THE JOURNAL OF CHEMICAL PHYSICS **140**, 204905 (2014)

# Precursory signatures of protein folding/unfolding: From time series correlation analysis to atomistic mechanisms

P. J. Hsu,[1,2] S. A. Cheong,[3] and S. K. Lai[1,2,a)]
[1]*Complex Liquids Laboratory, Department of Physics, National Central University, Chungli 320 Taiwan*
[2]*Molecular Science and Technology Program, Taiwan International Graduate Program, Academia Sinica, Taipei 115, Taiwan*
[3]*Division of Physics and Applied Physics, School of Physical and Mathematical Sciences, Nanyang Technological University, 21 Nanyang Link, Singapore 637371, Republic of Singapore*

(Received 7 January 2014; accepted 29 April 2014; published online 22 May 2014)

Folded conformations of proteins in thermodynamically stable states have long lifetimes. Before it folds into a stable conformation, or after unfolding from a stable conformation, the protein will generally stray from one random conformation to another leading thus to rapid fluctuations. Brief structural changes therefore occur before folding and unfolding events. These short-lived movements are easily overlooked in studies of folding/unfolding for they represent momentary excursions of the protein to explore conformations in the neighborhood of the stable conformation. The present study looks for precursory signatures of protein folding/unfolding within these rapid fluctuations through a combination of three techniques: (1) ultrafast shape recognition, (2) time series segmentation, and (3) time series correlation analysis. The first procedure measures the differences between statistical distance distributions of atoms in different conformations by calculating shape similarity indices from molecular dynamics simulation trajectories. The second procedure is used to discover the times at which the protein makes transitions from one conformation to another. Finally, we employ the third technique to exploit spatial fingerprints of the stable conformations; this procedure is to map out the sequences of changes preceding the actual folding and unfolding events, since strongly correlated atoms in different conformations are different due to bond and steric constraints. The aforementioned high-frequency fluctuations are therefore characterized by distinct correlational and structural changes that are associated with rate-limiting precursors that translate into brief segments. Guided by these technical procedures, we choose a model system, a fragment of the protein transthyretin, for identifying in this system not only the precursory signatures of transitions associated with $\alpha$ helix and $\beta$ hairpin, but also the important role played by weaker correlations in such protein folding dynamics. © *2014 AIP Publishing LLC*. [http://dx.doi.org/10.1063/1.4875802]

## I. INTRODUCTION

A vast literature on the computer simulation of proteins [1–5] addresses three main questions: (1) what are the three-dimensional structures (native conformations) of a protein molecule [6–9]?; (2) how stable these structures are in relative and absolute thermodynamic terms [10–12]?; and (3) how quickly these conformations fold and unfold [10–12]? A fourth question which concerns the mechanistic pathways behind folding and unfolding processes can also be addressed by simulations. A mechanistic understanding of protein folding and unfolding can help us better understand how small molecules regulate biological functions, how proteins can misfold and aggregate into amyloids that are believed to be associated with brain disorders such as Alzheimer's, Parkinson's, and Bovine Spongiform Encelphalitis,[13–17] and how ultimately these studies can help accelerate the drug discovery and drug design process. Before high-performance computing tools became readily available, and before the advent of high spatial- and temporal-resolution of experimental methods,[18,19] early theories of protein folding speculated on the rate-limiting step of the process (see the review by Liu *et al.*[20]). These include the *framework model*,[21–23] where the formation of complete secondary structures such as $\alpha$ helices or $\beta$ sheets is rate limiting; the *hydrophobic collapse model*,[21–23] where the collapse of the initial open protein structure into a compact disordered globule is rate limiting; and the *nucleation-condensation model*,[24,25] where only the nucleation of folding centers is rate-limiting, and thereafter there is a rapid spontaneous folding into the secondary and tertiary structures. To discuss these competing model-based theories, and also to explain the spread in folding times over many orders of magnitude, Onuchic and co-workers proposed the existence of a rugged funnel-shaped folding energy landscape.[21,26,27]

As our computational power grew, so did our understanding of protein folding mechanism (see reviews by Refs. 22, 28–35). Through molecular dynamics (MD) and Monte Carlo studies, we now know that the protein folding mechanism is universal[36] and robust.[37] We can now use MD to accurately determine native structures, as well as transition states. Based on these structures, and the sequences of contacts made during transitions, we can now even talk about coarse-grained mechanisms. The present state of the art belongs to

a)Author to whom correspondence should be addressed. Electronic mail: sklai@coll.phy.ncu.edu.tw

 

Lindorff-Larsen *et al.*,[36] who used the supercomputer *Anton* to simulate more than 400 folding and unfolding events for 12 proteins domains with residues between 10 and 80. They found that most of the 12 proteins studied fold through a universal cluster of three to four pathways. These advances notwithstanding, we feel it should be possible to take our understanding of protein folding/unfolding mechanisms further, to elucidate the minimal sequences of elementary moves (for example, bond rotations) that bring the molecule from one conformation to another. This kind of work is pressing, because very soon we might be able to compare such detail mechanisms against the state of the art in experiments.[23]

Nevertheless, we face three related problems when trying to automatically extract mechanisms from MD simulations of a protein molecule (Sec. II A), as opposed to the iterated visual inspection by a human expert. First, parts of the molecule may fold into an $\alpha$ helix at times, and into a $\beta$ sheet at other times. Second, these folded conformations have longer lifetimes if they are thermodynamically more stable, and shorter otherwise. Before it folds into a stable conformation, or after unfolding from a stable conformation, the protein will also wander from one random conformation to another leading thus to rapid fluctuations. Finally, rate-limiting precursors involve brief structural changes before the folding/unfolding events. In this study, we look for these precursory signatures of protein folding/unfolding within these rapid fluctuations through a combination of three techniques: (1) ultrafast shape recognition, (2) time series segmentation, and (3) time series correlation analysis. Using the ultrafast shape recognition (USR) technique (Sec. II B), we exploit the differences between *statistical* distance distributions of atoms in different conformations by calculating low-dimensional similarity indices $\zeta(t)$ from the high-dimensional trajectories. Observing the similarity time series, we found that stable conformations experiencing different bond and steric constraints and they appear in the form of plateaus with different high-frequency fluctuations. These temporal fingerprints allow us to pinpoint the times at which the protein makes transitions from one conformation to another using the method of time series segmentation [38–41] (Sec. II C). More importantly, velocity fluctuations of various atoms in the protein molecule also become correlated due to the same bond and steric constraints. Spatial fingerprints consisting of different sets of strongly correlated atoms in different conformations can be discovered as well using the time series clustering or correlation analysis[42] (Sec. II D). From this high-frequency fluctuation point of view, structural changes associated with rate-limiting precursors translate into brief segments that are characterized by sharp correlational changes. These short-lived segments are easily overlooked for they represent momentary excursions of the protein away from its stable conformation, making attempts to explore nearby conformations.

In this paper, the above-mentioned hypothesis on the rapid fluctuations was tested on a fragment of the protein transthyretin applying the combined three techniques. Transthyretin is a soluble amyloidogenic protein synthesized in the liver. When misfolding or aggregation occurs, transthyretin or its mutant may transform itself into an insoluble fibrillar structure and is believed to be the causes of amyloid diseases such as senile systemic amyloidosis, familial amyloid polyneuropathy, familial amyloid cardiomyopathy, and central nervous system-selective amyloidosis.[43–45] Fibril formation is also linked to other protein deposition diseases such as Alzhemier's disease, type II diabetes, and the transmissible spongiform encephalopathies.[46] Therefore, an understanding of how this fibrillar formation occurs is of medical importance. Recently, Jaroniec *et al.*[47,48] studied the transthyretin (*TTR*) structures of residues from 105 to 115, i.e., *TTR*(105-115) using the magic-angle spinning solid-state NMR. Other high-resolution structural studies of *TTR*(105-115) investigated not only the self-assembly,[49,50] but also its remarkably high pressure, thermal and chemical stability. These efforts pave the way for finding the molecule suitable for use in nanomaterial and biosensor designs.[51,52] Because of the implications in protein misfolding disorders, there have been several recent computer simulations devoted to studying *TTR*.[49,50,53–57] Of these existing simulation works, there were efforts studying the early aggregation process with full atomic models as in Ref. 50 or, investigating in more recent years, the universal characteristic of amyloid aggregates in the context of interstrand conformational rearrangements, and looking at dimeric aggregates as in Ref. 54. Subsequently, Porrini *et al.*[55] performed more extensive simulation works on the *TTR*(105-115) ogligomers, exploring how, during the fibril assembly, the growth and stability of cross-$\beta$ aggregates are influenced by protonation in low pH environment. Other simulation works have been reported also, including the search for evidence of $\alpha$-sheet secondary structure in the prefibrillar amyloidogenic intermediate at acidic conditions[56] and the study of protein *TTR*[53] but not the peptide fragment *TTR*(105-115). In this paper, we have chosen *TTR*(105-115) as a test case to illustrate the statistical time series method for simulation studies. Apart from our very recent successful diagnosis of the dynamics of this system using an approach that combines MD simulation data and diffusion theory method,[57] and the realization that only limited simulation papers (Refs. 50 and 54–56) were reported for this peptide fragment, we were motivated also by the fact that, at room temperature, the *TTR*(105-115) can fold either into a $\alpha$ helix or a $\beta$ hairpin.

This paper is organized as follows. In Sec. II, we give details on the equilibrium MD simulation of *TTR*(105-115) in an aqueous solution, then moving on to describe the USR, time series segmentation, and time series correlation methods. We describe in this section how these three techniques can be combined neatly to discover folding and unfolding precursors by analyzing statistically significant correlations that are stronger or weaker than average. In Sec. III, we present results of time series segmentation applied to USR similarity index of the head-tail partial structure. This is the first time the time series method has been combined with the USR method, and the coarse-grained results pointed out a stable $\alpha$-helix phase, a stable $\beta$-hairpin phase, and an unstable mixed $\alpha/\beta$ phase in the simulation time series. We then explain the physical implications of persistent and transient correlational changes in relation to these phases. Our analysis distinguishes the former to be fingerprints of the phases, while the latter are precursory signatures of transitions before the associated phases. It is another purpose of this work to

examine each correlational precursory and its association with the dominant atomistic changes and both are discussed alongside in Secs. III E 1–III E 3. We found in fact two universal classes of strong precursory correlations and four universal classes of weak correlations. These latter discoveries which were summarized in Secs. IV A and IV B provide an invaluable means to construct the mechanistic pathway at the atomic scale and they are certainly complementary to the time series analysis. To appreciate further the time series analysis, we devote a subsection, Sec. IV C, to a comparison of it with the widely used contact analysis method. Finally, in Sec. V, we give a summary of our work and an outlook for our future endeavors.

## II. METHODS

### A. Molecular dynamics simulations

In our explicit MD simulation using GROMACS, the *TTR*(105-115) peptide (consisting of the 13 residues $Ace_1(A_1)$, $Tyr_2(T_2)$, $Thr_3(T_3)$, $Ile_4(I_4)$, $Ala_5(A_5)$, $Ala_6(A_6)$, $Leu_7(L_7)$, $Leu_8$ $(L_8)$, $Ser_9$ $(S_9)$, $Pro_{10}$ $(P_{10})$, $Tyr_{11}(T_{11})$, $Ser_{12}(S_{12})$, $Nac_{13}(N_{13})$) was initially prepared as a linear chain, as shown in Fig. 1(a). The *TTR*(105-115) peptide is embedded in an aqueous solution consisting of 1340 water molecules and the system is placed in a $3.083 \times 3.134 \times 4.428\ nm^3$ simulation cell. All-atom force fields are applied to both polymer and water molecules.[58] We use OPLS/AA semi-empirical potentials to describe the interactions among atoms within the peptide as well as between peptide residues and water molecules. The TIP4P potentials are used to account for the interactions between water molecules. Because the water molecules are initialized randomly, we first apply the LBFGS algorithm to minimize their total energy keeping the peptide fixed. Still keeping the peptide fixed, we then equilibrate the water molecules by the MD simulation for 5 ns, before running an equilibrium MD simulation for 930 ns with a time step of 1-fs using the Nosé-Hoover algorithm.[59,60] Our simulation was done at constant pressure of 1 atm keeping the temperature at 298 K.[61,62] The trajectory is stored every 0.5 ps and more than $1.8 \times 10^6$ data points were generated for our USR analysis.

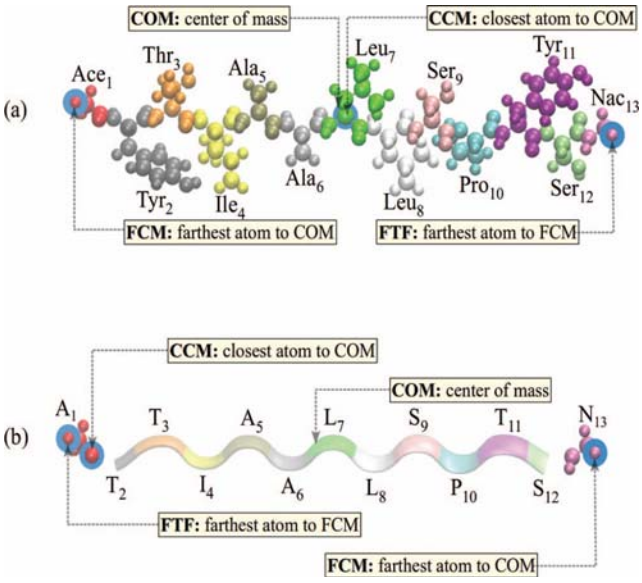


FIG. 1. The *TTR* polymer chain from $Ace_1$ to $Nac_{13}$ (shown in (a)) of which only 9 residues are distinct. Reference points used in calculating atomic distances at $t = 0$ for the entire *TTR*(105-115) peptide are defined in (a) and for the head-tail partial structure are defined in (b). The atoms in CPK forms are those used to calculate the reference points and atomic distances.

### B. Ultrafast shape recognition with partial structures

Shape matching and shape similarity methods were invented to address two main problems: (1) to classify entries in the protein data bank (PDB) based on their structures;[63–65] and (2) to accelerate drug design and discovery by understanding the shape complementary problem in ligand-receptor binding.[66–69] The second problem has recently received renewed interest, because experiments by Boström *et al.*[70] showed that structurally similar ligands can occupy the same region in the binding sites of the receptor.

Because of the large number of atoms, methods that infer structural characteristics from their coordinates a translation vector and a rotation matrix[71] to produce the best match between two protein molecules are computationally expensive. Such shape comparisons at atomic resolution are therefore not practical for understanding protein-folding kinetics within MD simulations. In principle, for both classification and shape complementarity problems only rough agreement between shapes is required. Recently, Ballester *et al.*[72] advanced a superposition-free algorithm, the USR technique, for ligand-based virtual screening. This method computes a rotation- and transition-invariant similarity, and is thousands of times faster than existing shape-matching techniques. A query that goes through billions of molecules in the PDB can be completed within few hours using an affordable computation resource. We were enticed by the speed of the USR technique to try it on the protein-folding problem.

The USR technique concerns basically with different conformations corresponding to different statistical distributions of atoms which may be translated into different statistical distributions of distances from a set of reference points. The statistical distribution of distances is rotation and translation invariant, and two distributions are different if their statistical moments are different. In this study, we use as reference points (1) the centroid of the atoms (COM), (2) the atom closest to COM (CCM), (3) the atom farthest from the COM (FCM), and (4) the atom farthest from the FCM (FTF) (Fig. 1). Cannon *et al.*[73] have applied the USR technique to virtual screening using the first four moments and shown that it is more accurate and efficient than using the first three or five moments. We therefore compute the first four moments for each of the four distributions of atomic distances. For atomic distances $r_i$, $i = 1, \ldots, n$ of $n$ atoms measured relative to a given reference point, the first four moments are: the *mean* distance,

$$\mu_1 = \frac{1}{n}\sum_{i=1}^{n} r_i, \tag{1}$$

which tells us how far the $n$ atoms are away from the reference point on average, the *variance*,

$$\mu_2 = \frac{1}{n}\sum_{i=1}^{n}(r_i - \mu_1)^2, \tag{2}$$

which measures the spread in the distribution of atomic distances, the *skewness*,

$$\mu_3 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{r_i - \mu_1}{\sqrt{\mu_2}}\right)^3, \tag{3}$$

which measures the degree of departure from symmetry of the distribution, and the *kurtosis*,

$$\mu_4 = \left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{r_i - \mu_1}{\sqrt{\mu_2}}\right)^4\right] - 3, \tag{4}$$

which measures the degree of peakedness of the distribution.

