JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

Hidden Protein Folding Pathways in Free-Energy Landscapes Uncovered by Network Analysis

Yanping Yin,[†] Gia G. Maisuradze,[†] Adam Liwo,[†,‡] and Harold A. Scheraga*[,†]

[†]Baker Laboratory of Chemistry and Chemical Biology, Cornell University, Ithaca, New York 14850-1301, United States

[‡]Faculty of Chemistry, University of Gdańsk, Sobieskiego 18, 80-952 Gdańsk, Poland

ABSTRACT: A network analysis is used to uncover hidden folding pathways in free-energy landscapes usually defined in terms of such arbitrary order parameters as root-mean-square deviation from the native structure, radius of gyration, etc. The analysis has been applied to molecular dynamics trajectories of the B-domain of staphylococcal protein A, generated with the coarse-grained united-residue force field in a broad range of temperatures (270 K $\leq T \leq$ 325 K). Thousands of folding pathways have been identified at each temperature. Out of these many folding pathways, several most probable ones were selected for investigation of the conformational transitions during protein folding. Unlike other conformational space network (CSN) methods, a node in the CSN variant implemented in this work is defined according to the native-likeness class of the structure, which defines the similarity of segments of the compared structures in terms of secondary structure, contact pattern, and local geometry as well as the overall geometric similarity of the conformation under consideration to that of the reference (experimental) structure. Our previous findings, regarding the folding model and conformations found at the folding-transition temperature for protein A (Maisuradze et al. *J. Am. Chem. Soc.*, **2010**, *132*, 9444), were confirmed by the conformational space network analysis. In the methodology and the analysis of the results, the shortest path identified by using the shortest-path algorithm corresponds to the most probable folding pathway in the conformational space network.

INTRODUCTION

Proteins have to fold into a unique ensemble of three-dimensional structures in order to perform their functions. To understand how proteins fold and function, knowledge of their free-energy landscapes (FELs)[1−4] is required. It is impossible to present an FEL as a function of all degrees of freedom of a protein. Consequently, it is very important to find the coordinates along which the intrinsic folding pathways can be viewed. The common choices for reaction coordinates are root-mean-square-deviation (rmsd) with respect to the native structure, radius of gyration, number of native contacts, and other order parameters. Recent studies[5−8] have shown that FELs projected on one or two order parameters are relatively simple and may contain artifacts. Naturally, questions have been raised[6,7,9] about the dimensionality of the FELs and the appropriate reaction coordinates, along which protein-folding progress can be described correctly. It has been demonstrated[6−11] that principal components are one of the alternatives that answers these questions. Another alternative is to present the FEL without projecting the free energy on chosen coordinates.

Recent significant developments in network research[12] have attracted attention for studying different complex systems, such as social interaction, the Internet, and protein folding. One of the approaches, in which a small network was constructed for the description of folding behavior, was developed by Krivov and Karplus[13] and applied to study the folding dynamics of the β-hairpin of protein G.[5] An unprojected FEL, termed a transition disconnectivity graph (TRDG), was introduced[13] for this purpose. Another example of the use of complex network theory for the analysis of conformational space with application to the folding of a 20-residue antiparallel β-sheet peptide was presented by Rao and Caflisch.[14] In that work, the nodes in the conformational space network (CSN) represent the conformations, and the links correspond to direct transitions between different nodes. They also presented the free energy surface of the alanine "dipeptide" by network analysis,[15] and the free-energy basins of the FEL were identified by partitioning the network into different clusters using a cluster-detection algorithm.[15] Recently, by studying the dynamics of a small peptide in its native state using an inherent structure (IS)[16] based approach, i.e., by focusing on local minima in the potential energy surface, Rao and Karplus[17] illustrated the correspondence between conformational changes, energy barriers, and transition kinetics by mapping the potential energy onto an FEL using an IS-based CSN. All these investigations have demonstrated that the network approach can be used to obtain free-energy landscapes without having to identify the essential degrees of freedom. Moreover, network analysis can capture all the transitions between different conformations; consequently, it is a very useful approach with which to analyze folding pathways.

In previous studies of protein folding by network analysis, a node was usually defined according to secondary structure,[14] rmsd,[14] or backbone dihedral angles (φ,ψ).[15] However, since the conformations in the same node should be structurally similar to each other, in this work a node in the CSN is defined according to the native-likeness of the structure, as introduced recently,[18,19] which considers more factors, not only secondary structure and rmsd but also the packing between pairs of secondary structure. Because all the conformational transitions

Received: November 10, 2011
Published: February 24, 2012

can be captured in the CSN, every possible folding pathway is also presented in the CSN, and it is possible to identify the most probable folding pathways in the CSN. Therefore, in this work, for the first time, the shortest-path algorithm is applied to the CSN in order to identify the most probable folding pathways.

In this paper, the folding trajectories from molecular dynamics (MD) simulations of the B-domain of staphylococcal protein A (1BDD), a 46-residue three-α-helical protein,[20] generated with the UNRES force field[18,19,21−27] are analyzed by network analysis. Protein A has been studied intensively,[7,8,28−45] because it is a small protein and folds rapidly to the native structure. Also, because of the loose native-like structure, which leads to quite a broad native state with several deep minima and the possibilities to fold by many pathways, 1BDD is quite a challenging system with which to test the network approach. Conformational space networks are constructed from 16 independent trajectories, each consisting of 1×10^8 MD steps (489 ns of UNRES time), at each of several temperatures (T = 270, 280, 300, 310, and 325 K).

METHODS

Simulation. Network analysis was applied here to the MD trajectories generated with the coarse-grained UNRES[18,19,21−27] force field applied to polypeptide chains. In the UNRES force field, a polypeptide chain is represented by a sequence of α-carbon (C^α) atoms linked by virtual bonds with united peptide groups and united side chains. Each united peptide group is located in the middle between two neighboring α-carbons. Only these united peptide groups and the united side chains serve as interaction sites, the α-carbons serving only to define the chain geometry. The energy function of the virtual-bond chain is expressed by eq 1:[27]

$$\begin{aligned} U = {} & w_{SC}\sum_{i<j} U_{SC_iSC_j} + w_{SC_p}\sum_{i\neq j} U_{SC_iP_j} \\ & + w_{pp} f_2(T)\sum_{i<j-1} U_{P_iP_j} + w_{tor} f_2(T)\sum_i U_{tor}(\gamma_i) \\ & + w_{tord} f_3(T)\sum_i U_{tord}(\gamma_i, \gamma_{i+1}) + w_b\sum_i U_b(\theta_i) \\ & + w_{rot}\sum_i U_{rot}(\alpha_{SC_i}, \beta_{SC_i}, \theta_i) + w_{bond} \\ & \sum_i U_{bond}(d_i) + \sum_{m=3}^{6} w_{corr}^{(m)} f_m(T) U_{corr}^{(m)} \\ & + w_{SS}\sum_i U_{SS;i} \end{aligned} \tag{1}$$

with the temperature-dependent factor

$$f_n(T) = \frac{\ln[\exp(1) + \exp(1)]}{\ln\left\{\exp\left[\left(\frac{T}{T_0}\right)^{n-1}\right] + \exp\left[-\left(\frac{T}{T_0}\right)^{n-1}\right]\right\}_{T_0=300K}} \tag{2}$$

The successive terms in eq 1 represent side chain−side chain, side chain−peptide, peptide−peptide, torsional, double-torsional, bond-angle bending, side-chain local, distortion of virtual bonds, multibody (correlation) interaction, and formation of disulfide bonds, respectively. More details of the theoretical basis of the UNRES force field and parameterization of the energy terms are described in previous papers.[18,19,21−27]

Canonical MD simulations for 1BDD were run at five different temperatures, 270, 280, 300, 310, and 325 K, with 16 trajectories at each temperature. The version of the force field used in this study was calibrated[27] with the 1GAB protein. The calibration procedure was based on a hierarchical-optimization method developed in our laboratory.[18,19,26,27] In this method, the energy-term weights are optimized so that the free energy of a given subensemble of conformations decreases with increasing native-likeness below the folding-transition temperature and increases above the folding-transition temperature and so that the free energies of all subensembles be equal at the folding-transition temperature. It should be noted that the calibration is only the last stage of force-field parametrization aimed at putting together the energy terms to give a folding force field, which are derived from free-energy surfaces of model systems[24,25] and PDB statistics.[22,23] Therefore, the obtained force field is transferable to other proteins, even though it was calibrated with one.[27] In particular, it can fold protein A, a protein which was not used in calibration and which has little sequence similarity to 1GAB. The time step in the MD simulations was 0.1 mtu (1 mtu = 48.9 fs is the "natural" time unit of MD), and the coupling parameter of the Berendsen thermostat[46] was 1 mtu.

Classification of Structures. In the conformational space network, each node represents a conformation, and a link between two nodes corresponds to the transition between two conformations. In this work, the node was defined according to the native-likeness of the structure,[18,19] which was represented by a series of class numbers, called a class code (see Appendix for details).[18,19] As an example, Table 1[19] shows the structural classification of 1BDD associated with a specific class number at each conformational level. For example, the native structure of 1BDD has a class code of 777.22.2. The class code has three levels. Level 1 represents the native-likeness of the elementary fragments, which are defined as the consecutive three helices. Therefore, the first three digits (777) of the class code correspond to the first level, with each digit representing a helix. It can be seen from Table 1 that a "1" in level 1 means that the fragment matches the native fragment only in secondary structure contact (without interaction between the helices) and a "7" in level 1 means that it is also similar to the native fragment in contact pattern and in geometric details. The next two numbers (22) correspond to level 2, which accounts for the similarity of contact pattern of packing of the pairs of fragments to that in the experimental structure. In Table 1, a "2" in level 2 shows the native packing without a sequence shift between a pair of elementary fragments. The packing between a given pair of elementary fragments is considered native if the number of native contacts (the number of side-chain contacts for 1BDD) between this pair of fragments is greater than 70% of the native contacts, and the rmsd of the segment consisting of this pair of fragments is lower than a certain threshold. The last number (2) represents level 3 and describes the overall similarity of the calculated to the experimental structure; a "2" is assigned if the rmsd is low, i.e., less than a chosen cutoff ("rmsd match"). A more extensive description of the class code is provided in the Appendix and in refs 18 and 19.

