We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

4,800
Open access books available

122,000
International authors and editors

135M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS BOOK CITATION INDEX INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

5

Analyses of Sequences of (β/α) Barrel Proteins Based on the Inter-Residue Average Distance Statistics to Elucidate Folding Processes

Masanari Matsuoka, Michirou Kabata, Yosuke Kawai and Takeshi Kikuchi
Department of Bioinformatics, College of Life Sciences, Ritsumeikan University
Japan

1. Introduction

It is well-known that many proteins fold into unique 3D structure (Anfinsen & Scheraga, 1975; Pain, 2000). The mechanisms by which an amino acid chain forms a complicated tertiary structure have been studied extensively (Sato et al,, 2006; Sato & Fersht, 2007; Sosnick & Barrick, 2011; Schaeffer & Daggett, 2011) but are still not understood in a comprehensive way (Bowman et al., 2011). It is well-recognized that all information on the 3D structure of a protein is coded in its amino acid sequence (Anfinsen & Scheraga, 1975; Pain, 2000). Hence we should be able to extract the information from the sequence. However, this is a significant, long-standing and unsolved problem in structural bioinformatics. Gross features of protein 3D structures or protein folds are characterized by some combination of secondary structural elements, and the ways of combination of secondary units are full of variety (Lesk, 2010). This situation does not allow us to construct a simple picture of the protein folding mechanisms. Among such a variety of protein folds, frequently appearing common folds, so-called superfolds (Orengo et al., 1994), are attractive targets for studying their folding mechanisms. The crucial point is that in general the fold of proteins tends to be more conservative than their sequences, i.e., sometimes proteins sharing the same fold show low sequence homology (Orengo et al., 1994; Jennings & Wright, 1993, Cavagnero et al., 1999; Nishimura et al., 2000). This fact implies the difficulty in approaching this problem using standard bioinformatics techniques such as multiple alignment techniques and so on.

We have been investigating this problem for several proteins using inter-residue average distance statistics of proteins. The tool we introduced is a kind of predicted contact map constructed from the sequence of a protein disregarding the knowledge of its 3D structure. We call this map the Average Distance Map (ADM) (Kikuchi et al., 1988; Kikuchi, 2002; Kikuchi 2011). The ADMs have seen used to analyze the folding problems of proteins in the fatty acid binding protein family (Ichimaru & Kikuchi, 2003; Kikuchi, 2011), the globin family (Ichimaru & Kikuchi, 2003; Nakajima et. al, 2005; Kikuchi, 2011), the c-type lysozyme family (Nakajima & Kikuchi, 2007; Kikuchi, 2010), IgG binding domains (Kikuchi, 2008; Kikuchi, 2011) and β-sandwich proteins (Ishizuka & Kikuchi, 2011). In this chapter, a new application of this technique to (β/α) barrel protein is presented. We further discuss the evolutionary variation of predicted folding units in a (β/α) barrel protein by analyzing sequences of homologues in a family.

The folding scenario of a protein may be like the following. One or several portions in a sequence form (a) partial hydrophobic collapse(s). Each portion may grow to a larger assembly with a native-like configuration or two or more portions may merge into a larger block to form a native-like configuration. Such regions assemble to form a folding transition state. Then, the final native structure emerges very quickly. The ultimate goal is to learn how a protein folds into a complicated fold via above scenario. Thus we would like to know which parts of the sequence form the native-like 3D configurations in the early stage of folding. These parts may correspond to a 3D structural formation portion at the folding transition state. In the present study, we try to predict a position of the initial hydrophobic collapse.

In our experiences so far, a protein in the T4-phage lysozyme family consists of two well-structured domains, and our method predicts two distinct folding units (Kawai et al., 2011). One clear folding unit and a short relatively weak one are predicted for a protein in the globin family (Kawai et al., 2011). The sequence of a β sandwich protein is predicted to contain two folding units, and this is interpreted to mean that these two portions merge into a larger block in the protein and form a native-like structure (Ishizuka & Kikuchi, 2011). Thus, it is interesting to see how our method predicts for the sequence of a (β/α) barrel protein and how its folding process can be interpreted from the predictions. It is also observed that location of such predicted regions is robust among evolutionally related proteins (Kawai & Kikuchi, unpublished). This suggests a robustness of the location of folding units during evolution of life on Earth. We seek to reveal this point for (β/α) barrel proteins.

2. (β/α) barrel protein

A (β/α) barrel protein shows a remarkable feature of the 3D scaffold which is constituted eight cyclically arranged successive (β/α) units with high symmetry. This fold is called the "TIM barrel" because the first discovered protein with this scaffold was triose phosphate isomerase (see Fig. 1). Since then, a huge number of proteins with the TIM barrel fold have

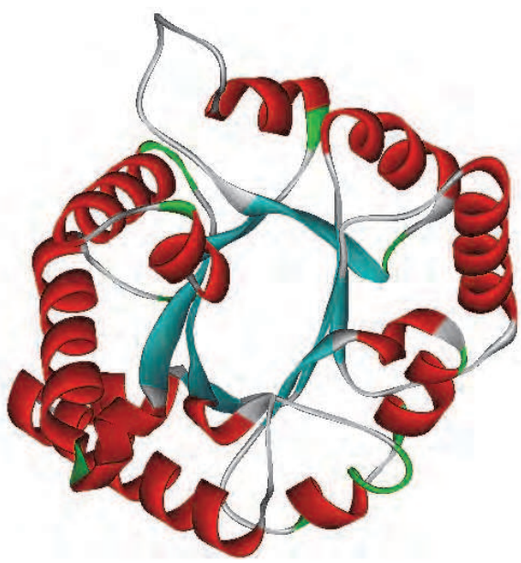


Fig. 1. An example of the 3D structure of a (β/α) barrel protein, triose phosphate isomerase (PDB: 1TIM). This is a typical (β/α) barrel protein consisting of eight (βα) units.

been found, and they show a variety of functions. Thus, elucidation of the folding mechanism and evolution of the (β/α) barrel scaffold is a quite interesting and challenging problem. There are many studies on folding mechanisms of several proteins with TIM barrel folds (Akanuma & Yamagishi, 2008; Gu et al., 2007; Silverman & Harbury, 2002; Seitz at al., 2007).

Among them, the recent studies on 3D structures and sequences of imidazole glycerol phosphate synthase (HisF) from Thermotoga maritime and N′-[(5'-phosphoribosyl) formimino]-5-aminoimidazole-4-carboxamide ribonucleotide isomerase (HisA) (Lang at al., 2000; Höcker et al., 2001) are remarkable. The 3D structures of HisF and HisA are presented in Fig. 2. According to these studies, 3D structural similarities are observed among the N terminal and C terminal halves of HisA and HisF (Lang et al., 2000). The sequence similarities among the sequences of the N terminal and C terminal halves of HisA and HisF are not high, but some conserved residues with similar properties are observed (Lang et al., 2000). These findings suggest that that the $(\beta/\alpha)_4$ half-barrel may fuse to yield a $(\beta/\alpha)_8$ complete barrel protein. This hypothesis that a half barrel of each protein can fold independently was confirmed by Höcker et al. (2001), and a $(\beta/\alpha)_8$ barrel protein designed as a fused identical half barrels can also form a stable folded structure (Seitz et al., 2007).

