Special Issue
"Memorial Issue for Prof. Nobuhiko Saitô"

Vol. 13, pp. 295–304 (2016)
doi: 10.2142/biophysico.13.0_295

Biophysics and Physicobiology

https://www.jstage.jst.go.jp/browse/biophysico/

Regular Article

Itinerary profiling to analyze a large number of protein-folding trajectories

Motonori Ota[1], Mitsunori Ikeguchi[2] and Akinori Kidera[2]

[1]*Graduate School of Information Science, Nagoya University, Nagoya, Aichi 464-8601, Japan*
[2]*Graduate School of Medical Life Science, Yokohama City University, Yokohama, Kanagawa 230-0045, Japan*

Received August 1, 2016; accepted September 5, 2016

Understanding how proteins fold through a vast number of unfolded states is a major subject in the study of protein folding. Herein, we present itinerary profiling as a simple method to analyze molecular dynamics trajectories, and apply this method to Trp-cage. In itinerary profiling, structural clusters included in a trajectory are represented by a bit sequence, and a number of trajectories, as well as the structural clusters, can be compared and classified. As a consequence, the structural clusters that characterize the foldability of trajectories were able to be identified. The connections between the clusters were then illustrated as a network and the structural features of the clusters were examined. We found that in the true folding funnel, Trp-cage formed a left-handed main-chain topology and the Trp6 side-chain was located at the front of the main-chain ring, even in the initial unfolded states. In contrast, in the false folding funnel of the pseudo-native states, in which the Trp6 side-chain is upside down in the protein core, Trp-cage had a right-handed main-chain topology and the Trp side-chain was at the back. The initial topological partition, as determined by the main-chain handedness and the location of the Trp residue, predetermines Trp-cage foldability and the destination of the trajectory to the native state or the pseudo-native states.

Key words: Trp-cage, computer simulation, profiling, folding network, molecular dynamics

Corresponding author: Motonori Ota, Graduate School of Information Science, Nagoya University, Furo-cho, Chikusa-ku, Nagoya, Aichi 464-8601, Japan.
e-mail: mota@is.nagoya-u.ac.jp

The dynamics of protein folding is one of the most fundamental problems in the life sciences. Extensive efforts in both theoretical and experimental studies have been directed toward understanding how a protein folds into a unique structure from an enormously large number of unfolded conformations [1]. Owing to accelerated growth in computational power, simulation of the folding of mini-proteins, or fast-folding proteins has become feasible to provide detailed trajectories of the entire folding process at an atomic level [2–9]. However, there is still a difficulty with large-scale computations in processing the large amount of trajectory data that arises from complex dynamics occurring with large degrees of freedom [9].

Large amounts of trajectory data are usually treated by projection onto two- or three-dimensional space, spanned by the principal components [10,11] or by reaction coordinates [12], such as the radius of gyration or the proportion of native contacts [13], and the dynamical process is examined. However, such a drastic reduction in dimensions tends to remove the important features of the folding dynamics.

◀ ***Significance*** ▶

Recent advances in computer technology now enable the performance of large-scale molecular dynamics simulations. As a result, a large number of protein-folding trajectories can be generated to be analyzed in detail. We have developed an itinerary profiling method to process trajectory data. The profiling and network methods were applied to the Trp-cage data and the characteristic folding pathways were determined. It demonstrates that these techniques are powerful tools which can be used to decipher the complex dynamical data of bio-molecules.

Moreover, the folding process is highly complex and even chaotic [14], and sometimes a simple average operation does not yield a physically relevant result. To handle these spatio-temporal complexities, methods for analyzing folding trajectories have been proposed based on various concepts, including network [15], graph [16], alignment [17,18] and manifold [19].

In this study, we have investigated the entire folding process of a 20-residue mini-protein, Trp-cage, based on 200 molecular dynamics (MD) simulations (each of 50 ns) obtained previously [17]. To overcome the difficulty in analyzing a large amount of complex simulation data, we used a combination of two bioinformatic approaches, profiling and network. Profiling has frequently been employed in analyses of the large-scale data derived in comparative genomics (phylogenetic profiling) [20,21] and in transcriptome (gene expression profiling) research [22,23]. In folding trajectories, profiling reduces each trajectory to a bit sequence of existence (=1) and nonexistence (=0) of a certain class of structures in a trajectory. This reduction is achieved by discarding information about the structural details of the molecule, and the number and order of occurrence of the structures in the trajectory. This coarse-grained representation, or the itinerary profile, enables us to simplify and characterize trajectories with chaotic dynamics, including non-monotonous motions, e.g., recurring motion frequently observed in the unfolded structures. The classification of the itinerary profiles of the folding trajectories allows us to identify representative trajectories and characteristic structural clusters that are important in determining the entire folding process.

Network representation is a powerful technique to represent the connections among numerous elements, e.g., gene regulatory networks [24] and protein interaction networks [25]. In the analysis of folding trajectories, network representation has played an important role in depicting the folding pathways [8,15,26], where the vertices represent structural clusters and the edges represent the transitions occurring during the folding process. However, when the raw data of 200 trajectories are directly mapped on a network, it is difficult to derive anything remarkable from the labyrinthine network. To avoid this complication, we identified representative trajectories and characteristic structural clusters in the profiling process to examine a network, with a layout optimized by the Kamada-Kawai algorithm [27], to give a comprehensive overview of the structural transitions toward the folded structure.

The small protein Trp-cage was used a model protein to demonstrate the use of our method. A designed mini-protein, Trp-cage (NLYIQ WLKDG GPSSG RPPPS; PDB: 1L2Y [28]), is one of the fastest folding proteins with the folding speed of approximately 4 μs [29]. Trp-cage has been studied extensively by folding simulations, in both explicit [6,30–34] and implicit solvents [4,17,35–37], using MD simulations [4,6,17,35–37], the generalized ensemble method [33,34,38,39], and transition path sampling [30,31]. These studies in turn have stimulated experimental studies using fluorescence and NMR spectroscopy [29,40–42]. The formation of the secondary structure and the hydrophobic cage, as well as the burial of Trp6 have been proposed as the elementary folding steps of Trp-cage [17,31,32,40–42]. In simulations, two distinctive folded forms, the native structure (1L2Y, Fig. 1a) and pseudo-native structures (Fig. 1b), have been found, whether or not the two forms were distinguished explicitly [17,32,35]. The pseudo-native structures are less

