Biophysical *Journal*
Article

Biophysical Society

Protein-Folding Analysis Using Features Obtained by Persistent Homology

Takashi Ichinomiya,[1,2,*] Ippei Obayashi,[3] and Yasuaki Hiraoka[4,3]
[1]Gifu University School of Medicine, Gifu, Japan; [2]The United Graduate School of Drug Discovery and Medical Information Sciences of Gifu University, Gifu, Japan; [3]Center for Advanced Intelligence Project, RIKEN, Tokyo, Japan; and [4]WPI-ASHBi, Kyoto University Institute for Advanced Study, Kyoto University, Kyoto, Japan

ABSTRACT Understanding the protein-folding process is an outstanding issue in biophysics; recent developments in molecular dynamics simulation have provided insights into this phenomenon. However, the large freedom of atomic motion hinders the understanding of this process. In this study, we applied persistent homology, an emerging method to analyze topological features in a data set, to reveal protein-folding dynamics. We developed a new, to our knowledge, method to characterize the protein structure based on persistent homology and applied this method to molecular dynamics simulations of chignolin. Using principle component analysis or nonnegative matrix factorization, our analysis method revealed two stable states and one saddle state, corresponding to the native, misfolded, and transition states, respectively. We also identified an unfolded state with slow dynamics in the reduced space. Our method serves as a promising tool to understand the protein-folding process.

SIGNIFICANCE To understand the protein-folding process, protein forms must be presented in a comprehensible way. In this article, we propose a method to represent the internal protein configuration using persistent homology, an emerging data analysis technique based on topology. Using this method, we simplified the complex dynamics of chignolin and identified two metastable and transition states as fixed points. Our method is applicable to other macromolecules and will help to understand the functions and dynamics of biomolecules such as proteins and DNA.

INTRODUCTION

Since the proposal of Levinthal's paradox in 1968, the folding of biomolecules, including proteins, has attracted the interest of numerous scientists (1). Molecular dynamics (MD) simulations have contributed to the understanding of the folding mechanisms (2). However, the atoms in the MD simulations have a large degree of freedom, and the essential folding dynamics must be extracted to comprehend the protein dynamics. Therefore, several methods have been proposed, such as principal component analysis (PCA) (3), relaxation mode analysis (4), time-structure-based independent component analysis (5,6), and manifold learning (7).

Previous studies attempted to identify the essential motion related to the large deformation that leads to protein folding. However, the definition of "large deformation" is ambiguous. For example, when a protein unfolds into nearly a straight line, a small bend at the center of the molecule will cause a large dislocation of the atoms at the end of the chain. In this case, the deformation in a Ramachandran plot (8) is small, but it is large in atom Cartesian coordinates. Moreover, the importance of deformations also depends on the protein structure. For example, in a small protein that has only one β-sheet, a small change in the bond angle at the hairpin of the molecule may disrupt the β-sheet structure. Thus, this small change in the angle results in a large deformation. Alternatively, if this protein is completely unfolded, then a slight change in the hairpin region bond angle does not cause a large deformation. These examples show the difficulties in defining a large deformation in a protein.

We propose using topological data analysis (TDA) to characterize the structure and deformation of a protein. Using TDA, we investigated the topological signatures such as loops or vacancies embedded in a data set. This approach yields successful results in many fields, including RNA-hairpin-folding analysis (9) or gene regulation networks (10). TDA has several advantages compared with standard protein structure analysis tools such as Ramachandran plots, distance matrices, and the atomic Cartesian coordinates. First, TDA captures changes in the global structure, whereas

Submitted October 28, 2019, and accepted for publication April 17, 2020.
*Correspondence: tk1miya@gifu-u.ac.jp
Editor: Alan Grossfield.
https://doi.org/10.1016/j.bpj.2020.04.032

other methods such as Ramachandran plots only consider local properties, such as bond angles. In contrast, "loops" or "vacancies" are formed by several atoms. Thus, TDA captures the nonlocal structure. Second, topological changes strongly depend on the atom conformation. For example, if a protein forms a straight chain, then there are no loops. If a small bend occurs at the center of this chain, then the atoms at the end of the chain exhibit large dislocations; however, loops do not form. Alternatively, a small change in the bond angle at the hairpin of a β-sheet can break the loops formed by the atoms in the β-sheet. Finally, TDA provides intuitive insights into protein dynamics. For example, loop emergence and disappearance are more clearly visualized using TDA than using the coordinated atomic motion.

Here, we applied persistent homology (PH) analysis (11) for TDA. PH is based on algebraic topology and has been applied to many problems in physics, chemistry, biology, and medicine (12–16). Although PH is a highly effective tool for the analysis of nonlocal structures, it has several inherent limitations. First, PH results are sometimes difficult to interpret. In the original PH analysis, we obtain two values called "birth" and "death" for each loop or cavity and make decisions based on the distribution of these values. Frequently, these two values are insufficient to understand the physical relevance provided by PH. For example, consider the folding of a protein that has two α-helices. If the birth and death values obtained from these α-helices are nearly identical, it is difficult to distinguish which α-helix is created first in the folding process. Recently, Escolar et al. developed a method to calculate "volume-optimal cycles," which enables identification of the atoms that form loops or cavities (17,18). This method is useful to explain PH results and has revealed hidden structures in glass and amorphous polymers (15,16). Another difficulty of PH lies in the fluctuation in the loop number. Even if the number of atoms is constant, the number of loops obtained by PH depends on the configuration of the atoms. However, standard machine-learning techniques such as PCA or k-means clustering require that all of the input data have the same dimension. These machine-learning techniques were avoided in previous studies using PH to analyze biomolecular structures (12,13). To overcome this difficulty, several methods such as persistent diagram vectorization (19), kernel methods (20), and persistent landscapes (21) have been proposed.

In this article, we propose a new, to our knowledge, technique to apply machine learning to PH analysis. The key concept is to construct a "topological feature vector" (TFV) using volume-optimal cycles. In this approach, we considered the volume-optimal cycles as the "text" that describes the protein structure. Each volume-optimal cycle is a collection of simplices (edges or faces), similar to a text being a collection of words. This concept enables the use of text-mining techniques. Next, we applied PCA and nonnegative matrix factorization (NMF) to reduce the TFVs obtained from MD simulations of chignolin. Finally, we compared the result with analyses based on atom-position and contact mapping. A previous study showed that chignolin has native, misfolded, unfolded, and intermediate structures (4,7). Therefore, we performed a full atomic MD simulation of chignolin in aqueous solution and analyzed the result using TFV. We observed that NMF of TFV provides essential information on protein structure and dynamics. Additionally, we found that the dynamics in the reduced space yielded two stable- and one saddle-fixed points, which correspond to native, misfolded, and transition states, respectively. The unfolded state did not correspond to a fixed point. However, the dynamics in the unfolded state were extremely slow.

The remainder of the article is structured as follows: in Methods, we describe the PH method, TFV construction, and dimension reduction by NMF. We also describe the details of the chignolin MD simulation. In Results and Discussion, we present the analysis results and compare them with analysis based on Cartesian coordinates and contact mapping. PH provides an intuitive description of the folded, misfolded, transition, and unfolded states. The challenges to overcome, as well as the future direction, are discussed in Conclusions.

METHODS

Our analysis process is composed of three procedures. First, we performed PH analysis and identified all loops with their volume-optimal cycles. Second, we constructed a TFV, which stores the edge contributions to the volume-optimal cycle formation. Third, we reduce the data set dimensionality using PCA or NMF. We explain each step in the following sections. The data set and scripts we used are uploaded on Open Science Framework: https://doi.org/10.17605/osf.io/hsp5w.

PH with volume-optimal cycles

The general mathematical definition of PH is described in terms of the filtration of simplicial complexes (11) or quiver representation (22). In this section, we explain the degree 1 PH of an α-complex composed of a point cloud, which was used to analyze protein folding.