A total number of 16 statistical moment descriptors was obtained from these four moments of distance distributions that are calculated with respect to each of the four reference points, i.e., $\mathbf{M}(t) = \{\mu_1^{\mathrm{COM}}, \ldots, \mu_4^{\mathrm{COM}}, \ldots, \mu_1^{\mathrm{FTF}}, \ldots, \mu_4^{\mathrm{FTF}}\} = \{\mathrm{M}_l(t)\}_{l=1}^{16}$, one set of $\mathbf{M}(t)$ at each MD time step $t$. For two structures, for example, the reference configuration at $t = 0$ and the instantaneous configuration at a MD time step $t$, we then define the similarity,

$$\zeta(t) = \left(1 + \frac{1}{16}\sum_{l=1}^{16}|\mathrm{M}_l(t) - \mathrm{M}_l(0)|\right)^{-1}. \tag{5}$$

This structural similarity $\zeta(t)$ takes on values between 0 ($\mathbf{M}(t)$ least similar to $\mathbf{M}(0)$) and 1 ($\mathbf{M}(t)$ identical to $\mathbf{M}(0)$), and helps to provide a global perspective of the dynamics of our target biomolecule.

Because computations of shape similarity in the context of statistical moments are fast, we can study the instantaneous protein fragment structure with reference to its structure at $t = 0$ (or any other more specific reference structure if it is prepared beforehand) for the whole time elapsed of the MD simulation. We can, in principle, apply the USR technique to the entire protein molecule to obtain $\zeta(t)$ from $t = 0$ to any desired equilibrium time. However, like other shape-recognizing techniques, the USR method accurately identifies highly similar conformations to the reference structure but fails to discriminate between conformations that are dissimilar. To discover precursors, we must follow the structure evolution pathway and identify only those residues that are important during folding or during precursory changes if such signatures can be sorted out. Therefore, for practical consideration, we apply the USR technique not to the entire polymer molecule but instead to partial structures. By partial structures or substructures, we mean in the USR technique we calculate the statistical moments for the atomic distance distribution of part of all the residues within the whole polymer molecule. In our model of *TTR*(105-115) which consists of 13 residues, including $A_1$, $T_2$, $T_3$, $I_4$, $A_5$, $A_6$, $L_7$, $L_8$, $S_9$, $P_{10}$, $T_{11}$, $S_{12}$, and $N_{13}$, the partial structures, for example, may compose of a number of residues, such as 13 partial structures based on 1 residue at a time with respect to those at $t = 0$, and up to 10 substructures based on 4 consecutive residues.

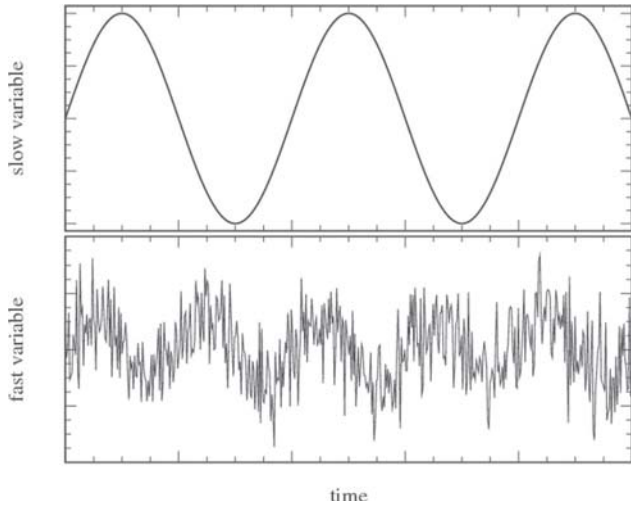


FIG. 2. Schematic diagrams showing slow (top) and fast (bottom) variables.

As a concrete illustration, consider the head-tail partial structure (Fig. 1(b)) consisting of only 12 atoms (residues $A_1$ and $N_{13}$) to which are connected between them the remaining 169 atoms constituting 11 residues (the color ribbon in Fig. 1(b)). In the calculation of the head-tail partial structure, only the atomic distances between head and tail residues are tracked. All other 11 residues in the molecule (Fig. 1(b)) are still there but are judiciously ignored in calculating atomic distances. Some remarks are in order about our choice of head-tail substructure to be used in the time series segmentation. In the first place we note that within the profile of stable conformations as slow manifolds, there are slow variables (order parameters) that characterize the slow manifolds, along with fast variables whose fluctuations look very much like noise. Naturally, the sets of slow variables for different stable conformations are different, although they can overlap. More importantly, the Haken's slaving principle tells us that the fast fluctuations are also statistically different in different stable conformations. The problem now is that we do not *a priori* know which variables are fast, and which variables are slow. If we do, then within a stable conformation, we should be able to obtain graphs that look like those shown in Fig. 2. In general, fast and slow variables are combinations of the microscopic variables. Therefore, all quantities we recorded from the peptide simulations will contain fast and slow dynamics. When fast and slow dynamics are mixed, as they would be in the end-to-end distance, it is generally more difficult to identify the slow manifolds and transitions between them. The similarity index $\zeta(t)$ introduced here, by design, has the advantage that the slow collective changes do not change the similarity index $\zeta(t)$.

In Fig. 3, we show the end-to-end distance $r_{\text{end-to-end}}$, the head-tail substructure $\zeta(t)$ and the whole peptide structure $\zeta(t)$ for *TTR*. The overall characteristics of all three are rather similar. Quantitatively the temporal fluctuations of the end-to-end distance $r_{\text{end-to-end}}(t)$ are generally larger than the whole or head-tail $\zeta(t)$, especially in mixed $\alpha/\beta$ regions (536-725 and 891-914 ns). If we were to carry out time series segmentation on the end-to-end distance time series, we would expect to

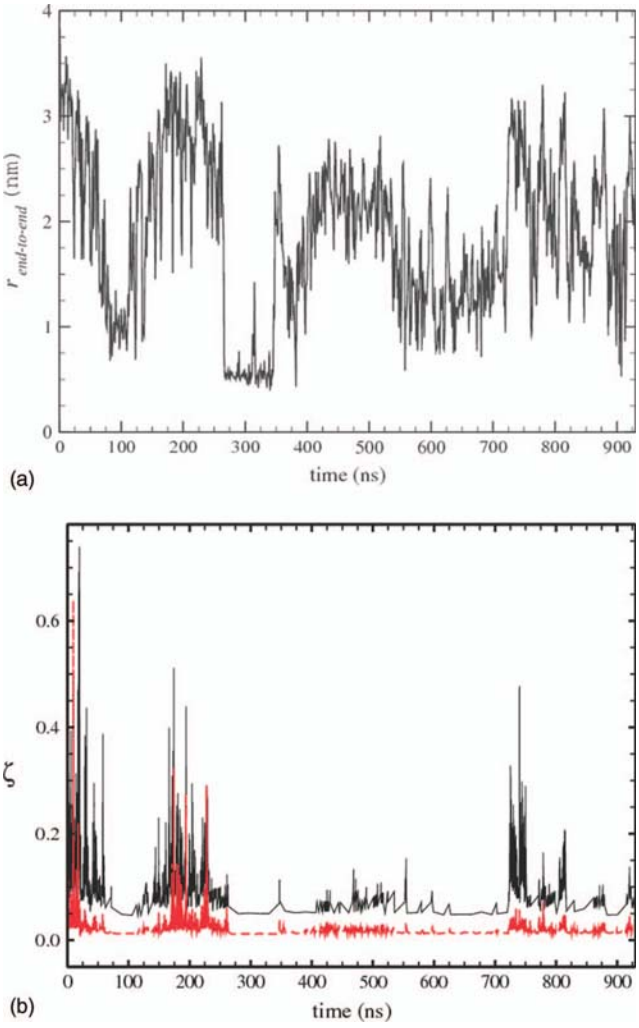


FIG. 3. (a) The end-to-end distance $r_{\text{end-to-end}}$ of C atom in $\text{Ace}_1$ and N atom in $\text{Nac}_{13}$, which is the head-tail distance in *TTR*(105-115), and (b) the similarity index functions of *TTR*(105-115) for the head-tail partial structure (red dashed line) and the whole structure (black) with 181 atoms included. Notice that the whole and head-tail $\zeta(t)$ are strikingly similar but the computing time to construct the head-tail $\zeta(t)$ is about 10-15 times less.

obtain different segment boundaries, perhaps less in number for the head-tail $\zeta(t)$. A study with relatively more segment boundaries will make the time series analysis somewhat intricate. It is, however, of less consequence to the present study since we are interested in stable conformations as slow manifolds which show up clearly in the head-tail $\zeta(t)$.

Since the head-tail time series similarity $\zeta(t)$ is strikingly similar to the $\zeta(t)$ with all 181 atoms included (Fig. 3(b)), this means that to understand the folding dynamics of the molecule, we can equally well apply the USR technique to the head-tail residues and was done in this work. An immediate advantage is that the computing time spent is substantially reduced. Other types of partial structures may be done in the same manner. In fact, besides the head-tail partial structure (Fig. 1(b)), we have tested partial structures with up to four consecutive residues, and arrived at the conclusion that partial structures with three consecutive residues, such as $\text{A}_1$-$\text{T}_2$-$\text{T}_3$, $\text{T}_2$-$\text{T}_3$-$\text{I}_4$, $\text{T}_3$-$\text{I}_4$-$\text{A}_5$, …, $\text{T}_{11}$-$\text{S}_{12}$-$\text{N}_{13}$ (cf. Fig. 1(b) consisting only residues $\text{A}_1$ and $\text{N}_{13}$) give the best signal-to-noise ratio to our correlation analysis. As the results here showed, the USR technique is not restricted by the size of the polymer molecule because only a few statistical moments of atomic distance distributions are used in the calculations. The method, as it stands, has the flexibility of analyzing partial structures without losing the essential information of the dynamics of polymer motions.

Consider then the head-tail (Fig. 1(b)) and the three consecutive residues (3-residues). We compute the similarity time series index of them all with respect to their respective initial configuration ($t = 0$), and thereafter, focus on the digital cross correlations of 3-residues which are used in our time series correlation study. Our correlation analysis will, therefore, be centered around the partial structures of 3-residues.

### C. Time series segmentation

The theoretical foundations for the method of time series segmentation go back to the seminal works by Haken [74–76] and Pecora and Carroll[77] after him, where high-frequency fluctuations are found to be slaved to the low-dimensional manifolds which a system settles into. Folded conformations are such low-dimensional manifolds, because the molecular trajectory visits a small region in phase space as a result of bond and steric constraints. This would imply that high-frequency fluctuations are not simply noise, but carried instead information of slow manifolds about which the protein is in. Since low-dimensional manifolds are generally well separated in phase space, their high-frequency fluctuations should also be distinguishable from each other statistically. We can therefore exploit this fact and determine much more accurately when the polymer molecule folds into an $\alpha$ helix or a $\beta$ hairpin, and more importantly, identify brief precursory segments before such folding events.

To partition or segment a time series into intervals within which are associated with different stable conformations (low-dimensional manifolds), we adapt a recursive entropic segmentation procedure first introduced for biological sequence segmentation by Bernaola-Galván *et al.*[78,79] It works as follows. First, we assume that the high-frequency fluctuations are statistically independent samples from a Gaussian distribution. This modeling choice does not pose any problem in practice, even though the autocorrelation time in MD simulations of polymer molecule is longer than the sampling time of 0.5 ps for $\zeta(t)$ (due to the way we prepared the *TTR* polymer), because we can still find the most significant segment boundaries. Within such an approximation, the likelihood to observe a given sequence of similarities $\vec{\zeta} = (\zeta_1, \zeta_2, \ldots, \zeta_i, \ldots, \zeta_n)$ would be given by

$$L_1 = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(\zeta_i - \gamma)^2}{2\sigma^2}\right], \tag{6}$$

if the similarity indicies $\zeta_i$ are all sampled from a Gaussian distribution with mean $\gamma$ and variance $\sigma^2$. However, we know that the protein fragment folds and unfolds over the course of the MD simulation, and hence the similarity time series $\vec{\zeta} = (\zeta_1, \zeta_2, \ldots, \zeta_i, \ldots, \zeta_n)$, for instance, the head-tail partial structure, must be statistically nonstationary, consisting of

many statistically stationary segments. The $\zeta(t)$ of this partial structure is shown in Fig. 3(b). Since we do not *a priori* know how many such segments are there, and where these segment boundaries lie, let us start by assuming that there might be two segments, with a segment boundary at $i = t$. The likelihood to observe $\vec{\zeta} = (\zeta_1, \zeta_2, \ldots, \zeta_i, \ldots, \zeta_n)$ in such a two-segment model is then given by

$$L_2(t) = \prod_{i=1}^{t} \frac{1}{\sqrt{2\pi\sigma_L^2}} \exp\left[-\frac{(\zeta_i - \gamma_L)^2}{2\sigma_L^2}\right] \prod_{i=t+1}^{n} \frac{1}{\sqrt{2\pi\sigma_R^2}} \times \exp\left[-\frac{(\zeta_i - \gamma_R)^2}{2\sigma_R^2}\right], \quad (7)$$

where the left segment $\vec{\zeta}_L = (\zeta_1, \zeta_2, \ldots, \zeta_t)$ is sampled from a Gaussian distribution with mean $\gamma_L$ and variance $\sigma_L^2$, while the right segment $\vec{\zeta}_R = (\zeta_{t+1}, \zeta_{t+2}, \ldots, \zeta_n)$ is sampled from another Gaussian distribution with mean $\gamma_R$ and variance $\sigma_R^2$. For any time series, we can *always* fit it to a model, either a 1-segment model or a 2-segment model. To decide on which model is better, we compute the logarithm of the ratio of likelihoods

$$\Delta(t) = \ln \frac{L_2(t)}{L_1} = n \ln \hat{\sigma} - t \ln \hat{\sigma}_L - (n-t) \ln \hat{\sigma}_R + \frac{1}{2}, \quad (8)$$

where $\hat{\sigma}$, $\hat{\sigma}_L$, and $\hat{\sigma}_R$ are the maximum-likelihood estimates of the standard deviations. The larger $\Delta(t)$ is, the better the 2-segment model fits the data compared to the 1-segment model. It can be shown that $\Delta(t)$ is $n$ times the Jensen-Shannon divergence (JSD).[80]

As a concrete means to see how the segmentation procedure operates, we redraw in Fig. 4(a) the time series of head-tail similarity index and indicate in this figure the assumed position of the segment boundary obtained from the $\Delta(t)$ depicted in Fig. 4(b). Note that the optimum time to segment $\vec{\zeta} = (\zeta_1, \zeta_2, \ldots, \zeta_i, \ldots, \zeta_n)$ into 2 segments that are statistically most distinct is at $t^* = 250$ ns, where $\Delta(t^*)$ peaks. To find more segment boundaries, we calculate $\Delta_L(t)$ (green) and $\Delta_R(t)$ (blue) separately, i.e., entirely within the left and right segments. Figure 4(b) also shows the positions of the segment boundaries that we segment further the left and right segments into subsegments at around $t_L^* = 20$ and $t_R^* = 720$ ns, respectively. This 1-to-2 segmentation procedure can then be iterated to find out more and more segment boundaries.

