Cooperativity and modularity in protein folding

(submitted to the Nobuhiko Saitô memorial issue of *Biophysics and Physicobiology***)**

Masaki Sasai,* George Chikenji, and Tomoki P. Terada

Department of Computational Science and Engineering and Department of Applied Physics, Nagoya University, Nagoya, Aichi 464-8603, Japan

*Electronic address: sasai@tbp.nuap.nagoya-u.ac.jp

Abstract

A simple statistical mechanical model proposed by Wako and Saitô has explained the aspects of protein folding surprisingly well. This model was systematically applied to multiple proteins by Muñoz and Eaton and has since been referred to as the Wako-Saitô-Muñoz-Eaton (WSME) model. The success of the WSME model in explaining the folding of many proteins has verified the hypothesis that the folding is dominated by native interactions, which makes the energy landscape globally biased toward native conformation. Using the WSME and other related models, Saitô emphasized the importance of the hierarchical pathway in protein folding; folding starts with the creation of contiguous segments having a native-like configuration and proceeds as growth and coalescence of these segments. The ϕ-values calculated for barnase with the WSME model suggested that segments contributing to the folding nucleus are similar to the structural modules defined by the pattern of native atomic contacts. The WSME model was extended to explain folding of multi-domain proteins having a complex topology, which opened the way to comprehensively understanding the folding process of multi-domain proteins. The WSME model was also extended to describe allosteric transitions, indicating that the allosteric structural movement does not occur as a deterministic sequential change between two conformations but as a stochastic diffusive motion over the dynamically changing energy landscape. Statistical mechanical viewpoint on folding, as highlighted by the WSME model, has been renovated in the context of modern methods and ideas, and will continue to provide insights on equilibrium and dynamical features of proteins.

Running title: Cooperativity and modularity

Key words: WSME model, energy landscape, statistical mechanics

Introduction

Understanding protein folding is a fascinating problem of biomolecular self-organization, and it is a prerequisite for comprehending the reactions and interactions of proteins. An important method for delineating the folding problem is through a simple statistical mechanical model. The model was proposed by Wako and Saitô in 1978 [1, 2] by extending classical models of helix-coil transitions [3, 4] to many-bodied heterogeneous cases. However, the model was not widely accepted until quantitative comparison between the model results and the experimental data became possible.

Around 1990–2000, three important advances changed the researchers' viewpoint. The first advance was the progress in statistical mechanics of complex systems such as spin glasses and neural networks. Accordingly, a complex system's behavior could be described as a competition between its tendency to be trapped into one of extensively many disordered states and its tendency to globally drift along the energy landscape toward an ordered functional state. Applying this notion to protein folding revealed that the global structure of the folding energy landscape is a key to explaining the experimental results [5]. The second advance was the experimental observation of the folding rates of systematically derived mutant proteins, which led to the ϕ-value analysis technique to reveal structures of the transition state ensemble of folding [6, 7]. The third advance was the drastic increase in computational power, which facilitated not only large-scale simulations with realistic models but also the quick and accurate evaluation of folding mechanisms with simplified models. Combining these advances, theoretical models of the energy landscape of folding were introduced to explain and predict the experimentally observed ϕ-values and other quantities, which led to the innovative cooperation between theories and experiments and promoted a paradigm shift in folding studies [8, 9]. The model developed by Wako and Saitô was "re-discovered" in 1999 by Muñoz and Eaton [10], and this model has since made a significant contribution to the advancement in folding studies.

A major advantage of this model is that the partition function can be exactly calculated from the model Hamiltonian [11, 12]; the exact calculation allows us to obtain a transparent picture on free-energy landscapes, pathways, and rates of folding. The model was at first criticized as quantitatively invalid [13]. However, such invalidity was due to the particular approximation used in the calculation and the problem disappeared when the exact solution

of the model was used. Since then, the Wako-Saitô-Muñoz-Eaton (WSME) model has been widely applied in calculating pathways [14–23] and kinetics [14, 19, 23–25] of folding as well as in explaining mechanical unfolding [26, 27], amyloidosis [28], and allosteric transitions and functions [29–32]. In this review, we discuss the physics behind the WSME model and its applications to folding and other intriguing biophysical problems.

The WSME Model and Cooperativity

In the WSME model, a protein conformation is described by a set of Ising-like variables, $\{m_i\}$. $m_i = 1$, when the dihedral angles of the backbone chain at the ith residue have similar values to those in the native conformation, and $m_i = 0$ otherwise. The WSME Hamiltonian is defined by a function of $\{m_i\}$ as

$$H_{\mathrm{WSME}}(\{m_i\}) = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \epsilon_{ij} \Delta_{ij} \prod_{k=i}^{j} m_k, \tag{1}$$

where N is the total number of residues in the protein and Δ_{ij} represents the pattern of native contacts: $\Delta_{ij} = 1$, when the residues i and j are in contact in the native conformation and $\Delta_{ij} = 0$ otherwise. $\epsilon_{ij} < 0$ represents the strength of the attractive native interactions, for which we may use $\epsilon_{ij} \approx -0.3$ to -1.5 kcal·mol^{-1} depending on the extent of the atomic contacts between the residues i and j in the native conformation [23]. The partition function is calculated as

$$Z_{\mathrm{WSME}}(n) = \mathrm{Tr}_n \exp\left(-H_{\mathrm{WSME}}(\{m_i\})/k_{\mathrm{B}}T - \sum_{i=1}^{N} \sigma_i m_i \right). \tag{2}$$

Here, $0 \leq n \leq 1$ is an order parameter of folding: $n = 0$ when the chain is completely disordered, $n = 1$ when the structure is identical to that determined via X-ray or NMR analysis. Tr_n is a sum under the constraint $M = \sum_{i=1}^{N} m_i = Nn$ as $\mathrm{Tr}_n = \sum_{m_1=0,1} \sum_{m_2=0,1} \cdots \sum_{m_N=0,1} \delta_{M,Nn}$, where $\delta_{M,Nn}$ is a Kronecker delta. $-\sigma_i$ represents the reduction of entropy upon structure ordering at the residue i, and we may use $\sigma_i k_{\mathrm{B}} \approx 2$–$3$ cal·mol^{-1}K^{-1} [23]. From Eq.2, we can calculate the free energy, $F(n) = -k_{\mathrm{B}}T \log Z_{\mathrm{WSME}}(n)$, which is the one-dimensional free-energy landscape represented as a function of n. The expression of Eq. 2 can be easily extended to the two-dimensional version, $Z_{\mathrm{WSME}}(n_1, n_2)$, with the corresponding free-energy landscape, $F(n_1, n_2)$, by introducing the two-dimensional folding order parameter (n_1, n_2) with $n_1 = \sum_{i=1}^{N_1} m_i/N_1$, $n_2 = \sum_{i=N_1+1}^{N} m_i/N_2$, and $N_1 + N_2 = N$ [15, 16, 20, 23]; the higher-dimensional representation is also feasible [20].

The WSME model is based on two major assumptions. First, it does not consider non-native interactions. Since only native interactions are explicitly considered in Eq. 1, the energy monotonously decreases as the chain approaches the native conformation, *i.e.,* the energy landscape has a global bias toward native conformation. This global bias has been considered as a characteristic of sequences selected by evolution to meet consistency between local and global structures [33] or to show minimally frustrated interactions [5]. The model with such a global bias was first considered by Gō and his colleagues [34–36], and the WSME model belongs to a class of such "Gō-like models".

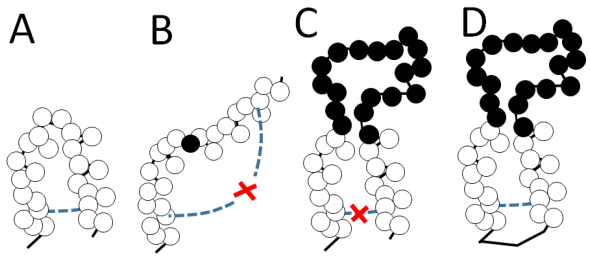

FIG. 1: The native interaction in the WSME model. Residues in the native-like configuration are shown with white circles, and residues in non-native configurations are shown with filled circles. A) The native interaction (a blue dashed line) between the residues within a contiguous native-like segment is taken into account in the WSME model. B) The interaction becomes ineffective when an intervening residue is in the non-native configuration. C) If the linker chain connecting two native-like segments is long enough, a number of residues with random configurations can compensate each other to allow two segments to reach the positions where native interactions are effective. This type of interaction, however, is not taken into account in the WSME model. D) Interactions as in C can be suitably calculated with the WSME Hamiltonian if we consider that the N- and C-termini are connected by a virtual link, as explained in the section "The WSME Model for Multi-domain Proteins".

