Special Issue
"Memorial Issue for Prof. Nobuhiko Saitô"

Vol. 13, pp. 263–279 (2016)
doi: 10.2142/biophysico.13.0_263

Biophysics and Physicobiology

https://www.jstage.jst.go.jp/browse/biophysico/

Regular Article

Characterization of protein folding by a Φ-value calculation with a statistical-mechanical model

Hiroshi Wako[1] and Haruo Abe[2]

[1]*School of Social Sciences, Waseda University, Shinjuku, Tokyo 169-8050, Japan*
[2]*Department of Electrical Engineering, Nishinippon Institute of Technology, Miyako, Fukuoka 800-0394, Japan*

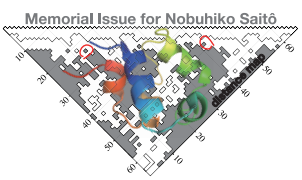
Received August 1, 2016; accepted September 20, 2016

The Φ-value analysis approach provides information about transition-state structures along the folding pathway of a protein by measuring the effects of an amino acid mutation on folding kinetics. Here we compared the theoretically calculated Φ values of 27 proteins with their experimentally observed Φ values; the theoretical values were calculated using a simple statistical-mechanical model of protein folding. The theoretically calculated Φ values reflected the corresponding experimentally observed Φ values with reasonable accuracy for many of the proteins, but not for all. The correlation between the theoretically calculated and experimentally observed Φ values strongly depends on whether the protein-folding mechanism assumed in the model holds true in real proteins. In other words, the correlation coefficient can be expected to illuminate the folding mechanisms of proteins, providing the answer to the question of which model more accurately describes protein folding: the framework model or the nucleation-condensation model. In addition, we tried to characterize protein folding with respect to various properties of each protein apart from the size and fold class, such as the free-energy profile, contact-order profile, and sensitivity to the parameters used in the Φ-value calculation. The results showed that any one of these properties alone was not enough to explain protein folding, although each one played a significant role in it. We have confirmed the importance of characterizing protein folding from various perspectives. Our findings have also highlighted that protein folding is highly variable and unique across different proteins, and this should be considered while pursuing a unified theory of protein folding.

Key words: free-energy profile, contact-order profile, circular permutation, framework model, nucleation-condensation mechanism

Understanding the problem of protein folding poses a great challenge, particularly from a biophysics perspective. The protein-folding problem involves three major issues: (1) how the three-dimensional (3D) structure of a protein is determined from its amino acid sequence, (2) how a protein folds from a random coil to a native structure, and (3) how the 3D structure is characterized from static and dynamic perspectives. The second problem, which concerns the fold-

Corresponding author: Hiroshi Wako, School of Social Sciences, Waseda University, 1-6-1 Nishi-Waseda, Shinjuku, Tokyo 169-8050, Japan.
e-mail: wako@waseda.jp



◀ ***Significance*** ▶

The Φ values were calculated for 27 proteins with a simple statistical-mechanical model of protein folding. We considered that the correlation between the calculated and experimentally observed Φ values should illuminate the protein-folding mechanism active for each protein, such as the framework model or the nucleation-condensation model. In addition, we investigated other properties associated with protein folding such as the free-energy profile, contact-order profile, and sensitivity to the parameters used in the Φ-value calculation of each protein. Through such analyses, we tried to characterize protein folding from various perspectives to derive a unified theory of protein folding.

ing process (or folding pathway) of a protein, is the focus of this paper.

Regarding the folding process, there are currently two major hypotheses: the three-step mechanism or framework model [1–4] and the nucleation-condensation model [5,6]. In the framework model, secondary-structure elements fold first, followed by the coalescence of preformed secondary-structure elements to yield the native structure. The occurrence of step-by-step folding from short- to medium- and then to long-range interactions is a key assumption. In contrast to this, the nucleation-condensation model assumes that the secondary and tertiary structures are formed in parallel, or even in reverse order. That is, the long-range interactions or tertiary structures can be formed before or at the same time as the secondary structures. Interpretations of these two hypotheses seem to differ among researchers. For convenience and clarity, in this paper we refer to these two hypotheses but with the following specific preconditions: the framework model does not allow longer-range interactions to form before shorter-range interactions, but the nucleation-condensation model does.

The above key principles have contributed significantly to the conceptual basis of protein folding. However, there is currently a critical limitation: the difficulty of quantitatively connecting theoretical predictions to experimental results [7]. Researchers have not been able to use an existing theory to consistently interpret their experimental results. Progress has been made towards eliminating this limitation by using improved simulations based on molecular dynamics and analytical calculations that use simple statistical-mechanical models. Since computer simulations of protein folding are time-consuming and can be performed only for one specific protein at a time, simple statistical-mechanical models of protein folding—which permit the direct analysis and fitting of experimental data—have been desired.

An Ising-like statistical-mechanical model of protein folding was first introduced by Wako & Saitô in 1978 [8,9]. This model was developed further by Gō & Abe [10], and Abe & Gō [11] demonstrated that it could accurately reproduce the results of computer simulations of two-dimensional lattice-protein folding. However, it was some time before a similar model could be applied to real proteins. This was first accomplished by Muñoz & Eaton in 1999 [12], who demonstrated that theoretically calculated folding rates could correlate well with experimentally observed ones.

In this paper, the Φ values for every residue of a specific protein calculated using a simple statistical-mechanical model are discussed and compared with the corresponding experimentally observed Φ values. Despite the simplicity of the statistical-mechanical model, the experimental Φ values were reasonably reproduced for many proteins, but not for others. Whereas the successful cases supported the assumptions of the model, the failures revealed the limitations of those assumptions. We will discuss protein-folding mechanisms while examining specific examples where the model succeeded and where it failed.

Φ-value analysis was introduced by Fersht and his colleagues [13,14] to provide information regarding the transition-state structures along the folding pathway of a protein by measuring the effects of amino acid mutations on folding kinetics. The Φ value is calculated as $\Delta\Delta G_{‡\text{-}D}/\Delta\Delta G_{N\text{-}D}$ for a two-state-folding protein, where $\Delta G_{‡\text{-}D}$ and $\Delta G_{N\text{-}D}$ are the free energy (FE) differences between the unfolded and transition states and between the unfolded and folded states, respectively, and ΔΔ denotes the changes in these FE differences brought about by a point mutation at a specific amino acid residue. A Φ value of 1 indicates that all of the mutated residue interactions are formed in the transition state whereas a Φ value of 0 means that the residue is not involved in stabilizing the transition state. Intermediate Φ values indicate that the interactions are partially formed or that there are two populations with mostly unfolded and mostly folded states. However, because the relationship between the actual Φ value and the extent of protein structural formation is not necessarily linear, the interpretation of intermediate Φ values is still controversial [15].

Φ values have been investigated experimentally for many two-state-folding proteins (see Table 1). Φ values obtained in such experiments have provided valuable information for refining protein-folding models such as the nucleation-condensation model and the framework model [16].

Φ values have also been studied from a theoretical perspective. Since folding simulations can provide information about the folding pathway of the protein, the transition-state structures in these simulations are associated with the experimentally observed Φ values [17–19]. More directly, Φ values have been calculated based on the FE profiles obtained from statistical-mechanical models of protein folding [12,20–27]. Although various models have been proposed, their concepts are essentially similar and assume that each residue is in one of two states: native (folded) or random-coil (unfolded). It is also assumed that one or more contiguous segments in the polypeptide chain consists of residues in the native state, and that only individual amino acids in a pair within a native-structure polypeptide segment and in contact in the native structure interact with each other (this is the so-called Gō model). Zero interactions are assumed to occur in segments consisting of residues still in the random-coil state. In these studies, the calculated Φ values generally showed a reasonable agreement with the experimental Φ values.

In this paper, we report the results of Φ-value calculations made using a previously developed, alternative, simple statistical-mechanical model of protein folding [28–30]. Although this model was developed earlier than those mentioned above, various studies have recently adopted this model or others that are similar but extended, to study the protein-folding problem [25,27,31–38]. We also determined the Φ values for lattice proteins using this alternative model with particular focus on the dependence of lattice proteins on amino acid sequence and folding topology [28,29]. In

line with these studies, we have calculated the Φ values of real proteins in the present study. For these calculations, we selected 27 proteins (each with 54 to 128 amino acid residues in length) with known experimentally derived Φ values.

In the next section, we briefly describe a statistical-mechanical model of protein folding that we developed. The calculated Φ values are shown and discussed in Results and Discussion along with the FE profiles, contact-order (CO) profiles, and sensitivities to the parameters used in the Φ-value calculation. In this study, we investigated protein folding from various perspectives rather than assessing the specific statistical-mechanical model used here, which we have already done in a previous paper [30].

Materials and Methods

Statistical-mechanical model of protein folding

The simple statistical-mechanical model of protein folding and unfolding used in this study was constructed with the intention of introducing the following picture of protein folding [8–11,28–30]. A protein folds in a stepwise manner along the polypeptide chain; this assumption conforms to the framework model. In this scheme, the first stage of folding includes short-range interactions that work dominantly to form small, native-like, or secondary structures such as the α-helix, β-strand, and hydrogen-bonded turn. Next, these structures grow gradually through medium-range interactions. Finally, these substructures coalesce into the native structure via long-range interactions.

It should be noted that our model is not exactly the same as the framework model. Whereas the framework model usually assumes a highly defined folding pathway in which specific local structures (e.g., secondary structures, super-secondary structures, domains) are formed in a hierarchical way, our model is more probabilistic, i.e., any local structures can form in any order according to their statistical weights, as described below. When we say that our model is based on the framework model, we refer to the important feature common to both the framework model and our own: longer-range interactions are not allowed to form without forming short-range interactions first.

The statistical-mechanical model of the folding and unfolding of a protein consisting of n residues in the above-mentioned picture is formulated as follows [28–30] (see Supplementary Text S1 for details).

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

(ii) The conformation of a protein at any stage in the folding process is represented by a sequence of two types of regions of various sizes, namely a "local structure" (or "island" in Wako & Saitô [8,9]) and a random-coil region, arranged alternately along the chain. The term "local structure" is used with a specific meaning in this model. A local structure and a random-coil region are defined as continuous regions in which all amino acid residues are in the native state and random-coil state, respectively. It should be noted that the number of local structures is not restricted to one in this model; this is important because some models have restricted the number of local structures formed to only one.

(iii) The key assumption of this statistical-mechanical model is that only Gō-type native interactions between amino acid residues within a local structure are considered [39]. The other interactions, such as those between the residues in different local structures and those within a random-coil region, are neglected.

(iv) For the FE within a random-coil region (where no interaction between the residues exists), it is assumed that only the chain entropy, which depends on the number of residues, contributes to the partition function. The random-coil state is the reference state; i.e., its statistical weight is set to unity.