Consider there are n atoms at $p_1 = (x_1, y_1, z_1)$, $p_2 = (x_2, y_2, z_2)$, …, $p_n = (x_n, y_n, z_n)$ in a three-dimensional space (Fig. 1). The PH of the α-complexes can be regarded as a topological structure when we place a ball of radius r at p_1, p_2, …, p_n. If $r = 0$, all of the balls are disconnected (Fig. 1 *a*). As we increase r, the balls coalesce, and a loop emerges at $r = b_1$ (Fig. 1 *b*). We call b_1 the "birth" of this loop, and the three edges, (p_3p_5), (p_3p_6), and (p_5p_6), surround this loop. This loop shrinks as r increases, and at $r = d_1$, the loop is fulfilled and disappears (Fig. 1 *c*). We call d_1 the "death" of this loop. In this case, the edges that surround the loop are unique; however, they are not always uniquely determined. For example, in the case of Fig. 1 *d*, the loop that emerged at $r = b_2$ is surrounded by five edges, (p_1p_2), (p_2p_4), (p_4p_6), (p_6p_3), and (p_3p_1), as depicted by the solid line. However, we can take another set of edges that surround this loop: (p_1p_2), (p_2p_4), (p_4p_6), (p_6p_5), (p_5p_3), and (p_3p_1), depicted as dashed lines. To avoid the ambiguity when defining the set of edges that surround the loop, the volume-optimal cycle is defined as the loop that has a minimal number of triangles inside. In our example, the first loop has five edges and can be divided into three triangles: ($p_1p_2p_4$), ($p_1p_4p_6$), and ($p_1p_6p_3$). Of course, there are many other

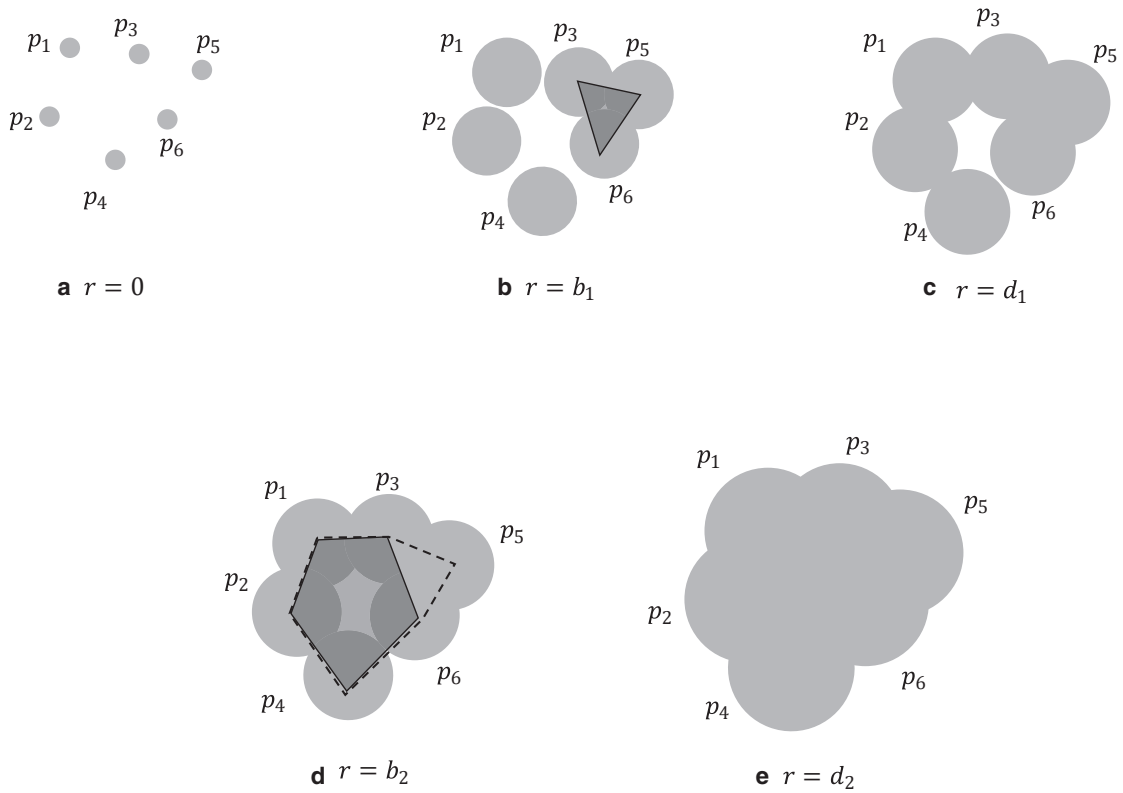

FIGURE 1 Example of PH analysis. If the radii of the balls are 0, then all atoms are disconnected (*a*). As we increase the size, the balls coalesce. At $r = b_1$, we obtain a loop surrounded by three edges: (p_3p_5), (p_3p_6), and (p_5p_6) (*b*). This loop is destroyed when we increase the size of the balls to $r = d_1$ (*c*). If we increase r further, a new loop appears (*d*). In this case, we can take several sets of edges that surround the empty space, depicted as solid and dashed lines. In this case, the volume-optimal cycle is (p_1p_2), (p_2p_4), (p_4p_6), (p_6p_3), and (p_3p_1), depicted by solid lines. This loop is destroyed at $r = d_2$ (*e*).

ways to decompose this loop into triangles; however, the number of triangles is uniquely determined. The second loop consists of six edges, and four triangles are needed to construct this loop. Therefore, we choose the first loop structure as the volume-optimal cycle.

We frequently designate loops that emerge in PH as "generators." From one atomic conformation, we obtain several generators. In PH, the generator birth and death distributions give important insights into the data set structure. To visualize this distribution, we use the scatter plot of births and deaths, which is called a persistence diagram. Another visualization method is persistent barcodes, in which horizontal lines represent generator births and deaths. We will present several barcode plots in Results and Discussion.

As we have mentioned in the Introduction, the number of generators strongly depends on the atom configuration. Even if the number of atoms is the same, the number of generators can differ. This fact makes it difficult to combine machine-learning techniques with PH. We attempted to overcome this challenge by introducing a TFV composed of the volume-optimal cycles (see below). The calculation of births, deaths, and volume-optimal cycles was performed by HomCloud ver.1.2.1 (https://www.wpi-aimr.tohoku.ac.jp/hiraoka_labo/homcloud/).

PH is strongly related to Betti numbers, which are topological invariants in mathematics. In topology, the k-th Betti number is defined as the rank of k-th homology groups. In our case, $k = 1$, it is the number of loops. Therefore, if we put balls with radius r at $p_1, p_2, \ldots, p_n$, the first Betti number of this set is the number of generators whose births and deaths are smaller and larger than r, respectively.

Before concluding this subsection, we discuss the use of higher-degree PH. In homology, "degree" is the dimension of "boundaries," and the PH with degree 2 is used to investigate the vacancies surrounded by triangles. Degree 2 PH often plays an important role in material science because it provides information on the voids. However, Xia and Wei found that PH with degree 2 gives little information on protein structure (12). Further, they revealed that both α-helices and β-sheets provide no void when analyzing C_α atoms as a point cloud. The native chignolin structure contains only one β-sheet and no tertiary structure. Thus, we choose to ignore higher-degree PH in this study.

Construction of TFV

Using the loop information, we defined a TFV, $\boldsymbol{v}$, which describes the point cloud topology. First, for each edge $E = (p_ip_j)$, we listed the generators g_k, whose volume-optimal cycles include E. We then calculated the "importance" of the edge E as the sum of the deaths of g_k. If an edge was not included in any volume-optimal cycles, we set the edge "importance" as 0 (Fig. 2). By this method, we obtained the TFV, whose dimension is $M = n(n-1)/2$.

Feature vector construction was similar to the "bag of words" and "term-frequency-inverse document frequency" methods, which are standard methods used in natural language processing (23). These methods regard a document as a multiset of terms and calculate the importance for each term. In our approach, edges were defined as the terms that describe the protein shape.

There are several possible methods to construct a TFV from the volume-optimal cycles. For example, we could create another TFV using births instead of deaths. Lifetime, the difference between death and birth, is also often used as an important PH variable. In this study, we examined the results of birth-based and death-based TFV. These qualitatively yielded the same result. Alternatively, we could use the products of deaths instead of sums of deaths. In this study, we used the sum of deaths for simplicity.

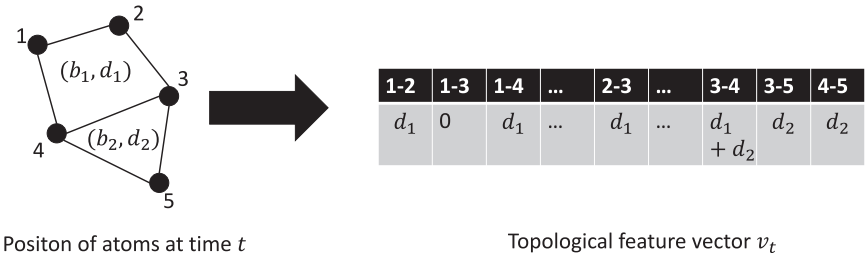

FIGURE 2 Schematic description of the TFV.

Indeed, there may be more a complex and sophisticated definition of the TFV (see Conclusions).

