Doctoral Dissertation (Censored)

博士論文 （要約）

Prediction of protein folding mechanisms by structure-based statistical mechanical models

(構造ベースの統計力学モデルによる
タンパク質のフォールディング機構の予測)

A Dissertation Submitted for the Degree of Doctor of Philosophy

December 2021

令和 3 年 12 月博士（理学）申請

Department of Physics, Graduate School of Science,

The University of Tokyo

東京大学大学院理学系研究科物理学専攻

Koji Ooka

大岡　紘治

Abstract

Elucidation of protein folding mechanisms is one of the most important problems in biophysics. Free energy landscapes of protein folding based on statistical mechanics provide comprehensive understanding of protein folding pathways and folding intermediates. Wako-Saitô-Muñoz-Eaton (WSME) model, a coarse-grained statistical mechanical model of proteins, is a promising strategy for predicting the free energy landscapes based on protein structures. The WSME model has successfully predicted experimentally observed folding mechanisms of small proteins. However, the WSME model is unsuitable for prediction of folding reactions of large multi-domain proteins that are stabilized by non-local interactions. To overcome this limitation, the WSME-L model having the linker term, representing the non-local interaction was developed and the exact analytical solution was established.

The WSME-L model was applied to hen egg-white lysozyme by introducing the linkers at the non-local disulfide bonds. The model predicted parallel folding pathways with several intermediates, which was consistent with the experimentally observed folding mechanisms of the disulfide-intact lysozyme. Furthermore, the kinetic behaviors of the folding pathways were consistent with the experimental observations.

Next, the free energy landscapes of four proteins homologous to hen egg-white lysozyme (canine milk lysozyme, human α-lactalbumin, goat α-lactalbumin, and bovine α-lactalbumin) were calculated. Although these proteins have similar backbone structure, folding experiments showed that they have different folding pathways depending on the derived species. The WSME-L model successfully predicted the dominant folding pathways and residue-specific folding processes of these proteins consistent with the experimental observations. Moreover, virtual mutational analysis suggested that the folding pathways of lysozyme and α-lactalbumin are sensitive to the distribution of the native contacts. Therefore, the WSME-L model can extract subtle differences of folding mechanisms encoded in native structures.

Finally, the WSME-L model was modified to be applicable to transient formation of non-local interactions by introducing a virtual linker and was applied to apomyoglobin that does not have disulfide bonds but forms non-local interactions by hydrophobic collapse early in the folding process. The model successfully predicted the folding pathway of apomyoglobin consistent with the experimental observation. Thus, the WSME-L model may pave the way for predicting the folding mechanisms of large multi-domain proteins.

Index

1. General Introduction

1.1. Protein folding problem

Proteins are biopolymers consisting of chains of a dozen to several hundred amino acids and fold into specific compact structures in appropriate solvents. Depending on amino acid sequences, proteins form various three-dimensional structures and express a variety of functions. Therefore, elucidating the mechanisms of how proteins fold into specific structures, i.e., the protein folding problem, is one of the most important problems in biophysics and life sciences. Solving the protein folding problem will provide an understanding of structures and functions of proteins, leading to the understanding of many biological phenomena, in which proteins are involved. Recently, proteins are used in medical applications, such as antibody drugs, and industrial applications, such as bioenergy production. Therefore, understanding the folding mechanisms will also facilitate these applications.

Many experimental and theoretical studies have been conducted to elucidate the protein folding mechanisms. Anfinsen proposed that protein structure is determined only by its amino acid sequence (Anfinsen 1973) and Levinthal proposed that proteins fold through specific pathways (Levinthal 1969). Thus, since the 1970s, characterization of intermediates in folding reactions has been extensively performed (Baldwin 2005) and, to date, many experimental techniques have been established. For example, circular dichroism (CD) spectroscopy and fluorescence spectroscopy have been used to measure folding free energies with thermodynamic analysis. Experimental techniques such as stopped-flow methods have also been developed for measuring folding kinetics and enabled to characterize the structures of the transition state by Φ-value analysis (Fersht et al. 1992). Using nuclear magnetic resonance (NMR) spectroscopy, detailed structures of folding intermediates can be measured by combining hydrogen/deuterium (H/D) exchange methods (Schmid & Baldwin 1979). These experiments showed that small single-domain proteins with less than 100 residues fold in a two-state manner (Jackson 1998). In contrast, large multi-domain proteins with more than 100 residues exhibit complex folding behavior, involving a folding intermediate(s) and one or more folding pathways (Kuwajima 1989; Ptitsyn 1995; Arai & Kuwajima 2000; Arai 2018). These intermediates observed in the folding processes of large proteins, called molten globules, exhibit compact structures without tight packing of side chains and have native-like secondary structures (Ohgushi & Wada 1983; Kuwajima 1989; Arai & Kuwajima 2000).

1.2. Gō model

In general, it is difficult to theoretically obtain overall picture of the protein folding reaction because the proteins are many-body systems consisting of complicated chemical bonds. However, from the viewpoint of statistical mechanics, if the free energies of possible protein states are calculated and the free energy landscape of protein folding is described, comprehensive understanding of folding pathways can be obtained (Bryngelson et al. 1995). Thus, to calculate the free energy landscape of protein folding, many theoretical methods have been developed. Molecular dynamics (MD) simulation based on ab initio approach is one of the successful methods (McCammon et al. 1977). In the simulation, the atoms that make up a protein are regarded as point particles and their time evolution is calculated using the Newton's equations of motion and force fields that determine interactions between particles. Due to the recent remarkable development of computers and efficient sampling methods, long-time folding simulations have been performed for many proteins (Lindorff-Larsen et al. 2011). However, folding simulations of large proteins with more than 100 residues are still difficult because of the extremely large conformational space that should be sampled (Perez et al. 2016; Gershenson et al. 2020). In addition, deep learning approaches to predict protein three-dimensional structures from amino acid sequence have recently made a great leap (Baek et al. 2021; Jumper et al. 2021). However, the state-of-the-art protein structure prediction methods do not provide the understanding of how proteins fold (Outeiral et al. 2021).