Another significant assumption in the model is that a native interaction occurs only within the "island" of a native-like configuration; the ϵ_{ij} term in Eq. 1 has a nonzero contribution to H_{WSME} only when the consecutive segment from residues i through j assume native-like configurations, satisfying $m_i m_{i+1} \cdots m_{j-1} m_j = 1$. This assumption is illustrated in Fig. 1, where intra-segment native interactions are effective (Fig. 1A), but interactions are ineffective when an intervening residue takes the "wrong" direction (Fig. 1B). This assumption

seems plausible when we consider that the residues should form a local ordered structure through compact atomic packing of residue side chains. Such local structural ordering should be represented as a cooperative many-residue correlation given by $m_i m_{i+1} \cdots m_{j-1} m_j = 1$ and not as a naive summation of pairwise correlations.

With these two assumptions, contiguous native-like segments are energetically stabilized. Therefore, as illustrated in Fig. 2, folding starts with the creation of short segments with the native-like configuration and proceeds through growth and coalescence of these segments into a larger region to assume the native conformation. We should note that there are combinatorially many ways of segment creation, growth, and coalescence, and the statistical weight of these different pathways is evaluated with the WSME model to explain the distribution of folding pathways observed in the ensemble of protein molecules.

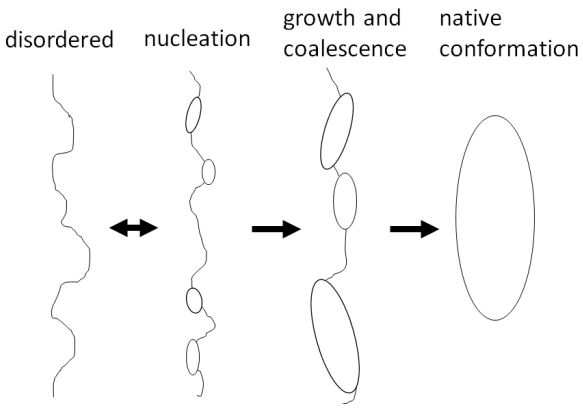

FIG. 2: The hierarchical process of protein folding. Folding starts with the creation of contiguous segments with a native-like configuration. After nucleation, folding proceeds as those segments grow and coalesce into larger regions to reach native conformation.

The WSME model quantitatively explains free-energy landscapes, pathways, ϕ-values, and kinetic rates of the folding of various proteins [14–23]. In Fig. 3, an example result is shown for the B domain of protein A (BdpA). As shown in Fig. 3A, BdpA is a small 60 residue, α-helical protein comprising three helices: H1, H2, and H3. BdpA demonstrates a two-state folding transition between the unfolded and native states [37]. The two-dimensional free-energy landscape $F(n_1, n_2)$ was calculated, where n_1 is the order parameter of folding for the N-terminal half, and n_2 is the one for the C-terminal half. In $F(n_1, n_2)$ of Fig. 3B, we find two basins: one at a small n_1 and a small n_2, which corresponds to the unfolded state, and the other at $(n_1, n_2) \approx (0.95, 0.96)$, which corresponds to the native state.

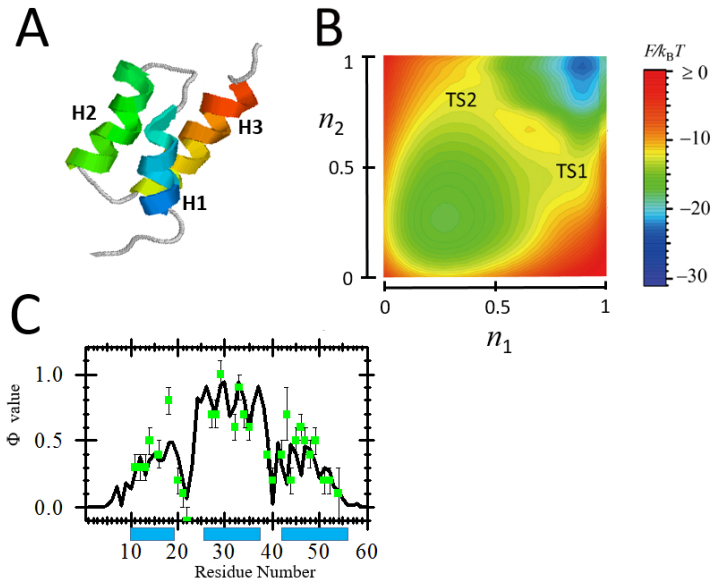

FIG. 3: Application of the WSME model to the B domain of *Staphylococcal* protein A (BdpA). A) Native conformation of BdpA (Protein Data Bank (PDB) code: 1bdd). B) Two-dimensional free-energy landscape, $F(n_1, n_2)$, calculated with the WSME model, where n_1 is the folding order parameter of the N-terminal half, and n_2 is the one of the C-terminal half. A contour is drawn every $0.5k_BT$. $F(n_1, n_2)$ has two basins: the unfolded state basin ($n_1 \approx 0.3, n_2 \approx 0.3$) and the basin of the native state ($n_1 \approx 1.0, n_2 \approx 1.0$). Two transition states, TS1 and TS2, are shown; there are two dominant pathways of folding, which proceed through TS1 and TS2. C) Comparison of the calculated and observed ϕ-values. The calculated values are shown with a line and the observed values [37] are green squares shown with error bars. Bars on the bottom represent the positions of α helices. Modified from Figs. 1, 3, and 5 of [15].

In this landscape, we find two saddles with similar free-energy heighs; therefore, BdpA has two dominant transition states, TS1 and TS2, in this representation. Along the pathway through TS1, the helix H1 folds earlier than H3, whereas along the pathway through TS2, H3 folds earlier than H1. The ϕ-values were calculated at TS1 and TS2 with the WSME model. Here, the ϕ-value represents the frequency of structure formation at each residue in the transition state ensemble. By averaging the ϕ-values at two TSs with the respective weights of the Boltzmann factor, the average ϕ-values are calculated and compared with the observed ones in Fig. 3C, which shows good agreement between the calculated and observed data. The existence of two TSs having almost equivalent free-energy heights is due to the

symmetrical native conformation of BdpA, as shown in Fig. 3A, and a subtle difference in the experimental conditions or settings of the simulation model should break this symmetry and change the relative heights of TS1 and TS2. The results of several simulation studies are conflicting on which helix, H1 or H3, folds earlier [38], but the WSME model provides a clear explanation of the reason for this disagreement; a symmetrical native conformation brings about the competing multiple pathways of folding and the detailed simulation condition or the parameter setting modulates the relative statistical importance of multiple pathways.

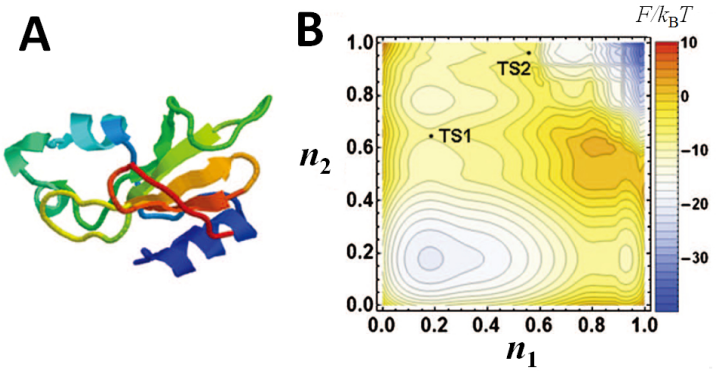

FIG. 4: Application of the WSME model to barnase. A) Native conformation of barnase from *Bacillus amyloliquefaciens* (PDB code: 1a2p). B) Two-dimensional free-energy landscape, $F(n_1, n_2)$, calculated with the WSME model, where n_1 is the order parameter of folding of the N-terminal half, and n_2 is the one of the C-terminal half. Contour is drawn in every $2k_BT$. $F(n_1, n_2)$ has four basins; basin of unfolded state ($n_1 \approx 0.2, n_2 \approx 0.2$), basin of native state ($n_1 \approx 1.0, n_2 \approx 1.0$), and two basins of intermediate states, I_1 ($n_1 \approx 0.2, n_2 \approx 0.8$) and I_2 ($n_1 \approx 0.9, n_2 \approx 0.2$). Saddles around the basin I_1 are much lower in free energy than those around I_2 are; therefore, a pathway through I_1 is a dominant pathway, and I_1 is a dominant intermediate. I_2 could be detected as an off-pathway intermediate. Along the dominant pathway, there are two transition states, TS1 and TS2. Modified from Fig. 14 of [20] with permission.

Another example is shown for barnase in Figs. 4 and 5. Barnase is a 110 residue $\alpha + \beta$ protein (Fig. 4A), and its folding proceeds via an intermediate state [6]. The two-dimensional free-energy landscape $F(n_1, n_2)$ was calculated by disregarding two structurally unresolved residues with $N_1 = 54$ and $N_2 = 54$; therefore, n_1 is the order parameter of folding for the N-terminal half and n_2 is the one for the C-terminal half. In $F(n_1, n_2)$ of Fig. 4B, a

dominant intermediate state is represented by a basin at a large n_2 and a small n_1 value, indicating that the C-terminal half is more structurally ordered than the N-terminal half is in the intermediate state. There are two transition states, TS1 between the unfolded and intermediate states, and TS2 between the intermediate and native states. In Fig. 5, the calculated ϕ-values at TS1 and TS2 are compared with the experimentally observed values [39, 40], showing a good agreement between the WSME results and the observed data. In barnase, as shown in Fig. 5, the ϕ-value shows a large change around the boundaries of the structural modules, which are defined by the geometrical pattern of the native contacts [41–45]. This interesting feature will be discussed later in the *Discussion* section.

