Part I

Weak correlation effect on the folding of transthyretin fragment studied by the shape similarity technique and time series methods

Abstract

We identified the folding signatures for the transthyretin fragment TTR(105-115), the model system chosen in this work for illustration, through a combination of three techniques: (1) shape recognition for shape similarity depiction, (2) time series segmentation for time domains discovery of TTR(105-115) peptide making transitions from one conformation to another, and (3) time series clustering for spatial fingerprints exploitation of the stable conformations. Guided by these techniques, the complex dynamics of trajectory was considerably simplified. Most importantly, using the substructure of three consecutive residues in the time series clustering revealed that weaker correlations between the consisting residues instigate the transitions between stable conformations, which we believed is universal for identifying the precursory events of polymer and protein folding.

1 Introduction

A mechanistic understanding of smaller amyloid-forming peptide can help us better understand how the fibril-packing proceeds [1, 2, 3, 4], how the deposit of amyloid fibrils associates with brain disorders such as Alzheimer's disease, type II diabetes, and transmissible spongiform encephalopathies [5, 6, 7], and ultimately helps accelerate the medical discovery [8, 9] and applications in nanomaterials [10, 11, 12, 13, 14]. In this work, we used known shapes to determine the time evolution of unknown shapes generated in a dynamical process of amyloid transthyretin fragment, TTR(105-115) [15]. We first translated the characteristics of the shape into a set of descriptors representing the signature of the shape. Next, the similarity function was defined to measure the deviation of the two sets of shape descriptors. Finally, a time series function of shape similarity was generated from the trajectory of molecular dynamics (MD) simulation. This method thus allows interpretation of the conformational change of a protein or a polymer without the necessity for inquiring into the complex force fields and the modeling systems. Furthermore, the shape similarity is independent of the size of the molecule, and is therefore possible to analyze substructures and identify more accurate inherent dynamics.

2 Methods

2.1 Shape recognition of the partial structures

In our MD simulation, all-atom force fields were applied to both peptide and water molecules [16]. The temperature and pressure were fixed at 298 K and 1 atm. [17, 18]. The trajectory was stored every 0.5 ps to yield more than 1.8×10^6 data points. To perform the dynamical analysis, we adopted the ultrafast shape recognition technique (USRT) developed by Ballester *et al.* [19] for two reasons. First, it renders the atomic distance distribution into a set of statistical moments such as the mean, variance, skewness, and kurtosis. Second, Cannon *et al.* [20] have proved its accuracy and efficiency using four moments, as well as four reference locations, in the screening of a large protein data bank.

The reference locations that form the four atomic distance distributions are: the center of mass (COM); the atomic location closest to the COM; the atomic location farthest from the COM (FCM); and the atomic location farthest from the FCM (FTF). The four moments are then performed for each distribution for a total of 16 moment descriptors,

$$\begin{aligned} M(t) &= \{\mu_1^{COM}, \ldots, \mu_4^{COM}, \ldots, \mu_1^{FTF}, \ldots, \mu_4^{FTF}\} \\ &= \{M_l(t)\}_{l=1}^{16} \end{aligned} \tag{1}$$

, one at each MD time step t. For two structures, for example the reference configuration at $t = 0$, and the instantaneous configuration at MD time step t, we define the similarity

$$\zeta(t) = \left(1 + \frac{1}{16}\sum_{l}^{16} |M_l(t) - M_l(0)|\right)^{-1} \tag{2}$$

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

In our work, the USRT was used to construct several time series functions of similarities between shapes along the 930 ns trajectory (1.8×10^6 data points) and the fully-extended initial conformation ($t = 0$), which served as the known shape. After the trajectory was complete, we excluded some specific atoms by atomic labeling for substructure analysis when calculating the reference locations and their corresponding atomic distances. For example, in our model calculation of the TTR(105-115) peptide (181 atoms), in an aqueous environment, the head-tail substructure considered only 12 atoms (those of the terminal residues A_1 and N_{13}), whereas the remaining atoms were excluded in the moment descriptors $M(t)$ and $M(0)$ (Fig. 1(b)). The similarity profile (Fig. 1(d)) proved to be similar to one that considered all 181 atoms, thereby substan-tially reducing computing time spent on calculating the dynamics of the complete molecule. This head-tail similarity function depicts the highest similarity ($\zeta \approx 1$) to the initial state at $t = 0$, when the distance between A_1 and N_{13} is greatest. However, when the head-tail similarity decreases to a very low value, it indicates the distance of A_1 and N_{13} deviates markedly from the distance at $t = 0$, suggesting that the TTR(105-115) peptide may have folded. Such prevision of shape similarity precluded regions of higher similarity and focused due attention on those regions with lower similarity (< 0.2). To focus more accurately on the lower similarity regions, we performed time series segmentation [21, 22, 23, 24] to obtain 29 precisely located boundaries (blue/red vertical lines, Fig. 1(d)). First, we assume that the high-frequency fluctuations are statistically independent samples from a Gaussian distribution. The likelihood to observe a given sequence of similarities $\zeta = (\zeta_1, \zeta_2, \ldots, \zeta_i, \ldots, \zeta_n)$ would be given by

$$L_1 = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big[-\frac{(\zeta_i - \gamma)^2}{2\sigma^2}\Big] \tag{3}$$

