Foldamer simulations: Novel computational methods and applications to poly-phenylacetylene oligomers

Sidney P. Elmer and Vijay S. Pande

Citation: The Journal of Chemical Physics **121**, 12760 (2004); doi: 10.1063/1.1812272

View online: http://dx.doi.org/10.1063/1.1812272

View Table of Contents: http://aip.scitation.org/toc/jcp/121/24

Published by the American Institute of Physics

Foldamer simulations: Novel computational methods and applications to poly-phenylacetylene oligomers

Sidney P. Elmer and Vijay S. Pande
Department of Chemistry, Stanford University, Stanford, California 94305-5080

(Received 21 May 2004; accepted 9 September 2004)

We apply several methods to probe the ensemble kinetic and structural properties of a model system of poly-phenylacetylene (pPA) oligomer folding trajectories. The kinetic methods employed included a brute force accounting of conformations, a Markovian state matrix method, and a nonlinear least squares fit to a minimalist kinetic model used to extract the folding time. Each method gave similar measures for the folding time of the 12-mer chain, calculated to be on the order of 7 ns for the complete folding of the chain from an extended conformation. Utilizing both a linear and a nonlinear scaling relationship between the viscosity and the folding time to correct for a low simulation viscosity, we obtain an upper and a lower bound for the approximate folding time within the range 70 ns$<\tau<$350 ns. This is in agreement with the experimentally measured folding time on the order of 160 ns. The kinetic model used to fit the kinetic behavior of the ensemble of trajectories provides a framework to describe the bulk folding mechanism. We were able to identify two unique clusters of conformations that provide a structural basis to account for the appearance of a kinetic intermediate in the mechanism. We discuss the implications of these findings in the context of helix-coil theory. [DOI: 10.1063/1.1812272]

I. INTRODUCTION

In recent years, there has been a remarkable increase in the amount of interest and attention given to nonbiological polymers which self-assemble into stable, unique native folds. These synthetic polymers, in part inspired by the biological folding properties of proteins, are generally called "foldamers."[1,2] Examples of some of the more popular foldamers include poly-phenylacetylene (pPA),[3] also known as phenylene ethynylene oligomers,[4,5] β-peptides,[6,7] γ-peptides,[8] peptoids,[9,10] and peptide nucleic acids,[11] in addition to other less familiar backbones.[12–15] Poly-phenylacetylene is unique compared to these other commonly studied polymers in that the monomer building blocks are not designed from derivatives of typical biological molecules such as amino acids or nucleic acids. While it may be reasonable to expect that polymers derived from biological building blocks could also form stable three-dimensional (3D) folds, this idea is not so obvious in the field of synthetic polymer chemistry.

Indeed, the fact that pPA folds into well-ordered helices reinforces the idea that helical structures transcend the realm of biopolymers. Although much has been learned about protein structure and evolution by studying the stability and kinetics of helical peptides and all-helical protein domains,[16] a rich body of knowledge remains untapped from an ever increasing source of helical foldamers. How are the folding dynamics similar, how are they different, and what general folding principles can we associate with all classes of helical polymers? While we cannot resolve these questions in this paper, we do present methods to probe and analyze the folding kinetics of any dynamical system studied using molecular dynamics computer simulations, such as data sets collected from Folding@Home (http://folding.stanford.edu).

This paper is divided into two primary analysis topics: clustering and kinetic modeling. In the Clustering and Structural Characterization sections, we discuss the standard *K*-means clustering method and then apply it to our system of interest. The resulting conformational classification is used to characterize the structural properties of the individual clusters. In the following section, Kinetics, we apply several kinetic methods to probe the kinetic behavior of our ensemble of folding trajectories, enabling us to extract a bulk folding mechanism to describe the folding behavior within the context of a simplified kinetic model. Finally, we seek to validate our computational model by a direct comparison to experimental rate measurements on the corresponding physical system.

II. METHODS

A. Dynamics and models

A molecular mechanics model for pPA has been described previously[3] for simulations using NAMD, which is based on the CHARMM force field. For the present study and all subsequent work on pPA, we use the GROMACS (Refs. 17 and 18) molecular dynamics suite (http://www.gromacs.org), version 3.1.3. Care was taken to ensure that the same energy function was used, along with the same force constants. The simulations were run using the Stochastic Dynamics algorithm as implemented in GROMACS. The simulation temperature was 300 K and the damping coefficient was 1 ps^{-1}. We used a switching function for the van der Waals forces. The switch began at 0.6 nm, with the cutoff at 0.7 nm. The neighbor list at 0.85 nm was updated every 8 fs, and the time step was 2 fs. Bonding and angle terms were harmonic with no constraints, and no periodic boundary conditions were imposed.

We use a very simple implicit solvent model to represent the solvophobic attraction between phenyl rings. Experimentally, solvophobicity is the driving force causing the polymer to fold,[4,19] and is purely a polymeric effect manifested in a poor solvent environment. We model these solvophobic forces via a generic short range potential (Lennard-Jones 6-12 potential) between distant residues in the chain. We vary the well depth to increase or decrease the strength of the attraction, similar to the solvent effects one observes experimentally in going from a polar to a nonpolar solvent. Thus, this potential includes all of the solvophobic terms and we do not include any other nonbonded terms, such as explicit electrostatic contributions to the energy function. This is a valid approximation since pPA is a relatively nonpolar molecule, and the π-π aromatic stacking interactions are secondary forces, subordinate to solvophobicity. This is evidenced by the fact that pPA unfolds experimentally in chloroform.[5] Moreover, the π-π aromatic stacking interactions are also short range and therefore it is reasonable to include them in our generic short range potential. Thus, the major forces in the model are sterics and solvophobicity.

The data set used for the analysis in this paper is comprised of over 2000 folding trajectories of a pPA 12-mer[3] containing carboxylate side chains, started from a fully extended chain. The chain consisted of 130 atoms and was allowed to diffuse unbiased (no external potential driving the chain) toward the folded state. The trajectories were terminated once the chain reached the folded state, which was defined to be within 0.1 nm conformational root mean square displacement (crms) of the helical native state (see Fig. 2, and the section, Structural Characterization).

B. Clustering

The *K*-means clustering algorithm[20] is a standard statistical tool for classifying objects when no prior information is known about the relationships between the objects in the set. By specifying a pairwise dissimilarity measure between the objects, one seeks to minimize the total within-cluster dissimilarity over the whole set. The result is a partitioning of the objects into groups (or clusters) with similar characteristics. We cluster the conformations using a modified *K*-means algorithm. Traditional *K*-means uses the Euclidean distance between two objects as the dissimilarity measure. However, a more natural distance metric for our system is the distance root mean square (drms) deviation

$$\mathrm{drms}(\mathbf{d}^{\mu},\mathbf{d}^{\nu})=\left[\frac{2}{L(L-1)}\sum_{i=1}^{L}\sum_{j=i+1}^{L}\|\mathbf{d}_{ij}^{\mu}-\mathbf{d}_{ij}^{\nu}\|^2\right]^{1/2}, \quad (1)$$

which is essentially a Euclidean norm of the atom-atom distances $\mathbf{d}_{ij}$ between two conformations μ and ν with L atoms. drms satisfies the triangle inequality; therefore, it is truly a distance metric by the strictest definition. This is an important consideration, since it directly affects the performance of the clustering algorithm.[20]

There are clustering methods that would be more ideal than *K*-means, such as hierarchical clustering.[20,21] However, the memory requirements for hierarchical clustering scale as $O(N^2)$, where N is the number of conformations, which becomes intractable for data sets containing more than several tens of thousands of conformations. The amount of memory needed to store the $N\times N$ matrix of pairwise distances quickly fills up the available memory of modern computers when data sets of this size are clustered. *K*-means is a reasonable alternative since its memory requirements scale linearly as $O(N)$. For our data set which contains over 350 thousand conformations from over 2000 trajectories, we are able to fit each conformation into memory and still have plenty of memory left to perform the clustering algorithm on a single computer.

There are, however, two major disadvantages to the *K*-means clustering algorithm. The obvious problem occurs with data sets where the true number of clusters is unknown. In this case, the choice to cluster the data into K partitions requires skill on the part of the user. Finding the true number of clusters in a data set is an ongoing research question that is actively being pursued. We use the method known as the gap statistic developed by Tibshirani.[22] Essentially, one generates a uniform distribution over the range of the principal components of the observed data points. Then this test distribution is clustered for increasing values of K and compared to the results of the true data set at the same value of K. The difference between the test distribution and the true data set at a given value of K is called the ‘‘gap.’’ The value of K^* which maximizes the gap is taken to be the value for the true number of clusters in the original data set. We illustrate the practical application of this method in the section, Structural Characterization.