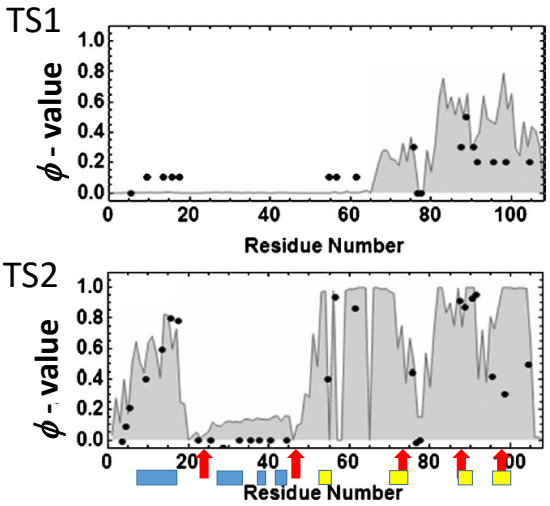

FIG. 5: Calculated and observed ϕ-values at the two transition states, TS1 and TS2, of barnase. Lines shaded with gray correspond to the calculated ϕ-values with the WSME model. Dots are the experimentally observed values [39, 40]. Red arrows are boundaries of modules defined by the pattern of atomic contacts in the native conformation [44, 45]. Bars shown on the bottom represent secondary structure elements, helices (blue) and strands (yellow). Modified from Fig. 15 of [20] with permission.

As in the above examples, the WSME model explained the experimentally observed data of many proteins, which strongly suggests that the two major assumptions made in developing the WSME model, dominance of native interactions and the local cooperative

formation of the native-like configuration, are indeed valid assumptions. The dominance of native interactions was also recently shown [21, 46] using folding trajectories of all-atom simulations performed by Shaw's group [47–49]. Comparing the folding trajectories of all-atom simulations and the WSME results, it was shown that the much simpler WSME model quantitatively explains the all-atom results [21]. The dominance of native interactions can be interpreted as following. When we consider the atomic details of a short molecular dynamics trajectory of the picosecond time-scale, there would be no distinction between native and non-native interactions; both have the same physical origin as electrostatic, hydrophobic, or van der Waals interactions. However, when we consider a microsecond or a longer process, the non-native interactions are only transiently formed within that process; also, the lifetime of native interactions is much longer due to the multi-residue cooperativity forming the local ordered structure. Then, we can approximate the long-term process using only the native interactions. The dominance of native interactions and the resulting globally biased energy landscape were first assumed by Gō and his colleagues to explain the two-state feature of folding transitions [33, 34]. It was re-formulated later to explain how the trapping into the non-native states is prevented as well as how the Levinthal paradox is resolved in the energy landscape perspective [5, 8]. Here, the dominance of native interactions in folding has been clearly supported by the results of the quantitative analyses of experimental data and all-atom simulations, and the WSME model has played an important role in these analyses.

By regarding the dominance of native interactions as the 0th order description, non-native interactions should determine the next order description. Thus, non-native interactions should bring about the off-pathway intermediates in the folding process or work as "friction" in the course of folding [50]; non-native interactions may destabilize the native conformation to some extent to make the structure flexible to meet functional requirements [51]. Understanding the role of non-native interactions in long-term dynamics remains as an important challenging problem.

In the WSME model, contiguous native-like segments are emphasized so that interactions such as those shown in Fig. 1B or 1C are neglected. Within a single-domain structure, this approximation seems reasonable. To make the native interaction between residues belonging to two segments separated by residues with the non-native configuration effective, as shown in Fig. 1C, the multiple intervening residues in the linker between two segments must follow multiple non-native directions to compensate for "incorrect" directions and to recover the

"correct" orientation between residues having the native interaction. This flexible structural adjustment of the linker chain is a necessary condition to make the interaction effective, but such flexible adjustment is rare in a single domain when the linker is short. Therefore, the assumption made for the WSME model is considered appropriate at least for describing the folding process of single-domain proteins. Indeed, the validity of the WSME model was shown for single-domain proteins [14, 15, 17–21], but further careful argument is necessary to describe multi-domain proteins, particularly when they have a nontrivial topological arrangement of domains, as discussed in the next section.

The eWSME Model for Multi-domain Proteins

Many proteins show all-or-none two-state transitions between the folded and unfolded states, but in 1978, Wako and Saitô [2] suggested the presence of an intermediate state for lysozyme based on the calculated heterogeneous size distribution of contiguous native-like segments. In the 1980s, clear experimental evidence was discovered for the folding intermediates, which were referred to as the molten globule states [52]. Particularly, the folding process of typical small multi-domain proteins, such as α-lactalbumin and lysozyme, was analyzed. It was shown that, in these example proteins, the intermediate state in the equilibrium three-state transition is very similar to the intermediate state that appears on the kinetic folding pathway, suggesting the pivotal role of the molten globule state in protein folding. Furthermore, the structure of the molten globule state is heterogeneous and composed of ordered and disordered parts, whereas the degrees of compaction and side-chain packing largely depend on the protein species. To obtain a unified picture of the diversity of the molten globule state, extending the WSME model to describe generic multi-domain proteins by taking account of native interactions, as illustrated in Fig. 1C, is strongly desired. The need for considering native interactions between residues separated by others with non-native configuration is evident particularly for proteins having topologically complex structures, as shown in Fig. 6.

Dihydrofolate reductase (DHFR), a 159 residue α/β protein, for example, has two domains, the discontinuous loop domain (DLD) and the adenosine-binding domain (ABD), as shown in Fig. 6A; the ABD is a continuous domain comprising the residues 38-106, and the DLD is a discontinuous domain comprising the N-terminal part (residues 1-37) and

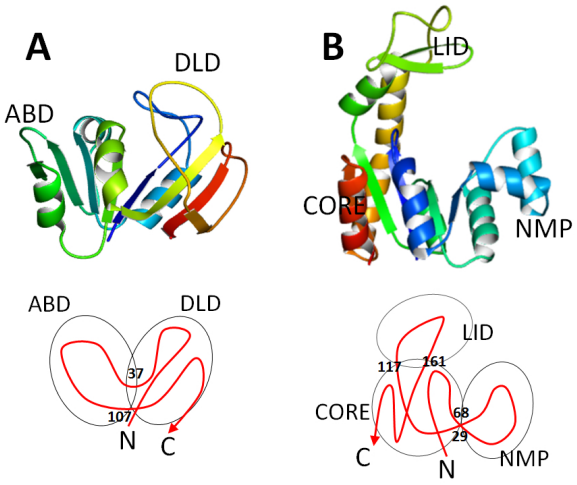

FIG. 6: Examples of multi-domain proteins with non-trivial topology. A) Dihydrofolate reductase (DHFR) (PDB code: 1rx1) has two domains, DLD and ABD. B) Adenylate kinase (AdK) (PDB code: 4ake) has three domains, CORE, NMP, and LID. Topological connectivity of the chain is illustrated at the bottom.

the C-terminal part (residues 107-159). Therefore, native interactions between the N- and C-terminal parts in the DLD are expected to form even when the intervening ABD is disordered, which is just the case illustrated in Fig. 1C. A convenient way to consider such interactions is to introduce a virtual link connecting the N- and C-termini (Fig. 1D) and applying the WSME Hamiltonian to this virtually closed ring to derive the partition function $Z_{\rm ring}$. Using $Z_{\rm ring}$, the extended WSME (eWSME) partition function is defined by

$$Z_{\rm eWSME}(n) = Z_{\rm WSME}(n) + (Z_{\rm ring}(n) - Z_{\rm WSME}(n))e^{S_{\rm ring}(n)/k_{\rm B}}, \qquad (3)$$

where $S_{\rm ring}(n) < 0$ is the entropic reduction arising from the constraint to place the N- and C-termini at a distance determined by the native conformation, which can be estimated assuming that the disordered parts of the chain under the n constraint behave as fragments with random configurations [23]. $Z_{\rm eWSME}$ is smoothly interpolated between $Z_{\rm WSME}$ and $Z_{\rm ring}$; $Z_{\rm eWSME} \approx Z_{\rm WSME}$, when the entropic reduction is significant, as $S_{\rm ring} \ll 0$, and $Z_{\rm eWSME} \approx Z_{\rm ring}$, when the entropic reduction is negligible, as $S_{\rm ring} \approx 0$. $Z_{\rm eWSME}$ incorporates both local multi-residue correlations as in $Z_{\rm WSME}$ and native interactions separated by intervening non-native residues with suitable statistical weights; also, it is exactly calculable.

