Analysis of the Local Sequences of Folding Sites in β Sandwich Proteins with Inter-Residue Average Distance Statistics

Yuko Ishizuka[1] and Takeshi Kikuchi*[,2]

[1]*Department of Chemistry and Bioscience, College of Industrial Technology, Kurashiki University of Science and the Arts, Kurashiki, Okayama, Japan*

[2]*Department of Bioinformatics, College of Life Sciences, Ritsumeikan University, Kusatsu, Shiga, Japan*

Abstract: The sequences of azurin and titin, β sandwich proteins, are analyzed based on inter-residue average distance statistics. A kind of predicted contact map based on inter-residue average distance statistics (Average Distance Map, ADM) is used to pinpoint regions of possible compact regions for two proteins. We compare predicted compact regions with the positions of the residues with experimental high ϕ values for these proteins reported in the literature. The results reveal that the regions predicted by ADMs correspond to the positions of residues with the high ϕ value. Furthermore, we perform random sampling of 3D conformations using these protein sequences with a potential derived from inter-residue average distance statistics. It is demonstrated that the residues with highest contact frequency during the simulations qualitatively correspond to the residues with the highest ϕ values for these proteins. Importantly, analysis with inter-residue average distance statistics predicts the properties of folding processes of the β sandwich proteins starting from only sequence information.

Keywords: β-sandwich protein, folding, inter-residue average distance statistics, ϕ values, azurin, titin.

INTRODUCTION

One of the ultimate goals of molecular biophysics or bioinformatics is to elucidate how the principle of protein folding is encoded in amino acid sequences. However, it is quite difficult to understand relationships between sequences and 3D structures of proteins, partly because structures of proteins are conserved better than their sequences as observed in so-called superfolds [1]. Among the various protein folds, the 3D structures of proteins with the β sandwich scaffold are rather complicated and the elucidation of the folding mechanisms is challenging [2]. Energetics of the β sheet structures has been extensively studied by several authors [3-10]. Recently, the regularity of β sandwich structures has been clarified [11-13] and the relationship of the regularity in β sandwich proteins and their folding mechanisms, has been recognized mainly through experimental ϕ value analyses. Portions with the regularity observed in β sandwich structures are related to the segments containing the residues with high ϕ values, i.e., the segments involved in the folding mechanism. Thus, β sandwich proteins are quite interesting for clarifying the relationship between 3D structures and sequences, and that is the reason why we treat β sandwich proteins in this work.

We focus on how we can extract information about folding mechanisms from the sequences of β sandwich proteins. In other words, the sequence specificity for the folding mechanisms of β sandwich proteins is considered. In particular, azurin and titin are taken as β sandwich proteins because detailed investigation of folding mechanisms, especially ϕ value analyses of azurin [14-19] and titin [20-24] have been performed.

There are also several theoretical and simulation studies on folding of azurin and titin. Zong *et al.* [25] predicted ϕ values of some residues in apo-azurin theoretically using the variational free energy functional; compared with the experimental data, the theoretical ϕ values coincide well with those from experiment (correlation coefficient of 0.90). However, their technique requires the native structure of a protein and does not clarify the characteristics of a local sequence related to the folding properties of a protein.

In the present work, we try to analyze the sequences of azurin and titin with the average distance map method and with simulations employing an inter-residue potential derived from inter-residue average distance statistics. It should be emphasized that the present work focuses on how folding information can be decoded in the sequences of azurin and titin. We demonstrate that average distance maps provide information on folding properties of proteins without any other information beside amino acid sequences. Actually, it was observed that differences in folding processes of homologous proteins within a family reflect on their average distance maps for the lipid binding protein family [26], the globin family [27] and the c-type lysozyme family [28].

On the other hand, Baker and coworkers [29, 30] and other authors [31, 32] found a strong correlation between values of contact order and folding rates of proteins exhibiting two-state folding. These works tried to predict protein folding rates from only sequence information, but all techniques predict just folding rates but not details of folding processes from amino acid sequences. Hence, the methods

*Address correspondence to this author at the Department of Bioinformatics, College of Life Sciences, Ritsumeikan University, 1-1-1 Nojihigashi, Kusatsu, Shiga 525-8577 Japan; Tel: +81-77-561-5909; Fax: +81-77-561-5203; E-mail: tkikuchi@sk.ritsumei.ac.jp

1875-0362/11

proposed so far are those that possess the quantitative predictability of folding rates and mechanisms but need native structures, or those that require just amino acid sequences but can predict only folding rates. In this paper, we attempt to extract more detailed information of folding mechanisms from only amino acid sequences of proteins.

MATERIALS AND METHODS

Proteins Used in This Work

Proteins treated in this paper are azurin (PDB code:1azu) and titin (PDB code:1tit), which are β sandwich proteins with the Greek key motif. The sequences with the indication

A Azurin

```
          *          *          *# #       *
--CSVDIQGNDQMQFNTNAITVDKSCKQFTVNLSHPGNLP
   BBBBB             BBB      BBBBBBB

     β1               β2         β3

         #
        # *          *          *          *
KNVMGHNWVLSTAADMQGVVTDGMASGLDKDYLKPDDSRV
        BBBB     AAAAAAAAAA AAAA
        β4          α1        α2

                                  #
        *     # #   *        #   *          *
IAHTKLIGSGEKDSVTFDVSKLKEGEQYMFFCTFPGHSAL
 BB           BBBBBB           BBBB
 β5             β6               β7


MKGTLTLK
 BBBBBB
   β8
```

B Titin

```
           *           *# #       *    # #    *
LIEVEKPLYGVEVFVGETAHFEIELSEPDVHGQWKLKGQP
   BBBB    BBBBB   BBBBBBBB        BBBBB

        β1              β2              β3


                     #
          *         # *           *# #       *
LTASPDCEIIEDGKKHILILHNCQLGMTGEVSFQAANAKS
       BBBBBB   BBBBBBB         BBBBBBB    BB
         β4       β5                β6


AANLKVKEL
BBBBBBBB
    β7
```

Fig. (1). Sequences of azurin (**A**) and titin (**B**) showing the positions of β strands and α helices. A symbol 'B' denotes an amino acid in a β strand, and a symbol 'A' donotes an amino acid in an α helix structure. A residue with # is involved in hydrophobic packing.