(v) The minimum size of a local structure is four consecutive residues. An amino acid residue i is considered to be in the native state if a region of four consecutive amino acid residues, $(i-1)$ to $(i+2)$, adopts the same conformation as the native structure ($i=2, 3, ..., n-2$). Otherwise, the amino acid residue is considered to be in the random-coil state. In other words, it is assumed that no interactions occur between amino acid residue i and amino acid residues $i\pm1$ and $i\pm2$.

The partition function for this statistical-mechanical model is given by the following recurrent relationship:

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

$$Z_{1,1} = 0,\ Z_{1,2} = Z_{1,3} = 1\,, \qquad (1)$$

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

The conformational energy of a local structure consisting of amino acid residues m to j is given as

$$E(m,j) = \sum_{m \le k,l \le j} U(\xi_k, \xi_l)\Gamma_{k,l}\,, \qquad (2)$$

where $U(\xi_k, \xi_l)$ is the interaction energy between amino acid residues ξ_k and ξ_l. The contact matrix $\Gamma_{k,l}=1$ if amino acid residues k and l are in contact with each other in the native conformation, and $\Gamma_{k,l}=0$ otherwise. Because the interactions between neighboring residues and residues separated by one residue are not considered, $\Gamma_{k,l}=0$ for $|k-l|\le 2$. $U(\xi_k, \xi_l)$ depends on amino acid types ξ_k and ξ_l. $\Gamma_{k,l}$, i.e., whether the residue pair is in contact or not 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.

Function $f(m, j)$ in Eq. (1) corresponds to the number of

possible conformations of a segment between the amino acid residues m and j in the random-coil state; thus, $k_B \ln f(m, j)$ is the chain entropy of the segment in the random-coil state, and $-k_B \ln f(m, j)$ is the entropy loss of the segment when it forms the local structure.

Following our previous paper [30], we assigned a single value $\varepsilon<0$ to $U(\xi_k, \xi_l)$, independent of the amino acid type, ξ_k or ξ_l (homogeneous contact-energy approximation). We set $\varepsilon=-0.10$ for all calculations in this study. For $f(m, j)$, because we did not have any established values representing a real protein, we tentatively used the same formulation values that had been used for the lattice protein in the previous paper, i.e., $f(m, j)=1.4084\times(4.750)^{j-m-2}$ or $\ln f(m, j)=(j-m-2)\ B+0.3425$, where $B=\ln 4.750=1.5581$. In addition, we examined another B value, i.e., $B=1.3$ for comparison in this study.

A contact matrix $\Gamma_{k,l}$ is defined as follows (as in the previous paper [30]): if a distance of at least one atom pair in two residues is shorter than a given cutoff distance D_c, this residue pair is considered to be in contact. We performed Φ-value calculations with four D_c values, 4.0, 5.0, 5.5, and 6.0 Å, in our previous study; and in that study, we found that the results were sensitive to the D_c value for some proteins but not for others. In the current study, however, two cases, $D_c=4.2$ and $D_c=5.5$ Å, were examined to study D_c dependence in the Φ-value calculations.

Free-energy profile

The energy (enthalpy) of the system, E_h, is expressed by the integer h in units of –0.01, i.e., $E_h=h\varepsilon_0$ and $\varepsilon_0=-0.01$, for computational convenience. Eventually, the partition function Z is given as a polynomial with two variables, t and u, as a function of the temperature T [29,30,40].

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

where

$$W(\eta, T) = \sum_{h}\Omega(\eta, h)u^{h}, \tag{4}$$

$$u = \exp(-\beta\varepsilon_0), \tag{5}$$

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 the recurrent equation, Eq. (1). $W(\eta, T)t^{\eta}$ is the sum of the statistical weights over all states with the given number of amino acid residues in the native state η and at a temperature T.

We define the FE for a given η from Eq. (4):

$$F(\eta, T) = -k_B T \ln W(\eta, T). \tag{6}$$

Eq. (6) is used to calculate the FE profiles of the proteins with η as the folding reaction coordinate. The enthalpy and entropy at a given temperature T can also be calculated (see Supplementary Text S1), but they are not discussed in this paper.

Folding and unfolding rates

The kinetics of the folding and unfolding processes of proteins (such as their folding and unfolding rates) were formulated as a motion along a one-dimensional FE profile, with the number of amino acid residues in the native state as a reaction coordinate, as established by Muñoz & Eaton [12] and extensively examined by Henry & Eaton [41]. Because we had the FE profile $F(\eta, T)$ of Eq. (6), we applied the method of Muñoz & Eaton. According to this method, using a simple approach that involved solving a system of differential equations describing reversible hopping between adjacent discrete values of reaction coordinates (η and $\eta+1$ in this study), the characteristic relaxation rate can be given as follows:

$$\begin{aligned}\frac{1}{k} \propto \tau &\equiv \int_0^{\infty} \frac{\langle\eta\rangle_{eq} - \langle\eta(t)\rangle}{\langle\eta\rangle_{eq} - \langle\eta(0)\rangle} dt \\ &= \{\langle\eta\rangle_{eq} - \langle\eta\rangle_0\}^{-1} \sum_{j=0}^{n-4} \frac{1}{p_{eq}(j)s_{j,j+1}} \\ &\quad \sum_{i=j+1}^{n-3} \{p_{eq}(i) - p_0(i)\} \sum_{\eta=0}^{j} p_{eq}(\eta)(\langle\eta\rangle_{eq} - \eta).\end{aligned} \tag{7}$$

Here, an equilibrium value of η (at temperature T),

$$\langle\eta\rangle_{eq} = \sum_{\eta=0}^{n-3} \eta p_{eq}(\eta) \tag{8}$$

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

$$p_{eq}(\eta) = Z^{-1} \exp\{-F(\eta, T)/k_B T\}. \tag{9}$$

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

The relaxation rate k is estimated as the mean rate of relaxation of the average number of native residues to its equilibrium value from a starting point of two initial conditions: first with the entire population in the completely unfolded state, i.e., $\eta=0$, and second with the entire population in the native state, i.e., $\eta=n-3$. In this paper, the former relaxation rate k above $1/T_m$ and the latter k below $1/T_m$ are regarded as the folding and unfolding rates, k_f and k_u, respectively. A transition temperature, T_m, is defined as $\ln k_f(T_m)=\ln k_u(T_m)$.

Φ value

We considered a single amino acid substitution for each residue of a given protein. For a single amino acid substitution at the kth residue (for example, if amino acid residue ξ_k in the wild-type protein is replaced by amino acid w_k), we simply assumed that $U(\xi_k, \xi_l)$ is transformed to $U(w_k, \xi_l)$ in

Eq. (2). Thus, the Φ values of the kth residue were calculated for the two cases: one where $U(w_k, \xi_l)=\varepsilon+0.01$ and the other where $U(w_k, \xi_l)=\varepsilon-0.01$, hereafter referred to as the "plus perturbation" and the "minus perturbation," respectively (recall that $U(\xi_k, \xi_l)=\varepsilon$ for each residue pair in the wild-type protein). The responses to the substitutions were examined based on changes in the logarithmic folding and unfolding rates, $\ln k_f$ and $\ln k_u$, respectively, determined in this study. The corresponding rate changes, $\Delta\ln k_f$ and $\Delta\ln k_u$, were used to calculate the Φ value defined by Eq. (10):

$$\Phi = \frac{\Delta\Delta F_{\ddagger-D}}{\Delta\Delta F_{N-D}} = \frac{\Delta \ln k_f}{\Delta \ln k_f - \Delta \ln k_u}, \quad (10)$$

where $\Delta\ln k_f = \ln k_f^{mut}(T_m) - \ln k_f^{wild}(T_m)$ and $\Delta\ln k_u = \ln k_u^{mut}(T_m) - \ln k_u^{wild}(T_m)$, because $\ln k_f^{wild}(T_m) = \ln k_u^{wild}(T_m)$ and $\Delta\ln k_f - \Delta\ln k_u = \ln k_f^{mut}(T_m) - \ln k_u^{mut}(T_m)$. The superscripts "wild" and "mut" denote the wild-type and mutated proteins, respectively.

From the two Φ values calculated by the plus and minus perturbations, a Φ value for a specific residue is defined as follows. (i) A Φ value greater than 1.3 or smaller than –0.3 is considered "irregular" and is then regarded as having no defined Φ value. (ii) If $1.0<\Phi<1.3$, then Φ is set to 1.0, and if $-0.3<\Phi<0.0$, then Φ is set to 0.0, because the Φ value is essentially defined as $0\le\Phi\le1$. Furthermore, (i) and (ii) are applied separately to the Φ values obtained in the plus and minus perturbations. (iii) The two Φ values are averaged. If only one Φ value is available, then it is defined as a Φ value of this residue. If neither of them is available, then the Φ value of this residue is classified as "not defined."

Contact-order profile

To investigate the calculated Φ values, we generated a CO profile—a concept introduced in the previous paper [28]. CO has been used to characterize the complexity of the folding topology of a protein in relation to its folding kinetics [42–46]. In line with this, the CO profile is defined as the cumulative number of native contacts $c_k=\sum_{i=1}^{k}\rho_i$ plotted against k, where ρ_i is the number of native contacts between two residues whose mutual distance along the polypeptide chain is i. The following relationship holds [28]:

$$\sum_{k=1}^{n-1} c_k = nc_{n-1} - \sum_{k=1}^{n-1} k\rho_k. \quad (11)$$

The second term on the right-hand side, $\Sigma k\rho_k$, corresponds to the area of the upper left region of the k vs. c_k curve and is equal to the CO of a given protein (see Fig. 4 below). The CO profile is more informative than the CO, as described below.

Proteins examined

Using the simple statistical-mechanical model, Φ values were calculated for 27 proteins for which Φ-value analysis had already been performed experimentally. The Protein Data Bank (PDB) entry codes, sizes, and fold classes of the proteins that we investigated are listed in Table 1. This protein set includes some proteins that belong to the same superfamily, such as 1SRM, 1SHG, and 1FYN, all of which are in the SH3 domain-containing superfamily. However, since the available data were limited, we did not omit data from homologous proteins; some would consider the data from such proteins to be redundant, but from a practical standpoint, their omission would make performing a statistical analysis far more difficult. Furthermore, comparisons between homologous proteins were expected to actually provide more useful information about the nature of protein folding. Accordingly, we used all the available data, disregarding perceptions of protein redundancy, in the following analyses. It is necessary, however, to remember this point in the statistical discussion below.

For each protein studied, the PDB entry code given in Table 1 is used as the protein name hereinafter, for convenience.

Results and Discussion

In this study, we were interested in characterizing protein folding from various perspectives. As is well known, protein size and fold class are primary characteristics of a protein that determine its folding dynamics. We can also characterize a protein by its Φ-value profile, FE profile, and CO profile. In addition, a Φ-value's sensitivity to the parameters used in the Φ-value calculation can be examined. The Φ-value calculation was performed for four sets of parameters, hereinafter referred to as D42_13, D42_16, D55_13, and D55_16, each of which denote parameter values (D_c, B) = (4.2, 1.3), (4.2, 1.5581), (5.5, 1.3), and (5.5, 1.5581), respectively (see Materials and Methods). The following data pairs were used as described here: (D42_13, D42_16) and (D55_13, D55_16) were used to investigate the sensitivity to the chain entropy parameter B, and (D42_13, D55_13) and (D42_16, D55_16) were used to investigate the sensitivity to the cutoff distance D_c.

