In: Protein Folding
Editor: Eric C. Walters
ISBN: 978-1-61728-990-3

Chapter 11

STUDY OF FOLDING/UNFOLDING KINETICS OF LATTICE PROTEINS BY APPLYING A SIMPLE STATISTICAL MECHANICAL MODEL FOR PROTEIN FOLDING

Hiroshi Wako[1] and Haruo Abe[2]
[1]School of Social Sciences, Waseda University, Tokyo 169-8050, Japan
[2]Department of Digital Engineering, Nishinippon Institute of Technology, Fukuoka 800-0394, Japan

ABSTRACT

The folding/unfolding kinetics of a three-dimensional lattice protein was studied using a simple statistical mechanical model for protein folding that we had developed earlier. The model considers the specificity of an amino acid sequence and the native structure of a given protein. We calculated the characteristic relaxation rate on the free energy surface starting from a completely unfolded structure (or native structure) that is assumed to associate with a folding rate (or an unfolding rate). The chevron plot of these rates as a function of the inverse temperature was obtained for four lattice proteins, *a1*, *a2*, *b1*, and *b2*, in order to investigate the dependency of the folding and unfolding rates on their native structures and amino acid sequences. Proteins *a1* and *a2* fold to the same native structure, but their amino acid sequences differ. The same is true for proteins *b1* and *b2*, but their native structure is different from that of *a1* and *a2*. To elucidate the roles of individual amino acid residues in protein folding/unfolding kinetics, we calculated the kinetic properties for all possible single amino acid substitutions of these proteins and examined their responses. The results are discussed with respect to the roles of short- and long-range interactions and formation of a folding nucleus in the kinetics of protein folding/unfolding.

1. INTRODUCTION

A protein molecule is a heteropolymer. Heterogeneity in the amino acid sequence is essential to the unique native structure of a protein. The 20 naturally occurring amino acids are characterized by their physicochemical properties such as hydrophobicity, polarity, acidity/basicity, bulkiness, and hydrogen-bonding ability. Accordingly, they have specific roles in the folding of a protein to its native structure. Since each amino acid is characterized by several properties, the same amino acid residues in different proteins can play different roles depending on their environment, such as the secondary structure they are involved in and the surrounding amino acid residues with which they interact. The many-body nature of the interactions between amino acid residues, however, makes it difficult to understand the roles of individual amino acid residues in protein folding.

Amino acid substitution is a useful method to explore the roles of individual amino acid residues in proteins. From the response to perturbation caused by amino acid substitution, it is possible to assess the role of the amino acid residue at the substituted site. For example, stabilizing contributions of an amino acid residue at the substituted site to the native structure have been evaluated by differential scanning calorimetry [1, 2]. The effect of amino acid substitutions is well correlated with their physicochemical properties such as hydrophobicity in some cases, but proteins tolerate the amino acid substitution by their intrinsic flexibility; thus, few effects on structural stability were observed in other cases. Consequently, it is usually difficult to predict changes in structural stability induced by amino acid substitution, although it may be possible to interpret the change caused by the substitution afterwards.

The Φ value is another characteristic of individual residues. It is calculated from the changes in the folding and unfolding rates (k_f and k_u) of a single amino acid substitution mutant from a wild-type protein and is used to evaluate the stabilizing contribution of an amino acid residue at the substituted site to the structure of the folding transition state [3, 4]. Changes in the various kinetic parameters, such as the folding and unfolding rates and *m*-values of mutants with single and double amino acid substitutions, have also been examined and discussed extensively [5-7].

In this paper, we study the abovementioned problem by applying a simple statistical mechanical model, which we had developed earlier, to a lattice protein [8-11]. The lattice protein was used because it is simple, and thus, its conformational space is well defined for a statistical mechanical analysis. Although the artifacts arising from the lattice protein and the simple statistical mechanical model are unavoidable to some extent, they are useful for deriving significant aspects of the protein-folding problem.

In the statistical mechanical study of protein folding, it is important for the model to explicitly incorporate heterogeneity in the amino acid sequence and a unique native protein structure, as described above. Since our model satisfies these conditions, we could investigate two different native structures and two different amino acid sequences that folded to the same native conformation for comparative analysis. In addition, we examined changes in thermodynamic properties for all possible single amino acid substitutions [10]. For example, from the relationship between conformational energy change, $\Delta E(\xi_i)$, and transition temperature change, $\Delta T_m(\xi_i)$, caused by the substitution of the amino acid ξ_i at the ith residue, it was found that although both $\Delta E(\xi_i)$ and $\Delta T_m(\xi_i)$ strongly depend on the amino acid sequences of the two proteins that fold to the same native structure, the slopes of linear

regression lines between $\Delta E(\xi_i)$ and $\Delta T_m(\xi_i)$ (i.e., susceptibility of $T_m(\xi_i)$ to $\Delta E(\xi_i)$) for all possible single amino acid substitution mutants at the given residue i of the two proteins are mainly determined by their native conformations and are less dependent on their amino acid sequences.

In a recent paper [11], we presented a method for calculating the folding and unfolding rates from the free energy profiles of lattice proteins, thereby obtaining their chevron plots. After this paper, we studied the change in the folding and unfolding rates induced by a single amino acid substitution to characterize individual residues from a kinetic viewpoint. A comparison of the changes in the two proteins that fold to the same native conformation but have different amino acid sequences was expected to reveal the dependence of kinetic properties of folding and unfolding on the amino acid sequences and native structures. Through this examination, we discuss the roles of short- and long-range interactions and the formation of a folding nucleus in protein folding/unfolding kinetics in this paper.

2. METHOD

2.1. Three-Dimensional Lattice Proteins

We consider a three-dimensional (3D) cubic-lattice protein consisting of n monomers, each of which is regarded as an amino acid residue of 20 types. We adopted four proteins, *a1*, *a2*, *b1*, and *b2*, which are shown in Figure 1. The number of amino acid residues (n) is 36 for *a1* and *a2*, and 48 for *b1* and *b2*. Their amino acid sequences are also shown in Figure 1. The native structures of proteins *a1* and *a2* are identical, but their amino acid sequences differ. The same is true for proteins *b1* and *b2*, but their native conformation is different from that of proteins *a1* and *a2*. Mirny et al. [12] and Abkevich et al. [13] designed proteins *a1* and *a2*, and *b1* and *b2*, respectively, to be folded to their native conformations.

These lattice proteins were originally designed for folding and unfolding simulations. In the simulation, two monomers are considered to interact with each other if and only if they occupy nearest-neighbor lattice points but are not covalently bonded along the polypeptide chain in a conformation. The total energy, E, of the given conformation in the simulation is given as

$$E = \sum_{\substack{1 \le i < j \le n \\ j - i \ne 1}} U(\xi_i, \xi_j)\delta_{ij}, \tag{2.1}$$

where $\delta_{ij} = 1$ if the monomers i and j are lattice neighbors, otherwise $\delta_{ij} = 0$; ξ_i is an amino acid residue type at position i, and $U(\xi_i, \xi_j)$ is the magnitude of the interaction energy between the amino acid residues ξ_i and ξ_j. For $U(\xi_i, \xi_j)$, we used the values from the statistical distributions of contacts in real proteins determined by Miyazawa and Jernigan [14].

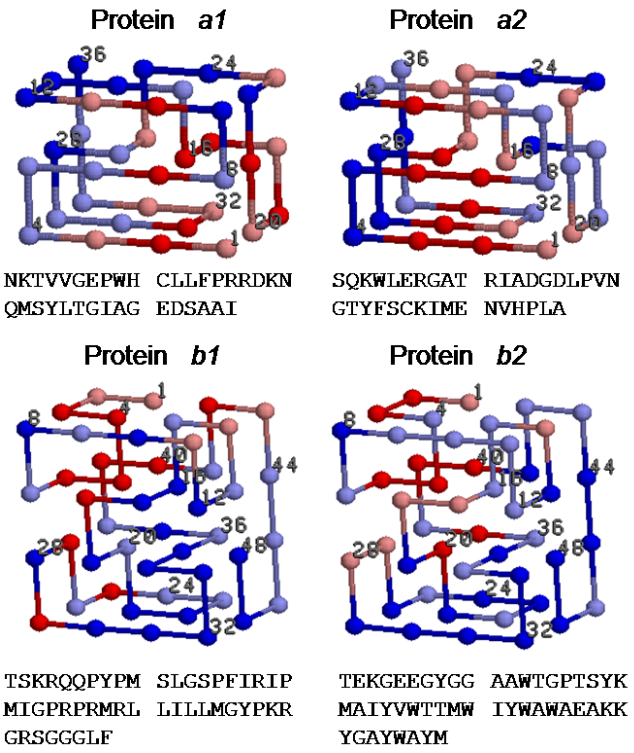

Figure 1. Native conformations of lattice proteins *a1*, *a2*, *b1*, and *b2* and their amino acid sequences. The number of amino acid residues is 36 for proteins *a1* and *a2*, and 48 for proteins *b1* and *b2*. The native conformations of proteins *a1* and *a2* are identical, but their amino acid sequences differ. The same is the case for proteins *b1* and *b2*. The amino acid residues are represented by beads and are colored to distinguish the following four groups: (1) hydrophobic and large amino acids (I, L, M, F, W, and Y) in blue, (2) hydrophobic and small amino acids (A, G, P, and V) in light blue, (3) polar and small amino acids (N, D, C, S, and T) in pink, and (4) polar and large amino acids (R, E, Q, H, and K) in red

2.2. Statistical Mechanical Model for Protein Folding and Unfolding

The simple statistical mechanical model used in this paper was constructed with the intention to introduce the stepwise protein folding paradigm [8-11, 15-18]. In this paradigm, a protein folds in a stepwise manner along the polypeptide chain (see Figure 2). In the first stage of folding, short-range interactions work dominantly to form small native-like structures such as α-helices, β-strands, and turns. In the next stage, these structures grow gradually through medium-range interactions. Finally, these substructures coalesce into the native structure by long-range interactions.

The statistical mechanical model for protein folding and unfolding, particularly for the cubic-lattice protein studied here, in the abovementioned paradigm is formulated as follows [8]:

(1) Each amino acid residue is assumed to be in either a native state or a random-coil state.

(2) A protein conformation at any stage of the folding process is represented by a sequence of two types of regions of various sizes, namely, a local structure and a random-coil region, arranged alternately along the chain. The term "local structure" is used with specific meaning in this paper. A local structure and a random-coil region are defined as continuous regions in which all amino acid residues are in the native state or random-coil state, respectively.

(3) The key assumption of this statistical mechanical model is that only the Go-type native interactions between amino acid residues within a local structure are considered (the illustrative examples are shown as being related to the free energy profile in Figure 2). The other interactions such as those between residues in the different local structures and those within a random-coil region are neglected (see the inset in Figure 2 for example).

(4) With regard to the free energy within a random-coil region (where no interaction between the residues is assumed to exist), it is assumed that only the chain entropy dependent on the number of residues contributes to the partition function. The random-coil state is the reference state, that is, its statistical weight is set to unity.

(5) The minimum size of a local structure is four residues in the lattice protein. Amino acid residue i is regarded to be in the native state if a segment of four consecutive amino acid residues, $(i - 1)$ to $(i + 2)$, takes the same conformation as the native structure ($i = 2, 3, \ldots, n - 2$). Otherwise, the amino acid residue is considered to be in the random-coil state.