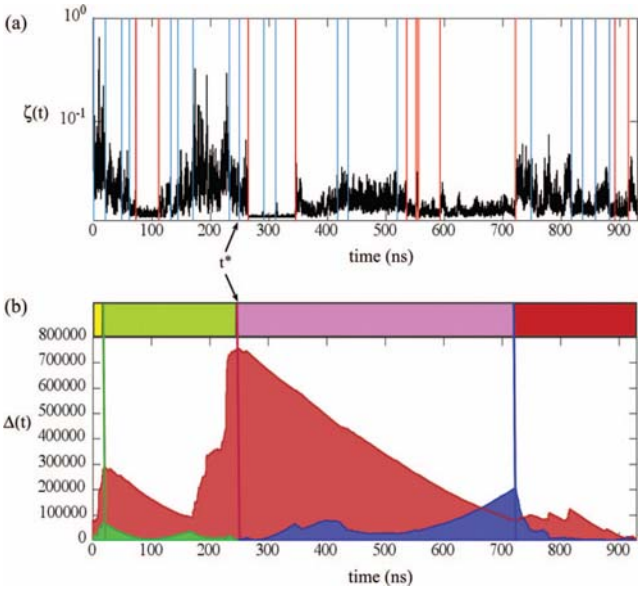


FIG. 4. (a) The shape similarity index $\zeta$ (black) of the head-tail residues of *TTR*(105-115), and the final 29 boundaries generated by the recursive JSD segmentation scheme (see text) marked by blue and red vertical lines. The red vertical lines delimit regions of lower similarity ($\zeta < 0.2$), and (b) the JSD for the head-tail residues of *TTR*(105-115), for the first (red) and second (green and blue) segmentation recursions.

As more and more segments are found, they become shorter and shorter (down to a lower limit of $2 \times 10^4$ data points in our study). With shorter segments, the JSD maxima $\Delta^* = \Delta(t^*)$ which are associated with new segment boundaries as well as some old segment boundaries also become smaller. At some point, $\Delta^*$ becomes so small that new segment boundaries are no longer statistically significant to be informative, and the recursive segmentation process will then be terminated. There are three systematic ways to do so in the literature. In the first, the hypothesis testing approach,[78,79] we check $\Delta^*$ of new segment boundaries. If all the new segment boundaries can be explained in terms of statistical fluctuations within stationary segments at a desired level of confidence, they are rejected and the segmentation terminates. In the second, the model selection approach,[81,82] we compute the Bayesian information criteria (BIC)[83] difference between the 1- and 2-segment models. If this BIC difference becomes positive (1-segment model has more explanatory power than the 2-segment model) for all new segment boundaries, they are rejected and the segmentation terminates. Finally, in the third, the intrinsic fluctuations approach,[84] we compare $\Delta(t)$ against a coarse-grained version of itself to compute the strength $\delta\Delta$ of intrinsic fluctuations within $\Delta(t)$. Denoting $\Delta_{\max} = \max[\Delta(t)]$ and thinking of $\Delta_{\max}/\delta\Delta$ as a signal to noise ratio, the segmentation terminates if the $(\Delta_{\max}/\delta\Delta)$s in all of segments fall below a chosen threshold.

In this work, we proposed a simpler segmentation scheme than the above three more rigorous approaches. Using this simpler segmentation procedure, we are able to obtain the terminal segment boundaries and when these are superimposed on $\zeta(t)$, one sees clearly the relevance between the segment boundaries in $\zeta(t)$ (Fig. 4(a)) and $\Delta(t)$ (Fig. 4(b)). We defer more detailed description of the results of segments to Sec. III B. Besides the folded conformations, we are also interested in brief time series segments that occur shortly before the folding and unfolding events, as these are likely to be precursors. The time series segmentation thus allows us to pinpoint when these occur, but is limited in it characterizing such precursors. We treat, therefore, the segmentation analysis of the head-tail time series of $\zeta(t)$ as an exploratory exercise to narrow down some time windows of interest, and proceed to use time series correlation analysis to characterize these folding and unfolding precursors.

### D. Time series correlation analysis

Since the conformation of the protein fragment is determined dynamically by interactions between its constituent

residues, we can, in principle, monitor all possible dihedral angle changes to track its structural evolution. It is, however, a formidable task or computationally laborious to examine the force fields explicitly and track all dihedral angles systematically starting from the stream of huge trajectory data. This is where the time series correlation analysis can be invaluable. The basic idea here is that the more strongly two residues interact, the more synchronize or anti-synchronize their motions will be. Therefore, the cross correlations between residues are relevant and good proxies for the forces that act between them. Given the similarity time series $\zeta(t)$ of a partial structure, we define the change in it to be

$$\Delta p\,(t_k) = \zeta\,(t_{k+1}) - \zeta\,(t_k) \tag{9}$$

and analyze correlations between the $\Delta p(t)$ instead of the time series of the similarity index $\zeta(t)$, because interaction forces are buried in the structural changes.

Based on the above premise, we then measure the synchronicity of structural changes by calculating the cross correlations between the changes of $\Delta p_i(t)$ and $\Delta p_j(t)$ for partial structures $i$ and $j$ respectively. Different types of cross correlations can be calculated as well. For example, the Pearson or linear cross correlation, or the Spearman or rank cross correlation, each has its own advantages and disadvantages. In this study, we use the *digital cross correlation* (DCC) previously defined by Lai *et al.*[42] in their study of the melting transition temperature of metallic clusters. In their calculations, they pointed out that the DCC is less biased towards small deviations, computationally inexpensive, and can be evaluated repeatedly for a long set of MD simulation data. Applying the same strategy to evaluate the DCC between $\Delta p_i(t)$ and $\Delta p_j(t)$, we compute first the *sign change* of $\Delta p_i(t)$ for the $i$th partial structure by

$$S_i(t_k) = \begin{cases} 1, & \Delta p_i(t_k) > 0 \\ 0, & \Delta p_i(t_k) = 0 \\ -1, & \Delta p_i(t_k) < 0 \end{cases}, \tag{10}$$

and do alike for $\Delta p_j(t)$ and hence $S_j(t)$. Then we construct the DCC matrix

$$C_{ij} = \frac{(1-\delta_{ij})}{2} \sum_{k=t/\Delta t}^{t'/\Delta t} \left| S_i(t_k) S_j(t_k) [S_i(t_k) + S_j(t_k)] \right| \tag{11}$$

for these partial structures within a given time window $[t,t']$.

Next, we examine the correlational structures at different times in the simulation. The choice of the sliding time window has to be made with caution. A period too long may average out the transition signatures, whereas too short a period will make the computed cross correlations statistically fluctuate strongly, thus becoming unstable from one time window to the next. We therefore has to scrutinize the profile of time series segmentation and guided by these results choose the window size. In principle, it should not be larger than the shortest segment. The essential point to grasp in this step is by looking at the head-tail $\zeta(t)$ in Fig. 4(a). Specifically, one can read from this figure that it is around 10 ns. Therefore, a 5-ns time window would be a reasonable value to slide forward at 1 ns at a time. After which the time evolution of $C(i,j) \equiv C_{i,j}$ is examined. According to Eq. (11), $C_{i,j}$ then has the minimum value equal to 0 and the maximum value $10^4$.

## III. RESULTS OF TIME SERIES ANALYSIS

### A. Similarity time series

From the MD trajectories, we calculated the time series of $\zeta(t)$ for two types of partial structures. In the preliminary exploration and time series segmentation stage of the study, we examine how distances and orientations of the head ($A_1$) and tail ($N_{13}$) residues deviate from the initial state, with atoms in other residues still there but were not explicitly included in the calculation of atomic distances. The time series of $\zeta(t)$ for this head-tail partial structure (Fig. 1(b)) has already been shown above in Fig. 3(b) (dashed red line) or Fig. 4(a) for the convenience of introducing the time series segmentation method. Note that in using the 12-atom head-tail partial structure time series $\zeta(t)$ instead of that of the whole peptide structure containing 181 atoms the computing time is enormously reduced by approximately 10-15 times and yet the two constructed $\zeta(t)$s are strikingly similar (Fig. 3(b)). The time series segmentation method is thus applied to the head-tail time series $\zeta(t)$ to obtain the segment boundaries (Fig. 4(a)). In this subsection, the eleven overlapping 3-residues $\zeta(t)$ defined in Sec. II B were calculated in the same manner as the head-tail time series index. These latter time series functions were then used for constructing the $C(i,j)$ color map. We deferred, however, to present the 3-residues $\zeta(t)$s in Appendix A to which the interested readers are referred to for further details.

### B. Segments of head-tail similarity time series

The choice of the head-tail partial structure allows an efficient pinpoint of different stages of peptide folding without having to deal with a large number of atoms. For this partial structure (Fig. 1(b)), we discovered 29 segment boundaries using the time series segmentation method with a cutoff $\Delta_0 = 200$ to ensure that there are no segments shorter than 1 ns. Of these 30 time series segments, the shortest one is 3.6 ns, while the longest is 129.1 ns. The average segment length is 33 ns. The segment boundaries are given in Fig. 4(a). Note that in this figure we have marked some segment boundaries red to highlight the time intervals with lower similarities $\zeta(t) < 0.2$ (basins). In Appendix B, we described in more details on how we proceeded and solved the technical implication[85] with our simpler segmentation procedure. We listed also there the numerical values of their JSD as well as the precise time locations of the segment boundaries.

We see from Fig. 4(a) that in the strongly fluctuating $\zeta(t)$ of the head-tail partial structure, the most prominent features are low-similarity plateaus and they are marked by the red segment boundaries. By delving into the conformations within these plateaus, we found that they are stable $\alpha$ helices (within, for example, $73.2 < t < 111.5$ ns) and $\beta$ hairpins (within, for example, $264.8 < t < 346.3$ ns). This demonstrates how accurately the time series segmentation determines the folding and unfolding times.

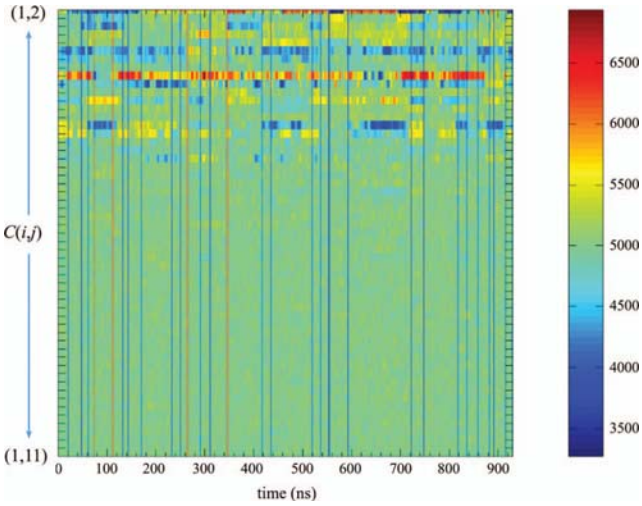


FIG. 5. The unfiltered color map of correlations over the 930 ns with sliding windows each of 5 ns for 3-residues. Because the windows are slided in 1-ns steps, the window number is also the beginning of the window at time $t$ = 1 ns for the first window, 2 ns for the second window, and so on. In this color map, the correlations are arranged such that those between nearest-neighbor partial structures $C(i,i+1)\equiv C_{i,i+1}$ are shown first, followed by those between next-nearest-neighbor partial structures $C(i,i+2)$, and so on. The tick labels along $y$-axis, from top to bottom, of $C(i,j)\equiv(i,j)$ run as (1,2), (2,3), (3,4), (4,5), (5,6), (6,7), (7,8), (8,9), (9,10), (10,11), (1,3), (2,4), (3,5), (4,6), (5,7), (6,8), (7,9), (8,10), (9,11), (1,4), (2,5), (3,6), (4,7), (5,8), (6,9), (7,10), (8,11), (1,5), (2,6), (3,7), (4,8), (5,9), (6,10), (7,11), (1,6), (2,7), (3,8), (4,9), (5,10), (6,11), (1,7), (2,8), (3,9), (4,10), (5,11), (1,8), (2,9), (3,10), (4,11), (1,9), (2,10), (3,11), (1,10), (2,11), (1,11).

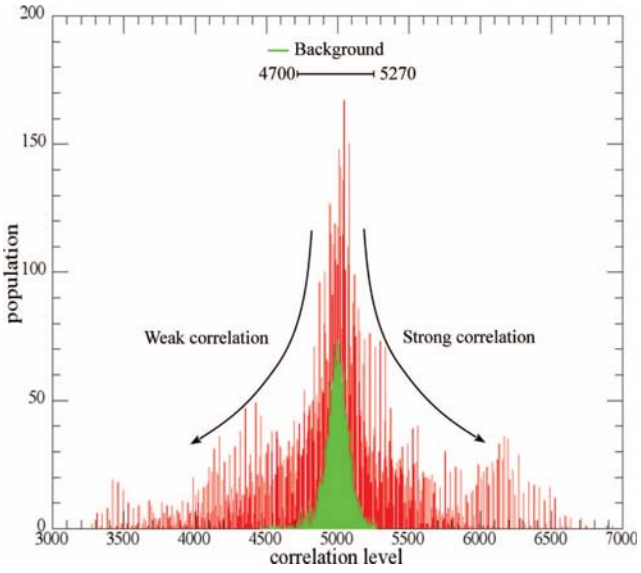


FIG. 6. The combined histogram of correlation levels from all pairs of 3-residues, distinguishing the strong and weak correlations from background correlation levels. The red histograms are correlations between neighboring or next neighboring 3-residues, whereas the green histograms are correlations between 3-residues farther away from each other, third nearest neighbors and beyond.

### C. Color map of cross correlations

In Fig. 5, we present the color maps of the DCCs between partial structures in different time windows for the eleven 3-residues ($A_1$-$T_2$-$T_3$, $T_2$-$T_3$-$I_4$, $T_3$-$I_4$-$A_5$,..,$T_{11}$-$S_{12}$-$N_{13}$). Compared with 1-, 2- and 4-residues (Appendix C), the intensity of color map in this figure reveals that the correlations of 3-residues have the highest contrast and this renders the structural evolution of 3-residues valuable information on the changes of the dihedral angle of the backbone. Further statistical analysis of the correlations reveals moreover that the uniform background signals of 3-residues are mostly contributed by $C(i,j)$, $j>i+2$, and form a Gaussian-like distribution spanning the range 4700-5270 (green histogram in Fig. 6). Strong correlations are not only separated, but also weaker-than-average correlations. This would mean that the weak/strong separation is the most significant feature of 3-residues partial structures, rather than those shown in Appendix C. For the *TTR*(105-115) peptide, the 3-residues partial structures therefore offer the best compromise between spatiotemporal resolution and stability of correlation features and will be employed below for detailed analysis of its dynamics.

### D. Correlation filtering

On the color map in Fig. 5, we superimpose the terminal segment boundaries as blue vertical lines, and the starts and ends of troughs in the head-tail time series of $\zeta$(t) as red vertical lines (we shall see later that these correspond to $\alpha$-helix, $\beta$-hairpin, and mixed conformations in Sec. III E 3).

We found that some cross correlations remained more or less unchanged throughout the simulation, while others showed significant changes at various times during the simulation. For 3-residues pairs that are third nearest neighbors and beyond, the color maps are in fact almost uniformly green ($C(i,j) \approx$ 5000). As explained earlier, we expect folding and unfolding precursors to be associated with brief and distinguishable correlational changes. Therefore, 3-residues pairs with nearly constant correlations are unlikely to have any connection with precursory changes, whereas 3-residues pairs with significant correlational changes are likely to be part of the precursors. Bearing this in mind, we now focus our attention on 3-residues pairs that exhibit stronger-than-average and weaker-than-average correlations. To do so, we need first of all to systematically separate strong (red in Fig. 5) and weak (blue in Fig. 5) correlations from background (green in Fig. 5) correlations. This insightful observation on correlational changes leads to the histogram of all correlations (red) in Fig. 6, and on it is superimposed the histogram of correlations between third nearest neighbors and beyond (green).