Dimension reduction by nonnegative matrix factorization

The dimensions of a TFV are generally high, making dimension reduction using PCA or another method useful. We primarily employed NMF to reduce the dimensionality of TFV (24). We assumed that the M-dimensional TFVs at time $t = t_1, t_2, \ldots, t_N$ are $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_N$, where $\boldsymbol{v}_i = (v_{1i}, v_{2i}, \ldots, v_{Mi})^t$ and $(\ldots)^t$ represents the transverse of the matrix, respectively. In NMF, we attempted to reduce the system into L-dimensional space, under the assumption that both coefficients and bases are nonnegative. We calculated $M \times L$ nonnegative matrix $\boldsymbol{W} = (w_{ij})$ and an $L \times N$ nonnegative matrix $\boldsymbol{H} = (h_{ij})$ that minimized $||\boldsymbol{V} - \boldsymbol{WH}||$, where $||\ldots||$ represents the Frobenius norm. Using this method, we can approximate $\boldsymbol{v}_i \approx \sum_{k=1}^{L} \boldsymbol{w}_k h_{ki}$, where $\boldsymbol{w}_k = (w_{1k}, w_{k2}, \ldots, w_{Mk})^t$ are the bases of the reduced space.

Compared with PCA, NMF has several advantages. First, when we reconstructed TFVs from the information in reduced spaces, NMF consistently generated nonnegative vectors. Both NMF and PCA attempt to approximate the feature vector $\boldsymbol{v}$ by the linear combination of several bases vectors $\boldsymbol{e}_i$: $\boldsymbol{v} \approx \sum c_i \boldsymbol{e}_i$. In NMF, we set $c_i \geq 0$ and $\boldsymbol{e}_i$ to be nonnegative, and approximated $\boldsymbol{v}^i$ was also nonnegative. Conversely, certain components in $\sum c_i \boldsymbol{e}_i$ can be negative in PCA. When $\boldsymbol{v}$ is defined as a nonnegative vector, understanding large negative components in $\sum c_i \boldsymbol{e}_i$ is difficult. Another advantage of NMF is that the bases can capture important local features. Though there is no theoretical explanation, the application of NMF to face-recognition problems shows that NMF can extract localized characteristics such as noses or eyes, whereas PCA captures nonlocal structures (24). In the case of protein-folding analysis, the ability of NMF to capture local structure is desirable. For example, we considered the folding of proteins with several secondary α-helix and β-sheet structures. In this case, it is natural to assume that the secondary structure formation does not occur simultaneously. In NMF, we expected several bases to represent secondary structures. However, if we used PCA for decomposition, each basis represents the complex structural change such as the disappearance of several helices and appearance of several sheets. Therefore, we need further investigation to understand the formation of these structures.

Though useful, NMF has several disadvantages. First, the NMF decomposition is not unique. Suppose that $\boldsymbol{W}$ and $\boldsymbol{H}$ are nonnegative matrices. If both $\boldsymbol{A}$ and $\boldsymbol{A}^{-1}$ are nonnegative matrices, then $\boldsymbol{W}' = \boldsymbol{WA}$ and $\boldsymbol{H}' = \boldsymbol{A}^{-1}\boldsymbol{H}$ are nonnegative, and we obtain another decomposition $\boldsymbol{V} \approx \boldsymbol{W}'\boldsymbol{H}'$. In practice, when the feature matrix is sparse and we initialize $\boldsymbol{W}$ and $\boldsymbol{H}$ by a nonnegative double singular value decomposition, then optimization with a coordinate descent solver generally yields small residue $||\boldsymbol{V} - \boldsymbol{WH}||$ with low computational costs (25). This method is deterministic and free from the problem caused by the nonuniqueness of NMF decomposition. Because our feature vector is sparse, we applied this initialization and optimization method. NMF also has the ambiguity of "scales." We can "rescale" the basis $\boldsymbol{w}_i$: $\boldsymbol{w}'_i = \alpha_i \boldsymbol{w}_i$ and $h'_{ij} = \alpha_i^{-1} h_{ij}$, where α_i-values are positive constants, which provides another decomposition. Here, we scaled $\boldsymbol{w}$-values so that $||\boldsymbol{w}|| = 1$. Thus, $\boldsymbol{w}$ can be assumed as a dimensionless vector, and h_{ij} has the same dimension as births and deaths.

Another disadvantage of NMF is the need to determine the rank of reduced space L a priori. Though there is no de facto standard to estimate rank L, several methods are proposed (26,27). In our study, we used the method proposed by Hutchins et al. (27), who showed that if the data set is random, the residual sums of squares (RSS) between $\boldsymbol{V}$ and $\boldsymbol{WH}$ decreases linearly with rank r, and proposed to use L at the inflection point. We performed these calculations using scikit-learn 0.19.1 and NMF 0.21.0 (28,29).

Chignolin molecular dynamics simulation

Using the method described in Mitsutake and Takano (4), we conducted MD simulations of aqueous chignolin near a transition temperature. We placed one chignolin molecule, two Na^+ atoms, and 3674 H_2O molecules in a cube and set the temperature and pressure at 450 K and 1 atm, respectively. After energy minimization and equilibration for 50 ns, we performed a 1-μs NPT-constant MD simulation. We captured snapshots of the molecules every 10 ps to create 100,000 samples. In this simulation, we used the ff99SB force field and TIP3P models for the water molecules. From each snapshot, we obtained the coordinates of 10 C_α atoms in chignolin and performed the PH analysis. The simulation was conducted using GROMACS 16.4 (30).

RESULTS AND DISCUSSION

In this study, we performed PH analysis of a point cloud composed of 10 C_α-atoms. We calculated the TFVs from snapshots of chignolin and reduced the configuration into low dimensional spaces by PCA and NMF.

Analysis using TFV

To carry out NMF analysis, we first determined the rank of reduced space. To determine the rank, we calculated RSS to determine the rank of reduced space L. To reduce the computational cost, we randomly selected 1000 samples from our data set and carried out NMF for $L = 1, 2, \ldots, 10$. The obtained RSS is shown in Fig. 3. In Fig. 3 *a*, when we used births to construct the TFV, the RSS rapidly decreased as L increased from 1 to 3 and slowly decreased for $L > 3$. This result indicates that $L = 2$ or 3 are the best reduced space ranks. When we used deaths to construct the TFV, the RSS shown in Fig. 3 *b* was obtained, again suggesting that $L = 2$ or 3 is the best rank of reduced space.

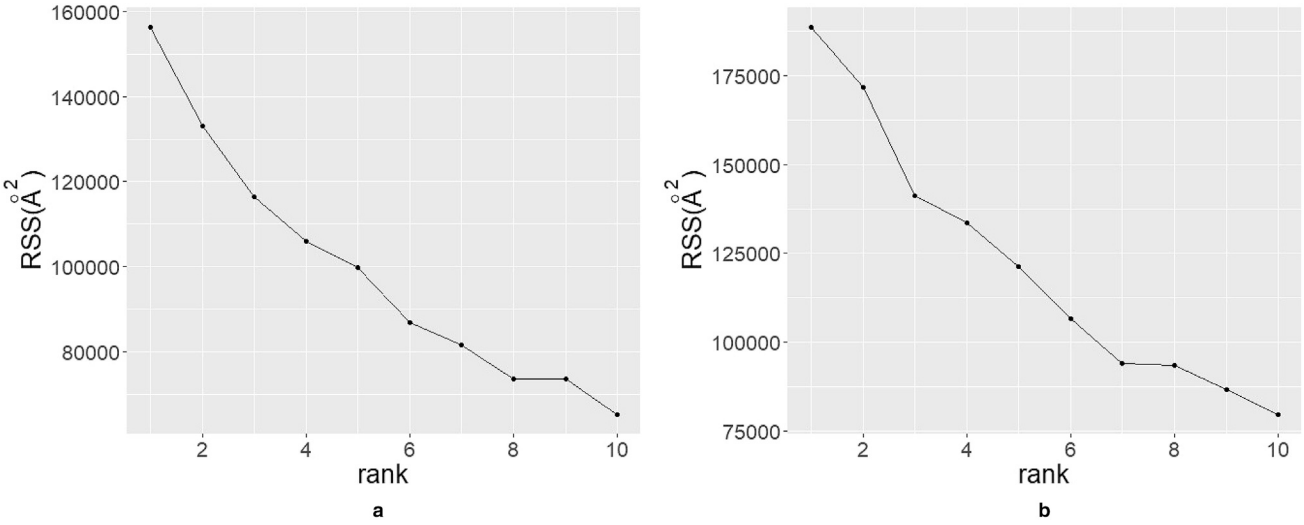

FIGURE 3 Plot of RSS for rank $L = 1, 2, \ldots, 10$: (*a*) TFV is constructed from births. (*b*) TFV is constructed from deaths.