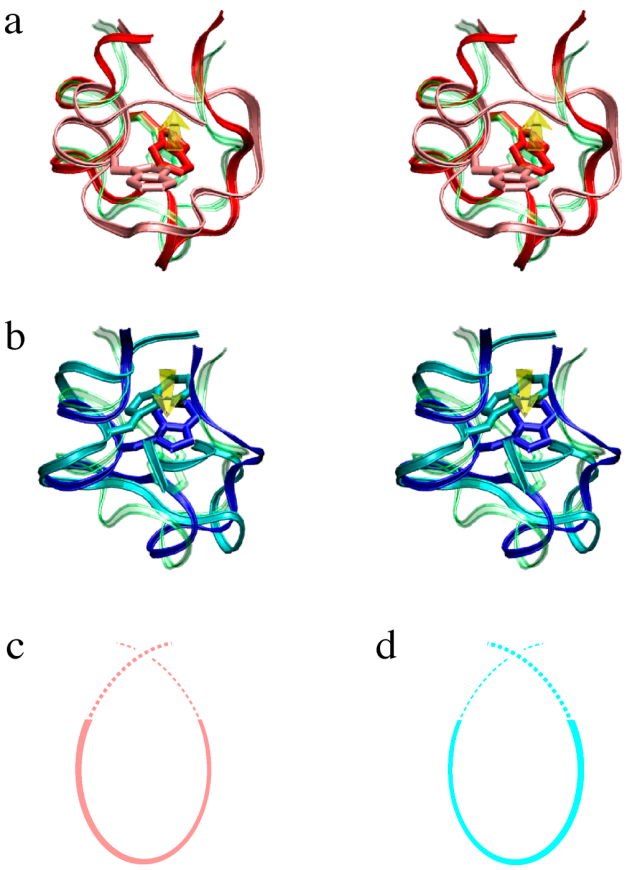

Figure 1 Final stage of the folding process of Trp-cage and its folded forms. (a) Stereo diagrams of the typical folding process toward the native structure observed in trajectory 126. In the unfolded structure (31.6 ns, pink) the main-chain forms a left-handed ring (c). Finally, the Trp6 side-chain enters the protein core from the front of the ring (the yellow arrow) and the left-handed ring becomes planar (32.4 ns, red). In the folded form, the position of the Trp6 residue, taking the rotamer type 4 ($|\chi 1|>120$ and $\chi 2>0$), overlaps well with that of the NMR structure (PDB: 1LY2, light green), where an α-helix (3–8), a G-helix (11–14), and a C-terminal poly-proline II helix constitute the hydrophobic cage that enwraps the central Trp residue. (b) Stereo diagrams of typical folding process toward a pseudo-native structure observed in trajectory 51. In the unfolded structure (33.2 ns, cyan) the main-chain forms a right-handed ring (d). Finally, Trp6 enters the protein core from the back of the ring (the yellow arrow) and the right-handed ring becomes a planar ring (33.4 ns, blue). In the folded form, the main-chain trace is similar to that of the NMR structure (light green), but Trp6 lies upside down in the protein core, compared with the NMR structure. (c) The left-handed ring. (d) The right-handed ring.

stable, and have the correct main-chain trace but a non-native Trp side-chain rotamer formed by turning the indole-ring upside down.

In a previous study [17], we analyzed the later stages of the folding process by the trajectory alignment method, in which folding trajectories (arrays of sequential snapshots of structures) were aligned by dynamic programming allowing gaps [43]. It was found that the two folded forms were attained through different pathways characterized by the main-chain and Trp side-chain motions. The Trp-cage protein first brings the chain termini together to form a ring-shaped structure. This ring shape is divided into two different types characterized by the handedness of the main-chain twist, i.e., either a left-hand (Fig. 1c) or right-hand (Fig. 1d) twisted conformation. In the final stage of folding, the Trp side-chain dives into the cage composed of the main-chain ring, from either the front (Fig. 1a) or the back (Fig. 1b) of the ring plane. Thus, the two pathways are characterized by the combination of main-chain twist and the motion of the Trp side-chain, i.e., left-hand/front and right-hand/back, implying that the motion of Trp6 is strongly correlated with the handedness of the main-chain twist. Furthermore, we found that the left-hand/front pathways led to the native structure (Fig. 1a), whereas the right-hand/back pathways resulted in the pseudo-native structures (Fig. 1b). Findings from the later stages of the folding process suggest that the folding funnel of Trp-cage is extremely rugged, and divided discretely by the handedness of the main-chain twist into pathways leading either to the native state or the pseudo-native states.

Herein, using profiling and network methods, we enhance this analysis, previously limited to the later stages of folding, to cover the entire folding process, including the initial state of fully denatured structures, and view the landscape discretized by the handedness of the main-chain twist.

Methods

Cluster analysis

A half million snapshots generated by 200×50 ns (10 μs in total) MD simulation [17], employing the modified AMBER99 force field [4] with the generalized Born implicit solvent [44] at 325 K, were divided into 2,036 clusters by the CD-hit like method [45]. We employed various root mean square deviation (RMSD) cutoffs (1.5–5.0 Å) according to the distance from the NMR structure [28] in the RMSD space (Supplementary Table S1). To calculate the RMSD between snapshots, we used 20 Cα atoms and the 3 Trp side-chain atoms, Cδ1, Cδ2, and Cε3. The 140 major clusters, in which the population was more than 0.1 %, were selected (Supplementary Table S1).

Folding event

At the folding event, the backbone RMSD from the NMR structure is less than 2.0 Å, and the averaged RMSD (400 ps) is less than 2.0 Å during the next 200 ps. The rotamer types of Trp6 were assigned for each snapshot by measuring the $\chi 1$ and $\chi 2$ angles [46]. Rotamer 4 ($|\chi 1|>120$ and $\chi 2>0$) corresponds to the native structure. Folding states were categorized into native or pseudo-native according to the rotamer type.

Itinerary profile

The itinerary profile is a bit sequence indicating whether a trajectory visits a structural cluster or not. The itinerary profiles for the 200 trajectories were represented in the form of a 140×200 matrix (see Fig. 2 in the Results Section). When the i-th structural cluster ($1 \le i \le 140$) is included in the itinerary of the j-th trajectory ($1 \le j \le 200$), the element (i, j) is marked by a dot. Each itinerary profile covers only the information of the first folding event along a 50 ns trajectory, starting from the fully extended structure and ending at the first folded structure attained, ignoring later unfolding events. If a trajectory does not reach a folded state, all data were employed in the itinerary profile.

