Vol. 18, pp. 50–59 (2021)
doi: 10.2142/biophysico.bppb-v18.006

Biophysics and Physicobiology
https://www.jstage.jst.go.jp/browse/biophysico/

Regular Article

Structure elements can be predicted using the contact volume among protein residues

Yasumichi Takase[1], Yoichi Yamazaki[1], Yugo Hayashi[1], Sachiko Toma-Fukai[1] and Hironari Kamikubo[1,2]

[1] *Division of Materials Science, Graduate School of Science and Technology, Nara Institute of Science and Technology, Ikoma, Nara 630-0192, Japan*
[2] *Institute of Materials Structure Science, High Energy Accelerator Research Organization (KEK), Tsukuba, Ibaraki 305-0801, Japan*

Received December 21, 2020; accepted February 15, 2021; Released online in J-STAGE as advance publication February 18, 2021

Previously, the structure elements of dihydrofolate reductase (DHFR) were determined using comprehensive Ala-insertion mutation analysis, which is assumed to be a kind of protein "building blocks." It is hypothesized that our comprehension of the structure elements could lead to understanding how an amino acid sequence dictates its tertiary structure. However, the comprehensive Ala-insertion mutation analysis is a time- and cost-consuming process and only a set of the DHFR structure elements have been reported so far. Therefore, developing a computational method to predict structure elements is an urgent necessity. We focused on intramolecular residue–residue contacts to predict the structure elements. We introduced a simple and effective parameter: the overlapped contact volume (CV) among the residues and calculated the CV along the DHFR sequence using the crystal structure. Our results indicate that the CV profile can recapitulate its precipitate ratio profile, which was used to define the structure elements in the Ala-insertion mutation analysis. The CV profile allowed us to predict structure elements like the experimentally determined structure elements. The strong correlation between the CV and precipitate ratio profiles indicates the importance of the intramolecular residue–residue contact in maintaining the tertiary structure. Additionally, the CVs between the structure elements are considerably more than those between a structure element and a linker or two linkers, indicating that the structure elements play a fundamental role in increasing the intramolecular adhesion. Thus, we propose that the structure elements can be considered a type of "building blocks" that maintain and dictate the tertiary structures of proteins.

Key words: contact map, folding element, residue–residue contact, precipitate ratio profile, protein folding

Corresponding author: Hironari Kamikubo, Division of Materials Science, Graduate School of Science and Technology, Nara Institute of Science and Technology, 8916-6, Takayama, Ikoma, Nara 630-0192, Japan. e-mail: kamikubo@ms.naist.jp

Introduction

Proteins possess large variations in their structures and functions. Currently, there are more than 140,000 structures registered in the Protein Data Bank (PDB) (www.rcsb.org) and 17,929 families in the Pfam [1,2]. Although a large number of protein structures are revealed, it is still unclear how a protein can spontaneously fold to its unique tertiary

◀ ***Significance*** ▶

This study aims to develop a computational method that can be used to predict structure elements (SEs) of proteins instead of experimental analysis. Additionally, the building block of the protein responsible for foldability is closely related to the extent of internal molecular packing. We believe that our study makes a significant contribution to the literature because it details a contact volume profiling method that can be implemented as a time- and cost-effective alternative to comprehensive Ala-insertion mutation analysis or circular permutation analysis.

structure, making it difficult to design artificial proteins. Although the folding mechanism has not been fully understood, there are significant improvements in the de novo design of artificial proteins proposed by Baker *et al.* [3–10]. They constructed computational models of the artificial proteins and successfully obtained proteins with the desired structure. However, success is still limited. Artificial proteins are highly stable but not functional in most cases, whereas naturally occurring proteins exhibit various kinds of functions and are marginally stable [6]. Therefore, to design functional proteins similar to the natural proteins, it is necessary to know an alternative way to design a moderately destabilized structure essential for functional expression.

Some pioneering studies have proposed a type of building block composed of proteins. Gō, in 1983, suggested that a protein is composed of compact structural motifs termed modules, which do not agree with secondary structures but correspond to exons [11]. Theoretical calculations based on the structures of proteins helped identify the segments of the structures that fold independently (foldon) [12]. There is a strong correlation between the foldons and the modules. Several studies have found that an amino acid sequence can be divided into several segments that constitute the original tertiary structure to retrieve its biological activities. This implies that these segments act as building blocks for structural assembly and biological activity [13,14]. Iwakura *et al.* investigated the importance of the contiguity of the sequence using circular permutation analysis of dihydrofolate reductase (DHFR) [15–17]. They examined the foldability upon possible circular permutations and proposed that the sequence of DHFR is composed of segments termed "folding elements (FE)" that are essential for foldability. Shiba *et al.* also examined the effect of the disconnection of the sequence on the solubility of DHFR using comprehensive Ala-insertion mutation analysis [18]. They identified segments that do not allow the Ala-insertion mutation, termed "structure element (SE)."

SE and FE are closely related to each other, and the resultant segments are found at similar positions within the sequence. While the circular permutation can be applied to proteins in which the N- and C-terminal regions are close to each other in the structure, the comprehensive Ala-insertion mutation analysis has no such limitation of application. Additionally, since the effect of the Ala-insertion on foldability is moderate compared to the circular permutation, every Ala-insertion mutant can be expressed in *E. coli*, whereas the circular permutation inside FE inhibits the expression in some cases. Therefore, the comprehensive Ala-insertion mutation analysis of DHFR allows quantitative estimation of the effect of the insertion. In this analysis, all possible Ala-insertion mutants of DHFR were prepared and the effect of the Ala-insertion mutation on the amount of insoluble fraction accumulated in a cell during the expression process was noted. Thus, we examined the ratio of the amount of precipitate formed after cell disruption to the total expressed protein along the sequence (precipitate ratio profile). By setting a criterion, the SEs can be identified from the precipitate ratio profile: SE1, I2-V10; SE2, L28-E48; SE3, I60-L62; SE4, W74-I82; SE5, E90-G96; SE6, G97-P105; SE7, K106-H114; SE8, E120-G121; SE9, D127-Y128; SE10, E129-P130; SE11, W133-S135; and SE12, Y151-E157.