, if the similarities ζ_i are all sampled from a Gaussian distribution with mean γ and variance σ^2. However, we know that the transthyretin fragment folds and unfolds over the course of the MD simulation, and hence the similarity time series $\zeta = (\zeta_1, \zeta_2, \ldots, \zeta_i, \ldots, \zeta_n)$ of the head-tail substructure must necessarily be statistically nonstationary, consisting of many statistically stationary segments (see Fig. 1(d)). Since we do not a priori know how many such segments there are, and where the 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 $\zeta = (\zeta_1, \zeta_2, \ldots, \zeta_i, \ldots, \zeta_n)$ in such a 2-segment model is then given by

$$\begin{array}{l} L_2(t) = \prod_{i=1}^{t} \frac{1}{\sqrt{2\pi\sigma_L^2}} \exp\Big[-\frac{(\zeta_i - \gamma_L)^2}{2\sigma_L^2}\Big] \\ \prod_{i=t+1}^{n} \frac{1}{\sqrt{2\pi\sigma_R^2}} \exp\Big[-\frac{(\zeta_i - \gamma_R)^2}{2\sigma_R^2}\Big] \end{array} \tag{4}$$

, where the left segment $\vec{\zeta}_L = (\zeta_1, \zeta_2, \ldots, \zeta_t)$ is sampled from a Gaussian distribution with mean γ_L and variance σ_L^2, while the right segment $\vec{\zeta}_R$ is sampled from a Gaussian distribution with mean γ_R and variance σ_R^2. For any time series, we can always fit it to a 1-segment model or a 2-segment model. To decide which model is better, we compute the logarithm of the ratio of likelihoods.

$$\Delta(t) = \ln \frac{L_2(t)}{L_1} = n \ln \hat{\sigma} - t \ln \hat{\sigma}_L - (n - t) \ln \hat{\sigma}_R + \frac{1}{2} \tag{5}$$

, 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 generalized *Jensen-Shannon divergence* (JSD) [25]. This 1-to-2 segmentation procedure can then be iterated to discover more and more segment boundaries. As more and more segments are found, they become shorter and shorter (down to a lower limit of 2×10^4 data points in our study). With shorter segments, the JSD maxima $\Delta^* = \Delta(t^*)$ associated with new segment boundaries as well as some old segment boundaries also become smaller. At some point, Δ^* become so small that the new segment boundaries are no longer statistically significant, and we terminate the recursive segmentation.

Alternatively, we can also terminate the process of recursive segmentation when the JSD maxima of all new segment boundaries fall below a simple cutoff $\Delta_0 = 200$. A JSD maximum $\Delta^* = 200$ 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 Δ^*) found by the three more rigorous approaches. At each stage of the recursive segmentation process, we also perform segmentation optimization, making certain to overcome the context sensitivity problem [26].

Among these 29 boundaries six segments with lower similarities were marked by red vertical lines (also shown in Figs. 2(a)-4(a) and 2(b)-4(b)). Since a boundary stands for the transition of one statistical model (before the boundary) to another statistical model (after the boundary), using a snapshot along that trajectory pathway to represent a single statistical model is possible. In addition, guided by the correlation maps described in the discussion sections, we represented the motion of the residue by a dashed arrow and indicated by a solid double-headed arrow the non-bonding interaction between two residues (Figs. 2(c)-4(c)). Finally, we added six extra snapshots to explain how these segments evolve. Thus, the six deduced segments of lower similarity and 12 snapshots revealed the mechanism of the precursory folding events.

2.2 Correlation filtering

The head-tail similarity led us to the six lower similarity regions. The mechanisms of forming these regions are still unknown. Because our peptide consists of 13 residues, labeling by Ace1(A_1), Tyr2(T_2), Thr3(T_3), Ile4(I_4), Ala5(A_5), Ala6(A_6), Leu7(L_7), Leu8 (L_8), Ser9 (S_9), Pro10 (P_{10}), Tyr11(T_{11}), Ser12(S_{12}), and Nac13(N_{13}) (Fig. 1(a)), we also tested similarity functions of substructures composed of residues, that is, 11 similarity functions based on three consecutive residues at a time with respect to those at $t = 0$, which is shortened as 3-residues in the following description. Then, we performed a digital cross correlation calculation by the sliding time window method for 11 similarity functions [27]. Given the similarity time series $\zeta(t)$ of a 3-residues, we define the similarity change time series to be

$$\Delta p(t_k) = \zeta(t_{k+1}) - \zeta(t_k) \tag{6}$$

To evaluate the digital cross correlation between $\Delta p_i(t)$ and $\Delta p_j(t)$, we first compute the *sign of change time series*

$$s_i(t_k) = \begin{cases} 1, & \Delta p_i(t_k) > 0; \\ 0, & \Delta p_i(t_k) = 0; \\ -1, & \Delta p_i(t_k) < 0 \end{cases} \tag{7}$$

of $\Delta p_i(t)$, and a similar time series for $\Delta p_j(t)$. We then construct the 11×11 digital cross correlation matrix

$$C(i,j) = (1 - \delta_{ij}) \sum_{k=t/\Delta t}^{t'/\Delta t} \frac{1}{2} |S_i(t_k) S_j(t_k) [S_i(t_k) + S_j(t_k)]| \tag{8}$$

for the 11 3-residues within a given time window $[t, t']$, with the running indices $i, j = 1, 2, \ldots, 11$ corresponding to $\{A_1, T_2, T_3\}$, $\{T_2, T_3, I_4\}$, ..., $\{T_{11}, S_{12}, N_{13}\}$ as defined in the y-axis of Fig. 2(a)-4(a) and 2(b)-4(b). For instance, for a time window, the correlation $C(5,6)$ represents the occurrence of a synchronous behavior between the similarity functions of $A_5 - A_6 - L_7$ and $A_6 - L_7 - L_8$ (Fig. 1(c)). The length of the time window was guided by the results of time series segmentation. In principle, it should not be larger than the shortest segment, which in our case is around 10 ns. Therefore, we use a 5-ns time window and slide it forward 1 ns at a time to examine the time evolution of $C(i,j)$.