By contrast, efficient sampling is possible using an artificial potential biased toward the native structure, known as Gō model. The Gō model considers an ideal protein and assumes that only the interactions formed in the native state contribute to stabilize the protein and all non-native interactions are excluded. This model is based on the consistency principle (Gō 1983) and the principle of minimal frustration (Bryngelson et al. 1995). The consistency principle states that local and non-local interactions are consistent with each other in an ideal protein, and the interactions formed in the native structure specifically stabilize the native state. In other words, the native state has the lowest energy among the possible conformations, and proteins have funnel-shaped energy landscape (Bryngelson et al. 1995) (Fig. 1). Protein molecules on the surface of the funnel fold toward the bottom corresponding to the native state. Cross section perpendicular to the energy axis represents the number of possible states. The funnel indicates that the number of protein states converges to one or small numbers after folding into the stable native state. Similarly, the principle of minimal frustration states that energetic frustrations corresponding to roughness of the funnel surface should be minimal

for a protein to be foldable. Since the ruggedness of the funnel surface is due to the inconsistency between the interactions in the native state and those formed during folding, both the consistency principle and the principle of minimal frustration are equivalent. Taken together, the energy landscape of ideal proteins has a shape of a perfect funnel. If no such bias toward the native structure was present, a protein would have to explore a tremendous amount of conformational space during the folding process and would not find the native state on biological timescales, suggesting that naturally occurring proteins have evolved to have a smooth funnel-like energy landscape biased toward the native state in order to quickly fold into the correct structure (Onuchic & Wolynes 2004). Hence, the Gō model considers only the native-like interactions. Note that although the Gō model requires the native structure, which is the answer to the protein structure prediction, finding the folding pathway(s) is a non-trivial problem even if the native structure is given.

The Gō model was originally developed as a simple lattice model for protein folding study (Taketomi et al. 1975). Amino acid residues are represented as points on a lattice, and stable interactions are formed only when residue pairs are in contact with each other in the native structure. This study proposed a consistency between local interactions that form secondary structure and non-local interactions that form tertiary structure in the native state of proteins (Gō & Taketomi 1978). After the lattice model was extensively studied, a coarse-grained off-lattice model using the Gō potential was constructed (Clementi et al. 2000). In this model, each amino acid residue is represented as a single bead that can be located anywhere in three-dimensional space, and the entire protein is treated as a chain of the beads constrained by a harmonic potential referenced to the native structure. An attractive interaction is implemented to the native pairs, and non-native pairs have repulsive interactions. Due to such bias toward the native state, the off-lattice Gō model allows for efficient folding simulations compared to MD simulations that use physically accurate potentials. The folding free energy landscapes were calculated for various proteins using this model and its extended version, and it was revealed that the predicted folding pathways and structures in the transition state are qualitatively consistent with experimental results (Koga & Takada 2001; Karanicolas & Brooks 2002; Kouza et al. 2006; Hills & Brooks 2008; Li et al. 2012).

Simple statistical mechanical models based on the Gō potential were also developed concurrently with the development of the off-lattice Gō model (Alm & Baker 1999; Galzitskaya & Finkelstein 1999; Muñoz & Eaton 1999). These models assume that, as folding reaction proceeds, total energy of the protein is decreased due to formation of the native interactions. In addition to this Gō-type potential, the reduction of conformational

entropy of the main chain attributed to structural constraint along the folding process is incorporated. These physical properties show energy-entropy compensation and provide free energy barrier. Despite their simplicity, the statistical models successfully explained the experimentally observed features such as folding rate of small single-domain proteins, indicating that both the consistency principle and the principle of minimal frustration hold for small proteins.

Since the Gō model completely ignores non-native interactions, it is difficult to fully capture the folding mechanism of proteins, which form non-native structures in folding intermediates such as β-lactoglobulin (Hamada et al. 1996; Arai et al. 1998; Chikenji & Kikuchi 2000). However, the success of the Gō model in predicting many aspects of protein folding has confirmed the importance of the role of native interaction (Zhou & Karplus 1997; Matysiak & Clementi 2004). All-atom MD simulations with physics-based force-field have validated the role of native and non-native interactions, and trajectories of long-time folding simulation of small proteins supported the assumption of the Gō model (Lindorff-Larsen et al. 2011; Best et al. 2013; Henry et al. 2013).

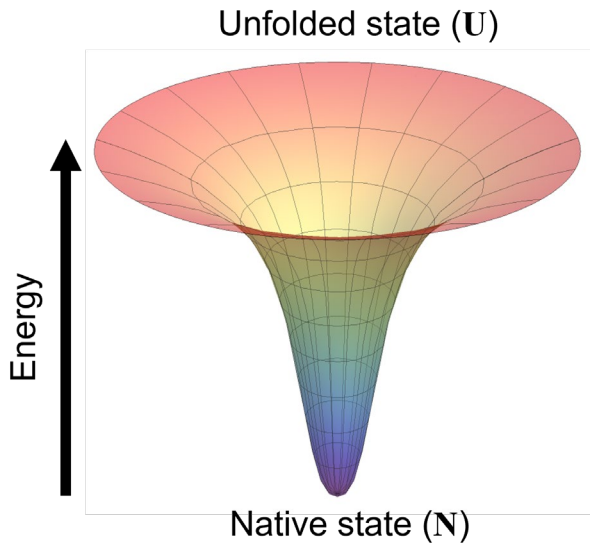

Figure 1. Perfect funnel-like energy landscape of protein folding.

This energy landscape is for an ideal protein based on the consistency principle and the principle of minimal frustration. From the unfolded state, protein molecules slide down the funnel surface and finally reach the native state.

1.3. WSME model

Wako-Saitô-Muñoz-Eaton (WSME) model is one of the promising models for calculating free energy landscapes of protein folding (Wako & Saitô 1978a; Wako & Saitô 1978b; Muñoz & Eaton 1999). This model was originally proposed by Wako and Saitô in 1970s and was later re-discovered by Muñoz and Eaton in 1999. The WSME model has three key features. First, this model is one of the Gō-type models and considers only the interactions formed in the native state. Second, this model is a simple, coarse-grained statistical mechanical model, and the exact solution of the partition function can be calculated (Bruscolini & Pelizzola 2002). Thus, the free energy landscapes can be readily obtained using a computer due to this simplicity. Third, this model assumes that the folding is initiated as the result of local interactions between neighboring residues and spreads to distal regions via the growth and docking of native segments. Such a folding picture is based on a framework model of protein folding (Baldwin 2005).