Since the disruption of only one segment of SE and FE using the mutation results in a substantial decrease in the solubility and foldability of DHFR, each SE and FE can be assumed to be a type of building block indispensable for the tertiary structure formation. Identifying the building block of proteins potentially gives us a clue to learn a new perspective on how the structural information is encoded into the native sequence of a protein. However, the protein design rules are still obscured because the SE and FE are identified only from the DHFR. The critical limitation arises from the considerable cost and time involved in the comprehensive mutation analysis. Therefore, the development of a prediction method for SEs instead of the experimental analysis is an urgent necessity for further application such as protein structure design.

To realize the prediction, we focused on the intramolecular residue–residue contacts (IRRC). IRRC is frequently used to analyze the topology of structures by performing contact network analysis or employing the protein contact map, a 2D representation of IRRC [19–23]. Additionally, IRRC is used to identify the semantic borders of the amino acid sequence; the module hypothesis mentioned above came from the contact map analysis [11]. It can be postulated that IRRC reflects not only the topological information but also information of the semantic borders of the sequence. In this study, to predict SEs, we introduced a simple and effective parameter to quantify IRRC, which is the contact volume (CV) of overlapped area among the residues. We calculated the CV profile along the sequence of DHFR from the crystal structure and examined the relationship between the CV profile and the precipitate ratio profile obtained from the previous comprehensive Ala-insertion mutation analysis. Finally, it was revealed that there is a strong correlation between the CV and precipitate ratio profiles. These results indicate that the SEs determined using the experimental analysis can be predicted by calculating the CV profile. Additionally, the building block of the protein responsible for foldability is closely related to the extent of internal molecular packing.

Materials and Methods

Contact volume calculation

The crystal structure of DHFR [PDB ID: 1rx4] obtained from the Protein Data Bank [1] was used for the CV calculation. Ligands and water molecules were removed from the structure. Additionally, hydrogen atoms missing in the structure were added using UCSF Chimera package [24]. CV was defined as the overlapped volume among atoms whose radii are assumed to be their van der Waals (vdW) radius added by 1.4 Å (Fig. 1A, B). The value of 1.4 Å corresponds to a water molecule's approximate radius [25–28]. A schematic representation of the CV is shown in Figure 1C, where atom a is in contact with atom b and atom c. The inner and outer circle radii represent the vdW radius and the vdW radius added by 1.4 Å, respectively. Although atom a is not in vdW-contact with atoms b and c, they are regarded as in contact when the radius of a water molecule is considered. An increase in the CV implies that penetrating the gap is more difficult for water molecules. In the case of Figure 1C, the CV of atom a with atom b (at the shortest distance from atom a) is calculated first (region 1, red). Then, the CV with atom c (at the second shortest distance from atom a) is obtained (region 2, blue), where the volume shared with region 1 (arrow) is excluded. The resultant CV of atom a is the sum of regions 1 and 2. CV among residues or CV among segments composed of multiple residues, were calculated in the same manner.

The Monte Carlo method was implemented for CV calculation [29]. In the Monte Carlo method, random points were generated in the atomic sphere. The fraction of the points in the overlapped volume was obtained. The CV was calculated by multiplying the volume of the atomic sphere by the fraction. To obtain the accurate number of random points, we calculated the CV values of DHFR with different number of random points three times. The mean value of all residues' standard deviation gradually decreases from 50 to 1000 points (Supplementary Fig. S1A). Since the standard deviation and the resultant CV values (Supplementary Fig. S1B) become close to equal after 1000 points, we concluded that the CV value is reliable by generating 1000 or more random points. In this study, the CV was calculated by generating 10000 random points.

The CV profile along the sequence of DHFR was obtained by averaging the CVs at each amino acid residue over the moving window. The moving window is often used to determine various physicochemical characteristics along the sequence [30–32]. The proper window size was determined by comparing the correlation coefficients and apparent match between the CV profile and the precipitate ratio profile when changing the window sizes of 2, 4, 6, and 8 (See Supplementary Fig. S2). The correlation coefficients between the CV profiles (lines) and the precipitate ratio profiles (bars) of the window sizes (2, 4, 6, and 8) are 0.53, 0.62, 0.63, and 0.63, respectively. The correlation coefficients become almost constant when the window size is four or larger. When the window size is 2, there is a considerable variation in the CV's values, and the precipitate ratio profile cannot be reproduced. The width of the CV profile becomes wider than that of the precipitate ratio profile when the window size is larger than 6, and it becomes difficult to predict the pSE. Therefore, we adopted a four-residue moving window in this study.

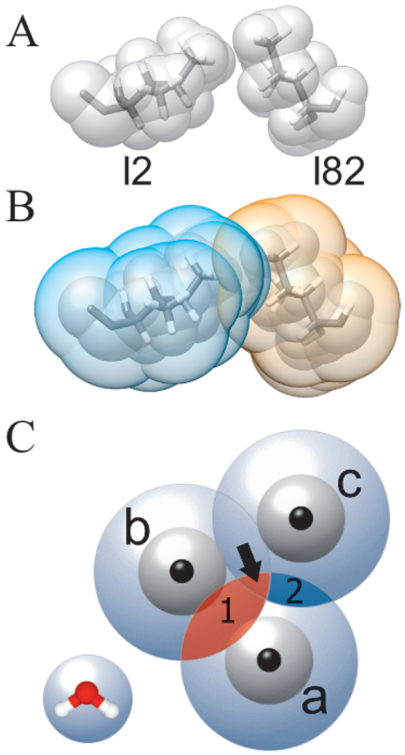

Figure 1 (A) The sphere models of I2 and I82 in DHFR, where the radii of the atoms are their vdW radius. (B) The contact when the average radius of a water molecule (1.4 Å) is added to the vdW radius (I2 blue, I82 orange). (C) Schematic representation of the region corresponding to the CV. The region colored by gray and light blue represents the atomic sphere with the vdW radius and the vdW radius added by 1.4 Å, respectively. The sphere with a radius of 1.4 Å approximating a water molecule is shown for reference. It is found that there is no space for water molecules to penetrate region 1 (red) and region 2 (blue) although these atoms are not in vdW-contact with each other. The arrow indicates part of the overlapped volume among atom a, b, and c. This CV was excluded when calculating the CV between atom a and c. The CV between atom a and atom c is shown in blue (region 2, blue).

iASA, riASA, and contact number profile calculation

Accessible surface area (ASA) and relative accessible surface area (rASA) were calculated by using NACCESS [33] in ProTSAV [34]. rASA is defined as the ratio (%) of the actual accessible surface area (ASA) of an amino acid residue to its standard accessible surface area (stdASA).