Table 1. Structural Classification of 1BDD Associated with a Specific Class Number Corresponding to Secondary Structure, Packing between a Given Pair of Secondary Structures, and Rmsd Match of the Whole Molecule[19]

level[a]	class number[b]	structural similarity
1	0	non-native fragment
	1	native secondary structure
	2	native hydrogen-bonding internal contacts, only after a sequence shift
	3	native secondary structure and hydrogen-bonding internal contacts after a sequence shift
	4	not used (see ref 17)
	5	not used (see ref 17)
	6	native hydrogen-bonding internal contacts only without a sequence shift
	7	native secondary structure and hydrogen-bonding internal contacts without a sequence shift
2	0	non-native packing
	1	native packing after a sequence shift
	2	native packing without a sequence shift
3	0	no rmsd match
	1	rmsd match after a sequence shift
	2	rmsd match without a sequence shift

[a]The levels in refs 18 and 19 have a totally different meaning from each other. The levels in ref 18 represent the hierarchy on the potential-energy surface. The energy levels of the structure in ref 18 decrease with increasing native-likeness of the structure. The conformational levels in ref 19 are defined to evaluate the native-likeness of conformations. [b]This class number differs slightly from that in ref 19. In ref 19, a class number "3" in level 2 and level 3 represents native packing or rmsd match without a sequence shift, and a "2" in level 2 and level 3 is not used.

Construction of the CSN. The class code is calculated for each conformation along the trajectory. The conformations with identical class code are grouped into the same node. If a transition between two conformations with different class codes is observed in the trajectory, a link is added between the two corresponding nodes. A weight w_i assigned to node i is equal to the number of conformations with a given class code grouped in node i. A weight w_{ij} assigned to the link between nodes i and j is equal to the number of times this transition is observed in either direction along the trajectory.

CSN, as well as many other real networks,[47−49] have been found to have a community structure[50] (which is also called a clustering[51] or modular structure[15]), which is defined as the presence of groups of nodes in a network that have dense links within each group and sparse links between the groups. This kind of group of nodes is often called a cluster in the network. Many algorithms have been developed to detect the clusters in the network. In this work, two cluster-detection algorithms, the Markov clustering (MCL)[52,53] and the modularity maximization (MM) algorithms,[51] have been applied to identify the clusters in the CSN. The MCL algorithm with a stochastic matrix simulates the behavior of stochastic walks on the network. A parameter I is usually used to tune the granularity of the clustering. At $I = 1$, the network is considered as one single cluster. As the value of I increases, more and more clusters are generated by the algorithm. In this paper, a small value of I is used to identify the largest clusters in the network. The MM algorithm is a hierarchical agglomerative algorithm that searches over possible partitionings of the network for ones with a high value of the modularity $Q = \sum_i (e_{ii} - a_i^2)$, which measures the quality of the partitioning of the network. The quantity e_{ii} is the fraction of links between nodes in cluster i, and $a_i = \sum_j e_{ji}$ is the fraction of links connected to nodes in cluster i.

Both the MCL and the MM algorithms generated two major clusters in the CSN for $T = 270$ and 280 K, as shown in Figure 1 for 270 K. The structures of the most populated nodes in each cluster were examined to see whether the node contains native-like structure. The nodes in one cluster contain mainly structures with an rmsd match, which shows that this cluster represents the native basin. The nodes in the native basin are colored in red (Figure 1). The nodes in the other cluster are colored in green and contain structures without an rmsd match, which include mirror-image structures and molten-globule (MG) structures.[54]

For the CSN at $T = 300$, 310, and 325 K, one major cluster and several small clusters are generated by both the MCL and the MM algorithms. All of the most populated nodes belong to the major cluster, indicating one big basin in the network. The structures of the most populated nodes in the major cluster are mainly native and MG structures.

After the free-energy basins were identified in the CSN at different temperatures, the most populated node in each basin, which represents the deepest minimum in the basin, was also found. From the information about the deepest minimum in the corresponding free-energy basin, the free-energy profiles at different temperatures can be sketched.

Coarse-Grained CSN. A coarse-grained CSN[17] is built from the original network by keeping the most populated nodes and links among these nodes. All others are deleted. In this paper, the top 10 most populated nodes are all kept in the coarse-grained CSN.

Shortest Path. Since all of the possible transitions between different conformations are captured by the network, thousands of the potential folding pathways are also presented in the CSN. In this work, the shortest path from the initial structures to the native-like structures is calculated. As a fundamental concept in graph theory,[55−57] the shortest path between two given nodes i and j is defined as a path connecting them with the shortest distance

$$d(i, j) = w_{ik} + w_{kl} + \cdots + w_{mn} + w_{nj} \quad (3)$$

where k,l,m, and n, etc. denote intermediate nodes on the path and w_{ik} is the weight of the link connecting nodes i and k. Among many algorithms, developed to solve the shortest-path problem,[58−61] Dijkstra's algorithm[59] is widely used to find the path with the lowest cost in the weighted network where the weight of the link between two nodes represents the cost of making this transition and links with high weights should be avoided in the shortest path. Newman[62] extended Dijkstra's algorithm by inverting the link weight to identify the shortest path in scientific collaboration networks, where the weight of a link represents the strength of the link between nodes, and links with high weights should now be included in the shortest path. In the conformational space network, the weight of a link represents the transition probability. Thus, the most probable pathway between the initial and the native-like structures should go along links that have as high a weight as possible. Therefore, the shortest path between a given pair of nodes identified by the implementation of Dijkstra's algorithm in ref 62 corresponds to the most probable pathway between these

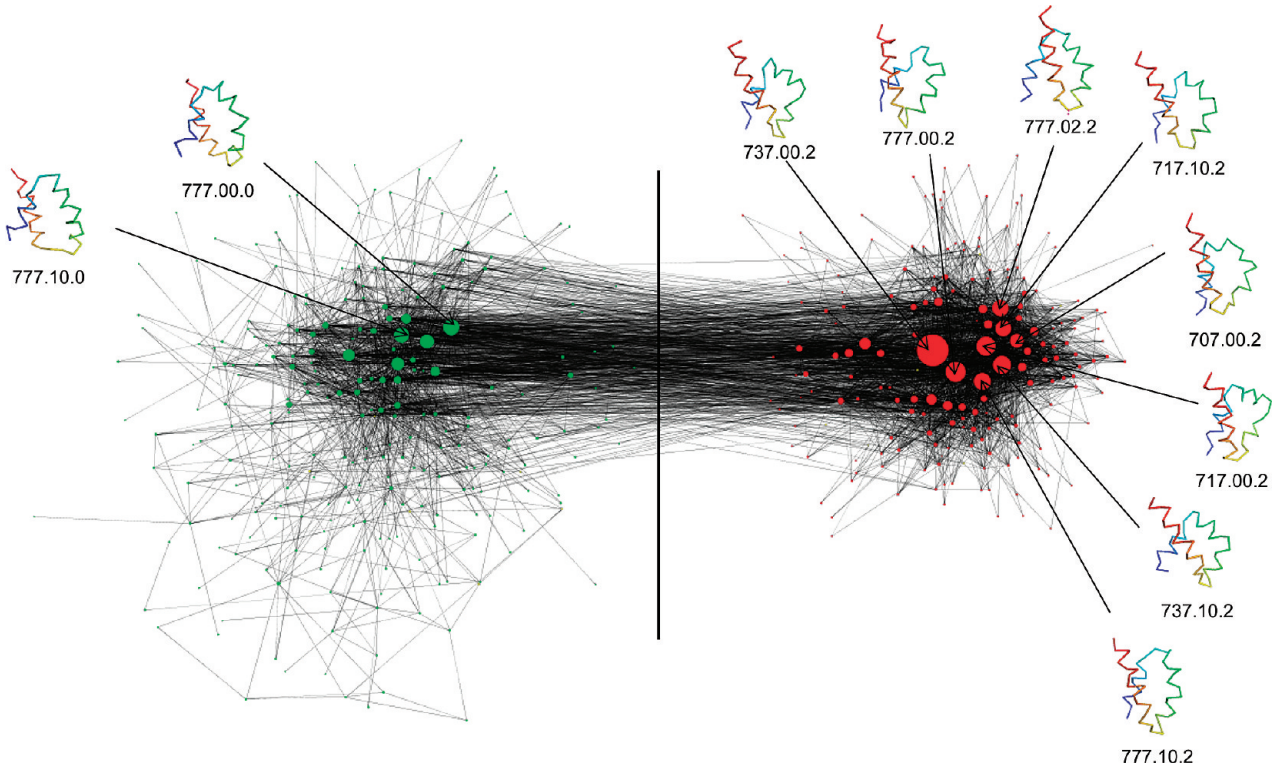

Figure 1. The 1BDD conformational space network with one metastable state (set of conformations linked by solid lines to green circles) and one basin (with examples of native-like conformations linked to red circles) at T = 270 K. The circles are nodes that represent conformations, and the very thin lines are links that represent transitions between conformations. The green nodes represent high-rmsd structures (higher than a chosen cutoff of 5.0 Å from the native), while the red nodes represent structures with a low rmsd, i.e., "rmsd match" (lower than the chosen cutoff). The magnitudes of the rmsd's are not indicated in the diagram. The size of the nodes is proportional to their population. The representative conformations and the class codes of the 10 most populated nodes are shown. The network is plotted by using the network visualization software VISONE (http://visone.info). The network visualization program is designed to show the size and location of the nodes (not only the red and green circles but also the numerous small data for low-populated conformations) in the network in a way that is easier for visual inspection. The distance between the metastable state and the native basin has been increased artificially on both sides of the vertical line to facilitate easy visualizations.

two nodes in the CSN. The distance along a path between node i and j in ref 62 is defined as

$$d(i, j) = \frac{1}{w_{ik}} + \frac{1}{w_{kl}} + \cdots + \frac{1}{w_{mn}} + \frac{1}{w_{nj}} \quad (4)$$

Consequently, $d(i,j)$ is lowest for links with high weights, thereby defining the most probable pathway. In this work, we applied Newman's implementation of Dijkstra's algorithm to identify the shortest path between the initial and the native-like structures in the CSN at different temperatures.

