BMC Molecular and Cell Biology

RESEARCH

Open Access

Conserved structural topologies in RNase A-like and trypsin-like serine proteases: a sequence-based folding analysis

K. M. Ahsanul Kabir[1], Takuya Takahashi[2] and Takeshi Kikuchi[2*]

Abstract

Background Protein folding is a complex process in which amino acid sequences encode the information required for a polypeptide chain to fold into its functional three-dimensional (3D) structure. Many proteins share common substructures and recurring secondary structure elements that contribute to similar 3D folding patterns, even across different protein families. This study examines two distinct groups of proteins, the RNase A-like fold and the trypsin-like serine protease fold, classified by SCOPe. These proteins share only some substructures that contribute to their folding cores. Despite minimal sequence identity, they exhibit partial structural similarities in their 3D topologies. We used a sequence-based approach, including inter-residue average distance statistics and contact frequency prediction, to explore these folding characteristics. Structural observations guided further analyses of conserved hydrophobic residue packing, highlighting key folding units within each fold.

Results Our analysis predicted two compact regions within each protein group. Interactions between these regions form a partially shared topology. We identified conserved hydrophobic residues critical to these interactions, suggesting a common mechanism for establishing these structural features. Despite overall structural differences between the RNase A-like and trypsin-like folds, our findings emphasize the presence of a shared partial folding core.

Conclusions The partially shared structural features in the RNase A-like and trypsin-like serine protease folds reflect a convergent folding mechanism. This mechanism underscores the evolutionary adaptation of protein folding, where distinct folds can still retain critical, conserved structural motifs. These findings highlight how proteins with overall different topologies can evolve to share key folding features, demonstrating the elegance and efficiency of protein evolution.

Keywords Protein folding, Ribonuclease, Chymotrypsin, Average distance map, F-value

*Correspondence:
Takeshi Kikuchi
tkikuchi@sk.ritsumei.ac.jp
[1]Research Organization of Science and Technology, Ritsumeikan University, 1-1-1 Nojihigashi, Kusatsu-City, Shiga, Japan
[2]Department of Bioinformatics, College of Life Sciences, Ritsumeikan University, 1-1-1 Nojihigashi, Kusatsu-City, Shiga, Japan

Background

Proteins are fundamental to a wide array of biological functions, and for each function, the protein must adopt a highly specific three-dimensional (3D) structure [1]. Over the course of billions of years, these complex structures have evolved to facilitate essential cellular activities and offer valuable insights for drug discovery and therapeutic interventions [2]. The primary sequence of amino acids in a protein contains the instructions for its final folded form. Despite recent breakthroughs in artificial intelligence (AI) for predicting protein structures from sequences [3, 4], it is still a significant and difficult problem to determine accurately how a sequence folds into a functional 3D form, as AI tools focus on final structures but do not reveal the folding pathways. Understanding these pathways is critical for elucidating folding mechanisms and addressing disorders related to protein misfolding. Therefore, complementary sequence-based methods are essential to uncover these dynamic folding processes, which remain beyond the scope of current AI approaches [5].

The prediction of protein fold is fundamentally different from understanding the protein folding problem and the biophysics of folding pathways. While protein folding studies focus on the physical principles that guide a polypeptide to its native structure, fold prediction deals with identifying recurring structural motifs across diverse proteins [6]. Many existing AI-driven approaches have advanced fold prediction; however, the evolutionary mechanisms that give rise to protein folds remain largely unexplored.

Protein folding is the intricate process by which a linear chain of amino acids transforms into a functional three-dimensional structure. While the idea that a protein's sequence determines its final shape, known as Anfinsen's dogma, has long been understood, the actual mechanisms behind this transformation have remained elusive [7, 8]. Given the enormous number of potential conformations, pinpointing the correct native structure without a clear guide is nearly impossible. The key to this folding puzzle lies in the specific interactions encoded within the amino acid sequence, which drive the protein to adopt its final shape [9].

Recent advancements have significantly enhanced our understanding of folding pathways through innovative experimental techniques and enhanced computational methods [10, 11]. Nevertheless, extracting folding information directly from amino acid sequences using conventional bioinformatics tools remains a significant challenge. Super-secondary structures, such as the Rossmann fold and four-helix bundle, are often indicative of potential folding sites within proteins. In this study, we identify a newly observed common structural motif in ribonuclease A from *Bos taurus* (PDB code: 6ETL) and a-chymotrypsin from *Bos taurus* (PDB code: 6CHA), despite their classification into different SCOPe (https://scop.berkeley.edu/) classification—RNase A-like and try psin-like serine proteases (Fig. 1). Our study sheds light on the formation and conservation of common structural topologies across the RNase A-like and trypsin-like serine protease folds. Understanding these conserved features is essential, as it provides insight into the fundamental principles of protein evolution and stability.

Ribonuclease is an enzyme that catalyzes the breakdown of RNA into smaller components, playing a crucial role in RNA metabolism and regulation. In contrast, chymotrypsin is a serine protease that hydrolyzes peptide bonds, specifically targeting the carboxyl side of large hydrophobic amino acids such as phenylalanine, tryptophan, and tyrosine. Despite their functional differences, both enzymes share a common structural motif characterized by specific β strands, indicating a shared folding topology beyond their SCOP classifications. These functional differences and their structural similarities make them compelling models for studying conserved folding topologies.

Careful examination of the 3D structures reveals the common topology constituted by b1, b4, b5, b6 and b7 in ribonuclease A and b1, b4, b5 and b6 in a-chymotrypsin in Fig. 1(A) and (C). As shown in Fig. 1, the b strands constituting the common topology and intervening b strands are named as ba, bb, bc, bd, be, bf (or be-f) and bg (According to the PDB annotation, b5 and b6, that is, be and bf in ribonuclease (PDB code: 6ETL) corresponds to b5, that is, be-f in a-chymotrypsin domain B (PDB code: 6CHA)).

This structural motif, involving specific β strands, suggests a shared folding topology that transcends their SCOP classifications. The common topology looks like a β hairpin is linked to a two-stranded b-sheet, as seen in Fig. 1(B) and (D). RNase A, a well-characterized enzyme consisting of 124 amino acid residues [12, 13], and α-chymotrypsin, a proteolytic enzyme with a 131-residue domain B [14, 15], provide ideal models for this investigation. Their sequences and secondary structures are illustrated in Fig. 2. It is a reasonably noteworthy problem whether this structural similarity is just by chance or by a folding mechanism common to these proteins. In other words, we consider whether a common folding mechanism exists beyond the SCOP classification category. Such a common partial topology might be a kind of super-secondary structure.

Our study aims to extract detailed folding information from the sequences of ribonuclease and chymotrypsin using inter-residue average distance statistics (ADM) and evolutionary analyses. Given their low sequence identity (12.38%), traditional bioinformatics approaches, such as sequence alignment, are insufficient for studying their

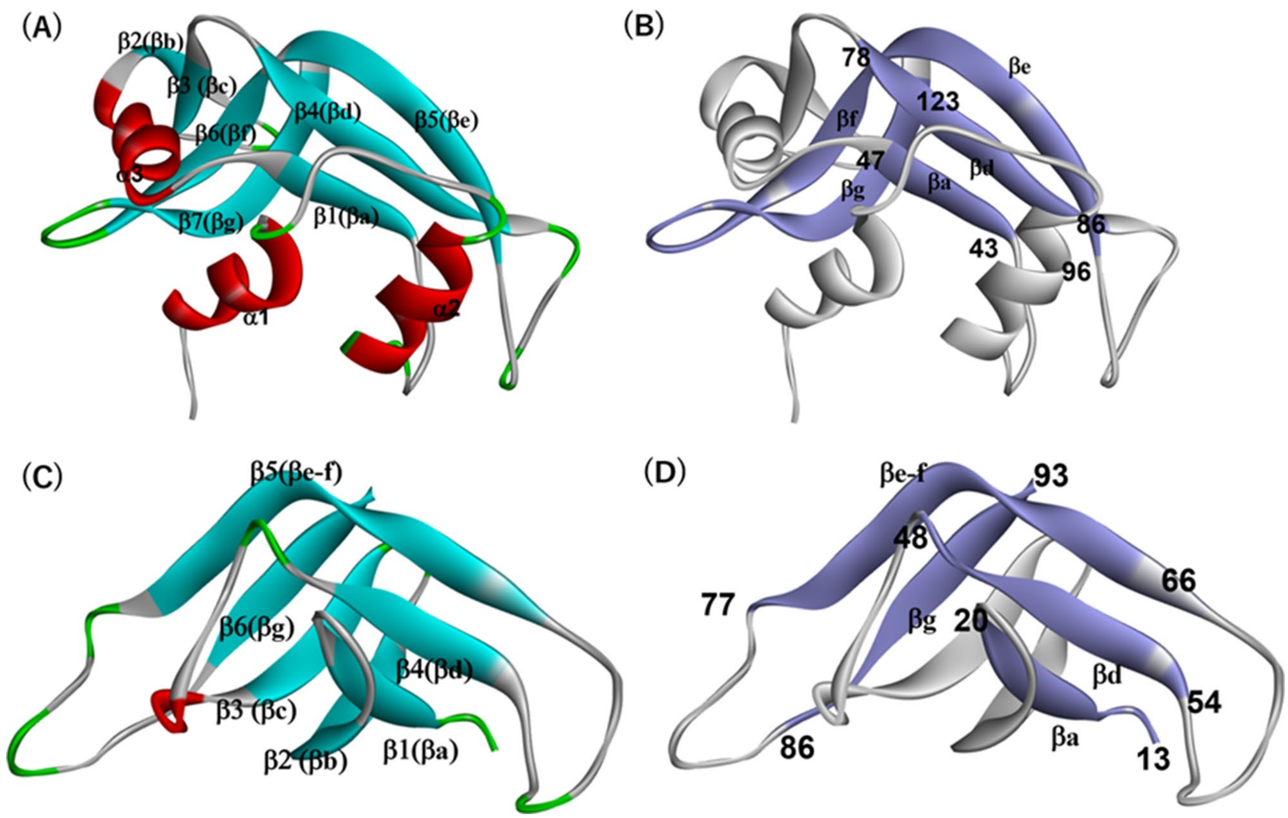

Fig. 1 (**A**) The 3D structure and secondary structures of ribonuclease A from *Bos taurus* (PDB code: 6ETL). (**B**) Partially common topology observed in ribonuclease A from *Bos taurus* observed in proteins of RNase A-like fold and trypsin-like serine proteases fold in SCOPe classification. The part corresponding to the partially common topology is colored by purple. (**C**) The 3D structure and secondary structures of a-chymotrypsin from *Bos taurus* domain B (PDB code: 6CHA). Only the structure of the core part (not the whole sequence) is shown (Fig. 2 can be referred to the entire sequence of 6CHA domain B). (**D**) Partially common topology observed in a-chymotrypsin from *Bos taurus* domain B in proteins of RNase A-like fold and trypsin-like serine proteases fold in SCOPe classification. The part corresponding to the partially common topology is colored by purple. A different number may be used for each b-strand according to PDB annotations. Thus, we also used a common name of each b-strand for all proteins as ba-bg

folding properties. For example, a sequence alignment technique provides a structural and/or functional unit in a protein but does not give information on an initial folding unit along the sequence of a protein. We focus on identifying key residues that contribute to the folding nucleus and examining the evolutionary pressures on these nucleation residues.

Employing predicted contact maps based on ADM, we have previously validated our methods by successfully extracting folding properties consistent with experimental data for various protein families [16–20]. We have confirmed in previous studies that our prediction methods successfully extract the folding properties that well to the data from experimental analyses (H/D exchange experiment of NMR technique, ϕ value analysis and so on) of the following proteins: fatty acid binding proteins and globin-like fold proteins [18], IgG binding and albumin binding domains [16], Globin E-to-H helix units [21], Ig-like fold proteins [22], ferredoxin-like fold proteins [23], beta-trefoil fold proteins [24, 25], and lysozyme-like superfamily proteins [26]. Applying these techniques to ribonuclease and chymotrypsin, we aim to elucidate their folding mechanisms and common topologies from their sequences. In particular, ADM predicted regions correspond well to structured regions in early stage of folding detected by the NMR H/D exchange experiments in the folding of myoglobin and leghemoglobin [27].