Relative inaccessible surface area (riASA) is defined by the ratio of the inaccessible surface area of amino acid residues (iASA=stdASA–ASA) to stdASA. Residue-residue contact map was obtained by using CMview [35], where the residues were considered to be in contact with each other when the distance between Cα atoms was shorter than 8 Å [36]. The contact number profile was calculated from the residue-residue contact map. The contact number at an amino acid residue derived from the adjacent 1–3 residues was excluded, following the CV profile calculation method (see below). The riASA, iASA, and contact number profile were finally obtained by averaging these values at each amino acid residue over the moving window of four residues.

Visualization

Visualization of the contact network was performed using Gephi (ver. 0.9.2) [37]. Molecular graphics were obtained using the UCSF Chimera package [24].

Results and Discussion

In this study, we used DHFR as a model protein, for which SEs and FEs were identified. SEs were determined by comprehensive Ala-insertion mutation analysis [18]. In this analysis, all possible Ala-insertion mutants of DHFR were prepared and the effect of the Ala-insertion mutation on the structure formation was examined by evaluating the amount of insoluble fraction integrated during the expression process. The precipitate ratio values of each Ala-insertion mutant were plotted against the amino acid residue number (Fig. 2A, Bar), termed precipitate ratio profile. The value is defined as the ratio of the amount of precipitate forming after cell disruption to the total amount of the expressed protein. The mutants with a precipitate ratio larger than 60% exhibited a smaller CD value than the wild type [18]. By setting 60% as a criterion value, the regions which do not allow the Ala-insertion to maintain the native structure called structure elements (SEs) were identified [18]. The resultant SEs are summarized in the introduction. To quantify the IRRC, we introduced the overlapped CV among residues in proteins. First, we calculated the CV of each DHFR residue along the amino acid sequence (Fig. 2A, solid line). On comparing these profiles, it was found that the peak positions were well-matched with each other, but there remained a large background in the CV profile. The contact map of the DHFR residues (Supplementary Fig. S3) shows that the adjacent three or more amino acid residues are in contact with each other, corresponding to diagonal lines colored red. Therefore, we assumed that the cause of the background were the contacts among the adjacent residues. The CV profile obtained after eliminating the adjacent contacts within the inter-residue distance ranging from 0 to 7 are shown in Supplementary Figure S4A–H. By increasing the inter-residue distance for removing the residues up to 3, the background in the original CV profile drops off, and the modified CV profiles become gradually more approximate to the precipitate ratio profile, where the peak heights and the background levels seem to be comparable. Conversely, by further increasing the inter-residue distance, some peaks of the CV profiles could no longer reproduce the precipitate ratio profile (shown by arrows in Supplementary Fig. S4E–H). Figure 2B shows the correlation coefficients between the precipitate ratio profile and the CV profiles when changing the inter-residue distance from 0 to 7. There is a large gap between the inter-residue distances of 3 and 4, in agreement with the gap between Supplementary Figure S4D and E.

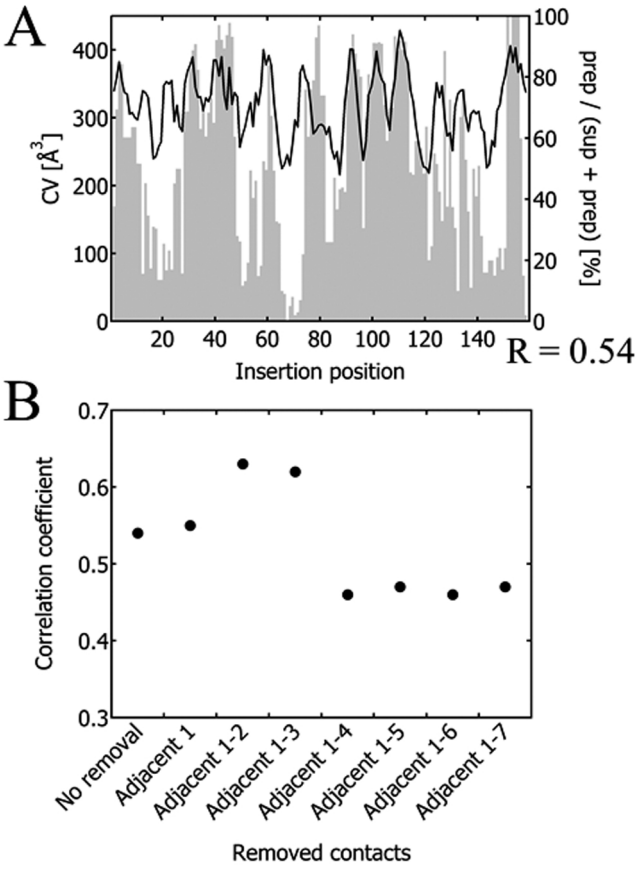

Figure 2 (A) The CV profile of DHFR (line) and the precipitate ratio profile of DHFR obtained using comprehensive Ala-insertion mutation analysis (bar) [18]. "Prep" and "sup" in the graph title are abbreviations for the protein in the precipitant and supernatant, respectively. The graph title of prep/(sup+prep) [%] indicates the ratio of the amount of precipitate to the total amount of the protein (precipitate ratio). (B) The correlation coefficient between the CV profile and the precipitate ratio profile after excluding the CV within the adjacent 0–7 residues.

After removing the contacts between three adjacent residues, the CV profile recapitulated the precipitate ratio

profile (Fig. 3A), except for around the 80th residue (indicated by asterisk). The discrepancy seen around the 80th residue is discussed below. The peak positions of both profiles aligned quite well and even the shape could be reproduced. For comparison, the CV profile calculated using vdW radius alone without considering the additional radius (1.4 Å) of a water molecule is shown in Figure 3B. The CV profile using vdW radius alone also reproduces the precipitate ratio profile's peak positions to some extent. However, when comparing the correlation coefficients and the intensities of some peaks, the addition of a water molecule radius reproduces the precipitate ratio profile with higher accuracy. This result suggests that the exclusion volume of water molecules is involved in the insertion effect. Concerning the regions forming vdW contacts, the two factors of vdW interaction and the exclusion volume should be considered. It is not easy to quantitatively discuss the effects of these two factors on protein stability. Therefore, we do not entirely deny the contribution of the vdW interaction. However, since the addition of water molecules improves the reproducibility of the precipitate ratio profile around profiles around the 35th and 100th residues (Fig. 3B, arrows), we concluded that it was better to add a water molecule radius to improve the prediction accuracy of pSE. These results suggest that meshing among the residues to exclude water molecules is associated with the destabilization due to Ala-insertion. It was also reported that molecular meshing to exclude solvent molecules plays an essential role in the formation of supramolecular complexes [38]. Eventually, the addition of the radius of a water molecule enables an evaluation of the molecular meshing effect, resulting in good alignment of the precipitate ratio profile and the CV profile.