The details of the original WSME model are as follows. First, an Ising-like two-state variable m_k is assigned to each residue of a protein. The index k is the residue number. These residue states correspond to the orientation of main-chain dihedral angles: $m_k = 1$ when the residue is in the native-like conformation and $m_k = 0$ when the residue takes unfolded state. The protein state $\{m\}$ is described as the set of the residue states ($m_1, m_2, \ldots, m_N$) where N is the total number of residues. In this model, the protein is coarse-grained and computationally grouped into 2^N states. Next, Hamiltonian of the WSME model is defined as follows:

$$H(\{m\}) = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \varepsilon_{i,j} m_{i,j} \tag{1}$$

where $\varepsilon_{i,j}$ is contact energy in the native state between residues i and j. $\varepsilon_{i,j}$ takes negative value when the stable interaction is formed in the native state. $m_{i,j}$ is defined as follows:

$$m_{i,j} = m_i m_{i+1} \cdots m_j = \prod_{k=i}^{j} m_k \tag{2}$$

and m_{ij} = 1 only when all residues between i and j are folded. Therefore, the native interactions between residues i and j are established only when all intervening residues fold cooperatively into their native conformations (Fig. 2).

The number of states W is defined as follows:

$$W(\{m\}) = \exp\left[\left(S_0 + \sum_{i=1}^{N} S_i m_i\right) / k_B\right] \tag{3}$$

where k_B is the Boltzmann constant, S_0 is the conformation entropy of the fully unfolded state, and S_i (<0) is the entropic reduction attributed to the formation of the native confirmation. Equation 3 shows that the number of states is explicitly described using the conformational entropy of the main chain, and the possible states decrease along the folding process. S_0 was set to 0 in later calculations because S_0 is constant and does not affect the result.

An order parameter of the magnetic phase transition is described by magnetization in the Ising model. From analogy with the magnetization, the order parameter of the degree of the native structure formation n is described as:

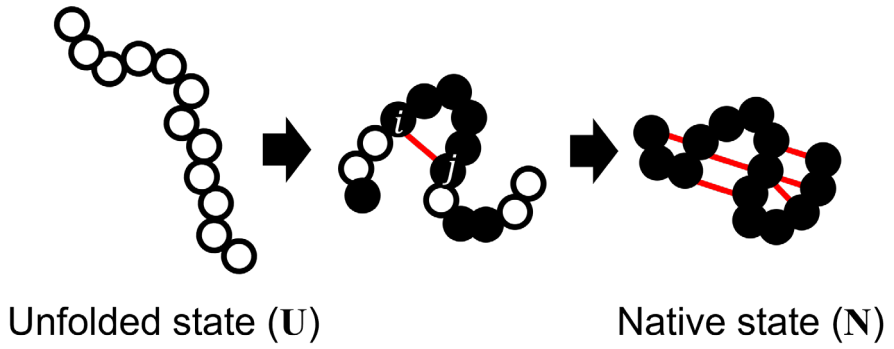

Figure 2. Schematic representation of the folding process in the WSME model.

Residues in native or unfolded conformations (m_k = 1 or 0) are shown in filled or open circles, respectively. The native contact between residues i and j (red line) is formed only when these residues are connected by a native stretch along the main chain, that is, when all intervening residues between them are in the native conformations

$$n = \frac{1}{N}\sum_{i=1}^{N} m_i \tag{4}$$

$n = 0$ when the protein is fully unfolded, and $n = 1$ when the protein is completely folded (Fig. 2). Then, the partition function Z restricted by the order parameter is denoted as follows:

$$\begin{aligned} Z(n) &= \mathrm{Tr}_n W(\{m\}) \exp\left[-\frac{H(\{m\})}{k_\mathrm{B}T}\right] \\ &= \mathrm{Tr}_n \exp\left[-\frac{1}{k_\mathrm{B}T}\left(\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} \varepsilon_{i,j} m_{i,j} - T\sum_{i=1}^{N} S_i m_i\right)\right] \end{aligned} \tag{5}$$

where T is the absolute temperature and Tr_n is the sum of all possible states under the constraint n. When M residues are folded, $n = M / N$ and the ${}_N C_M$ states are added up. The free energy $F(n) = -k_\mathrm{B}T \ln Z(n)$ is finally obtained from the partition function.

The WSME model has successfully explained the experimentally inferred folding mechanisms of small single-domain proteins (Muñoz & Eaton 1999; Itoh & Sasai 2006; Bruscolini & Naganathan 2011). An example is the folding of the B-domain of protein A (BdpA), which consists of three α-helices. Experimental results showed that the middle helix folds first and the C-terminal helix follows (Sato et al. 2004). Using the WSME model. the free energy landscapes consistent with the experimental results was obtained (Itoh & Sasai 2006). Moreover, this model has been successfully used to predict the folding free energy landscapes of multi-domain proteins with end-to-end tandem connection of small globular domains, for which the folding of each domain precedes the docking of both domains (Itoh & Sasai 2008; Itoh & Sasai 2009).

However, the WSME model is unsuitable for prediction of folding reactions of large multi-domain proteins that are stabilized by non-local interactions (Itoh & Sasai 2008; Inanami et al. 2014; Sasai et al. 2016; Gopi et al. 2019). For example, multi-domain proteins, which have an insertion of one domain in another domain requires non-local interactions to obtain a reasonable free energy landscape, because the folding of a discontinuous domain may occur through non-local hydrophobic collapse mechanisms

before the folding of the intervening domain (Arai & Kuwajima 2000; Arai 2018). To take into consideration the folding of a discontinuous domain even when the intervening continuous domain is disordered, Inanami et al. introduced a virtual linker between the N- and C-termini of *Escherichia coli* dihydrofolate reductase (DHFR) (Inanami et al. 2014; Muñoz 2014; Sasai et al. 2016). The virtual ring closure of the polypeptide chain enhanced the non-local interactions in the discontinuous domain and resulted in multiple minima in the free energy landscape of DHFR, which is consistent with multiple folding intermediates observed in experiments (Jennings et al. 1993; Arai et al. 2003a; Arai et al. 2003b; Arai et al. 2011). Thus, the introduction of a virtual linker in the WSME model enables the consideration of non-local interactions during the folding of multi-domain proteins. After the development of the extended WSME model, which has a virtual linker at the N- and C-termini of a protein, Sasai et al. proposed an idea of introducing multiple linkers at arbitrary positions to calculate free energy landscapes of the proteins that have more complicated domain arrangement (Sasai et al. 2016). However, the idea was only a proposal for model extension, and the construction of a concrete model, the establishment of the calculation method, and their application to actual proteins have not been accomplished.