The Matthew's correlation coefficient

To estimate the similarity of two bit sequences, the Matthew's correlation coefficient (MCC) was employed [47]. The arc-cosine was calculated to convert the correlation into the distance [48]. A perfect correlation (1) and anti-correlation (–1) correspond to 0 and 180, respectively. A dendrogram was obtained using the UPGMA method [49].

Network

All the networks were illustrated by the neato program in GraphVis (http://www.graphviz.org/) where the nodes are linked by the springs and the energy of this system was optimized by the Kamada-Kawai algorithm [27]. Basically, linked nodes are placed in the neighboring positions in the network.

The ring: front/back and left-handed/right-handed

The ring is a protein shape in which the N- and C-termini are placed close together [50]. The ring ratio was defined by the largest sequence separation between the close residue pair divided by the protein length. The close residue pair was the one with a Cα distance <7 Å. The ring ratio of the NMR structure is 0.85. Once the ring shape (defined by ring ratio ≥0.65) is formed, the ring plane and the front and back sides of the ring can be defined. The ring plane was defined by the following three points: the Cαs of Trp6 and Pro17, and the midpoint of the Cαs of Gly9 to Ser14. The position of the Trp side-chain was represented by the midpoint of Cδ2 and Cη2 and its location (d) from the ring plane was measured. The positive direction of d, corresponding to the front side of the ring, was determined by the corkscrew rule along the sequence (from the N to C termini). The handedness of the main-chain trace was determined by the projection of a vector from the Cα of Ser20 to the Cα of Asn1 onto the normal

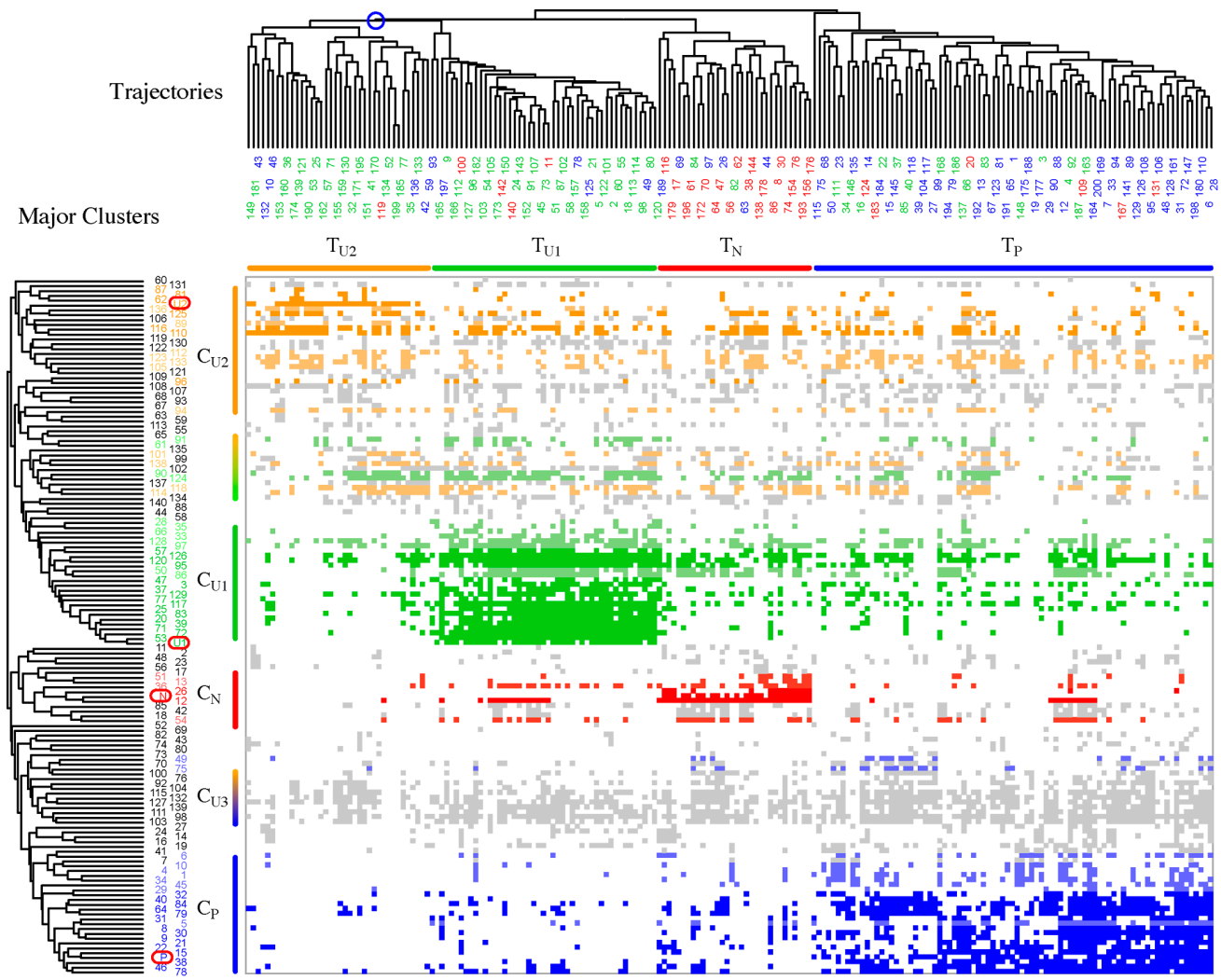

Figure 2 The itinerary profiles (central) and the dendrograms of trajectories (top) and major structural clusters (left). The red, blue, green numbers along the top of the itinerary profiles indicate the trajectories folding to the native state, those folding to the pseudo-native states, and the unfoldable trajectories, respectively. In the dendrogram, trajectories are divided into four groups T_N, T_P, T_{u1}, and T_{u2}, as indicated by the red, blue, green, and orange bars, respectively, on the top edge of the profiles. The branch indicated by the blue circle divides the unfoldable trajectories into T_{u1} and T_{u2}, The trajectories characteristic of the respective trajectory groups: 24 trajectories of T_N, 58 of T_P, 38 of T_{U1}, and 30 of T_{U2}, are represented by four row vectors. They can be considered as the representative trajectories for the four trajectory groups, and are denoted as N, P, U1, and U2, respectively, marked by the red circles at the left side. The structural clusters specific for a trajectory group are shown in dark (MCC≥0.5 for C_N, C_P, and C_{U1}. MCC≥0.2 for C_{U2}) and light (0.3 for C_N, C_P, and C_{U1}. 0.1 for C_{u2}) colored numbers at the left side. The red, blue, green, and orange numbers correspond to C_N, C_P, C_{U1}, and C_{U2}, respectively. The dots in the itinerary profiles are illustrated according to these colors. The color bars at the left side of matrix show the localization of specific structural clusters. Note that bars do not necessarily correspond to specific branches in the dendrogram.

vector of the ring plane. Positive and negative values correspond to left-handed and right-handed twists, respectively. The Trp side-chain position and the main-chain handedness in the NMR structure are given by 1.86 Å and 0.23, respectively.