ASA and rASA are familiar parameters to estimate the solvent accessibility of each residue [39]. The CV corresponds to the area buried inside the protein. We calculated relative inaccessible surface area (riASA=100–rASA) for the sake of easy comparison. The riASA is defined by the ratio of the inaccessible surface area of amino acid residues (iASA=stdASA–ASA) to stdASA, where stdASA refers the maximum possible solvent accessible surface area [39]. The obtained riASA and iASA profiles were compared with the precipitate ratio profile (Supplementary Fig. S5A and S5B). The peak positions of the riASA profile are consistent with those of the precipitate ratio profile. However, the riASA profile does not reproduce the intensity of the precipitate ratio peaks, unlike the concordance between the CV profile and precipitate ratio. In particular, the peaks around the 35th and 105th residues are so small that the peak positions cannot be determined (indicated by arrows in the figure). The correlation coefficient (0.59) between the riASA and precipitate ratio profiles is slightly lower than that between the CV and precipitate ratio profiles (0.62).

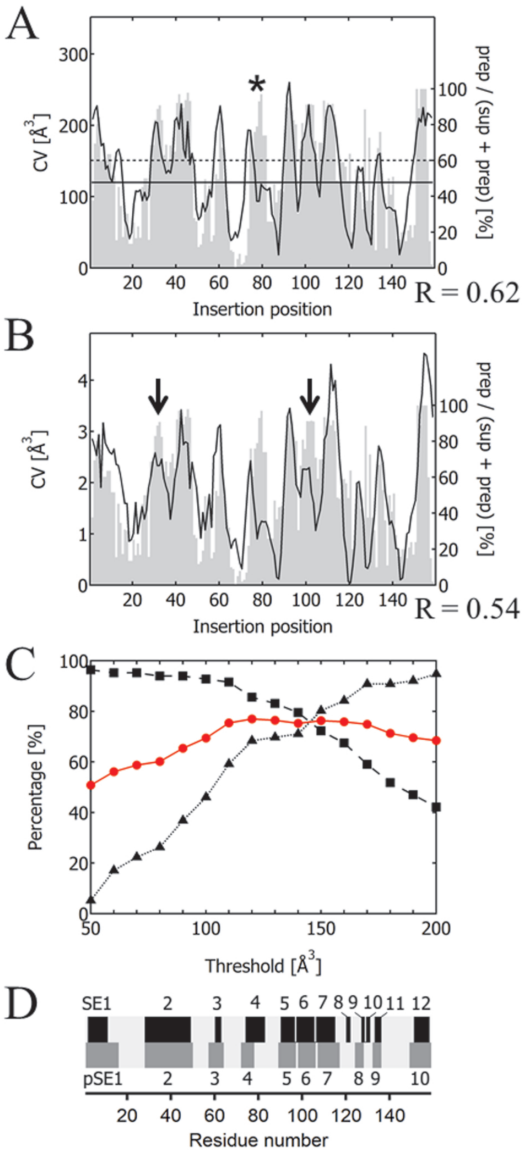

Figure 3 (A) The CV profile of DHFR after excluding the CV within the adjacent three residues (line) and the precipitate ratio profile of DHFR from the comprehensive Ala-insertion mutation analysis (bar) [18]. The correlation coefficient (R) is 0.62. Apparent discrepancy between the CV and the precipitate ratio profiles are indicated with an asterisk. "Prep" and "sup" indicate the amount of the protein in the precipitant and supernatant. The graph title of "prep/(sup+prep) [%]" indicates the ratio of the amount of precipitate to the total amount of the protein (precipitate ratio). The profiles are scaled to match the mean value. The precipitate ratio profile criterion to determine the structure elements (60%) and the CV profile criterion to predict the structure elements (120 A^3) are shown by the dashed and solid line, respectively. (B) The CV profiles when the mean radius of a water molecule is not considered (line), and the precipitate ratio profiles from comprehensive Ala-insertion mutation analysis (bar) [18]. The correlation coefficient (R) is 0.54. The profiles are scaled to match the mean value. (C) The proportion of the amino acid residues included in the structure elements (squares) and the linkers (triangles) that are correctly predicted using the indicated CV criteria (horizontal axis). The average proportion of these values is shown as circles. (D) The structure elements from the comprehensive Ala-insertion mutation analysis (top) and the predicted structure elements (bottom).

The values of riASA are normalized by stdASA of amino acid residues. In contrast, iASA indicates the absolute value of inaccessible surface area of amino acid residues (Supplementary Fig. S5B). The CV reflects the absolute volumes at an amino acid residue overlapped with other residues, indicating that the region may be buried in the protein. Therefore, iASA (total area buried) rather than riASA (ratio) would have a closer relationship to the CV. In fact, the iASA profile can reproduce the precipitate ratio profile peaks that could not be seen in the riASA profile (indicated by arrow in Supplementary Fig. S5B). However, the correlation coefficient (0.57) is smaller than that of the CV profile (0.62), indicating that the CV profile is a better indicator for predicting structure elements.

We also calculated the contact number profile of DHFR and compared it with the precipitate ratio profile (Supplementary Fig. S5C). The residue-residue contact map of DHFR was obtained using CMview [35] to calculate the contact number profile. Here, the residues were considered to contact each other when the distance between Cα atoms was shorter than 8 Å [36]. The contact number at an amino acid residue derived from the adjacent 1–3 residues was excluded, following the CV profile calculation method. Although many peaks of the contact number profile are well with those of the precipitate ratio profile, the peaks indicated by the arrow are not found to be reproduced. The correlation coefficient with the precipitate ratio profile (R=0.48) was also the lowest among the indices compared in this study.

All the indices discussed above are related to the residue-residue contacts in the protein and generally reproduce the precipitate ratio profile. This suggests that the intra-molecular contact significantly influence structure element determination. The parameter of iASA is a familiar index and reproduces the precipitate ratio profile to some extent. Therefore, it may be potentially used for the prediction of structure elements. However, it is difficult to calculate the individual contacts between the segments using iASA, as is evaluated in this study (see below). The correlation coefficient between the CV profile and the precipitate ratio profile is highest among these indexes. Thus, we can conclude that the CV is the most appropriate index for predicting and interpreting structure elements.