The two-dimensional free-energy folding landscape of DHFR calculated with this eWSME

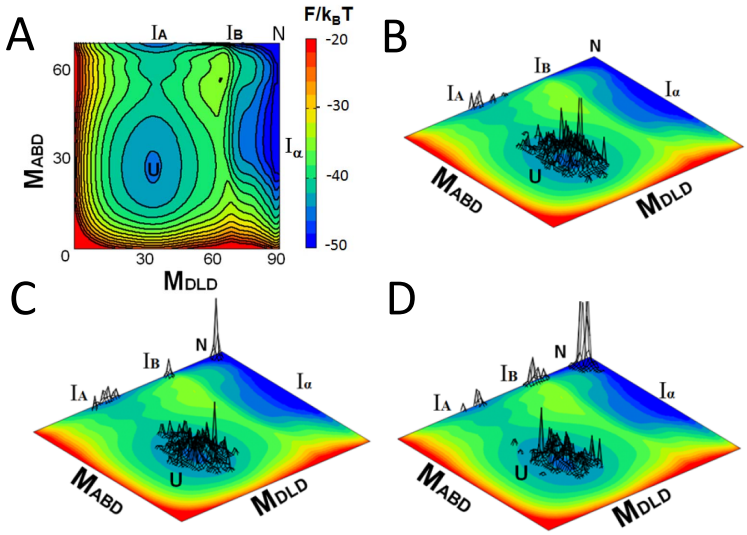

FIG. 7: Free-energy landscape and kinetics of DHFR folding calculated by the eWSME model. A) Free-energy landscape of DHFR folding represented in the two-dimensional space of M_{DLD} and M_{ABD}. The landscape has basins corresponding to the unfolded state U, the native state N, and the intermediates, $\mathrm{I_A}$, $\mathrm{I_B}$, and I_α. B–D) Evolution of the population of 200 molecules simulated with the Monte Carlo calculation at B) $3.3 \times 10^5 t_0$, C) $1.6 \times 10^6 t_0$, and D) $3.0 \times 10^6 t_0$, where t_0 is a unit of time in simulation. Reproduced from [23].

model is shown in Fig. 7A [23]. Here, the two-dimensional space is defined by the parameters $M_{\mathrm{DLD}} = \sum_{i \in \mathrm{DLD}} m_i$ and $M_{\mathrm{ABD}} = \sum_{i \in \mathrm{ABD}} m_i$. This landscape has basins at $(M_{\mathrm{DLD}}, M_{\mathrm{ABD}}) \approx (30, 30)$, which is the basin of the unfolded state (U); at $(M_{\mathrm{DLD}}, M_{\mathrm{ABD}}) \approx (30, 69)$ (the basin denoted by $\mathrm{I_A}$); at $(M_{\mathrm{DLD}}, M_{\mathrm{ABD}}) \approx (70, 69)$ (the basin $\mathrm{I_B}$); at $(M_{\mathrm{DLD}}, M_{\mathrm{ABD}}) \approx (90, 35)$ (the basin I_α); and at $(M_{\mathrm{DLD}}, M_{\mathrm{ABD}}) \approx (90, 69)$ (the basin of the native state, N). In $\mathrm{I_A}$, the ABD is folded and the DLD is unfolded, whereas, in I_α, the DLD is folded and the ABD is unfolded. The basin I_α has lower free energy than $\mathrm{I_A}$; however, I_α is separated from U by a higher free-energy barrier than $\mathrm{I_A}$. Therefore, we can expect that molecules starting from U pass through $\mathrm{I_A}$ to proceed along the pathway $\mathrm{U} \rightarrow \mathrm{I_A} \rightarrow \mathrm{I_B} \rightarrow \mathrm{N}$. This was confirmed by numerically following the kinetic change of $\{m_i\}$ with the Monte Carlo simulation using the following function to calculate the effective eWSME energy for the Metropolis criterion;

$$
\begin{aligned}
E_{\mathrm{eWSME}}(\{m_i\}) \; = & \\
& - \; k_{\mathrm{B}}T \log\left(e^{-H_{\mathrm{WSME}}/k_{\mathrm{B}}T} + (e^{-H_{\mathrm{ring}}/k_{\mathrm{B}}T} - e^{-H_{\mathrm{WSME}}/k_{\mathrm{B}}T})e^{S_{\mathrm{ring}}/k_{\mathrm{B}}}\right) \\
& + \; k_{\mathrm{B}}T \sum_i \sigma_i m_i .
\end{aligned}
\tag{4}
$$

The kinetic evolution of the DHFR molecules' population on the two-dimensional space is shown in Figs. 7B–7D. These panels show that the population indeed proceeds along the folding pathway U $\rightarrow$ $\mathrm{I_A}$ $\rightarrow$ $\mathrm{I_B}$ $\rightarrow$ N by sequentially visiting the intermediate states $\mathrm{I_A}$ and $\mathrm{I_B}$. This pathway agrees with the observed pathway and kinetics of folding [53]. This sequential pathway is preferred due to the high free-energy barrier between U and I_α, which prevents folding trajectories from branching to I_α. This barrier arises from the large entropy decrease, which brings together the discontinuous parts to form DLD. In other words, the topological complexity of DHFR is the reason for this simple sequential pathway of folding. It should also be noted that the free-energy barrier between N and I_α is predicted to be low, leading to structural fluctuations, including the partial unfolding/folding of the ABD that can be important for the function of DHFR in the native state.

We should note that the topological complexity of DHFR can be resolved by circular permutation. Connecting the N and C termini and disconnecting the linker part of the chain between DLD and ABD, both ABD and DLD become continuous domains comprising continuous parts of the chain. The free energy change due to this circular permutation was calculated by the eWSME model and shown in Fig. 8. This circular permutation increases the free energy at around $\mathrm{I_B}$ and lowers the free energy at the barrier between U and I_α. Then, the kinetic evolution of DHFR molecules' population branches into two pathways, U $\rightarrow$ $\mathrm{I_A}$ $\rightarrow$ $\mathrm{I_B}$ $\rightarrow$ N and U $\rightarrow$ I_α $\rightarrow$ N, as indicated by the Monte Carlo results of Figs. 8B–8D. In this way, the simplification of the DHFR topology through circular permutation brings about the complex folding behavior. This complex folding behavior is consistent with the observed folding kinetics of the circular permutant [54].

Further extension of the WSME model is possible for proteins with more complex topologies, and we here outline this idea. Adenylate kinase (AdK), for example, has three domains: CORE (residues 1-29, 68-117, and 161-214), NMP (residues 30-67), and LID (residues 118-167), as shown in Fig. 6B. We define the virtual ring closures at residues 29 and 68 (closure-1),

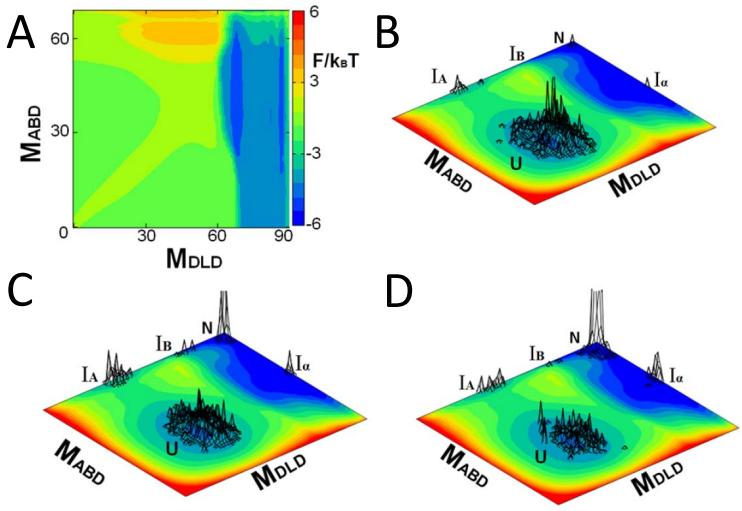

FIG. 8: Free-energy landscape and folding kinetics of the circular permutant of DHFR calculated by the eWSME model. A) Difference in the free-energy landscape between the wild type and the circular permutant of DHFR. B–D) Evolution of the population of 200 molecules simulated with the Monte Carlo calculation at B) $3.3 \times 10^5 t_0$, C) $1.6 \times 10^6 t_0$, and D) $3.0 \times 10^6 t_0$, where t_0 is a unit of time in simulation. Reproduced from [23].