Therefore, to make the WSME model generally applicable to any protein, it is necessary to develop a method for the introduction of multiple linkers at any positions in a protein. For example, an all-α type protein, apomyoglobin, folds through intermediates that have non-local interactions between distant helices (Jennings & Wright 1993; Tsui et al. 1999). Since multi-domain proteins are abundant in the proteomes and the average size of proteins is 300–400 residues (Milo & Phillips 2015), a key challenge that remains to be solved is to predict the folding mechanisms of multi-domain proteins.

In addition, the original WSME model cannot take into consideration a branched connection of a polypeptide chain introduced by a disulfide bond(s), which connects non-local segments of the chain by a covalent linker(s). The disulfide bond is an explicit non-local interaction that affects folding reactions. For instance, lysozyme is a multi-domain protein with four disulfide bonds, and the folding pathway has been experimentally investigated in the presence of intact disulfide bonds (Dobson et al. 1998). The WSME model cannot consider the disulfide bonds due to the model assumption. Therefore, it is necessary to develop a model that can account for such covalent non-local interactions.

1.4. Purpose

As described above, to calculate the free energy landscapes of various large, multi-domain proteins, it is necessary to develop an extended version of the WSME model that can account for multiple non-local linkers at arbitrary positions, not limited to a single linker between the N- and C-termini. In Chapter 2, I developed the WSME-L model to take into account specific non-local interactions by introducing multiple covalent linkers at arbitrary positions and constructed a calculation method for the partition function of the WSME-L model. Then, to verify whether the WSME-L model supports experimental results, the free energy landscapes of the disulfide-intact hen egg-white lysozyme (HEWL) was calculated.

In Chapter 3, I calculated the free energy landscapes of four proteins homologous to HEWL. Although lysozymes and α-lactalbumins are homologous to each other and have similar backbone structure, folding experiments in the presence of disulfide bonds showed that the folding pathways differ depending on the derived species (Nakamura et al. 2010). Thus, I examined whether the WSME-L model can explain the differences in the folding pathways of homologous proteins.

In Chapters 2 and 3, the WSME-L model was used for the proteins possessing disulfide bonds, which are explicit non-local interactions. In Chapter 4, to construct the model that can take into account transiently formed non-covalent, non-local interactions at an arbitrary position, not limited to the non-local interactions between the N- and C-termini, I provided a specific equation of this model with a virtual linker at an arbitrary position. Then, I calculated the free energy landscapes of apomyoglobin and examined whether the free energy landscapes calculated by the extended model are consistent with experimental results.

2. Development of the WSME-L model and prediction of lysozyme folding

本章については、5 年以内に雑誌等で刊行予定のため、非公開。

3. Prediction of different folding pathways of homologous proteins: lysozymes and α-lactalbumins

本章については、5 年以内に雑誌等で刊行予定のため、非公開。

4. Prediction of the folding pathway of apomyoglobin

本章については、5 年以内に雑誌等で刊行予定のため、非公開。

5. General Discussion

本章については、5年以内に雑誌等で刊行予定のため、非公開。

6. Conclusions

In this study, to predict protein folding mechanisms of large multi-domain proteins, the WSME-L model with linker terms representing non-local interactions was developed and the exact analytical solution was established by modifying the calculation method for the original model. First, this model was applied to a model protein, hen egg-white lysozyme (HEWL), by introducing the covalent linkers at the disulfide bonds. The free energy landscape predicted by the WSME-L model had parallel folding pathways with several intermediates, which was consistent with the experimentally observed folding mechanisms of the disulfide-intact lysozyme. The kinetic analysis based on the free energy landscape showed that the time-dependent behaviors of kinetic species were consistent with the experimentally observed behaviors. Moreover, theoretical Φ-value analysis integrated into the WSME-L model provided detailed insights into structure formation via the folding pathways, and the single linker analysis provided the understanding of the role of each disulfide bond.

Next, the WSME-L model with an extension of ion-binding energy was applied to predict the folding pathways of several proteins homologous to HEWL: canine milk lysozyme, human α-lactalbumin, goat α-lactalbumin, and bovine α-lactalbumin. Although these proteins have similar backbone structure, the folding experiments showed that they have different folding pathways depending on the derived species. The WSME-L model successfully predicted the dominant folding pathways and residue-specific structure formation consistent with the experimental observations. Moreover, virtual mutational analysis revealed the native contacts important in determining folding pathways and further predicted the mutations that can modulate the folding mechanisms. Thus, the WSME-L model can explain the subtle differences in the folding mechanisms encoded in native structures.

Finally, the WSME-L model was extended to consider non-covalent non-local interactions and was applied to apomyoglobin, which shows hydrophobic collapse early in the folding process. The free energy landscape calculated by the WSME-L model with a transient linker successfully predicted the folding pathway of apomyoglobin consistent with the experimental observation. Taken together, the WSME-L model can pave the way for predicting the folding mechanisms of large multi-domain proteins.

7. References

Abraham MJ, Murtola T, Schulz R, Páll S, Smith JC, Hess B, Lindahl E (2015) GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* 1-2: 19-25.

Alm E, Baker D (1999) Prediction of protein-folding mechanisms from free-energy landscapes derived from native structures. *Proc Natl Acad Sci U S A* 96(20): 11305-11310.

Anfinsen CB (1973) Principles that govern the folding of protein chains. *Science* 181(4096): 223-230.

Arai M (2018) Unified understanding of folding and binding mechanisms of globular and intrinsically disordered proteins. *Biophys Rev* 10(2): 163-181.