Materials and methods

Targeted proteins

This study focuses on two distinct classes of proteins as categorized by the SCOPe classification system: the RNase A-like fold and the trypsin-like serine proteases fold. For the RNase A-like fold, we selected three proteins: ribonuclease A from *Bos taurus* (PDB code: 6ETL), RNase ZF-3E from *Danio rerio* (PDB code: 2VQ9), and turtle egg white ribonuclease from *Chelonia mydas* (PDB code: 2ZPO). From the trypsin-like serine proteases fold, we selected the domain B of α-chymotrypsin from *Bos taurus* (PDB code: 6CHA_B), α1-tryptase from *Homo sapiens* (PDB code: 1LTO), and prostasin from *Homo sapiens* (PDB code: 3DFJ). These six proteins were

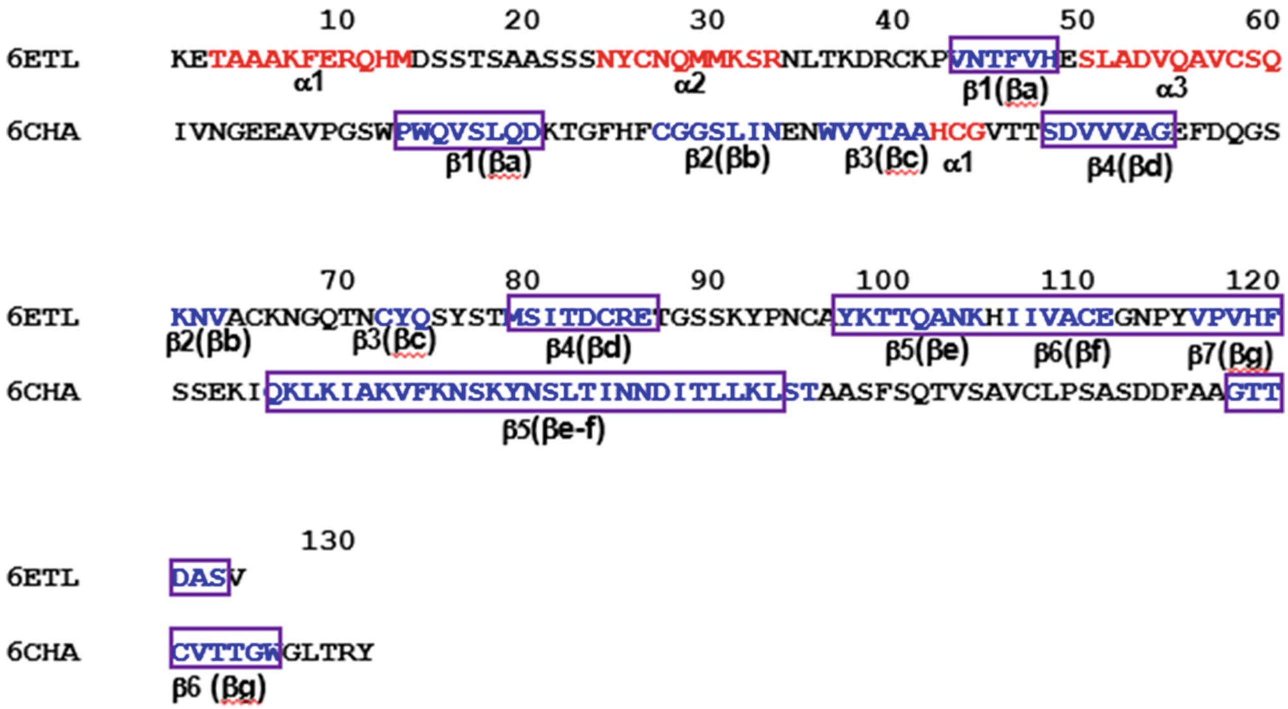
10
20
30
40
50
60
6ETL
KETAAAKFERQHMDSSTSAASSSNYCNQMMKSRNLTKDRCKPVNTFVHESLADVQAVCSQ
α1
α2
β1(βa)
α3
6CHA
IVNGEEAVPGSWPWQVSLQDKTGFHFCGGSLINENWVVTAAHCGVTTSDVVVAGEFDQGS
β1(βa)
β2(βb)
β3(βc)
α1
β4(βd)
70
80
90
100
110
120
6ETL
KNVACKNGQTNCYQSYSTMSITDCRETGSSKYPNCAYKTTQANKHIIVACEGNPYVPVHF
β2(βb)
β3(βc)
β4(βd)
β5(βe)
β6(βf)
β7(βg)
6CHA
SSEKIQKLKIAKVFKNSKYNSLTINNDITLLKLSTAASFSQTVSAVCLPSASDDFAAGTT
β5(βe-f)
130
6ETL
DASV
6CHA
CVTTGWGLTRY
β6 (βg)

Fig. 2 Sequences of trypsin-like serine proteases fold (PDB code: 6ETL), and a-chymotrypsin from *Bos taurus* domain B (PDB code: 6CHA). An a-helix or a b-strand is indicated by red or blue characters, respectively. A part constituting the common topology is enclosed by a purple rectangle

Table 1 Pairwise comparisons of sequence identity (SI, %), root-mean-square deviation (R, Å), and TM-scores [28] (TM) for proteins within and between the RNase A-like and Trypsin-like folds

		RNase A-like fold			Trypsin-like serine proteases fold		
PDBID		**6ETL**	**2VQ9**	**2ZPO**	**6CHA**	**1LTO**	**3DFJ**
6ETL	SI	100					
	R	0.0					
	TM	1.0					
2VQ9	SI	25.83	100				
	R	2.02	0.0				
	TM	0.81	1.0				
2ZPO	SI	39.83	37.29	100			
	R	1.81	1.89	0.0			
	TM	0.89	0.82	1.0			
6CHA	SI	**12.38**	**9.61**	**11.88**	100		
	R	**4.22**	**4.68**	**4.81**	0.0		
	TM	**0.30**	**0.31**	**0.27**	1.0		
1LTO	SI	**12.06**	**9.17**	**6.60**	37.27	100	
	R	**4.48**	**4.91**	**4.56**	1.62	0.0	
	TM	**0.22**	**0.20**	**0.22**	0.50	1.0	
3DFJ	SI	**11.21**	**10.38**	**13.59**	33.33	37.50	100
	R	**4.64**	**4.74**	**4.50**	1.25	1.78	0.0
	TM	**0.23**	**0.26**	**0.22**	0.50	0.90	1.0

Bold values denote the values for a protein pair from both folds

chosen to compare their folding properties derived from their amino acid sequences, despite belonging to different structural classes.

The sequence identity within each protein fold class is relatively low, making it challenging to study their folding properties using standard bioinformatics techniques. As indicated in Table 1, the maximum sequence identity within the trypsin-like serine proteases fold is 37%, while the minimum is 33%. For the RNase A-like fold, the maximum sequence identity is 39%, and the minimum is 25%. The sequence identity between the two different folds is even lower, ranging from 12% to as little as 6% (bold numbers in Table 1). Despite these low sequence identities, the proteins within each class share a similar

3D topology, characterized by a β-hairpin linked to a two-stranded β-sheet. In RNase A-like fold proteins, this topology is formed by the βa, βd, βe, βf, and βg, while in trypsin-like serine proteases fold proteins, it is formed by βa, βd, βe-f, and βg.

We also present the RMSD values and TM scores [28] between these proteins in Table 1. As the 3D structures in the trypsin-like serine proteases fold proteins for the calculations of these values are used in the 3D structures presented in Figs. 1 and S2. Within the same fold, comparisons consistently show higher TM scores (≥ 0.80) and lower RMSD values (≤ 2.02 Å), reflecting structural similarity within each fold. In contrast, different fold comparisons yield significantly lower TM-scores (≤ 0.31) and higher RMSD values (≥ 4.22 Å), indicating substantial structural divergence between the RNase A-like and Trypsin-like folds.

The common structural regions among these proteins are illustrated in Figs. S1-S2, and the specific locations of the segments contributing to these common structures are highlighted in Fig. S3. (The number of figures becomes too large in this study. Therefore, we put only the figures to explain the basis of the present study and those related to the results of ribonuclease A from *Bos taurus* (PDB code: 6ETL) and a-chymotrypsin from *Bos taurus* (PDB code: 6CHA) in the main text. The rest of the figures is put in the supplementary material.)

All protein structures and associated figures were generated using the software Discovery Studio V19.1.0.18287 (Dassaul Systémes Biovia Corp).

Average distance statistics map (ADM)

The average distance map (ADM) is constructed using a protein's amino acid sequence information. ADM is a predicted contact map that is made in the following way. We present each step of the method.

Definition of range along a protein sequence and computation of the interresidue average distances in each range

The inter-residue average distances and standard deviations were calculated using known 3D structures considering the amino acid types and sequence separation. Known structures of 42 proteins are used to calculate the average distances between Cα atoms of residues in each range [29]. The range is defined as the separation of two residues along the sequence of a protein. Symbolizing a range as M, M = 1 when $1 \leq k \leq 8$, where $k = |i - j|$, and i and j are the residue numbers in the sequence; M = 2 when $9 \leq k \leq 20$; M = 3 when $21 \leq k \leq 30$; M = 4 when $31 \leq k \leq 40$, and so on. For a protein with an unknown 3D structure, ADM is constructed by making a plot on a map when the average distance of a pair of residues is less than a cutoff value determined in advance [29].

Determination of cutoff values of each interresidue average distance in each range to make a kind of predicted contact map

A cutoff value is defined in each range so that the contact density of an entire contact map constructed based on the 3D structure of a protein is reproduced. The detailed procedure is as follows;

An average distance, d(A, B, M), where A and B are mentioned to amino acid types, of every residue pair in each range M was calculated. The cutoff distance for each range M can be defined in the following Eq.

$$\mathrm{P}\left(\mathrm{M}\right)_{\mathrm{c}} = \left(\frac{\mathrm{D}}{\mathrm{M}}\right)\mathrm{P}\left(\mathrm{M}\right)_{\mathrm{t}} \tag{1}$$

Here, $\mathrm{P}\left(\mathrm{M}\right)_{\mathrm{t}}$ is a number of statistically significant residue pairs, that is, 20 minus the number of statistically insignificant residue pairs (The definition of statistically significant residue pair can be referred by [29]). $\mathrm{P}\left(\mathrm{M}\right)_{\mathrm{c}}$ is the number of residue pairs with the average inter Ca atomic distances less than a cutoff distance at the range M. That is, the order of the average inter-residue Cα atomic contact distances in each range M specifies the cutoff value ($\mathrm{d_c}\left(\mathrm{A}, \mathrm{B}, \mathrm{M}\right)$). D is an adjustable parameter that gives the average plot density of entire ADM close to that of the contact map constructed based on the 3D structure of a protein. The details of the determination of the value D can be referred by Ref [29]. (From now, PDB code is referred to as the name of a protein.)

Definition of contact density of a map

The many interactions between two segments in a protein exhibit a high-density region on a contact map. A sudden change in the density of plots would be observed at a limit of such a region on a map. Suppose that we have a contact map. This map is divided into two parts, triangle and trapezoid parts, by a line parallel to the x-axis of the map, as shown in Fig. 3(A). Suppose that a map is divided at the residues i and i + 1. Then, let $\rho_{\mathrm{i}}^{\mathrm{v}}$ be the plot density of the triangle region, and $\tilde{\rho}_{\mathrm{i}}^{\mathrm{v}}$ be that of the trapezoidal region. The same procedure is applied with a line parallel to the y-axis of the map (Fig. 3(B)), that is, the $\rho_{\mathrm{j}}^{\mathrm{h}}$ and $\tilde{\rho}_{\mathrm{j}}^{\mathrm{h}}$ are defined for the regions divided by this line in the same way as the definitions of $\rho_{\mathrm{i}}^{\mathrm{v}}$ and $\tilde{\rho}_{\mathrm{i}}^{\mathrm{v}}$ (see Fig. 3 (B)). The density differences are defined as the following equations.

$$\varDelta\,\rho_{\mathrm{i}}^{\mathrm{v}} = \rho_{\mathrm{i}}^{\mathrm{v}} - \tilde{\rho}_{\mathrm{i}}^{\mathrm{v}} \tag{2}$$

$$\varDelta\,\rho_{\mathrm{j}}^{\mathrm{h}} = \rho_{\mathrm{j}}^{\mathrm{h}} - \tilde{\rho}_{\mathrm{j}}^{\mathrm{h}} \tag{3}$$