The results are summarized in Figure 1. We will discuss them below.

Φ-value profile

Φ-values are one of the few types of experimental data that can both A) give us information on the protein folding process, and B) be directly compared with theoretically calculated properties. In Figure 2, the Φ-value profiles calculated for every residue using the abovementioned statistical-mechanical model are shown with the corresponding experimentally observed Φ values for four example proteins: 1TEN, 1SHG, 3CI2, and 1AYE. Experimentally observed Φ values were available for some, but not all, of the residues. The Φ-value profiles for the other proteins are given in Supplementary Figure S1.

Table 1 Proteins examined in this study

Protein name[a)]	PDB code[b)]	No. of residues[c)]	Folding class[d)]	Reference[e)]
Engrailed homeodomain	1ENH	54	All α	[47]
c-Myb-transforming protein	1IDY	54	All α	[47]
IgG-binding domain of protein G	1PGB	56	α+β	[48]
Src SH3 domain	1SRM	56	All β	[49]
α-spectrin SH3 domain	1SHG	57	All β	[50]
Fyn SH3 domain	1FYN (84–142)	59	All β	[51]
B domain of protein A	1SS1	62	All α	[52]
DNA binding protein Sso7d	1BF4	63	All β	[53]
Chymotrypsin inhibitor 2	3CI2	64	α+β	[5]
IgG-binding domain of protein L	2PTL (15–78)	64	α+β	[54]
Cold shock protein	1CSP	67	All β	[55]
Ubiquitin	1UBQ	76	α+β	[56]
Procarboxypeptidase A2 actv. domain	1AYE (4A-83A)	78	α+β	[57]
Acyl-coenzyme A-binding protein	2ABD	86	All α	[58]
TI 127 Ig domain	1TIU	89	All β	[59]
Barstar	1BTB	89	α/β	[60]
TNfn3 domain of tenascin	1TEN	90	All β	[61]
FNfn10 domain of fibronectin	1TTF	94	All β	[62]
U1A	1URN	96	α+β	[63]
Ribosomal protein L23	1N88	96	α+β	[64]
Ribosomal protein S6	1RIS	97	α+β	[65]
Acylphosphatase	2ACY	98	α+β	[66]
FKBP12	1FKB	107	α+β	[67]
Barnase	1RNB	109	α+β	[68]
Villin 14T	2VIL	126	α+β	[69]
Azurin	1AZU	126	All β	[70]
CheY	3CHY	128	α/β	[71]

a) For some proteins, an abbreviated name is used for convenience. See the original paper for their full names.
b) The chain region is indicated in the parenthesis if the kinetic study and/or the calculation use only a part of the PDB data.
c) The number of residues used for calculating each Φ value is given. In some cases, the number of residues is different between the experimentally observed and calculated Φ values, because some residues are missing in the PDB data.
d) The fold class types are given based on the SCOP classification [72].
e) References for experimental Φ-value analysis are shown.

The theoretically calculated Φ-value profiles were assessed by their correlation coefficient (CC) with the experimentally observed Φ values. The results are shown in Table 2. The results from Garbuzynskiy *et al.* [73] are also shown, when available, for comparison. For Table 2, if the Φ-value experiments were carried out in more than one condition, we show only the case with the best CC. All the results are given in Supplementary Table S1.

In our previous study [30], we examined CC values to assess our statistical-mechanical model of protein folding. In this study, however, we shift our perspective to the characterization of protein folding using CC values, because they may imply whether or not a specific protein folds in the scheme assumed in our statistical-mechanical model. We consider here that differences in CC values reflect variation in active protein folding mechanisms. In Figure 1, proteins are classified into three groups according to the best CC value for each protein: $CC \geq 0.6$, $0.3 \leq CC < 0.6$, and $CC < 0.3$. We assess the groups as follows: the "$CC \geq 0.6$" group is a case where our model successfully reproduced the experimental Φ values, the "$CC < 0.3$" group is a case where our model failed, and the "$0.3 \leq CC < 0.6$" group is an intermediate case—where the experimental Φ values of some regions in a specific protein were successfully reproduced by our model, but others were not. In other words, the "$CC \geq 0.6$" group and the "$CC < 0.3$" group are rough indicators of the two major protein-folding schemes: the framework model and the nucleation-condensation model, respectively.

As mentioned in the Introduction, the concepts of the framework model and the nucleation-condensation model somehow remain ambiguous. In this paper, we confine our attention to their difference relating to the order in which short-, medium-, and long-range interactions are formed; that is, whereas longer-range interactions are assumed never to be formed before intervening shorter-range interactions in the framework model, long-range interactions can form before shorter-range interactions in the nucleation-condensation model.

(1) $CC \geq 0.6$. Out of 27 proteins examined, ten proteins were classified into this category. Two examples are shown in Figures 2A and B. See Supplementary Figure S1 for the remaining eight proteins.

Corr	Not sensitive	Sensitive to		
		+/− perturbation	Cutoff Distance	Chain entropy
CC > 0.6	1AZU (β) [C] 126 ((D)) 3CHY (α/β) [A] 128 ((D))		1SS1 (α) [A] 62 (S)	
			1ENH (α) [A] 54 (S) 1IDY (α) [A] 54 (S) 1N88 (α+β) [A] 96 ((D)) 1SRM (β) [B] 56 (D)	
		1URN (α+β) [B] 96 ((D))		
		1TEN (β) [C] 90 (S) 1SHG (β) [B] 57 (S)		
0.3 < CC < 0.6	3CI2 (α+β) [C] 64 (S) 1RNB (α+β) [A] 98 (S) 1RIS (α+β) [B] 109 (D) 2ACY (α+β) [B] 98 (S)		1BF4 (β) [A] 63 ((D)) 1TIU (β) [C] 89 (S) 1PGB (α+β) [C] 56 (S)	1FYN (β) [B] 59 ((D))
			1CSP (β) [A] 67 ((D))	
		2VIL (α+β) [A] 126 ((D))		
CC < 0.3	2ABD (α) [C] 86 ((D)) 1TTF (β) [C] 94 (D) 1UBQ (α+β) [C] 76 ((D)) 2PTL (α+β) [C] 64 (S) 1AYE (α+β) [B] 78 ((D)) 1BTB (α/β) [A] 89 ((D))	1FKB (α+β) [B] 107 (S)		

	#	Fold type				Size			CO profile			FE profile		
		α	β	α+β	α/β	S	M	L	A	B	C	(S)	(D)	((D))
CC > 0.6	10	3	4	2	1	5	0	5	5	3	2	5	1	4
0.3 < CC < 0.6	10	0	4	6	0	4	2	4	4	3	3	5	1	4
CC < 0.3	7	1	1	4	1	1	4	2	1	2	4	2	1	4

		#	α	β	α+β	α/β	S	M	L	A	B	C	(S)	(D)	((D))
Not Sensitive		12	1	2	7	2	2	4	6	3	3	6	4	2	6
Sensitive to	+/− perturbation	5	0	2	3	0	1	0	4	1	3	1	3	0	2
	cutoff distance	14	3	6	5	0	7	2	5	7	4	3	8	1	5
	chain entropy	9	2	5	2	0	5	1	3	4	4	1	5	1	3

Figure 1 Classification and characterization of proteins. The proteins were classified according to their correlation coefficient (CC) values between the calculated and experimentally measured Φ values, as well as their sensitivity to plus/minus perturbation, cutoff distance D_c, and chain entropy parameter B. The proteins were indexed according to four characteristics: fold class (α, β, α+β, or α/β), contact-order profile (A, B, or C), protein size (the number of residues), and free-energy profile [(S), (D), or ((D))]; see the text for details. The statistics of the top table are given in the lower two tables. The column headed "#" shows the number of proteins that belong to the relevant category. In the tables, the proteins have been grouped into three size classes, S, M, and L, according to the number of residues they contain: ≤64, 65–89, and ≥90, respectively.

Figure 2A for protein 1TEN shows that, although the Φ-value profile shows irregular trends, the experimentally observed Φ values were accurately reproduced. As described below, 1TEN is highly sensitive to plus/minus perturbations (it is also sensitive to other parameters used in the Φ-value calculation). The minus perturbation generated irregular Φ values for many residues. Rather than using the average of the two Φ values generated by the plus and minus perturbations, the Φ values generated by the plus perturbation were assigned to the Φ values of such residues. It is interesting that despite this potential confounding issue, 1TEN was found to belong to the high CC group.

Figure 2B for protein 1SHG is another example of CC≥0.6. Among the studied proteins, 1SHG showed high sensitivity to the parameters D_c and B. Whereas the CC was very high for D_c=5.5 Å, it was very low for D_c=4.2 Å. As for B, the CC for D42_13 was less than that for D42_16, whereas no difference was found between D55_13 and D55_16. The Φ-value profiles for the four parameter-sets differed significantly. However, because the number of experimentally observed Φ values was small relative to the number of residues, the statistical significance of the CC values was low. For example, although the calculated Φ values of the residues in the long N-terminal loop differed considerably depending on the parameter sets, such differences could not be assessed because of the lack of corresponding experimental data.

(2) 0.3≤CC<0.6. This category contains ten proteins. Only one example is given in Figure 2C for 3CI2. Whereas the Φ values of the residues in the N- and C-terminal regions were accurately reproduced, those in the middle of the polypeptide chain showed significant difference from the experimental Φ values. The statistical-mechanical model of a linear chain polymer has a tendency, in general, to have lower Φ values at both of the terminals because of a boundary effect, and to have higher Φ values in the middle of the chain because of strong cooperativity (owing to more interactions being present in the middle than at the terminals). Similar situations were observed in other proteins belonging to this category, such as 2ACY, 1BF4, and 1TIU (see Supplementary Fig. S1).

Figure 2C for protein 3CI2 also demonstrates a case where there was a variety of experimentally observed Φ values for a specific residue.