From Fig. 6, we see that $C(i,j)$ between 3-residues pairs farther than next nearest neighbors are narrowly distributed (the green sub-distribution), in the range $4700 < C < 5270$. Only $C(i,j)$ between neighboring and next neighboring 3-residues pairs have correlations outside of this range. Therefore, the range $4700 < C < 5270$ is treated as the band of background correlations based on which we define the (1) stronger-than-average correlations, $C > 5270$ (which follow a sub-distribution pattern that peaks around $C = 6200$) and the (2) weaker-than-average correlations, $C < 4700$ (which follow a sub-distribution pattern that peaks around $C = 3500$). With these definitions, all of 3-residues pairs, a total number of 52 pairs, whose $C(i,j)$ that fall always within the

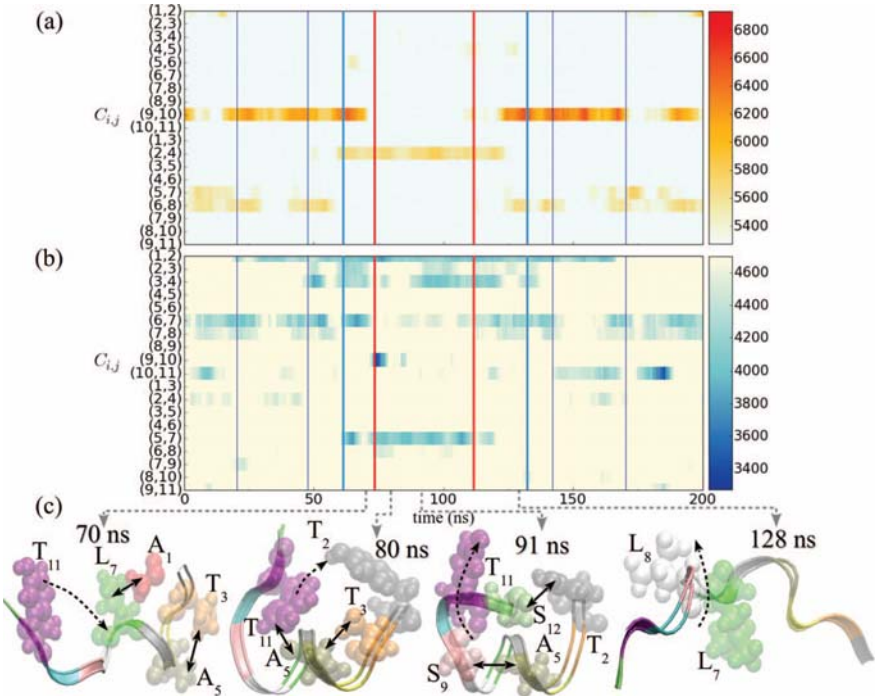


FIG. 7. The color map of correlations $C_{ij} \equiv C(i,j)$ from 0 to 200 ns for (a) above-average ($C_{ij} > 5270$) and (b) below-average correlations ($C_{ij} < 4700$) where $i, j = 1, 2, \ldots, 11$ denote the eleven 3-residues partial structures defined in text. The red and blue vertical lines are terminal segment boundaries calculated from the time series $\zeta(t)$ of the head-tail partial structure (Fig. 3(b)). Snapshots of *TTR*(105-115) shown in (c) are at various times (70, 80, 91, and 128 ns from left to right). In these snapshots, dashed arrows indicate the directions of motion of residues, while solid double-headed arrows indicate possible non-bonding interactions.

background band are neglected. Consequently only 19 pairs of 3-residues remain after this correlation filtering.

Before proceeding to our results, we note the following. The color maps showing stronger-than-average correlations (Figs. 7(a)–9(a)) are colored red and they describe correlations that are most different from background which is colored light cyan, whereas Figs. 7(b)–9(b) showing weaker-than-average correlations which are colored blue are most different from background which is colored yellow. Based on this reading convention, the $C(i,j)$ in any of Figs. 7(a)–9(a) is described enhanced (suppressed) if the color map shows its color approaching red (light cyan). The opposite meaning applies, however, to the $C(i,j)$ in any of Figs. 7(b)–9(b) where now we say the $C(i,j)$ is suppressed (enhanced) if its color approaches blue (light yellow). Accordingly, when reading changes in color, if $C(i,j)$ seen in Fig. 7(b), say has a blue color, is changing to light yellow, the correlation $C(i,j)$ is described as enhanced, whereas when approaching blue the $C(i,j)$ is suppressed. The color maps depicted in Figs. 7–9 will be dissected with the aid of this reading convention to gain insights into the folding and unfolding processes.

## E. Fingerprints and precursors

To illustrate how protein folding and unfolding precursors can be analyzed and identified, we return to Fig. 4(a) and focus on three specific time intervals: (1) around 100 ns when the peptide folds into an $\alpha$ helix, (2) around 300 ns when the peptide folds into a $\beta$ hairpin, and (3) between 500 and 700 ns when the peptide alternates in its morphology between $\alpha$ helix and $\beta$ hairpin. We will examine strong and weak correlations showing both persistent as well as transient changes, and explain how persistently strong or persistently weak correlations constitute unambiguous evidence for stable conformations as low-dimensional manifolds. Equally duly attention is paid to examining the transient strong and weak correlations which are classified as precursory signatures of protein folding/unfolding, and finally, we relook in Sec. IV at what they mean as precursor signatures from an atomistic perspective.

### 1. $\alpha$-helix conformation

In Figs. 7(a) and 7(b), we show pairing of correlation changes with few representative snapshots depicted in Fig. 7(c). Between 73 and 112 ns (Fig. 4(a)), the peptide reveals an $\alpha$-helix topology (Fig. 7(c)). We therefore expect to see precursors of the folding/unfolding events before the segment boundaries marked by red vertical lines. From the above-average correlations $C(i,j) \equiv C_{ij} > 5270$ (Fig. 7(a)), we see that $C(2,4)$ is enhanced at around 60 ns and thereafter remains so until after the $\alpha$-helix unravels. At around the same time, the correlation $C(9,10)$ shows a brief strong enhancement before falling off rapidly to background level at $\sim$70 ns. Within the $\alpha$-helix, we see from the below-average correlations $C_{ij} < 4700$ that $C(6,7)$ and $C(7,8)$ are persistently enhanced, $C(6,8)$ is weakly suppressed, and suppressions also in $C(2,3)$ and $C(3,4)$, whereas $C(1,2)$, and $C(5,7)$ are persistently suppressed. Persistent changes in correlations within a stable conformation are reminiscent of the Haken's slaving principle, whereby the character of high-frequency

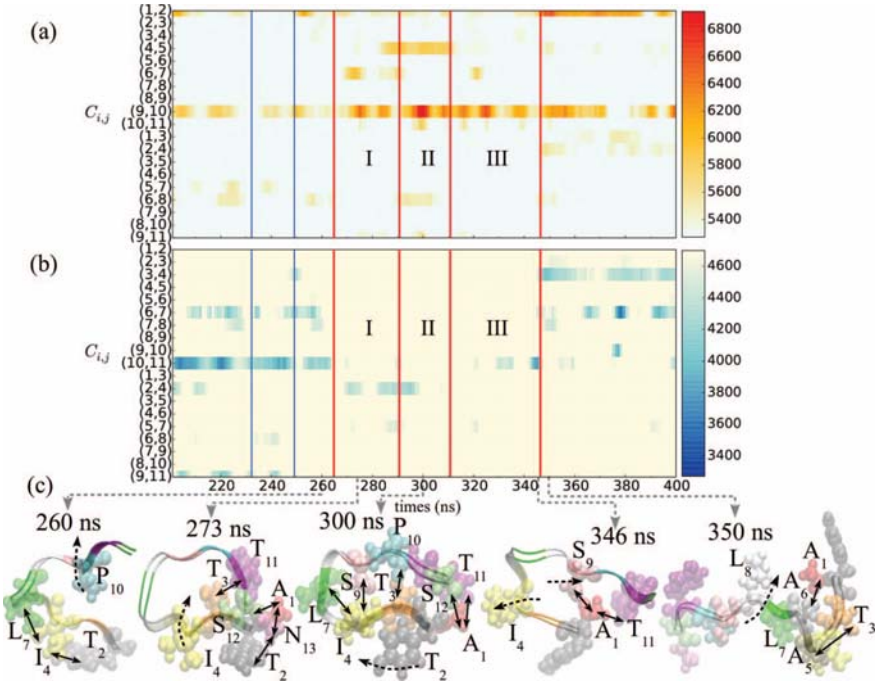


FIG. 8. The color map of correlations $C(i,j) \equiv C_{ij}$ from 200 to 400 ns, for (a) above-average ($C_{ij} > 5270$) and (b) below-average ($C_{ij} < 4700$) correlations. The vertical lines are terminal segment boundaries calculated from the time series $\zeta(t)$ of the head-tail partial structure. Note that the zone II marks the lifetime of the $\beta$-hairpin conformation. Snapshots of *TTR*(105-115) shown in (c) are at various times (260, 273, 300, 346, and 350 ns from left to right). In these snapshots, dashed arrows indicate the directions of motion of residues, while solid double-headed arrows indicate possible non-bonding interactions.

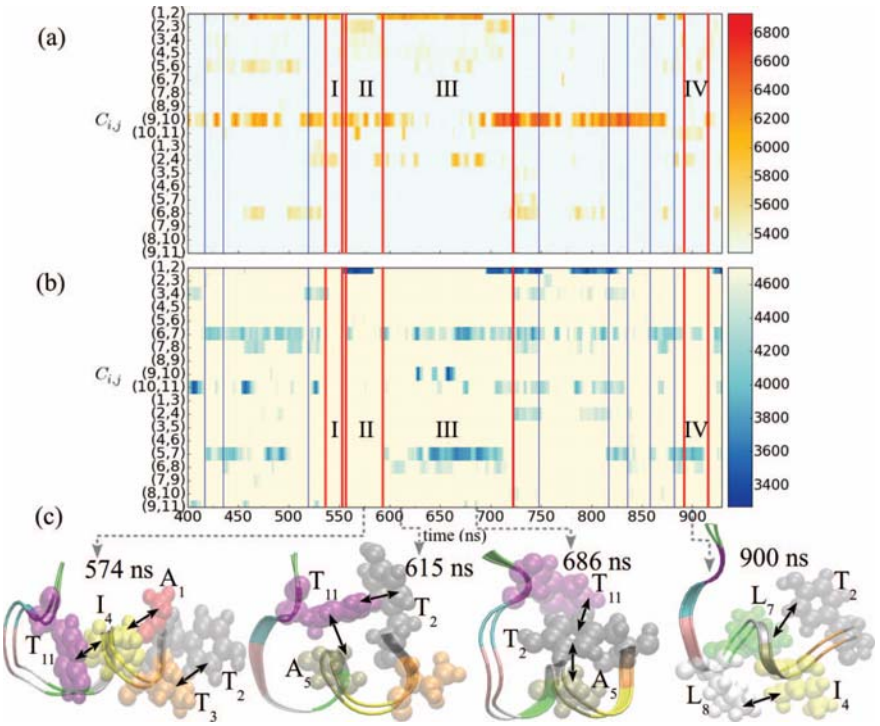


FIG. 9. The color map of correlations $C(i,j) \equiv C_{ij}$ from 400 to 930 ns for (a) above-average ($C_{ij} > 5270$) and (b) below-average ($C_{ij} < 4700$) correlations. The vertical lines are terminal segment boundaries calculated from the time series $\zeta(t)$ of the head-tail partial structure. Red vertical lines mark the beginning and end of mixed $\alpha$-helix and $\beta$-hairpin phases, whereas blue vertical lines mark transitions associated with transition states. Selected snapshots educed from MD trajectory are shown in (c), (from left to right) at 574, 615, 686, and 900 ns. In these snapshots, solid double-headed arrows indicate possible non-bonding interactions.

fluctuations is determined by the low-dimensional manifolds within which the peptide is in. We therefore think of these *persistent* correlational changes as *fingerprints* of the $\alpha$-helix phase.

Next we look for brief correlational changes before the folding and unfolding events. Before the peptide folds into an $\alpha$ helix, there are two short time series segments, i.e., (1) ~50 to ~60 ns, and (2) ~60 to ~70 ns deserved mentioning. Many brief correlational changes occur here, and we therefore think of these two time intervals as *precursory* segments. We notice in fact that the fingerprints in Fig. 7(a) displaying persistent enhancement of $C(2,4)$ and persistent suppression of $C(6,8)$, occur at the beginning of the second precursory segment. The other fingerprints, persistent suppression of $C(9,10)$ (Fig. 7(a)), persistent enhancement of $C(6,7)$ (Fig. 7(b)), and persistent suppression of $C(5,7)$ (Fig. 7(b)), appear only at the beginning of the $\alpha$-helix phase. Within these two precursory segments, we see moreover from Fig. 7(a) that a brief suppression of $C(9,10)$ in the first precursory segment, and it is followed by a brief enhancement in the second precursory segment. Furthermore, we also find from Fig. 7(b) brief suppressions followed by brief enhancements for $C(3,4)$ and $C(6,7)$ within the first precursory segment. These latter two correlations are in fact seen to show brief suppressions again in the second precursory segment, before it is raised to background levels at ~70 ns. Finally, a brief isolated enhancement of $C(5,6)$ can be recognized in Fig. 7(a) in the second precursory segment, at around 65 ns.

At around 112 ns, the $\alpha$ helix unfolds. Before this unfolding, $C(5,7)$ was briefly enhanced at ~100 ns (Fig. 7(b)) and thereafter suppressed right after 100 ns. In Fig. 7(a), we find also brief enhancements of $C(4,5)$ and $C(9,10)$ at around 110 ns. Followed later (>120 ns) is the revival of strong enhancement of $C(9,10)$ correlation (Fig. 7(a)). We therefore identify these transient correlational changes as precursors for the $\alpha$-helix unfolding process.

Based on these observations, the consistent picture that emerges is that the $\alpha$-helix folding starts at the head portion (enhancement of $C(2,4)$ at ~60 ns in Fig. 7(a)) and is followed by a folding at the tail portion (suppression of $C(9,10)$ at ~70 ns in Fig. 7(b)), before the peptide bending in the middle portion (suppression of $C(5,7)$ at ~75 ns in Fig. 7(b)) to form a stable $\alpha$ helix. Each of these persistent correlational changes is preceded by enhancement or suppression precursors. When the peptide unfolds, it unravels from the middle portion (enhancement of $C(5,7)$ at ~110 ns) which one can easily confirm from Fig. 7(b). Again, this correlational change is preceded by precursors.

The correlation analysis above shows that pairing of correlation changes between 3-residues can be realized as coordinated dihedral movements and could yield insight of precursors. More complete picture of folding/unfolding dynamics can be framed by examining the atomic displacements behind the correlational changes. The most important microscopic dynamics that can be drawn is the canonical precursors of folding. As shown in Figs. 7(a) and 7(b) around ~60-70 ns, we observe non-bonding interactions between $T_3$ and $A_5$ and between $A_1$ and $L_7$; the former gives rise to the enhancement of $C(2,4)$ whereas the latter leads to the brief enhancement of $C(5,6)$ seen in Fig. 7(a). Also they both produce transient weakening in $C(2,3)$, $C(3,4)$, $C(6,7)$, and $C(5,7)$ (Fig. 7(b)) as these residues recede from their respective neighbors. We classify these non-contacts as the *first kind of weak correlations*. One notices moreover that soon after the peptide folds into the $\alpha$ helix, the pairs, $T_{11}$ and $T_2$, $T_{11}$ and $A_5$, approach each other as the $\alpha$ helix becomes more compact. These compact movements in which $A_5$ is close to $T_3$ throughout the $\alpha$-helix phase explain why the stronger-than-average correlation $C(2,4)$ is persistently enhanced. Interactions between these residues lead also to brief suppressions of $C(6,8)$ and $C(9,10)$ which one can read from Fig. 7(b). In particular, we see that the stronger-than-average correlation $C(9,10)$ is strongly suppressed right after the $\alpha$ helix is formed. Note furthermore that the $T_{11}$ temporarily approaching $A_5$ before moving towards $T_2$ is the cause for this suppression. When the approach is complete, $T_{11}$ leaves room for $A_5$ and its interaction with $S_9$ begins, so that the most compact $\alpha$ helix is formed and, as a result, the close-packed coiling is enhanced. One conspicuous feature in this more compact conformation is that $I_4$, $L_7$, and $L_8$ are exposed to the solvent, which explains the weakening of $C(5,7)$ and $C(6,8)$ found in Fig. 7(b). Finally, the tail portion unfolds due to the dihedral rotation between $L_7$ and $L_8$ (Fig. 7(c) at 128 ns) and, as a consequence, correlations $C(6,7)$ and $C(7,8)$ fall below average (Fig. 7(b)) and $C(6,8)$ and $C(9,10)$ rise above average (Fig. 7(a)). As the tail unfolds, the coils in the head portion (residues up to $A_5$) stay intact. Finally, the head portion uncoils.