117 and 161 (closure-2), and 1 and 214 (closure-3). The WSME partition function $Z_{\rm ring}(i)$ is calculated by assuming only one closure for i =1, 2, or 3, $Z_{\rm ring}(ij)$ is calculated for two closures with ij =12, 23, or 31, and $Z_{\rm ring}(123)$ is calculated for three closures. Then, $Z_{\rm eWSME}$ is calculable from the WSME Hamiltonian as

$$\begin{aligned} Z_{\rm eWSME} &= Z_{\rm WSME} + (Z_{\rm ring}(1) - Z_{\rm WSME})A_1(1-A_2)(1-A_3) \\ &+ (Z_{\rm ring}(2) - Z_{\rm WSME})A_2(1-A_3)(1-A_1) \\ &+ (Z_{\rm ring}(3) - Z_{\rm WSME})A_3(1-A_1)(1-A_2) \\ &+ (Z_{\rm ring}(12) - Z_{\rm WSME})A_1A_2(1-A_3) + (Z_{\rm ring}(23) - Z_{\rm WSME})A_2A_3(1-A_1) \\ &+ (Z_{\rm ring}(31) - Z_{\rm WSME})A_3A_1(1-A_2) + (Z_{\rm ring}(123) - Z_{\rm WSME})A_1A_2A_3, \end{aligned} \tag{5}$$

where $A_i = \exp(S_{\rm ring}(i)/k_{\rm B})$ is a factor representing the entropy reduction due to the closure-i, which could be estimated by evaluating the probability that the two sites in a Gaussian chain are located at the closure distance from each other, under the constraint of a given

pattern of $\{m_i\}$. In this way, the eWSME model can be directly applied to proteins with various topologies, as exploring folding mechanisms of multi-domain proteins with a unified perspective is an important avenue of the folding studies.

The aWSME Model for Protein Allostery

The classical view of protein folding, wherein folding proceeds along a definite pathway [55], was replaced by the modern energy landscape picture, which describes protein folding as fluctuating diffusive motions over a globally biased energy landscape. Energy landscape methods have shown that the folding pathway and transition state ensemble are determined by the statistical features of the distributed fluctuating trajectories; these methods enabled the quantitative understanding of protein folding and guided methods of protein engineering [8]. The energy landscape perspective should be important not only for protein folding but also for protein conformational change, wherein fluctuations and diversity of trajectories are significant. Particularly, the energy landscape description should be necessary for understanding allosteric transitions [56–58].

An allosteric transition is a change in the distribution of a protein's structure triggered by a chemical or physical perturbation [59], which is often an essential step for proteins to exert their functions. Although the classical view of allosteric transition is based on the picture of a deterministic sequential structural change [60], motions in allosteric transition should bear flexible stochastic fluctuations that may allow diversely different transition trajectories, as in protein folding, which should be quantitatively assessed by energy landscape methods. For this purpose, the WSME model can be extended to describe the energy landscape of allosteric transitions.

Here, we assume that a protein shows two different low-energy conformations in the native state. To be more specific, we consider the case that one is the active (A) conformation, which has the higher affinity to bind a partner protein, and the other is the inactive (I) conformation, which has the lower affinity to bind it. The dominant conformation, around which the protein structure fluctuates, switches from I to A upon binding of a ligand or through chemical modification such as phosphorylation of the protein. We should note that the following theoretical scheme is applicable to cases other than this I-A structural change when the transition between two low-energy conformations is concerned with. We

assume that m_i can take three values, A, I, and D; m_i = A or I when the ith residue takes the configuration similar to that found in the A or I conformation, respectively, and m_i = D, when the residue takes a disordered non-native configuration. Here, for mathematical convenience, to calculate the partition function from the Hamiltonian, we use a redundant expression of either m_i = A or m_i = I for the residue with the configuration common to A and I [31].

The contact patterns in the native conformations are expressed as $\Delta_{ij}^{\rm A}$ and $\Delta_{ij}^{\rm I}$; $\Delta_{ij}^{\rm A(or\,I)} = 1$ when the residues i and j are in contact in the A(or I) conformation and $\Delta_{ij}^{\rm A(or\,I)} = 0$, otherwise. $\Delta_{ij}^{\rm C} = \Delta_{ij}^{\rm A}\Delta_{ij}^{\rm I}$ represents the contact pattern which is common to A and I. $\tilde{\Delta}_{ij}^{A} = \Delta_{ij}^{A}(1-\Delta_{ij}^{C})$ and $\tilde{\Delta}_{ij}^{\rm I} = \Delta_{ij}^{\rm I}(1-\Delta_{ij}^{\rm C})$ are the contact patterns which are specific to A and I, respectively. We define the functions $P_k^{\rm A}(m_k)$, $P_k^{\rm I}(m_k)$, and $P_k^0(m_k)$ by $P_k^{\rm A}({\rm A}) = 1$, $P_k^{\rm A}({\rm I}) = P_k^{\rm A}({\rm D}) = 0$, $P_k^{\rm I}({\rm I}) = 1$, $P_k^{\rm I}({\rm A}) = P_k^{\rm I}({\rm D}) = 0$, and $P_k^0(m_k) = P_k^{\rm A}(m_k) + P_k^{\rm I}(m_k)$. Then, the WSME Hamiltonian for allosteric transition (the aWSME Hamiltonian) is

$$\begin{aligned} H_{\rm aWSME}(\alpha,\{m_i\}) &= V_\alpha(\{m_i\}) \\ &+ \sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\epsilon_{ij}\left(\Delta_{ij}^{\rm C}\prod_{k=i}^{j}P_k^0(m_k) + \tilde{\Delta}_{ij}^{\rm A}\prod_{k=i}^{j}P_k^{\rm A}(m_k) + \tilde{\Delta}_{ij}^{\rm I}\prod_{k=i}^{j}P_k^{\rm I}(m_k)\right), \end{aligned} \tag{6}$$

where α distinguishes the ligand binding/unbinding or the phosphorylation/dephosphorylation and $V_\alpha(\{m_i\})$ represents the local interactions between the bound ligand and surrounding residues or those around the phosphorylated site [31]. The first term in the summation of the right-hand side of Eq. 6 is the energy decrease due to the many-residue correlation to form native-like segments, and the second and third terms represent the energy decrease due to the many-residue correlation to form A and I-like segments, respectively. We define the order parameter n of the folding and the order parameter x of allostery as $n = \sum_{i=1}^{N} P_k^0(m_i)/N$ and $x = M_{\rm A}/N_{\rm A}$, respectively. Here, $M_{\rm A}$ is the number of residues assuming the configuration specific to the A conformation, and $N_{\rm A}$ is the maximal number of M_A, so that $(x,n) = (0,1)$ is the I conformation, $(x,n) = (1,1)$ is the A conformation, and $(x,n) = (0,0)$ is the completely disordered state. The partition function $Z_{\rm aWSME}(\alpha,x,n)$ and the two-dimensional free-energy landscape $F_\alpha(x,n) = -k_{\rm B}T\log Z_{\rm aWSME}$ are exactly calculable from $H_{\rm aWSME}$. See [32] for a more detailed explanation of the model.

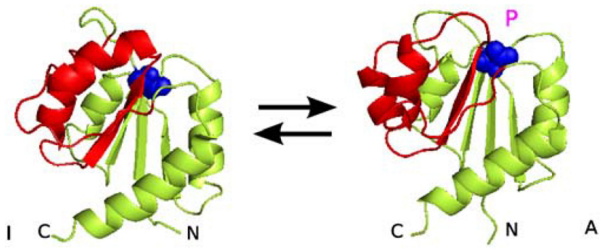

FIG. 9: Allosteric transition of NtrC. Upon phosphorylation of Asp54, the NtrC structure switches from a state around the inactive (I) conformation (PDB code: 1dc7) to another state around the active (A) conformation (PDB code: 1dc8). Asp54 is shown with blue colored spheres. "3445 face" (the region comprises helices and strands, $\alpha 3$, $\beta 4$, $\alpha 4$, and $\beta 5$) is colored red. Reproduced from [31].

Fig. 9 illustrates the allosteric transition of an example protein, the bacterial nitrogen regulatory protein C (NtrC). The distribution of the NtrC structures is dominated by the A conformation, when the residue Asp54 is phosphorylated, and by the I conformation, when dephosphorylated. Fig. 10 shows $F_\alpha(x, n)$ calculated with the aWSME model. Although the most stable structure in $F_{\rm dephos}(x, n)$ is the I conformation at $(x, n) \approx (0, 1)$, a low free-energy valley extends from I to A conformations with metastable basins at $(x, n) \approx (0.2, 0.97)$, (0.55, 0.97), and (0.75, 0.97), demonstrating that the dephosphorylated NtrC should exhibit large structural fluctuation. The NtrC molecules within the valley bear the A-like features, which transiently appear as fluctuations, though the most stable structure is the I conformation. As shown in Fig. 11, this structure fluctuation explains the observed $R_{\rm ex}$ values derived from the R_1, R_2, and the NOE relaxation data of NMR [61].

As shown in Fig. 10, when Asp54 is phosphorylated, a basin that does not exist in $F_{\rm dephos}(x, n)$ appears at $(x, n) = (0.95, 0.97)$ in $F_{\rm phos}(x, n)$. Therefore, the conformation close to A becomes most stable upon phosphorylation. The large fluctuation between A and I in the dephosphorylated state shows that the transition from I to A can be regarded as the selection of pre-existing A-like conformations, but the shift from (0.75, 0.97) to (0.95, 0.97) shows that the "induced-fit" works during the last step of this transition. Thus, the aWSME model reveals that the mixed mechanisms of conformation selection and induced fit regulate the allosteric transition of NtrC.

