Doctoral Dissertation

Prediction of the folding processes for immunoglobulin-like beta-sandwich and beta-trefoil proteins

September 2019

Doctoral Program in Advanced Life Sciences

Graduate School of Life Sciences

Ritsumeikan University

AUMPUCHIN Panyavut

Doctoral Dissertation Reviewed
by Ritsumeikan University

Prediction of the folding processes for immunoglobulin-like beta-sandwich and beta-trefoil proteins

(免疫グロブリン様ベータ － サンドイッチタンパク質およびベータ － トレフォイルタンパク質のフォールディング過程の予測)

September 2019

2019 年 9 月

Doctoral Program in Advanced Life Sciences

Graduate School of Life Sciences

Ritsumeikan University

立命館大学大学院生命科学研究科

生命科学専攻博士課程後期課程

AUMPUCHIN Panyavut

アウンプチン パンヤブット

Supervisor: Professor KIKUCHI Takeshi

研究指導教員：菊地 武司 教授

Holistic Abstract

Describing the folding mechanism of a protein is one of the goals in molecular bioinformatics study. Nowadays, it is enigmatic and difficult to extract folding information using standard bioinformatics techniques or even experimental protocols which can be time consuming. However, protein folding mechanisms have been intensively investigated with experimental as well as simulation techniques. Since a protein folds into its specific 3D-structure from a unique amino acid sequence, it is interesting to extract as much information as possible from the amino acid sequence of a protein. The mainly purpose of this thesis is to extract the folding mechanisms of variety target proteins, especially the structure formation in an initial state of folding, from its amino acid sequence by means of inter-residue average distance statistics methods Average Distance Map (ADM) and contact frequency analysis (F-value). Moreover, a coarse-grained Gō-model simulation was applied for some proteins to simulate the whole story of protein folding. The proteins used in this thesis are immunoglobulin-like beta-sandwich fold proteins and β-trefoil fold proteins.

Ig-like beta-sandwich fold protein is one of the most common structural families that play key roles in the immune system. In this study, the PDB codes 1TIT and 1TEN are used to represent the Ig domain and FN3 domain, respectively. The β strand 3 and 5 are suggested to be the significant strands for the initial folding processes in both proteins. This common feature coincides well with the experimental results and underscores the significance of the beta-sandwich proteins' common structure, namely, the key strands for folding and the Greek-key motif, which is located in the central region. The results show good correspondence to experimental data, ϕ-value and protection factor from H-D exchange experiments. The significance of conserved hydrophobic residues near F-value peaks for structural stability by using hydrophobic packing is confirmed. It is worth noting that the sequence-based techniques were able to predict the initial folding event just next to the denatured state. The long-range contacts formed in an early event of folding derived from Gō-model simulation also corresponded well to the experimental data and the conserved hydrophobic residues near the peaks of F-value.

The β-trefoil fold proteins are known to have the same 3D scaffold, namely, a three-fold symmetric scaffold, despite the proteins' low sequence identity among superfamilies. In this study, 26 high symmetric β-trefoil proteins and four irregular structure β-trefoil proteins from six superfamilies were analyzed by the same manner as Ig-like beta-sandwich fold protein. The ADM and F-value analyses predict that the N-terminal and C-terminal regions are conserved to form a compact region in an initial state of folding and that the hydrophobic residues at the central region can be regarded as an interaction center with other residues. These results coincide well with the experimental data obtained so far for folding of some of the β-trefoil proteins. In the case of irregular β-trefoil proteins which an experimental data has not been available so far, the 3D-based Gō-model simulations were performed to investigate the protein folding processes. It is interesting that the conserved hydrophobic residues near the peaks of F-value plot also observed to form the long-range interaction in the contact frequency map derived from a Gō-model simulation. Moreover, the results indicate that the irregular parts in the β-trefoil proteins do not hinder the protein formation. Conserved hydrophobic residues on the β5

strand are always the interaction center of packing between the conserved hydrophobic residues in both regular and irregular β-trefoil proteins. Suggesting that β strand 5 plays an important role in β-trefoil protein structure construction.

In summary, it can be confirmed that the average distance statistics methods, ADM and F-value analysis, were able to predict the initial folding event just next to the denatured state by using only amino acid sequence information and that a 3D-based Gō-model simulation can be used to investigate the whole process of protein folding.

Contents

Chapter 1

General Introduction

1.1Protein folding

Proteins are the end products of the decoding process that starts from cellular DNA that participate in all cellular process of living organisms, such as the protein collagen, which provides the structural support of our connective tissues. While, other proteins act as antibodies in immune responses, or perform mechanical work in our muscles, etc.[1] Protein folding is a process of polypeptide chain folds into their biologically active protein in its native structure. During translation processes, each protein is synthesized as a chain of amino acids or random coil which does not have a stable conformation. The chain folds into unique and elaborate 3D-structures with interactions that stabilize their structure to form a well-defined, folded protein. The correct final folded native structure of proteins are key to their functions.[2] Misfolding can lead to inactive or toxic proteins, protein aggregates that cause severe diseases, such as Alzheimer's or Parkinson's disease.[3]

To study the protein folding, the native structure of each protein is the most important one that requires as an input data in various methods, but the problems are that techniques need high proficiency and wherewithal to decode folding information and that not every protein is easy to express and purify *in vitro*. Even though, the native folded molecules are complex but are known in great detail thanks to NMR (Nuclear Magnetic Resonance) or X-ray crystallography.

The folding of a protein is a complex process, including four stages that lead to various specific functional protein structure. The multiple conformations of active protein structure are hinge on the amino acid sequence compositions.

The first stage called "Primary structure" refers to the linear of amino acid residues encoded from a nucleotides chain in DNA. This stage produces a single poly peptide chain with unstable structure, linear or random coil. Note that the post-translational modification such as phosphorylation and glycosylation are observed in a part of primary structure stage which can promote protein folding and stabilize the functional protein structure.[4] The second stage is "Secondary structure" which generated by formation of hydrogen bonds between backbone's atoms in the chain to construct the secondary structures, beta-sheets or alpha helices. The third stage is "Tertiary structure". The tertiary structure is constructed by the interactions and bonding of the amino acid side chains in the protein, that is, the folding of secondary structure helices or sheets into one another. Lastly, "Quaternary structure" refers to a functional protein which build by the interactions between different tertiary structure such as hemoglobin.[5] However, precisely proteins fold into their specific 3D-structure remains a fascinated question. It is well known that proteins can fold into two type of kinetic pathways, two-state folding and

multi-state folding. It is interesting that chain length is sufficient to classify the folding process type, that is, protein folds with two-state kinetics, if its length is smaller than 112 residues.[6]

1.2Protein classification

Due to structural similarities are most likely to indicate an evolutionary and/or functional similarity when sequence similarity is absent, protein folds are described in terms of the type and secondary structures arrangement. The protein classification can provide functional details through comparison to others[7]. Nowadays, 12 classes of protein structure are classified as published in structure classification databases, SCOPe 2.07,[8] such as all alpha proteins, all beta proteins, multi-domain proteins, and so on.

1.3Scope of the thesis

The aim of this study is to analyze the folding processes of the target protein, especially in an initial state of folding, with a focus on the conservation properties along related domains, which are classified into different superfamilies on the basis of sequence similarity but share the same fold. The methods used in this study include 3D-structure-based multiple sequence alignment (3D-MSA), sequence-based multiple sequence alignment (Seq-MSA), inter-residue average distance statistic-based methods (ADM and F value), and a coarse-grained Cα Gō-model simulation. The comparative results between 3D-structure-based and sequence-based techniques are also discussed in this study. It should be noted that the autonomous folding of a protein is also considered and the result of simulation will be compared with an available experimental data.

Chapter 2

Investigation Methods

2.1 Inter-residue average distance statistics methods

Due to the fact that each protein folds into its native structure by a folding mechanism with a specific amino acid sequence. Then, the relation between amino acid sequence and protein folding mechanism is a main goal of molecular bioinformatics approaches. These present prediction methods successfully extract the folding properties that well to the data from experimental analyses of following proteins: fatty acid binding proteins,[9] globin-like fold proteins,[10] IgG binding and albumin binding domains,[11] Ig-like fold proteins,[12,13] ferredoxin-like fold proteins,[14] beta-trefoil fold proteins,[15] and lysozyme-like superfamily proteins.[16]

2.1.1 Average Distance Map analysis (ADM)

Average distance map analysis method

The contact map based on inter-residue average distance statistics of 42 known protein structures as created and described in Kikuchi et al. (1988)[17] and Ichimaru and Kikuchi (2003)[9] was used in this study. The inter-residue average distances and standard deviations were calculated based on residue types and sequence separation. The details of this method are present by the following sections and Figure 1 and 2.

Grouping the residue pairs into ranges based on the distance between residues along the sequence and calculation of the average distance between residue types in each range.

Inter-residue average distances were calculated in each range along the amino acid sequence. When i and j are refer to the residue number along the sequence, a length is defined as the distance between residue *i* and *j* (k|*i*,*j*|). Then, the range is defined as M=1 when the distance between two residues along the amino acid chain are between 1-8 residues, and $9 \leq k \leq 20, 21 \leq k \leq 30,\ 31 \leq k \leq 40, 41 \leq k \leq 50$, and so on define ranges M = 2, 3, 4, …, respectively. An average distance, d(A, B, M), where A and B are refer to amino acid type, the distance separating residue pairs in each range M was calculated. Next the contact map of an unknown 3D protein structure was constructed, if

the average distances of the interested residue pairs are lower than determined cutoff value ($d(A, B, M) \leq d_c(A, B, M)$), then pairs will plot on the contact map. The cutoff value was descripted in the following section.

Definition of cutoff distance for ADM construction

The determination of cutoff distance in each range for the ADM construction of a given amino acid sequence is defined as follows. First, the contact density of the map must be considered. The contact density of ADM should be close to the density of real distance map (RDM), where the RDM represents the contact map constructed based on the actual 3D-structure of a protein when the inter-residue Cα atomic distance is less than 15 Å, which can be calculated by the formula (1)

$$\rho_{av} = C/N \tag{1}$$

Here ρ_{av} is the average contact density of the entire of the map, N is the number of residues, and C is an adjustable constant. In this study, C = 36.12 is used to reproduce a value of ρ_{av} under a 15 Å cutoff distance of RDM based on an earlier study by Kikuchi et al. (1988).[17] The number of average inter-residue Cα atomic contact distances that are less than 15 Å for each residue pair was investigated and ranked within each range. The cutoff distance for each range M is used to construct the ADM, which can be defined in the following equation.

$$P(M)_C = (\frac{D}{M})P(M)_t \tag{2}$$

$P(M)_C$ is the order of the average inter-residue Cα atomic contact distances in each range M, $P(M)_t$ is the constant number of pairs with statistically significant occurrence in range M, the residue pairs that occur less than 100 times in the data base of 42 proteins were considered as statistically not significant. D is an adjustable parameter that gives the average contact density of ADM close to the RDM.[17] A set of cutoff distance is determined to reproduce a different ratio of the number of plots to the whole area of the constructed contact map.

Construction of ADM with amino acid sequence of a protein with an unknown structure

The contact map based on average distance statistics is constructed in a similar

way as the usual 3D-structure-based coordinate contact map. Point on the map indicates a pair of residues at an average distance of separation less than the statistical cutoff value. The cutoff distance set was chosen for each protein to approximate plot density of RDM as calculated from equation (1).

Location of compact region

An area with a high density of points on a map is regarded. The limit of such a region on a map can be detected as sudden change in the density of points. When the contact map is considered in both axes, horizontal and vertical, the density difference for each residue between the triangular part and trapezoidal part, can be estimated from the following equations. (See Figure 3A and 3B) A peak and valley in the histogram of the values of $\Delta\rho_i^v$ show the boundary of the high contact region as displayed in Figure 1C along the vertical axis. In the same way, the boundary of the compact region can be detected by peak and valley in the histogram of $\Delta\rho_j^h$ as displayed in Figure 3C

$$\Delta\rho_i^v = \rho_i^v - \tilde{\rho}_i^v \tag{3}$$

$$\Delta\rho_j^h = \rho_j^h - \tilde{\rho}_j^h \tag{4}$$

Here ρ and $\tilde{\rho}$ are the contact densities of the triangular and trapezoidal regions, and h and v represent the residue measured from the horizontal and vertical axes, respectively.

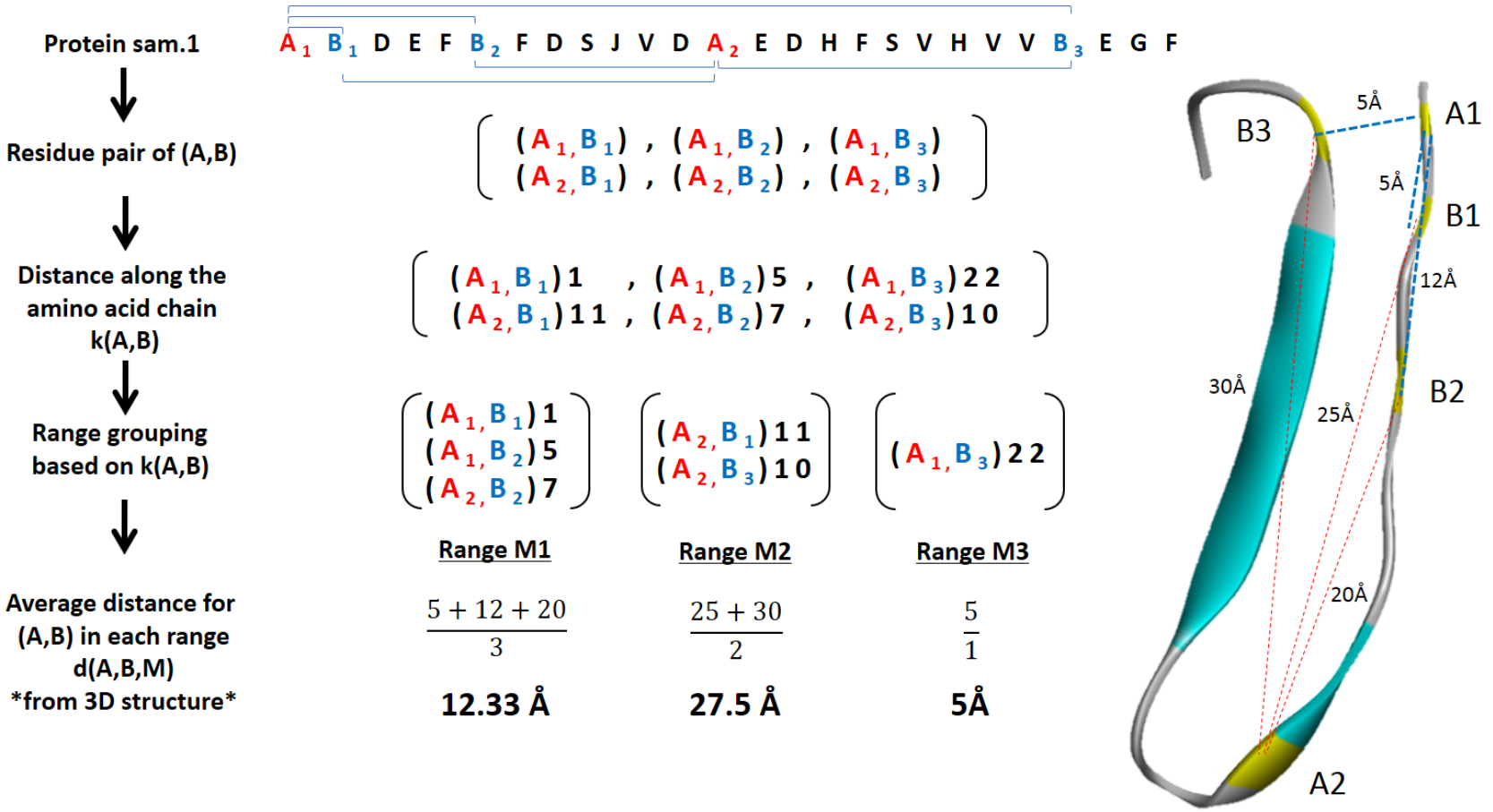


Figure 1 Grouping the residue pairs into a range. The upper-case alphabets (A, B, C, …, J) represent different types of amino acid residues. Part of protein structure on the right-hand side shows the position of residues A and B along the amino acid chain with the distance between each residue pair.

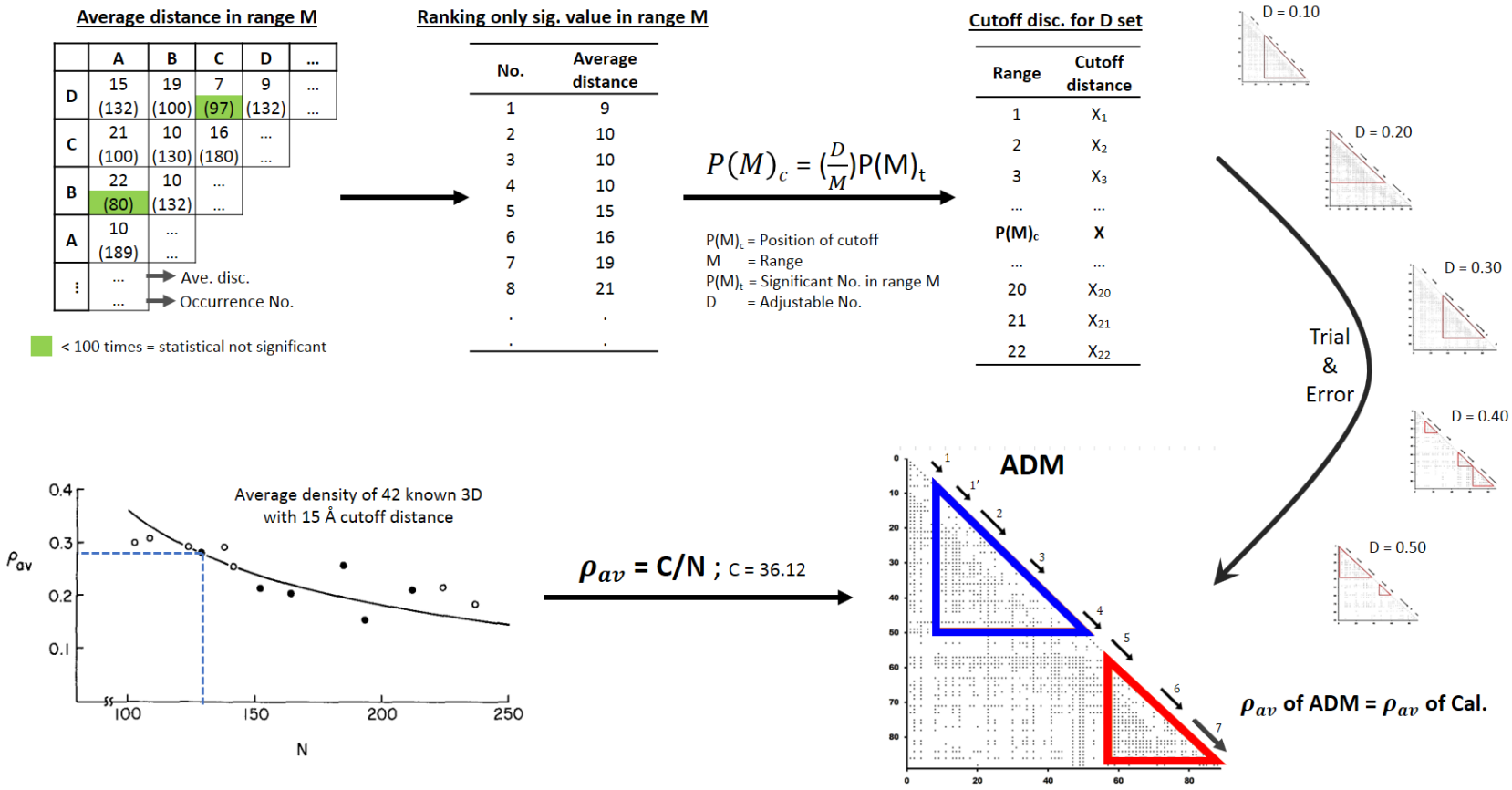


Figure 2 The cutoff distance calculation method. The upper part shows methods to determine the cutoff value from 42 known 3D-structure proteins. Small average distance maps on the right side were constructed from the determined cutoff value of each D set. The lower part represents the criteria to determine the ADM based on the average density calculated from the standard curve equation.

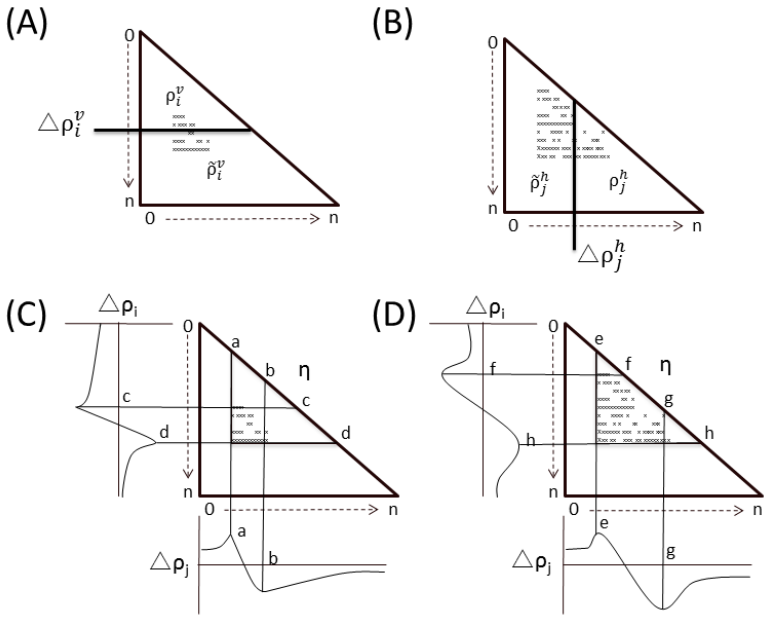


Figure 3 The map is divided by a line parallel to the x or y- axis to calculate the density difference at each residue. A peak and a valley appear at the boundaries of a high density region of the plots.

Thus, in Figure 3C, the contacts between segment a-b and c-d were detected in a given protein. The boundary of the compact region refers to the region where high density of contact plots along the diagonal of a contact map. As show in Figure 3D, the compact region indicated by lines through two peaks (e and h) that is the position where contact density suddenly changes. The summation of the differences in densities defined by lines parallel to the y-axis and the x-axis is called the h-value, and this value is considered to be a measure of the degree of compactness for the region. ($\eta = \Delta\rho_e^h + \Delta\rho_h^v$ in the example of Figure 3D). A region called "predicted compact region" (PdCR) for a given protein is regarded to be compact in the native structure of a protein and important in an early stage of folding. The only relative value is considered to design the location of folding unit. Such regions correspond to experimental analyzes of protein folding mechanisms[9-12,14-16].

2.1.2 Contact frequency analysis (F-value)

The contact frequency of each pair of residues in the random ensemble state of a protein is estimated. The potential energy used in this computational method derived from the average distance statistics of 42 known structural proteins as in ADM analysis. The frequency is used to determine the location where initial folding events occur, such as hydrophobic collapse. In other words, such a site in a random state ensemble is a possible folding initiation site. The potential is a harmonic potential to reproduce average distances and standard deviations in the statistics. The protein conformations were

simulated by Metropolis Monte Carlo simulation with potential energy derived from average distance statistic methods as used in ADMs. If assume the probability density with the potential energy between residue pair, $\mathbf{P}(\boldsymbol{\varepsilon}_{i,j})$, is equivalent to the standard Gaussian distribution calculated with its average distance and standard deviation, $\boldsymbol{\rho}(\bar{\boldsymbol{r}}_{i,j}, \boldsymbol{\sigma}_{i,j})$, that is,

$$\boldsymbol{P}(\boldsymbol{\varepsilon}_{i,j}) = \boldsymbol{\rho}(\bar{\boldsymbol{r}}_{i,j}, \boldsymbol{\sigma}_{i,j}), \tag{5}$$

where kT refers to acceptance ratio which set at 0.5, and $\bar{\boldsymbol{r}}_{i,j}$ refers to the distance between C$\boldsymbol{\alpha}$ atoms of the residues *i* and *j* and $\boldsymbol{\sigma}_{i,j}$ is the standard deviation. The potential energy $\varepsilon_{i,j}$ is expressed by average distance $\bar{r}_{i,j}$ and its standard deviation $\sigma_{i,j}$ as in equations (6) and (7),

$$\frac{\boldsymbol{exp}\left(-\frac{\boldsymbol{\varepsilon}_{i,j}}{\boldsymbol{kT}}\right)}{\boldsymbol{Z}} = \frac{\mathbf{1}}{\sqrt{\mathbf{2}\boldsymbol{\pi}}\boldsymbol{\sigma}_{i,j}} \boldsymbol{exp}\left\{-\frac{\left(\boldsymbol{r}_{i,j} - \bar{\boldsymbol{r}}_{i,j}\right)^{\mathbf{2}}}{\mathbf{2}\boldsymbol{\sigma}_{i,j}^{\mathbf{2}}}\right\} \tag{6}$$

$$\frac{\boldsymbol{\varepsilon}_{i,j}}{\boldsymbol{kT}} = \frac{\left(\boldsymbol{r}_{i,j} - \bar{\boldsymbol{r}}_{i,j}\right)^{\mathbf{2}}}{\mathbf{2}\boldsymbol{\sigma}_{i,j}^{\mathbf{2}}} - \mathbf{ln}\frac{\boldsymbol{Z}}{\sqrt{\mathbf{2}\boldsymbol{\pi}}\boldsymbol{\sigma}_{i,j}} \tag{7}$$

Z is the partition function and a significant value in a calculation is the difference between the energy values of conformations. Z does not appear in the calculation explicitly. Thus, Z is ignored in the calculations.

The contact frequency of each residue pair, g(*i*,*j*), in sampled conformations of a random state ensemble is calculated, and the value corresponding to the z-value in statistical theory is used. Here, D(M) is the standard deviation of the contact frequency g(*i*,*j*) with two residues separated by M residues (range M). These are expressed by equations (8) and (9).

$$\boldsymbol{D}(\boldsymbol{M}) = \sqrt{\frac{\sum_{|\boldsymbol{i}-\boldsymbol{j}|\in\boldsymbol{M}}\left(\frac{\sum_{|\boldsymbol{i}-\boldsymbol{j}|\in\boldsymbol{M}}\boldsymbol{g}(\boldsymbol{i},\boldsymbol{j})}{\sum_{|\boldsymbol{i}-\boldsymbol{j}|\in\boldsymbol{M}}} - \boldsymbol{g}(\boldsymbol{i},\boldsymbol{j})_{|\boldsymbol{i}-\boldsymbol{j}|\in\boldsymbol{M}}\right)^{\mathbf{2}}}{\sum_{|\boldsymbol{i}-\boldsymbol{j}|\in\boldsymbol{M}}}} \tag{8}$$

$$\mathbf{Q}(\boldsymbol{i},\boldsymbol{j}) = \frac{\boldsymbol{g}(\boldsymbol{i},\boldsymbol{j})_{|\boldsymbol{i}-\boldsymbol{j}|\in\boldsymbol{M}} - \frac{\sum_{|\boldsymbol{i}-\boldsymbol{j}|\in\boldsymbol{M}}\boldsymbol{g}(\boldsymbol{i},\boldsymbol{j})}{\sum_{|\boldsymbol{i}-\boldsymbol{j}|\in\boldsymbol{M}}}}{\boldsymbol{D}(\boldsymbol{M})} \tag{9}$$

$\boldsymbol{i}$ or $\boldsymbol{j}$ is the residue number. The summation of Q(*i*,*j*), normalized contact frequency, from *j* = 1 to N gives the relative contact frequency for residue *i* ($\boldsymbol{F_i}$) (N is the total number of residues).

$$\boldsymbol{F_i} = \sum_{\boldsymbol{j}} \boldsymbol{Q(i,j)} \tag{10}$$

The sum in eq. (10) is called the F-value. Residues around peaks in the plot of F-values are considered to be located in the center of many inter-residue contacts, such as a hydrophobic cluster, and thus a region near a peak of an F-value plot is assumed to be important for folding, especially in its initial stage. One of the major driving forces to construct the native structure is the hydrophobic interaction resulting in the burial of hydrophobic core residues.[18]

For the sampling of the conformations of the random state ensemble, Metropolis Monte Carlo simulations are conducted in this study using the simple Cα bead model with a bond length of 3.8 Å as a model of a protein. During a Monte Carlo simulation, the bond angle ($\boldsymbol{\theta}$) and the dihedral angle ($\boldsymbol{\phi}$) between residues *i* and *i*+1 are bent and rotated randomly followed by Metropolis judgment to decide on the acceptability of the new conformation. Every bond and dihedral angle of each residue *i*=1 … N-1 are varied in this way. Hundred Monte Carlo iterations were performed. One iteration included 60000 Monte Carlo steps. Then, an average of the F-values for the residue *i* was calculated. (The sampling of conformations was carried out from the very beginning.) In this study, a sequence of 20 glycine residues is flanked at both the N- and C-termini to avoid too high fluctuations at the ends as shown in previous studies.[12,14,15,19]

A peak is regarded as a significant site for an initial folding event. Thus, the definition of a peak is important. Peak and adjacent valleys are used to define a "real" peak when the difference in the values of adjacent valleys and a peak are more than the following cut-off value ($\boldsymbol{F_{cut}}$) as shown in equation 11. The relevant region of the F-value plot, peak and valley which have 1 residue distantly different, are unconcerned in the real peak judgments procedure. Moreover, a real peak was determined when the difference value of a peak and both side valleys are greater than an $\boldsymbol{F_{cut}}$ value.

$$\boldsymbol{F_{cut}} = \left[\frac{\boldsymbol{1}}{\boldsymbol{N-1}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{N-1}} (\boldsymbol{F_{i+1}} - \boldsymbol{F_i})^{\boldsymbol{2}}\right]^{\frac{\boldsymbol{1}}{\boldsymbol{2}}} \tag{11}$$

It has been confirmed that a hydrophobic residue within $\boldsymbol{\pm 5}$ residues of the F-value peak tends to form hydrophobic packing in the native structure of a protein.[16]

2.2 Evolution analysis methods

Multiple sequence alignment provides valuable information about many protein characteristics. This method is used to detect conserved regions which might be related to 3D-structure or functional property. Conserved positions are those whose residue type is conserved for all the aligned sequence's positions. Sequence conservation is a consequence of evolutionary pressure to maintain a given structure and/or their functions.[3,20] In this study, different alignment methods were used to investigate the sample data, and the conserved hydrophobic residues also considered. The following hydrophobic residues are regard in this study: Ala, Phe, Ile, Leu, Met, Val, Tyr and Trp. The hydrophobic positions are considered because their important role in the formation of a protein's core structure.[20]

In this study, the conservation during evolution was analyzed by using a multiple sequence alignment method with or without information on secondary structure. A multiple sequence alignment was created based on 3D-structures from the PDB database (https://www.rcsb.org/)[21] by using the Combinatorial Extension[22] program integrated within STRAP software[23] to determine the common regions along the query sequences of known protein structures. Due to published 3D-structure being few in number, the multiple sequence alignment based only on amino acid sequence from the Uniprot database (http://www.uniprot.org/)[24] were performed. Next, the ClustalW program integrated within MEGA version 7.0[25] was used to align obtained sequences to extract the conservation of predicted regions. A molecular phylogenetic tree was constructed by using the Neighbor-joining method.[26] The evolutionary distances were computed using the JTT matrix-based method.[27] The conserved hydrophobic residues are regarded as significant when more than 90% of residues are aligned. On the other hand, the conserved position of compact region derived from ADMs also indicated when the conservation ratio of compact regions exceed 70%.[16]

The hydrophobic packing is used to confirm an ability to form contacts with other residues in the native structure. The definition of hydrophobic packing is based on the distance between the heavy atoms of different hydrophobic residues. That is, when the distance of two heavy atoms in the native structure less than 5Å, these residues are regarded as forming hydrophobic packing.

2.3 Coarse-grained Gō model simulations

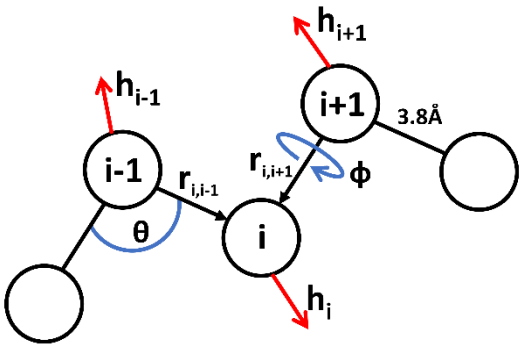


Figure 4 The Cα–bead model used in this coarse-grained Cα Gō model simulation with varying bond angle and torsional angle. The length between the Cα of each residue is specified as 3.8 Å. The vector h_i is defined by the vector $r_{i,i-1}+ r_{i,i+1}$.[28]

The protein folding characteristics of all selected domains were elucidated by means of a coarse-grained Cα Gō model that was developed in previous studies.[29,30] The following equation describes the potential energy of a sampled structure, Γ, during a simulation.

$$\begin{aligned}\mathbf{E}(\boldsymbol{\Gamma},\boldsymbol{\Gamma_0}) = {} & \boldsymbol{\Sigma_{angles}} \boldsymbol{K_\theta}(\boldsymbol{\theta_i} - \boldsymbol{\theta_{i0}})^2 \\ & + \boldsymbol{\Sigma_{dihedral}} \left\{ \boldsymbol{K_\phi^1}[-\boldsymbol{cos}(\phi_i - \phi_{i0})] + \boldsymbol{K_\phi^3}[-\boldsymbol{cos3}(\phi_i - \phi_{i0})] \right\} \\ & + \boldsymbol{\Sigma_{ij}^{NC}} \boldsymbol{\varepsilon C_{ij}} \left[\mathbf{5}\left(\frac{\boldsymbol{r_{ij0}}}{\boldsymbol{r_{ij}}}\right)^{\mathbf{12}} - \boldsymbol{B_{ij}} \cdot \mathbf{6}\left(\frac{\boldsymbol{r_{ij0}}}{\boldsymbol{r_{ij}}}\right)^{\mathbf{10}} \right] + \boldsymbol{\Sigma_{ij}^{NNC}} \boldsymbol{\varepsilon}\left(\frac{\mathbf{4}}{\boldsymbol{r_{ij}}}\right)^{\mathbf{12}}\end{aligned} \quad (12)$$

In equation (12), θ, ϕ, r_{ij}, NC, and NNC represent the bond angle, dihedral angle, inter-residue distance, native contacts, and non-native contacts, respectively. The first, second, third, and fourth terms correspond to the energies of the bond angle, dihedral angle, native interactions, and non-native interactions, respectively. The values related to the native structure are represented by subscript 0. In this study, the parameters are defined as follows: $\boldsymbol{K_\theta} = 20\varepsilon$, $\boldsymbol{K_\phi^1} = \varepsilon$, and $\boldsymbol{K_\phi^3} = 0.5\varepsilon$.[31] A native contact in the randomly simulated protein structure to be present when the distance of at least one heavy atom pair in two residues is less than the sum of van der Waals radii of two contacting heavy atoms + 1.4 Å.[31] The contact formed by the adjacent residues is disregarded in this study. A residue pair is defined as adjacent when the distance between the two residues is between 1-4 residues along the protein sequence, $|i - j| < 4$. The parameter C_{ij} indicates the strength of the scaled inter-residue interactions, which are calculated from the number of inter-heavy atom contacts divided by the average number of inter-heavy atom contacts per residue pair. The following equations show the calculation of $B_{ij}(\Theta_{ij})$, which indicates the closeness of the direction of two side chains in the native structure.

$$\boldsymbol{B_{ij}(\Theta_{ij})} = \begin{cases} 1 - \left(\Theta_{ij} - \Theta_{ij0}\right)^2 / a_{\Theta}^2, & if\ \Theta_{ij0} - \mathrm{a}_{\Theta} < \Theta_{ij} < \Theta_{ij0} + \mathrm{a}_{\Theta} \\ 0, & otherwise \end{cases}$$
$$\Theta_{ij} = \arccos\left(\frac{h_i \cdot h_j}{|h_i||h_j|}\right) \quad (13)$$

The relative orientation of the side chains of the two residues (Θ_{ij}) is calculated by using the relative angle between h_i and h_j, as shown in equation (13). The parameter h_i is defined as $r_{i,i-1} + r_{i,i+1}$, where $r_{i,i-1}$ denotes a vector between the ith and $(i-1)$th residues. h_i is used to define a bond vector mimicking the Cβ−Cα vector in the combination of $r_{i,i-1} \times r_{i,i+1}$.[32] The parameter B_{ij} value is between 0-1 and measures the ability of the two residues to make a contact. The residues are considered to make a contact when the B_{ij} value is close to 1, and the residues cannot make a contact when the B_{ij} value is close to 0. A cutoff value $a_{\Theta} = 0.6\pi$ was used as in our previous study.[33] According to undefinable to determine the vector h_i of the terminal residues, the B_{ij} value of both residues is set as 1. Then keep $\theta < \pi$ to prevent $h_i = 0$.

A replica exchange Monte Carlo simulation was used in the present study.[34] The various temperature ranges were applied to MC simulations. In this study, 32 temperatures (kBT/ε) are used for all selected samples. Where M refers to the domain length, M iterations of pivot moves for randomly selected residues followed by M iterations of crankshaft moves for randomly selected segments are included in one MC step. The segment size for a crankshaft move is randomly selected to avoid exceeding half the size of the domain length. It is worth noting that these methods aim to fulfill pivot and crankshaft moves, including small segment moves, end moves, and spike moves. After 10^5 equilibrium steps, 10^6 steps of the simulation were calculated, then exchange every 100 steps. The weighted histogram analysis method (WHAM) is used after the MC simulations to construct the free energy profiles from the trajectory at all temperatures.[35,36] The transition temperature for a protein was determined by the peak of its heat capacity curve.

Chapter 3

Study of folding mechanisms for Ig-like beta-sandwich proteins

3.1 Introduction

Titin has a huge molecular mass (~3MDa) and is responsible for striated-muscle elasticity.[37] It is a highly modular protein composed of ~300 domains.[38] The complete amino acid sequence of titin was recently analyzed and identified to two domain types, including Ig and FN3 domains,[39] linked in tandem. A single titin molecule extends from one end (Z disc) of the sarcomere to the middle (M line). The Ig domains are distributed along the entire length of the titin molecule and constitute the major part of I band region. While the FN3 domains are found in the A band region, which alternates with the Ig domain in a super-repeat pattern. Ig domains are found in various proteins including those for cell-cell recognition, cell-surface receptors, the immune system and muscle structure.

It is interesting to investigate how proteins with low sequence similarity can fold into a similar native structure. Particularly interesting is the case of Ig-like beta-sandwich fold. This is one of the most common protein structural folds mostly composed of 7 strands in two sheets. It includes 33 superfamilies, such as immunoglobulin (Ig) and fibronectin type III (FN3) superfamilies that have 5 and 2 families, respectively, in the current SCOPe 2.07 database.[8] Since these superfamilies are unrelated by sequence similarity and by evolution but share the common Ig fold called the "Greek key pattern" as the signature of these protein families.[40-44] Furthermore, the structural core was discovered to be four β-strands (β2, β3, β5 and β6),[45] key strands for folding, in both the Ig and FN3 domains.

In this present study, mainly two Ig-like beta-sandwich fold samples were investigated, the 27th Ig domain of the I band of the human titin (1TIT) and the FN3 domain of the human extracellular matrix protein tenascin (1TEN). Ig and FN3 domains have attracted extensive studies on mechanical properties, but the complete folding process is still uncertain due to the limitation of computational power.[46,47] Since the two domains have no discernible sequence identity (10.38%) but have similar structure and are composed of common features, such as the Greek key pattern and key strands for folding (Figure 1). The differences in transition state properties of folding processes and core residues, and the folding nucleus of these proteins have been investigated by experimental ϕ-value.[42,48] The mutations cause the transition state to become less native-like and cause the 1, 1' and 7 beta-strands to become completely unstructured. The high ϕ-value on the central strands (0.64-1.00) for TI I27 and Tanford β value (β_T) (> 0.9) suggesting that the transition state is very compact. On the other hand, different properties were found in 1TEN with moderate ϕ-value (0.39-0.60) and β_T about 0.7 in the transition state, indicating that about 30% more solvent is exposed than in the native structure.[42,48] For both proteins, the 3D-structure based sequence alignment shows the residues that make up the folding nucleus, corresponding to I20, Y36, I59 and V70 for the 1TEN

domain, and I23, W34, L58 and F73 for the 1TIT domain are located at the same location where the common structure core is.[41]

It was confirmed in previous studies that the prediction methods in our group are as successful in extracting the folding properties as the data from experimental analyses.[9-16] Even though our calculation methods have the potential to extract the different folding properties based only on amino acid sequence data and the extracted properties are in agreement with those determined by experimental analyses, the whole story of the folding processes could not determine due to limitations in the methods. Nowadays, to understand the whole story of the folding pathway of a studied domain, a 3D-structure-based Gō-model simulation can be used to clarify this missing information.

The aim of this study is to analyze the protein folding mechanism, with a focus on the conservation properties along related domains of the Ig-like beta-sandwich proteins, which are classified into different superfamilies on the basis of sequence similarity[45] but share the same fold. The methods used in this study include 3D-structure-based multiple sequence alignment (3D-MSA), inter-residue average distance statistic-based methods (ADM and F-value), and Gō-model simulation. The comparative results between 3D-structure-based and sequence-based techniques are also discussed in this study. It should be noted that the autonomous folding of a protein and the present results are compared to available experimental data.

3.2 Target proteins

The main proteins treated in this work are the Ig domain from titin protein (TI I27) (PDB code: 1TIT)[49] and the FN3 domain from tenascin protein (TNfn3) (PDB code: 1TEN),[50] which are members of beta-sandwich proteins with interlocked pairs and key strands for folding of beta-sandwich proteins.[45] The 3D-structure and topology of this protein, including common key strands and the Greek key pattern, are presented in the Figure 1. Furthermore, 152 Ig domains and 132 FN3 domains from human titin which published in the Uniprot database (UniProt ID: Q8WZ42)[24] were used as samples for evolutionary analyses.

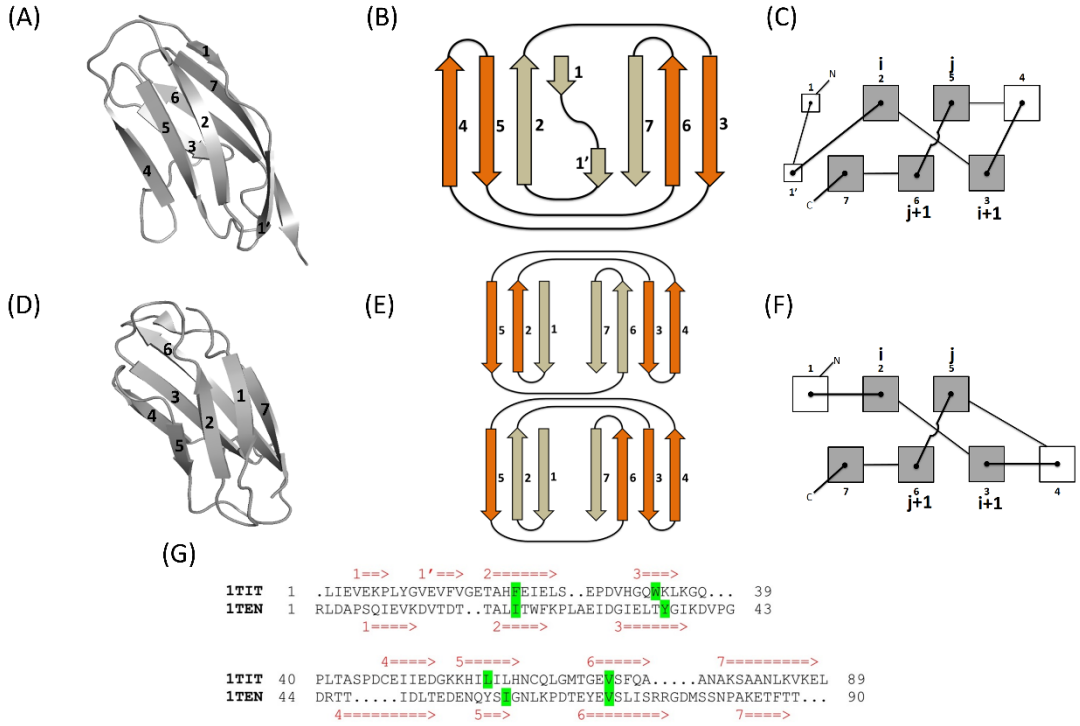


Figure 1 The structure of 1TIT and 1TEN. The ribbon diagram of the Ig domain structure of human titin (TI I27), PDB code: 1TIT (A),[49] and the FN3 domain structure of human tenascin (TNfn3), PDB code: 1TEN (D),[50] were constructed using PyMOL version 1.7.4. (B) and (E) show the strand arrangement of 1TIT and 1TEN with colored strands of the Greek key motif, and (C) and (F) show topology of 1TIT and 1TEN, respectively. A gray square denotes the conserved position in both domains. Multiple sequence alignment (G) of 1TIT and 1TEN were aligned based on native structure with green labeled hydrophobic core residues.