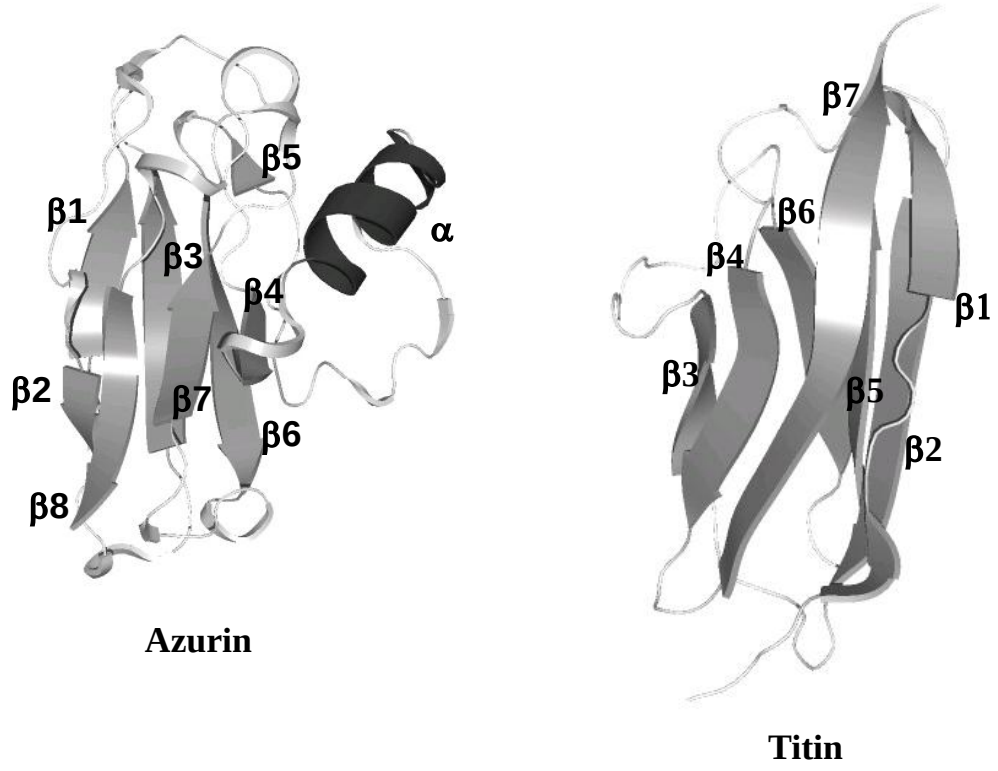

Fig. (2). 3D structures of azurin (**A**) and titin (**B**). The secondary structures are labeled by α and β.

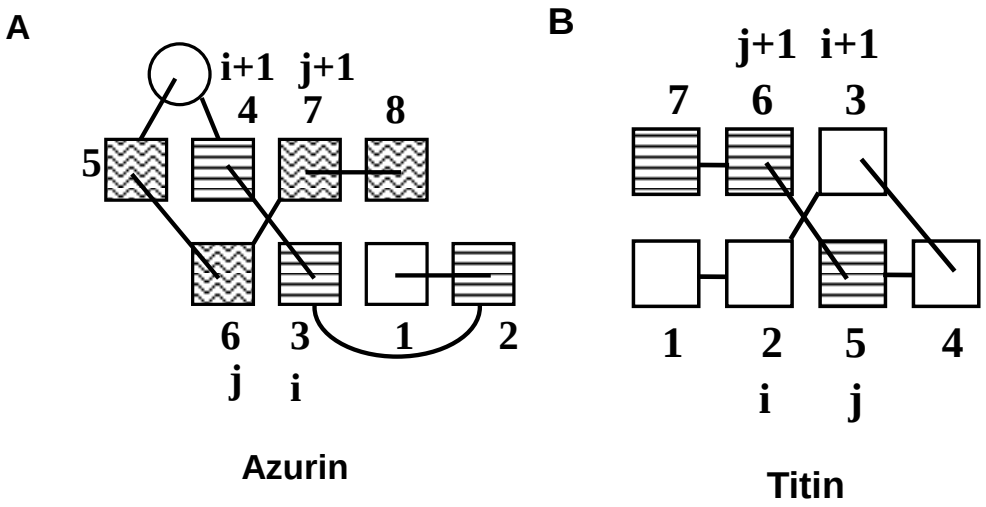

Fig. (3). Topology of azurin (**A**) and titin (**B**). A square and a circle denote a β strand and an α helix, respectively. The β strands with the same pattern denote that those strands are in a predicted compact region by ADM. i and j label the sequential numbers of β strands.

of the secondary structures and 3D structures of these proteins are presented in Figs. (**1**) and (**2**), respectively. The illustrations showing the topologies of these proteins are in Fig. (**3**).

Key Strands in the β Sandwich Fold

Kister *et al.* [11] observed the regularity in the 3D structures of β sandwich proteins. β sandwich proteins always contain two interlocked pairs of neighboring β strands; which are called key strands. We show the key strands in azurin and titin in Figs. (**4A**, **B**) according to the definition by Kister *et al.* [11]. The packing of hydrophobic residues in these strands is shown in Fig. (**5**).

It is observed that these hydrophobic residues in the key strands are well-conserved among β sandwich proteins [11]. The assignments of the positions of the secondary structures in the Protein Data Bank (www.rcsb.org/pdb/home/home.do) is used in this work. We define hydrophobic packing of residues when at least one of the heavy atoms in one residue is close to one of the heavy atoms in another residue within 5Å, and weak hydrophobic packing when these two atoms are within 8 Å.

ϕ Values of the β Sandwich Proteins

The ϕ values for azurin reported by Chen *et al.* [19], and those for titin by Fowler and Clarke [24] (under the condition of 1M GdmCl (guanidinium chloride)) are used in the present study.

Procedure of the Average Distance Map Method

The method used mainly in the present work is described in Kikuchi *et al.* [33] and let us give the brief procedure of the method. We refer this method as the Average Distance Map (ADM) method. The detailed procedure of the ADM method is described in Supplementary Material.

(1) Definition of Ranges by Separation Between Residues Along the Sequence of a Protein and the Calculations of Average Distances within Each Range

A range is defined as the length between two residues along a given sequence, and the average distances between Cα atoms of residues were calculated in each range using proteins with known structures. The details of the definition of range are described in Supplementary Material.

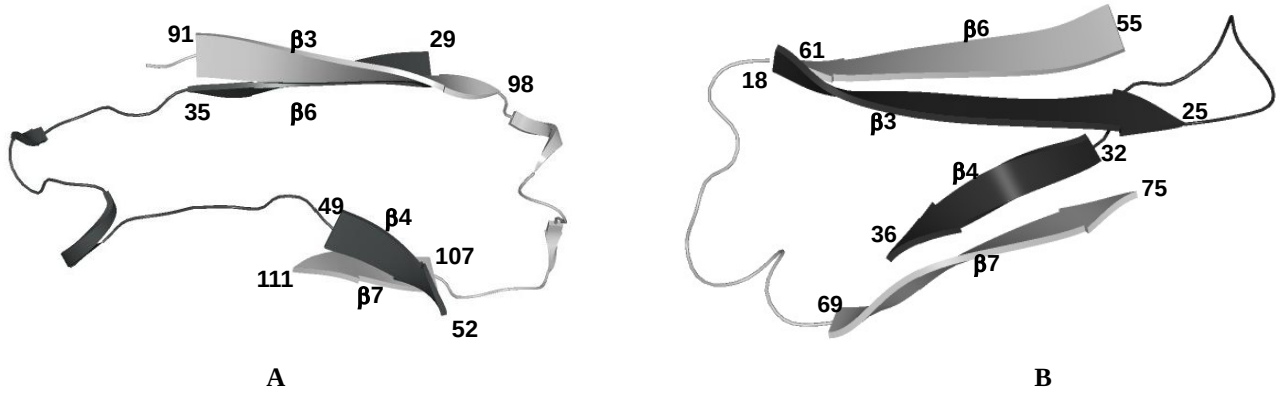

Fig. (4). Key strands in azurin (**A**) and titin (**B**). The N-terminal and C-terminal residues are numbered. A β strand is labeled by β1, β2 and so forth.

