Structural Significance of Intramolecular Interaction Network in Protein Structure

(分子内相互作用ネットワークのタンパク質構造における構造的意義)

Shohei Konno

今野　翔平

Graduate School of Chemical Sciences and Engineering

Hokkaido University

北海道大学大学院 総合化学院

2019

Acknowledgments

This thesis entitled “Structural Significance of Intramolecular Interaction Network in Protein Structure” was supervised by Professor Koichiro Ishimori (Department of Chemistry, Faculty of Science, Hokkaido University). The work in this thesis has been conducted from April 2013 to March 2019.

First of all, I would like to express my great gratitude to Professor Koichiro Ishimori. He gave me continuous guidance and fruitful discussion for my work and activities. I am also grateful to Dr. Takao Namiki (Department of Mathematics, Faculty of Science, Hokkaido University) his great supports and fruitful discussion. He supported me on network analysis of protein structures from the mathematical computational point of view.

I am also pleased to thank Dr. Takeshi Uchida and Dr. Hiroshi Takeuchi for their criticism and fruitful discussion and Dr. Tomohide Saio for his fruitful discussion and helpful advice for data collection of protein structures for network analysis, Secretary Maki Tanaka for accepting the troublesome office procedure, and the members of Structural Chemistry Laboratory for helps and assistance.

I wish to express sincere appreciation to Dr. Hirotoshi Kuroda (Department of Mathematics, Faculty of Science, Hokkaido University) for his assistance with running mathematical collaboration research in protein science. It should be emphasized that the study was supported by The Ministry of Education, Culture, Sports, Science and Technology through Program for Leading Graduate Schools (Hokkaido University ‘Ambitious Leader’s Program’). The program has financially supported my activities and encouraged me to run mathematical collaboration research in protein science.

At the review of this work, Professor Sadamu Takeda (Faculty of Science,

Hokkaido University), Tetsuya Taketsugu (Faculty of Science, Hokkaido University), Manabu Tokeshi (Faculty of Engineering, Hokkaido University) and Tamiki Komatsuzaki (Research Institute for Electronic Science, Hokkaido University) gave me valuable suggestions and guidance.

Finally, I express my deep appreciation to my family. They always encouraged and supported me.

March, 2019

Graduate School of Chemical Sciences and Engineering, Hokkaido University

Shohei Konno

List of publications

Chapter II

Shohei Konno, Kentaro Doi and Koichiro Ishimori, “Uncovering dehydration in cytochrome *c* refolding from urea- and guanidine hydrochloride-denatured unfolded state by high pressure spectroscopy”, *Biophys. Physicobiol.* **16**, 18-27 (2019).

Chapter III

Shohei Konno, Takao Namiki and Koichiro Ishimori, “Quantitative description and classification of protein structures by a novel robust amino acid network: interaction selective network (ISN)”, *Sci. Rep.*, to be submitted

List of Presentations

Oral Presentations

1. Shohei Konno, Kentaro Doi, Takeshi Uchida and Koichiro Ishimori

   "Titrating of dehydration on Cytochrome *c* folding"

   The 94th CSJ Annual Meeting (Nagoya, Japan) March 27-30, 2014

2. Shohei Konno, Kentaro Doi, Takeshi Uchida and Koichiro Ishimori

   "Dehydration mechanism of cytochrome *c* folding; analysis with hydrophobic effect"

   The CSJ Hokkaido Branch Summer Study Presentation 2014 (Tomakomai, Japan) July 12, 2014

3. Shohei Konno, Kentaro Doi, Takeshi Uchida and Koichiro Ishimori

   "Dehydration in protein folding revealed by high pressure spectroscopy"

   Hokkaido University - University of Strasbourg Joint Workshop by Graduate Students (Sapporo, Japan) March 12, 2015

Poster Presentations

1. Shohei Konno, Kentaro Doi, Takeshi Uchida and Koichiro Ishimori
   "Dehydration in cytochrome *c* folding revealed by high pressure spectroscopy"
   The 52th Annual Meeting of the Biophysical Society of Japan (Sapporo, Japan)
   September 25-27, 2014

2. Shohei Konno, Kentaro Doi, Takeshi Uchida and Koichiro Ishimori
   "Analysis of dehydration in cytochrome *c* folding for using pressure effect"
   The 4th CSJ Chemistry Festa (Tokyo, Japan) October 14-16, 2014

3. Shohei Konno, Kentaro Doi, Takeshi Uchida and Koichiro Ishimori
   "Analysis of dehydration in cytochrome *c* folding by high pressure spectroscopy"
   The 54th High Pressure Conference of Japan (Tokushima, Japan) November 21-24, 2014

4. Shohei Konno, Kentaro Doi, Takeshi Uchida and Koichiro Ishimori
   "Dehydration in protein folding revealed by high pressure spectroscopy"
   The 2nd International Symposium on AMBITIOUS LEADER'S PROGRAM
   (Sapporo, Japan) December 11, 2014

5. Shohei Konno, Kentaro Doi, Takeshi Uchida and Koichiro Ishimori
   "Dehydration in protein folding revealed by high pressure spectroscopy"
   Hokkaido University - Johannes Kepler University Joint Symposium on Chemica Sciences and Engineering (Linz, Austria) February 22, 2017

6. Shohei Konno, Takao Namiki and Koichiro Ishimori

"Complex Network Approach for Characterization of Protein Secondary Structure"

CBI Annual Meeting 2017 (Tokyo, Japan) October 3-5, 2017

Contents

Chapter I
General Introduction

1.1. Protein folding

Proteins are amino acid-based macromolecules that direct or participate in nearly every chemical reaction essential for life. The amino acid sequence of a protein determines its three-dimensional structure, which determines its mechanism of actions. As shown in Figure 1.1, genetic information in DNA is transferred into the amino acid sequence through RNA sequence. To acquire its biological functions, it is necessary to attain its folded three-dimensional (3D) structure. This process is termed "protein folding" which is transformation of one-dimensional linear information encoded in the amino acid sequence into a functional 3D structure. The protein folding is essentially a physical chemistry process for which the mechanism remains elusive [1, 2].

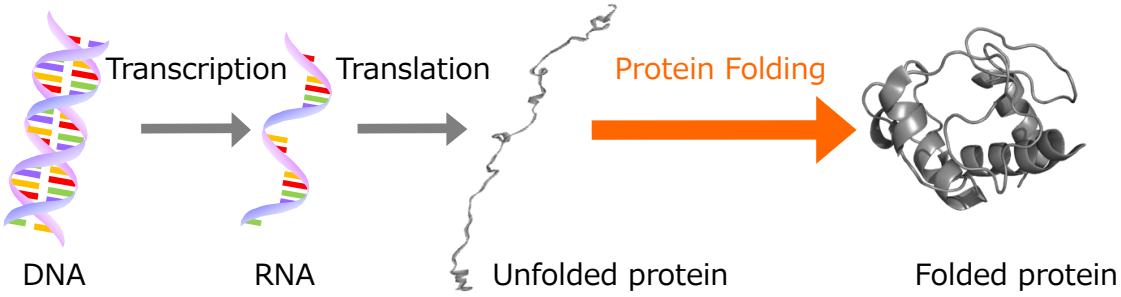

Figure 1. 1 The transfer of information from the DNA to the folded protein.
DNA sequence information is transferred to RNA sequence (Transcription). Subsequently, the sequence information is converted into the amino acid sequence as the polypeptide chain (Translation). To acquire its biological functions, proteins fold into a globular structure. This process is called protein folding.

The folded protein structures should be thermodynamically stable in aqueous solution to retain its folded structure whereas the protein flexibility is also essential for the enzyme activity [3-6]. Most globular proteins are marginally stable, with the Gibbs energy changes in the protein folding of about −40 kJ mol^{-1} [6, 7]. It has been suggested

that the marginal stability represents an adaptation for increased functionality, as the marginal stability would be correlated with increased protein flexibility [6]. The marginal stability results from the summation of numerous interactions composed of the protein folded structures between their amino acid residues and, moreover, with water molecules in a solvent. Due to the marginal stability, losing several favorable interactions induces destabilization of the folded structures. While the marginal stability of the protein structure is assumed to be the result of intramolecular interaction network in the protein, it has been unclear how the intramolecular interaction network forms what type of protein structures.

Due to the lack of our understanding of the intramolecular interaction network of a protein, the site-directed mutation of one amino acid residue of a protein, especially in the protein interior, sometimes results in failure to attain its folded structure. Even if attractive interactions are introduced in the protein interior by site-directed mutagenesis, non-adaptive substitution results in the destabilization of the global conformation of the folded structure and sometimes induces the loss of the folded structure [8]. Many efforts were taken to stabilize proteins by further optimizing the protein interior, but this turned out to be very difficult [9-11].

The protein interior is tightly packed and enriched in hydrophobic residues that are expelled from water. This is the result of the hydrophobic interaction which entails minimizing the surface area of hydrophobic amino acid residues exposed to the aqueous solvent [12-14]. However, it is difficult to capture the comprehensive description of the hydrophobic interaction in the protein folding. The contribution of the hydrophobic interaction to the protein stability per a methylene group are dependent on the proteins and was reported 2 to 7 kJ mol^{-1} [15].

1.2. Dehydration from hydrophobic groups in the protein folding (Chapter II)

The formation of the hydrophobic interaction in the protein folding is closely related to the dehydration which is the evacuation of water molecules from buried protein surface accompanied by the formation of the protein folded structures. Figure 1.2 displays the dehydration in the protein folding. The Gibbs energy change induced by the hydrophobic interaction in the protein folding is assumed to be proportional to the total solvent-inaccessible surface area of the hydrophobic groups in a protein [16, 17]. Therefore, the surface area involved in the dehydration in the protein folding expects to reflect the Gibbs energy change by the hydrophobic interaction. While it is difficult to directly examine the surface area involved in the dehydration, the partial molar volume (V) change for the protein folding reflects the overall volume change associated with the hydration/dehydration of the solvent [18-22]. This encourage me to detect the dehydration in the protein folding as the volume change.

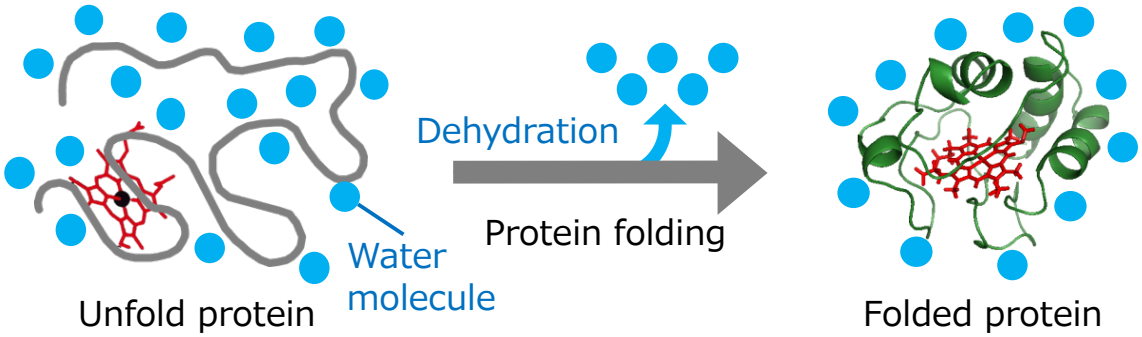

Figure 1. 2 The Dehydration in the protein folding.
The protein folding in aqueous solution accompanied by the release of water molecules from the hydrophobic surface of the protein. Polypeptide chains in unfolded protein interact with water molecules in a solvent. Formation of the protein structure results in the loss of the hydration of buried atoms in the protein. This process is termed dehydration.

Chapter II focused on the dehydration accompanied by the formation of the hydrophobic interaction to acquire the detailed description of the hydrophobic interaction. The partial molar volume change for the protein unfolding in cytochrome *c* (Cyt *c*) were experimentally determined by using the high-pressure absorption spectroscopy to estimate the hydration associated with the protein unfolding. The amount of the hydration to hydrophobic groups in the protein unfolding of Cyt *c* is calculated to discuss the thermodynamic significance of the dehydration in the Cyt *c* folding. Based on the analysis in Chapter II, the entropic contribution of the dehydration from the heme group in the Cyt *c* folding would be approximately -50 kJ mol^{-1}. This value is more negative than the total Gibbs energy change of the Cyt *c* folding, ΔG_{UN} (-37 kJ mol^{-1}) [17]. The results imply that the entropic contribution of the dehydration from the hydrophobic groups is critical for stabilizing protein structures.

1.3. Quantitative description of protein structures (Chapter III)

Although the Chapter II focus on dehydration accompanied by the formation of the hydrophobic interaction in the protein folding, it is also necessary to examine the relationship between the individual interaction and the global conformation of a protein. An emerging approach for quantitatively characterizing complicated structures is the complex network which is a mathematical model describing complex structures as 'vertices' and 'links', enabling the quantitative characterization of the geometry using the network parameters. A network structure is represented as an adjacency matrix. If there are N vertices in a network, its adjacency matrix, a_{ij}, is a $N \times N$ square matrix. If a link is established between ith and jth vertex, a_{ij}=1 otherwise 0. This approach has a possibility for representing the protein structures with less computationally intensive than the traditional molecular dynamic simulation analysis [23]. An approach of the network representation of the protein structures is amino acid network (AAN) [23-25]. As shown in Figure 1.3, an AAN is constructed based on the structural information of experimentally determined crystal structure of a protein. The 'vertices' correspond to each amino acid residue in the protein structures, while one amino acid residue composed of multiple atoms. The 'links' represent van der Waals contacts and/or chemical interactions between two amino acid residues. However, many previous AAN studies have been applied to elucidate the relationship between protein domains in allosteric regulation, because the network parameters reflect the difference of global, rather than local, conformation [23, 26]. Then, the application of the AAN for evaluating the relationship between the intermolecular interaction network and the global conformation of a protein is not explored.

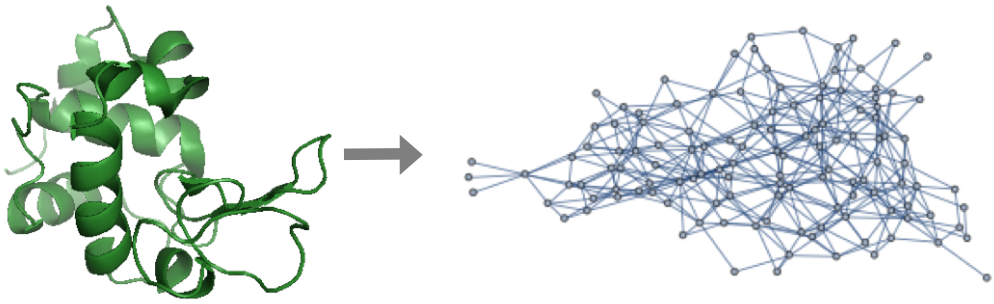

Figure 1. 3 The construction of an amino acid network (AAN) based on atomic coordinate of a protein structure.
An AAN of a protein is composed of 'vertices' and 'links'. The vertices correspond to one amino acid residues in the protein structures, while the links represent van der Waals contacts and/or chemical interactions between two amino acid residues. The links are established if two residues contact each other, which is determined by the atomic coordinates of corresponding protein structure deposited in the Protein Data Bank database.

In Chapter III, a new AAN with reflecting intramolecular interactions between the amino acid residues is introduced to establish a model to investigate the relationship between the intermolecular interaction network and the global conformation of a protein. From the viewpoint of characterization of the protein structures, previous studies have examined network properties whereas they focused on the comparison with the mathematical network models. It remains to be difficult to distinguish between protein secondary structures [27]. Hence, I developed a novel amino acid network—the interaction selective network (ISN)—that includes structural information about interactions in main and side chain atoms. The ISN allows to distinguish between α-helix and β-sheet protein structures by using network parameters. Although a conventional Cα (α-carbon) network (CAN) with a cutoff distance of 5.5 Å allowed us

to discriminate between the two secondary structures, the slight shifts of the cutoff distance led to a loss of discrimination, showing lower robustness of the model. This is because the CAN detects the interactions only between the main chain atoms. Conversely, the ISN shows higher robustness and the links in the ISN are based on interactions in *both* the main and side chains. The ISN reflects structural information of *both* secondary and tertiary structures. Therefore, the ISN is expected to be an effective model for evaluating the relationship between the intermolecular interaction network and the global conformation of a protein.

1.4. References

[1] Dill, K.A., Ozkan, S.B., Weikl, T.R., Chodera, J.D. & Voelz, V.A. The protein folding problem: when will it be solved? *Curr. Opin. Struct. Biol.* **17**, 342-346 (2007).

[2] Liu, S.-Q., Ji, X.-l., Tao, Y., Tan, D.-y., Zhang, K.-q. & Fu, Y.-X. 10 Protein Folding , Binding and Energy Landscape : A Synthesisn. in *Protein Eng.*), InTech, (2012).

[3] Rasmussen, B.F., Stock, A.M., Ringe, D. & Petsko, G.A. Crystalline ribonuclease A loses function below the dynamical transition at 220 K. *Nature* **357**, 423-424 (1992).

[4] Zavodszky, P., Kardos, J., Svingor, A. & Petsko, G.A. Adjustment of conformational flexibility is a key event in the thermal adaptation of proteins. *Proc. Natl. Acad. Sci. USA* **95**, 7406-7411 (1998).

[5] Tsou, C.-L. Active Site Flexibility in Enzyme Catalysis. *Ann. N. Y. Acad. Sci.* **864**, 1-8 (2006).

[6] Taverna, D.M. & Goldstein, R.A. Why are proteins marginally stable? *Proteins* **46**, 105-109 (2002).

[7] Privalov, P.L. & Khechinashvili, N.N. A thermodynamic approach to the problem of stabilization of globular protein structure: A calorimetric study. *J. Mol. Biol.* **86**, 665-684 (1974).

[8] Lim, W.A. & Sauer, R.T. Alternative packing arrangements in the hydrophobic core of lambda repressor. *Nature* **339**, 31-36 (1989).

[9] Schmid, F.X. Lessons about Protein Stability from in vitro Selections. *Chembiochem* **12**, 1501-1507 (2011).

[10] Finucane, M.D., Tuna, M., Lees, J.H. & Woolfson, D.N. Core-directed protein design. I. An experimental method for selecting stable proteins from combinatorial libraries. *Biochemistry* **38**, 11604-11612 (1999).

[11] Finucane, M.D. & Woolfson, D.N. Core-directed protein design. II. Rescue of a multiply mutated and destabilized variant of ubiquitin. *Biochemistry* **38**, 11613-11623 (1999).

[12] Kauzmann, W. Some Factors in the Interpretation of Protein Denaturation1. in *Adv. Protein Chem.* (C.B. Anfinsen, M.L.A.K.B. & John, T.E. eds.) vol. 14, pp. 1-63 Academic Press, (1959).

[13] England, J.L. & Haran, G. Role of Solvation Effects in Protein Denaturation: From Thermodynamics to Single Molecules and Back. in *Annu. Rev. Phys. Chem.* (Leone, S.R., Cremer, P.S., Groves, J.T., & Johnson, M.A. eds.) vol. 62, pp. 257-

277, (2011).

[14] Nozaki, Y. & Tanford, C. The solubility of amino acids and two glycine peptides in aqueous ethanol and dioxane solutions. Establishment of a hydrophobicity scale. *J. Biol. Chem.* **246**, 2211-2217 (1971).

[15] Pace, C.N., Fu, H.L., Fryar, K.L., Landua, J., Trevino, S.R., Shirley, B.A.*, et al.* Contribution of Hydrophobic Interactions to Protein Stability. *J. Mol. Biol.* **408**, 514-528 (2011).

[16] Makhatadze, G.I. & Privalov, P.L. Contribution of Hydration to Protein Folding Thermodynamics: I. The Enthalpy of Hydration. *J. Mol. Biol.* **232**, 639-659 (1993).

[17] Privalov, P.L. & Makhatadze, G.I. Contribution of Hydration to Protein Folding Thermodynamics: II. The Entropy and Gibbs Energy of Hydration. *J. Mol. Biol.* **232**, 660-679 (1993).

[18] Chalikian, T.V. & Filfil, R. How large are the volume changes accompanying protein transitions and binding? *Biophys. Chem.* **104**, 489-499 (2003).

[19] Chalikian, T.V. & Breslauer, K.J. On volume changes accompanying conformational transitions of biopolymers. *Biopolymers* **39**, 619-626 (1996).

[20] Silva, J.L., Foguel, D. & Royer, C.A. Pressure provides new insights into protein folding, dynamics and structure. *Trends Biochem. Sci.* **26**, 612-618 (2001).

[21] Royer, C.A. Revisiting volume changes in pressure-induced protein unfolding. *Biochim. Biophys. Acta Protein Struct. Molec. Enzy.* **1595**, 201-209 (2002).

[22] Roche, J. & Royer, C.A. Lessons from pressure denaturation of proteins. *J. Royal Soc. Interface* **15**, (2018).

[23] Di Paola, L., De Ruvo, M., Paci, P., Santoni, D. & Giuliani, A. Protein Contact Networks: An Emerging Paradigm in Chemistry. *Chem. Rev.* **113**, 1598-1613 (2013).

[24] Yan, W., Zhou, J., Sun, M., Chen, J., Hu, G. & Shen, B. The construction of an amino acid network for understanding protein structure and function. *Amino Acids* **46**, 1419-1439 (2014).

[25] Greene, L.H. Protein structure networks. *Brief. Funct. Genomics* **11**, 469-478 (2012).

[26] Di Paola, L. & Giuliani, A. Protein contact network topology: a natural language for allostery. *Curr. Opin. Struct. Biol.* **31**, 43-48 (2015).

[27] Alves, N.A. & Martinez, A.S. Inferring topological features of proteins from amino acid residue networks. *Physica A* **375**, 336-344 (2007).

Chapter II

Uncovering dehydration in cytochrome *c* refolding from urea- and guanidine hydrochloride-denatured unfolded state by high pressure spectroscopy

2.1. Abstract

To investigate the dehydration associated with protein folding, the partial molar volume changes for protein unfolding (ΔV_u) in cytochrome *c* (Cyt *c*) were determined using high pressure absorption spectroscopy. ΔV_u values for the unfolding of urea- and guanidine hydrochloride (GdnHCl)-denatured Cyt *c* were estimated to be 56 ± 5 and 29 ± 1 mL mol^{-1}, respectively. Considering that the volume change for hydration of hydrophobic groups is positive and that Cyt *c* has a covalently bonded heme, a positive ΔV_u reflects the primary contribution of the hydration of heme. Because of the marked tendency of guanidium ions to interact with hydrophobic groups, a smaller number of water molecules were hydrated with hydrophobic groups in GdnHCl-denatured Cyt *c* than in urea-denatured Cyt *c*, resulting in the smaller positive ΔV_u. On the other hand, urea is a relatively weak denaturant and urea-denatured Cyt *c* is not completely hydrated, which retains the partially folded structures. To unfold such partial structures, I introduced a mutation near the heme binding site, His26, to Gln, resulting in a negatively shifted ΔV_u (4 ± 2 mL mol^{-1}) in urea-denatured Cyt *c*. The formation of the more solvated and less structured state in the urea-denatured mutant enhanced hydration to the hydrophilic groups in the unfolding process. Therefore, I confirmed the hydration of amino acid residues in the protein unfolding of Cyt *c* by estimating ΔV_u, which allows to discuss the hydrated structures in the denatured states of proteins.

2.2. Introduction

In aqueous solutions, linear protein polypeptide chains decrease in entropy and collapse into a globule to minimize the surface area that is exposed to the solvent. The folded state is a low-entropy subensemble in all possible collapsed globular conformations for the protein chain. In contrast, the unfolded state is an ensemble that is not or much less structured and has higher entropy than the folded state does. For folding to be beneficial, the folded state must be sufficiently and energetically favorable to overwhelm the higher entropy associated with structural disorder within the globular phase. The major energetic driving force originates from the van der Waals, hydrogen bonding, and electrostatic interactions, both within the polypeptide chain and between the chain and its surrounding aqueous solvent. One particularly important subset of these interactions is the hydrophobic effect, which entails minimizing the surface area of hydrophobic amino acid residues exposed to the aqueous solvent [1, 2].

One problem related to estimating the hydrophobic effect on protein folding is the lack of understanding regarding the evaluation of the hydration effect. Makhatadze and Privalov [3, 4] have estimated the hydration effects of cytochrome *c* (Cyt *c*) unfolding using structural information on the hydrophilic and hydrophobic groups, as shown in Table 2.1, that have been exposed to water in the folded and unfolded states. They proposed that a large entropy decrease in conformational entropy will be compensated for by the evacuation of water molecules, known as dehydration, from the hydrophobic surface. However, experimental data examining the structural and functional significance of dehydration remain limited because of the difficulties in the spectroscopic characterization of water molecules. The present study focuses on the dehydration from

amino acid residues in the protein folding of Cyt *c* to determine the thermodynamic significance of the dehydration and discuss hydrated structures in the unfolded states.