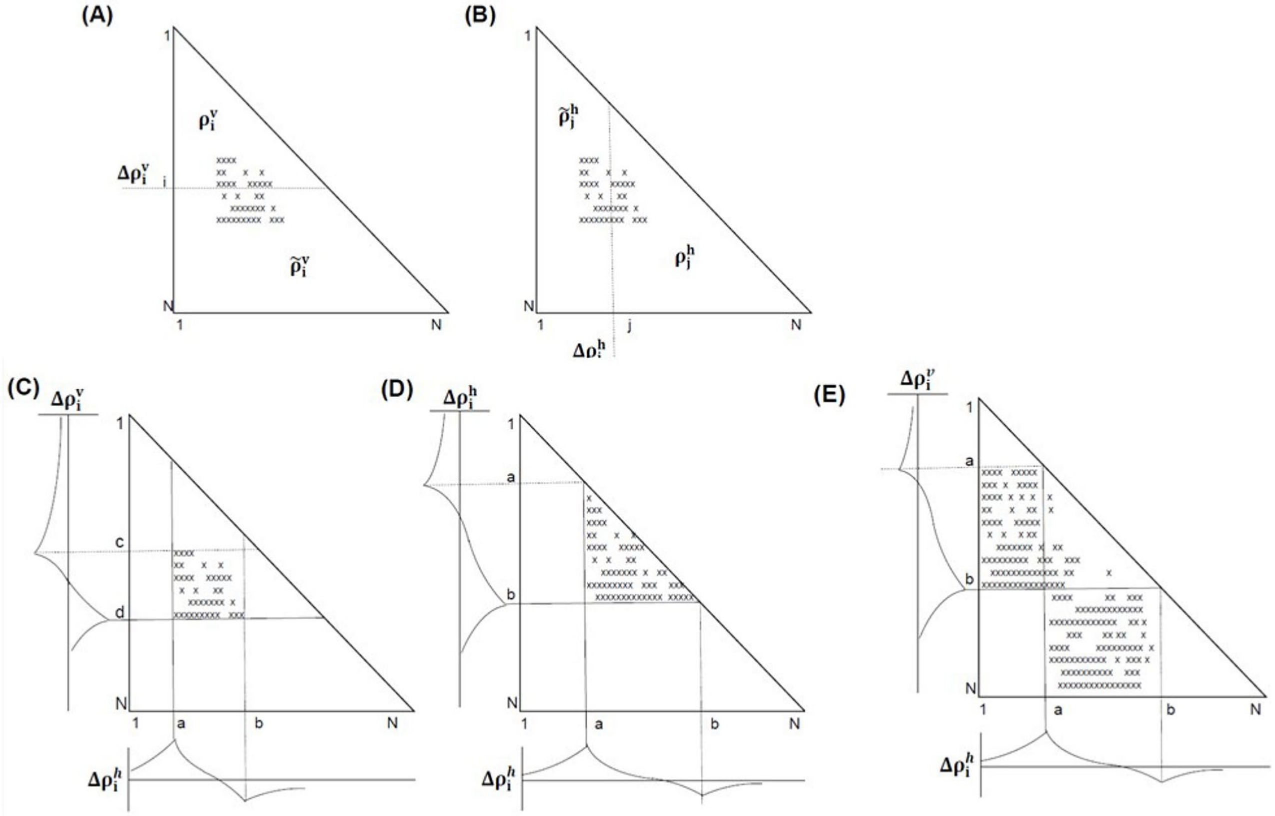

Fig. 3 An illustration of ADM analysis. (**A**) Corresponding to the step (3), ADM is divided into a triangle and trapezoid parts at i-th residue by a horizontal line. $\tilde{\rho}_i^v$ and ρ_i^V denote density of plots in the triangle and trapezoid parts respectively. $\Delta\rho_i^v$ means density difference between $\tilde{\rho}_i^v$ and ρ_i^v, that is, $\Delta\rho_i^v = \tilde{\rho}_i^v - \rho_i^v$. (**B**) Corresponding to the step (3), ADM is divided into a triangle and trapezoid parts at i-th residue by a vertical line. $\tilde{\rho}_j^h$ and ρ_j^h denote density of plots in the triangle and trapezoid parts respectively. $\Delta\rho_j^h$ means density difference between $\tilde{\rho}_j^h$ and ρ_j^h, that is, $\Delta\rho_j^h = \tilde{\rho}_j^h - \rho_j^h$. (**C**) Corresponding to the step (3), a part with high density plots is detected by a valley and a peak of the scanning plots of $\Delta\rho_i^v$ and $\Delta\rho_j^h$ in this figure. The high-density plot region in this figure denotes the interaction between the segments a-b and c-d along the sequence. (**D**) Corresponding to the step (4), a region with high density plots along the diagonal can be detected by peaks of the scanning plots of $\Delta\rho_i^v$ and $\Delta\rho_j^h$. This schematic figure shows that the region a-b forms a compact region. (E) Corresponding to the step (4), a region interacts with regions outside of this region also shows peaks at b and a in the scanning plots of $\Delta\rho_i^v$ and $\Delta\rho_j^h$. In this case, the region a-b is not a compact but constitutes a structured part

When we make a scanning plot $\Delta\rho_j^h$ for the map in Fig. 3(C), the plot will exhibit a peak and a valley at a and b as shown below of Fig. 3(C) (the figure is a schematic drawing). For the case of $\Delta\rho_i^v$, the plot will show a valley and a peak at c and d as presented in the left of Fig. 3(C). Thus, the interaction between segments defined by a-b and c-d in the given protein, as indicated in Fig. 3(C), is detected by this procedure. When we apply these procedures to a map with a high-density region along the diagonal, the scanning plots of $\Delta\rho_j^h$ and $\Delta\rho_i^v$ are like Fig. 3(D). As indicated in Fig. 3(D), two peaks of the plots of $\Delta\rho_j^h$ and $\Delta\rho_i^v$(i.e., a and b) where sudden changes of the contact densities are observed suggest a compact region in the given protein.

Pinpointing a location of a compact region

The summation of the differences in densities defined by lines parallel to the y-axis and the x-axis is called η-value, and this value is considered to be a measure of strength compactness of the region ($\eta = \Delta\rho_a^h + \Delta\rho_b^v$ in the example of Fig. 3(D)). A compact region predicted by ADM is regarded as a stable compact or structured region in the early stage of folding. In the case that the segment a-b interacts with the several parts of a protein outside of the segment a-b, also peaks at b in the vertical scan and at a in the horizontal scan as indicated by Fig. 3(E). In such a case, the region a-b is also identified as a structured region on ADM (see Fig. 3(E)). We take such a region as a structural region; for convenience, we call it a compact region though such a region is not compact. We regard a predicted compact region by ADM with a high η value as a compact or a structured region in the early stage of folding. It has been confirmed that compact regions correspond well to experimental results of protein folding mechanisms [18–26].

Contact frequency analysis (F-value)

The contact frequency of a residue with other residues in a random state is estimated using a potential derived from the present inter-residue average distance statistics to determine where initial folding events happen, such as hydrophobic collapse. In the present analysis, we employed a Cα bead model to represent a protein's structure. (The details of the procedure can be referred to [16].)

Monte Calro simulation of a protein Ca bead model

For a simulation of protein conformations, the Metropolis Monte Carlo method with the potential energy $\varepsilon_{i,\,j}$ derived from average distance $\bar{r}_{i,\,j}$ and its standard deviation $\sigma_{\ i,\ j}$ was used. The bond and dihedral angles of an initial conformation were randomly chosen. During a simulation, the bond and dihedral angles between the residue i and i + 1 are bent and rotated randomly, followed by the Metropolis judgment to decide whether the new conformation can be accepted or not. We started a simulation with a random distribution with the restriction derived from the average distance statistics. One step includes residues i = 1…N-1, that is, all the bond and dihedral angles are altered and judged.

Definition of interresidue potential based on average distance statistics

It is assumed that the probability density with the potential energy between two residues, $P(\varepsilon_{i,\,j})$, is equivalent to the probability density derived from the standard Gaussian distribution calculated with its average distance and standard deviation, $\rho\ (\bar{r}_{i,\,j}, \sigma_{i,\,j})$, as follows:

$$P\left(\epsilon_{i,j}\right) = \rho\ \left(\bar{r}_{i,j},\ \sigma_{\ i,j}\right) \tag{4}$$

This equation can be expressed by Eq. (4);

$$\frac{\exp\left(-\frac{\epsilon_{i,j}}{kT}\right)}{Z} = \frac{1}{\sqrt{2\pi\,\sigma_{\ i,j}}}\exp\left\{-\frac{\left(r_{i,j}-\bar{r}_{i,j}\right)^2}{2\sigma^{\ 2}_{\ i,j}}\right\} \tag{5}$$

Equation (4) leads to Eqs. (5) and (6):

$$-\frac{\epsilon_{i,j}}{kT} - \mathrm{In}Z = -\mathrm{In}\left(\sqrt{2\pi\,\sigma_{i,j}}\right) - \frac{\left(r_{i,j}-\bar{r}_{i,j}\right)^2}{2\sigma^2_{\ i,j}} \tag{6}$$

$$\frac{\epsilon_{i,j}}{kT} = \frac{\left(r_{i,j}-\bar{r}_{i,j}\right)^2}{2\sigma^2_{\ i,j}} - \mathrm{In}\frac{Z}{\sqrt{2\pi\,\sigma_{i,j}}} \tag{7}$$

Z is the partition function of the present system. Where kT is set so that the acceptance ratio is 0.5. Thus, this potential is expected to obtain ensembles reproducible in regard to inter-residue average distance statistics. However, it is noted that a significant value is the difference between those of two conformations, and Z does not appear in the calculation explicitly. Thus, Z is ignored in calculations.

Calculation of contact frequency and definition of F-value

From the results of simulations, the contact frequency, g(i, j), for each pair of residues is calculated with sampled structures generated using the potential energy function. (The contact is defined as a distance between Ca atoms of the residues i and j, r_{ij}, shorter than a threshold, r_C, i.e., $r_{i,j} < r_C$. The threshold is taken as 10 Å in this study.) Then, we normalize the residue contact frequencies, g(i, j), in the same range M as follows:

$$D\left(M\right) = \sqrt{\frac{\sum_{|\mu-v|}\left(\frac{\sum_{|\mu-v|\epsilon M} g(\mu,v)}{\sum_{|\mu-v|\epsilon M}} - g(\mu,v)_{|\mu-v|\epsilon M}\right)^2}{\sum_{|\mu-v|\epsilon M}}} \tag{8}$$

$$Q_{(i,j)} = \frac{g\left(i,j\right)_{|i-j|\in M} - \frac{\sum_{|\mu-v|\epsilon M} g(\mu,v)}{\sum_{|\mu-v|\epsilon M}.}}{D\left(M\right)} \tag{9}$$

Where μ or v is a residue number. Finally, we obtain the relative contact frequency, F, by summing the normalized contact frequencies, Q(i, j), from j = 1 to N for each residue i, where N is the total number of residues:

$$F_i = \sum{}_j Q\left(i,j\right) \tag{10}$$

We call F_i the F-value. Residues at peaks of the plot of F-values are expected to be located in the centre of many inter-residue contacts, such as at a hydrophobic cluster. A region around a peak in an F-value plot is plausible to be significant for folding, especially at the initial stage. We performed 100 simulations with 60,000 steps, calculating the average of the F-values for residue i (We calculate the sampled structure from the very beginning of the simulation). We attached a sequence of 10 glycine to both N- and C-termini to avoid too dynamic motions of residues at both ends.

Definition of a peak of F-value plot

A peak is defined when the difference in the values of a valley and a peak is more than the following cut-off value, F_{cut}:

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

Where F_i is the F-value of residue i, and N is the total residue number. That is, the difference between a valley and a peak should be more than the intrinsic fluctuation of a plot.

It is considered that a hydrophobic residue near the F-value plot for a protein tends to be buried in the initial stage of folding and folding core of the protein. We have confirmed that a hydrophobic residue near the F-value plot for a protein tends to form hydrophobic packing in the native structure of a protein for the IgG binding domain [16] and ferredoxin-like fold proteins [23]. Thus, we consider that the folding is initiated with residues near the F-value plot for a protein in this study.

Evolutionary analysis and hydrophobic packing

Our previous studies have suggested that experimental results of the folding processes of proteins can be explained well when we focus on hydrophobic residues [16–25], especially corresponding well to H/D exchange experiments. Thus, we also focus on the hydrophobic residues in the present work.

Evolutionarily conserved residues in a protein are responsible for its function, stability, and native structure formation [30–33]. In this study, we attempt to distinguish conserved hydrophobic residues for each of the targeted proteins. We regard the following residues as hydrophobic residues, Ala, Phe, Ile, Leu, Met, Val, Tyr, and Trp. When 90% of residues at an aligned site are hydrophobic in multiple alignments, this site is regarded as an "evolutionarily conserved hydrophobic site". We use the term "conserved hydrophobic residues" in this sense, and we use CHR as an abbreviation of "conserved hydrophobic residue".

To obtain homologues for each of the targeted proteins, homology searches were conducted using BLAST [34] using proteins in Table 1 as queries with the e-value is 0.01 to make sure to obtain evolutionary homologous sequences. UniPort database [35] is used to search the homologs. Then we excluded sequences from search results with a length less than 85% of the query sequence, with more than 90% of sequence identity, and with a gap of more than five residues within a secondary structure from the searched sequences. Multiple sequence alignment was conducted using the ClustalW program, which is integrated into the software MEGA [36]. We also constructed a molecular phylogenetic tree using the neighbor-joining method [37]. Finally, the evolutionary distances were computed using the JTT matrix-based method [38].

In this study, we also assume that conserved hydrophobic residues especially those around a peak of the F-value plot of a protein are significant for folding.

Conservation of predicted compact regions (PdCRs)

Let us use the abbreviation "PdCR" as a predicted compact region by ADM. The similarity of the location of PdCRs can be relevant to the conservation of PdCRs during molecular evolution. Once a sequence alignment of proteins in a family is performed, the conservation of PdCRs can be defined. The procedure to determine the conservation of regions predicted by ADMs is as follows. First, a multiple-sequence alignment is obtained. Then, a "site" is referred to as the common sequential number in the multiple alignments.

I. The number of residues at a given site commonly included or excluded from the PdCRs is calculated.
II. The ratio of this number to the number of aligned sequences is calculated.
III. We make a histogram of this ratio vs. the site number. (An example of a histogram can be seen in Fig. S8. (See the bottom of a corresponding figure)
IV. A region encompassing several residues with high ratios denotes a conserved PdCR during evolution.
V. Currently, the PdCR is regarded to be conserved when the conservation ratios exceed 70% in the same position of the aligned samples [26].

Definition of an inter-residue contact

When a heavy atom in a residue is close to a heavy atom in another residue within 5Å, these two residues are regarded to make contact. We believe the interaction between two residues in different secondary structures is a contact. However, the closed residues in the same secondary structure are not regarded as forming a contact.

Results

Analyses of RNase A-like fold and trypsin-like serine proteases fold proteins based on inter-residue average distance statistics

Tables 2, 3 and 4 summarize the results of the ADM analysis for each protein in this study, along with statistics on conserved hydrophobic residues and the detailed positions of F-value plot peaks.