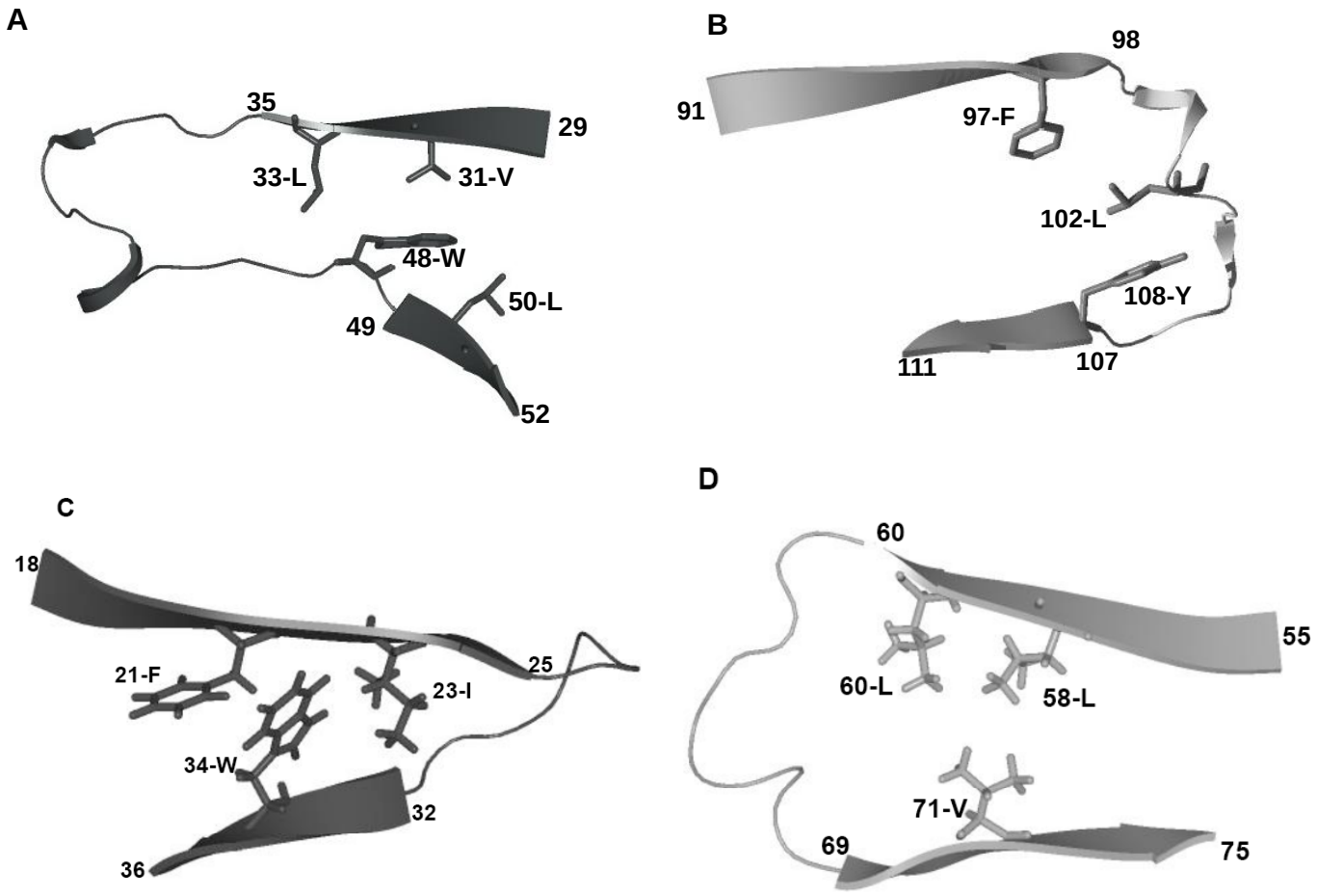

Fig. (5). The hydrophobic residues forming hydrophobic packing in these strands for azurin (**A**, **B**) and titin (**C**, **D**).

A contact map is constructed by making a plot (i.e., defining a contact) on a map for a protein with unknown 3D structure, if the average distance of a pair of residues in a range defined above is less than a cutoff value determined by the method described in the following way.

(2) Definition of Cutoff Distances for Construction of ADM

A cutoff distance value for the construction of ADM of a given sequence is defined in each range so that the contact density of the whole real distance map (RDM) of the protein is reproduced[25]. The RDM for a contact map is constructed based on the actual 3D structure. In the present study, a contact on the RDM is defined as an inter-residue Cα atomic distance less than 15 Å.

(3) Definition of Compact Regions

A compact region can be defined as a region of high density contacts along the diagonal of a map. The strength of the compactness of a compact region is measured by the η value. The details are described in Supplementary Material.

The region with the highest η value can be defined as the maximum of a compact region. Other regions with high η values can also be regarded as smaller compact regions [33].

Inter-Residue Potential Based on Average Distance Statistics

We define inter-residue effective potential based on average inter-residue distance statistics used to construct an ADM [34, 35].

(1) Model of a Protein

A protein is modeled such that each amino acid is represented by its Cα atom, whereas the detailed structure of each residue is ignored. Each peptide bond is represented as a virtual bond with the length of 3.8 Å. That is, a protein is represented as a beads-on-a-string model. In this model, variable parameters are the bond angle θ and the dihedral angle υ as indicated in Fig. (**6**).

(2) Effective Inter-Residue Potential

An effective inter-residue potential is defined in order to reproduce the average distances and their variances in the average distance statistics between residues defined above assuming a Gaussian distribution. When we define $\bar{r}_{AB}^{M}$ as the average distance between Cα atoms of residue types (kinds of amino acids) A and B in a range M and σ_{AB}^{M} as the standard deviation, the effective inter-residue potential between residue i and j, $\varepsilon_{ij}^{M}(r_{ij})$ can be expressed by Eq. (2).