The structural evolution of 3-residues can in principle reflect the changes of dihedral angle of the backbone. Statistical analysis of the correlations revealed that the uniform background signals of 3-residues are mostly contributed by $C(i,j)$, $j > i + 2$, and form a Gaussian-like distribution in the range of 4700 and 5270 (Fig. 1(e), green histogram). Not only strong correlations are separated, but also weaker-than-average correlations. The weak/strong separation is the most significant feature against the substructures other than 3-residues (e.g., 1-residue and 2-residues). To investigate this phenomenon, we filtered out the background signals, which are $4700 < C < 5270$. Only 19 correlation indices remained, from $C(1,2)$ to $C(9,11)$. We plotted these indices in our color maps (y-axis) where every pixel represents the starting time of the window (x-axis) (Figs. 2(a)-4(a) and 2(b)-4(b)). Each pixel was colored as the correlation intensity shown in the color bar. We thus combined the correlation intensity of the sliding window of 3-residues and the segmentation boundaries of the head-tail similarity with our color maps. The regions with lower head-tail similarity will be focused in the following sections.

2.3 Weak and strong correlation

The importance of the weak correlation is usually ignored or often discarded. To find the causes of the strong/weak separation, we review our computations of the correlation matrix, and propose the following explanation:

The digital cross correlation that we adopted measures how many times two similarity functions increase or decrease simultaneously. That is, for the stronger-than-average correlation, the substructures may undergo synchronous motion such as a close-packed formation or a global motion. In contrast, a weaker-than-average correlation may indicate that the individual residues are (i) experiencing diverse influences to/from the other residues, (ii) being affected by their self-twisting, (iii) possessing a long dangling residue, and (iv) self-approaching inside the substructure. The 3-residues may also interact with remote residues that come nearer, or just with its neighbors. Consequently, asynchronous motion can occur. Such weaker-than-average responses should only depend on the substructures that are selected for correlation computations and are less significant in 1- and 2-residues.

Using different substructures may yield new insights into the correlation. It is important therefore to have a proper choice of substructure that reflects most the dynamical information. Otherwise, the weak correlation may blend into the background level, and less information about the conformational change would be gained. As can be seen in the following sections, our hypothesis of weak/strong correlations revealed not only the folding/unfolding signatures, but also helped in the capture of their precursory events.

3 Results and discussion

The 930ns-long trajectory of the TTR(105-115) peptide was demarcated into six head-tail similarity intervals (Fig. 1(d)). Next, we filtered out most of the correlations lying within the uniform background levels. Then, we combined these results with the correlation color map, and focused on the lower similarity intervals. Within these intervals the disparate correlation patterns show distinguishable folding mechanisms. Three regions are presented: (1) at approximately 100 ns, when the peptide folded into an α helix, (2) at approximately 300 ns, when the peptide folded into a β hairpin, and (3) between 400 ns and 930 ns, when the peptide folded into an α helix, but preserved as well signatures of β hairpin.

In the color maps showing correlations stronger-than-average ($C > 5270$), correlations that are most different (largest) from background are colored red, while those that are least different (smallest) from background are colored blue (Figs. 2(a)-4(a)). Conversely, in the color maps showing weaker-than-average correlations ($C < 4700$), correlations that are most different (smallest) from background are colored red, while those that are least different (largest) from background are colored blue (Figs. 2(b)-4(b)).

Let us begin with the interval $73 - 112$ ns. Before this region (approximately 70 ns), the TTR(105-115) peptide undergoes coiling topology in the head part tending to form the α-helix (Fig. 2(c)). It was observed that A_1 exhibited a non-bonding interaction with L_7, and similarly for T_3 with A_5. Therefore, the synchronous motion of the related substructures (3-residues) is destroyed, resulting instead in weaker-than-average correlations (Fig. 2(b), $C(1,2)$, $C(2,3)$, $C(3,4)$, $C(6,7)$, and $C(5,7)$). This dynamical feature can be classified as the first type of weak correlation, as well as the representative precursory events of folding. However, the approach of these residues causes $(2,4)$ to lie in the center of the close-packed coil (Fig. 2(c)). At the same time, the tail part (S_9 to N_{13}) heads towards the head of the coil. The correlation color map captures both the close-packed (slightly enhanced $C(2,4)$) and the global motion (strongly enhanced $C(9,10)$) behaviors (Fig. 2(a), 70 ns). In the region spanning 73 to 112 ns, a significant switching from the enhanced to the short and heavily-suppressed $C(9,10)$ (approximately 80 ns) reflects the asynchronous response, because T_{11} temporarily approaches A_5 then moves towards T_2. When the movement is complete, T_{11} leaves room for A_5 to interact with S_9 (Fig. 2(c), 91 ns). At this moment the most compact α-helix formation is complete and the close-packed coiling is enhanced (seen to increase stronger-than-average $C(2,4)$ at approximately 91 ns). While leaving the folding region at approximately 112 ns, the weaker-than-average correlations start diminishing and then disappear, and the stronger-than-average correlations of global motion reappear ($C(9,10)$ and $C(6,8)$). Until the next folding region, from 265 to 346 ns, the polymer maintains a linear form (e.g., the snapshot at 128 ns). The suppression of $C(6,7)$ and many other weak correlations mostly occur outside the α-helix folding region, which can be referred by the twisting effect, a theme we discussed.