RNase A-like fold

The location of the corresponding common segments in ribonuclease A (6ETL) is illustrated in Fig. 1(B). Figures 4 and 5 present the ADM and F-value analyses for 6ETL, respectively, alongside data from the NMR H/D

Table 2 Summary of the ADM predicted regions in every study protein

Protein name	PDB ID	Sequence length	ADM predicted region with η value	Total number of ADM predicted residues	Ratio of residues in PdCRs	Dominant region
RNase A-like fold	6ETL	123	19–84: 0.219 92-118: 0.146[1)]	93	0.76	N terminal
	2VQ9	123	8–85: 0.227 93–109: 0.108	95	0.77	N terminal
	2ZPO	119	1–81: 0.372[2)] 92–107: 0.154	97	0.81	N terminal
Trypsin-like serine proteases fold	6CHA	131	26–53: 0.188 65 – 128: 0.180	92	0.70	N terminal
	1LTO	115	12–60: 0.164 65–112: 0.277	97	0.84	C terminal
	3DFJ	115	7–43: 0.232 76–114: 0.251	76	0.66	C terminal

[1)] The η value, 0.146, of this region is slightly smaller than that of 92–110 (η= 0.151) but the value is very close (within 0.85%, see [25]), and thus we take 92–118 as a PdCR

[2)] The η value of this region, 0.372, is slightly smaller than that of 1-58 (η= 0.385) but the value is very close (within 0.85%, see [25]), and thus we take 1–81 as a PdCR

Table 3 Statistics of conserved hydrophobic residues in study proteins

Protein name	PDB ID	Total number of Conserved hydrophobic residues	Conserved hydrophobic residues in ADM predicted regions	Highly (≥ 70%) conserved regions	Ratio of the conserved hydrophobic residues in predicted regions to the total number of conserved hydrophobic residues
RNase A-like fold	6ETL	11	11	11	1.00
	2VQ9	10	10	9	1.00
	2ZPO	11	9	9	0.82
Trypsin-like serine proteases fold	6CHA	22	18	17	0.82
	1LTO	18	11	14	0.61
	3DFJ	16	16	13	1.00

Table 4 The summary of the F-value peak positions

Fold	PDB ID	F-value peaks
RNase A-like fold	6ETL	26-Cys, 57-Val, 63-Val, 72-Cys, 95-Cys 108-Val
	2VQ9	13-Val, 56-Thr, 80-Val, 109-Cys
	2ZPO	55-Thr, 77-Ala, 103-Ile
Trypsin-like serine proteases fold	6CHA	38-Val, 69-Lys, 88-Ile, 106-Val, 120-Thr
	1LTO	38-Trp, 78-Ile, 103-Ile
	3DFJ	26-Cys, 36-Val, 58-Ala, 70-Val, 77-Ile, 93-Leu

exchange experiment by Neira et al. [39]. The ADM analysis predicts two predicted compact regions (PdCR), the N-terminal region from residue 19–84 is the prominent structured unit compared to their C-terminal region, which is from the residue 92 to 118 and their h-values (strength of compactness), 0.219 and 0.146 respectively. In Fig. 4(B), these ADM-predicted regions are highlighted in red and green in the 3D structures of 6ETL (see also Table 2), with the higher η-value region colored red (From here, the red and green colors are used to distinguish PdCRs).

The first PdCR (residues 19–84) includes the β1 and β4 strands from the common segments, while the second PdCR (residues 92–118) corresponds to β5, β6, and β7. The interaction between these two PdCRs is likely to contribute to the formation of the common structure.

Taking higher h value into account, the N-terminal site is expected to form an initially stable folding core during the folding. The F-value plot of 6ETL has six high peaks as presented in Table 4. The peaks are in α2, α3, β2, β3, and β6. The first PdCR includes the α2-α3 and β1-β4; on the other hand, the second PdCR includes the β5, β6 and a part of β7. The F-value analysis in Fig. 5 indicates that the α3 and β6 are the keys to their folding because they have high contact frequency. NMR H/D exchange result showed that the highly protected parts are the α3 and β6, as shown in Fig. 5. The orange bars with broken lines in Fig. 5 correspond to residues with high H/D exchange free energy from the study by Neira et al. [39] (see also the legend of Fig. 5 in detail). The number of such residues is 19. Among these 19 residues with high H/D exchange free energy, 12 residues are near the peaks of the F-value plot within ± 4 residues, with 11 of these residues clustered around the two highest peaks near conserved hydrophobic residues (CHRs, red dots in Fig. 5), namely 57-Val and 108-Val. This suggests that these residues are likely centers of folding during the early stages. Notably, α3 and β6, which are part of the first and second PdCRs, respectively, act as folding centers, with α3

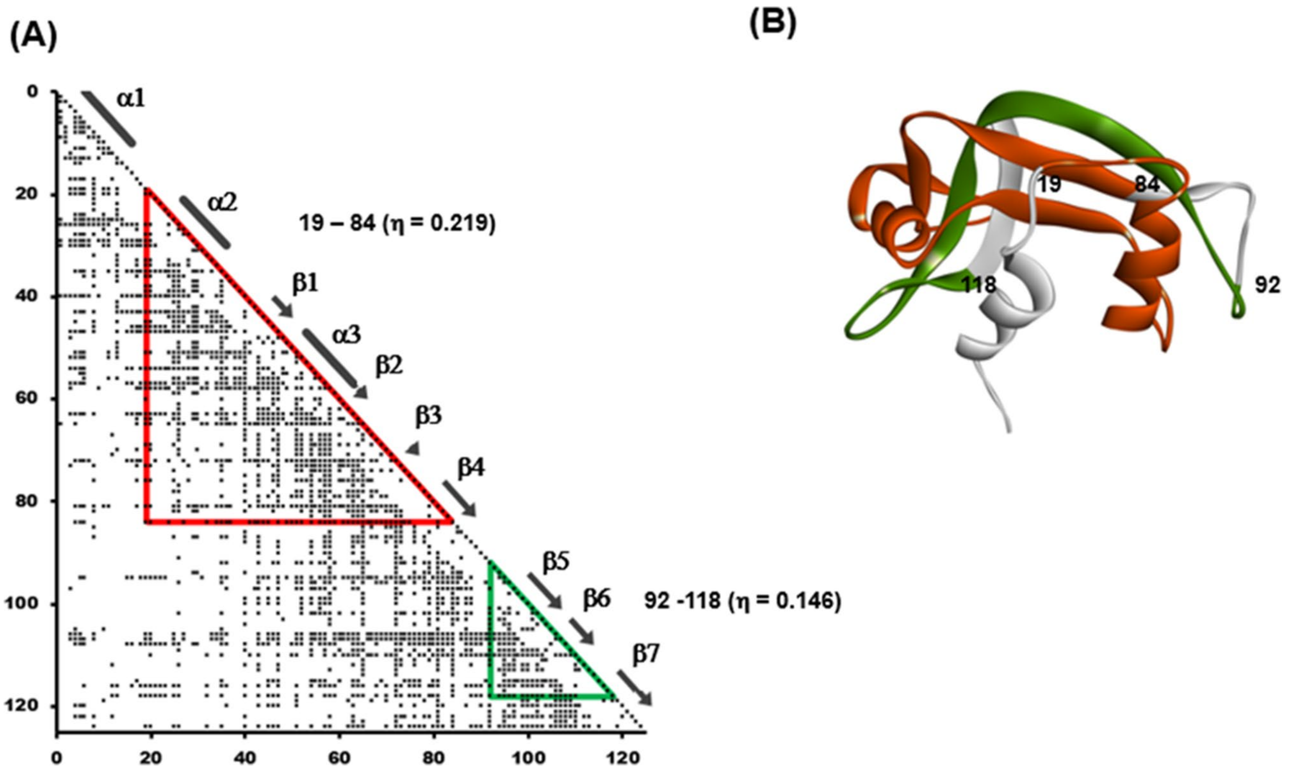

Fig. 4 (**A**) ADM for 6ETL. The location of a secondary structure is indicated by a bar or an arrow for a-helix or b-strand along the diagonal. The red (green) triangle denotes PdCR with a higher (lower) h value. (**B**) The regions of PdCRs in the 3D structure of 6ETL. The red (green) part denotes PdCR with a higher (lower) h value (See Table 2 for the assignment of the second PdCR)

serving as the leading folding initiation site due to its higher η-value. The number of these residues with high H/D exchange free energy are 19 and, among them, 8 residues are within the first PdCR and 8 residues are within the second PdCR, that is, two folding initiation sites identified by H/D exchange experiment are in respective PdCRs. The h value of the second one is smaller than the first one, and the peak of the F-value plot around 108-Val is relatively low. Therefore, the main center is the first ADM-predicted region, and the interaction with the first one may stabilize the second one.

In the RNase ZF-3E (PDB code: 2VQ9) protein, the corresponding common segments β1, β4, and β6 are presented in Fig. S1(B). Notably, the corresponding β5 and β6 strands in 6ETL are combined into β5 in 2VQ9 according to PDB annotations. ADM analysis of 2VQ9 reveals a highly compact N-terminal region (residues 8–85) with an η-value of 0.227, encompassing α1-α3 helices and β1-β4 (Fig. S4(A), (B)). The second PdCR (residues 93–109) corresponds to β5 with an η-value of 0.108. Similar to 6ETL, the first PdCR in 2VQ9 exhibits higher compactness, suggesting that it forms a stable folding core, while the second PdCR acts as a smaller structural unit stabilized by the interaction with the first PdCR. Fig. S4(C) shows the peaks of the F-value plot for 2VQ9, which are located near conserved hydrophobic residues. The two highest peaks (positions 56 and 80) in α3 and β4 are likely early-stage folding sites, forming part of the first PdCR. Additionally, a high F-value peak at 109-Cys in β5 indicates another compact predicted region in the C-terminal region, suggesting that the N-terminal region initiates folding, similar to 6ETL. As no experimental data on the folding of 2VQ9 is available, comparisons can only be made with the experimental data for 6ETL.

The ADM analysis of turtle egg white ribonuclease (PDB code: 2ZPO) shows a similar pattern to the other RNase A-like fold proteins, with two PdCRs identified (Fig. S5(A), (B)). The primary PdCR spans residues 1–81 with an η-value of 0.372, and the secondary PdCR spans residues 92–107 with an η-value of 0.154. The first PdCR is predicted to have a higher potential for initiating folding compared to the second PdCR. The F-value plot of 2ZPO, presented in Fig. S5(C), has three peak positions at the residues in Table 4. Three peak positions in α3, β4, and β6 indicate that these regions are critical to folding. The early-stage folding likely begins with the helix in the region 1–81 which then forms a core folding unit with the β6 strand, stabilized by residues 92–107.

Interestingly, despite the relatively low sequence identity among these RNase A-like fold proteins (25–40%), they all exhibit similar PdCRs according to ADM analyses. Each protein has two predicted compact regions,

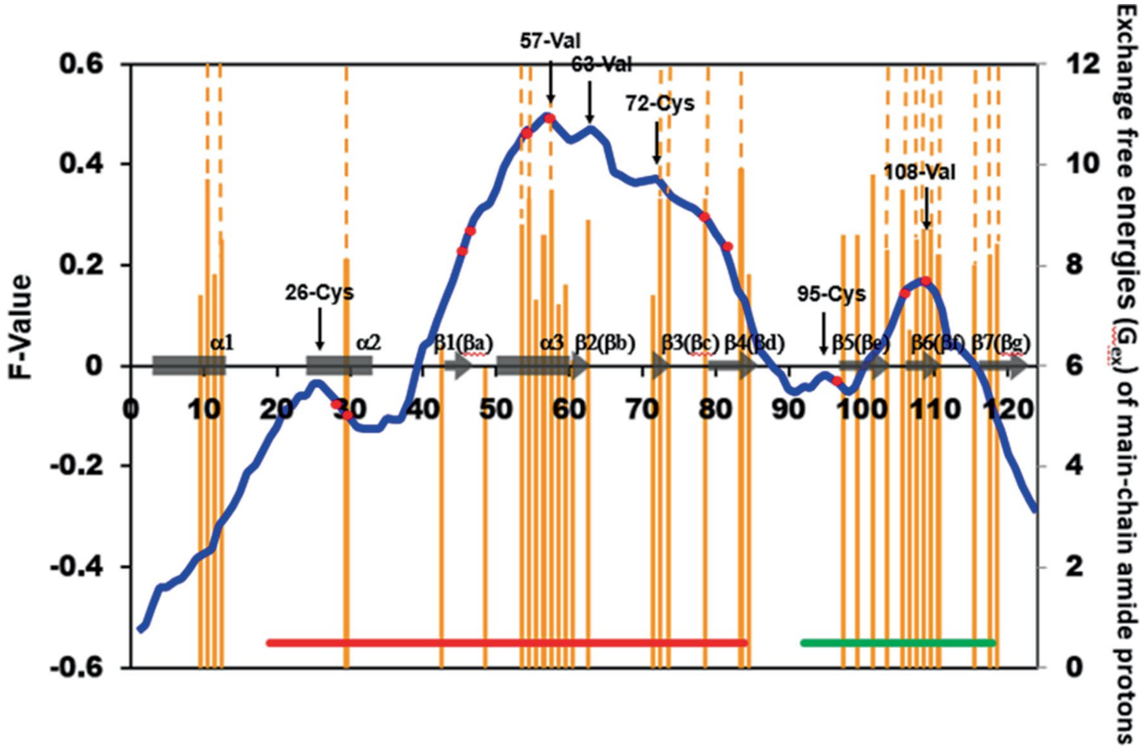

Fig. 5 F-value plot with the NMR H/D exchange experimental data [39] for 6ETL. The blue line denotes the F-value plot. A residue number indicates the position of a peak. The H/D exchange free energy of a residue is characterised by an orange bar. The value of exchange free energy for a residue was estimated from the experimentally obtained H/D exchange rate constant of a corresponding residue [39]. The value of exchange free energy for a residue is sometimes presented as > 9.5 and so on in [39]. The number of the orange lines with the broken lines is 19. To explain this situation, we use an orange broken line for the corresponding values. A red dot denotes the position of a CHR. The x-axis denotes the residue number. A grey bar and a grey arrow on the x-axis mean a position of a-helix and b-strand, respectively. In the lowest part of this figure, two PdCRs are indicated by the red and green bars

with the N-terminal region being more compact than the C-terminal region (Table 2). Importantly, the second PdCR in each protein contains β5 and β6 strands, which are part of the common 3D topology described in Figs. 1 and S1. The NMR H/D exchange experiment for 6ETL showed that the α3 helix, β3, and β6 strands exhibit the slowest exchange rates (high H/D protection), forming a highly stable hydrophobic core [39–41]. ADM analysis predicts two PdCRs, which include the α3 helix and β3 strand in the primary compact region and the β6 strand in the secondary compact region. Based on these findings, we can infer that the first PdCR is likely the folding initiation site for all these proteins, with the interaction between the PdCRs forming a stable folding core during the folding process, consistent with experimental results [39].