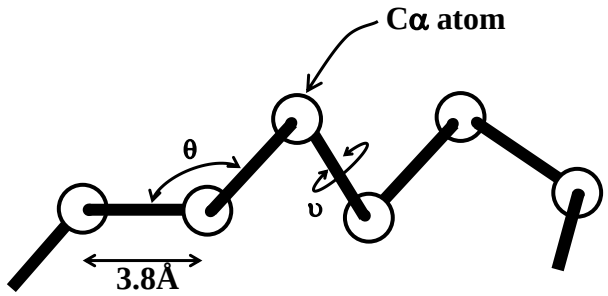

Fig. (6). Model of a protein as used in our analysis. Each amino acid is represented by its Cα atom and the detailed structure of each amino acid is ignored. A peptide bond is represented as a virtual bond with the length of 3.8 Å. The bond angle θ and dihedral angle υ are variable in the present model.

$$\varepsilon_{ij}^{M}(r_{ij}) = kT\frac{(r_{ij}-\bar{r}_{AB}^{M})^2}{2(\sigma_{AB}^{M})^2} - kT\ln Z + \frac{kT}{2}\ln 2\pi(\sigma_{AB}^{M})^2 \qquad (2)$$

where r_{ij} is the distance between Cα atoms of the residues i and j, and Z means the partition function. A and B are the residue types of i and j. k and T are the Boltzmann constant and temperature respectively. The constant terms in Eq. (2) can be regarded as the zero point. We put $\varepsilon_{ij}^{M}(r_{ij}) = \varepsilon_{HC}$ when $\bar{r}_{ij}^{M} \le r_{cut}$. We set r_{cut} = 1.9 Å and ε_{HC} = 50 kcal/mol. These values were obtained empirically [34, 36].

(3) Simulation

Sampling of random structures with the above potential and the standard Metropolis Monte Carlo method was performed. Only the dihedral and bond angles are variable parameters. In a Monte Carlo simulation, each dihedral angle, υ, and bond angle, θ, of a residue was changed within $-\gamma\pi \leqq \upsilon, \theta \leqq \gamma\pi$ followed by the Metropolis judgment [37]. γ is empirically fixed as 0.6, and the temperature parameter T was adjusted such that the acceptance ratio in the MC routine is around 0.5. In the present simulations, we set T = 300 and 210 for azurin and titin, respectively. This procedure was iterated for all residues. For a whole simulation, this routine is iterated 60000 times.

Definition of the Contact Frequency During a Simulation with the Present Effective Potential

During the simulations, we calculate the contact frequency, g(i, j), between a residue pair of the sequence, in other words, contact probability [36]. The contact is defined as a distance between Cα atoms of the residues i and j, r_{ij}, shorter than a threshold, r_c, i.e., $r_{ij} \le r_c$. The threshold is taken as 10 Å in this study. When G(i, j) values are calculated as the total number of contacts that a residue pair i and j forms during a simulation, g(i,j) is defined as G(i,j)/Nmc where Nmc is the total number of Monte Carlo iterations.

Then, a measure of high contact frequency is defined by Q(μ, ν) in the following equation (|μ−ν|=m).

$$D(m) = \sqrt{\frac{\sum_{|i-j|=m}(g(m)-g(i,j))^2}{N_m}} \text{ where,}$$

$$Q(\mu,\nu) = \frac{(g(\mu,\nu)-g(m))}{D(m)}$$

and

$$g(m) = \frac{\sum_{|i-j|=m} g(i,j)}{N_m}$$

where, N_m is the total number of residues pairs with separation m = |i-j|.

$p(\mu) = \sum_{\nu} Q(\mu,\nu)$ expresses a residue with high contact frequency. This value is not exactly same but can be compared to an experimentally observed ϕ value of the residue μ. We run 10 simulations for each protein, and took average values of the 10 simulations.

RESULTS

ADM Analyses for Azurin and Titin

The ADMs for azurin and titin are illustrated in Figs. (**7A** and **B**) respectively. In the same figures, the ADM for azurin predicts regions 19-61 with η = 0.202 and 80-127 with η = 0.355 and that for titin region 57-86 with η = 0.240 as compact.

(The detailed analyses of the ADM of azurin suggest that the two compact regions in this protein merge into one domain and the predicted compact region 57-86 in titin can be expanded to 8-86 as a domain according to the manner described in Supplementary Material and Ref. [33].) It is interesting to note that the predicted regions 19-61 for azurin includes the first and second key strands (see Fig. (**4A**)). In the same way, the predicted region 80-127 in azurin includes the third and fourth key strands (Fig. (**4A**)). For titin, the predicted region 57-86 includes the third and fourth key strands (Fig. (**4B**)). In both ADMs, the residues with the φ values $\ge \phi_{av} + s$ are indicated by arrows with the residue numbers where ϕ_{av} and s denote the average and the standard deviation of φ values, respectively (see also Fig. (**7**)). (A residue with $\phi_{av} + s > \phi \ge \phi_{av} + 0.5s$ is also indicated at the broken arrows.) For azurin, the residues with high φ values, i.e., $\phi_{av} + s \ge \phi$ are 31-V, 33-L and 50-L which are contained in the region 19-61 (see the arrows in Fig. (**7A**)). Thus, location of subdomains predicted by ADMs corresponds to the regions containing residues with higher φ values. For titin, the residues with φ values $\ge \phi_{av} + s$ are 23-I, 49-I, 58-L, 60-L, and 73-F (Fig. (**7B**)). In the same way, the residues with $\phi_{av} + s > \phi \ge \phi_{av} + 0.5s$ are 71-V and 75-A (see the broken arrows in Fig. (**7B**)). Except for 23-I and 49-I, these residues are included in the predicted compact region 57-86.

p(μ) Value Analysis

Fig. (**8**) presents the comparisons of p(μ) values with φ values for azurin (Fig. (**8A**)) and for titin (Fig. (**8B**)) with bold lines and broken lines, respectively.