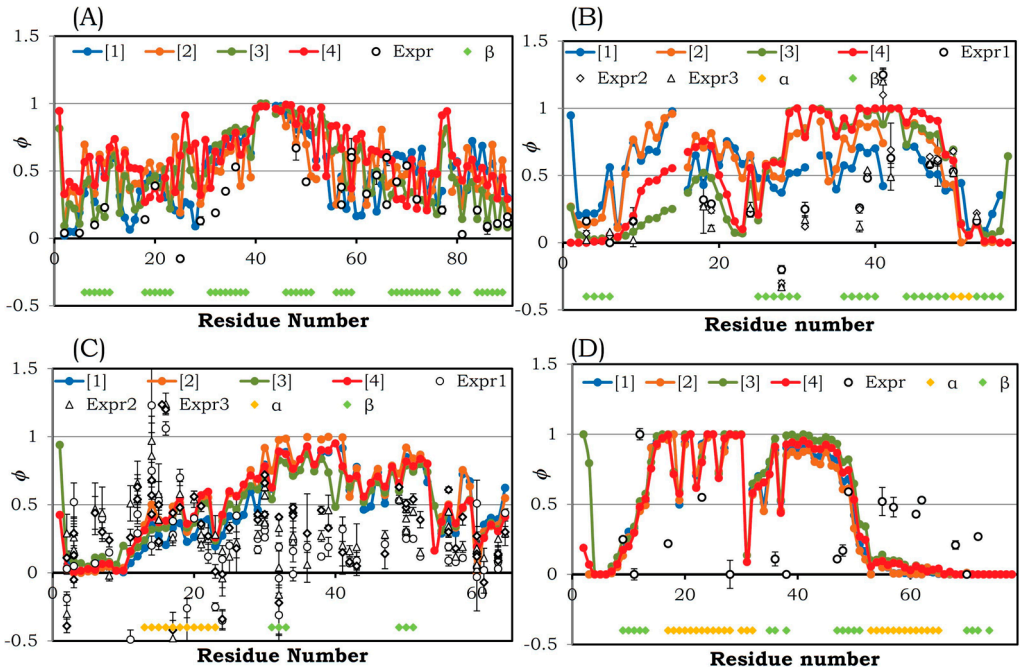


Figure 2 Φ-value profiles of proteins, (A) 1TEN, (B) 1SHG, (C) 3CI2, and (D) 1AYE. The Φ-value profiles of the other proteins are shown in Supplementary Figure S1. [1], [2], [3], and [4] indicate the parameter sets D42_13, D42_16, D55_13, and D55_16, respectively. Expr is an experimentally observed Φ value. Expr1, Expr2, etc. indicate that the Φ-value analysis was carried out in more than one experimental condition (see Supplementary Table S1 for details). α and β indicate the locations of an α-helix and a β-strand, respectively. The residues were numbered from "1" for all proteins regardless of their numbering in the PDB data. The corresponding residue numberings in the PDB data of proteins 1TEN, 1SHG, 3CI2, and 1AYE are 802–891, 6–62, 20–83, and 4A–83A, respectively.

(3) CC<0.3. This category is composed of the remaining seven proteins. These cases, which did not successfully reproduce experimentally determined Φ values, served to highlight the differences between the framework model and the nucleation-condensation model. Whereas our model is based on the framework model, the experiments suggested that the nucleation-condensation model is active for this protein.

Only one example is shown in Figure 2D for protein 1AYE. This protein has the secondary-structure sequence β1-α1-β2-β3-α2-β4. The experimental results suggested that while the folding nucleus is made by the packing of β1 and β3-α2 to form a β-sheet, β1-β3, in the transition state, the remaining parts, α1-β2 and β4, are completely unfolded. This means that the two separated regions, β1 and β3-α2, interact first before the formation of α1-β2 between them. Such a situation is beyond the scope of the assumptions of our model.

In other examples of CC<0.3 (Fig. 1; see also Supplementary Fig. S1), protein 2ABD has the secondary-structure sequence α1-α2-α3-α4. The experimental results suggested that the interaction between the N- and C-terminal α-helices, α1 and α4, initiates the folding process before the formation of the central part, α2-α3. Contradicting the experimental results, the calculated Φ values indicated that α2-α3 was formed but α1 and α4 were unstructured in the transition state. As a result, the CC for 2ABD was largely negative. It should be noted, however, that the experimental Φ values around residues 20–40 varied widely, and the authors of the study that generated those experimental values [58] pointed out that some of the data for these residues were not well suited for Φ-value analysis because the mutations probably significantly changed the folding pathway of the wild-type protein.

The 1BTB protein, which is another example from the "CC<0.3" group, has the secondary-structure sequence β1-α1-α2-β2-α3-α4-β3. The experimental results suggested that the β-sheet composed of β1 and β3 in the N- and C-terminals, respectively, is formed in the early stage of folding. By contrast, the Φ-value calculations indicated that whereas the C-terminal is structured in the early stage of folding, the N-terminal region is not.

In the case of protein 2PTL, which has the secondary-structure sequence β1-β2-α1-β3-β4, there may be another reason for observing CC<0.3. Whereas the N-terminal β-sheet of β1 and β2 was well predicted by the Φ-value calculations, the unstructured α-helix in the middle of the polypeptide chain could not be reproduced very well. As discussed in our previous paper [30] and also as pointed out above, our model has a strong tendency to describe the early formation of an α-helix in the middle part of the polypeptide chain, owing to the dominance of short-range interactions in an α-helix and the avoidance of boundary effects.

We should be careful, however, when using CC values for the above discussions, because CC value may not neces-

Table 2 Correlation between experimentally observed and theoretically calculated Φ values

Protein	N_r[a]	N_m[b]	Correlation coefficient, CC[c]					Description[d]
			D42_13	D42_16	D55_13	D55_16	Garbuzynskiy	
1ENH	54	13	0.69	0.41	0.90	0.63	N/A	
1IDY	54	18	0.61	0.62	0.59	0.63	N/A	
1PGB	56	25	0.34	0.33	0.08	0.09	0.76	
1SRM	56	35	0.65	0.65	0.65	0.63	0.63	
1SHG	57	14	0.05	0.35	0.77	0.77	0.82	pH 7
1FYN	59	9	0.52	0.35	0.52	0.34	N/A	
1SS1	62	31	0.25	0.27	0.70	0.71	N/A	2 M GdmCl
1BF4	63	21	0.36	0.39	0.35	0.37	0.81	
3CI2	64	40	0.23	0.33	0.27	0.30	0.46	Φ_F, 4 M GdmCl
2PTL	64	46	0.23	0.10	0.23	0.21	0.30	$1-\Phi_U$
1CSP	67	20	0.30	0.25	0.52	0.56	N/A	H_2O (kinetics)
1UBQ	76	20	0.03	0.01	–0.01	–0.01	N/A	Unfolding
1AYE	78	18	–0.22	–0.22	–0.21	–0.21	N/A	
2ABD	86	16	–0.74	–0.78	–0.76	–0.77	N/A	
1TIU	89	26	0.37	0.37	0.53	0.51	0.66	
1BTB	89	28	–0.09	–0.10	–0.11	–0.10	0.27	
1TEN	90	26	0.61	0.55	0.64	0.40	0.68	
1TTF	94	19	–0.14	–0.09	–0.15	–0.11	–0.22	0 D’
1URN	96	10	0.90	0.90	0.93	0.94	0.90	β'=0.5
1N88	96	16	0.64	0.63	0.47	0.59	N/A	
1RIS	97	20	0.13	0.11	0.06	0.03	0.49	1–4 M GdmCl
2ACY	98	22	0.47	0.43	0.45	0.42	N/A	
1FKB	107	22	–0.11	–0.35	–0.54	–0.54	0.32	3.9 M urea
1RNB	109	28	0.32	0.32	0.38	0.40	0.66	Water
2VIL	126	24	0.54	0.53	0.50	0.46	0.40	
1AZU	126	17	0.63	0.63	0.64	0.64	N/A	0 M GuHCl
3CHY	128	19	0.76	0.76	0.76	0.77	0.67	

a) The number of residues of a protein.
b) The number of mutated residues in a Φ-value analysis experiment. The residues with an irregular Φ value, i.e. Φ<–0.1 or Φ>1.1, are excluded.
c) The correlation coefficients between experimentally observed and theoretically calculated Φ values, referred to as CC in the text, are given for four parameter sets. The largest CC value for each protein is underlined. The results from Garbuzynskiy *et al.* [73] are also given for comparison, if available.
d) If the experiment was carried out in more than one condition, only one of them, that with the best CC value, is given in this table. The results not shown in this table are shown in Supplementary Table S1. See the corresponding reference in Table 1 for details about the experiment.

sarily be a reliable characteristic for assessing calculated Φ values. Whereas these calculations can generate a Φ value for any residue, the available experimental Φ values were limited. It is not clear how to assess irregular Φ values or residues for which Φ values cannot be obtained experimentally. For example, calculating the CC value after removing two Φ values of the C-terminal region from the 19 experimentally observed Φ values for protein 1TTF (see Supplementary Fig. S1) would result in changing the CC value from –0.09 to 0.28 in the D42_16 case. Moreover, Φ values for a specific protein frequently differ significantly, depending on the experimental conditions (see Supplementary Table S1).

FE profile

Since Φ-value profiles are calculated based on FE profiles, it is expected that FE profile characterization should be associated with protein-folding characterization. In Figure 3, the FE profiles of three proteins, 3CI2, 1RIS, and 1URN are shown for four parameter sets, D42_13, D42_16, D55_13, and D55_16 (FE profiles for the other proteins are given in Supplementary Fig. S2). Since the profiles shifted upward or downward depending on D_c, i.e., the number of interactions, we focused on the shapes of the profiles. It was found that the profile shapes were essentially similar if their chain entropy parameter B values were equal. A few exceptions were observed, such as 1CSP and 1N88. Conversely, a change in B value causes a change in shape of the FE profile. The major differences between the shapes of the FE profiles appeared in the region from the transition state to the folded state, i.e., from the central peak to the valley on the left side of the FE profile. As far as the unfolded region is concerned, the profile shape is conserved, whereas the distance from the local minimum to the peak at the transition state can change.

It was also found that, while the FE profiles of some proteins have a single peak at the transition state, others have more than one peak (usually double peaks) or a plateau. The latter case means that a transition state cannot be assigned to

a single folding reaction coordinate, but should be assigned to a range of coordinates. The secondary peak in a double-peaked profile is sometimes ambiguous, making it difficult to confidently judge its significance. We classified FE profiles into three types: single-peaked, double-peaked, and intermediate [denoted by (S), ((D)), and (D), respectively, in Fig. 1]. This classification could not be defined in a rigorous manner, but was carried out by visual inspection of the profiles. The FE profiles of 3CI2, 1RIS, and 1URN shown in Figure 3 were classified as single-peaked, intermediate, and double-peaked, respectively.

1CSP and 1N88 were exceptional cases, as mentioned above. Interestingly, while the FE profile of 1CSP with the parameter set D42_16 was single-peaked, the FE profiles calculated with the other parameter sets were double-peaked. This difference is reflected in the CC values: whereas the CC value was the lowest for D42_16, that for D55_16 was the highest. Accordingly, the FE profile of 1CSP was classified as double-peaked. For the FE profiles of 1N88, the shapes of the FE profiles for D42_16 and D55_16 differed considerably. While the former profile had a plateau, the latter was double-peaked. The CC values for the former profile were slightly better than those for the latter one.