Next, we observed the β-hairpin conformation of TTR(105-115) peptide in the region spreading from 265 to 346 ns (Fig. 3(c)). Before the conformation at approximately 260 ns, the contacts of I_4 with L_7 and with T_2 form the turning point of the β hairpin. This dynamical process is captured by the short suppression of $C(3,4)$, $C(6,7)$, and $C(7,8)$. We saw moreover a persistently weak $C(10,11)$, which is caused by the twisting of the tail part, and the latter starts to become stronger at approximately the same time (from yellow to light blue), indicating that tail is approaching the head part (global motion) (Fig. 3(b) and 3(c), 260 ns). After these precursory events at approximately 273 ns, the strong $C(9,10)$ and $C(6,7)$ assume the close-packed formation in the tail part where T_{11} is close to T_3, so do S_{12} and N_{13} is close to A_1 and T_2, respectively (Fig. 3(a) and 3(c), 273 ns). The head part, however, adjusts its topology to achieve the most compact β hairpin. The suppression of $C(2,4)$ (zone I) thus captures the asynchronous twisting motion between T_3, I_4, and A_5, which we refer to as the second type of weak

correlation. After I_4 rotates, it lies in a tight formation between T_3, L_7, S_9, and P_{10}, which is reflected by a strong $C(4,5)$ and the switching of the stronger-than-average $C(6,7)$ to $C(6,8)$ (zone II). Though the most compact β hairpin is formed in zone II, the dangling T_2 still prohibits the establishment of $C(2,4)$ (see Fig. 3(b), approximately from 280 to 300 ns). This free-dangling effect by the longer residue (see T_2 in Fig. 3(c), 300 ns) can be classified as the third type of weak correlation.

Scanning through the entire folding region (zones I-III), we observed oscillatory behavior of the stronger-than-average $C(9,10)$. Furthermore, we found that $C(10,11)$ frequently switches between strong and weak correlation. These features imply a sliding motion between the head and the tail, so that A_1 contacts back and forth with S_9 and T_{11}. The phenomenon becomes more frequently in zone III because of the rotation of I_4 that triggers the tail part to slide over the head part, and the latter can be realized by the weak $C(3,4)$ (Fig. 3(a), at the end of zone III). As a result, the β hairpin becomes unstable and starts to unfold.

The pattern of the color map for the time interval 400 to 930 ns is more complicated. Here we observed not only α-helix signatures (strong $C(2,4)$ and weak $C(5,7)$) but also β-hairpin characteristics (oscillatory $C(9,10)$ and $C(10,11)$) distributed over the zone I of lower head-tail similarity (Figs. 4(a) and 4(b)). Despite mixed α and β properties, the representative snapshot of zone I starting from 615 ns (Fig. 4(c)) appears similar to the aforementioned α-helix around 80 ns (Fig. 2(c)) and easily leads one to misjudge that the inherent properties of the two conformations are indistinguishable when examining along the trajectory pathway. The two conformations are in fact dynamically distinct, which reaffirms the importance of the correlation analysis that we have developed. We noticed moreover that the extremely suppressed $C(1,2)$ occurs before zone I and also approximately between 700 and 840 ns (Fig. 4(b)), because of the synchronous motion between $A_1 - T_2 - T_3$ and $T_2 - T_3 - I_4$ being destroyed when their common residue T_2 interacts with T_3, as illustrated at 574 ns of Fig. 4(c). Such self-nonbonding interaction can be recognized as the fourth type of weak correlation, which triggers the mixed characteristics of α and β in zone I. We found that $C(1,2)$ becomes extremely low only when T_2 is in contact with T_3, resulting in a compact head coil that is shorter than the stable α-helix formation seen at 80 ns. This leads to a longer tail that has some characteristics of a β hairpin (Fig. 4(c), 686 ns). In zone II, we observed another unstable α-helix, with a much shorter life span. In recapitulation, we found that the folding/unfolding events are mostly driven by the manifestations of the strong/weak correlations. Observations described above contain essential information of the complex dynamics and give new insights into the folding/unfolding mechanisms.

In summary, the USRT provides a mean of transforming the structural evolution into a time-series function, and yet preserving the dynamics of the trajectory. The methodology is applicable to any time-series analysis. Using the time series segmentation and time series clustering method, the complicated dynamics can be simplified into a few time intervals that cover the majority of the folding scenarios. Most importantly, we used three consecutive residues as our substructure to address the importance of the weak correlations between the consisting residues. The precursory events found in our model calculations indicate that longer residues play a dominant role in inducing the folding/unfolding process, because they can prolongate and reach more residues [28]. For example, Tyr2 is likely to initiate the coiling head of *TTR*(105-115) peptide and Tyr11 helps to curl the tail up to the head. Nevertheless, such dynamical processes and the patterns of the color maps strongly depend on model system under studied. On the other hand, the mechanisms of forming the strong/weak correlations rely on the assignment of the substructures. Future research may strike at finding further evidence of this weaker correlation phenomenon, perhaps in other more complex systems. A new perspective may be to look for a carefully assigned substructure. For example, a larger substructure such as three consecutive peptides, is possible for large proteins.

References

[1] Michael R Sawaya, Shilpa Sambashivan, Rebecca Nelson, Magdalena I Ivanova, Stuart A Sievers, Marcin I Apostol, Michael J Thompson, Melinda Balbirnie, Jed J W Wiltzius, Heather T McFarlane, Anders ØMadsen, Christian Riekel, and David Eisenberg. Atomic structures of amyloid cross-beta spines reveal varied steric zippers. *Nature*, 447(7143):453–457, 2007.

[2] Aneta T Petkova, Yoshitaka Ishii, John J Balbach, Oleg N Antzutkin, Richard D Leapman, Frank Delaglio, and Robert Tycko. A structural model for Alzheimer's beta-amyloid fibrils based on experimental constraints from solid state NMR. *Proc Natl Acad Sci USA*, 99(26):16742–16747, 2002.

[3] C Blake and L Serpell. Synchrotron X-ray studies suggest that the core of the transthyretin amyloid fibril is a continuous beta-sheet helix. *Structure London England 1993*, 4(8):989–998, 1996.

[4] Filip Meersman, Christopher Dobson, and Karel Heremans. Protein unfolding, amyloid fibril formation and configurational energy landscapes under high pressure conditions. *Chemical Society reviews*, 35(10):908–917, 2006.