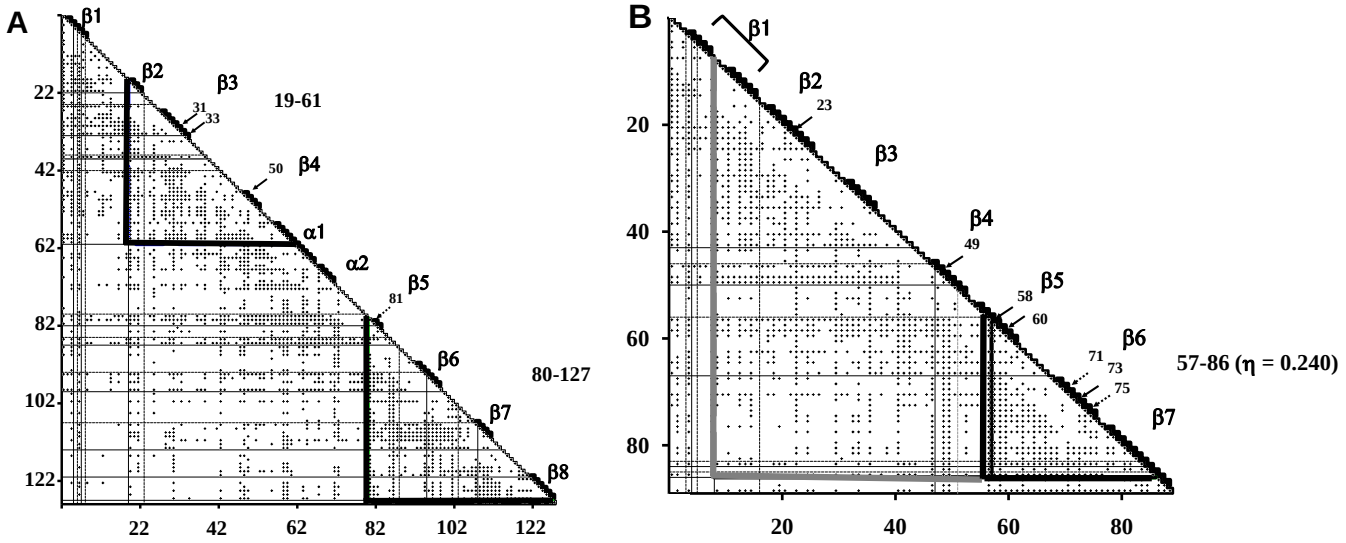

Fig. (7). ADMs for azurin (**A**) and titin (**B**). The regions enclosed by the black lines denote the compact regions predicted by ADM. An η value is also indicated in parentheses. The positions of peaks and valleys in the horizontal and vertical scanning plots are indicated by full and broken lines in a map respectively. Each α-helix or β-strand is labeled as α or β with a number denoting the order of each secondary structure. An arrow denotes a residue with a φ value $\geq \phi_{av} + s$ where ϕ_{av} and s denote the average and the standard deviation of φ values respectively. A broken arrow denotes a residue with a φ value, $\phi_{av} + s \geq \phi \geq \phi_{av} + 0.5s$.

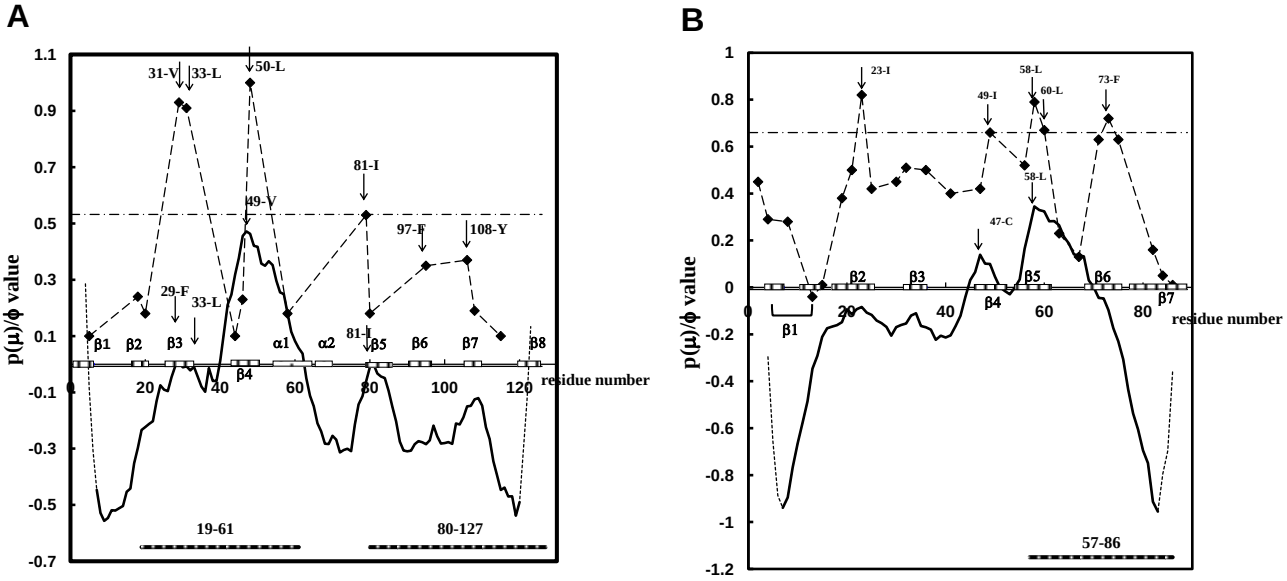

Fig. (8). Plots of p(μ) values for azurin (**A**) and titin (**B**). The profiles are shown in bold lines. The abscissa denotes the residue number. In each figure, the φ values are also plotted with the filled diamonds and broken line. The shaded and dotted bars on the abscissa denote the locations of β-strands and α-helix respectively. The residues with high φ values are marked by arrows. The chained line shows the value $\phi = \phi_{av} + 0.5s$ in each figure. The profiles of p(μ) values for the five residues at the N- and C-termini are shown as dotted lines. The full line parallel to (very close to) the abscissa in each figure shows the value of $p(\mu) = p(\mu)_{av} + 0.5s$ where $p(\mu)_{av}$ denotes the average value of $p(\mu)$. The compact regions predicted by ADMs are also shown at the bottom of figures.

As seen in Fig. (**8**), the locations of the peaks of p(μ) profiles generally locate in the regions of secondary structures. It is noted that the peaks of the φ values also appear on the secondary structures. That is, the maxima of the p(μ) plots qualitatively correspond to those of the φ plots in terms of secondary structures. (We disregard the profiles of p(μ) values at N and C terminal 5 residues because these parts might behave rather randomly and it might be difficult to extract the physical meaning from these profiles.) The most interesting results are that the highest peak of p(μ) values for

azurin are 49-V and that of ϕ values is 50-L, i.e., very close, and that for titin the highest peak of the p(μ) values is at 58-L and that one of the highest peaks in the profile of ϕ values at 58-L, i.e., exactly the same position.

The locations of the key strands are compared with the region containing the residues with the high ϕ values and that with the highest p(μ) values for azurin and titin respectively. As mentioned above, the residues with high ϕ values are 31-V, 33-L in the 3th β strand (β3) and 50-L in the β4 in azurin as shown in Fig. (**8A**). 31-V and 33-L form hydrophobic packing with 48-W between β3 and β4 as shown in Fig. (**5A**). 50-L forms weak hydrophobic packing with 31-V. 97-F and 108-Y show a moderate magnitude of ϕ values (Fig. (**8A**)), and these residues are on the key strands β6 and β7 respectively forming weak hydrophobic packing (or these two residues interact *via* 102-L. see Fig. (**5B**). On the other hand, residues with high p(μ) values with >Average + 0.5s are in the region of 40-63. (In the present case, Average + 0.5s ≈ 0. see the figure legend of Fig. (**8A**)), and this region contains β4 and α helix (α1). As seen above, the residues with the highest p(μ) is 49-V that correspond very well to 50-L with the highest ϕ value. It should be noted that the p(μ) profile in Fig. (**8A**) shows the small peaks at the 29-F and 33-L, and these residues are almost same as the high ϕ value residues in β4 and β5, i.e., 31-V and 33-L. The p(μ) profile also shows a peak at 81-I, and there is a peak in the ϕ values at this residue.