RESULTS AND DISCUSSION

Use of CSN To Analyze UNRES Trajectories. The representative structure of the top 10 most populated nodes at T = 270 K are listed in Figure 2A. The representative conformations, the most probable rmsd values (corresponding to the peak of the rmsd distribution), and the populations of the top 10 nodes with corresponding class code are also shown in Figure 2A. The coarse-grained network at T = 270 K is shown in Figure 2B. The nodes with native-like structures are shown as red-colored circles and labeled with the letter N and a number. The nodes with MG structures are shown as yellow-colored circles and labeled with the letters MG and a number. The nodes with topological mirror-image conformation are shown as green-colored circles and labeled with the letter D and a number. Each label corresponds to structures with a certain class code. For example, node N1, which is the most populated in the simulations at 270 K, corresponds to native-like structures with class code 737.00.2. The sizes of the nodes are proportional to their populations. The top 10 most populated nodes are included in the coarse-grained network. The top 10 nodes account for 84.92% of the total number of conformations. Eight of the top 10 nodes belong to the native basin. The other two, D1 and D2, of the top 10 nodes are mirror-image structures with only a 3.71% contribution to the total number of conformations. The class code for D1 is 777.00.0, which means that the structures in D1 have three fully-formed helices, no native packing, and no rmsd match. The class code for D2 is 777.10.0, which means that the structures in D2 have three fully-formed helices, a native packing after a sequence shift between the N-terminal helix and the middle helix, but no rmsd match. Both D1 and D2 contain mirror image and MG structures, but the number of mirror image structures is 5 times (in D1) and 10 times (in D2) higher, respectively, than that of the MG structures. Therefore, D1 and D2 are actually kinetic traps in the folding process at T = 270 K.

The cluster-detection algorithms identified one metastable state and one large basin, as shown in Figure 1. The metastable state is formed by mirror-image conformations (D1 and D2 nodes), and the large basin is formed by nodes with native-like structures. Among several folding pathways from the fully-unfolded conformation to the native state illustrated in the

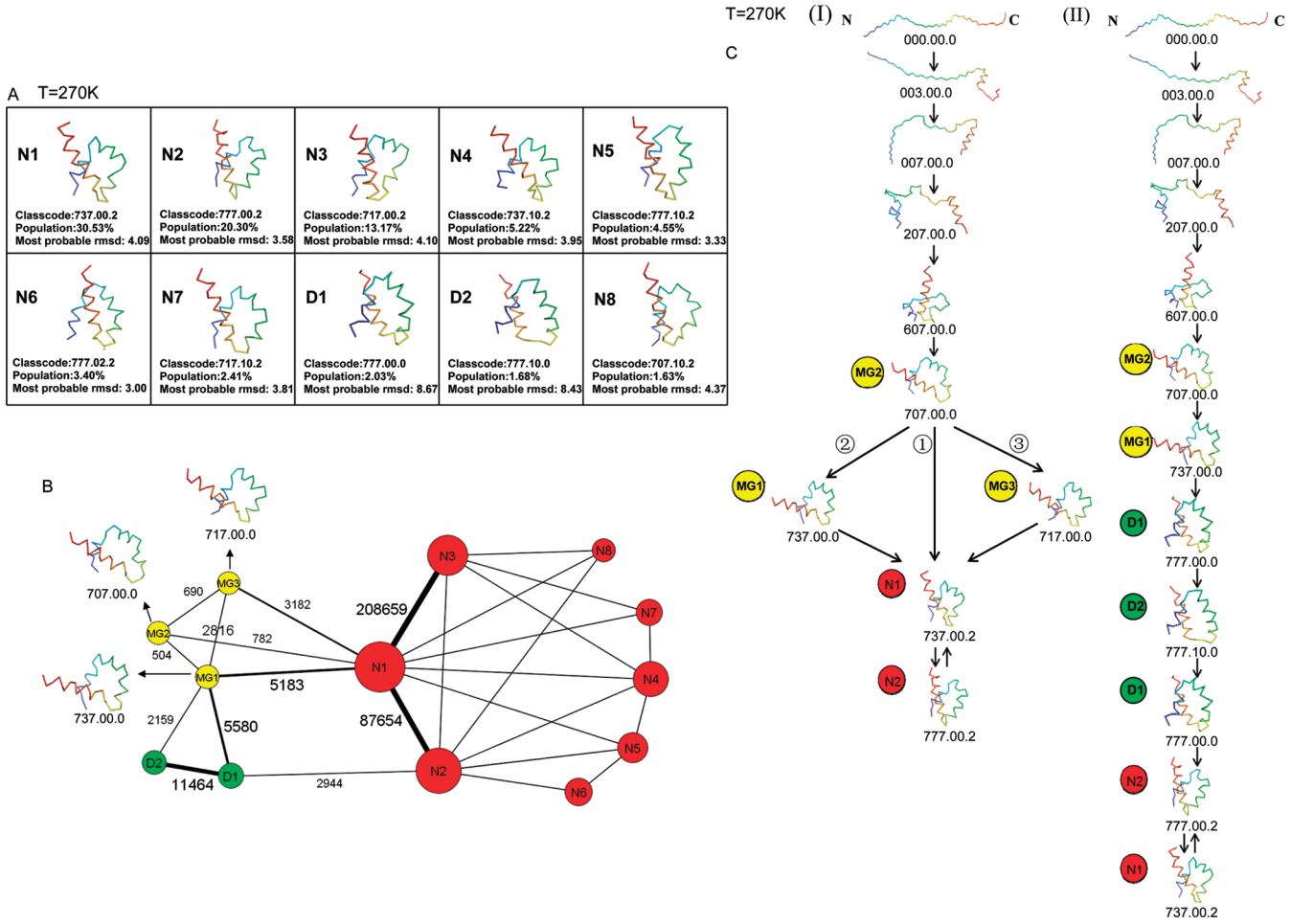

Figure 2. (A) Representative structures of the top 10 most populated nodes of Figure 1 at T = 270 K. (B) The coarse-grained CSN at T = 270 K. Nodes with native-like structures are in red-colored circles and labeled with the letter N and a conformation number. Nodes with MG structures are in yellow-colored circles and labeled with the letters MG and a conformation number. Nodes with mirror-image structures are in green-colored circles and labeled with the letter D and a conformation number. The size of a node is proportional to its population. The number of transitions of the most probable links within the metastable state and the native basin, and between the metastable state and the native basin, are also shown. The transition N1↔N3 is the one with the highest number 208659 in the native basin, and the transition D1↔D2 is the one with the highest number 11464 in the metastable state. The transition MG1↔N1 is the one with the highest number 5183 connecting the metastable state and the native basin. The lengths of the links do not reflect the relative numbers of transitions. (C) Possible folding pathways from the initial structures (with the N- and C-terminal residues labeled N and C) to the native-like structure with (II) and without (I) going through the kinetic traps at T = 270 K. D1 and D2, which contain mainly mirror-image structures, are the kinetic traps. In both type I and II pathways, the path from the initial structure to one of the MG structures, MG2, was obtained by using the shortest-path algorithm (see Methods Section). Among the paths from MG2 to one of the native-like structures, path ① in type I was generated by the shortest-path algorithm, and all the other paths could be identified from the coarse-grained CSN; see (2B).

coarse-grained CSN in Figure 2B, two main types of pathways can be identified (Figure 2C). In pathway type I [Figure 2C(I)], the protein folds without being trapped in a metastable state formed by the mirror-image topology. Instead, the protein first reaches the MG structure (MG2), forming the C-terminal helix first and the N-terminal helix later. The pathway from the initial structure to MG2 was calculated by the shortest-path algorithm. After reaching MG2, there are three possible pathways [labeled ①, ②, and ③ in Figure 2C(I)] to reach the native state. Path ① is the shortest one and goes directly from MG2 to N1 [MG2→N1↔N2, in Figure 2C(I)①], in which the middle helix is not initially formed and begins to form only after the native basin is reached. When following path ② MG2→MG1→N1↔N2 [in Figure 2C(I)②], the system goes from MG2 to MG1, first forming half of the middle helix and then jumping to the native basin with lower rmsd and with the half-formed middle helix retained. Path ③ starts with the transition from MG2 to MG3, which results in partial formation of the middle helix, and then the system jumps to the native basin [MG2→MG3→N1↔N2, in Figure 2C(I)③]. Pathway type II, from the initial fully-unfolded conformation to the native state, is shown in Figure 2C(II), in which the protein folds through the kinetic trap. After reaching MG2, the protein follows the MG2→MG1→D1↔D2↔D1 pathway, forming the middle helix in the kinetic trap before jumping to the native basin (D1→N2↔N1). It should be noted that, unlike the first and third helices, the middle helix of 1BDD is not very stable in the native basin.