Table 2.1 Thermodynamical energy changes in cytochrome *c* (Cyt *c*) unfolding. The Gibbs energy changes (ΔG), the enthalpy changes (ΔH) and the changes of the entropy term ($-T\Delta S$) in the Cyt *c* unfolding are estimated using the amounts of surface area of the hydrophilic and hydrophobic groups of Cyt *c* that have been exposed to water in the folded and unfolded states [3, 4].

Energy changes induced by	ΔG / kJ mol^{-1}	ΔH / kJ mol^{-1}	$-T\Delta S$ / kJ mol^{-1}
Overall	37	89	−52
Dehydration	−5003	−7409	2406
Dehydration from hydrophilic sites	181	−1029	1209
Dehydration from hydrophobic sites	5040	7498	−2458

A promising approach in observing hydration/dehydration is estimating the partial molar volume of a protein molecule in its aqueous solution (V), including the hydration of solvent-accessible protein atomic groups [2, 5-8]. As illustrated in Figure 2.1, V can be represented by the sum of the constitutive volume (V_m) [9, 10], the volume of structural voids (V_v) corresponding to the solvent-inaccessible core of the protein resulting from imperfect atomic packing [11], and the overall volume change associated with the hydration/dehydration of the solvent (ΔV_h) [12-16] as follows:

$$V = V_m + V_V + \Delta V_h. \qquad \text{(Eq. 1)}$$

The partial molar volume change in protein unfolding (ΔV_u) is therefore defined by the following equation:

$$\Delta V_u = \Delta V_m + \Delta V_V + \Delta\Delta V_h \qquad \text{(Eq. 2)}$$

where ΔV_m is the volume changes of V_m between the folded and unfolded states, ΔV_v is the changes of V_v, and $\Delta\Delta V_h$ is the observed hydration term changes corresponding to the volume changes caused by the hydration of polypeptide chains in the protein unfolding [2, 5]. Figure 2.2 displays Eq. 2 in detail. In the above equation, V_m does not change from the folded state to the unfolded state; this leads to the assumption that ΔV_m is zero. Because denatured polypeptide chains are exposed to the solvent, V_v for the unfolded states can be accurately approximated using zero, assuming that ΔV_v is equal to $-V_v$. Therefore, Eq. 2 can be simplified to Eq. 3.

$$\Delta V_u = -V_V + \Delta\Delta V_h \qquad \text{(Eq. 3)}$$

$\Delta\Delta V_h$ can be determined by the sum of V_v and ΔV_u. For the folded state, V_v can be determined from the protein crystal structures by rolling a virtual ball of water molecule around the van der Waals surface of the macromolecule [17]. If the system is at constant temperature T, ΔV_u can be expressed using the following thermodynamic relationship:

$$\Delta V_u = \left[\frac{\partial \Delta G_u}{\partial p}\right]_T = -RT\left[\frac{\partial \ln K}{\partial p}\right]_T \qquad \text{(Eq. 4)}$$

where ΔG_u is the Gibbs energy change in the protein unfolding, p is pressure, and K is the equilibrium constant between the folded and unfolded states:

$$K = \frac{[\text{Unfolded Cyt } c]}{[\text{Folded Cyt } c]} \qquad \text{(Eq. 5)}$$

Applying the two-state folded-to-unfolded transition model to the protein unfolding process [18], I determined K using high pressure ultraviolet/visible (UV/vis) spectroscopy and measured the pressure dependence of K to calculate ΔV_u and $\Delta\Delta V_h$.

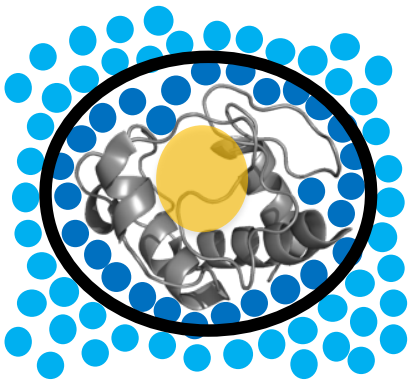

Figure 2.1 Partial molar volume of a protein in aqueous solution (V). V can be represented by the sum of the constitutive volume (V_m, gray polypeptide chain), the volume of structural voids (V_v, semitransparent orange circle) corresponding to the solvent-inaccessible core of the protein resulting from imperfect atomic packing and the overall volume change associated with the hydration/dehydration of the solvent (ΔV_h, deep blue filled cycles). Little blue filled cycles represent bulk water molecules.

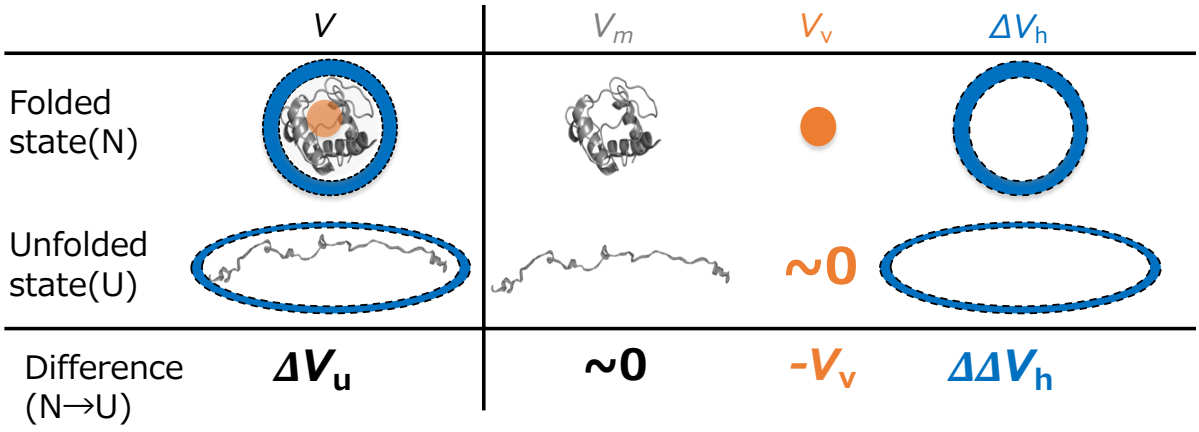

Figure 2.2 The partial molar volume changes from folded (N) to unfold state (U) (ΔV_u). As shown in Figure 2.1, the partial molar volume of a protein in aqueous solution (V) can be represented three terms as follows; the constitutive volume (V_m), the volume of structural voids (V_v) and the overall volume change associated with the dehydration of the solvent (ΔV_h). Here, the volume changes of V_m between N and U is assumed to be zero. V_v for the unfolded states can be accurately approximated using zero and assuming that the volume changes of V_v from N to U is equal to $-V_v$ for the folded state. $\Delta\Delta V_h$ is the observed hydration term changes corresponding to the volume changes caused by the hydration to the polypeptide chains in the protein unfolding.

To determine *K* between the folded and unfolded states, I used guanidine hydrochloride (GdnHCl) and urea as the chemical denaturants [19, 20]. GdnHCl is one of the commonly used denaturants that directly and strongly interacts with hydrophobic groups [21], and interactions between GdnHCl and hydrophobic amino acid residues facilitate the dehydration from hydrophobic amino acid residues, resulting in fewer hydrated *hydrophobic* groups in the unfolded states. In contrast, urea is a rather weak denaturant that interacts with hydrophilic amino acid residues to destabilize protein folded structures by forming hydrogen bonds to the peptide groups [21, 22]. Therefore, in the urea-denatured state, some of the hydrophilic amino acid residues interact with urea, not water molecules, leading to fewer hydrated *hydrophilic* groups. Therefore, although both urea and GdnHCl are typical denaturants, the hydrated structures in the denatured states are significantly different, and the present study characterizes urea- and GdnHCl-denatured unfolding of Cyt *c* to describe the hydrated structure of proteins. Together with the results showing that mutant Cyt *c* has perturbed hydrated structures in the denatured states, new insights into the dehydration in theprotein folding and the unfolded protein structures are provided.

2.3. Methods

Protein

To examine the hydration associated with protein unfolding, horse Cyt *c* was chosen for analysis, because the equilibrium constant between folded and unfolded states (*K*) could be monitored using the absorbance band of the covalently attached heme as the prosthetic group (shown in red in Figure 2.3) [3, 4]. In the protein unfolding of horse Cyt *c* at neutral pH, His18 is ligated to the oxidized heme iron in the unfolded state; however, the second axial ligand Met80 is replaced by a nonnative histidine ligand [23-27]. To estimate ΔV_u in the Cyt *c* unfolding, I monitored the replacement of the axial ligand using UV/vis spectroscopy under high pressure. The spectra of folded and urea-denatured unfolded Cyt *c* are shown in Figure 2.4. The Soret band around 400 nm is monitored to determine *K*.

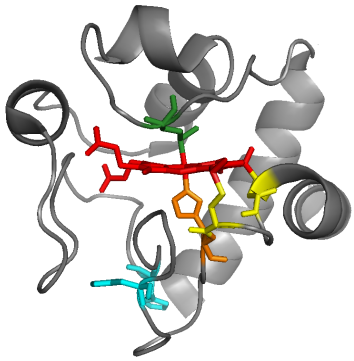

Figure 2.3 Ribbon diagram of horse Cyt *c* based on the crystal structure 1HRC. The heme in Cyt *c* (red) forms covalent bonds with Cys14 and Cys 17 (yellow). The side chains of the native heme ligands, Met80 (green) and His18 (orange), and potential nonnative ligands, His26 and His33 (cyan), are shown explicitly. Figure has been prepared using PyMol.

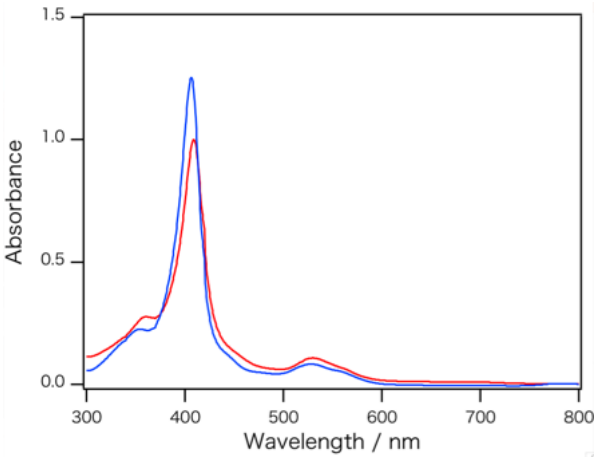

Figure 2.4 Absorption spectra of Cyt *c* in 0 M (red) and 9.2 M urea (blue). Red spectrum corresponds to folded state and blue spectrum exhibits urea-denatured unfolded state. The Soret band around 400 nm is monitored to determine *K*.

Sample preparation

To oxidize the residual ferrous form, folded horse heart Cyt *c* (Merck Millipore, Darmstadt, Germany) was treated with potassium ferricyanide and was purified on an Amicon ultrafiltration using 5-kDa cutoff membranes to remove excess oxidants. The expression vector for the Cyt *c* mutant containing Glu at the His26 position was constructed as shown in following sections. The protein was dissolved in 50 mM Tris-HCl buffer (pH 7.5) at various activities of urea (molecular biology grade from Kanto Chemical Co., Inc., Tokyo, Japan) or GdnHCl (biochemistry grade from Kanto Chemical Co.) before the experiments. The final protein concentration was 30 μM, which was spectrophotometrically estimated using molar extinction coefficient for oxidized Cyt *c*. The extinction coefficient used at 408 nm was 1.06×10^{-1} m^2 mol^{-1} [28].

Mutagenesis

Mutagenesis was conducted utilizing the PrimeSTAR mutagenesis basal kit from Takara Bio (Otsu, Japan). DNA oligonucleotides were purchased from Operon Biotechnologies (Tokyo, Japan). The mutated genes were sequenced (Operon Biotechnologies, Tokyo, Japan) to ensure that only the desired mutations were introduced (Eurofins Genomics) [29].

Protein expression and purification

The H26Q mutant of horse heart Cyt *c* was expressed in *Escherichia coli* and purified as well as previously described [29, 30]. The *Escherichia coli* strain Rosetta2(DE3)pLysS cells transformed with the plasmids containing the DNA of Cyt *c* were inoculated in 5 ml of LB medium and grown at 37 °C over 12 hours. This pre-cultured medium was added to 4 liters of LB medium, and the bacteria were further incubated at 37 °C. The expression of horse Cyt *c* was initiated by adding 0.8 mm isopropyl 1-thio-β-d-galactopyranoside to the culture when the cell density reached an absorbance of 0.6 at 600 nm. Then, 0.1 mm δ-aminolevulinic acid was added to promote heme biosynthesis. After incubation for an additional 24 hours, the cells were collected by centrifugation.

The cell pellet was resuspended in 50 mm Tris-HCl, pH 7.5, containing 1 g/liter lysozyme, 50 mg/liter DNase I, and 50 mg/liter RNase A and suspended for 3 hours to lyse the cell pellet completely. The supernatant of the crude extract was obtained by centrifugation at 18,000 rpm for 5 min and 40,000 rpm for 1 hour. This supernatant was purified by HiPrep 16/10 SP XL column (GE Healthcare, Uppsala, Sweden) with a linear salt gradient of 1–300 mm NaCl. The elution sample was

concentrated by Amicon ultrafiltration using 5-kDa cutoff membranes. Because the concentrated Cyt *c* was a mixture of ferric and ferrous forms, Cyt *c* was oxidized by 10-fold potassium ferricyanide with stirring for 1 hour to completely oxidize Cyt *c*. After being dissolved in 50 mm sodium phosphate buffer, pH 7.0, Cyt *c* was further purified by Mono S 10/100 GL column (GE Healthcare) with a linear salt gradient. The purified Cyt *c* fractions were pooled, concentrated, and applied to a HiLoad 16/60 Superdex 75 gel filtration column (GE Healthcare).

Measurement

A high pressure spectroscopy device is developed as illustrated in Figure 2.5. The UV/vis spectra of Cyt *c* from 800 to 250 nm were recorded using a JASCO model V-570 spectrophotometer. The measurements were performed using a cell connected to a circulating water bath to maintain the sample temperature at 25°C. A quartz cell filled with the sample was placed into the high pressure optical cell (PCI-500; Syn Corporation Ltd., Kyotanabe, Japan), and pressure was applied using the hand pump HP-501 (Syn Corporation Ltd.) through the water medium. The measurements were performed from 50 to 150 MPa at 25-MPa intervals.

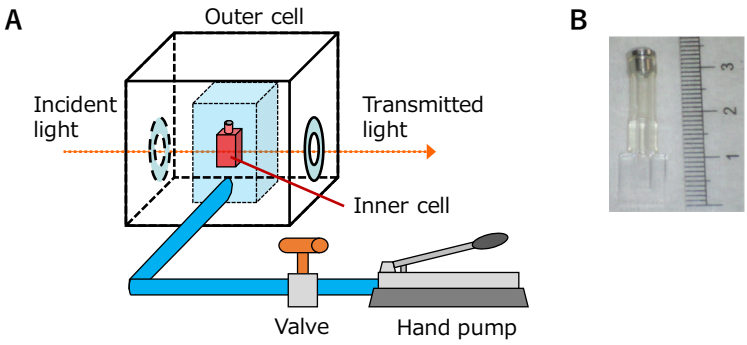

Figure2.5 (A) Schematic drawing of the high pressure spectroscopy device. Using a hand pump, pressure is applied via water to the inner cell filled with the sample. The valve is closed, following which the measurements were conducted. (B) Inner cell for the measurements of high pressure absorption spectra. This cell comprises two connected components: a quartz cell and a rubber tube that exerts the pressure on the sample.

Averaged hydrophobicity

To compare the hydrophobicity of proteins, the averaged hydrophobicity of the amino acid sequence of the protein are calculated. This included the summation of the hydrophobicity of amino acid residues in the sequence of the corresponding protein without including any prosthetic groups and exogenous ligands [31].

2.4. Results and Discussion

Determination of volume changes in the Cyt *c* unfolding

To estimate ΔV_u for unfolding to the urea- or GdnHCl-denatured states, I monitored the absorbance change at 402.5 and 404.5 nm by adding urea and GdnHCl under high pressures, respectively (50–150 MPa). In the Cyt *c* unfolding by the chemical denaturants [32, 33], intermediates with the His-Lys heme ligation are reported to be stabilized with increased pressure [34]. In my measurement, however, the Soret band was shifted from 409 nm to 407 nm by addition of the denaturant as shown in Figure 2.6, without appearance of a peak or shoulder from the His-Lys heme ligation around 405 nm [35-37]. Although I cannot exclude the possibility of the formation of the folding intermediates under my denaturation conditions, the contribution of the folding intermediates to the changes of the absorption spectra is rather small and it is more plausible that the transition I detected from the absorption spectra is from the folded state to the unfolded state. Figures 2.7 and 2.8 show experimental data analyses for the urea and GdnHCl denaturation, respectively. To determine the volume changes induced by the denaturants [38, 39], activities, not molarities, were used as the concentration of the denaturant [40-42]. As shown in Figures 2.7A and 2.8A, absorbance intensities of the Soret peak under various denaturant activities and pressures were fitted using the two-state model for each pressure. In, Figures 2.7B and 2.8B fitted absorbance spectra were normalized by the ratio of the folded states, to determine K [2, 5]. For each pressure, the absorbance was normalized by the denaturant activity. The absorbances without the denaturant and with the highest activity of the denaturant were adjusted to 1 and 0, respectively. Using Eq. 4, ΔV_u could be estimated based on the pressure dependence of K (Figures 2.7C and 2.8C). To obtain ΔV_u for the urea- and GdnHCl-denatured unfolding,

the liner extrapolation method was used [38, 39]. The dependence of ΔV_u on denaturant activity is illustrated in Figures 2.7D and 2.8D; ΔV_u values were 56 ± 5 and 29 ± 1 mL mol^{-1} for the unfolding processes of the urea and GdnHCl denaturation, respectively (Table 2.4).

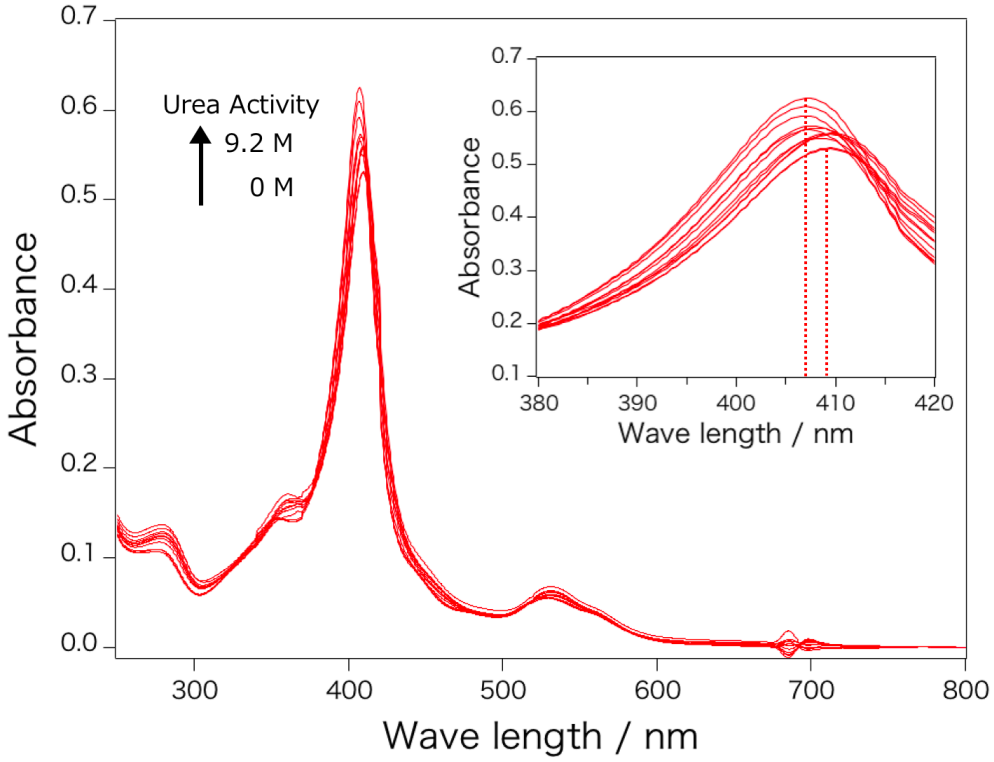

Figure 2.6 Absorption spectra of cytochrome *c* (Cyt *c*) at 50 MPa in the presence of various urea activities. (*Inset*) Enlarged spectra around the Soret band. As increasing urea activities, the absorbance maximum in the Soret band shifts from 409 nm (folded state) to 407 nm (unfolded state) [35].

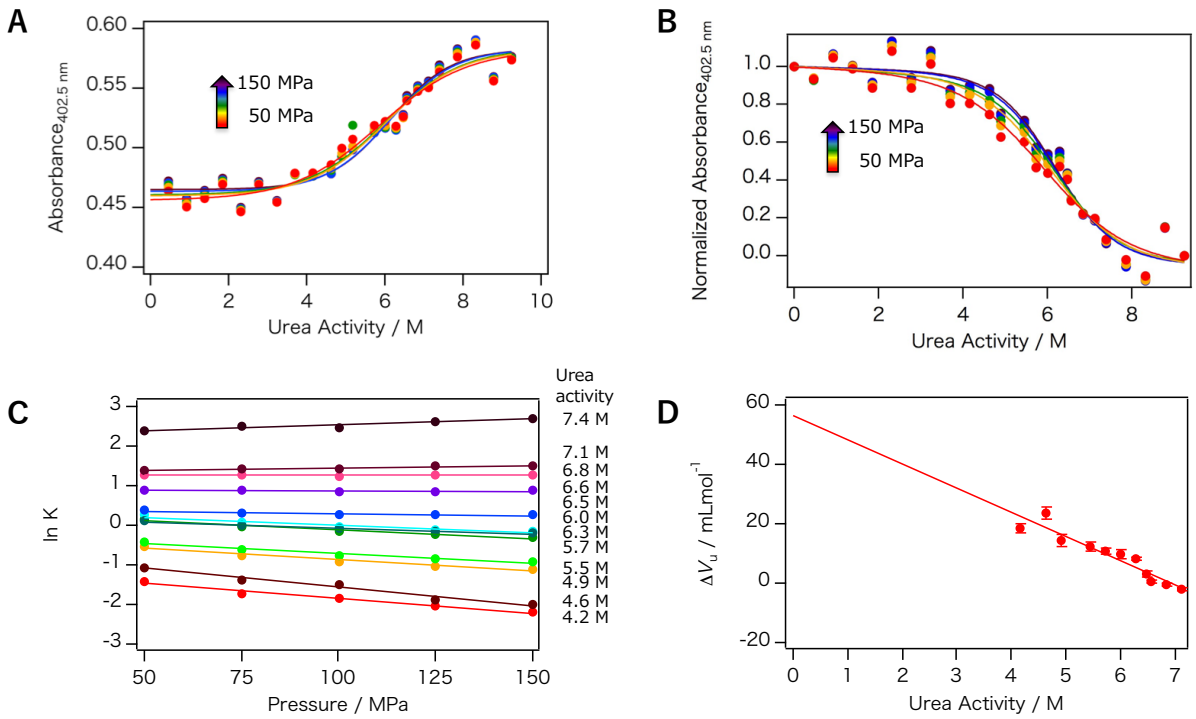

Figure 2.7 (A) Denaturation curves of wild-type Cyt *c* induced by urea. The curves are fitted using a two-state transition model by each pressure. (B) The denaturation curves normalized by the ratio of folded Cyt *c* molecules. (C) Pressure dependence of the equilibrium constant for the Cyt *c* folding on urea activities. ΔV_u for each urea activity is determined by calculating the slopes of the pressure dependence of the equilibrium constant. (D) ΔV_u for urea-denatured unfolding of Cyt *c*. ΔV_u is determined by the extrapolation to 0 M urea activity.

Table 2.2 ΔV_u for urea-induced unfolding of Cyt *c*. ΔV_u values were calculated using Eq. 4 and the slope of the corresponding activity in the ln *K*−Pressure plot in Figure 2.7C.