To investigate the effect caused by changing L, we compared the results obtained by TFVs constructed from births and deaths for $L = 2$, 3, and 4. Fig. 4 represents the dynamics in the reduced space for $t = 0$ to 100 ns. In this figure, h_k at time t represents the value of h_{ki}, where i is the TFV index obtained from the snapshot at time t. Fig. 4 *a* shows the dynamics when $L = 2$, and births were used to construct the TFV. Clearly, there are two phases: the first shows that 10 Å $\lesssim h_1 \lesssim$ 30 Å, whereas $h_2 \lesssim$ 10 Å; the other shows that 10 Å $\lesssim h_2 \lesssim$ 30 Å, whereas $h_1 \lesssim$ 10 Å. We also noted that short periods occur in which both 5 Å $\lesssim h_1, h_2 \lesssim$ 15 Å. Therefore, it seems that there are two or three phases. This result is not modified when deaths are used instead of births to define TFV, as shown in Fig. 4 *b*. The correlation between h_1 in Fig. 4, *a* and *b* was 0.9967, and the correlation between h_2 was 0.9968. Fig. 4, *c* and *d* show the plots when $L = 3$. When we compared Fig. 4, *a* and *c*, we found that at the second phase in Fig. 4 *a*, in which $h_1 \lesssim$ 10 Å and 10 Å $\lesssim h_2 \lesssim$ 30 Å, h_2 in Fig. 4 *c* is large: 10 Å $\lesssim h_2 <$ 30 Å, whereas $h_1, h_3 \lesssim$ 10 Å. Additionally, when 10 Å $\lesssim h_1 \lesssim$ 30 Å in Fig. 4 *a*, $h_2 \lesssim$ 10 Å in Fig. 4 *c*, although no clear difference was observed between h_1 and h_3. Third, when 5 Å $\lesssim h_1, h_2 \lesssim$ 15 Å in Fig. 4 *a*, 10 Å $\lesssim h_1 \lesssim$ 30 Å, whereas $h_2, h_3 \lesssim$ 10 Å in Fig. 4 *c*. However, the third observation seems controversial because of the short phase duration. From these observations, we found no additional stable phase by increasing the rank L from 2 to 3. However, the analysis for $L = 3$ may be useful to understand the details of the phase in which both h_1 and h_2 are small in Fig. 4 *a*. These observations are not modified when we set rank $L = 4$, shown in Fig. 4, *e* and *f*. Thus, we used rank $L = 2$ and deaths to construct TFVs.

Fig. 5 *a* shows the TFV distribution in the reduced space by NMF, $L = 2$. We observed two large peaks at (h_1, h_2) ∼(25 Å, 0 Å) and (0 Å, 25 Å), which we designated region A and B, respectively. The density was high along the straight line connecting these peaks, with the minimum at (h_1, h_2) ∼(10 Å, 15 Å). We also noted that the density in the area around (h_1, h_2) ∼(5 Å, 0 Å) was high. This result is consistent with the results obtained by PCA of TFVs shown in Fig. 5 *b*. Here, we clearly identified two clusters, which are distinguished by first principal components. The correlation between first principle component and h_1, h_2 are 0.941 and −0.927, respectively. Therefore, the right and left clusters in Fig. 5 *b* correspond to clusters at region A and B in Fig. 5 *a*, respectively. Though NMF and PCA gave qualitatively similar results, we only discuss the results of NMF in the following sections of this article because they more clearly reveal protein structures, whereas PCA indicated only the "difference" of these two clusters. As described in the Methods, we found negative components when we reconstructed the TFV from PCA, which exacerbates the difficulty in understanding the protein structure. The NMF result in Fig. 5 *a* indicates that the peaks of clusters are nearly on the h_1 and h_2 axes, respectively. Thus, we can infer the "typical" chignolin TFVs in each cluster by checking $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ directly.

To investigate the structures at the two density peaks, we examined the bases $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ obtained by NMF, as shown in Fig. 6. Chignolin has 10 C_α atoms, and the TFV dimension is 45, which is the number of C_α atom pairs. In the top panel of this figure, we plotted the value of each component in $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$. Darker squares represent that the corresponding edge component is large. We omitted the component in the lower left triangle. For example, in the case of $\boldsymbol{w}_1$, the color at the second row, ninth column is dark, which implies that the component of $\boldsymbol{w}_1$ that corresponds to the Tyr2-Trp9 pair is large. From this figure, when the protein was in region A in Fig. 5 *a*, the edge between Tyr2 and Trp9 made a large contribution to loop

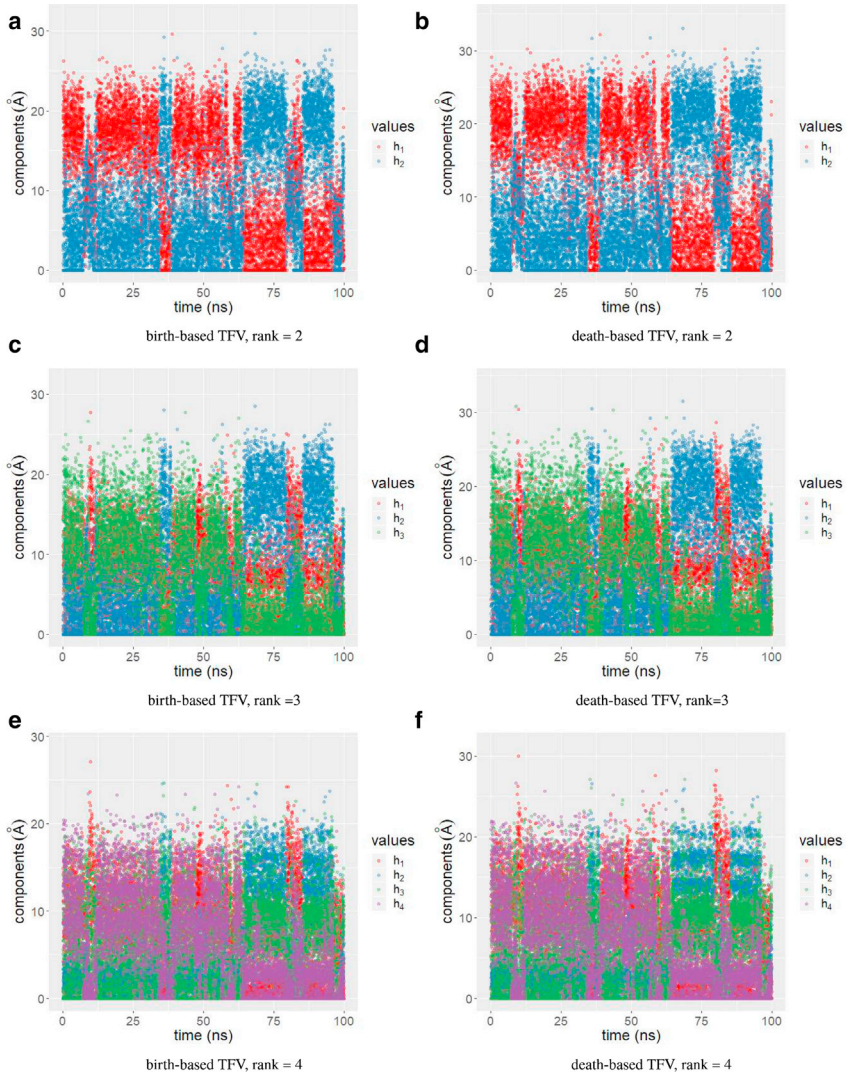

FIGURE 4 Dynamics of the TFV in reduced spaces. The ranks are as follows: (*a*) rank = 2, birth-based TFV; (*b*) rank = 2, death-based TFV; (*c*) rank = 3, birth-based TFV; (*d*) rank = 3, death-based TFV; (*e*) rank = 4, birth-based TV; and (*f*) rank = 4, death-based TFV.

formation because the corresponding feature vector was well approximated by $h_1 \boldsymbol{w}_1$. In the bottom panel of this figure, edges with corresponding components in $\boldsymbol{w}$ larger than 0.2 are indicated by blue lines. We note that $\boldsymbol{w}_i$-values are dimensionless, as discussed in Methods.

From this figure, we noted that the adjacent amino acid pairs, such as Gly1-Tyr2 or Asp3-Pro4, made a large contribution to the cycle formation. This is natural because the distances between adjacent amino acids are short because of chemical bonding. There was a large difference between $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ on the components corresponding to Tyr2-Trp9, Tyr2-Thr8, and Pro4-Gly7. In particular, the components corresponding to Tyr2-Trp9 and Tyr2-Thr8 showed clear differences; $\boldsymbol{w}_1$ (Tyr2-Trp9) = 0.43, whereas that of $\boldsymbol{w}_2$ (Tyr2-Trp9) = 0. $\boldsymbol{w}_1$ (Tyr2-Thr8) = 0, whereas $\boldsymbol{w}_2$ (Tyr2-Thr8) = 0.47. Therefore, h_1 becomes large if there is a loop whose volume-optimal cycle includes Tyr2-Trp9, and h_2 becomes large when a loop forms whose volume-optimal cycle includes Tyr2-Thr8. These results are depicted in the bottom panel of this figure. To determine which state corresponds to the native structure, prior knowledge of the native state is needed. When we applied PH to the native structure, we found that Tyr2-Trp9 was dominant in the native state. The TFV component for Tyr2-Trp9 was 1.249, whereas that for Tyr2-Thr8 was 0.130. Therefore, we conclude that the cluster at region A in Fig. 5 *a* corresponds to the native state, whereas the cluster at region B corresponds to the misfolded state. Concerning the state around (h_1, h_2) ~(5.0 Å, 0 Å), we noted that a small h_1 and h_2 implies that there are no loops. Because the feature vectors $\boldsymbol{v}$ were approximated as $\boldsymbol{v} \sim \sum h_i \boldsymbol{w}_i$, $h_1 = h_2 = 0$ suggests that $\boldsymbol{v} \sim 0$. Therefore, we hypothesized that the state with a small (h_1, h_2) is unfolded.