Following the above scheme, the partition function for this statistical mechanical model is obtained by summation of the statistical weights of the local structures and the random-coil regions over all possible arrangements. This is obtained by the repetitive use of the following recurrent equation:

$$Z_{1,j} = Z_{1,j-1} + \sum_{m=1}^{j-3} f(m,j)^{-1} \exp\{-\beta E(m,j)\} Z_{1,m+1} \quad (j = 4, 5, \ldots, n-1, n),$$

$$Z_{1,1} = 0,\ Z_{1,2} = Z_{1,3} = 1 \tag{2.2}$$

where $\beta = 1/k_B T$, k_B is the Boltzmann constant, and T is the absolute temperature. $Z_{1,j}$ is the auxiliary partition function of a hypothetical protein molecule consisting of the amino acid residues 1 to j. By definition, the partition function of the entire protein molecule is $Z_{1,n}$ ($\equiv Z(T)$) (see Figure 3 for illustrative representation of eq. (2.2)).

$E(m,j) = \sum_{m \le k,l \le j} U(\xi_k, \xi_l) \Gamma_{k,l}$ is the conformational energy of a local structure consisting of amino acid residues m to j, where $U(\xi_k, \xi_l)$ is the interaction energy between the amino acid residues ξ_k and ξ_l (see Section 2.1), and $\Gamma_{k,l} = 1$ if the amino acid residues k and l are in contact in the native conformation; otherwise, $\Gamma_{k,l} = 0$. $U(\xi_k, \xi_l)$ depends on the amino acid types ξ_k and ξ_l. The residue pair in contact, $\Gamma_{k,l}$, depends on the native structure. Consequently, $E(m, j)$ depends on the amino acid sequence and the native structure. In other words, the amino acid sequence and the native structure are considered through $E(m, j)$ in this model.

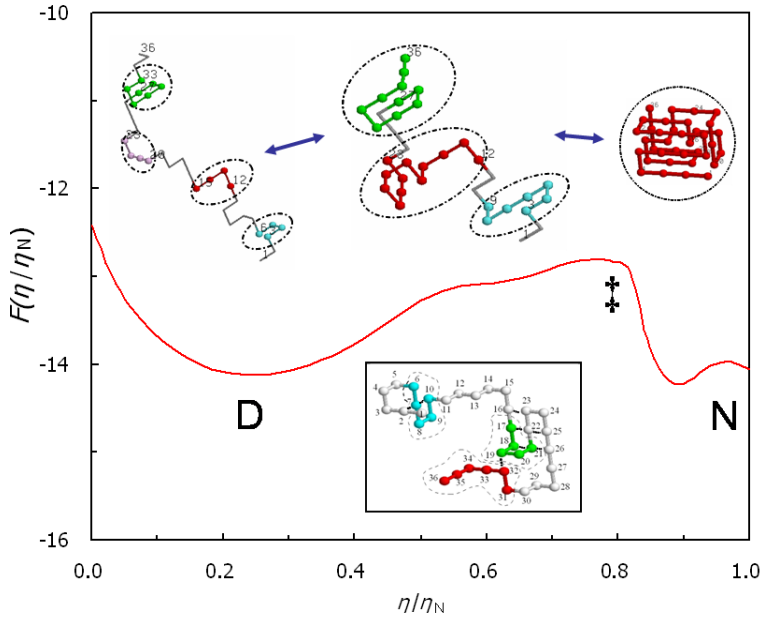

Figure 2. Schematic representation of folding and unfolding of a lattice protein and its free energy profile. The free energy profile $F(\eta/\eta_N)$ is plotted against a reaction coordinate η/η_N (where η is the number of amino acid residues in the native state and η_N is the maximum number of η) at the transition temperature ($T_m = 0.240$) for protein *a1*. The letters N, ‡, and D indicate the native-, transition- and denatured-state regions, respectively. Each conformation consists of the local structures (indicated by color and are enclosed) and the random-coil region (a white region). A local structure is defined as a continuous region where all amino acid residues are in the native state and a random-coil region is a continuous region where all residues are in the random-coil state. In the statistical mechanical model used in this paper, only the interaction energies between residues within a local structure are considered. The other interactions such as those between different local structures and those within a random-coil region are neglected. The conformation in the inset is an example of some of the contacts between residues in the different local structures and those between the local structure and the random-coil region ignored while calculating the conformational energies.

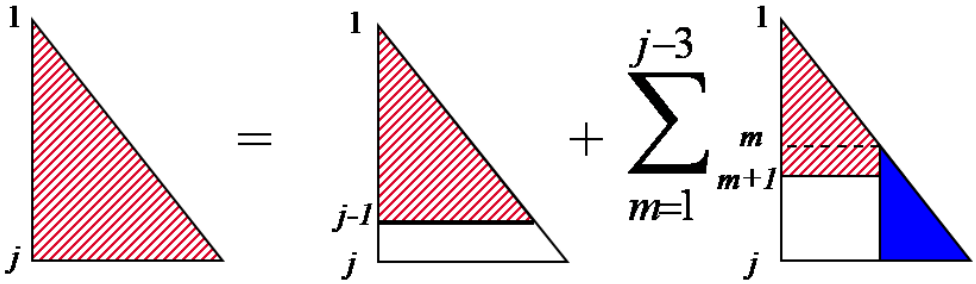

Figure 3. Illustrative representation of the recurrent equation (2.2). The auxiliary partition function $Z_{1,j}$ for a hypothetical molecule consisting of amino acid residues 1 to j is depicted by a hatched triangle. In the last term, the filled triangle indicates the local structure consisting of amino acid residues m to j. Summation is taken over all possible local structures to which the amino acid residue j belongs in the hypothetical molecule

Since $f(m, j)$ corresponds to the number of possible conformations of the segment consisting of the amino acid residues m to j in the random-coil state, $k_B \ln f(m, j)$ is the chain entropy of the segment in the random-coil state. Then, $k_B \ln f(m, j)^{-1}$ gives the entropy loss of the segment when it forms the local structure. We used the form $f(m, j) = 1.4084 \times (4.750)^{j-m-2}$ obtained for a cubic-lattice polymer in our previous work [8].

Since there are no distinctions between energy and enthalpy in this system, we simply regarded the conformational energy as the enthalpy. For computational convenience, the energy (enthalpy) of the system E_h is expressed by the integer h in the units of 0.01, i.e., $E_h = h\varepsilon_0$ and $\varepsilon_0 = 0.01$ (because the effect digit of residue-residue interactions given by Miyazawa and Jernigan [14] used in this model is the second digit after the decimal). Eventually, the partition function Z is given as a polynomial in two variables, t and u, as a function of temperature T, as follows [10]:

$$Z(T) = \sum_{\eta} \sum_{h} \Omega(\eta, h) t^{\eta} u^{h} \tag{2.3}$$

where

$$u = \exp(-\beta\varepsilon_0) \tag{2.4}$$

and t is a dummy parameter introduced to count the number of amino acid residues in the native state, η, and is set to unity in the last result. η runs from 0 to $n - 3$. The coefficient $\Omega(\eta, h)$ for given values of η and h can be calculated using recurrent equation (2.2).

The partition function $Z(T)$ can be rewritten as follows:

$$Z(T) = \sum_{h} W_1(h, T) u^{h} \tag{2.5}$$

$$W_1(h, T) = \sum_{\eta=0}^{n-3} \Omega(\eta, h) t^{\eta} \tag{2.6}$$

or

$$Z(T) = \sum_{\eta=0}^{n-3} W_2(\eta, T) t^{\eta} \tag{2.7}$$

$$W_2(\eta, T) = \sum_{h} \Omega(\eta, h) u^{h} \tag{2.8}$$

where $W_1(h, T)u^h$ and $W_2(\eta, T)t^\eta$ are the sums of the statistical weights over all states with the given enthalpy E_h ($=h\varepsilon_0$) and with the given number of amino acid residues in the native state η, at temperature T, respectively.

We can define the free energy for a given η from eq. (2.7) (the dummy parameter t in eq. (2.7) is set to unity here):

$$F(\eta, T) = -k_B T \ln W_2(\eta, T) \tag{2.9}$$

This formula was used to calculate the free energy profile for the four lattice proteins studied in this paper.

Eventually, the partition function, free energy, enthalpy, and entropy at a given temperature T are given as

$$Z(T) = \sum_{\eta=0}^{n-3} \exp\{-F(\eta,T)/k_B T\} \tag{2.10}$$

$$F(T) = -k_B T \ln Z(T) \tag{2.11}$$

$$H(T) = \sum_h E_h W_1(h,T) u^h / Z(T) \tag{2.12}$$

$$S(T) = \frac{1}{T}\{H(T) - F(T)\} \tag{2.13}$$

Here, the statistical mechanical model of protein folding is described somewhat specifically for the lattice protein. However, this model is applicable to a protein in a more general case, including real proteins. In fact, there are many studies on protein folding based on the same assumptions that are used in this paper [19-26].

2.3. Calculation of Folding and Unfolding Rates

The kinetics of the folding and unfolding process of proteins (such as folding and unfolding rates) were formulated by Muñoz and Eaton [22] as a motion along a one-dimensional free energy profile with the number of amino acid residues in the native state and were extensively examined by Henry and Eaton [23]. Since we can calculate the free energy profile $F(\eta, T)$ for the lattice proteins studied here as described above, we applied the method of Muñoz and Eaton to our model. According to their method, by using a simple approach that involves solving a system of differential equations describing reversible hopping between the adjacent discrete values of reaction coordinates (η and $\eta + 1$ in this study), the characteristic relaxation rate can be given as follows [22, 23]:

$$\frac{1}{k} \propto \tau \equiv \int_0^\infty \frac{\langle\eta\rangle_{eq} - \langle\eta(t)\rangle}{\langle\eta\rangle_{eq} - \langle\eta(0)\rangle} dt$$

$$= \{\langle\eta\rangle_{eq} - \langle\eta\rangle_0\}^{-1} \sum_{j=0}^{n-4} \frac{1}{p_{eq}(j) s_{j,j+1}} \sum_{i=j+1}^{n-3} \{p_{eq}(i) - p_0(i)\} \sum_{\eta=0}^{j} p_{eq}(\eta)(\langle\eta\rangle_{eq} - \eta) \tag{2.14}$$

Here, an equilibrium value of η at temperature T

$$\langle \eta \rangle_{\mathrm{eq}} = \sum_{\eta=0}^{n-3} \eta \, p_{\mathrm{eq}}(\eta) \tag{2.15}$$

can be calculated using $F(\eta, T)$, where $p_{\mathrm{eq}}(\eta)$ is the probability that a conformation has η amino acid residues in the native state:

$$p_{\mathrm{eq}}(\eta) = Z^{-1} \exp\{-F(\eta, T) / k_{\mathrm{B}} T) \tag{2.16}$$

$\langle \eta \rangle_{\mathrm{eq}}$ and $p_{\mathrm{eq}}(\eta)$ are functions of T, but T is omitted for clarification.