According to Figure 1, whereas in the "0.3≤CC<0.6" and "CC≥0.6" groups, more proteins had a single-peaked FE profile than a double-peaked FE profile (10 vs. 8), this was reversed in the "CC<0.3" group (2 vs. 4). However, this difference was not statistically significant. Consequently, whether the FE profile was single-peaked or double-peaked was not a crucial factor for the CC values. The Φ values were essentially defined based on the single-peaked FE profiles. However, since Φ values can be calculated in the fashion described above, regardless of the double-peaked FE profile at the transition state, it may be necessary to reconsider Φ-value calculation for double-peaked FE profiles. The theory of multidimensional representation of an FE profile of protein folding that was proposed by Itoh & Sasai [27] would be highly relevant to any further considerations of this issue.

At the outset of this study, we expected to obtain useful information about protein folding by determining the FE profiles of the evaluated proteins. In our previous paper [30], we demonstrated a close relationship between a protein's FE profile, its folding rate, and its fold class; however, in this study we realized that it is necessary to consider the fine structural changes of the FE profile to be able to predict changes in Φ values. For example, the FE profiles of 1ENH and 1IDY (Supplementary Fig. S2) seemed to be almost identical. Consequently, their calculated Φ-value profiles were grossly similar, but differed significantly in their fine details. FE profile changes that were induced by imposing perturbations were responsible for the observed changes in Φ values. However, it was hard to find such changes in the FE profiles, because such changes were not only small, but were also distributed over all regions of the profile. Similar

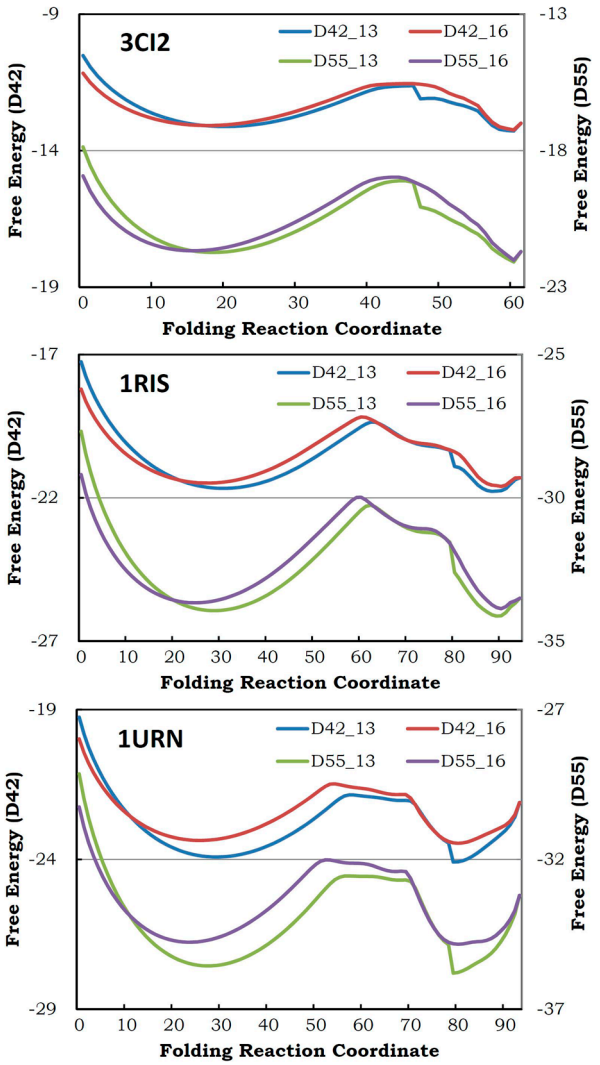


Figure 3 Free-energy profiles of proteins, (A) 3CI2, (B) 1RIS, and (C) 1URN. The vertical axis on the left is used for D42_13 and D42_16, while that on the right is used for D55_13 and D55_16. The horizontal axis traces the progress of the protein folding reaction from left to right, i.e., from a completely unfolded state to a completely folded state. The free-energy profiles of the other proteins are shown in Supplementary Figure S2.

situations existed for proteins in the SH3 domain-containing superfamily, namely 1SRM, 1SHG, and 1FYN (see Supplementary Figs. S1 and S2).

CO profile

CO is one property of proteins that is well-known to be correlated with folding kinetics. The CO profile, which was discussed above, is an extension of the CO; the cumulative number of native contacts c_k, is plotted against residue-residue distances along a chain, k. This profile provides information about the respective contribution ratios of the different interaction types (short-, medium-, and long-range) to the CO. Figure 4 shows the CO profiles of 27 proteins, in which c_k and k were normalized between zero and one to

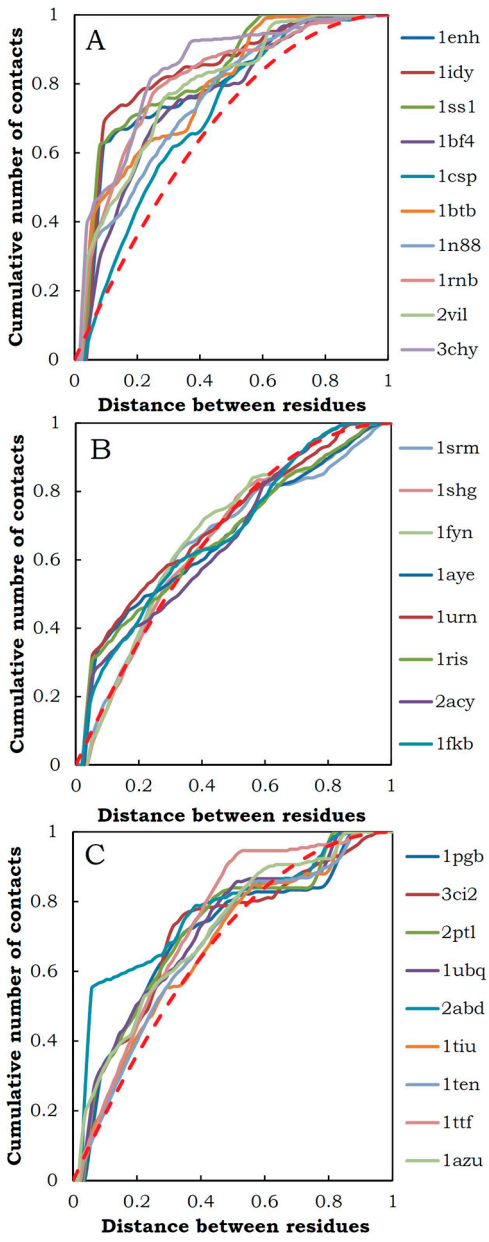


Figure 4 Contact-order profiles of 27 proteins. They were classified into three types, A, B, and C. The dashed curve is $y=-(1-x)^2+1$, where [x] and [y] are [the distance between two residues along the chain] and [the cumulative number of residue pairs with a distance less than or equal to x], respectively. This curve was used to derive the classification of the proteins (see the text for details).

adjust their protein sizes. It was found that the CO profiles could be classified into three types, referred to as Types A, B, and C (Fig. 4A, B, and C, respectively). In this classification, the reference curve (a dashed line in Fig. 4), $c_k=-(k-1)^2+1$ (with normalized c_k and k), was used. This reference curve can be utilized if the number of native contacts of a residue-residue distance along the chain i decreases in inverse proportion to i, i.e., $\rho_i \approx n-i$, where n is the number of residues of a given protein.

If a whole CO-profile curve was located above the reference curve, the protein was classified as Type A. If a CO-profile curve for small k was located above the reference curve, but for large k, it dropped below the reference curve, the protein was classified as either Type B or Type C. If the CO-profile curve cut across the reference curve around middle k and was close to the reference curve throughout, the protein was classified as Type B; if it cut across the reference curve around large k, the protein was classified as Type C. Short-range interactions were dominant in Type A proteins. The CO-profile curves of Type B proteins contained, as the reference curve, the short-, medium-, and long-range interactions that decreased in inverse proportion to residue-residue distance. Type C proteins were characterized by very long-range interactions such as interactions between N- and C-terminal regions. Since the upper left area of the curve corresponds to CO as pointed out by Eq. (11), the CO values of Type A proteins were the smallest and those of Type C were the largest.

Figure 1 shows that the CC values of Type A proteins were greater than those of Type C proteins (while the ratio of Type A to Type C was 9:5 in the "CC≥0.6" and "0.3≤CC<0.6" groups, it was 1:4 in the "CC<0.3" group). The CC values of the Type B proteins were of intermediate magnitude. Since the Type A proteins were rich in short-range interactions, it would be natural to expect that such proteins would be more likely to adopt protein-folding mechanisms consistent with the framework model. Another remarkable finding from this study was that Type A proteins were sensitive to the parameters used in the Φ-value calculation. The most plausible reason for this is that smaller proteins, which are more sensitive to these parameters, were more likely to be classified into Type A than into other groups. This finding indicates that CO-profile type, as a method of categorizing proteins, is probably not directly related to parameter sensitivity—although this cannot be completely ruled out.

The Type C proteins in the "CC<0.3" group were interesting because the long-range interactions, such as those between the N- and C-terminal regions, were formed at an early stage of the folding process without being preceded by the formation of secondary structures in the middle of the chain; that is, their protein-folding proceeded in accordance with the nucleation-condensation model. Unfortunately, the model studied in this paper cannot take into account the nucleation-condensation model. This is not because we deny that the nucleation-condensation mechanism can be active in protein folding, but simply because it is very difficult to construct a solvable statistical-mechanical model that can simulate this mechanism. However, we are interested in examining how proteins can fold in accordance with the nucleation-condensation model in future work.