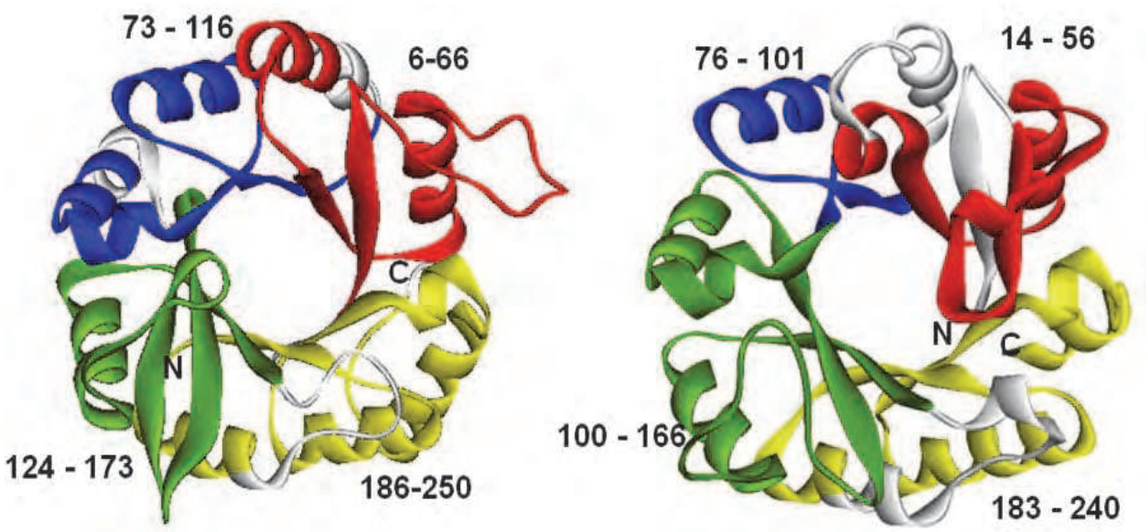


Fig. 2. 3D structures of 1THF(a) and 1QO2(b). Each region colored in red, blue, green or yellow denotes, respectively, the first, second, third or fourth folding unit predicted by each ADM (see text).

The two-fold symmetry may be further broken down into a four-fold symmetry, that is, a $(\beta\alpha)_2$ quarter-barrel is a symmetry unit. For example, the conserved GXD/GXG motif is repeatedly observed within α1-β2, α3- β4, α5-β5 and α7- β8 loops in HisF. Such symmetry is also observed for other $(\alpha\beta)_8$ barrel proteins based on the location of functional sites in them (Nagano at al., 2002). The phosphate binding sites in HisF corresponds to the four-fold symmetric active sites (see Fig. 3). Furthermore, Richter et al. (2010) designed a $(\alpha\beta)_8$-barrel protein composed of four identical $(\beta\alpha)_2$ quarter-barrel units that form a stable 3D structure by the introduction of disulfide bridges. Their results suggests that HisF evolved from an ancestral $(\beta\alpha)_2$ quarter-barrel via a $(\beta\alpha)_4$ half-barrel into the $(\alpha\beta)_8$-barrel (Richter et al., 2010).

On the other hand, the NMR study by Setiyaputra et al. (2011) demonstrated that the truncated phosphoribosylanthranilate isomerase (trPRAI), which is three-quarter-barrel-sized fragment of a $(\beta\alpha)_8$ barrel, forms the distinct 3D structure (Setiyaputra et al., 2011). This observation may suggest that the $(\beta\alpha)_2$ quarter-barrel is a kind of structural module in a $(\beta/\alpha)_8$ barrel protein but all 4 modules are not always indispensable.

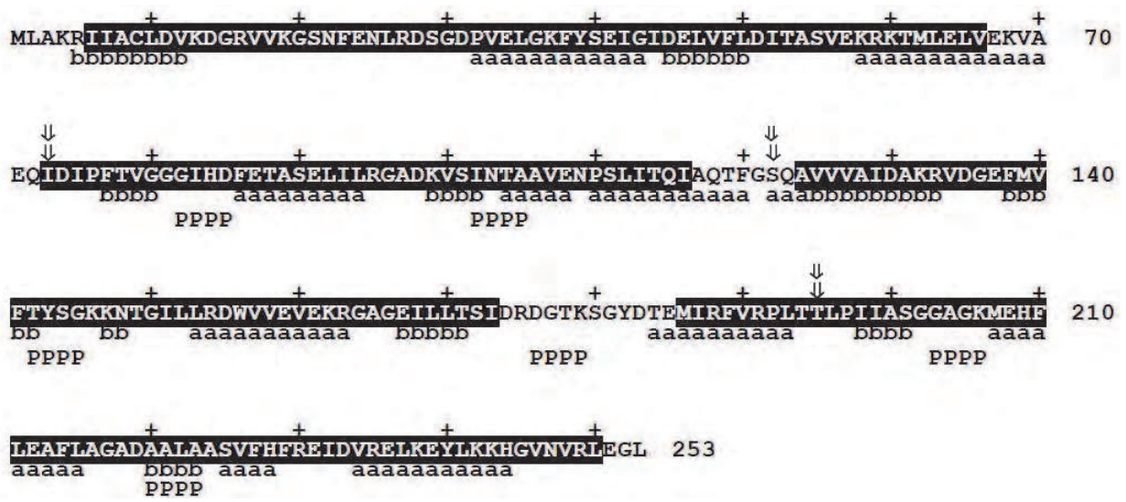


Fig. 3. Sequences of (a) HisF and (b) HisA. A residue in a β-strand is labeled by "b" and that in an α-helix by "a". A region with residues written by white letters with black background denotes the predicted region by ADM (see text). The symbol "PPPP" denotes an active site where phosphate binds in HisF and "(PPPP)" denotes a site in HisA that corresponds to a phosphate binding site in HisF. An arrow with a double line points to a boundary between two $(\beta\alpha)_2$ units.

Thus, HisF and HisA have attracted a lot of interest from researchers studying folding mechanisms and evolution. Hence we take HisF and HisA as examples of $(\beta/\alpha)_8$ barrel proteins in the present study.

3. Techniques used in the present work

3.1 Average distance map (ADM) method

The average distance map (ADM) method is a technique to predict structure-forming or compact portions in the amino acid sequence of a protein, and details of the method are described in (Kikuchi et al., 1988; Kikuchi, 2011). We have been confirming that ADMs contain variety of information on 3D structures and folding of proteins in spite of the simplicity of the method. The regions predicted by the ADM of a protein sequence correspond to; (1) nuclei of structural domains, (2) a nucleus of a structural domain and a portion forming a structure by interaction with the nucleus of the domain, or (3) two regions that form a stabilized structure by merging (Kawai, Matsuoka & Kikuchi, 2011). The essence of the procedure is as follows.