The representative structures of the top 10 most populated nodes at T = 280 K are listed in Figure 3A. The top 10 nodes account for 81.45% of the total conformations. Among the top 10 populated nodes, only D1 does not contain native-like structures. The nodes D1 and D2 at 280 K are also found to contain mainly the conformations with mirror-image topology

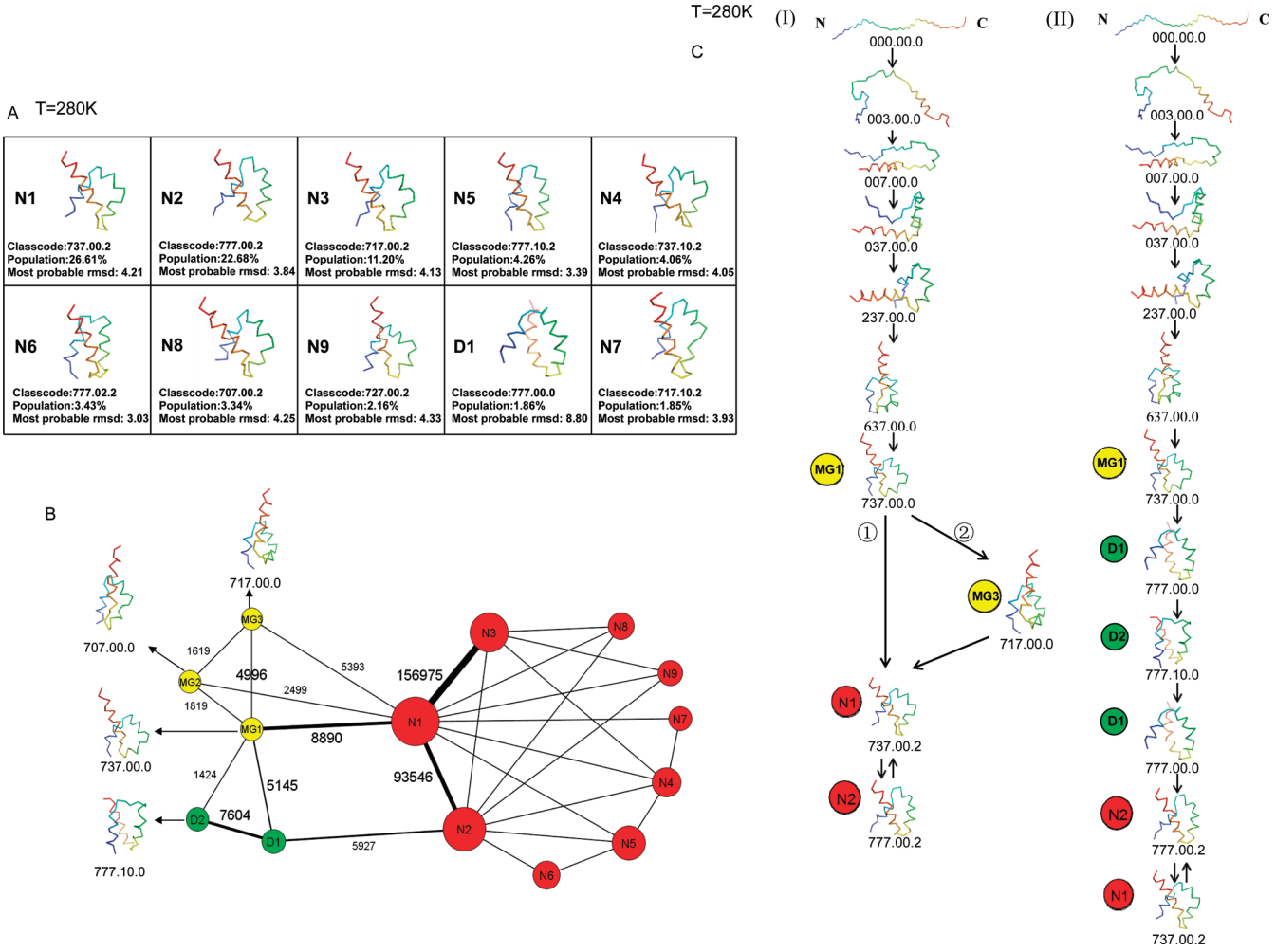

Figure 3. (A) and (B) same as Figure 2 but at T = 280 K. (C) Same as Figure 2C but at T = 280 K. In both type I and II pathways, the path from the initial structure to MG1 was obtained by using the shortest-path algorithm. Among the paths from MG1 to one of the native-like structures, path ① in type I was generated by the shortest-path algorithm, and all the other paths could be identified from the coarse-grained CSN; see (3B).

and fewer MG structures. As at 270 K, there are one metastable state and one basin at T = 280 K: the metastable state is formed by nodes D1 and D2, and the basin is formed by nodes containing native-like structures. The coarse-grained network at T = 280 K, illustrated in Figure 3B and showing several possible pathways, is very similar to that at 270 K. There are two different types of pathways from the fully-unfolded conformation to the native state at 280 K. Figure 3C(I) shows pathway type I without going through kinetic traps. The pathway from the initial structure to the MG structure MG1 is the shortest path between these two nodes. The C-terminal helix forms first, and the N-terminal helix forms fully later when reaching MG1. After reaching MG1, it either takes the shortest path (path ①) to the native-like structure by jumping to the native basin with a half-formed middle helix [MG1→N1↔N2, in Figure 3C(I)①], or first goes to MG3 (path ②), where the formation of the middle helix is degraded and then to the native basin [MG1→MG3→N1↔N2, in Figure 3C(I)②]. Figure 3C(II) shows pathway type II going through the kinetic trap. In particular, after reaching MG1 the system takes the pathway MG1→D1↔D2↔D1 forming the middle helix completely before jumping to the native basin (D1→N2↔N1).

The representative structures of the top 10 nodes and the coarse-grained network at T = 300 K are illustrated in Figure 4A and B, respectively. The top 10 nodes account for 53.36% of the total conformations and consist of three MG and seven native-like structures. No kinetic traps are found at T = 300 K. The MG and native-like structures are found in one large basin by the cluster-detection algorithm. Among several pathways, illustrated in the course-grained CSN in Figure 4B, Figure 4C shows three of the possible pathways from the initial fully-unfolded structure to the native-like structure. The pathway from the initial structure to MG5 was calculated by the shortest-path algorithm. This pathway shows that, unlike the pathways at 270 and 280 K, before reaching the MG5 structure, the protein first forms the C-terminal helix and then the middle helix. After reaching MG5, the system either takes the shortest path (path ①) to the native-like structure by jumping to the native-like node N10 with a not-fully-formed N-terminal helix (which becomes fully-formed in the native basin) [MG5→N10→N2, in Figure 4C ①], or first goes to MG6 from MG5 with almost complete formation of the first helix and then jumps to N2 [MG5→MG6→N2, path ② in Figure 4C], or first goes from MG5 to MG6 and then to MG4 forming the full N-terminal helix and later jumps to the native basin [MG5→MG6→MG4→N2, path ③ in Figure 4C].

Figure 5A shows the representative structures of the top 10 nodes at T = 310 K and Figure 5B shows the course-grained

dx.doi.org/10.1021/ct200806n | J. Chem. Theory Comput. 2012, 8, 1176–1189

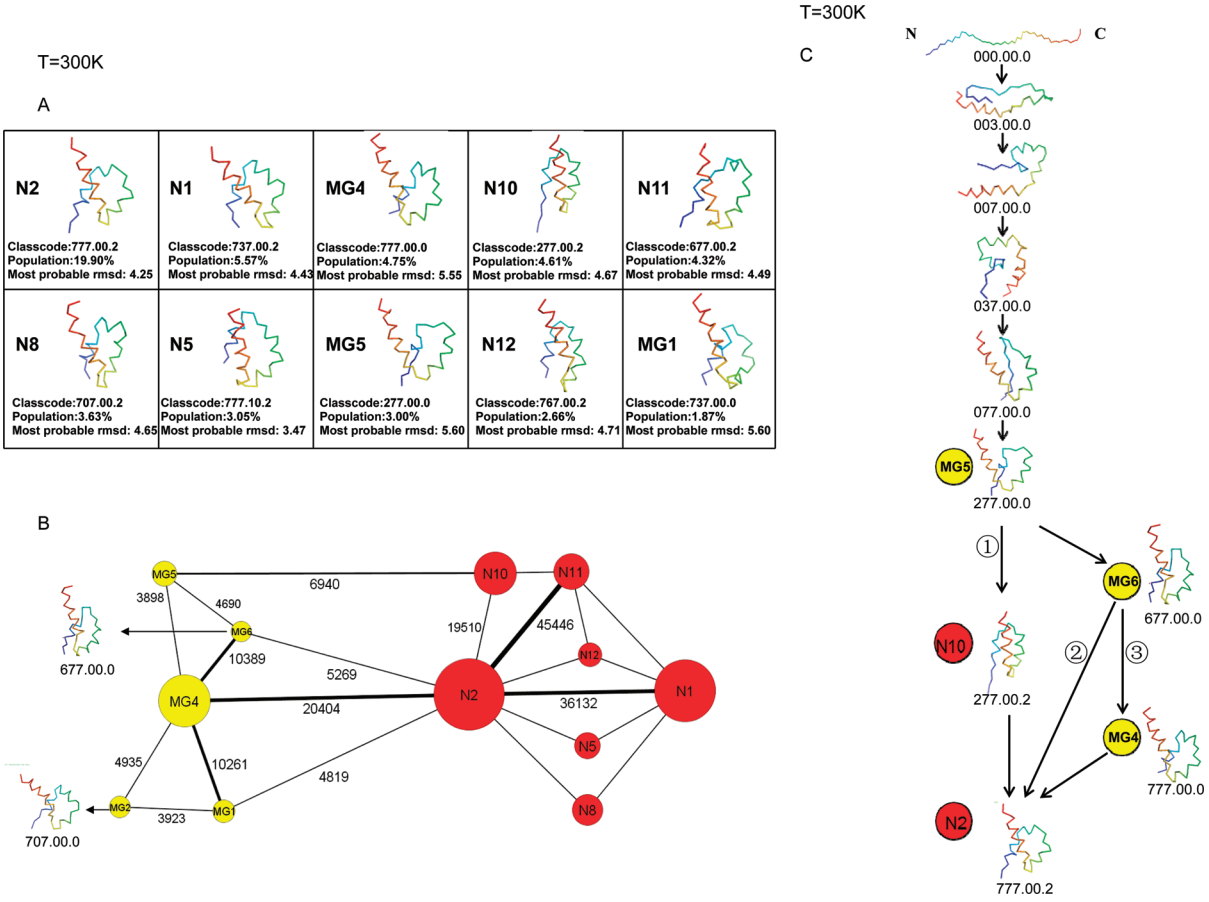

Figure 4. (A) and (B) same as Figure 2 but at T = 300 K. (C) Same as Figure 2C but at T = 300 K, and without kinetic traps. The path from the initial structure to MG5 was obtained by using the shortest-path algorithm. Among the paths from MG5 to N2, path ① was generated by the shortest-path algorithm, and paths ② and ③ could be identified from the coarse-grained CSN; see (4B).