The large structural fluctuation in the dephosphorylated state is due to the entropic gain

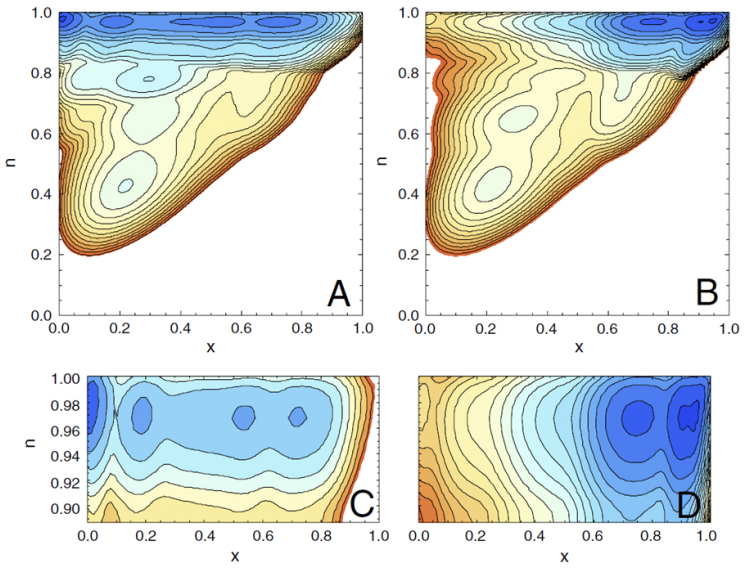

FIG. 10: Free-energy landscape $F_\alpha(x, n)$ of allosteric transition of NtrC calculated with the aWSME model. x is the order parameter of allosteric transition and n is the order parameter of folding transition. $(x, n) = (0, 1)$ is the I conformation, $(1, 1)$ is the A conformation, and $(0, 0)$ is the completely disordered state. A) $F_{\rm dephos}(x, n)$ in the dephosphorylated state and B) $F_{\rm phos}(x, n)$ in the phosphorylated state. C) and D) are closeups of A) and B), respectively, at $n \approx 1$. Contour is drawn for every $2k_{\rm B}T$. Reproduced from [31].

for the intermediate x. In the intermediate x regime, multiple A- or I-like segments coexist in the chain, and a large number of mosaic patterns of these segments are possible; this large number of structures is the reason for the large entropy in this regime. In other words, the multitude of fluctuating trajectories with similar energies is the reason for the flat free-energy landscape and large fluctuation along the x variance with $n \approx 1$. Such entropic gain is not taken into account by conventional simulations based on the classical picture assuming a unique definite transition pathway. Thus, the results of the WSME model reveal the importance of fluctuating movement over the energy landscape. It should be noted that in the problem of allostery, the landscape itself is modified by binding/unbinding of an effector such as the phosphate group, inducing the dynamical transition $F_{\rm dephos} \leftrightarrow F_{\rm phos}$. To emphasize this aspect, we would argue that the "dynamical energy landscape view" is important for analyzing protein allostery and functions.

Finally, we note that the aWSME model can be applied to the folding problem, when

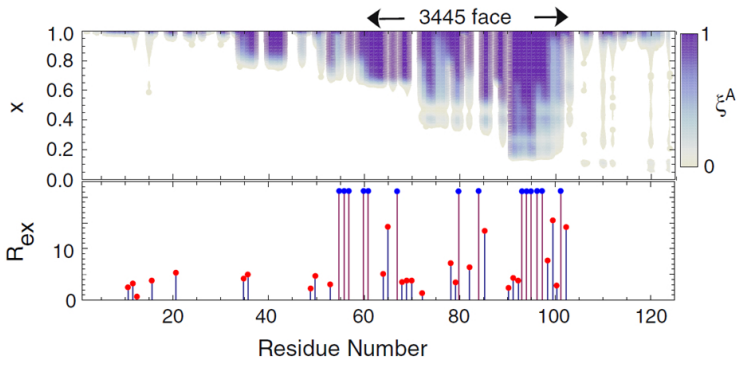

FIG. 11: Pre-existing structural fluctuation of NtrC. (Top) The parameter ξ^{A} showing the extent of the A-like structure development in the dephosphoryated state. ξ^{A} calculated with the aWSME model under the constraint of each fixed x and $n = 1$ is plotted in gray scale. Even in conformations near the I conformation with small x, the A-like structure appears as a fluctuation around the 3445 face. (Bottom) R_{ex} observed in the relaxation measurement of NMR in the dephosphorylated state [61] are shown with red dots. R_{ex} is larger than a threshold for the blue dots [61]. Reproduced from [31].

competition between the native conformation and an off-pathway intermediate state with a distinct non-native structure dominates the folding process [62, 63]. The aWSME model is applicable to this problem using these native and non-native conformations in place of the A and I conformations in the above analysis.

Discussion: Cooperativity and Modularity

Prof. Nobuhiko Saitô emphasized the importance of the hierarchal pathway of protein folding through the WSME model development and the related models of secondary structure formation [64–66]. In this hierarchical picture, "islands" or local native-like contiguous segments are spontaneously formed at the early stage of folding, and folding proceeds through growth and coalescence of these segments through long-range interactions. Saitô suggested that the segments formed first should typically be secondary structure elements (SSEs), such as α-helices or β-strands, and these SSEs are packed with hydrophobic interactions in the later stage of folding [64–66]. However, in many cases, the loop regions include as dense hydrogen-bonds or other interactions as in SSEs such that local structures

including loops can be energetically stabilized similarly to SSEs. Therefore, segments that include loops could also be formed during the early stage of folding. A well-known example of a loop, where the folding reaction initiates, is the distal hairpin loop of src SH3 [67]. The above discussion suggests that we should carefully examine the parts of the protein that fold during the early stage of the folding process. Importantly, the statistical weight of the different folding pathways can be compared with the WSME model by taking account the balance between energy and entropy so that the quantitative comparison between the experiments and the WSME results would facilitate solving this problem.

Local segments, which could be identified as units of a protein's substructure, have been defined and analyzed from several viewpoints. A notable approach is the geometrical analysis; using the contact pattern in the native conformation, "modules" were defined as units of the substructure [41]. Gō showed that the boundaries of these modules coincide with the boundaries of exons of example proteins [42, 43], which suggested that modern proteins were formed through shuffling of modules in the evolutionary history. Barnase, for example, comprises six modules, M1, M2, $\cdots$, M6, and their boundaries are at residues 24, 52, 73, 88, and 98 [44, 45]. In Fig. 5, these module boundaries are compared with the calculated and observed ϕ-values at two transition states, TS1 and TS2. Meanwhile, when we examine an ensemble of numerous protein molecules, those molecules diffusively move on the energy landscape to diversely trace different trajectories so that the transition state, in which the folding nucleus is formed, is not dominated by a unique structure, but should be described as an ensemble of many heterogeneous structures. The ϕ-values represent the average tendency to form the ordered structure at each residue in this transition state ensemble.

We found distinct dips in the calculated ϕ-values at residues 72–73 and 89–90 at TS1, and at 20–23, 46, 72–73, 77–78, and 87–89 at TS2, showing the rough correlation between the module boundaries and the ϕ-value boundaries. Through this comparison, we see that in the nucleus formation in TS1, M1 (residues 1–24) and M2 (residues 25–52) are disordered, M3 (residues 53–73) and M6 (residues 99–110) have small but finite probability of structure formation, and M4 (residues 74–88) and M5(residues 89–98) have intermediate levels of probability of folding. In another stage of nucleus formation in TS2, M1 has an intermediate level of probability of folding, M2 is disordered, and M3–M6 have higher probabilities of folding. Although the correspondence is not exact, this comparison suggests that module-like segments are formed at the transition states of barnase as cooperative structure formation

units.

Energetic analysis is another method to define the subunits. Using a knowledge-based potential, the units of cooperative folding, foldons, were defined as segments that show the maximal energy gap between ordered and disordered structures [68, 69]. For barnase, the foldons' boundaries do not exactly match with those of the modules; however, there is a correlation between them; foldon-1 corresponds to M1, foldon-2 corresponds to M2, and foldon-3 corresponds to a part extending from M3 to M6 [68]. With this terminology, foldon-3 is folded with a large probability, foldon-1 is folded with a modest probability, and foldon-2 is almost unfolded at TS2 of barnase.

Comparing multiple proteins showed that there are correlations among modules, exons, and foldons, but the correspondence is not perfect and deviations specific to proteins were reported [68, 69]. To elucidate the correlation and deviation of these differently defined local segments, the comprehensive comparison of different types of proteins is necessary. As shown in the above discussion, the ϕ-value analysis with the WSME model should be useful for interpreting the results of such a comparison.