1. Calculation of inter-residue average distances in proteins

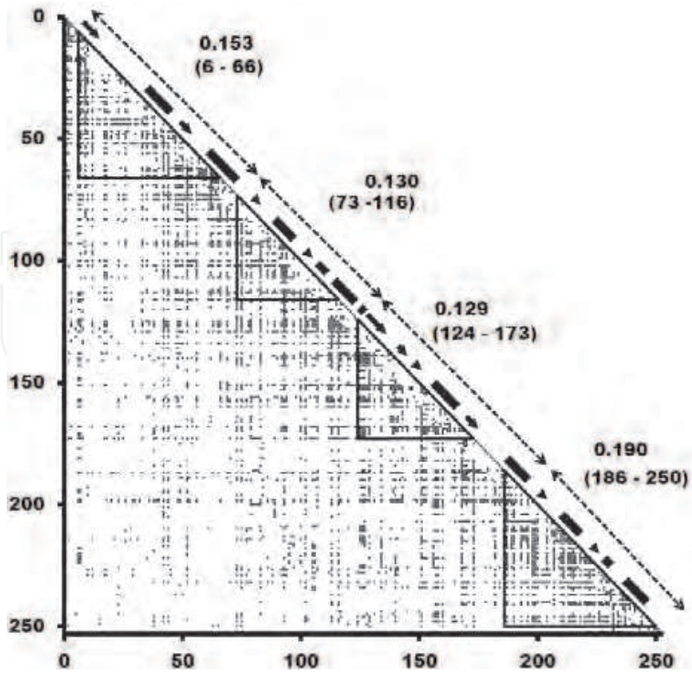
Inter-residue average distances were calculated using proteins with known 3D structures taking separation of two residues along the amino acid sequence of a protein into consideration. That is, the average distances were calculated for a residue pair within each group, i.e., 1≦k≦8, 9≦k≦20, 21≦k≦30, 31≦k≦40 and so on where k=|i-j|and i and j mean the i-th and j-th residues of the sequence. Each group of separation is referred to as a range, and each range is defined as; M=1 for 1≦k≦8, M=2 for 9≦k≦20, M=3 for 21≦k≦30 and so on (Refer to Kikuchi et al. (1988) for proteins used in the calculations of average distances). Here, an inter-residue distance means the distance between Cα atoms in Cartesian space.

2. Construction of a predicted contact map

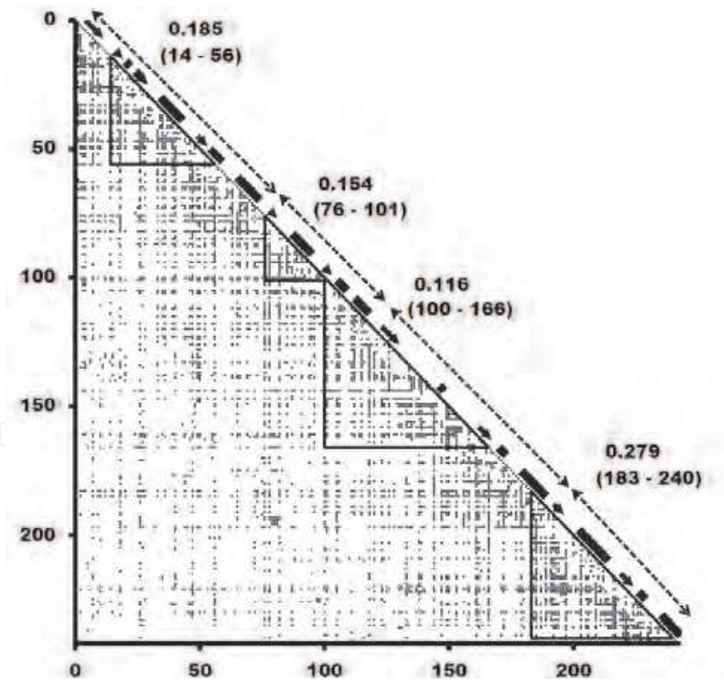
Our ADM analysis entailes making some plots showing when the average distance between a pair of residues in the sequence of a protein is less than some threshold value. A threshold value is defined for each range to reproduce density of plots on a contact map constructed from the spatial 3D structure of a protein under consideration. The prediction of density of plot on a contact map is made according to the way described by Kikuchi et al. (1988). Examples of ADMs are presented in Fig. 4. These ADMs were constructed based on the sequences of HisF and HisA.

3. Prediction of a compact region or a folding unit in a given sequence with ADM

When two segments in a protein form many contacts, the ADM should show be a region with a high density of contacts corresponding to plots on the contact map of a protein. Such a region shows a sudden change of the density of plots at its boundary on a map. Suppose that a map is divided into two parts by a line parallel to the ordinate of the map, thereby creating triangular and trapezoid parts. The difference in the density of plots between these two parts should be minimum or maximum at the boundary of the region with high density of plots. The same thing is also true when a map is divided into two parts by a line parallel to the abscissa. Let $\Delta\rho_i^v\left(\Delta\rho_j^h\right)$ be the differences of plot densities between two parts defined by lines parallel to the ordinate (abscissa) of a map. Then, the boundary of a region with high density contacts on a map can be detected by maxima and minima of the values of and as depicted in Fig. 5(a). In the example of Fig. 5(a), there are valleys at A and D , and peaks at B and C, and thus the interactions between the segments



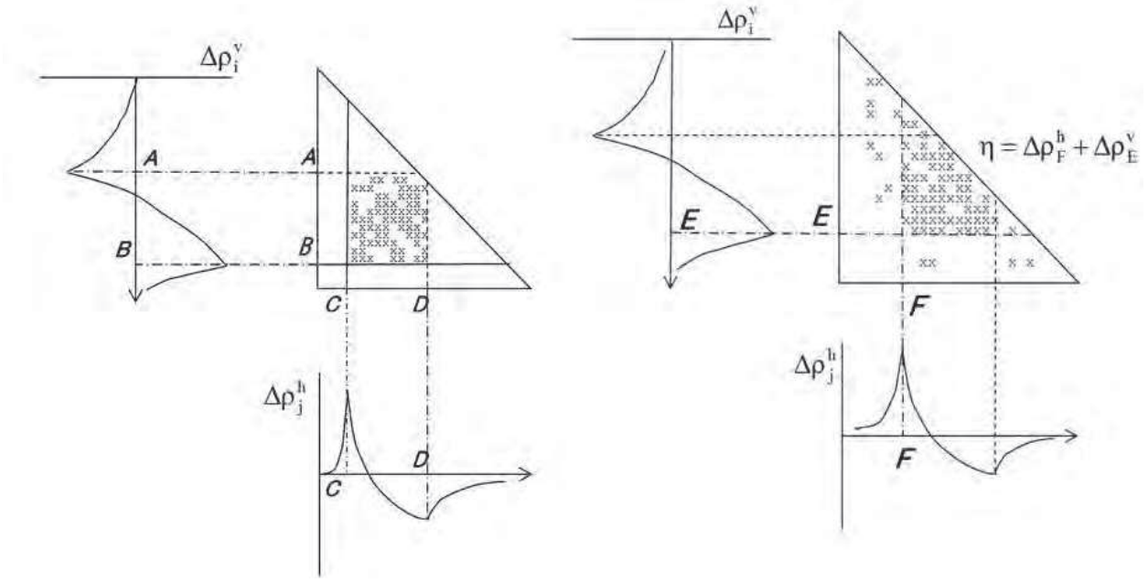
(a)



(b)

Fig. 4. ADMs for (a) HisF and (b) HisA. A bar along the diagonal denotes a position of a β-strand and an arrow does that of an α-helix. A region of a predicted folding unit is enclosed by solid lines in a map with a parenthesis and a numeral denotes the η value corresponding to the corresponding predicted region. A broken double arrow means the portion of a corresponding $(\beta\alpha)_2$ unit.