Secondary structure formation

Using the DSSP program [51], secondary structures were assigned for the NMR structure [28] and each snapshot in the trajectories. We considered only α-helix (H), G-helix (G), and β-strand (E) structures and all other conformations were regarded as coils (C). The secondary structure of the native Trp-cage is "CCHHH HHHCC GGGGC CCCCC", and the α-helix (residues 3–8) and G-helix (residues 11–14) regions were focused on. For these regions, the match and mismatch between the NMR structure and the snapshot were evaluated. A mismatch of H/G was counted as a three-quarter match. The total match was normalized by the length of region.

Results and discussion

Itinerary profile

The itinerary profiles for the 200 trajectories of the 140 clusters are shown in Figure 2, in which the corresponding cells are colored if the trajectory contained the cluster. The 200 trajectories were compared and classified in terms of the MCC [47] between a pair of itinerary profiles. The results are illustrated in a dendrogram of the 200 trajectories (Fig. 2, top), in which the columns of the matrix (Fig. 2, center) were permuted according to the distances defined by the MCC. The dendrogram indicates that there are four distinctive groups, T_N, T_P, T_{U1}, and T_{U2}. These groups are named for the destination structure at the end of the itinerary, based on the fact that 24 out of 32 trajectories classified as T_N have the native structure as the destination structure and 58 out of 83 trajectories in T_P have the pseudo-native structures as the destination. Both T_{U1} and T_{U2} mostly consist of the trajectories that did not fold during the entire 50 ns simulation time (unfoldable trajectories), but they are separable at the branch marked by a blue circle in the dendrogram; 38 trajectories out of 47 in T_{U1} and 30 trajectories out of 38 in T_{U2} are unfoldable. This classification demonstrates that trajectories represented in the form of itinerary profiles exhibit a high correlation with the destination structure at the end of the itinerary. Such a correlation between the folding paths and the destination structures has already been found for the later stages of the folding process [17]. Here, the same relationship was found for the entire folding process. The four sets of trajectories that are characteristic of the respective trajectory groups, 24, 58, 38, and 30 trajectories of T_N, T_P, T_{U1}, and T_{U2}, respectively, were marked by dots to form four row vectors in Figure 2. In the itinerary profiles, these four rows indicate the representative trajectories for the four groups of trajectories, and were denoted as N, P, U1, and U2 (red circles at the left side in Fig. 2). They can also be regarded as if they were four structural clusters in addition to the 140 structural clusters listed on the left side.

In the same manner as for the classification of the trajectories, the 140 structural clusters, together with the most characteristic structural clusters, N, P, U1, and U2, were classified to construct a dendrogram (Fig. 2, left; the rows of the matrix were permuted according to the distances). To identify more clearly the correspondence between the trajectories and the structural clusters in this matrix, we defined the groups of structural clusters appearing exclusively in one of the four trajectory groups as C_N, C_P, C_{U1}, and C_{U2}, in terms of the MCC with the sets of the most characteristic structural clusters, N, P, U1, and U2, respectively (MCC$\geq$0.3 for C_N, C_P, and C_{U1} and $\geq$0.1 for C_{U2}; the lower criterion was used for C_{U2} because U2 contains various types of structures and there are very few structures belonging exclusively to U2). Clear localization of dense regions of C_N, C_P, and C_{U1} in Figure 2 indicates that the trajectories of T_N, T_P, and T_{U1} exclusively contain the cluster groups C_N, C_P, and C_{U1}, respectively. In other words, the intermediate structures in these cluster groups are fingerprints characterizing their respective destination structures. In contrast, the C_{U2} clusters are ubiquitously distributed over all trajectory groups. The cluster group C_{U2} is considered to represent the initial folding state, because the structural clusters in C_{U2} were found in the first part of the folding process of trajectories belonging to either T_N or T_P. This is more clearly represented in the network structure given in the next section. In addition, there are structural clusters forming a group, located between C_N and C_P (see the bars at the left side of matrix), collectively called C_{U3}, although C_{U3} does not show a unique correlation with any of the trajectory groups, and shares a similar profile with C_{U2}. In the next section, C_{U3} will be discussed in terms of the folding pathway.

Table 1 Transitions among the characteristic structural clusters

Cluster groups	C_N	C_P	C_{U1}	C_{U2}	C_{U3}
C_N (6)	0.93	–0.27	0.20	–0.68	–1.13
C_P (25)		0.50	–0.41	–0.65	0.00
C_{U1} (29)			0.26	–0.21	–0.10
C_{U2} (18)				0.36	0.01
C_{U3} (10)					0.14

The numbers in parentheses (in the left column) are the number of structural clusters in each group. The connection of cluster groups X and Y were evaluated by the frequency of the connections defined as, $\log_{10}\frac{n(X,Y)N_a}{n_a(X)n_a(Y)}$, where n(X,Y) is the number of edges between the cluster groups X and Y, $n_a(X)=\sum_Y n(X,Y)$ and $N_a=\sum_X n_a(X)$. n(X,Y) was counted in the non-redundant way, i.e., 0 or 1. Therefore, even if the transition of a given cluster pair was observed frequently, we only noted the transition was "present". C_{U3} was defined by a branch in which the structual clusters 76, 92, 104, 115, 132, 127, 139, 111, 98 and 103 were included (see Fig. 2 and Fig. 3a).

The frequency of the transitions within and between the cluster groups, C_N, C_P, C_{U1}, and C_{U2}, are given in Table 1 in the form of the logarithm of the ratio between the observed frequency of the transitions and the frequency of a random transition expected from the size of clusters. Positive values were observed mostly in diagonal elements indicating that the major parts of the folding pathways consist of transitions within the same structural clusters. On the off-diagonal, only the transition between C_{U1} and C_N occurred more frequently (0.20) than a random transition. These frequencies suggest that the route of the major folding pathway to the native state starts from the initial folding state C_{U2}, surmounts the rather high barrier between C_{U2} and C_{U1}, and finishes by the transition from C_{U1} to C_N, i.e., $C_{U2}\rightarrow C_{U1}\rightarrow C_N$. The transition between C_{U2} and C_P was not frequent (–0.65), suggesting that there is another intermediate state connecting C_{U2} and C_P.