The second, less obvious problem that arises with the *K*-means algorithm is the location of the dividing hyperplanes that partition conformation space into the resulting clusters. For our application of clustering conformations into various macrostates, we would like to have the natural free energy barriers between free energy minima divide conformation space. *K*-means has no information about the underlying free energy landscape of our molecular system, so the dividing hyperplanes could very well partition conformation space at unnatural locations, especially when the number of clusters chosen is not the true number of clusters in the data set. Given these drawbacks to using the *K*-means algorithm to cluster our data, we ensured that the final clustering results we used for the subsequent kinetic analysis were sufficiently accurate, as discussed in the following section.

III. STRUCTURAL CHARACTERIZATION

A. Mean squared error (MSE)

As mentioned previously, *K*-means clustering minimizes the total within-cluster dissimilarity of the data set. This is measured to be the mean squared error, where the error is defined to be the distance from a given conformation, represented as the distance matrix $\mathbf{d}^i$, to the average distance matrix of the cluster to which the given conformation belongs,

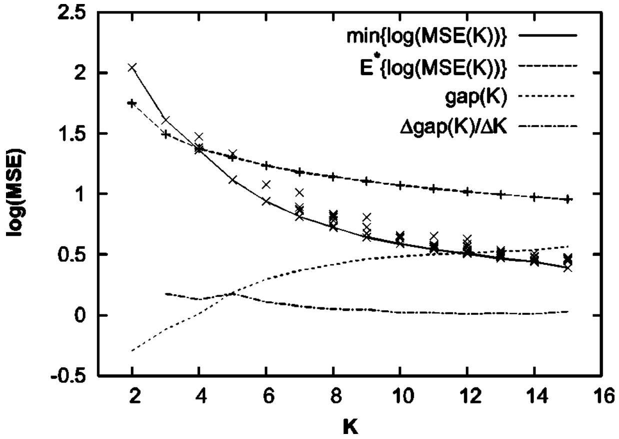


FIG. 1. The gap statistic helped identify the optimal number of clusters $K^* = 10$ in the data set. The gap(K) did not show a maximum, so we chose K^* to be the value of K at which gap(K) increased at a minimal rate, as shown by Δgap(K)/ΔK.

$$\mathrm{MSE}(K) = \frac{1}{2N}\sum_{i=1}^{N} \mathrm{drms}(\mathbf{d}^i, \overline{\mathbf{d}}^{C(i)})^2, \tag{2}$$

where $C(i) = k$ for $k = 1,...,K$ which specifies the cluster to which conformation i belongs, and N is the total number of conformations in the data set. Since K-means does not guarantee the global minimum to MSE(K), multiple trials are needed. The trial which gives the lowest MSE(K) as its solution is selected as the best clustering solution for that value of K. Figure 1 shows the results of the clustering for the 12-mers; the solid curve illustrates the minimum ln[MSE(K)] as we scan values for K. For each value of K, ten trials were initiated (the solution for each trial is represented by an ×). For a given trial, the initial cluster centers were chosen at random from the N conformations in the data set. We were able to reduce the representation of the conformations from an all-atom model (130 atoms) to a representation containing a single sphere for each residue (12 residues). This was accomplished by representing the position of a given residue as the centroid of its phenyl ring, ignoring the side chain, and acetylene linker atoms when calculating the centroid. The MSE(K) was subsequently minimized using the modified K-means algorithm with the residue-residue drms between two conformations as the pairwise dissimilarity measure.

B. Gap statistic

To calculate the gap statistic (see Methods section), we generated ten uniform distributions of 50 000 random conformations each. Generating a random conformation is not a trivial task. Unlike proteins which have two backbone degrees of freedom, well known as the dihedral angles ϕ and ψ, pPA has only one backbone degree of freedom, which is the dihedral angle between the planes spanned by adjacent phenyl rings. This simplifies the problem due to the fact that we only need to generate one random dihedral angle on the uniform unit circle per residue rather than two random dihedral angles per residue. However, it is very rare to obtain compact random conformations by taking random dihedral angles from a uniform unit circle. It is important that we generate compact random conformations in order to ensure good overlap between the uniform distributions and the real data set; the majority of the conformations are compact due to the solvophobic forces which cause the residues in distant parts of the chain to attract each other. So we need to use more sophisticated methods to generate a more representative uniform random distribution.

Tibshirani[22] suggests that one should generate random samples by sampling over the range of principal components of the data set. This is what we endeavored to do. However, there are issues that need to be considered when applied to polymers. First of all, the most common way to calculate the principal components of a polymeric system is to calculate the covariance matrix of the coordinates aligned relative to the average conformation (instead, we used the native conformation as the reference conformation, since the average conformation is pathological itself). The transformation matrix obtained from diagonalizing this coordinate covariance matrix, however, does not preserve bond lengths or bond angles. This is because principal component analysis[23] is a linear transformation, but polymer conformations are inherently nonlinear in the coordinate frame of reference. Therefore, transforming a conformation represented by the coordinates into principal component space moves the conformation off the underlying manifold.[24–26] One may sample uniformly over the range of the principal components, but upon back transforming this random sample into conformation space, obtain samples that are nowhere near representative of a typical conformation in the original data set, due to stretched or contracted bond lengths and bond angles.

We overcome this problem by reparameterizing conformation space in terms of dihedral angles. We then calculate the principal components by diagonalizing the covariance matrix of dihedral angles, aligned relative to the native conformation. This approach preserves bond lengths and bond angles and keeps the conformations on the manifold. In addition, it gave random conformations upon back transformation that were fairly compact. Again, however, there are issues with this approach in that the periodicity of the dihedral angle is not taken into account. For example, two conformations which differ only by the dihedral angle at a single residue, one with a dihedral angle of 0.1° and the other with a dihedral angle of 359.5°, would be very close to each other in an absolute sense. But, since the periodicity is not taken into account, the two close conformations would in fact be very far apart in principal component space since the distance between the two conformations would be 359.4° rather than 0.6° (values would need to be converted to the units in the principal component space). Despite this minor problem, we still used the principal components over the dihedral angles to generate the random uniform distributions to calculate the gap statistic.

We calculated the average, denoted $E^*\{X\}$, and the standard deviation for the ln[MSE(K)] from the results of the ten uniform random distributions scanned over K. These are also shown in Fig. 1, where the standard deviations are smaller than the objects (+) representing the average. The gap sta-

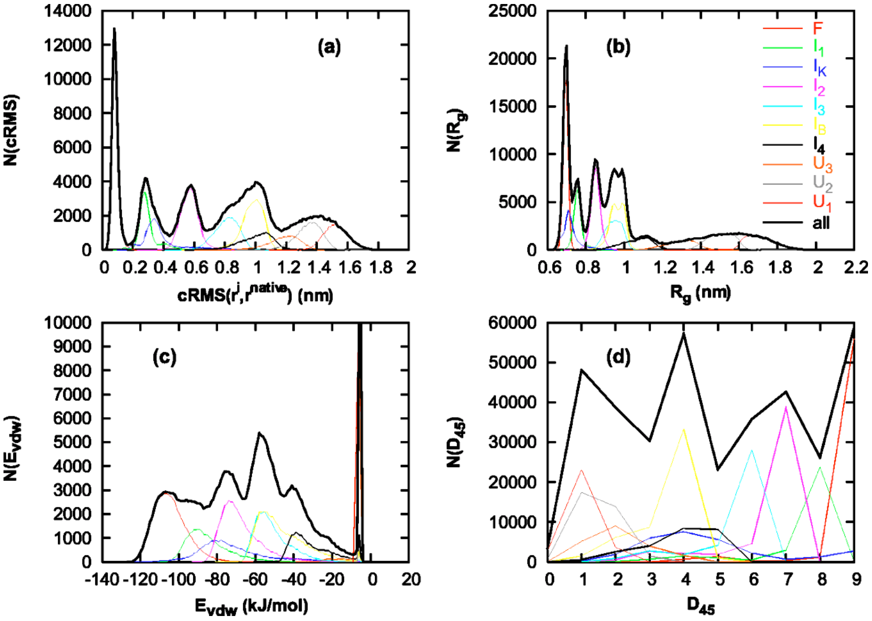


FIG. 2. Distributions for various conformational structural properties of individual clusters and the full data set. (a) crms: the distribution for the full data set shows that crms is capable of distinguishing the folded ensemble from all other conformations in the data set. (b) R_g. (c) E_{vdw}. (d) D_{45}. Each distribution is unimodal which demonstrates that the solution to the K-means clustering calculation accurately partitions conformation space into ten distinct clusters.

tistic is then the difference between the ln[MSE(K)] of the uniform distributions and the true data set,

$$\mathrm{gap}(K)=E^*\{\ln[\mathrm{MSE}(K)]\}-\ln[\mathrm{MSE}(K)]. \qquad (3)$$

The gap(K) unfortunately does not exhibit a clear maximum so it becomes more difficult to determine the true number of clusters in the 12-mer data set. We determined that the optimal K^* is the value of K at which there is a minimal increase of the gap. We show this with the curve for the slope of the gap(K), i.e., $\Delta\mathrm{gap}(K)/\Delta K$. We see that this curve approaches zero and plateaus at roughly $K^*=10$. Therefore, the solution for the cluster assignments of the 12-mer data set is taken to be the trial with the minimum MSE(10). (There happened to be four trials out of ten that all converged to this solution, which is a further validation that this is the proper solution. For $K\geqslant 11$, there were no trials at any given K that had multiple trials converge to the same solution.)

C. Structural characterization of individual clusters

In lieu of the disadvantages of the K-means clustering method discussed previously, we want to ensure that the partitions that divide the conformations into clusters are placed in relatively accurate places. K-means treats the data set as a collection of objects with no temporal relationships between those objects. The cluster to which a given conformation is assigned is solely determined by the position in conformation space relative to all other members of the set. As mentioned earlier, K-means has the potential to partition conformation space at unnatural locations; in order to counteract this weakness, we seek to make corrections to the cluster assignments by using the information contained in the dynamics of the individual trajectories. We emphasize that this is only a small perturbation to the solution of the K-means calculation and is only a valid approach when the partitioning of the solution is close to the natural free energy barriers of the underlying free energy landscape. We assume that we are close to this regime, as the gap statistic has helped us reach a solution that gives an accurate partitioning of conformation space, evidenced by the fact that we are at or close to the true number of clusters in the data set and multiple trials gave the same minimum solution.

We defined two criteria by which a conformation may be reassigned to a cluster different from the one assigned by K-means.

(1) We know the true folding time for each of the trajectories. Therefore, if K-means assigned any conformation in the trajectory to the folded cluster before the true folding time, that conformation was reassigned to its next closest cluster center.

(2) Due to the diffusive nature of the folding process, a trajectory will dwell in a free energy minimum until a random thermal fluctuation kicks it into another free energy minimum state where it will dwell for another extended period of time. Therefore, we would expect to see long stretches of conformations that are in the same cluster, followed by long stretches of conformations that are in another cluster. If the clustering partition is placed in a poor location, say through the middle of a free energy minimum or away from the free energy saddle point between free energy minima, we may see noise in the cluster assignments as the trajectory moves back and forth across the poorly placed partition. To assist in eliminating this noise, we create a moving window of 500 ps (11 conformations) within the given trajectory. If all conformations are in the same cluster except the one in the middle of the window, then that anomalous conformation is reassigned to be in the same cluster as all the other conformations in the window.

Using this filtering scheme, ~10 000 conformations (3%) were reassigned. Of these, half were knot conformations (cluster $\mathbf{I_K}$, see Fig. 2) that were incorrectly assigned to the folded cluster. For the subsequent cluster characterization and kinetic analysis, we use these filtered cluster assignments to classify the conformations in the data set.

We verify that our final clustering solution is accurate.

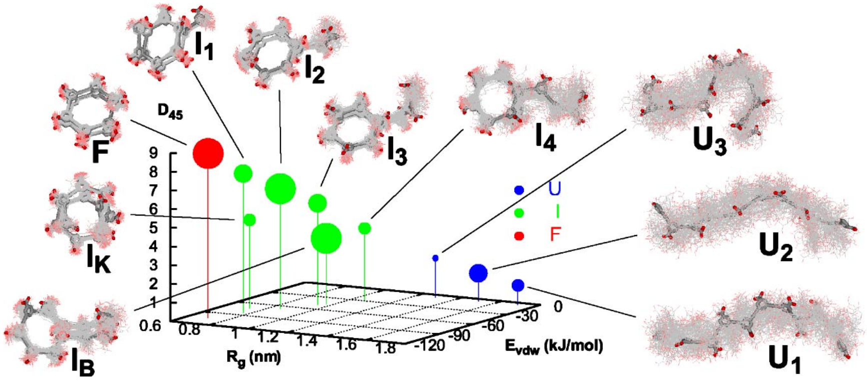


FIG. 3. Average conformational structural properties for the ten clusters in the data set. The size of the points represents the relative populations of each cluster. The raw data used to create this figure are tabulated in Table I. A NMR representation of each cluster is included to illustrate the general shape and flexibility of a conformation in each cluster. The conformation that is closest to the cluster average is shown in stick representation, while the closest 100 conformations are represented with lines.

We do this by looking at the distributions of various structural properties within each cluster (see Fig. 2): crms($\mathbf{r}^i,\mathbf{r}^{native}$), E_{vdw}, D_θ, and R_g. The crms measures the global structural similarities/differences between two structures. It is useful for measuring properties analogous to tertiary structure in biomacromolecules. In particular, crms($\mathbf{r}^i,\mathbf{r}^{native}$) is able to distinguish the native state ensemble from all other conformations in the data set. The van der Waals potential energy E_{vdw} is a measure of the number of contacts between nonlocal residues in the chain. Please note this is not the total potential energy; we have not included the bonding energies in this measure of potential energy. These contacts are not required to be native contacts. Indeed, as explained in previous work,[3] the pPA molecules that we are studying are homopolymers; hence, there is no energetic preference toward native contacts. The number of consecutive helical residues D_θ has also been described in previous work,[3] and is a measure of local structure, analogous to secondary structure in biomacromolecules. We classify a residue to be in a local helical state if the dihedral angle between adjacent phenyl rings is *cis*, where we define a *cis* dihedral angle to be a dihedral angle within ±45° from coplanar relative to each other. Hence, we will use D_{45} for the remainder of this paper. We also interpret D_{45} to represent the helical content of a conformation. The radius of gyration R_g is a measure of the compactness of the chain. Since the native state of pPA has a hollow core, the native state should have a higher radius of gyration than conformations in trapped states, which are typically compact with no regular structure.

If any of the clusters' structural parameter distributions are not unimodal, we have good reason to believe that the dividing hyperplanes that partition conformation space were assigned to poor locations by the *K*-means algorithm. Examples of situations that would produce multimodal distributions are: the clustering algorithm failed to divide two clusters, or a dividing hyperplane cut through the center of a cluster. Inspection of the distributions in Fig. 2 reveals that all of the clusters are very well distinguishable and are unimodal, demonstrating that the clustering solution is a reasonable partitioning of conformation space. The distribution for the full data set is also shown for reference.

The average structural properties of the conformations within each cluster were used to characterize the cluster as a whole. Figure 3 shows a 3D plot of the structural parameters R_g, E_{vdw}, and D_{45} for each of the ten clusters. Table I gives the same data in tabular form. Also shown for each cluster is an NMR representation of the conformation which is closest to the average structure (stick representation) with the 100 nearest conformations to this average structure (line repre-

TABLE I. Average structural and kinetic parameters calculated to identify unfolded clusters, folded clusters, and intermediates; shown visually in Fig. 3. The two intermediate clusters with low helical content $\mathbf{I_B}$ and $\mathbf{I_K}$ are shown separate from the other clusters to emphasize the peculiar structural features inherent in these clusters compared to the other clusters. Pop gives the population of conformations in the given cluster. Since each trajectory was terminated as soon as the folded state was reached, the population of cluster F is much lower than would be expected from an equilibrium population. We show two p_{fold} values and three mfpt values in order to compare methods for calculating these values.

Cluster	Pop	crms (nm)	E_{vdw} (kJ/mol)	D_{45}	R_g	p_{fold}[a]	p_{fold}[c]	mfpt[a] (ns)	mfpt[b] (ns)	mfpt[c] (ns)
$\mathbf{U_1}$	29 469	1.53	−5.10	1.0	1.71	0	0	7.70	5.55	7.31
$\mathbf{U_2}$	33 382	1.35	−5.11	1.5	1.51	0	0.14	7.46	5.18	7.23
$\mathbf{U_3}$	19 909	1.19	−6.33	2.1	1.30	0.01	0.27	7.33	4.88	6.99
$\mathbf{I_4}$	23 745	1.01	−29.96	3.9	1.07	0.38	0.57	6.66	4.15	6.32
$\mathbf{I_3}$	37 993	0.82	−50.40	5.5	0.95	0.81	0.73	5.46	3.59	5.80
$\mathbf{I_2}$	50 508	0.58	−67.22	6.4	0.86	0.98	0.85	3.76	2.33	4.40
$\mathbf{I_1}$	30 698	0.30	−83.99	7.4	0.76	0.99	0.92	2.12	0.96	2.45
F	59 876	0.09	−103.86	8.8	0.70	1	1	0.98	0.96	0
$\mathbf{I_B}$	50 744	0.99	−45.95	3.5	0.97	0.52	0.67	6.79	4.60	6.38
$\mathbf{I_K}$	28 196	0.39	−72.31	4.7	0.73	0.84	0.82	4.41	2.49	4.38

[a] Brute force method, first instance of state *i*.
[b] Brute force method, last instance of state *i*.
[c] Markovian state model.

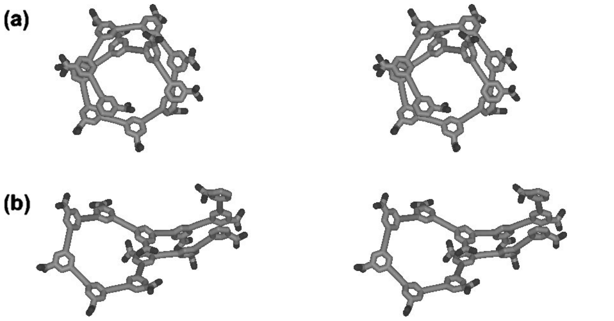


FIG. 4. Stereo view of clusters (a) $\mathbf{I_K}$ and (b) $\mathbf{I_B}$. The conformations shown are the conformations which are the nearest to the average conformation for the respective clusters. These two clusters are implicated as the source of multistate kinetics due to topologically unique structural properties which produce an observable kinetic intermediate macrostate.

sentations). We see a nice steady progression of increased helical content D_{45} (vertical axis) within the clusters as a typical folding trajectory progresses from the unfolded state; clusters $\mathbf{U_1}$, $\mathbf{U_2}$, and $\mathbf{U_3}$; through the intermediate state; clusters $\mathbf{I_4}$, $\mathbf{I_3}$, $\mathbf{I_2}$, and $\mathbf{I_1}$; and finally to the folded state; cluster **F**. This implies that a simple mechanism for folding can be described as a sequential propagation of the nucleus of the helix as more and more residues adopt a *cis* dihedral configuration. We pursue this idea of identifying the folding mechanism in the section on kinetics.

There are, however, two clusters that fall outside this regime of sequential helix propagation, which are illustrated in Fig. 4. They are primarily distinguishable by their low helical content compared to clusters with similar R_g and E_{vdw}. While these cannot formally be classified as traps in the context of our kinetic analysis to be shown hereafter (we show that there are no kinetic traps for the 12-mer data), they are the seeds for kinetically trapped conformations in longer chains which do exhibit kinetic traps,[27] as suggested by preliminary studies of a pPA 20-mer. In particular, cluster $\mathbf{I_B}$ has the form analogous to a β-turn/sheet in protein secondary structures. In order to get back on the folding pathway of sequential helix propagation, there must be several non-native contacts broken followed by a rearrangement of the ends of the chain about the turn region. This would likely become a trapped state as the ends get longer and more and more contacts would need to be broken to properly fold.

The other cluster $\mathbf{I_K}$ with low helical content could be described as a knot. Upon inspection of the structure of the average conformation of this cluster (see Fig. 4), one would quickly guess that it truly is a trapped state. Apparently, however, the knotted conformation can become disentangled quickly enough with respect to the time scale of folding that the conformation is not kinetically trapped. This brings up an interesting question: how does a chain fold from this knotted cluster? Does it unfold and restart along the folding intermediate pathway, or does it fold by some other means?

Approximately 40% of all the folding trajectories spent significant time ($>$100 ps) in the knotted conformations. Of these, the majority of the trajectories ($\sim$90%) that became knotted were able to fold without completely unfolding. Those which did unfold essentially started over to retry the folding process again, as though the chain were initiated from the unfolded state. 25% of the trajectories that became entangled in knotted conformations were able to fold directly from the knotted conformations. Apparently, the compact chain became swollen enough that one end of the knot could slip through the opening in the center of the molecule with the assistance of a thermal fluctuation, enabling the knot to become disentangled. The remaining 65% of the trajectories which became knotted, partially unfolded to a helical intermediate conformation. From there, the chain could reorder between the intermediate clusters and unfolded state before ultimately folding. In longer chains, preliminary studies suggest that these knotted conformations become kinetically trapped, resulting in very interesting kinetic behavior.[27] We will further discuss folding pathways and kinetic properties of the 12-mer system in the section, Kinetics. But an important idea to gain from this structural characterization of the clusters is that the chain can adopt a diverse set of conformations, even for relatively small chains such as the 12-mer. Some of these conformations are the seeds of truly kinetically trapped conformations in longer chains, which are knots and less stable secondary structures.

IV. KINETICS

Now that each conformation has been classified according to its similarities relative to its neighbors, we can look at the kinetic properties of the system as a whole. We use several techniques to characterize the kinetic properties of the pPA 12-mers. In the first method, we make no assumptions about the kinetic model by which the system folds. We calculate the probability of folding p_{fold} (Ref. 28) and the mean first passage time (mfpt) for each cluster by accounting for each conformational state along each trajectory. In the second method, we derive an analytical matrix equation describing the evolution of the concentrations of states over time. This method requires that we choose a kinetic model to describe the folding mechanism. Ideally, we would like to use a kinetic model with as few parameters as possible to capture the essential features of the folding mechanism.

A. p_{fold} and mfpt

A p_{fold} value for each cluster was calculated by observing each conformation in the data set individually. Starting from that conformation and looking forward from within its trajectory, did the trajectory fold or unfold first? If the trajectory folded first, we assign a 1 to that conformation, if the trajectory unfolded first, we assign a 0 to that conformation. Then, the average over all the conformations in a given cluster gives the p_{fold} for that cluster. Table I tabulates the p_{fold} and mfpt for each cluster. From this, we can get a sense of how far or close a given cluster is from the folded state. We see that the clusters that are extended or have no intramolecular contacts ($\mathbf{U_1}$, $\mathbf{U_2}$, and $\mathbf{U_3}$, as evidenced by a high energy E_{vdw}) have low p_{fold} values as would be expected. These clusters are grouped together to define the unfolded macrostate of the system. The cluster of conformations that contains the native ensemble has a p_{fold} of 1 as expected and this single cluster is defined to be the folded macrostate of the system. All other clusters have p_{fold} values intermediate between 0 and 1 and are grouped together to define the intermediate macrostate of the system. It is possible that there are more than three macrostates in the system, in which case, we would have to determine which clusters with intermediate p_{fold} values map to which additional macrostates. However, we show that our system of pPA 12-mers does indeed fit a kinetic model that includes only three macrostates; unfolded, intermediate, and folded states.

Table I includes two calculated brute force values for the mfpt from cluster *i* to the folded state. The first value is the result of averaging the time it takes to reach the folded state from the first instance of entering cluster *i*. For example, if the system starts in cluster $\mathbf{U_1}$, progresses to $\mathbf{I_4}$, but then unfolds back to $\mathbf{U_1}$, before finally folding, the time it took to complete the full circuit, $\mathbf{U_1}\rightarrow\mathbf{I_4}\rightarrow\mathbf{U_1}\rightarrow\mathbf{F}$, is included in the average. The second value only counts the time it took to complete the last phase of the circuit, $\mathbf{U_1}\rightarrow\mathbf{F}$, in the example above. As can be seen, the values for the mfpt in the second case are significantly smaller than the times in the first case, which suggests that the system does not monotonically approach the folded state. If there was a monotonic approach to the folded state, the values in both cases would be equal. Moreover, there can be significant reordering within the intermediate clusters and between the unfolded state, before a trajectory ultimately folds.

An alternative method for calculating an estimate for the p_{fold} and mfpt values is described by Singhal *et al.*[29] Essentially, the method represents a Markovian state model approach to calculate the kinetic properties of a system. We calculate the matrix of transition probabilities between cluster *i* and cluster *j*, P_{ij}. This is accomplished by parsing the full data set and counting the number of transitions from cluster *i* to cluster *j*, normalizing by the number of transitions out of cluster *i*. If the trajectory does not make a transition to a new cluster, we count this situation as a transition from cluster *i* to cluster *i*, which is necessary to ensure a 100% probability of moving from cluster *i* to anywhere, $\Sigma_j P_{ij} = 1$. Once we have the transition probability matrix we can calculate the p_{fold} and mfpt values for each cluster by self-consistently solving the following systems of equations, respectively:

$$p_{fold}(i) = \sum_{all\ j} P_{ij} p_{fold}(j), \tag{4a}$$

$$\mathrm{mfpt}(i) = \sum_{all\ j} P_{ij}[t_{ij} + \mathrm{mfpt}(j)]. \tag{4b}$$

Boundary conditions are set so that $p_{fold}(i) = 1$ if cluster *i* is the folded state and 0 if cluster *i* is the unfolded state; $\mathrm{mfpt}(i) = 0$ if cluster *i* is the folded state. For our system, t_{ij} is constant and equals the time between saved conformations 50 ps. The results for this method are included in Table I. We see that for some clusters, the Markovian estimate for the p_{fold} and mfpt parameters are fairly close to the brute force calculation. Other clusters do not agree as well, but the ordering of the clusters is correct. In general, the Markovian approach is shown to be a useful predictor for kinetic properties of this system.

B. Matrix theory

While the stochastic dynamics algorithm used to generate trajectories guarantees that our system samples from an equilibrium distribution of conformations according to the canonical (NVT) ensemble, it implies nothing about the kinetic properties of the system. In particular, it imposes no constraints on the connectivities or rates of transfer between states. Fundamentally, these are mechanistic issues that must be extracted from the data set.

A kinetic model that describes the folding mechanism involving two or more states separated by energetic barriers can be solved as a system of *n* coupled differential equations.[30,31] The model must obey the law of mass action, which requires that the time rate of change from state *A* to state *B* must be proportional to the concentration in state *A*. The rate constant, or proportionality constant, implied by this law represents the characteristic or average time of an exponential process between those two states, and will hereafter be referred to as the fundamental rate constant k_{BA}. For complex mechanisms, systems with more than two thermodynamic macrostates, the mass transfer between multiple macrostates is coupled together. The solution of the system of coupled differential equations effectively "decouples" these fundamental mass transfer processes into global, concerted kinetic phases described by the phase rate constants λ_i. Their corresponding phase time constants are $\tau_i = -\lambda_i^{-1}$. These rate constants λ_i's are the eigenvalues of the rate matrix constructed from the system of differential equations and are nontrivial functions of the fundamental rate constants.

Here is a general approach to solving any kinetic model of a thermodynamic system using the matrix method. We start from the law of mass action

$$\frac{d\mathbf{x}(t)}{dt} = \mathbf{K}\cdot\mathbf{x}(t), \tag{5}$$

where

$$\mathbf{x}(t) = \begin{bmatrix} c_1(t) \\ \vdots \\ c_n(t) \end{bmatrix}.$$

Next, we construct the matrix of rate constants **K**, which depends on the kinetic model. Finally, the concentrations of each state at any time may be found by solving this system of *n* linear differential equations[32]

$$\mathbf{x}(t) = e^{\mathbf{K}t}\mathbf{x}(0) = \mathbf{R}e^{\Lambda t}\mathbf{R}^{-1}\mathbf{x}(0), \tag{6}$$

where, $\mathbf{x}(0)$ gives the concentrations in each state at $t = 0$, $\mathbf{\Lambda}$ is a diagonal matrix of the eigenvalues $\lambda_1,\ldots,\lambda_n$ of the rate matrix **K**, and **R** is an orthogonal matrix of the corresponding eigenvectors $\mathbf{u_1},\ldots,\mathbf{u}_n$ of **K**. Since matter is neither created nor destroyed, one of the eigenvalues will always be 0; expressed another way, the elements in a given column of **K** must sum to 0 representative of a closed system. Under this formalism, the other $n-1$ eigenvalues are all negative. Hence, the elements of the diagonal matrix $\mathbf{\Lambda}$ in the argument to the exponent will always be less than or equal to 0, representative of typical decay processes. In addition, to minimize the complexity of the kinetic model, we make the assumption that all eigenvalues are nondegenerate $\lambda_i \neq \lambda_j$. The case where eigenvalues are allowed to be equal is not covered here, but may be found in more advanced texts.[32]

Equation (6) may be written in a more useful form

$$\mathbf{x}(t) = [\mathbf{u_1} \cdots \mathbf{u}_n] \begin{bmatrix} e^{\lambda_1 t} & 0 & \\ 0 & \ddots & 0 \\ & 0 & e^{\lambda_n t} \end{bmatrix} \mathbf{R}^{-1}\mathbf{x}(0)$$

$$= a_1 e^{\lambda_1 t}\mathbf{u_1} + \cdots + a_n e^{\lambda_n t}\mathbf{u}_n = \sum_{j=1}^{n} a_j e^{\lambda_j t}\mathbf{u}_j . \qquad (7)$$

The coefficients a_j are determined by the initial conditions through the relationship $\mathbf{a} = \mathbf{R}^{-1}\mathbf{x}(0)$. Therefore, the equation for the concentration of state i at any time t may be written,

$$c_i(t) = \sum_{j=1}^{n} a_j e^{\lambda_j t} u_{ij}, \qquad (8)$$

which is a linear combination of independent kinetic phases, with time constants $\tau_j = -\lambda_j^{-1}$ as mentioned previously. We give a practical example of the use of this matrix method applied to an irreversible two-state kinetic model in Appendix A.

In the case of a two-state system, as described in Appendix A, the parameter k is the rate constant for an exponential process. The reciprocal of k, $\tau = k^{-1}$ is the characteristic time constant for the exponential process, and in the folding community is known as the folding time. Experimentally reported folding times are typically values extracted from data fit to a two-state model, as shown in Appendix A. This raises the questions of how to characterize the folding time of a system that clearly does not follow two-state behavior? And how do we compare two systems when one is a fit to a two-state model, while the other includes additional states in the mechanism?

The mean first passage time describes the average time taken to reach the folded state, given that the trajectory was initially in the unfolded state. From the kinetic model using the matrix method, we are able to derive an analytical expression for the concentration of the folded state as a function of time [see Eq. (8), where $i = F$]. This also represents the cumulative folding time distribution when the concentrations are expressed in mole fractions. The mean first passage time may be determined from the cumulative folding time distribution according to the integral equation[29] mfpt $= \int_0^\infty t(dc_F(t)/dt)dt$. One may then show that the mfpt is a weighted sum of the inverse of the phase rate constants $\lambda_j < 0$,

$$\text{mfpt} = \sum_{j=2}^{n} \frac{b_j}{\lambda_j}, \quad \text{where} \quad \sum_{j=2}^{n} b_j = -1 \qquad (9)$$

and we have set $b_j = a_j u_{Fj}$. The term with the eigenvalue equal to 0, λ_1 drops out of the sum upon differentiation, avoiding the singularity and physically corresponds to the fact that we are modeling a closed system. In the limit of a two-state kinetic model, the mean first passage time reduces to the characteristic folding time described earlier, i.e., mfpt $= \tau = -\lambda^{-1}$. Thus, we may characterize the folding time scale based on the mean first passage time, without regard to the underlying kinetic mechanism. The question of how to compare two systems, one fit to a two-state model and the other multistate, is still difficult to address. We will discuss this topic further in the Discussion section where we compare the experimentally measured folding rate with our simulations.

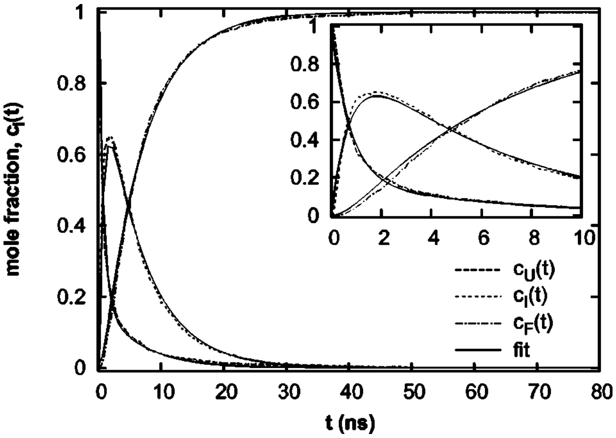
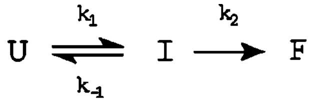


FIG. 5. Kinetic traces of the mole fraction of the three macrostates in the system. Also shown are the best fit curves to the kinetic model shown in scheme 1. The inset shows the short time behavior of the system; most notably, we see a distinct lag time that rises quadratically at $t < 1$ ns. The quantitative measure for goodness of fit is $\overline{\chi}^2 = 6.2 \times 10^{-5}$.

C. Kinetic mechanism

At this point in our analysis we have identified, with the assistance of p_{fold} and mfpt calculations, which clusters comprise the three macrostates of the system: unfolded, intermediate, and folded states. Figure 5 shows the kinetic trace of the mole fraction $c_i(t)$ for each state and the best fit curves to the kinetic model,



Scheme 1

containing three free parameters: k_1, k_{-1}, and k_2. This kinetic model is the minimalist kinetic representation that accurately includes the essential features of the dynamic system. In particular, from the inset to Fig. 5, which shows the short time kinetics, we see that there is a distinct lag time in the formation of folded chains. At a minimum, our kinetic model must include an intermediate state, since there is no lag time associated with a two-state system.[33] Moreover, there exists an equilibrium between the unfolded state and the intermediate state, which is needed in order to observe such a relatively high population of unfolded chains at long time periods, and to account for the reordering within intermediate clusters and between the unfolded state as mentioned previously. Therefore, we see that this kinetic model contains the essential kinetic features of the system. By fitting the data to this model, we can extract the kinetic parameters such as elementary rate constants for transfer between states. Then, from the elementary rate constants, we can calculate all other kinetic properties of the system determined by the kinetic model, including the mean first passage time, which we can use to check for internal consistency between the data and the kinetic model.

The rate matrix for the kinetic model shown in Scheme 1 is

$$\mathbf{K}=\begin{bmatrix} -k_1 & k_{-1} & 0 \\ k_1 & -(k_{-1}+k_2) & 0 \\ 0 & k_2 & 0 \end{bmatrix}, \quad \mathbf{x}(t)=\begin{bmatrix} c_U(t) \\ c_I(t) \\ c_F(t) \end{bmatrix}. \tag{10}$$

According to this kinetic model, the general solution for the equations for the concentrations for each of the states *U*, *I*, and *F* with all trajectories initially in the unfolded state $\mathbf{x}(0)=[1\ 0\ 0]^T$, where $\mathbf{x}(0)$ is a column vector (denoted by the superscript *T*), using the matrix method is

$$c_U(t)=\frac{\lambda_{slow}+k_{-1}+k_2}{\lambda_{slow}-\lambda_{fast}}e^{\lambda_{slow}t}-\frac{\lambda_{fast}+k_{-1}+k_2}{\lambda_{slow}-\lambda_{fast}}e^{\lambda_{fast}t}, \tag{11a}$$

$$c_I(t)=\frac{k_1}{\lambda_{slow}-\lambda_{fast}}(e^{\lambda_{slow}t}-e^{\lambda_{fast}t}), \tag{11b}$$

$$c_F(t)=1+\frac{\lambda_{fast}}{\lambda_{slow}-\lambda_{fast}}e^{\lambda_{slow}t}-\frac{\lambda_{slow}}{\lambda_{slow}-\lambda_{fast}}e^{\lambda_{fast}t}, \tag{11c}$$

where

$$\lambda_1=0, \tag{12a}$$

$$\lambda_{fast}=-\tfrac{1}{2}(k_1+k_{-1}+k_2+b), \tag{12b}$$

$$\lambda_{slow}=-\tfrac{1}{2}(k_1+k_{-1}+k_2-b), \tag{12c}$$

and

$$b=\sqrt{(k_1+k_{-1}+k_2)^2-4k_1k_2}. \tag{12d}$$

The kinetic modeling program Gepasi[34,35] (http://www.gepasi.org) was used to calculate the best fit parameters $\hat{k}_1$, $\hat{k}_{-1}$, and $\hat{k}_2$ by simultaneously fitting the three curves for the concentrations of states *U*, *I*, and *F*, see Fig. 5. A nonlinear least squares fit using the Levenberg–Marquardt algorithm as implemented in Gepasi produced the following parameters for the best fit equations:

$$\begin{aligned}\hat{k}_1&=1.106\pm0.003\ \ \text{ns}^{-1},\\ \hat{k}_2&=0.1844\pm0.0002\ \ \text{ns}^{-1},\\ \hat{k}_{-1}&=0.184\pm0.001\ \ \text{ns}^{-1}.\end{aligned} \tag{13}$$

Substituting the values for these best fit parameters into the equations for the phase rate constants in Eqs. (12a)–(12d) results in the following kinetic parameters:

$$\begin{aligned}\lambda_{slow}&=-0.1545\pm0.0002\ \ \text{ns}^{-1},\\ \lambda_{fast}&=-1.320\pm0.003\ \ \text{ns}^{-1}.\end{aligned} \tag{14}$$

We see that there is a clear separation of time scales between these two kinetic phases, where the fast phase exhibits folding events an order of magnitude faster than the slow phase. The mean first passage time for the folding of the 12-mer chains, extracted from this ensemble of folding trajectories, is calculated from Eq. (9) to be $\tau=7.23\pm0.01$ ns which is very close to the mfpt calculated from the Markovian method, mfpt=7.31 ns. It is interesting to observe that these results from matrix theory are closer to the Markovian value than the brute force value for the mfpt=7.7 ns (which also happens to be in good agreement). Therefore, we have demonstrated that the three-state mechanism shown in scheme 1 is consistent with the data.

The kinetic mechanism shown in Scheme 1 is widely known to be one of the simplest mechanisms[33] exhibiting an initial lag time in the formation of folded chains (see inset, Fig. 5). The behavior at very short times provides invaluable information for understanding the mechanism of folding.[33] There is obvious curvature to the plot of $c_F(t)$ at $t<1$ ns (see Fig. 5, inset). A Taylor series expansion around $t=0$ gives a polynomial in δt.

$$c_F(\delta t)\big|_{t=0}=\tfrac{1}{2}k_1k_2\delta t^2+O(\delta t^3). \tag{15}$$

Plotting the data at short times on a ln-ln plot gives a straight line with a slope equal to the exponent of the highest order term in the Taylor expansion. The exponent is calculated to be $\nu=2.3$. For the purposes of this paper, we truncate the value of ν to the nearest integer $\nu=2$ implying that the curve for $c_F(t)$ should rise quadratically at short time scales, which is indeed the case as shown in Eq. (15). Therefore, the information contained in the Taylor expansion further supports the folding mechanism shown in scheme 1.

V. COMPARISON BETWEEN SIMULATION AND EXPERIMENTS

We have shown that the kinetic behavior of the pPA 12-mer is well described by a three-state system with a kinetic intermediate. The chain collapses relatively quickly to an intermediate conformation, i.e., $k_1>k_{-1}$, k_2, after which it can either progress toward the folded state or revert back to the unfolded state with equal probability, i.e., $k_{-1}\approx k_2$. We have also used three separate methods to calculate the mean first passage time with the three methods giving folding times on the order of 7 ns.

The experimental folding time at 300 K was measured to be ≈160 ns via laser *T* jump.[36] Given the difficulty in interpreting the experimental results of the kinetic measurements of the folding time for the pPA 12-mer, it is hard to make a direct comparison between simulation and experiment. Indeed, to what effects may we attribute the discrepancy between the experimental folding time, measured to be ~160 ns, and the simulation folding time calculated to be ~7 ns? If one corrects for the low viscosity used in the simulations in a linear way,[37] then one would multiply the time by a factor of 50 (see Appendix B), resulting in a prediction of 350 ns. Since we ran our simulations at a viscosity nearly 1/50th the viscosity of a 1:1 mixture of THF/methanol (the solvent used in the experimental measurements), we are below the linear regime (>1/10th experimental solvent viscosity) precluding the use of a linear scaling relationship between the viscosity and the folding time. Therefore, we need to correct the multiplication factor in a nonlinear way (see Appendix B) to obtain a scaling factor of 10. This gives a prediction of 70 ns for the folding time. By scaling the simulation folding time to account for the low viscosity, we now have a lower and upper bound for our prediction for the folding time: 70 ns$<\tau<$350 ns, depending on whether we scale linearly or nonlinearly. This shows a reasonable agreement between our simulations and the experimental folding time.

There are several reasons for the lack of a more complete agreement. First, the computational model employed here is relatively simple and it is interesting to consider whether an explicit representation for the solvent may have a significant effect on the kinetics, such as via the drying effect.[38] Also, the experimental data was fit to a two-state model, whereas our simulations suggest more complex dynamics are involved. One additional reason for the discrepancy could be attributed to the two-state experimental interpretation, which we will discuss, next.

Yang *et al.*[36] have measured the folding rate of a pPA 12-mer using laser *T* jump. They used a two-state approximation to estimate the folding and unfolding rates as a function of temperature. They found, however, that this two-state approximation broke down at lower temperatures (below ~315 K), so they resorted to a lattice model to accurately interpret the folding kinetics of the 12-mer. We can still use the results of their two-state approximation to attempt to make a direct comparison between our simulations and their experimental measurements using the comparison methods outlined in this paper. In particular, the experimental folding time measured at 300 K, which again was obtained from a two-state approximation, was ~160 ns. In the process of extracting the kinetic parameters for the system from a two-state kinetic model, some extreme approximations are necessary.

First of all, by assuming that the system is two-state, the polymer chains can only exist in either the folded or unfolded state. When a measurement is taken via *T* jump the system is initially at equilibrium. Then, the system is perturbed by a laser pulse which heats the sample nearly instantaneously (compared to the time frame for folding). A new equilibrium is established and the decay rate of this equilibration process is measured by fluorescence at 350 nm. This wavelength is the monomer fluorescence band which predominates in the unfolded state.[5,36] Coincidentally, the folded state exhibits a quenched monomer band, with the appearance of a broad excimer band centered around 420 nm. The two-state approximation requires one to interpret the fluorescence signal as a linear combination of these two extremes only. Therefore in the context of a two-state model, as the signal decays upon heating, one would envision a given mole fraction from the folded state completely unfolding, hence increasing the intensity of the fluorescence at 350 nm.

We suggest an alternative way to interpret the signal measured from the equilibration process. The system, initially at equilibrium prior to the laser pulse which increases the temperature, has a significant population of chains in intermediate conformations, especially at intermediate temperatures, i.e., 300 K. Hence the decay signal most likely measures the rate of the transition from partially folded chains to either less partially folded chains or completely unfolded chains. This is the underlying process which manifests itself in the nonexponential decay profiles as reported in the experimental measurements. The decay rate extracted from the two-state model most likely represents the slowest phase of the equilibration process, which would be the complete unfolding of the chain from intermediate conformations. In this sense, the rate constants extracted from the two-state approximation of the decay process may very likely be the complete unfolding rates. Due to the nature of this system, the kinetics may be complex to characterize, and we agree with the assessment that the experimental folding rate of the 12-mer at 300 K occurs on the submicrosecond time scale; the value of 160 ns, therefore, represents a very rough estimate for the experimental folding time, and is sufficient for comparison purposes.

Despite the combination of low viscosity and simple computational model in our simulations, and the potential difficultly in interpreting the experimental rate measurements, we still have gained valuable insight into the folding behavior of a poly-phenylacetylene 12-mer due to a reasonable agreement between the folding times as measured by simulation and experiment.

VI. CONCLUSIONS

We have illustrated the use of several methods for analyzing the kinetic properties of a dynamic polymeric system. The model system studied in this paper is the nonbiological polymer pPA, 12-mer. Perhaps the most important step in the kinetic analysis is the conformational classification scheme. We used the *K*-means clustering algorithm modified to use the drms dissimilarity metric to identify and group similar structures together. We identified a topologically diverse collection of conformations from which we were able to extract a simple kinetic model to describe the bulk folding properties.[39] In particular, a simple mechanism for helix formation proceeds via a kinetic intermediate state which is a topologically diverse collection of knots, turns/sheets, and partially folded helical conformations, much like those expected from helix-coil theory.[40–42] However, helix-coil theory predicts a cooperative, two-state kinetic mechanism for helix formation. Even though intermediate conformations exist, they are either short lived or too high in energy to be visible experimentally in systems that are well described by helix-coil theory.[43]

We reconcile this apparent contradiction by noting that the intermediate state for the pPA 12-mer is much more structurally and topologically diverse than simple models used to derive the helix-coil theory. The presence of two unique intermediate clusters (see Fig. 4) with structures that are the seeds for kinetic traps requires the inclusion of additional kinetic effects into the underlying theory, rather than simply looking at helix formation as a simple nucleation-propagation mechanism. Most likely, including these same effects in a kinetic analysis of helical proteins and peptide fragments, as well, would lead to a more accurate view of the kinetics of helix formation in protein folding.[43] In contrast, pPA is more rigid than proteins, so these additional kinetic effects may only become apparent in rigid polymer backbones, eliminating the need to include them in protein analysis, which is typically two state and well described by helix-coil theory. In either case, the need for more sophisticated theories is apparent, in order to describe the general kinetic properties of helix formation in systems outside the traditional realm of biology.

Two of the intermediate conformational clusters were particularly interesting (see Fig. 4). We found a cluster of knotted conformations $\mathbf{I_K}$, which become kinetically trapped in longer pPA chains[27] as suggested by preliminary studies of a pPA 20-mer chain. The cluster of conformations identified as $\mathbf{I_B}$ contained a turn structure analogous to β turns in proteins. An interesting design problem would be to stabilize this turn structure so that it is sufficiently accessible kinetically and more energetically favorable than the helix structure. Gin and Moore[44] have proposed experimental evidence for a strand structure by incorporating a flexible tether into the pPA backbone. What other simple structural patterns may be designed as structural building blocks in order to build a hierarchical design and assembly paradigm utilizing pPA as the polymeric backbone? These complex structures would need to be designed to self-assemble in kinetically reasonable time scales, avoiding trapped states, and adopt fairly rigid native folds. If it were possible to design pPA in this way, the opportunities to design novel molecules in a synthetic environment based on nonaqueous solvents[19] having proteinlike properties[45–48] such as catalysis, signal transduction, or regulation could be vast.

ACKNOWLEDGMENTS

The authors would like to thank Mark Engelhardt, Nina Singhal, Bojan Zagrovic, Eric Sorin, and Afra Zomorodian for helpful discussions on kinetic models and experimental interpretation of rate measurements. This work was supported by the Center on Polymer Interfaces and Macromolecular Assemblies (CPIMA) as part of the NSF Materials Science and Engineering Center program under Grant No. DMR 9808677.

APPENDIX A: SIMPLE, TWO-STATE KINETICS

The simplest example of a kinetic model of a thermodynamic system is an irreversible, two-state system separated by an energetic barrier. Initially, the system is in state A $\mathbf{x}(0)=[1\ 0]^T$, where $\mathbf{x}(0)$ is a column vector (denoted by the superscript T). From the law of mass action, the mechanism

$$A \xrightarrow{k} B$$

gives the system of differential equations

$$dc_A(t)=-kc_A(t), \tag{A1a}$$

$$dc_B(t)=kc_A(t), \tag{A1b}$$

which may be written in matrix form

$$\frac{d}{dt}\begin{bmatrix} c_A(t) \\ c_B(t) \end{bmatrix}=\begin{bmatrix} -k & 0 \\ k & 0 \end{bmatrix}\begin{bmatrix} c_A(t) \\ c_B(t) \end{bmatrix}. \tag{A2}$$

The solution to the system of equations is

$$\begin{bmatrix} c_A(t) \\ c_B(t) \end{bmatrix}=\mathbf{R}e^{\mathbf{\Lambda}t}\mathbf{R}^{-1}\mathbf{x}(0)$$
$$=\begin{bmatrix} 0 & -1 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & e^{-kt} \end{bmatrix}\begin{bmatrix} 1 & 1 \\ -1 & 0 \end{bmatrix}\begin{bmatrix} 1 \\ 0 \end{bmatrix}=\begin{bmatrix} e^{-kt} \\ 1-e^{-kt} \end{bmatrix}. \tag{A3}$$

Here $\lambda_1=0$ and $\lambda_2=-k$ while the corresponding eigenvectors are the column vectors of matrix $\mathbf{R}$:$\mathbf{R}=[\mathbf{u_1}\,\mathbf{u_2}]$ where, $\mathbf{u_1}=[0\ 1]^T$ and $\mathbf{u_2}=[-1\ 1]^T$. For a two-state system initially in state A, the fraction of the system that will randomly cross over the energy barrier into state B, or ‘‘fold,’’ is $c_B(t)=1-\exp(-kt)$. The longer one waits, the greater the probability that a single system will be found in state B as a result of a random thermal fluctuation ‘‘bumping’’ the system over the energy barrier, resulting in a higher concentration of products as time increases. This is accompanied by a corresponding decrease in the concentration of reactants $c_A(t)=\exp(-kt)$. Since this is an irreversible process, mathematically, it is equivalent to a system with absorbing boundary conditions and demonstrates first-order kinetics.

APPENDIX B: VISCOSITY CORRECTION

TABLE II. Viscosity (in centipoise) of various solvents at 20 °C needed to correct the simulation folding time due to a low viscosity in the simulations.

Solvent	γ (cp)
H_2O[a]	1.002
MeOH[a]	0.597
THF[b]	0.48

[a] Reference 49.
[b] Reference 50.

The simulation folding time may be scaled due to the fact that the simulations were run at a very low viscosity. We would like to make an estimate of the simulation folding time when the viscosity is close to that of the experimental system. Yang *et al.*[36] used a 1:1 by volume mixture of THF/methanol solvent to perform the kinetic experiments on the pPA 12-mer. Table II shows the viscosities for water, methanol, and THF at 293 K. Assuming a linear relationship (a poor approximation, but sufficient for this back-of-the-envelop calculation) for the viscosity of the THF/methanol mixture, we find that the experimental viscosity of the 1:1 THF/methanol mixture is 0.54 cp. Water has a viscosity of 1.002 cp at 293 K, so the viscosity of the THF/methanol mixture should be roughly 54% that of pure water. In units of ps^{-1}, water's viscosity is 91 ps^{-1}, which translates to a viscosity of roughly 50 ps^{-1} for the THF/methanol mixture. Since our simulations were run at a viscosity of 1 ps^{-1}, we see that the scaling factor is 50, assuming a linear relationship between folding time and viscosity.

Zagrovic and Pande[37] demonstrated that below 1/10th the experimental solvent viscosity, this linear scaling relationship breaks down for a small peptide. Our pPA 12-mer is also small so we assume that the range of linearity is similar to that of their model peptide. Therefore, we use the relationship obtained for low viscosities from their results,

$$\ln\left(\frac{\tau}{\tau_{1:1\,\mathrm{THF/MeOH}}}\right)=0.21\ln\left(\frac{\gamma}{\gamma_{1:1\,\mathrm{THF/MeOH}}}\right)-0.633 \qquad \text{(B1)}$$

in the nonlinear regime. Coincidentally, the nonlinear relationship between the viscosity and folding time is $\tau\sim\gamma^{1/5}$ as demonstrated by the slope of 0.21 in the ln-ln relationship shown in Eq. (B1). Substituting our known quantities into Eq. (B1) and solving for the experimental folding time $\tau_{1:1\,\mathrm{THF/MeOH}}$,

$$\ln\left(\frac{7\ \mathrm{ns}}{\tau_{1:1\,\mathrm{THF/MeOH}}}\right)=0.21\ln\left(\frac{1\ \mathrm{ps}^{-1}}{50\ \mathrm{ps}^{-1}}\right)-0.633 \qquad \text{(B2)}$$

we obtain a prediction for the folding time $\tau_{1:1\,\mathrm{THF/MeOH}}=70\,\mathrm{ns}$ corresponding to a factor of 10.

[1] S. Gellman, Acc. Chem. Res. **31**, 173 (1998).

[2] D. J. Hill, M. J. Mio, R. B. Prince, T. S. Hughes, and J. S. Moore, Chem. Rev. (Washington, D.C.) **101**, 3893 (2001).

[3] S. Elmer and V. Pande, J. Phys. Chem. B **105**, 482 (2001).

[4] J. Nelson, J. Saven, J. Moore, and P. Wolynes, Science **277**, 1793 (1997).

[5] R. Prince, J. Saven, P. Wolynes, and J. Moore, J. Am. Chem. Soc. **121**, 3114 (1999).

[6] R. Cheng, S. Gellman, and W. Degrado, Chem. Rev. (Washington, D.C.) **101**, 3219 (2001).

[7] D. Seebach and J. Matthews, Chem. Commun. (Cambridge) **1997**, 2015.

[8] D. Seebach, M. Brenner, M. Rueping, and B. Jaun, Chem.-Eur. J. **8**, 573 (2002).

[9] R. Zuckermann, E. Martin, D. Spellmeyer *et al.*, J. Med. Chem. **37**, 2678 (1994).

[10] P. Armand, K. Kirshenbaum, A. Falicov, R. Dunbrack, K. Dill, R. Zuckermann, and F. Cohen, Folding Des. **2**, 369 (1997).

[11] M. Egholm, O. Buchardt, P. Nielsen, and R. Berg, J. Am. Chem. Soc. **114**, 1895 (1992).

[12] C. Cho, E. Moran, S. Cherry *et al.*, Science **261**, 1303 (1993).

[13] M. Green, N. Peterson, T. Sato, A. Teramoto, R. Cook, and S. Lifson, Science **268**, 1860 (1995).

[14] M. Soth and J. Nowick, Curr. Opin. Chem. Biol. **1**, 120 (1997).

[15] M. Hagihara, N. J. Anthony, T. J. Stout, J. Clardy, and S. L. Schreiber, J. Am. Chem. Soc. **114**, 6568 (1992).

[16] B. Zagrovic and V. S. Pande, Nat. Struct. Biol. **10**, 955 (2003).

[17] H. J. C. Berendsen, D. van der Spoel, and R. Vandrunen, Comput. Phys. Commun. **91**, 43 (1995).

[18] E. Lindahl, B. Hess, and D. Van Der Spoel, J. Mol. Model. [Electronic Publication] **7**, 306 (2001).

[19] D. Hill and J. Moore, Proc. Natl. Acad. Sci. U.S.A. **99**, 5053 (2002).

[20] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction: With 200 Full-Color Illustrations* (Springer, New York, 2001).

[21] P. S. Shenkin and D. Q. Mcdonald, J. Comput. Chem. **15**, 899 (1994).

[22] R. Tibshirani, G. Walther, and T. Hastie, J. R. Stat. Soc. Ser. B. Methodol. **63**, 411 (2001).

[23] I. Lotan and F. Schwarzer, J. Comput. Biol. **11**, 299 (2004).

[24] J. B. Tennenbaum, V. De Silva, and J. C. Langford, Science **290**, 2319 (2000).

[25] D. L. Donoho and C. Grimes, Proc. Natl. Acad. Sci. U.S.A. **100**, 5591 (2003).

[26] S. T. Roweis and L. K. Saul, Science **290**, 2323 (2000).

[27] S. Elmer and V. Pande (unpublished).

[28] R. Du, V. S. Pande, A. Y. Grosberg, T. Tanaka, and E. S. Shakhnovich, J. Chem. Phys. **108**, 334 (1998).

[29] N. Singhal, C. D. Snow, and V. S. Pande, J. Chem. Phys. **121**, 415 (2004).

[30] A. Fersht, *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding* (W. H. Freeman, New York, 1999).

[31] H. Gutfreund, *Kinetics for the Life Sciences: Receptors, Transmitters, and Catalysts* (Cambridge University Press, Cambridge, New York, 1995).

[32] G. Strang, *Linear Algebra and Its Applications*, 3rd ed. (Harcourt Brace Jovanovich, San Diego, 1988).

[33] A. R. Fersht, Proc. Natl. Acad. Sci. U.S.A. **99**, 14122 (2002).

[34] P. Mendes, CABIOS, Comput. Appl. Biosci. **9**, 563 (1993).

[35] P. Mendes, Trends Biochem. Sci. **22**, 361 (1997).

[36] W. Y. Yang, R. B. Prince, J. Sabelko, J. S. Moore, and M. Gruebele, J. Am. Chem. Soc. **122**, 3248 (2000).

[37] B. Zagrovic and V. Pande, J. Comput. Chem. **24**, 1432 (2003).

[38] P. R. Ten Wolde and D. Chandler, Proc. Natl. Acad. Sci. U.S.A. **99**, 6539 (2002).

[39] E. J. Sorin, B. J. Nakatani, Y. M. Rhee, G. Jayachandran, V. Vishal, and V. S. Pande, J. Mol. Biol. **337**, 789 (2004).

[40] S. Lifson, J. Chem. Phys. **34**, 1963 (1961).

[41] E. J. Sorin and V. Pande (unpublished).

[42] B. H. Zimm and J. K. Bragg, J. Chem. Phys. **31**, 526 (1959).

[43] V. Daggett and A. Fersht, Nat. Rev. Mol. Cell Biol. **4**, 497 (2003).

[44] M. S. Gin and J. S. Moore, Org. Lett. **2**, 135 (2000).

[45] K. Oh, K. S. Jeong, and J. S. Moore, Nature (London) **414**, 889 (2001).

[46] R. Prince, S. Barnes, and J. Moore, J. Am. Chem. Soc. **122**, 2758 (2000).

[47] A. Tanatani, M. J. Mio, and J. S. Moore, J. Am. Chem. Soc. **123**, 1792 (2001).

[48] A. Barron and R. Zuckermann, Curr. Opin. Chem. Biol. **3**, 681 (1999).

[49] *CRC Handbook of Chemistry and Physics*, 65th ed. (CRC, Boca Raton, FL, 1985), pp. f-37 and f-40.

[50] See terathane.invista.com/doc/files/743/THF_Data_Sheet.pdf