CO has previously been discussed in terms of the folding rate of a protein [42–46]. In our previous paper [30], we demonstrated that the folding rates calculated by the present model were also well correlated with the experimentally

Table 3 Root-mean-square difference between Φ-value profiles calculated with different parameter sets[a)]

Protein	Plus/minus perturbation				D_c		B	
	D42_13	D42_16	D55_13	D55_16	D42_13 D55_13	D42_16 D55_16	D42_13 D42_16	D55_13 D55_16
1ENH	0.06	0.05	0.04	0.02	0.17	0.13	0.31	0.29
1IDY	0.04	0.05	0.03	0.04	0.11	0.21	0.13	0.25
1PGB	0.02	0.02	0.01	0.01	0.16	0.15	0.04	0.04
1SRM	0.03	0.04	0.01	0.01	0.16	0.15	0.21	0.15
1SHG	0.18	0.03	0.14	0.01	0.36	0.23	0.19	0.16
1FYN	0.01	0.02	0.04	0.02	0.08	0.07	0.16	0.15
1SS1	0.07	0.04	0.03	0.04	0.25	0.30	0.12	0.08
1BF4	0.04	0.04	0.02	0.03	0.18	0.13	0.11	0.03
3CI2	0.03	0.02	0.19	0.03	0.09	0.08	0.10	0.11
2PTL	0.02	0.01	0.03	0.10	0.14	0.10	0.10	0.07
1CSP	0.04	0.04	0.02	0.01	0.33	0.23	0.20	0.10
1UBQ	0.04	0.02	0.03	0.03	0.10	0.12	0.12	0.09
1AYE	0.02	0.02	0.01	0.01	0.12	0.07	0.03	0.13
2ABD	0.04	0.11	0.02	0.03	0.06	0.08	0.09	0.05
1TIU	0.05	0.02	0.02	0.02	0.19	0.19	0.03	0.02
1BTB	0.01	0.05	0.02	0.02	0.15	0.12	0.03	0.03
1TEN	0.12	0.39	0.27	0.43	0.21	0.19	0.21	0.19
1TTF	0.04	0.06	0.04	0.03	0.13	0.13	0.04	0.02
1URN	0.21	0.13	0.04	0.08	0.21	0.12	0.13	0.12
1N88	0.06	0.05	0.08	0.02	0.16	0.14	0.14	0.17
1RIS	0.03	0.04	0.01	0.01	0.08	0.07	0.06	0.05
2ACY	0.01	0.02	0.01	0.02	0.10	0.11	0.04	0.04
1FKB	0.42	0.32	0.34	0.08	0.20	0.30	0.19	0.11
1RNB	0.01	0.04	0.03	0.02	0.10	0.11	0.02	0.04
2VIL	0.08	0.07	0.12	0.03	0.16	0.15	0.09	0.08
1AZU	0.01	0.04	0.02	0.01	0.10	0.11	0.02	0.01
3CHY	0.04	0.05	0.06	0.05	0.07	0.06	0.07	0.05

a) The underline indicates that a root-mean-square difference is greater than 0.1 for plus/minus perturbation and 0.15 for cutoff distance D_c and chain entropy parameter B.

determined folding rates of 72 proteins (with CC=0.81). Because Φ values were defined based on folding and unfolding rates, and such rates are related to the CO, the Φ-value profile should be related to the CO or the CO profile; and we did indeed find some relationships between them. However, it was also apparent that these are not the only factors that determine Φ values and protein folding.

Sensitivity to parameters

A single set of parameter values to be used in our model has not yet been established. In fact, Table 2 shows that the best CC values for individual proteins (see underlined figures) were obtained by using different parameter sets, i.e., D42_13, D42_16, D55_13, or D55_16. In other words, Φ-value calculations are sensitive to the parameters chosen. The cutoff distance D_c that defines a residue-residue contact is one of the most important parameters, because it defines the 3D structure of a specific protein in the model. Parameter B, chain entropy, is also important, because protein-folding transitions occur in the balance between enthalpy gain that occurs with residue-residue interactions and entropy loss—particularly that associated with chain entropy. In addition, we do not have a sufficient number of experimentally determined Φ values. Consequently, it is hard to determine the best parameter set for our model at present. We consider sensitivity of proteins to such parameters as being one of the intrinsic characteristics of protein folding.

Parameter sensitivity was assessed using the root-mean-square difference (RMSD) between two Φ-value profiles that were calculated with different parameter sets. For analysis of plus/minus perturbations, Φ-value profiles for plus and minus perturbations were compared. For analysis of the distance cutoff D_c, Φ-value profiles of [D42_13 and D55_13] and [D42_16 and D55_16] were compared. For analysis of chain entropy parameter B, Φ-value profiles of [D42_13 and D42_16] and [D55_13 and D55_16] were compared. The results are shown in Table 3. The relatively larger RMSD values are indicated by an underline.

Half of the proteins were sensitive to D_c. In particular, small proteins (smaller than 63 residues) such as 1ENH were sensitive to D_c. A natural expectation would be that changes in the residue-residue contact matrix (Γ_{ij}) caused by a change

in D_c would have a significant effect on smaller proteins (see Eq. (2)), and that relatively larger proteins, two examples being "all-β" proteins (such as 1CSP, 1TIU, and 1TEN) and "α+β" proteins (such as 1URN, 1N88, 1FKB, and 2VIL), would be sensitive to D_c. However, we found that many larger proteins and "α+β" proteins were classified into the "not sensitive" group.

Most proteins that were sensitive to chain entropy parameter B were also sensitive to D_c. Smaller proteins such as 1ENH, 1IDY, 1SRM, 1SHG, and 1FYN were easily influenced by changes in the parameters. It is understandable that "all-β" proteins such as 1SRM, 1TEN, 1SHG, 1FYN, and 1CSP would be sensitive to B, because chain entropy plays an important role in the formation of β-sheets. In the formation of a β-sheet, a larger loss of chain entropy must be overcome than in the formation of an α-helix.

It is possible to mutate any protein. Mutations may change a protein's 3D structure to some extent, and consequently may change the residue-residue contact matrix. A set of proteins with similar 3D structures can be regarded as an example of the outcome of such mutations. As discussed above, 1ENH and 1IDY have similar 3D structures, and their overall Φ-value profiles resembled each other. However, they differed significantly in some specific regions. In this study, we show that their calculated Φ-value profiles reproduce these experimentally observed differences, at least to some extent. We also show that the same is true for the SH3 domain-containing superfamily proteins, 1SRM, 1SHG, and 1FYN.

Beyond this, it is interesting that Φ-value profiles can provide information that is helpful for identifying residues that are sensitive to the model parameters. For example, Φ-value profiles allowed us to determine that residues in the N-terminal regions of 1SRM, 1SHG, and 1FYN were sensitive to model parameters (Supplementary Fig. S1), as were almost all residues in 1SHG. In the Φ-value experiments, the Φ values obtained for some specific residues sometimes differed significantly, depending on substituted amino acid types or on experimental conditions (see for example Fig. 2C). These results imply that mutations of certain residues in each protein have significant effects on protein folding. The sensitivity of specific residues to model parameters can reveal such characteristics of a protein.

Plus/minus perturbation

In this paper, Φ values were calculated by imposing perturbations of protein structure via single amino acid substitutions. These perturbations changed the predicted interactions of a specific residue with the other residues of the protein. The two perturbation types "plus" and "minus" were defined according to whether the interactions were weakened or strengthened by them, respectively. We hypothesized that when the degree of perturbation is small, the difference in the effect of the plus and minus perturbations should be small; and during the current study, this presumption held

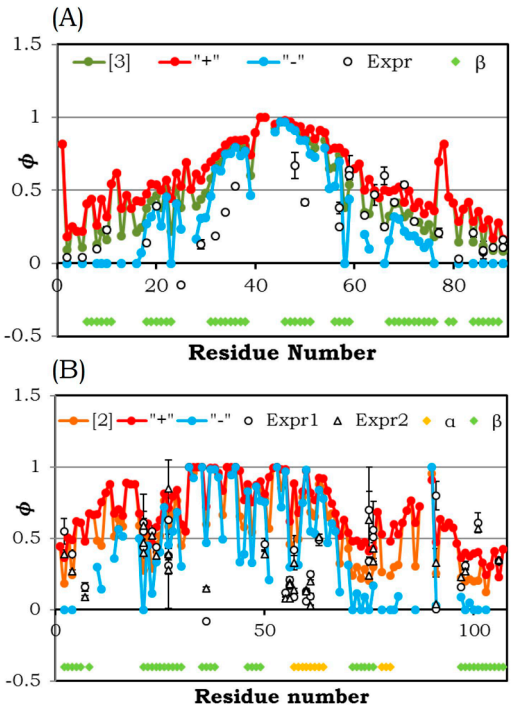


Figure 5 Φ-value profiles of plus/minus perturbations. (A) 1TEN, D42_16 and (B) 1FKB, D55_13. See also the caption of Figure 2 for the legend. "+" and "–" are Φ-value profiles obtained by plus and minus perturbations, respectively. Residues without data indicate that the calculated Φ values were irregular.

true for most proteins. However, it was found that the two types of perturbation produced significantly different results for some proteins: i.e., 1URN, 1TEN, 1SHG, 2VIL, and 1FKB (Fig. 1 and Table 2). Note that these are "all-β" or "α+β" proteins. In these proteins, some residues showed large differences between Φ values calculated in the presence of plus perturbations vs. minus perturbations. In some cases, either of the two perturbations gave an irregular Φ value, i.e., Φ much greater than one or much smaller than zero. Interestingly, when an irregular Φ value was obtained, it was almost always in the presence of a minus perturbation.

In particular, the RMSDs between Φ-value profiles were larger in 1TEN and 1FKB than in other proteins (Fig. 5). Irregular Φ values were obtained for many more residues in these two proteins than in the other proteins. Curiously, the CC values for these two proteins differed considerably. Whereas the CC value for 1TEN was very high, that for 1FKB was very low (see Table 2).

In the experimental situation, the substitution of an amino acid residue with a larger side chain is one of the strategies employed to introduce stronger interactions than exist in the wild-type protein. However, if such a residue disrupts the proper folding of the protein owing to its side chain size, the Φ value may be undeterminable or irregular. In the theoretical model, however, the plus/minus perturbations only change the ensemble of conformations at every step along the folding reaction coordinate. Our results suggest that

irregular Φ values can be obtained without disrupting proper folding.

Although we examined the FE profiles of 1TEN and 1FKB, we could not find remarkable differences between their plus/minus perturbations. In general, an irregular Φ value is obtained if $\Delta\ln k_f - \Delta\ln k_u$, which is the denominator of Eq. (10), is too small. One possible explanation is that subtle changes occur throughout the FE profiles, and occasionally a change in the folding rate $\Delta\ln k_f$ is very close to a change in the unfolding rate $\Delta\ln k_u$. However, it should be noted that 1TEN and 1FKB are "all-β" proteins and have complex folding topologies of β-strands. A protein's folding topology is regarded as complex if the folding process cannot easily be represented as a step-by-step formation from short-, to medium-, and then to long-range interactions. Although the CO is introduced to explain such complexity, more details regarding folding topology may be necessary to explain the remarkable differences observed between the plus/minus perturbations for these proteins. Interestingly, in the FE profile of 1FKB, the difference between the local minimum in the unfolding region and the local maximum in the transition region was considerably larger than that for the other proteins (see Supplementary Fig. S2). Obviously, this reflects the complex topology of folding that exists for 1FKB, which affects its irregular behavior in the Φ-value calculation.

Circular permutants

The Φ-value analysis experiment involving circular permutants of the ribosomal protein S6 [74] (PDB ID: 1RIS) provided an interesting example to discuss Φ-value calculation, particularly regarding the important role of the connectivity of the polypeptide chain. A circular permutant is created by dividing a given protein into two fragments, i.e., N- and C-terminal fragments, by an incision; then the N-terminal fragment is connected after the C-terminal fragment. In this analysis, circular permutants were assumed to have the same 3D structure as the wild-type protein, despite the differences in their connectivities.

1RIS has the secondary-structure sequence β1-α1-β2-β3-α2-β4. Five circular permutants, P^{13-14}, P^{33-34}, P^{54-55}, P^{68-69}, and P^{81-82}, were constructed, where the superscripts indicate the position of the incision. All possible incisions between the secondary-structure elements of 1RIS were examined [74].

