International Journal of Modern Physics B, Vol. 16, No. 13 (2002) 1807–1839

STATISTICAL MECHANICAL FOUNDATION FOR THE TWO-STATE TRANSITION IN PROTEIN FOLDING OF SMALL GLOBULAR PROTEINS

KAZUMOTO IGUCHI

70-3 Shinhari, Hari, Anan, Tokushima 774-0003, Japan
kazumoto@stannet.ne.jp.

Received 31 December 2001
Revised 31 January 2002

We discuss the statistical mechanical foundation for the two-state transition in the protein folding of small globular proteins. In the standard arguments of protein folding, the statistical search for the ground state is carried out from astronomically many conformations in the configuration space. This leads us to the famous Levinthal's paradox. To resolve the paradox, Gō first postulated that the two-state transition — all-or-none type transition — is very crucial for the protein folding of small globular proteins and used the Gō's lattice model to show the two-state transition nature. Recently, there have been accumulated many experimental results that support the two-state transition for small globular proteins. Stimulated by such recent experiments, Zwanzig has introduced a minimal statistical mechanical model that exhibits the two-state transition. Also, Finkelstein and coworkers have discussed the solution of the paradox by considering the sequential folding of a small globular protein. On the other hand, recently Iguchi have introduced a toy model of protein folding using the Rubik's magic snake model, in which all folded structures are exactly known and mathematically represented in terms of the four types of conformations: cis-, trans-, left and right gauche-configurations between the unit polyhedrons. In this paper, we study the relationship between the Gō's two-state transition, the Zwanzig's statistical mechanics model and the Finkelstein's sequential folding model by applying them to the Rubik's magic snake models. We show that the foundation of the Gō's two-state transition model relies on the search within the equienergy surface that is labeled by the contact order of the hydrophobic condensation. This idea reproduces the Zwanzig's statistical model as a special case, realizes the Finkelstein's sequential folding model and fits together to understand the nature of the two-state transition of a small globular protein by calculating the physical quantities such as the free energy, the contact order and the specific heat. We point out the similarity between the liquid-gas transition in statistical mechanics and the two-state transition of protein folding. We also study morphology of the Rubik's magic snake models to give a prototype model for understanding the differences between α-helices proteins and β-sheets proteins.

PACS number(s): 87.10.+e, 87.15.By, 64.60.Cn

1. Introduction

Protein folding problem is one of the most challenging problems in biophysics.[1] A denatured protein can fold into a native three-dimensional compact structure whose

information is encoded only in the amino acid sequence — the *Anfinsen's dogma*.[2] And the unfolded protein with randomly coiled polypeptide chains adapts the information to refold quickly and accurately into the unique native structure from astronomically many conformations in the configuration space — the *Levinthal's paradox*.[3] Stimulated by these works,[2,3] experiments in those days showed that small globular proteins can fold from the denatured state into the native state by a *two-state transition* (TST), although the pathways of the folding process were not determined precisely.[4]

About 25 years ago, Gō[4] first theoretically pointed out that the protein folding of real globular proteins follows not the Levinthal's paradox but the TST or the *all-or-none type transition*. The all-or-none type transition means a discontinuous transition such that at the midpoint of such a transition, the system consists of a 1:1 mixture of the native and denatured conformations, and not a mixture of partially denatured conformations with one half of a molecule remaining native and the other half becoming disordered. He asked the origin of the qualitative difference between the *helix-coil transition* (*HCT*)[5] and the folding and unfolding transitions in globular proteins, where the HCT is a continuous transition from random coil to helix structure such that the intermediate state of a partially coiled molecule can exist. Comparing with the concept of the HCT, using basic knowledge of biochemistry[1] and statistical mechanics,[6] he proposed that *the transitions in globular proteins are the TST type* such that the intermediate state is not so important.

To derive the standpoint of the TST, he and coworkers intensively studied the so-called *Gō's lattice model*.[7] And they found that if only native interactions are kept in interactions between the segments then the TST appears, while if the native and non-native interactions are included then the TST is not so sharp. Thus, they showed that the presence of the native interactions is very crucial and the non-native interactions are not so important for the TST of small globular proteins. The main conclusions that Gō drew were the following: The Gō's model is the one that the more the system approaches the native structure, the more the free energy of the system decreases. And therefore, by the competition between the energy and the entropy, an energy barrier may appear in between the native and denatured states in the folding process. Hence, this situation accounts for the TST of the protein folding.

In spite of his mentioning the importance of the TST and the native interactions,[4] at that time his model seemed to others to be recognized as an over-simplified and very unrealistic model so that his model received no much popularity nor reputation.[4] Very unfortunately to us, he left the subject of protein folding with summarizing what he has done and what he thought on the protein folding.[8]

On the contrary, after his work,[4,7,8] there have appeared many researches from both theoretical[9–19] and experimental sides[20,21] for over two decades, where the Gō's lattice model became a standard approach for the theoretical study of protein folding.[9–19,22] And, the theorists have concentrated to try to answer the Anfinsen's

dogma[2] and Levinthal's paradox,[3] using the Gō's lattice model.[4,7] These efforts have made the consensus that as first mentioned by Gō[4] such TST is relevant for the real protein folding problem and that the Levinthal's paradox might be resolved by the idea of the TST or by the concept of the so-called *funnel structure* of the free energy surface via configuration space, which has been introduced by Wolynes[13] and widely used for the study of protein folding by many authors.[11–13,22]

There have been mainly two categories of theoretical models: the *framework model*[9,10] and the *hydrophobic collapse model*.[11] In the former we assume *a priori* that the formation of each secondary structural elements such as α-helices and β-sheets or their mixture found in the native state precedes that of the tertiary structure, while in the latter we assume *a priori* that the hydrophobic condensation initially occurs without the formation of the secondary structures of α-helices and β-sheets or their mixture. However, by the above theoretical approaches one cannot answer which pathway is realized in the real protein folding. Therefore, one must usually choose one of them by hand as an assumption.

In 1991, Jackson and Fersht[23] discovered that a small protein — the 64-residue chymotrypsin inhibitor 2 (CI2) — folds and unfolds via simple two-state kinetics and that no intermediate accumulates at equilibrium or in the folding pathway. It was a remarkable fact that the TST might play a key role for the protein folding. This changed the situation in the study of the TST so that the Gō's model and the concept of the TST became the main theme in the protein folding problem. In this direction, many small proteins with less than 100 residues have been studied, using the so-called Φ-*value analysis*[24] and β-*analysis*.[23,25,26] And it was expected that the TST is a universal nature in the systems and raised a question that *the principles of the protein folding might be much simpler than what we have usually expected*.[27] Thus, by the quantitative agreements between the experimental results and the theoretical studies on the TST, the so-called *new view* of the protein folding has been established.[8]

More recently, Akiyama *et al.*[28] have found an evidence that even for a large protein with more than 100 residues such as Cytochrome *c*, the hydrophobic collapse precedes the formation of the secondary structure such as α-helices and β-sheets or their mixture. And it is true even for the *tandem protein* that the *N*- and *C*-terminals are linked to make a closed loop of a protein. Thus, the TST nature of hydrophobic condensation preceding to the formation of a secondary structure seems very important in nature.

Stimulated by the recent experimental progresses some theorists have studied the origin of the TST.[27,29–32] The main idea is that the TST is realized as a consequence of the search from *mismatches from the unique native three-dimensional structure*. Therefore, one needs not search the entire configuration space for astronomically many conformations but can take into account only the deviation from the native structure. Hence, it resolves the Levinthal's paradox. This idea goes back

to the original Gō's idea.[4,7,8] It is quite similar to the philosophy in the simple statistical mechanical model that exhibits the TST, which is known as the *Zwanzig model*.[29] This is also realized if the protein folding obeys the *sequential folding* as studied by Finkelstein and coworkers.[30]

Meantime, Baker and collaborators[27] have introduced the concept of the so-called *relative contact order* (RCO) in order to measure the relative importance of local and nonlocal contacts to a protein's native structure — the topology of connectivity of local and nonlocal interactions in the native structure. And they have related it to derive the relationship between the folding rate and the RCO, applying to the TST of small globular proteins. However, the RCO is different from the *contact order* (CO) introduced by Gō and collaborators,[4,8,7] which describes the deviation from the native structure, although they are related to each other in some respects.

On the other hand, Iguchi[33] has recently introduced a toy model of protein folding using the Rubik's magic snake model (RMSM), in which all structures such as the native and denatured structures are exactly known and mathematically represented in terms of the four types of conformations: cis-, trans-, left and right gauche- configurations between the unit polyhedrons. Using the RMSM, he has found that there are three types of the folded structures classified by the concept of chirality $C(= 0, \pm 1)$, where the folded structure of $C = 0$ is self-dual and the folded structure of $C = 1$ is the mirror image of that of $C = -1$, vice versa. He has also studied the ground state energy of the system in order to compare which of the folded structures takes the lowest energy and what gives rise to the energy gap between the folded structures and the unfolded structure. Since the RMSM is not an on-lattice model but a kind of off-lattice model, the conformations are very similar to real small globular proteins. Therefore, this model seems to have the advantage to consider the real folding problem that is not treated appropriately by the simplified lattice models.

The purpose of this paper is to give the statistical mechanical foundation for the TST in the small globular proteins, generalizing the Gō's TST idea and to combine it with the simple statistical mechanical models such as the Zwanzig model, the RMSM,[33] and the lattice models such as the Gō's lattice model.[9−19] By this approach using the RMSM, we will determine the contents of the parameters that are *a priori* adopted in the Zwanzig model. And we apply the Zwanzig model to the RMSM in order to obtain the free energy, the contact order and the specific heat in the folding problem of the RMSM. This may provides us an insight when we consider the real protein folding such as the morphology of protein folding. In the unfolding and folding transition of a real globular protein there are many important factors such as (i) the hydrogen bond interaction, (ii) the Coulomb interaction, (iii) the van der Waals interaction, (iv) the hydrophilic and hydrophobic interactions, (v) the conformational stiffness of the molecule. All of them may dominate the collapse of a protein. However, we take care of only (iv) and (v) in this paper for the sake of simplicity.

The organization of the present paper is the following. In Sec. 2, we introduce the Go's phenomenological theory for the TST.[4] In Sec. 3, we present the statistical formulation as a foundation for the Gō's phenomenological theory. In Sec. 4, we deduce the Zwanzig model from our formulation and discuss the basic properties of the Zwanzig model.[29] In Sec. 5, we give a generalization of the Zwanzig model. In Sec. 6, we discuss the RMSM[33] and using it, we give the explicit meaning of the parameters adopted in the Zwanzig model. We apply the Zwanzig statistical model for folding of the RMSM. In Sec. 7, we discuss the Gō's lattice model in the language of the TST. In Sec. 8, we discuss the problem of whether or not the hydrophobic condensation precedes the formation of the secondary structures in the course of protein folding. In Sec. 9, we discuss the equation of state for protein folding. In Sec. 10, we discuss the morphology of folded proteins. In Sec. 11, conclusion is made.

2. Gō's Phenomenological Theory

Following the Gō's discussion,[4] let $Q \equiv \{q_1, q_2, \cdots, q_N\}$ be a set of variables that describe a microscopic or instantaneous conformation of a protein molecule. Then, the partition function Z is given by

$$Z = \int dQ e^{-\beta[H(Q)-TS(Q)]}\,, \tag{1}$$

where $H(Q)$ $[S(Q)]$ means the enthalpy (entropy) of the system with the conformation Q and $\beta = 1/k_{\rm B}T$ with $k_{\rm B}$ the Boltzmann constant and T the temperature. To find the maximal contribution in the integrand, we must consider the entire search of the configuration space of Q, which leads us to the Levinthal's paradox.[3] Instead, Gō first postulated that there exist thermal fluctuations of Q that conserve the enthalpy surface $H = H(Q)$. Substituting

$$\int d(\beta H)\delta[\beta(H - H(Q))] = 1\,, \tag{2}$$

into Eq. (1) and integrating out over Q, Eq. (1) can be rewritten as $Z = \int d(\beta H)e^{-\beta[H-TS(H)]}$, where $S(H)$ means the entropy of the system with the enthalpy H, defined by

$$e^{S(H)/k_{\rm B}} = \int dQ\delta[\beta(H - H(Q))]e^{S(Q)/k_{\rm B}}\,. \tag{3}$$

Assuming simple *concave curve* property with positive and negative curvature for $S(H)$ and so is for $\Phi(H) \equiv S(H) - H/T$, Gō found possibilities of the TST and the non TST, respectively. The main reason why Gō was able to represent the entropy S as a function of the enthalpy H such as $S = S(H)$ is as follows: As a consequence of *consistency principle* of local and nonlocal interactions,[4,8,7] *if there is a tendency that the energy monotonically decreases as the protein structure approaches the native protein structure, then the energy can be regarded as an order parameter for the system.* However, he could not explicitly evaluate the partition

function and therefore the physical meaning of his enthalpy H as well as $\Phi(H)$ was not so clear.[4] Hence, he called the above theory a phenomenological theory.