All published 3D-structures of Ig and FN3 domains of human titin protein were selected as a known structure sample set in this study. In this case, the titin proteins from various living organisms were selected for the purpose of evolutionary analysis study, including human (Uniprot ID: Q8WZ42), mouse (Uniprot ID: A2ASS6), fruit fly (Uniprot ID: Q9I7U4), zebrafish (Uniprot ID: A5X6X5) and nematode (Uniprot ID: G4SLH0).

Therefore, three representations were selected for each domain type containing the 3D-structure information required to investigate the whole story of protein folding with Gō-like model simulation. These protein sets including three Ig domains (PDB codes: 1TIT, 2A38, and 3LCY) and three FN3 domains (PDB codes: 1TEN, 2NZI, and 1BPV). From the multiple domains presented in some of these PDB files, only a single domain was selected and named as 2A38_Z1, 3LCY_A165, 2NZI_A70 and 1BPV_A62.

3.3 Results

3.3.1 ADM analyses

The compact region predicted by ADM analyses for titin and tenascin proteins

The average distance map for titin (1TIT) and tenascin (1TEN) are shown in Figure 2, including the predicted compact region, PdCR. For 1TIT, two PdCRs were identified as the primary or auxiliary PdCR with high compact density of 0.24 and 0.16, respectively. The primary PdCR located near the C-terminal at residues I57-V86, which covers β5 to β7. The auxiliary PdCR located is near the N-terminal site at residue L8-I50, which includes β1' to β4. The only predicted folding region at position R1-L62 of 1TEN was pinpointed with a compact value of 0.35. The predicted region covers the β-strand of β1 to β5.

From the PdCR, the primary segments that fold in the initial state of these domains are quite different. In transition state construction, 1TIT seems to fold from the C-terminus, whereas 1TEN starts to fold from the N-terminus. The PdCRs of 1TIT were split into two parts, but the compact value of two regions is quite similar at 0.24 and 0.16. It could be considered that the transition state of 1TIT might construct from two individual segments and make contacts with each other. The previous study by Ishizuka and Kikuchi (2011) identified a PdCR at positions I57-V86 of 1TIT and suggested that the compact region can be expanded to L8-V86 according to the proximal compact value. However, it is interesting that the PdCR of each protein covers the key strands for folding (Figures 1C and 1F) and the Greek key motif (Figures 1B and 1E) which are the common features of immunoglobulin-like beta-sandwich protein. Regarding key strands formation in 1TIT, each PdCR includes two key strands, β2 and β3 and β5 and β6, respectively. Thus, it is predicted that each pair of key strands individually form compact structures, and those compact portions interact with each other. Whereas β2, β3 and β5 of 1TEN form a compact structure and then β6 interacts with it later. Furthermore, it is interesting to note that the result of ϕ-value analysis for 1TIT performed by Fowler and Clarke (2001)[42] shows significantly higher values for residues from β2, β3, β5 and β6 under consideration which are more than the cutoff. Summation of average values and standard deviation are shown in Figure 3A. The ϕ-value plot looks bimodal for the two PdCRs. On the other hand, analogous ϕ-value plot for 1TEN looks mono modal around β3- β6 along the sequence as presented in Figure 3C. This result also corresponds to only one PdCR.

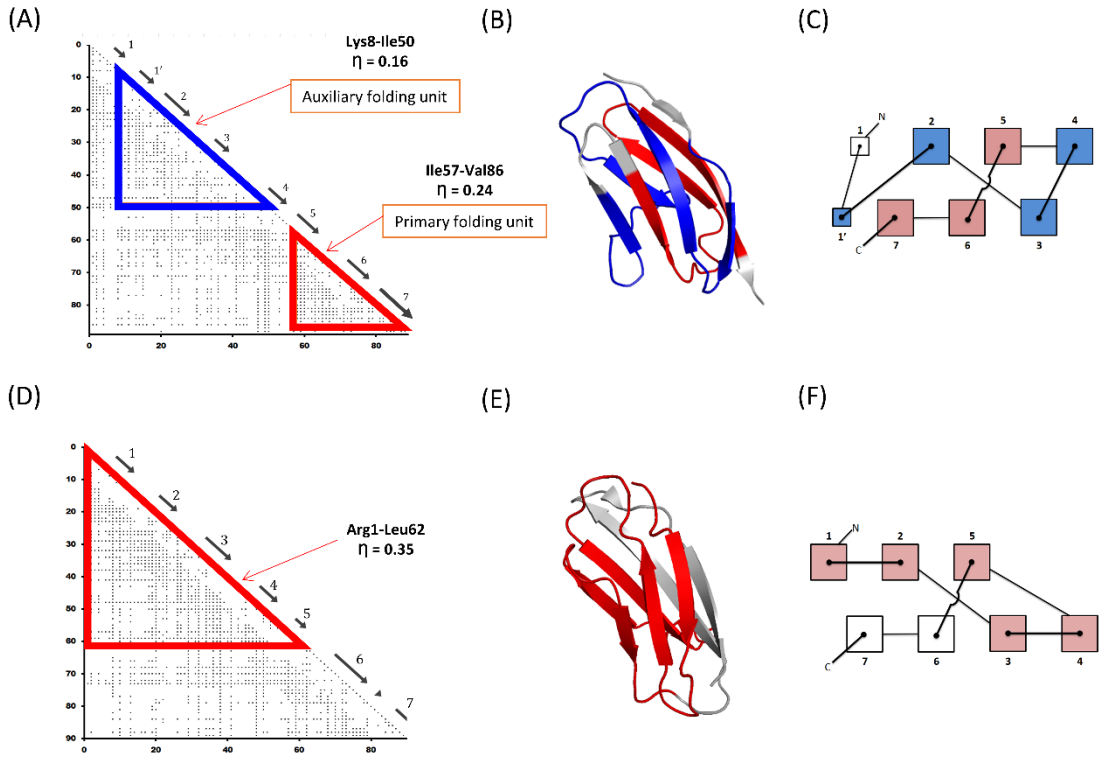


Figure 2 The average distance map (ADM) of (A) 1TIT and (D) 1TEN. The map of 1TIT predicts two compact regions, including primary PdCR and auxiliary PdCR which are colored red and blue in the native structure (B). Only one compact region (red) was predicted for 1TEN in its native structure (E). The black arrow marks each β strand. (C) and (F) show the topology with labeled predicted compact regions.

The ADM can predict the initial folding part of the protein but cannot explain the folding mechanism in that region. Then, the F-value analyses are used in the next step to extract the contact frequency in each residue that can occur in the initial event of folding.

3.3.2 F-value analyses

Due to nucleation-condensation mechanism analysis,[18,51,52] the contact interaction of key residues always occurs in all stages of folding, which is the hydrophobic core of the protein that collapses in a random fashion allowing the secondary structure elements to induce the native conformation. If the F-value analysis reflects the hydrophobic collapses to a denatured state, the high value of the F-value plot, i.e., the peak, should be considered as a probable folding nucleus of that protein. The F-value analysis could extract the residues with high possibility to make a contact in an early event as confirmed in previous studies.[10-12,14-16] In this study, the F-value were analyzed the 1TIT and 1TEN in comparing with experimental ϕ-value in each domain.

In 1TIT, the peaks from the F-value plots were determined at positions F14, W34, I49, I59 (Figure 3A), and the hydrophobic residues within ±5 residues of the peak are considered as a residue which might significantly important in the initial hydrophobic collapse event (show by red dot on the F-value plots).[16] Interestingly, the core residues, W34 and L58, are located near the peaks within ±5 residues (Table 1).

Table 1 Predicted compact region of 1TIT and 1TEN, and the position of F-value peaks. Red residues denote the nucleus residues.

	1TIT	1TEN
Folding unit	L8-I50(auxiliary) (β1'-β4), I57-V86(primary) (β5-β7)	R1 – L62 (β1-β5)
Compact value (η)	Auxiliary (0.16) & Primary (0.24)	0.35
Position of F-value peak	F14, W34, I49, I59	I20, I32, I59, V70, L72
HP near F-peak †	Y9, V11, V13 F14, V15, A19, V30, W34, L36, I49, I50, I57, L58, I59, L60	A18, L19, I20, W22, F23, A27, I29, I32, L34, Y36, Y57, I59, L62, Y68, V70, L72, I73

† Hydrophobic residue within ±5 residues of the F-value peak

The results of F-value calculations of the 1TIT and 1TEN domains are presented in Table 1. The results suggest that the hydrophobic residues near F-value peaks are involved in the folding mechanism of a protein. It is interesting that the selected hydrophobic residues were among the nucleus residues published in earlier study.[41]

Based on these computational analyses of the 1TIT domain, some early folding events could be discovered. The β5-β7 tend to be a primary segment for folding particularly the hydrophobic residue L58 in β5. The β1'-β4 were pointed as auxiliary segments, particularly core residues W34. However, the compact values of two PdCRs are not significantly different, 0.16 and 0.24. Each PdCR might fold individually before agglutination into the transition state conformation. The primary folding part tends to form contacts with the auxiliary parts involving hydrophobic residues, W34 and L58. In the previous study of 1TIT's structure by Fowler and Clark (2001),[42] the four core residues that formed the ring contact in the native structure include F21, W34, L58 and

V71, which are located within ±5 residues of the F-value peaks, except F21 and V71 on β2 and β6, respectively. V71 has a high F-value but was not included in the list of hydrophobic residues near the F-value peaks perhaps because the peak at position H61 has such a high value at the region around V71 could not show a peak. Nevertheless, the V71 residue should be considered as a high potential hydrophobic residue in the primary segment, that acts together with L58 to form the C-terminal region.

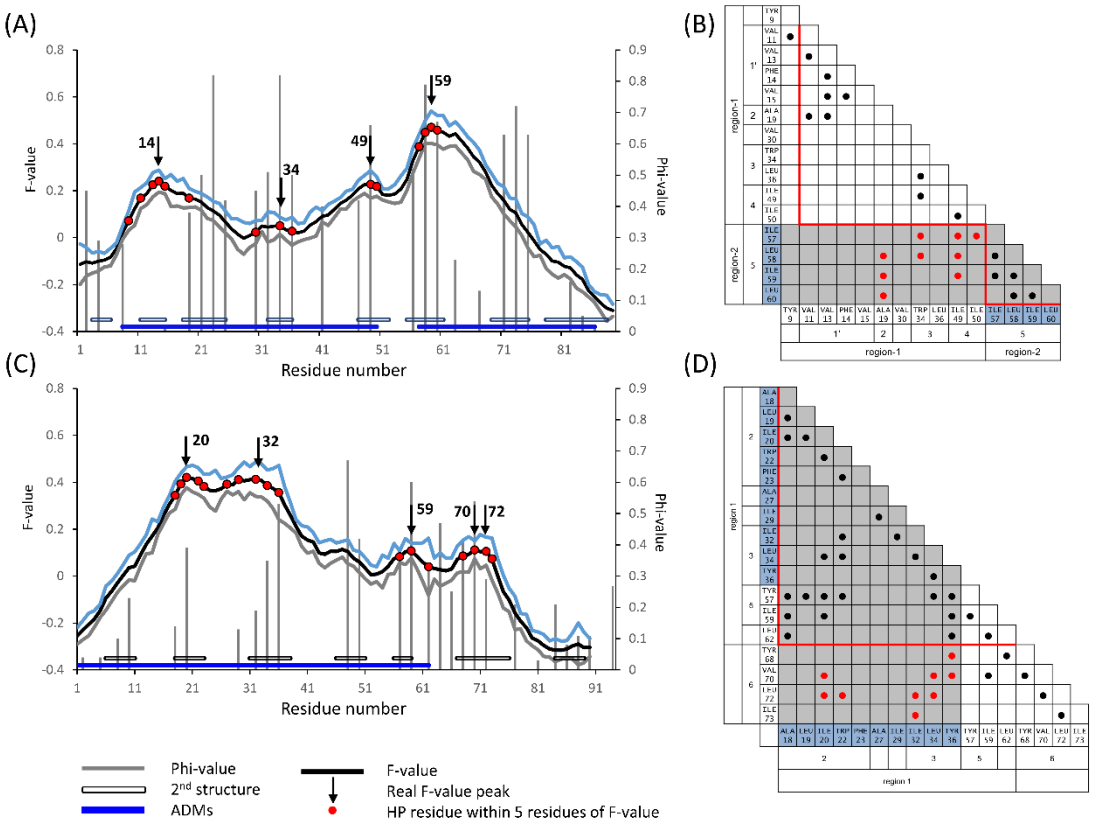


Figure 3 F-value plots of (A) 1TIT and (C) 1TEN with ADMs (the blue bar near the abscissa), experimental ϕ-value (gray bar) and standard error (blue and gray line). The hydrophobic residues (red dot) within 5 residues of F-value peaks (black arrow) are shown on the F-value plots. The open black bars near the abscissa represent each β strand. For (B) and (D), the blue region indicates the hydrophobic residues within 5 residues of the highest F-value peak. A dot in the contact map indicates a contact between hydrophobic residues, and red for a contact between a predicted region that formed by conserved residues near the highest F-value peak. "Region 1" means the first predicted compact region by ADM.

To compare to experimental results, ϕ-value analyses of 1TIT have done by Fowler and Clarke (2001).[42] As shown in Figure 3, the high ϕ-value residues demonstrate the importance of key strands for the 1TIT folding processes. It is worth noting that these prediction methods could predict the significant region. Moreover, Yagawa et al. (2010)[53] showed the result of PF (protection factor) values of H-D exchange of backbone NHs of each residue of 1TIT, which were calculated from the rates of exchange for a given NH under folded and unfolded conditions, and related to the free energy required to expose the amide to the solvent.[53,54] The result shows the stability of 1TIT native structure, especially on β5 where an intensive cluster of high PF values, which corresponds to the highest F-value peak and a compact region by ADMs (Figure 4).

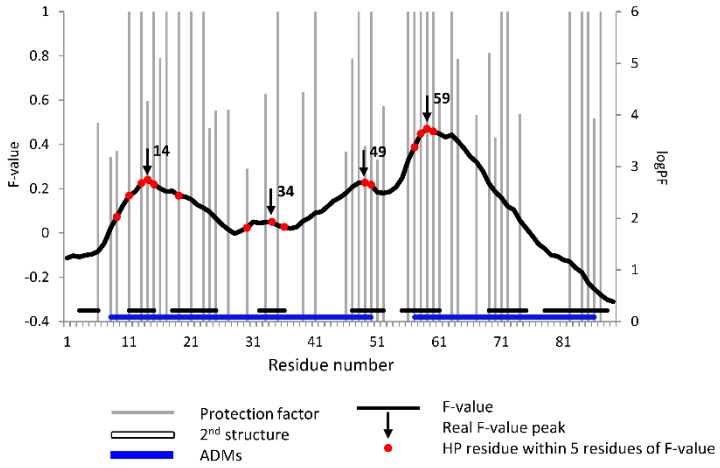


Figure 4 F-value plots and ADMs of 1TIT with protection factor (gray bar) from NMR-based amide H-D exchange experiment.[53]

The F-value peaks of 1TEN from a different superfamily are determined in positions I20, I32, I59, V70 and L72 (figure 3A), which are located on and or near the core residues, I20, Y36, I59 and V70, within ±5 residues of an F-value peak (Table 1). The folding core of 1TEN has the highest ϕ-value in each folding core unit. The ϕ-value of 1TEN has a high value above the cutoff, in the same manner as with 1TIT, on β3, β4, β5 and β6 which comprise the Greek key segment of this domain. It is interesting that F-value peaks are predicted on strands with high ϕ-value except for β4. Again, the F-value could extract residues with high possibility to be an amino acid that corresponds an early folding event.

In combination with ADM analysis, β1-β5 are included in the initial folding part that includes the effect of core residues on β3 and β5. In the next step of folding, the core residue on the β-strand might form contacts with an initial folding segment to assemble the native-like structure. The rest of the hydrophobic residues from the F-value analysis might provide temporary contacts in the denature state but not reflect the native-like contacts such as descripted by Kukic et al. (2017).[52]

These two proteins exhibit shared β-sheet fold structure,[55] but some of the

folding mechanisms are different as shown by the PdCP in Figure 2. The 1TIT starts folding from both termini, whereas 1TEN starts folding at the N-terminal site. Nevertheless, F-value analyses could extract the hydrophobic residues with high possibility to initiate the hydrophobic collapse in the initial state of folding. The key residues on the same key strands that make up the interconnecting contacts between strands stabilize a nucleation-condensation mechanism where long-range key residues interact in the formation of the transition state to make up the complicated native conformation.[42,48,56,57] Interestingly, the F-value peaks of both domains are located on the Greek key motif. That fact confirms the ability of these computational method to predict the high potential part for folding in an early event.

3.3.3 Evolution analyses

Next, ADM was used to extract the propensities of evolutionarily related proteins. To get the relevant result, multiple sequence alignment based on 3D-structure was investigated. The result of the structure-based multiple sequence alignment of 23 Ig domains and 6 FN3 domains of titin protein are shown in the Figures 5-6 and Table 2. Moreover, sequence homologies of human titin protein (Uniprot ID: Q8WZ42) are rather low in sequence-based multiple sequence alignments as shown in Figure A3 and A4 in appendix section: 10-43% for Ig domains and 18-49% for FN3 domains. The predicted compact regions are indicated by a red bar with the brighter color corresponding to higher compact density. Conserved hydrophobic residues are indicated by yellow letters in the predicted area and blue letters out of the predicted area. The blue arrow indicates the β-strand along the sequences. In addition, the conserved ratio of predicted regions (blue line) represents the ratio of residues at an aligned site included in PdCRs. The region with high value in the histogram can be regarded as a conserved PdCR during evolution. In this case, the 70% cutoff was applied to determine the conserved part (green line). The conserved hydrophobic residue site corresponds to a ratio of more than 90% of residues at an aligned site (red line).[10-12,14-16] It is interesting to find that the conserved part of this histogram appears at the key strands, that is, β3, β4, β5 and β6. Thus, the key strands and the conserved hydrophobic residues tend to be included in PdCRs evolutionarily.

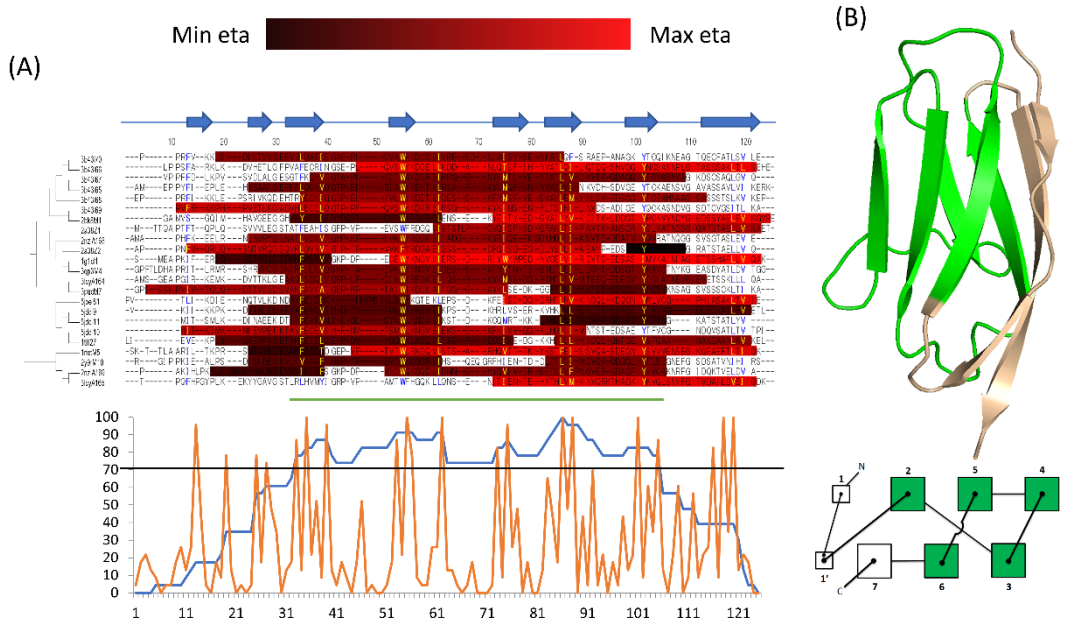


Figure 5 (A) 3D-based multiple sequence alignment of 23 known 3D-structures of titin Ig domain and (B) 1TIT topology with conserved predicted compact regions labeled in green. The predicted compact region is indicated by a red bar. The brighter red denotes higher compact density which are presented on the top of the figure. Conserved hydrophobic residues in the predicted compact region are indicated by a yellow letter whereas a blue letter for the residues out of a predicted compact region. The histogram of the conserved predicted compact region (blue line) and conserved hydrophobic ratio (red line) are presented on the bottom of the figure. A black line indicates 70% conservation and shows the conserved predicted region by a green bar. An arrow indicates the location of a β-strand in the sequence alignment. (See also Figure A1 in appendix section.)

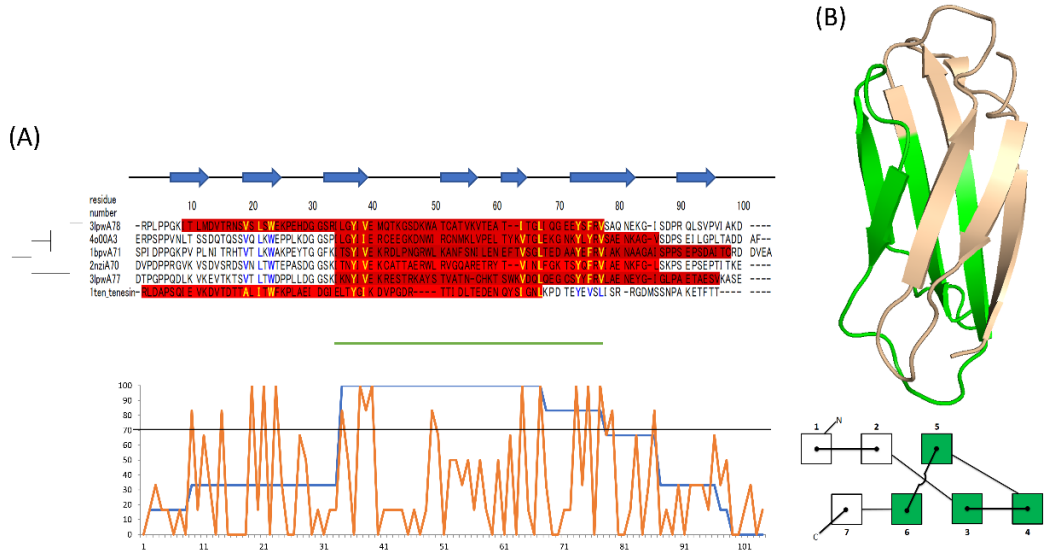


Figure 6 3D-based multiple sequence alignment of five known 3D-structures of titin FN3 domains with 1TEN from tenascin protein (A), and 1TEN topology with conserved predicted compact region labeled in green (B). See also Figure A2 in appendix section.)

The clustering of hydrophobic amino acids has traditionally been considered to be an important factor in protein core stability and can be directly derived from multiple sequence alignments.[20] In Figure 5 and Table 2, the 11 conserved hydrophobic residues were found under 90% conserved ratio, which includes all nucleus residues corresponding to F21, W34, L58 and V71 in 1TIT. It is interesting that almost every conserved hydrophobic residue is located in every β-strand, except β1'. Under the criteria of 70% cutoff, the conserved PdCR for 1TIT related domains covers residues E17-C63 (β2-β6).

Figure 6 shows the multiple sequence alignment and a histogram of known 3D-structures of FN3 domains in the titin domain chain. In this study, the amino acid sequence of 1TEN of tenascin protein, which has had extensive studies[41,48,58] on mechanical properties, was selected to compare with the FN3 domains from the titin protein, because there are no experimental data on the folding of these proteins. It is interesting that all conserved hydrophobic residues are located on the key strands for folding, and all core residues I20, Y36, I59 and V70 of 1TEN were detected as a conserved hydrophobic residue by native structure-based multiple sequence alignment as shown in Table 2. The PdCRs are conserved in the central region, β3-β6, which covers the entire of Greek key motif in both domains. This result suggests that the central region is the conserved folding initiation site.

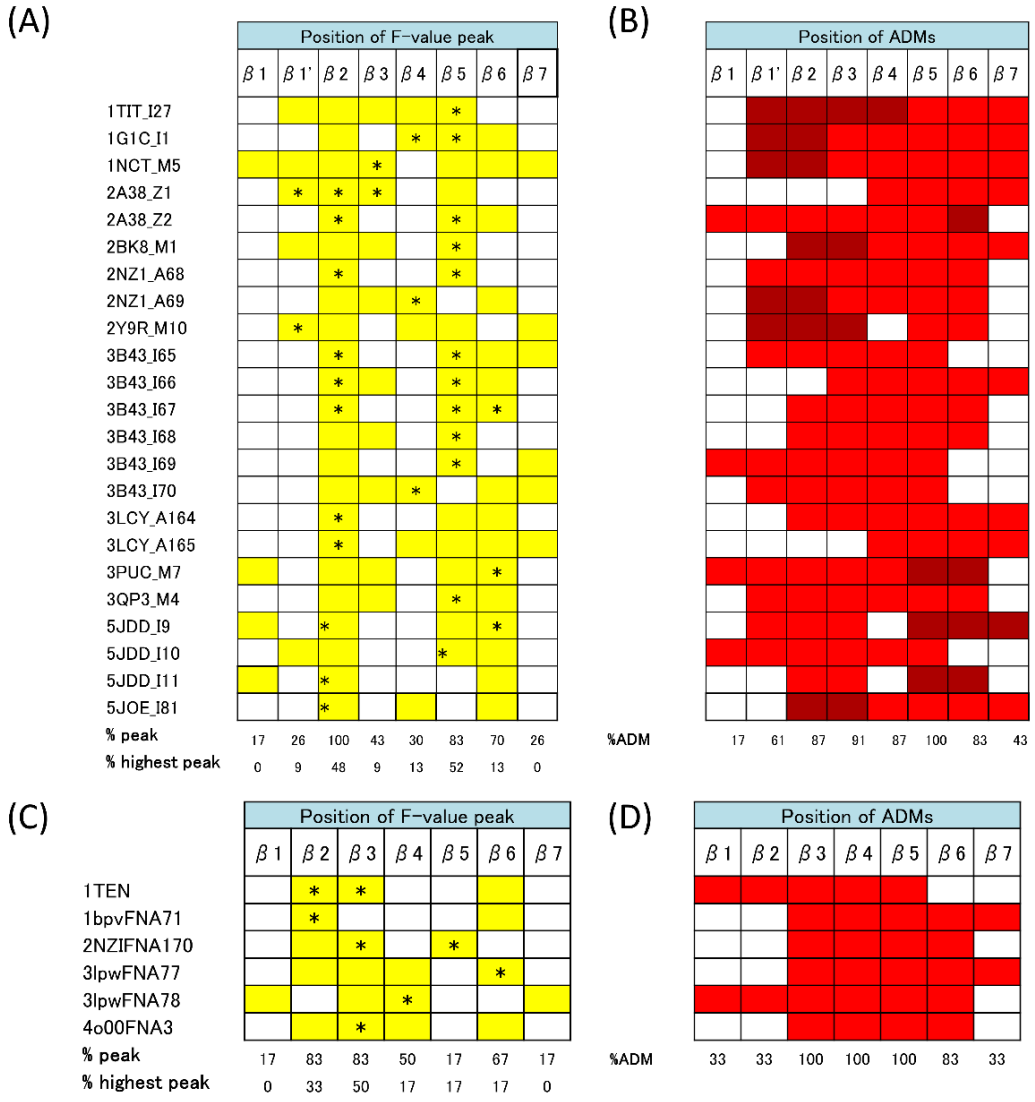


Figure 7 Position of F-value peaks and ADMs of known 3D-structures of Ig domains (A and B) and FN3 domains (C and D). The yellow regions represent the position of F-value peaks, and the highest value is marked with a star. The red regions denote the position of the PdCR.

Figure 7 presents the position of F-value peaks and PdCRs of known 3D-structures of Ig domains and FN3 domains. The yellow regions show the positions of F-value peaks, and the highest value is marked with a star. The positions of peaks tend to be conserved on β2 and β5 in the Ig domain with 100% and 83%, respectively, of occurrence ratio on the same β-strand (Figure 7A), and β2-β3 and β6 in FN3 domain with 83% for β2 and β3, and 67% for β6 (Figure 7C). The highest peak tends to be conserved on β2 and β5 for the Ig domain, and β3 for FN3 domain which is part of common key strands for folding and the Greek key motif strands. Figures 7B and 7D show the position of the PdCR from ADM analysis. The lighter red refers to the predicted region with higher compact value. The conservation of the both domain types tend to be in the central region. The F-value result and the position of conserved hydrophobic residues within ± 5 residues of the F-value peak of each domain are presented in the appendix section, Figures

B1–B2 and Figure C1-C2).

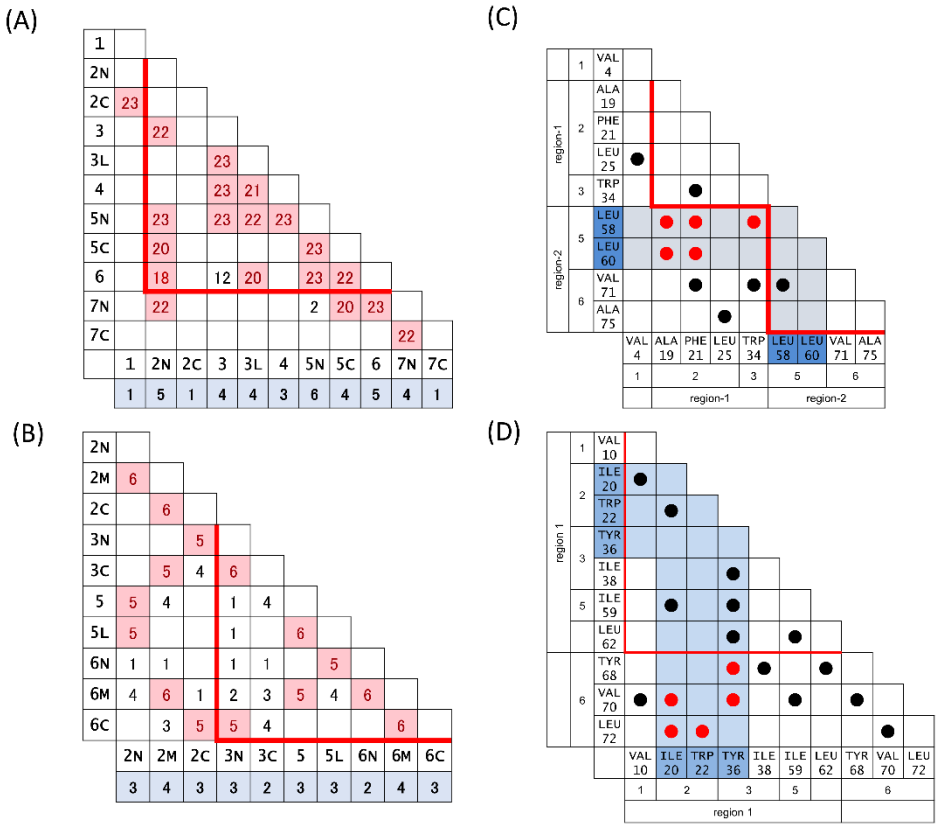


Figure 8 The appearance contact number of the conserved hydrophobic residue from structure-based sequence alignment of Ig domain (A) and FN3 domain (B). The appearance number greater than 70% is labeled in red. (C) and (D) show contacts between conserved hydrophobic residues derived from sequence-based multiple sequence alignment of Ig domains and FN3 domains, respectively. The blue region indicates the conserved residues within ±5 residues of the highest F-value peak. A dot in the contact map marks the contact between conserved hydrophobic residues, and red denotes the contact between predicted regions that are formed by conserved residues near the highest F-value peak. The red line denotes the conserved PdCR.

The frequency contacts in same position between the conserved hydrophobic residues for 23 known 3D structural Ig domains are presented in the Figure 8A, appearance contact number. The axial number represents the location of each conserved hydrophobic residue where N, C, M and L stand for N-terminal, C-terminal, middle and loop of each β-strand. If the appearance ratios over 70% is used, the contacts seem to be conserved at almost all contacted location, except the contacts formed by conserved hydrophobic residues on positions 3 to 6 and 5N to 7N. The conserved position number in the blue row shows under the axial number. The highest number of contacts is observed at position 5N where contacts with other hydrophobic residues on β2-β6 form, but they rarely make a contact with position 7N and not interact with position 1. However, position 5C shows in 20 out of 23 times to be in contact with position 7N. It is worth nothing the significance of the conserved hydrophobic residues on β5 through the stability of the

native structure. Even position 1 did not show such a contact with the conserved hydrophobic residues on β5 but the linking contact appears at position 2C to assemble the compact structure. The highest peak of F-value on β5 correlated with the highest number of predicted contacts of conserved hydrophobic residues.

If the results of 1TEN are compared with those of 1TIT (Figure 8), the contacts between the conserved hydrophobic residues of 1TEN seem to be less conserved. The highest number was observed at positions 2M and 6M. Position 2M shows high contact number with 3C and 6M. On the other hand, position 6M forms a high number of contacts with 2M and 5. In this case, if consider only the highest contact number positions, positions 2M and 6M, the conserved contact will occur between β2, β3, β5 and β6, which are key strands for folding. Figure 8B and our ADMs show the folding mechanism of the FN3 domain where early folding structure will occur in the central part, including β3-β6, with the interaction between the conserved hydrophobic residues in the PdCR, and then assemble the other strands to construct the native conformation.

Next, further analyses of sequences of distantly related domains are considered. The results show the amino acid sequence-based multiple sequence alignment of all domains in human titin protein (Uniprot ID: Q8WZ42), including the Ig domain and the FN3 domain, by using 1TIT and 1TEN as a query for the Ig and FN3 domains, respectively. To determine the conserved PdCR, 151 of 152 Ig domains were analyzed by multiple sequence alignment and ADMs. In this case, domain number 98 was eliminated to decrease the gap from the alignment method. Along the alignment, the sequence identities between domains vary from 5-73% in Ig domains (Figure A1) and 5-82% in FN3 domains (Figure A2) as shown in Table 2.

Table 2 The multiple sequence alignment and calculation results of different living organisms. The position of the conservation was shown by the query domain 1TIT (Ig domain) and 1TEN (FN3 domain).

		Known 3D-structure	ID: Q8WZ42 (Human)	ID: A2ASS6 (Mouse)	ID: Q9I7U4 (Fruit fly)	ID: A5X6X5 (Zebrafish)	ID: G4SLH0 (Nematode)
Ig domain	Conserved folding unit	A19-A75 (β2-β6)	E17-C63 (β2-β5)	F14 – N62 (β1'-β5)	I23-S26, D29-A75 (β2-β6)	F14-H61 (β1'-β5)	T28-Q64 (β2-β5)
	No. of Conserved HP	11	9	7	9	7	4
	Conserved HP	V4, F21, L25, W34, L41, I49, L58, L60, V71, L84, V86	V4, A19, F21, L25, W34, L58, L60, V71, A75	A19, F21, W34, L58, L60, V71, A75	A19, F21, L25, W34, L41, L58, L60, V71, A75	A19, F21, W34, L58, L60, V71, A75	F21, W34, L58, V71
	No. of conserved HP near F-peak †	5	5	5	5	5	3
	Conserved HP near F-peak †	F21, W34, I49, L58, L60	A19, F21, W34, L58, L60	A19, F21, W34, L58, L60	A19, F21, W34, L58, L60	A19, F21, W34, L58, L60	F21, W34, L58
	Sequence similarity	10%-43%	5% - 73%	5% - 72%	7%-60%	3%-67%	3%-40%
FN3 domain	Conserved folding unit	E33-L72 (β3-β6)	E33-R76 (β3-β6)	E33-R76 (β3-β6)	I32-S74 (β3-β6)	E33-D78 (β3-β6)	E33-D40, R45-R75 (β3-β6)
	No. of Conserved HP	10	10	9	10	9	11
	Conserved HP	A18, I20, W22, Y36, I38, I59, L62, Y68, V70, L72	V10, I20, W22, Y36, I38, I59, L62, Y68, V70, L72	V10, I20, W22, Y36, I38, I59, Y68, V70, L72	V13, I20, W22, L34, Y36, Y57, I59, L62, Y68, V70	V10, I20, W22, Y36, I38, L62, Y68, V70, L72,	A18, I20, W22, Y36, I38, Y57, I59, L62, Y68, V70, L72
	No. of conserved HP near F-peak †	9	8	7	9	6	9
	Conserved HP near F-peak †	A18, I20, W22, Y36, I59, L62, Y68, V70, L72	I20, W22, Y36, I59, L62, Y68, V70, L73	I20, W22, Y36, I59, Y68, V70, L74	I20, W22, L34, Y36, Y57, I59, L62, Y68, V70	I20, W22, Y36, L62, Y68, V70	A18, I20, W22, Y36, Y57, I59, L62, Y68, V70
	Sequence similarity	18-49%	10-82%	10-84%	13-34%	9-73%	10-37%

† Hydrophobic residue within ±5 residues of the F-value peak

Nine and twelve conserved hydrophobic residues were found with over 90% conserved ratio for the Ig and FN3 domains, respectively. However, only 10 of 12 conserved positions are the hydrophobic residue in 1TEN. The central region tends to be a conserved PdCR consisting of residues E17-C63, β2-β6, and E33-R76, β3-β6, of 1TIT and 1TEN, respectively. These regions are considered to be the conserved folding units during evolution. These domains fold generally from the central region with the help of conserved hydrophobic residues. The location of conserved PdCR in two alignment studies (3D-based and sequence-based multiple sequence alignment) show almost the same region at the middle part (Table 2), covering part of the common Greek key motif and interlocked pairs.

The position of conserved hydrophobic residues on F-value plots are shown in Figure 9. The 9 and 10 side chains of conserved hydrophobic residues of 1TIT and 1TEN, respectively, are shown in the native's cartoon structure with distinct colors of β-strands (Figures 9B and 9D). The most conserved hydrophobic residues have their chains pointed toward the core nucleus, which stabilizes the native conformation. Hydrophobic residues within ± 5 residues of the F-value peak are members of hydrophobic core residues.

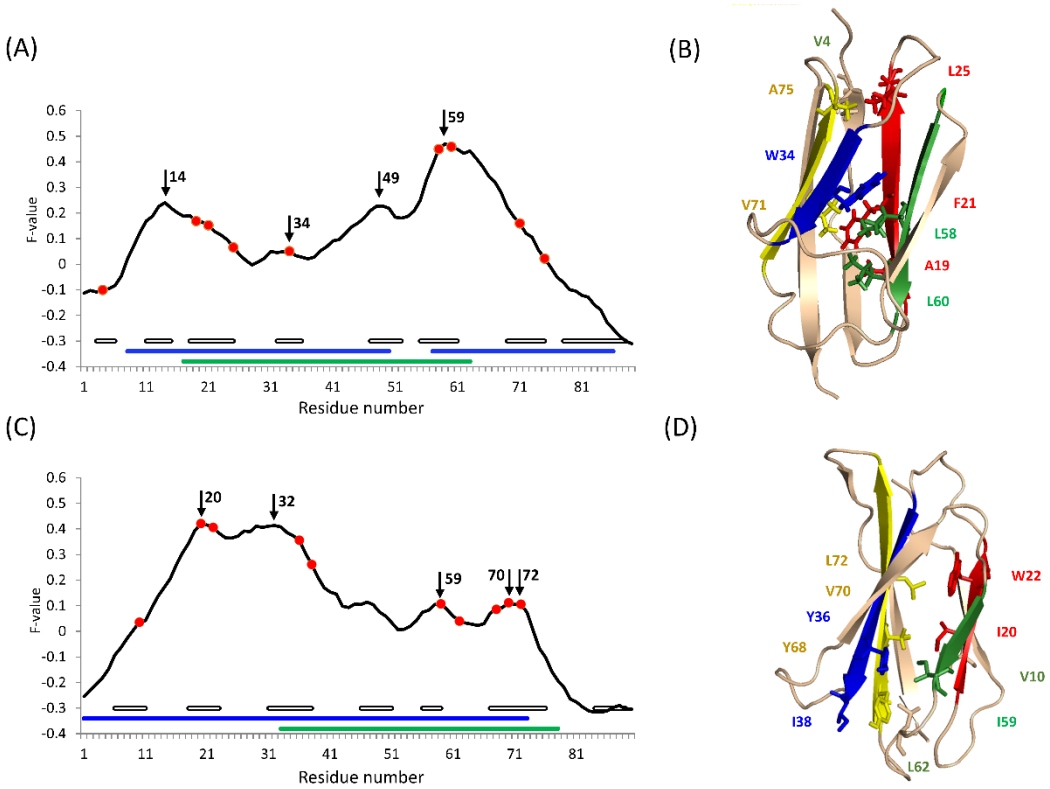


Figure 9 Position of conserved hydrophobic residues on F-value plots and 3D-structure of 1TIT (A and B) and 1TEN (C and D). The blue bar and green bar near abscissa indicate predicted compact region of query domain and the conserved predicted compact region from sequence-based multiple sequence alignment, respectively.

A conserved PdCR denotes a region that tends to be preserved in predicted compact regions during evolution and, of course, it may differ from PdCR(s) for an individual target protein. However, such a conserved PdCR contains the major (or all) key strands and indicates the robustness of the folding mechanism to form the common β-sandwich structure.[55]

Packing formed by conserved hydrophobic residues for each query protein is examined next. In Figure 8, a 1TIT (Figure 8C) makes contacts with conserved core residues in PdCRs in the native structure. Core residues L58 and L60 which within ± 5 residues of the highest F-value peak form contacts with other core residues in a different predicted region, F21 and W34. In 1TEN (Figure 8D), the conserved hydrophobic residues near the highest peak tend to form contacts with residues around the C-terminal. It is interesting that the core residues I20 and W36 in PdCR form contacts with V70 which is a hydrophobic core residue at the C-terminus. Due to this fact, the hydrophobic residues near the F-value peak are significant for stabilizing the native structure.

Evolutionary analysis of proteins from different organisms in the Uniprot database, including human (ID: Q8WZ42), mouse (ID: A2ASS6), fruit fly (ID: Q917U4), zebra fish (ID: A5X6X5) and nematode (ID: G4SLH0), was investigated based on sequence-based multiple sequence alignment as shown in the Table 2. The trend of their ADM results is quite similar. Notably the results of FN3 domains in all living things tend to show similarity in the position of β3-β6. On the other hand, Ig domains of each organism shown are slightly heterogeneity in the ADM results. Again, the conserved PdCRs of both domain types cover the strands of Greek key motif. The sequence similarity of titin domains, Ig domains and FN3 domains, in each living organisms are very different. But the computational results tend to predict the conserved PdCR in the same region, around central strands, which is similar to the result for human titin protein. When plotted, the conserved hydrophobic residues in each domain type on the F-value plots of the main studied domain, the conserved hydrophobic residues within ± 5 residues of F-value peaks of each domain type include hydrophobic core residues as shown by red residue numbers in Table 2. Moreover, the evolutionary analyses suggest conservation of folding properties between different living organisms. (The alignment results are shown in an appendix section (Figures A3-A12).)

3.3.4 Go model simulations

To overcome the limitation of these sequence-based prediction methods, which could decode the initial folding event, and to investigate the whole story of folding processes, a Gō-model simulation was conducted for all six selected samples. In this study, the free-energy profiles for each domain was calculated. As a reaction coordinate, Q value indicates the ratio of the number of the native contacts detected in each state to all native contacts. A long-range interaction is defined when the sequence separations are no less than 16 residues; $|i - j| \geq 16$.[59] Based on present sequence-based techniques, the highest peak of F-value may relate to an initial hydrophobic collapse, and thus, the residue interactions found close to the denatured state are reflected in the F-value result. As in the

ADM analysis, since the PdCR indicates a folding unit formed in the early events of folding, the result should correspond to the contacts observed at the Q value next to the denatured state. The contacts observed in every 0.05 range of the Q value are expressed by the contact frequency map to study the whole story of protein folding. The first and last valleys of the free-energy profile are considered as the denatured state and the native state, respectively, and the peak between these valleys is considered as the transition state.

Firstly, some information of six selected domains were analyzed, including amino acid sequence composition, sequence similarity, structure similarity and predicted initial folding unit derived from ADM and F-value analyses. This information is presented in Table 3 and 4.

Table 3 The ADMs results and hydrophobic residue composition for both entire amino acid sequence and in the PdCR of all selected domains.

Domain type	PDB code	Length	HP† residue	HP† ratio (%)	ADM			
					Folding unit	Compact value (η value)	HP† residue	HP† ratio (%)
Ig domain	1TIT	89	36	40.45	C47-V86	0.23	17	42.50
	2A38_Z1	101	38	37.62	I29-A99	0.29	27	38.03
	3LCY_A165	98	39	39.80	A35-Q96	0.39	25	40.32
FN3 domain	1TEN	90	30	33.33	R1-I73	0.40	27	36.99
	2NZI_A70	97	31	31.96	I34-T95	0.45	23	37.10
	1BPV_A62	104	39	37.50	I34-V102	0.31	28	40.58

† Hydrophobic residue

Table 4 Sequence identity and RMSD investigations.