For titin, the experimentally observed residues with the high ϕ values are 23-I, 49-I, 58-L, 60-L, and 73-F as shown in Fig. (**8B**). These residues are in the key strands β6 and β7, and 58-L and 60-L also form the hydrophobic packing with 71-V in β6 and β7 as shown in Fig. (**5D**). Furthermore, a peak at 23-I of the ϕ value profile (also on the key strand β3) and 23-I is involved in the hydrophobic packing between β3 and β4 as seen in Fig. (**5C**). On the other hand, residues with p(μ) values >Average+0.5s are 54-70 (Also in the present case, Average + 0.5s ≈ 0. see the figure legend of Fig. (**8A**) as in Fig. (**8B**) corresponding to the residues with high ϕ values and the interlocked pairs in key strands. This region contains β6 and a part of β7. The p(μ) profile also shows a peak at 47-C which is also very close to 49-I at which the ϕ profile denotes a peak.

p(μ) Values Fixing the Regions Predicted by ADMs to the Native 3D Structures

In the present study, the positions of the compact regions predicted by ADMs correspond to the folding transition structure forming region for the β sandwich proteins. Therefore, we are interested in how p(μ) values of each protein vary if the 3D structures of the predicted compact regions are fixed to the native structure. We also fix the 3D structures of the regular secondary structures to the native structure.

For azurin, we fixed the 3D structure of the region 19 – 61 and the conformations of the regions of all α helices and β strands (α1, β3 and β4). (Two regions are predicted by the ADM, i.e., 19 – 61 and 80 – 127. The η value of 80 – 127 is greater than that of 19 – 61; however, the highest peak is located in 19 – 61. This is a controversial result. In the present study, we take 19 – 61 to be fixed.) Then we performed the same calculations for p(μ) values. The result is shown in Fig. (**9A**).

The region with the broken line of the p(μ) profile in the figure denotes the region with fixed 3D native structure, and we mainly focus on change in the rest of the profile. A residue number underlined in the figure denotes a high ϕ value that we wish to focus on. The peaks of the p(μ) profile in β6 and β7 become higher compared with the profile in Fig. (**8A**), and the general tendency is getting closer to the ϕ profile. β6 and β7 form key strands, and this result suggests the participation of β6 and β7 with the folding consistent with the observation of the key strands. That is, this result suggests that β3 and β4 interacts with β6 and β7during the folding as long range interactions.

For titin, we fixed the 3D structure of the region 57-89 (predicted subdomain + C terminal two residues) and the conformations of all β strands (β6, β7 and β8) to the native structures. The simulation result is shown in Fig. (**9B**). The remarkable property is that the p(μ) values in the β3 region became higher compared with the profile in Fig. (**8B**) suggesting the participation of β3 in the folding; the correspond-

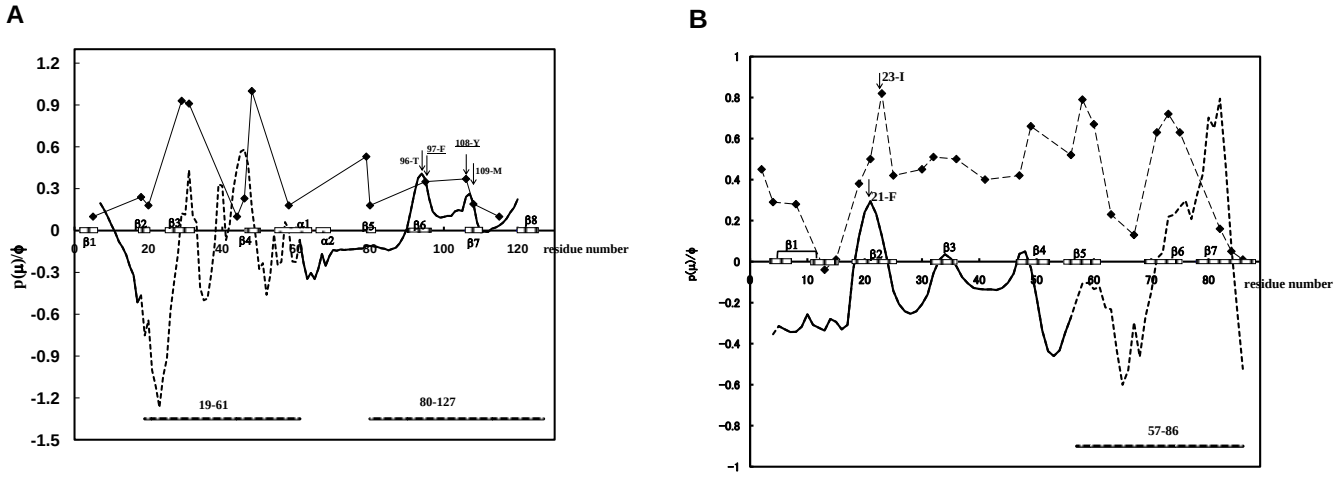

Fig. (9). Plots of p(μ) values for azurin (**A**) and titin (**B**) fixing the 3D structures of the compact regions predicted by ADMs (p(μ) profiles of these parts are drawn as broken lines in the figures) and the regular secondary structures to the native structures. An underlined residue name calls attention to the location of a peak in the ϕ profiles. ϕ profiles are indicated by filled diamonds and solid lines.

ing peak of the new p(μ) plot is 21-F which is close to 23-I which is a peak of the ϕ profile. β3 constitutes a key strand. Again, the long range interactions between β6-β8 and β3 are suggested in the folding of this protein.

For both azurin and titin, the positions of the new peaks appearing in Figs. (**9A** and **B**) are within three residues of the peaks observed in the experimental ϕ profiles.

DISCUSSION

For the β sandwich proteins treated in this study, azurin and titin, the ADMs predict the location of subdomains that include the residues with high ϕ values. These results mean that the regions in the sequences encode information on structure formation of these proteins are detected by ADMs.

Furthermore, the highest peak of the p(μ) plots corresponds well to that of the ϕ plots in the present cases. Therefore, a p(μ) plot specifies the location of the more significant areas in the region predicted by ADM. The whole profile of a p(μ) plot is not necessary to resemble to that of the corresponding ϕ plot. The most important point is the location of peaks. These results also mean that the average distance statistics contain information about initial 3D structure forming regions, i.e., folding sites. Thus, it is possible to specify which secondary structure elements are involved in the early folding event based on the knowledge of the location of secondary structures.

For example, let us suppose that we do not know the 3D structure and the secondary structures of azurin. Our ADM analysis predicts the compact regions 19-61 (η = 0.202) and 80-127 (η = 0.355) (Fig. (**7A**)). These two regions can be candidates for regions involved in structural formation. Although from the η value the region 80-127 is plausible to be compact during the early stage of folding, the p(μ) analysis predicts the region 44-59 to be a highly contacted region (Fig. (**8A**)). The knowledge of the location of the secondary structures allows us to predict that the β4 and a helix mainly pack as suggested from Fig. (**8A**). Furthermore, within the region 19-61, a peak of the p(μ) plot appears at β3. These observations lead us to predict that β3 and β4 pack together hydrophobically. These tentative predictions are consistent with the results of ϕ value and key strands analyses. A similar consideration can be applied to titin. From the sequence of titin, we predicted that the region 57 – 86 forms a compact region during a relatively early stage of folding (Fig. (**7B**)), and within this region, the segment 57 – 69 would be deeply involved in the folding (Fig. (**8B**)). If we know the location of β strands, it is also predictable that β6 and β7 pack together hydrophobically during folding (Fig. (**8B**)).