The SEs of DHFR were identified by setting 60% precipitate ratio as a criterion in a previous study [18] (Fig. 3A, dashed line). To evaluate the criterion used to predict SEs from the CV profile, we quantified the degree of the match when changing the criterion from 50 Å^3 to 200 Å^3 by 10 Å^3 (Fig. 3C). The prediction accuracy was assessed using the proportion of the SEs and the linkers covered by those predicted from the CV profile. With the increase in the criteria value, the proportions of the SEs (square) and the linkers (triangle) decreased and increased, respectively (Fig. 3C). The average of these proportions (circle) exhibits a maximum at 120 Å^3, where 86% of the amino acid residues comprising the SEs can be successfully predicted. The predicted structure elements (pSE) by setting the criterion of 120 Å^3 are as follows: M1-G15, L28-S49, R57-S63, V72-S77, P89-G96, R98-P105, A107-D116, H124-D127, D132-S135, and H149-R159 (Fig. 3D, bottom). For comparison, the SEs reported so far are also shown in the top row of Figure 3D. The segments extracted from the CV profile show good correspondence with the SEs.

However, upon careful comparison, it was found that pSE4 and pSE8 do not agree well with the SE. The structure around pSE4 is shown in Supplementary Figure S6A. The CV profile near pSE4 (indicated by asterisk in Fig. 3A) is composed of two peaks around T73 in β strand and I82 in α helix (colored by magenta in Supplementary Fig. S6A). The precipitate ratio profile peak is found in the turn structure between the β strand and the α helix (indicated by arrow in Supplementary Fig. S6A). It had been reported that the Ala-insertion mutation causes looping-out or shift, which propagates the structural distortion to the neighboring residues [40,41]. As mentioned above, the proper window size of four for averaging the CV profile supports that the structural perturbation reaches several neighboring amino acid residues due to the insertion. In the case of SE4, it can be considered that the Ala-insertion into the turn structure concomitantly influences the β strand and the α helix accompanied by relatively high CV values, which would trigger significant effects on the stability of the protein.

Supplementary Figure S6B shows the structure around pSE8. This region displays smaller precipitate ratio profile and the CV profile values than the other SEs and pSEs, which would make it difficult to predict SEs from the CV profile. SE8, SE9, and SE10 were determined by the increase in the precipitate of 120A121, 127A128, and 129A130. The neighboring mutants display less than 60% of the precipitate ratio. As mentioned above, comprehensive Ala-insertion mutation analysis utilizes structural perturbations, such as looping-out or shift due to the insertion. Therefore, it might be difficult to identify short structure elements such as SE8, SE9, and SE10, because the structural perturbation cannot be localized on a short segment accompanied by high CV.

Since the CV profile can predict the SE, it can be strongly proposed that one of the significant physico-chemical determinants of SEs is residue–residue contact restricting the penetration of water molecules. Next, we investigated the contacts among the SEs to reveal how SEs contribute to gain CV within the structure. We divided the sequence of DHFR into the segments of the pSEs and the segments connecting the pSEs (termed predicted linkers, pLN), and then calculated the total CV among the segments (Fig. 4A).

Figure 4A shows the CVs of a pair of contacts between

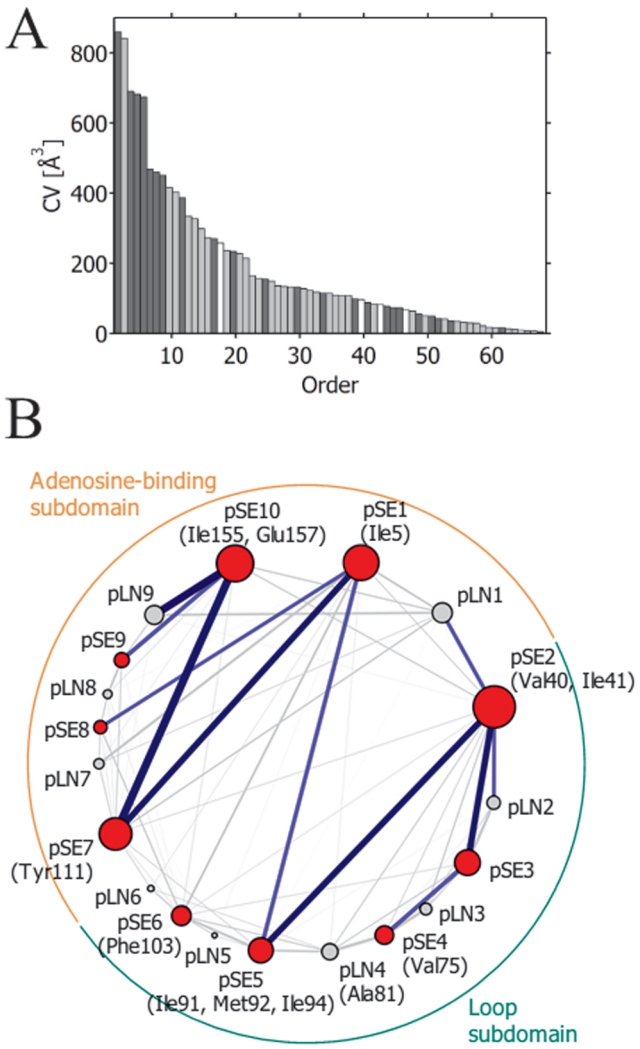

Figure 4 (A) The CV between pairs of the segments, shown in descending order from left to right. The contacts between structure elements, between a structure element and a linker, and between linkers are represented by dark gray, light gray, and white bars, respectively. (B) Contact network among the segments. The structure elements and linkers are shown as red and gray circles, respectively. The radii of these circles reflect the relative size of the total CV on the segment. Structure elements and linkers are named 'pSE1, pSE2, ...' and 'pLN1, pLN2, ...' from the N-terminal side, respectively. The contacts between the segments are indicated by lines. The thickness of the line is proportional to the CV formed between the segments. The top 11 contacts are colored, where the top five accompanied by a remarkably large CV are dark blue, while the rest are light blue. The loop subdomain and the adenosine-binding subdomain are shown as green and orange, respectively. The residues involved in the structure formation at the early stage of the folding are noted near the segment containing the residue.