The atomic displacements after the $\alpha$ helix unfolds at 112 ns are also of interest. Recalling from Fig. 7(b) that $C(3,4)$ and $C(5,7)$ begin to disappear at ~120 ns and, beyond this time, we see from Fig. 7(a) that C(9,10) and C(6,8), which are stronger-than-average before the $\alpha$ helix phase but fall to background level in the $\alpha$ helix phase, become stronger-than-average once again. Between 112 and 265 ns, the peptide maintains a linear form, before folding into a $\beta$ hairpin. The suppression of $C(6,7)$ and many other weak correlations (see Fig. 7(b)) mostly occur outside the $\alpha$-helix folding region. This kind of twisting effect is referred to as the *second kind of weak correlations* (to be discussed next in Sec. III E 2).

### 2. $\beta$-hairpin conformation

Between the terminal segment boundaries at 265 and 346 ns (Fig. 4(a)), the peptide folds into a U-shaped $\beta$ hairpin. The molecule turns around with $A_6$, $L_7$, and $L_8$ constituting the pivotal residues (Fig. 8(c)). Overall this conformation is characterized by persistent enhancements of $C(9,10)$ (Fig. 8(a)) and $C(10,11)$ (Fig. 8(b)) and a slight persistent enhancement of $C(4,5)$, $C(6,7)$ and $C(6,8)$ (Fig. 8(a)) as well as a slight persistent suppression of $C(2,4)$ (Fig. 8(b)). These are *fingerprints* that we identify as the $\beta$-hairpin phase.

In contrast to the $\alpha$-helix phase between 73 and 112 ns (Fig. 7), there are three statistically distinct time series segments within the $\beta$-hairpin phase conformation. We name these as zone I (265~290 ns), zone II (290~310 ns), and zone III (310~346 ns). In zone I, $C(6,7)$ and $C(9,10)$ are stronger-than-average correlations (Fig. 8(a)), while $C(2,4)$ is

weaker than average (Fig. 8(b)). In zone II, the main stronger-than-average correlations are $C(4,5)$, and $C(6,8)$ instead of $C(6,7)$. Finally, the stronger-than-average correlations $C(4,5)$ and $C(6,8)$ that occur in zone II both disappear in zone III, whereas $C(6,7)$ in this same zone (Fig. 8(a)) is enhanced once again at ∼320 ns. Ultimately, there is a brief strong suppression of $C(10,11)$ at the end of zone III (Fig. 8(b)), leading to an overall destabilization of the $\beta$ hairpin. The peptide completely unfolds at ∼350 ns. Immediately after 350 ns, soon after the peptide unfolds, notice that $C(2,4)$ becomes enhanced (Fig. 8(a)), implying that the head would tend to nucleate into a $\alpha$ coil. However, due to the stronger-than-average correlations of $C(1,2)$ and $C(9,10)$ around the same time (Fig. 8(a)) the entire molecule is prevented from further folding into an $\alpha$ helix.

As in the study of $\alpha$-helix, we turn next to look for brief correlational changes before the peptide starts folding at ∼265 ns. Again two precursory segments are detected, i.e., one during ∼235 to ∼250 ns, and another during ∼250 to 265 ns. Just before the start of the first precursory segment, we see in Fig. 8(b) a brief enhancement of $C(10,11)$ which changes to a brief suppression upon entering into the first precursory segment till the end of this segment. Within the first precursory segment, there occur also in Fig. 8(a) brief enhancements of $C(5,7)$ and $C(9,10)$ and, in addition, we find brief suppressions in $C(6,8)$ and $C(9,11)$ (Fig. 8(b)). At the start of the second precursory segment, we see the correlation $C(10,11)$ in Fig. 8(b) briefly enhances, and it is followed by a persistent suppression till the end of the segment. Thereafter within ∼250 to 265 ns are observed brief enhancements of $C(1,2)$, $C(6,8)$ and $C(9,10)$ in Fig. 8(a), and brief suppressions of $C(3,4)$ and $C(6,7)$ in Fig. 8(b). The most prominent transient correlational changes are perhaps strong oscillations in $C(9,10)$ (Fig. 8(a)) and $C(10,11)$ (Fig. 8(b)) before the peptide unfolds at 346 ns.

A summary of the above correlational changes can now be made about the folding/unfolding of the $\beta$ hairpin. In the initial stage, the head of the peptide remained loose as it starts to turn around with respect to the more rigid tail that is already linear. After the basic U-shape of the $\beta$ hairpin is formed, subtle re-organizations in the peptide topology proceeded to yield a more stable $\beta$ hairpin. In the unfolding stage the two arms of the $\beta$ hairpin did not suddenly open up, but instead gradually slide past each other until the tail and head lose contact. The peptide thereafter stretches out into a random coil. Overall, we would say that the $\beta$ hairpin is a more stable conformation for the *TTR*(105-115) peptide, even though there are more intermediate steps in the folding and unfolding processes.

Delving again into the atomic displacements, we find that at approximately 250 ns the contacts of $I_4$ with $L_7$ and $T_2$ lead to the formation of the $\beta$-hairpin with a turning point which is seen in Fig. 8(b) as brief suppression of $C(3,4)$. At the same time, we find also the twisting of tail part which signals that it approaches the head part. This picture of the tail-induced-head motion is recognized by the weak suppression of $C(10,11)$ starting to enhance momentarily (Fig. 8(b)). After these precursory events at approximately 273 ns, close packing of the tail (where $T_{11}$, $S_{12}$ and $N_{13}$ are close to $T_3$, $A_1$, and $T_2$, respectively) results in $C(9,10)$ and $C(6,7)$ going from slightly above background level before 273 ns to strongly above background level at around 273 ns (Fig. 8(a)), whereas $C(10,11)$ going from background level before 273 ns to slightly above background level at approximately 273 ns (Fig. 8(a)). Thereafter, the head part rearranges to achieve the most compact $\beta$ hairpin. The suppression of $C(2,4)$ seen at this moment in Fig. 8(b) is due to the asynchronous twisting between $T_3$, $I_4$, and $A_5$ and we may classify this scenario as the *second kind of weak correlation*. After $I_4$ rotates, it lies within a tight formation between $T_3$, $L_7$, $S_9$, and $P_{10}$. The consequence is a strong enhancement of $C(4,5)$ and the switching of stronger-than-average correlations in C(6,7) and C(6,8) from $C(6,7)$ to $C(6,8)$ (Fig. 8(a)). Although we now have in zone II the most compact $\beta$ hairpin, the dangling $T_2$ continues to prevent the enhancement of $C(2,4)$ for ∼10 ns after entering zone II (Fig. 8(b)). This free dangling by the larger residue (see $T_2$ in Fig. 1(a)) gives rise to the *third kind of weak correlations*.

To gain deeper insights into the atomic displacements, we proceed to look more closely at $C(i,j)$ in zones I-III. In zone I of the $\beta$-hairpin phase, the tail portion of the peptide forms a stable linear chain (snapshot at 273 ns, Fig. 8(c)) which is consistent with the stronger-than-average $C(9,10)$, whereas the shape of its head portion continues to fluctuate, as reflected by the weaker-than-average correlation $C(2,4)$. To explain these correlational changes associated with the $\beta$-hairpin folding, we refer to the snapshot at 273 ns and note that residues $A_1$ and $T_2$ are near $N_{13}$, while $T_3$ is near $T_{11}$. As a result, when $T_2$ drifts towards $I_4$ the latter starts to steer $T_3$ away so that enough room is spared for the two arms of the $\beta$ hairpin to approach. Because of this dynamical molecular rearrangement, the $T_3$ (the middle residue in partial structure $T_2$-$T_3$-$I_4$) and $A_5$ (the middle residue in partial structure $I_4$-$A_5$-$A_6$) are no longer in proximity of each other, and hence we find that $C(2,4)$ is suppressed (Fig. 8(b)). On the other hand, we see the positions of $A_6$, $L_7$, $La_8$, $S_9$ and $P_{10}$ stabilizing the tail end of the peptide, leading therefore to stronger enhancements of $C(6,7)$ and $C(9,10)$ which one can easily glean from Fig. 8(a).

Moving on to zone II at 300 ns, we observe from Fig. 8(c) that the residue $I_4$ comes to a new position close to $S_9$ and $L_7$ soon after the concerted rotation of $T_3$, $I_4$ and $A_5$. The turn of the U-shape now entirely "pivots" on $L_7$, instead of between residues $A_6$ and $L_7$ in zone I (snapshot at 273 ns, Fig. 8(c)). The consequence is that the two arms of the $\beta$ hairpin are drawn even closer and this explains the enhancements of stronger-than-average of correlations $C(4,5)$ and $C(6,8)$ which are the result of non-bonding interactions between $I_4$-$A_5$-$A_6$ and $A_5$-$A_6$-$L_7$ and between $A_6$-$L_7$-$L_8$ and $L_8$-$S_9$-$P_{10}$, respectively. Furthermore, in crossing from zone I to II, $T_2$ changes from a location originally near $N_{13}$ (snapshot at 273 ns) to now turning freely towards $I_4$ (dashed arrow in snapshot at 300 ns). Such a dynamical movement explains why $C(2,4)$, which is weaker-than-average in zone I, restores to background levels in zone II. The relative motion between the two arms of the $\beta$ hairpin is thus the cause of the enhancement of $C(10,11)$ in Fig. 8(a). Together with the persistently enhanced $C(9,10)$ seen in Fig. 8(a), we may now realize that the more rigid tail portion is connected to these concerted motions. Finally, one can understand correlational enhancements

between $S_9$, $P_{10}$, $T_{11}$, $S_{12}$ and $N_{13}$ from the spatial proximity of three sets of residues, namely, both $S_{12}$ and $T_{11}$ to $A_1$, $P_{10}$ to $T_3$ as well as $S_9$ to $I_4$. These enhanced correlations drive the tail and head closer (apart from $T_2$ at the head end which is dangling outward).

Going from zone II to zone III, the two arms of the $\beta$-hairpin slide once more, with the U-shape turn that was positioned within $A_6$-$L_7$-$L_8$ in zone II changing now to one residing on $I_4$. The result is the arm on the head end of the $\beta$ hairpin becoming shorter than that on the tail end. This difference in arm lengths has much influenced on residues beyond $T_{11}$ for they are neither in contact with the head portion nor as constrained as they were in zone II. Clear evidence of this sliding of arms can be found from $C(9,10)$ and $C(10,11)$ showing strong oscillations. The cause for these robust oscillations can thus be traced to $A_1$ and $T_2$ both sliding back and forth between $T_{11}$, $S_{12}$, and $N_{13}$. From the sequence of snapshots retrieved from the MD trajectory, we reconfirmed unambiguously that the head and tail residues are indeed decoupled and that the less rigid correlation of residues is associated with the longer tail arm.

In recapitulation, we may stress that, over the entire folding region (zone I-III), $C(9,10)$ and $C(10,11)$ in Figs. 8(a) and 8(b) display strong repetition of appearance and disappearance. This is the result of a sliding motion between the head and the tail, with $A_1$ going back and forth between $T_{11}$, $S_{12}$, and $N_{13}$. This kind of repeated picture is strongest in zone III due to the rotation of $I_4$ (marked by the disappearance of the strong $C(4,5)$, see zone III in Fig. 8(a)) allowing the tail part to slide over the head part. An immediate consequence is that the $\beta$ hairpin loses stability and starts to unfold. At 350 ns during the unfolding, we observed weak correlations $C(6,7)$ and $C(7,8)$ (Fig. 8(b)), and strong correlation $C(2,4)$ (Fig. 8(a)) because of the twisting action at residues $L_7$ and $L_8$, and the head coiling is accordingly back up.

### 3. Mixture of $\alpha$ and $\beta$ conformations

Between 400 and 930 ns (Fig. 4(a)), the head-tail shape similarity index remains low and oscillatory, without any noticeable plateau. We analyze the color map of correlations in four time series intervals, i.e., zone I (538∼553 ns), zone II (558∼592 ns), zone III (593∼721 ns), and zone IV (891∼917 ns). These four zones are chosen for diagnosis based on their lower $\zeta(t)$ values. Drawing on insights gained from preceding subsections, we now know that strong persistent correlation of $C(2,4)$ (see Fig. 7(a)) is a signature for coil topology in the head portion of the peptide. Such strong enhancement of $C(2,4)$ is in fact seen in Fig. 9(a) within zones I, III, and IV, and also near the end of zone II. On the other hand, the suppressions of $C(5,7)$ and $C(6,7)$ in zones III and IV (Fig. 9(b)) suggest competition to stabilize ($C(5,7)$) and destabilize ($C(6,7)$) the coil topology, due to the persistent repetition of suppression and enhancement in zone III and the persistent suppression throughout in zone IV in the former, and persistent suppressions in zones III and IV in the latter. Indeed, the sequence of snapshots (Fig. 9(c)) shows topologies which are intermediate between $\alpha$ helix and $\beta$ hairpin, with a typical picture displaying an $\alpha$-helix head and a $\beta$-hairpin tail (see, for example, the snapshot at 615 ns in zone III (Fig. 9(c))). Specifically, we find in zone III persistently stronger-than-average $C(1,2)$ and brief enhancements of $C(2,3)$, $C(4,5)$, $C(5,6)$ and $C(2,4)$. We had noted above that the persistently stronger-than-average $C(9,10)$ along with slight persistently stronger-than-average of $C(4,5)$, $C(6,7)$, and $C(6,8)$ are the fingerprints of a complete $\beta$ hairpin (Fig. 8(a)), but in Fig. 9(a) we rarely see these enhanced correlations shown up at the same time. Similarly, the simultaneous appearance of persistently enhanced $C(2,4)$ (Fig. 7(a)) and persistently suppressed $C(5,7)$ (Fig. 7(b)) are fingerprints of a complete $\alpha$ helix, but from Figs. 9(a) and 9(b) it is clear that they rarely occur together. Instead, Fig. 9(a) shows intermittent enhancements of $C(2,4)$ and $C(9,10)$, suggesting that the shape of the peptide is unstable during this period of mixed $\alpha/\beta$ phase. This mixed morphology is seen also in zone IV. In fact, the occurrence of the $\alpha$-helix head in zone III and its revival in zone IV (see Fig. 9(a)) indicate how rapid the folding and unfolding processes are. The sequence of snapshots in Fig. 9(c) clearly reveal this dynamical aspect of the *TTR*(105-115) peptide.

The analysis of the atomic displacements in this time interval is somewhat intricate. Here we find in the color map not only $\alpha$-helix signatures (stronger-than-average $C(2,4)$ and weaker-than-average $C(5,7)$) but also $\beta$-hairpin characteristics (repetition of appearance and disappearance of the stronger-than-average $C(9,10)$ and the weaker-than-average $C(10,11)$) spanned the four zones (Figs. 9(a) and 9(b)) with lower head-tail $\zeta(t)$ (Fig. 3(b)). Despite mixed $\alpha$ and $\beta$ features, the representative snapshot in zone III (Fig. 9(c) at 615 ns) appears similar to the $\alpha$ helix at 80 ns (Fig. 7(c)), suggesting that two conformations can be distinguishable not just from visualizing the structures alone. From the sequence of snapshots retrieved from the MD trajectory, we see that the head coil first formed as the peptide enters zone I of the mixed conformations and, upon entering zone II, proceeded to develop into a more complete $\alpha$ helix, but with a shorter head and a longer tail. Between 500 and 900 ns, one finds in fact that during the period in zone II and time windows at approximately ∼700-750 and ∼775-825 ns (Fig. 9(b)), $C(1,2)$ shows not only robust repetition of appearance and disappearance but also substantially suppressed (Figs. 9(a) and 9(b)). This latter panorama can be understood as due to the synchronous motion between $A_1$-$T_2$-$T_3$ and $T_2$-$T_3$-$I_4$ being destroyed by the residue $T_2$ interacting with the residue $T_3$ (see Fig. 9(c) at 574 ns). Because $T_2$ and $T_3$ are common residues in partial structures $A_1$-$T_2$-$T_3$ and $T_2$-$T_3$-$I_4$, we have a nonbonding interaction which can be categorized as a *fourth kind of weak correlation*. One should note, however, an interesting physical relevance of this extremely suppressed correlations of $C(1,2)$ (Fig. 9(b)) which is that when $T_2$ is in contact with $T_3$, the head coil becomes more compact and has a helical pitch that is shorter than that in the stable $\alpha$ helix formation found at 80 ns (Fig. 7(c)). This shorter head to which is joined a relatively longer tail nevertheless has some characteristics of a $\beta$ hairpin, which one can realize only after comparing the correlation color maps in Fig. 9. Otherwise, by merely scrutinizing the snapshots alone, we would have thought and in fact misled to consider the

structure of the peptide in zone III is just another type of $\alpha$ helix which, however, is not true.