Urea activity, M	ΔV_u , mL mol^{-1}	Standard deviation of ΔV_u , mL mol^{-1}
4.2	19	2
4.6	24	2
4.9	14	2
5.5	12	2
5.7	11	1
6.0	10	1
6.3	8.0	0.7
6.5	3.0	0.9
6.6	0.36	0.64
6.8	−0.53	0.47
7.1	−2.2	0.7

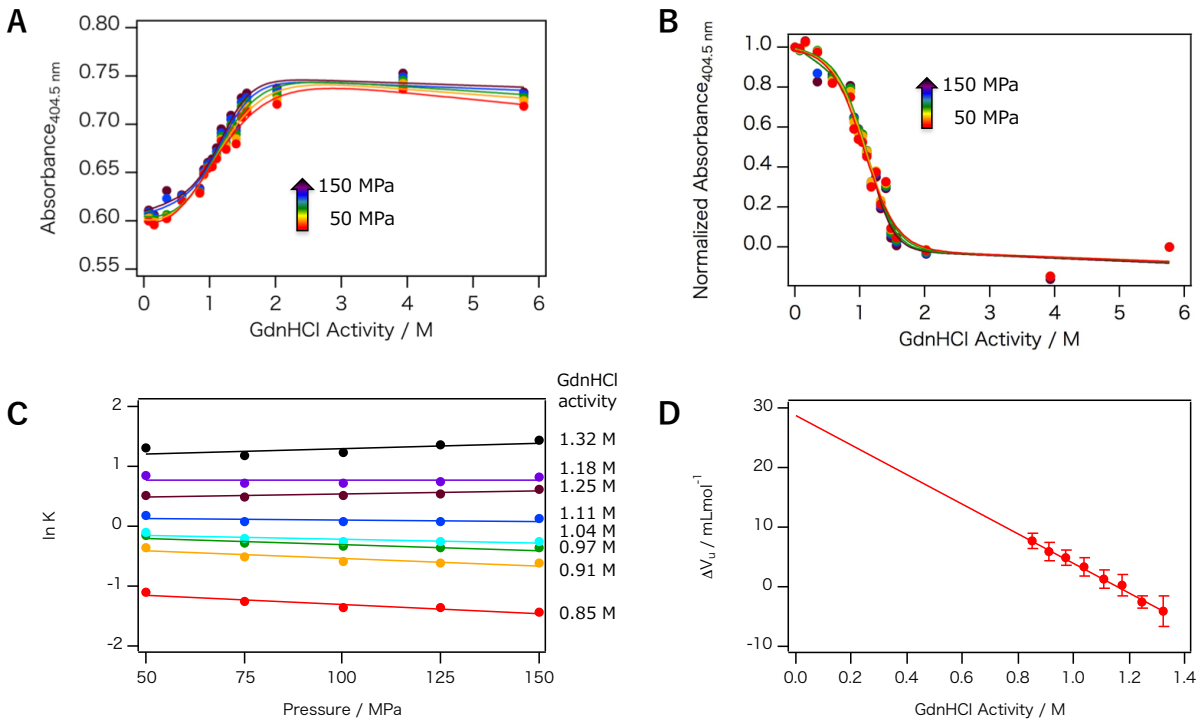

Figure 2.8 (A) Denaturation curves of wild-type Cyt *c* induced by guanidine hydrochloride (GdnHCl). The curves are fitted using a two-state transition model by each pressure (B) The denaturation curves normalized by the ratio of folded Cyt *c* molecules. (C) Pressure dependence of the equilibrium constant for the Cyt *c* folding on GdnHCl activities. ΔV_u for each GdnHCl activity is determined by calculating the slopes of the pressure dependence of the equilibrium constant. (D) ΔV_u for GdnHCl-denatured unfolding of Cyt *c*. ΔV_u is determined by the extrapolation to 0 M GdnHCl activity.

Table 2.3 ΔV_u for guanidine hydrochloride (GdnHCl)-induced unfolding of Cyt *c*. ΔV_u values were calculated using Eq. 4 and the slope of the corresponding activity in the ln *K*−Pressure plot in Figure 2.8C.

GdnHCl activity, M	ΔV_u , mL mol^{-1}	Standard deviation of ΔV_u , mL mol^{-1}
0.85	7.6	1.4
0.91	6.0	1.5
0.97	4.9	1.4
1.0	3.3	1.5
1.1	1.3	1.5
1.2	0.34	1.82
1.2	−2.6	1.1
1.3	−4.2	2.4

Table 2.4 Partial molar volume changes (ΔV_u) for denaturant-induced unfolding of Cyt *c*.

	ΔV_u, mL mol^{-1}	
	GdnHCl	Urea
Wild-type	29 ± 1	56 ± 5
H26Q mutant	20 ± 1	4 ± 2

On the basis of Eq. 3, the volume changes for the hydration ($\Delta\Delta V_h$) for urea and GdnHCl denaturation were estimated by the sum of ΔV_u and V_v. Using the structural data of Cyt *c*, 1HRC, [43] and the program for calculating the void volume in protein structure (Voss Volume Voxelator web server), V_v was determined to be approximately 80 mL mol^{-1} [17]. Thereafter, $\Delta\Delta V_h$ values for urea and GdnHCl denaturation were calculated as 140 and 110 mL mol^{-1}, respectively (Table 2.5).

Table 2.5 Partial molar volume changes for hydration ($\Delta\Delta V_h$) in denaturant-induced unfolding of Cyt *c*.

	$\Delta\Delta V_h$, mL mol^{-1}	
	GdnHCl	Urea
Wild-type	110	140
H26Q mutant	100	80

Structural significance of positive volume change in the Cyt *c* unfolding

One of the characteristic features in the protein unfolding of Cyt *c* is the *positive* ΔV_u values for both urea and GdnHCl denaturation. Although positive ΔV_u has reported in the thermal unfolding of Cyt *c* at acidic pH [44, 45], proteins showing a positive ΔV_u value at neutral pH have not yet been reported [6, 46-49]. As indicated in the previous section, the positive ΔV_u for the Cyt *c* unfolding originates from the larger positive $\Delta\Delta V_h$, compared to the smaller positive V_v. As illustrated in Figure 2.9, a positive $\Delta\Delta V_h$ is expected to be originated from hydration of hydrophobic groups because of the exclusion effects of the hydrophobic groups on the water molecules. The excluded volume for a water molecule located near hydrophobic groups is larger than that of a bulk water molecule; this implies that hydration of hydrophobic groups results in a positive $\Delta\Delta V_h$. In contrast, the hydration of one water molecule to a hydrophilic protein surface from a bulk solution is considered to contribute to negative volume changes (-3 or -4 mL mol^{-1}) [13]. Therefore, the positive ΔV_u suggests that the hydration of hydrophobic groups prevails over that of hydrophilic groups in the Cyt *c* unfolding.

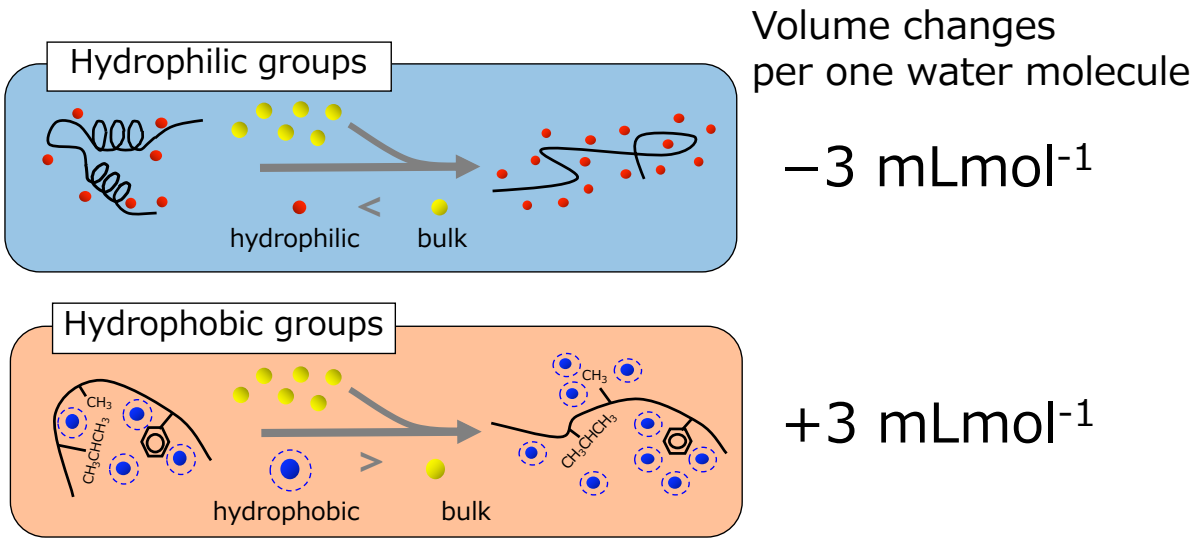

Figure 2.9 Volume changes induced by the hydration of hydrophilic and hydrophobic groups of polypeptide chains. $\Delta\Delta V_h$ for the hydration of the hydrophilic group was reported to be − 3 mL mol $^{-1}$ [13]. I hypothesized that $\Delta\Delta V_h$ for the hydration of one water molecule to the hydrophobic groups is also +3 mL mol^{-1}.

However, the average hydrophobicity of amino acid residues in the sequence of Cyt *c* does not deviate considerably from that of other proteins (Cyt *c*: −0.90; staphylococcal nuclease A [SNase A]: −0.86; *trp* repressor: −0.45; chymotrypsin inhibitor-2: −0.36). Therefore, no substantial difference would be present between the hydration of hydrophobic amino acid residues in Cyt *c* and that in other proteins. It should be noted here that Cyt *c* has a covalently attached heme group, and heme is an iron porphyrin complex containing several hydrophobic functional groups at the heme periphery. As previously reported [50, 51], heme serves as a hydrophobic core to initiate the refolding of Cyt *c*, and it is highly likely that the numerous water molecules located near heme in the unfolded state of Cyt *c* are dehydrated in the "collapse" phase of the refolding. Although reports on the quantitative analysis of the hydration of the heme

group are lacking, it can be safely concluded that the hydration of the hydrophobic heme group is the primary and major factor in the positive shift in ΔV_u for the Cyt *c* unfolding.

Hydration with heme in the Cyt *c* unfolding

To examine the contribution of the hydration of the heme to ΔV_u in the Cyt *c* unfolding, I compared volume changes in the protein unfolding of Cyt *c* with those of SNase A containing no prosthetic groups but a similar averaged hydrophobicity to Cyt *c*. A previous study has reported ΔV_u as -72 mL mol^{-1} for the urea-denatured unfolding of SNase A [46] , and V_v was determined to be 50 mL mol^{-1}, providing an estimate for $\Delta\Delta V_h$ of approximately -20 mL mol^{-1}. Assuming that the contribution of the hydration of the polypeptide to ΔV_u in the Cyt *c* unfolding is comparable to that in the SNase A unfolding, the contribution of hydration of heme to $\Delta\Delta V_h$ can be estimated to be 160 mL mol^{-1} for the urea-induced unfolding of Cyt *c*.

To determine the thermodynamic contribution of the hydration of the heme, the number of water molecules hydrated to the heme was estimated from $\Delta\Delta V_h$ [4]. The volume changes induced by the dehydration are illustrated in Figure 2.9. Although the volume change induced by the hydration of one water molecule to the hydrophobic group has not yet been experimentally and theoretically determined, $\Delta\Delta V_h$ for the hydration of the hydrophilic group was reported to be -3 mL mol^{-1} [13]. I hypothesized that $\Delta\Delta V_h$ for the hydration of one water molecule to the hydrophobic groups is +3 mL mol^{-1}, and on the basis of this hypothesis, I was able to estimate that approximately 50 water molecules can be hydrated to heme. To confirm the validity of the number of water molecules hydrated to the hydrophobic heme group, I calculated the maximum number of water molecules of the hydration layer of heme. Considering the surface area of heme

as approximately 400 $Å^2$ and the radius of a water molecule as 1.4 Å, the maximum number of water molecules involved in the hydration was determined to be 70. However, the actual number of hydrated water molecules will be less than 70, because heme in unfolded Cyt *c* is not completely exposed to the solvent. Therefore, it is plausible that approximately 50 water molecules are located near heme in the urea-denatured state of Cyt *c*, and these water molecules would be dehydrated in the Cyt *c* refolding process.

The number of hydrated water molecules of the heme allowed to quantitatively examine the thermodynamic contribution of the dehydration to Cyt *c* folding. The entropic contribution of the dehydration from the hydrophobic groups is approximately -1 kJ mol^{-1} ($-T\Delta S = -1.06$ kJ mol^{-1} at 25°C) per one water molecule (3). Therefore, the thermodynamic contribution of the dehydration from the heme group would be approximately -50 kJ mol^{-1}. Comparing the total Gibbs energy change of the Cyt *c* folding, ΔG_{UN} (-37 kJ mol^{-1}) [4], the entropic contribution of the dehydration from the heme is the major driving force for the Cyt *c* protein folding. If the dehydration from the heme is lost, the Cyt *c* folding would be a thermodynamically unfavorable reaction; this is further supported by the fact that Cyt *c* without the heme (apoCyt *c*) is unable to attain its native structure [52], and the hydrophobic interactions between the heme group and hydrophobic amino acid residues are crucial for the formation of the hydrophobic core for the initial stage of the Cyt *c* folding [50, 51].

Different volume changes between the two denaturants

Although the Cyt *c* unfolding induced by both urea and GdnHCl indicated a positive $\Delta\Delta V_h$, a more positive $\Delta\Delta V_h$ was estimated for the urea denaturation, which implies that the unfolded protein structure is different between the two denatured states.

Such different denatured states have also been encountered for the small-angle X-ray scattering measurements. The radius of gyration of GdnHCl-denatured Cyt *c* (30.3 Å) [53] is significantly larger than that of urea-denatured Cyt *c* (29.7 Å) [54]. Compared to urea, GdnHCl has been found to be a much effective denaturant, which is supported by the higher *m* values of GdnHCl [55] and the solubility of amino acids at the same denaturant concentration [20, 56, 57]. Therefore, the GdnHCl-unfolded state of Cyt *c* is less structured and more extended than that of the urea-denatured state, showing a large gyration radius.

Assuming hydration structures of urea- and GdnHCl-denatured unfolded Cyt *c* are shown in Figure 2.10. For GdnHCl-denatured Cyt *c*, guanidium ions have direct and strong interactions with hydrophobic groups in Cyt *c* [21]; certain hydrophobic sites in GdnHCl-denatured Cyt *c* are occupied by guanidium ions, and certain hydrated water molecules surrounding hydrophobic sites are expelled to the bulk. This results in fewer hydrated water molecules around the *hydrophobic* amino acid residues in the GdnHCl-denatured state and a more negative shift in $\Delta\Delta V_h$, compared with that in urea-denatured Cyt *c* (140 to 110 mL mol^{-1}). In contrast, urea is known as a relatively weak denaturant, which destabilizes protein folded structures by forming hydrogen bonds to the hydrophilic peptide groups [21, 22]. Such hydrogen bonds with urea impede the hydration of hydrophilic amino acid residues in the urea-denatured state. Therefore, urea reduces the number of hydrated water molecules near the *hydrophilic* groups in the denatured states, leading to a shift of $\Delta\Delta V_h$ to the positive side. In addition, urea-denatured Cyt *c* retains some secondary structure composed mainly of α-helix [21], and such an α-helical structure contains several hydrogen bonds between hydrophilic amide and carbonyl groups, and partially formed α-helices in the urea-denatured state reduce the number of

hydrated water molecules surrounding the *hydrophilic* amino acid residues, which further leads to a positive shift in $\Delta\Delta V_h$.

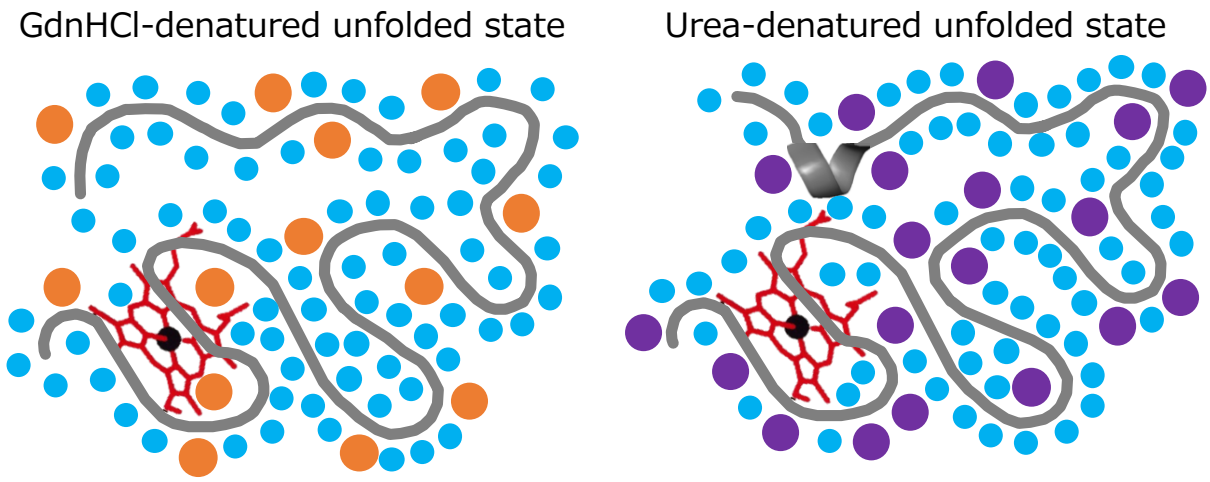

Figure 2.10 Assuming hydration structure of guanidine hydrochloride (GdnHCl)- and urea-denatured unfolded state of Cyt *c*. Polypeptide chain of Cyt *c* (gray), hydrated water molecule (light blue), guanidinium ion (orange) and urea (purple) are illustrated. GdnHCl interacts with hydrophobic sites in Cyt *c*; certain hydrophobic sites in GdnHCl-denatured Cyt *c* are occupied by guanidium ions, and certain hydrated water molecules surrounding hydrophobic sites are expelled to the bulk. This results in fewer hydrated water molecules around the *hydrophobic* amino acid residues in the unfolded state. In contrast, urea is a rather weak denaturant that interacts with hydrophilic amino acid residues to destabilize protein native structures by forming hydrogen bonds to the peptide groups. Therefore, in the urea-denatured unfolded state, some of the hydrophilic amino acid residues interact with urea, not water molecules, leading to fewer hydrated *hydrophilic* groups.

Mutation to perturb the hydrated structure of unfolded Cyt *c*

As clearly described in the previous section, the hydrated structure of the unfolded state in Cyt *c* depends on the denaturant. This difference in the hydrated structure could be attributed to the different ligation structures between urea- and GdnHCl-denatured states in Cyt *c*. In urea denaturation, Met80, one of the axial ligands for heme, is replaced by His33, whereas in GdnHCl denaturation, His26 is also ligated to the heme iron (Figure 2.11) [33, 58]; this enabled to speculate that the denatured structure

surrounding the heme axial ligands reflects the different $\Delta\Delta V_h$ values between the two denaturants. Therefore, I constructed the H26Q variant Cyt *c*, where His26 is replaced with Gln to cleave the iron–His26 ligation in the unfolded state, to mimic the urea-denatured state in the GdnHCl denaturation.

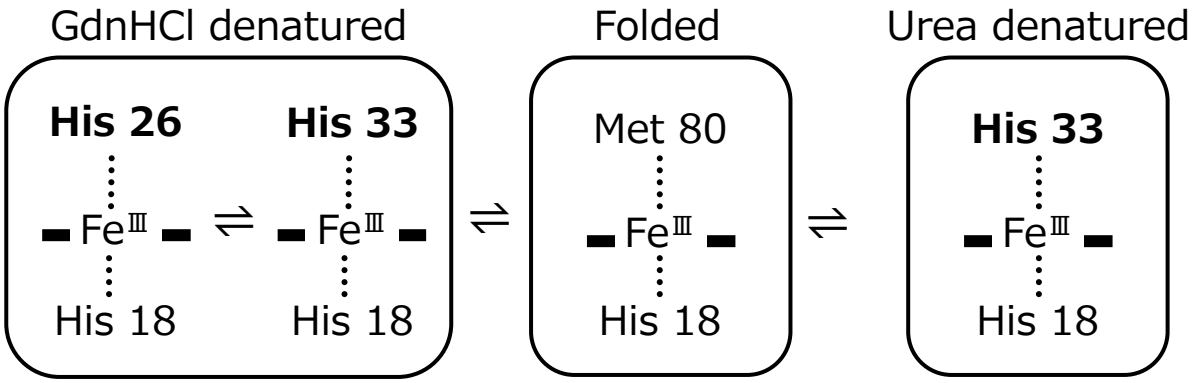

Figure 2.11 Heme coordination structures in urea- and guanidine hydrochloride (GdnHCl)-denatured Cyt *c*. The GdnHCl-denatured state presents equilibrium between two different coordination structures (left), whereas the urea-denatured state has a single coordination state (right).

I determined ΔV_u and $\Delta\Delta V_h$ for the mutants using urea (Figure 2.12) or GdnHCl (Figure 2.13) as listed in Tables 2.2 and 2.3. Although I expected that ΔV_u for the unfolding to the urea-induced unfolded state would be similar to that of the GdnHCl-induced unfolded state in the H26Q mutant, ΔV_u in the unfolding to urea-denatured Cyt *c* had a considerably positive shift from 60 ± 3 to 5 ± 1 mL mol^{-1}, whereas a mild positive shift from 29 ± 1 to 21 ± 1 mL mol^{-1} was observed for the unfolding to GdnHCl-denatured Cyt *c*. The hydrated structure of the unfolded state remained different in the H26Q mutant, and the mutational effects on the hydrated structure of the unfolded state were more enhanced in urea-denatured Cyt *c*.

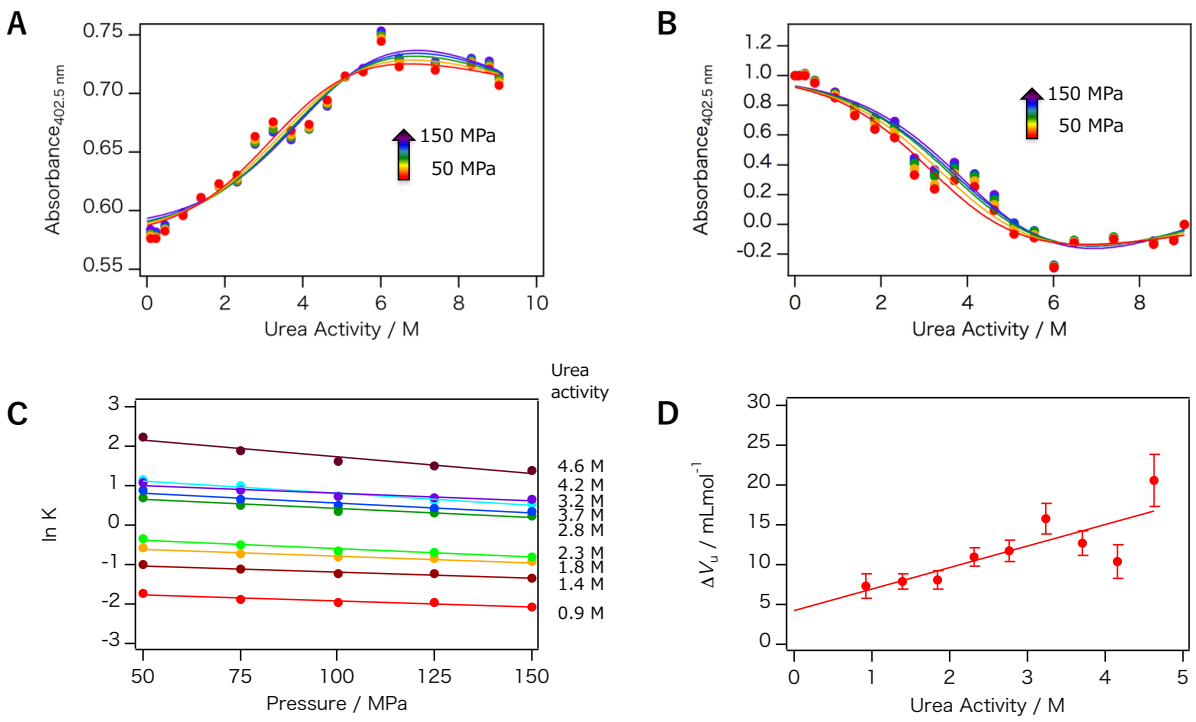

Figure 2.12 (A) Denaturation curves of the H26Q mutant induced by urea. The curves are fitted using a two-state transition model for each pressure. (B) The denaturation curves normalized by the ratio of folded Cyt *c* molecules**.** (C) Pressure dependence of the equilibrium constant for the Cyt *c* folding on urea activities. ΔV_u for each urea activity is determined by calculating the slopes of the pressure dependence of the equilibrium constant. (D) ΔV_u for urea-denatured unfolding of Cyt *c*. ΔV_u is determined by the extrapolation to 0 M urea activity.

Table 2.6 ΔV_u for urea-induced unfolding of H26Q mutant of Cyt *c*. ΔV_u values were calculated using Eq. 4 and the slope of the corresponding activity in the ln *K*−Pressure plot in Figure 2.12C.