[5] M Sunde and C C Blake. From the globular to the fibrous state: protein structure and structural conversion in amyloid formation. *Quarterly reviews of biophysics*, 31(1):1–39, 1998.

[6] J W Kelly. The alternative conformations of amyloidogenic proteins and their multi-step assembly pathways. *Current Opinion in Structural Biology*, 8(1):101–6, 1998.

[7] C M Dobson. The structural basis of protein folding and its links with human disease. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 356(1406):133–145, 2001.

[8] Fabrizio Chiti and Christopher M Dobson. Protein Misfolding, Functional Amyloid, and Human Disease. *Annual Review of Biochemistry*, 75(1):333–366, 2006.

[9] Sally L Gras. Amyloid Fibrils: From Disease to Design. New Biomaterial Applications for Self-Assembling Cross-β Fibrils. *Australian Journal of Chemistry*, 60(5):333, 2007.

[10] T P Knowles, A W Fitzpatrick, S Meehan, H R Mott, M Vendruscolo, C M Dobson, and M E Welland. Role of intermolecular forces in defining material properties of protein nanofibrils. *Science*, 318(5858):1900–1903, 2007.

[11] Senli Guo and Boris B Akhremitchev. Packing density and structural heterogeneity of insulin amyloid fibrils measured by AFM nanoindentation. *Biomacromolecules*, 7(5):1630–1636, 2006.

[12] Thomas Scheibel. Protein fibers as performance proteins: new technologies and applications. *Current opinion in biotechnology*, 16(4):427–433, 2005.

[13] Filip Meersman, Raúl Quesada Cabrera, Paul F McMillan, and Vladimir Dmitriev. Structural and mechanical properties of TTR105-115 Amyloid fibrils from compression experiments. *Biophysical Journal*, 100(1):193–197, 2011.

[14] Patrick Mesquida, E Macarena Blanco, and Rachel A McKendry. Patterning amyloid peptide fibrils by AFM charge writing. *Langmuir The Acs Journal Of Surfaces And Colloids*, 22(22):9089–9091, 2006.

[15] Christopher P Jaroniec, Cait E MacPhee, Nathan S Astrof, Christopher M Dobson, and Robert G Griffin. Molecular conformation of a peptide fragment of transthyretin in an amyloid fibril. *Proceedings of the National Academy of Sciences of the United States of America*, 99(26):16748–16753, 2002.

[16] David Van Der Spoel and Erik Lindahl. Brute-Force Molecular Dynamics Simulations of Villin Headpiece: Comparison with NMR Parameters. *The Journal of Physical Chemistry B*, 107(40):11178–11187, 2003.

[17] William G Hoover. Canonical dynamics: Equilibrium phase-space distributions. *Physical Review A*, 31(3):1695–1697, 1985.

[18] M Parrinello and A Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics*, 52(12):7182–7190, 1981.

[19] Pedro J Ballester, Isaac Westwood, Nicola Laurieri, Edith Sim, and W Graham Richards. Prospective virtual screening with Ultrafast Shape Recognition: the identification of novel inhibitors of arylamine N-acetyltransferases. *Journal of the Royal Society, Interface / the Royal Society*, 7(43):335–42, February 2010.

[20] Edward O Cannon, Florian Nigsch, and John B O Mitchell. A novel hybrid ultrafast shape descriptor method for use in virtual screening. *Chemistry Central journal*, 2:3, January 2008.

[21] Pedro Bernaola-Galván, Plamen Ivanov, Luís Nunes Amaral, and H Stanley. Scale Invariance in the Nonstationarity of Human Heart Rate. *Physical Review Letters*, 87(16):168105, 2001.

[22] J C Wong, H Lian, and S A Cheong. Detecting macroeconomic phases in the Dow Jones Industrial Average time series. *Physica aStatistical Mechanics and Its Applications*, 388(21):4635–4645, 2009.

[23] Yiting Zhang, Gladys Hui Ting Lee, Jian Cheng Wong, Jun Liang Kok, Manamohan Prusty, and Siew Ann Cheong. Will the US Economy Recover in 2010? A Minimal Spanning Tree Study. *Physica aStatistical Mechanics and Its Applications*, 390(11):2020–2050, 2011.

[24] Siew Ann Cheong, Robert Paulo Fornia, Gladys Lee, Jun Liang Kok, Woei Shyr Yim, Danny Yuan Xu, and Yiting Zhang. The Japanese Economy in Crises: A Time Series Segmentation Study. *SSRN Electronic Journal*, 2011.

[25] J Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991.

[26] Siew-Ann Cheong, Paul Stodghill, David J Schneider, Samuel W Cartinhour, and Christopher R Myers. The Context Sensitivity Problem in Biological Sequence Segmentation. *Transactions on Computational Biology and Bioinformatics*, page 39, 2009.

[27] S.K. Lai, Yu-Ting Lin, P.J. Hsu, and S.a. Cheong. Dynamical study of metallic clusters using the statistical method of time series clustering. *Computer Physics Communications*, 182(4):1013–1026, April 2011.

[28] Jörg Gsponer, Urs Haberthür, and Amedeo Caflisch. The role of side-chain interactions in the early steps of aggregation: Molecular dynamics simulations of an amyloid-forming peptide from the yeast prion Sup35. *Proceedings of the National Academy of Sciences of the United States of America*, 100(9):5154–9, 2003.