Later, to show the above point more explicitly, Gō and collaborators,[7] numerically calculated the entropy S in terms of the energy count m as $S = S(m)$ and found that the entropy $S(m)$ is given by

$$S(m)/k_{\mathrm{B}} = \ln \binom{N}{m} - (N-m)\ln 2\delta\,, \tag{4}$$

where $N \equiv n-2$ is the number of independent bond angles of the chain, $H \equiv -m\varepsilon_0$ is the enthalpy in units of ε_0 and δ is the effect of conformation between the segments in order to take care of the self-avoidance — the effective coordination number or the effective number of conformation between the nearest neighbor segments. Hence, Eq. (4) provides the partition number $W(m)$ as

$$W(m) = e^{S(m)/k_{\mathrm{B}}} = \binom{N}{m}(2\delta)^{-(N-m)}\,. \tag{5}$$

Thus, if the partition number $W(m)$ has the form of Eq. (5), then the statistical mechanics of the model may exhibit the TST property. This is the heart of the Gō's phenomenological theory.

3. Formulation

We think that the heart of his discussion above can be rephrased as a hypothesis that *there exist thermal fluctuations conserving an equienergy surface*. To understand this, we would like to revisit the concept of the sequential folding first studied by Finkelstein and Badretdinov.[30] Finkelstein and coworkers discussed that *a sequential folding is realized if no free energy barrier exists at each folding process* as follows: Denote by n the nth step in the folding process. Suppose that $\Delta S(n) < 0$ and $\Delta E(n) < 0$, where $\Delta S(n) < 0$ means that the entropy is decreasing one after another of the folding process as the chain is compactified to approach the native structure and $\Delta E(n) < 0$ means that the energy decreases by the formation of contacts between the segments as the chain is compactified. Since the free energy is given by $F = E - TS$, the free energy difference between the stages of the folding process is given by $\Delta F(n) = \Delta E(n) - T\Delta S(n)$. $\Delta F(n)$ increases if $\Delta S(n) < 0$, while $\Delta F(n)$ decreases if $\Delta E(n) < 0$. Therefore, the decrease in entropy is cancelled by the decrease in energy such that no energy barrier exists at each folding process. This provides a pathway for a sequential folding.

This logic is essentially the same as the idea of Gō in the previous section, if the step n of folding process can be regarded as the energy count n for the enthalpy H or if the energy $E(n)$ and the entropy $S(n)$ are monotonically decreasing as the step n of folding process is increasing. This is also realized if at the nth step of folding process there are n mismatches from the native structure, then the energy difference from the native structure can be represented as $E(n)$ and so is for the entropy $S(n)$. As was discussed by Zwanzig,[29] if there are n contacts at the nth stage of the folding

process, then the system statistically searches all conformations with keeping the number n of contacts as a conserved quantity since the energy conservation law acts during the search. In other words, this search for the conformations can be thought of as thermal fluctuations on the equienergy surface $E = E(n)$ of n contacts over the configuration space. This provides the form of the partition number as Eq. (5). Hence, the sequential folding model of Finkelstein and coworkers may also exhibit the TST property. Thus, the idea of Gō's all-or-none type transition,[4,7] the idea of Finkelstein's sequential folding[30] and the idea of Zwanzig's mismatch model[29] are the same.

Let us give a statistical mechanical foundation for the Gō's idea as well as the Finkelstein's and Zwanzig's ideas. Since the equienergy surface is not necessarily to be enthalpy, we can generalize his idea to an equienergy surface of $E(M) = E(Q;M)$, which means that fluctuations occur on the equienergy surface with conserving the number of mismatches or incorrectness M from the native structure as discussed by Gō,[4,7] Zwanzig,[29] and Finkelstein.[30] Following the same argument above, let us substitute instead of Eq. (2)

$$\sum_{M} \int d(\beta E(M))\delta[\beta(E(M) - E(M,Q))] = 1 \tag{6}$$

into Eq. (1) replaced $H(Q)$ by $E(Q)$, we obtain

$$Z = \sum_{M=0}^{N} W(M)\nu^{M} e^{-\beta E(M)} \equiv \sum_{M=0}^{N} e^{-\beta F(M)}\,, \tag{7}$$

where $F(M) \equiv E(M) - TS(M)$ such that $\nu^{M}W(M) = \exp[S(M)/k_{\mathrm{B}}]$, N is the maximum number of mismatches, and ν the number of ways that a mismatch occurs. This has a desired form of Gō, of Zwanzig and of Finkelstein. And it has also a quite analogous form to that of the grand partition function for a gas in statistical mechanics[6] where ν plays the role of the fugacity such as $z = e^{\beta\mu}$, μ the chemical potential of a particle. Therefore, we can easily derive the free energy F and the averaged number of the mismatches $\langle M\rangle$:

$$\beta F = -\frac{1}{N}\ln Z, \qquad \langle M\rangle = \nu\frac{\partial}{\partial\nu}\ln Z\,, \tag{8}$$

respectively. Here the expectation value $\langle M\rangle$ can be a good measure — the *order parameter* — for the protein folding. This was first defined by Gō and Taketomi.[7] If we define the *contact order*[4,7] as

$$\langle \mathrm{CO}\rangle = 1 - \frac{\langle M\rangle}{N}\,, \tag{9}$$

then $\langle \mathrm{CO}\rangle$ can be a good measure for considering the contact nature of the protein folding such that when there is no mismatch, we have $\langle \mathrm{CO}\rangle = 1$, while when there are the maximum number of mismatches, we have $\langle \mathrm{CO}\rangle = 0$. Thus, we show that the idea of Gō provides a good statistical mechanical foundation for the protein folding problem.

We note here that as we have claimed in the introduction, the contact order $\langle \mathrm{CO} \rangle$ that we are going to study in this paper is different from the relative contact order RCO that is recently introduced by Baker and coworkers.[27] Here, to make the distinction and to escape from confusion we have used the term RCO for their relative contact order although they used very misleadingly the term CO for their relative contact order.[27] This point will be discussed in detail in Sec. 7.

4. Zwanzig Model

Let us consider some examples. Let us assume that the energy $E(M)$ is given by

$$E(M) = MU - \varepsilon \delta_{M,0} \,, \tag{10}$$

where U is the energy cost for a mismatch and ε is an energy gap between the native state and the state with one mismatch of $M = 1$. This means that the energy $E(M)$ is not the Haqqmiltonian of the system but the energy difference between the native state and the deformed structures. If we ignore the interaction between mismatches in the deformed structures, then we can define

$$W(M) = \begin{pmatrix} N \\ M \end{pmatrix} = \frac{N!}{(N-M)!M!} \,, \tag{11}$$

then the partition function of Eq. (7) coincides with that of Zwanzig model, which is known to provide the TST property.[29] We now have

$$Z = e^{\beta\varepsilon} + (1 + \nu e^{-\beta U})^N - 1 \equiv Z_N \,, \tag{12}$$

$$\langle M \rangle = N \frac{\nu e^{-\beta U} (1 + \nu e^{-\beta U})^{N-1}}{e^{\beta\varepsilon} + (1 + \nu e^{-\beta U})^N - 1} \,. \tag{13}$$

Substituting Eq. (13) into Eq. (9), we obtain the contact order

$$\langle \mathrm{CO} \rangle = \frac{Z_{N-1}}{Z_N} \,. \tag{14}$$

This exhibits the TST behavior as follows: From Eq. (12), we find that if $T \to 0$ then $Z_N \to e^{\beta\varepsilon}$ and hence $\langle \mathrm{CO} \rangle \to 1$, and if $T \gg 1$ then $Z_N \to (1 + \nu)^N$ and hence $\langle \mathrm{CO} \rangle \to 1/(1 + \nu)$. The behavior of the exact contact order $\langle \mathrm{CO} \rangle$ is shown in Fig. 1(A) for the values of $\nu = 3$, $\varepsilon = 24$ and $U = 6$.

Let us next consider the free energy $F(M)$ that is dependent of the number of mismatches. By definition, we have

$$\begin{aligned} F(M) &= MU - \varepsilon \delta_{M,0} - k_{\mathrm{B}} T \ln(\nu^M W(M)) \\ &= MU - \varepsilon \delta_{M,0} - M k_{\mathrm{B}} T \ln \nu - k_{\mathrm{B}} T \ln \left[\frac{N!}{M!(N-M)!} \right] . \end{aligned} \tag{15}$$

By using the Stirling formula $N! \approx N(\ln N - 1)$, we obtain

$$\begin{aligned} F(M)/N &= \rho U - \frac{\varepsilon}{N} \delta_{\rho,0} - \rho k_{\mathrm{B}} T \ln \nu + k_{\mathrm{B}} T (\rho \ln \rho + (1 - \rho) \ln(1 - \rho)) \\ &\equiv f(\rho) \,, \end{aligned} \tag{16}$$

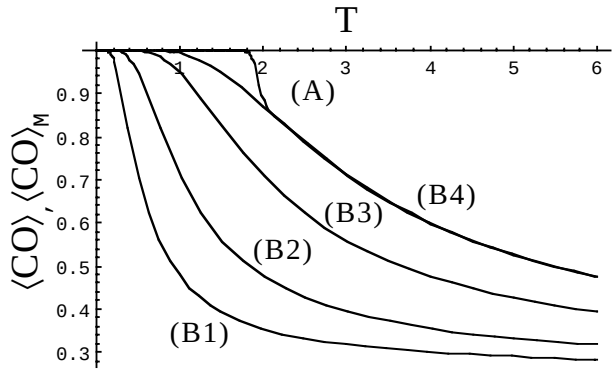


Fig. 1. The TST behavior of the contact order parameters of the exact $\langle \mathrm{CO} \rangle$ and $\langle \mathrm{CO} \rangle_M$. (A) The exact $\langle \mathrm{CO} \rangle$ given by Eq. (14) is shown for $\nu = 3$, $\varepsilon = 24$ and $U = 6$. (B1)–(B4) $\langle \mathrm{CO} \rangle_M$ given by Eq. (19) are shown for $U = 1, 2, 4, 6$, respectively. The horizontal axis means temperature T.

where $\rho \equiv M/N$. This starts with ε/N at $\rho = 0$, discontinuously goes up to make a barrier, and goes down and up continuously such that there is a minimum at $\rho = \rho_{\mathrm{M}}$. The behavior of this is shown in Fig. 2. There, the cases of $T = 0.9, 0.95, 1.0, 1.05, 1.1$ are drawn for $U = 2$, $\nu = 3$, $\varepsilon = 24$ and $N = 100$, respectively.

Let us find the minimum ρ_{M}. By taking the extremum condition $\frac{\delta f(\rho)}{\delta \rho} = 0$, we obtain

$$\rho_{\mathrm{M}} = \frac{1}{1 + \nu^{-1} e^{\beta U}} \,, \tag{17}$$

$$\langle \mathrm{CO} \rangle_{\mathrm{M}} \equiv 1 - \rho_{\mathrm{M}} = \frac{1}{1 + \nu e^{-\beta U}} \,. \tag{18}$$

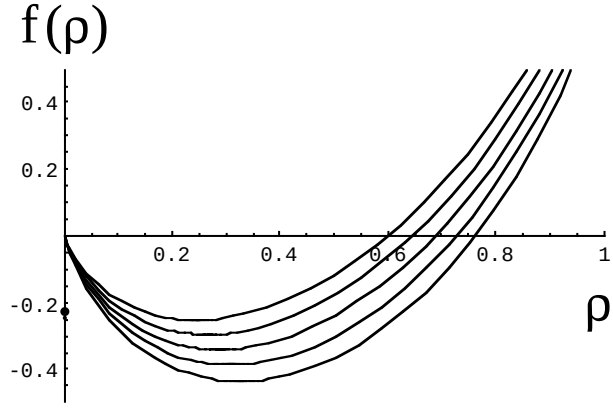


Fig. 2. The behavior of $f(\rho)$. This is defined by Eq. (16). From up to lower, $T = 0.9, 0.95, 1.0, 1.05, 1.1$ are drawn for $U = 2$ and $\nu = 3$, respectively. And a dot indicates the position of $-\varepsilon/N$ where $\varepsilon = 24$ and $N = 100$ are taken. As temperature is increased, the minimum at $\rho = \rho_{\mathrm{M}}$ moves following $\rho_{\mathrm{M}} = 1/(1 + \nu^{-1} e^{\beta U})$.