Φ values for the circular permutants were calculated by means of the same inter-residue contact information that was used for the wild-type 1RIS, but with differences in the respective connectivity, e.g., 14–97 and 1–13 for P^{13-14}. Figure 6 shows the Φ-value profiles for the wild-type 1RIS and its circular permutant, P^{13-14}. Φ-value profiles for the other circular permutants are given in Supplementary Figure S3. CC values for the wild-type proteins and circular permutants are shown in Table 4.

The two Φ-value profiles shown in Figure 6 are useful for illustrating the significance of chain connectivity [30]. The wild-type 1RIS has the secondary-structure sequence β1-α1-β2-β3-α2-β4; it forms β-sheets with long-range interactions between β1 and β3, and between β1 and β4. When the secondary-structure sequence in the circular permutant P^{13-14} changes to α1-β2-β3-α2-β4-β1, the interactions between β1 and β4 change to short- or medium-range ones, and the C-terminal region forms a compact domain, β3-α2-β4-β1, that contains β1 and β3. While the circular permutants P^{33-34} and P^{54-55} behave similarly to P^{13-14}, P^{68-69} and P^{81-82} do not. Accordingly, the Φ-value profiles of P^{13-14}, P^{33-34}, and P^{54-55} were different from those of P^{68-69}, P^{81-82}, and the wild-type

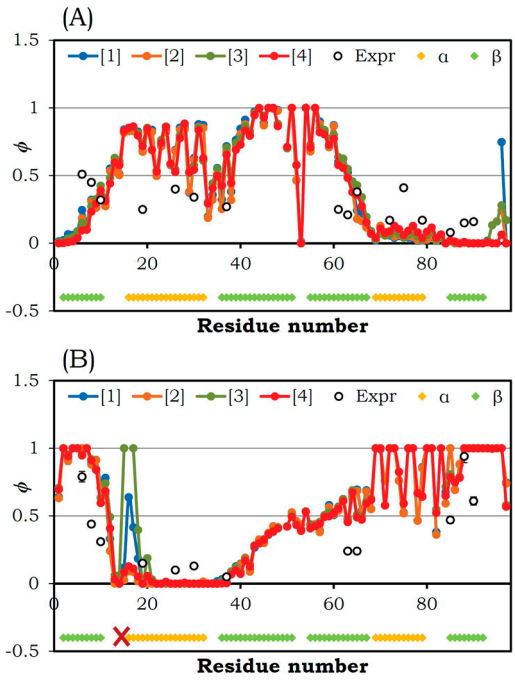


Figure 6 Φ-value profiles of the wild-type 1RIS and its circular permutant, P^{13-14}. See also the caption of Figure 2 for the legend. The cross symbol in (B) indicates the location of the scission point of the circular permutant. The Φ-value profiles of the other permutants are given in Supplemental Figure S3.

Table 4 Correlation between experimentally observed and theoretically calculated Φ values for wild-type 1RIS and its circular permutants[a)]

Protein	N_m	Correlation coefficient, CC			
		D42_13	D42_16	D55_13	D55_16
1RIS wt[b)]	16	0.32	0.29	0.25	0.22
1RIS P^{13-14}	12	0.84	0.85	0.88	0.89
1RIS P^{33-34}	15	0.43	0.44	0.51	0.51
1RIS P^{54-55}	16	0.51	0.50	0.60	0.60
1RIS P^{68-69}	15	–0.02	–0.05	–0.10	–0.13
1RIS P^{81-82}	15	–0.05	–0.08	–0.12	–0.16

a) See also the footnotes to Table 2.
b) The experimental data for wild-type 1RIS are cited from the different paper [74] from that in Table 2 [65].

protein, as shown in Figure 6 and Supplementary Figure S3. Owing to the reduction of the long-range interactions in the first group of 1RIS circular permutants, their CC values became markedly better than those of the second group of 1RIS circular permutants. In other words, whereas the former circular permutants underwent folding according to the framework model, the latter ones underwent folding according to the nucleation-condensation model.

Recently, Inanami *et al.* proposed a novel method to calculate the folding pathway of a multidomain protein in an extended form of the model discussed here [37]. Their target protein was dihydrofolate reductase (DHFR), which has two domains: one comprising a continuous middle part of the polypeptide chain (ABD domain), and the other comprising discontinuous N- and C-terminal parts (DLD domain). According to the limitations of our model, the DLD domain cannot fold without ABD folding. However, Inanami *et al.* enhanced the model by introducing the "virtual loop-closure mechanism" to partition function in an exactly calculable form, which allowed them to successfully reproduce DHFR folding behavior. Just as the circular permutant strategy can resolve topological complexities of protein folding, the methodology of Inanami *et al.* seems to open up a new method of reproducing protein folding by the virtual-closure mechanism, especially when the protein folding occurs according to the nucleation-condensation model.

Conclusion

When planning this project, we expected to derive a unified view of protein folding by studying the folding process from various perspectives. We anticipated that the further characterization of proteins would provide a coherent explanation of the information presented in the table in Figure 1. All of the aspects examined in this paper should play a significant role in future studies of protein folding, and they are associated with one another. However, such connections are complex and exhibit highly nonlinear characteristics. It is known that the relationship between the Φ value and the extent of structural formation is not necessarily linear, and thus the interpretation of intermediate Φ values is still controversial. In addition, the number of available data from Φ-value analyses was too small for reliable statistical analyses at the time of this study. Consequently, it is hard to understand the meanings of the Φ values clearly at present. In this study, however, we have confirmed the importance of characterizing protein folding from various perspectives. Our findings have also highlighted that protein folding is highly variable and individual among different proteins. This should be considered when pursuing a unified theory of protein folding.

Acknowledgements

This paper is respectfully dedicated to the late Professor Nobuhiko Saitô, who was a Ph.D. supervisor of H. W., and with whom the original statistical-mechanical model of protein folding, "island model", was developed. This work was supported by JSPS Grant-in-Aid for Scientific Research (C) (Grant No. 16K00407).

Conflicts of Interest

The authors have declared that no competing interests exist.

Author Contributions

H. W. and H. A. contributed equally to this work. Both authors discussed the results and implications, and commented on the manuscript at all stages.

References

[1] Anfinsen, C. B. The formation and stabilization of protein structure. *Biochem. J.* **128**, 737–749 (1972).
[2] Tanaka, S. & Scheraga, H. A. Hypothesis about the mechanism of protein folding. *Macromolecules* **10**, 291–304 (1977).
[3] Kim, P. S. & Baldwin, R. L. Specific intermediates in the folding reactions of small proteins and the mechanism of protein folding. *Annu. Rev. Biochem.* **51**, 459–489 (1982).
[4] Ptitsyn, O. B. Protein folding: hypotheses and experiments. *J. Protein Chem*. **6**, 273–293 (1987).
[5] Itzhaki, L. S., Otzen, D. E. & Fersht, A. R. The structure of the transition state for folding of chymotrypsin inhibitor 2 analysed by protein engineering methods: Evidence for a nucleation-condensation mechanism for protein folding. *J. Mol. Biol.* **254**, 260–288 (1995).
[6] Fersht, A. R. Nucleation mechanisms in protein folding. *Curr. Opin. Struct. Biol.* **7**, 3–9 (1997).
[7] Muñoz, V. A simple theoretical model goes a long way in explaining complex behavior in protein folding. *Proc. Natl. Acad. Sci. USA* **111**, 15863–15864 (2014).
[8] Wako, H. & Saitô, N. Statistical mechanical theory of the protein conformation. I. General considerations and the application to homopolymers. *J. Phys. Soc. Jpn.* **44**, 1931–1938 (1978).
[9] Wako, H. & Saitô, N. Statistical mechanical theory of the protein conformation. II. Folding pathway for protein. *J. Phys. Soc. Jpn.* **44**, 1939–1945 (1978).
[10] Gō, N. & Abe, H. Noninteracting local-structure model of folding and unfolding transition in globular proteins. I. Formulation. *Biopolymers* **20**, 991–1011 (1981).
[11] Abe, H. & Gō, N. Noninteracting local-structure model of folding and unfolding transition in globular proteins. II. Application to two-dimensional lattice proteins. *Biopolymers* **20**, 1013–1031 (1981).
[12] Muñoz, V. & Eaton, W. A. A simple model for calculating the kinetics of protein folding from three-dimensional structures. *Proc. Natl. Acad. Sci. USA* **96**, 11311–11316 (1999).
[13] Matouschek, A., Kellis, J. T., Serrano, L. & Fersht, A. R. Mapping the transition state and pathway of protein folding by protein engineering. *Nature* **340**, 122–126 (1989).
[14] Fersht, A. R., Matouschek, A. & Serrano, L. The folding of an enzyme. I. Theory of protein engineering analysis of stability and pathway of protein folding. *J. Mol. Biol.* **224**, 771–782 (1992).