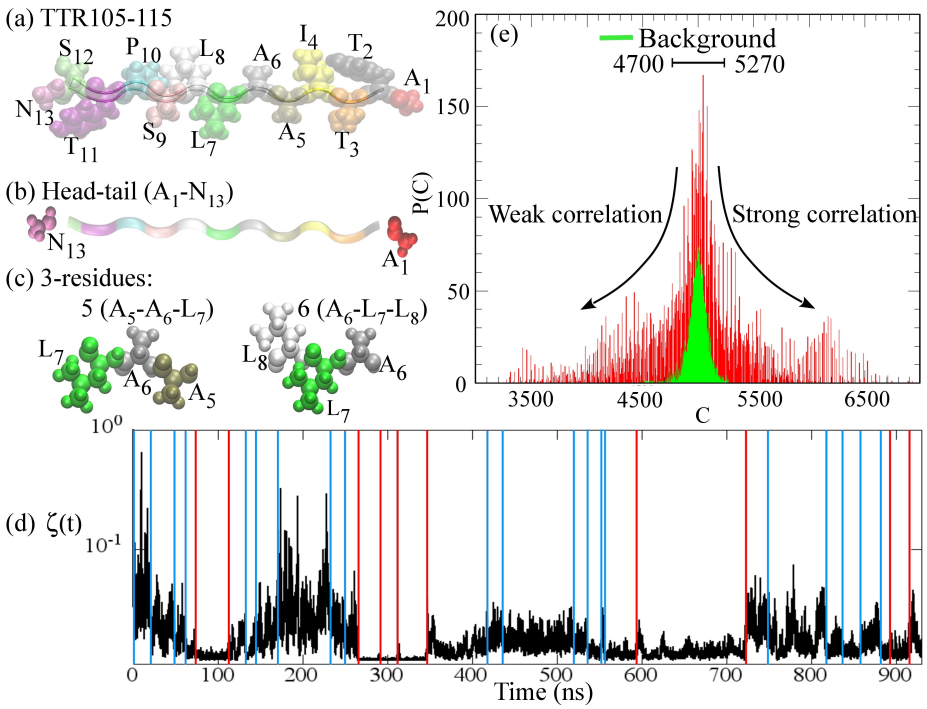

Figure 1: (a) The TTR(105-115) consisting of 13 residues (A_1 to N_{13}); (b) the head-tail substructure (A_1 and N_{13}) of the TTR(105-115); (c) the *5th* (A_5 to L_7) and the *6th* (A_6 to L_8) 3-redisues substructures; (d) shape similarity index ζ (black) of the head-tail residues of TTR(105-115), with vertical lines marking segment boundaries determined by the recursive JSD segmentation scheme; and (e) The combined histogram of correlation levels from all pairs of 3-residues.

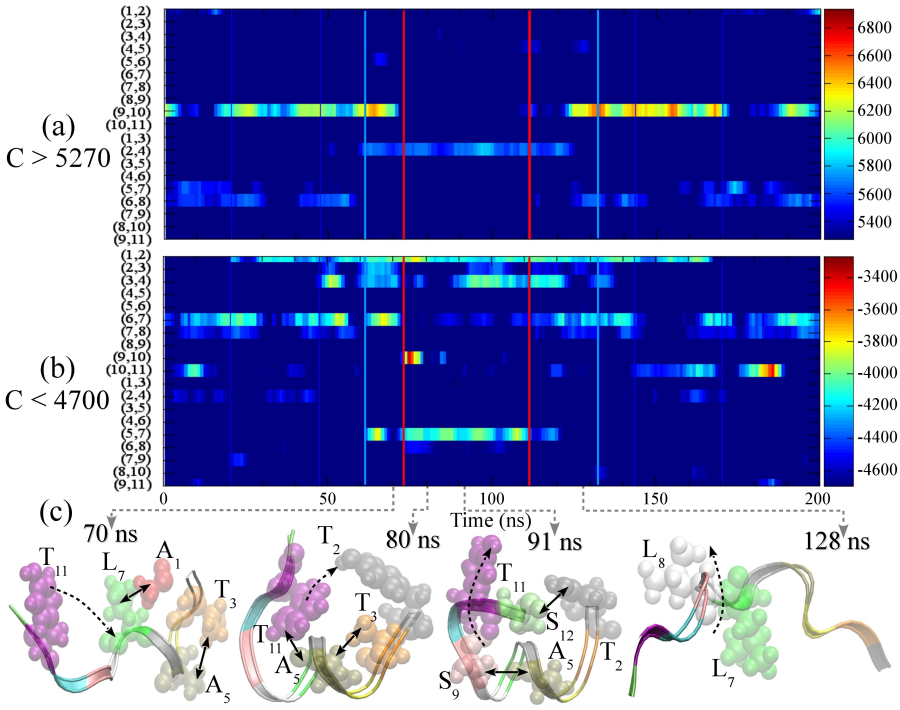

Figure 2: The 3-residues color map of correlations $C(i,j)$ from 0 to 200 ns for (a) $C > 5270$ and (b) $C < 4700$ overlapped by the boundaries of head-tail similarity (blue/red vertical lines); and (c) snapshots of TTR(105-115) at various times (70, 80, 91, and 128 ns from left to right). The correlation intensities are shown in the color bars.