network at T = 310 K. The top 10 nodes account for 41.11% of the total conformations and consist of six MG structures and four native-like structures. There is no kinetic trap at T = 310 K. The MG structures and the native-like structures lie in one large basin. Among several pathways, illustrated in the coarse-grained CSN in Figure 5B, two of the possible pathways from the initial to native-like structures are shown in Figure 5C. The shortest path between the fully-unfolded structure and the MG structure MG5 shows that the N-terminal helix starts to partially form early; however, the C-terminal helix fully forms first followed by the middle helix. After reaching MG5, the system either takes the shortest path (path ①) to the native-like structure by jumping directly to the native-like node N10 with a not-fully-formed N-terminal helix (which later forms fully in the native state) [MG5→N10→N2, in Figure 5C ①]. After reaching MG5, the other possible path ② goes from MG5→MG6→MG4, fully forming the N-terminal helix and then jumping to the native state [MG4→N2, in Figure 5C ②].

The representative structures of the top 10 most populated nodes and the coarse-grained CSN at the folding-transition temperature 325 K are shown in Figure 6A and B, respectively. No native minimum is found among the top 10 most populated nodes. All of the most populated minima contain either unfolded or partially ("residually") folded structures. D3 (class code 200.00.0) is the most populated node in the network. D4 is the second most populated node with class code 000.00.0, which is also the class code of the initial structure in the simulation. Therefore, the initial structures are grouped into D4. The protein goes from the initial structure to the deepest minimum by following the transition D4→D3, which has the highest number of transitions in the whole network at this temperature. This finding that, at the folding-transition temperature (325 K), protein A does not adopt its native, three-dimensional folded conformation and the only conformations found in the most populated nodes are conformations with only residual secondary structure, contradicts one of the widely used models for the description of single-domain protein folding— the two-state model.[63]

Effect of Temperature on Folding Pathways. From the coarse-grained networks at different temperatures, it can be seen that mirror-image topology appears more frequently at lower temperature (270 and 280 K). Therefore, at these temperatures, protein A folds either directly using a downhill folding scenario or through the kinetic trap. With increasing temperature (300 and 310 K), the kinetic trap disappears, and hence the protein folds by downhill folding. Also, with increasing temperature, protein A gradually becomes unfolded, and at the folding-transition temperature (325 K), the conformational ensemble of protein A is a collection of residually folded structures. This observation is consistent with our previous studies.[8,45] In order to show these changes with temperature, Figure 7 sketches the free-energy profiles for different temperatures. The representative structures with corresponding class codes of the most populated nodes at different temperatures are also shown in Figure 7. The most populated nodes at T = 270 and 280 K have the same class

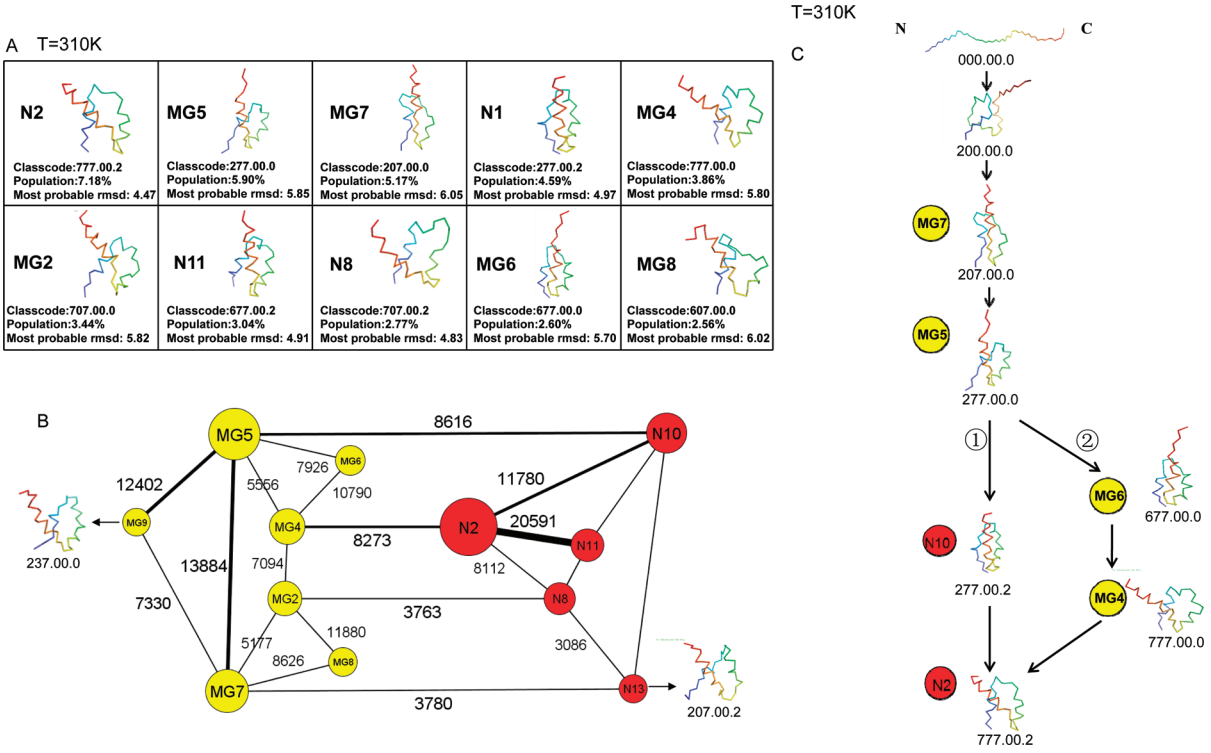

Figure 5. (A) and (B) Same as Figure 2 but at T = 310 K. (C) Same as Figure 2C but at T = 310 K, and without kinetic traps. The path from the initial structure to MG5 was obtained by using the shortest-path algorithm. Among the paths from MG5 to N2, path ① was generated by the shortest-path algorithm, and path ② could be identified from the coarse-grained CSN; see (5B).

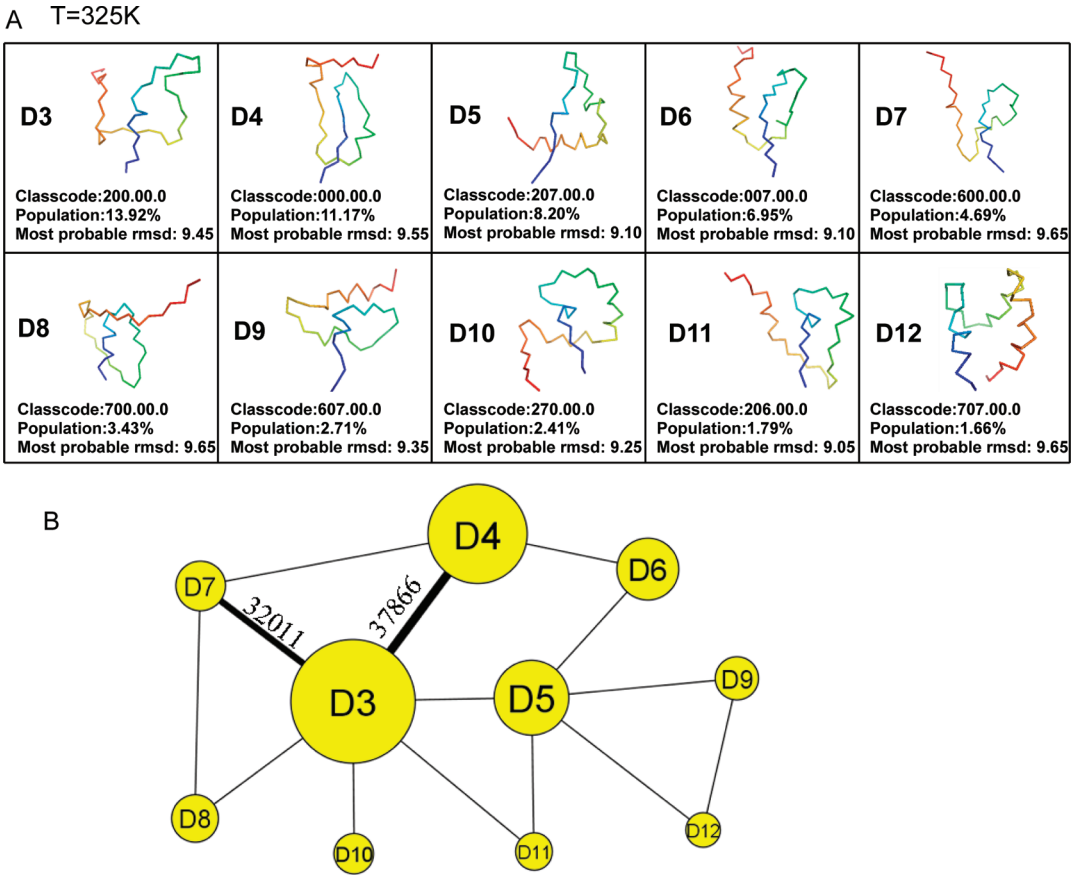

Figure 6. (A) and (B) same as Figure 2 but at T = 325 K. The nodes with residually folded structures are in yellow-colored circles and labeled with the letter D and a conformation number.