Substituting Eq. (17) into Eq. (16), we find

$$f(\rho_{\mathrm{M}}) = -k_{\mathrm{B}}T\ln(1+\nu e^{-\beta U})\,. \tag{19}$$

This yields

$$Z_N \approx e^{-N\beta f(\rho_{\mathrm{M}})} = (1+\nu e^{-\beta U})^N\,. \tag{20}$$

Comparing Eq. (20) with the exact partition function Eq. (12), the factor of Eq. (20) is dominant in Eq. (12) at high temperature. Therefore, the high temperature behavior is dominated by Eq. (20). Here we note that our partition function of Eq. (20) is essentially equivalent to the partition function obtained by Gō and Taketomi [see Eq. (8) in Gō and Taketomi's paper], stemmed out from the numerical simulations and ρ_{M} of Eq. (17) is equivalent to θ of Gō and Taketomi [see Eq. (9) in Gō and Taketomi's paper], where our ν is equivalent to their 2δ.[7]

We also note that the form of ρ_{M} (or $\langle \mathrm{CO}\rangle_{\mathrm{M}}$) is equivalent to that of the Fermi–Dirac distribution function for particles (holes)[6] and that the ρ_{M} (or $\langle \mathrm{CO}\rangle_{\mathrm{M}}$) does not depend on the total number of mismatches, which strongly supports the efforts using the contact order analysis.[27] From Eq. (18), we find that if $T \to 0$ then $\langle \mathrm{CO}\rangle_{\mathrm{M}} \to 1$ and if $T \gg 1$ then $\langle \mathrm{CO}\rangle_{\mathrm{M}} \to 1/(1+\nu)$. Hence, $\langle \mathrm{CO}\rangle_{\mathrm{M}}$ clearly exhibits the TST (Fig. 1). In Fig. 1, the exact $\langle \mathrm{CO}\rangle$ of Eq. (14) is shown by (A) for $\nu = 3$, $\varepsilon = 24$ and $U = 6$. And $\langle \mathrm{CO}\rangle_{\mathrm{M}}$ of Eq. (18) are shown by (B1)–(B4) for $U = 1, 2, 4, 6$, respectively.

We would like to note here the following: (i) First, since ρ_{M} of Eq. (17) and the local minimum $f(\rho_{\mathrm{M}})$ of the free energy at ρ_{M} are temperature dependent, the minimum $f(\rho_{\mathrm{M}})$ decreases as temperature T increases. Therefore, comparing the position of the free energy $f(0) = -\varepsilon/N$ at $\rho = 0$ (which is the free energy for the native structure) with $f(\rho_{\mathrm{M}})$, there is a critical temperature T_{m} (i.e. the melting temperature) given by $f(0) = f(\rho_{\mathrm{M}})$. This leads us to

$$\frac{\varepsilon}{N} = k_{\mathrm{B}}T_{\mathrm{m}}\ln(1+\nu e^{-U/k_{\mathrm{B}}T_{\mathrm{m}}})\,. \tag{21}$$

For simple limits, we find that if $U = 0$ then $T_{\mathrm{m}} = \varepsilon/(k_{\mathrm{B}}N\ln(1+\nu))$ and if $U = \infty$ then $T_{\mathrm{m}} = \infty$. Thus, there are two minima in the free energy curve of Eq. (16) at $\rho = 0$ which describes the folded native state and $\rho = \rho_{\mathrm{M}}$ which describes the unfolded coil state, respectively. Then, we find that if $T > T_{\mathrm{m}}$ then $f(0) > f(\rho_{\mathrm{M}})$ such that the ground state is the unfolded coil state, and if $T < T_{\mathrm{m}}$ then $f(0) < f(\rho_{\mathrm{M}})$ such that the ground state is the folded native state.

(ii) Second, the free energy surface $f(\rho)$ via $\rho = M/N$ of Eq. (16) can be thought of as the *funnel structure* of the free energy surface via configuration space in the sense of Dill and Chan[22] and of Wolynes.[13] This is because if we regard our free energy $f(\rho)$ as a free energy represented by the radial distance ρ in the multi (ν^N)-dimensional configuration space, then it certainly describes the funnel structure of the free energy surface discussed by Dill and Chan.[22] This situation may support the importance of the realization of the funnel structure of the free energy surface

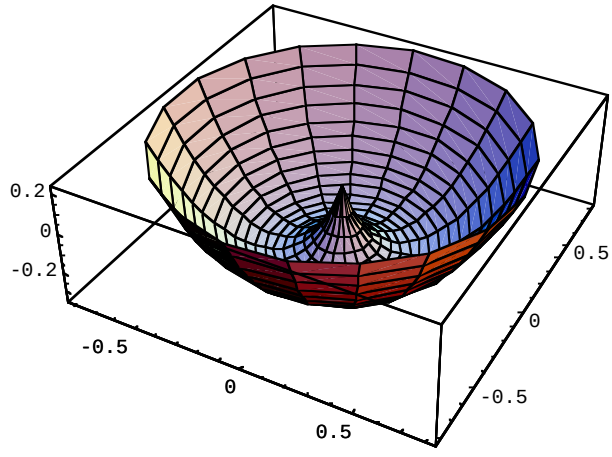


Fig. 3. The funnel structure of $f(\rho)$. The vertical axis means the free energy $f(\rho)$ defined by Eq. (16), while the horizontal plane means the configuration space whose radial distance is given by ρ. The case of $T = 1, U = 2$ and $\nu = 3$ is drawn. There is a minimum of the position of $f(0) = -\varepsilon/N$ at the center with a energy barrier very near the origin, where $\varepsilon = 24$ and $N = 100$ are taken. The champagne bottle minimum in the free energy is given by $f(\rho_{\rm M}) = -k_{\rm B}T\ln(1 + \nu e^{-\beta U})$ at $\rho = \rho_{\rm M} \equiv 1/(1 + \nu^{-1}e^{\beta U})$. $f(0) = f(\rho_{\rm M})$ defines the melting temperature $T_{\rm m}$. If $T > T_{\rm m}$ then $f(0) > f(\rho_{\rm M})$ such that the ground state is the unfolded coil state and if $T < T_{\rm m}$ then $f(0) < f(\rho_{\rm M})$ such that the ground state is the folded native state.

as well as the discussions of the funnelists. The funnel structure is shown in Fig. 3. The case of $T = 1$, $U = 2$, $\nu = 3$, $\varepsilon = 24$ and $N = 100$ is drawn.

Now, let us consider the specific heat C_p. From Eq. (7) the internal energy E is given by

$$E = \langle E(M)\rangle = -\frac{\partial}{\partial\beta}\ln Z\,. \tag{22}$$

Therefore, the specific heat is given by

$$C_p = \frac{\partial E}{\partial T} = -k_{\rm B}\beta^2\frac{\partial E}{\partial\beta} = k_{\rm B}\beta^2\frac{\partial^2}{\partial\beta^2}\ln Z\,. \tag{23}$$

Using Eq. (10), we find that

$$C_p = U\frac{\partial\langle M\rangle}{\partial T} = -Uk_{\rm B}\beta^2\frac{\partial\langle M\rangle}{\partial\beta} = UNk_{\rm B}\beta^2\frac{\partial\langle{\rm CO}\rangle}{\partial\beta}\,, \tag{24}$$

where we have used in the last line the relation $\langle{\rm CO}\rangle = 1 - \langle M\rangle/N$. Substituting the exact expression $\langle{\rm CO}\rangle$ of Eq. (14) into Eq. (24), we find

$$C_p = UNk_{\rm B}\beta^2\langle{\rm CO}\rangle\left(\frac{1}{Z_{N-1}}\frac{\partial Z_{N-1}}{\partial\beta} - \frac{1}{Z_N}\frac{\partial Z_N}{\partial\beta}\right)\,. \tag{25}$$

We show the behavior of the specific heat in Fig. 4. As examples, the case of $U = 2.5, \varepsilon = 24, \nu = 3$ is shown for the values of $N = 8, 24, 36, 72, 120$.

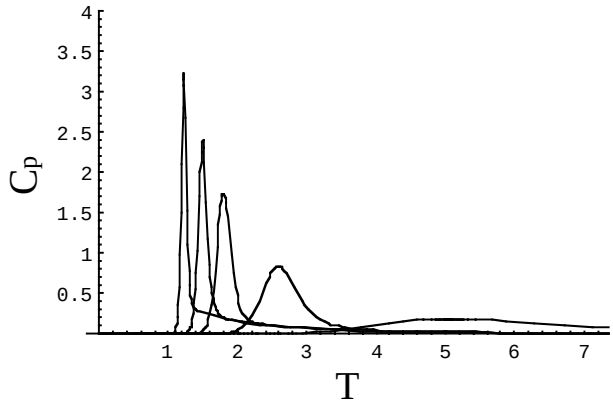


Fig. 4. The specific heat in the Zwanzig model. The case of $U = 2.5, \varepsilon = 24, \nu = 3$ is shown for the values of $N = 8, 24, 36, 72, 120$.

Let us consider the folding rate at last. Following the analysis of Zwanzig,[29] we can also calculate the folding rate $k_{\rm f}$ that is the inverse of the folding time $\tau_{\rm f} = 1/k_{\rm f}$ as

$$k_{\rm f} = k_0 \frac{N\nu e^{-\beta U}}{(1+\nu e^{-\beta U})^N - 1} \frac{1}{P_0(eq)}\,, \tag{26}$$

where $P_0(eq)$ is the occupancy of the correctly folded configuration given by $P_0(eq) = e^{\beta\varepsilon}/Z_N$ and $k_0 = 10^9\ {\rm s}^{-1}$. Since the factor $1/P_0(eq)$ consists of Z_N, the folding rate of $k_{\rm f}$ is proportional to the partition function Z_N. Thus, the behavior is dominated by that of the partition function Z_N.

The behavior of log of the folding rate $k_{\rm f}/k_0$ is shown in Fig. 5. The cases of $N = 6, 10, 20, 50, 80, 100, 150, 200$ are drawn for $\varepsilon = 24$, $U = 6$ and $\nu = 3$, respectively. This result shows that initially at higher temperatures the folding rate (the folding time) is very high (very low) and around the melting temperature $T_{\rm m}$ the folding rate (the folding time) becomes very small (very large), and finally at very low temperature the folding rate (the folding time) converges to a constant. Thus, the folding rate (the folding time) has a dip (peak)-structure. This means that near the critical temperature the transition becomes slow. This is a kind of critical slowing-down in the protein folding near the melting temperature since the transition is a TST.

5. Generalized Zwanzig Model

Let us a bit generalize the original Zwanzig model. If the hydrophobic collapse energy is present not only in the native structure but also in the deformed structures, then it is reasonable to assume

$$E(M) = MU - \varepsilon\left(1 - \frac{M}{N}\right)\,. \tag{27}$$

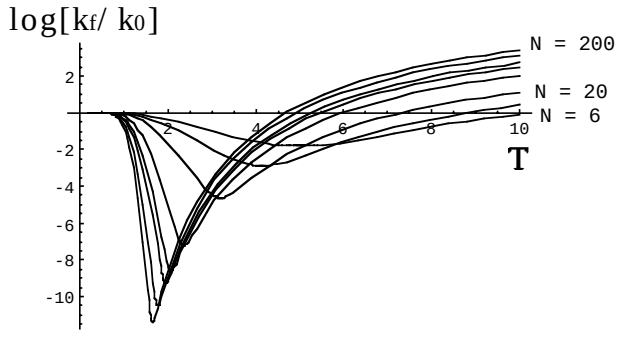


Fig. 5. The behavior of the folding rate $\ln(k_{\rm f}/k_0)$. This is defined by Eq. (26). The vertical line means $\ln(k_{\rm f}/k_0)$ while the horizontal axis means T. From lower to up, the cases of $N = 6, 10, 20, 50, 80, 100, 150, 200$ are drawn for $\varepsilon = 24$, $U = 6$ and $\nu = 3$, respectively.

Applying Eq. (27) into Eq. (7) together with Eq. (11), we find

$$Z_N = e^{\beta\varepsilon}(1 + \nu e^{-\beta U_0})^N \,, \tag{28}$$

which yields

$$\frac{F}{N} = -\frac{\varepsilon}{N} - k_{\rm B}T \ln(1 + \nu e^{-\beta U_0}) \,, \tag{29}$$

$$\langle {\rm CO} \rangle = 1 - \frac{\langle M \rangle}{N} = \frac{1}{1 + \nu e^{-\beta U_0}} = \frac{Z_{N-1}}{Z_N} \,, \tag{30}$$

where we have defined as

$$U_0 = U + \frac{\varepsilon}{N} \,. \tag{31}$$

The above Eq. (30) has the same form as that of Eq. (18), while the specific heat is given by Eq. (25). We can also generalize the argument of Zwanzig for the folding rate and get

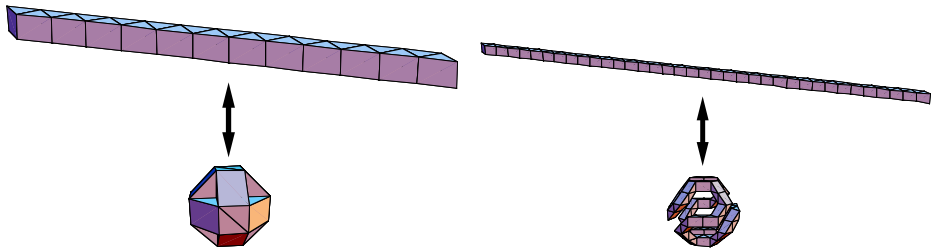
$$k_{\rm f} = k_0 N \nu e^{-\beta U_0} \frac{Z_N e^{-\beta\varepsilon}}{Z_N - e^{\beta\varepsilon}} \,, \tag{32}$$

where Z_N here is given by Eq. (28).

