Identifying metastable states of biomolecules by trajectory mapping and density peak clustering

Chuanbiao Zhang,[1] Shun Xu,[2] and Xin Zhou[3,*]

[1] *College of Physics and Electronic Engineering, Heze University, Heze 274015, China*
[2] *Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China*
[3] *School of Physical Sciences, University of Chinese Academy of Sciences, Beijing 100190, China*

(Received 11 April 2019; revised manuscript received 21 June 2019; published 3 September 2019)

Efficiently and accurately analyzing high-dimensional time series, such as the molecular dynamics (MD) trajectory of biomolecules, is a long-standing and intriguing task. Two different but related techniques, i.e., dimension reduction methods and clustering algorithms, have been developed and applied widely in this field. Here we show that the combination of these techniques enables further improvement of the analyses, especially with very complicated data. Specifically, we present an approach that combines the trajectory mapping (TM) method, which constructs slow collective variables of a time series, with density peak clustering (DPC) [A. Rodriguez and A. Laio, Science **344**, 1492 (2014)], which identifies similar data points to form clusters in a static data set. We illustrate the application of the TMDPC approach with hundreds of microseconds of all-atomic MD trajectories of two proteins, the villin headpiece and protein G. The results show that TMDPC is a powerful tool for achieving the metastable states and slow dynamics of these high-dimensional time series due to the efficient consideration of the time successiveness and the geometric distances between data points.

DOI: 10.1103/PhysRevE.100.033301

I. INTRODUCTION

In the analysis of high-dimensional data, such as pattern recognition [1] and the understanding of molecular dynamics (MD) trajectories [2], two different but closely related techniques are usually performed: clustering algorithms and dimension reduction methods. Both of these methodologies evaluate the similarity among the data points and then obtain a simplified description of the data by neglecting small deviations among points. While clustering algorithms focus on grouping similar data points to decrease the size of the data, the dimension reduction methods attempt to achieve low-dimensional features from high-dimensional data. In recent decades, many clustering algorithms (e.g., k-means clustering [3] and density-based clustering [4]) and dimension reduction methods [e.g., principal component analysis (PCA) [5], independent component analysis [6], time-structure-based independent component analysis [7], isometric feature mapping [8], and diffusion maps [9]] have been developed and widely applied in the analysis of a variety of data and have greatly improved our ability to understand complicated data.

A common situation is that the efficiency and accuracy of these data analysis methods inevitably decrease as the dimension of the data points increases. Thus, endeavors to improve the performance of these methods in more complicated data have continued. Rodriguez and Laio introduced a powerful clustering algorithm, called density peak clustering (DPC) [10], to efficiently identify aggregated points as clusters based on the density of points. However, similar to other methods, the capability of DPC quickly decreases as the dimension of the space where these data points are located increases. Therefore, it is natural to combine dimension reduction techniques with clustering algorithms to achieve the capability of analyzing high-dimensional data [2]. In addition, for many specific data, in addition to the geometric similarity between points, there may be other features that characterize the relations among the data. For example, for the MD trajectory, the frames, i.e., the conformations, are time ordered as in a time series. Two time-consecutive conformations are intrinsically more likely to be located in the same metastable state, even though their geometric similarity is not as high. Therefore, conformations should not be treated by the MD trajectory as only a set of all individual conformations, but the time orders of these conformations in the trajectory should be considered.

In previous works, techniques, such as the trajectory mapping (TM) technique [11–14], have been developed to efficiently construct a few collective variables in slow dynamics from the MD trajectory by taking into account both the time successiveness of the conformations and the usual geometric similarity between conformations. Here we further combine the TM with the DPC to produce the TMDPC approach to further improve the analyses of the MD trajectory of more complicated biomolecules. In the TMDPC approach, we first use the TM to construct some collective variables to focus on the slow dynamics of the MD trajectory (or other time series). Then we apply the DPC to group the aggregated data points in the slow-variable space to achieve the metastable states of the system. We demonstrate the application of the TMDPC approach in the atomistic MD trajectories of two proteins: a 125-μs trajectory of the villin headpiece [15] and a 444-μs trajectory of protein G [16] with explicit water molecules. The results show that the TMDPC approach can more efficiently capture the slow-dynamics characteristic of the systems than applying the original TM or DPC method alone.