Arai M, Ikura T, Semisotnov GV, Kihara H, Amemiya Y, Kuwajima K (1998) Kinetic refolding of β-lactoglobulin. Studies by synchrotron X-ray scattering, and circular dichroism, absorption and fluorescence spectroscopy. *J Mol Biol* 275(1): 149-162.

Arai M, Iwakura M (2006) Peptide fragment studies on the folding elements of dihydrofolate reductase from *Escherichia coli*. *Proteins* 62(2): 399-410.

Arai M, Iwakura M, Matthews CR, Bilsel O (2011) Microsecond subdomain folding in dihydrofolate reductase. *J Mol Biol* 410(2): 329-342.

Arai M, Kataoka M, Kuwajima K, Matthews CR, Iwakura M (2003a) Effects of the difference in the unfolded-state ensemble on the folding of Escherichia coli dihydrofolate reductase. *J Mol Biol* 329(4): 779-791.

Arai M, Kondrashkina E, Kayatekin C, Matthews CR, Iwakura M, Bilsel O (2007) Microsecond hydrophobic collapse in the folding of *Escherichia coli* dihydrofolate reductase, an α/β-type protein. *J Mol Biol* 368(1): 219-229.

Arai M, Kuwajima K (2000) Role of the molten globule state in protein folding. *Adv Protein Chem* 53: 209-282.

Arai M, Maki K, Takahashi H, Iwakura M (2003b) Testing the relationship between foldability and the early folding events of dihydrofolate reductase from Escherichia coli. *J Mol Biol* 328(1): 273-288.

Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, Wang J, Cong Q, Kinch LN, Schaeffer RD, Millan C, Park H, Adams C, Glassman CR, DeGiovanni A, Pereira JH, Rodrigues AV, van Dijk AA, Ebrecht AC, Opperman DJ, Sagmeister T, Buhlheller C, Pavkov-Keller T, Rathinaswamy MK, Dalwadi U, Yip CK, Burke JE, Garcia KC, Grishin NV, Adams PD, Read RJ, Baker D (2021) Accurate prediction of protein structures and interactions using a three-track neural network. *Science* 373(6557):

871-876.

Baldwin RL (2005) Early days of studying the mechanism of protein folding. *Protein Folding Handbook*: 3-21.

Baryshnikova EN, Melnik BS, Finkelstein AV, Semisotnov GV, Bychkova VE (2005) Three-state protein folding: experimental determination of free-energy profile. *Protein Sci* 14(10): 2658-2667.

Berendsen HJC, Postma JPM, van Gunsteren WF, DiNola A, Haak JR (1984) Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics* 81(8): 3684-3690.

Best RB, Hummer G, Eaton WA (2013) Native contacts determine protein folding mechanisms in atomistic simulations. *Proc Natl Acad Sci U S A* 110(44): 17874-17879.

Bieri O, Kiefhaber T (2001) Origin of apparent fast and non-exponential kinetics of lysozyme folding measured in pulsed hydrogen exchange experiments. *J Mol Biol* 310(4): 919-935.

Bruscolini P, Naganathan AN (2011) Quantitative prediction of protein folding behaviors from a simple statistical model. *J Am Chem Soc* 133(14): 5372-5379.

Bruscolini P, Pelizzola A (2002) Exact solution of the Munoz-Eaton model for protein folding. *Phys Rev Lett* 88(25 Pt 1): 258101.

Bryngelson JD, Onuchic JN, Socci ND, Wolynes PG (1995) Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins* 21(3): 167-195.

Bushmarina NA, Blanchet CE, Vernier G, Forge V (2006) Cofactor effects on the protein folding reaction: acceleration of α-lactalbumin refolding by metal ions. *Protein Sci* 15(4): 659-671.

Bussi G, Zykova-Timan T, Parrinello M (2009) Isothermal-isobaric molecular dynamics using stochastic velocity rescaling. *J Chem Phys* 130(7): 074101.

Case DA, Aktulga HM, Belfon K, Ben-Shalom IY, Brozell SR, Cerutti DS, Cheatham TE, I, Cisneros GA, Cruzeiro VWD, Darden TA, Duke RE, Giambasu G, Gilson MK, Gohlke H, Goetz AW, Harris R, Izadi S, Izmailov SA, Jin C, Kasavajhala K, Kaymak MC, King E, Kovalenko A, Kurtzman T, Lee TS, LeGrand S, Li P, Lin C, Liu J, Luchko T, Luo R, Machado M, Man V, Manathunga M, Merz KM, Miao Y, Mikhailovskii O, Monard G, Nguyen H, O'Hearn KA, Onufriev A, Pan F, Pantano S, Qi R, Rahnamoun A, Roe DR, Roitberg A, Sagui C, Schott-Verdugo S, Shen J, Simmerling CL, Skrynnikov NR, Smith J, Swails J, Walker RC, Wang J, Wei H, Wolf RM, Wu X, Xue Y, York DM, Zhao S, Kollman PA (2021). Amber 2021. University of California, San Francisco.

Case DA, Ben-Shalom IY, Brozell SR, Cerutti DS, Cheatham TEI, Cruzeiro VWD, Darden TA, Duke RE, Ghoreishi D, Gilson MK, Gohlke H, Goetz AW, Greene D, Harris

R, Homeyer N, Huang Y, Izadi S, Kovalenko A, Kurtzman T, Lee TS, LeGrand S, Li P, Lin C, Liu J, Luchko T, Luo R, Mermelstein DJ, Merz KM, Miao Y, Monard G, Nguyen C, Nguyen H, Omelyan I, Onufriev A, Pan F, Qi R, Roe DR, Roitberg A, Sagui C, Schott-Verdugo S, Shen J, Simmerling CL, Smith J, SalomonFerrer R, Swails J, Walker RC, Wang J, Wei H, Wolf RM, Wu X, Xiao L, York DM, Kollman PA (2018). AMBER 2018, University of California, San Francisco.

Chaudhuri TK, Arai M, Terada TP, Ikura T, Kuwajima K (2000) Equilibrium and kinetic studies on folding of the authentic and recombinant forms of human α-lactalbumin by circular dichroism spectroscopy. *Biochemistry* 39(50): 15643-15651.