Let us digress and delve further into its physical significance. As shown in Fig. 7(c), the peptide snapshot at 80 ns is within the $\alpha$-helix phase and the peptide snapshot here at 615 ns is within the mixed $\alpha/\beta$ phase. These two structures are aligned, so that $A_1$ and $N_{13}$ are directed upwards and they differ only in their respective turning point being half a residue apart. Judging just by the snapshots alone, we would never have thought that the $\alpha$-helix structure at 615 ns is less stable and participates in rapid interchangeability with the $\beta$-hairpin structure. In fact if one were to compare the $C(i,j)$ color map at 615 ns with that at 80 ns, one will see easily that the correlations (in particular the $\alpha$-helix fingerprints) are very similar. They, however, differ in $C(1,2)$ where it enhanced at 615 ns (Fig. 9(a)) but not so at 80 ns (Fig. 7(a)) and in $C(5,7)$ where it suppressed at 80 ns (Fig. 7(b)) but not so at 615 ns (Fig. 9(b)). Since the instant 615 ns is in the middle of zones I-III (spanning 538-721 ns) of mixed phases, it is instructive to compare also the correlational color map at 615 ns with that at 300 ns, which falls deep within the $\beta$-hairpin phase, even though the structure at 615 ns looks very different from a $\beta$ hairpin. Interestingly, we find in Fig. 9(a) at 615-ns enhanced correlations $C(9,10)$ and $C(10,11)$, which are fingerprints of $\beta$-hairpin (Fig. 8(a)). Without the aid of the correlation color map, we would never have arrived at this conclusion that a half-residue shift in the morphology of the $\alpha$ helix is associated with $\beta$-hairpin fingerprints, i.e., tipping of the peptide in favor of the $\beta$ hairpin. Thus, because of its mixed characteristics, the $\alpha$ helix with a shortened head is not stable, and the peptide switches rapidly between $\alpha$-helix and $\beta$-hairpin structures.

We mention in passing about a closer examination of the snapshots in zone III (Fig. 9(c)). We find that the strong suppression of $C(5,7)$ is due to bending in the middle of the $\alpha$ helix, whereas the suppression of $C(9,10)$ is the result of contact between the head and the tail.

#### 4. Summary of fingerprints and precursors

In Table I, we summarize the correlational fingerprints and folding/unfolding precursors of the $\alpha$-helix and $\beta$-hairpin conformations of the *TTR*(105-115). In this table, "+" means enhancement, while "−" denotes suppression. Fingerprints are persistent changes to the cross correlations, whereas precursors are transient changes to the cross correlations. From Table I, we see that the fingerprints allow us to easily discriminate between the $\alpha$-helix and $\beta$-hairpin conformations. The folding and unfolding precursors for the two conformations are also different, in the sense that the folding precursors for the $\beta$ hairpin are not the unfolding precursors of the $\alpha$ helix, and the folding precursors for the $\alpha$ helix are not the unfolding precursors of the $\beta$ hairpin. This would mean that the $\alpha$-helix and $\beta$-hairpin conformations do not compete directly with each other for stability, but each has its own distinct mechanistic folding pathways. Parenthetically, we should perhaps emphasize that, for although we observed persistent changes in correlations within a stable conformation whereby arriving at the fact that the character of high-frequency fluctuations is determined by the low-dimensional manifolds and hence fingerprints and precursors of folding/unfolding, we cannot, however, say conclusively that the peptide studied confirms the Haken's slaving principle since only one folding/unfolding event is analyzed in the present simulation work. It would be on a more general and solid ground if our method can be tested further on a few well-behaved peptides such as, for instance, Trpzip2,[86,87] a peptide known to exhibit many times the $\alpha$ folding. This, however, is outside the scope of the present study.

## IV. ATOMIC-RESOLUTION AND CLASSIFICATION OF PRECURSORS

With just the correlational fingerprints and precursors, we cannot obtain a complete picture of folding/unfolding dynamics. In this section, we will explain how the correlational precursors can guide us discovered the atomic-resolution of folding mechanisms from snapshots of the MD simulation. It is by now fair to say that without the correlational precursors, we will not know *a priori* what to look out for, where to look for them, and when to look for them. Because correlational precursors are between localized 3-residues partial structures, we therefore know where to look for them. These precursors are also localized in time before the folding and unfolding events, so we know when to look for them. Finally, the signs of the correlational changes suggest what to look out for.

In Sec. III E, we presented simultaneously our scrutinization of the precursory correlational changes and the identification of the underlying atomic displacements. Based on insights gained there (see Secs. III E 1–III E 3), we proceed furthermore to classify systematically in Sec. IV A two classes of correlation enhancements and, in Sec. IV B, four classes of correlation suppressions that make up these

TABLE I. Correlational fingerprints and folding/unfolding precursors of the $\alpha$-helix and $\beta$-hairpin conformations of a fragment of the protein transthyretin *TTR*(105-115). Plus (minus) sign indicates enhanced (suppressed) correlation irrespective of the strong than average or the weaker than average in color chart.

| | $C(1,2)$ | $C(2,4)$ | $C(3,4)$ | $C(4,5)$ | $C(5,7)$ | $C(6,7)$ | $C(6,8)$ | $C(7,8)$ | $C(9,10)$ | $C(9,11)$ | $C(10,11)$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\alpha$-helix fingerprints | − | + | + | | − | + | − | | − | | |
| $\alpha$-helix folding | | + | − | | − | − | + | | + | | |
| $\alpha$-helix unfolding | | | | | − | − | | | | | |
| $\beta$-hairpin fingerprints | | − | | + | | + | + | | + | | + |
| $\beta$-hairpin folding | + | | − | | | − | + | − | + | + | − |
| $\beta$-hairpin unfolding | | | | | | + | | | +/− | | +/− |

precursory changes. Then, in Sec. IV C, we make a comparison between the present time series analysis and the widely used contact analysis method.

### A. Stronger-than-average precursors

In Secs. III E 1–III E 3 above, we briefly discussed the four precursors associated with weaker-than-average correlations. Although we did not refer to them as such, we also discovered two precursors which are associated with stronger-than-average correlations.

In the context of the USR technique and time series correlation analysis of the folding and unfolding processes, the stronger-than-average correlation between two 3-residues partial structures implies that the time series similarity indices of the two partial structures are strongly synchronized in their temporal variations. This can only happen when the two 3-residues partial structures are part of the same slow manifold, for example, an $\alpha$ helix or a $\beta$ hairpin, dictating how alike they are in their high-frequency fluctuations. Therefore, prior to folding into either of these two stable conformations, we expect enhancements of correlations between the soon-to-be synchronized parts of the peptide. Once enhanced, these strong correlations will persist while the peptide remains in the stable conformation.

To understand the folding/unfolding processes that have led into such stable conformations, we look at transient enhancements of correlations. We identify two kinds of enhancement precursors. The *strong correlations of first kind* is associated with the compactification of the stable conformations. For the $\alpha$ helix, this compactification precursor occurs in the enhancement of $C(2,4)$ seen in Fig. 7(a), and always it occurs within the $\alpha$-helix phase (Fig. 7(c) at 91 ns). It appears also to be the precursor for unfolding. For the $\beta$ hairpin, the compactification precursor is found both within and outside of the $\beta$-hairpin phase. Within the $\beta$-hairpin phase, the compactification precursor is the enhancements of $C(6,7)$ and $C(9,10)$ occurred in Fig. 8(a) (Fig. 8(c) at 300 ns), whereas outside the $\beta$-hairpin phase, the compactification precursor is the enhancement of $C(1,2)$ seen in Fig. 8(a) also (Fig. 8(c) at 350 ns). The *strong correlations of second kind* is associated with the global motion of a large part of the peptide. For both the $\alpha$ helix and $\beta$ hairpin, this global motion precursor appears only outside their respective phase which is $C(9,10)$ (see Fig. 7(a)) for the former (Fig. 7(c) at 70 ns) and $C(1,2)$ and $C(9,10)$ (see Fig. 8(a)) for the latter (Fig. 8(c) at 260 and 350 ns, respectively).

### B. Weaker-than-average precursors

Intuitively it is quite natural not to expect that precursors associated with weaker-than-average correlations are important. After all, a weaker-than-average correlation between partial structures $i$ and $j$ suggests that the two partial structures are separately interacting with different partial structures, say $k$ ($\neq i,j$), and their correlations naturally fall below average. If, however, we think of folding mechanisms in lock-and-key terms, then as the lock and key become synchronized, there must be a corresponding drop in the correlation between the lock and its supporting structures. From this point of view, the weaker-than-average correlations may be just as important as stronger-than-average correlations. In atomic terms, we find that a weaker-than-average correlation which manifests asynchronous motion between two partial structures, may be caused by the individual residues experiencing (a) diverse influences from the other residues (Fig. 7(c) at 70 and 91 ns), (b) twisting motion (Fig. 7(c) at 70 ns and Fig. 8(c) at 260 and 350 ns), (c) dangling movement (Fig. 8(c) at 300 ns), and (d) self-interaction(s) within the partial structure(s) (Fig. 9(c) at 574 ns). We remark that such weaker-than-average correlations are, however, dependent on partial structures chosen for correlational analysis as discussed above in Secs. III C and III D.

### C. Comparison of cross correlation analysis and contact analysis

To appreciate further the cross correlation analysis, it is appropriate at this point to compare it with the widely used contact analysis.[88–92] First of all, we note that a full contact analysis is a more restricted cross correlation analysis, and that one normally picks a much smaller set of contacts to track during the simulation. This selection somehow is *ad hoc*, and not being guided by the simulation data. Our cross correlation analysis is entirely guided by simulation data, and more importantly, it looks at cross correlational changes (Eqs. (10) and (11)) beyond close contacts (see also the comment in Fig. 3(a) in Sec. II B).

Before commenting on the use of the contact analysis, we note first of all that the contact analysis is based on the idea of computing the smallest contact distance between residues and therefore will be less obvious for studying those residues that are not in contact. Consider again the end-to-end $r_{\text{end-to-end}}$ for representative atoms in the head-tail residues shown in Fig. 3(a). According to the widely used contact analysis, one would construct the so-called 2D contact map for analyzing its structure. For concreteness, let us return to the color map in Fig. 7 focusing on the stable conformation into which $\alpha$-helix topology falls into, i.e., between 73 and 112 ns (red vertical lines). As described there, this time duration delimits the segment boundaries which we obtained by applying the segmentation method to the head-tail time series $\zeta(t)$. These segment boundaries are, of course, different from those obtained by performing the time series segmentation on Fig. 3(a). For simplicity and convenience in the following discussion, let us *assume* that the end-to-end time series $r_{\text{end-to-end}}$ dictates equally well the distance-to-distance $r_{\text{distance-to-distance}}$ for all residues as in the peptide.

Figure 10 is the contact map which is constructed following the procedure described in Refs. 88–92. It can be seen that ($\text{Ala}_5$, $\text{Ala}_6$, $\text{Leu}_7$) is in proximity to ($\text{Ace}_1$,$\text{Tyr}_2$,$\text{Thr}_3$) and ($\text{Tyr}_{11}$,$\text{Ser}_{12}$,$\text{Nac}_{13}$) (red rectangle or square) due to their pivotal locations at the center of the turn-point. The same conclusion can be drawn to ($\text{Ace}_1$,$\text{Tyr}_2$,$\text{Thr}_3$) and ($\text{Tyr}_{11}$,$\text{Ser}_{12}$,$\text{Nac}_{13}$). The contact map thus shows that they are nearby. Note, however, that this contact map was obtained after taking the time average of the time elapsed, and thus does not therefore contain the underlying dynamics between

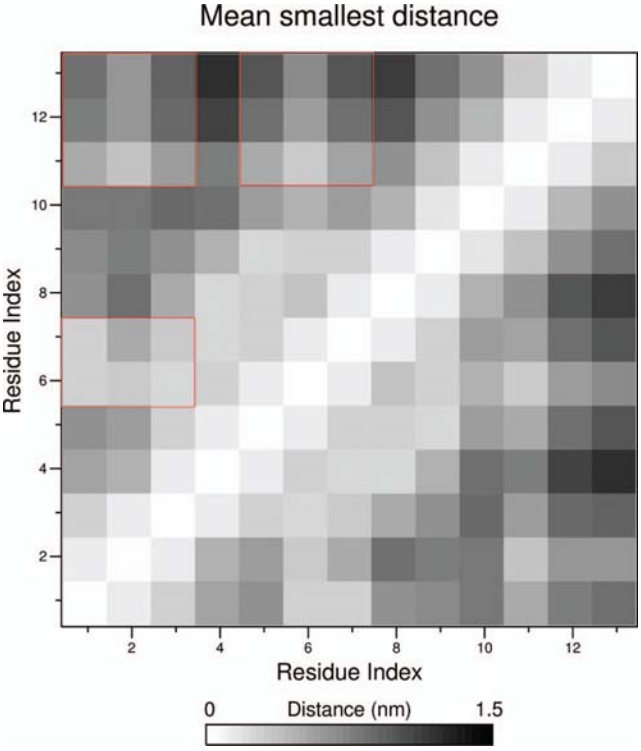


FIG. 10. Color map within stable conformation of $\alpha$-helix topology between 73 and 112 ns constructed by the contact analysis method. The matrix element $(i,j)$ refers to residue index of residues $i$ and $j$ and its value, the mean smallest distance, is calculated by the time average of distance between the atom in $i$th residue closest to the atom in $j$th residue at each time over the total time elapsed. The $(i,j)$ value is presented in grey color according to the color panel showing the distance (0-1.5 nm). Note that the diagonal matrix element $(i,i)$ has the value zero and is colored white. Two residues that are far apart therefore have $(i,j)$ darker in grey color. We have stipulated a maximum truncated distance of 1.5 nm for best contrast in reading. The above results are based end-to-end distance of Fig. 3(a).

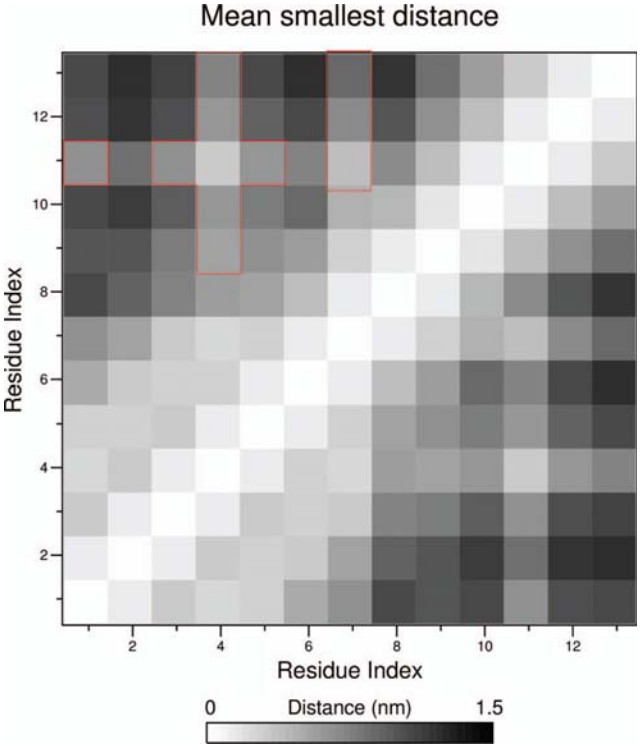


FIG. 11. Same as Fig. 10 except for the time duration 552-593 ns.

residues. To recover the dynamics, one must refrain from taking the time averaging but carrying out instead a time series analysis as we did for the 3-residues $\zeta(t)$. In principle, both would provide information on the dynamics of the peptide. The two methods, contact analysis and cross correlation analysis, do differ in the way the 2D matrix obtained. For the former, it is rather difficult to resolve $r_{\text{distance-to-distance}}$ for nearest neighbor residues and hence the elements in the contact map, whereas for the latter, because we work on the correlation of *sign changes* $\Delta p(t)$ (Eq. (9)), it goes beyond close contacts. A representative case is the weak correlation of fourth kind (see the snapshot at 574 ns in Fig. 9(c) showing self-contact effect of residues $\text{Tyr}_2$ and $\text{Thr}_3$). The contact map in the time interval 552-593 ns (Fig. 11) lends further support to our observation.