The relaxation rate k is estimated as the mean rate of relaxation of the average number of native residues to its equilibrium value, starting with the entire population in the completely unfolded state of $\eta = 0$:

$$p_0(\eta) = \delta_{\eta,0} \tag{2.17}$$

$$\langle \eta \rangle_0 = \sum_{\eta=0}^{N-3} \eta \, p_0(\eta) = 0 \tag{2.18}$$

The hopping rates from j to $j + 1$ and from $j + 1$ to j are assumed as

$$s_{j,j+1} = \gamma \left(\frac{p_{\mathrm{eq}}(j+1)}{p_{\mathrm{eq}}(j)} \right)^{\kappa} \text{ and } s_{j+1,j} = \gamma \left(\frac{p_{\mathrm{eq}}(j+1)}{p_{\mathrm{eq}}(j)} \right)^{\kappa - 1}, \tag{2.19}$$

respectively, in order to satisfy the detailed balance $s_{j,j+1} / s_{j+1,j} = p_{\mathrm{eq}}(j+1) / p_{\mathrm{eq}}(j)$. Following Henry and Eaton [23], we set κ to 0.5 and γ to 1 (they suggested that the choice for κ had little effect on the results).

However, this characteristic rate depends on the initial condition because of its approximation. We can consider another initial condition; all populations are in the native state of $\eta = n - 3$ as follows [11]:

$$\frac{1}{k} \propto \tau \equiv \int_0^{\infty} \frac{\langle \eta \rangle_{\mathrm{eq}} - \langle \eta(t) \rangle}{\langle \eta \rangle_{\mathrm{eq}} - \langle \eta(0) \rangle} dt$$

$$= \{ \langle \eta \rangle_{\mathrm{eq}} - \langle \eta \rangle_0 \}^{-1} \sum_{j=0}^{n-4} \frac{1}{p_{\mathrm{eq}}(j+1) s_{j+1,j}} \sum_{i=0}^{j} \{ p_{\mathrm{eq}}(i) - p_0(i) \} \sum_{\eta=j+1}^{n-3} p_{\mathrm{eq}}(\eta) (\langle \eta \rangle_{\mathrm{eq}} - \eta) \tag{2.20}$$

$$p_0(\eta) = \delta_{\eta,n-3} \tag{2.21}$$

$$\langle \eta \rangle_0 = \sum_{\eta=0}^{n-3} \eta p_0(\eta) = n-3 \tag{2.22}$$

In Figure 4, the illustrative examples of two cases of k are given (referred to as k_f calculated by eq. (2.14) and k_u calculated by eq. (2.20)). The logarithmic rates ln k_f and ln k_u are plotted against $1/T$. Their behaviors differ considerably. In this estimation, the characteristic relaxation rate for a given temperature is approximated by the smallest of the two rates (although it is possible to consider the other initial conditions that lead to different relaxation rates, we considered the two extreme cases and chose the smaller one for a given temperature). As a result, k_f above $1/T_m$ and k_u below $1/T_m$ are assumed as the folding and unfolding rates, respectively, in this approximation. For the estimation of ln k_f and ln k_u near the transition temperature T_m, we simply assumed the linearity of the logarithmic rates against the inverse temperature around $1/T_m$. We estimated ln k_f and ln k_u by linearly extrapolating ln k_f from higher to lower $1/T$ values and ln k_u from lower to higher $1/T$ values, respectively. The intersecting point of the two extrapolated lines is defined as the transition temperature T_m, i.e., ln $k_f(T_m)$ = ln $k_u(T_m)$ (see the inset in Figure 4). (For more discussions about drawing a chevron plot, see reference [11].)

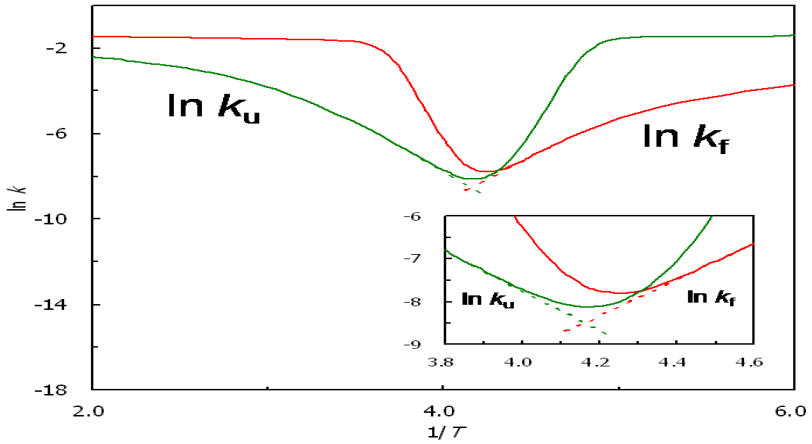

Figure 4. Illustrative examples of logarithmic folding and unfolding rates, ln k_f and ln k_u. The red and green solid curves are intact ln k_f and ln k_u values obtained from eqs. (2.14) and (2.20), respectively, for protein *a1*. The red and green dashed lines around $1/T_m = 4.17$ are obtained for ln k_f and ln k_u, respectively, by assuming their linearity. In the inset, the magnified view around $1/T_m$ is shown. Eventually, the lowest lines and curves are assumed to be a chevron plot (see text for details)

2.4. Single Amino Acid Substitutions

We considered all possible single amino acid substitutions for the four proteins (36 × 19 mutants for proteins *a1* and *a2*, respectively, and 48 × 19 mutants for proteins *b1* and *b2*, respectively). By the single amino substitution at the ith residue (for example, the amino acid residue w_i in the wild-type protein is replaced by the amino acid ξ_i), we simply assumed that $U(w_i, \xi_j)$ is transformed to $U(\xi_i, \xi_j)$ in eq. (2.1). The responses to the substitutions were examined on the basis of changes in the logarithmic folding and unfolding rates, ln k_f and

ln k_u, in this paper. The changes $\Delta\ln k_f$ and $\Delta\ln k_u$ are used to calculate the Φ value defined by eq. (2.23) below.

2.5. Φ Value Analysis

Φ value analysis, introduced by Fersht et al. [3, 4], is the most commonly used method to interpret the changes $\Delta\ln k_f$ and $\Delta\ln k_u$ in single amino acid substitution studies. The Φ value is defined as the ratio of the change in free energy of activation for folding, $\Delta\Delta F_{‡\text{-D}}$, to the equilibrium free energy of folding, $\Delta\Delta F_{\text{N-D}}$, between a wild-type and single amino acid substitution mutant (see Figure 5). Owing to the substituted amino acid working as a reporter of structural changes, it is possible to evaluate the importance of a mutated residue in stabilizing the folding transition state structure. If no structure forms at the position of the mutation in the transition state, there is no difference in free energy between the mutant and wild-type, and $\Phi = 0$. If the structure completely forms at that position in the transition state, the free energy difference is as large as in the folded state, and $\Phi = 1$. Consequently, Φ value analysis indicated which amino acid residues are in the native state in the transition state of the folding and which are not.

The Φ value is defined as

$$\Phi = \frac{\Delta\Delta F_{‡\text{-D}}}{\Delta\Delta F_{\text{N-D}}} = \frac{\Delta \ln k_f}{\Delta \ln k_f - \Delta \ln k_u} \tag{2.23}$$

where $\Delta\ln k_f = \ln k_f^{mut}(T_m) - \ln k_f^{wild}(T_m)$ and $\Delta\ln k_u = \ln k_u^{mut}(T_m) - \ln k_u^{wild}(T_m)$ [4, 27]. Since $\ln k_f^{wild}(T_m) = \ln k_u^{wild}(T_m)$, $\Delta\ln k_f - \Delta\ln k_u = \ln k_f^{mut}(T_m) - \ln k_u^{mut}(T_m)$. These relationships are explained pictorially in Figure 5b.

3. RESULTS

In the first paper in our study series [8], we showed that the above statistical mechanical model could well reproduce the folding/unfolding transition curves obtained by the Monte Carlo simulations without any adjusting parameters, even though the interaction energies for non-native contacts were taken into account in the Monte Carlo simulations. In the next paper [9], the equilibrium thermodynamic properties such as enthalpy, entropy, and free energy, were calculated for the four proteins *a1*, *a2*, *b1*, and *b2*, and the results were extensively discussed by comparing to the Monte Carlo simulations. In particular, we examined the degree to which the non-native contacts, which were not considered in the present model but were considered in the simulation, affected the folding and showed that such contacts are not negligible, but have minor contributions to the thermodynamic properties of the folding. In the subsequent paper [10], in order to elucidate the roles of individual residues in the stability of the native structure, the susceptibility of the residues to single amino acid substitutions were studied, analyzing the changes in the thermodynamic properties, in particular, the change in transition temperatures for all possible single amino acid substitutions of the four proteins. By comparing the two proteins with different amino acid sequences but identical native structures, we suggested that the susceptibility of the residue to the amino acid

substitution (the slope of the linear regression lines between the two changes for a given residue) is mainly determined by their native conformations and are less dependent on their amino acid sequences, despite the fact that both changes in conformational energy and transition temperature strongly depend on the amino acid sequence. Finally, in the preceding paper [11], the statistical mechanical model was developed into the folding/unfolding kinetics problem, and the chevron plots for the four proteins were discussed.

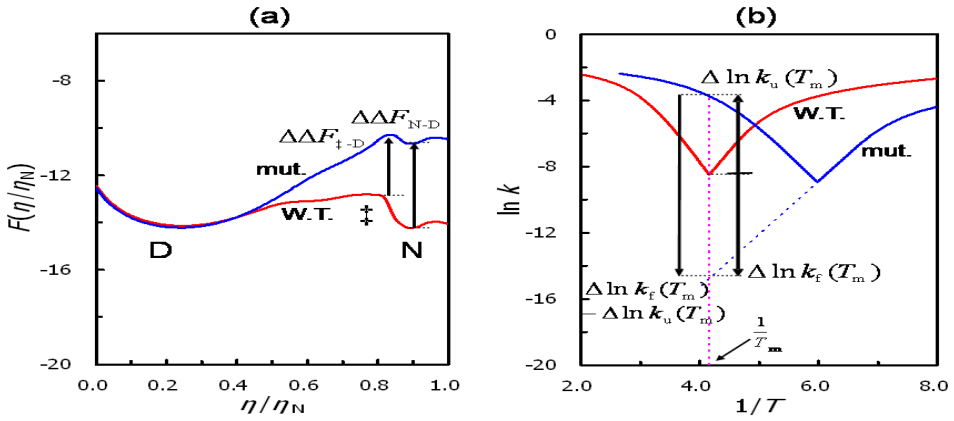

Figure 5. Illustrative example of free energy profiles and chevron plots of a wild-type and a mutant for calculating the Φ value. (a) Free energy profiles for the wild-type (red) and the R16M mutant (blue) of protein *a1* plotted against η/η_N at the transition temperature of the wild-type protein *a1* ($T_m = 0.240$). The letters N, D, and ‡ indicate the native-, denatured-, and transition-state regions, respectively. $\Delta\Delta F_{‡\text{-}D}$ is the change in activation free energy for folding and $\Delta\Delta F_{N\text{-}D}$ is the change in equilibrium free energy of folding. It should be noted that the free energy levels of the denatured states of the wild-type and mutant are equal in this example. (b) Chevron plots, or logarithmic folding ($\ln k_f$) and unfolding rates ($\ln k_u$), for wild-type (red) and R16M mutant (blue) of protein *a1* plotted against the inverse temperature. The Φ value defined by eq. (2.23) is the ratio of $\Delta\ln k_f(T_m)$ to $\Delta\ln k_f(T_m) - \Delta\ln k_u(T_m)$ (see text for details). In this example $\Delta\ln k_f(T_m) = -6.265$, $\Delta\ln k_u(T_m) = 4.739$, and $\Phi = 0.569$

In this paper, we focus our attention on the folding/unfolding kinetics of the four proteins to further develop the preceding study [11]. First, we will review the chevron plots of the four proteins briefly in Section 3.1, and then examine the changes in the chevron plots caused by the single amino acid substitutions. On the basis of this analysis, we will discuss the roles of short- and long-range interactions and formation of a folding nucleus in the folding/unfolding kinetics.