We also note that the above partition function of Eq. (28) coincides with that of Gō and Taketomi[7] since if we apply Eq. (27) into Eq. (7) then we obtain the same form as Eqs. (4) and (5) in the partition function. This coincidence between our theory of the generalized Zwanzig model and the result of numerical simulations of Gō and coworkers[4,7] implies that the important factor for the protein folding is the native interactions that are the interactions among the residues maintaining a native protein structure.

6. Rubik's Magic Snake Model

Let us next consider a more realistic model that is known as the RMSM[33] (Fig. 6). Explicitly using the RMSM, we are going to understand how the concepts in the



(a) (b)

Fig. 6. The unfolded and folded structures of Rubik's magic snake models. (a) The Rubik's magic snake model with 24 polyhedrons is shown; (b) The Rubik's magic snake model with 72 polyhedrons is shown.

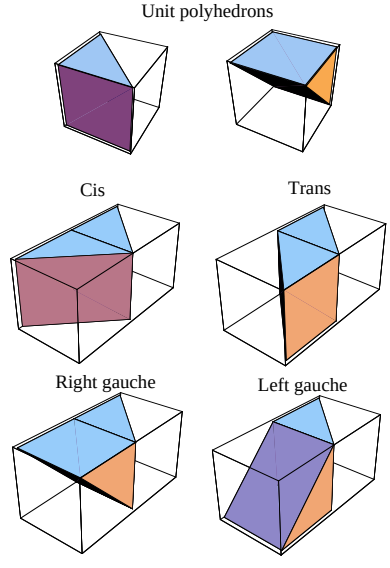


Fig. 7. The unit polyhedrons and basic conformations of Rubik's magic snake models.

Zwanzig model work for the protein folding problem and to find the contents of the parameters *a priori* adopted by Zwanzig.[29]

First, let us consider explicitly the geometry of the folded structure of the model with 24 polyhedrons [Fig. 6(a)], where each polyhedron is constructed by five faces of polygons: two isosceles triangles, two squares and one rectangle. We denote by N_{s} the total number of polyhedrons in the RMSM. In this model, there are four conformations between the segments: cis (c), trans (t), left gauche (g^-), and right gauche (g^+) conformations (Fig. 7), where the defining relations are given by $c^4 = g^+g^-g^+g^-g^+g^- = 1$ (Fig. 8).[33]

Let us consider only the HP-model[11] where if the hydrophobic (H) and hydrophilic or polar (P) segments are taken into account, then only three types of

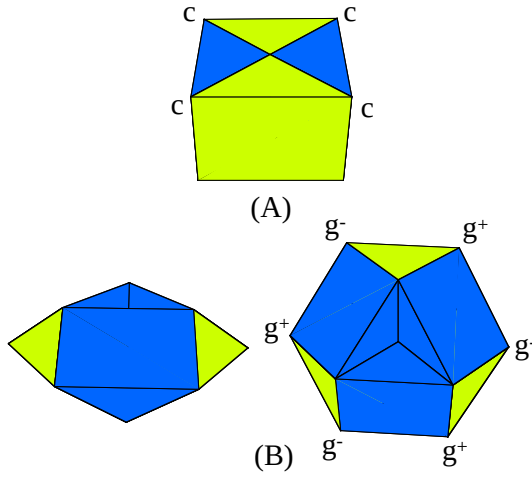


Fig. 8. The defining relations for the conformations of Rubik's magic snake models. (A) The defining relation $cccc = c^4 = 1$ is shown; (B) The defining relation $g^+g^-g^+g^-g^+g^- = (g^+g^-)^3 = 1$ is shown.

energies, E_{HH}, E_{HP}, E_{PP}, are assigned. And for the RMSM, let us assume that the sequence of the hydrophobic and hydrophilic segments (the HP sequence) is given by[33]

$$HPHPHPHPHPHPHPHPHPHPHPHP \tag{33}$$

for the sake of simplicity. Here we may simply denote it by $(HP)_{12}$ and the hydrophilic (hydrophobic) segments are located on even (odd) numbers of polyhedrons. Therefore, the total number of hydrophobic (hydrophilic) segments $N_{\rm H}$ ($N_{\rm P}$) is given by

$$N_{\rm H} = N_{\rm P} = \frac{N_{\rm s}}{2} \tag{34}$$

for our RMSM here.

Let us denote by $h_{\rm H}$ ($h_{\rm P}$) the hydrophobic energy cost between the hydrophobic (hydrophilic) segments and the surrounding medium such as water. Denote by E_{H_i,H_j} the hydrophobic interactions between the i- and jth polyhedrons where we have assumed that $E_{H_i,P_j} = E_{P_i,P_j} = 0$. And denote by U_t ($U_{g^\pm}$) the rotational energy cost for the trans (gauche) configuration. Here all quantities are assumed to be positive.

In the unfolded conformation[33] the chain structure is linear such that the structure is denoted by the symbol t^{23}. Therefore, all hydrophobic and hydrophilic segments are faced with the surrounding medium. This gives the total energy

$$E^{\rm unfold} = 23U_t + 12(h_{\rm H} + h_{\rm P})\,. \tag{35}$$

Here we have neglected the electronic energy in the chain, which has been discussed in Ref. 33.

Similarly, in the folded conformation with chirality C, the conformations are given by

$$g^+g^-g^+g^-g^-g^+g^-g^+g^+g^-g^+g^-g^-g^+g^-g^+g^+g^-g^+g^-g^-g^+g^- \tag{36}$$

for $C = 0$,

$$g^-g^+g^-g^+g^+g^-g^-g^+g^+g^-g^+g^-g^-g^+g^-g^+g^+g^-g^-g^+g^+g^-g^+ \tag{37}$$

for $C = 1$ and its mirror image for $C = -1$, respectively.[33] By using these conformations, the total energy is given by

$$E^{\text{fold}}(C) = E_{\text{c}}(C) + 23U_g + 12h_{\text{P}}\,, \tag{38}$$

where $E_{\text{c}}(C)$ is the hydrophobic collapse energy for the folded structure with chirality C. We now have for the folded structures with chirality of $C = 0$ and of $C = \pm 1$

$$\begin{aligned} E_{\text{c}}(C=0) = {} & E_{H_1,P_{24}} + E_{H_1,H_{19}} + E_{H_3,H_9} + E_{H_5,H_{23}} \\ & + E_{H_7,H_{13}} + E_{H_{11},H_{17}} + E_{H_{15},H_{21}}\,, \end{aligned} \tag{39}$$

$$\begin{aligned} E_{\text{c}}(C=\pm 1) = {} & E_{H_1,P_{24}} + E_{H_1,H_{13}} + E_{H_3,H_9} + E_{H_5,H_{23}} \\ & + E_{H_7,H_{19}} + E_{H_{11},H_{17}} + E_{H_{15},H_{21}}\,, \end{aligned} \tag{40}$$

respectively, where $E_{H_1,P_{24}} = 0$ in our assumption. This provides the difference between $E_{\text{c}}(C = 0)$ and $E_{\text{c}}(C = \pm 1)$

$$\begin{aligned} \Delta E_{\text{c}}(C) &= E_{\text{c}}(C=\pm 1) - E_{\text{c}}(C=0) \\ &= E_{H_1,H_{13}} + E_{H_7,H_{19}} - E_{H_1,H_{19}} - E_{H_7,H_{13}}\,. \end{aligned} \tag{41}$$

Consider the energy difference between the unfolded and folded structures, $\Delta E(C) \equiv E^{\text{unfold}} - E^{\text{fold}}(C)$. We now obtain

$$\Delta E(C) = 23(U_t - U_g) + 12h_{\text{H}} - E_{\text{c}}(C)\,, \tag{42}$$

where we have assumed $U_{g^+} = U_{g^-} \equiv U_g$. The first term represents the energy cost when a gauche conformation in the folded structure is converted into a trans conformation in the unfolded structure. The second term represents the hydrophobic energy cost by the 12 hydrophobic polyhedrons in the surrounding medium such as water. And the third term represents the hydrophobic condensation energy between the hydrophobic polyhedrons in the folded structure with chirality C. This means that each time when a conformational flip (i.e. rotation) between the nearest neighbor polyhedrons from a gauche conformation of $g^{\pm}$ to a trans conformation of t, then the energy cost of $U_t - U_g$ appears and if the fliped polyhedron is hydrophobic, then the energy cost appears since the hydrophobic pair-interaction between the hydrophobic polyhedrons in the folded structure is cut such that the hydrophobic polyhedron faces with the surrounding medium. In this way, a transition from the gauche conformation to the trans conformation with a hydrophobic energy cost plays the same role of a mismatch in the Zwanzig model.[29] In other words, the

denaturation or departure from the native state can be counted as a mismatch in the folded structure. Thus, we can recognize the number 12 in Eq. (42) as the number of mismatches in the Zwanzig model. Hence, Eq. (42) can be thought of as the difference between the total configuration energies of the deformed structure with the maximum number of mismatches of $N = N_{\rm H} = N_{\rm s}/2 = 12$ and of the native folded structure for the RMSM. This is the meaning of the terminology "mismatch" in our RMSM. Therefore, if the number of mismatches is M, then we can approximately give the difference between the total configuration energies of the deformed structure and native folded structure $E(M)$ as

$$E(M) = M(2(U_t - U_g) + h_{\rm H}) - E_{\rm c}(C)\delta_{M,0}\,. \tag{43}$$

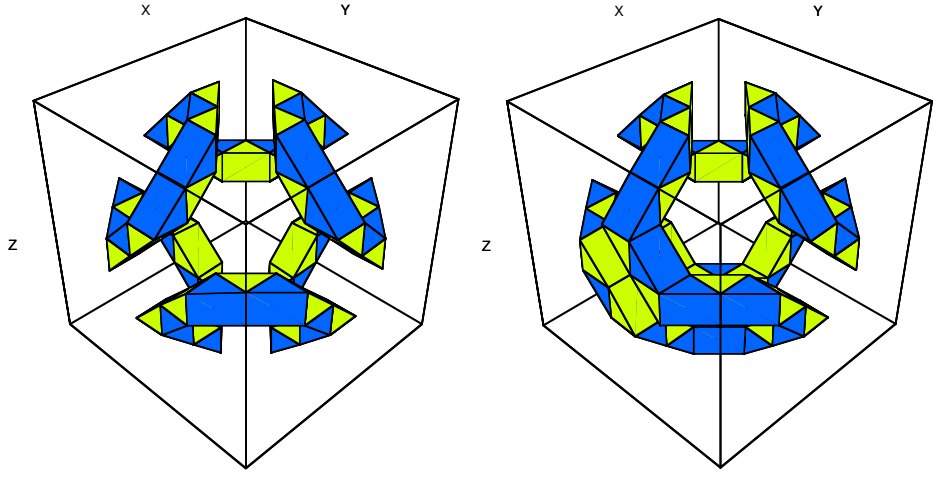
Hence, we can identify the parameters U and ε of the Zwanzig model as

$$U = 2(U_t - U_g) + h_{\rm H}, \qquad \varepsilon = E_{\rm c}(C) \tag{44}$$

in the RMSM. This provides the physical meaning of the parameters in the Zwanzig model, which was not so clear in his argument as was *a priori* defined.[29]

The important difference between the RMSM and the Zwanzig model is that the former can have the three types of exact native structures of chirality $C = 0, \pm 1$, and the latter does not take care of such real folded structures. This holds true even for the RMSM with 72 polyhedrons[33] (Fig. 9), where we have assumed the HP-sequence for the model is given by $(HP)_{36}$.

