.60$, and the regression line is $\Delta \Delta G_{ND}(\sigma^N_{j \alpha}, \sigma^N_{i} \rightarrow \sigma_i) = -3.23(\Delta G_{ND})_{\sigma}/L + 1.14$. The free energies are in kcal/mol units.
Figure S11: $\hat{T}_s / \hat{T}_g$ is plotted against $T_m / \hat{T}_g$ for each protein domain. A dotted curve corresponds to Eq. S33 $\hat{T}_s / \hat{T}_g = 2(T_m / \hat{T}_g) / ((T_m / \hat{T}_g)^2 + 1)$. Plus and open circle marks indicate the values estimated with $r_{\text{cutoff}} \sim 8$ and 15.5 Å, respectively. The $T_s$ and $T_g$ must satisfy $T_s < T_g < T_m$ for proteins to be able to fold into unique native structures.
Figure S12: Folding free energies, $\langle \Delta G_{ND} \rangle_\sigma$, predicted by the present method are plotted against their experimental values, $\Delta G_{ND}(\sigma_N)$. The free energies are in kcal/mol units.