At a larger scale, local cooperative structures, foldons or modules, are assembled into the native conformation in a further cooperative way. A question in this scale is how such long-range cooperative assembly is realized. Here, the geometrical analysis sheds light on this problem. One of the present authors developed an efficient non-sequential structure alignment software, MICAN [70], and demonstrated that the spatial arrangement of SSEs of numerous different proteins can be precisely superposed on each other if we disregard both the chain direction in SSEs and the manner those SSEs are connected by chains [70, 71]. An example of a non-sequential structure alignment by MICAN is shown in Fig. 12. Indeed, approximately 80 % of the fold representatives defined in the SCOP database [72] share the same spatial arrangement of SSEs with other folds [71]. Because it is widely accepted that proteins with different folds are very unlikely to be evolutionarily related, this frequent sharing of the same SSE arrangement suggests that particular SSE arrangements were evolutionarily selected as liquid-crystal-like configurations, which satisfy the chemical or physical requirements for interactions. With the same SSE arrangement, the non-local interactions in native conformations can be similarly stable, but local interactions can exhibit significantly different stabilities, depending on the connectivity of the SSEs. In addition, differences in the chain connectivity can modify the entropy reduction process along the folding funnel.

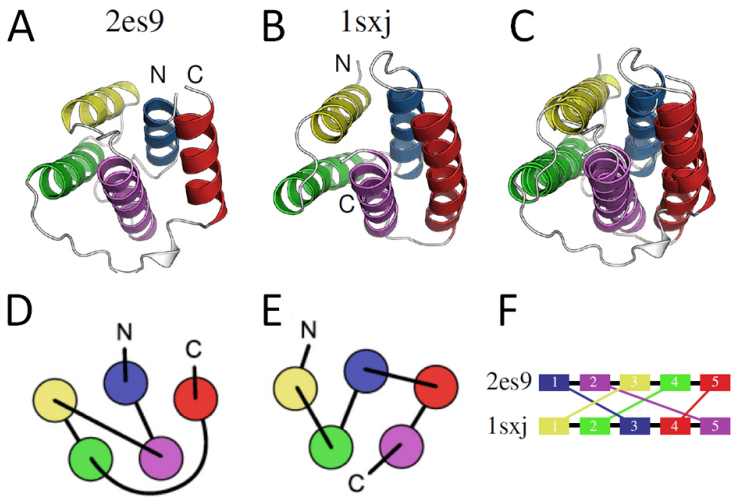

FIG. 12: An example of a non-sequential structure alignment. A) Structure of Q8ZRJ2 (PDB code: 2es9), B) structure of the eukaryotic clamp loader (PDB code: 1sxj), and C) the superimposition of Q8ZRJ2 and the eukaryotic clamp loader obtained by the non-sequential alignment program MICAN [70]. In A–C, the structurally equivalent regions are drawn with the same color. It can be clearly seen that all helices are well superimposed if both the chain direction and the connectivity are ignored. D and E are two-dimensional diagrams of protein topology of Q8ZRJ2 (A) and eukaryotic clamp loader (B), respectively. F) Correspondence relation of helices obtained by MICAN. Reproduced from [73] with permission.

To elucidate the relative importance of local versus non-local interactions as well as the role of entropy in the SSE assembly, it would be interesting to compare folding pathways for a set of proteins that share the same SSE arrangement but have different topologies. For such a purpose, the WSME model would play an important role, as implicated by the successful description of the folding pathways of both the wild type and the circular permutant of DHFR [23].

Conclusively, we address the implications of the coarse-grained modeling studies discussed in this review. Protein folding is a complex molecular process, affected by various atomic interactions; non-native interactions, particularly non-native disulfide bonds, slow down the folding process. Isomerization of proline or other residues affects the folding/unfolding rates. Cooperative exclusion of water molecules and the concomitant hydrophobic packing in each local part affect the height and position of the barrier in the free-energy landscape of folding.

Some of these features, such as the effects of non-native interactions and proline isomerization, have been explicitly considered in the kinetic description using the WSME model [23]. Here, we emphasize that important aspects of these atomic features are represented in a coarse-grained way, which are compatible with the core assumption of the WSME model that is the cooperativity in forming local structural modules and assembling those local structures, as indicated by the agreement between the WSME results and the observed data. Therefore, the analyses of modularity and cooperativity with the WSME model provide guidelines on how to represent the effects of atomic interactions in a coarse-grained way to construct models of complex problems, such as allostery dynamics [57]. Therefore, coarse-graining methods should provide insights on protein evolution, development of techniques for protein structure prediction, and protein engineering. Finally, this approach using simplified statistical mechanical models, which was pioneered by Saitô, should continue to play an important role in this modern field of protein biophysics.

Acknowledgment

This study was supported by JSPS KAKENHI Grant Number JP16H02217, CREST of the Japan Science and Technology Agency, and Riken Pioneering Project "Cellular Evolution".

Conflicts of Interest

The authors declare no competing financial interest.

Author Contribution

M.S., G.C., and T.P.T. co-wrote the manuscript.

[1] Wako, H. & Saitô, N. Statistical mechanical theory of the protein conformation. I. General considerations and the applications to homopolymers. *J. Phys. Soc. Jpn.* **44**,1931–1938 (1978).

[2] Wako, H. & Saitô, N. Statistical mechanical theory of the protein conformation. II. Folding pathway for protein. *J. Phys. Soc. Jpn.* **44**, 1939–1945 (1978).

[3] Lifson, S. & Roig, A. On the theory of helix-coil transition in polypeptides. *J. Chem. Phys.* **34**, 1963–1974 (1961).

[4] Poland, D. & Scheraga, H. A. Phase transitions in one dimension and the helix-coil transition in polyamino acids. *J. Chem. Phys.* **45**, 1456-1463 (1966).

[5] Bryngelson, J. D., Onuchic, J. N., Socci, N. D. & Wolynes, P. G. Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins* **21**, 167–195 (1995).

[6] Fersht, A. Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding (Freeman, New York, 1999).

[7] Fersht, A. Transition-state structure as a unifying basis in protein-folding mechanisms: Contact order, chain topology, stability, and the extended nucleus mechanism. *Proc. Natl. Acad. Sci. USA* **97**, 1525–1529 (2000).

[8] Onuchic, J. N. & Wolynes, P. G. Theory of protein folding. *Curr. Opin. Str. Biol.* **14**, 70–75 (2004).

[9] Daggett, V. & Fersht, A. The present view of the mechanism of protein folding. *Nat. Rev. Mol. Cell Biol.* **4**, 497–502 (2003).

[10] Munôz, 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).

[11] Bruscolini, P. & Pelizzola, A. Exact solution of the Muñoz-Eaton model for protein folding. *Phys. Rev. Lett.* **88**, 258101 (2002).

[12] Pelizzola, A. Exactness of the cluster variation method and factorization of the equilibrium probability for the Wako-Saitô-Muñoz-Eaton model of protein folding. *J. Stat. Mech.*, P11010 (2005).

[13] Karanicolas, J. & Brooks III, C. L. The importance of explicit chain representation in protein folding models: An examination of Ising-like model. *Proteins* **53**, 740–747 (2003).

[14] Henry, E. R. & Eaton, W. A. Combinatorial modeling of protein folding kinetics: free energy

profiles and rates. *Chem. Phys.* **307**, 163–185 (2004).

[15] 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).

[16] Itoh, K. & Sasai, M. Cooperativity, connectivity, and folding pathways of multidomain proteins. *Proc. Natl. Acad. Sci. USA* **105**, 13865–13870 (2008).

[17] 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).

[18] Nelson, E. D. & Grishin, N. V. Folding domain B of protein A on a dynamically partitioned free energy landscape. *Proc. Natl. Acad. Sci. USA* **105**, 1489–1493 (2008).

[19] Yu, W., Chung, K, Cheon, M., Heo, M., Kyou-Hoon Han, K.-H., Ham, S. & Chang, I. Cooperative folding kinetics of BBL protein and peripheral subunit-binding domain homologues. *Proc. Natl. Acad. Sci. USA* **105**, 2397–2402 (2008).

[20] Itoh, K. & Sasai, M. Multidimensional theory of protein folding. *J. Chem. Phys.* **130**, 145104 (2009).

[21] Henry, E. R., Best, R. B. & Eaton, W. A. Comparing a simple theoretical model for protein folding with all-atom molecular dynamics simulations. *Proc. Natl. Acad. Sci. USA* **110**, 17880–17885 (2013).

[22] 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).

[23] 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).

[24] Zamparo, M. & Pelizzola, A. Rigorous results on the local equilibrium kinetics of a protein folding model. *J. Stat. Mech.*, P12009 (2006).

[25] Zamparo, M. & Pelizzola, A. Kinetics of the Wako-Saitô-Muñoz-Eaton model of protein folding. *Phys. Rev. Lett.* **97**, 068106 (2006).

[26] Imparato, A., Pelizzola, A., Zamparo, M. Ising-like model for protein mechanical unfolding. *Phys. Rev. Lett.* **98**, 148102 (2007).

[27] Imparato, A. & Pelizzola, A. Mechanical unfolding and refolding pathways of ubiquitin. *Phys. Rev. Lett.* **100**, 158104 (2008).