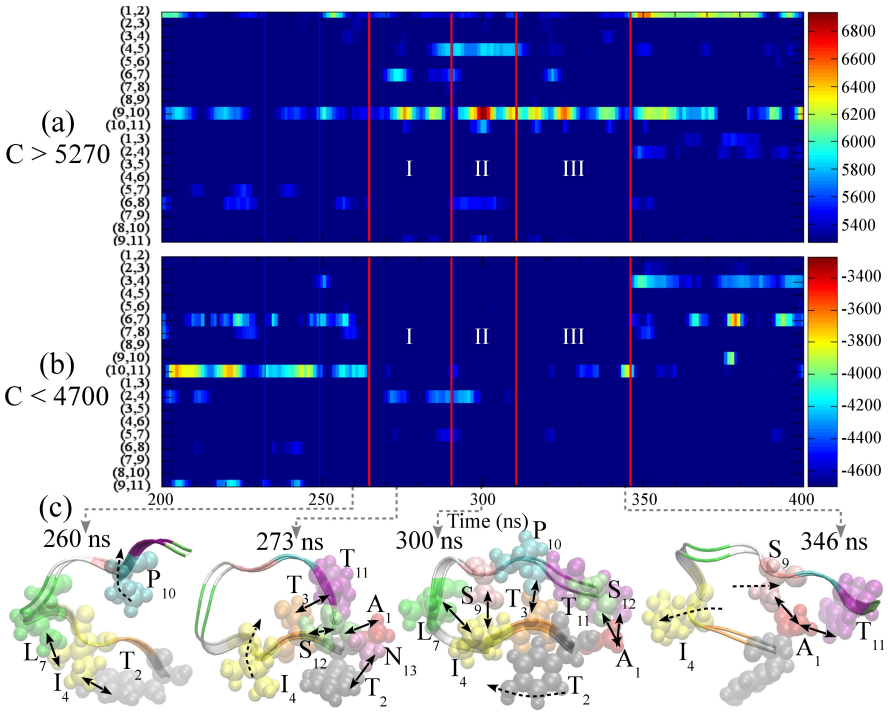

Figure 3: The 3-residues color map of correlations $C(i,j)$ from 200 to 400 ns for (a) $C > 5270$ and (b) $C < 4700$ overlapped by the boundaries of head-tail similarity (red vertical lines); and (c) snapshots of TTR(105-115) at various times (260, 273, 300, and 346 ns from left to right). The correlation intensities are shown in the color bars.

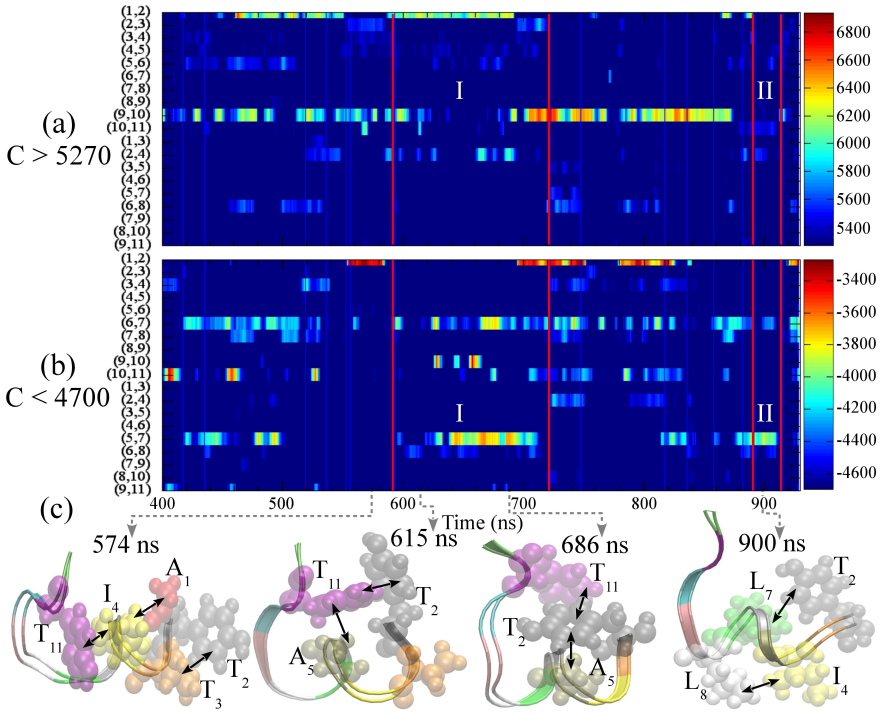

Figure 4: The 3-residues color map of correlations $C(i,j)$ from 400 to 930 ns for (a) $C > 5270$ and (b) $C < 4700$ overlapped by the boundaries of head-tail similarity (red vertical lines); and (c) snapshots of TTR(105-115) at various times (574, 615, 686, and 900 ns from left to right). The correlation intensities are shown in the color bars.

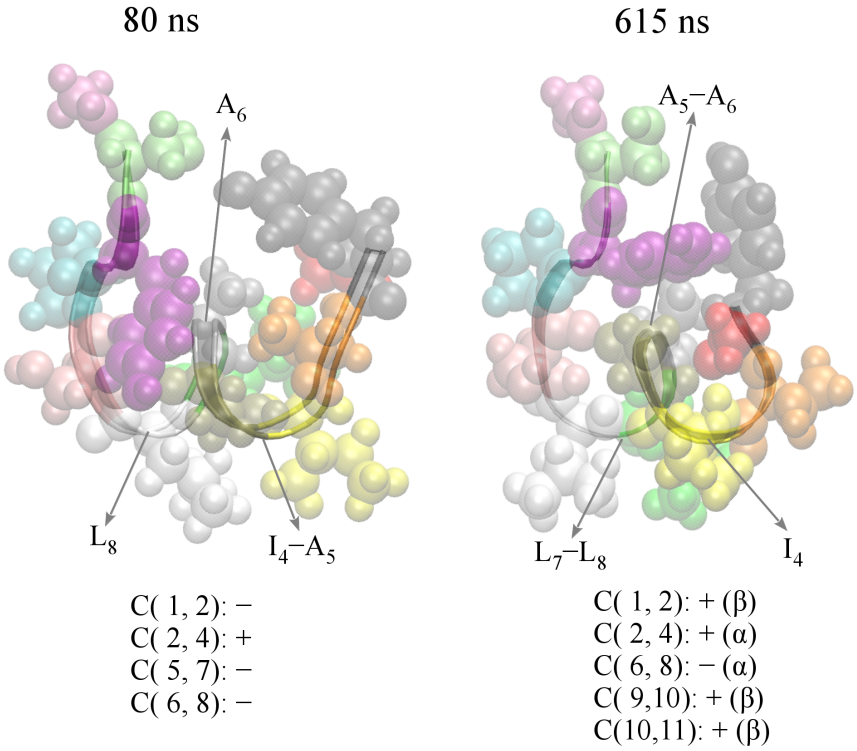

Figure 5: The snapshots of the TTR(105-115) peptide at 80 ns (within the α-helix phase) and at 615 ns (within the mixed phase). Also shown are the correlation pairs that are enhanced and suppressed. For the correlation pairs at 615 ns, we also indicate whether they are α or β signatures.