A-B and C-D are observed as a high density region of contacts on a map. Sometimes, a compact region in a protein can be observed as a high density region of contacts near the diagonal of a map. Such a region can be detected by peaks in the values in and as shown in Fig. 5(b). Peaks at E and F are observed in Fig. 5(b), and this predicts the compact region E-F in the sequence. It is convenient to define η value = $\Delta\rho_E^v + \Delta\rho_F^h$ in Fig. 5(b) as a measure of compactness of the region E-F. For several proteins, we have confirmed that such a compact region on ADM can also be regarded as a folding unit in the sequence of a protein (Ichimaru & Kikuchi, 2003; Nakajima et. al, 2005; Nakajima & Kikuchi, 2007; Kikuchi, 2011).



(a) (b)

Fig. 5. (a) When a map is divided into two parts by a line parallel to the ordinate of the map creating triangular and trapezoid parts, the difference of the density of plots between these two parts should be minimum or maximum at the boundary of the region with high density of plots. The same thing is also true when a map is divided into two parts by a line parallel to the abscissa. $\Delta\rho_i^v$ $\left(\Delta\rho_j^h\right)$ denote the differences of plot densities between two parts (the triangular and trapezoid parts) defined by lines parallel to the ordinate (abscissa) of a map. A peak and a valley appear at the boundaries of a high dense region of a plot of or. This hypothetical map suggests the interaction between the segments A-B and C-D. (b) A compact region or a domain in a given protein can be observed as a highly dense region of plots along the diagonal of a map. This figure shows a schematic drawing of a compact region at F-E. We define η as a measure of the compactness of the region, i.e. $\eta = \Delta\rho_r^h + \Delta\rho_q^v$.

In Fig. 4, we show the predicted folding units with η values for (a) HisF and (b) HisA in respective ADMs. Such a region corresponds to that enclosed by triangle in each map.

3.2 Multiple alignment analyses with homologues of HisF and HisA

It is interesting to see how predicted folding units in the sequences of the homologues of HisA and HisF appear in the sequences, i.e., whether predicted folding units are common among homologues. If commonality of folding units were observed, folding units of a

protein could be considered to be robust during evolution, and thus the folding process should be conserved during evolution. In the present study, evolutionary analyses with multiple alignments for homologues are also tried in the following procedure.

1. Homologous sequences of HisF and HisA were searched within the Uniprot and Swiss-prot databases. We used BLAST (Altschul et al, 1990) as a search algorithm and collected homologues less than 0.01 of their e value.
2. The multiple alignments of the collected sequences were made with ClustalW.
3. Phylogenetic trees of the collected sequences of homologues of HisF and HisA obtained in the way above were made based on the multiple alignments using the Neighbor-Joint method (Saitou & Nei, 1987) with 100 times bootstrapping.
4. For all sequences, the predictions of the positions of folding units by ADMs were made with AutoADM (Kawai et al., unpublished).

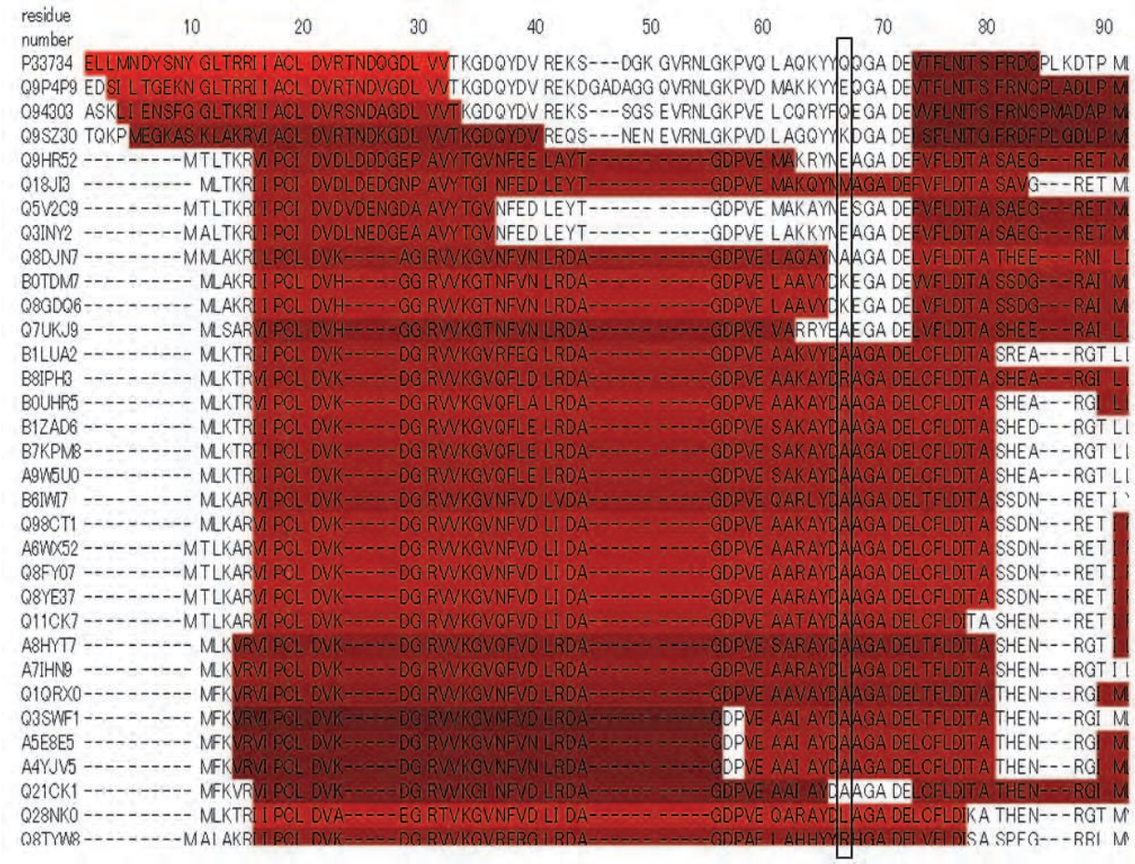


Fig. 6. A part of aligned sequences of aligned sequences in a multiple alignment. A brighter red region indicates a larger η value region. The number of residues within predicted folding units is counted at each site in a multiple alignment, e.g., the number of residues in red zones in the region enclosed by a black rectangle is counted and a histogram of such numbers is constructed for all sites of the multiple alignment.

A part of aligned sequences is presented in Fig. 6. In this figure, a region colored by red in each sequence denotes a predicted folding unit in the protein. A brighter red region means a larger η value region.

5. The number of residues within predicted folding units is counted at each position of a residue in a multiple alignment. That is, a histogram of the number of residues within the predicted region for an each position in a multiple alignment is made.