			RMSD					
			Ig domain			FN3 domain		
			1TIT	2A38_Z1	3LCY_A165	1TEN	2NZI_A70	1BPV_A62
Sequence identity	Ig domain	1TIT		1.65	2.83	9.34	11.99	10.81
		2A38_Z1	27.27		0.89	9.90	13.28	8.12
		3LCY_A165	14.12	20.21		6.586	3.43	9.87
	FN3 domain	1TEN	13.92	10.53	11.25		3.04	2.26
		2NZI_A70	6.25	10.26	8.54	21.59		1.21
		1BPV_A62	11.25	11.39	12.20	20.45	39.18	

3.3.4.1 ADM of six selected domains

As mentioned above that the predicted average distance maps indicate the region with highest compact value, while some overlapping regions with lower compact values are observed. That is, a PdCR may be flanked by a region with a slightly lower compact value that may contain residues forming many contacts with the core predicted region but few contacts with other parts. Therefore, the flanking region(s) should be considered as a part of the PdCR if the overlapped region(s) contain the compact value within 85% of the core value,[17] which call the 85% rule. The larger predicted region will be regarded as part of the PdCR when the length does not exceed 70% of the whole sequence. The new ADMs with the 85% rule of 1TIT, 1TEN and other four known 3D-structure proteins are presented in Figure 10.

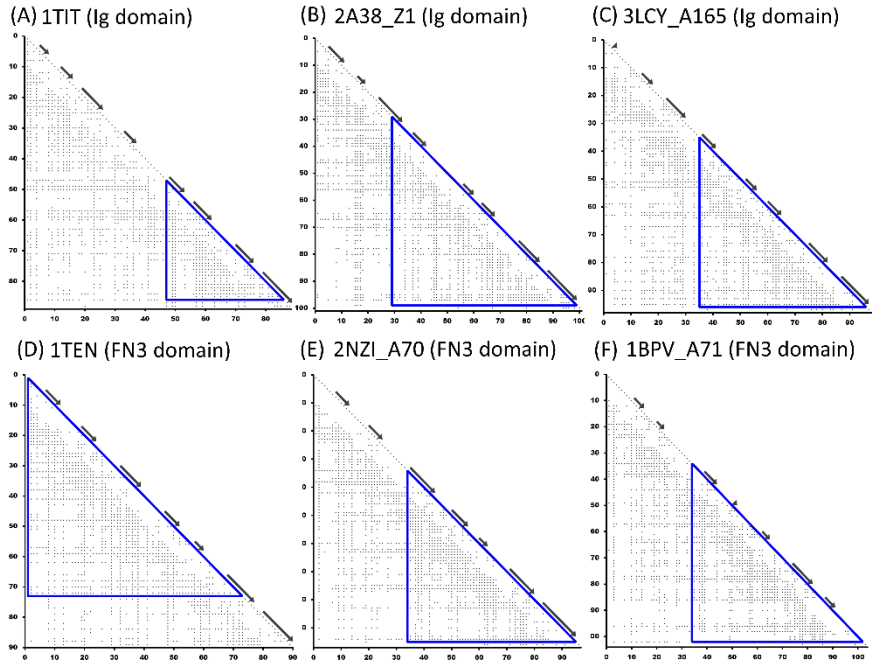


Figure 10 The average distance map with 85% rule of 1TIT (A), 2A38 (B), 3LCY (C), 1TEN (D), 2NZI (E), and 1BPV (F), respectively. The predicted compact region, PDCR, is indicated by the blue triangle. The black arrow marks each β-strand.

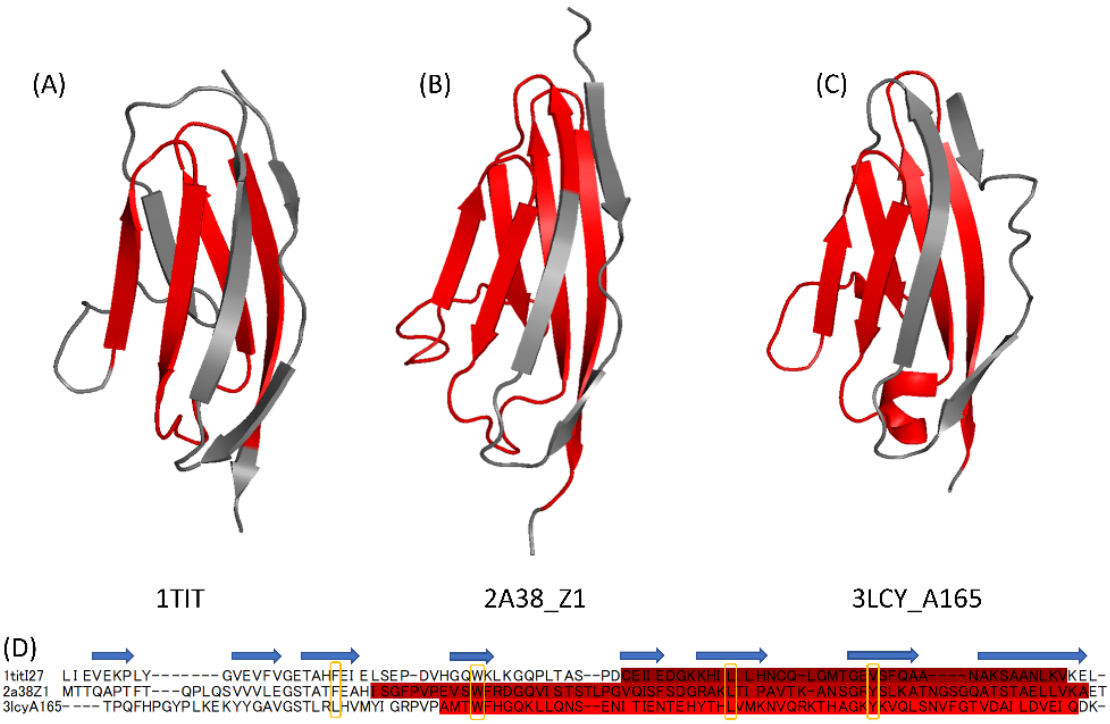


Figure 11 The structure of three selected Ig domains, 1TIT (A), 2A38 (B), and 3LCY (C). The PdCR is colored red in the native structure and in the 3D-structure-based multiple sequence alignment (D). The blue arrow indicates the location of a β-strand. The nucleus residues are enclosed by yellow squares.

Figure 11 presents the native conformation and multiple sequence alignment of three selected Ig domains. The PdCR was indicated in the 1TIT predicted contact map with a high compact density of 0.23 at residue C47-V86 (β4 to β7). For 2A38_Z1, the region covering the residues between I29-A99 (β3 to β7) was predicted to be a compact region. At the same time, the PdCR of 3LCY_A165 was detected at position A35-Q96, which covers β3 to β7. According to these PdCR results, the first folding site seems to fold from the C-terminus, then accumulates other segments to form the native conformation.

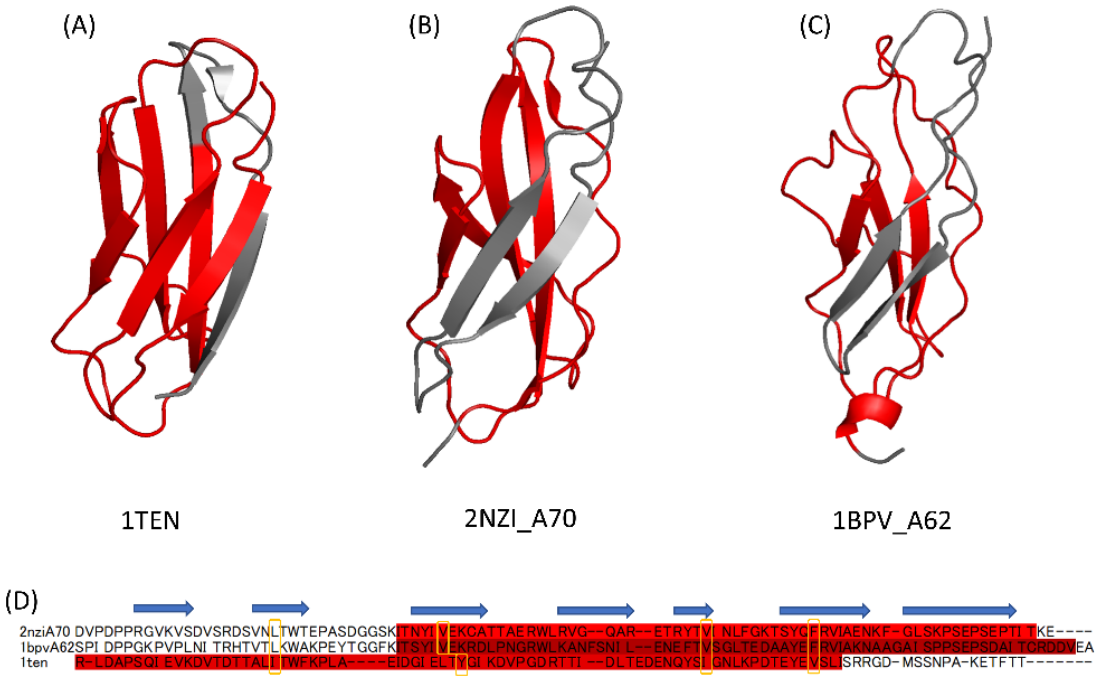


Figure 12 The structure of three selected FN3 domains, 1TEN (A), 2NZI (B), and 1BPV (C). The PdCR is colored red in the native structure and 3D-structure-based multiple sequence alignment (D). The blue arrow indicates the location of a β-strand. The nucleus residues are enclosed by yellow squares.

The native structure and multiple sequence alignment of three selected FN3 domains are presented in Figure 12 with the PdCR labelled in red. The PdCR was detected in the region covering residues R1-I73 (β1 to β6), I34-T95 (β3 to β7), and I34-V102 (β3 to β7) of 1TEN, 2NZI_A70, and 1BPV_A62, respectively. The central region between β3 to β6 tends to conserve in this group, and thus, the initial folding unit may start to form the contact in this area. The ADM analyses of the Ig domains and FN3 domains points to the conserved initial folding unit being located in the central region same as the previous evolutionary analyses. It is interesting that the PdCR of each domain covers the key strands for folding and the Greek key motif, which are the common features of immunoglobulin-like beta-sandwich proteins. The nucleus residues on the same key strands that form the inter-domain interaction between strands stabilize a nucleation-condensation mechanism where long-range key residues interact in the formation of the transition state to constitute the complicated native conformation.[42,48,56,57] Furthermore, for 1TIT, β5 is covered by the PdCR and is the highest protected region, according to the

cluster of protected residues. Even though the conserved PdCR of the Ig domain presents on the C-terminal side, the ϕ-value analysis shows less stability for β7. Therefore, β7 may form many contacts within the PdCR but does not interact with another part.

3.3.4.2 F-value of six selected domains

The F-value profiles of six selected domains are shown in Figure 13, with plus and minus SD indicated by blue and gray lines, respectively. The green and orange bars just above the x-axis represent the secondary structure and the PdCR. The results of F-value analyses are also shown in Table 5. The F-value profiles of 1TIT and 1TEN were descripted above, section 3.4.2.

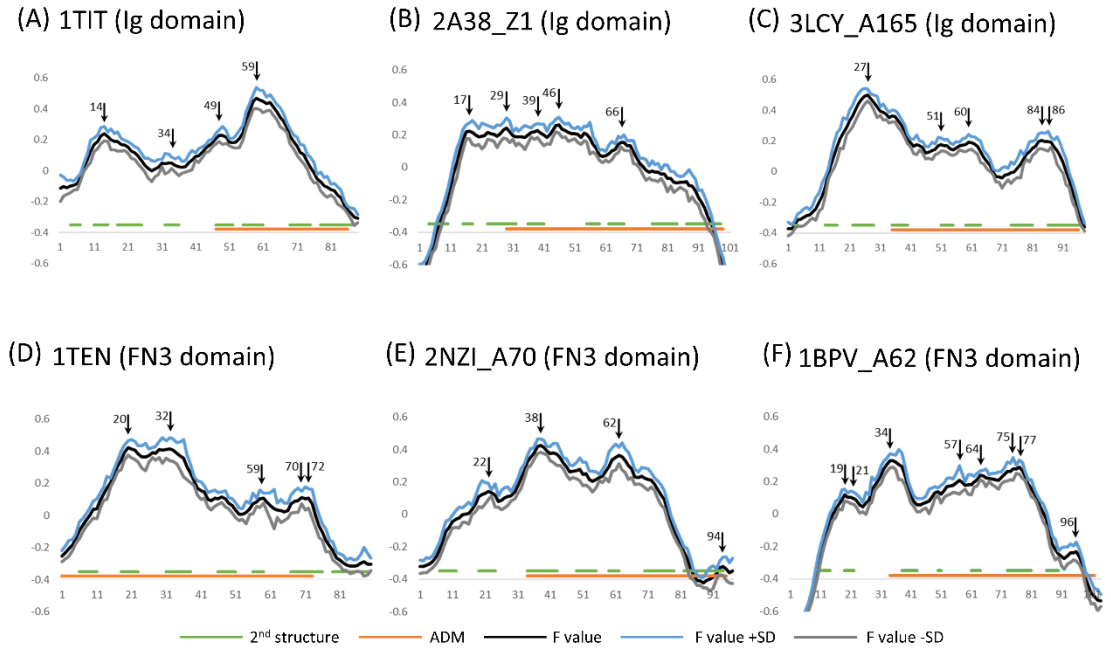


Figure 13 F-value plots of six selected domains with the position of the peaks indicated by black arrows. The orange and green bars near the x-axis represent the PdCR and beta strands. In all figures, the value of F, F + SD, and F - SD were plotted for each residue as black, blue, and gray lines, respectively.

In the F-value profile of 2A38_Z1, high values are observed from β1'-β3 as shown in Figure 13B, and therefore the highest value region in this domain could not pinpoint. However, five residues, V17, I29, W39, I46, and T66, are noted as a peak. Residue numbers F25, W39, L65, and Y78 are considered as nucleus residues from multiple sequence alignment with two other Ig domains. Three out of four residues besides Y78 were detected near the peak of F-value within ±5 residues. These three residues are located on β2, β3, and β5. The PdCR of 2A38_Z1 is predicted at residue I29-A99, which covers some part of the high F value shown in Figure 13B. It is possible that

some residues in the N-terminal region may start to form contacts with other residues in the initial event but cannot form a stable structure.

The 3LCY_A165's F-value profile (Figure 13C) indicates five peaks at residues M27, I51, T60, F84, and T86. The highest peak at residue M27 is located on β2, which is one of the strand compositions of the key folding strands and Greek-key structure, a common feature of Ig-like beta-sandwich proteins.[55] From this result, the first stable structure may start to fold from the C-terminus via the effect from hydrophobic residues near the indicated peaks, particularly L62, which is one of the nucleus residues from multiple sequence alignment and is observed near an F-value peak, T60.

Table 5 The position of the peaks on the F-value plots. Red residues denote the nucleus residues found near the peak of F-value within ±5 residues.

Domain type	PDB code	Position of F-value peak	HP† near F-peak
Ig domain	1TIT	F14, W34, I49, I59	Y9, V11, V13 F14, V15, A19, V30, W34, L36, I49, I50, I57, L58, I59, L60
	2A38_Z1	V17, I29, W39, I46, T66	L12, V15, V16, V17, L18, F25, A27, I29, F32, V34, V37, W39, F40, V45, I46, L51, A63, L65, I67, A69, V70
	3LCY_A165	M27, I51, T60, F84, T86	L22, L24, V26, Y28, I29, I51, I53, Y59, L62, V63, L80, V83, F84, V87, A89, I90, L91
FN3 domain	1TEN	I20, I32, I59, V70, L72	A18, L19, I20, W22, F23, A27, I29, I32, L34, Y36, Y57, I59, L62, Y68, V70, L72, I73
	2NZI_A70	T22, I38, V62, I94	V19, L21, W23, A27, I34, Y37, I38, V39, A43, Y60, V62, I63, L65, F66
	1BPV_A62	V19, L21, I34, F75, V77, I96	I14, V19, L21, W23, A24, F32, I34, Y37, I38, V39, A71, A72, Y73, F75, V77, I78, A79, A82, A95, I96

† Hydrophobic residue

The FN3 domains were investigated in the same manner as the Ig domains (Figure 13D–13F). In 2NZI_A70, four peaks of the F-value plot are present at residues T22, I38, V62, and I94, as shown in Figure 7E. However, the last peak at residue I94 can be disregarded due to the fluctuation of the terminal region. Three of four nucleus residues were detected near the peaks: L21, V39, and V62, which are located on β2, β3, and β5, respectively. Based on our prediction, the initial folding event seems to start from the interaction between β3 and β5 and stabilizes the C-terminal region. Then the whole structure is put together with the contacts of the hydrophobic residues in the PdCR and another nucleus on β2, specifically, L21.

Figure 13F depicts the contact frequency profile of 1BPV_A62. Six peaks of the F-value plot are indicated at residues V19, L21, I34, F75, V77, and I96. The last peak, I96, can be disregarded due to the unstable terminal region as 2NZI_A70. Three of four nucleus residues, L21, V39, and F75, were detected near the peaks of the F-value plot within ±5 residues, which are located on β2, β3, and β6, respectively. As a result, V39

and F75 on β3 and β6 have the potential to interact with each other and then construct the steady folding unit as predicted by the ADM. Afterwards, the N-terminus starts interacting with the initial folding unit to become a native-like structure.

3.3.4.3 Gō-model simulation results

Now, the Gō-model simulation results for the Ig-domain superfamily are discussed as a part of an initial folding mechanism analysis and compare these results with those from present sequence-based methods and available experimental results. Subsequently, the whole folding processes for each sample is also investigated. An intermediate state in 1TIT, 2A38_Z1, and 3LCY_A165 was not observed in the free-energy profiles derived from the Gō-model simulation, that is, these selected domains suggest to fold into the native structure with two-state processes. This result reflects the fact that small domain proteins with 100 amino acids or fewer have generally been shown to fold in two-state kinetic processes.[60] The free-energy profiles of these domains indicate the same location of denatured and native states as Q = 0.05 and 0.90. A large single barrier presents among these two states. Due to this result, the location with the highest energy is considered as a transition state of folding.

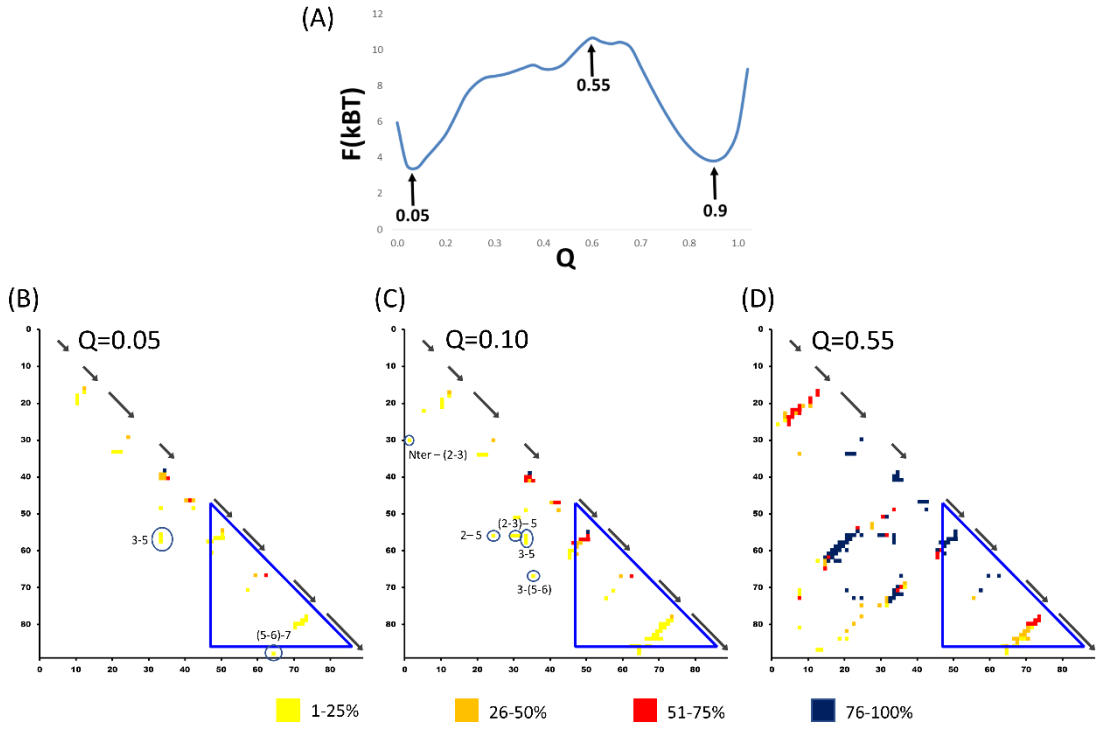


Figure 14 Free-energy profile of 1TIT at the respective Q values (A). The contact frequency maps of Q = 0.05, 0.10, and 0.55 represent the contacts detected by a Gō-model simulation of the denatured state (B), late denatured state (C), and transition state (D), with the blue triangle of the PdCR predicted by ADM analysis. The cluster of long-range contacts is indicated by a blue circle.

In the case of 1TIT, as shown in Figure 14A, the transition state was detected at Q = 0.55. The first long-range interactions were observed at Q = 0.05, the denatured state, with contacts among β3 and β5, and the loop of β5-6 and β7. This result corresponds well to the highest peak of the F-value plot shown in Figure 13A that predicted β5 as the significant segment for forming contacts in the initial state of folding. At Q = 0.10, the cluster of long-range interactions formed by the loop of β5-6 and β7 is observed in the C-terminus. As a result, the primary PdCR from ADM analysis is related to 3D-based simulation result. The contacts formed by the residues on β5, which are located near the highest peak of the F-value plot within ±5 residues, start to make contacts with other residues on β2, a loop of β2-3, and β3, which underscores the accuracy of our sequence-based method. The folding processes of 1TIT are presented by a topology derived from the contact map as shown in Figure D1. A noteworthy fact is that the contact of β2 and β4 does not present in every state of folding. The native-like structure was observed from Q = 0.15 except for β1, β1', and β7, which do not form the native contacts until fully formed at Q = 0.40. From Q = 0.15 to 0.25, the unsteady contacts among β2 and β6 were also detected until Q = 0.30. This result corresponded well to the ϕ-value analysis[42] in terms of the N-terminus and C-terminus of the domain showing low stability, which reflects the fluctuation of these regions in the folding processes. 1TIT seems to fold to the native conformation faster than other domains, according to the detection of all the strands interactions.

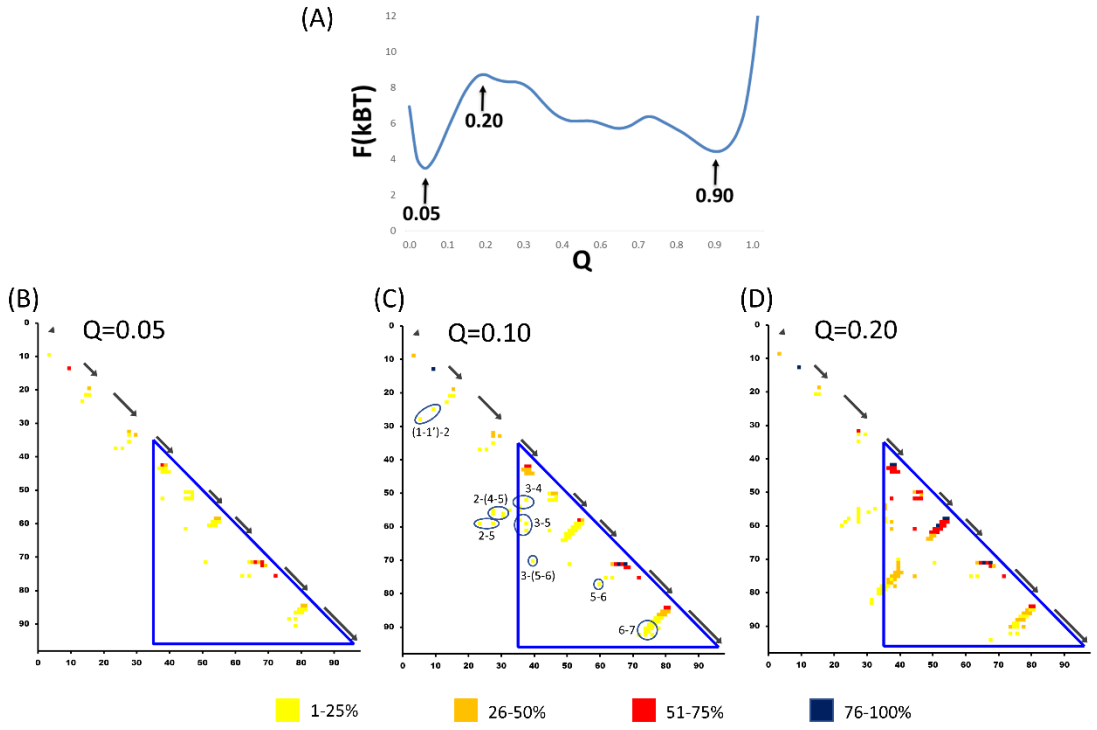


Figure 15 Free-energy profile of 2A38_Z1 at the respective Q values (A). The contact frequency maps of Q = 0.05, 0.10, and 0.70 represent the contacts detected by a Gō-model simulation of the denatured state (B), late denatured state (C), and transition state (D), with the blue triangle of PdCR predicted by ADM analysis. The cluster of long-range contacts is indicated by a blue circle.

For 2A38_Z1, the location of transition state is indicated at Q = 0.70. The initial long-range interaction of 2A38_Z1 is similar to that of 1TIT, in which the clusters of contacts between β3 and β5 present in the very early state at Q = 0.05. The ability of a nucleus residue to lead the protein folding was confirmed by the contact of two nucleus residues on β3 and β5, which were detected in the early events. It is comparable with the F-value indicating that β3 is located near the predicted peak. At Q = 0.10, contacts among β3, β4, β5, and β6 were detected, which is different than for 1TIT at the same Q value, as shown by the topology prediction in Figure D2. Furthermore, the residues on the long loop β3-4 start forming contacts with β5 and β6, and the loop β5-6, concurrently interacting with β7. The compactness of this PdCR (I29-A99) is reflected by these contacts covering three clusters of long-range interactions (Figure 9C). In this state, where β1, β1', and β2 only formed contacts with adjacent strands, the long-range native-like contacts are not observed. The contacts of β2 with β3, β4, β5, and β6 have been formed since Q = 0.15. Even though β2–β6 formed native contacts and stabilized the central region despite the fluctuation of the terminal strands, β1, β1', and β7, are observed as well. From Q = 0.35, β7 tends to form the contacts with other strands in the native conformation, except with β1 and β1'. The 2A38_Z1 has been fully formed since Q = 0.70. It is interesting that contacts of β2 and β4 are detected in this domain that are not present in 1TIT and 3LCY_A165.

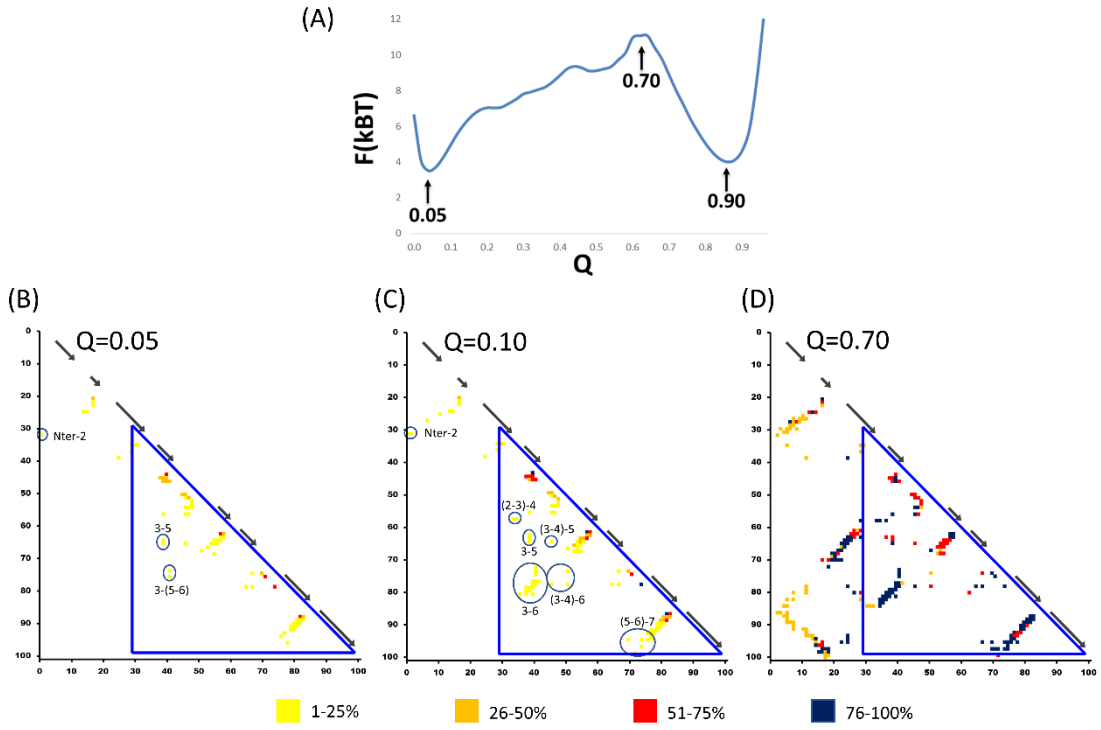


Figure 16 Free-energy profile of 3LCY_A165 at the respective Q values (A). The contact frequency maps of Q = 0.05, 0.10, and 0.20 represent the contacts detected by a Gō-model simulation of the denatured state (B), late denatured state (C), and transition state (D), with the blue triangle of PdCR predicted by ADM analysis. The cluster of long-range contacts is indicated by a blue circle.

The 3LCY_A165 Gō-model results are presented in Figure 16, with the detected transition state at Q = 0.20. This domain presents only adjacent contacts at the state with Q = 0.05, in contrast to the previous two domains, 1TIT and 2A38_Z1. However, the long-range contacts, β2 to β5 and β3 to β5, are detected for Q = 0.10, which is similar to the contacts formed in 1TIT at the same state of Q value. These initial long-range contact formations corresponded well to our PdCR, particularly when the predicted region was extended to A35-Q96. Additionally, contacts formed by β2 occurred in this state as well. Due to the observed big cluster contacts, it seems related to the position with a high F-value around β2 and β3. Next, at Q = 0.15-0.25, β6 starts to form contacts with β3 but not with β2 until Q = 0.30. The fluctuation of the contacts formed by β1, β1', and β7 are also observed in this domain. At Q = 0.60, 3LCY_A165 are almost completely structured except the contact between β1 and β2. The native-like structure has been evident since Q = 0.65, as shown in Figure D3.

The Gō-model simulation results for the FN3 domains will now be presented. Regarding to the position of the transition states detected in this study, only one transition state is detected in 1TEN at Q = 0.35, while two transition-state peaks are present in 2NZI_A70 and 1BPV_A62. The free-energy profile of 2NZI_A70 shows the transition-state peaks at Q = 0.15 and 0.50, however, there are unclear valleys between these peaks. The first peak indicates as a pre-transition state and the second as a post-transition state. In 1BPV_A62, two transition states are indicated at position Q = 0.15 and 0.55, and a clear valley of an intermediate state is seen at Q = 0.4. Based on this, 1TEN and 2NZI suggest to fold into the native conformation by a two-state folding pathway.

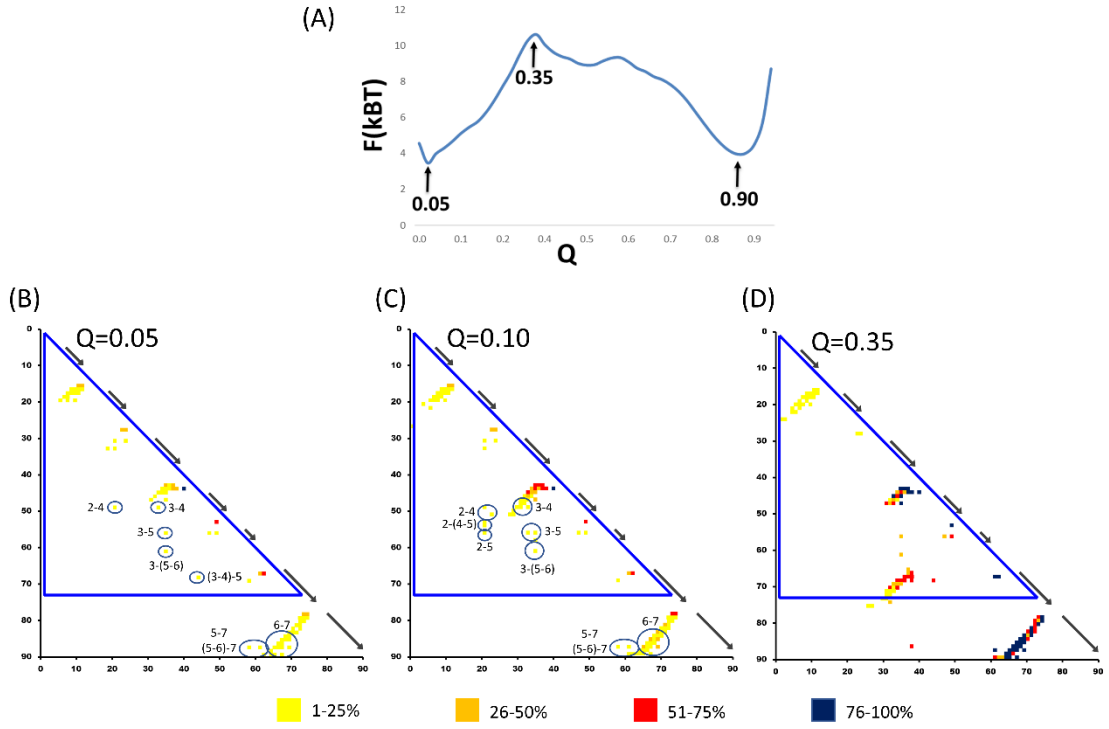


Figure 17 Free-energy profile of 1TEN at the respective Q values (A). The contact frequency maps of Q = 0 .05, 0.10, and 0.35 represent the contacts detected by a Gō-model simulation of the denatured state (B), late denatured state (C), and transition state (D), with the blue triangle of PdCR predicted by ADM analysis. The cluster of long-range contacts is indicated by a blue circle.

First, for 1TEN, the long-range interactions in the initial state at Q = 0.05 are shown for the β-strand of β2-β4, β3-β4, β3-β5, β5-β7, and β6-β7. The new interactions between β2-β5 and the additional contacts of β3-β4 that are detected at Q = 0.10 and 0.15 as presented by the contact map and the predicted topology in Figures 17C and Figure D4 in appendix section. These contact clusters can be compared with the F-value result, which shows a high value for β2 and β3. Moreover, the ADM predicted region indicated a PdCR at residue R1-L62 which covers the long-range interaction clusters in the central region. Interestingly, the stability of the central region was also investigated by ϕ-value analysis as shown in Figure 3C. According to this study, both sequence-based and 3D-based techniques can extract some folding information from 1TEN and tend to correspond to experimental results. However, the contacts formed by β2 do not seem robust in the folding processes as presented for Q values ranging from 0.20 to 0.50. The contacts between β3 to β6 start to form at Q = 0.20. At Q = 0.25, there are no such contacts between β1-β3 as indicated by the blue solid line between the β-strands. The contacts formed with β5 fluctuate at Q = 0.30-0.35, including the interactions with adjacent strands, β4 and β6. Moreover, the native contact between β2-β5 is not stable until Q = 0.55. The native-like structure starts to fold at Q = 0.65.

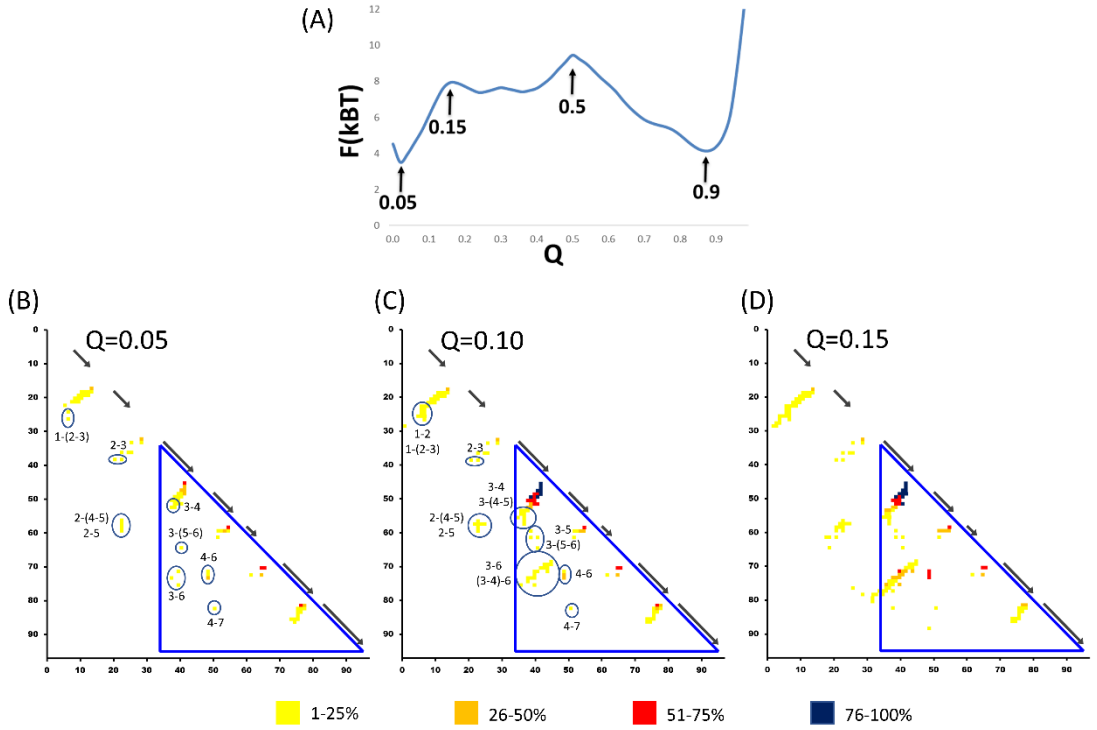


Figure 18 Free-energy profile of 2NZI_A70 at the respective Q values (A). The contact frequency maps of Q = 0.05, 0.10, and 0.15 represent the contacts detected by a Gō-model simulation of the denatured state (B), late denatured state (C), and pre-transition state (D), with the blue triangle of PdCR predicted by ADM analysis. The cluster of long-range contacts is indicated by a blue circle.

In Figure 18B, the contacts formed at Q = 0.05 of 2NZI_A70 are presented by a type of contact map. In this state, numerous long-range contacts are detected which differ

from one another, including the Ig domains. However, there is no such long-range contact among β6 and β7 that is observed. It can be concluded that the initial structure of 2NZI_A70 is rapidly formed and robust in the early stage, as shown in Figure D5 in appendix section. Then the contacts between β3 and β5 are observed at Q = 0.10. The swiftly increasing contacts of β3 stabilizes the central region, similar to the predicted folding unit result from the ADM analysis. Furthermore, the high number of β3 contacts correspond well to the highest peak of the F-value plot, which pinpoints the same region. Then the new contacts of β3-β7 and β2-β6 are detected at Q = 0.15 and Q = 0.20, respectively. The relative Q value between the pre-transition state and the post-transition state, Q = 0.25 to 0.55, detected the fluctuation of the terminal segments, the N-terminal and C-terminal. Finally, the native-like structure of 2NZI_A70 is detected from Q = 0.60.

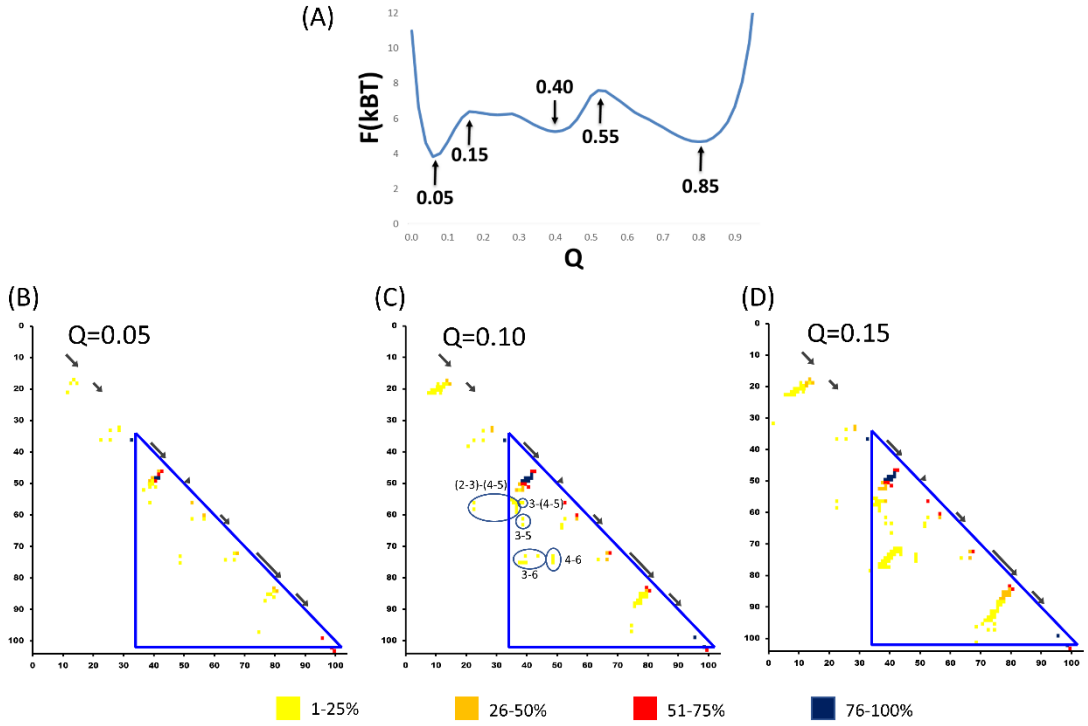


Figure 19 Free-energy profile of 1BPV_A62 at the respective Q values (A). The contact frequency maps of Q = 0.05, 0.10 and 0.15 represent the contacts detected by a Gō-model simulation of the denatured state (B), late denatured state (C), and first transition state (D), with the blue triangle of PdCR predicted by ADM analysis. The cluster of long-range contacts is indicated by a blue circle.

The last Gō model results for the FN3 domains, 1BPV_A62, are presented in Figures 19 and Figure D6 in appendix section. The interaction of adjacent strands from 1BPV_A62 is observed to start at Q = 0.05, except for β2 to β3 and β4 to β5. Abundant contacts for loop β2-3 and β3 with other residues are observed in the central region at Q = 0.10, corresponding to the highest peak from the F-value analysis, and the compactness of this region seems to correspond with the PdCR as well. At Q = 0.10-0.25, contacts among β3-β5 and β3-β6 are detected. The contacts between β2-β5 are permanently detected starting at Q = 0.40. This corresponds well to the free-energy profile, which shows the intermediate state at this Q value. Lastly, the completely native-like structure

of 1BPV_A62 was observed starting at Q = 0.50. It is interesting that there is no such contact observed between β4-β5, unlike the case of the other domains. The whole folding pathway of the selected samples was analyzed in this study by Gō model simulation. The long-range native contacts formed in the initial state of folding were detected at Q = 0.05 and 0.10.

According to these results, some residues tend to form a strong interaction, while some residues show only transient contacts. The example of transient contacts is presented in 1TEN. The contact between the residue on loop β3-4 and β5 can be detected in the state when Q = 0.05 and no longer observed at Q = 0.10, then this interaction reforms at Q = 0.35. Due to this kind of situation, the predicted structures derived from contact frequency maps (Figures D1-D6 in appendix section) present unstable conformations that can be found throughout the folding processes which reflect the fluctuation of the protein structure during native structure construction. According to the Gō model results, the fluctuation is frequently observed during the formation processes, particularly at the terminal sites, which corresponds well to ϕ-value analyses that show low value, unstable interaction at both terminal ends.[42,48]

3.4 Discussions

Nowadays, a Gō-model simulation, which needs 3D-structure-based data as an important material, have been used to predict the whole story of folding processes. Nevertheless, the contact map generally used in Gō-model simulation is generated from the experimental study which not available for all proteins. To overcome this problem, the folding properties in the initial state of folding were analyzed based on only amino acid sequence data. In this study, an average distance map of unknown structure protein was constructed from average distance statistic methods to extract the folding mechanism by the means of high contact density in the contact map.

The ADM and F-value analyses required only amino acid sequence data to extract the initial folding event and to distinguish between proteins with different folding routes, as shown by the present results. In other words, if the ADMs and F-value profiles show different properties, then these proteins suggest to fold by different routes. These techniques can be used to identify the folding event of a target protein and distinguish it from that of other similarly structured conformations. According to the contacts predicted by ADM, the present results suggest that nonlocal contacts tend to form early during the folding process.[61,62] The non-native contacts formed during the folding process have been intensively studied,[63] indicating that some interactions in the initial state can be broken and constantly reform into new residues. As shown by the ADMs, a predicted contact derived from inter-residue average distance statistics is not restricted to the native structure of the protein sample: this kind of predicted contact map can be predicted all possible contacts, which may include some non-native contacts in the early stage. Hence, the PdCR refers to a region with a high number of residue interactions which comprises all possible contacts. ADMs and F-value have suggested the folding process includes a position of initial hydrophobic collapse leading to the transition state for folding. Due to the study of structural similarity between transition and native state of 1TIT, the Tanford β value (β_T),[42,64] which relates the denaturant dependence of the unfolding processes, provides an estimate of solvent-exposed surface area. 1TIT has β_T greater than 0.9, indicating that the transition conformations are very close to the native state.[42]

In the ADM study of 1TIT, the top three regions, I57-V86, C47-V86, and L8-V86, are indicated by high compact values of 0.240, 0.228, and 0.205, respectively (data not shown). The predicted region of I57-V86 is considered to be a primary folding unit, and L8-I50 is considered to be a secondary folding unit. The compact region can be expanded to L8-V86 according to the proximal compact value.[12,13] However, this region covers 76 out of 89 residues of the whole domain of 1TIT, about 85% of the domain length, which covers almost all the protein. When take the 70% criterion of PdCR length into account, only one predicted region at C47-V86 with only 44% of the domain length is regarded as an initial folding unit.