Hereafter, we use the order parameter η (the number of residues in the native state) normalized by η_N, where $\eta_N = n - 3$, the maximum value of η, to treat the proteins with different residue numbers (i.e., $n = 36$ of proteins *a1* and *a2*, and $n = 48$ of proteins *b1* and *b2*) together. In addition, temperature T in the argument of the properties is omitted for clarification. For example, $F(\eta, T)$ is denoted as $F(\eta/\eta_N)$.

3.1. Folding and Unfolding Rates

The free energy $F(\eta/\eta_N)$ is plotted along the reaction coordinate η/η_N ($0 \le \eta/\eta_N \le 1$) at the transition temperature T_m for the individual proteins to give the free energy profiles.

Additionally, the logarithmic folding and unfolding rates ln k_f and ln k_u as a function of $1/T$ (i.e., the chevron plots) for proteins *a1*, *a2*, *b1*, and *b2* [11] are shown in Figure 6.

In the free energy profiles for proteins *a1* and *a2*, $F(\eta_‡/\eta_N) - F(\eta_D/\eta_N) = 1.3$ and 1.7, and $F(\eta_‡/\eta_N) - F(\eta_M/\eta_N) = 1.4$ and 1.8, respectively, where $F(\eta_‡/\eta_N)$ is the local maximum value in the transition-state region, and $F(\eta_D/\eta_N)$ and $F(\eta_M/\eta_N)$ are the local minimum values in the denatured-state and native-state regions, respectively. The transition-state region of protein *a1* has two peaks and is broader than that of *a2*. On the other hand, the native-state region of protein *a2* is broader than that of *a1*.

The chevron plots for proteins *a1* and *a2* are similar to each other. In detail, however, the transition temperature T_m of protein *a2* is higher than that of *a1*, suggesting that protein *a2* is more stable than protein *a1*. The unfolding rate of *a2* is smaller than that of *a1* below the $1/T_m$ of protein *a1*, whereas the folding rates of the two proteins are close to each other. As mentioned above, the barrier height against the unfolding for protein *a2* is slightly larger than that for *a1*, and the native-state region in the free energy profile of protein *a2* (Figure 6a) is broader than that of *a1*. These facts may be the reason why the unfolding rate of protein *a2* is smaller than that of *a1*. On the other hand, the folding arms of the chevron plots of proteins *a1* and *a2* are rather close to each other.

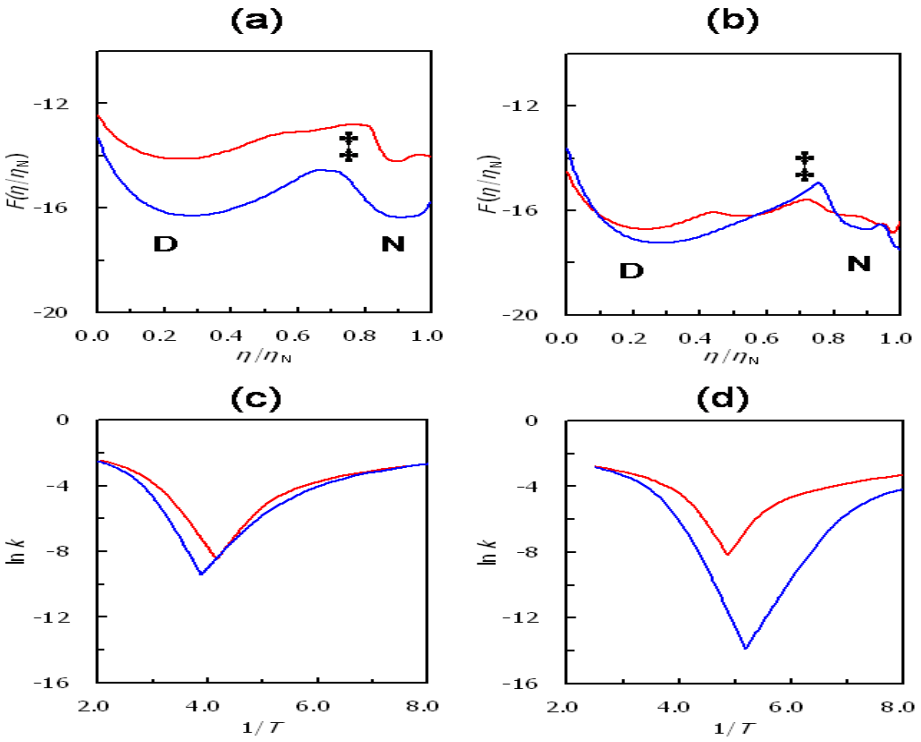

Figure 6. Free energy profiles and chevron plot for proteins *a1*, *a2*, *b1*, and *b2*. (a) The red and blue lines are the free energy profiles for protein *a1* at $T_m = 0.240$ and protein *a2* at $T_m = 0.257$, respectively. The letters N, ‡ and D indicate the native-, transition- and denatured-state regions, respectively. (b) The same as (a) but for protein *b1* at $T_m = 0.206$ (red line) and protein *b2* at $T_m = 0.193$ (blue line). (c) The red and blue lines are the logarithmic folding and unfolding rates, ln k_f and ln k_u, for proteins *a1* (red line) and protein *a2* (blue line), respectively. (d) The same as (c) but for protein *b1* (red line) and protein *b2* (blue line).

In contrast, the difference between the free energy profiles for proteins *b1* and *b2* is remarkable: $F(\eta_{\ddagger}/\eta_N) - F(\eta_D/\eta_N) = 1.1$ and 2.3, and $F(\eta_{\ddagger}/\eta_N) - F(\eta_M/\eta_N) = 1.3$ and 2.6, respectively. Both barriers against the folding and unfolding for protein *b2* are considerably greater than those for protein *b1*. In the free energy profile for protein *b2*, there are two peaks: the prominent one and the small but distinct one close to the native state (see Figure 6b). In contrast, the profile for protein *b1* has several peaks and is flatter than protein *b2*.

The chevron plots for proteins *b1* and *b2* are considerably different, reflecting the difference between their free energy profiles. The folding and unfolding rates of protein *b2* are much smaller than those for protein *b1*, because the $F(\eta_{\ddagger}/\eta_N) - F(\eta_D/\eta_N)$ and $F(\eta_{\ddagger}/\eta_N) - F(\eta_M/\eta_N)$ differences for protein *b2* are significantly larger than those for protein *b1*.

3.2. The Φ Value and Folding Nucleus

A slope, Φ_i, of a linear regression line between $\Delta\ln k_f(\xi_i)$ and $\Delta\ln k_f(\xi_i) - \Delta\ln k_u(\xi_i)$ of 19 mutants with an amino acid substitution at residue i is plotted against the residue number together with their correlation coefficient r_i for the four proteins in Figure 7. Φ_i is defined as the mean Φ value given by eq. (2.23) over various amino acid substitutions at residue i. The absolute values of r_i, $|r_i|$, close to unity indicate that their Φ values do not depend on the kind of amino acid substitutions. Although $|r_i|$ is close to unity for most of the residues, it should be noted that some residues for which $|r_i|$ is considerably small are exceptions. It should also be noted that some Φ values are largely negative. This point will be discussed in Section 4.5.

In Figure 7, the residues with $\Phi_i \geq 0.3$ and $r_i > 0.8$ are emphasized in the line plots and 3D structures. The region in which the residues have relatively high Φ values has a high probability of forming a native-like structure in the transition state. Such a region is usually identified as a folding nucleus. (This definition for a folding nucleus, which is based on high Φ values, is controversial, and thus discussed in Sections 4.1 and 4.4.)

In protein *a1*, defining the folding nucleus based on the Φ values results in a substructure consisting of two segments, residues 15–27 and 32–36. However, since residues 15–36, including the intersegmental residues 28–31, form one compact substructure in the native structure, it may be reasonable to consider residues 15–36 as a folding nucleus. The Φ values for residues 28–31 are small, presumably because their interactions with residues 2–5, which are formed in the native structure, are not formed at the transition state. Remember that Φ_i is close to unity if residue i has the same interactions as in the native structure. This means that Φ_i is not necessarily close to unity, even if residue i is in the native state by itself. Consequently, it is possible to consider that residues 28–31 take the native structure and are included in the folding nucleus, but do not interact with residues 2–5 in the transition state.

In protein *a2*, one segment, consisting of residues 8–27, is believed to compose the folding nucleus, although the Φ values of residues 19–20 in protein *a2* are small. The small Φ values of residues 19–20 come from the long-range contacts between residues 20 and 31, and 19 and 32, which are not formed in the transition state, but are formed in the last stage of folding, similar to residues 28–31 in protein *a1*.