For although the contact map method can be used to yield information of static positions of residues or reconstructed for 3D static structure in proteins,[91,92] the method is, however, inadequate and in fact rather difficult to use in cases of analyzing the local dynamics, especially for residues that are nearest neighbors whose intramolecular interactions are presumably more important. The same picture happens in $\beta$ hairpin and $\alpha/\beta$ mixed phases. Contact analysis cannot give the kinds of information such as stronger-than-average $C(4,5)$ and $C(6,8)$ in the former and $C(1,2)$ in the latter. We should mention moreover that all our weaker-than-average correlations describe nearest and next nearest neighbor correlations for pairs of 3-residues because the 3-residues that are farther away has been filtered out (Sec. III D). Such near-distance interactions cannot therefore be gleaned easily by the strategy of contact analysis.

## V. CONCLUSIONS

In this work, we have illustrated based on data recorded from an equilibrium MD simulation of a fragment of the protein transthyretin *TTR*(105-115) how a combination of USR technique, time series segmentation method, and time series correlation analysis can be used to identify precursors of folding and unfolding. We first showed in this model study how the USR technique was applied to calculate partial structures, separating local structural changes from the global motion of the molecule. An efficient 16 statistical moment descriptors were employed to construct with respect to a same reference partial structure the time series of shape similarity index $\zeta(t)$. Specifically, we focused on how the statistical segmentation was applied to the head-tail similarity time series $\zeta(t)$ to accurately determine when the molecule folded into stable conformations of $\alpha$-helix and $\beta$-hairpin. Then, by the sliding time windows technique, we compute the cross correlations $C(i,j)$ between the sign change of the change in $\zeta(t)$ for the time series of 3-residues partial structures $i$ and $j$. We showed in particular how to extract the correlational fingerprints of the $\alpha$-helix and $\beta$-hairpin conformations within their respective

time segment which was obtained by the time series segmentation. We described also how precursors can be read from the correlation color maps, referring to the time series segments preceding the stable conformations.

As far as we can tell, this is the first time the USR technique, time series segmentation method, and time series correlation analysis have been combined and used to study the dynamics of a peptide, especially to determine the mechanistic pathways associated with folding and unfolding. One main contribution in this paper is to extend using USR for structural screening to diagnose the dynamics of molecule, reducing our reliance on snapshots of the molecular motion. To gain insights into the mechanisms, we employed the time series methods such as segmentation and correlation to delimit the time intervals for a detailed dissection. We showed, in particular, that fingerprints which are reminiscent of the Haken's slaving principle (high-frequency fluctuations entrained by the low-dimensional manifolds) can be identified in the correlations. More importantly, the combination of time series segmentation and time series correlation allowed us to unambiguously identify brief correlational changes preceding and succeeding the folding and unfolding events. This made it possible for us to identify the precursory traits in folding/unfolding conformations, and map out the coarse-grained mechanisms for these processes. Specifically, the precursors discovered in our present study suggest that polymer folding into a $\alpha$ helix starts with the head nucleated into a $\alpha$ turn, and unfolding is the result of a $\beta$ turn nucleating in the middle portion of the polymer. This $\beta$-turn nucleation may then lead to folding into a $\beta$ hairpin, which will unfold when the $\beta$ turn overcomes a steric barrier that prevents the two arms of the $\beta$ hairpin from sliding past each other. Unlike tracking the folding and unfolding processes by eye or by testing a set of prescribed contacts as in the widely used contact analysis,[88–92] our methodology applies equally well to large molecules for which little or nothing is known about the folding/unfolding mechanisms. The same methodology can be applied to analyze other regime-shift problems. In those cases, even though the correlations will be different, we should still be able to understand more deeply the underlying physics. It is hope that the methodology presented here not only is applicable to biological macromolecules, but also in other molecular-like systems such as nanotubes and molecular clusters.

Finally, it appears worthwhile to make two relevant remarks on the present statistical approach to study the dynamics of peptide. First, when we combine the correlation analysis supplemented by retrieving the snapshots from simulation data, we gain the most insights into the processes of folding and unfolding. Because the correlations are between localized 3-residues partial structures, these correlational precursors tell us where in the molecule and when in time to look for telltale mechanistic signatures. A major finding in our study is the important roles played by both strong and weak precursors in the mechanistic pathways. Two strong precursors and four weak precursors were discovered for our model *TTR*(105-115). For other peptides, these precursory numbers may be different. However, because these are folding precursors of the $\alpha$ helix and $\beta$ hairpin, which are basic building blocks making up the tertiary structure of proteins, we anticipate some level of universality. That is to say, if we change a peptide, the only change should be the distribution of strong and weak correlations; the interplay between strong and weak correlations in driving folding and unfolding is expected to be universal.

Second, we did not analyze trajectory segments that are not assigned to folded states in the present work, beyond looking for short-lived precursors, and definitely not in the context of intermediates. This issue is unquestionably an interesting theme that will in need of further studies and urge for line of inquiry, i.e., what is the physical meaning of these non-folded trajectory segments? From our correlation analysis, the unravelings of both the $\alpha$ and $\beta$ conformations were very rapid and complete. There is no evidence for a partial $\alpha$ helix after the latter phase conformation unfolded, although there are fingerprint correlations that periodically re-emerge after the unfolding. Our point of view is that the intermediate states, generally speaking, are related to unfolded states displaying, by and large, linear or semi-linear topology. We notice, in particular, that when the polymer enters or leaves folded states, its correlation is vastly different from those in folded or unfolded phase. That it happens is mainly because we used partial structures, digital correlation and sliding window to capture precisely the local dynamics and, from the segmentation boundaries, we are able to study the correlational changes before and after the stable conformations. It is likely that fingerprints and precursory of intermediate states occur before and after stable conformation. This topic is certainly worth investigating in its own right, but the study requires more careful analysis to see how much richer the story of protein folding mechanics can be by examining the non-folded trajectory segments before and after the folding and unfolding events. However, we believe, any new insights that emerge from such a study should justify a separated follow-up work for near future endeavor.

To summarize, we see that the present effort to combine time series correlation and atomic displacement analysis helps us discover many features that are difficult to obtain by using molecular snapshots alone. In particular, we realize the important roles played by the tyrosine residues $T_2$ and $T_{11}$, which have the longest side chains in the *TTR*(105-115) peptide. From the time series analysis, we found that folding/unfolding events are mostly triggered by the appearance of strong/weak correlations. Such observations can guide us to learn more of the folding/unfolding mechanisms from the complex MD simulation.

## ACKNOWLEDGMENTS

We thank the National Science Council for financial support (NSC101-2112-M-008-013-MY2). P.J.H. would like to thank the Taiwan International Graduate Program for a Ph.D. scholarship.

## APPENDIX A: 3-RESIDUES TIME SERIES SIMILARITY INDICES

In the time series correlation stage of the study, we work with 11 overlapping 3-residues partial structures in addition to the time series of $\zeta(t)$ for the head-tail partial structure to better understand the folding mechanism. Partial structures of

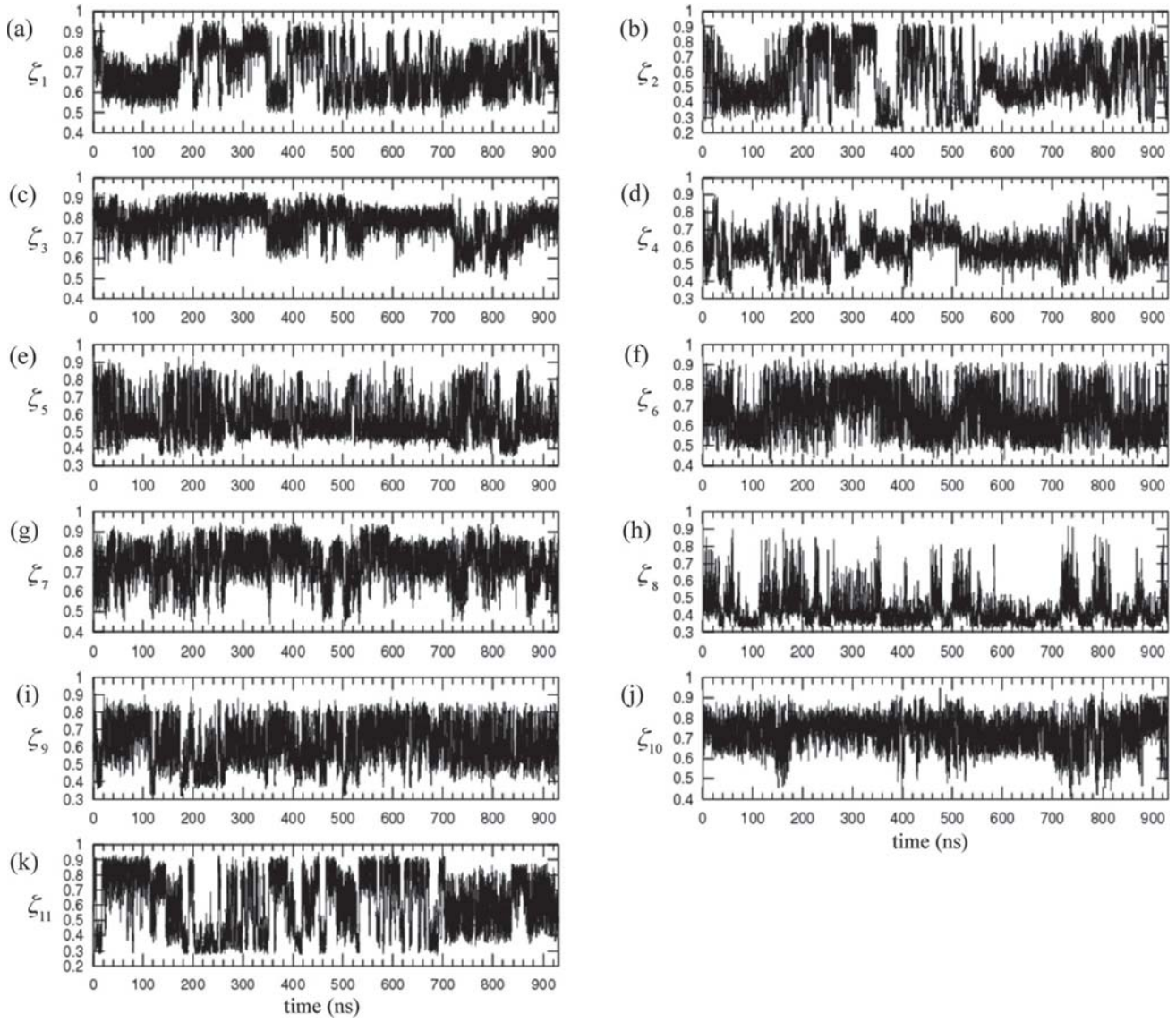


FIG. 12. The time series similarity indices of eleven 3-residues recorded with reference to their initial configurations. They are (a) $\zeta_1$ or $A_1$-$T_2$-$T_3$, (b) $\zeta_2$ or $T_2$-$T_3$-$I_4$, (c) $\zeta_3$ or $T_3$-$I_4$-$A_5$, (d) $\zeta_4$ or $I_4$-$A_5$-$A_6$, (e) $\zeta_5$ or $A_5$-$A_6$-$L_7$, (f) $\zeta_6$ or $A_6$-$L_7$-$L_8$, (g) $\zeta_7$ or $L_7$-$L_8$-$S_9$, (h) $\zeta_8$ or $L_8$-$S_9$-$P_{10}$, (i) $\zeta_9$ or $S_9$-$P_{10}$-$T_{11}$, (j) $\zeta_{10}$ or $P_{10}$-$T_{11}$-$S_{12}$, (k) $\zeta_{11}$ or $T_{11}$-$S_{12}$-$N_{13}$.

these 11 time series of $\zeta(t)$ computed according to Eq. (5) are displayed in Fig. 12. One sees readily that the 3-residues partial structures are more similar to their initial structures compared to the head-tail partial structure. Furthermore, we see that the 11 time series of $\zeta(t)$ roughly fall into three categories, associated with the head, middle, and tail of the molecule. These time series of $\zeta(t)$ are not segmented. Instead, they were analyzed in Sec. III C for the correlations between all pairs of

TABLE II. The 29 segment boundaries and their corresponding JSD values obtained by recursive segmentation.

| Time (ps) | $\Delta^*$ | Time (ps) | $\Delta^*$ | Time (ps) | $\Delta^*$ |
|---|---|---|---|---|---|
| 20423.5 | 50709.0435 | 264843.5 | 184614.2047 | 593115.0 | 17420.0404 |
| 47576.0 | 6758.1810 | 290593.0 | 9163.9965 | 722230.5 | 296080.5455 |
| 61284.0 | 22845.9471 | 310813.0 | 42809.8414 | 747670.0 | 14603.8986 |
| 73202.0 | 60220.5089 | 346318.0 | 148901.1384 | 817159.5 | 25253.5409 |
| 111526.0 | 121731.6304 | 416971.5.0 | 34018.0166 | 835879.0 | 18180.1141 |
| 132343.0 | 6702.0271 | 435042.0 | 9799.8174 | 858286.0 | 52371.3826 |
| 142021.5 | 29122.1572 | 519278.5.0 | 4107.7753 | 882099.5 | 29654.3702 |
| 170248.0 | 48583.2080 | 536259.0 | 29411.1070 | 891504.5 | 3086.2180 |
| 232197.0 | 25540.0758 | 552454.5 | 34559.9942 | 914769.5 | 44427.8081 |
| 249101.0 | 5653.0885 | 556065.5 | 65624.5964 | | |

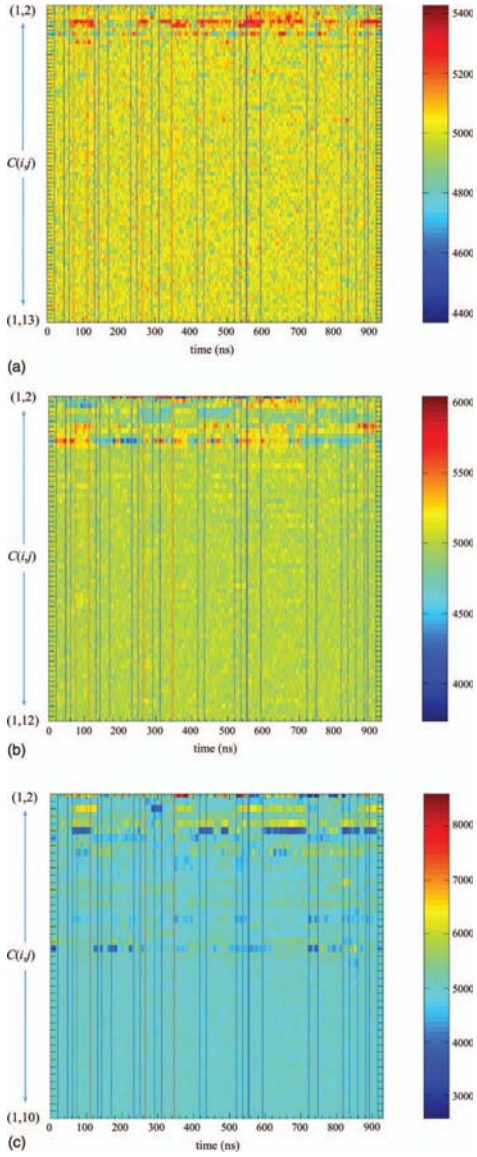


FIG. 13. The unfiltered color map of correlations over the 930 ns with sliding windows each of 5 ns for (a) 1-residue (top), (b) 2-residues (middle), and (c) 4-residues (bottom). Because the windows are slided in 1-ns steps, the window number is also the beginning of the window at time $t = 1$ ns for the first window, 2 ns for the second window, and so on. In these color maps, the correlations are arranged such that those between nearest-neighbor partial structures $C(i,i+1) \equiv C_{i,i+1}$ are shown first, followed by those between next-nearest-neighbor partial structures $C(i,i+2)$, and so on. The tick labels along $y$-axis, from top to bottom, of $C(i,j) \equiv (i,j)$ run as (1,2), (2,3), (3,4), (4,5), (5,6), (6,7), (7,8), (8,9), (9,10), (10,11), (11,12), (12,13),(1,3), (2,4), (3,5), (4,6), (5,7), (6,8), (7,9), (8,10), (9,11), (10,12), (11,13), (1,4), (2,5), (3,6), (4,7), (5,8), (6,9), (7,10), (8,11), (9,12), (1,5), (2,6), (3,7), (4,8), (5,9), (6,10), (7,11), (8,12), (1,6), (2,7), (3,8), (4,9), (5,10), (6,11), (7,12), (1,7), (2,8), (3,9), (4,10), (5,11), (6,12), (1,8), (2,9), (3,10), (4,11), (5,12), (1,9), (2,10), (3,11), (4,12), (1,10), (2,11), (3,12), (1,11), (2,12), (1,12) for (b) and (1,2), (2,3), (3,4), (4,5), (5,6), (6,7), (7,8), (8,9), (9,10), (1,3), (2,4), (3,5), (4,6), (5,7), (6,8), (7,9), (8,10), (1,4), (2,5), (3,6), (4,7), (5,8), (6,9), (7,10), (1,5), (2,6), (3,7), (4,8), (5,9), (6,10), (1,6), (2,7), (3,8), (4,9), (5,10), (1,7), (2,8), (3,9), (4,10), (1,8), (2,9), (3,10), (1,9), (2,10), (1,10) for (**c**).