[*] xzhou@ucas.ac.cn

II. METHODS

A. Trajectory mapping: Constructing slow variables

The TM approach can robustly identify slow motions from MD trajectories. The feature of the TM is that it treats each segment of the trajectory in time, rather than treating an individual conformation as a data point, and then simplifies these data points, i.e., trajectory segments, based on principal component analysis. The principal components give slow-dynamics variables of the MD trajectory. Here we provide a brief description of the TM approach, which mainly includes three steps. More details in the TM approach have been described in previous works [11–14].

(i) Many functions of the Cartesian coordinates of atoms are chosen to describe the conformations in the MD trajectory. These functions are called basis functions. Usually, collective variables that are relevant to large changes in conformations, such as dihedral angles, pair distances of heavy atoms, and the root-mean-square deviation (RMSD), are more appropriate as basis functions than local-motion-relevant variables, such as bond lengths and bond angles. Before using these different types of collective variables as basis functions, we first subtract the mean value of each variable and normalize it by dividing it by its standard deviation. Therefore, each basis function is a dimensionless variable with a mean of zero and a unit standard deviation.

(ii) Each of the original MD trajectories is cut into segments with a preset time length τ and the conformations in each segment are averaged as a mapped data point in the space spanned by these basis functions. This τ-length average can filter out fast motions whose timescales are much shorter than τ. The parameter τ is a free parameter that can be determined by our desired timescale in the specific problem. We first set τ to be large enough (e.g., a fraction of the length of the whole trajectory) to obtain a simpler picture of slow dynamics at longer timescales. Then more detailed dynamics at shorter timescales can be obtained by using a smaller τ. We verified that the constructed slow variables are robust to different values of τ [13].

(iii) The principal components (PCs) are obtained from the τ-length-trajectory-mapped points by the standard PCA technique. These PCs give the slow variables of the system. The first few PCs correspond to the slowest motions. A method for choosing the number of PCs is provided by the plot of the eigenvalues sorted in decreasing order.

B. Density peak clustering

Density peak clustering is based on two assumptions [10]. First, the local density of a cluster center is higher than the local density of its neighbors. Second, the distance between a cluster center and any other points with a higher local density is large. To identify the cluster center, we only need to calculate two quantities for each data point i: its local density ρ_i and its distance δ_i from the nearest points with higher density. The local density ρ_i is defined as

$$\rho_i = \sum_j \chi(d_{ij} - d_c), \tag{1}$$

where $\chi(x) = 1$ when $x < 0$ and $\chi(x) = 0$ otherwise. Thus, ρ_i is the number of points that are closer than a preset cutoff value d_c to point i. The distance δ_i is defined as

$$\delta_i = \min_{j:\rho_j > \rho_i} (d_{ij}), \tag{2}$$

where δ_i corresponds to the minimum distance between point i and any other point with a higher density. If we chose an appropriate d_c, the points on the top right corner of the decision graph (the scatter plot of ρ_i and δ_i) correspond to cluster centroids. As proposed by Rodriguez and Laio [10], the plot of $\gamma_i = \rho_i\delta_i$ sorted in decreasing order also helps choose the number of clusters. The final results are robust to different d_c. After obtaining the cluster center points, the remaining points can be assigned to the closest centroid.

C. The TMDPC method

The TMDPC method mainly involves three steps. First, the TM is used to construct slow variables with a preset parameter τ. Second, the original trajectory is cut into short segments of length τ_α, the τ_α-length-trajectory segments are mapped to the slow-variable space, and each segment is averaged into one point. Third, the DPC is used to identify clusters of the τ_α-length-averaged points in the slow-variable space. These clusters correspond to the metastable states of the system. Here the τ_α average is applied to reduce the amount of calculation in DPC and to further filter possibly remaining fast components of motions. The length τ_α can be set approximately two to three orders of magnitude smaller than τ, and the results of DPC are robust to the selection of τ_α.