The unfolded chain is given by t^{71} and the folded structures are given by the



(a) (b)

Fig. 9. The Rubik's magic snake model with 72 polyhedrons. (a) The folded structure of chirality $C = 0$ is shown; (b) The folded structure of chirality $C = 1$ is shown. And there is the mirror image of (b), which is the folded structure of chirality $C = -1$. The drack (bright) polyhedrons indicate the hydrophilic (hydrophobic) residues.

iteration:

$$g^{\pm} \to tg^{\pm}t \tag{45}$$

for using in the RMSM with 24 polyhedrons of Eqs. (32) and (33), respectively. This scheme provides

$$\begin{aligned} &t^2g^+t^2g^-t^2g^+t^2g^-t^2g^-t^2g^+t^2g^-t^2g^+t^2g^+t^2g^-t^2g^+t^2g^- \\ &\quad\times t^2g^-t^2g^+t^2g^-t^2g^+t^2g^+t^2g^-t^2g^+t^2g^-t^2g^-t^2g^+t^2g^-t^2 \end{aligned} \tag{46}$$

for $C = 0$,

$$\begin{aligned} &t^2g^-t^2g^+t^2g^-t^2g^+t^2g^+t^2g^-t^2g^-t^2g^+t^2g^+t^2g^-t^2g^+t^2g^- \\ &\quad\times t^2g^-t^2g^+t^2g^-t^2g^+t^2g^+t^2g^-t^2g^-t^2g^+t^2g^+t^2g^-t^2g^+t^2 \end{aligned} \tag{47}$$

for $C = 1$ and its mirror image for $C = -1$, respectively.[33]

In these cases, since the HP-sequence is obtained by the inflation rules:

$$H_j \to H_{3j-2}P_{3j-2}H_{3j}\,, \tag{48}$$

$$P_j \to P_{3j-2}H_{3j-2}P_{3j}\,, \tag{49}$$

where $j =$ odd (even) for H (P) in the RMSM of 24 polyhedrons,[33] we have the hydrophobic energies between the polyhedrons as

$$E_{H_1,P_{24}} \to E_{H_1,P_{72}}\,, \tag{50}$$

$$E_{H_i,H_j} \to E_{H_{3i-2},H_{3j}} + E_{P_{3i-2},P_{3j-2}} + E_{H_{3i},H_{3j-2}}\,, \tag{51}$$

$$E_{P_i,P_j} \to E_{P_{3i-2},P_{3j}} + E_{H_{3i-2},H_{3j-2}} + E_{P_{3i},P_{3j-2}}\,, \tag{52}$$

$$h_{H_j} \to h_{H_{3j-2}} + h_{P_{3j-1}} + h_{H_{3j}}\,, \tag{53}$$

$$h_{P_j} \to h_{P_{3j-2}} + h_{H_{3j-1}} + h_{P_{3j}}\,. \tag{54}$$

From Eqs. (50)–(54), if we assume $E_{H_i,P_j} = E_{P_i,P_j} = 0$ and $E_{H_i,H_j} = E_{HH}$, then we just renormalize the hydrophobic energies by the number of contacts between polyhedrons as $E_{H,H} \to 2E_{H,H}$, $h_H \to 2h_H$ and the total number of mismatches as $N = 12 \to N = 36$. This argument can be applied to the RMSM with $N_{\rm s}$ polyhedrons, successively. This situation gives us the total number of mismatches

$$N = N_{\rm H} = \frac{N_{\rm s}}{2} \tag{55}$$

in the RMSM with the HP sequence of $(HP)_{N_{\rm s}/2}$. To take care of the right and left gauche configurations, we just discriminate U_g as $U_{g^{\pm}}$. Thus, we can also draw a similar consequence of the TST behavior for the inflated RMSM as well.

We note here the following: $\Delta U \equiv U_g - U_t > 0$ means the energy gap between the trans and gauche conformations between the nearest neighbor polyhedrons such that $l_P \equiv l_0 e^{\beta\Delta U}$ defines the *persistence length* for the molecule with l_0 being a

constant of order of a few Å.[34] Since, by definition, $U = h_{\rm H} - 2(U_g - U_t)$ must be positive, the relation

$$h_{\rm H} > 2(U_g - U_t) = 2\Delta U \tag{56}$$

must hold true. This indicates that the hydrophobic energy cost must exceed twice the difference in the conformation energy between the trans and the gauche configurations of the polyhedrons. Otherwise (i.e. U becomes negative), the chain can be regarded as a *stiff chain* that is hard to bend and therefore it is irrelevant for folding. On the other hand, from Eqs. (39) and (40), if we identify as $E_{H_i,H_j} = E_{HH}$, then we find

$$E_{\rm c}(C) = \frac{N_{\rm H}}{2} E_{HH} = \frac{N_{\rm s}}{4} E_{HH}\,, \tag{57}$$

for the RMSM with the total number $N_{\rm s}$ of polyhedrons.

To know the role of the hydrophobic energy $h_{\rm H}$ in the Zwanzig model, we show some examples of the contact order $\langle {\rm CO} \rangle$ in Fig. 10 and of the specific heat C_p in Fig. 11, respectively. Figure 10 shows the calculated $\langle {\rm CO} \rangle$ of Eq. (12) and Fig. 11 shows the calculated specific heat of Eq. (25) for the RMSM of $N_{\rm s} = 24$ polyhedrons adjusting with the Zwanzig model of $N = 12$, respectively. The parameters are taken as $\nu = 3$, $\varepsilon = 24$, $U = 2(U_t - U_g) + h_{\rm H}$, where we have assumed $U_t - U_g = 0$ for the sake of simplicity and $h_{\rm H}$ has been taken for (A) $h_{\rm H} = 2$, (B) $h_{\rm H} = 3$, (C) $h_{\rm H} = 4$, (D) $h_{\rm H} = 6$, respectively.

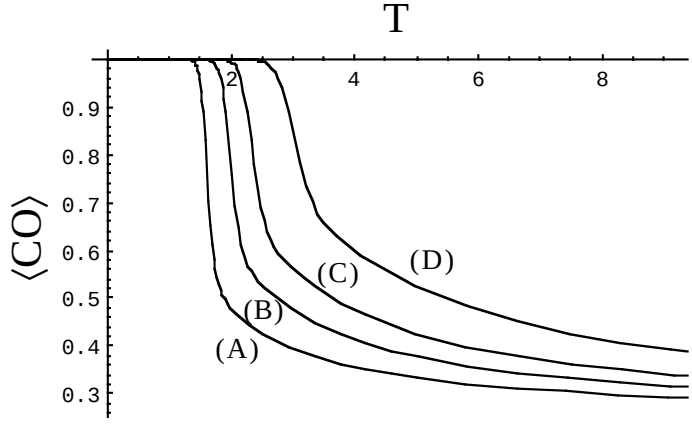


Fig. 10. The TST behavior of the contact order parameter $\langle {\rm CO} \rangle$ for the RMSM when the hydrophobic energy cost $h_{\rm H}$ is changed. The vertical axis means $\langle {\rm CO} \rangle$ while the horizontal axis means temperature T. $\langle {\rm CO} \rangle$ of Eq. (12) is calculated for the RMSM of $N_{\rm s} = 24$ polyhedrons adjusting with the Zwanzig model of $N = 12$. The parameters are taken as $\nu = 3$, $\varepsilon = 24$, $U = 2(U_t - U_g) + h_{\rm H}$, where we have assumed $U_t - U_g = 0$ for the sake of simplicity and $h_{\rm H}$ has been taken for (A) $h_{\rm H} = 2$, (B) $h_{\rm H} = 3$, (C) $h_{\rm H} = 4$, (D) $h_{\rm H} = 6$, respectively. Since the melting temperature $T_{\rm m}$ is given by Eq. (21) and therefore, as U is increasing, $T_{\rm m}$ is increasing, we find that the larger the $h_{\rm H}$, the larger the $T_{\rm m}$.

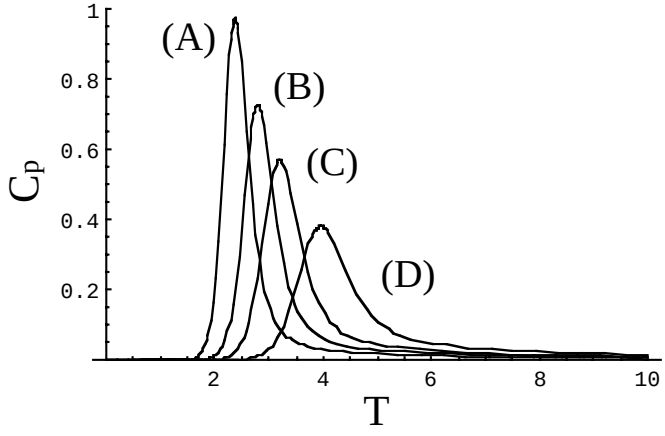


Fig. 11. The TST behavior of the specific heat C_p for the RMSM when the hydrophobic energy cost $h_{\rm H}$ is changed. The vertical axis means C_p in units of $Nk_{\rm B}U$, while the horizontal axis means temperature T. C_p of Eq. (25) is calculated for the RMSM of $N_{\rm s} = 24$ polyhedrons adjusting with the Zwanzig model of $N = 12$. The parameters are taken as $\nu = 3$, $\varepsilon = 24$, $U = 2(U_t - U_g) + h_{\rm H}$, where we have assumed $U_t - U_g = 0$ for the sake of simplicity and $h_{\rm H}$ has been taken for (A) $h_{\rm H} = 2$, (B) $h_{\rm H} = 3$, (C) $h_{\rm H} = 4$, (D) $h_{\rm H} = 6$, respectively. Since the melting temperature $T_{\rm m}$ is given by Eq. (21) and therefore, as U is increasing, $T_{\rm m}$ is increasing, we find that the larger the $h_{\rm H}$, the larger the $T_{\rm m}$.

Seen from the results, as $h_{\rm H}$ becomes large, the folding starts at high temperature. As was discussed in Sec. 4, as U is increasing, the melting temperature $T_{\rm m}$ is increasing, where the melting temperature $T_{\rm m}$ is defined by Eq. (21). Since U includes the hydrophobic energy cost $h_{\rm H}$ such as Eq. (44), we find that the larger the $h_{\rm H}$, the larger the melting temperature $T_{\rm m}$. This tendency coincides with the numerical calculations in Fig. 10. So, the larger the hydrophobic energy cost between the hydrophobic segments of the protein and the surrounding medium, the higher the folding temperature. This suggests that proteins with high contamination of hydrophobic segments can fold at higher temperature than proteins with low contamination of hydrophobic segments can do.

In the same way, let us know the role of the total number N of mismatches in the Zwanzig statistical model using the RMSM with $N_{\rm s}$ polyhedrons where $N = N_{\rm s}/2$. We show some examples of the contact order $\langle {\rm CO} \rangle$ in Fig. 12 and of the specific heat C_p in Fig. 13. There, $\langle {\rm CO} \rangle$ of Eq. (12) is calculated for the RMSM of $N_{\rm s}$ polyhedrons with $N_{\rm s} = 8, 12, 24, 36, 48, 60, 72, 96, 120, 240$, and C_p of Eq. (26) is calculated for the RMSM of $N_{\rm s}$ polyhedrons with $N_{\rm s} = 8, 24, 48, 72, 120$, respectively, where $\nu = 3$, $U = 2.5$, $\varepsilon = N_{\rm s}E_{HH}/4$ with $E_{HH} = 24$ are used, respectively.

The results show that as $N_{\rm s}$ becomes large, the TST behavior becomes eminent to be a more discontinuous transition. This crucial nature comes from the linear $N_{\rm s}$-dependence in the hydrophobic collapse energy ε [Eq. (57)], which is due to the nature of the RMSM.[33] Hence, the phase transition of the RMSM of 72 polyhedrons is much more like a discontinuous TST than that of the RMSM of 24 polyhedrons.

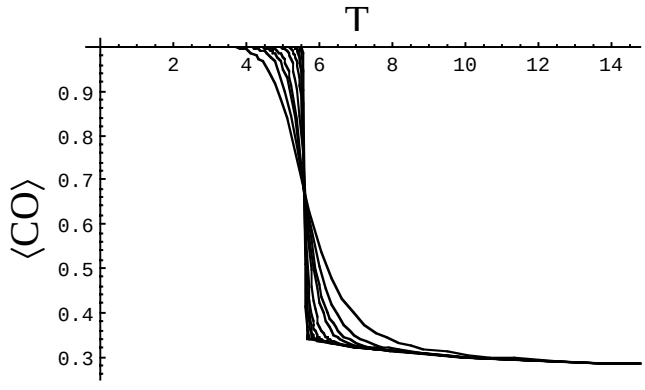


Fig. 12. The TST behavior of the contact order parameter $\langle \mathrm{CO} \rangle$ for the RMSM when the size of the RMSM is increased. The vertical axis means $\langle \mathrm{CO} \rangle$ while the horizontal axis means temperature T. $\langle \mathrm{CO} \rangle$ of Eq. (12) is calculated for the RMSM of N_s polyhedrons with $N_\mathrm{s} = 8, 12, 24, 36, 48, 60, 72, 96, 120, 240,$ respectively. Here $\nu = 3$, $U = 2.5$, $\varepsilon = N_\mathrm{s} E_{HH}/4$ with $E_{HH} = 24$ are used, respectively. As N_s becomes larger, the TST behavior becomes a more discontinuous transition.

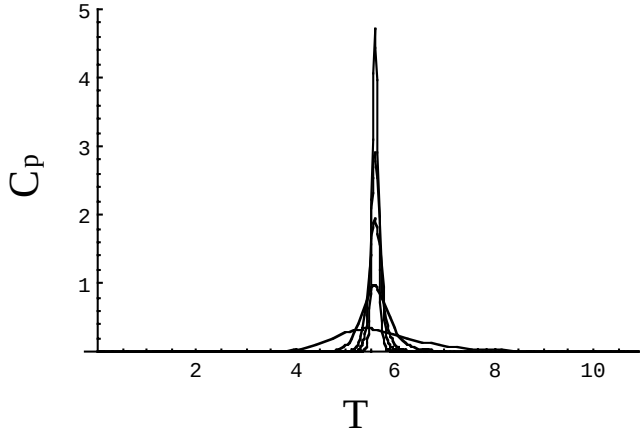


Fig. 13. The TST behavior of the specific heat C_p for the RMSM when the size of the RMSM is increased. The vertical axis means C_p in units of $Nk_\mathrm{B}U$, while the horizontal axis means temperature T. C_p of Eq. (25) is calculated for the RMSM of N_s polyhedrons with $N_\mathrm{s} = 8, 24, 48, 72, 96,$ respectively. Here $\nu = 3$, $U = 2.5$, $\varepsilon = N_\mathrm{s} E_{HH}/4$ with $E_{HH} = 24$ are used, respectively. As N_s becomes larger, the TST behavior of the specific heat becomes like a Dirac δ-function.