This hypothesis was supported by the molecule snapshots. In Fig. 7, we plotted examples of the chignolin configuration in the native, misfolded, and unfolded structures.

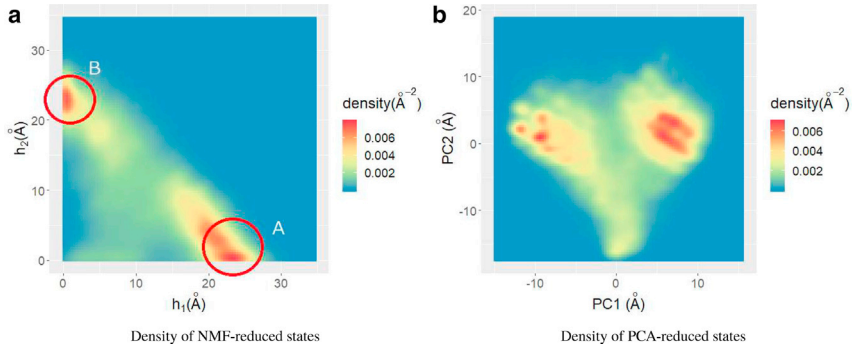

FIGURE 5 Density of states in reduced spaces. (*a*) shows the NMF reduction. (*b*) shows the PCA reduction. To see this figure in color, go online.

This figure presents results consistent with our hypotheses. Fig. 7, *d–f* show the distance between amino acids in each state. In this plot, we calculated the distance between C_α atoms and inferred that this was the "distance" between corresponding amino acids. In the folded state, the distance between Tyr2-Trp9 is small, whereas the distance between Tyr2-Thr8 is small in misfolded states. We also noted that both the distances between Tyr2-Trp9 and Tyr2-Thr8 are small enough to assume these amino acids are "contacted," as discussed in the next subsection. In the unfolded state, the distance map shown in Fig. 7 *f* has no clear structure.

Finally, we compared the TFV-based analysis and other PH analysis. When applying PH to proteins, Xia and Wei proposed the molecular topological fingerprint (MTF) method (12,13) and claimed that "accumulated bar length" of persistent barcodes are useful in identifying protein structure. To investigate the MTF, we plotted a "barcodes" diagram for the sample data in Fig. 7, *g–i*. In the barcode plot, each cycle is represented by a horizontal line, which begins at birth and ends at death. In this plot, cycles are sorted in ascending birth order. At $t = 160$ ns, we have eight loops; however, several loops had lifetimes too short for observation in Fig. 7 *g*. Fig. 7, *h* and *i* represent examples of barcodes for misfolded and unfolded states, respectively. In the misfolded state, we observed nine cycles; however, it was difficult to distinguish the native and misfolded states. We observed four loops with very short lifetimes in the unfolded state, which can be easily distinguished from other states. Xia and Wei proposed to use accumulated bar length, i.e., the sum of bar length of all cycles, to describe the structure of protein. In Fig. 8, we show the density plot of h_1, h_2 and accumulated bar length. Clearly, the accumulated bar length for both clusters was ~1.5–2.0 Å. Thus, we cannot distinguish these two peaks by accumulated bar length. The advantage of TFV analysis compared with other PH analyses relies on the volume-optimal cycles, which offer much more information than persistent barcodes. For example, we show samples of volume-optimal cycles for native and misfolded states in Fig. 9. From the list of native state cycles shown in Fig. 9 *a*, we found that as the amino acid radius of r increases, the edge Tyr2-Tr9 emerge first, and by increasing r further, the Gly1-Gly10, Pro4-Gly7, Asp3-Thr8, Asp3-Gly7, Asp3-Thr6, Tyr2-Thr8, and Gly1-Trp9 edges emerge, in this order. Compared with the MTF for the 2JOX protein presented by Xia and Wei (12), the chignolin's MTF is more complex. For 2JOX, the edges emerged between the closest adjacent amino acids, with one amino acid having only one edge in contact with the other strand of the β-sheet. In our case, several amino acids contact two or more neighbors. For example, Gly1 contacts

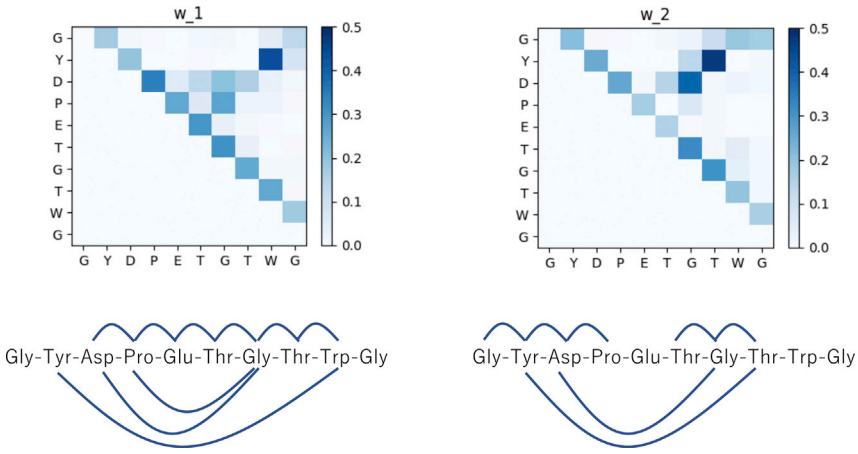

FIGURE 6 Components of $\boldsymbol{w}$ for rank = 2. We also represent the edges whose corresponding components are larger than 0.2 by blue lines. To see this figure in color, go online.