SEs (dark gray), between a SE and a linker (light gray), and between linkers (white) in descending order. It can be observed that the contacts between SEs tend to possess a more extensive CV than the other kinds of contacts. In particular, the top five contacts have an extensive CV compared to the others. The network of the contacts among segments is shown in Figure 4B. The pSE and pLN are represented by red and gray circles, respectively. The size of the circles is proportional to the CV of each segment. As expected, pSEs exhibit a larger CV than pLNs. The thickness of the lines connecting the segments reflects the CV between segments. The top 11 contacts are colored blue, where the five contacts with distinctively large CVs are emphasized by dark blue. Interestingly, all pSEs except for pSE6 are involved in the top 11 inter-pSE contacts. These results indicate that pSEs selectively form contacts with each other; thus, individual pSEs gain a relatively large CV.

DHFR consists of two subdomains, the loop subdomain and the adenosine-binding subdomain [42] (Fig. 4B). The five inter-pSE contacts with the larger CV (dark blue line) are divided into two networks (pSE1-pSE7-pSE10 and pSE3-pSE2-pSE5) within each domain. Jones and Matthews (1995) have proposed two hydrophobic clusters allocated in each subdomain at the early stages of folding and revealed essential residues responsible for the structure formation [43], including Ile5, Tyr111, Ile155, and Glu157 in the loop subdomain and Val40, Ile41, Val75, Ala81, Ile91, Met92, Ile94, and Phe103 in the adenosine-binding subdomain. These residues are included in FEs determined using the circular permutation analysis, providing the rationale that FEs are indispensable regions for protein folding [17]. Furthermore, Arai *et al.* reported that FE1, 2, 7, and 10 (including the above hydrophobic residues) were shown to coalesce with each other during the early stage of the folding reaction [44,45]. Our study revealed that the contacts between pSE1-pSE7-pSE10 and pSE3-pSE2-pSE5 exhibited a more extensive CV than the other inter-pSE contacts. Interestingly, most of the residues responsible for the protein folding were found in these networks (see Fig. 4B). The networks of pSE1-pSE7-pSE10 and pSE3-pSE2-pSE5 match well with those of FE1-FE7-FE10 and FE2, respectively. These consistencies imply that the analysis of the inter pSE CV also enable us to predict the regions responsible for the protein folding reaction, where pSEs with large pSE-pSE CV indicate regions responsible for the early folding stage.

Conclusions

In this study, we revealed that the quantitative evaluation of CV among the residues could accurately reproduce the degree of structural destabilization due to the Ala-insertion mutation. Therefore, we can predict the SEs by using the CV profile without performing time-consuming experiments. A previous study showed that segments composed of consecutive amino acid residues responsible for the structure formation could be identified using methods such as comprehensive Ala-insertion mutation

analysis or circular permutation analysis [15–18]. However, the physicochemical factors involved in destabilization due to mutations remain unclear. Since CV can quantitatively reproduce the degree of the destabilization, it is proposed that CV is one of the determinants for the protein's stability. Specifically, it can be considered that the mutation of the SEs would disrupt the local structure, causing extensive loss of the intramolecular contact (or CV), enough to unfold. Furthermore, the more extensive CV selectively found in the pairs of the pSEs suggests that SEs enhance the protein's stability through molecular meshing.

Figure 5 summarizes the pSE-contact network in the DHFR structure. The surface represents each pSE at the vdW radius plus the approximate radius of a water molecule (1.4 Å). The two networks (pSE1-pSE7-pSE10 in the loop subdomain and pSE3-pSE2-pSE5 in the adenosine-binding subdomain) accompanying the top five inter-pSE CVs are drawn in the top. The pSEs (pSE8, pSE9 in the loop subdomain, pSE4 in the adenosine-binding subdomain) with the subsequent inter-pSE CVs are shown at the bottom. In Figure 5, we can easily assess that each pSE gains a large CV arising from their approximate meshing with each other, resulting in the cluster formation of the primary groups within each subdomain. Subsequently, the pSEs in the secondary group are attached to these clusters via the pSE-pSE meshing, and the two networks are connected through the pSE1-pSE5 to form the whole molecular packing. The structural aspect reminds us that pSEs behave like parts of Lego blocks, and the meshing of pSEs mediates the building up of the block to form the definitive structure. If this scheme is correct, protein structure can be designed by meshing pairs of the SEs and then combining them to construct the desired structure. In the future, we will use CV to predict the SEs of many reported proteins and build a database of the patterns of block combinations to propose a new method to design protein structures.

Acknowledgements

This research was inspired by a series of studies on the comprehensive Ala-insertion mutation analysis conducted by Professor Mikio Kataoka. We are deeply indebted to him for fruitful discussions during this research. This work was partly supported by Grants-in-Aid from the JSPS for Scientific Research [Category S, No. JP17H06165 (H.K.); Innovative Areas, No. JP25102003 (H.K.)].