dx.doi.org/10.1021/ct200806n | J. Chem. Theory Comput. 2012, 8, 1176–1189

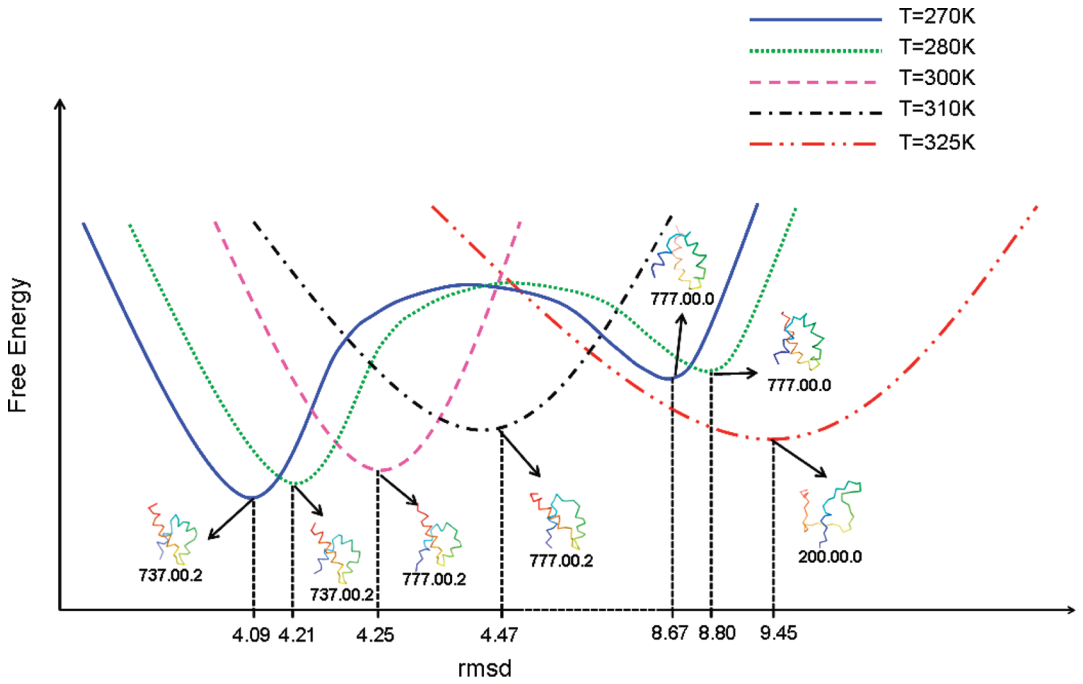

Figure 7. Structures at minima of free energy and a sketch of free energy profiles (see Methods Section) at T = 270, 280, 300, 310, and 325 K. Two minima are shown at 270 and 280 K; the higher ones being the kinetic trap.

code, and the most populated nodes at T = 300 and 310 K have the same class code. The most populated node at T = 325 K contains only a partially folded (or "residually folded") structure. With increasing temperature, the most probable rmsd of the most populated node is shifted to higher values.

Figure 8 shows plots of the helix content $\bar{Q}$[64] of three helices averaged over the 16 trajectories at T = 270, 280, 300, 310, and 325 K. It can be seen that, at all temperatures, the initial formation (in the first few MD steps) of the N-terminal helix is a bit faster than that of the middle and C-terminal helices, but the helix content of the C-terminal helix overgrows that of the N-terminal helix later, and the maximum helix content of the C-terminal helix is higher than that of middle and N-terminal helices. Figure 8a and b shows that the middle helix starts forming later than the two end helices at T = 270 and 280 K (from 0.05 to 0.5 ns). The middle and N-terminal helices seem to compete in formation at lower temperatures, especially at 270 K; consequently, it is hard to distinguish the differences between the speed of formation of these helices from the plot of averaged $\bar{Q}$. However, it should be noted that, after 0.5 ns, the middle helix starts forming faster and surpasses the N-terminal helix, although the maximum helix content of the middle helix and the N-terminal helix is still very close. To determine the order of helix full-formation, the folding pathways (along with helix content) for each trajectory were examined one by one (rather than in terms of the average value $\bar{Q}$). The significant number of folding pathways with different orders of helix formation was observed at 270 and 280 K; however, the order of formation of full helices: C- and N-terminal helices and the middle helix is slightly more probable than the other orders. This scenario of formation of the helices observed in Figure 8a and b agrees with the shortest path found at 270 and 280 K (Figures 2C and 3C).

It can be seen from Figure 8c and d that, even though the N-terminal helix starts forming faster than the middle helix at the beginning, the helix content of the middle helix quickly overgrows that of the N-terminal helix, which shows that the whole middle helix forms before the N-terminal helix. Therefore, at T = 300 and 310 K, the most probable order of formation of helices is the C-terminal, the middle, and the N-terminal. This order of formation of the three helices agrees with the shortest path found at T = 300 and 310 K (Figures 4C and 5C). At all temperatures, the order of formation of helices, reflected in the plot of the helix content (Figure 8a-d), is consistent with the shortest path found in the CSN, which demonstrates that the shortest path identified by using the shortest-path algorithm corresponds to the most probable folding pathway in the conformational space network.

Figure 8e shows that none of the three helices are fully formed at the melting temperature of 325 K. Based on the results illustrated in Figure 8, the order of stability of helices at all temperatures is the following: C-terminal, middle, and N-terminal. This sequence of stability is in harmony with earlier experimental data.[30]

From the most probable pathways at T = 270, 280, 300, and 310 K, it can be seen that the C-terminal helix always fully forms first at these temperatures, which agrees with some of the earlier experimental[30] and theoretical[31,34−36,41,43,44] studies. Investigating the folding of protein A by different experimental methods, Bai et al.[30] detected early formation of the C-terminal and middle helices. Alonso and Daggett[31] studied the unfolding process of protein A with all-atom MD simulations in explicit solvent, and found that the C-terminal helix denatured later than the N-terminal and middle helices, which suggested that the C-terminal helix forms earlier than the other two helices. Ghosh et al.[34] computed the folding pathways of protein A with the stochastic difference equation and found that the C-terminal helix forms first, followed by the N-terminal and middle helices. Jang et al.[35] investigated the folding process of protein A with all-atom MD simulation in implicit solvent and found that the C-terminal helix forms first in the early stage of folding. Garcia and Onuchic[36] performed all-atom replica

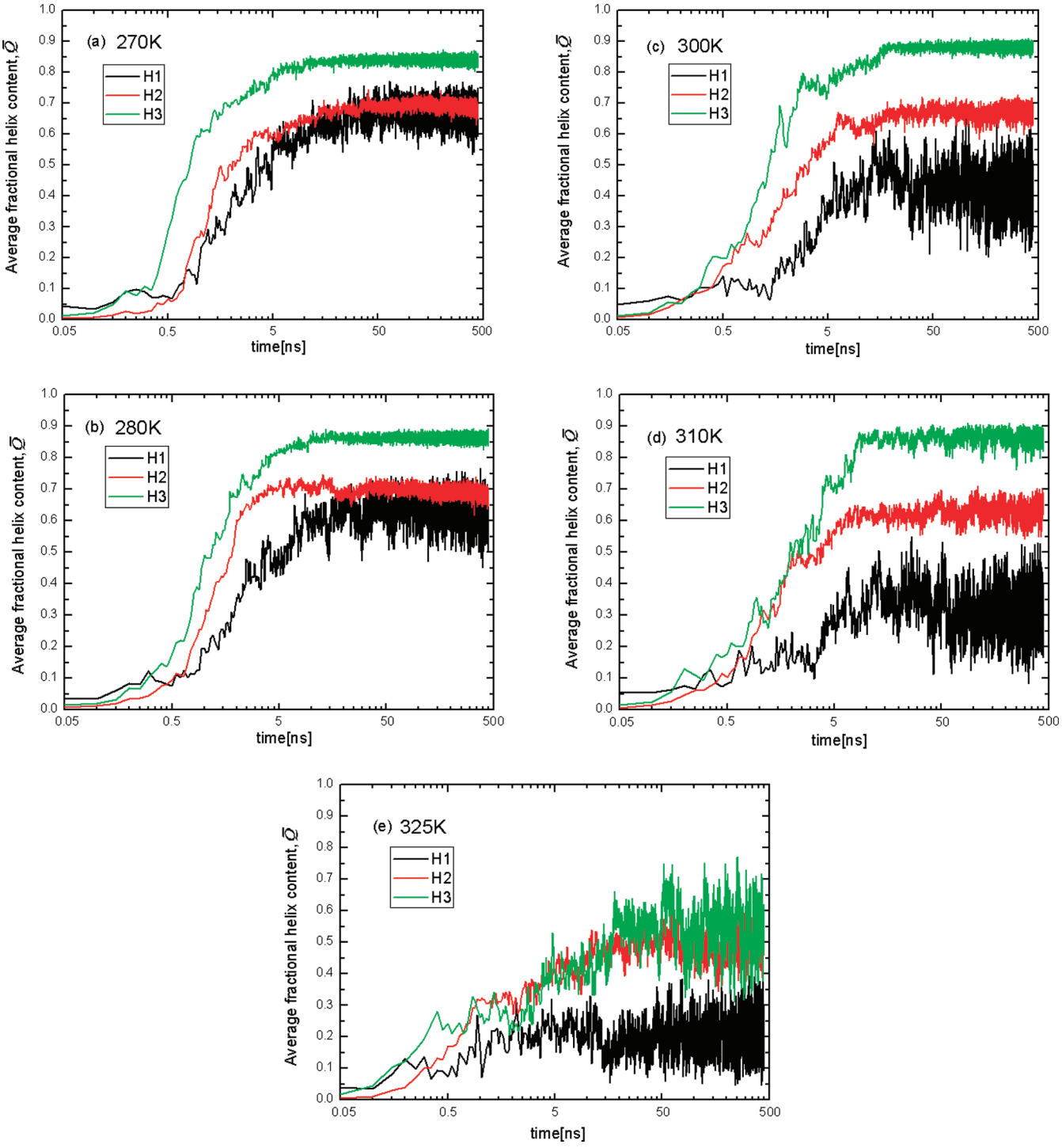

Figure 8. Plots of the averaged helix content ($\bar{Q}$) of the N-terminal helix (H1), middle helix (H2), and C-terminal helix (H3), averaged over the 16 trajectories at T = 270 (a), 280 (b), 300 (c), 310 (d), and 325 K(e).