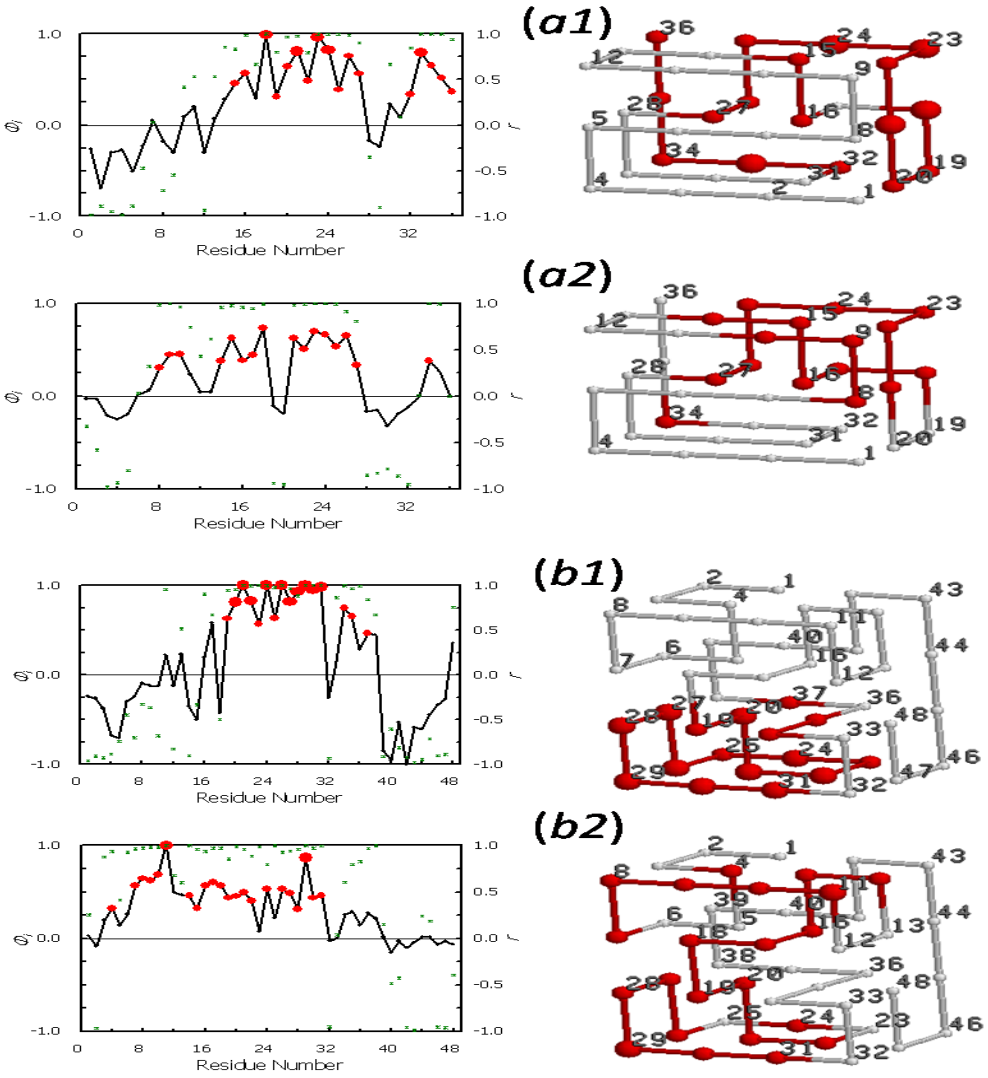

Figure 7. Φ_i and correlation coefficient r_i (panels on the left), and residues with high Φ_i values (3D structures on the right) for the four proteins *a1*, *a2*, *b1*, and *b2*. Φ_i (the line plot and the left vertical axis) is a slope of the linear regression line between $\Delta\ln k_f(\xi_i)$ and $\Delta\ln k_f(\xi_i) - \Delta\ln k_u(\xi_i)$ of 19 single amino acid substitution mutants at residue i, which is considered to be a mean Φ value over various amino acid substitutions at this position. The correlation coefficient r_i (the asterisk symbol and the right vertical axis) between $\Delta\ln k_f(\xi_i)$ and $\Delta\ln k_f(\xi_i) - \Delta\ln k_u(\xi_i)$ of the 19 mutants are shown together. The large and small closed circles on the line plot indicate the residues with $\Phi_i \geq 0.8$ and $r_i > 0.8$, and $0.3 \leq \Phi_i < 0.8$ and $r_i > 0.8$, respectively. These residues are also indicated on the 3D structures of the lattice proteins by color.

The segment consisting of residues 15–27 is common to both proteins *a1* and *a2* and forms the significant portion of the folding nuclei of these two proteins. The structural compactness of this segment can be observed in Figure 7. This common nucleus extends toward the C-terminus in protein *a1* and toward the N-terminus in protein *a2*. Obviously, the difference in their amino acid sequences is responsible for this difference; however, there is

another difference between the two proteins. While the Φ values of the residues in the folding nucleus in protein *a1* (except residues 28–31) are close to unity, those of the residues in the folding nucleus in protein *a2* are smaller than unity. Furthermore, while the unique nucleus formation in the transition state is suggested in protein *a1*, the mixture of various substructures is assumed in protein *a2*. Native contact occurs between the substituted amino acid residue and other residues in some substructures but not in others. It may be appropriate to define residues 8–27 as a nucleus composed of various substructures, in other words, an ensemble of relatively stable substructures of residues 8–27 in a statistical mechanical sense for protein *a2*.

The folding nuclei defined from the Φ values in proteins *b1* and *b2* differ significantly. In protein *b1*, the folding nucleus essentially consists of one segment, residues 19–37. Residues 32–33 in protein *b1* are included in the folding nucleus, even though their Φ values are rather small (the long-range contacts between residues 32 and 47, and 33 and 48 may be formed in the last stage of folding). On the other hand, in protein *b2*, the folding nucleus consists of residues 4–31, including residues 5–6, 12–13, 23, and 25, whose Φ values are small (the long-range contacts between residues 5 and 40, 6 and 39, 12 and 33, 13 and 44, 23 and 46, and 25 and 38 are responsible for the small Φ values). However, if it is reasonable to define a folding nucleus as a compact region in the 3D structure of a protein, then it may be better to consider two folding nuclei that consist of residues 4–18 and 19–31, or a mixture of various substructures included within the segment of residues 4–31 in protein *b2* (see the 3D structure of *b2* in Figure 7). Proteins *b1* and *b2* share residues 19–31, a portion of their folding nuclei. The conformation of this segment in the native structure is compact, as observed in Figure 7.

Similar to proteins *a1* and *a2*, it is observed that, while the Φ values of the folding nucleus residues in protein *b1*, with the exception of residues 32–33, are close to unity, those of the folding nucleus residues in protein *b2* are rather small. Likewise, while the unique nucleus formation in the transition state is suggested in protein *b1*, the folding nucleus in protein *b2* may be defined as an ensemble of various substructures, similar to that in protein *a2*.

3.3. Changes in the Folding and Unfolding Rates

The changes in the logarithmic folding and unfolding rates, $\Delta\ln k_f(\xi_i)$ and $\Delta\ln k_u(\xi_i)$, of 19 mutants with a substituted amino acid type ξ_i at the ith residue are calculated. They are linearly related to the changes in the total energy of the native conformation $\Delta E(\xi_i)$ for most cases (data not shown here). Therefore, the slopes χ_i^f and χ_i^u obtained by the linear regression lines between $\Delta\ln k_f(\xi_i)$ and $\Delta E(\xi_i)$, and $\Delta\ln k_u(\xi_i)$ and $\Delta E(\xi_i)$ of the 19 mutants with amino acid substitutions at residue i, respectively, are good indexes to characterize the individual residues from a kinetic viewpoint. χ_i^f and χ_i^u are plotted against residue number i for the four proteins *a1*, *a2*, *b1*, and *b2* in Figs. 8 and 9. The correlations of $\Delta\ln k_f(\xi_i)$ and $\Delta\ln k_u(\xi_i)$ with $\Delta E(\xi_i)$ are poor for some residues. In order to indicate such residues, the correlation coefficients are shown together, at right vertical axes in Figs. 8 and 9. χ_i^f and χ_i^u are considered the susceptibilities of $\ln k_f(\xi_i)$ and $\ln k_u(\xi_i)$ to the changes in the interaction energy between amino acid residues $\Delta E(\xi_i)$. The negative χ_i^f indicates that $\ln k_f(\xi_i)$ increases when a more stabilizing interaction is introduced by the single amino acid substitution, that is,

stabilizing mutations accelerate folding. On the other hand, the positive χ_i^{u} indicates that stabilizing mutations decelerate unfolding.

Figure 8 shows that χ_i^{f} for the residue in the folding nucleus is negative. This means that the stabilizing mutations within the nucleus speed up the folding. On the other hand, χ_i^{f} for the residues in the terminals and turns (e.g., residues 19–20 in proteins *a1* and *a2*, and 12–13 and 32–33 in proteins *b1* and *b2*) is small. This indicates that these residues contribute little to the folding rate.

There are some residues with a positive χ_i^{f} outside the folding nucleus, although they are relatively small in general. This seems peculiar because it means that introduction of a destabilizing interaction speeds up the folding, or conversely, that introduction of a stabilizing interaction slows down the folding. For such residues, χ_i^{u} is largely positive and the Φ value is negative. As discussed in Section 4.5, for some of these residues, the amino acid substitutions affect the free energy profile across the states (i.e., native, transition, and denatured states) macroscopically, and the overall distribution of statistical weights of individual conformations varies microscopically.

With regard to the unfolding rates, χ_i^{u} is positive for almost every residue, although there are a few exceptions that have very small negative values. This is reasonable because it means that the stabilizing mutation decelerates unfolding. However, it should be pointed out that in proteins *a1* and *b1*, while χ_i^{u} is largely positive for the residues near the terminals, it is relatively small for the residues in the folding nuclei. On the other hand, in proteins *a2* and *b2*, χ_i^{u} is largely positive not only for the residues near the terminals, as in proteins *a1* and *b1*, but also for the residues in the folding nuclei. These facts are probably related to the smaller Φ values of the residues in the folding nuclei of proteins *a2* and *b2* than those of proteins *a1* and *b1*, as described above.

4. DISCUSSION

4.1. Folding Nucleus and Short-Range Interactions

A folding nucleus is a spatially localized substructure of the native state formed in the transition state. Formation of the folding nucleus is necessary for subsequent rapid folding to the native state and essential for the two-state nature of the transition. The formation of the native structure of a local region smaller than the nucleus is unfavorable because the interactions within such a region are not sufficient to stabilize the specific 3D structure in overcoming the chain-entropy loss. There should be some critical size to compensate for the chain-entropy loss. Such a substructure is defined as a folding nucleus. In the equilibrium thermodynamics of protein folding, the instability of the transition state related to the nucleation is responsible for the two-state nature of the transition. On the other hand, in folding kinetics, the nucleation is a rate-limiting step, where once the folding nucleus has formed, the folding proceeds rapidly.

Roughly speaking, our results show that the residues in the folding nucleus have negative χ_i^{f}, and for the residues outside the nucleus, the χ_i^{f} is small (Figure 8). The negative χ_i^{f} indicates that the stabilizing interaction accelerates the folding, and the small χ_i^{f} indicates that the change in the interaction contributes little to the folding rates. In the case of a real protein,

Northey et al. [5] studied the effect of single and double amino acid substitutions in the hydrophobic core of the Fyn SH3 domain (which can be assumed as a folding nucleus) on the folding and unfolding rates and showed that more hydrophobic residues generally accelerate folding, as long as the residues do not have a disruptive effect on the native protein structure.

Since a folding nucleus is a compact substructure consisting of consecutive residues, short-range interactions play a dominant role in its stabilization. To characterize the short- and long-range interactions in the folding/unfolding kinetics, we examined the vertex residues of the rectangular-cuboid-shaped lattice proteins studied here (i.e., residues 1, 4, 9, 12, 19, 23, 34, and 36 in proteins *a1* and *a2*, and residues 2, 8, 11, 25, 29, 32, 43, and 46 in proteins *b1* and *b2*). Since these residues can interact with only one other residue, the effect of an amino acid substitution can be interpreted more directly than for residues interacting with more than one residue. Out of these residues, residues 23 and 34 in proteins *a1* and *a2*, and 8, 11, and 29 in proteins *b1* and *b2* have short-range interactions. Residue 18 in proteins *a1* and *a2* is also interesting in this sense, although it is not a vertex residue, as it interacts with two other residues, 21 and 23, in close range.