Trypsin-like serine proteases fold

ADM prediction for 6CHA domain B exhibits two compact regions. One is from residue 26 to 53, and the other is from residue 65 to 128, illustrated in Fig. 6(A). The highly compact region in the N-terminal has the h-value of 0.188 (first PdCR) that includes α1 helix and β2-β4 (bb-bd). Another one has the h-value of 0.180 (second PdCR), which contain β5 (be-f) and β6 (bg) (as mentioned b5 (be-f) in 6CHA corresponds to b5 (be) and b6 (bf) in 6ETL, that is, one b strand in the PDB annotation). The compactness of the predicted regions is quite similar, and their N-terminal region is slightly higher than the second PdCR. As shown in this figure, the N-terminal compact region covers β2(bb), β3(bc), α1 and β4(bd) secondary structures. The F-value plot is in Fig. 7. The highest F-value peak appears at β3 at residue number 38-Val. The second PdCR ranges from β5(be-f) to b7 and has four high peak positions as Table 4 presents. These are possible folding sites. From the results of the F-value analysis, we can expect that the β3(bc), b5(be-f), and β6(bg) strands will make the folding core in the common topology. The meaning of the peaks at 106-Val and 120-Thr will be discussed later.

Though the results of our present analyses should be examined by objective experimental data, to our best knowledge, no experimental analyses on folding of a protein in Trypsin-like serine proteases fold have been performed so far.

Figure S6(A), (B) and Table 2 represent the ADM of 1LTO having two PdCRs, the first PdCR includes residues 12–60 with η value 0.164, and the second PdCR includes

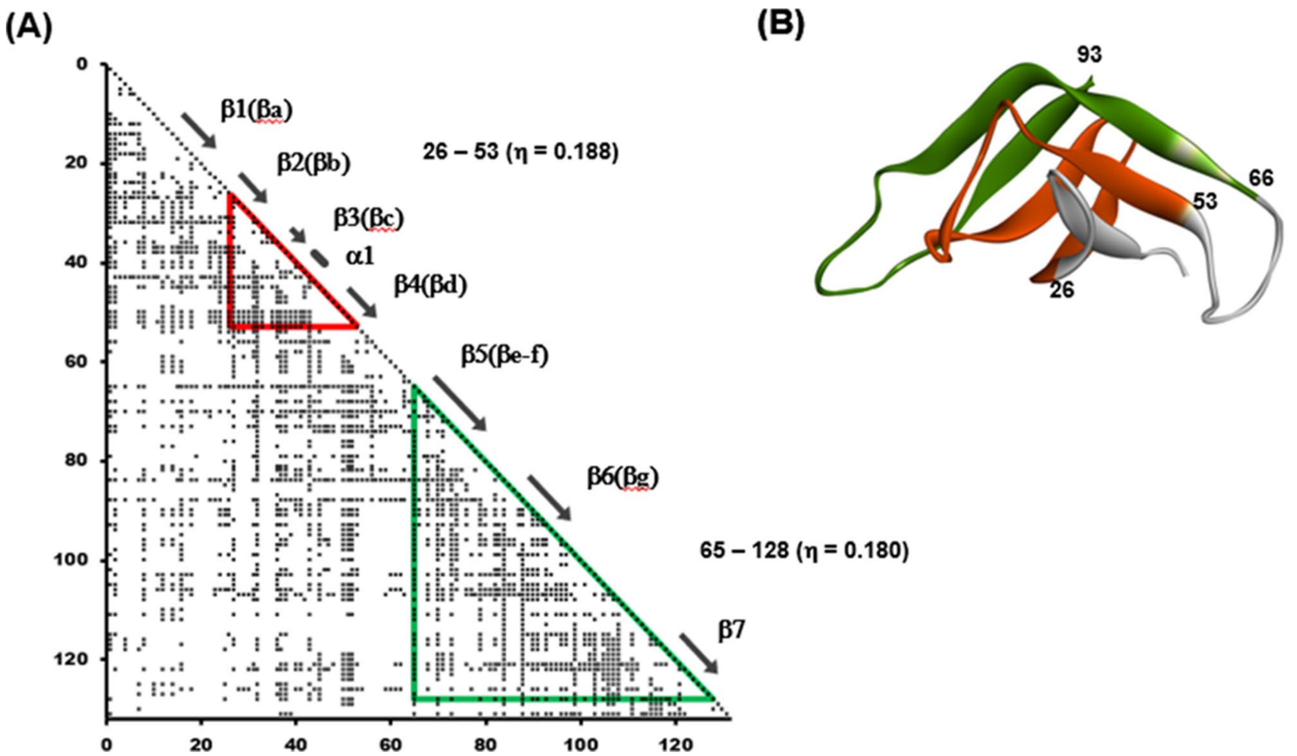

Fig. 6 (**A**) ADM for 6CHA (For the construction of ADM, the whole sequence of 6CHA is used). The location of a secondary structure is indicated by a bar or an arrow for a-helix or b-strand along the diagonal. The red (green) triangle denotes PdCR with a higher (lower) h value. (**B**) The regions of PdCRs in the 3D structure of 6CHA (Only the structure of the core part (not the whole sequence) is shown). The red (green) part denotes PdCR with a higher (lower) h value

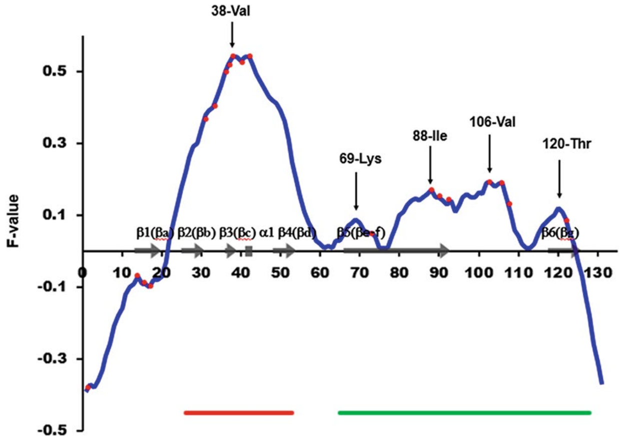

Fig. 7 F-value plot for 6CHA. The blue line denotes the F-value plot. An arrow with the corresponding residue indicates the position of a peak. A red dot represents the position of a CHR. At the bottom of the figure, the x-axis denotes the residue number. A grey bar and a grey arrow on the x-axis mean a position of a-helix and b-strand, respectively. In the lowest part of this figure, two PdCRs are indicated by the red and green bars

residues 65–112 (η value of 0.277). The C-terminal site contains α3 helix and β6-β7strands, which are expected to be stable in the early stage of folding based on the h values. The F-value for 1LTO shown in Figure S6 (C) indicates that there are three prominent peaks at 38-Trp (in β4), 78-Ile (in b6) and 103-Ile (near b7) (see also Table 4), predicting that the β4 and β7 strands with many CHRs are significant for their folding of this protein. The hydrophobic residues in β4 and β7 are highly conserved during the evolution, that is, a lot of CHRs in β4 and β7. Thus, β4 and β7 strands seem to be very important for the folding of this protein. Those high peaks at β4 and β7 are in two different PDCRs. It is plausible that these conserved hydrophobic residues near an F-value peak are significant in stabilizing the interaction between two PdCRs. From here, we use the abbreviation "CHRnF" as conserved hydrophobic residues near an F-value peak.

ADM and F-value results of 3DFJ are shown in Figs. S7 (A)-(C) (See also Table 2). Two PdCRs are predicted by ADM; the first PdCR covers 7–43 with a 0.232 h value, and the second one is 76–114 with h value of 0.251. F-value analysis in Fig. S7 (C) exhibits six peaks in the data plot, and many CHRnFs are distributed around the peak at 26-Cys and 36-Val in the first PdCR and 93-Leu in the second PdCR (Table 4). Thus, these CHRnFs would be prominent for the folding of 3DFJ. Considering these results, we can assume that the folding core will be made by β3-β4 and β8 strands.

In conclusion, our analysis of ADM predictions, F-value data, and conserved hydrophobic packing suggests that the folding properties of all studied proteins within the Trypsin-like serine proteases fold are remarkably consistent. These findings highlight the critical role

Table 5 The summary of the sequence for the evolutionary analyses

	RNase A-like fold			Trypsin-like serine proteases fold		
	6ETL	**2VQ9**	**2ZPO**	**6CHA**	**1LTO**	**3DFJ**
Average sequence identity (%)	38.4	41.8	33.7	31.9	32.5	31.9
Maximum sequence identity (%)	89.3	89.6	88.6	89.1	89.3	89.4
Minimum sequence identity (%)	11.8	16.2	8.7	12.2	16.7	11.2

Table 6 Table of conserved hydrophobic residues

Protein name	PDB ID	Conserved Hydrophobic Residues
RNase A-like fold	6ETL	29-Met. 30-Met, 46-Phe, 47-Val, 54-Val, 57-Val, 79-Met, 81-Ile, 97-Tyr, 106-Ile, 108-Val
	2VQ9	26-Met, 46-Phe, 47-Ile, 54-Val, 57-Val, 78-Phe, 80-Val, 96-Tyr, 105-Ile, 107-Val
	2ZPO	46-Phe, 47-Val, 54-Ile, 57-Ile, 76-Phe, 78-Leu, 94-Tyr, 103-Ile, 105-Ile, 113-Val, 115-Tyr
Trypsin-like serine proteases fold	6CHA	1- Ile, 14-Trp, 16-Val, 18-Leu, 31-Leu, 32-Ile, 36-Trp, 37-Val, 38-Val, 40-Ala, 41-Ala, 74-Phe, 88-Ile, 90-Leu, 91-Leu, 93-Leu, 103-Val, 106-Val, 108-Leu, 122-Val
	1LTO	14-Trp, 16-Val, 18-Leu, 33-Leu, 34-Ile, 38-Trp, 39-Val, 40-Leu, 42-Ala, 43-Ala, 92-Ile, 93-Ala, 94-Leu, 95-Leu, 97-Leu, 101-Val, 107-Val, 112-Leu
	3DFJ	16-Val, 18-Ile, 30-Leu, 31-Val, 35-Trp, 36-Val, 37-Leu, 39-Ala, 40-Ala, 91-Ile, 92-Ala, 93-Leu, 94-Leu, 96-Leu, 106-Ile, 111-Leu,

of conserved hydrophobic residues and their interactions in guiding the folding process and stabilizing the protein structure.

Evolutionary analysis of the RNase A-like fold and trypsin-like serine proteases fold proteins

We examined the evolutionary conservation of predicted folding units (PdCRs) in RNase A-like and trypsin-like serine proteases fold proteins using multiple sequence alignments combined with ADM results for homologous sequences obtained via BLAST search. Figs S8(A)-(F) show these alignments, with PdCRs highlighted in red. A histogram below each figure, indicated by a blue line, shows the ratio of residues at aligned sites within PdCRs to the total number of aligned sequences. Higher ratios suggest that these regions are evolutionarily conserved folding units. Table 1 represents the sequence identities of around 33–40% within the same group but only about 10% identity between proteins from different groups.

In ribonuclease A (6ETL), the alignment of 76 sequences (average sequence identity ~ 38%) shows that regions with higher η-values are concentrated in the N-terminal region illustrated in Fig. S8(A). Conserved folding units include β1, α3, β2, β4, and regions containing β5 and β6. All 11 conserved hydrophobic residues are within PdCRs, with 7 near peaks in the F-value plot (Table 5). It is important to note that the properties of PdCRs and F-value profiles exhibit considerable variability among the proteins analyzed here. Despite these differences, the overarching features, such as critical folding nuclei, remain consistent across homologous proteins. This variability may reflect nuances in the specific folding properties of homologous proteins, as discussed in references [18–19], and [22].