exchange MD simulations with an explicit solvent to study the folding mechanism of protein A and found that the C-terminal helix forms first, followed by the middle helix and then the N-terminal helix. Khalili et al.[41,43] carried out MD simulations with the UNRES force field to study the folding pathways of protein A and found that the order of helix formation is the C-terminal, the N-terminal, and the middle helices. Jagielska and Scheraga[44] used all-atom MD simulations in implicit solvent at different temperatures to investigate the folding of protein A and found that the middle helix forms significantly later than the C-terminal helix and later than the N-terminal helix at lower temperature. They also found that, with increasing temperature, the speed of formation of the middle helix increases, and at higher temperatures, the middle helix forms right after the formation of the other two helices. Since the same force field was used at each temperature in ref 44, the authors were able to conclude that the order of helix formation in protein A depends on the temperature used in the experimental and theoretical studies with a given force field.

dx.doi.org/10.1021/ct200806n | J. Chem. Theory Comput. 2012, 8, 1176–1189

However, the folding pathways found in this paper do not agree with some other experimental[39] and theoretical[28,29,42] studies, which observed the early formation of the middle helix. Sato et al.[39] used experimental Φ values to analyze the transition state for the folding of protein A and found that the middle helix forms before the C- and N-terminal helices. Brooks and co-workers[28,29] examined the free-energy landscape of protein A with umbrella sampling and concluded that the C-terminal helix forms after the formation of the N-terminal and middle helices. Cheng et al.[42] carried out all-atom MD simulations for the folding of protein A in implicit solvent and found that the middle helix forms first, followed by the N-terminal helix, and then the C-terminal helix.

It should be noted that, in the present work, the middle helix was found to form before the N- and C-terminal helices in several trajectories at T = 300 and 310 K, but this pathway was not the dominant one in the 16 trajectories from which the CSN were built; consequently, it was not identified by the shortest-path algorithm. Figure 8 illustrates the tendency of the middle helix to be formed faster with the increase of temperature. Also, as concluded in ref 44, the folding pathways can change with the change of temperature. Since the temperature has a different meaning for each force field and different force fields were used in the computations of refs 28, 29, 42, and 44, it is not possible to attribute the discrepancies among refs 28, 29, 42, and 44 simply to possible differences in temperature.

The rmsd of protein A (averaged over 16 trajectories for each temperature) as a function of average helix content $\bar{Q}$ over three helices (and over 16 trajectories for each temperature) illustrates the coupling between secondary and tertiary structure formation (see Figure 9). In particular, the rmsd

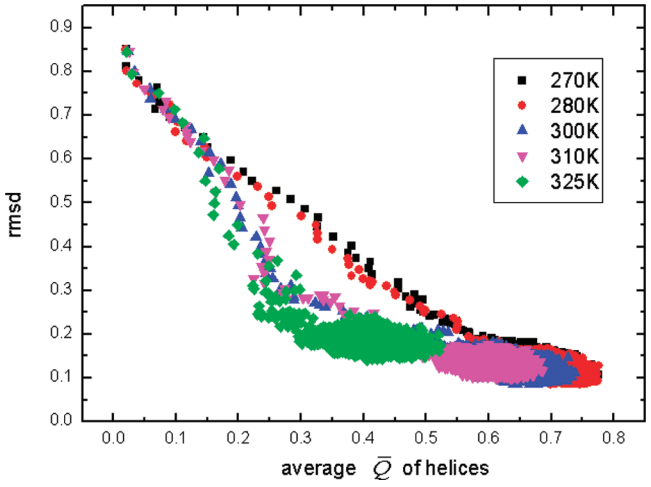

Figure 9. Plot of rmsd of the whole molecule (averaged over 16 trajectories for each temperature) vs average helix content ($\bar{Q}$) over three helices (and over 16 trajectories for each temperature). The rmsd is divided by the number of residues to keep it in the same range as $\bar{Q}$.

decreases with growing $\bar{Q}$ at all temperatures, thus secondary and tertiary structures form simultaneously, which is consistent with the results reported earlier.[31,36,39,42,44] It should be noted that, at lower temperatures (T = 270 and 280 K), rmsd decreases linearly with increasing $\bar{Q}$, whereas at higher temperatures (T = 300, 310, and 325 K), rmsd decreases quickly first and then slowly later. Such behavior of rmsd at higher temperatures indicates that, with the increase of temperature, the tertiary structure starts forming faster than a secondary structure.

CONCLUSIONS

In spite of many studies performed on protein A, to the best of our knowledge, the folding dynamics of protein A was investigated by an unbiased approach, the conformational space network, for the first time, enabling us to identify the large spectrum of folding pathways, hidden in the FELs along the order parameters, at different temperatures. Moreover, it was shown that the folding pathway changes with temperature for a given force field, as also concluded in ref 44.

In detail, our findings are the following:

(i) At lower temperatures (270 and 280 K), protein A can fold either directly following the downhill folding scenario or through an indirect route involving an intermediate (kinetic trap). The order of full formation of helices calculated by the shortest-path algorithm during the folding dynamics at T = 270 and 280 K is the following: C-terminal, N-terminal, and middle helices, which agrees with the pathway shown by the plot of the time dependence of the helix content of the three helices. However, it should be noted that the formation of the N-terminal and middle helices occurs almost simultaneously; even in the interval of 0.5−5 ns, the middle helix forms faster than the N-terminal helix. In downhill folding, the middle helix fully forms after the protein jumps to the native basin. In the folding pathway through the kinetic trap, the middle helix forms in the kinetic trap.

(ii) The kinetic trap disappears at higher temperatures (300 and 310 K), and protein A follows downhill folding. The order of formation of helices calculated by the shortest-path algorithm during the folding dynamics at T = 300 and 310 K is the following: C-terminal, middle, and N-terminal helices, which is different from the order at lower temperature. This pathway agrees with the pathway shown by the plot of the time dependence of the helix content of the three helices. At higher temperatures, the protein does not always jump to the native basin with all helices fully-formed, which is also observed at lower temperatures. In particular, the N-terminal helix forms fully either in the native basin or before reaching the native basin at higher temperatures.

(iii) At the folding-transition temperature (325 K), none of the three helices are fully formed, and the conformational ensemble of protein A is a collection of residually folded structures and not a 50−50% mixture of native and non-native conformations.[45]

APPENDIX

The structures are classified based on levels of description (Table 1[19]). Level 1 describes elementary fragments, which are usually identified as elementary units with defined secondary structure in the experimental structure (e.g., single α-helices, β-strands, or β-hairpins, etc.) or are characterizable by other means (e.g., loops). For protein A, the elementary fragments are α-helices.

The elementary fragments are defined as follows:

(1) An α-helix is a fragment in which all of the virtual-bond dihedral angles γ are within $30^\circ \leq \gamma \leq 60^\circ$ and every

dx.doi.org/10.1021/ct200806n | J. Chem. Theory Comput. 2012, 8, 1176−1189

peptide group is in electrostatic contact with its third neighbor. Two peptide groups are considered to be in electrostatic contact if their average electrostatic interaction energy is lower than −0.3 kcal/mol.

(2) A two-stranded antiparallel β-sheet is a fragment in which all virtual-bond dihedral angles γ, except those at turn residues, are greater in absolute value than 90° and an electrostatic-contact pattern characteristic of an antiparallel β-sheet is observed (i.e., if peptide group i is in electrostatic contact with peptide group j, then peptide group $i + 1$ is in electrostatic contact with peptide group $j - 1$, etc.). This type of element can involve either a contiguous part of the chain (a β-hairpin) or a noncontiguous part.

(3) A two-stranded parallel β-sheet is a fragment in which all virtual-bond dihedral angles γ are greater in absolute value than 90° and an electrostatic-contact pattern characteristic of a parallel β-sheet is observed (i.e., if peptide group i is in electrostatic contact with peptide group j, then peptide group $i + 1$ is in electrostatic contact with peptide group $j + 1$, etc.). This type of element always involves two noncontiguous parts of the chain.

(4) A strand is a fragment in which all virtual-bond dihedral angles γ are greater in absolute value than 90° and an electrostatic-contact pattern characteristic of a single chain in a parallel or antiparallel β-sheet is observed.

(5) An elementary fragment with irregular structure is identified based on the values of the virtual-bond-valence, and virtual-bond-dihedral angles as well as the local geometry of the side-chain center.