After obtaining the metastable states, we can extract various kinetic information by directly identifying the transition events along the simulation trajectories. Here we mainly focus on the transition rates among states. As proposed by Gong and Zhou [11], the transition rate $k_{\beta\alpha}$ from state S_α to state S_β can be estimated by

$$k_{\beta\alpha} = \frac{N_{\beta\alpha}^{\text{trans}}}{t_\alpha}, \tag{3}$$

where t_α is the survival time of state α and $N_{\beta\alpha}^{\text{trans}}$ denotes the total number of transition events from state S_α to S_β.

III. RESULTS

A. Villin headpiece

We first use the TMDPC approach to analyze the C-terminal fragment of the villin headpiece (HP35) with a double norleucine mutant (Nle/Nle) [15]. As shown in Fig. 1(a), this protein contains 35 residues and mainly consists of three helices. Due to its small size and fast-folding feature, HP35 has been extensively studied in experiments and molecular dynamics simulations [16–19]. To show the performance and the potential of the TMDPC approach, we adopt a 125-μs equilibrium MD trajectory of HP35 at 360 K, which was performed by Lindorff-Larsen *et al.* [20]. This trajectory involves approximately 625 000 snapshots. We downloaded this trajectory from D. E. Shaw Research [20]. The time series of the C$_\alpha$ RMSD is shown in Fig. 1(b). We can see that the whole trajectory undergoes the folding-unfolding process several times.

We take all 33 pairs of dihedral angles in the peptide backbone to describe the system. Because the dihedral angles

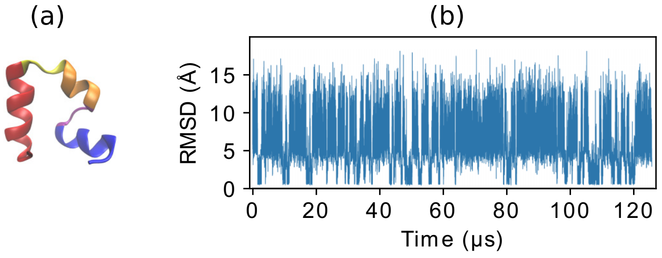

FIG. 1. (a) Representative structures of the folded state of HP35. This protein mainly consists of three helices. The first helix is shown in blue, the second helix is orange, and the third helix is red. (b) Time series of the C_α RMSD between each conformation and the folded structure.

are periodic, we transform those angles into their cosine and sine functions and obtain a total of 132 dimensionless basis functions. To filter out fast motions, we set $\tau = 10\ \mu$s; i.e., we map each 50 000 time-neighboring conformation (trajectory segment) to one point and obtain 6250 segments by sliding the time window 20 ns for each conformation (trajectory segments can partially overlap). As illustrated in Fig. 2(a), two eigenvalues are significantly greater than zero, indicating that the system contains two main slow variables in the $\tau = 10\ \mu$s timescale; the two corresponding slow variables, called B_1 and B_2, are shown in Fig. 2(b). Here the slow-dynamics trajectories are smoothed by time-window averaging with a length of $\tau_\alpha = 100$ ns. Then DPC is used to cluster these τ_α-length-averaged points in the (B_1, B_2) space. We calculate the ρ and δ of each point and show the decision graph in Fig. 2(c). There are three points located at the top, indicating that this system contains three clusters. In addition, the plot of