Folding network

The detailed features of the transitions between the structural clusters can be illustrated more explicitly in a network

connecting the structural clusters in C_N, C_P, C_{U1} and C_{U2} and others (uncategorized clusters), which were optimally configured by the Kamada-Kawai algorithm [27] (Fig. 3a). The network is roughly divided into three regions. The largest region (Fig. 3a, the top dashed circle) includes abundant clusters of C_{U2} (orange), thus named "the C_{U2} region". The other two regions are those protruding from the C_{U2} region in Figure 3a; the left region (the left dashed circle) is characterized by C_P clusters (blue, "the C_P region") and the right region (the right dashed circle) mostly contains C_{U1} clusters (green "the C_{U1} region"). At the end of the C_{U1} region, C_N (red) and the native state (the red rectangular node labeled as "N") are situated. The pseudo-native states (blue rectangular nodes labeled by "P and a number"; where the number corresponds to the rotamer number of the side-chain of Trp6 [46]) are positioned at the end of the C_P region. This network structure is fully consistent with the transition frequencies given in Table 1, and confirms the major folding pathway, $C_{U2} \rightarrow C_{U1} \rightarrow C_N$.

Structural clusters of C_{U3} are localized between the C_{U2} and the C_P regions, as shown by the red dashed circle in Figure 3a. These clusters can be interpreted as the intermediate states connecting C_{U2} and C_P. The log ratios of the frequencies between C_{U3} and C_{U2} and between C_{U3} and C_P were calculated to be 0.01 and 0.00, respectively (Table 1). Although these values are in the order of that for a random transition, reflecting the non-specific character of the itinerary profile in C_{U3}, they are much larger than those among the other pairs of the cluster groups (the off-diagonal elements in Table 1), except for that between C_{U1} and C_N. Therefore, the major folding pathway to the pseudo-native structures can be said to be via C_{U3}, i.e., $C_{U2} \rightarrow C_{U3} \rightarrow C_P$. However, because of the C_{U2}-like randomness in C_{U3}, the role of C_{U3} in the network is not the same as that of C_{U1}, and thus it was concluded that the pathways to the native structure and the pseudo-native structures are not formed symmetrically in the folding network, with the former much more definitely determined than the latter.

It is possible in the network to view the structural features in the folding process. The very first event determining the destination of the folding pathway, going either to "N" or "P" nodes, is the differentiation of the main-chain handedness in the initial folding process. We evaluated the average degree of handedness for each structural cluster (Fig. 1; see the Methods Section for the definition). Figure 3b indicates that handedness clearly divides the clusters into two groups, corresponding to the true pathway going to the native structure (left-handed, red nodes) and the false pathway going to the pseudo-native structures (right-handed, blue nodes). It was noted that this separation had already occurred in the initial folding states in the C_{U2} region and was retained in the later folding stages. This means that the destination structure is determined at a very early stage of the folding through the handedness of the main-chain trace. To explain the above observation more quantitatively, we calculated the MCC of handedness (left-handed (=1) or right-handed (=0)) between two neighboring structural clusters appearing along a trajectory (the same and the reverse transitions were ignored). We found the correlation coefficient for structures in the C_{U2} region to be 0.20, although it is lower than the MCC for structures in the other regions, which have a MCC value of 0.52 on average. This implies that Trp-cage determines the handedness of its main-chain structure in the C_{U2} region, with allowance for only occasional changes. After the protein has attained C_{U1} or C_P, the handedness becomes largely fixed, leading to either the pathway to the native or the pseudo-native states.

Another characteristic feature of Trp-cage is the position of the side-chain of Trp6 relative to the surface formed by the ring-shaped structure of the main-chain, i.e., whether it is at the front or the back (Fig. 1; see the Methods Section for the definition), which strongly correlates with the main-chain handedness. In Figure 3c, the average positions of the side-chains of Trp6 for each cluster are plotted and the structural clusters can be seen to be clearly divided into two groups, according to whether the Trp is positioned front (red nodes) or back (blue nodes), which correlate well with the left-handed and right-handed conformations of the main-chain, respectively. This indicates that the position of the side-chain of Trp6 is another determinant of the destination of the folding, either to the native state or the pseudo-native states, in addition to the main-chain handedness. The MCC values of the side-chain position of Trp6 between neighboring structures appearing along a trajectory were found to be similar to those for handedness, 0.32 within the C_{U2} region, and 0.59 outside the C_{U2} region. In the later stages of folding, we found two typical pathways, left-hand/front leading to the native structure and right-hand/back leading to the pseudo-native structures [17]. Thus, the network analysis revealed that the separation of the two folding pathways starts at a very early stage of the folding.

Another event in the folding process, the formation of the α-helix at the N-terminal region (residues 3–8), also separates the structural clusters into two groups, but in a manner different from the handedness or the position of Trp6. The degree of formation of the N-terminal helix divides the clusters between those in the initial stage of folding (the C_{U2} region) and the other later stages (Fig. 3d). This indicates that the α-helix is formed after C_{U2}, independent of the destination structure [41]. The formation of the α-helix correlates with the radius of gyration (Fig. 3e) and the q-value (the ratio of the native contacts, Fig. 3f), suggesting that the factors separating the initial folding stage and the later stages are the α-helix formation [41] and the hydrophobic collapse [40].

A very late event in the folding process is the formation of the G-helix in the middle of the sequence (residues 11–14). Most of the clusters in the C_{U2} and C_{U1} regions do not contain the G-helix, but a few clusters around N and P nodes indicate some G-helix content (Fig. 3g).