The plot denoted by an orange line in Fig. S8 indicates a ratio of hydrophobic residues to the total number of sequences at an aligned site. We regard a more than 90% ratio that the current site shows the conservation of hydrophobic residues. The conserved hydrophobic residues in PdCRs are indicated by yellow letters, and out of PdCRs are indicated by blue letters in this figure. For 6ETL, the conservation of hydrophobic residues is observed at 11 sites in Fig. S8(A), and all of the conserved hydrophobic residues are in PdCRs of 6ETL. If we consider a region with more than 70% ratio of conservation of predicted regions tentatively (indicated by an orange line in the bottom of Fig. S8(A)) as an evolutionarily conserved predicted unit, then 11 out of the 11 conserved hydrophobic residues are included in the evolutionarily conserved predicted unit (The blue line above the orange line in the bottom of Fig. S8(A)). Moreover, 7 of these residues are near the peaks of the F-value plot of 6ETL within ± 5 residues represented in Fig. 5; Tables 6 and 7.

For RNase ZF-3E (2VQ9), analysis of 78 sequences (average sequence identity ~ 40%) reveals that PdCRs include β1, β2, β4, α3, and β5-β6 (Fig. S8(B)). The histogram indicates that over 70% of conserved predicted folding units involve these regions, which contain 9 out of 10 conserved hydrophobic residues, 6 of which are near F-value peaks (Fig. S4(C), Tables 3, 6 and 7).

In turtle egg white ribonuclease (2ZPO), analysis of 29 sequences (average sequence identity ~ 33%) shows that high η-values are primarily in the N-terminal region, with PdCRs covering α2-β4 and β5-β6 as shown in Fig. S8(C). More than 70% of conserved folding units involve β1-β3 and β4-β6, with 6 out of 11 CHRnFs located within these regions (Fig. S5(C), Table 7).

The sequence alignment of α-chymotrypsin domain B (6CHA) is shown in Fig. S8(D), incorporating 110 sequences with an average sequence identity of about 31%. The alignment results suggest that both N-terminal and C-terminal regions exhibit similar levels of

Table 7 Conserved hydrophobic residues near F-Value peak (CHRnFs)

Fold	PDB ID	F-value peak	Conserved hydrophobic residues near F-value peak
RNase A-like fold	6ETL	26-Cys 57-Val 63-Val 72-Cys 95-Cys 108-Val	29-Met, 30-Met 54-Val, 57-Val 97-Tyr 106-Ile, 108-Val
	2VQ9	13-Val 56-Thr 80-Val 109-Cys	54-Val, 57-Val 78-Phe, 80-Val 105-Ile, 107-Val
	2ZPO	55-Thr 77-Ala 103-Ile	54-Ile, 57-Ile 76-Phe, 78-Leu 103-Ile, 105-Ile
Trypsin-like serine proteases fold	6CHA	38-Val 69-Lys 88-Ile 106-Val 120-Thr	36-Trp, 37-Val, 38-Val, 40-Ala, 41-Ala 74-Phe 88-Ile, 90-Leu, 91-Leu, 93-Leu 103-Val, 106-Val, 108-Leu 122-Val
	1LTO	38-Trp 78-Ile 103-Ile	34-Ile, 38-Trp, 39-Val, 40-Leu, 42-Ala, 43-Ala 101-Val, 107-Val
	3DFJ	26-Cys 36-Val 58-Ala 70-Val 93-Leu	30-Leu, 31-Val 35-Trp, 36-Val, 37-Leu, 39-Ala, 40-Ala 91-Ile, 92-Ala, 93-Leu, 94-Leu, 96-Leu

conservation, covering β1-β4 and β5-β7, respectively. Highly conserved regions are observed in β2, β3, α1, β4, β6, and β7, with more than 70% conservation. Of the 22 conserved hydrophobic residues identified, 18 are located within these highly conserved regions, all of which correspond to peaks in the F-value plot (Fig. 7; Tables 6 and 7).

Figure S8(E) shows the multiple sequence alignment for α1-tryptase (1LTO), which included 117 sequences with an average sequence identity of approximately 32% (Table 5). ADM analysis predicts two folding units: one in the N-terminal region covering β3 to α2, and another in the C-terminal region covering β6 to β7. Of the 18 conserved hydrophobic residues, 14 are located within these evolutionarily conserved units, and 9 are near F-value peaks within ± 5 residues (Fig. S6(C), Tables 6 and 7).

Finally, the analysis of prostasin (3DFJ) shown in Fig. S8(F) includes 76 sequences with an average sequence identity of 31%, with maximum and minimum identities of 89% and 11%, respectively. The regions with higher η-values are primarily located in the C-terminal region. The conservation of hydrophobic residues is observed at 16 sites, all within PdCRs. The histogram indicates that β4, α1, β6, and β8 are evolutionarily conserved folding sites, with 13 out of 16 conserved hydrophobic residues located within these regions, adjacent to F-value peaks (Fig. S7(C), Tables 6 and 7).

Hydrophobic packing of evolutionarily conserved residues

In this section, we analyze the interactions of conserved hydrophobic residues within predicted folding units (PdCRs) based of the 3D structures of the studied proteins. The basic assumption is that conserved hydrophobic residues near F-value peaks (CHRnFs) are buried early in protein folding, serving as folding initiation residues [23–24, 26]. The folding process is thought to progress mainly through interactions among CHRnFs and other conserved hydrophobic residues (CHRs).

We start by identifying CHRnFs near the highest F-value peak within ± 5 residues [26] for each protein and then examine the compact regions (PdCRs) where these residues are located. We investigate the hydrophobic contacts within these PdCRs and between them, summarizing the results as contact maps (e.g., Fig. 8 for 6ETL). In the main text, we focus on interactions between secondary structure elements containing CHRnFs or CHRs, with detailed descriptions provided in the figure legends.

RNase A-like fold proteins

For ribonuclease A (6ETL), the PdCRs are located at residues 19–84 and 92–118, with the first PdCR exhibiting a higher η-value (Table 2). Figure 5 indicates that the highest F-value peak, 57-Val in α3, is a CHRnF within the first PdCR, suggesting α3 as the folding center. Figure 8 shows that hydrophobic contacts form among CHRnFs

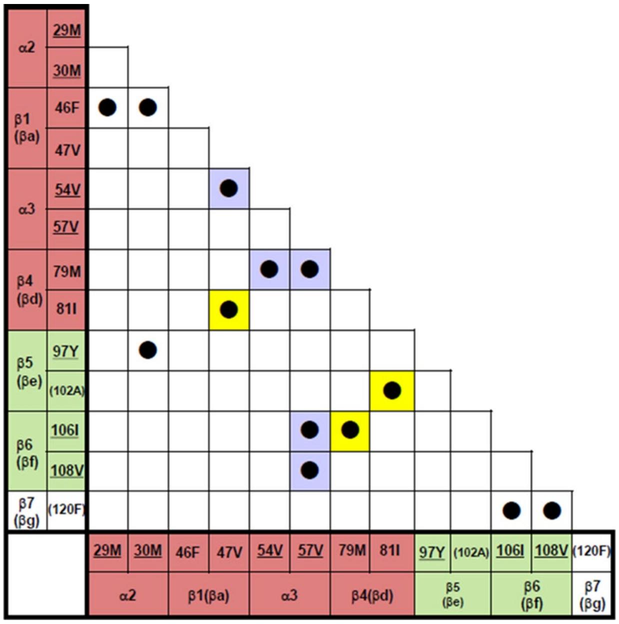

Fig. 8 Kind of contact map for 6ETL. Conserved hydrophobic residues are mainly presented. A red or green colored residue means a concerned residue is in a PdCR with the same color as in Fig. 4. The secondary structure containing a concerned residue is also indicated. A residue with an underline denotes CHRnF. A residue within a parenthesis is not CHR nor CHRnF. A contact colored by yellow means stabilization of a part of the common structure and a contact colored by violet means a contact between a residue in the common structure and a3. The CHRnFs in the second PdCR, 92–118, that is, 106-Ile and 108-Val make hydrophobic packing with 120-Phe

and CHRs in α2, α3, β1 (βa), and β4 (βd) within this region. The second PdCR, spanning residues 92–118, has a peak of the F-value plot at 108-Val as shown in Fig. 5 (a CHRnF in β6). Contacts between α2 and β6, as well as among α3, β4, β5 (βe), and β6 (βf), indicate interactions between the first and second PdCRs, stabilizing the common structure (Figs. 8 and 9(A)-(E)). The α3 helix can be regarded as a center of contacts that support the formation of the common structure. In Fig. 8, contacts colored yellow indicate those that directly stabilize the common structure, while contacts colored violet signify those that provide stabilization through their interaction with α3.

For RNase ZF-3E (2VQ9), PdCRs are at residues 8–85 and 93–109 (Table 2), with 56-Thr in α3 being the highest F-value peak, and CHRnF near to this residue is 57-Val (Fig. S4). Similar to 6ETL, α3 is predicted as the folding center, with hydrophobic contacts forming among CHRs in α2, α3, β1 (βa), and β4 (βd) within the first PdCR (Figs. S9, S10(A), (B)) stabilizing a part of the common structure (Fig. S9). The second PdCR, with a peak at 109-Cys (the CHRnF is 107-Ile in β6(bf). See Fig. S4(C).), also shows interactions between β6(bf) and β7(bg) within this PdCR. Furthermore, a contact between b1 and β6, that is, the interactions between PdCRs is also observed (Fig. S9). The contacts among a2, a3, b4(bd), b5(be) and b6(bf) also stabilizing the interaction between two PdCRs and leading to stabilize the common structure under the support of a3 (Figs. S9, S10(C)-(D)).

In turtle egg white ribonuclease (2ZPO), PdCRs are located at residues 1–81 and 92–107 (Table 2), with 55-Thr in α3 identified as the highest F-value peak. CHRnF, 54-Ile, in α3, regarded as a folding center in 2ZPO, forms contacts with β1 (βa) and β4 (βd), stabilizing the structure within the first PdCR with higher h-value (Figs. S11, S12(A)-(B)). Fig. S11 also indicates the contacts among a2, a3, b1(ba) and b4(bd) within the first PdCR stabilizing a part of the common structure. The second PdCR, with an F-value peak at 103-Ile (CHRnF) in β6(bf), shows interactions between β6(bf) and β7(bg), further stabilizing the structure (Figs. S12C-D). Interactions among β1(ba), β4(bd), β6(bf) directly stabilize the common structure, and α3, similar to 6ETL and 2VQ9, reinforces the common structure (Figs. S11 and 12(A), (C)).

Trypsin-like serine proteases fold

In α-chymotrypsin domain B (6CHA), PdCRs are located at residues 26–53 and 65–128, with a slightly higher η-value in the first PdCR (Table 2). As presented in Fig. 7, the highest F-value peak at 38-Val (a CHRnF in β3) suggests that β3(bc) is the folding center. Hydrophobic contacts among β2(bb), β3(bc), and β4(bd) stabilize the four stranded partial β-barrel structure (Figs. 10 and 11(A)), with additional interactions between β5(be-f) and β6(bg) contributing to the stability of the second PdCR (CHRnF is 90-Ile in b5(be-f) in this PdCR. Figures 10 and 11(A)-(C)). The common structure is stabilized by direct interactions among β1 (βa), β4 (βd), β5 (βe-f), and β6 (βg), as indicated by contacts colored yellow in Fig. 10. Additionally, β3 (βc) plays a supportive role in these interactions, as reflected in the violet-colored contacts in Fig. 10 (see also Fig. 11(D)).

106-Val and 122-Val are CHRnFs (see Table 6). However, a detailed examination of the 3D structure of 6CHA B and C reveals that the segment 96–128 form interactions with domain C, as shown in Fig. S17. Hence, these CHRnFs seem to contribute to the interactions between domains B and C, not within the 6CHA B domain. Although the peaks in the F-value plot of Fig. S17 suggest the tendency to form some interactions with other hydrophobic residues, these peaks do not mean the actual interactions within 6CHA domain B.