Chedad A, Van Dael H (2004) Kinetics of folding and unfolding of goat α-lactalbumin. *Proteins* 57(2): 345-356.

Chikenji G, Kikuchi M (2000) What is the role of non-native intermediates of β-lactoglobulin in protein folding? *Proc Natl Acad Sci U S A* 97(26): 14273-14277.

Clementi C, Nymeyer H, Onuchic JN (2000) Topological and energetic factors: what determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? An investigation for small globular proteins. *J Mol Biol* 298(5): 937-953.

Darden T, York D, Pedersen L (1993) Particle mesh Ewald - An $N \cdot \log(N)$ method for Ewald sums in large systems. *J Chem Phys* 98(12): 10089-10092.

Daura X, Mark AE, van Gunsteren WF (1999a) Peptide folding simulations: no solvent required? *Comp Phys Commun* 123(1): 97-102.

Daura X, van Gunsteren WF, Mark AE (1999b) Folding-unfolding thermodynamics of a β-heptapeptide from equilibrium simulations. *Proteins* 34(3): 269-280.

Dobson CM, Evans PA, Radford SE (1994) Understanding how proteins fold: the lysozyme story so far. *Trends Biochem Sci* 19(1): 31-37.

Dobson CM, Sali A, Karplus M (1998) Protein folding: A perspective from theory and experiment. *Angew Chem Int Ed Engl* 37(7): 868-893.

Eliezer D, Wright PE (1996) Is apomyoglobin a molten globule? Structural characterization by NMR. *J Mol Biol* 263(4): 531-538.

Faccin M, Bruscolini P, Pelizzola A (2011) Analysis of the equilibrium and kinetics of the ankyrin repeat protein myotrophin. *J Chem Phys* 134(7): 075102.

Fersht AR, Matouschek A, Serrano L (1992) The folding of an enzyme. I. Theory of protein engineering analysis of stability and pathway of protein folding. *J Mol Biol* 224(3): 771-782.

Finkelstein AV, Badretdinov AY (1997) Physical reason for fast folding of the stable spatial structure of proteins: A solution of the Levinthal paradox. *Mol Biol* 31(3): 391-

398.

Forge V, Wijesinha RT, Balbach J, Brew K, Robinson CV, Redfield C, Dobson CM (1999) Rapid collapse and slow structural reorganisation during the refolding of bovine α-lactalbumin. *J Mol Biol* 288(4): 673-688.

Galzitskaya OV, Finkelstein AV (1999) A theoretical search for folding/unfolding nuclei in three-dimensional protein structures. *Proc Natl Acad Sci U S A* 96(20): 11299-11304.

Gershenson A, Gosavi S, Faccioli P, Wintrode PL (2020) Successes and challenges in simulating the folding of large proteins. *J Biol Chem* 295(1): 15-33.

Gō N (1983) Theoretical studies of protein folding. *Annu Rev Biophys Bioeng* 12: 183-210.

Gō N, Taketomi H (1978) Respective roles of short- and long-range interactions in protein folding. *Proc Natl Acad Sci U S A* 75(2): 559-563.

Gopi S, Aranganathan A, Naganathan AN (2019) Thermodynamics and folding landscapes of large proteins from a statistical mechanical model. *Curr Res Struct Biol* 1: 6-12.

Guez V, Roux P, Navon A, Goldberg ME (2002) Role of individual disulfide bonds in hen lysozyme early folding steps. *Protein Sci* 11(5): 1136-1151.

Hamada D, Segawa S, Goto Y (1996) Non-native α-helical intermediate in the refolding of β-lactoglobulin, a predominantly β-sheet protein. *Nat Struct Biol* 3(10): 868-873.

Henry ER, Best RB, Eaton WA (2013) Comparing a simple theoretical model for protein folding with all-atom molecular dynamics simulations. *Proc Natl Acad Sci U S A* 110(44): 17880-17885.

Hess B (2008) P-LINCS: A parallel linear constraint solver for molecular simulation. *J Chem Theory Comput* 4(1): 116-122.

Hills RD, Jr., Brooks CL, 3rd (2008) Subdomain competition, cooperativity, and topological frustration in the folding of CheY. *J Mol Biol* 382(2): 485-495.

Ikeguchi M, Kuwajima K, Sugai S (1986) Ca^{2+}-induced alteration in the unfolding behavior of α-lactalbumin. *J Biochem* 99(4): 1191-1201.

Inanami T, Terada TP, Sasai M (2014) Folding pathway of a multidomain protein depends on its topology of domain connectivity. *Proc Natl Acad Sci U S A* 111(45): 15969-15974.

Itoh K, Sasai M (2006) Flexibly varying folding mechanism of a nearly symmetrical protein: B domain of protein A. *Proc Natl Acad Sci U S A* 103(19): 7298-7303.

Itoh K, Sasai M (2008) Cooperativity, connectivity, and folding pathways of multidomain proteins. *Proc Natl Acad Sci U S A* 105(37): 13865-13870.

Itoh K, Sasai M (2009) Multidimensional theory of protein folding. *J Chem Phys* 130(14).

Itoh K, Sasai M (2010) Entropic mechanism of large fluctuation in allosteric transition. *Proc Natl Acad Sci U S A* 107(17): 7775-7780.

Itoh K, Sasai M (2011) Statistical mechanics of protein allostery: Roles of backbone and side-chain structural fluctuations. *J Chem Phys* 134(12): 125102.

Jackson SE (1998) How do small single-domain proteins fold? *Fold Des* 3(4): R81-91.

Jennings PA, Finn BE, Jones BE, Matthews CR (1993) A reexamination of the folding mechanism of dihydrofolate reductase from Escherichia coli: verification and refinement of a four-channel model. *Biochemistry* 32(14): 3783-3789.

Jennings PA, Wright PE (1993) Formation of a molten globule intermediate early in the kinetic folding pathway of apomyoglobin. *Science* 262(5135): 892-896.

Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics* 79(2): 926-935.

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Zidek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D (2021) Highly accurate protein structure prediction with AlphaFold. *Nature* 596(7873): 583-589.

Karanicolas J, Brooks CL, 3rd (2002) The origins of asymmetry in the folding transition states of protein L and protein G. *Protein Sci* 11(10): 2351-2361.