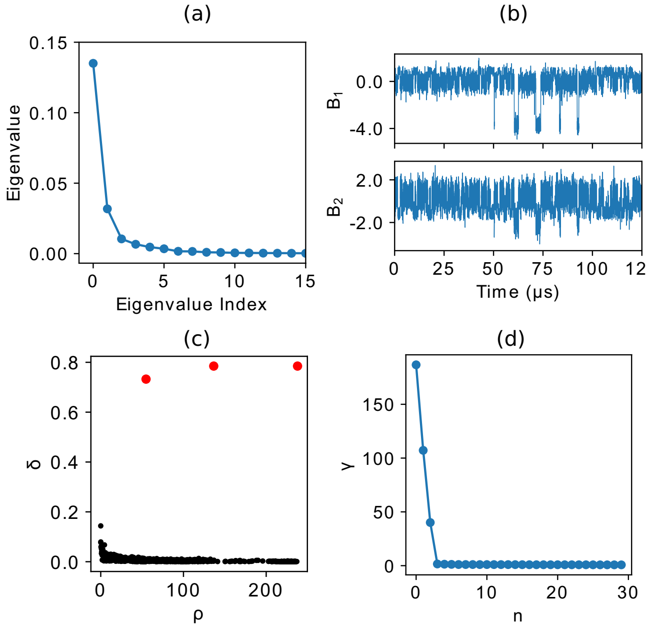

FIG. 2. (a) Eigenvalue of the variance-covariance matrix of the τ-averaged points of HP35. (b) Time series of slow variables constructed by the TM algorithm. (c) Decision graph. The three red points at the top correspond to the cluster centers. (d) Value of γ in decreasing order for the data in (c).

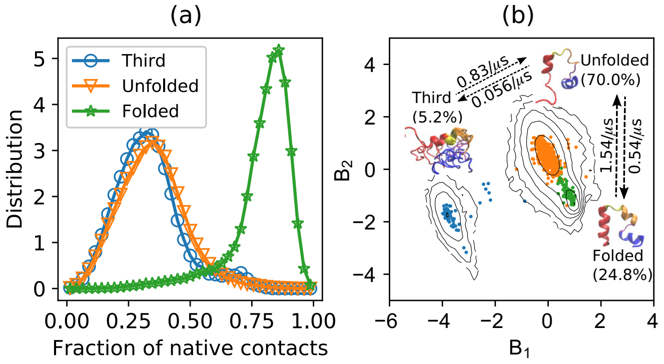

FIG. 3. (a) Distribution of the fraction of native contacts of the three states of HP35. (b) The τ_α-averaged conformations are projected onto slow-variable space constructed by the TM algorithm. The three colors correspond to the three states. Here the representative structures of the three states are shown, including their populations (in %) and transition rates (in μs). For each conformation, the first helix is shown in blue, the second helix is yellow, and the third helix is red.

$\gamma = \rho\delta$ shows that there are three values that are significantly greater than zero, which also indicates that this trajectory contains three clusters [Fig. 2(d)]. The τ_α-length-averaged points belonging to the three clusters are illustrated with three different colors [shown in Fig. 3(b)] in the (B_1, B_2) space. The boundaries between these three clusters are quite clear.

The fraction of native contacts [20] of these three states are shown in Fig. 3(a): One state is the folded state F and the other two states, called U and T, are not completely folded. As illustrated in Fig. 3(b), in state U, the second helix and the first half of the third helix are unfolded, while the first helix and the second half of the third helix are folded. In state T, the first half of the third helix and the second helix are combined together to yield one helix. After obtaining the metastable states, we directly identify the transition events among these states along the MD trajectory and the result is shown in Fig. 3(b). There is only one folding pathway: $T \rightleftharpoons U \rightleftharpoons F$. In state T, the junction between the second and the third helix forms a helix structure, which makes the state T unable to jump to the folded state directly. State T connects to only the unfolded state U and slightly changes its population. Thus, the folding-unfolding process is insignificantly affected by state T, and a two-state model can be used to approximately describe the folding-unfolding mechanism of HP35 [16] if state T is ignored. Using the transition rate calculation method proposed by Gong and Zhou [11], we obtained the folding rate of HP35: 1.54 μs^{-1}. This rate is in good agreement with the experimental value, approximately 1 μs^{-1} [15], and our previous result: 1.52 μs^{-1} [13].