If the hydrophobic collapse energy ε is kept constant in the RMSM as for the case of the original Zwanzig model,[29] then as the number of mismatches N is increased, the tendency becomes *opposite* to the case of the RMSM with the HP sequence of $(HP)_{N_\mathrm{s}/2}$ and it turns out to be more smooth change as a function of T. This is because as N becomes large, the role of the hydrophobic collapse energy ε becomes relatively small in the Zwanzig model [Eq. (10)]. This means that the system can be regarded as a chain with low contamination of the hydrophobic segments such

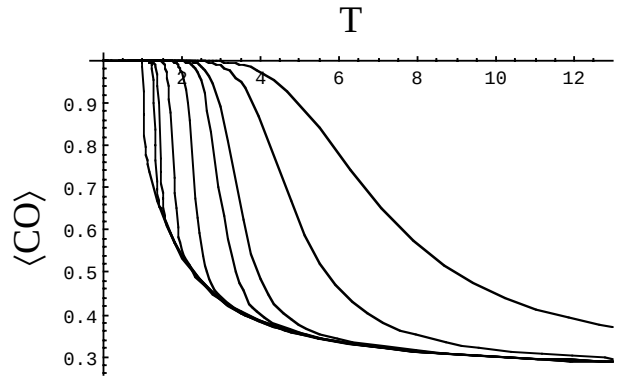


Fig. 14. The TST behavior of the contact order parameter $\langle \mathrm{CO} \rangle$ for the RMSM when the hydrophobic collapse energy ε is kept constant. The vertical axis means $\langle \mathrm{CO} \rangle$ while the horizontal axis means temperature T. $\langle \mathrm{CO} \rangle$ of Eq. (12) is calculated for the Zwanzig model with the total mismatches $N = 6$ (right most), $8, 12, 24, 36, 48, 60, 72, 96, 120, 240$ (left most), respectively. Here $\nu = 3$, $U = 2.5$, $\varepsilon = E_{HH} = 24$ are used, respectively.

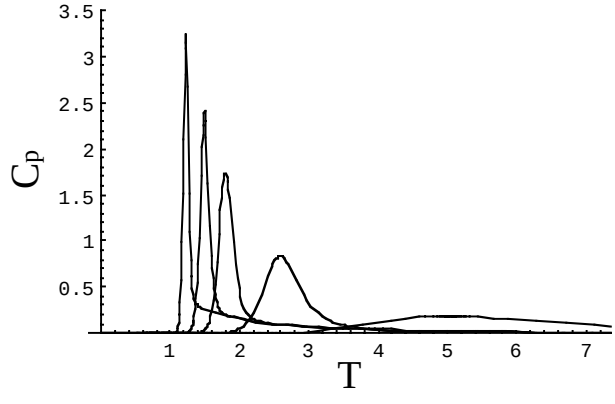


Fig. 15. The TST behavior of the specific heat C_p for the RMSM when the hydrophobic collapse energy ε is kept constant. The vertical axis means C_p in units of $Nk_{\mathrm{B}}U$, while the horizontal axis means temperature T. C_p of Eq. (25) is calculated for the Zwanzig model with the total mismatches $N = 8$ (right most), $24, 48, 72, 120$ (left most), respectively. Here $\nu = 3$, $U = 2.5$, $\varepsilon = E_{HH} = 24$ are used, respectively.

that the density $N_{\mathrm{H}}/N_{\mathrm{s}}$ of hydrophobic segments in the chain becomes relatively small. This is shown in Figs. 14 and 15. Figure 14 shows the calculated $\langle \mathrm{CO} \rangle$ of Eq. (12) while Fig. 15 shows the calculated C_p of Eq. (25) for the Zwanzig model of the total mismatches N with $N = 8, 12, 24, 36, 48, 60, 72, 96, 120, 240$, respectively, where $\nu = 3$, $U = 2.5$, $\varepsilon = E_{HH} = 24$ are used, respectively.

Thus, the above results suggest that the geometry of the folded protein adjusting with the HP sequence is very important for folding phenomena.

7. Gō's Lattice Model

Let us now consider the standard lattice model[4,9,11–19] in order to understand the role of the contact order in this model. Let us denote by r_i's $(1 \leq i \leq L)$ the positions of the L residues of a protein of length L, where the residues are placed on the vertices of a two-dimensional square lattice or a three-dimensional cubic lattice and L is in our previous notation $L = N_{\rm s}$, the total number of polyhedrons in the chain. Then this model is called the *Gō's lattice model*.[4] In this model, the Hamiltonian is given by the following standard lattice Hamiltonian:

$$H_L = \sum_{i<j=1}^{L} \Delta(r_i - r_j) E_{\sigma_i \sigma_j} \,, \tag{58}$$

where $\Delta(r_i - r_j) = 1 (= 0)$ if r_i and r_j are in a contact situation (otherwise) and $E_{\sigma_i \sigma_j}$ is the interaction energy depending on the types σ_i of segments in contact. For example, for the HP-model, as the same as before, only three types of energies, E_{H_i,H_j}, E_{H_i,P_j}, E_{P_i,P_j}, can be assigned.[11]

Setting as $Q \equiv \{r_1, r_2, \cdots, r_L\}$ in Eq. (1), we now need calculate the partition function Z for Eq. (58). Again, we meet a hopeless situation that we cannot calculate the partition function explicitly in the configuration space, resulting in the Levinthal's paradox. However, defining the number of contacts of the structures,[4,7] K, by

$$K = \sum_{i<j=1}^{L} \Delta(r_i - r_j) \,, \tag{59}$$

we can apply the same argument that thermal fluctuations occur on the equienergy surface of $E(K) = E(R, K)$ and derive the same equations as Eqs. (7) and (8) with replacing M by K for this case. Defining the contact order parameter by $\mathrm{CO} = K/N$,[4,7] we find

$$\langle \mathrm{CO} \rangle = \frac{\langle K \rangle}{N} = \frac{1}{N} \, \nu \frac{\partial}{\partial \nu} \ln Z_N \,, \tag{60}$$

where N is the total number of contacts. This number K of contacts is different from the number M of mismatches by $K = N - M$.

Now, the hardest part in the problem is to find the exact form of the number of states, $W(K)$. However, if we can assume that the positions of the hydrophobic contacts do not interact with each other, then the emergence of such contacts can be regarded as independent events. Therefore, we can assume the same partition as Eq. (11) with replacing M by K. In this setting, the mathematical situation is the same as before. Hence, we can expect the TST behavior of the protein folding for the lattice models as well, and see how the concept of the contact order works in the protein folding problem.[27,30–32] The above argument is in the same philosophy of the Bragg–Williams approximation for the Ising model.[35] We see many similarities between the magnetism and the protein folding.

Let us now make a comment on the relationship between the $\langle \mathrm{CO} \rangle$ and the relative contact order RCO defined by Baker *et al.*[27] While the total number N of contacts is defined by

$$N = \sum_{i<j=1}^{L} \Delta(r_i - r_j)\,, \tag{61}$$

they defined RCO as

$$\mathrm{RCO} = \frac{1}{NL} \sum_{i<j=1}^{N} \Delta S(r_i - r_j)\,. \tag{62}$$

Here $\Delta S(r_i - r_j)$ stands for the sequential distance or the sequence separation between the ith segment at the position r_i and the jth segment at the position r_j along the course of the protein chain. It is not the spatial distance in the three-dimensional space such as $|r_i - r_j|$ but the number of difference between i and j such that $\Delta S(r_i - r_j)$ is almost approximated by

$$\Delta S(r_i - r_j) \approx |i - j|\,. \tag{63}$$

Since the sequential distance acts only when a contact is formed, Eq. (62) can be rewritten using $\Delta(r_i - r_j)$ as

$$\mathrm{RCO} = \frac{1}{L^2} \sum_{i<j=1}^{L} \Delta S(r_i - r_j)\Delta(r_i - r_j)\,. \tag{64}$$

Comparing the expression of Eq. (64) with that of Eq. (58), $\Delta S(r_i - r_j)$ can be thought of as a kind of *correlation function* between the ith and the jth residues. Since RCO is given as an averaged quantity, RCO can be regarded as a *susceptibility* of the folded protein such as specific heat C_p, compressibility κ, susceptibility χ, etc. Therefore, RCO can be regarded as the *topological susceptibility* τ. As is well-known in statistical mechanics, an order parameter $\langle m \rangle$ is not the susceptibility $\langle (m - \langle m \rangle)^2 \rangle$.[6,35] The contact order CO is an order parameter while the relative contact order RCO is a susceptibility. Thus, the contact order CO given by Eq. (60) is different from the relative contact order RCO given by Eq. (64). This kind of argument is missing in the previous literature.[27]

The difference between CO and RCO in the above discussion is strengthened as follows: The RCO discriminates the topological structure — the topological nature of the native interactions in the folded structures. If the native structure consists of no native interactions as in the case of no hydrophobic residues such that the native structure looks like a linear chain, then $\mathrm{RCO} = 0$. On the other hand, if the native structure consists of the native interactions as in the case that all residues are hydrophobic such that the native structure looks like an oil droplet, then $\mathrm{RCO} = 1$. Native structures of real globular proteins are in between the above two limits such that $0 < \mathrm{RCO} < 1$. Thus, RCO is an index for what the native structure is.

On the other hand, the contact order CO is meaningful once a native structure is assumed such that it defines a set of the native interactions. Whatever the native structure, we can count the energy difference between the denatured structure and the native structure in terms of the language of contacts or mismatches. Thus, CO is a measure — an order parameter — for the folding process itself. In this way, CO and RCO are different quantities for investigating different aspects of the protein folding.

8. Hydrophobic Condensation

In this way, we are tempted to conclude that the TST of hydrophobic condensation is a universal nature in protein folding no matter how the hydrophobic residues are distributed in the sequences. However, at this moment, we cannot answer whether or not the global hydrophobic collapse precedes the formation of the secondary structures in the process of folding.

To consider this problem, let us go back to see Figs. 10–15. From these figures, we see the following tendency of folding: (a) If the hydrophobic energy cost h_{H} between hydrophobic segments and the surrounding medium of water is increased, then the transition temperature T_{m} is increased (see Figs. 10 and 11). (b) If there is enough density of hydrophobic segments in the folding chain so that the hydrophobic collapse energy ε is sufficiently large such as $\varepsilon/N_{\mathrm{s}} = \text{const.}$, then as the length of the chain is increased, the transition becomes a discontinuous TST. This means that the shorter (longer) the length of the chain, the more continuous (discontinuous) the transition. (See Figs. 12 and 13.) (c) If there is low density of hydrophobic segments in the folding chain so that the hydrophobic collapse energy ε is relatively small, then as the length of the chain is increased, the transition temperature is lowered and the transition is broad and continuous not like a discontinuous TST (see Figs. 14 and 15).

From this, we can conclude that very small protein segments such as α-helices with short length can fold continuously as a broad transition around the transition temperature, while very large protein segments that are much larger than the size of α-helices can fold discontinuously as a TST. In other words, if there are some secondary structures such as α-helices in a small globular protein with size of 80 residues, then the transition of the secondary structures are broad and continuous transition while the global hydrophobic collapse transition may be a more discontinuous TST.

Therefore, we can have the following folding scenario: When folding is processing, first small secondary structures in the protein chain can start to fold and end up with completing the secondary structures at last, since the transition of small segments is slow and continuous. On the other hand, the global hydrophobic collapse or condensation can be very quick to complete. Thus, we may draw a speculation that *in the course of folding, the hydrophobic collapse may precedes the formation of the secondary structures in the chain.*

9. Equation of State for Protein Folding

Let us next consider the equation of state of protein folding. For this purpose, we need know the relationship between the free energy F and the contact order $\langle \mathrm{CO} \rangle$, which can be thought of as a the equation of state for protein folding.

In statistical mechanics of a gas, since the system can be regarded as the one consisting of the infinite number of particles, we take usually the *thermodynamic limit* where the total number of particles $N \to \infty$ and the total volume of the system $V \to \infty$ with the density $d \equiv N/V$ being kept finite.[35] In the thermodynamic limit, we consider the equation of state of the system. This is carried out as follows: Pressure P and the total number N of particles are represented in terms of the fugacity $z = e^{\beta\mu}$ where μ is the chemical potential for the gas, respectively, such that

$$\frac{P}{k_{\mathrm{B}}T} = \frac{1}{V} \ln \Xi \equiv F(z, V)\,, \tag{65}$$

$$d = z \frac{\partial}{\partial z} \frac{1}{V} \ln \Xi \equiv z \frac{\partial}{\partial z} F(z, V)\,, \tag{66}$$

where Ξ is the grand partition function of the system. Eliminating z from both, we find the equation of state:

$$P = P(d, T)\,. \tag{67}$$

By the Lee–Yang theorem,[35] if there is a singularity in the partition function Ξ in the positive real axis of z, then there is a phase transition in the system. Thus, the existence of a singularity in $F(z, V)$ is very important for the study of a phase transition.