Kiefhaber T (1995) Kinetic traps in lysozyme folding. *Proc Natl Acad Sci U S A* 92(20): 9029-9033.

Kobashigawa Y, Demura M, Koshiba T, Kumaki Y, Kuwajima K, Nitta K (2000) Hydrogen exchange study of canine milk lysozyme: stabilization mechanism of the molten globule. *Proteins* 40(4): 579-589.

Koga N, Takada S (2001) Roles of native topology and chain-length scaling in protein folding: a simulation study with a Gō-like model. *J Mol Biol* 313(1): 171-180.

Koshiba T, Yao M, Kobashigawa Y, Demura M, Nakagawa A, Tanaka I, Kuwajima K, Nitta K (2000) Structure and thermodynamics of the extraordinarily stable molten globule state of canine milk lysozyme. *Biochemistry* 39(12): 3248-3257.

Kouza M, Li MS, O'Brien E P, Jr., Hu CK, Thirumalai D (2006) Effect of finite size on cooperativity and rates of protein folding. *J Phys Chem A* 110(2): 671-676.

Kuwajima K (1989) The molten globule state as a clue for understanding the folding and

cooperativity of globular-protein structure. *Proteins* 6(2): 87-103.

Lai JK, Kubelka GS, Kubelka J (2015) Sequence, structure, and cooperativity in folding of elementary protein structural motifs. *Proc Natl Acad Sci U S A* 112(32): 9890-9895.

Levinthal C (1969) How to fold graciously. *Mössbauer Spectroscopy in Biological Systems Proceedings*: 22-24.

Li W, Terakawa T, Wang W, Takada S (2012) Energy landscape and multiroute folding of topologically complex proteins adenylate kinase and 2ouf-knot. *Proc Natl Acad Sci U S A* 109(44): 17789-17794.

Lindorff-Larsen K, Piana S, Dror RO, Shaw DE (2011) How fast-folding proteins fold. *Science* 334(6055): 517-520.

Maier JA, Martinez C, Kasavajhala K, Wickstrom L, Hauser KE, Simmerling C (2015) ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J Chem Theory Comput* 11(8): 3696-3713.

Matagne A, Chung EW, Ball LJ, Radford SE, Robinson CV, Dobson CM (1998) The origin of the α-domain intermediate in the folding of hen lysozyme. *J Mol Biol* 277(5): 997-1005.

Matagne A, Jamin M, Chung EW, Robinson CV, Radford SE, Dobson CM (2000) Thermal unfolding of an intermediate is associated with non-Arrhenius kinetics in the folding of hen lysozyme. *J Mol Biol* 297(1): 193-210.

Matagne A, Radford SE, Dobson CM (1997) Fast and slow tracks in lysozyme folding: insight into the role of domains in the folding process. *J Mol Biol* 267(5): 1068-1074.

Matsushita K, Kikuchi M (2013) Frustration-induced protein intrinsic disorder. *J Chem Phys* 138(10): 105101.

Matysiak S, Clementi C (2004) Optimal combination of theory and experiment for the characterization of the protein folding landscape of S6: how far can a minimalist model go? *J Mol Biol* 343(1): 235-248.

McCammon JA, Gelin BR, Karplus M (1977) Dynamics of folded proteins. *Nature* 267(5612): 585-590.

Miller BR, 3rd, McGee TD, Jr., Swails JM, Homeyer N, Gohlke H, Roitberg AE (2012) MMPBSA.py: An efficient program for end-state free energy calculations. *J Chem Theory Comput* 8(9): 3314-3321.

Milo R, Phillips R (2015). Cell biology by the numbers. New York, NY, Garland Science.

Miranker A, Robinson CV, Radford SE, Aplin RT, Dobson CM (1993) Detection of transient protein folding populations by mass spectrometry. *Science* 262(5135): 896-900.

Mizuguchi M, Masaki K, Nitta K (1999) The molten globule state of a chimera of human

α-lactalbumin and equine lysozyme. *J Mol Biol* 292(5): 1137-1148.

Moriwaki Y, Terada T, Tsumoto K, Shimizu K (2015) Rapid heme transfer reactions between NEAr transporter domains of *Staphylococcus aureus*: A theoretical study using QM/MM and MD simulations. *PLoS One* 10(12): e0145125.

Muñoz V (2014) A simple theoretical model goes a long way in explaining complex behavior in protein folding. *Proc Natl Acad Sci U S A* 111(45): 15863-15864.

Muñoz V, Eaton WA (1999) A simple model for calculating the kinetics of protein folding from three-dimensional structures. *Proc Natl Acad Sci U S A* 96(20): 11311-11316.

Muttathukattil AN, Singh PC, Reddy G (2019) Role of disulfide bonds and topological frustration in the kinetic partitioning of lysozyme folding pathways. *J Phys Chem B* 123(15): 3232-3241.

Nakamura T, Makabe K, Tomoyori K, Maki K, Mukaiyama A, Kuwajima K (2010) Different folding pathways taken by highly homologous proteins, goat α-lactalbumin and canine milk lysozyme. *J Mol Biol* 396(5): 1361-1378.

Nakao M, Arai M, Koshiba T, Nitta K, Kuwajima K (2003) Folding mechanism of canine milk lysozyme studied by circular dichroism and fluorescence spectroscopy. *Spectroscopy-An International Journal* 17(2-3): 183-193.

Narayan A, Naganathan AN (2018) Switching protein conformational substates by protonation and mutation. *J Phys Chem B* 122(49): 11039-11047.

Nelson ED, Grishin NV (2006) Scaling approach to the folding kinetics of large proteins. *Phys Rev E Stat Nonlin Soft Matter Phys* 73(1 Pt 1): 011904.

Nishimura C (2017) Folding of apomyoglobin: Analysis of transient intermediate structure during refolding using quick hydrogen deuterium exchange and NMR. *Proc Jpn Acad Ser B Phys Biol Sci* 93(1): 10-27.

Ohgushi M, Wada A (1983) 'Molten-globule state': a compact form of globular proteins with mobile side-chains. *FEBS Lett* 164(1): 21-24.

Onuchic JN, Wolynes PG (2004) Theory of protein folding. *Curr Opin Struct Biol* 14(1): 70-75.