Urea activity, M	ΔV_u , mL mol^{-1}	Standard deviation of ΔV_u , mL mol^{-1}
0.93	7.3	1.6
1.4	8.0	1.0
1.9	8.1	1.2
2.3	11	1
2.8	12	1
3.2	16	2
3.7	13	2
4.2	10	2
4.6	21	3

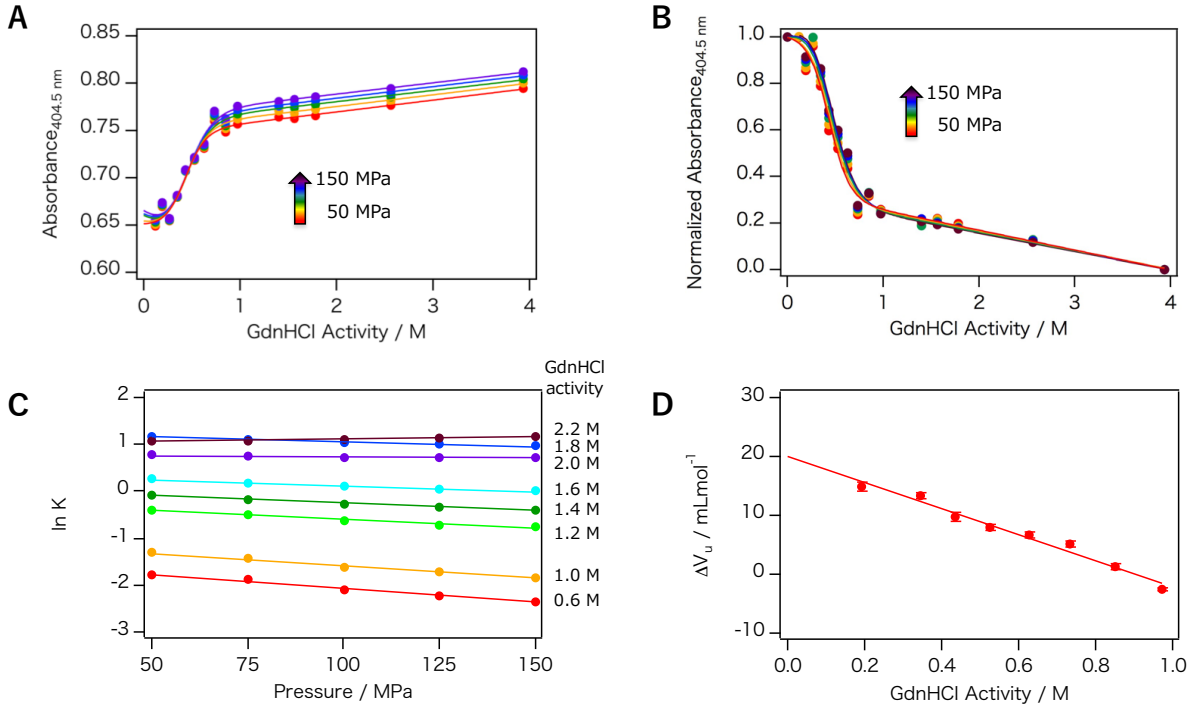

Figure 2.13 (A) The denaturation curves of the H26Q mutant induced by guanidine hydrochloride (GdnHCl). The curves are fitted using a two-state transition model for each pressure. (B) The denaturation curves normalized by the ratio of folded Cyt *c* molecules. (C) Pressure dependence of the equilibrium constant for the Cyt *c* folding on GdnHCl activities. ΔV_u for each GdnHCl activity is determined by calculating the slopes of the pressure dependence of the equilibrium constant. (D) ΔV_u for GdnHCl-denatured unfolding of Cyt *c*. ΔV_u is determined by the extrapolation to 0 M GdnHCl activity.

Table 2.7 ΔV_u for guanidine hydrochloride (GdnHCl)-induced unfolding of H26Q mutant of Cyt *c*. ΔV_u values were calculated using Eq. 4 and the slope of the corresponding activity in the ln *K*−Pressure plot in Figure 2.13C.

GdnHCl activity, M	ΔV_u , mL mol^{-1}	Standard deviation of ΔV_u , mL mol^{-1}
0.19	15	1
0.35	13	1
0.43	9.6	0.7
0.53	7.9	0.5
0.63	6.7	0.5
0.74	5.2	0.5
0.85	1.3	0.6
0.97	−2.5	0.3

Comparing the midpoint denaturant activity $[urea]_{50\%}$ (the activity of urea at which 50% of molecules are denatured) for the denaturant-induced unfolding between the H26Q variant and wild-type Cyt *c*, a more prominent shift was observed for the unfolding of the urea-denatured state: from 7.0 to 4.0 M in the urea-denatured state and from 1.3 to 0.5 M in the GdnHCl-denatured state. This is in contrast to the pH-induced unfolding probed by Trp fluorescence, showing that the protein stability of the mutant is nearly the same as that of wild type Cyt *c* [58]. Such discordance would arise from the structural differences between the denaturant- and lowering pH-induced unfolded states [32, 36, 58]. The lower midpoint activities in the mutant correspond to the lower stability of the ligation of Met80 in the mutant, indicating that the mutation at His26 to Gln promotes the replacement of the heme axial ligand, Met80, by His33 at lower urea activity. Hence, the mutant facilitates the urea-denatured unfolding, which causes more solvated and less structured α-helix components in urea-denatured Cyt *c*. Considering that several of the α-helix components are hydrophilic groups consisting of hydrogen bonds,

hydrophilic amino acid residues in the urea-denatured state are more hydrated in the H26Q mutant and the number of hydrated water molecules near the hydrophilic amino acid residues in the urea-denatured state will be increased. The more hydrated structure surrounding the hydrophilic amino acid residues in the urea-denatured state corresponds to the negative shift of $\Delta\Delta V_h$ of the H26Q mutant (Table 2.5).

The mutation caused a much less drastic negative shift in $\Delta\Delta V_h$ for the GdnHCl denaturation compared with that for urea denaturation. Because GdnHCl is a more effective denaturant than urea, GdnHCl-denatured Cyt *c* already lost most of the α-helical structures and the mutation at His26 did not induce further drastic denaturation, leading to the small shift of $\Delta\Delta V_h$ by the mutation. Although the unfolded state of the urea-denatured mutant Cyt *c* was a more solvated and less structured state, $\Delta\Delta V_h$ for the urea-denatured unfolding of the mutant remains different from that for the GdnHCl-denatured mutant. This difference in $\Delta\Delta V_h$ between the two denaturants is a result of the residual folded structure including hydrophobic interactions in urea-denatured Cyt *c* and/or hydrophilic interactions in GdnHCl-denatured Cyt *c*.

2.5. References

[1] England, J.L. & Haran, G. Role of Solvation Effects in Protein Denaturation: From Thermodynamics to Single Molecules and Back. in *Annu. Rev. Phys. Chem.* (Leone, S.R., Cremer, P.S., Groves, J.T., & Johnson, M.A. eds.) vol. 62, pp. 257-277, (2011).

[2] Kauzmann, W. Some Factors in the Interpretation of Protein Denaturation1. in *Adv. Protein Chem.* (C.B. Anfinsen, M.L.A.K.B. & John, T.E. eds.) vol. 14, pp. 1-63 Academic Press, (1959).

[3] Makhatadze, G.I. & Privalov, P.L. Contribution of Hydration to Protein Folding Thermodynamics: I. The Enthalpy of Hydration. *J. Mol. Biol.* **232**, 639-659 (1993).

[4] Privalov, P.L. & Makhatadze, G.I. Contribution of Hydration to Protein Folding Thermodynamics: II. The Entropy and Gibbs Energy of Hydration. *J. Mol. Biol.* **232**, 660-679 (1993).

[5] Chalikian, T.V. & Breslauer, K.J. On volume changes accompanying conformational transitions of biopolymers. *Biopolymers* **39**, 619-626 (1996).

[6] Silva, J.L., Foguel, D. & Royer, C.A. Pressure provides new insights into protein folding, dynamics and structure. *Trends Biochem. Sci.* **26**, 612-618 (2001).

[7] Royer, C.A. Revisiting volume changes in pressure-induced protein unfolding. *Biochim. Biophys. Acta Protein Struct. Molec. Enzy.* **1595**, 201-209 (2002).

[8] Roche, J. & Royer, C.A. Lessons from pressure denaturation of proteins. *J. Royal Soc. Interface* **15**, (2018).

[9] Imai, T., Kovalenko, A. & Hirata, F. Solvation thermodynamics of protein studied by the 3D-RISM theory. *Chem.Phys. Lett.* **395**, 1-6 (2004).

[10] Chalikian, T.V. & Filfil, R. How large are the volume changes accompanying protein transitions and binding? *Biophys. Chem.* **104**, 489-499 (2003).

[11] Frye, K.J. & Royer, C.A. Probing the contribution of internal cavities to the volume change of protein unfolding under pressure. *Protein Sci.* **7**, 2217-2222 (1998).

[12] Harpaz, Y., Gerstein, M. & Chothia, C. Volume changes on protein folding. *Structure* **2**, 641-649 (1994).

[13] Gerstein, M. & Chothia, C. Packing at the protein-water interface. *Proc. Natl. Acad. Sci. USA* **93**, 10167-10172 (1996).

[14] Dill, K.A. Dominant forces in protein folding. *Biochemistry* **29**, 7133-7155 (1990).

[15] Heremans, K. High pressure effects on proteins and other biomolecules. *Annu. Rev. Biophys. Bioeng.* **11**, 1-21 (1982).

[16] Hummer, G., Garde, S., Garcia, A.E., Paulaitis, M.E. & Pratt, L.R. The pressure dependence of hydrophobic interactions is consistent with the observed pressure denaturation of proteins. *Proc. Natl. Acad. Sci. USA* **95**, 1552-1555 (1998).

[17] Voss, N.R. & Gerstein, M. 3V: cavity, channel and cleft volume calculator and extractor. *Nucleic Acids Res.* **38**, W555-W562 (2010).

[18] Horovitz, A., Matthews, J.M. & Fersht, A.R. α-Helix stability in proteins: II. Factors that influence stability at an internal position. *J. Mol. Biol.* **227**, 560-568 (1992).

[19] Alonso, D.O.V. & Dill, K.A. Solvent denaturation and stabilization of globular proteins. *Biochemistry* **30**, 5974-5985 (1991).

[20] Smith, J.S. & Scholtz, J.M. Guanidine Hydrochloride Unfolding of Peptide Helices: Separation of Denaturant and Salt Effects. *Biochemistry* **35**, 7292-7297 (1996).

[21] Hedoux, A., Krenzlin, S., Paccou, L., Guinet, Y., Flament, M.P. & Siepmann, J. Influence of urea and guanidine hydrochloride on lysozyme stability and thermal denaturation; a correlation between activity, protein dynamics and conformational changes. *Phys. Chem. Chem. Phys.* **12**, 13189-13196 (2010).

[22] Lim, W.K., Rosgen, J. & Englander, S.W. Urea, but not guanidinium, destabilizes proteins by forming hydrogen bonds to the peptide group. *Proc. Natl. Acad. Sci. USA* **106**, 2595-2600 (2009).

[23] Babul, J. & Stellwagen, E. The existence of heme-protein coordinate-covalent bonds in denaturing solvents. *Biopolymers* **10**, 2359-2361 (1971).

[24] Babul, J. & Stellwagen, E. Participation of the protein ligands in the folding of cytochrome *c*. *Biochemistry* **11**, 1195-1200 (1972).

[25] Tsong, T.Y. Acid induced conformational transition of denatured cytochrome *c* in urea and guanidine hydrochloride solutions. *Biochemistry* **14**, 1542-1547 (1975).

[26] Muthukrishnan, K. & Nall, B.T. Effective concentrations of amino acid side chains in an unfolded protein. *Biochemistry* **30**, 4706-4710 (1991).

[27] Elove, G.A., Bhuyan, A.K. & Roder, H. Kinetic Mechanism of Cytochrome *c* Folding: Involvement of the Heme and Its Ligands. *Biochemistry* **33**, 6925-6935 (1994).

[28] Margoliash, E. & Frohwirt, N. Appendix—Spectrum of horse-heart cytochrome *c*. *Biochem. J.* **71**, 570-572 (1959).

[29] Sato, W., Hitaoka, S., Inoue, K., Imai, M., Saio, T., Uchida, T*., et al.* Energetic Mechanism of Cytochrome *c*-Cytochrome *c* Oxidase Electron Transfer Complex Formation under Turnover Conditions Revealed by Mutational Effects and

Docking Simulation. *J. Biol. Chem.* **291**, 15320-15331 (2016).

[30] Jeng, W.Y., Chen, C.Y., Chang, H.C. & Chuang, W.J. Expression and characterization of recombinant human cytochrome *c* in E-coli. *J. Bioenerg. Biomembr.* **34**, 423-431 (2002).

[31] Kyte, J. & Doolittle, R.F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**, 105-132 (1982).

[32] Russell, B.S., Melenkivitz, R. & Bren, K.L. NMR investigation of ferricytochrome *c* unfolding: Detection of an equilibrium unfolding intermediate and residual structure in the denatured state. *Proc. Natl. Acad. Sci. USA* **97**, 8312-8317 (2000).

[33] Russell, B.S. & Bren, K.L. Denaturant dependence of equilibrium unfolding intermediates and denatured state structure of horse ferricytochrome *c*. *J. Biol. Inorg. Chem.* **7**, 909-916 (2002).

[34] Dreydoppel, M., Becker, P., Raum, H.N., Groger, S., Balbach, J. & Weininger, U. Equilibrium and Kinetic Unfolding of GB1: Stabilization of the Native State by Pressure. *J. Phys. Chem. B* **122**, 8846-8852 (2018).

[35] Brems, D.N. & Stellwagen, E. Manipulation of the observed kinetic phases in the refolding of denatured ferricytochromes *c*. *J. Biol. Chem.* **258**, 3655-3660 (1983).

[36] Goto, Y., Takahashi, N. & Fink, A.L. Mechanism of acid-induced folding of proteins. *Biochemistry* **29**, 3480-3488 (1990).

[37] Latypov, R.F., Cheng, H., Roder, N.A., Zhang, J. & Roder, H. Structural characterization of an equilibrium unfolding intermediate in cytochrome *c*. *J. Mol.Biol.* **357**, 1009-1025 (2006).

[38] Santoro, M.M. & Bolen, D.W. A test of the linear extrapolation of unfolding free energy changes over an extended denaturant concentration range. *Biochemistry* **31**, 4901-4907 (1992).

[39] Yao, M. & Bolen, D.W. How Valid Are Denaturant-Induced Unfolding Free Energy Measurements? Level of Conformance to Common Assumptions over an Extended Range of Ribonuclease A Stability. *Biochemistry* **34**, 3771-3781 (1995).

[40] Stokes, R. Thermodynamics of aqueous urea solutions. *Aust. J. Chem.* **20**, 2087-2100 (1967).

[41] Makhatadze, G.I., Fernandez, J., Freire, E., Lilley, T.H. & Privalov, P.L. Thermodynamics of aqueous guanidinium hydrochloride solutions in the temperature range from 283.15 to 313.15 K. *J. Chem. Eng. Data* **38**, 83-87 (1993).

[42] Makhatadze, G.I. & Privalov, P.L. Protein interactions with urea and guanidinium chloride. *J. Mol.Biol.* **226**, 491-505 (1992).

[43] Bushnell, G.W., Louie, G.V. & Brayer, G.D. High-resolution three-dimensional structure of horse heart cytochrome *c*. *J. Mol. Biol.* **214**, 585-595 (1990).

[44] Nakamura, S. & Kidokoro, S. Volumetric Properties of the Molten Globule State of Cytochrome *c* in the Thermal Three-State Transition Evaluated by Pressure Perturbation Calorimetry. *J. Phys. Chem. B* **116**, 1927-1932 (2012).

[45] Dewa, M., Tayauchi, M., Sakurai, M. & Nitta, K. Compression refolding of cytochrome *c*. *Protein Pept. Lett.* **5**, 265-268 (1998).

[46] Vidugiris, G.J.A., Markley, J.L. & Royer, C.A. Evidence for a molten globule-like transition state in protein folding from determination of activation volumes. *Biochemistry* **34**, 4909-4912 (1995).

[47] Desai, G., Panick, G., Zein, M., Winter, R. & Royer, C.A. Pressure-jump studies of the folding/unfolding of trp repressor. *J. Mol. Biol.* **288**, 461-475 (1999).

[48] Mohana-Borges, R., Silva, J.L., Ruiz-Sanz, J. & de Prat-Gay, G. Folding of a pressure-denatured model protein. *Proc. Natl. Acad. Sci. USA* **96**, 7888-7893 (1999).

[49] Pappenberger, G., Saudan, C., Becker, M., Merbach, A.E. & Kiefhaber, T. Denaturant-induced movement of the transition state of protein folding revealed by high-pressure stopped-flow measurements. *Proc. Natl. Acad. Sci. USA* **97**, 17-22 (2000).

[50] Kimura, T., Sakamoto, K., Morishima, I. & Ishimori, K. Dehydration in the Folding of Reduced Cytochrome *c* Revealed by the Electron-Transfer-Triggered Folding under High Pressure. *J. Am. Chem. Soc.* **128**, 670-671 (2006).

[51] Akiyama, S., Takahashi, S., Ishimori, K. & Morishima, I. Stepwise formation of alpha-helices during cytochrome *c* folding. *Nat. Struct. Biol.* **7**, 514-520 (2000).

[52] Fisher, W.R., Taniuchi, H. & Anfinsen, C.B. On the Role of Heme in the Formation of the Structure of Cytochrome *c*. *J. Biol. Chem.* **248**, 3188-3195 (1973).

[53] Segel, D.J., Fink, A.L., Hodgson, K.O. & Doniach, S. Protein denaturation: A small-angle X-ray scattering study of the ensemble of unfolded states of cytochrome *c*. *Biochemistry* **37**, 12443-12451 (1998).

[54] Hsu, I.J., Shiu, Y.J., Jeng, U.S., Chen, T.H., Huang, Y.S., Lai, Y.H., *et al.* A solution study on the local and global structure changes of cytochrome *c*: An unfolding process induced by urea. *J. Phys. Chem. A* **111**, 9286-9290 (2007).

[55] Myers, J.K., Pace, C.N. & Scholtz, J.M. Denaturant m values and heat capacity changes: Relation to changes in accessible surface areas of protein unfolding *Protein Sci.* **5**, 981-981 (1995).

[56] Nozaki, Y. & Tanford, C. The Solubility of Amino Acids and Related Compounds in Aqueous Urea Solutions. *J. Biol. Chem.* **238**, 4074-4081 (1963).

[57] Nozaki, Y. & Tanford, C. The solubility of amino acids, diglycine, and triglycine in aqueous guanidine hydrochloride solutions. *J. Biol. Chem.* **245**, 1648-1652 (1970).

[58] Colón, W., Wakem, L.P., Sherman, F. & Roder, H. Identification of the Predominant Non-Native Histidine Ligand in Unfolded Cytochrome *c*. *Biochemistry* **36**, 12535-12541 (1997).

Chapter III
Quantitative description and classification of protein structures by amino acid networks

3.1. Abstract

While protein classes based on their secondary structures are widely used, quantitative characterization of the protein structures remains difficult. To quantitatively categorize protein structures, I developed a quantitative coarse-grained model of protein structures with a novel amino acid network—the interaction selective network (ISN)—that includes structural information about interactions in main and side chain atoms. The ISN enables us to distinguish between α-helix and β-sheet protein structures using two network parameters: average vertex degree (k) and average clustering coefficient (C). Although a conventional Cα (α-carbon) network with a cutoff distance of 5.5 Å allowed for the discrimination between the two secondary structures on the k–C plot, the slight shifts in the cutoff distance led to a loss of discrimination, showing lower robustness of the model, due to the definition of the links by the interactions only between the main chain atoms. Conversely, the links in the ISN are based on interactions in *both* the main and side chains, thus reflecting structural information from both secondary and tertiary structures, leading to a wider range for the cutoff distance for the discrimination, and correspondingly higher robustness. Thus, the ISN provides a quantitative and more robust description of three-dimensional protein structures.

3.2. Introduction

Proteins are biological macromolecules made up of liner chains of amino acid residues that fold into the corresponding unique three-dimensional (3D) structures comprising secondary structure elements, whereby they acquire their own functions regulated by their 3D structures. The search to understand protein structure geometries has led to the development of many experimental and theoretical methods [1-5]. Classification based on the protein structural data is one approach to comparing 3D protein structures. Two of the most prominent protein structure classification schemes, SCOP (Structural Classification Of Proteins) and CATH (Class, Architecture, Topology, Homologous superfamily) [6-10], have been widely utilized.

SCOP is the oldest structural manual classification database in which the protein structures are classified into several 'classes' and 'folds'. On the first level of the hierarchy, the 'class' is sorted into four major classes—all-α, all-β, α+β and α/β—describing the content of these secondary structural elements in the domain. These classes are distinguished by dominant secondary structural elements—α-helix and β-sheet—which are detected by the geometry of hydrogen bonds into domains [11]. In the newer classification database, CATH, the class assignment is automatically assigned according to the ratio of the secondary structural compositions, whereas it is manually classified in the case of the protein tertiary structures [12]. However, it should be noted here that quantitative comparison of the protein secondary structures cannot utilize these approaches, since the secondary structural compositions show no clear boundaries for the specific structures as they are continuously distributed. Moreover, because these databases characterize the protein structures by separating several hierarchies,

understanding the protein 3D structures by integrating both the secondary and tertiary structures become difficult.

Network characterization, which represents the protein 3D structure as an amino acid network (AAN), is one promising approach to providing quantitative insights into the classification of protein 3D structures. The network is a mathematical model describing complex structures as 'vertices' and 'links', enabling the quantitative characterization of their geometry using the network parameters. Figure 3.1 shows a network model and calculation of the network parameters. The 'vertices' correspond to amino acid residues in the protein structures, while the 'links' represent van der Waals contacts and/or chemical interactions between two amino acid residues. I treat the AAN as a coarse-grained model of protein 3D structures characterized by vertices and links. The AAN also offers great computational advantages. Traditional molecular dynamic simulation for the analysis of protein 3D structures demanded enormous computational resources [13]. The machine learning used in the prediction of protein secondary structures from amino acid sequences requires complex iterative calculations [14]. Compared with these approaches, the AAN is a substantially less computationally intensive model of protein 3D structures using information on protein geometry [5, 15].

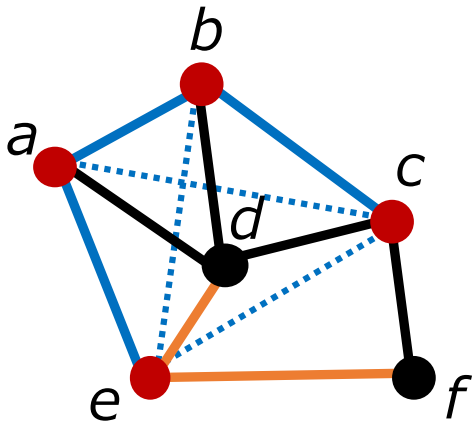

Figure 3.1 Network representation and network parameters. In an amino acid network (AAN), 'vertex' represents each amino acid residue and 'link' corresponds to an interaction between amino acid residues. Here the network comprises six vertices—*a, b, c, d, e, f*—and links between these vertices. One of the network parameters, 'vertex degree of an amino acid residue *d*', k_d, is the number of links connected to *d*. In the figure, *d* has four links and is connected to four red vertices; *a, b, c, e* ($k_d = 4$). Vertex degree for each vertex (from *a* to *f*) is 3, 3, 3, 4, 3, and 2, respectively. 'Average degree,' k is the average vertex degree of all vertices in the protein structures ($k = 3$). 'Clustering coefficient of *d*,' C_d, is the fraction of the links among the nearest neighbors of *d* to the maximum number of possible links among them. Here, *d* has four neighbors colored in red. There are three links with full blue lines between these vertices, while the maximum number of possible links between four vertices is 6 (sum of the number of links with blue continuous and dashed lines), and then $C_d = \frac{3}{6} = 0.50$. 'Average clustering coefficient', C is the average clustering coefficient of all vertices in the protein structures. The clustering coefficient of each vertex (from *a* to *f*) is $\frac{2}{3}, \frac{2}{3}, \frac{1}{3}, \frac{3}{6}, \frac{1}{3}, 0$, respectively, and then $C = \frac{21}{6} / 6 = 0.42$. 'Distance between *d* and *f*', L_{df}, is the number of links on the shortest path between *d* and *f* (orange links) and then $L_{df} = 2$. 'Average distance', L, is the average of the distances between all vertex pairs. The total number of vertex pairs is $\binom{6}{2} = \frac{6\times5}{2\times1} = 15$ and the sum of the distance between all vertex pairs is calculated as

$$\sum_{\substack{i=a,j=b \\ i\neq j}}^{f} L_{ij} = L_{ab} + L_{ac} + L_{bc} + \cdots = 1 + 2 + 1 + \cdots = 21$$

Therefore, $L = \frac{21}{15} = 1.4$.