For α1-tryptase (1LTO), PdCRs are at residues 12–60 and 65–112 (Table 2), with the highest F-value peak at 38-Trp (CHRnF) in β4(bc) as presented in Fig. S6(C) (the predicted folding center). Hydrophobic contacts among β2(ba), β4(bc), and β5(bd) stabilize the four stranded partial b barrel structure and thus the first PdCR, while interactions between β6 and β7 stabilize the second

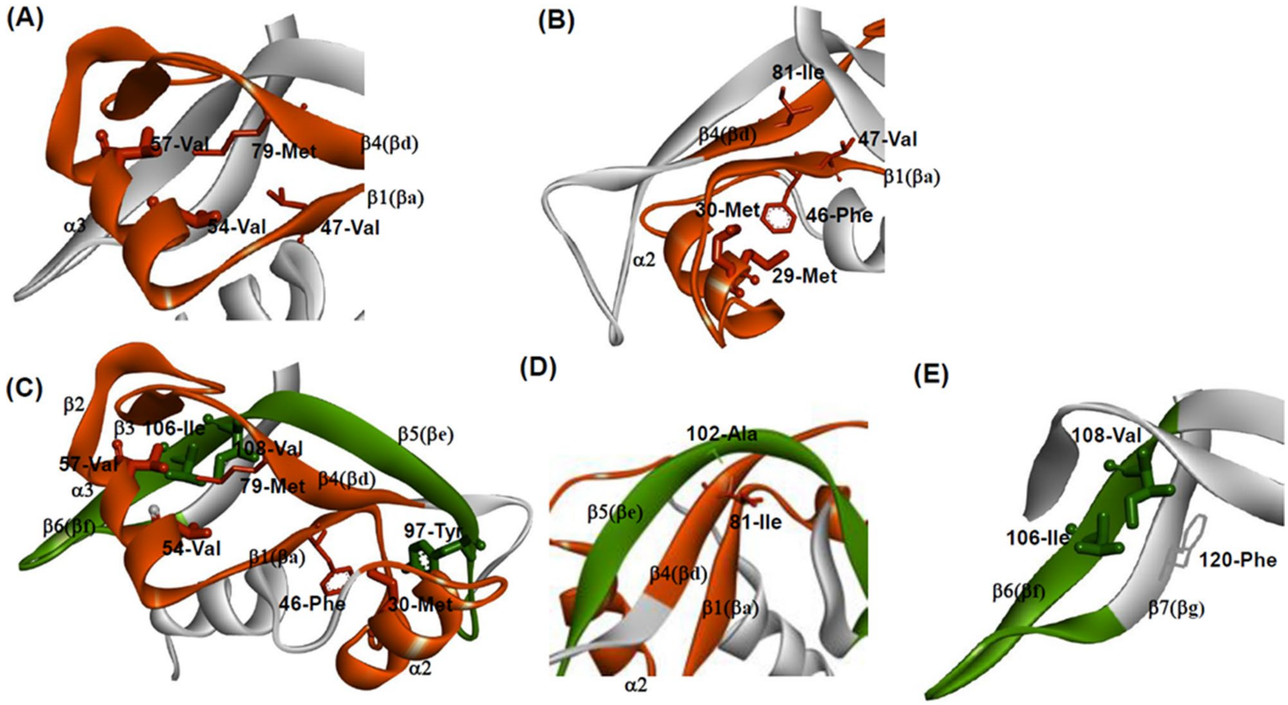

Fig. 9 (**A**) Hydrophobic packing formed by CHRnFs (indicated by a bold stick representation), 54-Val and 57-Val, and CHRs (indicated by a thin stick representation), 47-Val and 79-Met stabilizing the interactions among a3, b1 and b4 within the first PdCR in 6ETL. (**B**) Hydrophobic packing formed by CHRnFs, 29-Met and 30-Met, and CHR, 46-Phe stabilizing the interactions between a2 and b1 within the first PdCR. The interaction between CHRs, 47-Val and 81-Ile, stabilizes interaction between b1 and b4 within the first PdCR. (**C**) Hydrophobic packing formed by CHRnFs, 54-Val and 57-Val, 106-Ile and 108-Val stabilizing the interactions between a3 and b6, that is, the interaction between the first and second PdCRs. (**D**) 102-Ala makes a contact with 81-Ile and this contact contributes to the stabilization of the b-sheet between b4 and b5. 102-Ala is relatively close to a peak of the F-value plot at 108-Val (Fig. 5) separating by 6 residues. 102-Ala is not CHR but expected to be involved in the early stage of folding though it may be not so strong. (**E**) Hydrophobic packing is formed by CHRnFs, 106-Ile,108-Val and a hydrophobic residue 120-Phe within the second PdCR. 120-Phe, a residue which is not CHR nor CHRnF, is indicated by a line representation. The corresponding residue appears in both 2VQ9 and 2ZPO (See Figs. S9, S10, S11 and S12). It is considered that the 3D structure of the region 92–118 is stabilized mainly by the interaction with the region 19–84. That is, PdCR 92–118 is the case in Fig. 3(E); that is, the region 92–118 interacts with the major part of the other part from this region (see the legend of Fig. 9(E)). See the legend of Fig. 3(E)

PdCR (Figs. S13, S14(A)-(C)). The common structure is formed by interactions among β2(ba), β5(bc), β6(bd), and β7(bg), with β4(bc) supporting these contacts (Figs. S13, S14(D)).

In prostasin (3DFJ), PdCRs are located at residues 7–43 and 76–114, with a higher η-value in the second PdCR (Table 2). The second highest F-value peak in the first PdCR is at 36-Val (CHRnF) in β3(bc), with contacts among β2(bb), β3(bc), and β4(bd) stabilizing this region (Figs. S15, S16A-C). The CHRs around the F-value peak at 26-Cys and 36-Val is predicted as a folding center of this protein. The common structure is stabilized by interactions among β2(bb), β5(bd), β6(be-f), and β7(bg), with β4(bc) supporting these contacts as shown in Figs. S15 and S16, consistent with the patterns observed in 6CHA and 1LTO.

Discussion

In this study, we investigated the folding mechanisms of two distinct protein classes—RNase A-like fold proteins (α + β class) and trypsin-like serine proteases fold proteins (all β class)—utilizing a sequence-based analytical approach. Although these protein families exhibit very low sequence identity (as illustrated in Table 1), our analysis revealed striking structural similarities, particularly in their C-terminal regions. These regions, which feature two β hairpin-like structures, appear to play a crucial role in guiding the folding process. Our primary aim was to explore whether these common structural motifs arise from shared folding mechanisms, despite the proteins' divergent evolutionary origins. By applying sequence-based techniques, we identified potential conserved elements that contribute to the folding pathways of these seemingly disparate proteins, providing new insights into the underlying principles that govern their structural formation.

The sequence-based evolutionary analyses (Fig. S8) revealed strong conservation of hydrophobic residues across all the studied proteins, even though their sequence identities are low. Notably, these conserved hydrophobic residues (CHRs) are predominantly located on helices and β-strands.

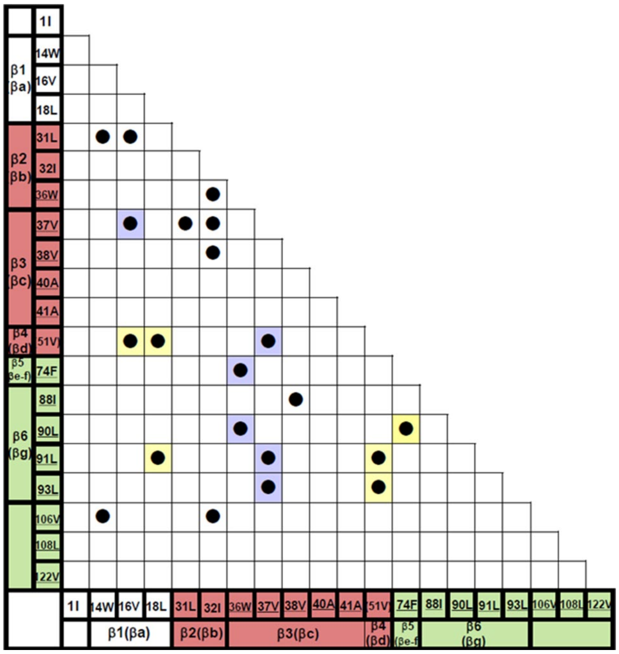

Fig. 10 Kind of contact map for 6CHA. Mainly conserved hydrophobic residues are presented. A red or green colored residue means a concerned residue is in a PdCR with the same color as in Fig. 6. The secondary structure containing a concerned residue is also indicated. A residue with an underline denotes CHRnF. A residue within a parenthesis is not CHR nor CHRnf. A contact colored by yellow means stabilization of a part of the common structure and a contact colored by violet means a contact between a residue in the common structure and b3(bc)

ADM analyses further showed that each protein possesses two PdCRs (Predicted Compact Regions), with the N-terminal regions generally exhibiting higher compactness than the C-terminal regions. Among the six proteins studied, four displayed higher compactness in their N-terminal regions, one (3DFJ) had similar η values in both regions, and only one (1LTO) exhibited greater compactness in the C-terminal region. This suggests that these proteins typically contain two folding units, with the N-terminal regions being more stable in the early stages of folding.

Interestingly, the C-terminal PdCRs consistently contain hairpin structures, regardless of whether the protein belongs to the RNase A-like fold or the trypsin-like serine proteases fold. For RNase A-like fold proteins, the N-terminal PdCRs are composed of three helices and two antiparallel β-sheets (e.g., βa-βd and βb-βc), as illustrated in Fig. 1. In contrast, for trypsin-like serine proteases fold proteins, the N-terminal PdCR corresponds to a four-stranded partial β-barrel stabilized by CHRnFs (conserved hydrophobic residues near F-value peaks) and CHRs. These CHRnFs and CHRs are structurally analogous across the three proteins studied, as shown in Figs. 11, S14, and S16.

We propose that CHRnFs are likely buried early in the folding process, forming folding cores in conjunction with other CHRs. In RNase A-like fold proteins, α3 contains several CHRnFs, which play a crucial role in stabilizing the common structure through interactions with βa, βd, βe, βf, and βg. In trypsin-like serine proteases fold proteins, βc serves a similar role, as illustrated in Figs. 8 and 10, S9, S11, S13, and S15.

Figure 12 provides a schematic representation of this process. Figures 12(a) and (b) illustrate the interactions between α3 in 6ETL and βc in 6CHA with βf and/or βg. The oval part in Fig. 12(c) represents the structural components α3, βb, and βc in RNase A-like fold proteins and βb and βc in trypsin-like serine proteases fold proteins. It is predicted that these interactions contribute to the formation of the β-sheet comprising βa, βd, βe, βf, and βg — the common structural motif shared by both protein classes.

Although the common structure does not constitute an autonomous or highly robust structural entity, it plays a significant role in the overall stability of these proteins by interacting with other structural elements. Consequently, while it may not qualify as a standalone supersecondary structure, we argue that such structural motifs deserve recognition and definition within the broader context of protein folding.

Conclusion

This study provides new insights into the folding mechanisms of RNase A-like fold proteins and trypsin-like serine proteases, two distinct classes of proteins with low sequence identity yet remarkable structural similarities. Our sequence-based analysis highlights a shared folding topology, particularly in the C-terminal β hairpin-like structures, suggesting a conserved mechanism that transcends their evolutionary divergence. Despite their classification into different SCOPe categories, these proteins demonstrate commonalities in their folding pathways, driven by key structural elements that stabilize the partial three-dimensional architecture. Understanding these conserved folding principles is crucial for bridging sequence-based predictions with experimental validation. While structural bioinformatics has made significant progress, our findings emphasize the importance of integrating sequence-derived insights with experimental folding studies to elucidate fundamental protein stability mechanisms. By exploring these conserved topologies, we contribute to a broader understanding of how evolutionary forces shape protein architecture, ultimately aiding in the design of novel proteins and therapeutic targets.

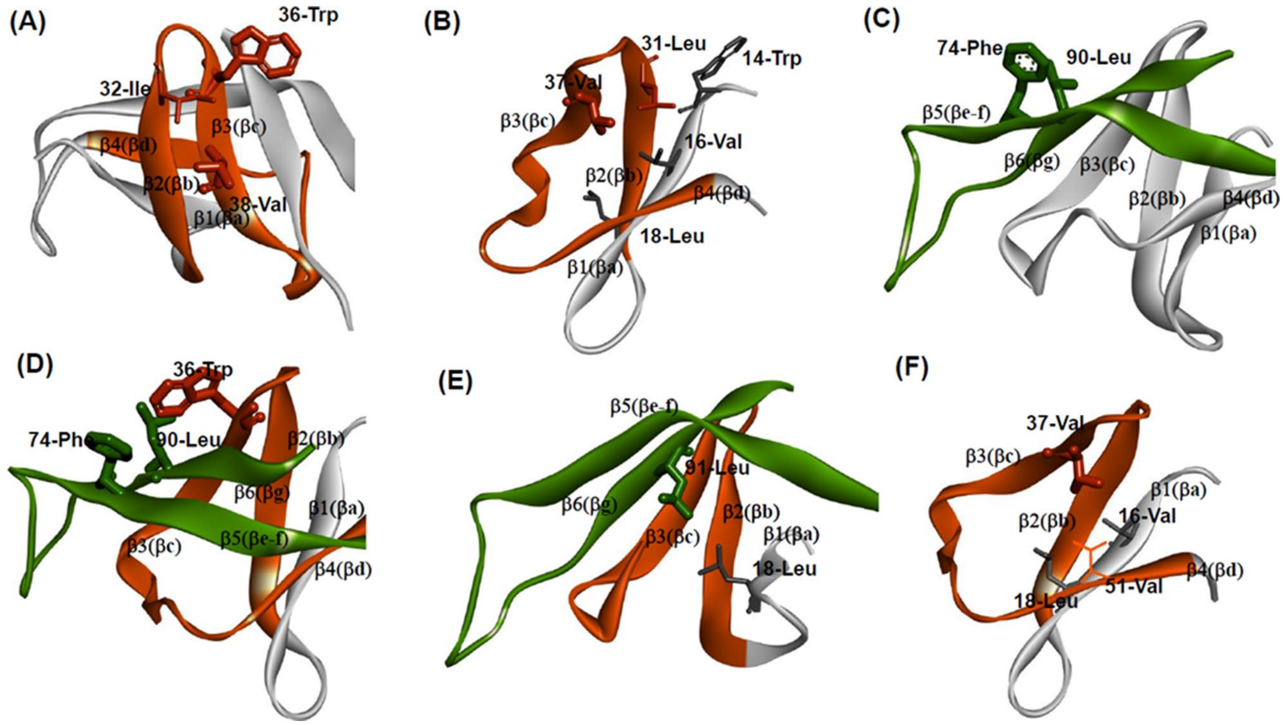

Fig. 11 (**A**) Hydrophobic packing formed by CHRnFs (indicated by a stick representation), 36-Trp and 38-Val, and CHR (indicated by a line representation), 32-Ile stabilizing the interactions between b2 and b3 within the first PdCR in 6CHA. (**B**) Hydrophobic packing formed by CHRnF, 37-Val, and CHR, 31-Leu stabilizing the interactions between b2 and b3 within the first PdCR. 31-Leu and 37-Val are also interacting hydrophobic residue, 14-Trp, 16-Val and 18-Leu stabilizing the interactions among b1, b2 and b3 within the first PdCR. (**C**) Hydrophobic packing formed by CHRnFs, 74-Phe and 90-Leu stabilizing the interactions between b5 and b6 within the second PdCR. (**D**) Hydrophobic packing formed by CHRnFs, 36-Trp, 74-Phe and 90-Leu stabilizing the interaction between the first and second PdCRs. (**E**) Hydrophobic packing formed by CHRnF, 91-Leu and 18-Leu stabilizing the interaction between b1 and b6. (**F**) Hydrophobic packing is formed by CHRnF, 37-Val, and hydrophobic residue, 51-Val stabilizing the interaction between b1 and b3. There is no CHRs in b4, and b4 seems not to be actively involved in the folding of the first PdCR. However, CHRnF, 37-Val makes a contact with 51-Val in b4. This contact stabilizes the interaction between b3(bc) and b4(bd) although this contact would not be so strong during the folding. Furthermore, CHRs, 16-Val and 18-Leu make hydrophobic contact with 51-Val indicating the stabilization of b1(ba) and b4(bd), that is, a part of the common structure