For comparison, in the 132-dimensional original dihedral angle space, we calculate the ρ and δ of each individual conformation and use DPC to cluster them. Since the whole 125-μs trajectory consists of 625 000 conformations, it is difficult to calculate the distance matrix between these conformations. To reduce the amount of calculation, we select one conformation every 10 ns and obtain 12 500 conformations. The decision graph of those 12 500 conformations is shown in Fig. 4(a). Density peak clustering only detects one cluster,

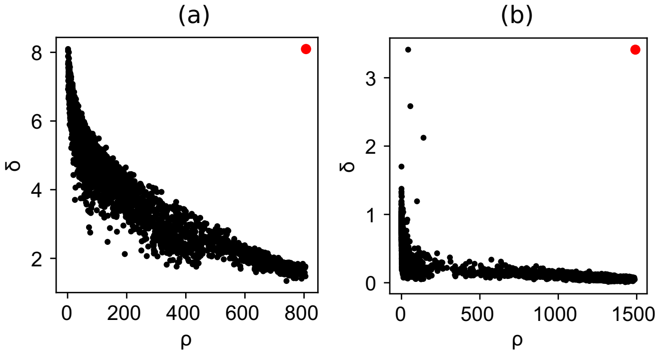

FIG. 4. (a) Decision graph of the samples constructed in the 132-dimensional dihedral angle space. (b) Decision graph of the samples constructed in the PCs space. The PCs are constructed by the standard PCA in the 132-dimensional dihedral angle space without τ averaging.

meaning it cannot identify any metastable states because the high-dimensional original basis functions not only involve the slow processes but also contain a large number of fast processes, such as the fast twisting of the dihedral angle. These fast processes blur the state boundaries, and the samples in each state do not cluster well around the state center, making it difficult to identify the centers. We also perform PCA directly on the dihedral angles without using τ averaging; then we calculate the ρ and δ of each conformation in the low-dimensional space spanned by the PCs. The decision graph is shown in Fig. 4(b). It is hard to identify the number of centers from this figure. This finding indicates that the direct PCA is not sufficient to identify a metastable state, and the τ-averaging process in the TM can efficiently filter fast motions and achieve slow variables to describe the metastable states of the system.

To check the robustness of the TMDPC approach, we adopt different values of τ and τ_α to perform TMDPC. There is no rigid standard for selecting τ and τ_α. As shown in Fig. 5, we can obtain consistent results with different values of τ and τ_α. The TM constructs the slow variables in the selected τ (and longer) timescale by filtering the motion much faster than τ. The τ_α average can greatly reduce the amount of calculation and further filter the motions faster than τ_α without alerting the slow motions at the τ level. Thus, the results of the TMDPC are robust to different τ and τ_α.

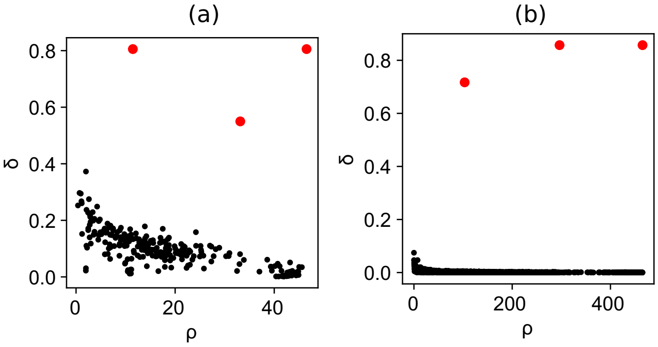

FIG. 5. The number of clusters is insensitive to τ and τ_α. For HP35, we can obtain three cluster centroids from the decision graph based on different values of τ and τ_α: (a) $\tau = 5$ μs and $\tau_\alpha = 500$ ns and (b) $\tau = 40$ μs and $\tau_\alpha = 50$ ns.

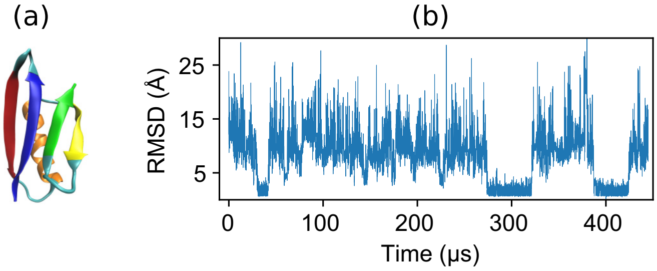

FIG. 6. (a) Representative structures of the folded state of protein G. This protein mainly consists of four β sheets and an α helix. The first β sheet ($\beta1$) is shown in blue, $\beta2$ is red, $\beta3$ is yellow, $\beta4$ is green, and the helix is orange. (b) Time series of the C$_\alpha$ RMSD between each conformation and the folded structure.