In this study, a new AAN was constructed and used for classification of protein 3D structures deposited in the Protein Data Bank (PDB) [16] by network parameters. Of

the two well-known types of reported AANs, one is the Cα (α-carbon) network(CAN), in which the links are established if the distance between two Cα atoms is less than a cutoff distance, R_c, which is empirically determined from 7.0 Å to 8.5 Å [5, 17]. Another previously used network is the atom distance network (ADN), where the links are defined by the atom interactions between all but the hydrogen atoms in the amino acids. The ADN provides information about all van der Waals contacts between amino acid residues [5, 18]. Using these AANs to characterize protein 3D structures, previous studies have clarified several network properties of AANs [19, 20].

Many previous AAN studies have been applied to elucidate the relationship between protein domains in allosteric regulation, because the network parameters reflect the difference of global, rather than local, conformation. However, only a few studies have utilized AANs for the classification of protein 3D structures by their secondary structures, while an extremely limited number of studies have focused on distinguishing α-helix and β-sheet structures and classifying protein 3D structures by AAN [17, 20]. Alves and Martinez, for instance, analyzed 160 low homology proteins using an AAN and classified them into four structural classes—all-α, all-β, α+β and α/β [17]—but also showed that these four classes share similar network geometry, making it difficult to discriminate between α-helix and β-sheet structures. Such previous AAN studies present two serious problems in applying network approaches to distinguish protein 3D structures.

One problem is the ambiguity in determining the optimum value of R_c. Although R_c is a key factor in the development of network geometry and the characterization of network properties, in previous studies using CANs, a wide range of R_c values (mainly 7.0 Å to 8.5 Å) have been employed, implying ambiguity of the classification system. Previous CANs also cannot quantitatively characterize protein secondary structures [5,

17]. Links in CANs do not include the chemical properties of interactions, such as hydrogen bonds, and hydrophobic or other interactions. The links in CANs are established depending only on the distance between two Cα atoms. The lack of structural information about side chains is another problem. In CANs, the links are determined only by the distance between two Cα atoms [21] and include no structural and chemical properties of side chains.

On the other hand, in ADNs, the optimum R_c can be uniquely determined as the distance of chemical interactions by the protein crystal structure. Although the links in ADNs are established based on the interactions involving both main chain and side chain atoms, all interactions are independent of the chemical properties of these interactions and thus are uniformly treated. Therefore, ADNs are also assumed to lose information about the various types of interactions. I therefore developed a new AAN—termed the interaction selective network (ISN)—which includes both information about chemical properties of interactions and successfully identified α-helix and β-sheet protein structures, in order to quantitatively characterize protein secondary structures and classify protein 3D structures. The links of our ISN are determined by the distance of certain atom pairs (apart from hydrogen atoms) between two amino acid residues. Only atom pairs involved in hydrogen bonds, hydrophobic interactions, disulfide bonds, ionic interactions and covalent bonds are used. The ISN as a model of protein 3D structures has the advantage that interactions involved in *both* main chain and side chain atoms are reflected. The ISN has thus enabled us to discriminate between the all-α and the all-β protein structures by their geometry, based on the ratio of the secondary structure elements, especially the α-helix. I also confirmed the difficulty of characterizing the protein secondary structures with previously used CANs and ADNs. However, I also found that

a CAN using R_c = 5.5 Å was able to distinguish between all-α and all-β protein structures, but a small deviation of R_c from 5.5 Å resulted in a less clear discrimination, implying that the CAN is a less robust network than the ISN in terms of R_c. The ISN is, therefore, a more quantitative and robust AAN, which I expect will be widely used for quantitatively characterizing and classifying protein 3D geometries.

3.3. Methods

Construction of amino acid network (AAN)

The AANs are based on the structural data of folded-state protein as available in PDB [16] and defined as follows: each amino acid residue of the protein molecule is taken as a vertex in AANs. If two amino acid residues contact each other in a protein structure, a link is established. All interactions between the residues are identified as links. Three types of AANs with differing definition of the links are used. In the CAN, two residues are connected (a link between two vertices is established) if the distance between their Cα atoms is less than the threshold distance (R_c = 8.5 Å). The links of the ADN are determined by the distance of the closest atom pairs between two amino acid residues. Because data sets of the protein structures determined by X-ray crystallography do not contain the atomic coordinates of hydrogen atoms, hydrogen atoms are ignored. If the distance is less than R_c = 5.0 Å, these two amino acid vertices are connected; otherwise, they are disconnected. R_c is defined as the distance of van der Waals contacts [22, 23]. In our novel introduction of the ISN, the links are determined by the distance as defined in the ADN. However, to establish the links, only atom pairs involved in hydrogen bonds, hydrophobic interactions, disulfide bonds, ionic interactions and covalent bonds are used. Hydrogen atoms, as in the ADN, are ignored. R_c in the ISN is determined based on the cutoff value used in Protein Interaction Calculator (PIC) [24]. PIC is a server computing various interactions in a protein structure based upon the coordinate set of the 3D structure of the protein. The R_c value of a hydrogen bond was set to be 3.8 Å between oxygen and/or nitrogen atoms, a value determined by exploring the optimum R_c for distinguishing all-α and all-β proteins in this study; this value of R_c is larger than that of previous studies (3.5 Å), where the maximum donor-acceptor distance was used [25, 26]. In the case of

hydrophobic interactions, R_c was 5.0 Å between the side chain carbon atoms in the following hydrophobic residues: Ala, Val, Leu, Ile, Met, Phe, Trp, Pro and Tyr [27]. The R_c value of a disulfide bond was 2.2 Å between the sulfur atoms, while that of the ionic bond was 6.0 Å between the side chain nitrogen and oxygen atoms in the following ionic residues: Arg, Lys, His, Asp and Glu. The covalent bond is established if there are two consecutive amino acid residues on the amino acid sequence.

These AANs are characterized by following several parameters, as shown in Table 3.1. The network parameters widely used for previous AANs are average vertex degree *k*, average clustering coefficient *C*, and average path length *L*, as shown in Figure 3.1. The *k* value of a protein structure is defined as

$$k = \frac{1}{N}\sum_{i=1}^{N} k_i$$

where *N* is the number of amino acid residues, and the vertex degree, k_i, is the connected number of amino acid residues, *i*, with other vertices in a protein structure. C_i, the clustering coefficient of the amino acid residue, *i*, is defined as the fraction of links that exist among the nearest neighbors of the amino acid residue, *i*, to the maximum number of possible links among them. Therefore,

$$C_i = \frac{2n_i}{k_i(k_i\text{-}1)}$$

where n_i is the number of links that actually exist, and k_i (k_i -1) is the number of all possible links for the nearest neighbors. *C* of a network can be calculated by

$$C = \frac{1}{N}\sum_{i=1}^{N} C_i$$

L is defined as

$$L = \frac{2}{N(N\text{-}1)} \sum_{i=1}^{N\text{-}1} \sum_{j=i+1}^{N} L_{ij}$$

where L_{ij} is the number of links on the shortest path between the *i*th and *j*th amino acid vertices. Additional parameters in network, assortativity and maximum vertex degree, are calculated. Assortativity is the correlation of the degree between vertices adjacent to each other. The maximum vertex degree is the largest *k* in the network.

Table 3.1 Network parameters in this study.

Network parameters	Symbol	Description
Number of vertices	N_{Vertices}	The number of vertices in the network.
Number of links	N_{Links}	The number of links in the network.
Average vertex degree	k	Average degree for all vertices. Degree of a vertex is the number of links connected to the vertex.
Average clustering coefficient	C	Average of clustering coefficient for all vertices. Clustering coefficient of a vertex is the fraction of links that exist among the nearest neighbours of each residue to the maximum number of possible links among them.
Average distance	L	Average of network distance for all vertex pairs. Network distance is the number of links on the shortest path between vertices.
Vertex assortativity	---	The correlation of the degree between vertices adjacent to each other.
Maximum vertex degree	---	The maximum degree in the network.

Data sets of protein structures

Protein structures are obtained from the PDB database [16]. This study focused on four broad protein structural classes—all-α, all-β, α+β and α/β, according to the SCOP classification [7]. The SCOP classifies single αβ class of CATH into two classes, α+β and α/β. We, therefore, used the SCOP classification data to display the difference between α+β and α/β classes by the network parameters. An example of protein structure for each class is shown in Figure 3.2. All-α proteins comprise predominantly α-helices, while all-β proteins, β-sheets. α+β proteins haveα-helices and β-strands that are largely segregated, whereas α/β proteins are largely interspersed [7]. For data set, to ensure the quality of structural data used for AAN construction, high-resolution (higher than 2.0 Å) X-ray

crystallographic structures of proteins as classified all-α, all-β, α+β andα/β by the SCOP are chosen. The data set was limited in non-modified monomer proteins without any ligands or nucleic acids. This is due to the difficulty in the definition of non-amino acid components in AAN. Since PDB entries frequently contain only a portion of the protein sequence, the protein structures with over 90% coverage of a protein sequence are selected [16, 28]. Table 3.2 shows that the number of protein structures satisfied the above criteria. List of PDB ID of these data sets are shown in Table 3.3. The number of such structures depended on the AAN I examined. Although the SCOP database has approximately 10,000 protein domains for all-α, all-β, α+β and α/β classes, high-resolution (higher than 2.0 Å) X-ray crystallographic protein structures without any ligands are less than 1,000 structures. Some protein structures were not converted into the specific AAN structures due to errors in the construction of the AAN. This is caused by the lack of the atomic coordinates at the middle of the sequence in the PDB entry files. In this study, these irregular data sets should be removed.

Above data set also used as the learning data set For the test data, two selection criteria, non-modified proteins and structures with over 90% coverage of a protein sequence, are removed from the above criteria due to increasing the number of protein structures of data set. The number of protein structures and list of PDB ID in test data set is shown in Tables 3.4 and 3.5.

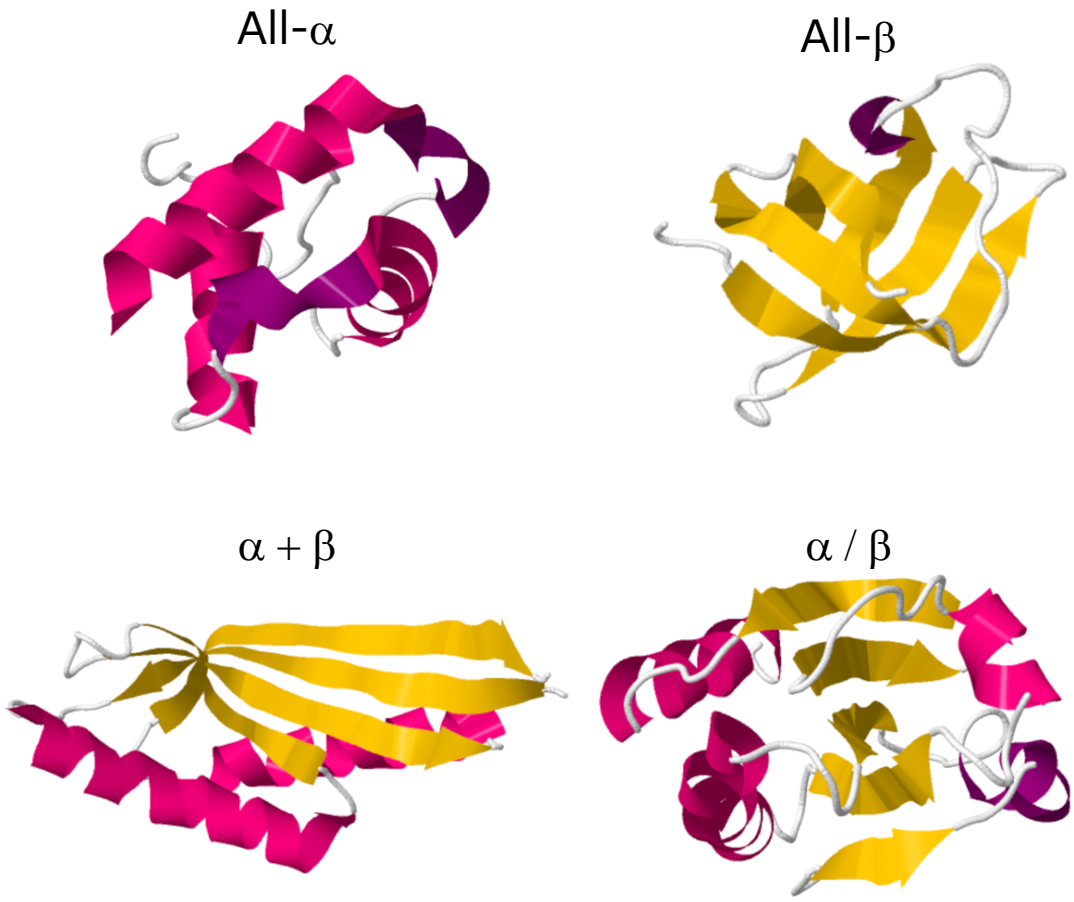

Figure 3.2 Four major protein classes according to the SCOP classification. An example of protein structures for all-α (PDB ID:1uj8, ORF3 in two crystal forms, a member of Isc machinery of *E. coli*), all-β (1mjc, Crystal structure of CSPA, the major cold shock protein of *Escherichia. coli*), α+β (1j27, Crystal structure of a hypothetical protein, TT1725, from *thermus thermophilus* HB8 at 1.7Å resolution) and α/β (1thx, Thioredoxin-2) is shown. Protein structures are deposited in the Protein Data Bank (PDB). All-αproteins comprise predominantly α-helices, while all-βproteins, β-sheets. α+β proteins haveα-helices and β-strands that are largely segregated, whereas α/β proteins are largely interspersed.

Table 3.2 The number of protein structures used for the Cα network (CAN), the atom distance network (ADN), and the interaction selective network (ISN). The data set also used as the learning data set.

	CAN	ADN	ISN
All-α	59	60	52
All-β	83	84	71
α+β	155	157	131
α/β	161	163	135

Table 3.3 Lists of PDB IDs of protein structures for constructing amino acid networks (AANs). PDB IDs with including data set for the interaction selective network (ISN), the Cα network(CAN) and atom distance network (ADN) are listed with the patterns of listed data set. Lists of PDB ID of protein structures for constructing amino acid networks (AANs). The data set also used as the learning data set.

Class	Including data set for			List of PDB ID
	ISN	CAN	ADN	
All-α	Yes	Yes	Yes	1uj8, 1hyp, 1klx, 2end, 1tzv, 1ng6, 2oeb, 1ljp, 1y6i, 1vap, 1xqo, 1ouv, 2f23, 2ijq, 1nc5, 1xsz, 2d5j, 1nkd, 1rpo, 2oo2, 2pmr, 1cei, 1r7j, 3wrp, 1y7y, 2za7, 2fd5, 2a8f, 1kso, 2fbq, 1bm9, 1ztd, 1lj9, 1c02, 1unk, 2b0j, 3c90, 1sgm, 1r1u, 1t06, 1otk, 1o0w, 1ufb, 2a61, 2cwl, 1b43, 1fp3, 2gz4, 2gmy, 2o70, 1np3
	No	Yes	Yes	1ks9, 2fez, 1h13, 1sqg, 1ayi, 1cuk, 1xg7, 1ynb
	No	No	Yes	---
	Yes	No	Yes	1us6
	No	Yes	No	---
	Yes	Yes	No	---
	Yes	No	No	---

Class	Including data set for			List of PDB ID
	ISN	CAN	ADN	
All-β	Yes	Yes	Yes	1ucs, 1ame, 1mjc, 1tvq, 1ifb, 1ifc, 1lpj, 1kqx, 2eif, 1md6, 1w8v, 2cpl, 1awq, 1dyw, 1amm, 1gcs, 4gcr, 1qoi, 1kn3, 1r8n, 1sqw, 1xnd, 2jic, 2sfa, 1pvx, 1row, 2cv3, 3app, 1ist, 1qcx, 1wos, 1ueb, 1bhe, 1iz6, 1pp3, 1vav, 1uok, 1pxz, 1m1g, 1awr, 1yrz, 1gvp, 1a1x, 1whi, 2dp9, 1lib, 1is5, 2p84, 1c3k, 1gpr, 1ep0, 1v8h, 1opa, 1bkz, 1y2t, 1xxo, 1wlt, 1nxm, 1gd7, 1cq3, 2cga, 1i81, 1mgq, 1ku8, 1i8d, 1e15, 1th7, 1a12, 1h64, 1yif, 1k32
	No	Yes	Yes	2fr2, 1hk0, 1wba, 1xvm, 2fez, 1pev, 1pe9, 1vqb, 1lcl, 1ei5, 1cjd, 1mr7
	No	No	Yes	1snt
	Yes	No	Yes	---
	No	Yes	No	---
	Yes	Yes	No	---
	Yes	No	No	---

Class	Including data set for			List of PDB ID
	ISN	CAN	ADN	
α+β	Yes	Yes	Yes	1ulr, 1zpw, 1j27, 1p1l, 1ln4, 1ew4, 2ppn, 1ra4, 2czw, 2fc3, 1acf, 1tp6, 1dkj, 1q2y, 1t3x, 2d4p, 1vfq, 1j74, 2esk, 1oz9, 1bd8, 1jyh, 1hka, 1u9a, 1u9b, 2a4d, 1lyd, 2lzm, 3lzm, 1rl6, 153l, 1gbs, 1mol, 1sqw, 2dyj, 1tua, 1iv9, 1cqm, 1nwa, 1xkr, 1dzf, 1rc9, 1bol, 1uoh, 2baa, 1zwz, 1pxw, 1o0x, 2ohw, 1ako, 1uek, 2zjd, 1oqw, 1gcu, 1ltu, 1mla, 2f23, 1nij, 1xiz, 1xix, 1wos, 1vf8, 1gnd, 1g61, 2p4x, 1xsz, 1ll7, 1m1g, 1gx3, 1nfj, 1ris, 2pii, 1bxy, 1dhn, 1qto, 1s7i, 1twu, 2fl4, 1npk, 1bv1, 2b3m, 1r7l, 2hiq, 1eo6, 1reg, 1rwz, 1gy6, 1rxz, 1x25, 1zo2, 1sei, 2f5g, 1kcm, 1qah, 1wkj, 1y9w, 2nck, 1vh5, 1z4e, 1sh8, 1mk4, 2cwk, 1ihb, 1b8p, 1mkb, 2gdg, 1cgq, 1gif, 1dpt, 1otf, 1mp9, 1jzo, 1gfl, 1rw0, 1o0w, 1te5, 1n5b, 1erz, 2fe5, 1tkl, 1iz9, 1el6, 1hfo, 1mmi, 3d1e, 1ok7, 2gdq, 1iv1, 1e15, 1ftr, 1vdh
	No	Yes	Yes	1ptf, 1box, 2o0q, 1hhl, 1zhv, 1t3y, 1icx, 1ojq, 2f1n, 1gqz, 1jbb, 1et6, 2nml, 1q8r, 1tlu, 2g3t, 1ode, 1vq3, 1vpk, 2o5u, 1bkp, 1ty0, 2a10, 1jd1
	No	No	Yes	1vjk, 1w9g
	Yes	No	Yes	---
	No	Yes	No	---
	Yes	Yes	No	---
	Yes	No	No	---

Class	Including data set for			List of PDB ID
	ISN	CAN	ADN	
α/β	Yes	Yes	Yes	1h75, 1thx, 1tmy, 1wou, 1e6l, 2chf, 2fi9, 1iuk, 1iul, 2d59, 1srv, 1o7u, 1o85, 1o8w, 1o8x, 1ezk, 2rn2, 1goa, 2nx2, 1q2u, 2pth, 1yzf, 1v77, 1i39, 1g8a, 1eug, 1h4y, 1tib, 1a8q, 1gcu, 1bqc, 1k6a, 1mla, 1ak1, 1ede, 2had, 2pky, 2yxp, 2cyg, 1tm2, 1qwk, 1nij, 1lzl, 1xfk, 1mpb, 1omp, 1rh9, 1vf8, 1gnd, 1yxy, 1cwy, 1p1x, 1wd7, 1uok, 1jr2, 1gzj, 1onr, 1t1u, 1ll7, 1ert, 1j7g, 1tfu, 1j85, 2b0a, 2dvp, 1p5f, 1vc1, 2fuk, 1byi, 1v8e, 3erj, 1gqn, 2dst, 3ct6, 1fsf, 1ye5, 1v7l, 1b8p, 1q98, 1j33, 2b0j, 1ald, 1g2q, 2r1u, 1dsb, 6xia, 2gub, 2car, 1v5x, 2ab0, 1iu8, 1e5m, 1jzo, 1auo, 1jfl, 1n7k, 1h7e, 1ypi, 1tcd, 1ktn, 1a4u, 1udr, 1a7u, 1wdj, 1oqf, 1s2t, 1xxu, 1l7a, 1bsl, 1iz9, 1b43, 1dys, 2ahb, 1jkm, 2gdq, 1hkq, 1ock, 1a88, 1lu9, 1ade, 1j2w, 1e15, 1hxh, 1hm5, 1hrd, 1np3, 1xmp, 1t5o, 1zah, 2gtd, 1woh, 1jiq, 1k32
	No	Yes	Yes	1m5t, 1r26, 1o6d, 1a2j, 1io2, 1t1j, 1ks9, 1mtz, 1cv2, 1jw4, 1sqg, 1uax, 1vic, 1ytl, 2ex0, 2i5d, 1f2j, 1fvk, 1ucf, 1q5x, 1mjf, 1tc5, 1aug, 1vr6, 1ecp, 1epx, 1vl2
	No	No	Yes	1qk8
	Yes	No	Yes	2bgk, 1ojx
	No	Yes	No	---
	Yes	Yes	No	1l2l
	Yes	No	No	---

Table 3.4 The number of protein structures used for the Cα network (CAN), the atom distance network (ADN), and the interaction selective network (ISN) in the test data set.

	CAN	ADN	ISN
All-α	247	254	207
All-β	557	557	454
α+β	647	664	551
α/β	399	415	344

Table 3.5 Lists of PDB IDs of protein structures for constructing amino acid networks (AANs) in the test data set. PDB IDs with including data set for the interaction selective network (ISN), the Cα network(CAN) and atom distance network (ADN) are listed with the patterns of listed data set.