Furthermore, our method can specify the hydrophobic residues forming packing pairs in key strands. We observed that such a hydrophobic residue possessing a high ϕ value is also located within a few residues from those with a high p(μ) value as indicated in Fig. (**8**). Inversely, the hydrophobic residues in the segment with the highest p(μ) value can be considered as residues capable of forming packing pairs in a β sandwich protein.

Thus, the indices derived from average distance statistics correspond well to folding properties including folding transition states. It is remarkable that average distance statistics reflect some properties of the folding transition state of a protein, although average distance statistics probably include properties of a denatured conformation ensemble of a protein near the folding transition state. It should be emphasized again that the predictions based on ADMs and p(μ) values have been done without any 3D structure information.

As learned from Fig. (**9**), we can further predict a detailed folding mechanism if we fix the 3D structure of a predicted folding site and regular secondary regions to the native structure. That is, the main feature of the p(μ) profile becomes closer to the ϕ profile. In other words, a simulation with fixed partial structures reflects the folding process consistent with the folding transition state suggested from the ϕ value analyses. In particular, it is interesting that the peaks of the p(μ) profile within the segments corresponding to the key strands are getting higher, and thus key strands involved in folding are predictable from the present method. Thus, our method has potential to predict the location of the whole folding transition state area in a protein if we have knowledge of the partial 3D structure of a transition structure formation region.

The next interesting problem is whether the 3D structure of such a compact region in folding can be modeled. At least, for β sandwich proteins, 3D structures of transition

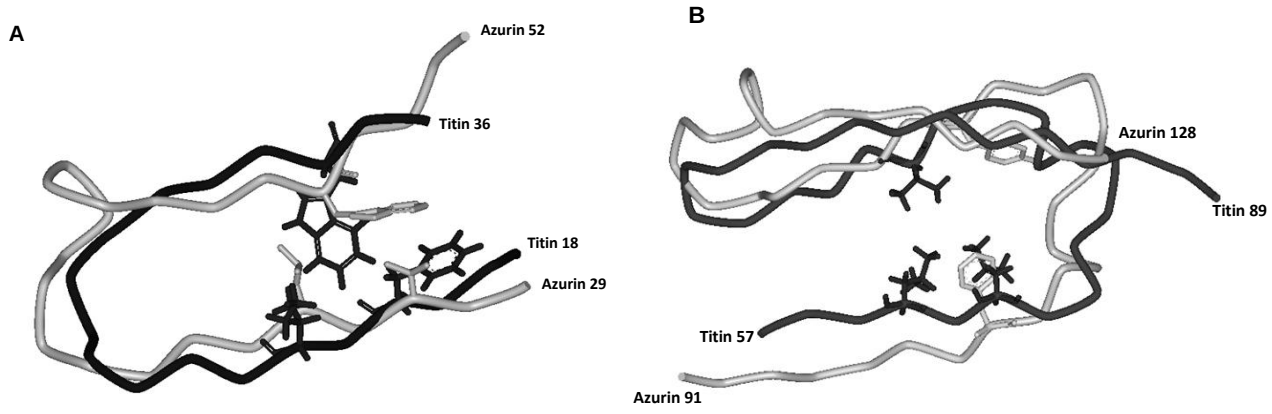

Fig. (10). A. Superposition of the 3D structures of azurin 29-52 (light gray) and titin 18-36 (dark gray). The rmsd value is 2.72 Å. **B**. Superposition of the 3D structures of azurin 91-128 (light gray) and titin 57-89 (dark gray). The rmsd value is 4.08 Å. The interlocked pairs of hydrophobic residues are also shown.

state formation areas might be similar. As an attempt, we compare the 3D structures between 29-52 corresponding to the packing pair of β3 and β4 in azurin and 18-36 corresponding to the packing pair of β3 and β4 in titin, and the region 91-128 of azurin containing the C-terminal 3 β strands within the predicted compact region 80-129 by the ADM and the region 57-89 predicted as a compact region by the ADM for titin. Illustrations of superposed structures are presented in Figs. (**10A** and **B**).

The alignments were performed so that the positions of the hydrophobic residues forming the interlocked pairs and secondary regions are superposed. The rmsd values of the Cα atoms are 2.72 Å and 4.08 Å, respectively, in the former superposition and in the latter. Thus, these rmsd values suggest that it is possible to model the 3D structure of the transition structure formation region of a β sandwich protein from an appropriate template structure.

ACKNOWLEDGEMENT

This work was supported by Grand-in-Aids for Scientific Research (C) (no. 19510202) from the Ministry of Education, Culture, Science, Sports, and Technology (MEXT) of Japan.

SUPPLEMENTARY MATERIAL

Supplementary material is available on the publishers Web site along with the published article.

REFERENCES

[1] C.A. Orengo, D. T. Jones, and J. M. Thornton, "Protein superfamilies and domain superfolds", *Nature*, vol. 372, pp. 631-634, 2002.

[2] P. Bradley, and D. Baker, "Improved beta-protein structure prediction by multilevel optimization of nonlocal strand pairings and local backbone conformation", *Proteins*, vol. 65, pp. 922-929, 2006.

[3] K. C. Chou, G. Némethy, S. Rumsey, R. W. Tuttle, and H. A. Scheraga, "Interactions between two beta-sheets: energetics of beta/beta packing in proteins", *J. Mol. Biol.*, vol. 188, pp. 641-649, 1986.

[4] K. C. Chou, G. Némethy, and H. A. Scheraga, "Role of interchain interactions in the stabilization of right-handed twist of β-sheets", *J. Mol. Biol*., vol. 168, pp. 389-407, 1983.

[5] K. C. Chou, G. Némethy, and H. A. Scheraga, "Effects of amino acid composition on the twist and the relative stability of parallel and antiparallel β-sheets", *Biochemistry*, vol. 22, pp. 6213-6221, 1983.

[6] K. C. Chou, G. Némethy, and H. A. Scheraga, "Review: energetics of interactions of regular structural elements in proteins", *Acc. Chem. Res.*, vol. 23, pp. 134-141, 1990.

[7] K. C. Chou, M. Pottle, G. Némethy, Y. Ueda, and H. A. Scheraga, "Structure of β-sheets: origin of the right-handed twist and of the increased stability of antiparallel over parallel sheets", *J. Mol. Biol.*, vol. 162, pp. 89-112, 1982.

[8] K. C. Chou, and H. A. Scheraga, "Origin of the right-handed twist of β-sheets of poly-L-valine chains", *Proc. Natl. Acad. Sci. USA*, vol. 79, pp. 7047-7051, 1982.