In this study, replication of F-value analyses was increased to 100 times to increase the smoothness of the F-value plots compared with the previous studies.[15,16] In accordance with the increasing of smoothness, the F-value plots around β6 shows only ripples that could not be described as a peak. However, β6 should be considered as a significant region for folding due to contact with conserved hydrophobic residues shown

in Figure 8, which displays contacts between the primary compact region and a different β sheet.

The folding behavior of 1TIT was studied by Fowler and Clarke.[42] The central strands β2-β6 are fully formed in the transition state. This result corresponds well to a present Gō-model simulation of 1TIT that shows the stability of the central strands during native structure construction (Figure D1). The ϕ-value analysis of the transition state of 1TIT (Figure 3A), shows two core residues, W34 and L58, in agreement with the experimental result.[42] This fact suggests that W34 and L58 tend to form the native-like contacts in the initial stage of folding. The protection factor result[53] indicated that β5 is the steadiest strand. In comparison to this study, ADMs and F-values shown high value on the C-terminal site where including L58.

The PdCR of 1TEN was detected by ADM at the N-terminus covering β1-β5. It is interesting that the experimental ϕ-value shows the central strands are more ordered than the edge strands.[48,58] β1 is considered as a part of the initial folding unit. It is possible that β1 might form non-native contacts in the initial state and adjust to native-like contacts with β7 later. However, the peak of F-value analyses appears at the 4 key strands which correspond well to the stability of central strands from experimental work.[42] Furthermore, Paci *et al.*[58] found that the contacts between the nucleus residues in β2 and β6 are essentially lost in the transition state. It is interesting that the contact between β2 and β6 cannot be detected by a Gō-model simulation in the early state of folding (Figure D4).

Even though 1TIT and 1TEN share the key strand for folding[41,42,55] which stabilizes and drives the folding pathway to build up the native conformation, the overall properties of the domains are significantly different. Experiments suggested that 1TIT is significantly more structured than 1TEN in the transition state of folding as judged by the mean of ϕ-value, Tanford β value, RMSD, SASA and radius of gyration.[41,48,58] The Greek key motif of 1TIT and 1TEN includes β3-β6 and β2-β5, respectively. Interestingly, the calculated F-value shows the highest peak on the Greek key strand(s) in both domains, β5 for 1TIT and β2-β3 for 1TEN.

Using the criteria of 70% conserved PdCR and 90% conserved hydrophobic residue of residues at an aligned site, the same conserved hydrophobic residues from evolutionary analyses of known and unknown 3D-structures of human titin protein were detected. The predicted compact regions tend to be conserved in the central region, where three of four key strands for folding and all four Greek key motif strands reside. In addition, our study of conservation properties of titin protein from different living organisms showed surprising conservation in the same position as in human titin, even though the sequence similarities are quite different. 1TIT and 1TEN were used as representative for each domain type, and it is possible that conserved PdCR may differ from a query domain due to the different amino acid composition. Following from these results, the conserved PdCR of Ig domain and FN3 domain of titin protein could reflect the conservation of common features for folding of Ig-like proteins.

It is possible that the initial folding unit in some proteins would be broken along the folding paths. It has been extensively studied that the non-native contacts during folding which are not taken into consideration in the present study should be also

significant.[63] Non-native contacts formed by conserved hydrophobic residues may take a significant role during folding. Furthermore, it was also demonstrated by native-centric Gō-model simulation that higher-order (more nonlocal) contacts tend to form early during the folding process.[61,62] Such contacts may be formed by conserved hydrophobic residues.

The present Gō-model simulation is purely native centric, and there are no attractive non-native interactions. It is interesting that the folding properties extracted from only amino acid sequences agreed well with the contacts detected in the initial state of folding, just next to the denatured state detected by Gō model simulations, as presented in Figures 14-19. According to the results for 1TIT, the PdCR, C47-V86, does not cover the long-range contacts formed by residues on β3 but corresponded well to the higher calculated contact density (0.05, data not shown) of the PdCR in the contact frequency map simulated by a Gō model at Q = 0.10. In contrast, the density of the region covering the contacts formed by β3 is lower, with a contact density value of 0.03 (data not shown). With respect to the hydrophobic residue's composition in the amino acid chain, as shown in Table 2, the hydrophobic ratio for entire sequences are slightly lower than the ratio calculated in the PdCRs. The classification of intrinsic disorder proteins based on amino acid sequence composition emphasizes the significance of hydrophobic residues to stabilize the core of folded globular protein.[65] The higher hydrophobic ratio of the PdCR reflects the higher possibility of forming the compact structure in the early state of folding.

3D-structure-based Gō model simulation was also conducted to capture some common features and investigate the whole story of folding processes for all the domains. The free-energy profiles of all samples show only one energy barrier, except for 1BPV_A62, which means these selected domains fold into the native structure with two-state processes, while 1BPV_A62 folds with multi-state folding kinetic processes with one detected intermediate state. These results corresponded well to the fact that generally, small domain proteins, with 112 amino acids or fewer, have been shown to fold in two-state kinetic processes.[6] With regard to the contacts formed in the initial state, some of the nucleus residues start interacting with other residues in the central region. Contacts between the nucleus residues on β3 and β5 of all domains, namely, W34 and L58 in 1TIT, W39 and L65 in 2A38_Z1, W38 and L62 in 3LCY_A165, Y35 and I58 in 1TEN, V39 and V62 in 2NZI_A70, and V39 and V64 in 1BPV_A62, were detected at Q = 0.10. This interaction corresponds to the common feature that the folding mechanism of Ig-like beta-sandwich proteins starts folding from the nucleus residues, which located on β2, β3, β5, and β6, key strands for folding. The ADMs and F-value analyses pointed to the central region as an initial folding unit, which agrees with the present Gō-model simulations, which also detected the cluster of contact in the central region in the early event of folding. This suggests that the central region plays a significant role by folding first, then accumulating other parts to become a native structure. It is interesting that our prediction methods, including sequence-based and 3D-based techniques, could extract folding properties that correspond well to experimental data.[42,48,53]

3.5 Conclusion

The initial folding processes of Ig-like beta-sandwich proteins have been investigated in this study by means of inter-residue average distance statistic methods, ADM, and F-value analysis as well as by 3D-based Gō-model simulation. In this present study, the initial folding unit of titin protein, including the Ig domain and FN3 domain, was investigated. Furthermore, an evolutionary analysis also performed by using sequence-based and structure-based multiple sequence alignment for every domain in the titin chain. The conservation of predicted folding units and hydrophobic residues was studied. Interestingly, the predicted results correspond well to the previous experimental ϕ-value and protection factor based on H-D exchange. The present Gō-model simulation studies confirm that the central region of six samples from Ig domains and FN3 domains was predicted to be an initial folding unit that forms the compact structure in the early event. This common feature is in line with the available experimental results of 1TIT and 1TEN, which also detected the stability of the central unit and the fluctuation of both terminal ends. Interestingly, the results underscore the importance of the common structure of these proteins, especially the key strands for folding and the Greek-key motif. Moreover, the difference in the folding pathways and the whole story of the protein folding processes can be described by the present Gō-model simulations. Consequently, the techniques used in this study are capable to decode the protein folding mechanism, especially in the initial state of folding.

Chapter 4

Study of folding mechanisms for beta-trefoil fold proteins

4.1 Introduction

The 3D-structure of a beta-trefoil fold protein shows the remarkable property of a pseudo three-fold symmetry without clear hydrophobic packing, and it exists ubiquitously in the protein structure space as a member of the superfold proposed by Orengo et al.[66] exhibiting various functions. The first discovered example of such a protein with this fold was soybean trypsin inhibitor, the 3D-structure of which was determined by Sweet et al.[67] The specific properties of the 3D-structure of this protein were described in detail by McLachlan.[68] Murzin et al.[69] proposed calling this specific 3D-structure the "beta-trefoil fold".

Figure 1A represents an example of the 3D-structure of a beta-trefoil protein, that is, 29-kDa galactose-binding lectin with the PDB code of 2RST, and each structural unit or subdomain exhibiting the three-fold symmetry is also shown in Figure 1B. A schematic drawing of the topology of a beta-trefoil protein consisting of 12 β-strands with intervening loops forming the trefoil shape is presented in Figure 1C. The six β-strands, β1, β4, β5, β8, β9 and β12, form the barrel structure (β-strands colored orange) and the rest of the β-strands constitute three β-hairpins (hairpin triplets), β2-β3, β6-β7 and β10-β11 (β-strands colored green). However, there are some beta-trefoil proteins containing irregular structures. In such a protein, the three-fold symmetry is partly disturbed. The schematic drawing of two irregular beta-trefoil proteins is presented in Figure 2A and 2B.

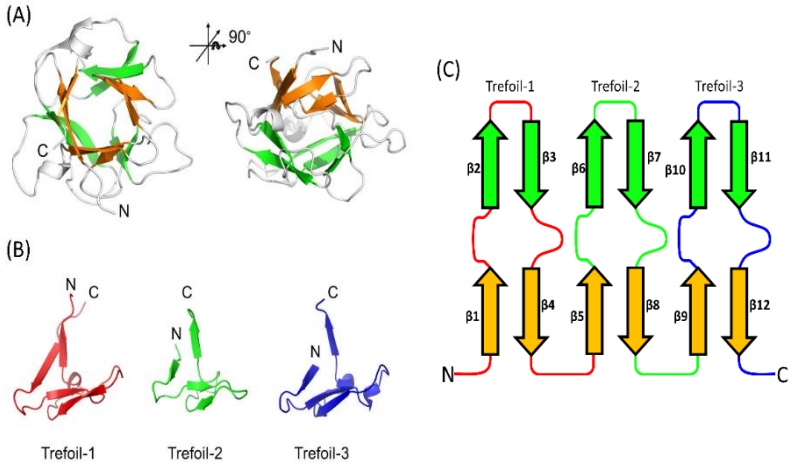


Figure 1 (A) 3D-structure of a protein with the beta-trefoil fold (29-kDa galactose-binding lectin (PDB ID:2RST) as an example). β-Strands colored orange form a barrel structure and β-strands indicated by green color are so-called cap strands. (B) Fundamental structural unit in the beta-trefoil fold. The red, green, and blue units are the segments M1-V53, I54-N89, and T90-E132 in 2RST. (C) Schematic drawing of beta-trefoil topology.

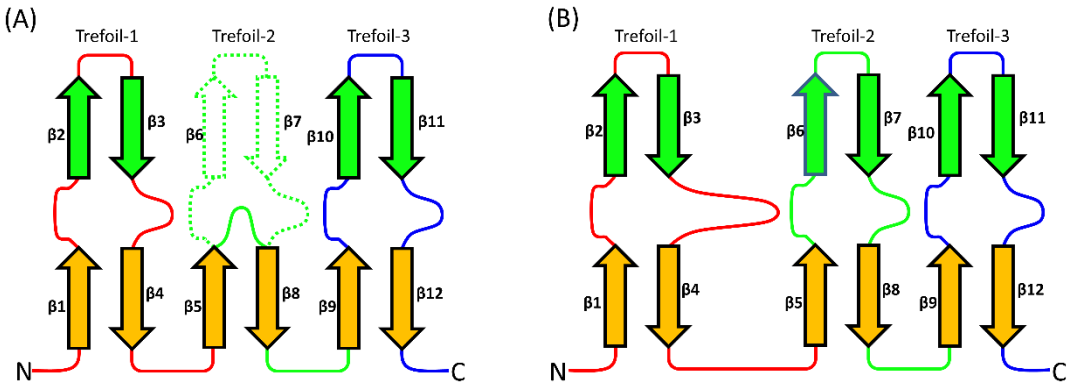


Figure 2 The schematic drawing of irregular beta-trefoil protein. (A) presents a beta-trefoil protein with a deletion of β6 and β7, while (B) presents the partly insertion between β3 and β4.

According to the SCOPe database, the beta-trefoil fold can be grouped into eight superfamilies.[8] The sequence identity among proteins from different superfamilies in the beta-trefoil fold is rather low as shown in Table 1. It is quite interesting how proteins from different superfamilies in the beta-trefoil assemble into the same beta-trefoil topology in spite of the rather low sequence identity. This suggests that only a small number of amino acids is determinant of the beta-trefoil fold. Such amino acids may be conserved during evolution.

Table 1: Sequence identity among the selected proteins from six superfamilies, cytokine (PDB ID: 2K8R), ricin B-like lectins (PDB ID: 2RST), STI-like (PDB ID: 3BX1_D), actin-crosslinking proteins (PDB ID: 1HCD), MIR domain (PDB ID: 1T9F), and DNA-binding protein LAG-1 (CSL) (PDB ID: 1TTU), in the beta-trefoil fold.

		RMSD					
		1HCD	1TTU	2RST	1T9F	2K8R	3BX1
Sequence identity	1HCD		9.35	7.76	13.65	3.75	10.56
	1TTU	12.22		15.41	11.66	14.17	5.66
	2RST	12.61	9.47		1.61	6.15	13.83
	1T9F	9.35	14.75	5.88		8.76	12.99
	2K8R	14.41	9.78	10.58	8.57		12.71
	3BX1	10.48	8.20	9.57	13.61	12.15	

Thus, a beta-trefoil protein attracts the interests of many researchers. McLachlan (1979),[68] Murzin et al. (1992),[69] Ponting et al. (2000),[70] Lee et al. (2011),[71,72] Broom et al. (2012),[73,74] Longo et al. (2012)[75,76] and Xia et al. (2016)[77] pointed out that the beta-trefoil structure might be constructed by triplication of the gene produced by gene duplication from a trefoil unit during evolution. From the analyses of sequences and 3D-structures of beta-trefoil proteins by Murzin et al.[69] and Ponting et al.,[70] and from the

success to design protein sequences exhibiting the beta-trefoil structure by Lee et al. (2011),[71,72] Broom et al. (2012),[73,74] Longo et al. (2012)[75,76] and Xia et al. (2016),[77] it is believed that the present beta-trefoil proteins evolved from a common ancestor of a homotrimer protein. Longo et al.[75,76] and Xia et al.[77] revealed based on ϕ-value analyses that folding nuclei of fibroblast growth factor 1 (FGF-1), that is, β2-β5 are significant units in the initial stage of folding. They designed and produced several kinds of artificial beta-trefoil proteins: a protein with three-repeat partial sequences, "Symfoil" using the Top-Down Symmetric Deconstruction method, a beta-trefoil protein formed by a homotrimer called "Monofoil",[71,72,77] and a protein having the same folding nuclei to those in fibroblast growth factor 1 (FGF-1) detected by ϕ-value analysis. The latter protein is called "Phifoil".[75-77]

Folding experiments of the following beta-trefoil proteins were performed extensively: fibroblast growth factor 1 (FGF-1), interleukin-1 beta (IL-1β) in cytokine superfamily and hisactophilin-1 (His) in the actin-crosslinking proteins superfamily. The ϕ-value analyses[75-77] for FGF-1, H/D exchange experiment by NMR[78-81] and Gō model simulations[82,83] for FGF-1, IL-1β and His were performed to identify their folding mechanisms. These studies revealed the differences in overall folding pathways among these proteins, while these proteins commonly start to fold at the central β-strands.[75,79] Li et al.[84,85] and Feng et al.[86] conducted multiple sequence alignments of selected beta-trefoil proteins based on their 3D-structures and proposed key residues in the three symmetrical units to form the beta-trefoil fold.

How the information on the folding mechanism to form such the beta-trefoil topology is encoded in the amino acid sequence of a protein is a very interesting problem. Moreover, the problem whether a β-trefoil protein with an irregular structure folds into its native structure via folding pathway same to a β-trefoil protein with high symmetry is remarkable. It is interesting to identify the significant residues to form the beta-trefoil unit from their sequences by using a kind of contact map and contact frequency prediction of a residue in a protein in random state based on inter-residue average distance statistics. These methods can predict the folding properties of proteins. Using these techniques, significant parts for folding in an amino acid sequence can be detected with relatively high accuracy. Furthermore, conserved hydrophobic residues also investigate in representative proteins in the beta-trefoil fold. However, the sequence identities of these proteins are rather low, and making accurate multiple alignment based on only sequences is difficult. Thus, the information of the 3D-structures was used as input data for 3D-based multiple sequence alignment. The information of conserved hydrophobic residues in combination with the results based on the average distance statistics is used to detect the significant residues to form the beta-trefoil structure in spite of the low sequence identity. The hydrophobic packing that formed by conserved hydrophobic residues in the native structures is used to extract the residues significant for folding.

4.2 Target proteins

The beta-trefoil fold proteins were selected according to the classification of SCOPe 2.05.[8] A protein with one domain was selected including a protein with an irregular structure. Sequences in a same superfamily were aligned with MAFFT[87] and classified into several groups of sequences. A sequence in a group was selected arbitrarily as the representative according to the experimental data available. As a result, the 30 proteins belonging to six superfamilies (Table 2) were selected for the present study.

In this section, fibroblast growth factor 1 (FGF-1) (PDB ID: 2K8R), interleukin-1 beta (IL-1β) (PDB ID: 6I1B), and hisactophilin-1 (His) (PDB ID: 1HCD), were treated because these proteins have been investigated extensively to reveal their folding mechanisms by experimental techniques. Furthermore, three high symmetric proteins and four irregular proteins which an experimental data on folding have not yet been reported are also selected, that is, 29-kDa galactose-binding lectin (PDB ID: 2RST), alpha-amylase/subtilisin inhibitor (PDB ID: 3BX1) and protein R12E2.13 from C. elegans (PDB ID: 1T9F), and CSL bound to DNA (PDB ID: 1TTU), Tetanus toxin (PDB ID: 1A8D), Clostridium neurotoxin type B (PDB ID: 1EPW), and Botulinum neurotoxin serotype A (PDB ID: 3BTA).

4.2.1 Experimental data on protein folding

Fibroblast growth factor 1 (FGF-1) (2K8R): FGF-1, also called heparin-binding growth factor 1, is classified in the fibroblast growth factors family in the cytokine superfamily in SCOPe. This protein binds to heparin at the residues K105-K121, called turn 11.[75,76,80] It has been reported that this functional site folds in the late stage of the folding process (the foldability-function tradeoff hypothesis).[75,76,83] In preceding studies, residues in β2 and β5-β8 are protected in the early stage of folding revealed by H/D exchange experiments of NMR.[79,80] Longo et al.[75,76] and Xia et al.[77] reported that the folding nucleus detected by the ϕ-value analyses consists of residues L16-L58 in β2-β6.

Interleukin-1 beta (IL-1β) (6I1B): This protein is a kind of inflammatory cytokine and classified to the interleukin-1 family of the cytokine superfamily according to SCOPe. The studies of the H/D exchange experiments[78,81] were done previously. According to these studies, protection of β6-β10 occurs in the early stage of folding followed by closure of the barrel structure formed by β1-β4 and β11-β12.[78,81]

Hisactophilin-1 (His) (1HCD): Hisactophilin-1 (His) is a kind of actin binding protein, and its sequence contains 31 histidine residues out of total 118 residues, that is, this is a histidine-rich protein in comparison with other beta-trefoil proteins. Another characteristic of this protein is that it contains shorter loops and β-strands compared with other beta-trefoilproteins.[78] This is classified into the histidine-rich actin-binding protein (hisactophilin) family in the actin-crosslinking proteins superfamily in the SCOP database. H/D exchange experiments and Gō model simulations were also performed for this protein so far by Liu et al.[78] and Chavez et al.[82] The results of the H/D exchange

experiments show that this protein folds at β4-β8 at the beginning and β1-β3 and β10-β12 are structured in the end of the folding,[78] while Gō model simulations suggest that the folding proceeds at the central β-strands (Trefoil-2) and the C-terminal region (Trefoil-3).[82]

Table 1 Target proteins in this study.

Superfamily	PDB ID	UniProt ID	Protein name (UniProtKB)	Sequence length
Cytokine	2K8R	P05230	Fibroblast growth factor 1	133
	1Q1U	P61328	Fibroblast growth factor 12	138
	2FDB_M	P55075	Fibroblast growth factor 8	147
	1QQK	Q02195	Fibroblast growth factor 7	129
	1J0S	Q14116	Interleukin-18	157
	6I1B	P01584	Interleukin-1 beta (*Homo sapiens* [*Human*])	153
	1MD6	Q9QYY1	Interleukin-36 receptor antagonist protein	154
	2KKI	P01583	Interleukin-1 alpha	151
	2WRY	O73909	Interleukin-1 beta (*Gallus gallus* [*Chicken*])	155
	2P39	Q9GZV9	Fibroblast growth factor 23	142
	2P23	O95750	Fibroblast growth factor 19	136
Ricin B-like lectins	2RST	O96048	29-kDa galactose-binding lectin	132
	1SR4_A	O06522	Cytolethal distending toxin subunit A	167
	1SR4_C	O06524	Cytolethal distending toxin subunit C	154
	1KNM	P26514	Endo-1,4-beta-xylanase A	129
	1DQG	Q61830	Macrophage mannose receptor 1	134
STI-like	3BX1_D	P07596	Alpha-amylase/subtilisin inhibitor	181
	1TIE	P09943	Trypsin inhibitor DE-3	166
	1R8N	P83667	Kunitz-type serine protease inhibitor DrTI	185
	1WBA	P15465	Albumin-1	171
	2GZB	P83051	Kunitz-type proteinase inhibitor BbCI	164
	3ZC8	D2YW43	Trypsin inhibitor	182
	3TC2	Q8S380	KTI-B protein	181
	1A8D	P04958	Tetanus toxin	205
	1EPW	P10844	Clostridium neurotoxin type B	211
	3BTA	P0DPI1	Botulinum neurotoxin serotype A	204
Actin-crosslinking proteins	1HCD	P13231	Hisactophilin-1	118
MIR domain	1T9F	O61793	Protein R12E2.13 (*Caenorhabditis elegans*)	176
	3HSM	P11716	Ryanodine receptor 1	164
DNA-binding protein LAG-1 (CSL)	1TTU	Q8MXE7	CSL bound to DNA	161

4.3 Results

4.3.1 ADM analyses

The average distance map of high symmetric beta-trefoil fold proteins is shown in Figure 3, including fibroblast growth factor 1 (FGF-1) (2K8R), interleukin-1 beta (IL-1β) (6I1B), hisactophilin-1 (His) (1HCD), 29-kDa galactose-binding lectin (2RST), alpha-amylase/subtilisin inhibitor (3BX1), and protein R12E2.13 from *C. elegans* (1T9F).

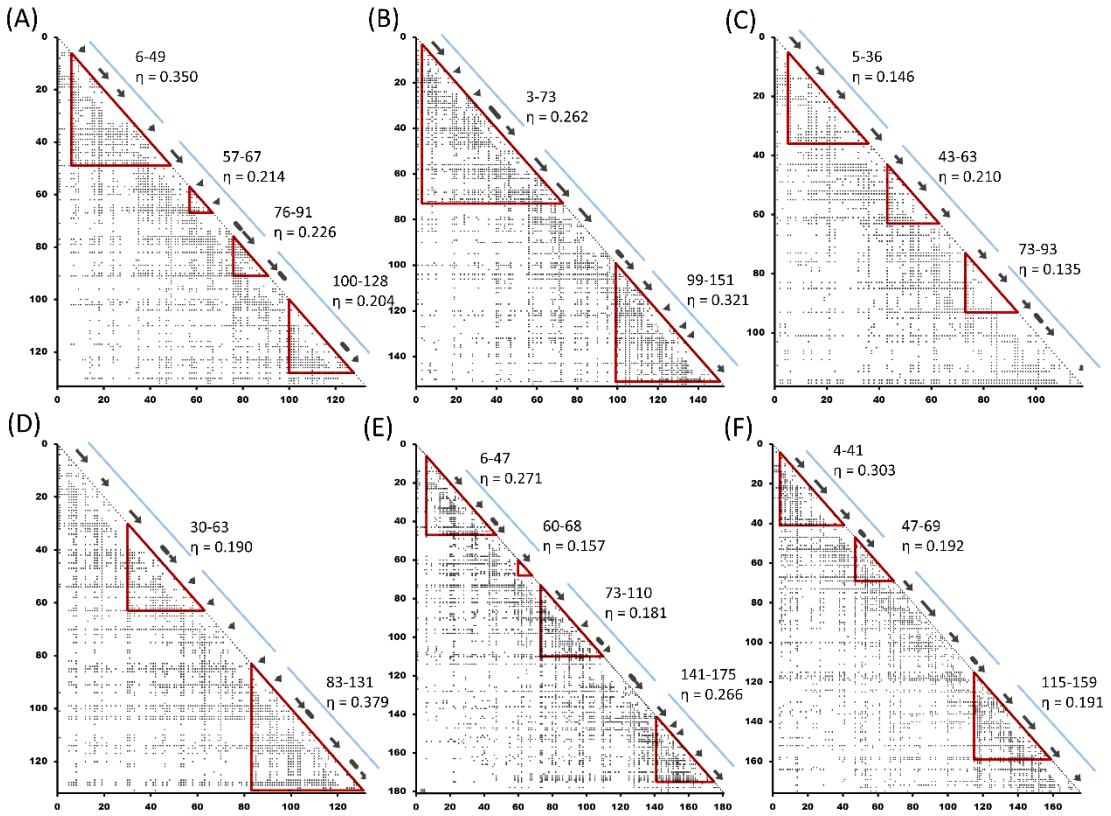


Figure 3 Average Distance Map (ADM) for six symmetric beta-trefoil proteins, (A) fibroblast growth factor 1 (FGF-1) (2K8R), (B) interleukin-1 beta (IL-1β) (6I1B), (C) hisactophilin-1 (His) (1HCD), (D) 29-kDa galactose-binding lectin (2RST), (E) alpha-amylase/subtilisin inhibitor (3BX1), and (F) protein R12E2.13 from C. elegans (1T9F). A black bar and a black arrow along the diagonal denote an α-helix and a β-sheet, respectively. A blue bar means the position of a trefoil unit. A red triangle represents a predicted compact region (PdCR) indicated by average distance statistics-based method.

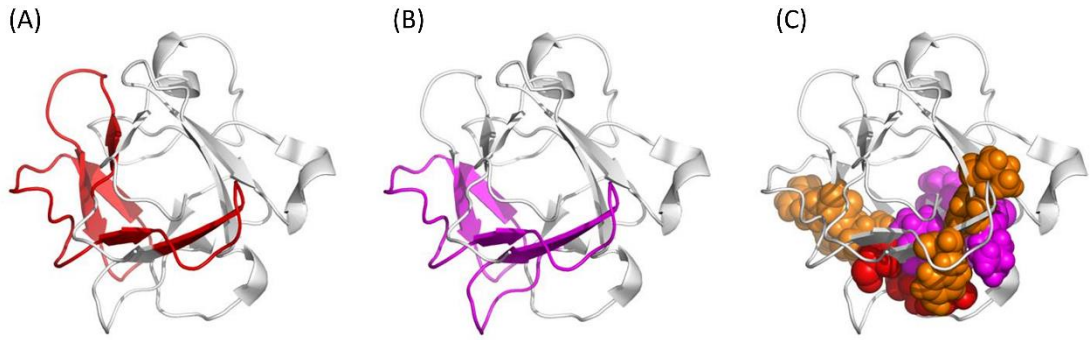


Figure 4 3D-structure of fibroblast growth factor 1 (FGF-1, PDB ID:2K8R). (A) PdCR 6-49 by ADM colored red. (B) Folding nucleus obtained by the ϕ-value analysis[75-77] indicated by the purple region. (C) The residues with high protection factor values in the H/D exchange experiment[80] are colored red, purple, or orange in the order from high to low protection factor values in the space-filling model.

Fibroblast growth factor 1 (FGF-1) (2K8R): In Figure 3A, the PdCRs predicted by the ADM are residues L6-I49 corresponding to β1-β5, residues Y57-Y67 corresponding to β6-β7, residues C76-I91 corresponding to β8-β9, and residues W100-L128 corresponding to β10-β12. It is noticed that the PdCRs L6-I49, Y57-Y67, C76-I91, and W100-L128 roughly correspond to the first, second, and third trefoil units, respectively. The PdCRs L6-I49, Y57-Y67, C76-I91, and W100-L128 are indicated as region-1, region-2, region-3 and region-4. In particular, the PdCR L6-I49 (a region colored red in Figure 4A) shows the highest η-value of 0.350, and this is expected to form stable compact structure in the initial stage of folding. All conserved hydrophobic residues are contained in the PdCRs (the number of all conserved hydrophobic residues is 15; the number of the residues within the predicted compact regions is 100 and the total number of residues is 133 as indicated in Table 3).

It is interesting that the predicted primary folding unit (residues L6-I49) corresponds well to residues L16-L58 where were identified as the folding nucleus by Long et al.[75,76] and Xia et al.[77] as shown in Figure 4B (in this figure the folding nucleus obtained by the ϕ-value analysis[75-77] is indicated by the purple region and in Figure 4C the residues with high protection factor values in the H/D exchange experiment[80] are colored red, purple, and orange in the order from high to low protection factor values in the space-filling model.) Thus, it is expected that a region predicted by ADM with the highest η-value forms a stable compact or a structured region in the early stage of folding.

Interleukin-1 beta (IL-1β) (6I1B): In Figure 3B, the PdCR V3-L73 (region-1) includes β1-β6 with η-value of 0.262 and the PdCR F99-V151 (region-2) including β8-β12 with that of 0.321 suggesting that the region F99-V151 is more stable in the early stage of folding. That is, the first PdCR corresponds to trefoil-1 and the N-terminus of trefoil-2, whereas the second PdCR corresponds to the C-terminus of trefoil-2 and trefoil-3. The PdCR by ADM for 6I1B contains 14 conserved hydrophobic residues out of 15 (the number of the residues within the PdCRs is 124, and the total number of residues is 153) as indicated in Table 3.

Hisactophilin-1 (His) (1HCD): In Figure 3C, the PdCR by ADM for 1HCD

includes 13 hydrophobic residues (the number of the residues within the PdCRs is 74 and the total number of residues is 118). Figure 3C presents that the PdCRs by ADM involve residues A5-V36 (η = 0.146, region-1) including β1-β4, V43-L63 (η = 0.210, region-2) including β5-β7, and L73-I93 (η = 0.135, region-3) and including β8-β10. These three PdCRs roughly correspond to trefoil-1 to trefoil-3. The central PdCR V43-L63 exhibits the highest η-value and is expected to be stable in the early stage of folding.

Galactose-binding lectin (2RST): The result of ADM analysis for 29-kDa galactose-binding lectin (2RST) is shown in Figure 3D. ADM predicts regions I30-I63 including β3-β6 (η = 0.190, region-1) and A83-S131 including β8-β12 (η = 0.379, region-2). Figure 3D suggests that the C-terminal part would be stable in early stage folding because of the larger η-value. The PdCRs for 2RST include 11 conserved hydrophobic residues as shown in Table 3 (the number of the residues within the predicted compact regions is 83 and the total number of residues is 132).

Alpha-amylase/subtilisin inhibitor (3BX1): Figure 3E presents result of ADM analysis for alpha-amylase/ subtilisin inhibitor (3BX1). The PdCRs by ADM are regions V6-V47 including β1-β3 (η = 0.217, region-1) and V60- A68 including β4 (η = 0.157, region-2) and I73-I110 including β5-β7 (η = 0.181, region-3) and L141-F175 including β9-β12 (η = 0.266, region-4). The PdCRs for 3BX1 include 14 conserved hydrophobic residues as shown in Table 3 (the number of the residues within the predicted compact regions is 124 and the total number of residues is 181 in Table 3). It is suggested from this figure that the C-terminal L141-F175 would be stable in early stage folding because of the larger η-value

Protein R12E2.13 (1T9F): Figure 3F indicates the result of ADM for protein R12E2.13 from C. elegans (1T9F). Regions F4-V41 cover β1-β3 (η = 0.303, region-1) and I47-C69 cover β4 (η = 0.192, region-2) and W115-V159 cover β8-β11 (η = 0.191, region-3). The PdCRs for 1T9F include 10 conserved hydrophobic residues as shown in Table 3 (the number of the residues within the predicted compact regions is 106 and the total number of residues is 176). The ratio of the conserved hydrophobic residues in the PdCRs to those in the whole sequence is 0.67, whereas the ratio of the number of residues within the predicted compact regions to the number of residues in the whole sequence is 0.6. Thus, the conserved hydrophobic residues tend to be included in the PdCRs also in this protein. It is suggested from Figure 3F that the N-terminal F4-V41 would be stable in the early stage of folding due to the larger η-value, and region I47-C69 may merge with the N-terminal region to fold.

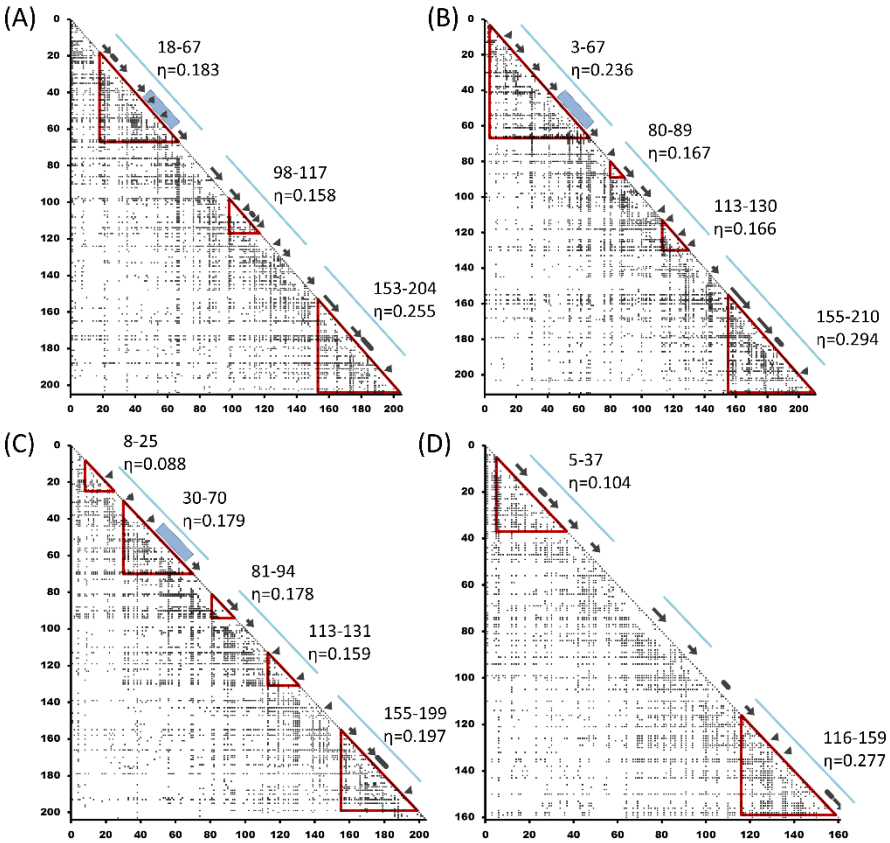


Figure 5 Average Distance Map (ADM) for four irregular beta-trefoil proteins, (A) Tetanus toxin (1A8D), (B) Clostridium neurotoxin type B (1EPW), (C) Botulinum neurotoxin serotype A (3BTA), and (D) CSL bound to DNA (1TTU). A black bar and a black arrow along the diagonal denote an α-helix and a β-sheet, respectively. A blue bar means the position of a trefoil unit. The blue rectangle along the diagonal indicates the inserted part. A red triangle represents a predicted compact region (PdCR) indicated by average distance statistics-based method.

Tetanus toxin (1A8D): Figure 5A indicates the result of ADM analysis. A compact region having the largest η-value of 0.255 is detected at A153-N204, containing β10-β12. Actually, the compact region I173-N204 exhibits the largest η-value of 0.258. However, that of A153-N204 is 0.255 and almost the same. Thus, the position of residue A153-N204 was predicted as a compact region in this study. There is a compact region with the second largest η-value at residue number Y18-I67, containing β1-β4. This compact area includes the inserted segment. These are the major PdCRs and correspond to trefoil unit 3 and unit 1, respectively. These regions are expected to form a stable compact region in the early stage of folding.

Clostridium neurotoxin type B (1EPW): Figure 5B indicates the result of ADM analysis of 1EPW. A region with residues I155-T210 is predicted as a PdCR with the largest η-value. This region contains β10-β12. The PdCR with the second largest η-value is Y3-I67, containing β1-β4. This compact area also contains the insertion part as shown by the blue rectangle along the diagonal. These regions also correspond to units 3 and 1 of the β trefoil, respectively. These regions are expected to form a stable compact region in the early stage of folding.

Botulinum neurotoxin serotype A (3BTA): Figure 5C indicates the result of ADM analysis. The PdCR with the largest η-value is I155-W199, which covers β10-β12 similar to 1EPW. The region with the second highest η-value is Y30-I70, with β2-β4. This compact area also contains the insertion part. Again, these regions correspond to trefoil unit 3 and unit1, respectively.

CSL bound to DNA (1TTU): Figure 5D presents the result of the ADM. Regions C116-I159, including β9-β12 and C5-A37 containing β1-β3, are predicted to be PdCRs with the highest and the second highest η-values. These correspond to unit 3 and unit 1 of the β trefoil, respectively, which are expected to form stable compact regions in the early stage of folding. No compact area is found in the center, which includes the lacking site.

The conserved hydrophobic residues statistics of irregular beta-trefoil proteins (1A8D, 3BTA, 1EPW and 1TTU) also present in Table 3. The conserved hydrophobic residues tend to be included in the PdCRs. The ratio of hydrophobic residues in the PdCRs to those of the whole sequence shows rather high except 1TTU which presents only 0.53 due to the deletion part in the second trefoil unit that caused the undetectable of PdCR in this region.

Table 3 Statistics of conserved hydrophobic residues.

PDB ID	Sequence length	Total number of residues in the predicted compact regions by ADM	Ratio of total number of residues in the predicted compact regions by ADM to the total number of residues	Number of conserved hydrophobic residues in the predicted compact regions by ADM	Ratio of the number of conserved hydrophobic residues in the predicted compact regions by ADM to the total number of conserved hydrophobic residues	Number of conserved hydrophobic residues out of the predicted compact regions by ADM	Total number of packing formed by conserved hydrophobic residues
2K8R	133	100	0.75	15	1.00	0	43
1Q1U	138	106	0.77	14	0.93	1	44
2FDB_M	147	82	0.56	12	0.80	3	40
1QQK	129	71	0.55	10	0.67	5	38
1J0S	157	124	0.79	15	1.00	0	36
6I1B	153	124	0.81	14	0.93	1	44
1MD6	154	112	0.73	13	0.87	2	37
2KKI	151	111	0.74	13	0.87	2	42
2WRY	155	104	0.67	11	0.73	4	40
2P39	142	114	0.80	13	0.87	2	40
2P23	136	93	0.68	13	0.87	2	39
2RST	132	83	0.63	11	0.73	4	44
1SR4_A	167	102	0.61	10	0.67	5	38
1SR4_C	154	125	0.81	14	1.00	0	31
1KNM	129	74	0.57	10	0.67	5	41
1DQG	134	114	0.85	14	0.93	1	43
3BX1_D	181	124	0.69	14	0.93	1	33
1TIE	166	86	0.52	12	0.80	3	35
1R8N	185	130	0.70	13	0.87	2	38
1WBA	171	139	0.81	15	1.00	0	36
2GZB	164	89	0.54	12	0.80	3	38
3ZC8	182	117	0.64	11	0.73	4	39
3TC2	181	125	0.69	14	0.93	1	36
1HCD	118	74	0.63	13	0.87	2	36
1T9F	176	106	0.60	10	0.67	5	40
3HSM	164	107	0.65	15	1.00	0	36
1A8D	205	122	0.60	11	0.73	4	33
1EPW	211	149	0.71	13	0.87	2	38
3BTA	204	137	0.67	13	0.87	2	34
1TTU	161	77	0.48	8	0.53	7	31

4.3.3 Evolution analyses

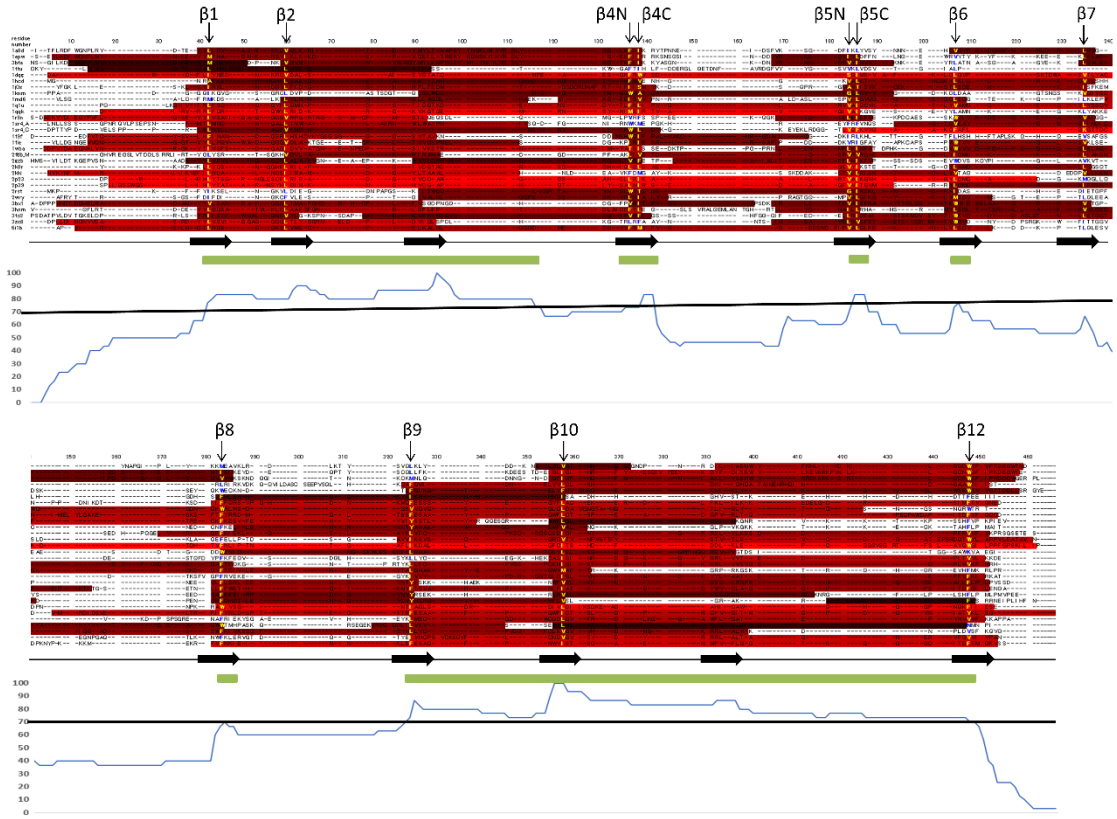


Figure 6 3D-structure-based multiple sequence alignment of 30 beta-trefoil proteins, four irregular beta-trefoil proteins and 26 structurally symmetric beta-trefoil proteins and a histogram of the conserved PdCR derived from ADMs. The portions shown in red are regions predicted by ADM. The brighter red means a higher η-value. The conserved hydrophobic residues within the predicted region are shown in yellow and outside the predicted region in blue. Text and arrow above an alignment indicates the position of conserved hydrophobic residues. Black arrow under an alignment indicates position of β-strand. Conserved PdCR areas above 70% are indicated by a green bar on the histogram (a clear picture presents in Figure A13 in appendix section).

The result of the structure-based multiple sequence alignment of 30 sequences shown in Figure 6 by means of STRAP,[23] in order to identify the conserved hydrophobic residues and elucidate the common regions by ADM predictions. The sequence identities of these proteins are rather low (3-37%), so making the accurate multiple alignment is difficult. Thus, the 3D-structure information is used to make multiple alignment. The sequence identities showing around 25-35% identity within the same superfamily but only about 10% identity between proteins from different superfamilies. A predicted compact region is indicated by a red bar. Brighter red denotes higher η-value. The conserved hydrophobic residues in the alignment are indicated by a yellow letter in a predicted compact region and by a blue letter out of a predicted compact region. However, almost all pairs of beta-trefoil proteins exhibit RMSD of about 3Å indicating high similarity of their 3D-structures in spite of the low sequence identities.