[28] Zamparo, M., Trovato, A. & Maritan, A. Simplified exactly solvable model for β-amyloid aggregation. *Phys. Rev. Lett.* **105**, 108102 (2010).

[29] Itoh, K. & Sasai, M. Dynamical transition and proteinquake in photoactive yellow protein. *Proc. Natl. Acad. Sci. USA* **101**, 14736–14741 (2004).

[30] Itoh, K. & Sasai, M. Coupling of functioning and folding: photoactive yellow protein as an example system. *Chem. Phys.* **307**, 121–127 (2004).

[31] Itoh, K. & Sasai, M. Entropic mechanism of large fluctuation in allosteric transition. *Proc. Natl. Acad. Sci. USA* **107**, 7775–7780 (2010).

[32] Itoh, K. & Sasai, M. Statistical mechanics of protein allostery: roles of backbone and side-chain structural fluctuations. *J. Chem. Phys.* **134**, 125102 (2011).

[33] Gō, N. Theoretical studies of protein folding. *Annu. Rev. Biophys. Bioeng.* **12**, 183–210 (1983).

[34] Taketomi, A., Ueda, Y. & Gō, N. Studies on protein folding, unfolding and fluctuations by computer simulations 1. The effect of specific amino acid sequence represented by specific inter-unit interactions. *Int. J. Peptide Protein Res.* **7**, 445–459 (1975).

[35] Gō, N. & Abe, H. Noninteracting local-structure model of folding and unfolding transition in globular proteins. I. Formulation. *Biopolymers* **20**, 991–1011 (1981).

[36] 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).

[37] 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).

[38] Wolynes, P. G. Latest folding game results: Protein A barely frustrates computationalists. *Proc. Natl. Acad. Sci. USA* **101**, 6837–6838 (2004).

[39] 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).

[40] Salvatella, X., Dobson, C. M. & Fersht, A. R. , Vendruscolo, M. Determination of the folding transition states of barnase by using Φ_{I}-value-restrained simulations validated by double mutant Φ_{IJ}-values. *Proc. Natl. Acad. Sci. USA* **102** 12389–12394 (2005).

[41] Gō, M. Modular structural units, exons, and function in chicken lysozyme *Proc. Natl. Acad. Sci. USA* **80**, 1964–1968 (1983).

[42] Gō, M. & Nosaka, M. Protein architecture and the origin of introns. *Cold Spring Harb. Symp. Quant. Biol.* **52**, 915–924 (1987).

[43] Gō, M. Correlation of DNA exonic regions with protein structural units in haemoglobin. *Nature* **291**, 90–92 (1981).

[44] Yanagawa, H. Yoshida, K., Torigoe, C., Park, J.-S., Sato, K., Shirai, T. & Gō, M. Protein anatomy: Functional roles of barnase module. *J. Biol. Chem.* **268**, 5861–5865 (1993).

[45] Noguti, T., Sakakibara, H. & Gō, M. Localization of hydrogen-bonds within modules in barnase. *Proteins* **16**, 357–363 (1993).

[46] Best, R. B., Hummer, G. & Eaton, W. A. Native contacts determine protein folding mechanisms in atomistic simulations. *Proc. Natl. Acad. Sci. USA* **110**, 17874–17879 (2013).

[47] Lindorff-Larsen, K., Piana, S., Dror, R. O. & Shaw, D. E. How fast-folding proteins fold. *Science* **334**, 517–520 (2011).

[48] Piana, S., Lindorff-Larsen, K. & Shaw, D. E. Protein folding kinetics and thermodynamics from atomistic simulation. *Proc. Natl. Acad. Sci. USA* **109**, 17845?17850 (2012).

[49] Piana, S., Lindorff-Larsen, K. & Shaw, D. E. Atomic-level description of ubiquitin folding. *Proc. Natl. Acad. Sci. USA* **110**, 5915–5920 (2013).

[50] Borgia, A., Wensley, B. G., Soranno, A., Nettels, D., Borgia, M. B., Hoffmann, A., Pfeil, S. H., Lipman, E. A., Clarke, J. & Schuler, B. Localizing internal friction along the reaction coordinate of protein folding by combining ensemble and single-molecule fluorescence spectroscopy. *Nat. Commun.* **3**, 1195 (2012).

[51] Ferreiro, D. U., Komives, E. A. & Wolynes, P. G. Frustration in biomolecules. *Quart. Rev. Biophys.* **47** 285–363 (2014).

[52] Arai, M. & Kuwajima, K. Role of the molten globule state in protein folding. *Adv. Protein Chem.* **53**, 209–282 (2000).

[53] Arai, M., Iwakura, M., Matthews, C. R. & Bilsel, O. Microsecond subdomain folding in dihydrofolate reductase. *J. Mol. Biol.* **410**, 329–342 (2011).

[54] Texter, F. L., Spencer, D. B., Rosenstein, R. & Matthews, C. R. Intramolecular catalysis of a proline isomerization reaction in the folding of dihydrofolate reductase. *Biochemistry* **31**, 5687–5691 (1992).

[55] Baldwin, R. L. The nature of protein folding pathways: The classical versus the new view. *J. Biomolecular NMR*, **5**, 103–109 (1995)

[56] Boehr, D. D., McElheny, D., Dyson, H. J. & Wright, P. E. The dynamic energy landscape of dihydrofolate reductase catalysis. *Science* **313**, 1638–1642 (2006).

[57] Terada, T. P., Kimura, T. & Sasai, M. Entropic mechanism of allosteric communication in conformational transitions of dihydrofolate reductase. *J. Phys. Chem. B* **117**, 12864–12877 (2013).

[58] Tsai, C.-J. & Nussinov, R. A unified view of "How allostery works". *PLoS Comput. Biol.* **10**, e1003394 (2014).

[59] Motlagh, H. N., Wrabl, J. O., Li, J. & Hilser, V. J. The ensemble nature of allostery. *Nature* **508**, 331–339 (2014).

[60] Vreede, J., Juraszek, J. & Bolhuis, P. G. Predicting the reaction coordinates of millisecond light-induced conformational changes in photoactive yellow protein. *Proc. Natl. Acad. Sci. USA* **107**, 2397–2402 (2010).

[61] Volkman, B. F., Lipson, D., Wemmer, D. E. & Kern, D. Two-state allosteric behavior in a single-domain signaling protein. *Science* **291**, 2429–2433 (2001).

[62] Hamada, D., Segawa, S. & Goto, Y. Non-native α-helical intermediate in the refolding of bold β-lactoglobulin, a predominantly bold β-sheet protein. *Nat. Struct. Biol.* **3**, 868–873 (1996)

[63] Borgia, A., Kemplen, K. R., Borgia, M. B., Soranno, A., Shammas, S., Wunderlich, B., Nettels, D., Best, R. B., Clarke, J. & Schuler, B. Transient misfolding dominates multidomain protein folding. *Nat. Commun.* **6**, 8861 (2015).

[64] Wako, H., Saitô, N. & Scheraga, H. A. Statistical mechanical treatment of α-helices and extended structures in proteins with inclusion of short- and medium-range interactions. *J. Protein Chem.* **2**, 221–249 (1983).

[65] Saitô, N., Shigaki, T., Kobayashi, Y. & Yamamoto, M. Mechanism of protein folding: I. General considerations and refolding of myoglobin. *Proteins* **3**, 199–207 (1988).

[66] Saitô, N. & Kobayashi, Y. Physical foundation of protein architecture. *Int. J. Modern Phys. B* **13**, 2431–2529 (1999).

[67] Grantcharova, V. P., Riddle, D. S., Santiago, J. V. & Baker, D. Important role of hydrogen bonds in the structurally polarized transition state for folding of the src SH3 domain. *Nat. Struct. Biol.* **5**, 714–720 (1998).

[68] Panchenko, A. R., Luthey-Schulten, Z. & Wolynes, P. G. Foldons, protein structural modules, and exons. *Proc. Natl. Acad. Sci. USA* **93**, 2008–2013 (1996).

[69] Panchenko, A. R., Luthey-Schulten, Z., Cole, R. & Wolynes, P. G. The foldon universe: A survey of structural similarity and self-recognition of independently folding units. *J. Mol. Biol.* **272**, 95–105 (1997).

[70] Minami, S., Sawada, K. & Chikenji, G. MICAN: a protein structure alignment algorithm that can handle Multiple-chains, Inverse alignments, Cα only models, Alternative alignments, and Non-sequential alignments. *BMC Bioinform.* **14**, 24 (2013).

[71] Minami, S., Sawada, K. & Chikenji, G. How a spatial arrangement of secondary structure elements is dispersed in the universe of protein folds. *PLoS ONE* **9**, e107959 (2014).

[72] Andreeva, A., Howorth, D., Chandonia, J., Brenner, S., Hubbard, T., Chothia, C & Murzin, A. G. Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res.* **36**, D419–D425 (2008).

[73] Minami, S. *Ph.D thesis. Nagoya University* (2015).