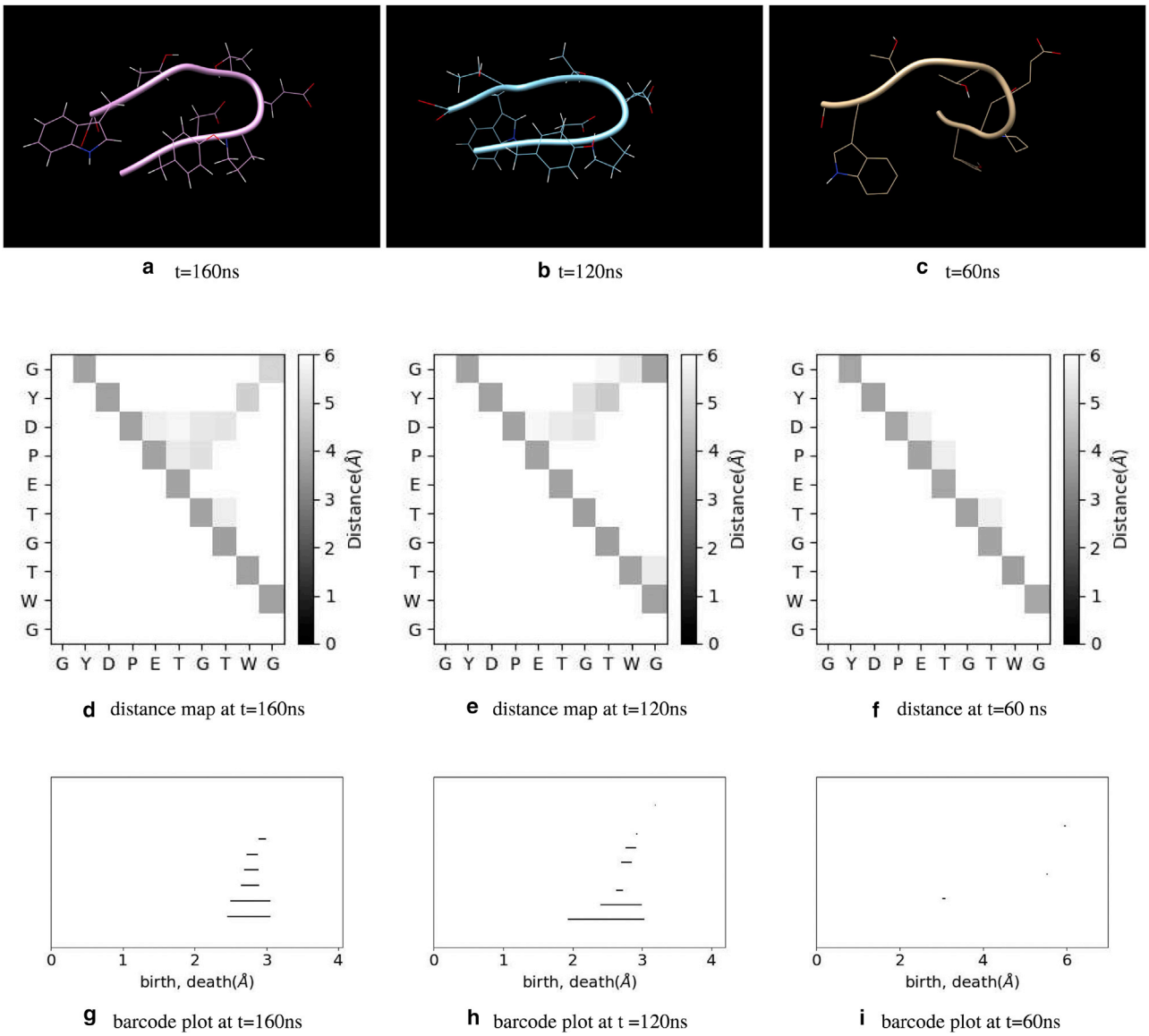

FIGURE 7 (*a*)–(*c*) Example of protein configuration in native, misfolded, and unfolded state. (*a*) shows the native state, $t = 160$ ns, $(h_1, h_2) = (22.7,3.03)$. (*b*) shows the misfolded state, $t = 120$ ns, $(h_1, h_2) = (3.81, 23.1)$. (*c*) shows the unfolded state, $t = 60$ ns, $(h_1, h_2) = (3.60, 6.44)$. (*d*)–(*f*) show examples of distance maps for native (*d*), misfolded (*e*), and unfolded states (*f*). (*g*)–(*i*) show persistent barcode plots for native (*g*), misfolded (*h*), and unfolded states (*i*). To see this figure in color, go online.

Trp9 and Gly10, whereas Asp3 contacts Gly7, Thr6, and Thr8. The list of volume-optimal cycles in Fig. 9 *a* also showed that four of eight cycles include the edge Tyr2-Trp9. However, in the misfolded state at $t = 120$ ns, we observed that four of nine cycles included the edge Tyr2-Thr8, as shown in Fig. 9 *b*. These observations are consistent with the result presented in Fig. 6. We also noted that births and deaths are not sufficient to distinguish the cycles in the native and misfolded states. For example, cycles Gly1-Tyr2-Trp9-Gly10 in Fig. 9 *a* and Gly1-Tyr2-Thr8-Gly10 in Fig. 9 *b* have nearly the same births and deaths. The volume-optimal cycle reveals the difference between these "similar" generators.

Comparison with the Cartesian coordinates and contact map results

It was instructive to investigate the relationship between our TFV analysis and other previous methods. Here, we compare our result with those of Cartesian coordinate and contact map-based analysis.

First, we compared our results and those based on the Cartesian coordinates of C_α atoms, as described by Mitsutake and Takano (4). In this analysis, we constructed the $3n$-dimensional vector $\boldsymbol{R} = (x_1, y_1, z_1, x_2, y_2, z_2, \ldots, x_n, y_n, z_n)$, where (x_i, y_i, z_i) represents the position of the i-th C_α atom. We then reduced dimensionality by PCA. NMF is not available in this case because Cartesian

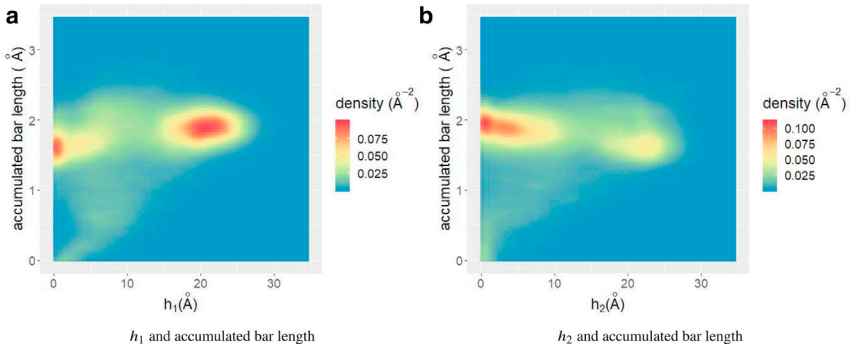

FIGURE 8 Density plot of accumulated bar length and (*a*) h_1 and (*b*) h_2. To see this figure in color, go online.

coordinates can be negative. The density plots in the reduced space are represented in Fig. 10. We identified two clear peaks of density in Fig. 10 *c*. These results are consistent with the result of Mitsutake and Takano (4). Importantly, the first principal component does not contribute to the cluster identification. These results indicate that the PCA is strongly affected by structural changes, which are not related to the transition between the folded and misfolded states. A possible explanation is the large degree of freedom in the unfolded state. In unfolded states, most amino acids can move freely, which results in large Cartesian coordinate fluctuations. In TFV analysis, we had few cycles in the unfolded state, and the corresponding TFV became nearly 0. Therefore, we were able to avoid this large fluctuation in the unfolded state by using TFV.

To confirm the cluster consistency between the TFV-based analysis and Cartesian coordinate-based analysis, we investigated the relationship between the h_i in Fig. 5 *a* and the PCA result. Fig. 11 shows the scatter plot of the second and third principal components, in which the point color indicates the score h_1 and h_2 obtained by NMF. Comparing this figure with Fig. 10 *c*, h_1 returns large values for the cluster PC3 $\sim$0, whereas h_2 are large for the cluster at PC3 ~ -4. Therefore, we concluded that the clusters obtained by Cartesian coordinate PCA and TFV-based analysis coincide.

Next, we conducted PCA and NMF analysis based on contact maps using the method proposed by Ernst et al. (31). In this method, we calculated the distance between C_α atoms and created the contact map D_{ij} as

$$D_{ij} = \begin{cases} 1 & \text{if the distance between } i-\text{th and} \\ & j-\text{th } C_\alpha \text{ is smaller than 8 Å,} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

D_{ij} is nonnegative, so we could apply both PCA and NMF. In Fig. 12, we show the density plot obtained by PCA and NMF of D_{ij}. In both cases, it was difficult to identify the folded or misfolded states.

Compared with the analysis based on contact maps, TFV selects the shortest distance automatically. When the distances between Tyr2-Tyr8 and Tyr2-Trp9 are both shorter than 8 Å, $D_{ij} = 1$ for both pairs. In this case, we cannot distinguish the folded and misfolded states by contact map. Of 100,000 samples, 80,366 samples showed that both Tyr2-Tyr8 and Tyr2-Trp9 are contacted, and we failed to distinguish the native and misfolded states. However, TFV depends on the distance difference; if the distance between Tyr2-Trp9 is smaller than that between Tyr2-Tyr8, the edge between Tyr2-Tyr8 is not included in the volume-optimal cycle because at the birth of the cycle, including Tyr2-Tyr8, Tyr2-Trp9 is not contacted. In other words, TFV is sensitive to the difference in distance between atoms, whereas contact maps are sensitive to the absolute distance. This explains why TFV successfully detected two clusters that were not captured by the contact-map-based method.

Before concluding this subsection, we discuss a previous study by Mitsutake and Takano describing the structure of native and misfolded state (4) in which they investigated the distance between many atom pairs and found that the hydrogen bond between Asp3 and Gly7 is strongly related

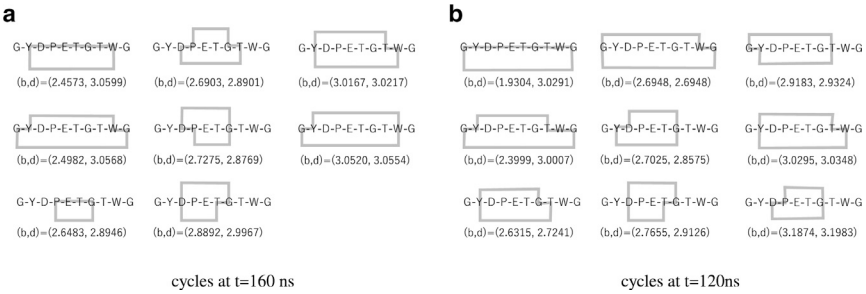

FIGURE 9 Volume-optimal cycles at native ($t =$ 160 ns) (*a*) and misfolded ($=$ 120 ns) (*b*) states. Gray lines indicate the edge included in the volume-optimal cycle. (b, d) represent the birth and death time of each cycle, respectively. The unit of b and d are Å.