Class	Including data set for			List of PDB ID
	ISN	CAN	ADN	
All-α	Yes	Yes	Yes	1A62, 1AA2, 1BEA, 1BKR, 1BZ4, 1CEM, 1DVO, 1EG3, 1EG4, 1ELK, 1ENJ, 1ENK, 1EYH, 1EZ3, 1FAZ, 1FPO, 1FQI, 1GAK, 1GS9, 1GXN, 1GXQ, 1HYP, 1I2T, 1IAP, 1J8M, 1JMW, 1K0M, 1K6K,1KLX, 1KW4, 1LJP, 1LKI, 1LWB, 1MIX, 1MN8, 1MZL, 1N7N, 1N7O, 1N7P, 1NC5, 1NG6, 1NTY, 1OAI, 1OPC, 1OUV, 1OXJ, 1P2F, 1PBV, 1PBW, 1Q5Z, 1R0D, 1R69, 1RJ1, 1S29, 1SBX, 1T6O, 1T95, 1TQG, 1TZV, 1U2K, 1UJ8, 1VAP, 1VIN, 1W7B, 1WER, 1WWI, 1X91, 1XQO, 1XSZ, 1Y6I, 1YU5, 1YU8, 1Z96, 1ZLB, 2CWY, 2CXD, 2D5J, 2END, 2ES9, 2F23, 2FF4, 2FQ3, 2GV2, 2ICT, 2IJQ, 2IOL, 2IXM, 2J5Y, 2J9V, 2LIS, 2OEB, 2P5K, 2Q0Z, 2QJZ, 2QY9, 2RH3, 2RJV, 2RJY, 2VE8, 2YGS, 1A8O, 1AIE, 1AIL, 1BGF, 1BHD, 1BM9, 1C02, 1C26, 1CEI, 1CUN, 1D9C, 1DJ8, 1FC3, 1FIP, 1FP3, 1G2Y, 1G2Z, 1G39, 1G8E, 1G8Q, 1GTO, 1GU9, 1HF8, 1IJY, 1IRQ, 1IZM, 1JB6, 1K8U, 1KSO, 1LDD, 1LJ9, 1M4R, 1NKD, 1NOG, 1NP3, 1O0W, 1OTK, 1Q2H, 1QQF, 1QSJ, 1R1T, 1R1U, 1R7J, 1RK4, 1ROP, 1S4K, 1SGM, 1SH5, 1T06, 1UFB, 1UFI, 1UNK, 1UTG, 1UZ3, 1VLS, 1WU9, 1X2I, 1XSV, 1XWG, 1Y6X, 1Y7Y, 1YIB, 1YIG, 1ZKR, 1ZTD, 2A61, 2A8F, 2B0J, 2CWL, 2D8E, 2DB7, 2F6L, 2FD5, 2FQ4, 2G7O, 2G7S, 2G9E, 2GEN, 2GMY, 2GPO, 2GS4, 2GZ4, 2IB0, 2IJK, 2O4D, 2O70, 2OFY, 2OKU, 2OO2, 2PMR, 2SPC, 2UTG, 2ZA7, 3C90, 3LYN, 3WRP
	No	Yes	Yes	1H13, 1H14, 1KS9, 1KU3, 1L8R, 1LIS, 1ORC, 1SH6, 1SQG, 1U00, 1VKU, 1WLZ, 1XW2, 1XWT, 1YZM, 2FEZ, 2HUJ, 2IOS, 1AYI, 1B6Q, 1BAZ, 1CI4, 1DKX, 1DKZ, 1GMG, 1ILK, 1MD0, 1RFY, 1RPO, 1TJV, 1V2Z, 1XG7, 1YNB, 2A6B, 2FBN, 2FBQ, 2GOM, 2HZT, 2O8P, 2PEQ, 2Q0T
	No	No	Yes	1E6I, 1EA8, 1H7I, 1FBQ, 1FBS, 1FBU, 1QJA, 1VKA
	Yes	No	Yes	1GVJ, 1UPG, 1US6, 1UTU, 2IU5, 2ZBS
	No	Yes	No	1NIG, 2CJJ, 1C3C, 1H99, 1O3U
	Yes	Yes	No	2CKX, 1BGF
	Yes	No	No	---

Class	Including data set for			List of PDB ID
	ISN	CAN	ADN	
All-β	Yes	Yes	Yes	1A58, 1A62, 1AAJ, 1AGJ, 1AME, 1AMM, 1AMX, 1ARB, 1AUN, 1AW7, 1AWQ, 1AWR, 1B7I, 1B7J, 1BCK, 1BFG, 1BHE, 1BJ7, 1BK1, 1C5H, 1CDY, 1CKA, 1CKB, 1CQY, 1CRZ, 1CWC, 1CWF, 1CWH, 1CWI, 1CWJ, 1CWK, 1CWL, 1CWM, 1CWO, 1CYN, 1CZT, 1DDJ, 1DDV, 1DDW, 1DSL, 1DUA, 1DUE, 1DYW, 1E5P, 1EDQ, 1EG3, 1EG4, 1EKL, 1ENX, 1EQV, 1EUJ, 1EUR, 1EVH, 1EY0, 1EY4, 1EY5, 1EY6, 1EY8, 1EY9, 1EYA, 1EYC, 1EYD, 1EZ6, 1EZ8, 1F00, 1F2M, 1F2Y, 1F2Z, 1FC6, 1FC7, 1FC9, 1FGL, 1FKN, 1FL0, 1FNF, 1FSO, 1G1K, 1GCS, 1GMU, 1GSM, 1GZI, 1H6T, 1HOE, 1HV0, 1HV1, 1I0C, 1IDK, 1IFB, 1IFC, 1IFG, 1IHZ, 1II3, 1IOB, 1IST, 1IU1, 1IV8, 1IZ6, 1J2A, 1J48, 1JAB, 1JB3, 1JPE, 1KAA, 1KAB, 1KEX, 1KHI, 1KMT, 1KN3, 1KQX, 1KYF, 1KYU, 1L8F, 1LB6, 1LMI, 1LOP, 1LPJ, 1LPL, 1M1G, 1M1S, 1MBM, 1MD6, 1MFG, 1MFL, 1MHN, 1MIK, 1MIX, 1MJC, 1N7E, 1N7N, 1N7O, 1N7P, 1NG2, 1NG5, 1NKR, 1NOA, 1NTY, 1OA4, 1OGM, 1OLR, 1OPS, 1P3C, 1P4P, 1PDQ, 1PP3, 1PSO, 1PVX, 1PWT, 1PXZ, 1PZC, 1Q3L, 1Q7F, 1QCX, 1QKX, 1QOI, 1QRZ, 1QTP, 1QTS, 1QWX, 1QZN, 1R8N, 1RI6, 1ROW, 1RWR, 1SGZ, 1SHG, 1SLL, 1SNO, 1SQW, 1SSH, 1STN, 1SYC, 1SYE, 1SYG, 1T2P, 1THV, 1TS3, 1TUD, 1TVQ, 1TWB, 1UAI, 1UCS, 1UEB, 1UNP, 1UOK, 1UTI, 1UXZ, 1VAV, 1VBS, 1VCA, 1VDN, 1W8V, 1WGB, 1WKA, 1WOS, 1WP5, 1WWC, 1X6L, 1XN2, 1XN3, 1XNC, 1XND, 1XO7, 1XYO, 1XYP, 1YHF, 1YNA, 1YRZ, 1YTQ, 1Z0H, 1Z78, 1ZCE, 21BI, 2A6Z, 2AME, 2BEN, 2BIT, 2BJQ, 2BVV, 2BXX, 2CNY, 2CNZ, 2CO4, 2CPL, 2CV3, 2EIF, 2ERF, 2ES3, 2F15, 2FCB, 2FF4, 2FU0, 2H3L, 2I1B, 2I6V, 2J43, 2JIA, 2JIC, 2MSI, 2MSJ, 2NUZ, 2OEI, 2PCY, 2PIE, 2R99, 2RA2, 2SFA, 2SIL, 2SPG, 2YXF, 31BI, 3APP, 3APR, 3BDU, 3DJ9, 3EZM, 3MSI, 3TSS, 4GCR, 4I1B, 4MSI, 4PEP, 5MSI, 6MSI, 7AME, 7MSI, 8AME, 9AME, 1A12, 1A1X, 1AB0, 1AE2, 1AE3, 1ALY, 1AT0, 1B0W, 1B2P, 1B8E, 1BFT, 1BKB, 1BKZ, 1BMZ, 1BRE, 1BWW, 1BZ8, 1BZD, 1BZE, 1C1F, 1C3K, 1C48, 1CDC, 1CQ3, 1D2O, 1DQ0, 1DUN, 1DUP, 1DVQ, 1E15, 1E65, 1EP0, 1EZG, 1F39, 1F41, 1FCG, 1FH2, 1FIV, 1G6G, 1GD7, 1GKH, 1GPR, 1GVP, 1H64, 1H6U, 1HIX, 1HQQ, 1HXL, 1HXZ, 1HY2, 1HZ9, 1HZA, 1I07, 1I81, 1I8D, 1IS5, 1IX2, 1JHC, 1JVK, 1K32, 1KFF, 1KHX, 1KL3, 1KL4, 1KL5, 1KNB, 1KU8, 1KZQ, 1L7J, 1LGV, 1LIB, 1LVE, 1MGQ, 1N7F, 1NLR, 1NQB, 1NXM, 1NXZ, 1O6A, 1OBQ, 1OBU, 1OGH, 1OKI, 1OPA, 1OV3, 1PM4, 1Q5U, 1QAC, 1QB5, 1QWD, 1REI, 1RST, 1RSU, 1SFP, 1SHF, 1SLE, 1SLG, 1SMX, 1SOK, 1STM, 1STR, 1STS, 1SVP, 1SWA, 1SWB, 1SWL, 1SWO, 1TCZ, 1TD4, 1TH7, 1TSH, 1TTA, 1TTB, 1TTC, 1TTR, 1TVD, 1U1S, 1U1T, 1UDZ, 1UJ1, 1UZ2, 1V1H, 1V6Z, 1V8H, 1VIE, 1VPN, 1VSC, 1VWA, 1VWB, 1VWC, 1VWD, 1VWE, 1VWF, 1VWG, 1VWH, 1VWM, 1VWN, 1VWO, 1VWP, 1W9A, 1WHI, 1WHO, 1WLG, 1WLT, 1WTL, 1XB9, 1XDH, 1XHN, 1XXO, 1Y2T, 1YFN, 1Z8K, 2A13, 2AS0, 2CGA, 2DLB, 2DP9, 2DUC, 2ED6, 2FHQ, 2G4G, 2GEC, 2GQV, 2HPE, 2J2J, 2J71, 2MCG, 2NPH, 2OUJ, 2OX7, 2OYZ, 2P84, 2PSG, 2PYT, 2Q4I, 2Q4N, 2QGB, 2RHE, 2TRH, 2TRY, 3D3R, 3MCG

Class	Including data set for			List of PDB ID
	ISN	CAN	ADN	
All-β	No	Yes	Yes	1ACX, 1B9K, 1BAS, 1BE9, 1EY7, 1FNA, 1G9O, 1GYU, 1GYV, 1HCV, 1I04, 1I1B, 1I1J, 1IHJ, 1KDC, 1L6P, 1L7R, 1L6P, 1L7R, 1MSI, 1NPU, 1PE9, 1PEV, 1PGS, 1PHT, 1QAU, 1RFE, 1TEN, 1TJ6, 1TP3, 1TP5, 1TQ3, 1TS9, 1TSF, 1U00, 1U6D, 1UH9, 1V9T, 1VAI, 1VLO, 1VPR, 1WBA, 1XVM, 1ZGK, 1ZMF, 2BEN, 2BTL, 2BYG, 2FEZ, 2FNE, 2FR2, 2G30, 2I1B, 2RMC, 3SEB, 1A30, 1AOH, 1CJD, 1CZY, 1D00, 1D01, 1D0A, 1DKX, 1DKZ, 1E3F, 1EEQ, 1EI5, 1FHN, 1FON, 1FUX, 1IGQ, 1LCL, 1MR7, 1MY5, 1MY7, 1N99, 1OHQ, 1PQE, 1SEM, 1TD0, 1TY0, 1U3Z, 1VQA, 1VQB, 1VQC, 1VQD, 1VQE, 1VQF, 1VQG, 1VQH, 1VQI, 1VQJ, 1XE0, 1YIF, 2D0N, 2F9H, 2GIY, 2PAB, 2Q8O, 3CBR, 3D7P, 4TSV
	No	No	Yes	1HBQ, 1KDA, 1KDB, 1OGM, 2BHG, 1UPI, 2P23
	Yes	No	Yes	2VO8
	No	Yes	No	1FHG, 1G4M, 1MJS, 1RZ2, 1A64
	Yes	Yes	No	1MS3, 1MS5, 2Q0Z, 2OUJ
	Yes	No	No	1SND

Class	Including data set for			List of PDB ID
	ISN	CAN	ADN	
α+β	Yes	Yes	Yes	132L, 133L, 134L, 135L, 153L, 180L, 1ACF, 1AGI, 1AHC, 1AKI, 1AKO, 1AW7, 1B1J, 1B6V, 1BB3, 1BD8, 1BHF, 1BK7, 1BM8, 1BMB, 1BOL, 1BVX, 1BWH, 1BWI, 1BWJ, 1C44, 1CEW, 1CKH, 1CM2, 1CM3, 1CNS, 1CQM, 1CS8, 1D4T, 1D4W, 1DIX, 1DKJ, 1DZF, 1EDQ, 1EIC, 1EID, 1EIE, 1EKG, 1EM7, 1EQ6, 1EW4, 1F00, 1F0W, 1F10, 1FLY, 1FS3, 1FU0, 1FVA, 1G61, 1GBS, 1GCU, 1GMU, 1GND, 1GOU, 1GX3, 1HEL, 1HEM, 1HEN, 1HEO, 1HEP, 1HEQ, 1HER, 1HKA, 1HSW, 1HSX, 1HUF, 1HZT, 1I1Z, 1I20, 1I7K, 1IGD, 1IOQ, 1IOR, 1IOS, 1IOT, IQ4, 1IR7, 1IR8, 1IR9, 1IV7, 1IV9, 1IXL, 1IZP, 1IZQ, 1IZR, 1J27, 1J74, 1JHS, 1JIS, 1JIT, 1JIY, 1JJ1, 1JJ3, 1JOS, 1JVW, 1JWR, 1JYH, 1JYR, 1K1A, 1K1B, 1K50, 1KAF, 1KF5, 1KF7, 1KF8, 1KHP, 1KHQ, 1KOE, 1KXW, 1KXX, 1KXY, 1KYF, 1KYU, 1L01, 1L02, 1L04, 1L05, 1L06, 1L07, 1L08, 1L09, 1L10, 1L12, 1L13, 1L14, 1L15, 1L16, 1L17, 1L18, 1L19, 1L20, 1L21, 1L22, 1L23, 1L24, 1L25, 1L27, 1L28, 1L29, 1L30, 1L31, 1L32, 1L33, 1L34, 1L35, 1L37, 1L38, 1L39, 1L40, 1L41, 1L42, 1L43, 1L44, 1L45, 1L46, 1L47, 1L49, 1L50, 1L51, 1L52, 1L54, 1L60, 1L97, 1LAA, 1LHH, 1LHI, 1LHJ, 1LHK, 1LHL, 1LHM, 1LIT, 1LKL, 1LL7, 1LLN, 1LMN, 1LN4, 1LOU, 1LOZ, 1LP8, 1LSA, 1LSB, 1LSC, 1LSD, 1LSE, 1LSF, 1LSM, 1LSN, 1LSY, 1LTU, 1LYD, 1LYO, 1LYS, 1LYY, 1LYZ, 1LZ1, 1LZA, 1LZD, 1LZT, 1M1G, 1MLA, 1MOL, 1MWP, 1MX4, 1MX6, 1NIJ, 1NWA, 1O0X, 1OGW, 1OQW, 1OSD, 1OZ9, 1P1L, 1P56, 1P7S, 1PGB, 1PIP, 1PQK, 1PRZ, 1PV5, 1PXW, 1Q2Y, 1QS1, 1QS9, 1QSW, 1QTH, 1QTP, 1QTS, 1QXT, 1QY3, 1QYO, 1R4B, 1R9H, 1RA4, 1RAT, 1RBX, 1RC9, 1REX, 1RFP, 1RHA, 1RL0, 1RL6, 1ROA, 1SC0, 1SHA, 1SHB, 1SMB, 1SQH, 1SQW, 1T07, 1T2I, 1T3X, 1T95, 1TAY, 1TBY, 1TCY, 1TDY, 1TF1, 1TIG, 1TJE, 1TP6, 1TS3, 1TUA, 1U79, 1U9A, 1U9B, 1UBI, 1UBQ, 1UCO, 1UEK, 1UIC, 1UID, 1UIE, 1UIF, 1UIG, 1ULR, 1UOH, 1VAJ, 1VCC, 1VDQ, 1VDS, 1VDT, 1VED, 1VF8, 1VFQ, 1WDV, 1WNA, 1WOS, 1X6L, 1XIX, 1XIZ, 1XKR, 1XPS, 1XPT, 1XSZ, 1Y0O, 1Y6L, 1YCK, 1YH2, 1YI6, 1YPC, 1YQB, 1ZPW, 1ZWZ, 1ZZK, 214L, 256L, 2A4D, 2AIF, 2AUB, 2BAA, 2BF5, 2BQG, 2BQH, 2BQI, 2BQK, 2BQM, 2C8O, 2C8P, 2CDS, 2CW4, 2CZW, 2D4O, 2D4P, 2DYJ, 2ESK, 2ESO, 2ESQ, 2F23, 2FAZ, 2FC3, 2FZP, 2G8Q, 2GS5, 2HRX, 2IGD, 2LHM, 2LYM, 2LYZ, 2LZM, 2NR7, 2NWD, 2O0P, 2OHW, 2P4X, 2POF, 2PPN, 2PV1, 2RAT, 2RER, 2YVB, 2YVT, 2ZD2, 2ZEQ, 2ZJD, 2ZQ4, 3BZP, 3BZR, 3BZT, 3CTK, 3LYM, 3LYZ, 3LZM, 3RAT, 3RSD, 3TSS, 4LYT, 4LYZ, 4RAT, 5LYT, 5LYZ, 5RAT, 6LYT, 6LYZ, 6RAT, 7RAT, 8RAT, 9RAT, 137L, 1A3A, 1A68, 1B8P, 1BHH, 1BUO, 1BV1, 1BV4, 1BXY, 1BYR, 1CGQ, 1DHN, 1DPT, 1DSX, 1E15, 1EL6, 1EO6, 1ERZ, 1ESR, 1EYM, 1F2L, 1F46, 1F9Q, 1F9R, 1FTR, 1GFL, 1GIF, 1GXJ, 1GY6, 1GYB, 1HFO, 1HLW, 1HQ8, 1I7N, 1IDP, 1IFT, 1IHB, 1IV1, 1IZ9, 1J2V, 1J8B, 1KCM, 1KPA, 1KPB, 1KPT, 1KR4, 1KS2, 1M4J, 1M5S, 1MBY, 1MMI, 1MP9, 1MSC, 1N5B, 1NAP, 1NFJ, 1NPK, 1O0W, 1O22, 1O5J, 1OCV, 1OK7, 1OTF, 1PCF,

Class	Including data set for			List of PDB ID
	ISN	CAN	ADN	
α+β (Continued)	Yes	Yes	Yes	1Q6H, 1Q8R, 1QAH, 1QDV, 1QTO, 1QU9, 1QVE, 1QW2, 1QWI, 1R29, 1R7L, 1RWZ, 1RXZ, 1SEI, 1SH8, 1SPH, 1T1D, 1T4A, 1T4D, 1T82, 1TE5, 1TFE, 1TKI, 1TLU, 1TMI, 1TVX, 1TWU, 1TY0, 1U07, 1U69, 1U7I, 1VDH, 1VFJ, 1VGG, 1VH5, 1VLA, 1VPK, 1VQ3, 1W5R, 1WKJ, 1WM3, 1WZ3, 1X25, 1XS0, 1Y5H, 1Y9W, 1YAL, 1YER, 1YLX, 1YQH, 1YVO, 1Z4E, 1ZO2, 2A10, 2B3M, 2B5R, 2CB5, 2CHC, 2CHS, 2CU6, 2CVL, 2CWK, 2D3G, 2D7V, 2E2C, 2F5G, 2FB5, 2FIU, 2FL4, 2FUJ, 2G3T, 2GDG, 2GDQ, 2GE7, 2GR8, 2GUK, 2HIQ, 2HNG, 2IKB, 2IPR, 2IVY, 2J6B, 2J7Z, 2JER, 2NCK, 2NML, 2NMU, 2NWV, 2O5U, 2O7M, 2OMO, 2PII, 2PV2, 2RFR, 2RK5, 3D1E, 3IL8
	No	Yes	Yes	1AFU, 1B9K, 1BNF, 1BOX, 1BRI, 1BSA, 1BSB, 1BSC, 1BSE, 1DEU, 1DTJ, 1DZO, 1EOE, 1F32, 1FLQ, 1FLU, 1FLW, 1FN5, 1FUS, 1FZY, 1G24, 1GQZ, 1GZ2, 1HHL, 1HZ6, 1ICX, 1IS0, 1JBB, 1KH0, 1LCJ, 1LZ4, 1MHX, 1MI0, 1OJQ, 1PGX, 1PNE, 1PTF, 1RHB, 1T3Y, 1UKF, 1UN3, 1VDP, 1VHF, 1VHS, 1VLO, 1X6P, 1X6Q, 1X6R, 1X6X, 1X6Y, 1X6Z, 1XTE, 1YHW, 1YPA, 1YPB, 1ZHV, 2F1N, 2FHT, 2FO3, 2G30, 2O0Q, 2PST, 3BU3, 3BU6, 3LYZ, 3SEB, 1A09, 1A1A, 1AH6, 1BKP, 1FNJ, 1FNK, 1GY7, 1JB2, 1JD1, 1JYA, 1JYQ, 1JZO, 1KWB, 1MFF, 1MK4, 1MKB, 1MZG, 1Q8B, 1R0V, 1RG0, 1RIS, 1RU0, 1RW0, 1S7I, 1TKL, 1U0K, 1VAX, 1VGY, 1VKN, 1WM2, 2AAG, 2AAL, 2C2I, 2CI2, 2G3A
	No	No	Yes	1E21, 1M1H, 1UMW, 1UYL, 1VJK, 1VR9, 1W41, 1W42, 1W9G, 2BSZ, 2Q43, 1H8X, 1I4J, 2DM9
	Yes	No	Yes	1H9O, 1W3E, 2V94, 1O7Z, 1O80, 1ODE
	No	Yes	No	1K4N, 2CXA
	Yes	Yes	No	1P4O
	Yes	No	No	1QYU

Class	Including data set for			List of PDB ID
	ISN	CAN	ADN	
α/β	Yes	Yes	Yes	1A8Q, 1AK1, 1AKZ, 1ARL, 1ATZ, 1B00, 1B31, 1BN6, 1BQC, 1C4W, 1CEX, 1CHD, 1CNV, 1CRZ, 1CUA, 1CUB, 1CUC, 1CUF, 1CUG, 1CUH, 1CUJ, 1CUS, 1CUU, 1CUX, 1CUY, 1CWY, 1CZ1, 1DIN, 1DUS, 1DZF, 1E0W, 1E6L, 1E6M, 1EDE, 1EDG, 1EDQ, 1EQP, 1EUG, 1EZK, 1F21, 1F9M, 1FAA, 1FBN, 1FC6, 1FC7, 1FC9, 1FG4, 1G8A, 1GCU, 1GND, 1GOA, 1GOB, 1GOC, 1GZJ, 1H1N, 1H4Y, 1H6T, 1H75, 1I1X, 1I39, 1I60, 1IIB, 1IUK, 1IUL, 1IV8, 1J8M, 1JFR, 1JFU, 1JL1, 1JLN, 1JR2, 1JXB, 1JYK, 1K0M, 1K6A, 1KVA, 1KVB, 1KVC, 1LAV, 1LAW, 1LL7, 1LU4, 1LZL, 1M21, 1MF7, 1MLA, 1MPB, 1N3Y, 1NA5, 1NAR, 1NIJ, 1NM8, 1NY1, 1O13, 1O7U, 1O85, 1O8W, 1O9G, 1ONR, 1OTM, 1P15, 1P1X, 1P2F, 1PGV, 1PRY, 1Q0U, 1Q2U, 1Q7S, 1QK8, 1QWK, 1QZ0, 1QZM, 1R88, 1RBR, 1RBS, 1RBT, 1RBU, 1RBV, 1RDB, 1RH9, 1RHS, 1SRV, 1SU9, 1T1U, 1T7N, 1THX, 1TIB, 1TM2, 1TMY, 1TR9, 1TUX, 1U24, 1U9C, 1UC7, 1UOK, 1UXO, 1V77, 1VF8, 1W94, 1WD7, 1WDE, 1WEH, 1WOU, 1WSJ, 1X42, 1X6L, 1XFK, 1XYZ, 1XZA, 1XZE, 1XZF, 1XZG, 1XZH, 1XZI, 1XZJ, 1YXY, 1YZF, 1ZON, 2A4V, 2AH5, 2APJ, 2B3S, 2BV9, 2CHF, 2CVB, 2CXH, 2CYG, 2D59, 2EXO, 2F9F, 2F9S, 2FG1, 2FHP, 2FI9, 2HAD, 2HVM, 2IIY, 2NX2, 2ORA, 2PKY, 2PLC, 2PTD, 2PTH, 2QXT, 2QXU, 2QY9, 2R48, 2RN2, 2YWO, 2YXP, 3E9O, 3ENB, 3TGL, 4EUG, 1A4U, 1A7U, 1A88, 1ADE, 1AIU, 1AJR, 1ALD, 1AUO, 1B43, 1B8P, 1BI5, 1BSL, 1BYI, 1DQZ, 1DSB, 1DTS, 1DYS, 1E15, 1E5M, 1EDT, 1ERT, 1ERV, 1ERW, 1ES9, 1F5Z, 1FSF, 1G2Q, 1GQN, 1H6U, 1H7E, 1HM5, 1HQK, 1HRD, 1HXH, 1I45, 1I7N, 1I89, 1I8B, 1ILV, 1IU8, 1IZ9, 1J22, 1J23, 1J2W, 1J33, 1J7G, 1J85, 1JFL, 1JIQ, 1JKM, 1JWX, 1JZO, 1K32, 1KS2, 1KTN, 1KZ1, 1L7A, 1LU9, 1LX7, 1M6J, 1N7K, 1NP3, 1NS5, 1NSW, 1NTH, 1NXV, 1NXZ, 1O63, 1OCK, 1OQF, 1P5F, 1PBN, 1PDO, 1PDV, 1Q5X, 1Q98, 1QMV, 1QVZ, 1R0S, 1R12, 1S2T, 1SUR, 1T4D, 1T5O, 1T6T, 1TC5, 1TCD, 1TFU, 1TVN, 1U0V, 1UCF, 1UDR, 1UFO, 1UIU, 1V5X, 1V6Z, 1V7L, 1V8E, 1VC1, 1VGY, 1VHC, 1VHV, 1WDJ, 1XK6, 1XK7, 1XMP, 1XVW, 1XWG, 1XXU, 1YAC, 1YE5, 1ZAH, 2AB0, 2AS0, 2B0A, 2B0J, 2B61, 2CAR, 2DST, 2DVP, 2FUK, 2GDQ, 2GTD, 2I5D, 2I5I, 2I9I, 2O57, 2OB5, 2R1U, 2VB7, 2VL3, 3BED, 3CT6, 3ERJ, 3PGA, 6XIA
	No	Yes	Yes	1A2J, 1A3H, 1AGY, 1C3P, 1CV2, 1EWX, 1FFA, 1FFB, 1FFC, 1FFD, 1FFE, 1FTO, 1G1F, 1IJB, 1IO2, 1IXK, 1JW4, 1KID, 1KNG, 1KS9, 1L7R, 1M5T, 1MTZ, 1NDB, 1O20, 1O6D, 1OMP, 1QO2, 1R26, 1SQG, 1T15, 1T1J, 1UAX, 1VI4, 1VIC, 1VL5, 1YTL, 2EXO, 2ISB, 2LAO, 1AC1, 1ACV, 1AC1, 1ACV, 1BED, 1ECP, 1EPX, 1F2J, 1FVK, 1IHC, 1JL2, 1O4W, 1PE0, 1PRX, 1R3R, 1RK4, 1SOA, 1VGA, 1VHX, 1VKN, 1VR6, 1WOH, 1YPI, 2AHB, 2F7W, 2FA8, 2GUB, 2JK2
	No	No	Yes	1LL7, 1W94, 1Y9U, 1FBA, 1USG,
	Yes	No	Yes	1GOK, 1GOM, 1O8X, 1OEM, 1GP1, 1O9N, 1O9Q, 1OBJ, 1OCH, 1OJX, 1QCZ, 1QKK, 2BGK
	No	Yes	No	1J7X, 1R0V,
	Yes	Yes	No	---
	Yes	No	No	1ZCU,

3.4. Results

Distinguishing protein secondary structures by the interaction selective network (ISN)

In order to determine optimum R_c to distinguish between α-helix and β-sheet protein structures, the ISN with R_c of hydrogen bonds ranging from 1.0 Å to 10.0 Å are constructed, then calculated their network parameters (Table 3.1). Although all of the network parameters as listed in Table 3.1 are examined, significant correlation was only observed between k and C. Figure 3.3 shows the effect of R_c for hydrogen bonds on the classification of the all-α and all-β proteins (the number of structures are shown in Table 3.2). Under R_c = 3.2 Å, the distributions of k and C for the all-α and all-β proteins overlapped (Figure 3.3A). Ranging from R_c = 3.4 Å to R_c = 3.8 Å, the distributions were segregated with increasing R_c (Figures 3.3B and 3.3C), while with R_c > 3.8 Å the distributions were less segregated (Figure 3.3D). In the region from R_c = 3.4 Å (Figure 3.3B) to R_c = 3.8 Å (Figure 3.3C), k and C of the all-α proteins were increased, whereas those of the all-β proteins were almost unchanged (Figures 3.3B and 3.3C), reflecting the different geometry between α-helix and β-sheets.