[9] K. C. Chou, and L. Carlacci, "Energetic approach to the folding of α/β barrels", *Proteins*, vol. 9, pp. 280-295, 1991.

[10] K. C. Chou, L. Carlacci, and G. M. Maggiora, "Conformational and geometrical properties of idealized β-barrels in proteins", *J. Mol. Biol.*, vol. 213, pp. 315-326, 1990.

[11] A. E. Kister, A. V. Finkelstein, and I. M. Gelfand, "Common features in structures and sequences of sandwich-like proteins", *Proc. Natl. Acad. Sci. USA*, vol. 99, pp. 14137-14141, 2002.

[12] A. S. Fokas, T. S. Papatheodorou, A. E. Kister, and I. M. Gelfand, "A geometric construction determines all permissible strand arrangements of sandwich proteins", *Proc. Natl. Acad. Sci. USA*, vol. 102, pp. 15851-15853, 2005.

[13] Y. -S. Chiang, T. I. Gelfand, A. E. Kister, and I. M. Gelfand, "New classification of supersecondary structures of sandwich-like proteins uncovers strict patterns of strand assemblage", *Proteins*, vol. 68, pp. 915-921, 2007.

[14] I. Pozdnyakova, and P. Wittung-Stafshede, "Approaching the speed limit for greek key β-barrel formation: Transition-state movement tunes folding rate of zinc-substituted azurin", *Biochim. Biophys. Acta*, vol. 1651, pp. 1-4, 2003.

[15] B. Rizzuti, V. Daggett, R. Guzzi, and L. Sportelli, "The early steps in the unfolding of azurin", *Biochemistry*, vol. 43, pp. 15604-15609, 2004.

[16] C. J. Wilson, and P. Wittung-Stafshede, "Role of structural determinants in folding of the sandwich-like protein *Pseudomonas aeruginosa* azurin", *Proc. Natl. Acad. Sci. USA*, vol. 102, pp. 3984-3987, 2005.

[17] C. J. Wilson, and P. Wittung-Stafshede, "Snapshots of a dynamic folding nucleus in zinc-substituted *Pseudomonas aeruginosa* Azurin", *Biochemistry*, vol. 44, pp. 10054-10062, 2005.

[18] C. J. Wilson, D. Apiyo, and P. Wittung-Stafshedeand, "Solvation of the folding-transition state in *Pseudomonas aeruginosa* azurin is modulated by metal", *Protein Sci.*, vol. 15, pp. 843-852, 2006.

[19] M. Chen, C. J. Wilson, Y. Wu, P. Wittung-Stafshede, and J. Ma, "Correlation between protein stability cores and protein folding kinetics: a case study on *Pseudomonas aeruginosa* Apo-Azurin" *Structure*, vol. 14: pp. 1401-1410, 2006.

[20] J. Clarke, E. Cota, S. B. Fowler, and S. J. Hamill, "Folding studies of immunoglobulin-like β-sandwich proteins suggest that they share a common folding pathway", *Structure*, vol. 7, pp. 1145-1153, 1999.

[21] E. Cota, S. J. Hamill, S. B. Fowler, and J. Clarke, Two proteins with the same structure respond very differently to mutation: the role of plasticity in protein stability. *J. Mol. Biol.*, vol. 302, pp. 713-725, 2000.

[22] R. B. Best, S. B. Fowler, J. L. T. Herrera, A. Steward, E. Paci, and J. Clarke, "Structure of transition state revealed by combining atomic force microscopy, protein engineering and molecular dynamics simulations", *J. Mol. Biol*., vol. 330, pp. 867-877, 2003.

[23] C. F. Wright, K. Lindorff-Larsen, L. G. Randles, and J. Clarke, "Parallel protein-unfolding pathways revealed and mapped", *Nat. Struct. Biol.*, vol. 10, pp. 658-662, 2003.

[24] S. B. Fowler, and J. Clarke, "Mapping the Folding Pathway of an Immunoglobulin Domain: Structural Detail from Phi Value Analysis and Movement of the Transition State", *Structure*, vol. 9, pp. 347-450, 2001.

[25] C. Zong, C. J. Eilson, T. Shen, P. G. Wolynes, P. Wittung-Stafshede, "ϕ-value analysis of apo-azurin folding: comparison between experiment and theory", *Biochemistry*, vol. 4, pp. 6458 – 6466, 2006.

[26] T. Ichimaru, and T. Kikuchi, "Analysis of the differences in the folding kinetics of structurally homologous proteins based on predictions of the gross features of residue contacts", *Proteins*, vol. 51, pp. 515-530, 2003.

[27] S. Nakajima, E. Álvarez-Salgado, T. Kikuchi, and R. Arredondo-Peter, "Predicting Folding pathways and kinetics among plant hemoglobins by using an average distance map method", *Proteins*, vol. 61, pp. 500-506, 2005.

[28] S. Nakajima, and T. Kikuchi, "Analysis of the differences in the folding mechanisms of c-type lysozymes based on contact maps constructed with interresidue average distances", *J. Mol. Model.* vol. 13, pp. 587-594, 2007.

[29] K. W. Plaxco, K. T. Simons, I. Ruczinski, and D. Baker, "Topology, stability, sequence, and length: defining the determinants of two-state protein folding kinetics", *Biochemistry*, vol. 39, pp. 11177 – 11183, 2000.

[30] D. Baker, "A surprising simplicity to protein folding", *Nature*, vol. 405, pp. 39-42, 2000.

[31] M. M. Gromiha, and S. Selvaraj, "Comparison between long-range interactions and contact order in determining the folding rate of two-state proteins: application of long-range order to folding rate prediction", *J. Mol. Biol.*, vol. 310, pp. 27-32, 2001.

[32] P. D. Dixit, and T. R. Weikl, "A simple measure of native-state topology and chain connectivity predicts the folding rates of two-state proteins with and without crosslinks", *Proteins*, vol. 64, pp. 193-197, 2006.

[33] T. Kikuchi, G. Némethy, and H. A. Scheraga, "Prediction of the location of structural domains in globular proteins", *J. Protein Chem.*, vol. 7, pp. 427-471, 1988.

[34] T. Kikuchi, "Inter-Cα atomic potentials derived from the statistics of average interresidue distances in proteins: application to bovine pancreatic trypsin inhibitor", *J. Comput. Chem.*, vol. 17, pp. 226-237, 1996.

[35] T. Kikuchi, "Study of protein fluctuation with an effective inter-Cα atomic potential derived from average distances between amino acids in proteins", *J. Comput. Chem.*, vol. 20, pp. 713-719, 1999.

[36] T. Kikuchi, "Analysis of 3D structural differences in the IgG-binding domains based on the interresidue average-distance statistics", *Amino Acids*, vol. 35, pp. 541-549, 2008.

[37] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of state calculations by fast computing machines", *J. Chem. Phys.*, vol. 21, pp. 1087-1092, 1953.

Received: December 28 2010 Revised: February 21, 2011 Accepted: March 08, 2011