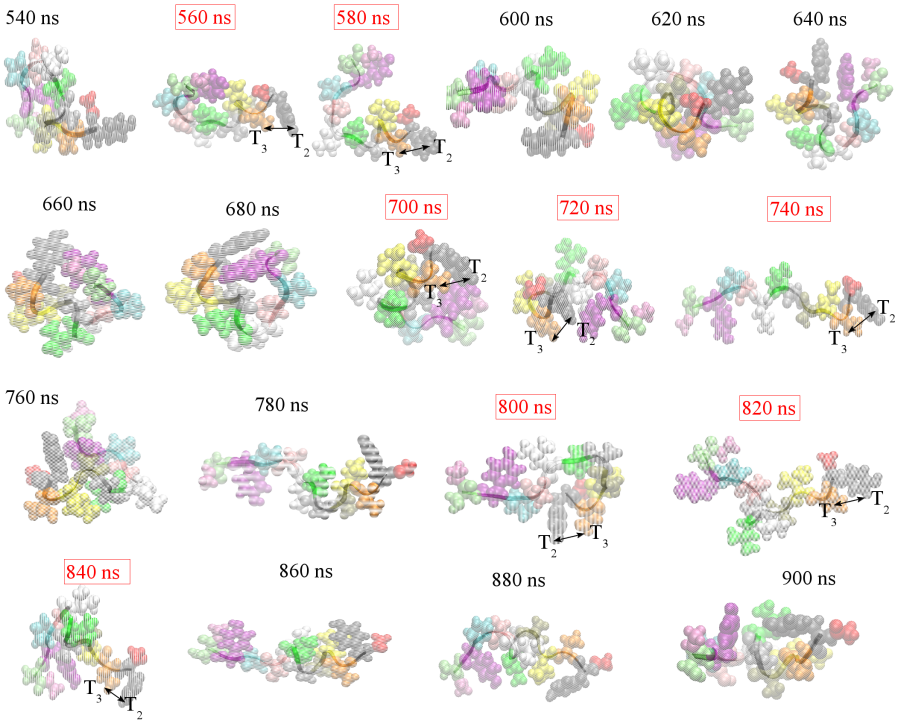

Figure 6: The snapshots from 540 ns to 900 ns. The extremely weak C(1,2) occurs when T_2 is in contact with T_3 (red blocks).

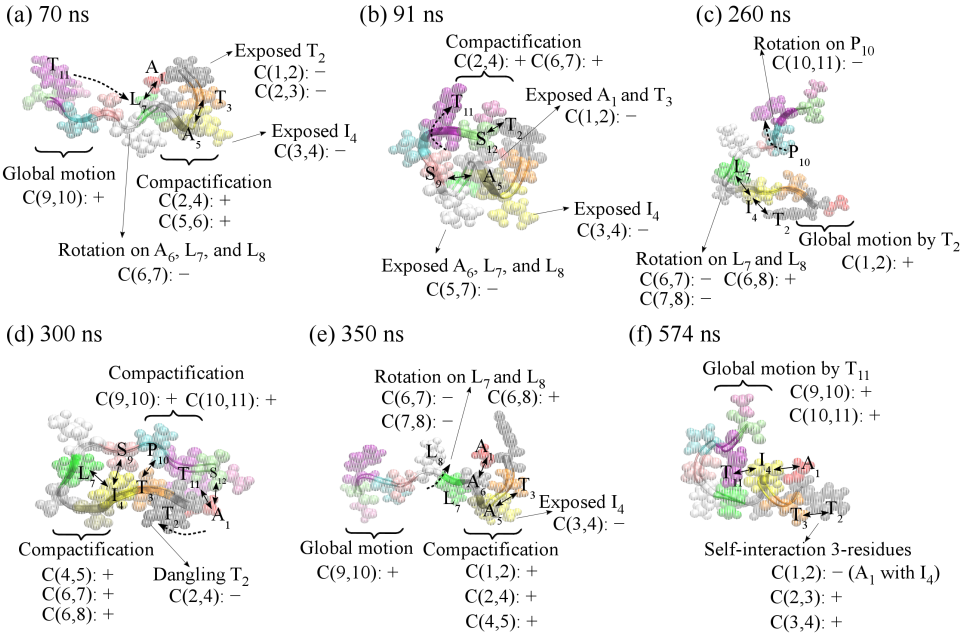

Figure 7: The six representative snapshots from (a) 70 ns to (f) 574 ns illustrating the mechanisms of stronger-than-average and weaker-than-average precursors. The former is mostly found in the global motion ((a), (c), (e), and (f)), and the compactification ((a), (b), (d), and (e)). The latter is possibly induced by exposed ((a), (b), and (e)) or dangling long residues (d), rotating residues ((a), (c), and (e)), or self-interaction within a partial structure (f). The plus signs indicates enhanced correlations and the minus signs denotes suppressed correlations.

Time (ps)	Δ^*	Time (ps)	Δ^*	Time (ps)	Δ^*
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		

Table 1: The 29 segment boundaries and their corresponding *JSD* values in the final segmentation.