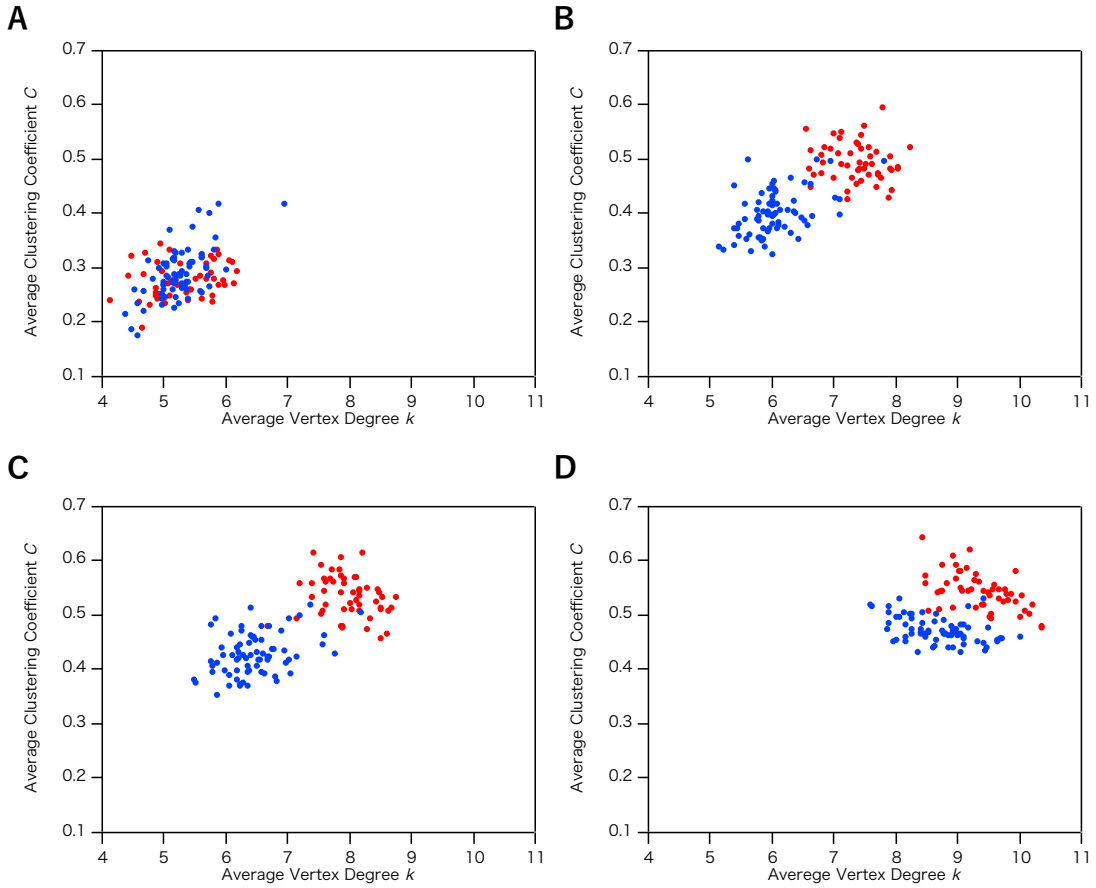

Figure 3.3 Average clustering coefficient (C) and average vertex degree (k) in an interaction selective network (ISN). The cutoff value (R_c) of the hydrogen bond is (A) 3.2 Å, (B) 3.4 Å, (C) 3.8 Å, and (D) 5.0 Å. All-α and all-β proteins are colored in red and blue, respectively.

In addition to hydrogen bonds, the ISN comprises four types of interactions—hydrophobic interactions, disulfide bonds, ionic interactions and covalent bonds. Among five types of interactions, hydrogen bonds and hydrophobic interactions are primary interactions in protein structures, thus most of the links in AANs are, therefore, established by hydrogen bonds and hydrophobic interactions. To examine the contribution of R_c to these interactions on the network geometry of AANs, ISNs with a wide range of R_c of hydrogen bonds and hydrophobic interactions are constructed. As shown in Figure 3.4A, changing R_c of hydrogen bonds from 2.5 Å to 4.0 Å induced

dramatic increases in C and N_{Links}, whereas these parameters remained almost unperturbed by changing R_c of hydrophobic interactions around 5.0 Å (Figure 3.4B). Such low sensitivity of the network parameters to R_c for hydrophobic interactions indicates that the AAN geometry is independent of R_c, which in turn suggests that R_c of hydrogen bonds is the dominant factor for the network geometries of whole protein structures.

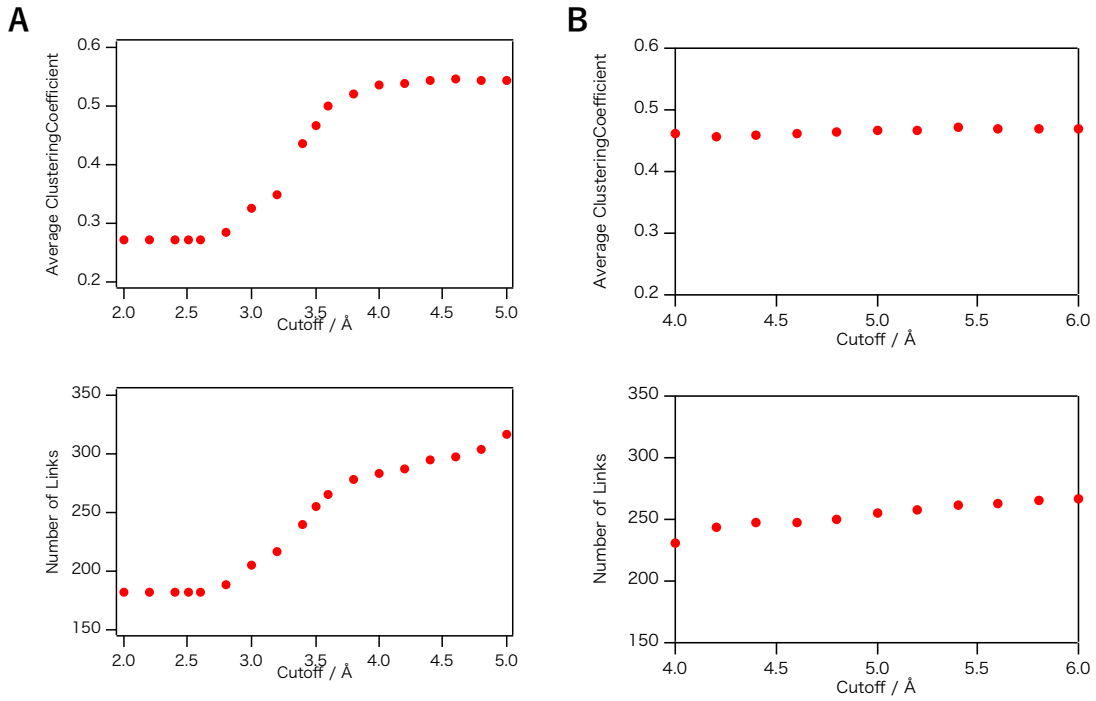

Figure 3.4 Average clustering coefficient (C) and cutoff value (R_c) (top panels), and number of links (N_{link}) and cutoff value (R_c) (bottom panels) in an interaction selective network (ISN). (A) The hydrogen bonds and (B) hydrophobic interactions. The protein structure of ORF3 in two crystal forms, a member of Isc machinery of *E. coli* (PDB ID: 1uj8), was used for the calculations.

Characterizing protein structural classes by the interaction selective network (ISN)

In the previous section, I focused on the all-α and all-α protein classes, without examining the classification of the additional two classes, α+β and α/β. Figure 3.5A

illustrates the protein structures categorized into these four classes as examined by the ISN, where the distribution of k and C for the protein structures is shown. The all-α proteins have higher k and C values than the all-β proteins; therefore, the k–C plot enables us to classify the all-α (red circles) and all-β protein structures (blue circles). It should be noted here that not all protein structures are consistent with the SCOP classification. While SCOP classified these structures into the all-α or all-β classes, the X-ray structural analysis reported that these inconsistent protein structures have both α-helix and β-sheet structures [29]. For instance, type III antifreeze protein RD1 from an Antarctic eelpout (PDB ID: 1ucs) is classified as an all-β protein in SCOP, whereas k and C (k = 8.19, C = 0.506) for this protein in the ISN can be plotted in the region for α/β proteins, not all-β proteins (Figure 3.5A). The ratio of the secondary structural components of the α-helix and β-sheet, based on the X-ray structure of this protein [29], are 20% and 25%, respectively, indicating that the analysis characterizes the protein secondary structure more accurately.

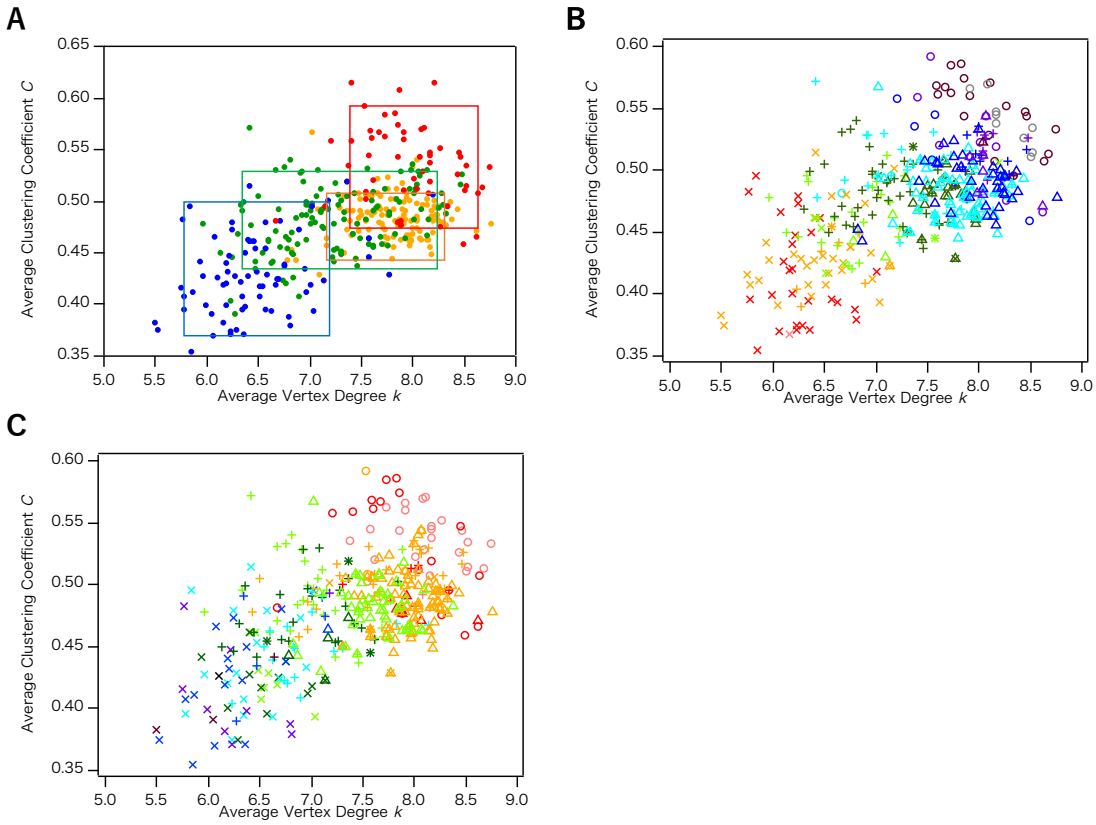

Figure 3.5. Average clustering coefficient (C) and average vertex degree (k) from the analysis of the interaction selective network (ISN). Protein structures are categorized as (A) all-α (red), all-β (blue), $\alpha+\beta$ (green), and α/β (orange); (B) α-helix, and (C) β-sheet contents. Secondary structure content; 0% (pink), 1%–10% (red), 11%–20% (orange), 21%–30% (yellow green), 31%–40% (green), 41%–50% (light blue), 51%–60% (blue), 61%–70% (purple), 71%–80% (deep purple), 81%–90% (grey), and 91%–99% (black). In (B) and (C), these protein structures are classified as all-α ($\circ$), all-β ($\times$), $\alpha+\beta$ ($+$), and α/β ($\triangle$).

I also found that the ratio of the secondary structure content determines the distribution of the k–C plot. Figures 3.5B and 3.5C indicate the dependence of k and C on the secondary structure contents. Larger k and C values of the ISN were calculated by increasing the α-helix contents, whereas the corresponding decrease in the β-sheet contents were less clear. These results suggest that the ISN detects the geometry of the α-

helix components and reflects the ratio of α-helix in the network parameters, *k* and *C*, allowing us to discriminate between the all-β and the all-β protein structures.

Figure 3.5A also show that some of the protein classes overlapped in the *k*–*C* plot of ISN. The α/β proteins (mainly from 7.2 to 8.3 for *k* and from 0.44 to 0.52 for *C*) had higher *k* values, as did the all-α proteins (mainly from 7.4 to 8.6 for *k* and from 0.48 to 0.59 for *C*) and medial *C* values, which were observed for the all-α proteins and the all-β proteins (mainly from 5.8 to 7.4 for *k* and from 0.37 to 0.50 for *C*). On the other hand, the α+β proteins (mainly from 6.3 to 8.3 for *k* and from 0.43 to 0.53 for *C*) showed a wide range of distribution of *k* and *C*, which overlapped with those for the all-α (from 7.4 to 8.3 for *k* and from 0.48 to 0.53 for *C*) as well as those for the all-β proteins (from 6.3 to 7.4 for *k* and from 0.43 to 0.50 for *C*). Although the different distributions of *k* and *C* between the α+β and the α/β proteins were roughly proportional to the α-helix content (Figure 3.5B), these overlaps on the *k*–*C* plots suggest that the geometry of the secondary structure elements should also be considered for proteins having both α-helix and β-sheet structures.

Comparison of the interaction selective network (ISN) with previously used amino acid networks (AANs)

To confirm the validity of the ISN for characterizing protein 3D structures, the network properties of the ISN are compared with those of the CAN and the ADN. For protein structures, *k* and *C* were calculated and classified according to their protein classes. In contrast to the ISN, the CAN (R_c = 8.5 Å) and the ADN (R_c = 5.0 Å) share a similar distribution of the network parameters in protein structures, regardless of the different classes (plots of *C* and *k* are shown in Figures 3.6A and 3.6B). Although the ADN

includes interactions involving both main and side chain atoms, the distribution of the *k*–*C* plot is significantly different from that of the ISN. This difference in distribution arises from the definition of the links in the two networks. The treatment of van der Waals interactions comprises the largest difference. In the ADN, all van der Waals interactions are counted as links, but the links in the ISN are defined only by van der Waals interaction between hydrophobic residues. The ISN, therefore, detects only critical interactions as links, without recognizing weakly interacting contacts with the neighbor residues in their 3D structures. This property enables us to discriminate the difference in geometry between protein secondary structures.

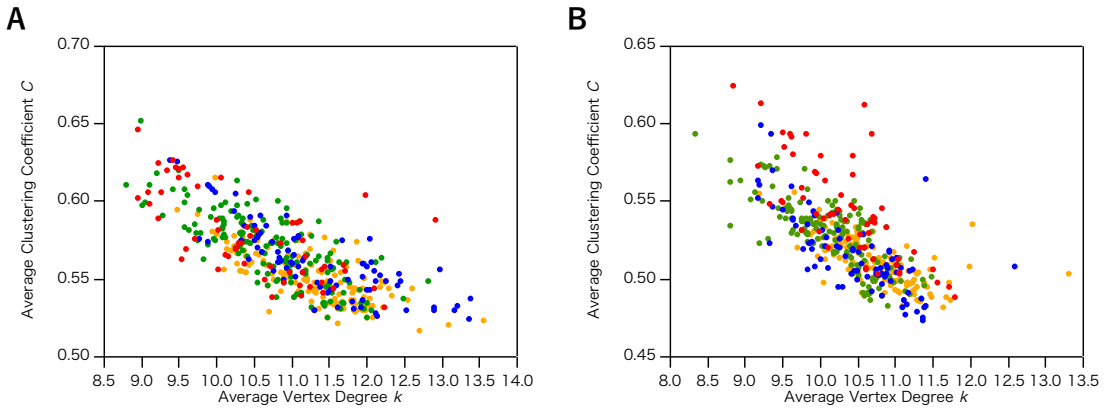

Figure 3.6 Average clustering coefficient (*C*) and average vertex degree (*k*). (A) CAN (R_c = 8.5 Å) and (B) ADN (R_c = 5.0 Å): all-α (red), all-β (blue), $\alpha+\beta$ (green), and α/β (orange).

Other notable network parameters are the numbers of vertices, N_{Vertices}, and links, N_{Links}. As clearly shown in Figure 3.7, I detected a higher correlation between N_{Vertices} and N_{Links}. Figure 3.7A shows the linear relationships with clear differences in the slopes of the four protein classes—all-α, α/β, α+β and all-β protein structures, in descending order of the slopes. Steeper slopes correspond to the protein structures that have more links per vertex than the other structures. An increased N_{Links} implies a more

crowded network, corresponding to larger *k* and *C* values, consistent with the fact that the hydrogen bond network of the α-helix is a spatial structure with many interacting atoms, whereas the β-sheet is a more planar structure with a lower number of interacting atoms. On the other hand, in the case of the CAN and the ADN, it is difficult to detect the correlations between the slopes and the protein classes, as shown in Figures 3.7B and 3.7C, respectively.

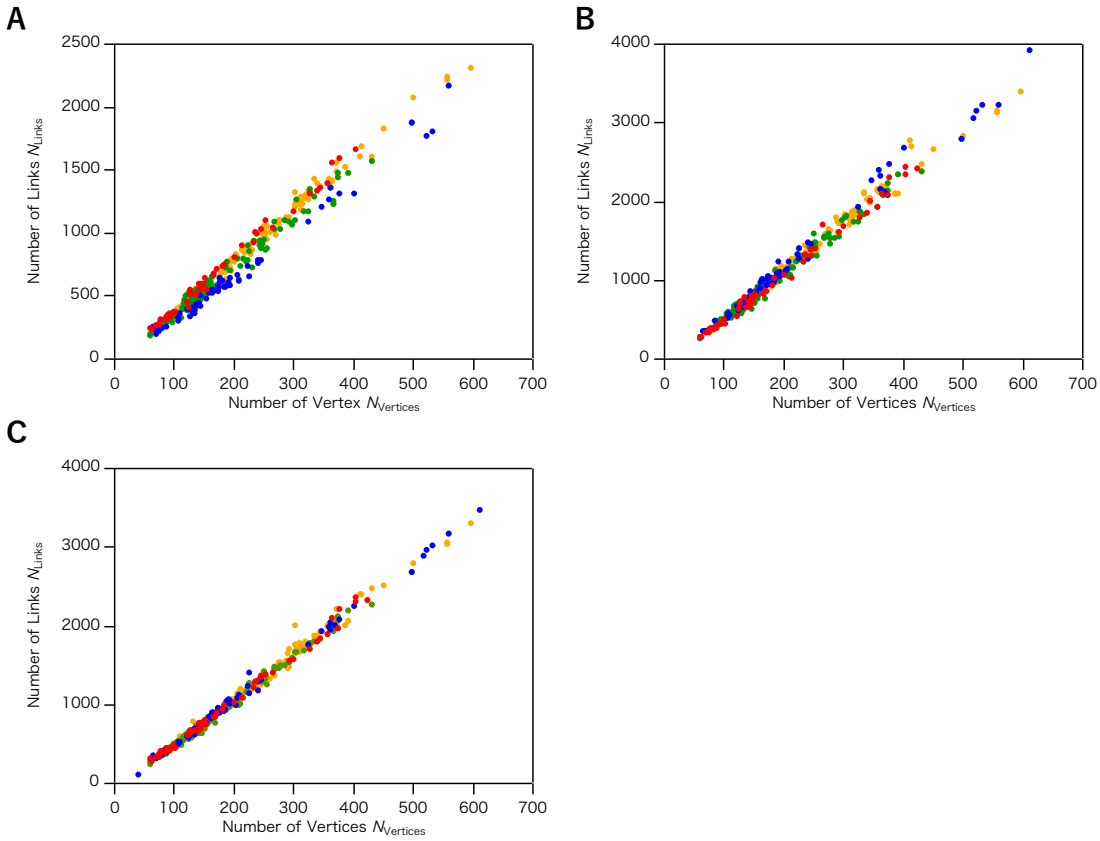

Figure 3.7 The number of links (N_{Links}) as a function of the number of vertices ($N_{Vertices}$). (A) ISN, (B) CAN (R_c = 8.5 Å), and (C) ADN (R_c = 5.0 Å). The protein structures are classified into all-α (red), all-β (blue), α+β(green), and α/β (orange).

Reexamination of discrimination between all-α and all-β protein structures by Cα network (CAN)

Previous sections demonstrated that I successfully classified the protein structure using the ISN, but Bagler and Sinha also reported that they could distinguish all-α and all-β protein structures using the CAN (R_c = 7.0 Å) [20]. Their results display a clear separation of all-α and all-β proteins in the *L*–*C* plots, in contrast to result of previous section (R_c = 8.5 Å, Figure 3.6A). To confirm the effect of R_c on the distribution of network parameters *k*, *C* and *L* in the CAN, the CAN (R_c = 5.0 Å ~ 8.5 Å) is reconstructed and the *k*–*C* and *L*–*C* plots are shown in Figure 3.8. The distributions of

k and *C* indicate that all-α and all-β protein structures were not discriminated at R_c = 7.0 Å (Figure 3.8A), while lowering R_c to under 6.0 Å enabled us to separate the two classes (Figure 3.8B). The two classes are clearly separated at R_c = 5.5 Å (Figure 3.8C). On the other hand, when lowering R_c to under 5.5 Å, *C* dramatically decreased with decreasing R_c, resulting in the loss of clustering structures (Figure 3.8D). Although the *C* and *L* provided in both this and previous studies are similar, this study indicates that the plots of *C* and *L* overlapped between all-α and all-β proteins in the CAN (R_c = 7.0 Å) (Figure 3.8E), in contrast to the previous study [20]. This discrepancy between the two results is due to differences in the data sets. The previous study used 20 protein structures for each protein class [20], whereas 59 all-α and 83 all-β protein structures are used in this study (Table 3.2), showing that data set in this study is sufficiently large to display the overlapping of plots of between all-α and all-β proteins. From the above discussion, this study demonstrates that the CAN at R_c = 5.5 Å can successfully discriminates between all-α and all-β proteins, but fails to do so at R_c = 7.0.