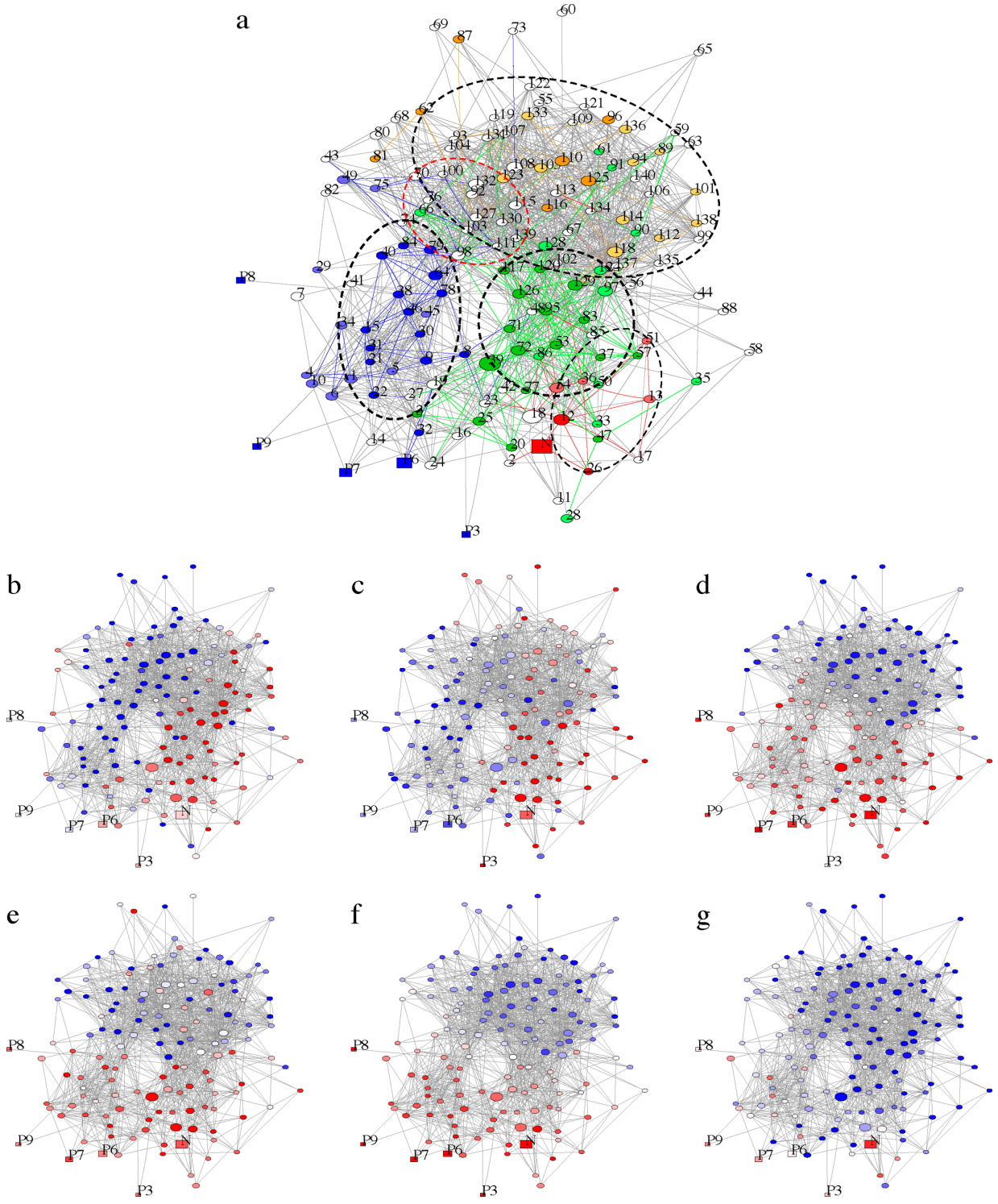

Figure 3 The folding network. (a) The network of structural clusters. C_N, C_P, C_{U1}, and C_{U2} are shown in red, blue, green, and orange, respectively. The specific edges for T_N, T_P, T_{U1} and T_{U2} are also shown in the same colors. We created bit sequences of edges (pairs of clusters) that show the corresponding trajectory includes the edge or not. Using these profiles of edges and N, P, U1, and U2, the correlated edges were determined as those with MCC≥0.2. (b–g) The network colored according to the structural features of each cluster (the average values of snapshots in the cluster). (b) The handedness of the main-chain [red: maximum value = 6.0 Å (left-handed), blue: minimum value = –6.0 Å (right-handed)]. (c) The position of the Trp side-chain [red: 3.5 Å (front), blue: –3.5 Å (back)]. The helix formation [red: 0.85 (about 5 of 6 residues formed on average), blue: 0.40 (2.4 of 6 residues formed)]. (e) The radius of gyration [red: 6.4 Å (compact), blue: 8.0 Å (expanded)]. (f) The q-value [red: 0.8 (abundant native contacts), blue: 0.3 (few native contacts)]. (g) The G-helix formation [red: 0.85 (3.4 of 4 residues formed), blue: 0.6 (2.4 of 4 residues formed)]. The coloring was empirically adjusted to highlight the folding process clearly.

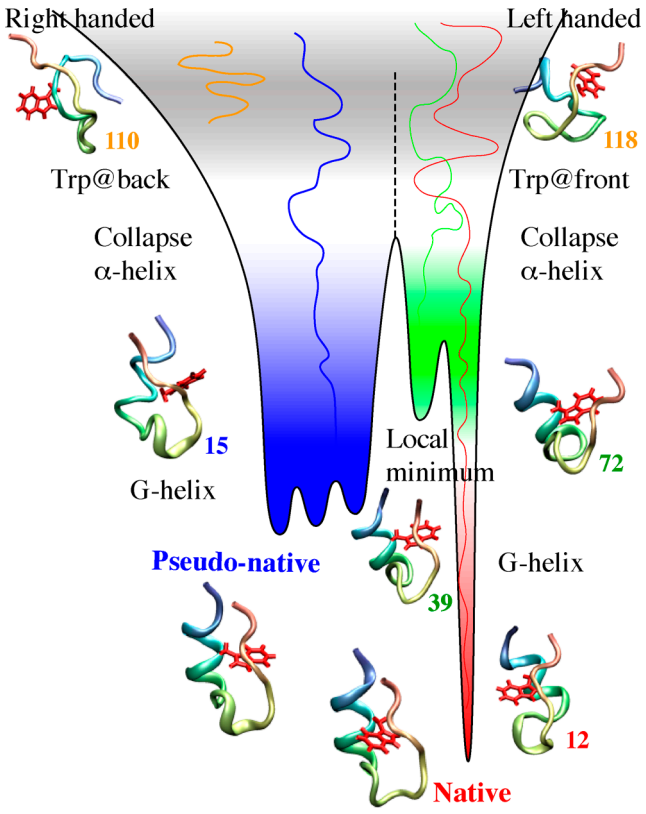

Figure 4 A schematic of the folding funnel [52]. The red, blue, green, and gray patterns indicate the regions of C_N, C_P, C_{U1}, and C_{U2}, respectively. The events occurring at the corresponding positions of the funnel are shown, illustrated with typical structures. The N- and the C-termini of the structures are shown in blue and pink, respectively, and the Trp side-chain is drawn as a red stick model. The vertical dashed line in the middle is the topological partition that divides the folding funnel into the true and the false funnels.