B. Protein G

We use TMDPC to analyze the simulation data of protein G, which has been studied in some experimental and computational works [16,20–23]. As shown in Fig. 6(a), protein G contains 56 residues, mainly consisting of four β sheets and an α helix. Lindorff-Larsen *et al.* performed a 444-μs equilibrium MD simulation of protein G at 350 K. The generated trajectory involves approximately 2 220 000 snapshots with a time interval of 200 ps. We downloaded this trajectory data

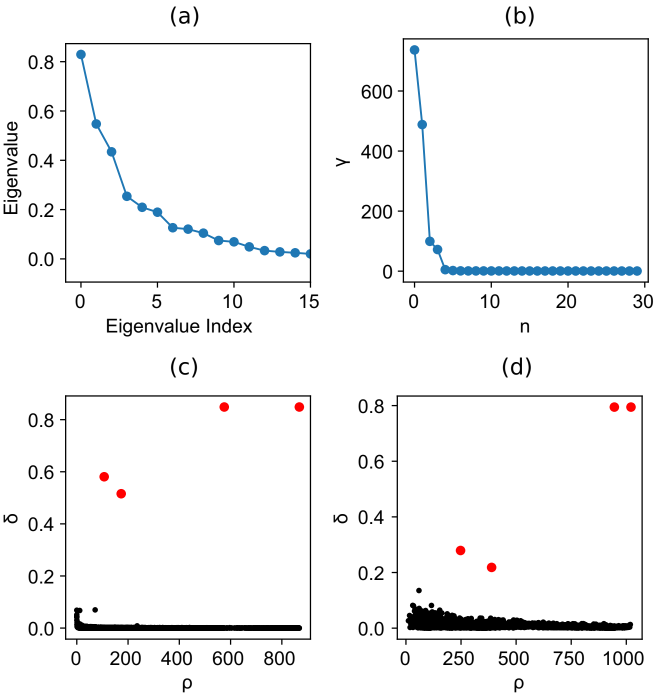

FIG. 7. (a) Eigenvalue of the variance-covariance matrix of the τ-averaged points of protein G. (b) Value of γ in decreasing order for the data in (c). Also shown is the decision graph of the segment-averaged conformations based on the first (c) three and (d) five slow variables.

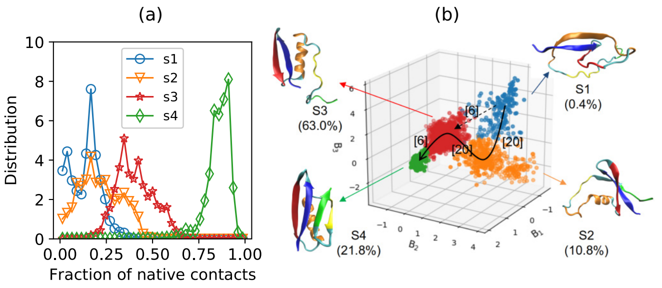

FIG. 8. (a) Distribution of the fraction of native contacts of three states of protein G. (b) The τ_α-averaged conformations are projected into the slow-variable space; the four colors correspond to the four metastable states. The number in parentheses indicates the popularity of each state. The solid black arrow indicates the main folding path. The number in square brackets indicates the transition time between two states.

from D. E. Shaw Research [20]. The time series of the C$_\alpha$ RMSD is shown in Fig. 6(b); the whole trajectory undergoes the folding-unfolding process several times.