Twelve conserved hydrophobic residues are identified when the conservation of hydrophobic residue exceeds 90%, and all conserved hydrophobic residues tend to locate on a β-strand as shown in Figure 6. Almost all β-strand contains one or two conserved hydrophobic residues except β-strand 3 and 11. These 12 conserved hydrophobic residues were labeled by the numbers of the β-strands. For example, the conserved hydrophobic residue in the β1 is labeled by β1. The β-strand 4 and 5 contain two conserved hydrophobic residues. In the case of the β4, a conserved hydrophobic residue located in the N-terminal side of β-strand is labeled as β4N and in the same way that in the C-terminal side is labeled as β4C. The result confirms that these conserved hydrophobic residues correspond well to the conserved residues proposed by Murzin et al. and Feng et al.[69,86] Murzin et al. defined the conservation of hydrophobic residues in two proteins from the Cytokine superfamily and one protein from STI-like superfamily. Thus, multiple alignment of sequences from various superfamilies leads to the similar results. Feng et al.[86] performed the structure-based sequence alignment for beta-trefoil fold proteins. A similar result is obtained in this present study, and the results reveal clearer conservation by focusing on hydrophobic residues Ala, Val, Leu, Ile, Met, Phe, Trp and Tyr.

A histogram denoting the conserved PdCR derived from ADMs. There is variety in the predicted patterns of the PdCR in each protein. However, the histogram indicates some conserved regions of the compact regions with more than 70% conservation as shown in Figure 6, that is, β1-β3, β4, β5, β6, β8 and β9-β12. The threshold of 70% conservation is indicated by a green line at the histogram. These regions are corresponding moderately well to trefoil-1, trefoil-2 and trefoil-3, respectively. Although a portion with low conservation between β5-β8 in trefoil-2 is also observed, this portion is a relatively long loop and seems not to be conserved during evolution.

4.3.2 F-value analyses

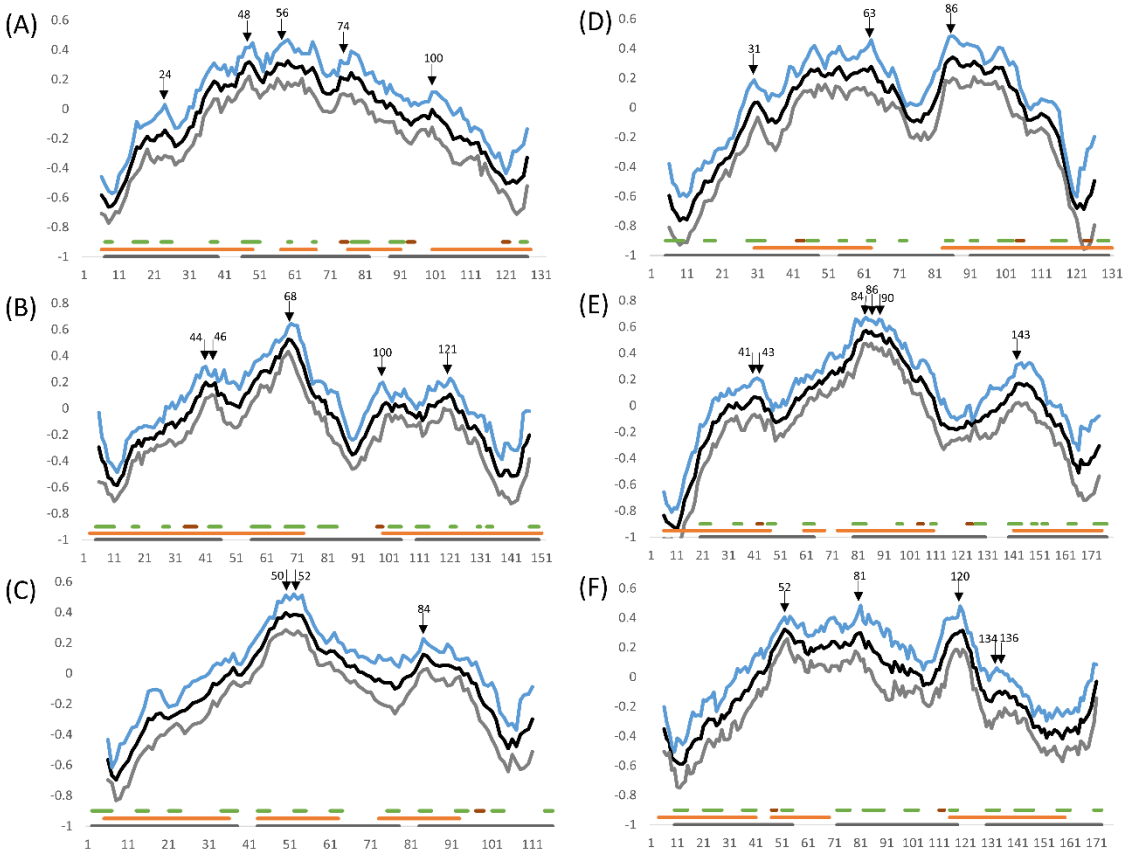


Figure 7 F-value plots of six high symmetric beta-trefoil proteins, (A) 2K8R, (B) 6I1B, (C) 1HCD, (D) 2RST, (E) 3BX1, and (F) 1T9F with the position of the peaks indicated by black arrows. The orange, black, red and green bars near the x-axis represent the PdCR, trefoil unit, α-helix and β strand, respectively. In all figures, F, F + SD, and F - SD for each residue are plotted as black, blue, and gray lines, respectively.

Fibroblast growth factor 1 (FGF-1) (2K8R): The high peaks in the F-value plot appear around β5-β8 as presented in Figure 7A. This area coincides with the highly protected region measured by NMR, that is, β5-β8 as shown in Figure 8. The highest peak of F-value plot at residue Y48 appeared in this region that close to the highest protection value of the H/D protection factors. Therefore, the peak in the F-value plot can be considered as a site to be structured in the early stage of folding based on the comparison with the H/D exchange experiment in this protein. It should be noted that the conserved hydrophobic residues β5N, β5C, β6, β8 and β10 are near the peaks of the F-value plot. Considering this result, the PdCRs L6-I49 and Y57-Y67 by ADM can be regarded as a compact region in the early stage of folding due to the high η-value of position L6-I49

and the highest peak of F-value at residues Y48 and Q56.

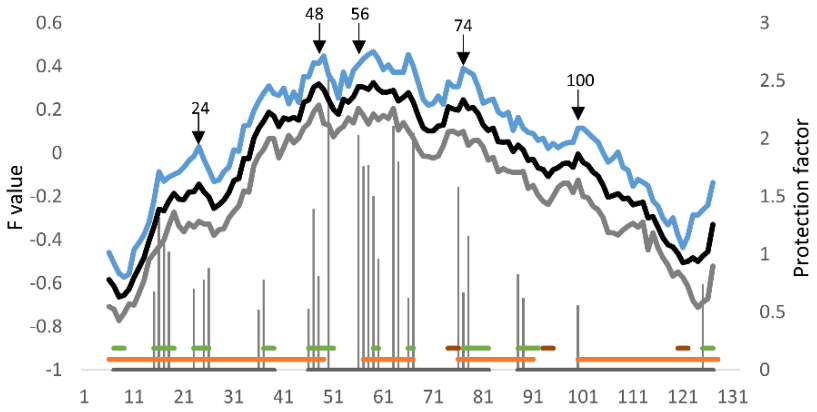


Figure 8 F-value plots and H/D exchange experiment[80] of Fibroblast growth factor 1 (FGF-1) (2K8R) with the position of the peaks indicated by black arrows. The histogram means the protection factor value obtained by the H/D exchange experiment.[80] The orange, black, red and green bars near the x-axis represent the PdCR, trefoil unit, α-helix and β strand, respectively. In all figures, F, F + SD, and F - SD for each residue are plotted as black, blue, and gray lines, respectively.

Next, packing formed by conserved hydrophobic residues for each protein is examined. In Figure 8, the following description applies. A conserved hydrophobic residue is indicated by a number or a number with N or C. Conserved hydrophobic residues near the highest peak in the F-value plot are placed in the blue cells. A black circle means a contact, and a red circle means a contact formed by a residue in a blue cell with a residue in a different predicted compact region. Conserved hydrophobic residues near the highest peak in the F-value plot tend to form contacts with conserved hydrophobic residues within various predicted compact regions and also linking predicted compact regions.

As shown in Figure 7A, the highest peaks in the F-value plot of 2K8R are around position β5N, β5C and β6. Conserved hydrophobic residues β5N and β5C are in the predicted compact region L6-I49 by ADM, and residue β6 exists in the predicted compact region Y57-Y67. Figure 9A shows a kind of contact map for just the conserved hydrophobic residues. That is, a plot is made when two conserved hydrophobic residues form a packing. The residues near the peak in the F-value plot, β5N, β5C and β6 form contacts widely from predicted region-1 to region-3 (the contacts are labeled by red circle).

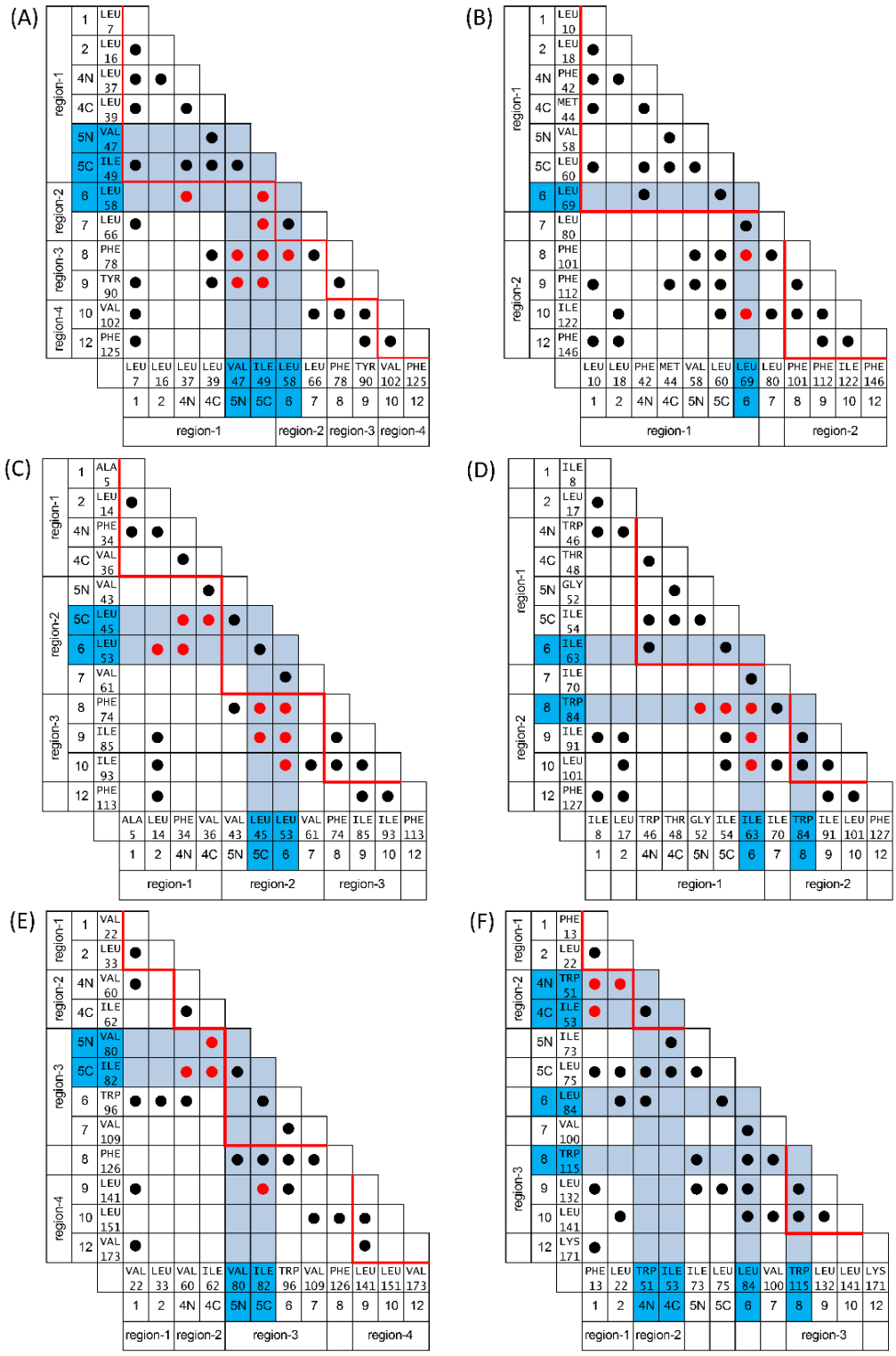


Figure 9 The contact maps present the interaction between conserved hydrophobic residues of six high symmetric beta-trefoil proteins, (A) 2K8R, (B) 6I1B, (C) 1HCD, (D) 2RST, (E) 3BX1, and (F) 1T9F. The blue region indicates the conserved residues within ± 5 residues of the highest F-value peak. A dot in the contact map marks the contact between conserved hydrophobic residues, and red denotes the contact between predicted regions. The red line denotes the conserved PdCR.

Wang and Yu[80] demonstrated in their study that β5, β6 and β7 (including V47, Y48, I49, Y57, L58, L65, L66 and Y67) form a hydrophobic core with Y57, L58 and A59 as center of the hydrophobic core in the native FGF-1 structure. Our prediction shows that the center of folding of 2K8R is β5N, β5C and β6 with the conserved hydrophobic residues V47, I49 and L58 and coincides well to the results of Wang and Yu. It is noted that the heparin binding site of this protein, K105-K121, no such conserved hydrophobic residue, and this fact seems to reflect that this part is not strongly involved in the folding indicating the foldability-function tradeoff hypothesis.[75,76,83]

Interleukin-1 beta (IL-1β) (6I1B): The highest peak is on β6 as shown in Figure 7B. According to Liu et al. and Capraro et al.,[78,81] the folding of 6I1B occurs at β6-β10, and the present results show that β8-β12 would be stable in the early stage of folding (the η-value of the region F99-V151 including β8-β12 is higher) reflecting the results of Liu et al. and Capraro et al.,[78,81] although β6 is included in the first compact region. The highest peak in the F-value plot is in β6 indicating the strong involvement of this β-strand in folding. Suggesting that the significant parts for the folding in this protein are region F99-V151 and β6, and this speculation corresponds to the result of Liu et al. and Capraro et al.[78,81] The conserved hydrophobic residues β4N, β4C, β6, β8 and β10 are near the peaks of the F-value plot for 6I1B as shown in Figure 7B and Table G5 in appendix section.

The packing formed by these conserved hydrophobic residues is presented in Figure 9B. This figure shows that a conserved hydrophobic residue β6 form packing within the predicted region-1 and also with the conserved hydrophobic residues in predicted region-2, position β8 and β10, indicating the significance of this residue for the 3D-structure formation of 6I1B (these contacts are indicated by a red circle).

Hisactophilin-1 (His) (1HCD): The F-value plot of hisactophilin shows the highest peak at the β6 in the second predicted compact region as shown in Figure 7C, suggesting the frequent contact formations in the early stage of folding. The results of both the ADM and the F-value plot analyses reflect those of the H/D exchange experiments that indicate the formation of the β4-β8 at the initial state of folding. The conserved hydrophobic residues β5C, β6, β9, β10 and β12 are near the peaks of the F-value plot as shown in Figure 7C and Table G5 in appendix section. The packing formed by these residues is presented in Figure 9C. Residues β5C and β6 on predicted region-2 form packing with predicted region-1 and region-2. This result emphasizes an ability of F-value prediction method.

Galactose-binding lectin (2RST): The F-value plot of this protein presents in Figure 7D. The high peaks in the F-value plot appear around conserved hydrophobic residue β6 and β8. A peak near residue β6 is again observed in this protein as the same for the other three proteins (fibroblast growth factor 1 (FGF-1), interleukin-1 beta (IL-1β) and hisactophilin-1 (His)) suggesting this residue is significant for the initial folding. Among the conserved hydrophobic residues in 2RST, β6 and β8, are near the highest peaks as presented in Figure 7D. Conserved residue β6 is in the predicted compact region I30-I63 (region-1) and residue β8 is located in the predicted compact region A83-S131 (region-2). Those conserved hydrophobic residues make the hydrophobic interactions within region-1 and region-2 or between these two regions as shown in Figure 9D. Thus, the conserved hydrophobic residues in these regions are considered to be significant for

the 3D-structure of this protein.

Alpha-amylase/subtilisin inhibitor (3BX1): The F-value profile of 3BX1 is shown in Figure 7E. The highest peak in the F-value plot appears around conserved residues β5N, β5C and β9. Region I73-I110 would also form a stable compact region but would be weaker compared to the region L141-F175. The conserved hydrophobic residues near the highest peaks on β5 may interact with the residues in the region L141-F175 and region I73-I110 may merge with region L141-F175. Among the conserved hydrophobic residues in 3BX1, the highest peaks of the F-value plot are observed around β5N and β5C. These residues are in the predicted compact region I73-I110 (region-3). Figure 9E indicates that these conserved hydrophobic residues form packing with other conserved hydrophobic residues in region-2 and region-3.

Protein R12E2.13 (1T9F): In Figure 7F, the high peaks of F-value plot appear around conserved residue β4N, β4C, β6, β8, β9 and β10. However, there are three highest peaks indicated for this protein according to a similar value of F-value result. Among the conserved hydrophobic residues in 1T9F, the conserved hydrophobic residues near the highest peaks of the F-value plot are residues β4N, β4C, β6 and β8. Conserved residue β4N and β4C are included in the predicted compact region I47-C69 (region-2) and CHR-β8 is in the predicted compact region W115-V159 (region-3). Figure 9F presents the packing formed by these conserved hydrophobic residues. These four conserved residues near the highest peak form packing with other conserved hydrophobic residues.

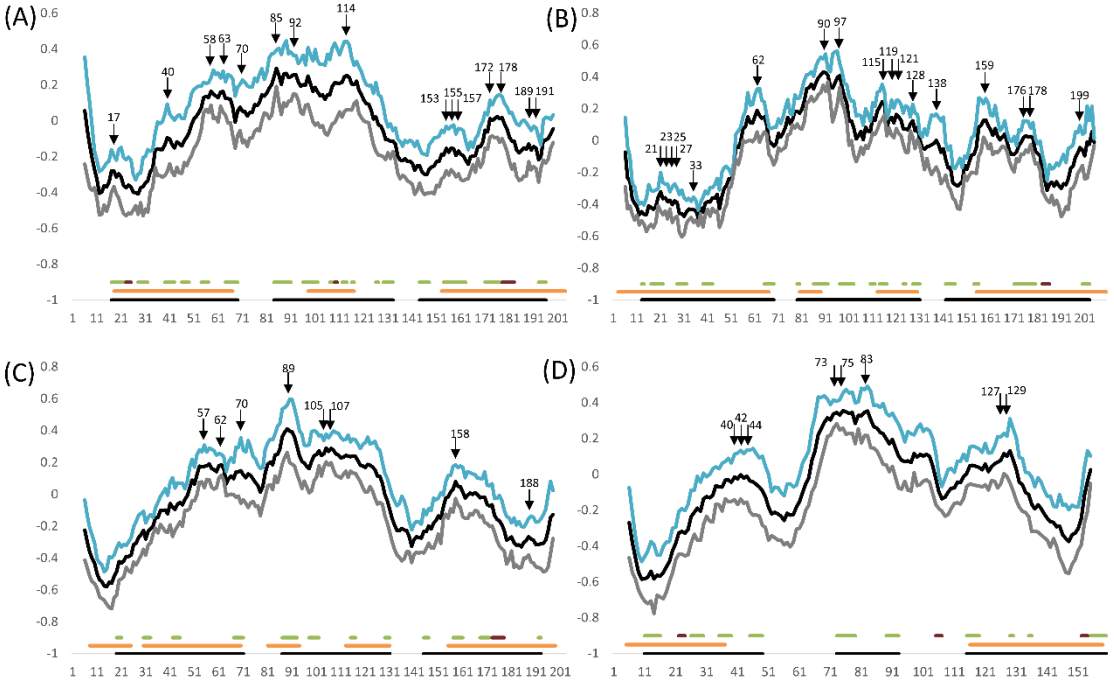


Figure 10 F-value plots of four irregular beta-trefoil proteins, (A) 1A8D, (B) 1EPW, (C) 3BTA and (D) 1TTU with the position of the peaks indicated by black arrows. The orange, black, red and green bars near the x-axis represent the PdCR, trefoil unit, α-helix and β strand, respectively. In all figures, F, F + SD, and F - SD for each residue are plotted as black, blue, and gray lines, respectively.

Next, four irregular beta-trefoil proteins, including three proteins from STI-like superfamily (1A8D, 1EPW and 3BTA) with an insertion part between β strands 3 and 4 as marked by a blue rectangle along the diagonal of predicted ADM in Figure 5, and one protein from DNA-binding protein LAG-1 (CSL) superfamily (1TTU) with a deletion part of β strands 6 and 7, were analyzed by the mean of F-value analyze and conserved hydrophobic packing.

Tetanus toxin (1A8D): The F-value profile of 1A8D is shown in Figure 10A. The highest peak of F-value plot appears close to β strand 5, 6 and 7. Considering that the conserved hydrophobic residues β5N, β5C and β7 are close to the highest F-value peak within ±5 residues. The results suggest that these conserved residues of the central unit are important for the structure formation of this protein. Residue β5C forms packing with β4N and β4C in the predicted region-1, and with residue β6 in region-2. On the other hand, residue β7 interacts with β10 in region-3 and thus, these residues are thought to be significant for packing with the predicted regions in the C-terminal part.

Clostridium neurotoxin type B (1EPW): Figure 10B presents an F-value profile of 1EPW. A region with residues I155-T210 is predicted as a compact region with the largest η-value. The predicted region with the second largest η-value is Y3-I67. These regions also correspond to units 3 and 1 of the beta-trefoil protein. These regions are expected to form a stable compact region in the early stage of folding. The highest peak was indicated on the β strand 5 as presented in Figure 10B. This figure illustrates that

residue β5C forms packing with β1, β4N, and β4C in the predicted region-1, and β8 and β10 in region-3 and region-4, respectively. Thus, CHR- β5C is supposed to be a significant residue for the packing with the predicted regions in the C-terminal and N-terminal parts.

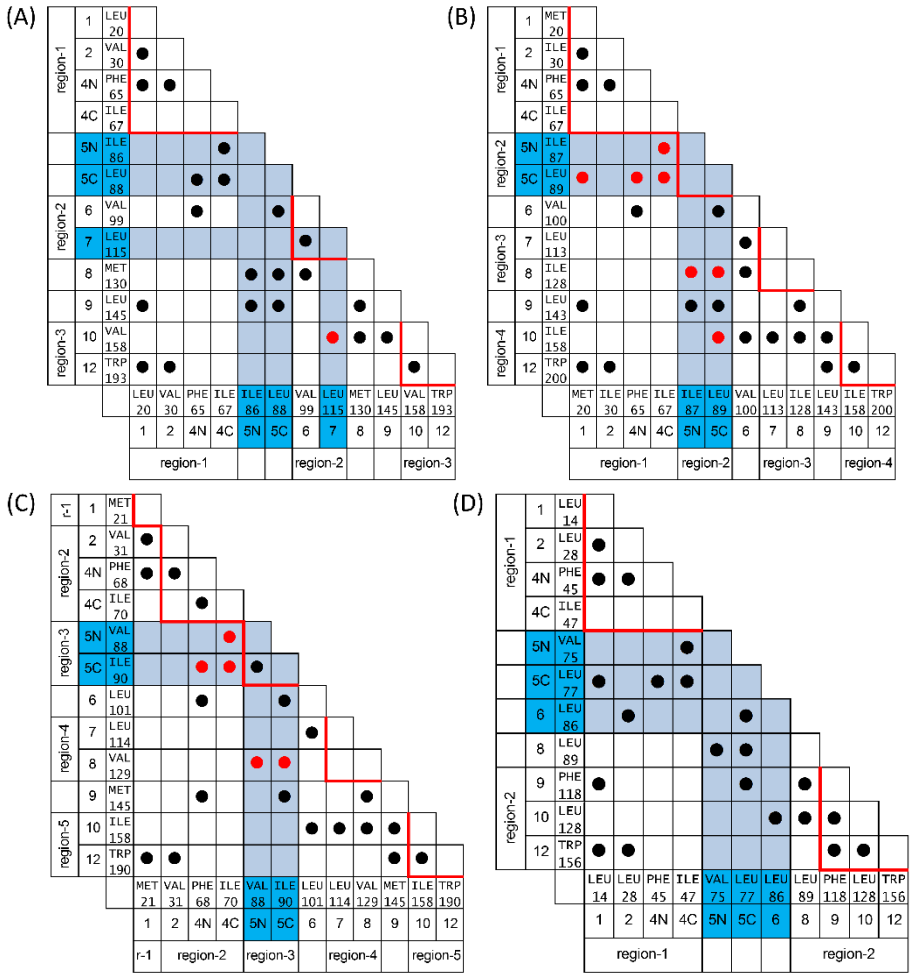


Figure 11 The contact maps present the interaction between conserved hydrophobic residues of four irregular beta-trefoil proteins, (A) 1A8D, (B) 1EPW, (C) 3BTA and (D) 1TTU. The blue region indicates the conserved residues within ±5 residues of the highest F-value peak. A dot in the contact map marks the contact between conserved hydrophobic residues, and red denotes the contact between predicted regions. The red line denotes the conserved PdCR.

Botulinum neurotoxin serotype A (3BTA): The F-value profile and conserved hydrophobic packing of 3BTA present in Figure 10C and 11C, respectively. The F-value peaks are determined at residue number Y57, L62, I70, Y89, A105, Q107, I158 and C188. The highest peak was detected at residue Y89, which is located on β strand 5 (PdCR region-3) and close to conserved hydrophobic residues β5N and β5C as also presented in 1A8D and 1EPW. Furthermore, these two residues formed packing between PdCR region-2 and region-4 in the native state of folding as shown in Figure 11C. Therefore, these residues are considered to be significant for packing with the predicted regions in the C-terminal part.

CSL bound to DNA (1TTU): The F-value peaks of 1TTU are detected at residues

T40, W42, A44, S73, V75, G83, Y127 and C129 as presented in Figure 10D. The highest peaks of the F-value plot are observed at residues S73, V75 and G83, which are located near conserved hydrophobic residues β5N and β5C within ±5 residues. It is interesting that even though this protein contains a deletion site, β strands 6 and 7, but the residue L86 was detected as a conserved hydrophobic residue at the conserved position β6. Due to this result, a conserved position β6 is also considered as a conserved hydrophobic residue that close to the highest peak of F-value plot. The hydrophobic packing of conserved hydrophobic residues is presented in Figure 11D. According to this figure, conserved residue β5C interacts with β1, β4N and β4C in predicted region-1, and with β9 in predicted region-2. That is, residue β5C plays a role in connecting the N-terminal and C-terminal regions and is considered to be significant in the formation of the whole 3D-structure. On the other hand, residue β6 forms packing with β2 in predicted region-1 and with β10 in predicted region-2. Conserved residue β6 can be also a significant residue in the 3D-structure formation.

To compare the hydrophobic packing in the symmetric beta-trefoil proteins with that in the proteins with irregular structures (Figure 12). The commonly appearing contacts in both protein groups are highlighted by red. As a result, 17 common contacts are confirmed (contacts appearing more than 90% were counted). Five of 17 contacts are formed by residue β5C which form packing with different trefoil unit, that is, β4N and β4C in trefoil unit-1, and β9 in trefoil unit-3. On the other hand, residue β9 also detected as a residue which forms packing with other two trefoil units, that is, trefoil unit-1 and trefoil unit-2. Thus, conserved residues β5C and β9 are considered to play a role as the core of the beta-trefoil protein fold.

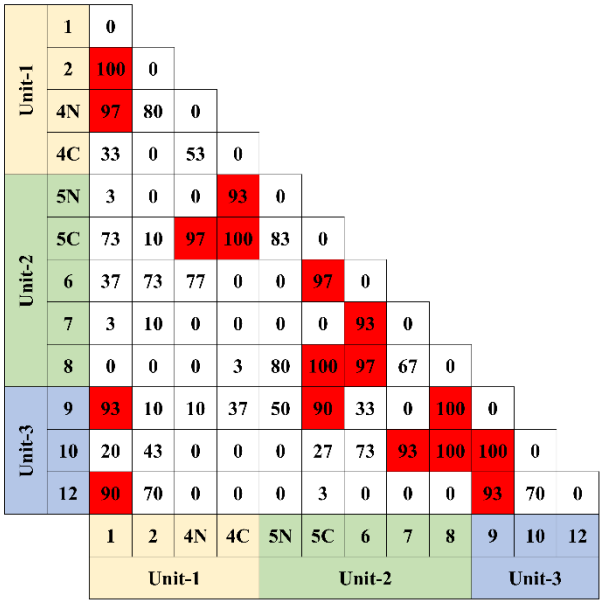


Figure 12 Occurrence ratio of contacts formed by conserved hydrophobic residues derived from multiple sequence alignment of 30 beta-trefoil proteins. A number shown on the x-axes and y-axes denotes a conserved hydrophobic residue. The occurrence ratio is indicated by a percentage in each cell. Red color indicates a contact which greater than 90% conservation.

It is interesting to see the interaction between β6 and β7, β6 and β8, and β7 and β10 conserved in 26 beta-trefoil proteins with the high symmetry structures, but no longer conserved in four irregular beta-trefoil proteins due to the deletion of β-strand 6 and β-strand 7 which affect to the missing of conserved hydrophobic residue β7 (Figure 1E in appendix section). However, instead of this packing, an interaction between β6 and β10 is observed. It is thought that this interaction between β6 and β10 in 1TTU compensates for the missing interaction of conserved residue β7 to stabilize the common structure of beta-trefoil protein.

According to the results of beta-trefoil proteins derived from amino acid sequence-based techniques, the characteristic tendency of the beta-trefoil fold in which an F-value plot shows high values at residues in the central trefoil unit was also observed in both structural types of beta-trefoil proteins. Suggesting that irregular site does not affect to the folding mechanism of beta-trefoil protein. These results emphasize the significance of a conserved hydrophobic residue near a peak in an F-value plot to make packing to connect predicted compact regions to form a whole protein 3D-structure.

4.3.4 Go model simulations

According to an unavailable of experimental result of irregular beta-trefoil protein, a 3D-structure-based Gō-like model simulation can be used to simulate the folding pathway which necessary to emphasize the correctness of present sequence-based techniques, ADM and F-value analysis. In this study, a representative of each irregular beta-trefoil protein was selected, that is, 1TTU as a representative of a deletion type and 1A8D as a representative of an insertion type. The results of Gō-like model simulation are present by the mean of a free-energy profile and a contact frequency map. A free-energy profile presents a free-energy value derived from the simulation against relative Q value, the ratio of the number of the native contacts detected in each state to all native contacts. The first and last valleys of the free-energy profile are considered as the denatured state and the native state, respectively, and the peak between these valleys is considered as the transition state.

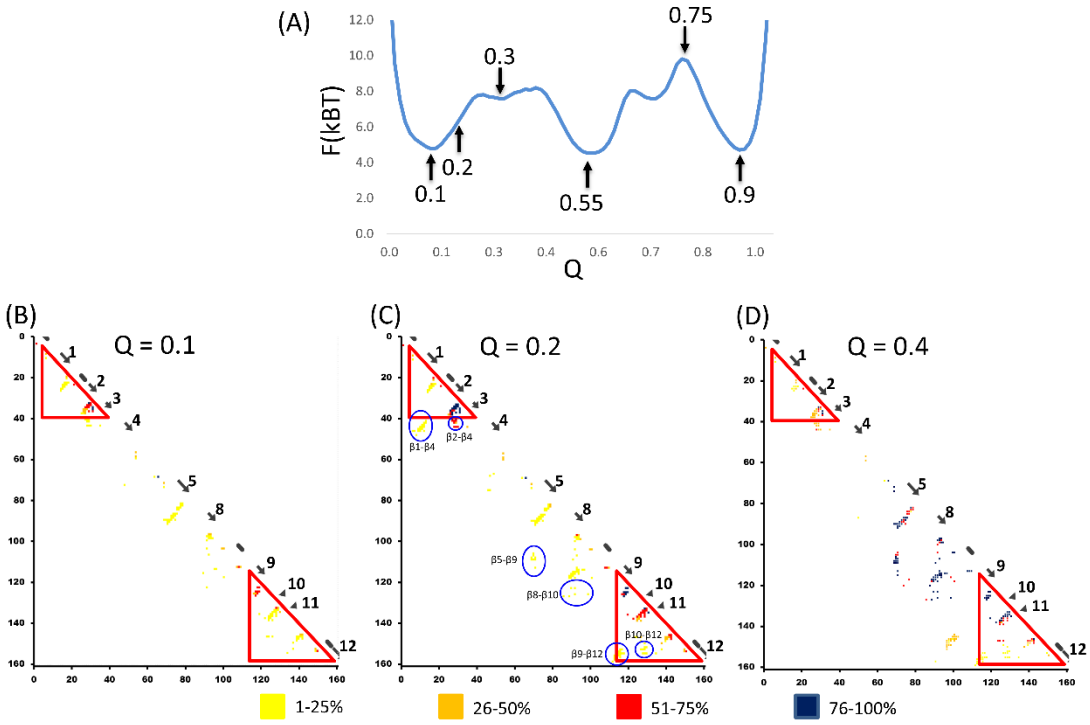


Figure 13 Free-energy profile of 1TTU at the respective Q values (A). The contact frequency maps of Q = 0.1, 0.2, and 0.4 represent the contacts detected by a Gō-model simulation of the denatured state (B), late denatured state (C), and first transition state (D), with the red triangle of PdCR predicted by ADM analysis.

The Gō-like model results of 1TTU are present in Figure 12. The free-energy profile of 1TTU (Figure 13A) indicates the location of denatured, intermediate and native states as Q = 0.1, 0.55 and 0.9. Two major transition states are identified at Q = 0.4 and 0.75. In the denatured state (Q = 0.1) presents only local contact and some interaction with adjacent β strand. In Figure 13B, the cluster of contacts around β2 and β3, and β9, β10 and β11 are frequently observed at Q = 0.2. These highly frequent contacts correspond to the ADM predicted regions C5-A37 and C116-I159, which corresponds to the state just before the first transition state. Moreover, the long-range contacts formed by β3, β4, β5, β8, β9 and β10 are also observed in this state as presented in Figure 13B. It is interesting that these strands are also detected near the peak of in F-value plot (Figure 10D). The contacts in state Q = 0.4 clearly illustrated the formation of the C-terminal PdCR and the robust of contacts formed by this region and the residues from β5 and β8 in the central beta-trefoil unit. According to these results, ADM and F-value analysis can be used to detect the compact region and the interactions in the early stage of folding.

After an initial state of folding of 1TTU, the interactions detected at Q = 0.4 are more frequent observed in an intermediate state, Q =0.55, as shown in Figure F1 in appendix section. The interactions formed by trefoil unit-1 to another unit were detected in the state of Q = 0.75, second transition state, then the whole domain tend to form a native structure at Q = 0.9.

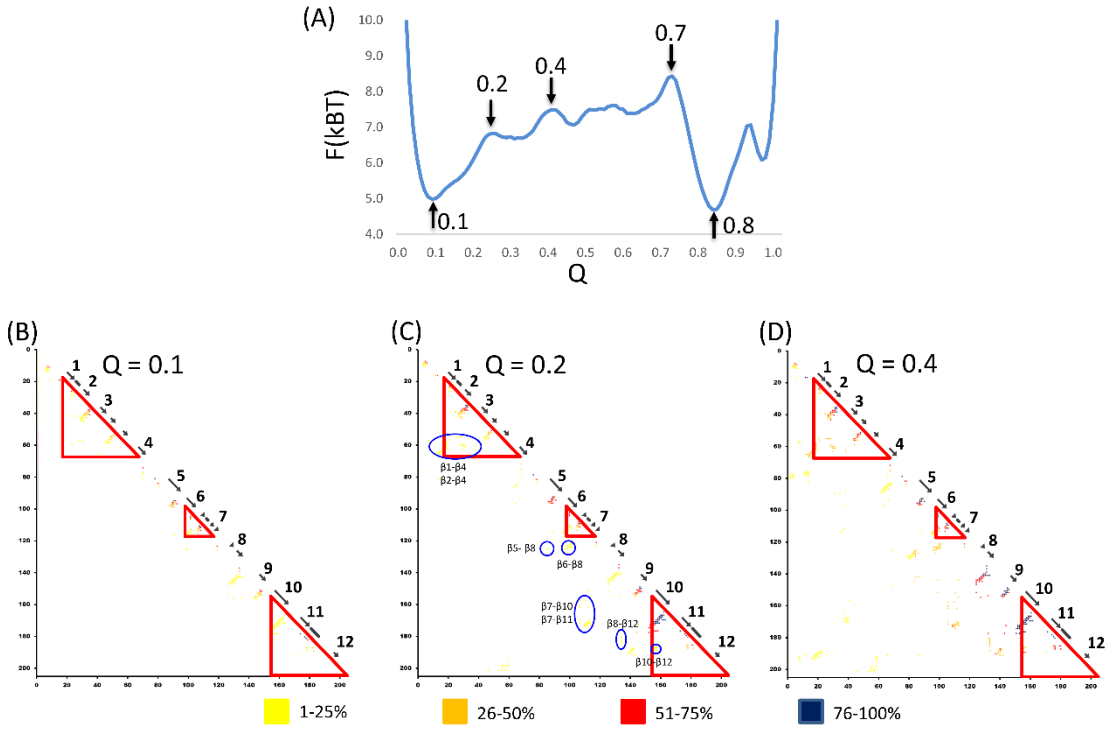


Figure 14 Free-energy profile of 1A8D at the respective Q values (A). The contact frequency maps of Q = 0.1, 0.2, and 0.4 represent the contacts detected by a Gō-model simulation of the denatured state (B), first transition state (C), and second transition state (D), with the red triangle of PdCR predicted by ADM analysis.

The results of Gō-model simulation for 1A8D are present in Figure 14. The free energy profile is presented in Figure 14A. The location of denatured state and native state are indicated at Q = 0.1 and 0.8, respectively. While three peaks at Q = 0.2, 0.4 and 0.7 are indicated as first, second and third transition state. As in 1TTU, the contact map at Q = 0.15 was also investigated to find the interaction in an initial state of folding, but no such long-range interaction was found in this state (Figure F2 in appendix section). Due to this result, the contact map at Q = 0.2 was investigated to detect the contact formed in the early state of folding. The N-terminal and C-terminal regions containing highly frequent contacts correspond to the ADM-predicted regions Y18-I67 and A153-N204. The long-range contacts formed by β4, β5, β6, β7, β8, β10 and β12 are detected which near the position of F-value peaks (Figure 9A).

The second transition state presents at Q = 0.4. The contacts formed by trefoil unit-1 with another unit are observed in this state. Then all contacts were detected at Q = 0.7 as corresponded to the third transition state. (The contact frequency map of state Q = 0.15, 0.7 and 0.8 are present in Figure F2 in appendix section.)

4.4 Discussions

The structure-based sequence alignment for beta-trefoil proteins was analyzed and conserved hydrophobic residues were identified in this study. Interestingly, almost every β-strand contains one or two conserved hydrophobic residues in spite of low sequence identity among sequences under the criteria of 90% conservation. This fact may indicate that these equally distributed hydrophobic residues need to form the symmetrical beta-trefoil fold. Similar results were already obtained by Murzin et al. and Feng et al.,[69,86] and they revealed the relationships between conserved residues and 3D-structures or interaction energies. Feng et al.[86] also found "symmetric key structural residues" specific for the beta-trefoil structure based on structure-based multiple sequence alignment of domains in five two-domain proteins with two beta-trefoil structures. Furthermore, they elucidated that these symmetric key structural residues are well conserved in the majority of beta-trefoil proteins. Conserved hydrophobic residues identified in this study correspond well to the symmetric key structural residues obtained by Feng et al.[86] indicating conserved hydrophobic residues derived from the present alignment are also key residues to make packing within the beta-trefoil fold.

Furthermore, the results of ADM and F-value analyses reflect the results of folding experiments for proteins 2K8R, 6I1B and 1HCD. Longo et al.[75,76] succeeded in designing Phifoil, beta-trefoil designed by folding nucleus symmetric expansion, based on a folding nucleus of 2K8R deduced from the results of the η-value analyses, and this folding nucleus comprises β-strand 2 to β-strand 6. This nucleus corresponds to the ADM predicted compact regions L6-I49 and Y57-Y67 which includes some of highest peaks in the F-value plot. That is, our predicted compact regions correspond well to the folding nucleus defined from ϕ-value analyses. For 6I1B, the present study predicts the C-terminal part is significant for the folding reflecting the experimental data,[78,81] although the highest peak in the F-value plot is included in the N-terminal predicted compact region by ADM. The region around this peak is also included in the folding region identified by Liu et al. and Capraro et al.[78,81]

The present study performed the analyses by the ADMs in the combination with the structure-based sequence alignment, and the results predict the compact regions stable in the early stage of folding for beta-trefoil proteins. Although compact regions of beta-trefoil proteins look to show a variety of locations of the predicted regions, the modest conservation of the compact regions is also observed as shown in Figure 6. The fact that the predicted stable compact regions with the highest η value for 2K8R, 6I1B and 1HCD correspond well to the experimentally obtained folding units denotes a predicted compact region by ADM can be regarded as a kind of unit of folding in a beta-trefoil protein. Based on this, it is considered that the results in Figure 6 indicate the variety of the folding mechanisms of the beta-trefoil proteins. On the other hand, from the F-value analyses β-strands 5 and 6 are always the center of folding for 2K8R, 6I1B and 1HCD consistent with the experimental results. This property is always observed in the results of the F-value plots for other beta-trefoil proteins for which folding mechanisms have not yet been examined experimentally. Further folding mechanism, that is, with which part, N-

terminal or C-terminal part, β-strands 5 and 6 interact, depends on each protein. Thus, the present analyses can be applied to predict the folding properties for beta-trefoil proteins in general. Furthermore, the information of the location of conserved hydrophobic residues in combination with the results of ADM and an F-value plot provides the significant residues for hydrophobic packing during folding. The conserved hydrophobic residues tend to be included in the predicted compact regions by ADMs. This indicates that the interactions between them tend to occur within a compact region or between compact regions to stabilize the native structure of beta-trefoil protein.

The irregular beta-trefoil proteins, three proteins from STI-like superfamily and one protein from DNA-binding protein LAG-1 (CSL) superfamily, are also investigated in this study. In the three STI-like superfamily proteins, the ADM-predicted regions always contain the insertion parts, suggesting that this insertion does not disturb the N-terminal trefoil-unit formation. On the other hand, the native contacts in 1TTU from DNA-binding protein LAG-1 (CSL) superfamily show that the conserved hydrophobic residue at position β6 interacts with β10 in the C-terminal side. This contact is not frequently observed in the beta-trefoil proteins with high symmetry. That is, conserved residue β6 assumes the role of the deficient β7.

A Gō-model simulation in the present study demonstrates that the conserved hydrophobic residues that form contacts are mainly within an ADM-predicted region in the early stage of folding. These conserved residues are also located around F-value peaks. Contacts by the conserved residues between ADM-predicted regions start to form with increasing Q value. This suggests that the conserved hydrophobic residues play a significant role in protein folding. Moreover, the Gō-model simulation results of 1TTU and 1A8D confirmed the conservation of beta-trefoil protein folding mechanism, that is, the contacts between beta-trefoilunit-2 and unit-3 are detected in an early event of folding.

4.5 Conclusion

Information on the location of conserved hydrophobic residues in combination with the results of ADM and an F-value plot reveal the significant residues for hydrophobic packing during folding. In other words, an initial folding site in a protein can be defined as a site with conserved hydrophobic residues near a high F-value peak in a predicted region by ADM with the highest η value. Several such sites form contacts within a predicted region by ADM and form a larger structure. Although almost every β-strand contains one or two conserved hydrophobic residues and these equally distributed hydrophobic residues seem to be significant to form the symmetrical beta-trefoil fold, the conserved hydrophobic residues in trefoil unit-2 may be more significant. Conserved residues position β5N, β5C and β6 contain conserved hydrophobic residues near the highest or second highest peak of an F-value plot in a protein from almost every superfamily. The results of folding experiments for several beta-trefoil proteins indicates the significance of β-strands 5 and 6 for folding which corresponded well to these sequence-based prediction methods. The results from coarse-grained Gō-model simulations in an initial state of folding coincide well with the predictions made by the

ADMs and F-value analyses based on the amino acid sequence information. Furthermore, these results also correlated to the experimental results of symmetric beta-trefoil proteins. Again, ADMs and F-value analyses can decode the folding information from the amino acid sequences of not only the beta-trefoil proteins with high symmetry but also of those with irregular structures.

Chapter 5

Thesis Conclusion

The relationship between the amino acid sequences and structures of proteins is one of the bioinformatics goals and have been intensively studied so far. It is interesting to investigate how proteins with low sequence similarity can fold into a similar native structure. Protein folding mechanisms have been intensively investigated with experimental as well as simulation techniques. Even though, protein folding process have attracted extensive studies on mechanical properties, but the complete folding process is still uncertain due to the limitation of experimental method and computational power. According to this fact, to extract the folding property of a target protein especially in an initial state of folding is one of the most important factors in discovering the whole story of protein folding processes.

The initial folding processes of different fold types in all beta protein class, Ig-like β-sandwich fold and beta-trefoil fold, have been investigated in this study by means of multiple sequence alignment and inter-residue average distance statistic methods, ADM, and F-value analysis as well as by 3D-based Gō-model simulation. However, the sequence identities of these proteins are rather low, and making accurate sequence-based multiple alignments is difficult. Therefore, the 3D-structures were used to make multiple alignments. The information of conserved hydrophobic residues in combination with the results based on the average distance statistics is used to detect the significant residues to form the compact region in an early event of folding.