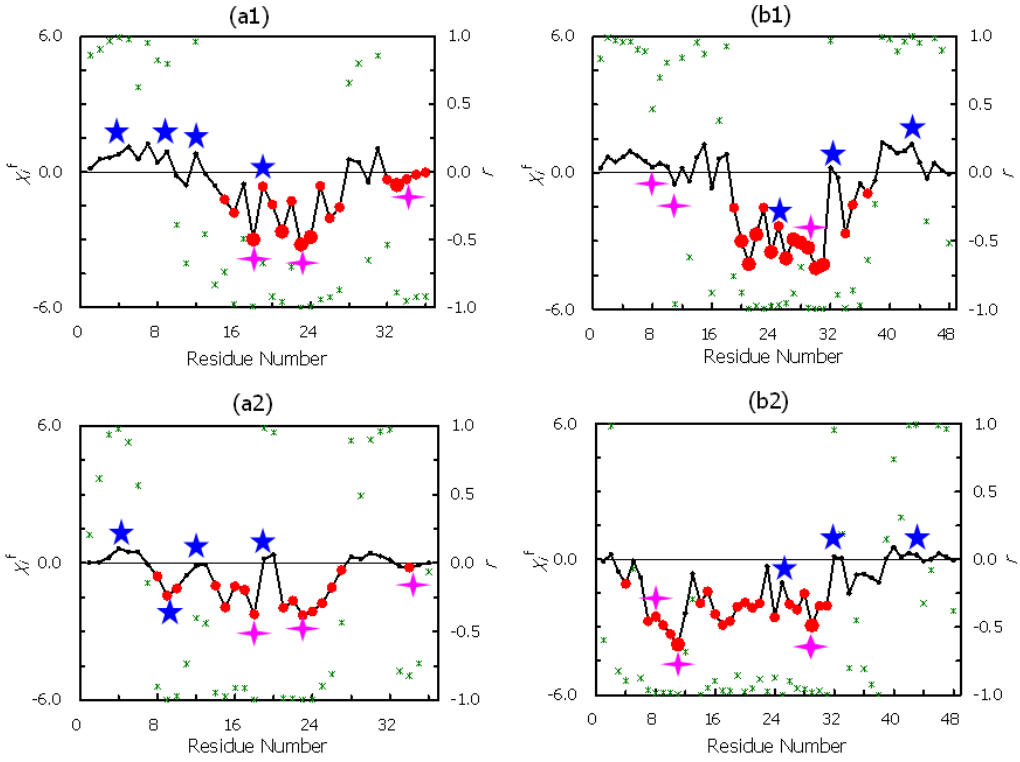

Figure 8. Susceptibility of folding rates to single amino acid substitution, χ_i^{f}, which is the slope obtained by the linear regression lines between $\Delta\ln k_{\mathrm{f}}(\xi_i)$ and $\Delta E(\xi_i)$ (see text in Sections 2.5 and 3.3). The asterisk indicates a correlation coefficient r_i between $\Delta\ln k_{\mathrm{f}}(\xi_i)$ and $\Delta E(\xi_i)$. The residues with a high Φ value (defined in Figure 7) are indicated by a closed circle on the line plot. The vertex residues with short- and long-range interactions are marked with four- and five-pointed star symbols, respectively (the terminal residues are ignored; see text for detail).

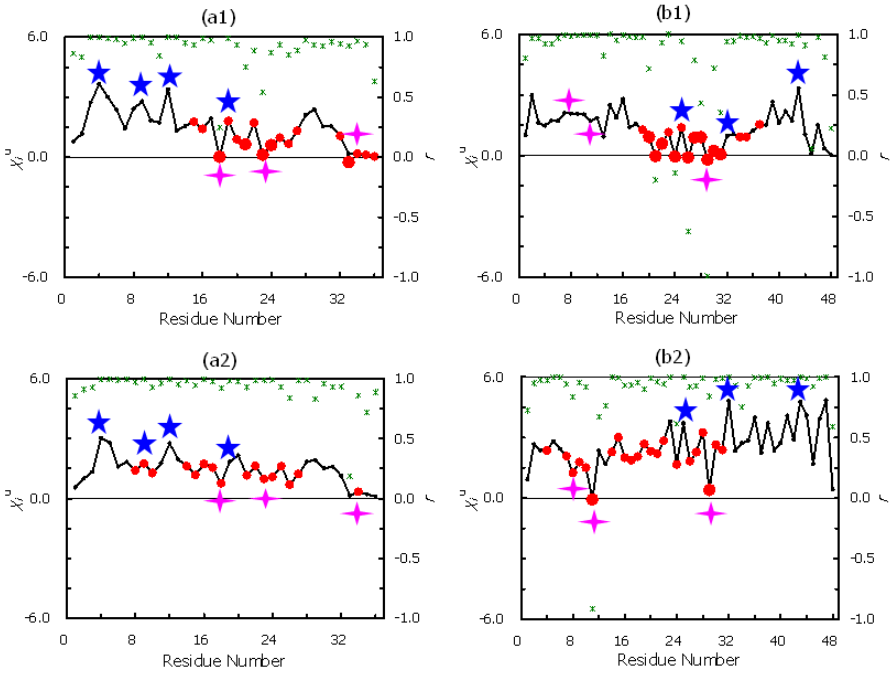

Figure 9. Susceptibility of unfolding rates to single amino acid substitution, χ_i^{u}, which is the slope obtained by the linear regression lines between $\Delta\ln k_{\mathrm{u}}(\xi_i)$ and $\Delta E(\xi_i)$ (see text in Sections 2.5 and 3.3). See also the caption of Figure 8.

Residues 18 and 23 are in the folding nucleus in both proteins *a1* and *a2*, and their χ_i^{f} are largely negative (residue 34 in proteins *a1* and *a2* is an exception, most likely because it is very close to the C-terminal). Residues 8, 11, and 29 in protein *b2* are in the folding nucleus and their χ_i^{f} are largely negative as well. While the χ_i^{f} of residue 29 in the folding nucleus in protein *b1* is also largely negative, the χ_i^{f} of residues 8 and 11, which are outside the folding nucleus, are small. These results clearly show that short-range interactions can accelerate the folding only when they are in the folding nucleus. The accelerating effect of the short-range interaction on folding was already discussed by Go and Taketomi for a 2D lattice protein [28].

Incidentally, comparing Figs. 7 and 8, the χ_i^{f} is negatively correlated with Φ_i. If $\Delta\ln k_{\mathrm{f}}(\xi_i) - \Delta\ln k_{\mathrm{u}}(\xi_i) \sim -\Delta\Delta F_{\mathrm{N\text{-}D}}$ is approximately equal to $-\Delta E(\xi_i)$; simply stated, if the change in the interaction energy dominantly affects $\Delta\Delta F_{\mathrm{N\text{-}D}}$ and the change in the entropy does little, $\Phi_i = \Delta\ln k_{\mathrm{f}}(\xi_i) / (\Delta\ln k_{\mathrm{f}}(\xi_i) - \Delta\ln k_{\mathrm{u}}(\xi_i)) \sim -\Delta\ln k_{\mathrm{f}}(\xi_i) / \Delta E(\xi_i) \sim -\chi_i^{\mathrm{f}}$. The large Φ values not only indicate that the residues are included in the folding nucleus but are also related to the folding rate in such a manner. If this is true (in reality, there are exceptions, particularly outside the folding nucleus), it is not surprising that the residues in the folding nucleus, which are assigned due to their large Φ values, have largely negative χ_i^{f}. On the basis of the study of a simple lattice protein, Ozkan et al. [29] insisted that the Φ value correlates with the acceleration/deceleration of folding induced by mutations, rather than with the degree of nativeness of the transition state.

4.2. Contact Order and Long-Range Interactions

It is interesting to note that the two proteins that are folded to the same conformation, *b1* and *b2*, have very different kinetic properties, because the folding rate is usually discussed in relation to the absolute and relative contact orders (ACO and RCO, respectively) of the native conformation [11, 30, 31] and more extensively to its backbone topology [32-34]. The contact order (CO) is defined as

$$\mathrm{CO} = \frac{1}{A} \sum_{i<j} \sum |i - j| \, \Gamma_{i,j} \tag{4.1}$$

where i and j are residue numbers; $\Gamma_{i,j} = 1$ if i and j are in contact in the native conformation, otherwise $\Gamma_{i,j} = 0$; the summation is performed over all residue pairs; A is the total number of contacts between amino acid residues for ACO and the total number of contacts between amino acid residues times the total number of amino acid residues for RCO, but we omitted this division in this paper, because it is identical in proteins *b1* and *b2*.

Although proteins *b1* and *b2* have an identical CO and backbone topology, there are differences in the amino acid sequences and consequently in the interaction energy values between the amino acid residues. From this viewpoint, we examined how the interaction energies between amino acid residues in the native conformation are distributed with respect to their distances along the chain. The distances between the amino acid residues along the chain, i.e., the distributions of short-, medium-, and long-range interactions, are considered a key point of the CO.

In Figure 10, the sum of the interaction energies between the amino acid residues whose distance along the chain is k, i.e., $\varepsilon_k = \sum_i U(\xi_i, \xi_{i+k}) \Gamma_{i,i+k}$, and their cumulative values $e_k = \sum_{i=1}^{k} \varepsilon_i$ are plotted against k, where $U(\xi_i, \xi_j)$ is the interaction energy between the amino acid residues ξ_i and ξ_j in contact (see Section 2.1). The cumulative number of native contacts $c_k = \sum_{i=1}^{k} \rho_i$ is also plotted for comparison, where $\rho_i = \sum_j \Gamma_{j,j+i}$ is the number of native contacts between the residues whose mutual distance along the chain is i (where the term "native contact" refers to an amino-acid-residue contact in the native conformation). The ε_k, e_k, and c_k values shown in Figure 10 are normalized by the total interaction energy $\varepsilon_{\mathrm{total}}$, the minimum value e_{n-1}, and the maximum value c_{n-1}, respectively.

The three curves of the cumulative relative interaction energies (two e_k values) and the cumulative relative number of native contacts (one c_k value) for proteins *a1* and *a2* are close to each other, as shown in Figure 10a. In detail, however, the short- and medium-range interactions for protein *a2* are slightly larger than those for protein *a1*. Our earlier work [9] showed that small local structures are formed at higher temperatures in protein *a2* but not in protein *a1*. In contrast, although e_k for protein *b1* is close to c_k, e_k for *b2* is considerably different from the two other curves, as shown in Figure 10b. Similarly, the short- and medium-range interactions for protein *b2* are slightly larger than those for protein *b1* (see Figure 10b). Furthermore, our earlier work [9] showed that the small local structures are