4. Analyses for HisA and HisF

4.1 Predicted folding units in HisF and HisA by ADMs

The ADM for HisF presented in Fig. 4(a) predicts four folding regions; 6-66, 73-116, 124-173 and 186-250 (regions enclosed by solid lines) as illustrated in this figure. Each region contains a phosphate binding site as shown in Fig. 3 (and also in Fig. 7). That is, this ADM predicts that HisF contains four folding units. The η values of these regions are 0.153, 0.130, 0.129 and 0.190, respectively, and thus the η values of these portions are relatively similar suggesting that each of these regions constitutes a folding unit with equal significance. From Figs. 3(a) and 4(a) it is easily confirmed that these predicted parts correspond to β1-α1- β2-α2, β3- α3-β4- α4, β5- α5- β6 and α6- β7- α7- β8- α8 in the 3D structure of HisF, and each of these regions corresponds well to the positions of each $(\beta\alpha)_2$ unit. The structures of these parts are colored in red, blue, green and yellow in Fig. 2(a).

On the other hand, the folding units of HisA are predicted as 14-56, 76-101, 100-166 and 183-240 according to its ADM as shown in Fig. 4(b). In Fig. 3 (and also in Fig. 7), the corresponding phosphate binding sites in HisF are presented, and the second, third and fourth predicted folding units contain such regions. Thus, HisA can be regarded as consisting of four folding units. The η values of these regions are 0.185, 0.154, 0.116 and 0.279, respectively. The η values of these regions are relatively similar except the larger value of the fourth region. This result may suggest the stronger significance of the fourth region and other three parts would contribute equally to the scaffold of the protein. These parts correspond to β1- α1- β2- α2, β3-α3- β4, α4- β5- α5- β6 and α6- β7- α7- β8- α8 in the 3D structure of HisA, and thus again each of these regions corresponds well to the positions of each of $(\beta\alpha)_2$ unit as confirmed in Figs. 3(b) and 4(b). The 3D structures of these parts are depicted in Fig. 2(b).

4.2 Multiple sequence alignment analyses of homologues of HisF and HisA and the location of folding units

Collected in Fig. 7 are 184 and 232 homologues of HisF and HisA. For these we made ADM predictions and multiple alignments for these sequences. For a detailed explanation of Fig. 7, see the caption. A colored bar below the alignment shows a possible common folding unit in this group of proteins. Each of these regions was defined as a region where the majority of sequences in the multiple alignment shows the positions of folding units, i.e., the values in the histogram are relatively large. The histogram of the number of residues within predicted folding units is shown at the bottom of the figure.

Taking the four predicted folding units in HisF by the ADM into account, careful observation of the multiple alignment in Fig. 7 reveals conservation or robustness of the region of a predicted folding unit. For example, we can observe in Fig. 7 that the N terminal part of each sequence always shows a predicted folding unit, i.e. red zone in each sequence in Fig. 7. This conservation or robustness does not mean conservation or robustness of residues in the sequence but rather that of the properties of residues in the sequence. From the histogram, a common folding unit can be assigned in the N terminal part as a region with high values in the

histogram as symbolized by the red bar below of the alignment. The boundaries of the red bar were defined by the height of the histogram around this region and visual inspection of the location of the first folding units in the alignment, i.e., the red zones. We call a region defined in this way a common folding unit. Thus, some robustness of the first folding unit can be observed and this region corresponds well to the first $(\beta\alpha)_2$ units as seen in Fig. 7. The fourth region symbolized by the yellow bar also shows relatively large robustness and corresponds to the fourth $(\beta\alpha)_2$ unit. This region shows another common folding unit. We confirm that such analyses with the multiple alignment in this manner reveal clearly the location of folding units and these regions correspond well to the portion of $(\beta\alpha)_2$ units.

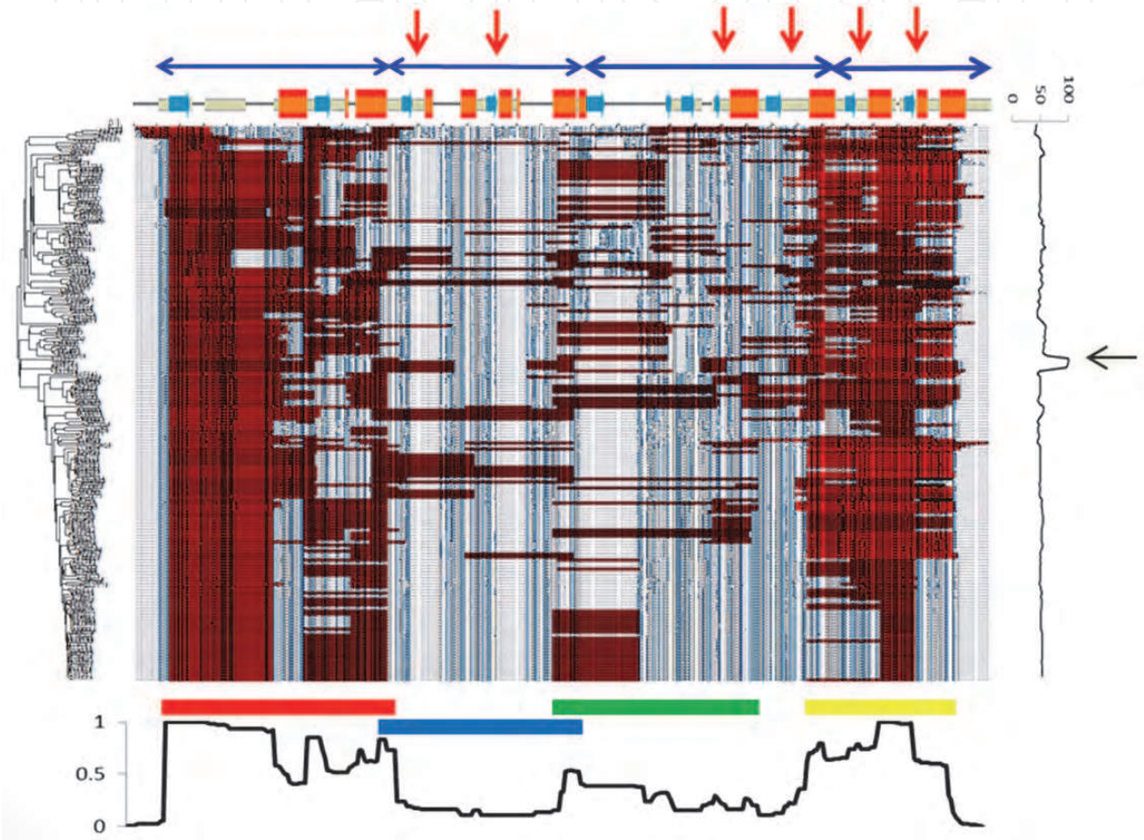


Fig. 7. Whole multiple alignment for HIsF and its homologues. The red arrows at the top of the figure denote a position of phosphate binding sites. A dark blue two headed arrows just below the red arrows indicate a position of $(\beta\alpha)_2$ barrel unit. The bars just below these blue arrows show secondary structures, i.e., a blue bar denotes a β strand and red bar an α helix. The phylogenetic tree is presented on the left side of the figure. A colored bar at the bottom of the alignment means a possible common folding unit in this group of proteins. Each of these regions was defined as a region at where the majority of sequences in the multiple alignment shows the positions of folding units, i.e., the values in the histogram are relatively large. The boundaries of a region can be defined by peaks of the histogram at the bottom. The histogram graphs the number of residues within predicted folding units. A graph on the right side of the figure denotes a plot of sequence homology (%) to that of HisF. The average homology of homologue sequences to that of HisF is about 50%. The arrow at the peak means the 100% homology with HisF sequence.