The Ig-like β-sandwich protein used in this study is titin protein. Titin is an important protein that is responsible for striated-muscle elasticity. It is a highly modular protein composed of ~300 domains, mostly Ig and FN3 domains, linked in tandem. 1TIT and 1TEN were selected as the representative of each domain type according to an available of experimental results. The central region of all domains from Ig domains and FN3 domains was predicted to be an initial folding unit that forms the compact structure in the early event, due to the present results of sequence-based and 3D-structure-based techniques. This common feature is in line with the available experimental results of 1TIT and 1TEN, which also detected the stability of the central unit and the fluctuation of both terminal ends. Interestingly, the results underscore the importance of the common structure of these proteins, in particular, the key strands for folding and the Greek-key motif. Moreover, the difference in the folding pathways and the whole story of the protein folding processes can be described by the present Gō-model simulations.

The beta-trefoil fold protein is found in Kunitz inhibitors and fibroblast growth factors. The overall fold has an approximate three-fold symmetry and consists of a six-stranded-barrel capped by a triangular hairpin triplet. To investigate the protein folding mechanisms, 26 high symmetric beta-trefoil proteins and four irregular structure beta-trefoil proteins were selected from six superfamilies. However, the folding experiments of only some proteins were performed extensively. It is confirmed that a conserved

hydrophobic residue is always located in a β-strand. In particular, β-strands 5 and 6 are significant for the initial folding from the analyses based on the inter-residue average distance statistics. These results coincide well with the experimental data obtained so far for folding of some of the beta-trefoil proteins. N-terminus and C-terminus tend to be the compact region in an initial state of folding according to the high conservation of PdCR at both terminal sites. Due to the lacking experimental data of irregular beta-trefoil protein, a Gō-model simulation was used to investigate the folding processes. The long-range contacts formed in an initial state of folding are corresponded well to the residues close to the peaks of the F-value profile. It is also confirmed that the conserved hydrophobic residues defined in this study contribute to form hydrophobic packing in beta-trefoil proteins in general.

The present sequence-based techniques, ADM and F-value analyses, results show well related to the experimental data so far. I believe that these techniques can be applied to predict such kind of folding processes from other proteins with various topologies, such as intrinsic disorder protein or knot protein, which much more complex structure and the folding mechanisms have not been clearly discovered. Furthermore, according to the increasing experimental data of protein folding mechanisms (ϕ-value and H/D exchange experiments). Then, it is possible to improve the accuracy of the present inter-residue average distance statistics methods by using available experimental results as a restrain value to generate new statistical methods.

References

1. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, and Walter P 2002 *Molecular biology of the cell.* Garland, Newyork.
2. Rizzuti B, Daggett V (2013) Using simulations to provide the framework for experimental protein folding studies. *Archives of Biochemistry and Biophysics.* 531(1):128-135.
3. Dobson CM (2003) Protein folding and misfolding. *Nature.* 426:884-890.
4. Khoury GA, Baliban RC, Floudas CA (2011) Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database. *Scientific reports.* 1(90):1-5.
5. Boys BL, Konermann L (2007) Folding and Assembly of Hemoglobin Monitored by Electrospray Mass Spectrometry Using an On-line Dialysis System. *Journal of the American Society for Mass Spectrometry.* 18(1):8-16.
6. Huang JT, Cheng JP (2008) Differentiation between two‐state and multi‐state folding proteins based on sequence. *Proteins: Structure, Function, and Bioinformatics.* 72(1):44-49.
7. Russell RB (2002) Classification of protein folds. *Molecular biotechnology.* 20(1):17-28.
8. Fox NK, Brenner SE, Chandonia J (2013) SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic acids research.* 42(D1):D304-D309.
9. Ichimaru T, Kikuchi T (2003) Analysis of the differences in the folding kinetics of structurally homologous proteins based on predictions of the gross features of residue contacts. *Proteins: Structure, Function, and Bioinformatics.* 51(4):515-530.
10. Matsuoka M, Fujita A, Kawai Y, Kikuchi T (2014) Similar structures to the E-to-H helix unit in the globin-like fold are found in other helical folds. *Biomolecules.* 4(1):268-288.
11. Matsuoka M, Sugita M, Kikuchi T (2014) Implication of the cause of differences in 3D-structures of proteins with high sequence identity based on analyses of amino acid sequences and 3D-structures. *BMC Research Notes.* 7(1):654-666.
12. Ishizuka Y, Kikuchi T (2011) Analysis of the local sequences of folding sites in β sandwich proteins with inter-residue average distance statistics. *The Open Bioinformatics Journal.* 5(1):59-68.
13. Aumpuchin P, Kikuchi T (2019) Prediction of folding mechanisms for Ig-like beta sandwich proteins based on inter-residue average distance statistics methods. *Proteins: Structure, Function, and Bioinformatics.* 87(2):120-135.
14. Matsuoka M, Kikuchi T (2014) Sequence analysis on the information of folding initiation segments in ferredoxin-like fold proteins. *BMC Struct Biol.* 14(15):1-15.
15. Kirioka T, Aumpuchin P, Kikuchi T (2017) Detection of folding sites of beta-trefoilfold proteins based on amino acid sequence analyses and structure-based sequence alignment. *J Proteomics Bioinform.* 10(9):222-235.
16. Nakashima T, Kabata M, Kikuchi T (2017) Properties of amino acid sequences of lysozyme-like superfamily proteins relating to their folding mechanisms. *J Proteomics Bioinform.* 10(4):94-107.
17. Kikuchi T, Némethy G, Scheraga HA (1988) Prediction of the location of

structural domains in globular proteins. *Journal of protein chemistry.* 7(4):427-471.
18. Guo Z, Thirumalai D (1997) The nucleation-collapse mechanism in protein folding: evidence for the non-uniqueness of the folding nucleus. *Folding and Design.* 2(6):377-391.
19. Kikuchi T (2008) Analysis of 3D structural differences in the IgG-binding domains based on the inter-residue average-distance statistics. *Amino acids.* 35(3):541-549.
20. Tsigelny IF 2002 *Protein structure prediction: Bioinformatic approach.* Vol 3. La jolla, CA: Internat'l University Line.
21. Berman HM, Westbrook J, Feng Z, et al. (2000) The Protein Data Bank. *Nucleic Acids Research.* 28(1):235-242.
22. Shindyalov IN, Bourne PE (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein engineering.* 11(9):739-747.
23. Gille C, Frömmel C (2001) STRAP: Editor for STRuctural Alignments of proteins. *Bioinformatics.* 17(4):377-378.
24. Pundir S, Martin MJ, O'Donovan C (2017) UniProt protein knowledgebase. In: *Protein Bioinformatics.* Vol 1558. Springer, 41-55.
25. Kumar S, Stecher G, Tamura K (2016) MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Molecular biology and evolution.* 33(7):1870-1874.
26. Saitou N, Nei M (1987) The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular biology and evolution.* 4(4):406-425.
27. Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. *Bioinformatics.* 8(3):275-282.
28. Sugita M, Kikuchi T (2013) Incorporating into a Cα Gō model the effects of geometrical restriction on Cα atoms caused by side chain orientations. *Proteins Struct Funct Bioinf.* 81.
29. Sugita M, Kikuchi T (2014) Analyses of the folding properties of ferredoxin-like fold proteins by means of a coarse-grained Go model: relationship between the free energy profiles and folding cores. *Proteins.* 82(6):954-965.
30. Sugita M, Kikuchi T (2013) Incorporating into a Cα Go model the effects of geometrical restriction on Cα atoms caused by side chain orientations. *Proteins: Structure, Function, and Bioinformatics.* 81(8):1434-1445.
31. Sugita M, Matsuoka M, Kikuchi T (2015) Topological and sequence information predict that foldons organize a partially overlapped and hierarchical structure. *Proteins: Structure, Function, and Bioinformatics.* 83(10):1900-1913.
32. Sulkowska JI, Cieplak M (2008) Selection of optimal variants of Go-like models of proteins through studies of stretching. *Biophysical journal.* 95(7):3174-3191.
33. Sugita M, Matsuoka M, Kikuchi T (2015) Topological and sequence information predict that foldons organize a partially overlapped and hierarchical structure. *Proteins.* 83(10):1900-1913.
34. Mitsutake A, Sugita Y, Okamoto Y (2001) Generalized-ensemble algorithms for molecular simulations of biopolymers. *Biopolymers.* 60(2):96-123.
35. Ferrenberg AM, Swendsen RH (1988) New Monte Carlo technique for studying phase transitions. *Phys Rev Lett.* 61(23):2635-2638.

36. Ferrenberg AM, Swendsen RH (1989) Optimized Monte Carlo data analysis. *Phys Rev Lett.* 63(12):1195-1198.
37. Somkuti J, Mártonfalvi Z, Kellermayer MSZ, Smeller L (2013) Different pressure–temperature behavior of the structured and unstructured regions of titin. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics.* 1834(1):112-118.
38. Tonino P, Kiss B, Strom J, et al. (2017) The giant protein titin regulates the length of the striated muscle thick filament. *Nature Communications.* 8(1):1041.
39. Labeit S, Kolmerer B (1995) Titins: Giant proteins in charge of muscle ultrastructure and elasticity. *Science.* 270(5234):293-296.
40. Clarke J, Cota E, Fowler SB, Hamill SJ (1999) Folding studies of immunoglobulin-like β-sandwich proteins suggest that they share a common folding pathway. *Structure.* 7(9):1145-1153.
41. Geierhaas CD, Paci E, Vendruscolo M, Clarke J (2004) Comparison of the transition states for folding of two Ig-like proteins from different superfamilies. *Journal of Molecular Biology.* 343(4):1111-1123.
42. Fowler SB, Clarke J (2001) Mapping the folding pathway of an immunoglobulin domain: structural detail from phi value analysis and movement of the transition state. *Structure.* 9(5):355-366.
43. Alzari PM (1998) Domains, immunoglobulin-type In: Roitt I, ed. *Encyclopedia of Immunology (Second Edition).* Oxford: Elsevier, 775-778.
44. Richardson JS (1981) The anatomy and taxonomy of protein structure. In: Anfinsen CB, Edsall JT, Richards FM, eds. *Advances in Protein Chemistry*. Vol 34. USA: Academic Press, 167-339.
45. Bork P, Holm L, Sander C (1994) The immunoglobulin fold: Structural classification, sequence patterns and common core. *Journal of Molecular Biology.* 242(4):309-320.
46. Gao M, Lu H, Schulten K (2001) Simulated refolding of stretched titin immunoglobulin domains. *Biophysical Journal.* 81(4):2268-2277.
47. Rivas-Pardo JA, Eckels EC, Popa I, Kosuri P, Linke WA, Fernández JM (2016) Work done by titin protein folding assists muscle contraction. *Cell Reports.* 14(6):1339-1347.
48. Hamill SJ, Steward A, Clarke J (2000) The folding of an immunoglobulin-like Greek key protein is defined by a common-core nucleus and regions constrained by topology *Journal of Molecular Biology.* 297(1):165-178.
49. Improta S, Politou AS, Pastore A (1996) Immunoglobulin-like modules from titin I-band: Extensible components of muscle elasticity. *Structure.* 4(3):323-337.
50. Leahy DJ, Hendrickson WA, Aukhil I, Erickson HP (1992) Structure of a fibronectin type III domain from tenascin phased by MAD analysis of the selenomethionyl protein. *Science.* 258(5084):987-991.
51. Guo Z, Thirumalai D (1995) Kinetics of protein folding: Nucleation mechanism, time scales, and pathways. *Biopolymers.* 36(1):83-102.
52. Kukic P, Pustovalova Y, Camilloni C, Gianni S, Korzhnev DM, Vendruscolo M (2017) Structural characterization of the early events in the nucleation–condensation mechanism in a protein folding process. *Journal of the American Chemical Society.* 139(20):6899-6910.
53. Yagawa K, Yamano K, Oguro T, et al. (2010) Structural basis for unfolding

pathway‐dependent stability of proteins: Vectorial unfolding versus global unfolding. *Protein Science.* 19(4):693-702.
54. Best RB, Vendruscolo M (2006) Structural interpretation of hydrogen exchange protection factors in proteins: Characterization of the native state fluctuations of CI2. *Structure.* 14(1):97-106.
55. Kister AE, Finkelstein AV, Gelfand IM (2002) Common features in structures and sequences of sandwich-like proteins. *Proceedings of the National Academy of Sciences USA.* 99(22):14137-14141.
56. Steward A, McDowell GS, Clarke J (2009) Topology is the principal determinant in the folding of a complex all-alpha Greek key death domain from human FADD. *Journal of Molecular Biology.* 389(2):425-437.
57. Lindorff-Larsen K, Røgen P, Paci E, Vendruscolo M, Dobson CM (2005) Protein folding and the organization of the protein topology universe. *Trends in Biochemical Sciences.* 30(1):13-19.
58. Paci E, Clarke J, Steward A, Vendruscolo M, Karplus M (2003) Self-consistent determination of the transition state for protein folding: application to a fibronectin type III domain. *Proceedings of the National Academy of Sciences USA.* 100(2):394-399.
59. Sugita M, Kikuchi T (2014) Analyses of the folding properties of ferredoxin-like fold proteins by means of a coarse-grained Gō model: Relationship between the free energy profiles and folding cores. *Proteins: Structure, Function, and Bioinformatics.* 82(6):954-965.
60. Jackson SE (1998) How do small single-domain proteins fold? *Folding and Design.* 3(4):81-91.
61. Hu J, Chen T, Wang M, Chan HS, Zhang Z (2017) A critical comparison of coarse-grained structure-based approaches and atomic models of protein folding. *Physical Chemistry Chemical Physics.* 19(21):13629-13639.
62. Zhang Z, Chan HS (2012) Transition paths, diffusive processes, and preequilibria of protein folding. *Proceedings of the National Academy of Sciences.* 109(51):20919-20924.
63. Chen T, Song J, Chan HS (2015) Theoretical perspectives on nonnative interactions and intrinsic disorder in protein folding and binding. *Curr Opin Struct Biol.* 30:32-42.
64. Tanford C (1968) Protein denaturation. In: Anfinsen CB, Anson ML, Edsall JT, Richards FM, eds. *Advances in Protein Chemistry.* Vol 23. USA: Academic Press, 121-282.
65. Dyson HJ, Wright PE (2005) Intrinsically unstructured proteins and their functions. *Nature Reviews Molecular Cell Biology.* 6:197-208.
66. Orengo CA, Jones DT, Thornton JM (1994) Protein superfamilles and domain superfolds. *Nature.* 372(6507):631-634.
67. Sweet RM, Wright HT, Janin J, Chothia CH, Blow DM (1974) Crystal structure of the complex of porcine trypsin with soybean trypsin inhibitor (Kunitz) at 2.6 Å resolution. *Biochemistry.* 13(20):4212-4228.
68. McLachlan AD (1979) Three-fold structural pattern in the soybean trypsin inhibitor (Kunitz). *Journal of molecular biology.* 133(4):557-563.
69. Murzin AG, Lesk AM, Chothia C (1992) beta-trefoilfold: patterns of structure and sequence in the kunitz inhibitors interleukins-1β and 1α and fibroblast growth

factors. *Journal of molecular biology.* 223(2):531-543.
70. Ponting CP, Russell RB (2000) Identification of distant homologues of fibroblast growth factors suggests a common ancestor for all beta-trefoilproteins. *Journal of molecular biology.* 302(5):1041-1047.
71. Lee J, Blaber M (2011) Experimental support for the evolution of symmetric protein architecture from a simple peptide motif. *Proceedings of the National Academy of Sciences.* 108(1):126-130.
72. Lee J, Blaber SI, Dubey VK, Blaber M (2011) A polypeptide "building block" for the beta-trefoilfold identified by "top-down symmetric deconstruction". *Journal of molecular biology.* 407(5):744-763.
73. Broom A, Ma SM, Xia K, et al. (2015) Designed protein reveals structural determinants of extreme kinetic stability. *Proceedings of the National Academy of Sciences.* 112(47):14605-14610.
74. Broom A, Doxey AC, Lobsanov YD, et al. (2012) Modular evolution and the origins of symmetry: reconstruction of a three-fold symmetric globular protein. *Structure.* 20(1):161-171.
75. Longo L, Lee J, Blaber M (2012) Experimental support for the foldability–function tradeoff hypothesis: Segregation of the folding nucleus and functional regions in fibroblast growth factor‐1. *Protein Science.* 21(12):1911-1920.
76. Longo LM, Kumru OS, Middaugh CR, Blaber M (2014) Evolution and design of protein structure by folding nucleus symmetric expansion. *Structure.* 22(10):1377-1384.
77. Xia X, Longo LM, Sutherland MA, Blaber M (2016) Evolution of a protein folding nucleus. *Protein Science.* 25(7):1227-1240.
78. Liu C, Gaspar JA, Wong HJ, Meiering EM (2002) Conserved and nonconserved features of the folding pathway of hisactophilin, a β‐trefoil protein. *Protein Science.* 11(3):669-679.
79. Chi Y-H, Kumar TKS, Chiu M, Yu C (2002) Identification of rare partially unfolded states in equilibrium with the native conformation in an all β-barrel protein. *Journal of Biological Chemistry.* 277(38):34941-34948.
80. Wang H-M, Yu C (2011) Investigating the refolding pathway of human acidic fibroblast growth factor (hFGF-1) from the residual structure (s) obtained by denatured-state hydrogen/deuterium exchange. *Biophysical journal.* 100(1):154-164.
81. Capraro DT, Roy M, Onuchic JN, Gosavi S, Jennings PA (2012) β-Bulge triggers route-switching on the functional landscape of interleukin-1β. *Proceedings of the National Academy of Sciences.* 109(5):1490-1493.
82. Chavez LL, Gosavi S, Jennings PA, Onuchic JN (2006) Multiple routes lead to the native state in the energy landscape of the beta-trefoilfamily. *Proceedings of the National Academy of Sciences.* 103(27):10254-10258.
83. Gosavi S (2013) Understanding the folding-function tradeoff in proteins. *PLoS One.* 8(4):e61222.
84. Li M, Huang Y, Xiao Y (2008) Effects of external interactions on protein sequence‐structure relations of beta‐trefoil fold. *Proteins: Structure, Function, and Bioinformatics.* 72(4):1161-1170.
85. Li M, Huang Y, Xu R, Xiao Y (2005) Nonlinear analysis of sequence symmetry of beta-trefoil family proteins. *Chaos, Solitons & Fractals.* 25(2):491-497.

86. Feng J, Li M, Huang Y, Xiao Y (2010) Symmetric key structural residues in symmetric proteins with beta-trefoil fold. *PloS one.* 5(11):e14138.
87. Katoh K, Misawa K, Kuma Ki, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic acids research.* 30(14):3059-3066.

Acknowledgements

I would like to express my infinite gratitude to my advisor Professor Takeshi Kikuchi for his patience, motivation, and supporting of my Ph.D. study. His guidance helped me in all the time of research and writing of this thesis. However, the completion of this thesis was impossible without the support of many people. Beside my advisor, I would like to thank Professor Fumio Hirata and Assistant Professor Masatake Sugita, for their insightful comments and suggestions which incented me widen my research from various perspectives. I thank my fellow labmates, especially Ms. Kimura Risako, for the stimulating discussions, selfless help and for all the fun we have had in the last three years. Finally, I would like to give a big thank to my family: my parents and to my brothers for supporting me and made my studies possible.

Publications

1. Aumpuchin P, Kikuchi T (2019) Prediction of folding mechanisms for Ig-like beta-sandwich proteins based on inter-residue average distance statistics methods. Proteins: Structure, Function, and Bioinformatics. 87(2):120-135.
2. Kirioka T, Aumpuchin P, Kikuchi T (2017) Detection of folding sites of β-trefoil fold proteins based on amino acid sequence analyses and structure-based sequence alignment. J Proteomics Bioinform. 10(9):222-235.

Appendix A

Multiple sequence alignment of Ig domains and FN3 domains

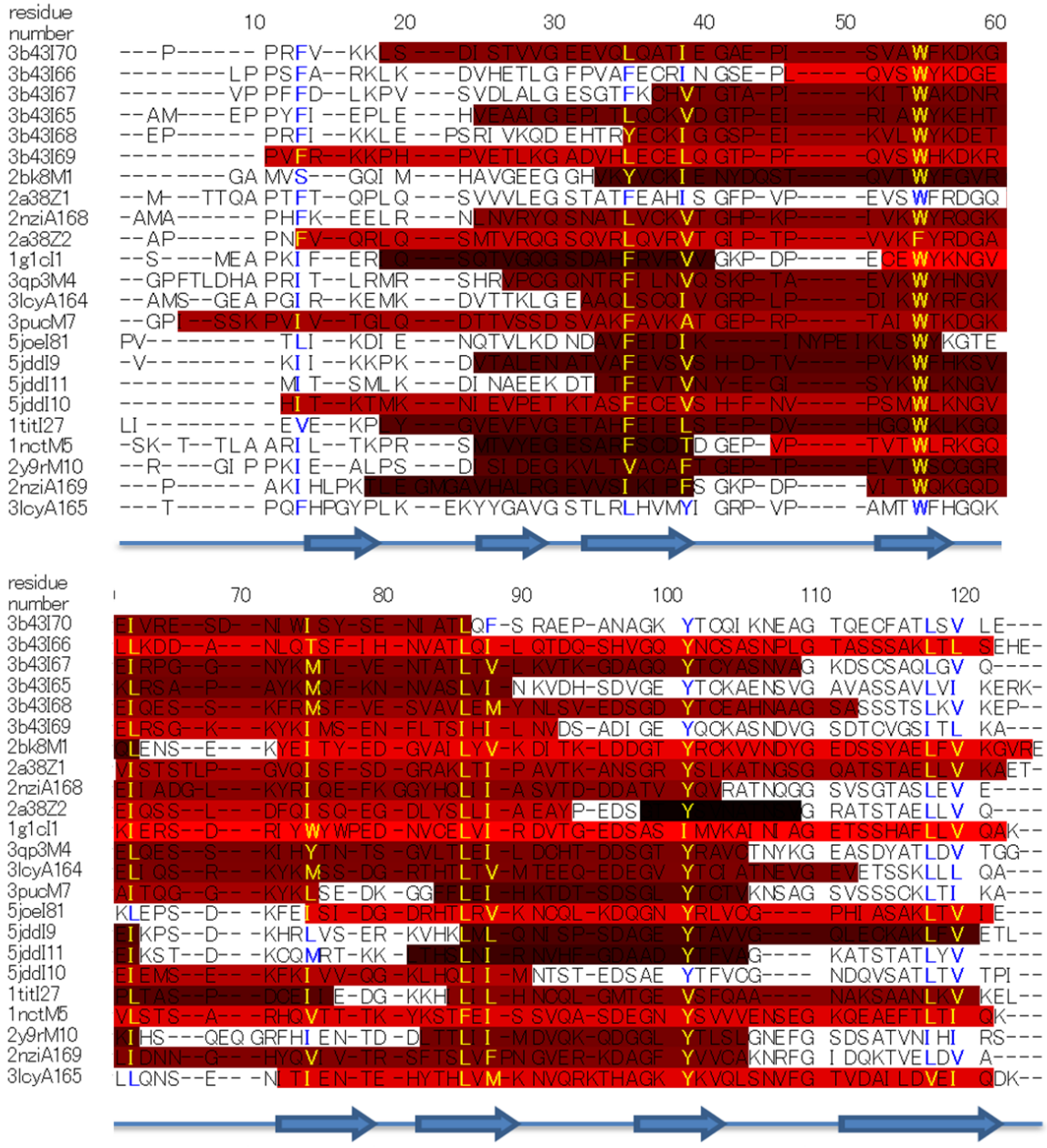


Figure A1: Structure-based multiple sequence alignment for 23 Ig-domains yielding the ADM prediction. The predicted compact region is indicated by a red bar. The brighter red denotes a higher compact density. The conserved hydrophobic residues in the predicted compact region are yellow, and a blue letter is a residue out of the predicted compact region.

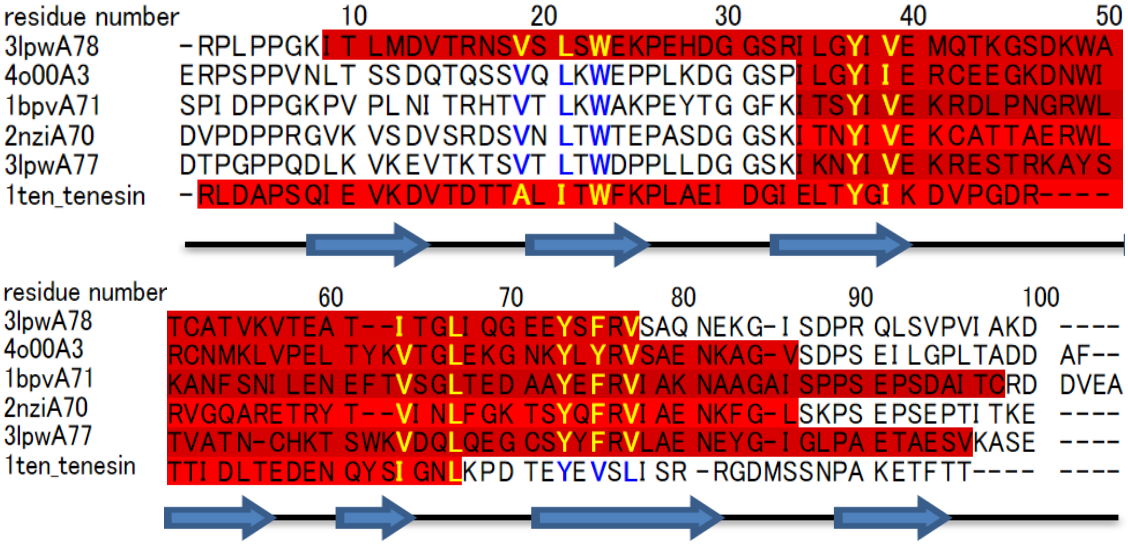


Figure A2: Structure-based multiple sequence alignment for 6 FN3 domains showing the result of ADM prediction.

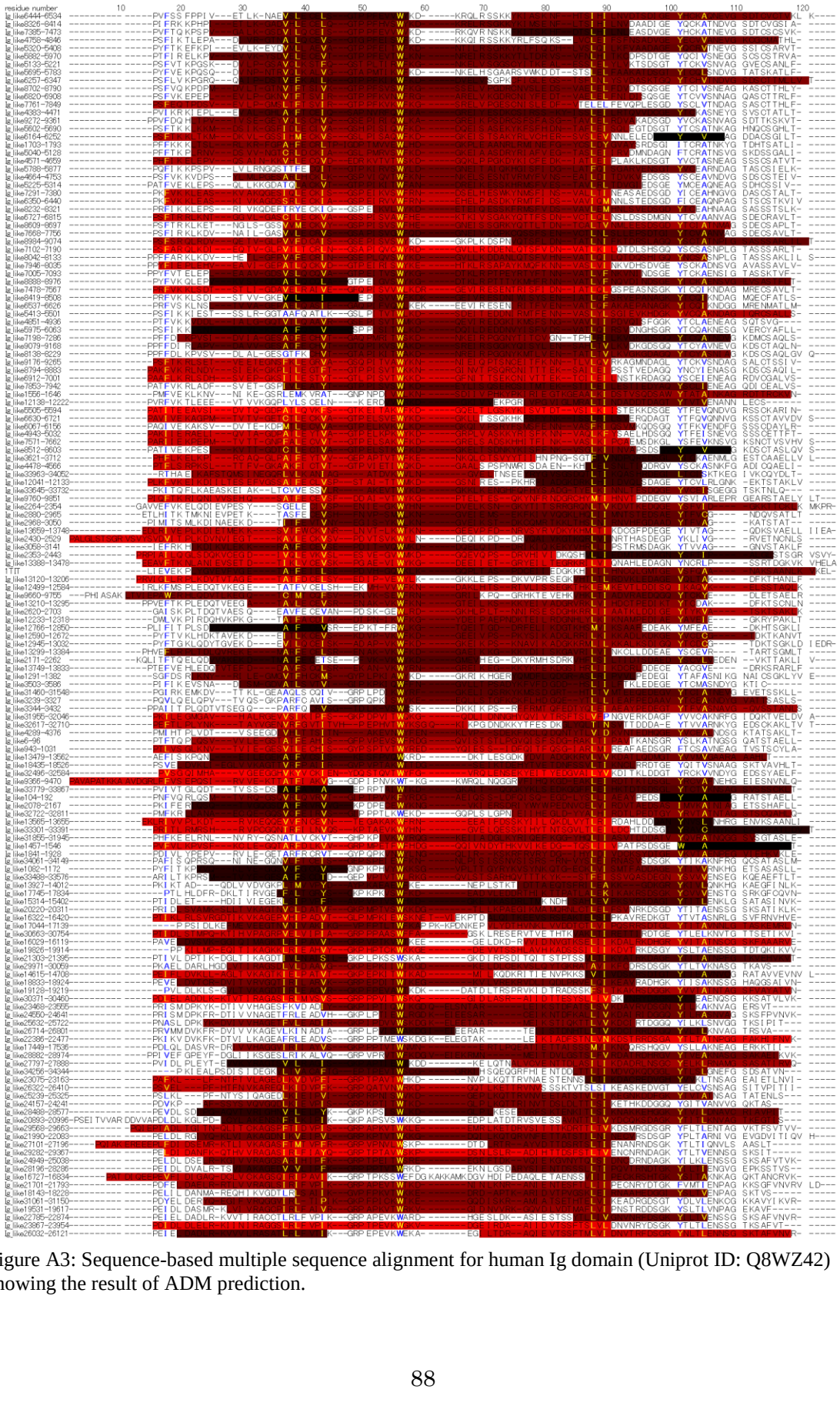


Figure A3: Sequence-based multiple sequence alignment for human Ig domain (Uniprot ID: Q8WZ42) showing the result of ADM prediction.

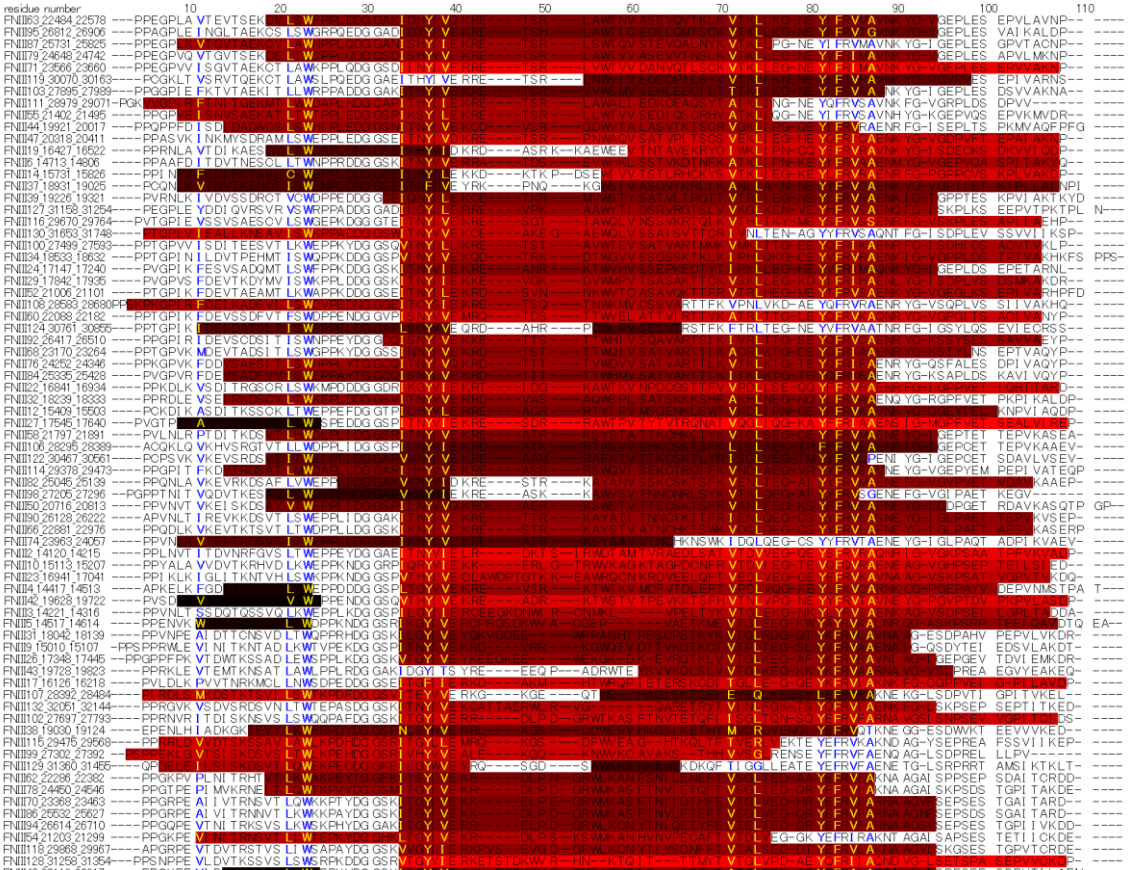


Figure A4: Sequence-based multiple sequence alignment for human FN3 domain (Uniprot ID: Q8WZ42) showing the result of ADM prediction.

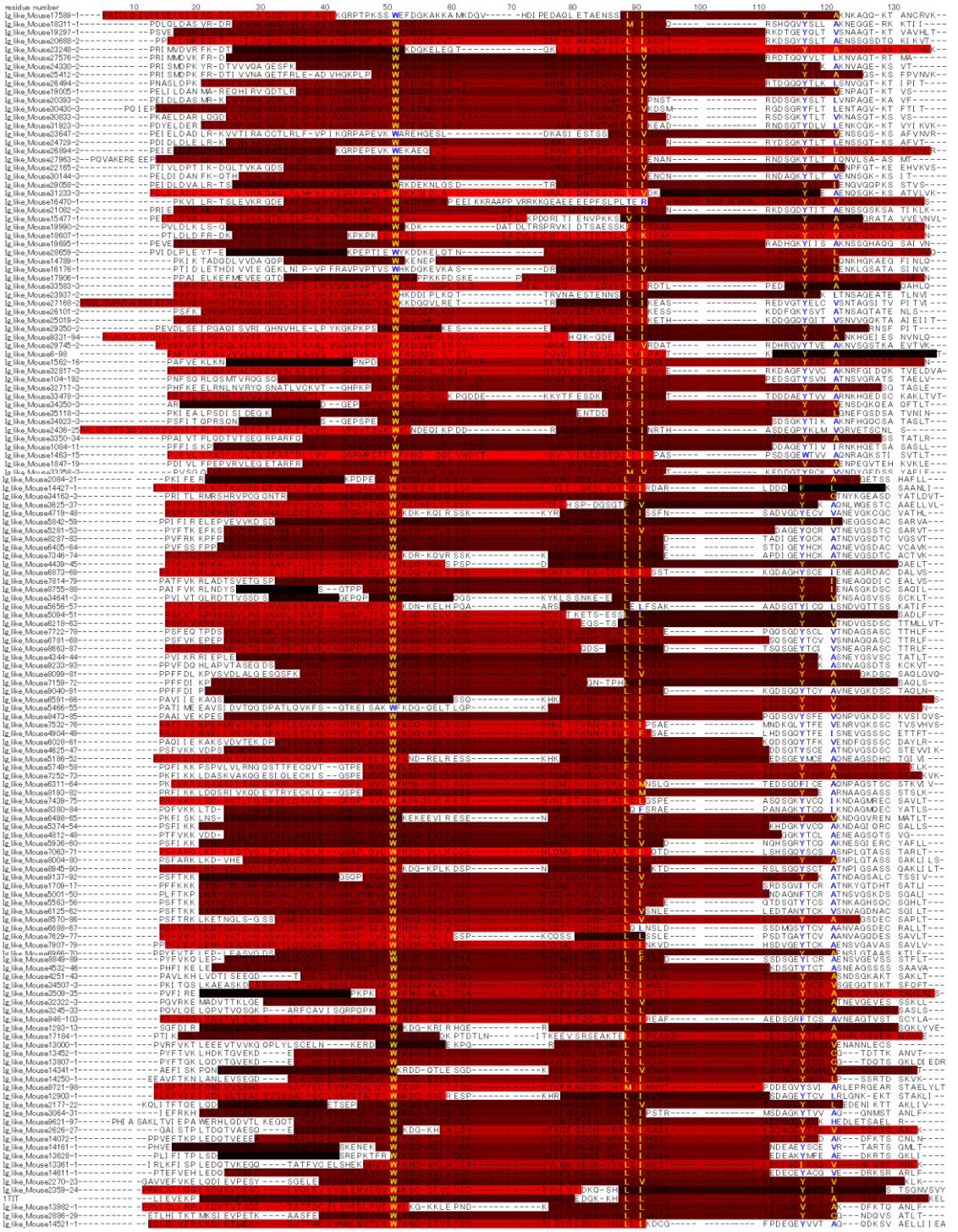


Figure A5: Sequence-based multiple sequence alignment for mouse Ig domain (Uniprot ID: A2ASS6) showing the result of ADM prediction.

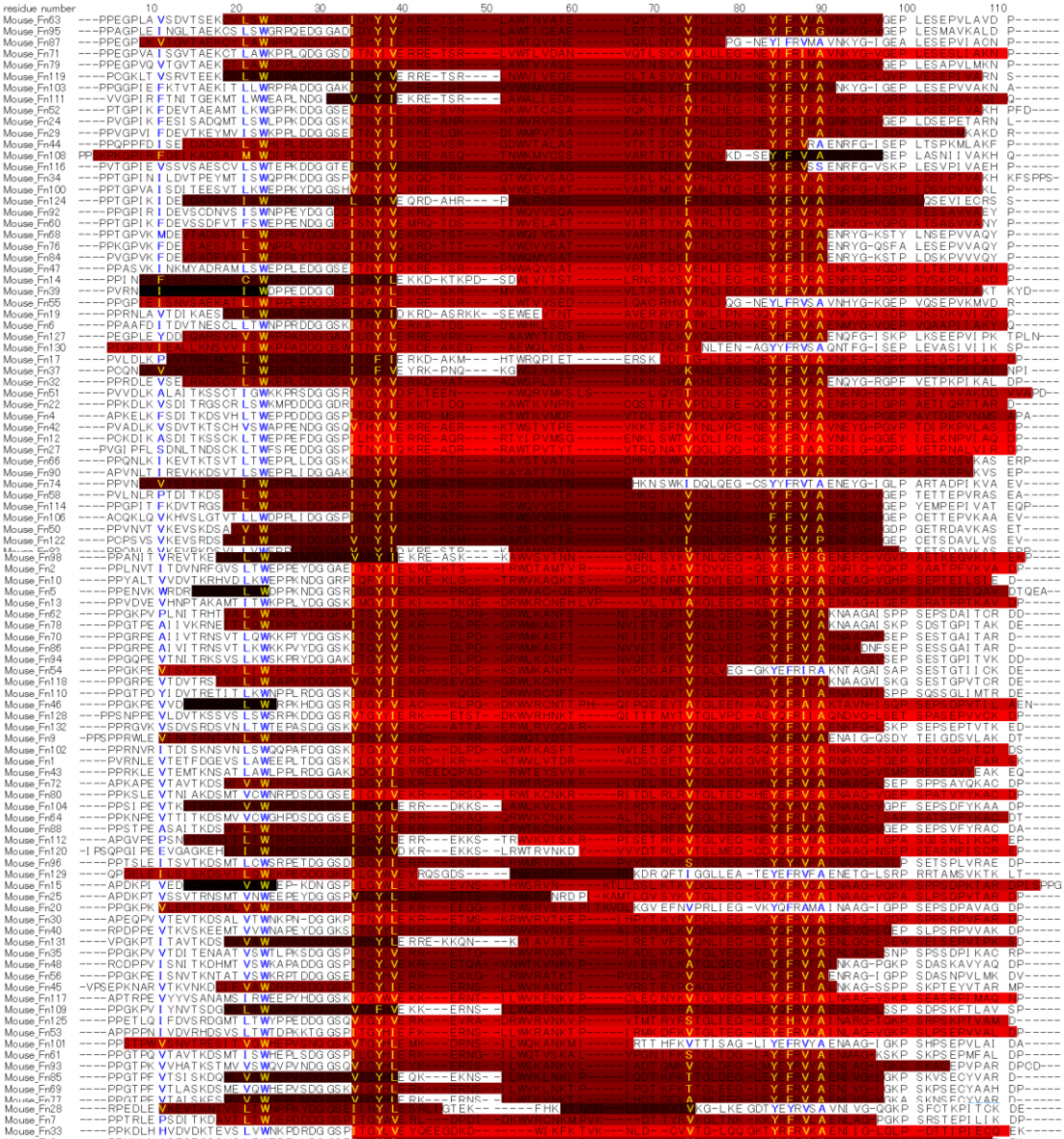


Figure A6: Sequence-based multiple sequence alignment for mouse FN3 domain (Uniprot ID: A2ASS6) showing the result of ADM prediction.

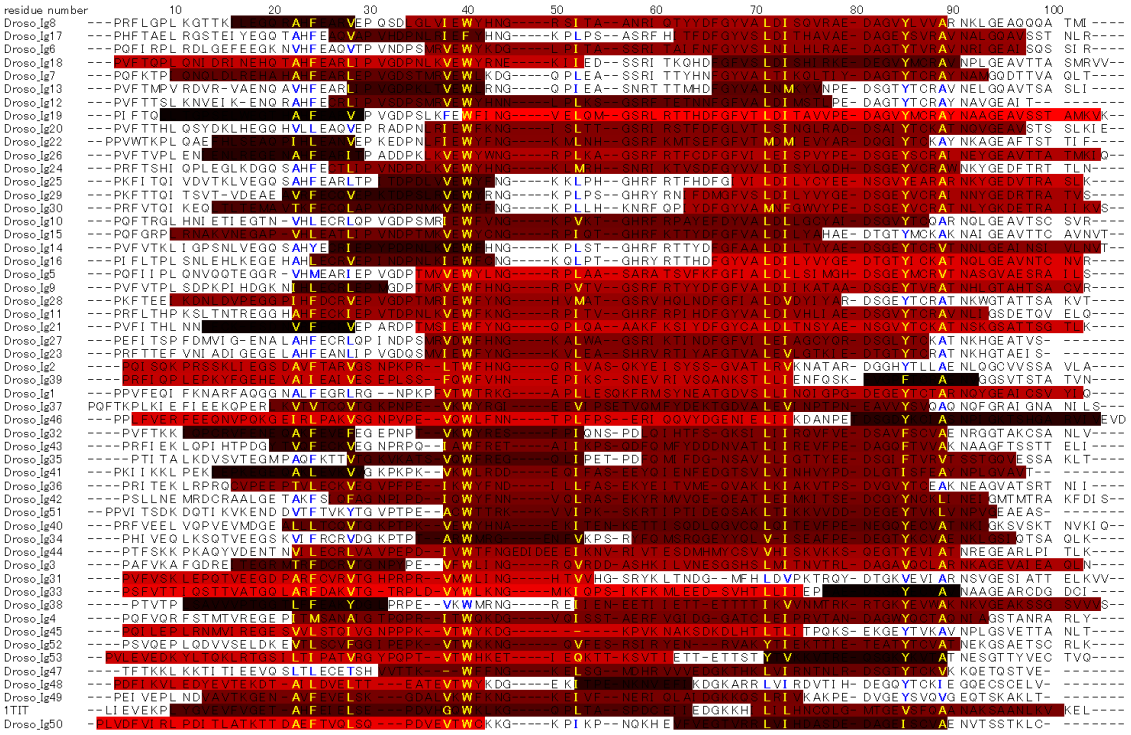


Figure A7: Sequence-based multiple sequence alignment for fruit fly Ig domain (Uniprot ID: Q9I7U4) showing the result of ADM prediction.

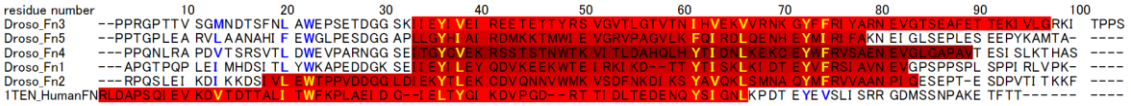


Figure A8: Sequence-based multiple sequence alignment for fruit fly FN3 domain (Uniprot ID: Q9I7U4) showing the result of ADM prediction.