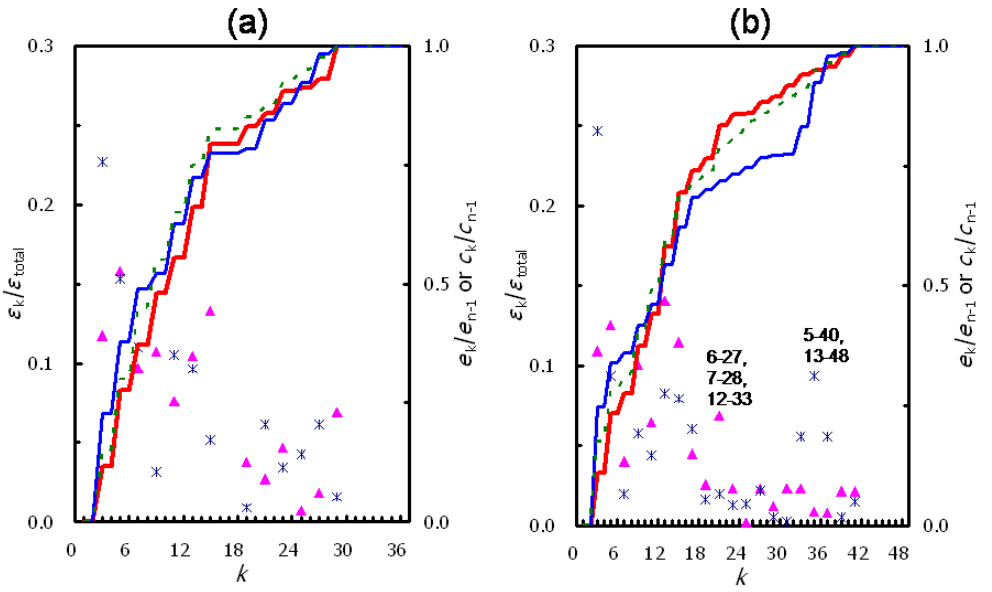

Figure 10. Interpretation of contact order. The interaction energy, ε_k, of the residue pairs with distance k along the chain ($k = 1, 2,\ldots, n-1$) in the native conformation is shown by a closed triangle and an asterisk for proteins *a1* and *a2* in (a), respectively, and for proteins *b1* and *b2* in (b), respectively. ε_k is normalized using the total interaction energy, ε_{total}. The cumulative relative interaction energy, e_k/e_{n-1}, is shown by the solid line: the red and blue lines in (a) are for proteins *a1* and *a2*, respectively, and those in (b) for proteins *b1* and *b2*, respectively, where $e_k = \sum_{i=1}^{k} \varepsilon_i$ and e_{n-1} is the minimum value of e_k. The cumulative relative frequency of the number of native contacts, c_k/c_{n-1}, is shown by dashed lines, where $c_k = \sum_{i=1}^{k} \rho_i$ and c_{n-1} is the maximum value of c_k (see text in Section 4.2 for details).

formed at higher temperatures in protein *b2* but not in protein *b1*. Significant differences occur in long-range interactions. In particular, the interactions of the protein *b1* residue pairs 6 and 27, 7 and 28, and 12 and 33, which are separated by 20 amino acids along the chain, are larger than those in protein *b2*. Conversely, the interactions in the protein *b2* residue pairs 5 and 40, and 13 and 48, which are separated by 34 amino acids, are considerably larger than those in protein *b1*. Simply put, the native structure of protein *b2* is stabilized only after the longer-range interactions between residues 5 and 40, and 13 and 48 are formed. This situation (an opposing effect of large entropy loss and large enthalpy gain by the formation of the longer-range interactions) is responsible for the considerably smaller folding and unfolding rates of protein *b2* (see Figure 6).

Essentially, the cumulative curves shown in Figs. 10a and b are related to the CO, considered to be one of the dominant factors in determining the folding rate. The following identity holds [11]:

$$\sum_{k=1}^{n-1} e_k = n e_{n-1} - \sum_{k=1}^{n-1} k\varepsilon_k \tag{4.2}$$

If the interaction energy is independent of the amino acid types, we can set $\varepsilon_k = \varepsilon_0 \rho_k$ (ε_0 is a constant) and eq. (4.2) can be written as

$$\sum_{k=1}^{n-1} c_k = n c_{n-1} - \sum_{k=1}^{n-1} k\rho_k \tag{4.3}$$

(Although in Figs. 10a and b, the abovementioned properties are normalized by their maximum values and this normalization is ignored). The left-hand side of eqs. (4.2) and (4.3), $\sum e_k$ and $\sum c_k$, correspond to the area of the lower-right region of the curves in Figs. 10a and b, and the second term on the right-hand side of eqs. (4.2) and (4.3), $\sum k\varepsilon_k$ and $\sum k\rho_k$, correspond to the area of the upper left region of the curves. Incidentally, the second term on the right-hand side of eq. (4.3) is a CO defined in eq. (4.1):

$$\sum_k k\rho_k = \sum_k \sum_i k\Gamma_{i,i+k} = \sum_{i<j} \sum |\, j-i \,|\, \Gamma_{i,j} \,. \tag{4.4}$$

Thus, $\sum k\varepsilon_k$ corresponding to $\sum k\rho_k$ may be referred to as an energy-weighted CO.

Accordingly, a larger CO results in a smaller left-hand side and smaller folding rate. Similarly, a smaller left-hand side in eq. (4.2) results in a smaller folding rate. This is the case for proteins *b1* and *b2*.

Although the CO is usually used for discussing the folding rates, this result indicates that the interaction energies defined in the second term of eq. (4.2) is a primary characteristic that contributes toward the folding rate. For real proteins, the folding rate is discussed in relation to the CO. However, as mentioned above, the folding rate should be examined in relation to the energy-weighted CO, $\sum k\varepsilon_k$, instead of the CO, $\sum k\rho_k$, defined using only the distance between the residues along the chain. Thus, a CO should be used for approximation.

4.3. Unfolding Rate and Long-Range Interactions

As shown in Figure 9, the unfolding rate is decelerated when stabilizing interactions are introduced to the residues outside the folding nucleus by a single amino acid substitution. Since such residues contribute to the native conformation by forming longer-range interactions, it is interesting to examine the role of the long-range interactions played in the unfolding. Again, the vertex residues referred to in Section 4.1 are useful to examine this feature. Out of these residues, residues 4, 9, 12, and 19 in proteins *a1* and *a2* and residues 25, 32, and 46 in proteins *b1* and *b2* form long-range interactions and have relatively large χ_i^{u}. The residues outside the folding nucleus, such as residues 4, 9, and 12 in proteins *a1* and *a2* and residue 46 in proteins *b1* and *b2*, are more remarkable. The long-range interactions of these residues reinforce the folding nucleus in the last stage of folding. Conversely, the unfolding has to start with the breakdown of these long-range interactions, followed by entry

into the transition state. As a result, the introduction of favorable interactions by an amino acid substitution at these sites decelerates the unfolding rate.

In proteins *a2* and *b2*, even the residues in the nucleus have relatively large χ_i^{u} values, whereas the χ_i^{u} values of such residues in proteins *a1* and *b1* are small. This is probably related to the fact that these residues have fractional Φ values. As discussed in Section 4.1, while the nuclei in proteins *a1* and *b1* are defined definitively, nuclei definitions for proteins *a2* and *b2* are assumed to be a mixture of various substructures. Here, it is further suggested that the long-range interactions work in a more complicated manner during the formation of a folding nucleus that consists of various substructures.

4.4. Fractional Φ Values

In this paper, we define a folding nucleus as a continuous region containing residues with relatively high Φ values. However, this definition is questioned by some researchers [7, 35, 36]; the term "folding nucleus" is a somehow subjective concept, which may not hold true for all proteins. In fact, high Φ values need not be associated with a nucleus, and low Φ values are found in the nuclei of real proteins. In addition, the interpretation of a fractional Φ value is rather difficult, because there can be a variety of reasons for their cause.

Table 1. Residue-residue interactions that significantly differ between proteins *a1* and *a2* and proteins *b1* and *b2* [a)]

Interactions more favorable in protein *a2* than *a1*			Interactions more favorable in protein *a1* than *a2*		
(14, 27)	13	–1.14	(2, 31)	29	0.70
(16, 27)	11	–0.72	(15, 24)	9	0.71
(3, 6)	3	–0.71	(9, 22)	13	0.58
(3, 30)	27	–0.71	(19, 32)	13	0.47
(6, 11)	5	–0.66	(2, 7)	5	0.45
(6, 27)	21	–0.59	(16, 31)	15	0.44
(27, 30)	3	–0.59	(17, 32)	15	0.43
(4, 29)	25	–0.57			
(13, 48)	35	–1.05	(20, 35)	15	0.56
(45, 48)	3	–1.05	(6, 27)	21	0.52
(2, 39)	37	–0.84	(22, 47)	25	0.50
(37, 40)	3	–0.72	(19, 34)	15	0.50
(16, 41)	25	–0.71	(36, 41)	5	0.47
(13, 16)	3	–0.66	(19, 28)	9	0.41
(20, 37)	17	–0.64	(19, 30)	11	0.37

a) The values given in this table are a residue pair (i, j), the number of residues between the residues along the chain, $j - i$, and an energy difference $E(a2) - E(a1)$ or $E(b2) - E(b1)$, respectively.

The fractional Φ value is usually considered an indication of either the partial formation of structure in a transition state or a mixture of substructures, some of which have interactions fully formed at the mutation site and others that have the interactions fully broken. Furthermore, the Φ value is considered monotonically related to how native-like the

transition-state conformation is or what percentage of the structure is fully formed. However, it is not quite that simple. Although Φ value analysis fundamentally assumes that a folding pathway and a relationship between structure and energy are not significantly altered by mutation, in some mutants, the statistical weights of individual conformations are considerably altered from the wild-type due to interaction changes between a substituted residue and other residues. As a result, the folding pathway and the relationship between structure and energy may be significantly altered in some cases.

To examine the fractional Φ value of the residues in the folding nucleus of the lattice proteins considered here, the interactions between residues significantly different between proteins *a1* and *a2* and between proteins *b1* and *b2* were compared and are shown in Table 1. In protein *a1*, the interactions between residues within the folding nucleus 15–36, such as 15 and 24, 19 and 32, 16 and 31, and 17 and 32 are more favorable than *a2*. Such favorable interactions in the nucleus enhance the stability of the nucleus, thus resulting in a higher Φ value for the residues in the nucleus of protein *a1*. In contrast, these interactions weaken in protein *a2*. Alternatively, the interaction of residue 27 with other residues, namely 14 and 16, are strengthened. As a result, the folding nucleus moves to residues 8–27. It should be noted that the interaction of residues within the N-terminal region, such as those between residues 3 and 6 and 6 and 11, as well as the interaction between N-terminal residues and residues distant along the chain, such as 3 and 30, 6 and 27, and 4 and 29, are more favorable in protein *a2* than *a1*. This suggests that these interactions are weaker in protein *a1*. These dispersed interactions stabilizing protein *a2* increase the relative possibilities for various conformations and are presumably the reason why residues in the folding nucleus of protein *a2* have fractional Φ values.

Similarly, highly favorable interactions such as those between residues 20 and 35, 19 and 34, 19 and 28, and 19 and 30 are found within the folding nucleus encompassing residues 19–37, and enhance the stability of the nucleus in protein *b1*. In contrast, instead of being concentrated in the folding nucleus, favorable interactions are dispersed in protein *b2*. Some interactions are short-range, like those between residues 45 and 48, 37 and 40, and 13 and 16, and other interactions long-range, like those between residues 13 and 48, 2 and 39, and 16 and 41. Similar to protein *a2*, the dispersed favorable interactions increase the relative possibilities for various conformations, thus the residues in the protein *b2* folding nucleus have fractional Φ values.

4.5. Negative Φ Values

Finally, we discuss irregular or non-classical Φ values. As shown in Figure 7, the Φ values of some residues are negative. In addition, for some specific single amino acid substitutions, the Φ values become greater than unity (data not shown). Although the Φ value is usually assumed to be defined between 0 and 1, such irregular cases are possible theoretically, and in fact, it has been pointed out that 10–20% of reported Φ values for real proteins are out of the normal range [29]. Therefore, it is interesting to examine the cases with irregular negative Φ values in the lattice proteins considered here.