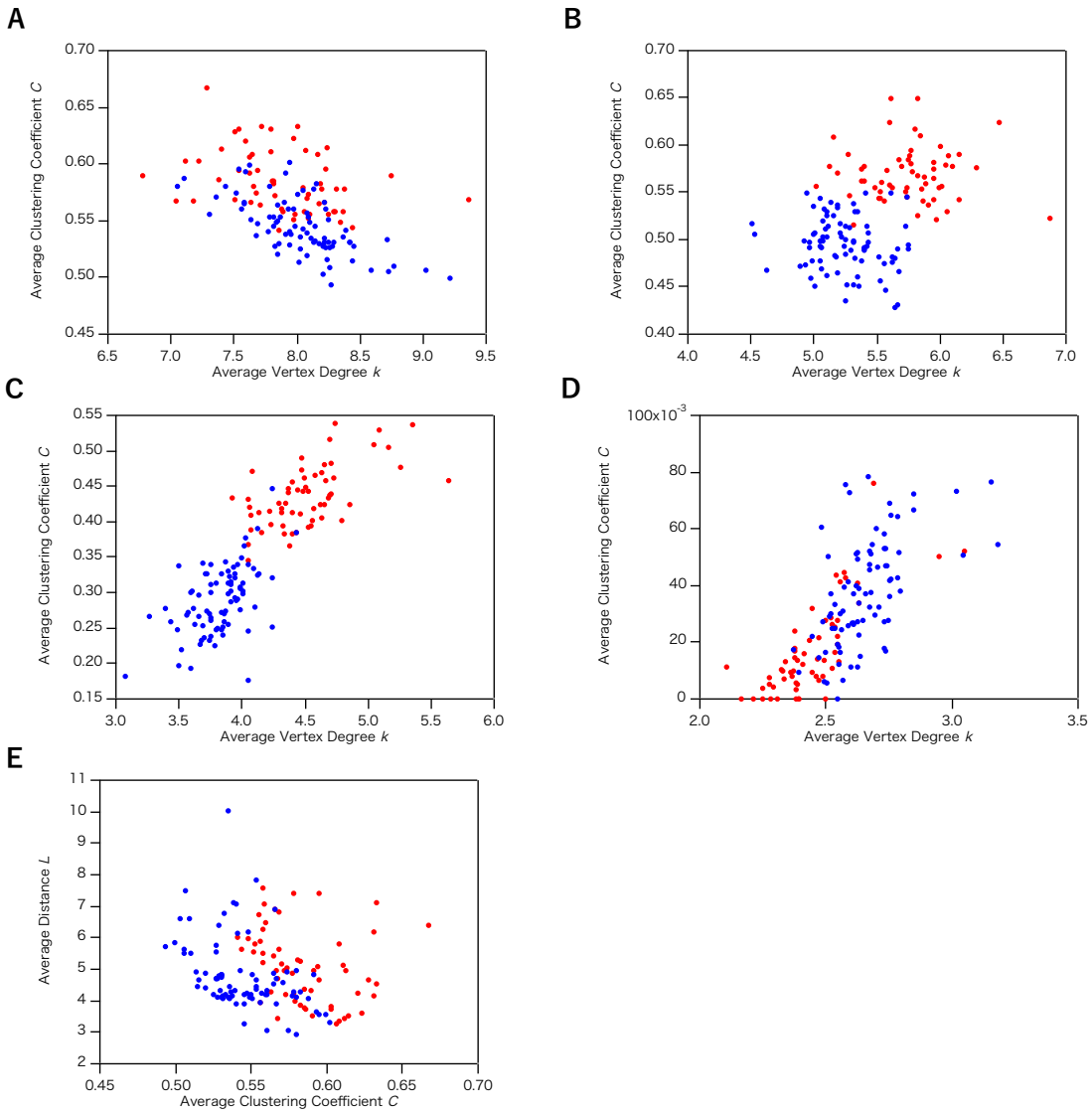

Figure 3.8. Average clustering coefficient C–average vertex degree k plots (A)–(D) and average distance L–C plots (E) of CAN whose distance of cutoff is (A), (E) 7.0 Å, (B) 6.0 Å, (C) 5.5 Å, and (D) 5.0 Å. All-α protein structures are represented as red circles and all-β protein structures are shown as blue circles.

3.5. Discussion

I discuss the results on the ISN and the CAN based on the all-α and all-β protein domains classified by SCOP classification. To ensure the validity of discrimination of all-α and all-β protein structures by the ISN, I confirmed that the discrimination is not depended on the data set. Because the distribution of *k*-*C* plot of all-α and all-β proteins appears to overlapped, the logistic regression was performed to Figure 3.3(C). The regression line is $C = 0.945 - 0.0622\ k$ as shown in Figure 3.9(A). Here, this data set is used as learning data set. Data set with larger number of protein structures was used for test data set. The number of protein structures and PDB ID of the test data set are shown in Tables 3.4 and 3.5, respectively. *k*-*C* plot of the ISN with the test data set is shown in Figure 3.9(B). Discriminant probability on all the data set using the discriminant line: all-α proteins are 50/52=0.96 and all-β proteins are 68/71=0.96 for the learning data set. All-α proteins are 195/208=0.9375 and all-β proteins are 454/461=0.985 for the test data set. The discrimination line determined by the learning data set discriminates between all-α and all-β proteins in the test data set. This demonstrates that the ISN analysis is not biased (above analysis was performed by Professor Takao Namiki, Hokkaido University).

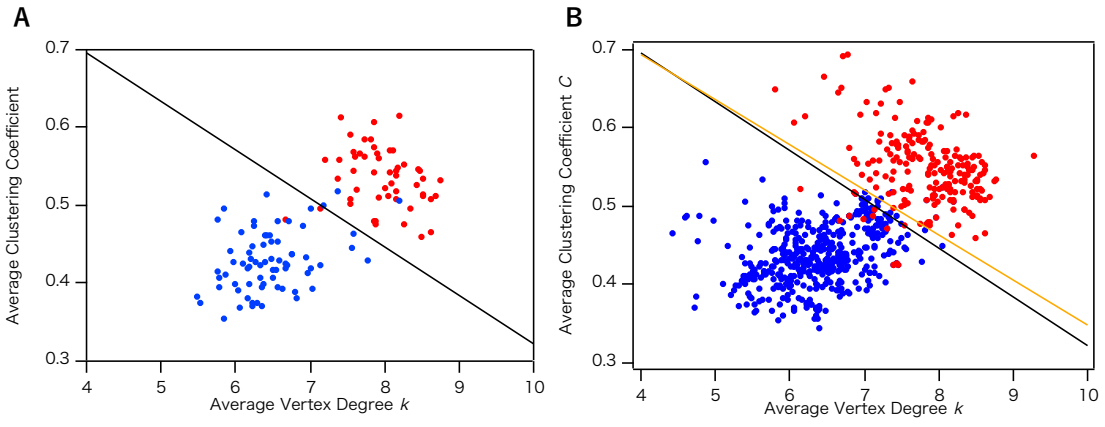

Figure 3.9. Average clustering coefficient C–average vertex degree k plots of all-α (red) and all-β (blue) proteins with discriminant line by logistic regression. (A) Learning data set is the data set as well as Figure 3.3(C) and the discriminant line is determined as $C = 0.945 - 0.0622\ k$. (B) Test data set and the discriminant lines. Black and orange lines are the discriminant lines determined by the learning data set and test data set ($C = 0.924 - 0.0576\ k$), respectively.

As clearly shown above, both the ISN and the CAN (R_c = 5.5 Å) were able to distinguish between all-α and all-β protein structures. However, the ISN and the CAN differ in terms of three points, as follows: the robustness, the number of links (N_{links}), and the distribution of the network parameters. Robustness refers to the capability of maintaining similar network geometry with a wide range of parameters. Here, 'robustness' is used as the capability of distinguishing all-α and all-β protein structures within a wide range of R_c values. Less robust networks are perturbed by small deviations of the network parameters, such as R_c. A notable difference between the CAN and the ISN resides in the comparison between the distributions of the k–C plots. The protein secondary structures were distinguished with a wide range of R_c from 3.4 Å to 5.0 Å in the ISN, while the CAN was able to distinguish between all-α and all-β protein structures only around R_c = 5.5 Å. The CAN lost clear discrimination when R_c was lower than 5.0 Å, where C became approximately 0 and k was between 2 and 3 (Figure 3.8D), showing that the CAN is less robust than the ISN in terms of R_c.

A comparison of N_{Links} between the CAN and the ISN also indicated a significant difference between the two amino acid networks (AAN). The k value of protein structures in the CAN (R_c = 5.5 Å) ranged from 3 to 5 (Figure 3.8C), thus exhibiting three to five links per vertex. This number of links per vertex corresponds to the number of interactions around Cα atoms. Among these links, two were derived from the links to the neighbor residues by covalent bonds. For the remaining one to three links,

most would be hydrogen bonds involved in the main chain atoms. As shown in Figure 3.10, if a hydrogen bond is established between two residues by their main chain atoms (orange dashed lines in Figure 3.10), their Cα atoms (green cycles in Figure 3.10) are also close to each other. These hydrogen bonds reflect the protein secondary structures and, therefore, most of the links in the CAN (R_c = 5.5 Å) would include interactions in main chain atoms. In the case of the ISN, five to nine links were exhibited per vertex (Figure 3.5A), implying an additional three to seven interactions besides the two covalent bonds. These additional three to seven links cannot be accounted for only by interactions involved in main chain atoms, indicating that some of the links are derived from interactions involved in side chain atoms. The links in the ISN, therefore, comprise interactions involved in *both* main and side chain atoms and the ISN includes structural information about the secondary and tertiary structures, corresponding to the interactions mediated by main and side chain atoms, respectively. This is in contrast to the CAN (R_c = 5.5 Å), wherein links only reflect the secondary structures from interactions in the main chain atoms.

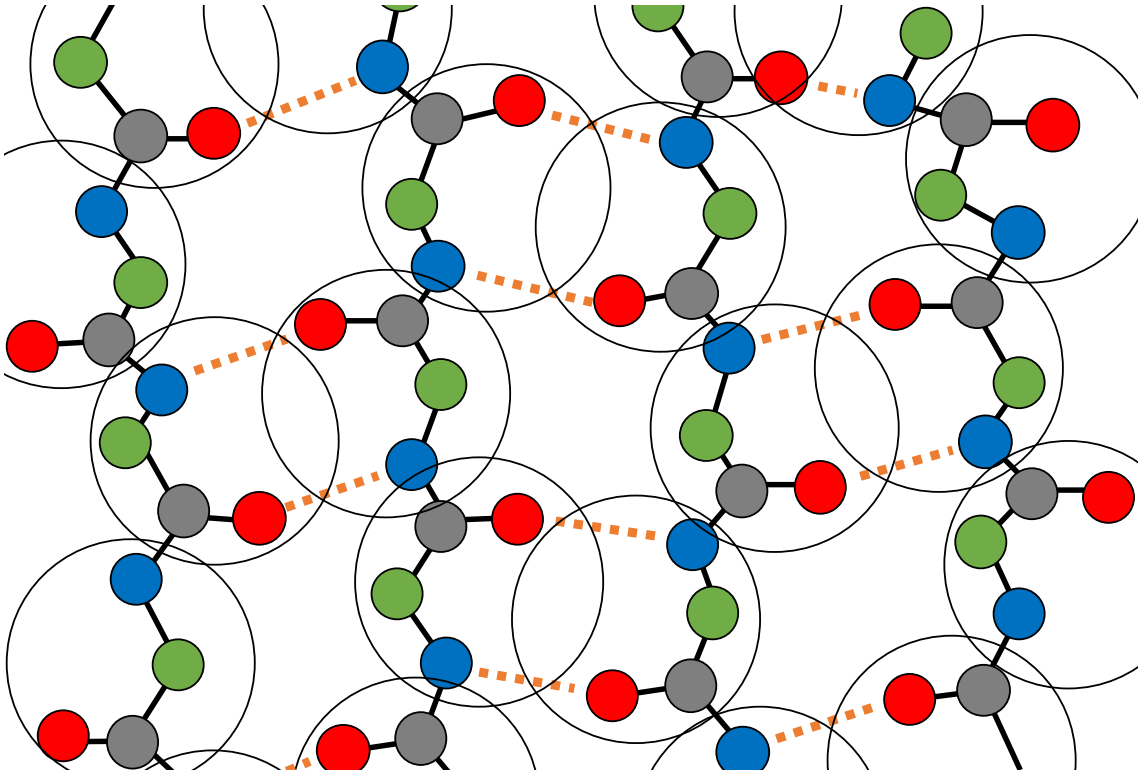

Figure 3.10. A schematic presentation of anti-parallel β-sheet structure in a protein. Filled circles correspond to atoms and their colors show the type of atoms; red (oxygen), blue (nitrogen), green (carbon alpha) and gray (other carbon). Black lines represent covalent bonds and orange dashed lines displays hydrogen bonds. Each amino acid residue is encircled. Hydrogen bonds composed of the protein secondary structure are close to Cα atoms.

It should be noted here that the distribution of network parameters in ISN, k and C, is markedly different from that of the CAN. In the CAN (R_c = 5.5 Å), the all-α and all-β protein structures indicate a linear, rather than a scattered, distribution (Figure 3.8C). For example, at $C \sim 0.25$, the range of k is 3.2 to 4.2. On the other hand, the distribution of all-α and the all-β protein structures are widely spread in the ISN (Figure 3.5A). In contrast, the CAN, when $C \sim 0.40$, has a range of k from 5.5 to 7.1, which is larger than that in the ISN, thus indicating that the CAN (R_c = 5.5 Å) defines protein structures in the same class to similar network geometry, whereas in ISN, protein structures in the same class can be assigned to different network geometry.

As discussed above, the CAN (R_c = 5.5 Å) is based on the information about protein secondary structures reflecting hydrogen bonds between main chain atoms. On the other hand, the ISN includes additional information about interactions involved in the side chain atoms. Since the side chain atoms often mediate van der Waals contacts and hydrophobic interactions between two residues located on separated positions on the primary sequence, thus reflecting the tertiary structures of proteins, the links in the ISN include the structural information from the tertiary structures of proteins. The scattered distribution in the *k*–*C* plots in the ISN, compared to that in the CAN, also suggests that, not only the structural information from the secondary elements, but also additional structural information from the tertiary structures are reflected in the distribution. The CAN (R_c = 5.5 Å), therefore, describes the protein 3D structures only according to the secondary structures, whereas the ISN characterizes protein structures based on both their secondary and their tertiary structures.

3.6. References

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

[2] Alm, E. & Baker, D. Matching theory and experiment in protein folding. *Curr. Opin. Struct. Biol.* **9**, 189-196 (1999).

[3] Matsumura, M., Signor, G. & Matthews, B.W. Substantial increase of protein stability by multiple disulphide bonds. *Nature* **342**, 291-293 (1989).

[4] Shoichet, B.K., Baase, W.A., Kuroki, R. & Matthews, B.W. A relationship between protein stability and protein function. *Proc. Natl. Acad. Sci. USA* **92**, 452-456 (1995).

[5] Yan, W., Zhou, J., Sun, M., Chen, J., Hu, G. & Shen, B. The construction of an amino acid network for understanding protein structure and function. *Amino Acids* **46**, 1419-1439 (2014).

[6] Csaba, G., Birzele, F. & Zimmer, R. Systematic comparison of SCOP and CATH: a new gold standard for protein structure analysis. *BMC Struct. Biol.* **9**, (2009).

[7] Murzin, A.G., Brenner, S.E., Hubbard, T. & Chothia, C. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536-540 (1995).

[8] Andreeva, A., Howorth, D., Chandonia, J.M., Brenner, S.E., Hubbard, T.J.P., Chothia, C.*, et al.* Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res.* **36**, D419-D425 (2008).

[9] Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B. & Thornton, J.M. CATH - a hierarchic classification of protein domain structures. *Structure* **5**, 1093-1108 (1997).

[10] Greene, L.H., Lewis, T.E., Addou, S., Cuff, A., Dallman, T., Dibley, M.*, et al.* The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. *Nucleic Acids Res.* **35**, D291-D297 (2007).

[11] Kabsch, W. & Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577-2637 (1983).

[12] Michie, A.D., Orengo, C.A. & Thornton, J.M. Analysis of domain structural class using an automated class assignment protocol. *J. Mol. Biol.* **262**, 168-185 (1996).

[13] Sung Joon, P.G., Chikenji; Takatsugu, Hirokawa; Kentaro, Tomii; Shoji, Takada Present and Future Prospects of Protein Structure Prediction. *Journal of JSAI* **20**, 479-485 (2005).

[14] Yoo, P.D., Zhou, B.B. & Zomaya, A.Y. Machine learning techniques for protein secondary structure prediction: An overview and evaluation. *Curr. Bioinform.* **3**, 74-86 (2008).

[15] Di Paola, L., De Ruvo, M., Paci, P., Santoni, D. & Giuliani, A. Protein Contact Networks: An Emerging Paradigm in Chemistry. *Chem. Rev.* **113**, 1598-1613 (2013).

[16] Rose, P.W., Prlic, A., Altunkaya, A., Bi, C.X., Bradley, A.R., Christie, C.H., *et al.* The RCSB protein data bank: integrative view of protein, gene and 3D structural information. *Nucleic Acids Res.* **45**, D271-D281 (2017).

[17] Alves, N.A. & Martinez, A.S. Inferring topological features of proteins from amino acid residue networks. *Physica A* **375**, 336-344 (2007).

[18] Bartoli, L., Fariselli, P. & Casadio, R. The effect of backbone on the small-world properties of protein contact maps. *Phys. Biol.* **4**, L1-L5 (2007).

[19] Vendruscolo, M., Dokholyan, N.V., Paci, E. & Karplus, M. Small-world view of the amino acids that play a key role in protein folding. *Phys. Rev. E* **65**, (2002).

[20] Bagler, G. & Sinha, S. Network properties of protein structures. *Physica A* **346**, 27-33 (2005).

[21] Greene, L.H. & Higman, V.A. Uncovering network systems within protein structures. *J. Mol. Biol.* **334**, 781-791 (2003).

[22] Aftabuddin, M. & Kundu, S. Hydrophobic, hydrophilic, and charged amino acid networks within protein. *Biophys. J.* **93**, 225-231 (2007).

[23] Aftabuddin, M. & Kundu, S. Weighted and unweighted network of amino acids within protein. *Physica A* **369**, 895-904 (2006).

[24] Tina, K.G., Bhadra, R. & Srinivasan, N. PIC: Protein Interactions Calculator. *Nucleic Acids Res.* **35**, W473-W476 (2007).

[25] Overington, J., Johnson, M.S., Sali, A. & Blundell, T.L. Tertiary structural constraints on protein evolutionary diversity: templates, key residues and structure prediction. *Proc. Royal Soc. Lond.* **241**, 132-145 (1990).

[26] Baker, E.N. & Hubbard, R.E. Hydrogen bonding in globular proteins. *Prog. Biophys. Mol. Biol.* **44**, 97-179 (1984).

[27] Kyte, J. & Doolittle, R.F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**, 105-132 (1982).

[28] Bateman, A., Martin, M.J., O'Donovan, C., Magrane, M., Apweiler, R., Alpi, E., *et al.* UniProt: a hub for protein information. *Nucleic Acids Res.* **43**, D204-D212 (2015).

[29] Ko, T.P., Robinson, H., Gao, Y.G., Cheng, C.H.C., DeVries, A.L. & Wang, A.H.J.

The refined crystal structure of an eel pout type III antifreeze protein RD1 at 0.62-A resolution reveals structural microheterogeneity of protein and solvation. *Biophys. J.* **84**, 1228-1237 (2003).

Chapter IV
Conclusion

In the present thesis, I aimed to explore the significance of the hydrophobic interaction in the intramolecular interaction network in the protein structure. The hydrophobic interaction is viewed as the principal force that drives the protein toward globular collapse and engenders a solvent-shielded molecular interior [1-3]. However, it remains difficult to capture the comprehensive description of the hydrophobic interaction in the protein folding [4]. To acquire the detailed description of the hydrophobic interaction, I focused on the dehydration accompanied by the formation of the hydrophobic interaction in the protein folding. By performing the high-pressure absorption spectroscopy, the amount of the hydration of the hydrophobic heme group in cytochrome *c* (Cyt *c*) unfolding were estimated as the volume chnage. The analysis demonstrated that the entropic contribution of the dehydration from the heme group in the Cyt *c* folding is correspondingly important for stabilizing the protein structure. In addition, to examine the relationship between the intramolecular interaction network and the global conformation of a protein, a novel amino acid network (AAN) model, interaction selective network (ISN) was established. The ISN reflects intramolecular interactions between the amino acid residues. The results and discussion in this thesis, and future perspectives are highlighted below.

4.1 Uncovering dehydration in cytochrome *c* refolding from urea- and guanidine hydrochloride-denatured unfolded state by high pressure spectroscopy (Chapter II)

To elucidate the entropic contribution of the dehydration from the hydrophobic groups in the protein folding, the high-pressure absorption spectroscopy measurements were performed. I successfully determined the partial volume changes for the unfolding of Cyt *c* induced by the denaturant and experimentally confirmed that the hydration to the hydrophobic groups, the heme, in the Cyt *c* unfolding. The hydration to the heme is the primary factor to determine the volume change induced by the hydration ($\Delta\Delta V_h$), which is also a thermodynamic determinant for the protein folding in Cyt *c*. This suggests that the entropic contribution of the dehydration from the hydrophobic groups is critical for stabilizing protein structures.

4.2 Quantitative description and classification of protein structures by a novel robust amino acid network (Chapter III)

To elucidate the relationship between the intramolecular interaction network and the global conformation of a protein, I developed a novel amino acid network (AAN) model, interaction selective network (ISN). The ISN reflects hydrophobic interaction network in a protein. While previous AAN models cannot distinguish between protein secondary structures, the ISN allows to discriminate between α-helix and β-sheet protein structures by using two network parameters, the average degree (k) and the average clustering coefficient (C). By exploring the optimum cutoff ditance (R_c) for the discrimination of protein structures, this study demonstrated that the ISN is more

robust than the conventional AANs, and the wider range of the distribution of k and C in the ISN suggests that the ISN has additional structural information about the protein tertiary structures. Therefore, the ISN provides a more quantitative and robust description of the protein structures, by reflecting *both* the secondary and the tertiary protein structures. The results suggest that the ISN can be applied to elucidating the relationship between the intramolecular interaction network and the global conformation of a protein.

4.3 Conclusion remarks

As summarized above, I attempted to explore the significance of the hydrophobic interaction in the intramolecular interaction network in the protein structure. The present thesis suggested that the dehydration from the hydrophobic groups accompanied by the formation of the hydrophobic interaction is critical for the protein folding. Moreover, the ISN which reflects hydrophobic interaction network in a protein, is established. The ISN can be used for elucidating the relationship between the intramolecular interaction network and the global conformation of a protein.

Further studies should focus on elucidating this relationship by both experimental and theoretical approaches. For the experimental approach, the site-specific amino acid substitutions induce perturbations of the intramolecular interaction network in a protein. Comparison of $\Delta\Delta V_{\mathrm{h}}$ with various type of mutants expect to uncover the entropic contribution of specific residues to stabilizing global conformation of a protein. For the theoretical approach, comparing proteins with different $\Delta\Delta V_{\mathrm{h}}$ by using the ISN expects to enable us to evaluate the significance of individual interaction

for the global conformation of a protein. Combining the experimental and the theoretical approaches expects to establish a novel quantitative model to assess the structural significance of the intramolecular interaction network in a protein, which contributes to uncovering the principal of stabilizing the protein three-dimensional structure.

References

[1] Kauzmann, W. Some Factors in the Interpretation of Protein Denaturation1. in *Adv. Protein Chem.* (C.B. Anfinsen, M.L.A.K.B. & John, T.E. eds.) vol. 14, pp. 1-63 Academic Press, (1959).

[2] England, J.L. & Haran, G. Role of Solvation Effects in Protein Denaturation: From Thermodynamics to Single Molecules and Back. in *Annu. Rev. Phys. Chem.* (Leone, S.R., Cremer, P.S., Groves, J.T., & Johnson, M.A. eds.) vol. 62, pp. 257-277, (2011).

[3] Nozaki, Y. & Tanford, C. The solubility of amino acids and two glycine peptides in aqueous ethanol and dioxane solutions. Establishment of a hydrophobicity scale. *J. Biol. Chem.* **246**, 2211-2217 (1971).

[4] Pace, C.N., Fu, H.L., Fryar, K.L., Landua, J., Trevino, S.R., Shirley, B.A.*, et al.* Contribution of Hydrophobic Interactions to Protein Stability. *J. Mol. Biol.* **408**, 514-528 (2011).