The third region of Fig. 7 seems to have varied frequently during the evolution from the ancestral HisF, and only a modestly high region can be observed in the histogram. That is symbolized by the green bar. Furthermore, we cannot observe the second folding unit in the histogram, and we just see predicted folding units, i.e., red zones, in some sequences of homologues by visual inspection in this region. Temporarily, we assign two common folding units at the regions symbolized by the blue and green bars.

In Fig. 9(a), we show the 3D structures of the common folding units in HisF. Each color of the region in the structure corresponds to the color of each bar in Fig. 7. The reason why the second region implies the low robustness of sequence properties is not clear. It may relate to the folding property or functional property.

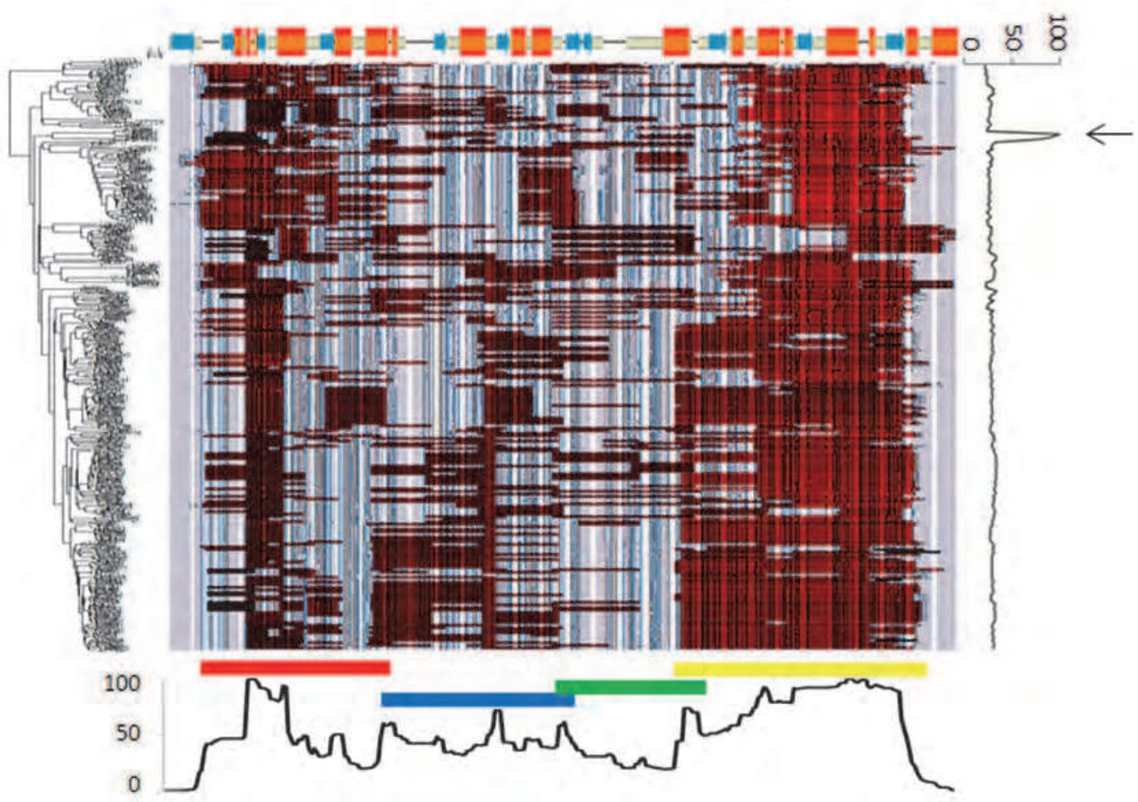
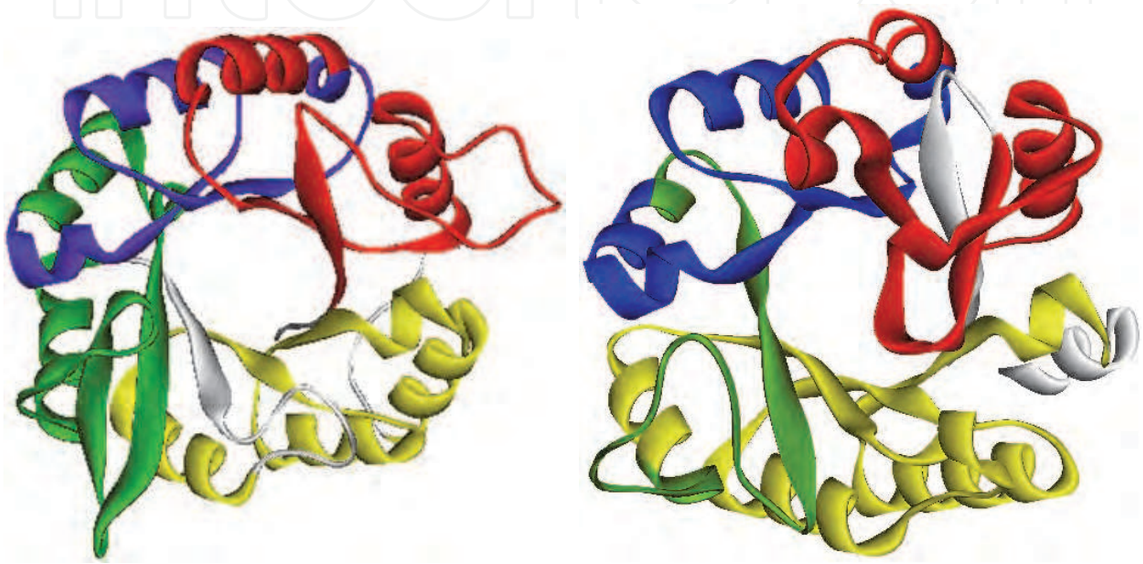


Fig. 8. Whole multiple alignment for HisA and its homologues. See the caption of Fig. 7 for detailes. A graph on the right side of the figure denotes a plot of sequence homology (%) with that of HisF. The arrow at the peak means the 100% homology with HisA sequence.

The whole multiple alignment for HisA and its homologues is presented in Fig. 8. The meaning of each symbol is the same as in Fig. 7. In HisA, a strong robustness is observed in the C terminal part of the sequences, and the region is symbolized by the yellow bar just below the multiple alignment in Fig. 8. This region is longer than the fourth $(\beta\alpha)_2$ unit suggesting that this part is the main region of the folding of HisA, and we define this region as a common folding unit. The first region also shows modest robustness (strong robustness of the shorter part in this region) indicating the existence of a common folding unit but not as strong as in HisF. This observation suggests some significance of the C terminal region

for folding mechanism of a protein in this group. Compared with the case of HisF, the histogram for HisA in Fig. 8 shows the modestly high region corresponding to the second predicted folding unit by ADM, and this region is symbolized by the blue below the alignment in Fig. 8. The corresponding third region cannot be assigned by the histogram, but by visual inspection predicted folding regions (red zones) appear in several sequences of homologues. So we make an additional assignment of the third common folding unit symbolized by the green bar. The 3D structures of the common folding units in HisA are presented in Fig. 9(b).