The Φ value is negative if $\Delta\ln k_f > 0$ and $\Delta\ln k_f - \Delta\ln k_u < 0$, or $\Delta\ln k_f < 0$ and $\Delta\ln k_f - \Delta\ln k_u > 0$, according to eq. (2.23). In either case, $|\Delta\ln k_f| < |\Delta\ln k_u|$. This means that the relationship between the native and transition states is more greatly altered than that between

the denatured and transition states. Some examples are shown in Figure 11. Typically, there are three cases:

i) The change in the free energy profile is restricted around the transition- and native-state regions. Nothing changes in the denatured-state region (see the V5L and V5C mutants of protein *a1* in Figure 11 as an example). If $\Delta\ln k_f$ is a small negative value, the Φ value is close to zero and this case may be accepted as a normal value within the scope of assumption of the Φ value. The V5L mutant is an example of this: $\Delta\ln k_f = -0.01125$, $\Delta\ln k_u = -0.45058$, and $\Phi = -0.026$. Another case is when $\Delta\ln k_f$ is significantly large and Φ value is largely negative. The V5C mutant is an example of this: $\Delta\ln k_f = 0.41378$, $\Delta\ln k_u = 2.0995$, and $\Phi = -0.245$. In these cases, since the amino acid substitution has a larger effect on the native state than the transition state, the residue at the substituted site is not included in the folding nucleus and contributes to protein stability by its long-range interactions with residues in the nucleus. This interaction works after nucleation.

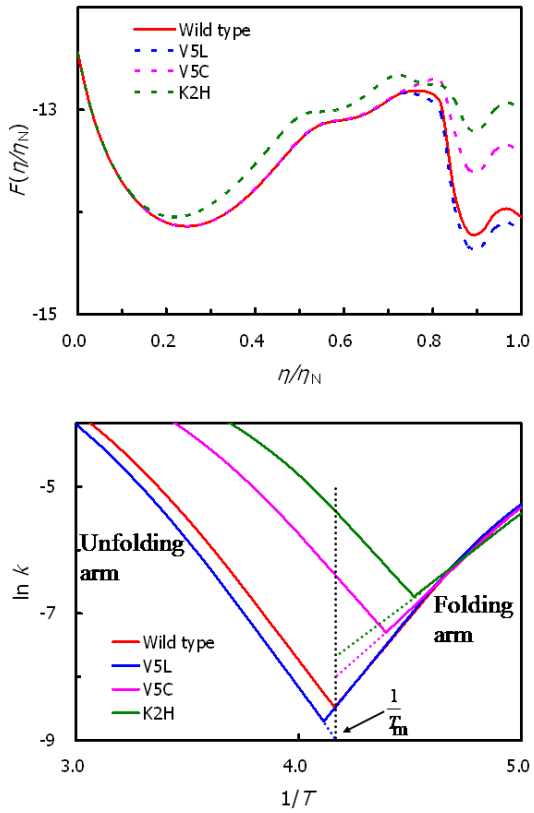

Figure 11. Examples of changes in free energy profiles and chevron plots by single amino acid substitutions. The free energy profiles (upper panel) and chevron plots (lower panel) are shown for three mutants of protein *a1*, V5L, V5C, and K2H, together with those for the wild-type. Since $\Delta\ln k_f$ is calculated by linearly extrapolating the folding arm to the transition temperature of the wild-type protein, it is necessary to pay attention to a change in the slope of the folding arm.

ii) The free energy profile varies in entire regions from denatured to native states (see the K2H mutant of protein *a1* in Figure 11 as an example; the R4L mutant of protein *b1* is also an example; however, its data are not shown). The Φ value analysis usually assumes that the free energy level of the denatured state is not affected by amino acid substitutions; however, this is not always the case. It is possible to generate extensive changes in the relative probabilities of conformations, thus altering the free energy profile entirely. In the protein *a1* K2H mutant and protein *b1* R4L mutant, most of the conformations, in particular the native conformations, are destabilized compared to the wild-types (K2H: $\Delta\ln k_f = 0.772$, $\Delta\ln k_u = 3.110$, and $\Phi = -0.330$; and R4L: $\Delta\ln k_f = 1.199$, $\Delta\ln k_u = 2.919$, and $\Phi = -0.697$).

iii) There is a little or no change in the free energy profile (the F48L mutant of protein *b1* is an example; its free energy profile is not shown here, but it is almost indistinguishable from that of the wild-type). In this case, the residue at the substituted site has a little or no contribution to folding/unfolding. However, when both the numerator and denominator in eq. (2.23) are very small, the Φ value sometimes becomes not only negative but also irregularly largely positive (F48L: $\Delta\ln k_f = -0.05683$, $\Delta\ln k_u = -0.05247$, and $\Phi = 13.0$). In this case, the Φ value may not have a significant meaning.

Ozkan et al. [29] showed that from a simple lattice protein, a negative Φ value results when a mutation destabilizes a slow flow channel to the native conformation, causing a backflow into a faster flow channel in the energy landscape. In such a case, destabilizing mutations can accelerate folding. However, because the folding kinetics are considered on the coarse-grained one-dimensional free energy profile in our model, we cannot discuss such parallel microscopic flow processes.

5. CONCLUSION

In this paper, we studied the folding/unfolding kinetics of proteins by applying a simple statistical mechanical model to lattice proteins. We learned the following aspects through the study:

(1) The folding nucleus is a compact substructure in the native conformation and defined as the segment of residues with higher Φ values. However, for some proteins, the folding nucleus is an ambiguous concept, because various substructures are assumed to compose the folding nucleus, rather than a unique compact structure. As shown in the comparative analysis of the two proteins with identical native structures but different amino acid sequences, the folding nucleus depends not only on the native structure but also on the amino acid sequence. This suggests that geometrical compactness is not the only requirement.

(2) The stabilizing short-range interactions within a folding nucleus increase folding rate but affect the unfolding rate less.

(3) The stabilizing long-range interaction of a residue outside the folding nucleus with a residue within a folding nucleus contributes a little to the formation of the folding

nucleus and consequently has little effect on the folding rate but significantly affects the unfolding rate.

(4) The folding rate should be correlated primarily with the energy-weighted CO rather than the one defined using only the distance between residues along the chain.

(5) The change in some of the interactions of a residue outside a folding nucleus due to a single amino acid substitution can change the free energy profile entirely. In this case, kinetic characteristics such as the folding/unfolding rates and Φ value are significantly affected.

(6) The stabilizing interactions are concentrated within a folding nucleus in some proteins and dispersed across the protein in others. The fractional Φ values in a folding nucleus are found in the latter case.

(7) The negative Φ values result from several reasons. Typically, there are three cases: (i) the free energy profile of a mutant varies only around the transition- and native-state regions, (ii) it varies in entire regions, and (iii) it varies very little. In any of these cases, the unfolding rate is affected more largely than the folding rate.

It is possible for a single residue to have both short- and long-range interactions with other residues. In addition, interactions are formed within a folding nucleus, outside a folding nucleus, and between residues within and outside the folding nucleus. Because of the many-body nature of these interactions, the above summary is clearly satisfied in some cases, but realized in a rather complicated manner in others.

It is inevitable that these results may involve some artifacts caused by the simple statistical mechanical model and lattice proteins. Nonetheless, we believe that these results present important aspects to studying the folding/unfolding kinetics of real proteins and that we can learn something significant through them.

REFERENCES

[1] Takano, K; Ota, M; Ogasahara, K; Yamagata, Y; Nishikawa, K; Yutani, K. *Protein Eng*., 1999, 12, 663-672.
[2] Matthew, BW. *Adv. Protein Chem*., 1995, 46, 249-278.
[3] Matouschek, A; Kellis, JT; Serrano, L; Fersht, AR. *Nature*., 1989, 340, 122-126.
[4] Fersht, AR; Matouschek, A; Serrano, L. *J. Mol. Biol*., 1992, 224, 771-782.
[5] Northey, JGB; Di Nardo, AA; Davidson, AR. *Nat. Struct. Biol*., 2002, 9, 126-130.
[6] Sánchez, IE; Kiefhaber, T. *J. Mol. Biol*., 2003, 327, 867-884.
[7] Zarrine-Afsar, A; Davidson, AR. *Methods*., 2004, 34, 41-50.
[8] Abe, H; Wako, H. *J. Phys. Soc. Jpn*., 2004, 73, 1143-1146.
[9] Abe, H; Wako, H. *Phys. Rev. E*., 2006, 74, 011913.
[10] Wako, H; Abe, H. *J. Phys. Soc. Jpn*., 2007, 76, 104801.
[11] Abe, H; Wako, H. *Physica A*., 2009, 383, 3442-3454.
[12] Mirny, LA; Abkevich, VI; Shakhnovich, EI. *Folding Des*., 1996, 1, 103-116.
[13] Abkevich, VI; Gutin, AM; Shakhnovich, EI. *Folding Des*., 1996, 1, 221-230.
[14] Miyazawa, S; Jernigan, RL. *Macromolecules*., 1985, 18, 534-552.
[15] Wako, H; Saitô, N. *J. Phys. Soc. Jpn*., 1978, 44, 1931-1938.
[16] Wako, H; Saitô, N. *J. Phys. Soc. Jpn*., 1978, 44, 1939-1945.

[17] Gō, N; Abe, H. *Biopolymers*., 1981, 20, 991-1011.
[18] Abe, H; Gō, N. *Biopolymers*., 1981, 20, 1013-1031.
[19] Miyazawa, S; Jernigan, RL. *J. Stat. Phys*., 1983, 30, 549-559.
[20] Segawa, S; Kawai, T. *Biopolymers*., 1986, 25, 1815-1835.
[21] Muñoz, V; Henry, ER; Hofrichter, J; Eaton, WA. *Proc. Natl. Acad. Sci*., USA. 1998, 95, 5827-5879.
[22] Muñoz, V; Eaton, WA. *Proc. Natl. Acad. Sci*., USA. 1999, 96, 11311-11316.
[23] Henry, ER; Eaton, WA. *Chem. Phys*., 2004, 307, 163-185.
[24] Itoh, K; Sasai, M. Proc. *Natl. Acad. Sci*., USA. 2004, 101, 14736-14741.
[25] Zamparo, M; Pelizzola, A. *Phys. Rev. Lett*., 2006, 97, 068106.
[26] Imparato, A; Pelizzola, A; Zamparo, M. *Phys. Rev. Lett*., 2007, 98, 148102.
[27] Finkelstein, AV; Ptitsyn, OB. *Protein Physics*; *Academic Press: Amsterdam*, 2002, 251-262.
[28] Gō, N; Taketomi H. *Int. J. Peptide Protein Res*., 1979, 13, 235-252.
[29] Ozkan, SB; Bahar, I; Dill, KA. *Nat. Struct. Biol*., 2000, 8, 765-769.
[30] Plaxco, KW; Simons, KT; Baker, D. *J. Mol. Biol*., 1998, 277, 985-994.
[31] Plaxco, KW; Simons, KT; Ruczinski, I; Baker, D. *Biochemistry*., 2000, 39, 11177-11183.
[32] Gromiha, MM; Selvaraj, S. *J. Mol. Biol*., 2001, 310, 27-32.
[33] Kamagata, K; Arai, M; Kuwajima, K. *J. Mol. Biol*., 2004, 339, 951-965.
[34] Kuznetsov, IB; Rackovsky, S. *Proteins*., 2004, 54, 333-341.
[35] Fersht, AR. *Curr. Opin. Str. Biol*., 1994, 5, 79-84.
[36] Hubner, IA; Shimada, J; Shakhnovich, EI. *J. Mol. Biol*., 2004, 336, 745-761.