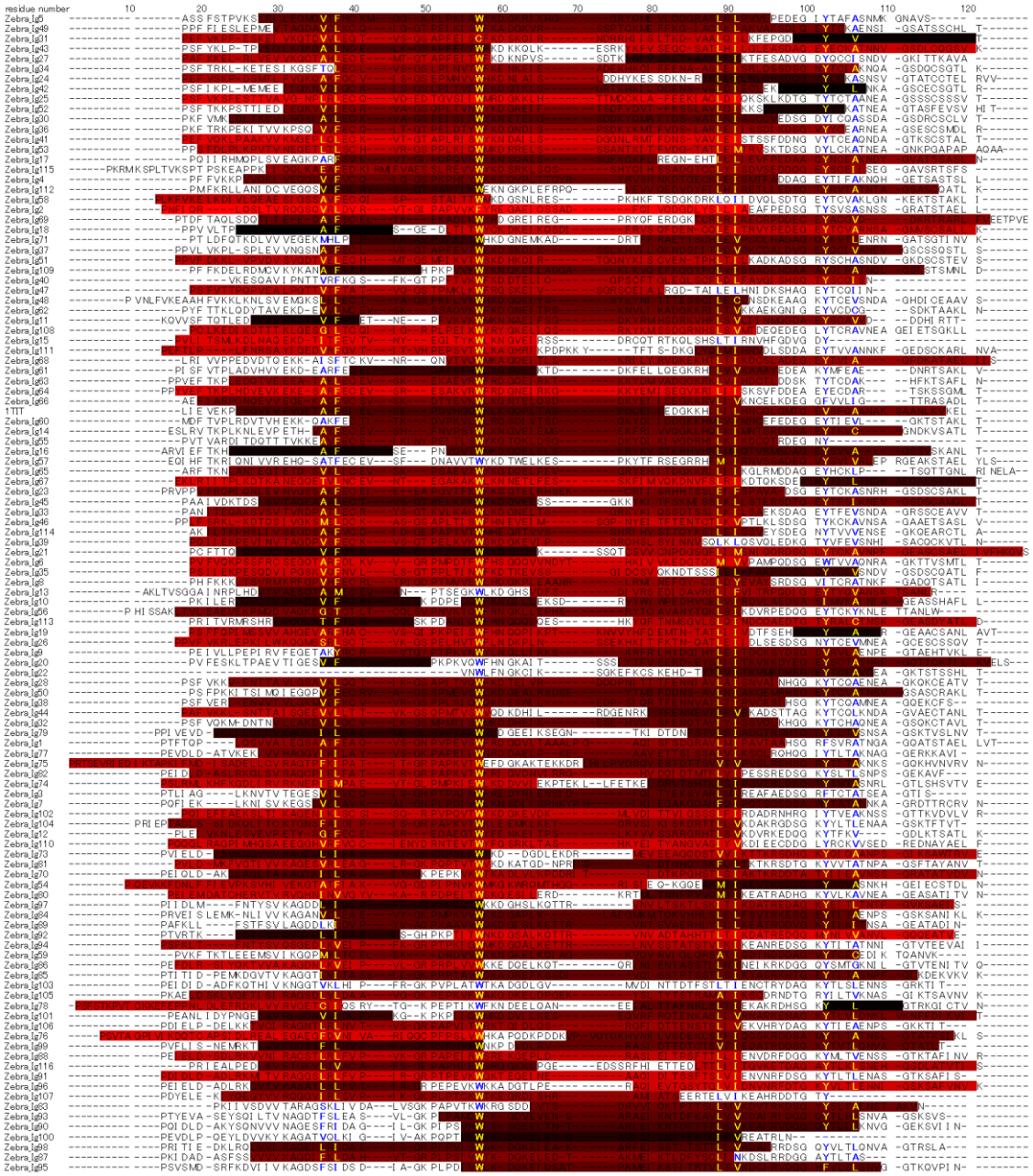


Figure A9: Sequence-based multiple sequence alignment for zebrafish Ig domain (Uniprot ID: A5X6X5) showing the result of ADM prediction.

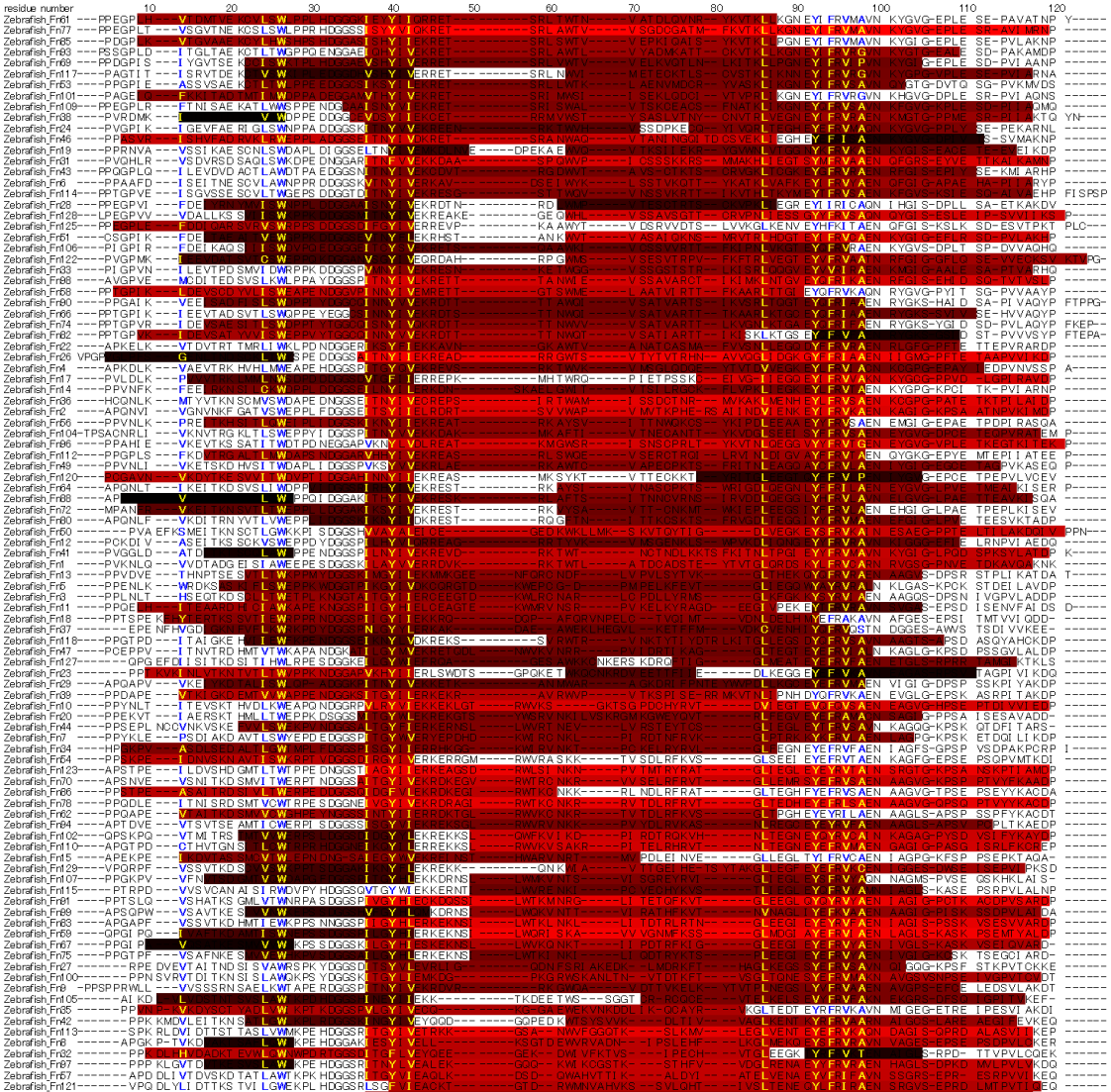


Figure A10: Sequence-based multiple sequence alignment for zebrafish FN3 domain (Uniprot ID: A5X6X5) showing the result of ADM prediction.

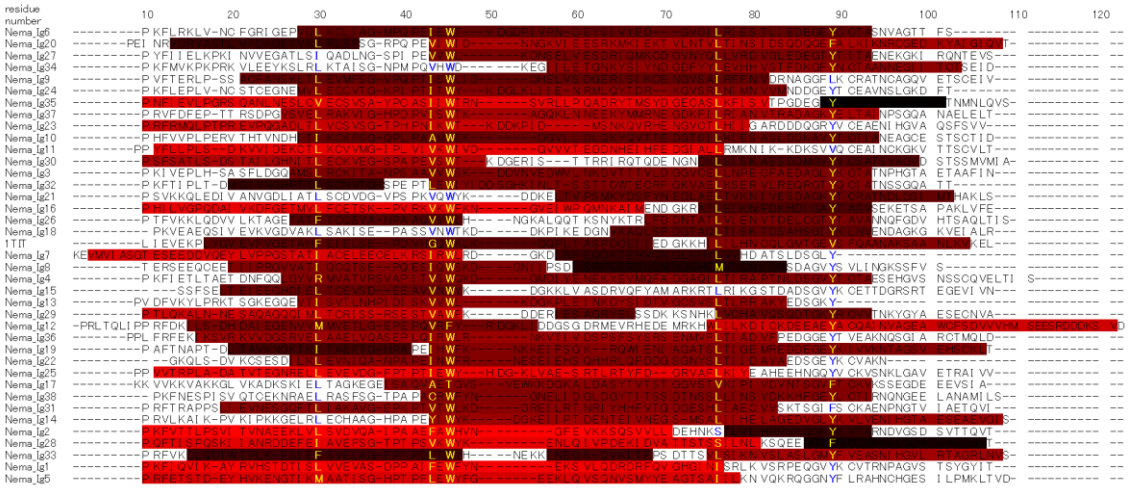


Figure A11: Sequence-based multiple sequence alignment for nematode Ig domain (Uniprot ID: G4SLH0) showing the result of ADM prediction.

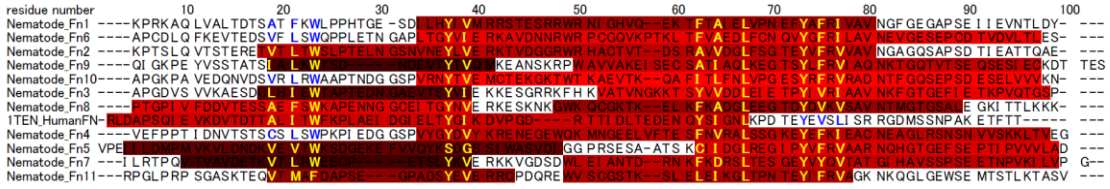


Figure A12: Sequence-based multiple sequence alignment for nematode FN3 domain (Uniprot ID: G4SLH0) showing the result of ADM prediction.

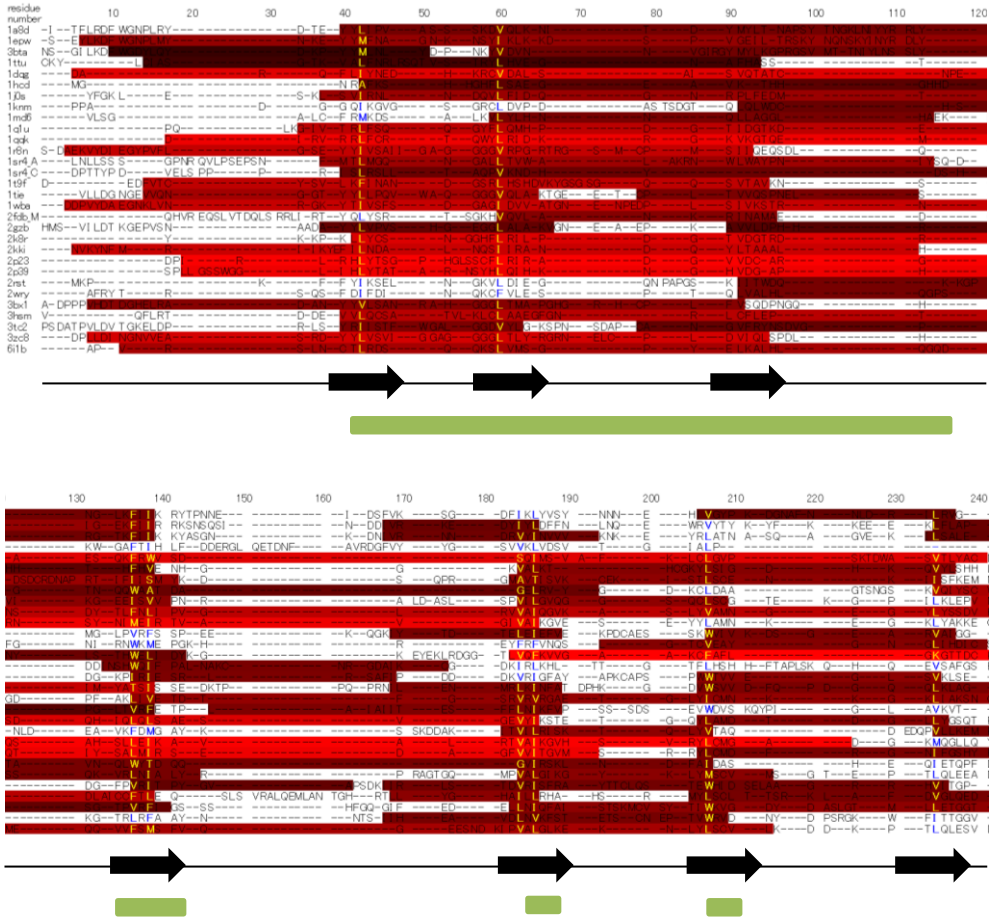
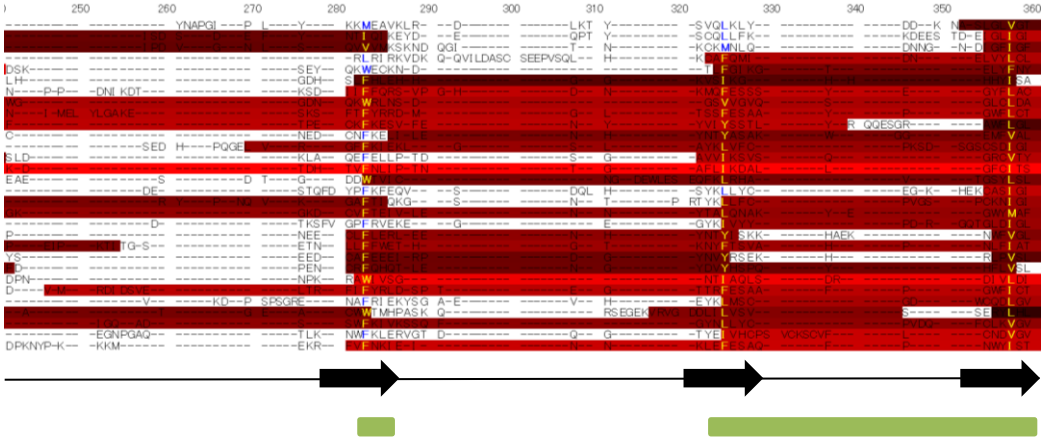
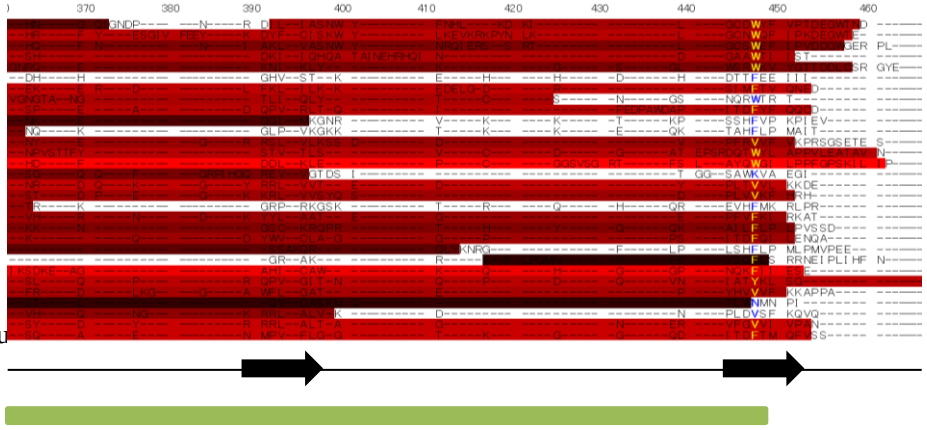


Figure A13: Structure-based multiple sequence alignment for four irregular beta-trefoilproteins and 26 structurally symmetric beta-trefoilproteins with the ADM prediction results. The predicted compact region is indicated by a red bar. The brighter red denotes a higher compact density. The conserved hydrophobic residues in the predicted compact region are yellow, and a blue letter is a residue out of the predicted compact region. A black arrow represents beta-strand. A green bar indicates the conserved predicted compact region when the conservation over than 70% of the aligned site.





Figu

Appendix B

F-value analyses of Ig domains and FN3 domains

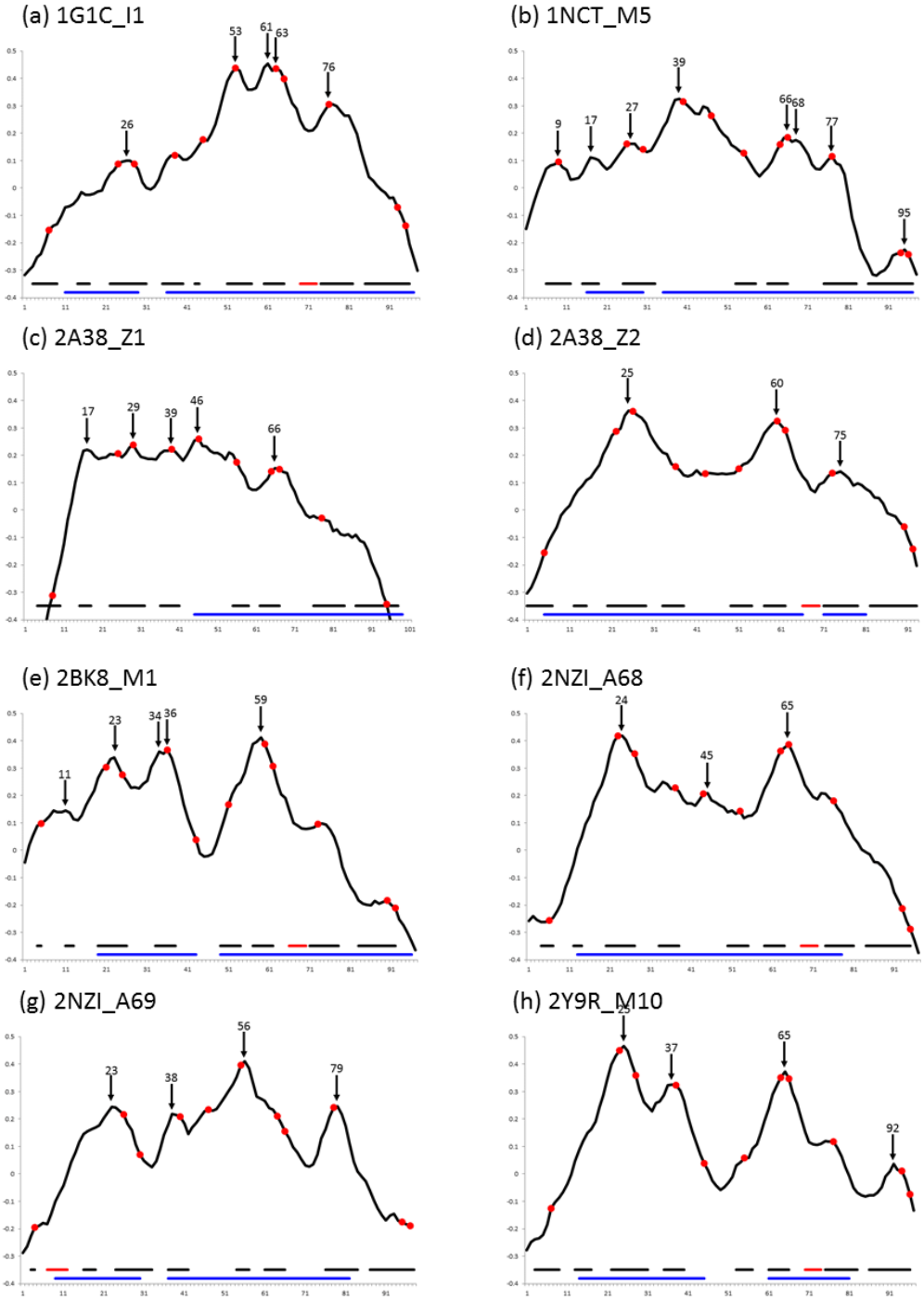


Figure B1: F-value results of known 3D-structure Ig domains. The black line corresponds to F-value plots with conserved hydrophobic residues (red dot) and peak position (black arrow). The black, red and blue bars indicate β-strand, helix and ADM result, respectively.

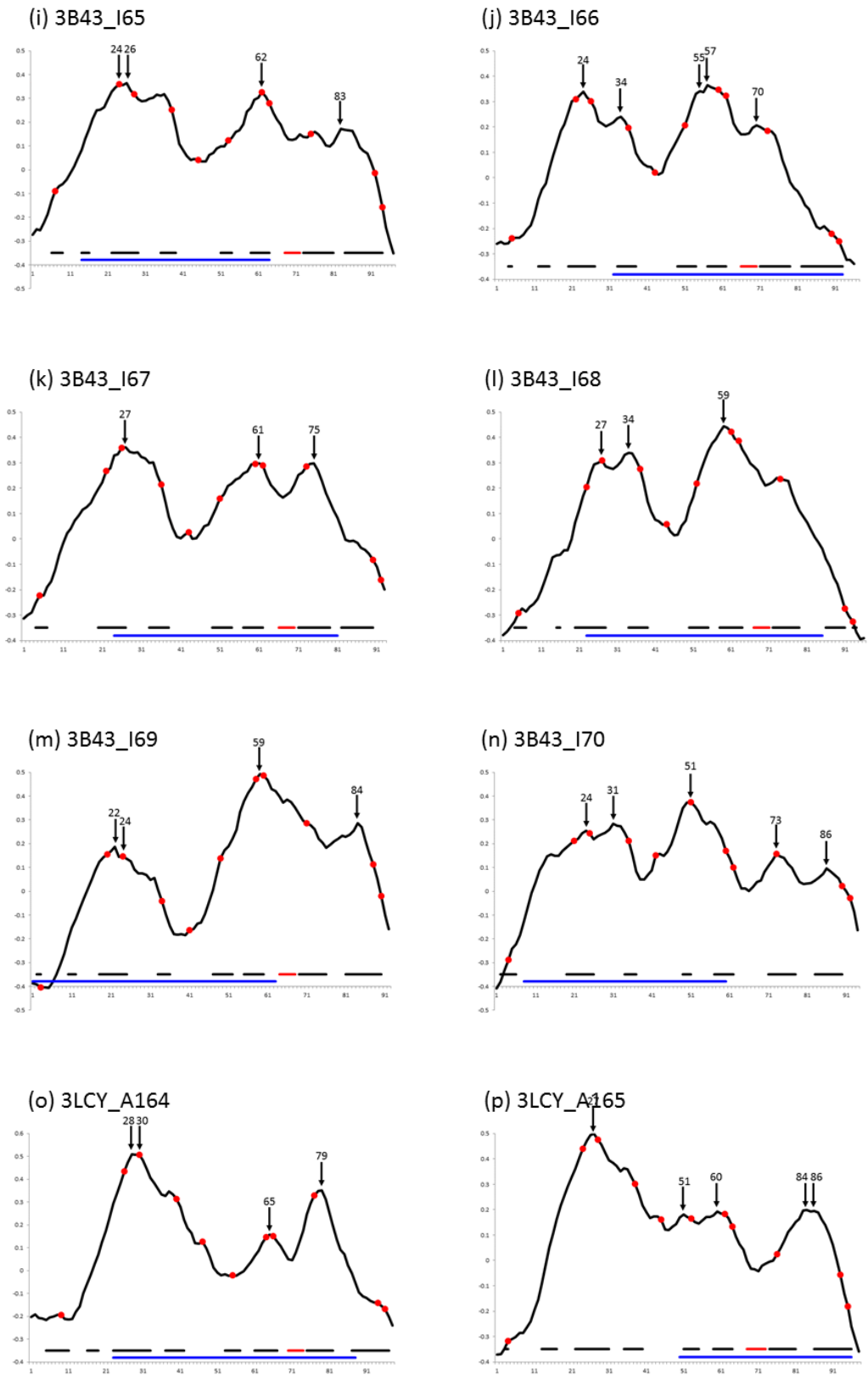


Figure B1: Continued

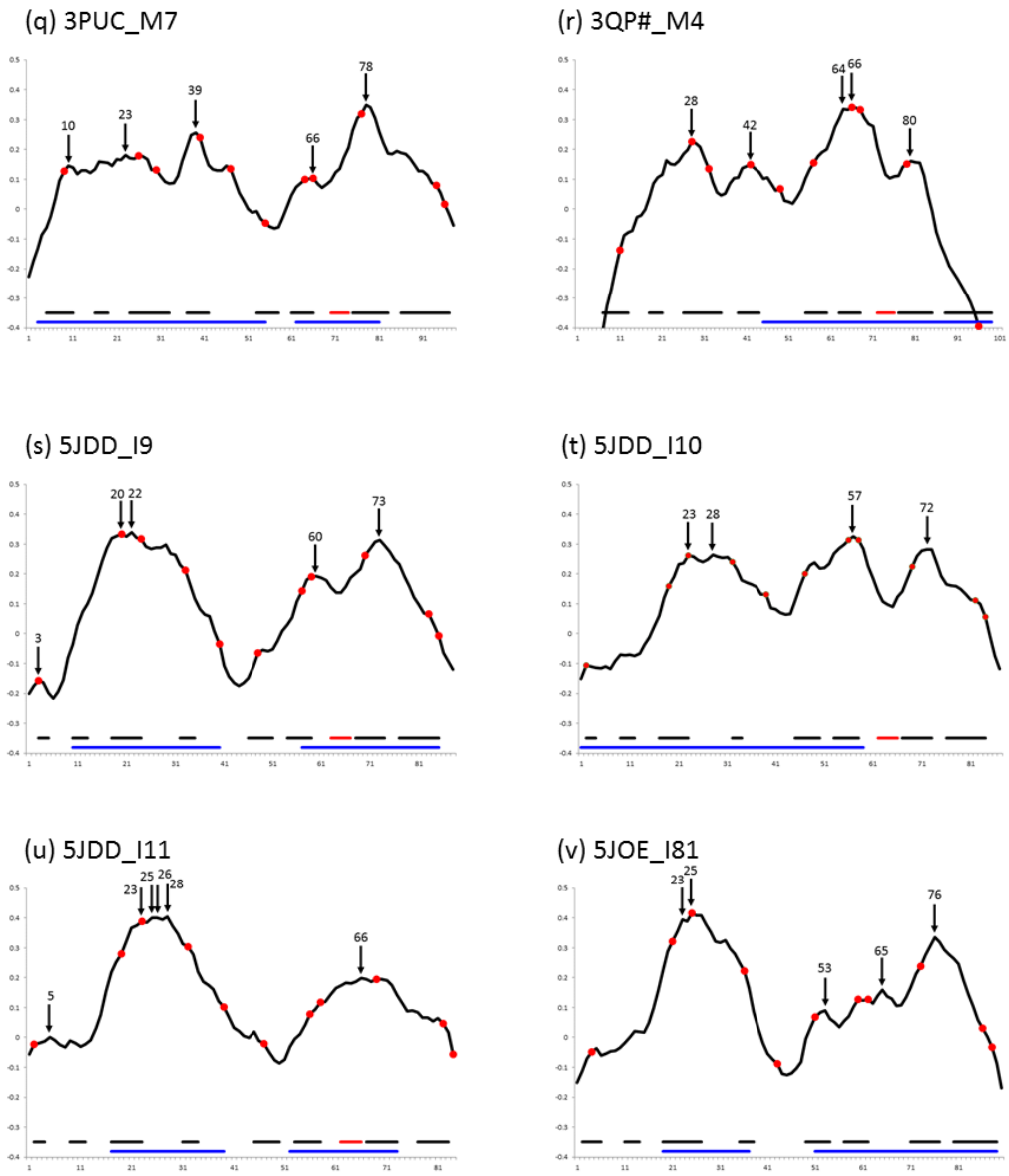


Figure B1: Continued

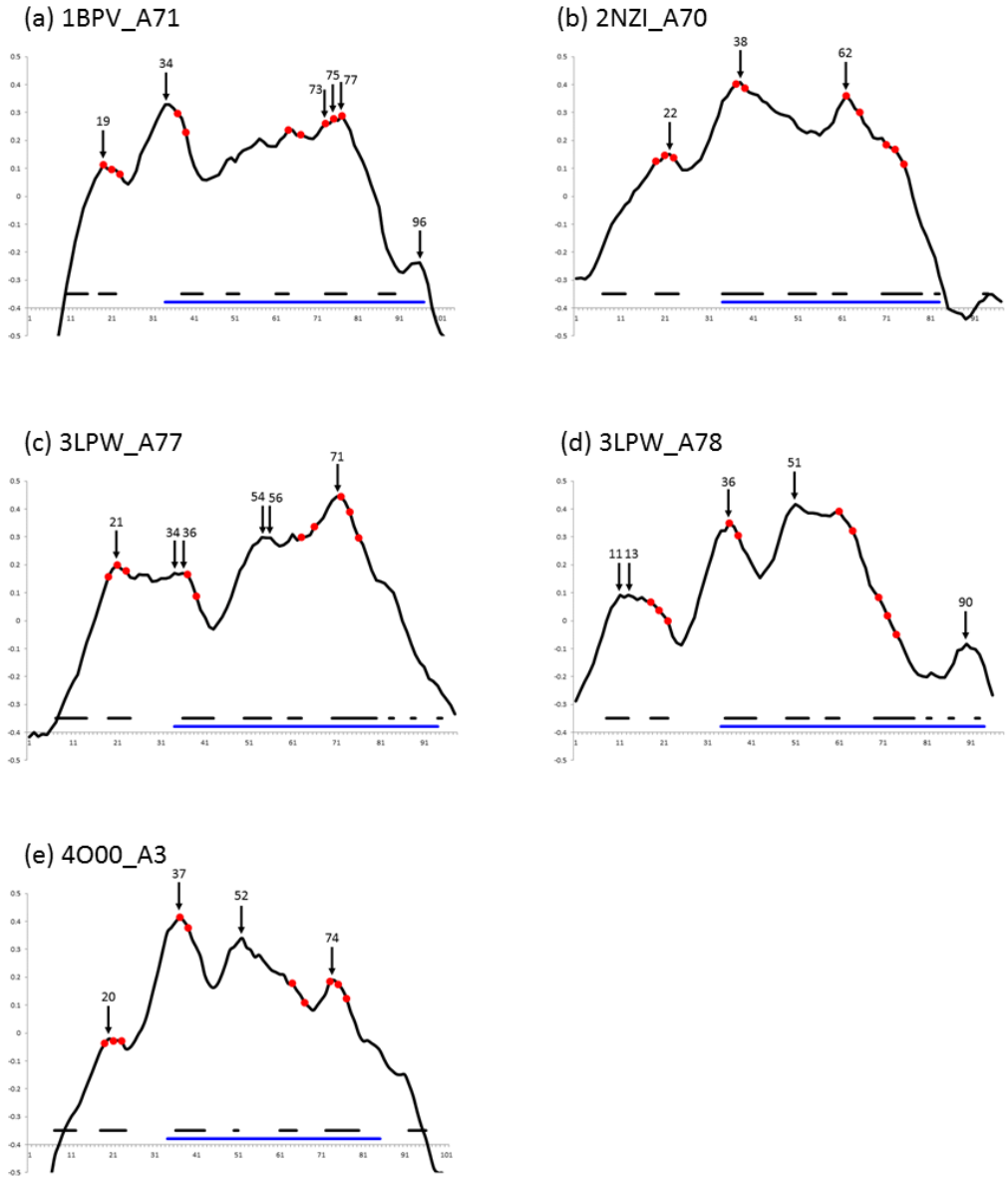


Figure B2: F-value analyses for each known 3D-structure FN3-domain. The black histogram corresponds to F-value plots with conserved hydrophobic residues (red dot) and peak position (black arrow). The black, red and blue bars represent β-strand, helix and ADM result, respectively.

Appendix C

Native structure of Ig domains and FN3 domains with predicted compact region and conserved hydrophobic residues

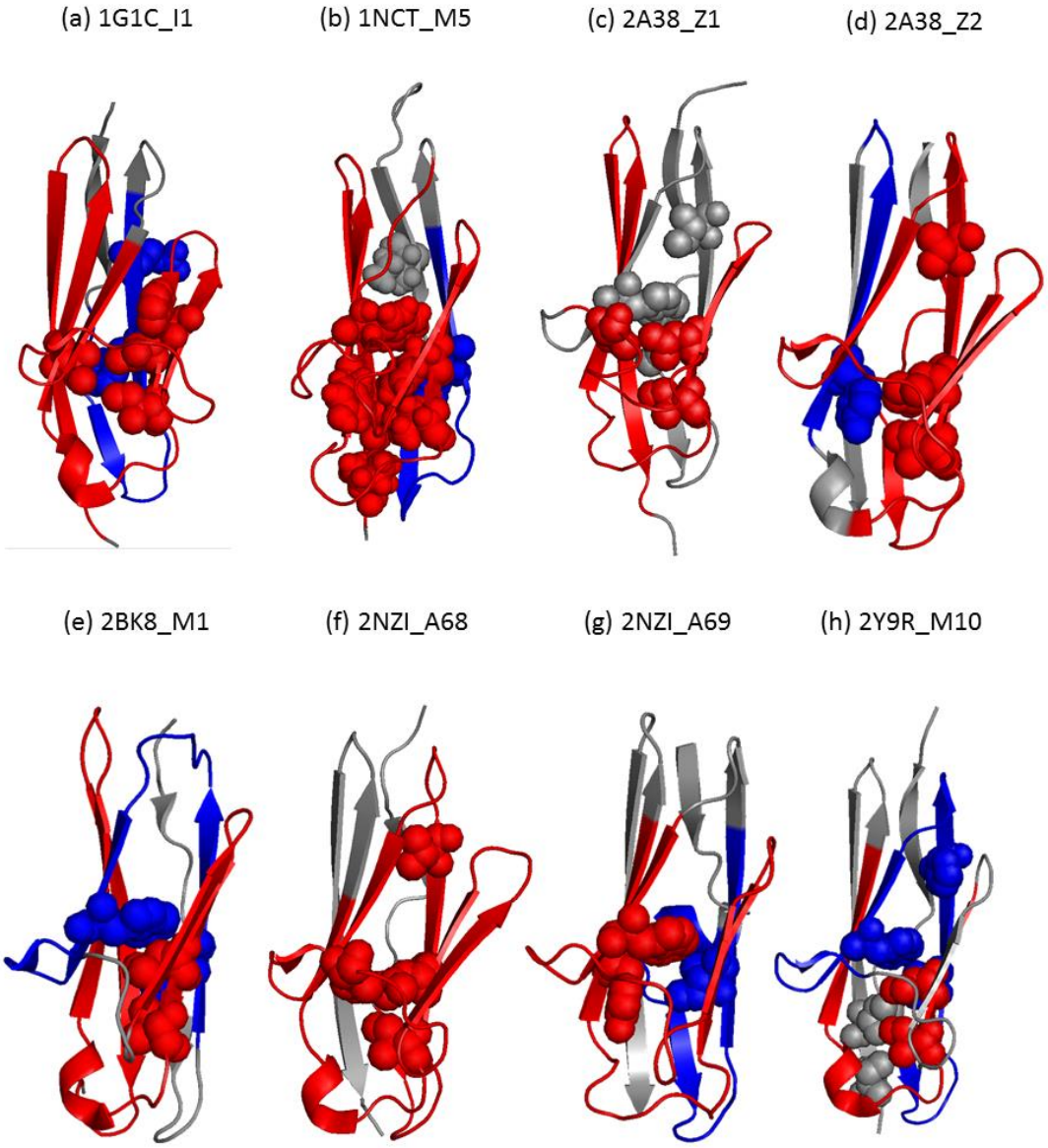


Figure C1: Position of conserved hydrophobic residue within 5 residues of the F-value peaks of the Ig domains.

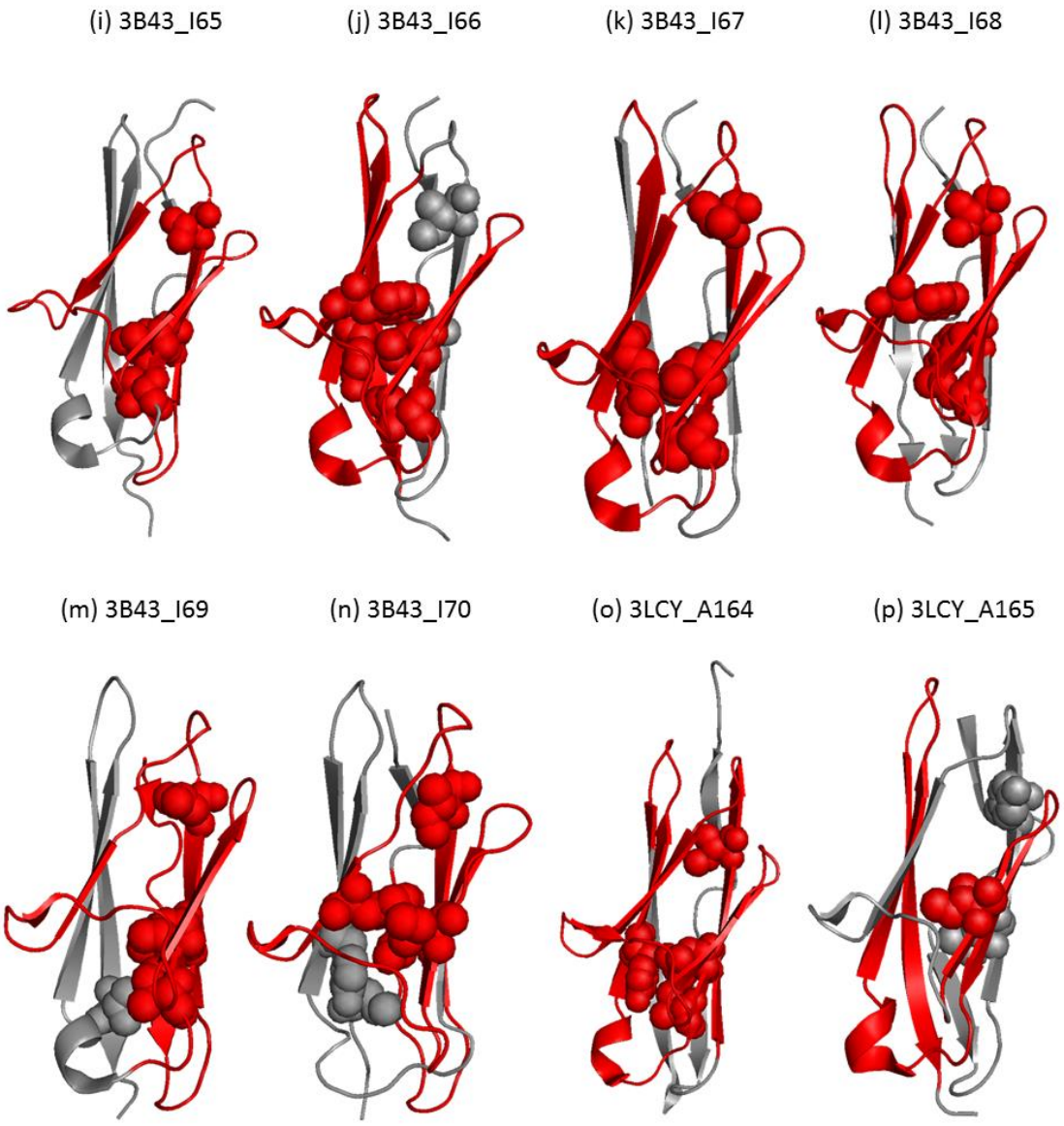


Figure C1: Continued

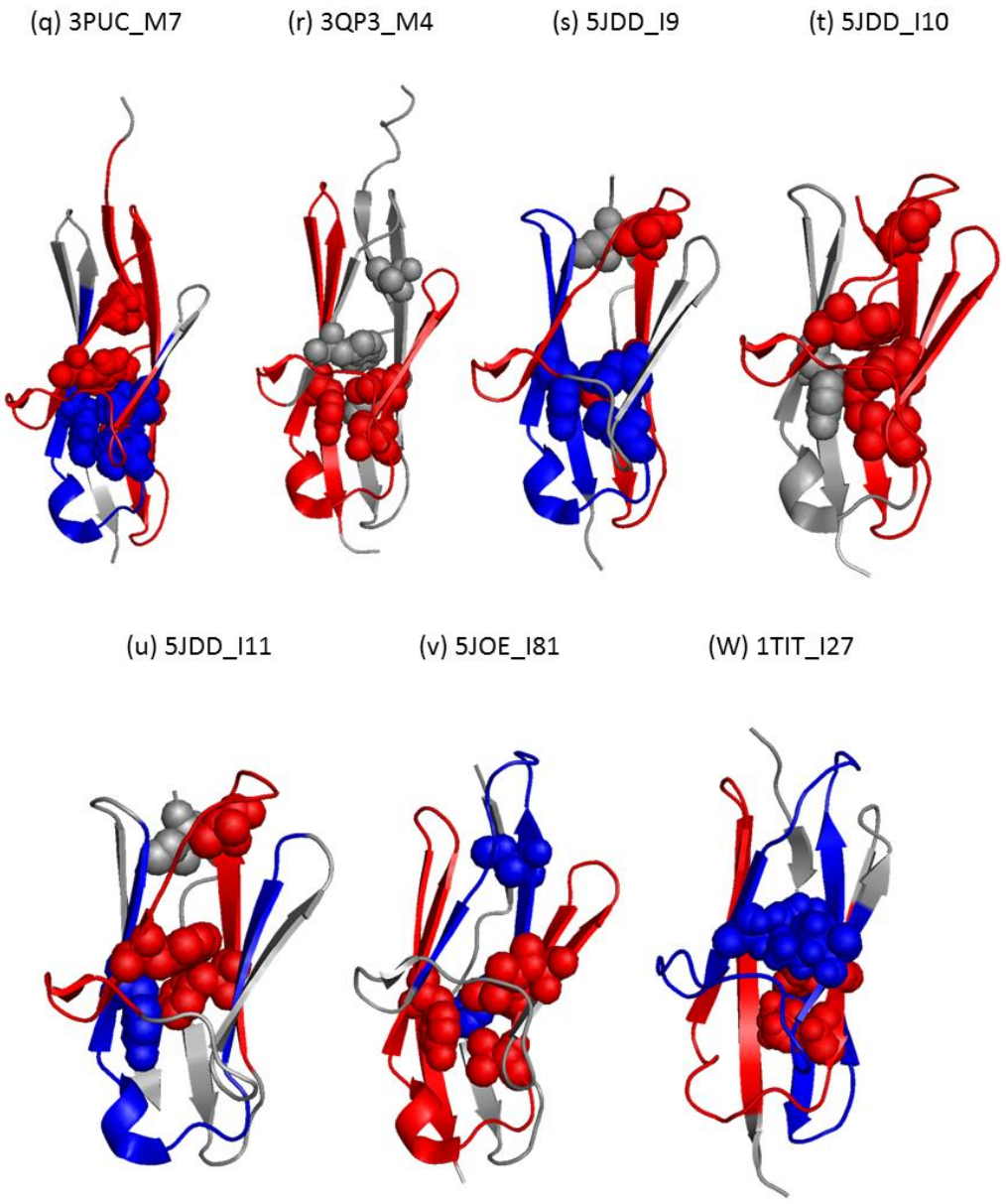


Figure C1: Continued

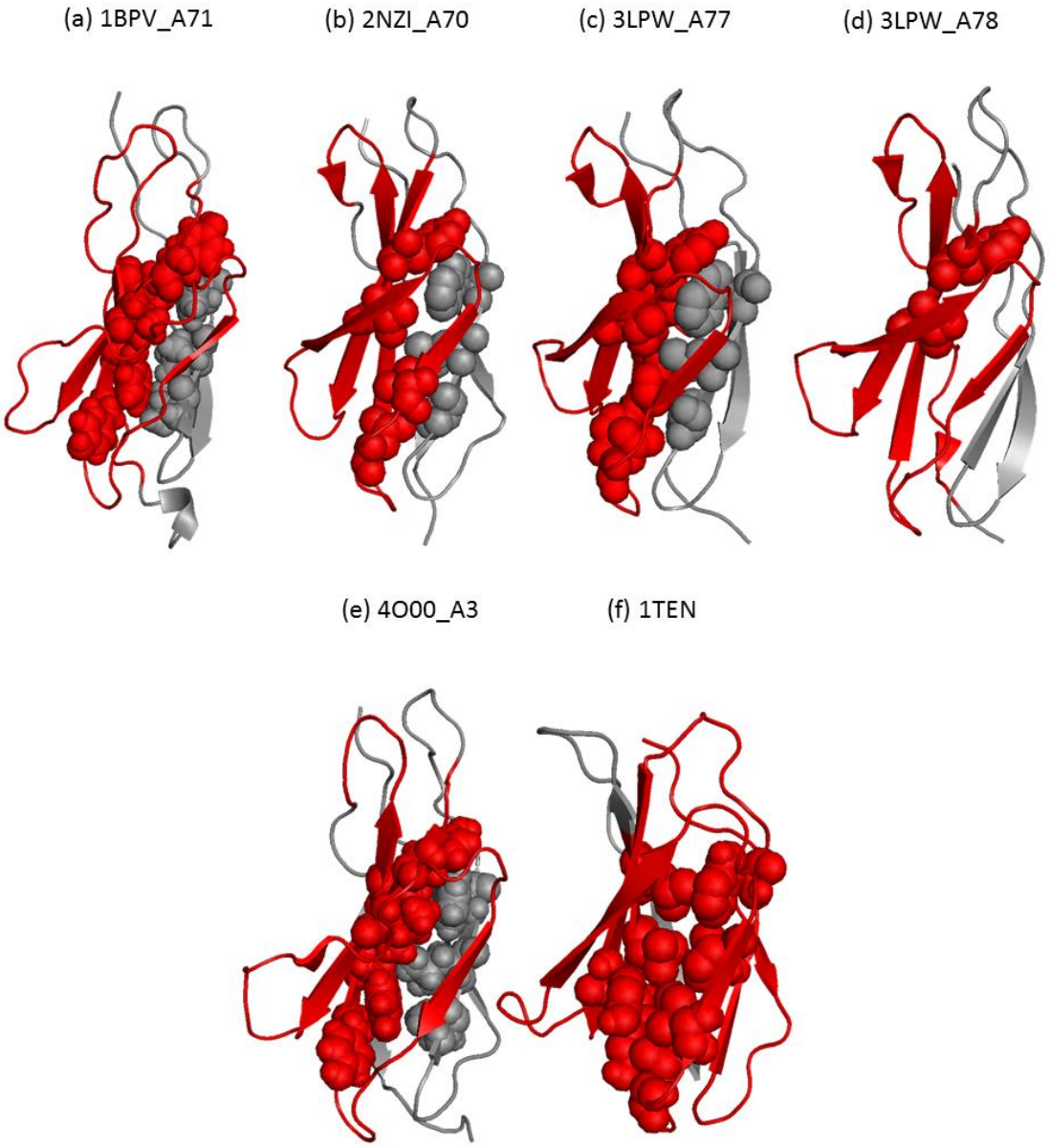
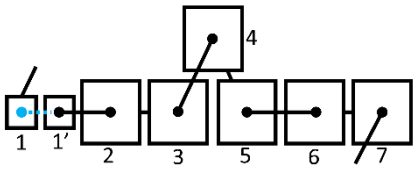
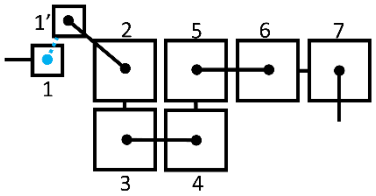
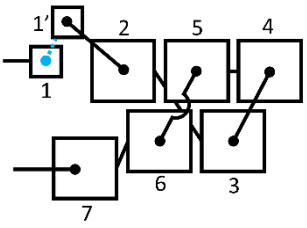
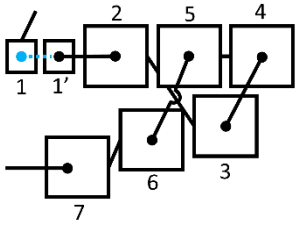


Figure C2: Position of conserved hydrophobic residues within 5 residues of the F-value peaks of the FN3 domains.