Outeiral C, Nissley DA, Deane CM (2021). Current protein structure predictors do not produce meaningful folding pathways. bioRxiv**:** 2021.2009.2020.461137.

Parrinello M, Rahman A (1981) Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics* 52(12): 7182-7190.

Perez A, Morrone JA, Simmerling C, Dill KA (2016) Advances in free-energy-based simulations of protein folding and ligand binding. *Curr Opin Struct Biol* 36: 25-31.

Ptitsyn OB (1995) Molten globule and protein folding. *Adv Protein Chem* 47: 83-229.

Radford SE, Buck M, Topping KD, Dobson CM, Evans PA (1992) Hydrogen exchange in native and denatured states of hen egg-white lysozyme. *Proteins* 14(2): 237-248.

Radford SE, Dobson CM (1995) Insights into protein folding using physical techniques: studies of lysozyme and α-lactalbumin. *Philos Trans R Soc Lond B Biol Sci* 348(1323): 17-25.

Ramboarina S, Redfield C (2003) Structural characterisation of the human α-lactalbumin molten globule at high temperature. *J Mol Biol* 330(5): 1177-1188.

Redfield C, Schulman BA, Milhollen MA, Kim PS, Dobson CM (1999) α-lactalbumin forms a compact molten globule in the absence of disulfide bonds. *Nat Struct Biol* 6(10): 948-952.

Saeki K, Arai M, Yoda T, Nakao M, Kuwajima K (2004) Localized nature of the transition-state structure in goat α-lactalbumin folding. *J Mol Biol* 341(2): 589-604.

Sasai M, Chikenji G, Terada TP (2016) Cooperativity and modularity in protein folding. *Biophys Physicobiol* 13: 281-293.

Sato S, Religa TL, Daggett V, Fersht AR (2004) Testing protein-folding simulations by experiment: B domain of protein A. *Proc Natl Acad Sci U S A* 101(18): 6952-6956.

Schmid FX, Baldwin RL (1979) Detection of an early intermediate in the folding of ribonuclease A by protection of amide protons against exchange. *J Mol Biol* 135(1): 199-215.

Schulman BA, Redfield C, Peng ZY, Dobson CM, Kim PS (1995) Different subdomains are most protected from hydrogen exchange in the molten globule and native states of human α-lactalbumin. *J Mol Biol* 253(5): 651-657.

Segel DJ, Bachmann A, Hofrichter J, Hodgson KO, Doniach S, Kiefhaber T (1999) Characterization of transient intermediates in lysozyme folding with time-resolved small-angle X-ray scattering. *J Mol Biol* 288(3): 489-499.

Taketomi H, Ueda Y, Gō N (1975) Studies on protein folding, unfolding and fluctuations by computer simulation. I. The effect of specific amino acid sequence represented by specific inter-unit interactions. *Int J Pept Protein Res* 7(6): 445-459.

Tsui V, Garcia C, Cavagnero S, Siuzdak G, Dyson HJ, Wright PE (1999) Quench-flow experiments combined with mass spectrometry show apomyoglobin folds through and obligatory intermediate. *Protein Sci* 8(1): 45-49.

Van Dael H, Haezebrouck P, Joniau M (2003) Equilibrium and kinetic studies on folding of canine milk lysozyme. *Protein Sci* 12(3): 609-619.

van Gunsteren WF, Burgi R, Peter C, Daura X (2001) The key to solving the protein-folding problem lies in an accurate description of the denatured state. *Angew Chem Int Ed Engl* 40(2): 351-355.

Wako H, Abe H (2016) Characterization of protein folding by a Φ-value calculation with a statistical-mechanical model. *Biophys Physicobiol* 13: 263-279.

Wako H, Saitô N (1978a) Statistical mechanical theory of protein conformation. 1. General considerations and application to homopolymers. *J Phys Soc Jpn* 44(6): 1931-1938.

Wako H, Saitô N (1978b) Statistical mechanical theory of protein conformation. 2. Folding pathway for protein. *J Phys Soc Jpn* 44(6): 1939-1945.

Wildegger G, Kiefhaber T (1997) Three-state model for lysozyme folding: triangular folding mechanism with an energetically trapped intermediate. *J Mol Biol* 270(2): 294-304.

Wright PE, Dyson HJ (1999) Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J Mol Biol* 293(2): 321-331.

Yu W, Chung K, Cheon M, Heo M, Han KH, Ham S, Chang I (2008) Cooperative folding kinetics of BBL protein and peripheral subunit-binding domain homologues. *Proc Natl Acad Sci U S A* 105(7): 2397-2402.

Zamparo M, Pelizzola A (2006) Kinetics of the Wako-Saito-Munoz-Eaton model of protein folding. *Phys Rev Lett* 97(6): 068106.

Zhou Y, Karplus M (1997) Folding thermodynamics of a model three-helix-bundle protein. *Proc Natl Acad Sci U S A* 94(26): 14429-14432.

8. Acknowledgments

本研究の遂行にあたり指導教員である新井宗仁先生には研究の方針や考え方、研究をより深く掘り下げるための知識など、たくさんのご指導をいただきました。また学術的なご指導だけでなく研究のための環境の準備や、研究発表についてのご助言など様々な面から研究を支えていただきました。心より感謝申し上げます。

本論文の審査にあたり、主査の伊藤創祐先生、副査の岡田康志教授、樺島祥介教授、岡田真人教授、竹内一将准教授には本研究について細部にわたり議論していただきました。深く感謝申し上げます。

名古屋大学大学院工学研究科応用物理学専攻の笹井理生教授、寺田智樹准教授には本研究について充実したご助言をいただきました。深く感謝申し上げます。

新井研究室助教 林勇樹先生には研究に対するご助言やご指導をいただき、さらに研究生活について様々な相談にのっていただきました。心より感謝申し上げます。

研究室の後輩の吉村匡隆君には MD シミュレーションの計算プログラム作成を助けていただきました。感謝申し上げます。

新井研究室の皆様には様々な面で研究生活を支えていただきました。皆様のおかげで有意義で充実した研究生活を送ることができました。感謝申し上げます。

本研究は特別研究員奨励費 JP20J11762 の助成を受けたものです。