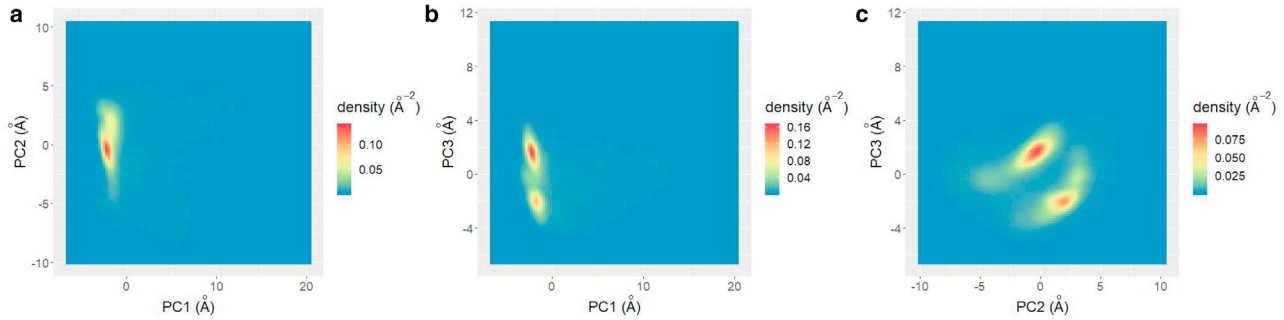

FIGURE 10 The density plots in the reduced space obtained by PCA based on Cartesian coordinates of C_α atoms. (*a*) shows the first and second components. (*b*) shows the first and third components. (*c*) shows the second and third components. To see this figure in color, go online.

to the difference between the native and misfolded states. Fig. 6 shows that the contribution of the edge between Asp3 and Gly7 is larger in $\boldsymbol{w}_2$ than in $\boldsymbol{w}_1$, which is consistent with their results. However, our analysis showed that the edge between Tyr2 and Thr8 is more remarkable in $\boldsymbol{w}_2$, which is not mentioned previously. Thus, further studies are needed to determine the difference between our work and previous studies.

Dynamics in reduced space

To investigate the transition between native, misfolded, and unfolded states, we plotted the average flow in Fig. 13. First, we calculated $\delta\boldsymbol{h}(t) = \boldsymbol{h}(t + \delta t) - \boldsymbol{h}(t)$, where $\delta t = 10$ ps. Next, we divided the two-dimensional space into grids of size 1.5 × 1.5 Å and calculated the average of $\delta\boldsymbol{h}(t)$ for each grid. Fig. 13 *a* shows the flow in the entire two-dimensional space. The results indicate that there are two stable solutions at (h_1, h_2) ∼(20 Å, 5 Å) and (5 Å, 20 Å), respectively. These two stable points correspond to folded and misfolded states, as discussed above. The positions of these fixed points differ slightly from the density peak, which is due to the constraint of $h_1, h_2 \geq 0$. Though the density peak is on the h_1 and h_2 axes, these points cannot be fixed points because any configuration change drives the system away from the axes. We also found that the flow along the line $h_1 + h_2 \sim 25$ is strong, whereas the flow at small $h_1 + h_2$ is very weak. To investigate the dynamics in this area more clearly, we plotted the flow at 0 Å $\leq h_1, h_2 \leq$ 15 Å in Fig. 13 *b*. These results strongly suggest that there is a saddle point at (h_1, h_2) ∼(12 Å, 12 Å). Therefore, this position is likely the transition state. These results also demonstrate that there is no fixed point corresponding to the unfolded state. Once the chignolin molecule reaches the thermal noise-induced unfolded state, it remains unfolded for an extended period of time. However, this is not because of the attraction to the stable fixed point, but rather the slow dynamics in reduced space.

CONCLUSIONS

In this study, we analyzed the chignolin folding process using PH and proposed TFV as a feature to characterize protein structure. TFV allows us to combine PH with machine-learning methods, such as PCA or NMF. In particular, we show that NMF analysis of TFV provides essential information on folding dynamics. By investigating flow in the reduced space, we found that there are two stable fixed points that correspond to the native and misfolded states, one saddle point corresponding to transient state, and one unfolded state. The difference between the native and misfolded states lies in the edge differences between Tyr2-Trp9 and Tyr2-Thr8. The unfolded state has no fixed point, although the protein remains unfolded for a long time because of the slow dynamics.

Our results show that featurization of protein structure by TDA is promising. Several previous studies attempted to

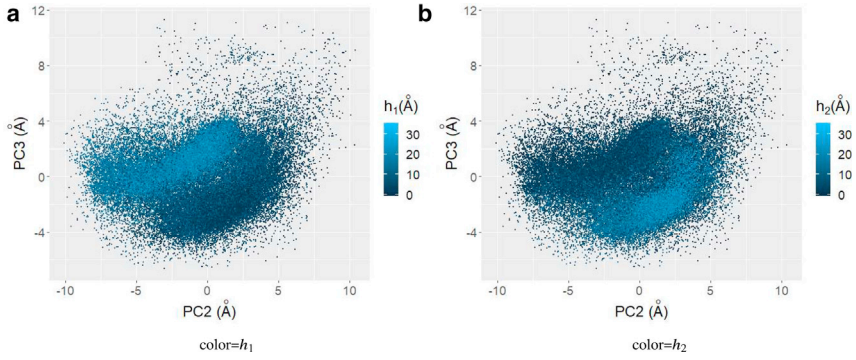

FIGURE 11 Scatter plot in second and third Cartesian coordinate principal components. Color indicates the values of (*a*) h_1 and (*b*) h_2 obtained by TFV and NMF, respectively. To see this figure in color, go online.

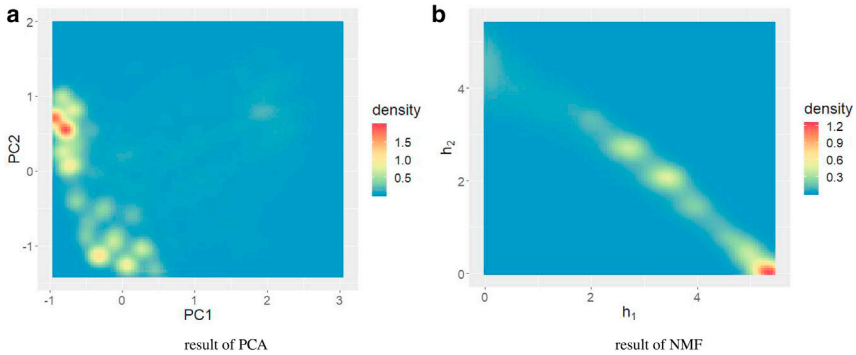

FIGURE 12 Result of (*a*) PCA and (*b*) NMF of contact map D_{ij}. To see this figure in color, go online.

apply TDA to biomolecule dynamics. For example, Yao et al. applied Mapper, a major TDA method, to investigate RNA folding (9). Xia and Wei proposed MTF to characterize protein structure (12,13). Gameiro et al. investigated the relationship between PH and protein compressibility (32). Compared to these previous works, there are several advantages to the TFV-based analysis. First, TFV uses volume-optimal cycles, which include significantly more information than does persistent barcodes. As shown in Results and Discussion, volume-optimal cycles can distinguish different structures that have the same births and deaths. Another advantage is applicability to machine learning. As we discussed in the Introduction, the fluctuation in number of cycles causes difficulty when applying machine learning with PH. However, the TFV dimension only depends on the number of amino acids; therefore, we can easily apply machine-learning methods such as PCA, NMF, and deep learning. Notably, TFV calculations require large computational cost. Therefore, it would be difficult to calculate TFV using all atoms in macromolecules.