The situation in statistical mechanics of the protein folding is analogous to that in statistical mechanics of the gas. The correspondence between the former and the latter is as follows: $N \leftrightarrow \langle M \rangle$; $V \leftrightarrow N$; $P \leftrightarrow -F$; $d \equiv N/V \leftrightarrow \rho \equiv \langle M \rangle / N$ (or $\langle \mathrm{CO} \rangle$), where we have

$$\beta F = -\frac{1}{N} \ln Z \equiv -g(\nu, N)\,, \tag{68}$$

$$\rho = 1 - \langle \mathrm{CO} \rangle = \nu \frac{\partial}{\partial \nu} \frac{1}{N} \ln Z = \nu \frac{\partial}{\partial \nu} g(\nu, N)\,. \tag{69}$$

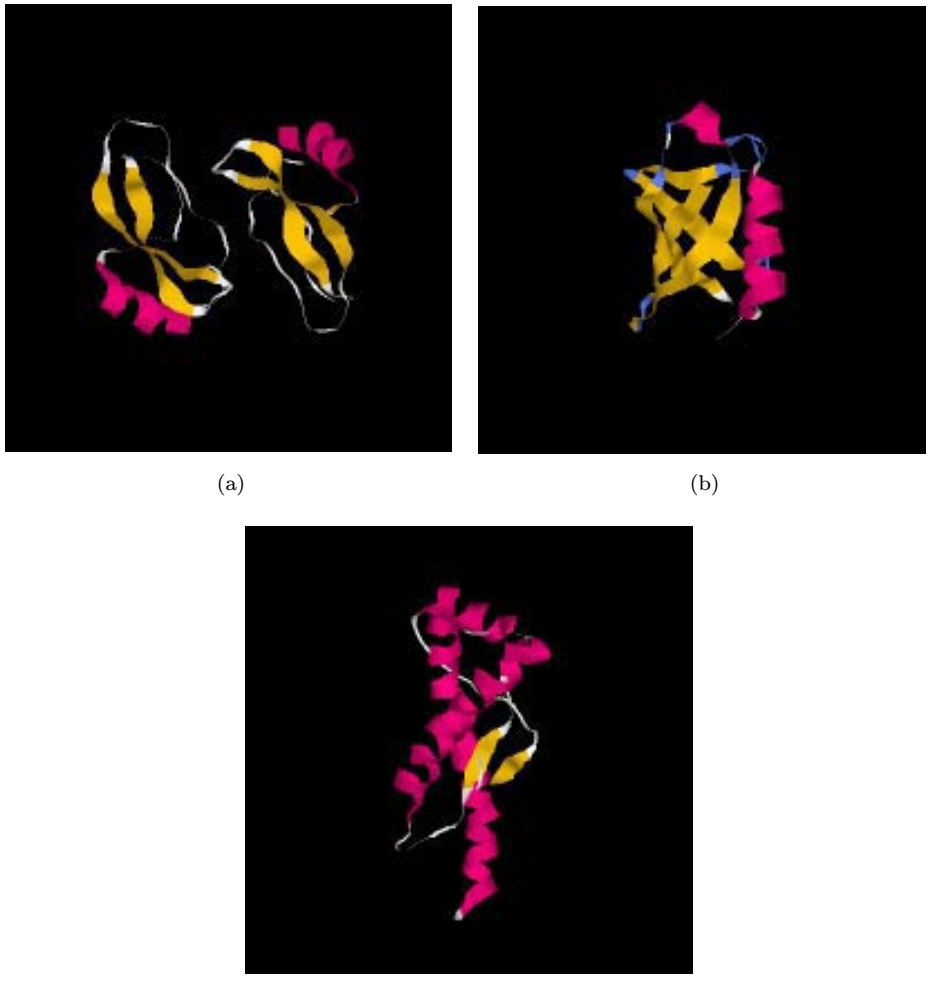
And hence, we have the correspondence: Eqs. (65) and (66) $\leftrightarrow$ Eqs. (68) and (69), respectively. Now, as in the case of a gas, by eliminating ν, we get the equation of state:

$$F = F(\rho, T)\,. \tag{70}$$

In this way, the theoretical framework of statistical mechanics of protein folding and that of statistical mechanics of gas are very similar to each other, apart from that in the former we consider only finite systems while in the latter we consider the infinite systems in the thermodynamic limit. This nature is recently discussed by Pande *et al.*[36]

10. Morphology of Protein Folding

Let us consider *morphology of protein folding*. Recently, the important role of folding and misfolding of small globular proteins has been known. The Alzheimer's disease is caused by a small protein called the *Alzheimer's amyloid* protein [Fig. 16(a)]. The Parkinson's disease is caused by a small protein called *parkin*, which is the *ubiquitin* protein [Fig. 16(b)]. And the Creutzfeldt-Jacob's disease is caused by a small protein called an anomalous *prion* protein that is believed to be a misfolded prion protein [Fig. 16(c)].[37]



(a) (b)

(c)

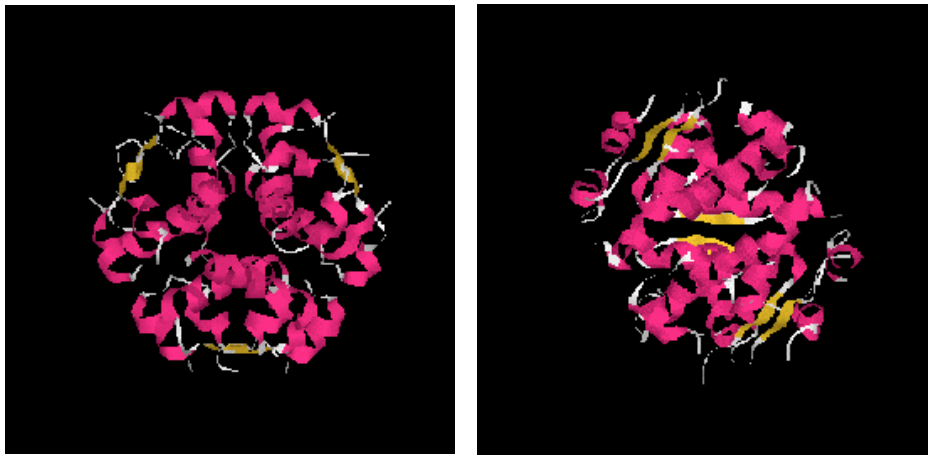
Fig. 16. The three-dimensional folded protein structures of the famous diseases. (a) The Altzheimer's amyloid protein causing Altzheimer's disease. (b) The ubiquitin protein causing the Parkinson's disease. (c) The prion protein causing the Creutzfeldt–Jacob disease. Dark, gray and white parts show α-helices, β-sheets and γ-turns, respectively. Usually, turns are called "β-turns" but we prefer to call them "γ-turn" without the double use of β.

As is seen in Fig. 16, these proteins consist of secondary structures such as *α-helices*, *β-sheets* and *γ-turns*, respectively. Thus, the morphology of the folded structures among such secondary parts seems very important to know the nature of a protein. This is another important protein problem other than the protein problem to know the fast folding process. While the latter is a statistical mechanical problem as discussed in the previous sections, the former is a topological problem of proteins as a problem of the protein architecture. Thus, the problem of morphology of proteins may be another target for protein physicists.

In this respect, it is in general well-known that there are several classes of folded structures of small globular proteins.[38,39] (i) One type is small globular proteins such as Cytochrome c and Acyl-coenzyme A binding protein (ACBP), in which the secondary structures are made by only α-helices — *α-helices proteins*.[38] This type has tendency that proteins are more *symmetrical*, having a symmetry axis. (ii) Another type is small globular proteins such as SRC homology 3 (SH3) domain from α-spectrin and cold shock protein B, in which the secondary structures are made by only β-sheets — *β-sheets proteins*.[39] This type has tendency that proteins are not so symmetrical but more like *twisted* geometry. (iii) Third type is small globular proteins made by their mixture. The above three examples for the famous diseases in Fig. 16 are of this type. This type has tendency that proteins are perfectly *asymmetric*.

The above first two types look very similar to the folded structures of the RMSM of 72 polyhedrons with chirality $C = 0, \pm 1$ (Fig. 9). That is, α-helices proteins look like the RMSM of 72 polyhedrons with $C = 0$ [see Fig. 9(a)], while β-sheets proteins look like the RMSM of 72 polyhedrons with $C = \pm 1$ [see Fig. 9(b)]. Therefore, we may model such real small globular proteins by the RMSM of 72 polyhedrons with different chirality C. Indeed, the most similar protein structure to our RMSM of 72 polyhedrons with $C = 0$ is the three-dimensional structure of the *insulin* protein (Fig. 17) although there is a difference between them. In the insulin protein there are about 12 separated chains that form the folded structure, while in our RMSM of 72 polyhedrons with $C = 0$ there is only one chain for the folded structure. However, the connectivity in the global folded structures is the same [please compare the structure of Fig. 9(a) with that of Fig. 17(a)]. And the most similar protein structure to our RMSM of 72 polyhedrons with $C = \pm 1$ is the three-dimensional structure of the *SH3-domain* protein (Fig. 18) although there is a difference between them.

In the SH3-domain protein there are about 3 twisted turns and sheets that form the folded structure, while in our RMSM of 72 polyhedrons with $C = \pm 1$ there are only two twisted parts for the folded structure. However, the connectivity in the global folded structures is very similar to each other [Please compare the structure of Fig. 9(b) with that of Fig. 18(a)]. In this way, the study of the morphology of the RMSM with different chirality seems valuable for understanding the morphology of real small globular chains in order to recognize the difference between the α-helices (rich) proteins and β-sheets (rich) proteins.



(a) (b)

Fig. 17. The three-dimensional folded protein structure of insulin protein (the pdb code, 1WAV). (a) The structure seen from the three-fold symmetry axis. (b) The side view of the structure. Dark, gray and white parts show α-helices, β-sheets and γ-turns, respectively.

(a) (b)

Fig. 18. The three-dimensional folded protein structure of SRC-homology 3 (SH3) domain protein (the pdb code, 1ABQ). (a) The structure seen from the top. (b) The side view of the structure. Drak, gray and white parts show α-helices, β-sheets and γ-turns, respectively.

Now, let us consider the stability of small proteins, using the RMSM of 72 polyhedrons. The hydrophobic collapse energies $E_{\rm c}(C)$ for the RMSM of 24 polyhedrons are given by Eq. (39) for $C = 0$ and Eq. (40) for $C = \pm 1$, respectively. And their difference is given by Eq. (41). As noted before, the hydrophobic energies are scaled as Eqs. (48) and (49) for the RMSM of 72 polyhedrons. Therefore, we obtain the

inflation rules as

$$\begin{aligned} &E_{H_1,H_{19}} \to E_{H_1,H_{57}} + E_{H_3,H_{55}}; E_{H_3,H_9} \to E_{H_7,H_{27}} + E_{H_9,H_{25}}; \\ &E_{H_5,H_{23}} \to E_{H_{13},H_{69}} + E_{H_{15},H_{67}}; E_{H_7,H_{13}} \to E_{H_{19},H_{39}} + E_{H_{21},H_{37}}; \\ &E_{H_{11},H_{17}} \to E_{H_{31},H_{51}} + E_{H_{33},H_{49}}; E_{H_{15},H_{21}} \to E_{H_{43},H_{63}} + E_{H_{45},H_{61}}; \\ &E_{H_1,H_{13}} \to E_{H_1,H_{39}} + E_{H_3,H_{37}}; E_{H_7,H_{19}} \to E_{H_{19},H_{57}} + E_{H_{21},H_{55}} . \end{aligned} \tag{71}$$

Using Eq. (71) in Eqs. (39) and (40), we obtain

$$\begin{aligned} E_{\rm c}(C=0) = {} & E_{H_1,P_{72}} + E_{H_1,H_{57}} + E_{H_3,H_{55}} + E_{H_7,H_{27}} + E_{H_9,H_{25}} \\ & + E_{H_{13},H_{69}} + E_{H_{15},H_{67}} + E_{H_{19},H_{39}} + E_{H_{21},H_{37}} \\ & + E_{H_{31},H_{51}} + E_{H_{33},H_{49}} + E_{H_{43},H_{63}} + E_{H_{45},H_{61}} \end{aligned} \tag{72}$$

$$\begin{aligned} E_{\rm c}(C=\pm 1) = {} & E_{H_1,P_{72}} + E_{H_1,H_{39}} + E_{H_3,H_{37}} + E_{H_7,H_{27}} + E_{H_9,H_{25}} \\ & + E_{H_{13},H_{69}} + E_{H_{15},H_{67}} + E_{H_{19},H_{57}} + E_{H_{21},H_{55}} \\ & + E_{H_{31},H_{51}} + E_{H_{33},H_{49}} + E_{H_{43},H_{63}} + E_{H_{45},H_{61}} , \end{aligned} \tag{73}$$

respectively. Therefore, the difference between $E_{\rm c}(C=0)$ and $E_{\rm c}(C=\pm 1)$ is given by

$$\begin{aligned} \Delta E_{\rm c}(C=\pm 1) = E_{\rm c}(C=\pm 1) - E_{\rm c}(C=0) = {} & E_{H_1,H_{39}} + E_{H_3,H_{37}} + E_{H_{19},H_{57}} \\ & + E_{H_{21},H_{55}} - E_{H_1,H_{57}} - E_{H_3,H_{55}} - E_{H_{19},H_{39}} - E_{H_{21},H_{37}} . \end{aligned} \tag{74}$$

The above energy difference may vanish if we take the simple assumption that $E_{H_iH_j} = E_{HH} = \text{const.}$, which means that the ground state energies of the folded structures of $C=0$ and $C=\pm 1$ are degenerate. However, in real globular proteins, this type of simple assumption does not work because local coordinations of the residues are always different even for the same hydrophobic residues. Therefore, the difference may appear to discriminate the ground states of the folded structures[33] such that the different HP-sequences may favor different folded structures according to the match between the HP-sequences and the folded structures as an example of the consistency principle. The difference of native interactions between the RMSM of 72 polyhedrons of $C=0$ and $C=\pm 1$ is schematically shown in Fig. 19.