Appendix D

The folding processes of Ig domain and FN3 domain

Q value	Structure
0.05	
0.10	
0.15 – 0.20	
0.25	

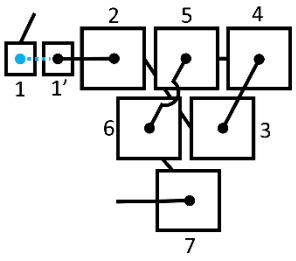
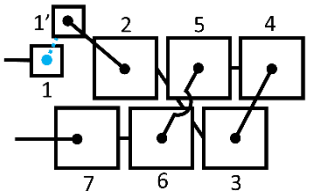
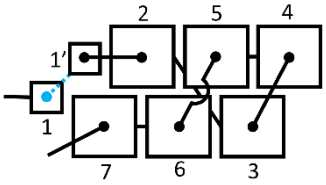
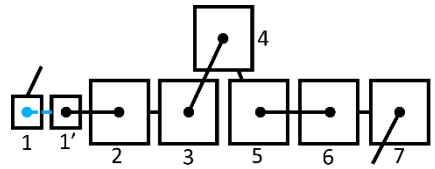
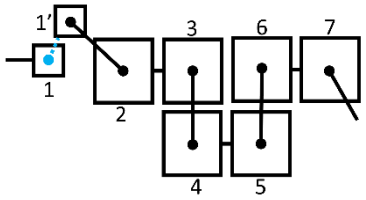
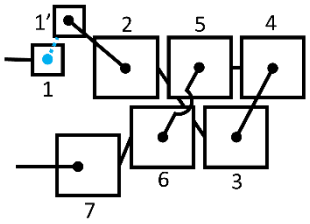
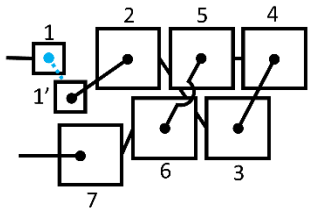
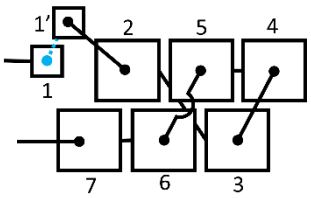
Q value	Structure
0.30	
0.35	
0.40	

Figure D1: The arrangement of β strands of 1TIT during folding processes constructed from a kind of contact frequency map derived from Go model simulations. Q value corresponds to the ratio of the number of the native contact detected in each state to all native contacts. A blue connecting line indicates no contact formed among adjacent β strands.

Q value	Structure
0.05	
0.10	
0.15 - 0.25	
0.30	
0.35	

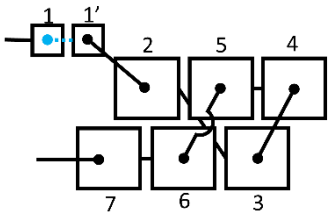
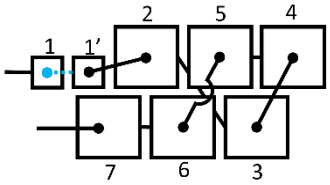
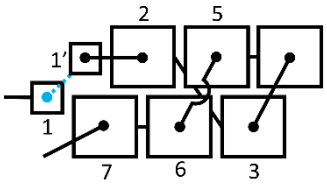
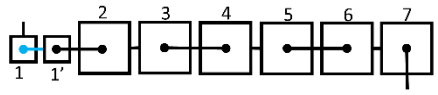
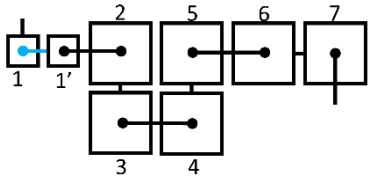
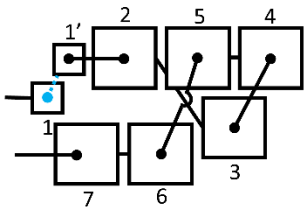
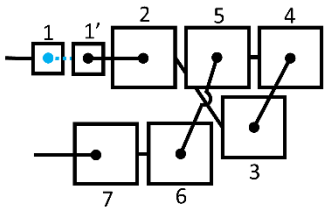
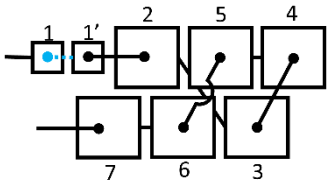
Q value	Structure
0.40 - 0.50	
0.55 - 0.65	
0.70	

Figure D2: The arrangement of β strands of 2A38_Z1 during folding processes constructed from a kind of contact frequency map derived from Go model simulations. Q value corresponds to the ratio of the number of the native contact detected in each state to all native contacts. A blue connecting line indicates no contact formed among adjacent β strands.

Q value	Structure
0.05	
0.10	
0.15	
0.20 - 0.25	
0.30	

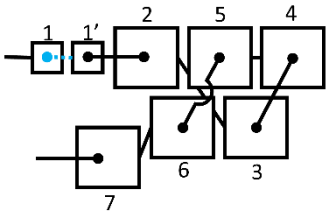
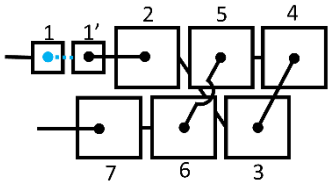
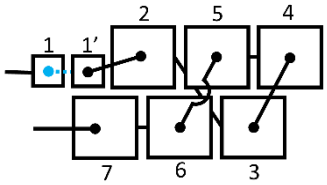
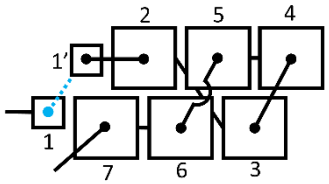
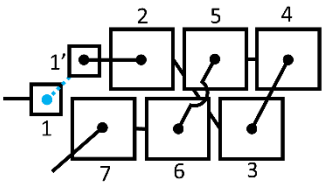
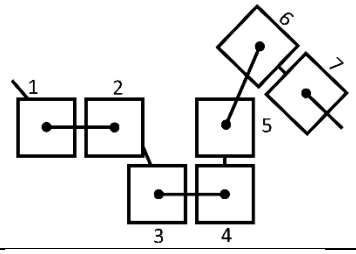
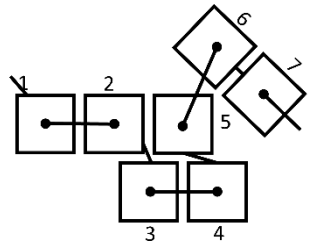
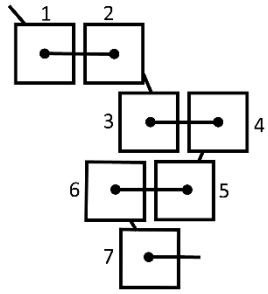
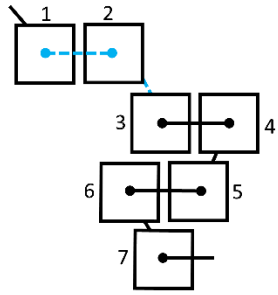
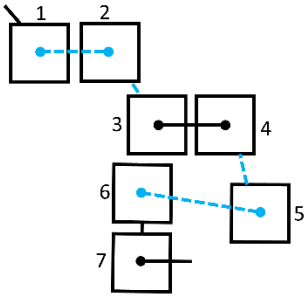
Q value	Structure
0.35	
0.40 - 0.50	
0.55	
0.60	
0.65	

Figure D3: The arrangement of β strands of 3LCY_A165 during folding processes constructed from a kind of contact frequency map derived from Go model simulations. Q value corresponds to the ratio of the number of the native contact detected in each state to all native contacts. A blue connecting line indicates no contact formed among adjacent β strands.

Q value	Structure
0.05	
0.10 - 0.15	
0.20	
0.25	
0.30	

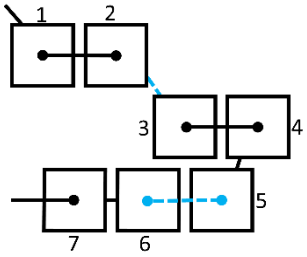
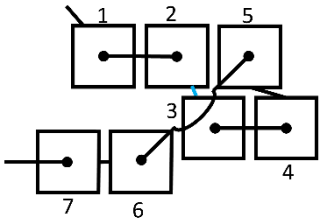
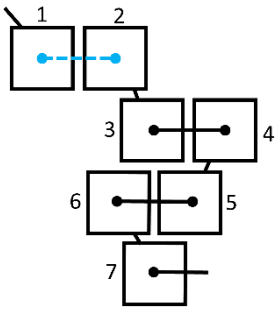
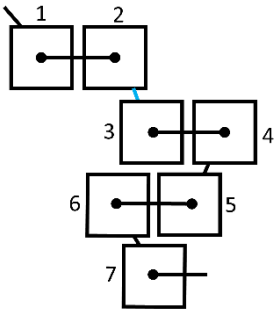
Q value	Structure
0.35	
0.40	
0.45	
0.50	

Figure D4: The arrangement of β strands of 1TEN during folding processes constructed from a kind of contact frequency map derived from Go model simulations. Q value corresponds to the ratio of the number of the native contact detected in each state to all native contacts. A blue connecting line indicates no contact formed among adjacent β strands.

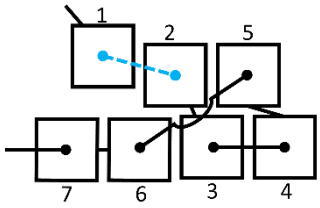
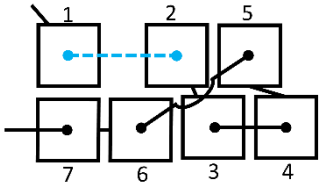
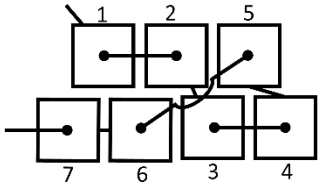
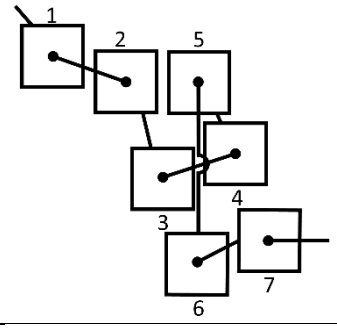
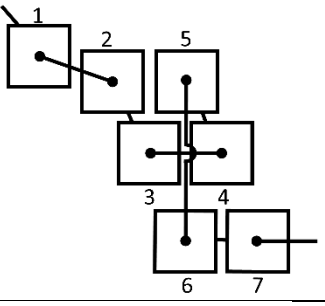
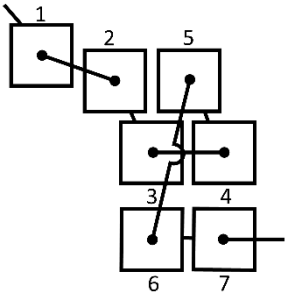
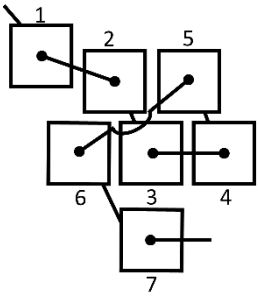
Q value	Structure
0.55	
0.60	
0.65	

Figure D4: Continued

Q value	Structure
0.05	
0.10	
0.15	
0.20	

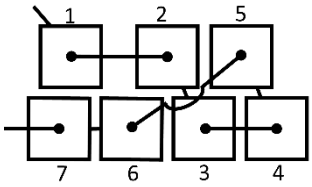
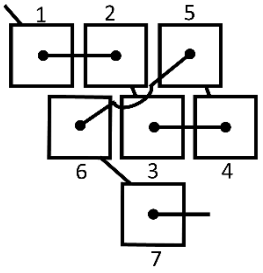
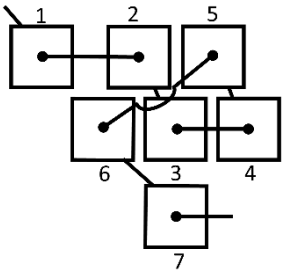
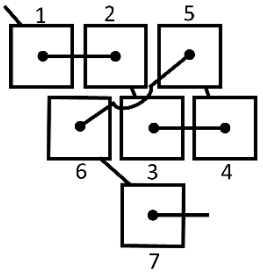
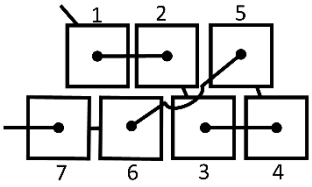
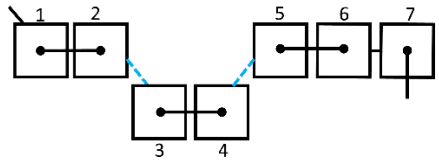
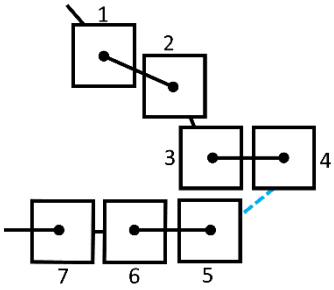
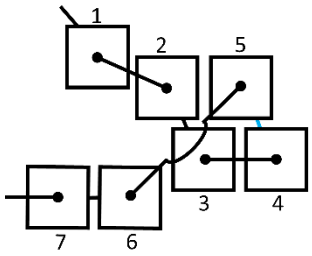
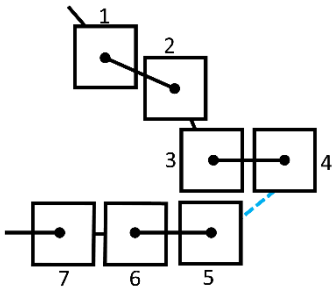
Q value	Structure
0.25	
0.30	
0.35 - 0.50	
0.55	
0.60	

Figure D5: The arrangement of β strands of 2NZI_A70 during folding processes constructed from a kind of contact frequency map derived from Go model simulations. Q value corresponds to the ratio of the number of the native contact detected in each state to all native contacts. A blue connecting line indicates no contact formed among adjacent β strands.

Q value	Structure
0.05	
0.10 - 0.25	
0.30	
0.35	

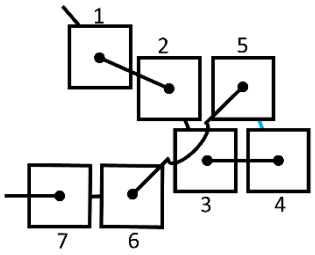
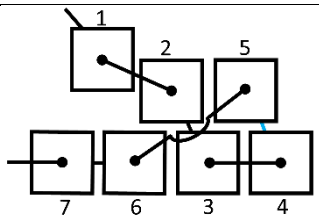
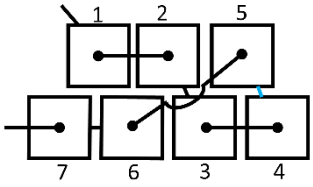
Q value	Structure
0.40	
0.45	
0.50	

Figure D6: The arrangement of β strands of 1BPV_A62 during folding processes constructed from a kind of contact frequency map derived from Go model simulations. Q value corresponds to the ratio of the number of the native contact detected in each state to all native contacts. A blue connecting line indicates no contact formed among adjacent β strands.

Appendix E

Conserved hydrophobic packing of beta-trefoilprotein

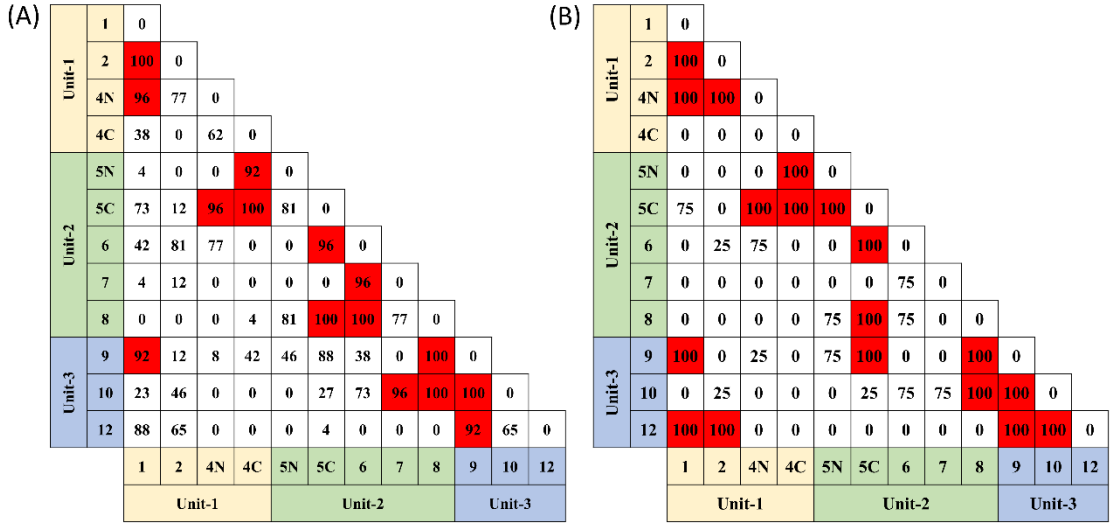


Figure E1: Occurrence ratio of contacts formed by conserved hydrophobic residues of (A) 26 high symmetric beta-trefoil proteins and (B) four irregular beta-trefoil proteins. A number shown on the x-axes and y-axes denotes a conserved hydrophobic residue. The occurrence ratio is indicated by a percentage in each cell. Red color indicates 100%.

Appendix F

Contact frequency map derived from Gō-model simulation

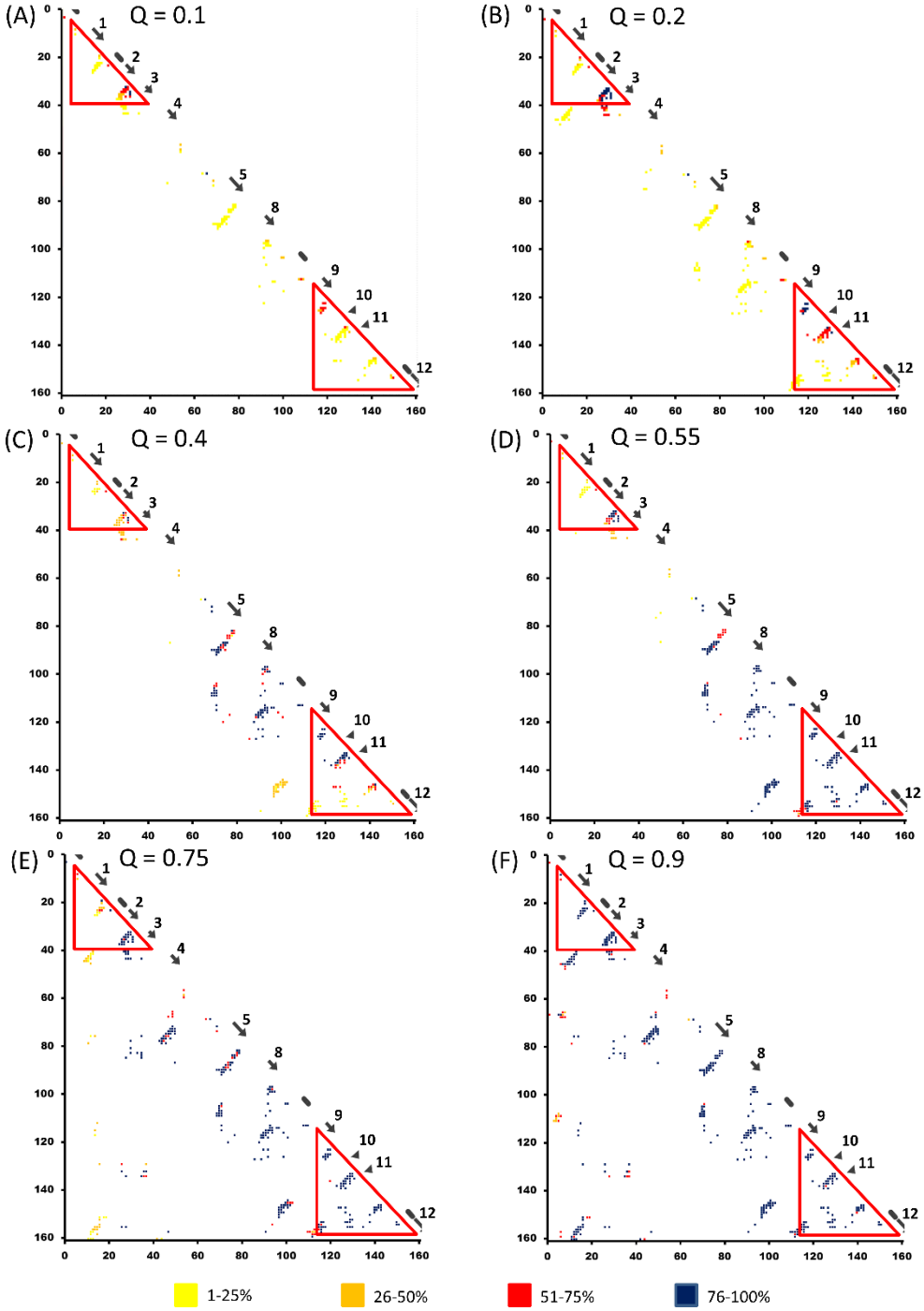


Figure F1: Contact frequency map derived from Gō-model simulation of 1TTU. A red triangle represents a PdCR derived from ADM analysis.

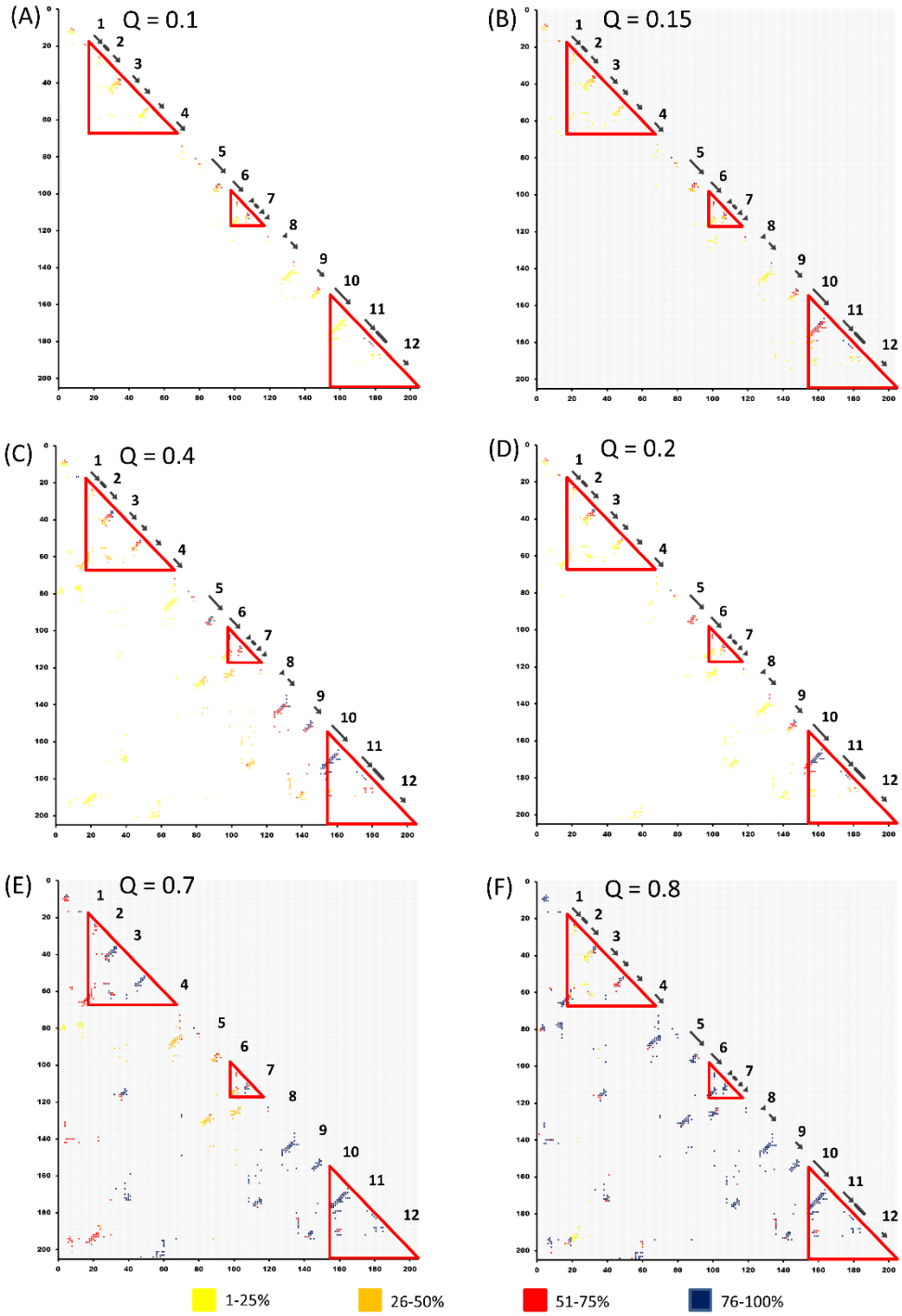


Figure F2: Contact frequency map derived from Gō-model simulation of 1A8D. A red triangle represents a PdCR derived from ADM analysis.

Appendix G

Tables

Table G1: Results of the ADMs for the proteins treated by structure-based multiple sequence alignment.

Domain	PDB ID	Description	Location on titin domain chain (Uniprot ID: Q8WZ42)	Compact region (□-value)
Fibronectin type III (FN3)	1TEN			1-62 (0.346)
	1BPV	A71	FN3 62	34-98 (0.282)
	2NZI	A70	FN3 132	34-83 (0.352)
	3LPW	A77	FN3 66	34-94 (0.297)
	3LPW	A78	FN3 67	8-72 (0.304)
	4O00	A3	FN3 3	34-85 (0.311)
Immunoglobulin (Ig)	1G1C	I1	Ig 10	11-29 (0.115), 36-97 (0.447)
	1NCT	M5	Ig 143	16-30 (0.100), 35-97 (0.391)
	1TIT	I27	Ig 85	8-50 (0.157), 57-86 (0.240)
	2A38	Z1	Ig 1	45-99 (0.299)
	2A38	Z2	Ig 2	5-66 (0.351), 71-81 (0.021)
	2BK8	M1	Ig 143	19-43 (0.137), 49-96 (0.412)
	2NZI	A68	Ig 141	13-78 (0.242)
	2NZI	A69	Ig 142	9-30 (0.116), 37-82 (0.227)
	2Y9R	M10	Ig 152	14-45 (0.160), 61-81 (0.206)
	3B43	I65	Ig 62	14-64 (0.197)
	3B43	I66	Ig 63	32-93 (0.414)
	3B43	I67	Ig 64	24-81 (0.212)
	3B43	I68	Ig 65	23-85 (0.272)
	3B43	I69	Ig 66	1-63 (0.357)
	3B43	I70	Ig 67	8-60 (0.236)
	3LCY	A164	Ig 140	23-88 (0.222)
	3LCY	A165	Data missing	51-96 (0.395)
	3PUC	M7	Ig 149	3-55 (0.291), 62-81 (0.149)
	3QP3	M4	Ig 146	20-83 (0.218)
	5JDD	I10	Ig 16	1-59 (0.319)
	5JDD	I11	Ig 17	17-39 (0.175), 52-73 (0.107)
	5JDD	I9	Data missing	10-40 (0.149), 57-85 (0.140)
	5JOE	I81	Data missing	19-37 (0.158), 51-89 (0.398)

Table G2: Hydrophobic residues near the peaks of the F-value plots for proteins treated.

Domain	PDB ID	Peaks of a F-value	Hydrophobic residues near a peak in a F-value plot
Fibronectin type 3 (FN3)	1TEN	20, 32, 58, 70	L1-A18, β2-I21, β2-W22, β3-Y36, β3-I38, β5-I59, L5-L62, β6-Y68, β6-V70, β6-L72
	1BPV	19, 34,73, 75, 77, 96	β2-V19, β2-L21, L2-W23, L2-Y37, β3-V39, β5-V64, L5-L67, β6-Y73, β6-F75, β6-V77
	2NZI	22, 38, 62	β2-V19, β2-L21, β2-W23, β3-Y37, β3-V39, β5-V62, L5-L65, β6-Y71, β6-F73, β6-V75
	3LPW	21, 34, 36, 54, 56, 71	β2-V19, β2-L21, β2-W23, β3-Y37, β3-V39, β5-V63, L5-L66, β6-Y72, β6-F74, β6-V76
	3LPW	11, 13, 36, 51, 90	β2-V18, β2-L20, β2-W22, β3-Y36, β3-V38, β5-I61, L5-L64, β6-Y70, β6-F72, β6-V74
	4O00	37, 52, 74	β2-V19, β2-L21, β2-W23, β3-Y37, β3-I39, β5-V64, L5-L67, β6-Y73, β6-Y75, β6-V77
Immunoglobulin (Ig)	1G1C	26, 53, 61, 63, 76	β1-I7, β2-F24, β2-V28, β3-W38, L3-I45, β4-W53, β5-L63, β5-I65, β6-I76, β7-L93, β7-V95
	1NCT	9, 17, 27, 39, 66, 68, 77, 95	β1-I9, β2-F26, β2-T30, β3-W40, L3-L47, β4-V55, β5-F64, β5-I66, β6-Y77, β7-L94, β7-I96
	1TIT	16, 35, 51, 61	β1-V4, β2-F21, β2-L25, β3-W34, L3-L41, β4-I49, β5-L58, β5-L60, β6-V71, β7-L84, β7-V86
	2A38	17, 29, 39, 46, 66	β1-F8, β2-F25, β2-L29, β3-W39, L3-I46, β4-I56, β5-L65, β5-I67, β6-Y78, β7-L95, β7-V97
	2A38	25, 60, 75	β1-F5, β2-L22, β2-V26, β3-F36, L3-I43, β4-I51, β5-L60, β5-I62, β6-Y73, β7-L90, β7-V92
	2BK8	11, 23, 34, 36, 59	β1-S5, β2-Y21, β2-I25, β3-W36, L3-L43, β4-I51, β5-L60, β5-V62, β6-Y73, β7-L90, β7-V92
	2NZI	24, 45, 65	β1-F6, β2-L23, β2-V27, β3-W37, L3-I44, β4-I53, β5-L63, L5-I65, β6-Y76, β7-L93, β7-V95
	2NZI	23, 38, 56, 79	β1-I4, β2-I26, β2-F30, β3-W40, L3-I47, β4-V55, β5-L64, β5-F66, β6-Y78, β7-L95, β7-V97
	2Y9R	25, 37, 65, 92	β1-I7, β2-V24, β2-F28, β3-W38, L3-I45, β4-I55, β5-L64, β5-I66, β6-Y77, β7-I94, β7-I96
	3B43	24, 26, 62, 83	β1-F7, β2-L24, β2-V28, β3-W38, L3-L45, β4-M53, β5-L62, β5-I64, β6-Y75, β7-L92, β7-I94
	3B43	24, 34, 55, 57, 70	β1-F5, β2-F22, β2-I26, β3-W36, L3-L43, β4-T51, β5-L60, β5-I62, β6-Y73, β7-L90, β7-L92
	3B43	27, 61, 75	β1-F5, β2-F22, β2-V26, β3-W36, L3-I43, β4-M51, β5-L60, β5-V62, β6-Y73, β7-L90, β7-V92
	3B43	27, 34, 59	β1-F5, β2-Y23, β2-I27, β3-W37, L3-I44, β4-M52, β5-L61, β5-M63, β6-Y74, β7-L91, β7-V93
	3B43	22, 24, 59, 84	β1-F3, β2-L20, β2-L24, β3-W34, L3-L41, β4-I49, β5-I58, β5-I60, β6-Y71, β7-I88, β7-L90
	3B43	24, 31, 51, 73, 86	β1-F4, β2-L21, β2-I25, β3-W35, L3-I42, β4-I51, β5-L60, β5-F62, β6-Y73, β7-L90, β7-V92

3LCY	28, 30, 65, 79	β1-I9, β2-L26, β2-I30, β3-W40, L3-L47, β4-M55, β5-L64, β5-V66, β6-Y77, β7-L94, β7-L96
3LCY	27, 51, 60, 84, 86	β1-F4, β2-L24, β2-Y28, β3-W38, L3-L45, β4-I53, β5-L62, β5-M64, β6-Y76, β7-V93, β7-I95
3PUC	10, 23, 39, 66, 78	β1-I9, β2-F26, β2-A30, β3-W40, L3-I47, β4-L55, β5-L64, β5-I66, β6-Y77, β7-L94, β7-V96
3QP3	28, 42, 64, 66, 80	β1-I11, β2-F28, β2-V32, β3-W42, L3-L49, β4-Y57, β5-L66, β5-I68, β6-Y79, β7-L96, β7-V98
5JDD	57, 72	β1-I2, β2-F19, β2-V23, β3-W32, L3-I39, β4-I47, β5-L56, β5-I58, β6-Y69, β7-L82, β7-V84
5JDD	3, 20, 22, 60, 73	β1-I3, β2-F20, β2-V24, β3-W33, L3-I40, β4-L48, β5-L57, β5-L59, β6-Y70, β7-L83, β7-V85
5JDD	5, 23, 25, 26, 28, 66	β1-I2, β2-F19, β2-V23, β3-W32, L3-I39, β4-M47, β5-L56, β5-I58, β6-Y69, β7-L82, β7-V84
5JOE	23, 25, 53, 65, 76	β1-I4, β2-F21, β2-I25, β3-W36, L3-L43, β4-I51, β5-L60, β5-V62, β6-Y73, β7-L86, β7-V88

Table G3: Position of conserved hydrophobic residues near the peaks of the F-value plots for proteins treated.

Domain	PDB ID	Peaks of a F-value	Position of conserved hydrophobic residues near a peak in a F-value plot
Fibronectin type 3 (FN3)	1TEN	20, 32, 58, 70	β1L, β2N, β2C, β3, β5, β5L, β6N, β6M, β6C
	1BPV	19, 34,73, 75, 77, 96	β2N, β2C, β2L, β3N, β3C, β6N, β6M, β6C
	2NZI	22, 38, 62	β2N, β2M, β2C, β3N, β3C, β5N, β5L
	3LPW	21, 34, 36, 54, 56, 71	β2N, β2M, β2C, β3N, β3C, β5L, β6N, β6M, β6C
	3LPW	11, 13, 36, 51, 90	β2C, β3N, β3C
	4O00	37, 52, 74	β2N, β2M, β2C, β3N, β3C, β6N, β6M, β6C
Immunoglobulin (Ig)	1G1C	26, 53, 61, 63, 76	β2N, β2C, β4, β5N, β5C, β6
	1NCT	9, 17, 27, 39, 66, 68, 77, 95	β1, β2N, β2C, β3, β5N, β5C, β6, β7N, β7C
	1TIT	16, 35, 51, 61	β2, β3, β4, β5N, β5C,
	2A38	17, 29, 39, 46, 66	β2N, β2C, β3N, β3L, β4, β5N, β5C,
	2A38	25, 60, 75	β2N, β2C, β5N, β5C, β6
	2BK8	11, 23, 34, 36, 59	β2N, β2C, β3, β5N, β5C
	2NZI	24, 45, 65	β2N, β2C, β3, β3L, β5, β5L
	2NZI	23, 38, 56, 79	β2, β3, β4, β6
	2Y9R	25, 37, 65, 92	β2N, β2C, β3, β5N, β5C, β7N, β7C
	3B43	24, 26, 62, 83	β2N, β2C, β5N, β5C
	3B43	24, 34, 55, 57, 70	β2N, β2C, β3, β4, β5N, β5C, β6
	3B43	27, 61, 75	β2N, β2C, β5N, β5C, β6
	3B43	27, 34, 59	β2N, β2C, β3, β5N, β5C
	3B43	22, 24, 59, 84	β2N, β2C, β5N, β5C, β7
	3B43	24, 31, 51, 73, 86	β2N, β2C, β3, β4, β6, β7

3LCY	28, 30, 65, 79	β2N, β2C, β5N, β5C, β6
3LCY	27, 51, 60, 84, 86	β2N, β2C, β4, β5N, β5C,
3PUC	10, 23, 39, 66, 78	β1, β2, β3, β5N, β5C, β6
3QP3	28, 42, 64, 66, 80	β2N, β2C, β3, β5N, β5C, β6
5JDD	57, 72	β5N, β5C, β6
5JDD	3, 20, 22, 60, 73	β1, β2N, β2C, β5N, β5C, β6
5JDD	5, 23, 25, 26, 28, 66	β1, β2N, β2C, β3, β6
5JOE	23, 25, 53, 65, 76	β2N, β2C, β4, β5N, β5C, β6

Table G4: Results of the ADMs for the proteins treated.

Superfamily	PDB ID	Compact region (□-value)
Cytokine	2K8R	6-49 (0.350), 57-67 (0.214), 76-91 (0.226), 100-128 (0.204)
	1Q1U	5-97 (0.298), 105-117 (0.122)
	2FDB_M	32-44 (0.212), 53-121 (0.235)
	1QQK	1-48 (0.380), 80-102 (0.194)
	1J0S	9-52 (0.245), 60-81 (0.225), 99-156 (0.305)
	6I1B	3-73 (0.262), 99-151 (0.321)
	1MD6	17-32 (0.193), 35-44 (0.221), 56-69 (0.225), 82-153 (0.291)
	2KKI	1-37 (0.320), 60-73 (0.223), 82-94 (0.190), 101-147 (0.250)
	2WRY	28-49 (0.190), 60-75 (0.220), 89-154 (0.295)
	2P39	3-61 (0.379), 65-80 (0.195), 85-108 (0.177), 116-130 (0.069)
	2P23	3-66 (0.374), 81-95 (0.174), 102-115 (0.147)
Ricin B-like lectins	2RST	30-63 (0.191), 83-131 (0.379)
	1SR4_A	21-50 (0.251), 78-94 (0.210), 112-166 (0.286)
	1SR4_C	16-47 (0.168), 61-153 (0.404)
	1KNM	30-54 (0.167), 70-118 (0.270)
	1DQG	1-71 (0.356), 87-129 (0.265)
STI-like	3BX1_D	6-47 (0.217), 60-68 (0.157), 73-110 (0.181), 141-175 (0.266)
	1TIE	9-32 (0.259), 39-50 (0.196), 56-67 (0.199), 88-97 (0.224), 135-162 (0.305)
	1R8N	3-49 (0.309), 73-84 (0.191), 95-108 (0.196), 119-175 (0.240)
	1WBA	1-58 (0.314), 71-82 (0.204), 90-119 (0.196), 131-169 (0.248)
	2GZB	19-37 (0.160), 45-60 (0.137), 64-80 (0.199), 124-160 (0.245)
	3ZC8	3-48 (0.337), 68-93 (0.188), 137-181 (0.317)
	3TC2	21-37 (0.199), 47-68 (0.158), 83-168 (0.225)
	1A8D	18-67 (0.183), 98-117 (0.158), 153-204 (0.255)
	1EPW	3-67 (0.236), 80-89 (0.167), 113-130 (0.166), 155-210 (0.277)
	3BTA	8-25 (0.088), 30-70 (0.179), 81-94 (0.178), 113-131 (0.159), 155-199 (0.197)
Actin-crosslinking proteins	1HCD	5-36 (0.146), 43-63 (0.210), 73-93 (0.135)
MIR domain	1T9F	4-41 (0.303), 47-69 (0.192), 115-159 (0.191)
	3HSM	10-47 (0.292), 68-76 (0.221), 83-111 (0.188), 126-137 (0.210), 141-159 (0.080)
DNA-binding protein LAG-1 (CSL)	1TTU	5-37 (0.104), 116-159 (0.277)

Table G5: Residues on the peaks of the F-value plots for proteins treated.

Superfamily	PDB ID	Peaks of a F-value	Conserved hydrophobic residues near a peak in a F-value plot
Cytokine	2K8R	24, 48, 56, 74, 100	β5N, β5C, β6, β8, β10
	1Q1U	26, 46, 71, 90, 106	β4C, β5N, β5C, β7, β9, β10
	2FDB_M	37, 57, 74, 76, 98, 117	β2, β4N, β4C, β6, β8, β10
	1QQK	44, 75, 82, 86, 88	β5N, β5C, β8, β9
	1J0S	21, 48, 65, 100, 120, 122	β2, β4N, β4C, β5N, β5C, β8, β9, β10
	6I1B	44, 46, 68, 100, 121	β4N, β4C, β6, β8, β10
	1MD6	26, 65, 84, 94, 115, 117	β6, β9, β10
	2KKI	33, 63, 82, 101, 103, 122	β5N, β5C, β7, β8, β10
	2WRY	21, 29, 71, 73, 101, 124, 126	β2, β6, β8, β10
	2P39	15, 56, 94, 96, 98, 108	β5N, β5C, β9
	2P23	31, 62, 81, 92, 103	β6, β8, β9
Ricin B-like lectins	2RST	31, 63, 86	β6, β8
	1SR4_A	42, 89, 113	β7, β9, β10
	1SR4_C	37, 68, 70, 72, 101, 131, 133	β3, β4C, β5, β8
	1KNM	58, 75, 92, 94	β6, β7, β9
	1DQG	22, 47, 49, 80, 90, 112, 114	β2C, β4N, β4C, β8, β9, β12
STI-like	3BX1_D	41, 43, 84, 86, 90, 143	β5N, β5C, β9
	1TIE	28, 39, 41, 66, 89, 126, 135	β2, β6, β9, β10
	1R8N	26, 44, 56, 96, 133, 168	β1, β4N, β6, β9, β12
	1WBA	26, 28, 30, 32, 46, 48, 50, 91, 130	β2, β6, β9
	2GZB	23, 31, 65, 67, 123, 125, 130, 134	β1, β2, β4C, β9, β10
	3ZC8	32, 42, 44, 69, 89, 91, 103, 105, 122, 139, 144	β2, β6, β7, β8, β9
	3TC2	30, 70, 95	β4N, β4C,
	1A8D	17, 40, 58, 63, 70, 85, 92, 114, 153, 155, 157, 172, 178, 189, 191	β1, β4N, β4C, β5N, β5C, β7, β10, β12
	1EPW	21, 23, 25, 27, 33, 62, 90, 97, 115, 119, 121, 128, 138, 159, 176, 178, 199	β1, β2, β4N, β4C, β5N, β5C, β6, β7, β8, β9, β10, β12
	3BTA	57, 62, 70, 89, 105, 107, 158, 188	β4N, β4C, β5N, β5C, β6
Actin-crosslinking proteins	1HCD	50, 52, 84	β5C, β6, β9, β10, β12
MIR domain	1T9F	52, 81, 120, 134, 136	β4N, β4C, β6, β8, β9, β10
	3HSM	39, 41, 69, 74, 82, 104, 133	β4N, β5N, β5C, β6, β9
DNA-binding protein LAG-1 (CSL)	1TTU	40, 42, 44, 73, 75, 83, 127, 129	β4N, β4C, β5N, β5C, β10