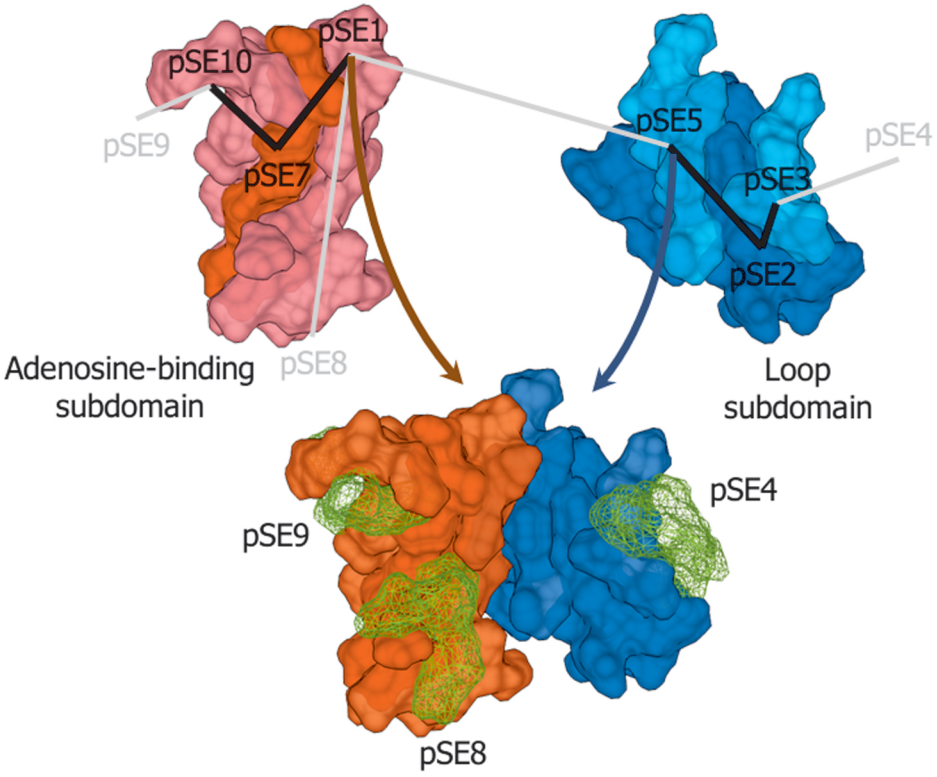

Figure 5 The pSE-pSE contacts within the top 11 inter-segment CV are drawn in the structure of DHFR. pSEs involved in the inter-pSE contacts with the top five (pSE1-pSE7-pSE10, and pSE3-pSE2-pSE5) and the subsequent (pSE4, pSE8, pSE9) inter-pSE CV are shown at the top and the bottom, respectively.

Conflicts of Interest

All authors declare that they have no conflict of interest.

Author Contributions

Y.T. and H.K. conceived the project and designed the research. Y.T., Y.Y., Y.H., S.T.F., and H.K. performed the research. Y.T. and H.K. wrote the paper. All authors approved the final version of the manuscript.

References

[1] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000). DOI: 10.1093/nar/28.1.235

[2] El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., *et al.* The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**, D427–D432 (2019). DOI: 10.1093/nar/gky995

[3] Koga, N., Tatsumi-Koga, R., Liu, G., Xiao, R., Acton, T. B., Montelione, G. T., *et al.* Principles for designing ideal protein structures. *Nature* **491**, 222–227 (2012). DOI: 10.1038/nature11600

[4] Huang, P.-S., Oberdorfer, G., Xu, C., Pei, X. Y., Nannenga, B. L., Rogers, J. M., *et al.* High thermodynamic stability of parametrically designed helical bundles. *Science* **346**, 481–485 (2014). DOI: 10.1126/science.1257481

[5] Lin, Y.-R., Koga, N., Tatsumi-Koga, R., Liu, G., Clouser, A. F., Montelione, G. T., *et al.* Control over overall shape and size in de novo designed proteins. *Proc. Natl. Acad. Sci. USA* **112**, E5478–E5485 (2015). DOI: 10.1073/pnas.1509508112

[6] Huang, P.-S., Boyken, S. E. & Baker, D. The coming of age of de novo protein design. *Nature* **537**, 320–327 (2016). DOI: 10.1038/nature19946

[7] Marcos, E., Basanta, B., Chidyausiku, T. M., Tang, Y., Oberdorfer, G., Liu, G., *et al.* Principles for designing proteins with cavities formed by curved β sheets. *Science* **355**, 201–206 (2017). DOI: 10.1126/science.aah7389

[8] Dou, J., Vorobieva, A. A., Sheffler, W., Doyle, L. A., Park, H., Bick, M. J., *et al.* De novo design of a fluorescence-activating β-barrel. *Nature* **561**, 485–491 (2018). DOI: 10.1038/s41586-018-0509-0

[9] Marcos, E., Chidyausiku, T. M., McShan, A. C., Evangelidis, T., Nerli, S., Carter, L., *et al.* De novo design of a non-local β-sheet protein with high stability and accuracy. *Nat. Struct. Mol. Biol.* **25**, 1028–1034 (2018). DOI: 10.1038/s41594-018-0141-6

[10] Silva, D.-A., Yu, S., Ulge, U. Y., Spangler, J. B., Jude, K. M., Labão-Almeida, C., *et al.* De novo design of potent and selective mimics of IL-2 and IL-15. *Nature* **565**, 186–191 (2019). DOI: 10.1038/s41586-018-0830-7

[11] Gō, M. Correlation of DNA exonic regions with protein structural units in haemoglobin. *Nature* **291**, 90–92 (1981). DOI: 10.1038/291090a0

[12] Panchenko, A. R., Luthey-Schulten, Z. & Wolynes, P. G. Foldons, protein structural modules, and exons. *Proc. Natl. Acad. Sci. USA* **93**, 2008–2013 (1996). DOI: 10.1073/pnas.93.5.2008

[13] Shiba, K. & Shimmel, P. Functional assembly of a randomly cleaved protein. *Proc. Natl. Acad. Sci. USA* **89**, 1880–1884 (1992). DOI:10.1073/pnas.89.5.1880

[14] Hiraga, K., Yamagishi, A. & Oshima, T. Mapping of unit boundaries of a protein: exhaustive search for permissive sites for duplication by complementation analysis of random fragment libraries of tryptophan synthase α subunit. *J. Mol. Biol.* **335**, 1093–1104 (2004). DOI:10.1016/j.jmb.2003.11.029

[15] Iwakura, M. & Nakamura, T. Effects of the length of a glycine linker connecting the N-and C-termini of a circularly permuted dihydrofolate reductase. *Protein Eng.* **11**, 707–713 (1998). DOI: 10.1093/protein/11.8.707

[16] Nakamura, T. & Iwakura, M. Circular permutation analysis as a method for distinction of functional elements in the M20 loop of Escherichia coli dihydrofolate reductase. *J. Biol. Chem.* **274**, 19041–19047 (1999). DOI: 10.1074/jbc.274.27.19041

[17] Iwakura, M., Nakamura, T., Yamane, C. & Maki, K. Systematic circular permutation of an entire protein reveals essential folding elements. *Nat. Struct. Biol.* **7**, 580–585 (2000). DOI: 10.1038/76811

[18] Shiba, R., Umeyama, M., Tsukasa, S., Kamikubo, H., Yamazaki, Y., Yamaguchi, M., *et al.* Systematic alanine insertion reveals the essential regions that encode structure formation and activity of dihydrofolate reductase. *Biophysics* **7**, 1–10 (2011). DOI: 10.2142/biophysics.7.1

[19] Barah, P. & Sinha, S. Analysis of protein folds using protein contact networks. *Pramana* **71**, 369–378 (2008). DOI: 10.1007/s12043-008-0170-5