3-residues partial structures, being guided by the time series segments of the head-tail partial structure for the analysis of precursors.

## APPENDIX B: TIME SERIES SEGMENTATION SCHEME

In this appendix, we present an alternative segmentation scheme. To begin with, we note that one can terminate the process of recursive segmentation when the JSD maxima of all new segment boundaries fall below a simple cutoff $\Delta_0$, say 200. This means that a JSD maximum of $\Delta^* = \Delta_0 = 200$ implies a 1% statistical difference per bit between the 1- and 2-segment models. This simpler procedure yields the strongest segment boundaries (i.e., those with the largest terminal $\Delta^*$) compared with those found by the three more rigorous approaches mentioned in Sec. II C. At each stage of the recursive segmentation process, we also perform segmentation optimization, ensuring the context sensitivity problem[85] has been overcome. Using this segmentation scheme, we obtained 29 terminal segment boundaries for the head-tail time series $\zeta(t)$ of $TTR$(105-115), and they are listed numerically in Table II. These terminal segment boundaries which are depicted in Fig. 4(a) will be used in this study.

## APPENDIX C: UNFILTERED CROSS CORRELATION COLOR MAPS OF 1-, 2- AND 4-RESIDUES

In this appendix, we present the color maps of the DDCs between substructures in different time windows for thirteen 1-residue ($A_1$, $T_2$, …, $N_{13}$, Fig. 13(a)); twelve 2-residues ($A_1$,-$T_2$, $T_2$,-$T_3$, …, $S_{12}$,-$N_{13}$, Fig. 13(b)); and ten 4-residues ($A_1$-$T_2$-$T_3$-$I_4$, $T_2$-$T_3$-$I_4$-$A_5$, …,$P_{10}$-$T_{11}$-$S_{12}$-$N_{13}$, Fig. 13(c)) which we calculated using Eq. (11) following the procedure described in Sec. II D. Notice that the correlation features terminated at $C(11,12)$ for the 1- and 2-residues, whereas, for the 4-residues, this larger-size partial structure terminated at $C(7,10)$. Upon analyzing these color maps and the correlation filtering described in Sec. III D, we ruled out the 1- and 2-residues since no separation of strong and weak correlations were obtained. The 3- and 4-residues both show the latter separation. We have finally chosen the 3-residues, aside from the fact that it has the lowest residues number showing this separation feature, its contrast is relatively better and computationally attractive. The 3-residues is therefore used in this study.

[1]M. Karplus and G. A. Petsko, Nature (London) **347**, 631 (1990).
[2]M. Karplus and J. A. McCammon, Nat. Struct. Biol. **9**, 646 (2002).
[3]M. Karplus and J. Kuriyan, Proc. Natl. Acad. Sci. U.S.A. **102**, 6679 (2005).
[4]U. H. E. Hansmann and Y. Okamoto, Curr. Opin. Struct. Biol. **9**, 177 (1999).
[5]R. D. Taylor, P. J. Jewsbury, and J. W. Essex, J. Comput. Aided Mol. Des. **16**, 151 (2002).
[6]B. Rost, J. Struct. Biol. **134**, 204 (2001).
[7]J. Schonbrun, W. J. Wedemeyer, and D. Baker, Curr. Opin. Struct. Biol. **12**, 348 (2002).
[8]C. A. Floudas, H. K. Fung, S. R. McAllister, M. Mönnigmann, and R. Rajgaria, Chem. Eng. Sci. **61**, 966 (2006).
[9]Y. Zhang, Curr. Opin. Struct. Biol. **18**, 342 (2008).
[10]J.-E. Shea and C. L. Brooks III, Annu. Rev. Phys. Chem. **52**, 499 (2001).
[11]C. D. Snow, E. J. Sorin, Y. M. Rhee, and V. S. Pande, Annu. Rev. Biophys. Biomol. Struct. **34**, 43 (2005).

[12]H. A. Scheraga, M. Khalili, and A. Liwo, Annu. Rev. Phys. Chem. **58**, 57 (2007).
[13]S. B. Prusiner, Science **278**, 245 (1997).
[14]R. N. Rosenberg, Neurology **54**, 2045 (2000).
[15]D. J. Selkoe and M. B. Podlisny, Annu. Rev. Genomics Hum. Genet. **3**, 67 (2002).
[16]T. Foltynie, S. Sawcer, C. Brayne, and R. A. Barker, J. Neurol., Neurosurg. Psychiatry **73**, 363 (2002).
[17]M. J. Farrer, Nat. Rev. Genet. **7**, 306 (2006).
[18]J. Liu, L. A. Campos, M. Cerminara, X. Wang, R. Ramanathan, D. S. English, and V. Munoz, Proc. Natl. Acad. Sci. U.S.A. **109**, 179 (2012).
[19]H. S. Chung, I. V. Gopich, K. McHale, T. Cellmer, J. M. Louis, and W. A. Eaton, J. Phys. Chem. A **115**, 3642 (2011).
[20]S. Q. Liu, X. L. Ji, Y. Tao, D. Y. Tan, K. Q. Zhang, and Y. X. Fu, *Protein Engineering*, edited by P. Kaumaya (In Tech., Croatia, 2012), p. 207.
[21]J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes, Proteins **21**, 167 (1995).
[22]V. Daggett and A. R. Fersht, Trends Biochem. Sci. **28**, 18 (2003).
[23]E. R. Morris and M. S. Searle, "UNIT 28.2 overview of protein folding mechanisms: Experimental and theoretical approaches to probing energy landscapes," Curr. Protoc. Protein. Sci. (published online, 2012).
[24]A. R. Fersht, Curr. Opin. Struct. Biol. **7**, 3 (1997).
[25]A. R. Fersht, Proc. Natl. Acad. Sci. U.S.A. **97**, 1525 (2000).
[26]P. E. Leopold, M. Montal, and J. N. Onuchic, Proc. Natl. Acad. Sci. U.S.A. **89**, 8721 (1992).
[27]D. J. Wales, Philos. Trans. R. Soc., A **363**, 357 (2005).
[28]C. M. Dobson and M. Karplus, Curr. Opin. Struct. Biol. **9**, 92 (1999).
[29]J. N. Onuchic and P. G. Wolynes, Curr. Opin. Struct. Biol. **14**, 70 (2004).
[30]C. M. Dobson, Semin. Cell Dev. Biol. **15**, 3 (2004).
[31]S. Gianni, C. D. Geierhaas, N. Calosci, P. Jemth, G. W. Vuister, C. Travaglini-Allocatelli, M. Vendruscolo, and M. Brunori, Proc. Natl. Acad. Sci. U.S.A. **104**, 128 (2007).
[32]M. Brunori, S. Gianni, R. Giri, A. Morrone, and C. Travaglini-Allocatelli, Biochem. Soc. Trans. **40**, 429 (2012).
[33]Y. Ivarsson, C. Travaglini-Allocatelli, M. Brunori, and S. Gianni, Eur. Biophys. J. **37**, 721 (2008).
[34]K. A. Dill, S. B. Okzan, M. S. Shell, and T. R. Weikl, Annu. Rev. Biophys. **37**, 289 (2008).
[35]M. Tsytlonok and L. S. Itzhaki, Arch. Biochem. Biophys. **531**, 14 (2013).
[36]K. Lindorff-Larsen, S. Piana, R. O. Fror, and D. E. Shaw, Science **334**, 517 (2011).
[37]J. K. Weber and V. S. Pande, Biophys. J. **102**, 859 (2012).
[38]P. Bernaola-Galván, P. C. Ivanov, L. A. N. Amaral, and H. E. Stanley, Phys. Rev. Lett. **87**, 168105 (2001).
[39]J. C. Wong, H. Lian, and S. A. Cheong, Physica A **388**, 4635 (2009).
[40]Y. Zhang, G. Lee, J. C. Wong, J. L. Kok, M. Prusty, and S. A. Cheong, Physica A **390**, 2020 (2011).
[41]S. A. Cheong, R. P. Fornia, G. Lee, J. L. Kok, W. S. Yim, D. Y. Xu, and Y. Zhang, Econ. E-J. **6**, 2012 (2012).
[42]S. K. Lai, Y. T. Lin, P. J. Hsu, and S. A. Cheong, Comput. Phys. Commun. **182**, 1013 (2011).
[43]M. Pokrzywa, I. Dacklin, D. Hultmark, and E. Lundgren, Eur. J. Neurosci. **26**, 913 (2007).
[44]R. E. Steward, R. S. Armen, and V. Daggett, Protein Eng., Des. Sel. **21**, 187 (2008).
[45]M. Yang, B. Yordanov, Y. Levy, R. Brüschweiler, and S. Huo, Biochemistry **45**, 11992 (2006).
[46]R. Tycko, Curr. Opin. Struct. Biol. **14**, 96 (2004).
[47]C. P. Jaroniec, C. E. MacPhee, N. S. Astrof, C. M. Dobson, and R. G. Griffin, Proc. Natl. Acad. Sci. U.S.A. **99**, 16748 (2002).
[48]C. P. Jaroniec, C. E. MacPhee, V. S. Bajaj, M. T. McMahon, C. M. Dobson, and R. G. Griffin, Proc. Natl. Acad. Sci. U.S.A. **101**, 711 (2004).
[49]J. Sørensen, D. Hamelberg, B. Schiøtt, and J. A. McCammon, Biopolymers **86**, 73 (2007).
[50]E. Paci, J. Gsponer, X. Salvatella, and M. Vendruscolo, J. Mol. Biol. **340**, 555 (2004).
[51]P. Mesquida, E. M. Blanco, and R. A. McKendry, Langmuir **22**, 9089 (2006).
[52]F. Meersman, R. Q. Cabrera, P. F. McMillan, and V. Dmitriev, Biophys. J. **100**, 193 (2011).
[53]M. Lei, M. Yang, and S. Huo, J. Struct. Biol. **148**, 153 (2004)
[54]D. W. Li, L. Han and S. Huo, J. Phys. Chem. B **111**, 5425 (2007).
[55]M. Porrini, U. Zachariae, P. E. Barran, and C. E. MacPhee, J. Phys. Chem. Lett. **4**, 1233 (2013).
[56]R. S. Armen, D. O. V. Alonso, and V. Daggett, Structure **12**, 1847 (2004).
[57]P. J. Hsu, S. K. Lai, and A. Rapallo, J. Chem. Phys. **140**, 104910 (2014).
[58]D. van der Spoel and E. Lindahl, J. Phys. Chem. B **107**, 11178 (2003).
[59]S. Nosé, Mol. Phys. **52**, 255 (1984).
[60]W. G. Hoover, Phys. Rev. A **31**, 1695 (1985).
[61]M. Parrinello and A. Rahman, J. Appl. Phys. **52**, 7182 (1981).
[62]S. Nosé and M. L. Klein, Mol. Phys. **50**, 1055 (1983).
[63]J. F. Gibrat, T. Madej, and S. H. Bryant, Curr. Opin. Struct. Biol. **6**, 377 (1996).
[64]S. E. Brenner, C. Chothia, and T. J. P. Hubbard, Curr. Opin. Struct. Biol. **7**, 369 (1997).
[65]J. L. Sussman, D. Lin, J. Jiang, N. O. Manning, J. Prilusky, O. Ritter, and E. E. Abola, Acta Crystallogr., Sect. D: Biol. Crystallogr. **54**, 1078 (1998).
[66]Ajay and M. A. Murcko, J. Med. Chem. **38**, 4953 (1995).
[67]T. Lengauer and M. Rarey, Curr. Opin. Struct. Biol. **6**, 402 (1996).
[68]M. K. Gilson and H. X. Zhou, Annu. Rev. Biophys. Biomol. Struct. **36**, 21 (2007).
[69]S. Henrich, Outi M. H. Salo-Ahen, B. Huang, F. Rippmann, G. Cruciani, and R. C. Wade, J. Mol. Recognit. **23**, 209 (2010).
[70]J. Boström, A. Hogner, and S. Schmitt, J. Med. Chem. **49**, 6716 (2006).
[71]J. O. Ebalunode and W. Zheng, Curr. Top. Med. Chem. **10**, 669 (2010).
[72]P. J. Ballester, I. Westwood, N. Laurieri, E. Sim, and W. G. Richards, J. R. Soc., Interface **7**, 335 (2010).
[73]E. Cannon, F. Nigsch, and J. Mitchell, Chem. Cent. J. **2**, 3 (2008).
[74]H. Haken, Naturwissenschaften **67**, 121 (1980).
[75]A. Wunderlin and H. Haken, Z. Phys. B **44**, 135 (1981).
[76]H. Haken, Physica D **97**, 95 (1996).
[77]L. M. Pecora and T. L. Carroll, Phys. Rev. Lett. **64**, 821 (1990).
[78]P. Bernaola-Galván, R. Román-Roldán, and J. L. Oliver, Phys. Rev. E **53**, 5181 (1996).
[79]R. Román-Roldán, P. Bernaola-Galván, and J. L. Oliver, Phys. Rev. Lett. **80**, 1344 (1998).
[80]J. Lin, IEEE Trans. Inf. Theory **37**, 145 (1991).
[81]W. Li, in *Proceedings of the Fifth Annual Conference on Computational Molecular Biology* (RECOMB 2001) (ACM, Montreal, 2001), p. 204.
[82]W. Li, Phys. Rev. Lett. **86**, 5815 (2001).
[83]G. E. Schwarz, Ann. Stat. **6**, 461 (1978).
[84]S. A. Cheong, P. Stodghill, D. J. Schneider, S. W. Cartinhour, and C. R. Myers, "Extending the recursive Jensen-Shannon segmentation of biological sequences," arXiv:0904.2466 [q-bio.GN] (unpublished).
[85]S. A. Cheong, P. Stodghill, D. J. Schneider, S. W. Cartinhour, and C. R. Myers, "The context sensitivity problem in biological sequence segmentation," arXiv:0904.2668 [q-bio.GN] (unpublished).
[86]C. Chen and Y. Xiao, Bioinformatics **24**, 659 (2008).
[87]Y. Xiao, C. Chen and Y. He, Int. J. Mol. Sci. **10**, 2838 (2009).
[88]L. Holm and C. Sander, Science **273**, 595 (1996).
[89]P. D. Lena, M. Vassura, L. Margara, P. Fariselli, and R. Casadio, Algorithms **2**, 76 (2009).
[90]F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weight, Proc. Natl. Acad. Sci. U.S.A. **108**, E1293 (2011).
[91]M. Vendruscolo, E. Kussell, and E. Domany, Folding Des. **2**, 295 (1997).
[92]A. Vullo, I. Walsh, and G. Pollastri, BMC Bioinf. **7**, 180 (2006).