Thus, applying the above to the Zwanzig model discussed in the previous sections, we are able to consider the folding nature of the RMSM of 72 polyhedrons with different chirality as a prototype model to distinguish the physical nature between the α-helices proteins and β-sheets proteins. This suggests why there are two types of small globular proteins in the native structures such as α-helices (rich) proteins and β-sheets (rich) proteins in nature. And as can be seen in Fig. 9, the structures of the α-helices (rich) proteins and the β-sheets (rich) proteins are not so different from each other, because they can be transformed into one another by the

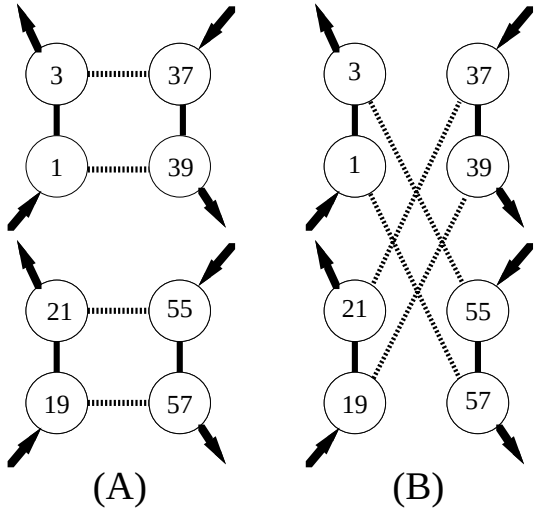


Fig. 19. The schematic diagrams for the difference of the native interactions between the RMSM of 72 polyhedrons with $C = 0$ and $C = \pm 1$. (A) The native interactions in the folded structure of the RMSM of 72 polyhedrons with $C = 0$; (B) The native interactions in the folded structure of the RMSM of 72 polyhedrons with $C = \pm 1$. Thick arrows mean the directions of the nearest neighbor polyhedrons along the course of the chain. Dashed lines indicate the interaction bonds between the hydrophobic polyhedrons. Circles with numbers stand for the numbered hydrophobic polyhedrons. Here the other native interactions are not drawn since they are the same for both structures. Thus the folded structures of the RMSM of 72 polyhedrons with $C = 0$ and with $C = \pm 1$ consist of the recombination of the native interaction bonds.

recombination of the native interaction bonds, schematically illustrated in Fig. 19. This may suggest why some small globular proteins such as the prion protein can easily mutate into an anomalous and misfolded protein.

11. Conclusion

In conclusion, we have discussed the statistical mechanical foundation for the TST in the protein folding. First, we have given the statistical mechanical formulation for protein folding, which relates the Zwanzig model[29] with the Finkelstein's sequential folding model[30] in the spirit of Gō's TST idea[4,7,8] that may resolve the Anfinsen's dogma[2] and the Levinthal's paradox.[3] We have showed that the TST nature of hydrophobic condensation in the Zwanzig model[29] comes from the principles of statistical mechanics that thermal fluctuations with equienergy surface are very important. We have introduced the contact order and the free energy for protein folding as corresponding to the density and the pressure in the gas system, respectively. And using them, we have derived the equation of state for protein folding. Second, we have showed that if the energy difference from the native state is given by $E(M)$, then we derive the Zwanzig model,[29] where we have derived the free energy, the contact order and the specific heat, respectively. Third, we have used the RMSM in order to give the contents of the parameters adopted by

Zwanzig,[29] and applied the Zwanzig model to the folding of the RMSM. Here we have derived the contact order for the RMSM and showed several examples. Fourth, we have discussed the so-called Gō's lattice model in relation to the contact order and the Zwanzig model. Fifth, using the results for the RMSM, we have discussed the hydrophobic condensation problem and concluded that the hydrophobic collapse may precede the formation of the secondary structures such as α-helices etc. Sixth, we have discussed the morphology of folded proteins why in nature there are mainly two types of folded structures such as α-helices proteins and β-sheets proteins. In this way, the statistical formulation for protein folding using the concepts of mismatches and contact order is very useful in considering many problems in protein folding. And it generally exhibits the TST nature in protein folding for small globular proteins modeled by the RMSM.

Acknowledgment

I would like to thank Satoshi Takahashi for sending me his notes and a preprint prior to publication, Nobuhiro Gō for sending me reprints, Masayo Ajiro for assistance in collecting many relevant papers, and Kazuko Iguchi for her continuous financial support and encouragement.

References

1. G. E. Schulz and R. H. Schirmer, *Principles of Protein Structure* (Springer-Verlag, New York, 1979); T. E. Creighton, *Proteins* (Freeman, New York, 1993).
2. C. B. Anfinsen, *Science* **181**, 223 (1973).
3. C. Levinthal, *J. Chim. Phys.* **65**, 44 (1968).
4. N. Gō, *Int. J. Peptide Protein Res.* **7**, 313 (1975); *Adv. Biophys.* **9**, 65 (1976); *Ann. Rev. Biophys. Bioeng.* **12**, 183 (1983).
5. B. H. Zimm and J. K. Bragg, *J. Chem. Phys.* **31**, 526 (1959).
6. L. D. Landau and E. M. Lifshitz, *Statistical Physics*, 3rd Edition Part 1 (Pergamon, New York, 1986).
7. N. Gō and H. Taketomi, *Proc. Natl. Acad. Sci.* USA **75**, 559 (1978); H. Taketomi, Y. Ueda and N. Gō, *Int. J. Peptide Protein Res.* **7**, 445 (1975); Y. Ueda, H. Taketomi and N. Gō, *Biopolymers* **17**, 1531 (1978); N. Gō and H. Taketomi, *Int. J. Peptide Protein Res.* **13**, 235, 447 (1979).
8. N. Gō, The consistency principle revisited, in *Old and New Views of Protein Folding*, eds. K. Kuwajima and M. Arai (Elsevier Science, New York, 1999).
9. P. S. Kim and R. L. Baldwin, *Annu. Rev. Biochem.* **51**, 459 (1982).
10. N. Saitô and Y. Kobayashi, *Int. J. Mod. Phys.* **B13**, 2431 (1999).
11. K. A. Dill, *Biochemistry* **24**, 1501 (1985); H. S. Chan and K. A. Dill, *Physics Today*, 24 (1993).
12. A. Šali *et al.*, *Nature* **369**, 248 (1994); E. I. Shaknovich, *Phys. Rev. Lett.* **72**, 3907 (1994); E. L. Kussel and E. I. Shaknovich, *ibid.* **83**, 4437 (1999).
13. P. G. Wolynes *et al.*, *Science* **267**, 1619 (1995); J. Wang *et al.*, *Phys. Rev. Lett.* **76**, 4861 (1996); E. D. Nelson *et al.*, *ibid.* **79**, 3534 (1997).
14. P.-A. Lindgård and H. Bohr, *Phys. Rev. Lett.* **77**, 779 (1996).
15. H. Li *et al.*, **Science 273**, 666 (1996); H. Li *et al.*, *Phys. Rev. Lett.* **79**, 765 (1997).

16. T. Haliloglu *et al.*, *Phys. Rev. Lett.* **79**, 3090 (1997).
17. H. J. Bussemaker *et al.*, *Phys. Rev. Lett.* **79**, 3530 (1997).
18. C. Micheletti *et al.*, *Phys. Rev. Lett.* **80**, 2237 (1998); *Proteins* **32**, 80 (1998); C. Clementi *et al.*, *ibid.* **81**, 3287 (1998); C. Micheletti *et al.*, *Phys. Rev. Lett.* **87**, 088102-1 (2001).
19. M. Vendruscolo *et al.*, *Phys. Rev. Lett.* **82**, 656 (1999).
20. S. Takahashi *et al.*, *Nature. Str. Biol.*, **4**, 45 (1997); S.-R. Yeh *et al.*, *ibid.* **4**, 51 (1997).
21. J. U. Bowie *et al.*, *Science* **253**, 164 (1991); M. Levitt and C. Chothia, *Nature* **261**, 552 (1992); J. U. Bowie and D. Eisenberg, *Curr. Opin. Struc. Biol.* **3**, 437 (1993); M. Wilmanns and D. Eisenberg, *Proc. Natl. Acad. Sci.* USA **90**, 1379 (1993).
22. K. A. Dill and H. S. Chan, *Nature Str. Biol.* **4**, 10 (1997); R. E. Burton *et al.*, *ibid.* **4**, 305 (1997).
23. S. E. Jackson and A. R. Fersht, *Biochemistry* **30**, 10428 (1991); *ibid.* **30**, 10436 (1991).
24. A. R. Fersht, *Curr. Opin. Str. Biol.* **5**, 79 (1994); *ibid.* **7**, 3 (1997); A. G. Ladurner and A. R. Fersht, *J. Mol. Biol.* **273**, 330 (1997); S. E. Jackson, Folding & Design **3**, R81 (1998).
25. T. Schindler and F. X. Schmid, *Biochemistry* **35**, 16833 (1996); R. E. Burton *et al.*, *ibid.* **37**, 5337 (1998).
26. B. Nölting, *Protein Folding Kinetics* (Springer-Verlag, Berlin, 1999). This is one of the review books for the latest developments in this direction. References are therein.
27. K. W. Plaxco *et al.*, *J. Mol. Biol.* **277**, 985 (1998); E. Alm and D. Baker, *Curr. Opin. Str. Biol.* **9**, 189 (1999); *Proc. Natl. Acad. Sci.* USA **96**, 11305 (1999); D. Baker, *Nature* **405**, 39 (2000).
28. S. Akiyama *et al.*, Progressive Formulation of α-helices after the Hydrophobic Collapse of Backbone Structures in the Folding Process of Cytochrome *c*, preprint, 2000.
29. R. Zwanzig, *Proc. Natl. Acad. Sci.* USA **92**, 9801 (1995). Also see, R. Zwanzig, *ibid.* **89**, 20 (1992); *ibid.* **94**, 148 (1997).
30. A. V. Finkelstein and A. Ya. Badretdinov, *Mol. Biol.* **31**, 391 (1997); Folding & Design **2**, 115 (1997); O. V. Galzitskaya and A. V. Finkelstein, *Proc. Natl. Acad. Sci.* USA **96**, 11299 (1999).
31. V. Muñoz and L. Serrano, Folding & Design **1**, R71 (1996); W. A. Eaton *et al.*, *Curr. Opin. Str. Biol.* **7**, 10 (1997); V. Muñoz and W. A. Eaton, *Proc. Natl. Acad. Sci.* USA **96**, 11311 (1999).
32. J. R. Gillespie and D. Shortle, *J. Mol. Biol.* **268**, 170 (1997).
33. K. Iguchi, *Mod. Phys. Lett.* **B12**, 499 (1998); *Int. J. Mod. Phys.* **B13**, 325 (1999); Mathematica programs generating the figures are shown in the following home page: http://www.stannet.ne.jp/kazumoto/RubikMagicSnakeS0j. html for $C = 0$ and http://www.stannet.ne.jp/ kazumoto/RubikMagicSnakeS1j. html for $C = 1$.
34. P.-G. de Gennes, *Scaling Concepts in Polymer Physics* (Cornell University Press, Ithaca, 1979).
35. K. Huang, *Statistical Mechanics*, 2nd Edition (John Wiley and Sons, New York, 1987).
36. V. S. Pande *et al.*, *Curr. Opin. Struc. Biol.* **8**, 68 (1998).
37. http://www.pdb.org. Altzheimer's amyloid is data-banked by the code "1AAP". Parkinson's ubiquitin is data-banked by the code "1UBQ". Creutzfeldt-Jacob's prion is data-banked by the code "1FKC".
38. S. E. Jackson, Folding & Design **3**, R81 (1998).
39. A. P. Capaldi and S. E. Radford, *Curr. Opin. Struc. Biol.* **8**, 86 (1998).