[15] Fersht, A. R. & Sato, S. Φ-value analysis and the nature of protein-folding transition states. *Proc. Natl. Acad. Sci. USA* **101**, 7976–7981 (2004).
[16] Nölting, B. & Agard, D. A. How general is the nucleation-condensation mechanism? *Proteins* **73**, 754–764 (2008).
[17] Alonso, D. O. & Daggett, V. Staphylococcal protein A: Unfolding pathways, unfolded states, and differences between the B and E domains. *Proc. Natl. Acad. Sci. USA* **97**, 133–138 (2000).
[18] Settanni, G., Rao, F. & Caflisch, A. Φ-value analysis by molecular dynamics simulations of reversible folding. *Proc. Natl. Acad. Sci. USA* **102**, 628–633 (2005).
[19] Banachewicz, W., Religa, T. L., Schaeffer, R. D., Daggett, V. & Fersht, A. R. Malleability of folding intermediates in the homeodomain superfamily. *Proc. Natl. Acad. Sci. USA* **108**, 5596–5601 (2011).
[20] Portman, J. J., Takada, S. & Wolynes, P. G. Variational theory for site resolved protein folding free energy surfaces. *Phys. Rev. Lett.* **81**, 5237–5240 (1998).
[21] Galzitskaya, O. V. & Finkelstein, A. V. A theoretical search for folding/unfolding nuclei in three-dimensional protein structures. *Proc. Natl. Acad. Sci. USA* **96**, 11299–11304 (1999).
[22] Alm, E. & Baker, D. Prediction of protein-folding mechanisms from free-energy landscapes derived from native structures. *Proc. Natl. Acad. Sci. USA* **96**, 11305–11310 (1999).
[23] Alm, E., Morozov, A. V., Kortemme, T. & Baker, D. Simple physical models connect theory and experiment in protein folding kinetics. *J. Mol. Biol.* **322**, 463–476 (2002).
[24] Zong, C., Wilson, C. J., Shen, T., Wolynes, P. G. & Wittung-Stafshede, P. Φ-Value analysis of apo-azurin folding: Comparison between experiment and theory. *Biochemistry* **45**, 6458–6466 (2006).
[25] Itoh, K. & Sasai, M. Flexibly varying folding mechanism of a nearly symmetrical protein: B domain of protein A. *Proc. Natl. Acad. Sci. USA* **103**, 7298–7303 (2006).
[26] Kubelka, J., Henry, E. R., Cellmer, T., Hofrichter, J. & Eaton, W. A. Chemical, physical, and theoretical kinetics of an ultrafast folding protein. *Proc. Natl. Acad. Sci. USA* **105**, 18655–18662 (2008).
[27] Itoh, K. & Sasai, M. Multidimensional theory of protein folding. *J. Chem. Phys.* **130**, 145104 (2009).
[28] Abe, H. & Wako, H. Folding/unfolding kinetics of lattice proteins studied using a simple statistical mechanical model for protein folding, I: Dependence on native structures and amino acid sequences. *Physica A* **383**, 3442–3454 (2009).
[29] Wako, H. & Abe, H. Folding/unfolding kinetics of lattice proteins by applying a simple statistical mechanical model for protein folding. in *Protein Folding* (Walters, E.C. ed.) pp. 349–376 (Nova Sci. Pub. Inc., New York, 2011).
[30] Wako, H. & Abe, H. Calculation of free-energy profiles, folding rates and Φ values by means of a simple statistical-mechanical model of protein folding. in *Advances in Protein Folding Research* (Hale, M. ed.) pp. 19–63 (Nova Sci. Pub. Inc., New York, 2015).
[31] Itoh, K. & Sasai, M. Dynamical transition and proteinquake in photoactive yellow protein. *Proc. Natl. Acad. Sci. USA* **101**, 14736–14741 (2004).
[32] Imparato, A., Pelizzola, A. & Zamparo M. Ising-like model for protein mechanical unfolding. *Phys. Rev. Lett.* **98**, 148102 (2007).
[33] Itoh, K. & Sasai, M. Entropic mechanism of large fluctuation in allosteric transition. *Proc. Natl. Acad. Sci. USA* **107**, 7775–7780 (2010).
[34] Bruscolini, P. & Naganathan, A. N. Quantitative prediction of protein folding behaviors from a simple statistical model. *J. Am. Chem. Soc.* **133**, 5372–5379 (2011).
[35] Caraglio, M. & Pelizzola, A. Effects of confinement on thermal stability and folding kinetics in a simple Ising-like model. *Phys. Biol.* **9**, 016006 (2012).
[36] Sivanandan, S. & Naganathan, A. N. A disorder-induced domino-like destabilization mechanism governs the folding and functional dynamics of the repeat protein IκBα. *PLoS Comput. Biol.* **9**, e1003403 (2013).
[37] Inanami, T., Terada, T. P. & Sasai, M. Folding pathway of a multidomain protein depends on its topology of domain connectivity. *Proc. Natl. Acad. Sci. USA* **111**, 15969–15974 (2014).
[38] Abe, H. & Wako, H. Analyses of simulations of three-dimensional lattice proteins in comparison with a simplified statistical mechanical model of protein folding. *Phys. Rev. E* **74**, 011913 (2006).
[39] Taketomi, H., Ueda, Y. & Gō, N. Studies on protein folding, unfolding and fluctuations by computer simulation. I. The effect of specific amino acid sequence represented by specific inter-unit interactions. *Int. J. Pept. Protein Res.* **7**, 445–459 (1975).
[40] Abe, H. & Wako, H. Application of a statistical-mechanical model for protein folding to a three-dimensional lattice protein. *J. Phys. Soc. Jpn.* **73**, 1143–1146 (2004).
[41] Henry, E. R. & Eaton, W. A. Combinatorial modeling of protein folding kinetics: free energy profiles and rates. *Chem. Phys.* **307**, 163–185 (2004).
[42] Plaxco, K. W., Simons, K. T. & Baker, D. Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.* **277**, 985–994 (1978).
[43] Plaxco, K. W., Simons, K. T., Ruczinski, I. & Baker, D. Topology, stability, sequence, and length: defining the determinants of two-state protein folding kinetics. *Biochemistry* **39**, 11177–11183 (2000).
[44] Gromiha, M. M. & Selvaraj, S. Comparison between long-range interactions and contact order in determining the folding rate of two-state proteins: application of long-range order to folding rate prediction. *J. Mol. Biol.* **310**, 27–32 (2001).
[45] Kamagata, K., Arai, M. & Kuwajima, K. Unification of the folding mechanisms of non-two-state and two-state proteins. *J. Mol. Biol.* **339**, 951–965 (2004).
[46] Kuznetsov, I. B. & Rackovsky, S. Class-specific correlations between protein folding rate, structure-derived, and sequence-derived descriptors. *Proteins* **54**, 333–341 (2004).
[47] Gianni, S., Guydosh, N. R., Khan, F., Caldas, T. D., Mayor, U., White, G. W. N., *et al.* Unifying features in protein-folding mechanisms. *Proc. Natl. Acad. Sci. USA* **100**, 13286–13291 (2003).
[48] McCallister, E. L., Alm, E. & Baker, D. Critical role of β-hairpin formation in protein G folding. *Nat. Struct. Biol.* **7**, 669–673 (2000).
[49] Riddle, D. S., Grantcharova, V. P., Santiago, J. V., Alm, E., Ruczinski, I. & Baker, D. Experiment and theory highlight role of native state topology in SH3 folding. *Nat. Struct. Biol.* **6**, 1016–1024 (1999).
[50] Martinez, J. C. & Serrano, L. The folding transition state between SH3 domains is conformationally restricted and evolutionarily conserved. *Nat. Struct. Biol.* **6**, 1010–1016 (1999).
[51] Northey, J. G. B., Di Nardo, A. A. & Davidson, A. R. Hydrophobic core packing in the SH3 domain folding transition state. *Nat. Struct. Biol*. **9**, 126–130 (2002).
[52] Sato, S., Religa, T. L., Daggett, V. & Fersht, A. R. Testing protein-folding simulations by experiment: B domain of protein A. *Proc. Natl. Acad. Sci. USA* **101**, 6952–6956 (2004).
[53] Guerois, R. & Serrano, L. The SH3-fold family: experimental evidence and prediction of variations in the folding pathways. *J. Mol. Biol.* **304**, 967–982 (2000).
[54] Kim, D. E., Fisher, C. & Baker, D. A breakdown of symmetry

in the folding transition state of protein L. *J. Mol. Biol.* **298**, 971–984 (2000).
[55] Garcia-Mira, M. M., Boehringer, D. & Schmid, F. X. The folding transition state of the cold shock protein is strongly polarized. *J. Mol. Biol.* **339**, 555–569 (2004).
[56] Went, H. M. & Jackson, S. E. Ubiquitin folds through a highly polarized transition state. *Protein Eng. Des. Sel.* **18**, 229–237 (2005).
[57] Villegas, V., Martínez, J. C., Avilés, F. X. & Serrano, L. Structure of the transition state in the folding process of human procarboxypeptidase A2 activation domain. *J. Mol. Biol.* **283**, 1027–1036 (1998).
[58] Kragelund, B. B., Osmark, P., Neergaard, T. B., Schiødt, J., Kristiansen, K., Knudsen, J., *et al.* The formation of a native-like structure containing eight conserved hydrophobic residues is rate limiting in two-state protein folding of ACBP. *Nat. Struct. Biol.* **6**, 594–601 (1999).
[59] Fowler, S. B. & Clarke, J. Mapping the folding pathway of an immunoglobulin domain: structural detail from phi value analysis and movement of the transition state. *Structure* **9**, 355–366 (2001).
[60] Nölting, B. & Andert, K. Mechanism of protein folding. *Proteins* **41**, 288–298 (2000).
[61] Hamill, S. J., Steward, A. & Clarke, J. The folding of an immunoglobulin-like Greek key protein is defined by a common-core nucleus and regions constrained by topology. *J. Mol. Biol.* **297**, 165–178 (2000).
[62] Cota, E., Steward, A., Fowler, S. B. & Clarke, J. The folding nucleus of a fibronectin type III domain is composed of core residues of the immunoglobulin-like fold. *J. Mol. Biol.* **305**, 1185–1194 (2001).
[63] Ternström, T., Mayor, U., Akke, M. & Oliveberg, M. From snapshot to movie: φ analysis of protein folding transition states taken one step further. *Proc. Natl. Acad. Sci. USA* **96**, 14854–14859 (1999).
[64] Hedberg, L. & Oliveberg, M. Scattered Hammond plots reveal second level of site-specific information in protein folding: ϕ' ($\beta^{\ddagger}$). *Proc. Natl. Acad. Sci. USA* **101**, 7606–7611 (2004).
[65] Otzen, D. E. & Oliveberg, M. Conformational plasticity in folding of the split β-α-β protein S6: evidence for burst-phase disruption of the native state. *J. Mol. Biol.* **317**, 613–627 (2002).
[66] Chiti, F., Taddei, N., White, P. M., Bucciantini, M., Magherini, F., Stefani, M., *et al.* Mutational analysis of acylphosphatase suggests the importance of topology and contact order in protein folding. *Nat. Struct. Biol.* **6**, 1005–1009 (1999).
[67] Fulton, K. F., Main, E. R. G., Daggett, V. & Jackson, S. E. Mapping the interactions present in the transition state for unfolding/folding of FKBP12. *J. Mol. Biol.* **291**, 445–461 (1999).
[68] Serrano, L., Matouschek, A. & Fersht, A. R. The folding of an enzyme. III. Structure of the transition state for unfolding of barnase analysed by a protein engineering procedure. *J. Mol. Biol.* **224**, 805–818 (1992).
[69] Choe, S. E., Li, L., Matsudaira, P. T., Wagner, G. & Shakhnovich, E. I. Differential stabilization of two hydrophobic cores in the transition state of the villin 14T folding reaction. *J. Mol. Biol.* **304**, 99–115 (2000).
[70] Wilson, C. J. & Wittung-Stafshede, P. Snapshots of a dynamic folding nucleus in zinc-substituted *Pseudomonas aeruginosa* azurin. *Biochemistry* **44**, 10054–10062 (2005).
[71] López-Hernández, E. & Serrano, L. Structure of the transition state for folding of the 129 aa protein CheY resembles that of a smaller protein, CI-2. *Fold. Des.* **1**, 43–55 (1996).
[72] Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540 (1995).
[73] Garbuzynskiy, S. O., Finkelstein, A. V. & Galzitskaya, O. V. Outlining folding nuclei in globular proteins. *J. Mol. Biol.* **336**, 509–525 (2004).
[74] Haglund, E., Lindberg, M. O. & Oliveberg, M. Changes of protein folding pathways by circular permutation. Overlapping nuclei promote global cooperativity. *J. Biol. Chem.* **283**, 27904–27915 (2008).