To correctly construct the slow variables of this system by the TM, we adopt 1759 collective variables as basis functions, which consist of 1540 distances between atom pairs among the 56 C$_\alpha$ atoms, 216 trigonometric functions of the 54 dihedral angles in the backbone of the protein, the radius of gyration, the C$_\alpha$ RMSD to the native structure, and the number of hydrogen bonds in the system. The whole trajectory is divided into short trajectory segments, each 10 μs long ($\tau = 10$ μs). Due to the complexity of the protein, we found that the eigenvalues decrease to zero very fast, but a few eigenvalues are significantly greater than zero [Fig. 7(a)]. We choose the first three slow variables to perform DPC. As shown in Fig. 8(b), the average conformations of the trajectory ($\tau_\alpha = 100$ ns) are projected into the three-dimensional slow-variable space. The ρ and δ of those averaged conformations are calculated in the slow-variable space. The decision graph [Fig. 7(c)] and the plot of γ [Fig. 7(b)] indicate that the system contains four metastable states. We also perform DPC with more (e.g., five) slow variables, which still yields four metastable states [Fig. 7(d)]. This finding means that the first three slow variables are sufficient to describe the main dynamics of protein G at the τ scale, and more PCs do not change the identification of metastable states.

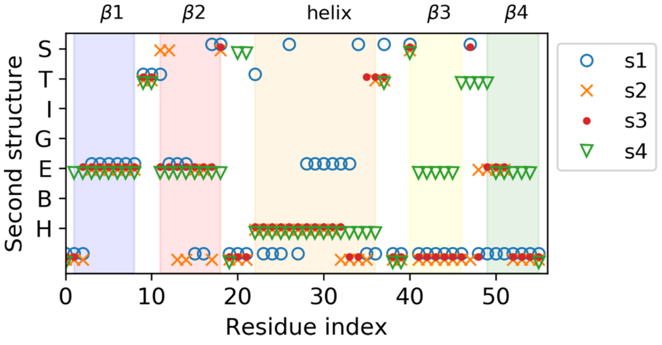

FIG. 9. Most likely secondary structure of each residue in the four metastable states of protein G. The $\beta1$ range is shown in blue, $\beta2$ is red, the helix is orange, $\beta3$ is yellow, and $\beta4$ is green. The second structure is assigned by the DSSP algorithm [24]. Specifically, H represents the α helix, B represents the β bridge, E represents the β sheet, G represents the three-turn helix, I represents the π helix, T represents the hydrogen-bonded turn, and S represents the bend.

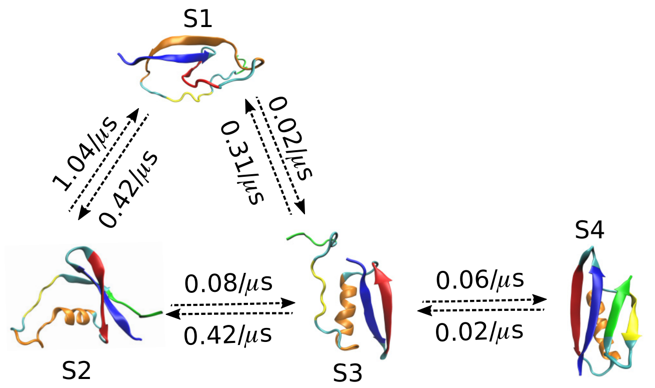

FIG. 10. Transition network of protein G. Here the representative structures of the four states and the transition rates (in μs) are shown.

From the distribution of the fraction of native contacts of the four states [Fig. 8(a)], we find that state $s4$ is the folded state. To obtain the structural characteristics of the other three metastable states, we calculate the most likely secondary structure of each residue in different states. As shown in Figs. 9 and 8(b), the feature of state $s3$ is that $\beta3$ and $\beta4$, i.e., the third and the fourth β sheet, respectively, are unfolded. The feature of state $s2$ is that $\beta2$ and the latter half of the helix are unfolded. In state $s1$, the helix of the native structure is unfolded and forms a β sheet instead. The four-state transition network is obtained and portrayed in Fig. 10. The main folding path is $s1 \rightarrow s2 \rightarrow s3 \rightarrow s4$, and the direct transitions from $s1$ to $s3$ are much smaller than those from $s1$ to $s2$. As presented in Ref. [20], from the collective variable based on the RMSD, only two states, i.e., the folded state and the unfolded state, can be distinguished. The four-state model constructed by TMDPC provides us with more details about the folding dynamics.