(a) (b)

Fig. 9. Region assigned as common folding units in (a) HisF and (b) HisA from Figs. 7 and 8, respectively.

Our ADM analyses predict the existence of four folding units in HisF and HisA, which correspond to $(\beta\alpha)_2$ units proposed by Richiter et al., 2010, only from their sequences. For HisA, the significance of the C terminal unit predicted by the ADM for its 3D structure formation can be speculated based on the η value of this region. The combination of the multiple alignment analyses with the ADM analyses reveals that the N and C terminal common folding units are robust and would be significant for the folding process of these proteins commonly. On the other hand, the second and third regions are easily varied and the participation of these units to folding mechanisms would not be common among the homologues.

Thus, we can predict the four folding units from properties of sequences of $(\beta\alpha)_8$ proteins through ADMs. It is rather difficult to obtain such information by standard techniques of sequence analyses (Richter et al., 2010). Furthermore, our method specifies a kind of plasticity of each region from the present multiple alignment analysis. We think that this information may useful for predicting the folding properties of a protein, this is discussed below.

5. Concluding remarks

As we see above, the ADM method predicts positions of possible folding units in a protein. Regions predicted by the ADMs in homologues of a protein reveal the conservation or robustness of a predicted region during evolution. Proteins treated in this chapter show high (8-fold) symmetry. The folding process of such proteins with high symmetry may be rather different from other folds. The present analyses along with the preceding investigations (Lang et al., 2000; Höcker et al., 2001; Richter et al., 2010) imply that $(\beta\alpha)_8$ proteins such as HisF and HisA consist of four independent $(\beta\alpha)_2$ units and these proteins start to fold at each $(\beta\alpha)_2$ unit. However, huge number of proteins with the $(\beta\alpha)_8$ fold exist, so the folding mechanisms of $(\beta\alpha)_8$ proteins may show wide variations. Actually, various folding mechanisms of $(\beta\alpha)_8$ proteins have been reported (Akanuma & Yamagishi, 2008; Gu et al., 2007; Silverman & Harbury, 2002; Seitz at al., 2007; Setiyaputra et al. (2011)). Such variation of folding may be related to the nonrobustness of the second and third common folding units in HisF and HisA. We will perform ADM analyses on additional $(\beta\alpha)_8$ proteins.

As mentioned in the Introduction, the folding scenario of proteins may be put in the following categories:

1. A one domain protein

A portion in the sequence of a protein starts to form a hydrophobic collapse, and this part grows to a whole sequence to form the native structure. Protein G, Protein A and so on may belong to this category. In this case, the ADM for a protein tends to predict the location of one folding core in the sequence (Kikuchi et al, 1988; Kikuchi, 2008).

2. A protein composed by two (or more) distinct domains

Two (or more) portions in the sequence of a protein start simultaneously to form hydrophobic collapses, and these parts grow independently to domains in the native structure. T4-lysozyme, papain and so on belong to this category (Kikuchi et al, 1988; Kawai, Matsuoka and Kikuchi, 2011). In this case, the ADM for a protein tends to predict the location of distinct folding cores corresponding to the domains in the sequence.

3. A protein formed by a main part with an interacting short fragment

A main portion in the sequence of a protein starts to form hydrophobic collapse, and this part grows to the partial native structure, and the rest (relatively short fragment compared with the main part) interacts with the main portion followed by the formation of the final native structure. According to our investigations with the ADM method, proteins with the Globin fold, fatty-acid binding protein and so on may belong to this category (Ichimaru & Kikuchi, 2003; Nakajima et. al, 2005; Kawai, Matsuoka & Kikuchi, 2011). In this case, the ADM for a protein tends to predict the location of the main part in the sequence.

4. A protein containing long range interactions along its sequence

Two (or more) portions in the sequence of a protein starts to form partial (or imperfect) hydrophobic collapses followed by aggregation of these two (or more) portions to form a more perfect hydrophobic core. This part grows to the final native structure. β-Sandwich proteins such as titin, azurin (Ishizuka & Kikuchi, 2011) and so on, proteins in the c-type

lysozyme fold (Nakajima & Kikuchi, 2007), αβ sandwich proteins (Matsuoka & Kikuchi, unpublished) such as ribosomal protein S6, and so on may belong to this category. In this case, the ADM for a protein tends to predict the location of folding cores which should aggregate into the larger hydrophobic core.

5. A protein with a highly symmetrical structure

Three (or more) portions in the sequence in the sequence of a protein start to form hydrophobic collapses independently, and each portion forms the native structure. The formation of the native structure may occur cooperatively, i.e., once a symmetrical unit starts to form the partial native structure, the rest of the protein tends to form the native structure. This case corresponds to the present work.

The scenarios of protein folding mechanisms are proposed based on the investigations of many researchers (Fersht, 1997; Sato et al,, 2006; Sato & Fersht, 2007; Sosnick & Barrick, 2011; Schaeffer & Daggett, 2011) and the results from our ADM analyses. A portion of a protein sequence contains a folding unit predicted by ADM, and the folding process proceeds by one of the scenarios above, we consider that with ADM analysis it is possible to predict the folding process of a related protein from only a sequence. Of course, there many unclear properties in ADM analyses for a protein, and we are continuing to elucidate the meanings of outcomes from the ADM analyses.

As presented in this chapter, ADM analyses combined with multiple alignment analyses provide fruitful results on evolutionary changes in the folding process of proteins in a family.

Finally, it would be quite nice if our techniques were able to contribute to ab initio protein 3D structure prediction. Unfortunately, the present stage is still far from this goal, but we are trying to do it by analyzing the relationships between the 3D structure of the folding unit of a protein and common properties of residues in this portion among homologues of a protein. We hope that we will be close to this goal in the near future.

6. References

Akanuma, S. & Yamagishi, A. (2008). Experimental evidence for the existence of a stable half-barrel subdomain in the $(\beta\alpha)_8$-barrel fold. J. Mol. Biol., 382, 458–466.

Altschul, SF., Gish, G., Miller, W., Myers, EW. & Lipman, DJ. (1990) Basic local alignment search tool. *J. Mol. Biol.*, 215, 403-410.

Anfinsen, CB & Scheraga, HA. (1975). Experimental and theoretical aspects of protein folding. *Adv. Protein Chem.*, 29, 205-300.

Bowman, GR., Voelz, VA. & Pande, VS. (2011). Taming the complexity of protein folding. *Curr Opin. Str. Biol.*, 21, 4-11.

Cavagnero, S., Dyson, HJ & Wright, PE. (1999). Effect of H helix destabilizing mutations on the kinetic and equilibrium folding of apomyoglobin. *J. Mol. Biol.*, 285, 269-282.