The method developed in this study is powerful; however, further development is also possible. First, several approaches can be taken to construct TFVs from the volume-optimal cycles. Here, we used the sum of the deaths as the edge weight, but we could alternatively use the sum of births, the product of deaths, or a combination of births and deaths. For chignolin, there is no qualitative difference when we use the sum of the births in place of the deaths because the births or deaths of every cycle are the same order, as shown in Fig. 7, *g* and *h*. However, if we analyze more complex molecules, the analysis may depend on the TFV definition. Chignolin is a small molecule with only one β-sheet and no tertiary structure. If we need to investigate a more complex protein with tertiary structures, then the TFV definition may affect the analysis results. Another improvement could be achieved by the selection of the TFV analysis method. In this study, we applied NMF to reduce the structure into two dimensions; however, when analyzing more complex proteins, the reduced dimensions will increase, so requiring careful determination of the appropriate reduced space dimension. In this case, other analysis methods may be more appropriate. Once the TFV is calculated, we can apply several data mining and time series analysis methods such as hierarchical clustering, PCA, Fourier analysis, relaxation mode analysis, and independent component analysis. In time series analysis, Markov models are also promising, as it is a powerful tool for time series analysis of protein dynamics (7). The difficulty in applying Markov modeling lies in the definition of the states. In our study, we identified the misfolded and folded states; however, it is difficult to identify the boundary between them. Application of a hidden Markov model (33) may solve this problem because this method classifies each state automatically. Moreover, we can also apply text-mining methods. In our approach, an edge is regarded as a "term" to describe the protein shape. In this analogy, the set of generators is a document that

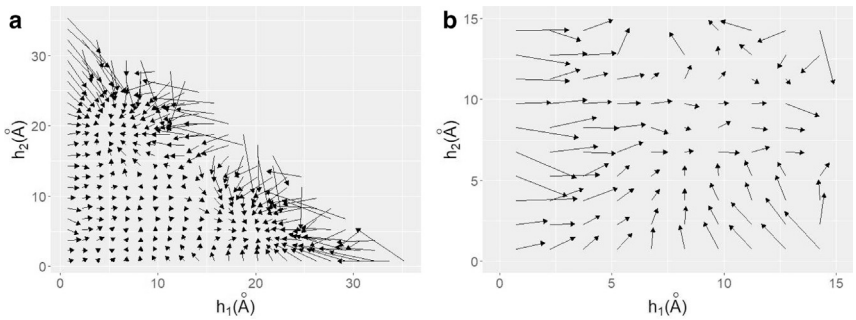

FIGURE 13 Average flow $\delta \boldsymbol{h}(t) = \boldsymbol{h}(t + \delta t) - \boldsymbol{x}(t)$. Panel (*a*) shows flow in the entire two-dimensional space. Panel (*b*) shows flow at 0 Å $\leq h_1, h_2 \leq$ 15 Å.

describes the protein shape, and the volume-optimal cycles are the sentences in the document. Therefore, we will be able to use text-mining methods such as topic models, term network analysis, and deep learning.

The method we developed here is applicable not only to protein folding but also to other problems in physics, chemistry, and engineering. For example, we could capture protein binding with small molecules, which will contribute to new drug development. Another interesting application is active matter dynamics such as schools of fish or flocks of birds. Although active matter dynamics is keenly studied, quantitatively analyzing the shapes of clusters of active matter is difficult. Our method will provide insights regarding this problem.

AUTHOR CONTRIBUTIONS

T.I. performed research. I.O. and Y.H. contributed to the development of analytical tools. T.I., I.O., and Y.H. wrote the article.

ACKNOWLEDGMENTS

The authors thank Hiroya Nakao, Hiromichi Suetani, Takenobu Nakamura, and Kenji Fukumizu for helpful comments. We thank Editage (www.editage.com) for English language editing.

This work was financially supported by Japan Science and Technology Agency Core Research for Evolutionary Science and Technology grant JPMJCR15D3, Japan.

REFERENCES

1. Levinthal, C. 1968. Are there pathways for protein folding? *J. Chim. Phys.* 65:44–45.
2. Lane, T. J., D. Shukla, …, V. S. Pande. 2013. To milliseconds and beyond: challenges in the simulation of protein folding. *Curr. Opin. Struct. Biol.* 23:58–65.
3. Kitao, A., F. Hirata, and N. Gō. 1991. The effects of solvent on the conformation and the collective motions of protein: normal mode analysis and molecular dynamics simulations of melittin in water and in vacuum. *Chem. Phys.* 158:447–472.
4. Mitsutake, A., and H. Takano. 2015. Relaxation mode analysis and Markov state relaxation mode analysis for chignolin in aqueous solution near a transition temperature. *J. Chem. Phys.* 143:124111.
5. Naritomi, Y., and S. Fuchigami. 2011. Slow dynamics in protein fluctuations revealed by time-structure based independent component analysis: the case of domain motions. *J. Chem. Phys.* 134:065101.
6. Schwantes, C. R., D. Shukla, and V. S. Pande. 2016. Markov state models and tICA reveal a nonnative folding nucleus in simulations of NuG2. *Biophys. J.* 110:1716–1719.
7. Fujisaki, H., K. Moritsugu, …, H. Suetani. 2018. Conformational change of a biomolecule studied by the weighted ensemble method: use of the diffusion map method to extract reaction coordinates. *J. Chem. Phys.* 149:134112.
8. Ramachandran, G. N., and V. Sasisekharan. 1968. Conformation of polypeptides and proteins. *Adv. Protein Chem.* 23:283–438.
9. Yao, Y., J. Sun, …, G. Carlsson. 2009. Topological methods for exploring low-density states in biomolecular folding pathways. *J. Chem. Phys.* 130:144115.
10. Nicolau, M., A. J. Levine, and G. Carlsson. 2011. Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proc. Natl. Acad. Sci. USA.* 108:7265–7270.
11. Edelsbrunner, H., D. Letscher, and A. Zomorodian. 2002. Topological persistence and simplification. *Discrete Comput. Geom.* 28:511–533.
12. Xia, K., and G.-W. Wei. 2014. Persistent homology analysis of protein structure, flexibility, and folding. *Int. J. Numer. Methods Biomed. Eng.* 30:814–844.
13. Xia, K., and G.-W. Wei. 2015. Multidimensional persistence in biomolecular data. *J. Comput. Chem.* 36:1502–1520.
14. Carlsson, G. 2014. Topological pattern recognition for point cloud data. *Acta Numer.* 23:289–368.
15. Hiraoka, Y., T. Nakamura, …, Y. Nishiura. 2016. Hierarchical structures of amorphous solids characterized by persistent homology. *Proc. Natl. Acad. Sci. USA.* 113:7035–7040.
16. Ichinomiya, T., I. Obayashi, and Y. Hiraoka. 2017. Persistent homology analysis of craze formation. *Phys. Rev. E.* 95:012504.
17. Escolar, E. G., and Y. Hiraoka. 2016. Optimal cycles for persistent homology via linear programming. *In* Optimization in the Real World. K. Fujisawa, Y. Shinano, and H. Waki, eds. Springer, pp. 79–96.
18. Obayashi, I. 2018. Volume-optimal cycle: tightest representative cycle of a generator in persistent homology. *SIAM J. Appl. Algebr. Geom.* 2:508–534.
19. Kimura, M., I. Obayashi, …, Y. Hiraoka. 2018. Non-empirical identification of trigger sites in heterogeneous processes using persistent homology. *Sci. Rep.* 8:3553.
20. Kusano, G., K. Fukumizu, and Y. Hiraoka. 2016. Persistence weighted Gaussian kernel for topological data analysis. In Proceedings of The 33rd International Conference on Machine Learning, pp. 1–24.
21. Bubenik, P., and J. Scott. 2014. Categorification of persistent homology. *Discrete Comput. Geom.* 51:600–627.
22. Zomorodian, A., and G. Carlsson. 2005. Computing persistent homology. *Discrete Comput. Geom.* 33:249–274.
23. Manning, C., P. Raghavan, and H. Schütze. 2008. Introduction to Information Retrieval. Cambridge University Press, Cambridge, England.
24. Lee, D. D., and H. S. Seung. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature.* 401:788–791.
25. Boutsidis, C., and E. Gallopoulos. 2008. SVD based initialization: a head start for nonnegative matrix factorization. *Pattern Recognit.* 41:1350–1362.
26. Brunet, J.-P., P. Tamayo, …, J. P. Mesirov. 2004. Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl. Acad. Sci. USA.* 101:4164–4169.
27. Hutchins, L. N., S. M. Murphy, …, J. H. Graber. 2008. Position-dependent motif characterization using non-negative matrix factorization. *Bioinformatics.* 24:2684–2690.
28. Pedregosa, F., G. Varoquaux, …, E. Duchesnay. 2011. Scikit-learn: machine learning in {P}ython. *J. Mach. Learn. Res.* 12:2825–2830.
29. Gaujoux, R., and C. Seoighe. 2010. A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics.* 11:367.
30. Berendsen, H. J., D. van der Spoel, and R. van Drunen. 1995. GROMACS: a message-passing parallel molecular dynamics implementation. *Comput. Phys. Commun.* 91:43–56.
31. Ernst, M., F. Sittel, and G. Stock. 2015. Contact- and distance-based principal component analysis of protein dynamics. *J. Chem. Phys.* 143:244114.
32. Gameiro, M., Y. Hiraoka, …, V. Nanda. 2015. A topological measurement of protein compressibility. *Jpn. J. Ind. Appl. Math.* 32:1–17.
33. Baum, L. E., and T. Petrie. 1966. Statistical inference for probabilistic functions of finite state Markov chains. *Ann. Math. Stat.* 37:1554–1563.