Folding funnel

The folding process of Trp-cage is summarized schematically in Figure 4 [52]. The first event in the folding is the construction of the ring-shaped topology of the main-chain, characterized by the main-chain handedness and the position of Trp6. At this stage, the destination of the folding, either to the native or the pseudo-native states, is determined. We called this separation of the folding pathways "true funnel" and "false funnel" for folding leading to the native and pseudo-native structures, respectively [17]. After the initial folding state, C_{U2}, the correct sequential events occur in the true folding funnel, i.e., the hydrophobic collapse and α-helix formation, followed by the G-helix formation, through $C_{U2} \rightarrow C_{U1} \rightarrow C_N$. At the later stages of folding to the native state, the side-chain of Trp6 dives into the protein core [28] from the front, reshaping the ring-shaped main-chain structure from the left-handed form to the native planar form. This sequence of events was not so clearly observed in the false funnel leading to the pseudo-native states, but folding occurred mainly through $C_{U2} \rightarrow C_{U3} \rightarrow C_P$. At the last stage, the side-chain of Trp6 enters into the core from the back.

Conclusions

We analyzed the entire folding process of Trp-cage, using a large amount of trajectory data [17]. To tackle the difficulty in analyzing a large amount of complex simulation data, we designed a method using itinerary profiling and a network to analyze the folding data. This method was applied to determine the characteristic folding pathways for Trp-cage (summarized in Fig. 4). These techniques are powerful tools which can be used to decipher the complex dynamical data of bio-molecules.

Acknowledgment

I (M. O.) was a student of late Prof. Nobuhiko Saitô, and am very grateful to him. Almost a quarter century ago, after completing M. S. degree, I joined the Tonen Corporation and started a theoretical study of proteins. At the time, Prof. Saitô had retired from Waseda University, but still continued with his research. I was kindly allowed to attend the progress meeting of his group, once or twice in a month, and learned the fundamentals of protein research, especially how to interpret data and evidence, and criticize papers. I learned the importance of hydrophobic interactions and chain entropy from him [53], and this principle was graven on my mind, and works as a compass to locate the direction of adventurous voyages across the oceans of protein research.

Conflicts of Interest

The authors declare no competing financial interest.

Author Contribution

M. O. conceived the research. M. O., M. I. and A. K. performed the research. M. O. and A. K. wrote the paper.

References

[1] Pain, R. H. *Mechanisms of Protein Folding (Frontiers in Molecular Biology)* (Oxford University Press, Oxford, 2000).
[2] Duan, Y. & Kollman, P. A. Pathway to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science* **282**, 740–744 (1998).
[3] Snow, C. D., Nguen, H., Pande, V. S. & Gruebele, M. Absolute comparison of simulated and experimental protein-folding dynamics. *Nature* **420**, 102–106 (2002).
[4] Simmerling, C., Strockbine, B. & Roitberg, A. E. All-atom structure prediction and folding simulations of a stable protein. *J. Am. Chem. Soc.* **124**, 11258–11259 (2002).
[5] Lei, H., Wu, C., Liu, H. & Duan, Y. Folding free-energy landscape of villin headpiece subdomain from molecular dynamics simulations. *Proc. Natl. Acad. Sci. USA* **104**, 4925–4930 (2007).
[6] Lindorff-Larsen, K., Piana, S., Dror, R. O. & Shaw, D. E. How fast-folding proteins fold. *Science* **334**, 517–520 (2011).
[7] Piana, S., Lindorff-Larsen, K. & Shaw, D. E. Protein folding kinetics and thermodynamics from atomistic simulation. *Proc.*