IV. CONCLUSION

In this paper, we proposed a useful approach, TMDPC, to analyze high-dimensional time series data. The TMDPC method first adopts the TM algorithm to obtain the slow variables of the system and then uses the DPC algorithm to identify the metastable states in the slow-variable space. The TM algorithm takes advantage of the temporal successiveness of conformations to construct slow variables rather than considering only the geometric similarity between conformations. Compared to applying only a few variables with experience, such as the RMSD, the slow variables constructed by the TM algorithm can better reflect the dynamic correlation in the time series. The DPC can use the distribution of the segment-averaged trajectory in slow-variable space to automatically identify metastable states with different shapes and densities. We studied the folding process of villin headpieces and protein G by TMDPC. It is worth mentioning that the extension to other kinds of time series data using the TMDPC approach is direct and does not change.

ACKNOWLEDGMENTS

This study was supported by the National Natural Science Foundation of China (Grant No. 11574310), the Program of Scientific Research Informatization, CAS (Grant No. XXH13506-404), and the Doctoral Program of Heze University (Grant No. XY18BS16). We thank D. E. Shaw Research for sharing their protein-folding trajectories.

[1] U. Maulik and S. Bandyopadhyay, Pattern Recogn. **33**, 1455 (2000).

[2] F. Sittel and G. Stock, J. Chem. Phys. **149**, 150901 (2018).

[3] A. K. Jain, Pattern Recogn. Lett. **31**, 651 (2010).

[4] H.-P. Kriegel, P. Kröger, J. Sander, and A. Zimek, WIRES Data Min. Knowl. **1**, 231 (2011).

[5] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed., Springer Series in Statistics (Springer, New York, 2002).

[6] A. Hyvärinen and E. Oja, Neural Netw. **13**, 411 (2000).

[7] C. R. Schwantes and V. S. Pande, J. Chem. Theory Comput. **9**, 2000 (2013).

[8] J. B. Tenenbaum, V. de Silva, and J. C. Langford, Science **290**, 2319 (2000).

[9] B. Nadler, S. Lafon, R. R. Coifman, and I. G. Kevrekidis, Appl. Comput. Harmon. Anal. **21**, 113 (2006).

[10] A. Rodriguez and A. Laio, Science **344**, 1492 (2014).

[11] L. Gong and X. Zhou, J. Phys. Chem. B **114**, 10266 (2010).

[12] L. Gong, X. Zhou, and Z. Ouyang, PLoS One **10**, e0125932 (2015).

[13] C. Zhang, J. Yu, and X. Zhou, J. Phys. Chem. B **121**, 4678 (2017).

[14] C. Zhang, F. Ye, M. Li, and X. Zhou, Sci. China Phys. Mech. **62**, 67012 (2018).

[15] J. Kubelka, T. K. Chiu, D. R. Davies, W. A. Eaton, and J. Hofrichter, J. Mol. Biol. **359**, 546 (2006).

[16] K. A. Beauchamp, R. McGibbon, Y.-S. Lin, and V. S. Pande, Proc. Natl. Acad. Sci. USA **109**, 17807 (2012).

[17] J. Kubelka, E. R. Henry, T. Cellmer, J. Hofrichter, and W. A. Eaton, Proc. Natl. Acad. Sci. USA **105**, 18655 (2008).

[18] A. Jain and G. Stock, J. Phys. Chem. B **118**, 7750 (2014).

[19] W. Du and P. Bolhuis, Biophys. J. **108**, 368 (2015).

[20] K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw, Science **334**, 517 (2011).

[21] J. Shimada and E. I. Shakhnovich, Proc. Natl. Acad. Sci. USA **99**, 11175 (2002).

[22] L. Lapidus, S. Acharya, C. Schwantes, L. Wu, D. Shukla, M. King, S. DeCamp, and V. Pande, Biophys. J. **107**, 947 (2014).

[23] C. Schwantes, D. Shukla, and V. Pande, Biophys. J. **110**, 1716 (2016).

[24] W. G. Touw, C. Baakman, J. Black, T. A. H. te Beek, E. Krieger, R. P. Joosten, and G. Vriend, Nucleic Acids Res. **43**, D364 (2015).