[20] Hu, G., Yan, W., Zhou, J. & Shen, B. Residue interaction network analysis of Dronpa and a DNA clamp. *J. Theor. Biol.* **348**, 55–64 (2014). DOI: 10.1016/j.jtbi.2014.01.023

[21] Di Paola, L. & Giuliani, A. Protein contact network topology: A natural language for allostery. *Curr. Opin. Struct. Biol.* **31**, 43–48 (2015). DOI: 10.1016/j.sbi.2015.03.001

[22] Li, W.-Y., Wei, H.-Y, Sun, Y.-Z., Zhou, H., Ma, Y. & Wang, R.-L. Exploring the effect of E76K mutation on SHP2 cause gain-of-function activity by a molecular dynamics study. *J. Cell. Biochem.* **119**, 9941–9956 (2018). DOI: 10.1002/jcb.27316

[23] Amala, A. & Emerson, I. A. Understanding contact patterns of protein structures from protein contact map and investigation of unique patterns in the globin-like folded domains. *J. Cell. Biochem.* **120**, 9877–9886 (2019). DOI: 10.1002/jcb.28270

[24] Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., *et al.* UCSF Chimera—A visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004). DOI: 10.1002/jcc.20084

[25] Chothia, C. The nature of the accessible and buried surfaces in proteins. *J. Mol. Biol.* **105**, 1–12 (1976). DOI: 10.1016/0022-2836(76)90191-1

[26] Nicholls, A., Sharp, K. A. & Honig, B. Protein folding and association: Insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins* **11**, 281–296 (1991). DOI: 10.1002/prot.340110407

[27] Eyal, E., Najmanovich, R., Mcconkey, B. J., Edelman, M. & Sobolev, V. Importance of Solvent Accessibility and Contact Surfaces in Modeling Side-Chain Conformations in Proteins. *J. Comput. Chem.* **25**, 712–724 (2004). DOI: 10.1002/jcc.10420

[28] Chen, J., Sawyer, N. & Regan, L. Protein-protein interactions: General trends in the relationship between binding affinity and interfacial buried surface area. *Protein*

Sci. **22**, 510–515 (2013). DOI: 10.1002/pro.2230
[29] Metropolis, N. & Ulam, S. The Monte Carlo Method. *J. Am. Stat. Assoc.* **44**, 335–341 (1949). DOI: 10.1080/01621459.1949.10483310
[30] Voigt, C. A., Martinez, C., Wang, Z.-G., Mayo, S. L. & Arnold, F. H. Protein building blocks preserved by recombination. *Nat. Struct. Biol.* **9**, 553–558 (2002). DOI: 10.1038/nsb805
[31] Dyson, H. J., Wright, P. E. & Scheraga, H. A. The role of hydrophobic interactions in initiation and propagation of protein folding. *Proc. Natl. Acad. Sci. USA* **103**, 13057–13061 (2006). DOI: 10.1073/pnas.0605504103
[32] Webb, B. & Sali, A. Comparative protein structure modeling using MODELLER. *Curr. Protoc. Bioinformatics* **47**, 1–32 (2014). DOI: 10.1002/0471250953.bi0506s47
[33] Hubbard, S. J. & Thornton, J. M. *'NACCESS', Computer Program* (Department of Biochemistry and Molecular Biology, University College, London, 1993).
[34] Singh, A., Kaushik, R., Mishra, A., Shanker, A. & Jayaram, B. ProTSAV: A protein tertiary structure analysis and validation server. *Biochim. Biophys. Acta* **1864**, 11–19 (2016). DOI: 10.1016/j.bbapap.2015.10.004
[35] Vehlow, C., Stehr, H., Winkelmann, M., Duarte, M. J., Petzold, L., Dinse, J., *et al.* CMView: Interactive contact map visualization and analysis. *Bioinformatics* **27**, 1573–1574 (2011). DOI: 10.1093/bioinformatics/btr163
[36] Jinag, Z., Zhang, L., Chen, J., Xia, A. & Zhao, D. Effect of amino acid on forming residue–residue contacts in proteins. *Polymer* **43**, 6037–6047 (2002). DOI: 10.1016/S0032-3861(02)00501-3
[37] Bastian, M., Heymann, S. & Jacomy, M. Gephi: an open source software for exploring and manipulating networks. In *Third international AAAI conference on weblogs and social media* (2009).
[38] Zhan, Y.-Y., Ogata, K., Kojima, T., Koide, T., Ishii, K., Mashiko, T., *et al.* Hyperthermostable cube-shaped assembly in water. *Commun. Chem.* **1**, 14 (2018). DOI: 10.1038/s42004-018-0014-2
[39] Lee, B. & Richards, F. M. The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **55**, 379–400 (1971). DOI: 10.1016/0022-2836(71)90324-x
[40] Heinz, D. W., Baase, W. A., Dahlquist, F. W. & Matthews, B. W. How amino-acid insertions are allowed in an α-helix of T4 lysozyme. *Nature* **361**, 561–564 (1993). DOI: 10.1038/361561a0
[41] Vetter, I. R., Baase, W. A., Heinz, D. W., Xiong, J. P., Snow, S. & Matthews, B. W. Protein structural plasticity exemplified by insertion and deletion mutants in T4 lysozyme. *Protein Sci.* **5**, 2399–2415 (1996). DOI: 10.1002/pro.5560051203
[42] Sawaya, M. R. & Kraut, J. Loop and subdomain movements in the mechanism of Escherichia coli dihydrofolate reductase: Crystallographic evidence. *Biochemistry* **36**, 586–603 (1997). DOI: 10.1021/bi962337c
[43] Jones, B. E. & Matthews, C. R. Early intermediates in the folding of dihydrofolate reductase from Escherichia coli detected by hydrogen exchange and NMR. *Protein Sci.* **4**, 167–177 (1995). DOI: 10.1002/pro.5560040204
[44] Arai, M., Maki, K., Takahashi, H. & Iwakura, M. Testing the relationship between foldability and the early folding events of dihydrofolate reductase from Escherichia coli. *J. Mol. Biol.* **328**, 273–288 (2003). DOI: 10.1016/S0022-2836(03)00212-2
[45] Arai, M. & Iwakura, M. Probing the interactions between the folding elements early in the folding of Escherichia coli dihydrofolate reductase by systematic sequence serturbation analysis. *J. Mol. Biol.* **347**, 337–353 (2005). DOI: 10.1016/j.jmb.2005.01.033

(*Edited by Haruki Nakamura*)