Natl. Acad. Sci. USA **109**, 17845–17850 (2012).
[8] Piana, S., Lindorff-Larsen, K. & Shaw, D. E. Atomic-level description of ubiquitin folding. *Proc. Natl. Acad. Sci. USA* **110**, 5915–5920 (2013).
[9] Lane, T. J., Shukla, D., Beauchamp, K. A. & Pande, V. S. To milliseconds and beyond: challenges in the simulation of protein folding. *Curr. Opin. Struct. Biol.* **23**, 58–65 (2013).
[10] de Groot, B. L., Daura, X., Mark, A. E. & Grubmüller, H. Essential dynamics of reversible peptide folding: memory-free conformational dynamics governed by internal hydrogen bonds. *J. Mol. Biol.* **309**, 299–313 (2001).
[11] Kamiya, N., Higo, J. & Nakamura, H. Conformational transition states of a β-hairpin peptide between the ordered and disordered conformations in explicit water. *Protein Sci.* **11**, 2297–2307 (2002).
[12] Best, R. B. & Hummer, G. Reaction coordinates and rates from transition paths. *Proc. Natl. Acad. Sci. USA* **102**, 6732–6737 (2005).
[13] Zhou, R. Exploring the protein folding free energy landscape: coupling replica exchange method with P3ME/RESPA algorithm. *J. Mol. Graph. Model.* **22**, 451–463 (2004).
[14] Rylance, G. J., Johnston, R. L., Matsunaga, Y., Li, C. B., Baba, A. & Komatsuzaki, T. Topographical complexity of multidimensional energy landscapes. *Proc. Natl. Acad. Sci. USA* **103**, 18551–18555 (2006).
[15] Rao, F. & Caflisch, A. The protein folding network. *J. Mol. Biol.* **342**, 299–306 (2004).
[16] Krivov, S. V. & Karplus, M. Hidden complexity of free energy surfaces for peptide (protein) folding. *Proc. Natl. Acad. Sci. USA* **101**, 14766–14770 (2004).
[17] Ota, M., Ikeguchi, M. & Kidera, A. Phylogeny of protein-folding trajectories reveals a unique pathway to native structure. *Proc. Natl. Acad. Sci. USA* **101**, 17658–17663 (2004).
[18] Oroguchi, T., Ikeguchi, M., Ota, M., Kuwajima, K. & Kidera, A. Unfolding pathways of goat alpha-lactalbumin as revealed in multiple alignment of molecular dynamics trajectories. *J. Mol. Biol.* **371**, 1354–1364 (2007).
[19] Das, P., Moll, M., Stamati, H., Kavraki, L. E. & Clementi, C. Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction. *Proc. Natl. Acad. Sci. USA* **103**, 9885–9890 (2006).
[20] Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. & Yeates, T. O. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. USA* **96**, 4285–4288 (1999).
[21] Yang, S., Doolittle, R. F. & Bourne, P. E. Phylogeny determined by protein domain content. *Proc. Natl. Acad. Sci. USA* **102**, 373–378 (2005).
[22] Armstrong, S. A., Staunton, J. E., Silverman, L. B., Pieters, R., den Boer, M. L., Minden, M. D., *et al. MLL* translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nat. Genet.* **30**, 41–47 (2001).
[23] Zhong, J. F., Zhao, Y., Sutton, S., Su, A., Zhan, Y., Zhu, L., *et al.* Gene expression profile of murine long-term reconstituting vs. short-term reconstituting hematopoietic stem cells. *Proc. Natl. Acad. Sci. USA* **102**, 2448–2453 (2005).
[24] Peter, I. S. & Davidson, E. H. A gene regulatory network controlling the embryonic specification of endoderm. *Nature* **474**, 635–639 (2011).
[25] Barabasi, A. L. & Oltvai, Z. N. Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* **5**, 101–113 (2004).
[26] Honda, S., Akiba, T., Kato, Y. S., Sawada, Y., Sekijima, M., Ishimura, M., *et al.* Crystal structure of a ten-amino acid protein. *J. Am. Chem. Soc.* **130**, 15327–15331 (2008).
[27] Kamada, T. & Kawai, S. An algorithm for drawing general undirected graphs. *Inf. Process. Lett.* **31**, 7–15 (1989).
[28] Neidigh, J., Fesinmeyer, R. & Andersen, N. Designing a 20-residue protein. *Nat. Struct. Biol.* **9**, 425–430 (2002).
[29] Qiu, L., Pabit, S. A., Roitberg, A. E. & Hagen, S. J. Smaller and faster: the 20-residue Trp-Cage protein folds in 4 μs. *J. Am. Chem. Soc.* **124**, 12952–12953 (2002).
[30] Juraszek, J. & Bolhuis, P. G. Sampling the multiple folding mechanisms of Trp-cage in explicit solvent. *Proc. Natl. Acad. Sci. USA* **103**, 15859–15864 (2006).
[31] Juraszek, J. & Bolhuis, P. G. Rate constant and reaction coordinate of Trp-cage folding in explicit water. *Biophys. J.* **95**, 4246–4257 (2008).
[32] Kannan, S. & Zacharias, M. Role of tryptophan side chain dynamics on the Trp-cage mini-protein folding studied by molecular dynamics simulations. *PLoS ONE* **9**, e88383 (2014).
[33] Paschek, D., Hempel, S. & Garcia, A. E. Computing the stability diagram of the Trp-cage miniprotein. *Proc. Natl. Acad. Sci. USA* **105**, 17754–17759 (2008).
[34] Zhou, R. Trp-cage: folding free energy landscape in explicit water. *Proc. Natl. Acad. Sci. USA* **100**, 13280–13285 (2003).
[35] Snow, C. D., Zagrovic, B. & Pande, V. S. The trp cage: folding kinetics and unfolded state topology via molecular dynamics simulations. *J. Am. Chem. Soc.* **124**, 14548–14549 (2002).
[36] Chowdhury, S., Lee, M. C. & Duan, Y. Characterizing the rate-limiting step of Trp-cage folding by all-atom molecular dynamics simulations. *J. Phys. Chem. B* **108**, 13855–13865 (2004).
[37] Chowdhury, S., Lee, M. C., Xiong, G. & Duan, Y. *Ab initio* folding simulation of the trp-cage mini-protein approaches NMR resolution. *J. Mol. Biol.* **327**, 711–717 (2003).
[38] Pitera, J. W. & Swope, W. Understanding folding and design: replica-exchange simulation of "trp-cage" miniproteins. *Proc. Natl. Acad. Sci. USA* **100**, 7587–7592 (2003).
[39] Nagasima, T., Kinjo, A. R., Mitsui, T. & Nishikawa, K. Wang-Landau molecular dynamics technique to search for low-energy conformational space of proteins. *Phys. Rev. E* **75**, 066706 (2007).
[40] Neuweiler, H., Doose, S. & Sauer, M. A microscopic view of miniprotein folding: enhanced folding efficiency through formation of an intermediate. *Proc. Natl. Acad. Sci. USA* **102**, 16650–16655 (2005).
[41] Culik, R. M., Serrano, A. L., Bunagan, M. R. & Gai, F. Achieving secondary structural resolution in kinetic measurements of protein folding: a case study of the folding mechanism of Trp-cage. *Angew. Chem. Int. Ed. Engl.* **50**, 10884–10887 (2011).
[42] Meuzelaar, H., Marino, K. A., Huerta-Viga, A., Panman, M. R., Smeenk, L. E., Kettelarij, A. J., *et al.* Folding dynamics of the Trp-cage miniprotein: evidence for a native-like intermediate from combined time-resolved vibrational spectroscopy and molecular dynamics simulations. *J. Phys. Chem. B* **117**, 11490–11501 (2013).
[43] Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453 (1970).
[44] Tsui, V. & Case, D. A. Theory and applications of the generalized Born solvent model in macromolecular simulations. *Biopolymers* **56**, 275–291 (2000).
[45] Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
[46] Dunbrack, R. L. & Karplus, M. Backbone-dependent rotamer library for proteins. Application to side-chain prediction. *J. Mol. Biol.* **230**, 543–574 (1992).
[47] Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* **405**, 442–451 (1975).

[48] Nagano, N., Ota, M. & Nishikawa, K. Strong hydrophobic nature of cysteine residues in proteins. *FEBS Lett.* **458**, 69–71 (1999).
[49] Sokal, R. R. & Michener, C. D. A statistical method for evaluating systematic relationships. *Univ. Kansas Sci. Bull.* **28**, 1409–1438 (1958).
[50] Koike, R., Kinoshita, K. & Kidera, A. Ring and zipper formation is the key to understanding the structural variety in all-β proteins. *FEBS Lett.* **533**, 9–13 (2003).
[51] Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).
[52] Wolynes, P. G., Onuchic, J. N. & Thirumalai, D. Navigating the folding route. *Science* **267**, 1619–1920 (1995).
[53] Saitô, N., Shigaki, T., Kobayashi, Y. & Yamamoto, M. Mechanism of protein folding: I. General considerations and refolding of myoglobin. *Proteins* **3**, 199–207 (1988).