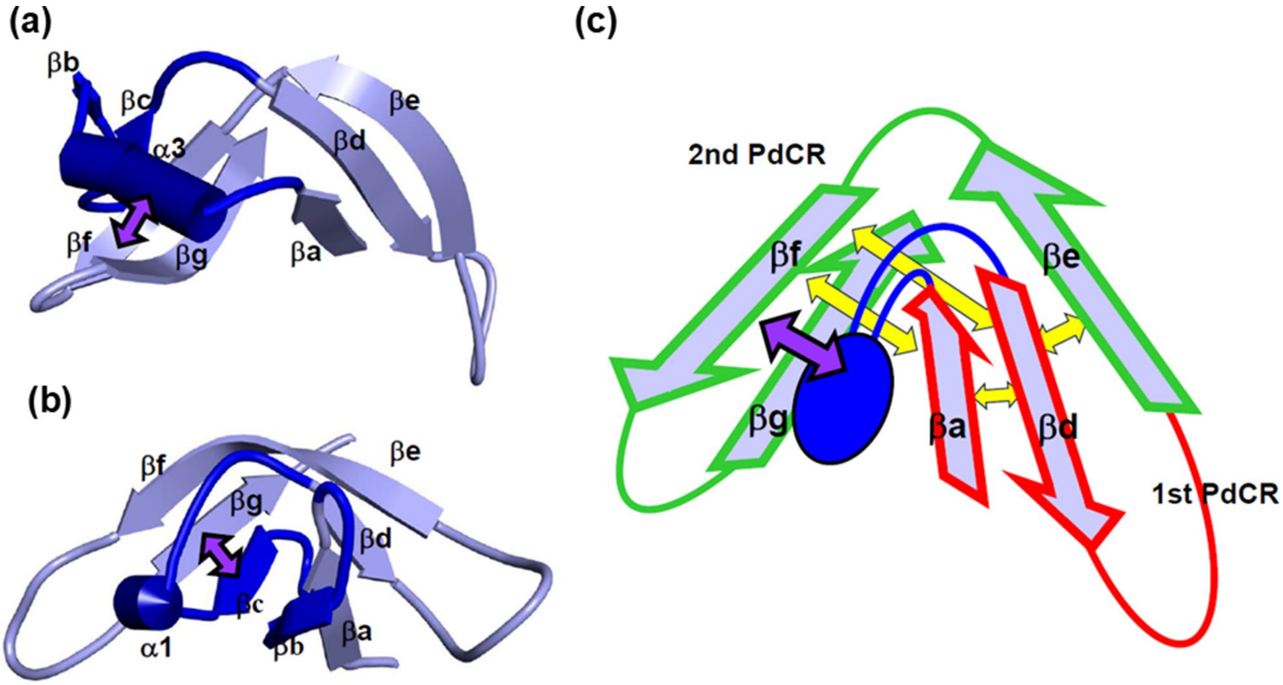

Fig. 12 The common structures appeared in 6ETL (Rnase A-like fold) (**a**) and 6CHA (trypsin-like serine proteases fold) (**b**). The region colored by light violet is a common structure. The deep blue part means a3, bb and bc in 6ETL and bb and bc in 6CHA. (**c**) Schematic drawing of the common structure with the region preceding part to bd (a3, bb and bc) in an RNase A-like fold protein and bb, bc and a1 in a trypsin-like serine proteases fold protein). An arrow denotes a b-strand. An arrow enclosed by a red or a green lines denote that this arrow is in the first or the second PdCRs respectively. The deep blue part indicates a3, bb and bc in an RNase A-like fold protein and bb, bc and a1 in a trypsin-like serine proteases fold protein. A yellow double arrow denotes an interaction between two b strands in the common structure. The violet double arrow denotes the interactions between the blue oval part and residues bf and/or bg. These interactions are considered to support the common structure

Abbreviations

ADM	Average Distance Map
CHR	Conserved Hydrophobic Residue
PdCR	Conservation of Predicted Compact Regions
CHRnF	Conserved Hydrophobic Residues near an F-value peak

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12860-025-00542-y.

Supplementary Material 1

Author contributions

TK conceived and designed the basis of this study. KMAK performed all the calculations and data analyses. TK and KMAK wrote the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by Japan Society for Promotion of Science as JSPS KAKENHI (Grant No. 21K12130).

Data availability

The datasets analyzed in this study were obtained from publicly available sources, specifically the Protein Data Bank (PDB). The computational scripts and workflows used for this study are available from the corresponding author upon reasonable request.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 27 September 2024 / Accepted: 19 May 2025
Published online: 28 May 2025

References

1. Dobson CM. Protein folding and misfolding. Nature. 2003;426(6968):884–90.
2. Whisstock JC, Lesk AM. Prediction of protein function from protein sequence and structure. Q Rev Biophys. 2003;36(3):307–40.
3. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D. Highly acculate protein structure prediction with alphafold. Nature. 2021;596(7973):583–9.
4. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Potapenko A, Bridgrand A, Meyer C, Kohl SAA, Ballard A, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P. Hassabis D Applying and improving AlphaFold at CASP14. Proteins. 2021;89(12):1711–1721.

5. Gao M, Zhou H, Skolnick J. DESTINI: A deep-learning approach to contact-driven protein structure prediction. Sci Rep. 2019;9:3514.
6. Chen S-J, Hassan M, Robert L, Jernigan RL, Jia K, Kihara D, Kloczkowski A, Sergei Kotelnikov S, Kozakov D, Liang J, Liwo A, Matysiak S, Meller J, Cristian Micheletti C, Mitchell JC, Sayantan Mondalo, Nussinov R, Okazaki K, Padhorny D, Skolnick J, Sosnick TR, Stan G, Vakser I, Zou X, Rose GD. Protein folds vs. protein folding: Differing questions, different challenges. Proc. Natl. Acad. Sci. 2023;120(1):e2214423119.
7. Levinthal C. Are there pathways for protein folding? J Chim Phys. 1968;65:44–5.
8. Dill KA, MacCallum JL. The protein-folding problem, 50 years on. Science. 2012;338(6110):1042–6.
9. Anfinsen CB. The formation and stabilization of protein structure. Biochem J. 1972;128(4):737–49.
10. Karplus M, Weaver DL. Protein-folding dynamics. Nature. 1976;260(5550):404–6.
11. Peran I, Holehouse AS, Carrico IS, Raleigh DP. Unfolded states under folding conditions accommodate sequence-specific conformational preferences with random coil-like dimensions. Proc. Natl. Acad. Sci. 2019;116(25):12301–12310.
12. Raines RT, Ribonuclease A. Chem Rev. 1998;98(3):1045–66.
13. White FH Jr., Anfinsen CB. Some relationships of structure to function in ribonuclease. Ann N Y Acad Sci. 1959;81(3):515–23.
14. Wilcox PE, Chymotrypsinogens—chymotrypsins. Methods Enz. 1970;19:64–108.
15. Berger A, Schechter I. Mapping the active site of Papain with the aid of peptide substrates and inhibitors. Philos Trans R Soc Lond B Biol Sci. 1970;257(813):249–64.
16. Kikuchi T. Analysis of 3D structural differences in the IgG-binding domains based on the interresidue average-distance statistics. Amino Acids. 2008;35(3):541–9.
17. Kawai Y, Matsuoka M, Kikuchi T. Analyses of protein sequences using inter-residue average distance statistics to study folding processes and the significance of their partial sequences. Prot Pep Lett. 2011;18(10):979–90.
18. Ichimaru T, Kikuchi T. Analysis of the differences in the folding kinetics of structurally homologous proteins based on predictions of the gross features of residue contacts. Proteins. 2003;51(4):515–30.
19. Nakajima S, A´lvarez-Salgado E, Kikuchi T, Arredondo-Peter R. Prediction of folding pathway and kinetics among plant hemoglobins using an average distance map method. Proteins. 2005;61(3):500–6.
20. Matsuoka M, Sugita M, Kikuchi T. Implication of the cause of differences in 3D structures of proteins with high sequence identity based on analyses of amino acid sequences and 3D structures. BMC Res Notes. 2014;7:654.
21. Matsuoka M, Fujita A, Kawai Y, Kikuchi T. Similar structures to the E-to-H helix unit in the globin-like fold are found in other helical folds. Biomolecules. 2014;4(1):268–88.
22. Aumpchin P, Kikuchi T. Prediction of folding mechanisms for Ig-like beta sandwich proteins based on inter-residue average distance statistics methods. Proteins. 2019;87(2):120-35.
23. Matsuoka M, Kikuhci T. Sequence analysis on the information of folding initiation segments in ferredoxin-like fold proteins. BMC Str Biol. 2014;14(15):15.
24. Kirioka T, Aumpuchin P, Kikuchi T. Detection of folding sites of β-trefoil fold proteins based on amino acid sequence analyses and structure-based sequence alignment. J Proteom Bioinf. 2017;10(9):222–35.
25. Kimura R, Aumpuchin P, Hamaue S, Shimomura T, Kikuchi T. Analyses of the folding sites of irregular β-trefoil fold proteins through sequence-based techniques and Gō-model simulations. BMC Mol Cell Biol. 2020;21(1):28.
26. Nakashima T, Kabata M, Kikuchi T. Properties of amino acid sequences of Lysozyme-Like superfamily proteins relating to their folding mechanisms. J Proteom Bioinf. 2017;10(4):94–107.
27. Nishimura C, Prytulla S, Dyson HJ, Wright PE. Conservation of folding pathways in evolutionarily distant globin sequences. Nat Struct Biol. 2000;7(8):679–86.
28. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. 2005;33(7):2302–9.
29. Kikuchi T, Némethy G, Scheraga HA. Prediction of the location of structural domains in globular proteins. J Protein Chem. 1988;7(4):427–71.
30. Dasmeh P, Serohijos AWR, Kepp KP, Shakhnovich EI. Positively selected sites in cetacean myoglobins contribute to protein stability. PLOS Comput Biol. 2013;9(3):e1002929.
31. Mirny L, Shakhnovich E. Evolutionary conservation of the folding nucleus. J Mol Biol. 2001;308(2):123–9.
32. Rorick MM, Wagner GP. Protein structural modularity and robustness are associated with evolvability. Gen Biol Evol. 2011;3:456–75.
33. Liao H, Yeh W, Chiang D, Jernigan RL, Lustig B. Protein sequence entropy is closely related to packing density and hydrophobicity. Prot Eng Des Sel. 2005;18(2):59–64.
34. Altschul SF, Gish W, Myers EE, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–10.
35. Consortium TU. UniProt: a worldwide hub of protein knowledge. Nucl Acids Res. 2018;47(D1):D506–15.
36. Kumar S, Stecher G, Tamura K. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol Biol Evol. 2016;33(7):1870–4.
37. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 1987;4(4):406–25.
38. Jones DT, Taylor WR, Thornton JM. The rapid generation of mutation data matrices from protein sequences. Bioinformatics. 1992;8(3):275–82.
39. Neira JL, Rico M. Folding studies on ribonuclease A, a model protein. Fold Des. 1997;2(1):R1–11.
40. Santoro J, González C, Bruix M, Nieto JL, Herranz J, Rico M. High-resolution three-dimensional structure of ribonuclease A in solution by nuclear magnetic resonance spectroscopy. J Mol Biol. 1993;229(3):722–34.
41. Wang A, Robertson AD, Bolen DW. Effects of a naturally occurring compatible osmolyte on the internal dynamics of ribonuclease A. Biochemistry. 1995;34(46):15096–104.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.