Fersht, AR. (1997). Nucleation mechanisms in protein folding. *Curr. Opin. Struct. Biol.*, 7, 3–9.

Gu, Z., Rao, MK., Forsyth WR., Finke, JM. & Mathews CR. (2007) Structural analysis of kinetic folding intermediates for a tim barrel protein, indole-3-glycerol phosphate synthase, by hydrogen exchange mass spectrometry and Gō model simulation. J. Mol. Biol., 374, 528–546.

Höcker, B., Beismann-Driemeyer, S., Hettwer, S., Lustig, A. & Sterner R. (2001). Dissection of a (βα)8-barrel enzyme into two folded halves. Nature Str. Biol., 8, 32-36.

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, 51, 515-530.

Ishizuka, Y. & Kikuchi, T. (2011). Analysis of the local sequences of folding sites in β sandwich proteins with the interresidue average distance statistics. The Open Bioinf. J., *5,* 59-68.

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

Kawai, Y., Matsuoka, M. & Kikuchi, T. (2011). Analyses of protein sequences using inter-residue average distance statistics to study folding processes and the significance of their partial sequences. Protein & Peptide Let. 18, 979-990.

Kikuchi, T., Némethy, G. Scheraga, HA. (1988). Prediction of the location of structural domains in globular proteins. *J. Protein Chem.*, 7, 427-471.

Kikuchi, T. (2002). Contact maps derived from the statistics of average distances between residues in proteins. Application to the prediction of structures and active sites of proteins and peptides, In: *Recent Research Developments in Protein Engineering.* Pandalai SG. (ed.), 1-48, Research Signpost, Kerala, India, ISBN 81-7736-147-3.

Kikuchi, T. (2008). Analysis of 3D structural differences in the IgG binding domains based on the interresidue average distance statistics. *Amino Acids*, 35, 541-549.

Kikuchi T. (2011). Decoding amino acid sequences of proteins using inter-residue average distance statistics to extract information on protein folding mechanisms, In: *Protein Folding*. Walters EC. (ed.), 465-487, Nova Science Publishers. Inc., NY, USA, ISBN 978-1-61728-990-32011.

Lang, D., Thoma, R., Henn-Sax, M., Sterner R. & Wilmanns, M. (2000). Structural Evidence for evolution of the β/α barrel scaffold by gene duplication and fusion. Science, 289, 1546-1550.

Lesk, AM. (2010). Introduction to protein science, architecture, function, and genetics (2nd edition), Oxford University Press, ISBN 978-0-19-954130-0, Oxford, UK.

Nagano, N., Orengo, OA & Thornton, JM. (2002). One fold with many functions: the evolutionary relationships between tim barrel families based on their sequences, structures and functions. J. Mol. Biol., 321, 741–765.

Nakajima, S., Álvarez-Salgado, E., Kikuchi, T., Arredondo-Peter, R. (2005). Prediction of folding pathway and kinetics among plant hemoglobins by using an average distance map method. *Proteins*, 61, 500-506.

Nakajima, S. & Kikuchi, T. (2007). Analysis of the Differences in the folding mechanisms of c-type lysozymes based on contact maps constructed with interresidue average distances. *J. Mol. Model.*, 13, 587-594.

Orengo, CA., Jones, DT. & Thornton, JM. (1994). Protein superfamilies and domain superfolds. *Nature*, 372, 631-634.

Nishimura, C., Prytulla, S., Dyson, HJ. & Wright, PE. (2000). Conservation of folding pathways in evolutionarily distant globin sequences. *Nature Struct. Biol.*, 7, 679-686.

Pain, RH. (ed.), (2000). *Mechanisms of Protein Folding* (2nd Edition), Oxford University Press, ISBN 0-19-963788-1, Oxford, UK.

Richter, M., Bosnali, M., Carstensen, L., Seitz, T., Durchschlag, H., Blanquart, S., Rainer Merk, R. & Sterner R. (2010). Computational and experimental evidence for the evolution of a (βα)8-barrel protein from an ancestral quarter-barrel stabilised by disulfide bonds. *J. Mol. Biol.*, 398, 763–773.

Saitou, N., & Nei, N. (1987). A neighbor-joining method: a new method for constructing phylogenetic tree. *Mol. Biol. Evol.*, 44, 406-425.

Sato, S., Religa, TL. & Fersht, AR. (2006). Φ-Analysis of the folding of the B domain of protein a using multiple optical probes. *J. Mol. Biol.*, 360, 850-864.

Sato S. & Fersht, AR. (2007). Searching for multiple folding pathways of a nearly symmetrical protein: Temperature dependent Φ-value analysis of the b domain of protein A. *J. Mol. Biol.*, 372, 254-267.

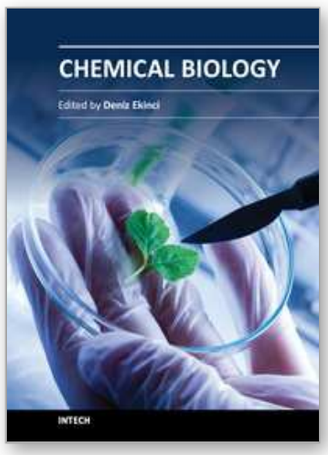
Schaeffer, RD. & Daggett, V. (2011). Protein folds and protein folding. *Protein Eng. Des. Sel.*, 24, 11–19.

Seitz, T., Bocola, M., Claren, J. & Sterner, R. (2007). Stabilisation of a (βα)8-barrel protein designed from identical half barrels. *J. Mol. Biol.*, 372, 114–129.

Silverman, JA. & Harbury PB. (2002). The equilibrium unfolding pathway of a (β/α) 8 barrel. *J. Mol. Biol.*, 324, 1031–1040.

Setiyaputra, S., Mackay, JP. & Patrick, WM. (2011). The structure of a truncated phosphoribosylanthranilate isomerase suggests a unified model for evolution of the $(\beta\alpha)_8$ barrel fold. *J. Mol. Biol.*, 408, 291–303.

Sosnick, TR. & Barrick, D. (2011). The folding of single domain proteins—Have we reached a consensus? *Curr Opin. Str. Biol.*, 21, 12–24.



Chemical Biology

Edited by Prof. Deniz Ekinci

ISBN 978-953-51-0049-2
Hard cover, 444 pages
Publisher InTech
Published online 17, February, 2012
Published in print edition February, 2012

Chemical biology utilizes chemical principles to modulate systems to either investigate the underlying biology or create new function. Over recent years, chemical biology has received particular attention of many scientists in the life sciences from botany to medicine. This book contains an overview focusing on the research area of protein purification, enzymology, vitamins, antioxidants, biotransformation, gene delivery, signaling, regulation and organization. Particular emphasis is devoted to both theoretical and experimental aspects. The textbook is written by international scientists with expertise in synthetic chemistry, protein biochemistry, enzymology, molecular biology, drug discovery and genetics many of which are active chemical, biochemical and biomedical research. The textbook is expected to enhance the knowledge of scientists in the complexities of chemical and biological approaches and stimulate both professionals and students to dedicate part of their future research in understanding relevant mechanisms and applications of chemical biology.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen