PROTEIN FOLDING TRAJECTORY ANALYSIS USING PATTERNED CLUSTERS

J. FENG

Department of Computer Science, New York University, New York, USA
E-mail: jiawu@cs.nyu.edu

L. PARIDA AND R. ZHOU

Computational Biology Center
IBM T J Watson Research Center, Yorktown Heights, USA
E-mail: {parida,ruhongz}@us.ibm.com

Understanding how protein folds into a functional and structural configuration is arguably one of the most important and challenging problems in computational biology. Currently, the protein folding mechanism is often characterized by calculating the free energy landscape versus the reaction coordinates such as the fraction of native contacts, the radius of gyration, the principal components and so on. In this paper, we present a combinatorial algorithmic approach towards understanding the global state changes of the configurations. The approach is based on cluster computation, each cluster being defined by a pattern of a combination of various reaction coordinates. We present an algorithm of time complexity $O((N + nm) \log n)$ where N is the size of the output and $n \times m$ is the size of the input. To date, this is the best time complexity for the problem. We next demonstrate that this approach extracts crucial information about protein folding intermediate states and mechanism. (1) The method recovers states previously obtained by visually analyzing free energy contour maps. (2) It also succeeds in extracting meaningful patterns and structures that had been overlooked in previous works, which provide a better understanding of the folding mechanism (of a β-hairpin protein). These new patterns also interconnect various states in existing free energy contour maps versus different reaction coordinates. (3) The approach does not require the free energy values, yet it offers analysis comparable and sometimes better than the methods that use free energy landscapes, thus validating the choice of reaction coordinates.

1. Introduction

Understanding protein folding is one of the most challenging problems in molecular biology [1]. The interest is not just in obtaining the final fold (generally referred to as structure prediction) but also understanding the folding mechanism and folding kinetics involved in the actual folding process. Many native proteins fold into unique globular structures on a very short time scale. The so-called fast folders can fold into the functional structure from random coil in microseconds to milliseconds. Recent advances in experimental techniques that probe proteins at different stages during the folding process have shed light on the nature of the folding kinetics and thermodynamics [2, 3]. However, due to experimental limitations, detailed protein folding pathways remain unknown. Computer simulations performed at various levels of complexity, can be used to supplement experiment and fill in some of the gaps in our knowledge about folding mechanisms. Meanwhile, effective analyses of the trajectory data from the protein folding simulations, either by molecular dynamics or Monte Carlo, remains yet another challenge due to the large number of degrees of freedom and the huge amount of trajectory data. Currently, the protein folding mechanism is often characterized by calculating the free energy landscape versus the so-called reaction coordinates [4, 5]. We and others have used various reaction coordinates [4, 5]. Searching for better reaction coordinates is still of great interest in protein folding mechanism studies. These analyses have provided important information for a better understanding of protein folding. However, it often requires a priori knowledge about the system under study and the free energy contour maps usually result in too much information reduction due to their limit in dimensionality which is often as low as two or three. Thus better or complementary analysis tools are in great demand.

It is also known that the folding process of many proteins takes the amino acid coil through different states before stabilizing on the final folded state. Therefore, a first step towards understanding the folding process is to identify these states. In this paper, we propose the use of a combinatorial pattern discovery technique to protein folding trajectory data from simulation experiments. A novel aspect of the algorithm is that it incorporates arbitrary and possibly different distribution functions of the data in each dimension and guarantees complete and accurate solution to the patterned clustering problem. The procedure involves computations of clusters of the data: each cluster has a signature pattern describing all the elements of the cluster. The simplicity of the pattern leads to easy interpretation of and thus better understanding of the underlying processes. By appropriate redundancy checks the number of clusters is made manageably small. The results of this method are threefold. Firstly the method is validated by comparing its results with previously published results with a free energy landscape analysis. Secondly, the method succeeds in extracting meaningful new patterns and structures (from a folded state) that had been overlooked before. These new structures provide a better understanding of the folding mechanism of a β-hairpin, which is used as a case study in this paper. These new patterns also interconnect various states in existing free energy contour maps versus different reaction coordinates. This success encourages us to postulate that the automatic discovery will lead to much greater understanding of the folding process. Thirdly, the method validates the choice of reaction coordinates since the pattern discovery analysis based on these reaction coordinates compares well with the previous free energy based approaches.

2. The Problem Description

Well known simulation methods exist to carry out the folding of a protein. However it is often not sufficient to obtain a succinct understanding of the folding process. The task here is to understand the folding mechanism by recognizing intermediate states that the folding process goes through. For example, the folding of a small protein, a β-hairpin, could be understood at a global level in terms of a few states. Although we would aim to understand the folding of every protein in this simplistic form, the current state-of-the-art is far from this goal.

At each step of the simulation process, a configuration of the solvated protein can be computed. However, the simulation may be carried for nanoseconds to microseconds in units of femtoseconds (10^{-15}) and so the number of such intermediate configurations could easily be millions in number. Hence the task is to identify and capture representative intermediate configurations. Since working in the structure space of the protein is extremely complex, researchers often identify a few key characteristic features of the protein, or often so-called reaction coordinates, and study the trends and variations in these reaction coordinates [5, 6].

In this paper we utilize a four step process towards understanding the folding of a protein. The first step involves the in-silico simulation that gives rise to a large collection of data points, each point being an array of the characteristic features of the folding protein at that time point. For example, the radius of gyration or the number of hydrogen bonds could be such features. In Section 4 we study the β-hairpin folding as a show case and describe seven such characteristic features that we have used previously in the study of this particular protein.

In the second step, we study these data points to extract the characteristic set of features which we call patterned clusters. Again, in the case of the β-hairpin, the data points are seven dimensional, corresponding to the characteristic features of the protein at each time interval. See Figure 1 for a small portion of the data as an example. In the third step, these patterns are filtered to retain the most significant ones. It is very difficult to model the significant patterns in this domain, so we have combined the second and third steps and use appropriate parameters to filter out possibly insignificant patterns. For instance, if a pattern occurs less than k times, then the pattern is possibly not salient. Also we exercise control by the use of meaningful $\delta()$ functions (see the next section for details).

The fourth step is of analyzing the patterns: this involves extracting the structure of the configuration using the time coordinates and studying the correlation of the different structures. For instance, one could observe that the hydrophobic core is formed before the beta-strand hydrogen bonds, or vice versa; and one can interconnect various free energy states in different free energy contour maps by monitoring the high dimensional (multi-column) patterns. These findings can provide a better understanding of the protein folding mechanism. Further, the time correlation between various patterns or states could be studied. For example, it is extremely useful to know which pattern or state precedes the other and by how much time.

3. On Patterned Clusters

Due to space constraints, the theoretical and algorithmic details such as proof of correctness and complexity analysis will appear in the full version of the paper.

Definition 3.1 *(δ-cluster, maximal δ-cluster) Given $\delta() : \mathcal{R} \rightarrow \mathcal{R}^+$, $v_i \in \mathcal{R}$, $1 \leq i \leq n$ and a quorum k. A δ-cluster is collection of i with $v_i \in V_c$, $|V_c| \geq k$ such that if $v_1, v_2 \in V_c$, then $|v_1 - v_2| \leq \frac{1}{2}(\delta(v_1) + \delta(v_2))$. Further, V_c is maximal if there exists no V'_c such that $V_c \subset V'_c \subseteq V$ and V'_c is a δ-cluster.*

Definition 3.2 *(cross δ-cluster, maximal cross δ-cluster) Given $\delta_j() : \mathcal{R} \rightarrow \mathcal{R}^+$, quorum k*

and $v_{ij} \in \mathcal{R}$, $1 \leq j \leq m, 1 \leq i \leq n$. A cross δ-cluster is collection i and j with $v_{ij} \in V_c$ such that for each j, $\{v_{ij} \in V_c | 1 \leq i \leq n\}$ is a δ_j-cluster. Further, V_c is maximal if there exists no additional i' or j' with the corresponding V_c' with $V_c \subset V_c' \subseteq V$ such that V_c' is a cross δ-cluster.

Here we present an output sensitive algorithm that computes all the maximal patterned (cross) δ-clusters. The algorithm has two main steps: **Step 1: Maximal δ_j-cluster computation.** For each j, $1 \leq j \leq m$, compute the maximal δ_j-cluster, V_l^j. For simplicity let the number of these be L and the clusters be $V_{cl}, 1 \leq l \leq L$. The pseudocode, *Compute-Cluster()*, describes the maximal δ-cluster computations, for each j. To avoid clutter, the end-of-input check is not included in the code. **Step 2: Maximal patterned (cross) δ-cluster computation.** The algorithm in this step is based on the Set Intersection Problem (SIP) described in [7] in the context of computing redundant motifs from irredundant ones. Let the elements be numbered $1 \ldots L$. Assume a function Generate-Set($\mathcal{V}_c$) which creates $\mathcal{V}_c$, a subset of $V_{c1}, V_{c2}, \ldots, V_{cn}$ in an appropriate data structure $\mathcal{D}$ (say a tree). A query of the form if a subset $\mathcal{V}_c$ exists in $\mathcal{D}$ takes $O(\log n)$ time. The pseudocode, *Generate-Set()*, describes the algorithm. The initial call is Generate-Set($\{V_{c1}, V_{c2}, \ldots, V_{cL}\}, n, 0$). The maximal patterned δ-clusters are $\{v_{ij} | V_{cj} \in \mathcal{V}_k$ and $v_{ij} \in V_{cj}\}$, for each computed $\mathcal{V}_k$ stored in $\mathcal{D}$.

```
Compute-Cluster()                                    [Step 1]
    (1) Sort $v_i$'s as $v_1, \ldots, v_n$
    (2) $i \leftarrow 1$, $l \leftarrow i + 1$
    (3) If $|v_i - v_l| \leq \frac{1}{2}(\delta(v_i) + \delta(v_l))$
        (4) $l \leftarrow l + 1$, goto Step 3
    (5) Else $C^i = \{v_j | i \leq j < l\}$,
            $i \leftarrow i + 1$, goto Step 3
```

```
Generate-Set($\mathcal{V}_k, i, flag$)                [Step 2]
    (1) If (flag =0) add $\mathcal{V}_k$ to $\mathcal{D}$
    (2) If ($i \leq 0$) then exit
    (3) Let $\mathcal{V}_k' = \{V_{cl} \in \mathcal{V}_k | i \in V_{cl}\}$
    (4) If $\mathcal{V}_k'$ exists in $\mathcal{D}$
            Generate-Set($\mathcal{V}_k, i - 1, 1$)
    (5) Else Generate-Set($\mathcal{V}_k', i-1, 0$)
```

Complexity of the Algorithm. The time taken by the complete algorithm is $O(nm \log n + N \log n)$ where N is the size of the output and nm is the size of the input.

4. Case Study: Folding of β-hairpin

A small but important protein system has been selected as an example to demonstrate our approach to understanding the folding process. This small protein is a 16-residue β-hairpin (GEWTYDDATKTFTVTE) from the C-terminus of protein G (2gb1.pdb). Its folding mechanism and folding free energy states have been studied extensively in previous works [5, 6]. The current study will use our new approach to analyzing the existing trajectories from the previous molecular dynamics simulations in explicit solvent. This β-hairpin has received much attention recently from both experimental and theoretical fronts [3, 2, 8, 9, 10, 4, 11, 12, 13]. However, there are still a number of important aspects that remain controversial, such as the relative importance and time sequential order between the beta-strand hydrogen bonds formation and the hydrophobic core formation; and whether or not the existence of alpha-helical intermediates during the folding.

Figure 1. Raw data from the REM sampling of the β-hairpin folding in explicit water. (1) N_{HB}^{β}: the number of native beta-strand hydrogen bonds, (2) R_g^{core}: radius of gyration of the hydrophobic core residues (TRP43, TYR45, PHE52 and VAL54). (3) ρ: radius of gyration of entire protein (R_g), (4) fraction of native contacts, (5) PC-1: the first principal component from Principal Component Analysis (6) PC-2: the second principal component, and (7)$RMSD$: the backbone root mean square deviation (RMSD) from the native structure.

J_1	J_2	J_3	J_4	J_5	J_6	J_7
N_{HB}^{β}	R_g^{core}	R_g	ρ	PC-1	PC-2	$RMSD$
5.000	5.175	8.653	1.000	-7.819	-34.008	0.000
4.474	5.328	8.361	0.953	-7.972	-35.772	1.595
4.354	5.416	8.471	0.988	-7.899	-36.399	1.379
4.053	5.257	8.298	0.893	-8.373	-35.536	1.708
3.776	5.186	8.381	0.857	-7.777	-35.415	1.624
2.155	5.390	7.816	0.778	-2.277	-27.017	3.672
4.842	6.043	7.312	0.778	2.144	-33.772	5.208
0.000	8.303	10.033	0.242	-27.075	43.521	10.163
3.797	5.990	7.514	0.728	-3.084	-30.185	4.838
2.898	5.483	7.775	0.778	-2.888	-26.254	3.904

4.1. *Simulation Parameters (Step 1)*

In this study, an all-atom model is used for the description of the protein solvated in water. The Optimized Potential for Liquid Simulations - All-Atom (OPLS-AA) force field [14] with an explicit solvent model, Simple Point Charge (SPC) model is used. A total of 64 replicas of the solvated system consisting of 4342 atoms is simulated with temperatures spanning from 270 K to 695 K. For each replica, a 3 nanosecond molecular dynamic simulation is run with replica exchanges attempted every 400 femtoseconds. The reader is directed to [5, 6] for details of this simulation. For each conformation, seven different reaction coordinates are used: see Figure 1 for details. There are a total of about 20,000 conformations saved for each replica. Figure 1 lists a small portion of the data for the replica at 310 *Kelvin* (37 Celsius), which is the biological temperature.

These simulations have revealed a hydrophobic-core driven folding mechanism from free energy contour map analysis [5]. Since this is a well studied system and a large amount of data is available, comparisons with other analysis tools, such as the free energy contour map analysis, might be easier and more straightforward. Various reaction coordinates obtained from previous runs serve as the starting point.

4.2. *Discovery Parameters (Steps 2 & 3)*

The δ function of the cluster detection problem is defined as a constant. Thus $\delta(x) = c$, for some constant $c \in \mathcal{R}$ for each x. The δ functions for each column of Figure 1 is given as follows: $\delta_1(x) = 0.2$, $\delta_2(x) = 0.6$, $\delta_3(x) = 0.35$, $\delta_4(x) = 0.15$, $\delta_5(x) = 5.0$, $\delta_6(x) = 16.5$, $\delta_7(x) = 1.0$ for all x. Further, the quorum k is defined to be 2000. Figure 2 lists some representative patterns of size two with these parameters. The time sequences are not shown due to the space constraints. These simple patterns can be directly compared with the previous free energy states in the 3-D free energy contour maps. These are 3-D plots of free energy versus a pair of reaction coordinates or data columns of Figure 1.

Figure 2. Simple patterns of size two. These patterns can be easily compared to the 3-D free energy landscapes using a pair of corresponding reaction coordinates.

Size	Cluster Pattern	
2	J_1 = 4.886 $\underline{+}$ 0.2	J_2 = 5.448 $\underline{+}$ 0.6
2	J_1 = 2.875 $\underline{+}$ 0.2	J_2 = 5.448 $\underline{+}$ 0.6
2	J_2 = 4.979 $\underline{+}$ 0.6	J_4 = 0.816 $\underline{+}$ 0.15
2	J_2 = 5.871 $\underline{+}$ 0.6	J_4 = 0.686 $\underline{+}$ 0.15
2	J_2 = 4.979 $\underline{+}$ 0.6	J_3 = 8.144 $\underline{+}$ 0.35

One might often want to study detailed patterns or structures in some predefined subregions such as the structures in the unfolded ensemble. More and more evidences have shown that the protein structures in unfolded states are not fully extended, but often have well-defined structures instead [15]. This can also avoid the problem that important patterns in these less populated areas are being overlooked due to a smaller population than the predefined quorum k. Thus, some less populated free energy states in free energy landscapes can be recovered by reducing the quorum. Hence another set of parameters have been used and here we confine our search to data points with $N_{HB}^{\beta} = 0.0$ and $R_g^{core} >$ 5.0 Å (see Figure 1 for definitions of these reaction coordinates) with $k = 100$. Yet another set of parameters have included $N_{HB}^{\beta} =$ 0.0 and $R_g^{core} >$ 9.0 Å with $k = 50$. A subset of the results are shown in Figure 5. Thus this approach might be useful for hierarchical pattern searches which gradually zoom into the predefined subsets of data.

4.3. *Analysis of Results (Step 4)*

To obtain a representative structure(s) from a set of configurations c_i, the set is partitioned into a minimum number of groups G_j such that for each G_j there exists a representative $c_i^j \in G_j$ and for each $c_k \in G_j$, the structure corresponding to c_k is at most 1 Å RMSD from c_i^j. Thus each G_j will be represented by a structure corresponding to c_i^j [5, 16].

Recovering known free energy states. Obviously, the first question of importance is: Can we recover the previously found free energy states in the new approach? The “time sequence” of each pattern is then used to extract the corresponding conformations of the protein. Figure 3(a) shows a representative or most populated structure for the first pattern in Figure 2. This structure mimics the representative structure from the folded state (F state) in the free energy contour map versus N_{HB}^{β} and R_g^{core} very well. Thus this pattern resembles the F state of the free energy contour map. Similarly, the second pattern of Figure 2 resembles the partially folded state, P state, in the same free energy landscape. The structures for the two patterns are shown in Fig. 3. Thus our approach recovers the most populated states in the free energy landscape analysis.

The third and fourth patterns in Figure 2 also resemble the F state and P state, respectively, in the same free energy contour map versus N_{HB}^{β} and R_g^{core}. Numerous other patterns have shown similar results, i.e., recovering various previously found free energy states in the free energy contour maps versus different reaction coordinates. It should be noted though that many patterns might be redundant, either because the $\delta()$ function values given

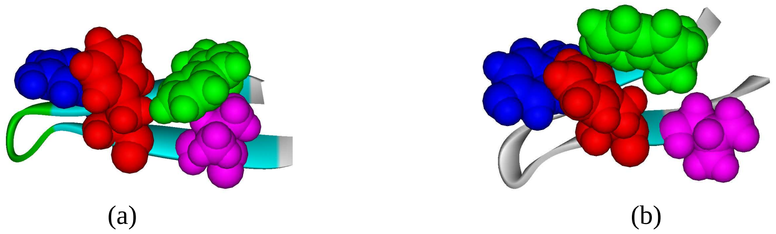

Figure 3. Representative structures for two patterns are shown here. In the schematic diagram the hydrophobic residues TRP43, TYR45, PHE52, and VAL54 are represented by spacefill and the rest of the residues are represented by ribbons. (a) Pattern 1 in Figure 2 captures the folded state (F state) in free energy contour map analysis (b) Pattern 2 in Figure 2 captures the partially folded state (P state) in the same free energy contour map.

Figure 4. Complex patterns of size up to six.

Size	Cluster Pattern			
3	J_2= 5.375 ± 0.6	J_3= 7.971 ± 0.35	J_5= -5.881 ± 5.0	
3	J_2= 5.375 ± 0.6	J_4= 0.743 ± 0.15	J_5= -5.881 ± 5.0	
3	J_1= 4.903 ± 0.2	J_4= 0.796 ± 0.15	J_6= -33.574 ± 16.5	
4	J_1= 4.903 ± 0.2	J_2= 5.375 ± 0.6	J_4= 0.870 ± 0.15	J_6= -33.574 ± 16.5
4	J_1= 4.903 ± 0.2	J_2= 5.375 ± 0.6	J_5= -5.881 ± 5.0	J_6= -33.574 ± 16.5
5	J_3= 8.144 ± 0.35	J_4= 0.815 ± 0.15	J_5= -5.881 ± 5.0	
	J_6= -33.574 ± 16.5	J_7= 3.292 ± 1.0		
5	J_3= 8.144 ± 0.35	J_4= 0.902 ± 0.15	J_5= -3.855 ± 5.0	
	J_6= -33.574 ± 16.5	J_7= 3.292 ± 1.0		
6	J_1= 4.950 ± 0.2	J_3= 8.013 ± 0.35	J_4= 0.848 ± 0.15	
	J_5= -5.881 ± 5.0	J_6= -33.574 ± 16.5	J_7= 3.292 ± 1.0	
6	J_2= 5.748 ± 0.6	J_3= 8.013 ± 0.35	J_4= 0.848 ± 0.15	
	J_5= -5.881 ± 5.0	J_6= -33.574 ± 16.5	J_7= 3.800 ± 1.0	

for reaction coordinates are too wide, or some of the reaction coordinates are highly correlated. For example the fifth pattern of Figure 2 is $R_g^{core} = 4.979 \pm 0.6$, $R_g = 8.144 \pm 0.35$. Clearly, these two reaction coordinates are highly correlated, since R_g^{core} measures the radius of gyration of 4 key residues out of the total 16 which is measured by R_g. However for many other cases it may not be so obvious.

Interconnecting various free energy contour maps. More complicated patterns with many reaction coordinates are also found in the current approach which had been previously undetected. While in the traditional free energy contour map analysis, typically one or two reaction coordinates are used at each time, since a 2-D or 3-D free energy contour map is usually plotted. It is extremely difficult to visualize high dimensional free energy landscapes in order to identify the free energy basins or barriers. Figure 4 lists some of these complicated patterns with up to 6 reaction coordinates. Of course, as pointed out earlier, some reaction coordinates might be correlated, so the data in each reaction coordinate may not be totally independent. Nevertheless, it still reveals some interesting new findings. First of all, these patterns can interconnect various free energy states in different free en-

Figure 5. Clusters with (1) $J_1 = 0.0, J_2 \geq 5.0, k = 50$ and (2) $J_1 = 0.0, J_2 \geq 10.0, k = 100$. To avoid clutter the J_1 values are not shown.

Sz	Cluster Pattern			
1	J_2= 5.448 $\underline{+}0.5$			
2	J_3= 10.218 $\underline{+}0.2$	J_4= 0.050 $\underline{+}0.15$		
2	J_3= 10.773 $\underline{+}0.2$	J_5= -21.188 $\underline{+}15$		
3	J_3= 10.208 $\underline{+}0.2$	J_4= 0.050 $\underline{+}0.15$	J_7= 9.299 $\underline{+}0.8$	
4	J_2= 9.632 $\underline{+}0.5$	J_3= 10.302 $\underline{+}0.2$	J_5= -21.188 $\underline{+}15$	J_7= 9.299 $\underline{+}0.8$
5	J_2= 9.951 $\underline{+}0.5$	J_4= 0.050 $\underline{+}0.15$	J_5= -21.188 $\underline{+}15$	
	J_6= 36.517 $\underline{+}15$	J_7= 9.872 $\underline{+}0.8$		

ergy contour maps. This might not be so obvious in free energy contour maps themselves. For example the sixth pattern in Figure 4 interconnects the following two free energy contour maps: PC-1 and PC-2 and ρ and R_g in Figures 3(a) and 3(b) respectively in [5]. The states corresponding to the free energy well (of value $\approx$ -9KT) near PC-1 $= -5.9$, PC-2$= -33.6$ in the first contour map and $\rho = 0.82$, $R_g = 8.1$ in the second contour map are indeed the same free energy state consisting of the same structures. In this particular case, they all represent the folded state (F state).

Better understanding folding mechanism. More importantly, the new approach reveals important structures overlooked previously, which might help understand the folding mechanism better. Eaton and coworkers [3] proposed a ”hydrogen bond zipping” mechanism for this β-hairpin in which folding initiates at the turn and propagates toward the tails by making beta-strand hydrogen bonds one by one, so that the hydrophobic core, from which most of the stabilization derives, form relatively late during the folding. In our previous study, we proposed a different folding mechanism that this β-hairpin undergoes a hydrophobic core collapse first, then makes native β-strand hydrogen bonds to make over the free energy loss due to the loss of H-bonds between the backbone atoms and water. Figure 6(a) shows a representative structure for the eighth pattern in Figure 4. The structure shows that all the five native β-strand H-bonds have been formed, but the hydrophobic core is not completely aligned yet. The loop region also bends towards the hydrophobic core to somewhat offset the non-perfect hydrophobic core. These structures with H-bonds formed but hydrophobic core not perfectly aligned (RMSDs up to 4 Å) implies that the hairpin can also have a path to form β-strand hydrogen bonds before the core is finalized. The current findings indicate that the final hydrophobic core and β-strand hydrogen bonds might be formed almost simultaneously. This can also be seen from the low free energy barrier in free energy landscapes as discussed before [5]. Interestingly, Thirumalai *et al.* also found that the lag time between collapse and hydrogen bond formation is very short and the two processes occur nearly simultaneously [17].

Finally, the patterns of subsets of data in less populated states, such as the unfolded state, are studied in detail by zooming into these regions with a smaller quorum k and a different set of $\delta()$. As mentioned earlier, more and more evidences have shown that the protein structures in unfolded states are not fully extended, but often have well-defined

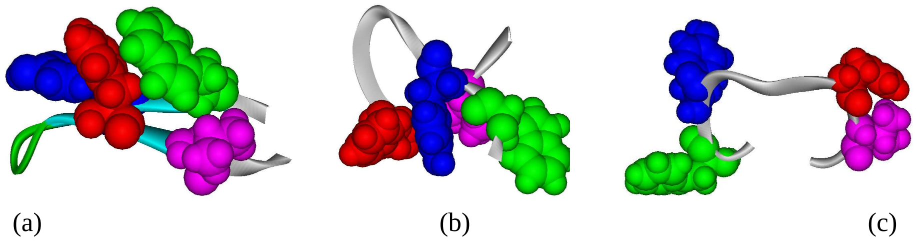

Figure 6. (a) Pattern 6 of Figure 4 which represents a new class of structures previously overlooked in free energy landscape analysis. (b) Pattern 1 of Figure 5 which captures the H state (hydrophobic core formed but no beta-strand H-bonds) in free energy contour map analysis (c) Pattern 2 in Figure 2 captures the unfolded state (U state) in the same free energy contour map. The hydrophobic residues TRP43, TYR45, PHE52, and VAL54 are represented by spacefill and the rest are represented by ribbons.

structures instead [15]. The first pattern in Figure 5 resembles the previous H-state in free energy contour map versus N_{HB}^{β} and R_g^{core}, where the hydrophobic core is largely formed but no native beta-strand H-bonds have been made yet. Figure 6(b) shows a representative structure of this pattern, which mimics the structures from previous H-state very well. Figure 6(c) shows a representative structure for the sixth pattern in Figure 5. This is the most populated structure of this β-hairpin in unfolded state. Even though not much structural features are found in this structure, it is certainly not fully extended either. Since this is a very small protein with only one secondary structure in the native state, not much has been identified in the unfolded state; for larger and more complicated protein systems, such as lysozyme, more structural features might be expected in the unfolded state [15].

5. Conclusion & Ongoing Work

In this paper we have presented a method to enhance our understanding of protein folding mechanisms. At the heart of this method is a combinatorial pattern discovery algorithm that analyzes multi-dimensional data from the simulation of the protein folding trajectory. The approach is based on cluster computation, each cluster being defined by a pattern of the reaction coordinates. A small but important protein system, a β-hairpin from C-terminus of protein G, is then used to demonstrate this approach. It is shown that the method not only reproduces the previously found free energy states (most populated states) in free energy contour maps, but also reveals new information overlooked previously in free energy landscape analysis about the intermediate structures and folding mechanism. It is also shown to be useful in making interconnections between various 3-D free energy contour maps versus different reaction coordinates and also explain the mechanisms of the folding process. The method also validates the choice of reaction coordinates as the analysis without using free energy values compares well the ones that use them. The success with β-hairpin is very encouraging and we are currently exploring the application of this method to other larger protein molecules.

References

1. C. M. Dobson, A. Sali, and M. Karplus. Protein folding: a perspective from theory and experiment. *Angrew Chem. Int. Edit. Engl.*, 37:868–893, 1998.
2. F. J. Blanco, G. Rivas, and L. Serrano. A short linear peptide tha folds in a native stable β-hairpin in aqueous solution. *Nature Struc. Bio.*, 1:584–590, 1994.
3. V. Munoz, P. A. Thompson, J. Hofrichter, and W. A. Eaton. Folding dynamics and mechanism of β-hairpin formation. *Nature*, 390:196–199, 1997.
4. A. E. Garcia and K. Y. Sanbonmatsu. Exploring the energy landscape of a β hairpin in explicit solvent. *Proteins*, 42:345–354, 2001.
5. R. Zhou, B. J. Berne, and R. Germain. The free energy landscape for β-hairpin folding in explicit water. *Proc. Natl. Acad. Sci.*, 98:14931–14936, 2001.
6. R. Zhou. Free energy landscape of protein folding in water: explicit vs. implicit solvent. *Proteins*, 53:148–161, 2003.
7. Laxmi Parida. Some results on flexible-pattern discovery. *Combinatorial Pattern Matching (CPM2000), LNCS*, 1848:33–45, 2000.
8. V. S. Pande and D. S. Rokhsar. Molecular dynamics simulations of unfolding and refolding of a β-hairpin fragment of protein g. *Proc. Natl. Acad. Sci. USA*, 96:9062–9067, 1999.
9. B. Zagrovic, E. J. Sorin, and V. S. Pande. β-hairpin folding simulation in atomistic detail. *J. Mol. Biol.*, 313:151, 2001.
10. A. R. Dinner, T. Lazaridis, and M. Karplus. Understanding β-hairpin formation. *Proc. Natl. Acad. Sci. USA*, 96:9068–9073, 1999.
11. D. Roccatano, A. Amadei, A. Di Nola, and H. J. Berendsen. A molecular dynamics study of the 41-56 β-hairpin from b1 domain of protein g. *Protein Sci.*, 10:2130–2143, 1999.
12. A. Kolinski, B. Ilkowski, and J. Skolnick. Dynamics and thermodynamics of β-haipin assembly: insights from various simulation techniques. *Biophys. J.*, 77:2942–2952, 1999.
13. B. Ma and R. Nussinov. Molecular dynamics simulations of a β-hairpin fragment of protein g: balance between side-chain and backbone forces. *J. Mol. Bio.*, 296:1091, 2000.
14. W. L. Jorgensen, D. Maxwell, and J. Tirado-Rives. Development and testing of the opls all-atom froce field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.*, 118:11225–11236, 1996.
15. J. Klein-Seetharaman, M. Oikawa, S. B. Grimshaw, J. Wirmer, E. Duchardt, T. Ueda, T. Imoto, L. J. Smith, C. M. Dobson, and H. Schwaldbe. Long-range interactions within a nonnative protein. *Science*, 295:1719–1722, 2002.
16. R. Zhou. Trp-cage: Folding free energy landscape in explicit water. *Proc. Natl. Acad. Sci.*, 100:13280–13285, 2003.
17. D. K. Klimov and D. Thirumalai. Mechanism and kinetics of β-hairpin formation. *Proc. Natl. Acad. Sci. USA*, 97:2544–2549, 2000.