The above definitions do not exhaust all possibilities; one is free to define other types of structural elements, such as, for example, a 3_{10}-helix, a β-helix, or a collagen helix.

The elementary fragments are compared as follows: (1) The secondary structure is considered native if at least 70% of the chain of the considered conformation has the same secondary structure as in the native conformation. In Table 1,[19] a "1" in level 1 corresponds to native secondary structure. (2) The hydrogen-bonding contact pattern is considered native if the number of contacts between the peptide groups in the compared structure matching the native peptide group contacts is greater than 70% of the native contacts (this is called a match). Shifting the sequence by ±3 residues is allowed to obtain a match, but it results in a decreasing class number. In a β-hairpin, such a shift corresponds to shifting the position of the β-turn. In Table 1, a "2" in level 1 corresponds to native hydrogen-bonding contacts after a sequence shift, and a "6" in level 1 corresponds to native hydrogen-bonding contacts only without a sequence shift.

Level 2 consists of pairs of elementary fragments. Each class number in level 2 represents the packing between a given pair of elementary fragments. In level 2 of protein A, only two class numbers (each of which could be any of the three numbers in level 2 of Table 1) are used. The first number corresponds to the packing between the N- and C-terminal helices. The second number corresponds to the packing between the middle and C-terminal helices.

The packing of elementary fragments is compared as follows: (1) The number of side-chain contacts (for helix-to-helix and helix-to-strand packing) or the number of peptide group contacts (for β-strand packing) corresponding to the native contact between a given pair of fragments is computed. If it is greater than 70% of the native contacts, the packing is considered native. Shifting the sequence by ±3 residues is allowed to obtain a match. (2) The rmsd of the segment consisting of the compared pair of fragments from the corresponding fragment of the experimental structure is computed. If the rmsd is less than the threshold value (0.1 Å per residue), the segment is considered to be geometrically conformable with the native segment.

If conditions 1 and 2 hold, then the packing is considered to be native-like; otherwise it is considered to be non-native. In Table 1, a "2" in level 2 corresponds to native packing without a sequence shift.

Level 3 pertains to the whole molecule. A single number from the ones in level 3 of Table 1 of the class code represents the whole molecule. If the rmsd value of the whole molecule is lower than a cutoff value (5.0 Å for protein A in this work but different for each protein), a rmsd match is obtained. In Table 1, a "2" in level 3 corresponds to an rmsd match without a sequence shift.

AUTHOR INFORMATION

Corresponding Author

*has5@cornell.edu

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by grants from the National Institutes of Health (GM-14312) and the National Science Foundation (MCB-1019767) and conducted by using the resources of (a) our 588-processor Beowulf cluster at the Baker Laboratory of Chemistry and Chemical Biology, Cornell University, (b) the National Science Foundation Terascale Computing System at the Pittsburgh Supercomputer Center, (c) the John von Neumann Institute for Computing at the Central Institute for Applied Mathematics, Forschungszentrum Juelich, Germany, (d) the Beowulf cluster at the Department of Computer Science, Cornell University, (e) the Informatics Center of the Metropolitan Academic Network (IC MAN) in Gdańsk, and (f) the Interdisciplinary Center of Mathematical and Computer Modeling (ICM) at the University of Warsaw.

REFERENCES

(1) Frauenfelder, H.; Sligar, S. G.; Wolynes, P. G. The energy landscapes and motions of proteins. *Science* **1991**, *254*, 1598−1630.

(2) Wales, D. J.; Scheraga, H. A. Global optimization of clusters, crystals, and biomolecules. *Science* **1999**, *285*, 1368−1372.

(3) Brooks, C. L. III; Onuchic, J. N.; Wales, D. J. Taking a walk on a landscape. *Science* **2001**, *293*, 612−613.

(4) Wales, D. J. *Energy landscapes*; Cambridge University Press: Cambridge, U.K., 2003; p 681.

(5) Krivov, S.; Karplus, M. Hidden complexity of free energy surfaces for peptide (protein) folding. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 14766−14770.

(6) Altis, A.; Otten, M.; Nguyen, P. H..; Hegger, R.; Stock, G. Construction of the free energy landscape of biomolecules via dihedral angle principal component analysis. *J. Chem. Phys.* **2008**, *128*, 245102.

(7) Maisuradze, G. G.; Liwo, A.; Scheraga, H. A. How adequate are one- and two-dimensional free energy landscapes for protein folding dynamics? *Phys. Rev. Lett.* **2009**, *102*, 238102.

(8) Maisuradze, G. G.; Liwo, A.; Scheraga, H. A. Relation between free energy landscapes of proteins and dynamics. *J. Chem. Theory Comput.* **2010**, *6*, 583−595.

(9) Hegger, R.; Altis, A.; Nguyen, P. H.; Stock, G. How complex is the dynamics of peptide folding? *Phys. Rev. Lett.* **2007**, *98*, 028102.
(10) Zhou, R.; Berne, B. J.; Germain, R. The free energy landscape for β hairpin folding in explicit water. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 14931−14936.
(11) Zhou, R.; Parida, L.; Kapila, K.; Mudur, S. PROTERAN: animated terrain evolution for visual analysis of patterns in protein folding trajectory. *Bioinformatics* **2007**, *23*, 99−106.
(12) Newman, M. E. J. The structure and function of complex networks. *SIAM Rev.* **2003**, *45*, 167−256.
(13) Krivov, S.; Karplus, M. Free energy disconnectivity graphs: Application to peptide models. *J. Chem. Phys.* **2002**, *117*, 10894−10903.
(14) Rao, F.; Caflisch, A. The protein folding network. *J. Mol. Biol.* **2004**, *342*, 299−306.
(15) Gfeller, D.; De Los Rios, P.; Caflisch, A.; Rao, F. Complex network analysis of free-energy landscapes. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 1817−1822.
(16) Stillinger, F.; Weber, T. Hidden structure in liquids. *Phys. Rev. A* **1983**, *28*, 2408−2416.
(17) Rao, F.; Karplus, M. Protein dynamics investigated by inherent structure analysis. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 9152−9157.
(18) Liwo, A.; Arlukowicz, P.; Czaplewski, C.; Oldziej, S.; Pillardy, J.; Scheraga, H. A. A method for optimizing potential-energy functions by a hierarchical design of the potential-energy landscape: Application to the UNRES force field. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 1937−1942.
(19) Oldziej, S.; Liwo, A.; Czaplewski, C.; Pillardy, J.; Scheraga, H. A. Optimization of the UNRES force field by hierarchical design of the potential-energy landscape. 2. Off-lattice tests of the method with single proteins. *J. Phys. Chem. B* **2004**, *108*, 16934−16949.
(20) Gouda, H.; Torigoe, H.; Saito, A.; Sato, M.; Arata, Y.; Shimada, I. Three-dimensional solution structure of the B domain of staphylococcal protein A: comparisons of the solution and crystal structures. *Biochemistry* **1992**, *31*, 9665−9672.
(21) Liwo, A.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Scheraga, H. A. Prediction of protein conformation on the basis of a search for compact structure: Test on avian pancreatic polypeptide. *Protein Sci.* **1993**, *2*, 1715−1731.
(22) Liwo, A.; Oldziej, S.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Scheraga, H. A. A united-residue force field for off-lattice protein-structure simulations. I. Functional forms and parameters of long-range side-chain interaction potentials from protein crystal data. *J. Comput. Chem.* **1997**, *18*, 849−873.
(23) Liwo, A.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Oldziej, S.; Scheraga, H. A. A united-residue force field for off-lattice protein-structure simulation. II: Parameterization of local interactions and determination of the weights of energy terms by Z-score optimization. *J. Comput. Chem.* **1997**, *18*, 874−887.
(24) Ołdziej, S.; Kozłowska, U.; Liwo, A.; Scheraga, H. A. Determination of the potentials of mean force for rotation about $C^{\alpha}...C^{\alpha}$ virtual bonds in polypeptides from the *ab initio* energy surfaces of terminally-blocked glycine, alanine, and proline. *J. Phys. Chem. A* **2003**, *107*, 8035−8046.
(25) Liwo, A.; Oldziej, S.; Czaplewski, C.; Kozlowska, U.; Scheraga, H. A. Parametrization of backbone-electrostatic and multibody contributions to the UNRES force field for protein-structure prediction from ab initio energy surfaces of model systems. *J. Phys. Chem. B* **2004**, *108*, 9421−9438.
(26) Oldziej, S.; Lagiewka, J.; Liwo, A.; Czaplewski, C.; Chinchio, M.; Nanias, M.; Scheraga, H. A. Optimization of the UNRES force field by hierarchical design of the potential-energy landscape. 3. Use of many proteins in optimization. *J. Phys. Chem. B* **2004**, *108*, 16950−16959.
(27) Liwo, A.; Khalili, M.; Czaplewski, C.; Kalinowski, S.; Oldziej, S.; Wachucik, K.; Scheraga, H. A. Modification and optimization of the united-residue (UNRES) potential energy function for canonical simulations. I. Temperature dependence of the effective energy function and tests of the optimization method with single training proteins. *J. Phys. Chem. B* **2007**, *111*, 260−285.
(28) Boczko, E. M.; Brooks, C. L. III. First principles calculation of the free energy surface for folding of a three helix bundle protein. *Science* **1995**, *269*, 393−396.
(29) Guo, Z.; Brooks, C. L. III; Boczko, E. M. Exploring the folding free energy surface of a three-helix bundle protein. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 10161−10166.
(30) Bai, Y.; Karimi, A.; Dyson, H. J.; Wright, P. E. Absence of a stable intermediate on the folding pathway of Protein A (B domain). *Protein Sci.* **1997**, *6*, 1449−1457.
(31) Alonso, D. O. V.; Daggett, V. Staphylococcal protein A: Unfolding pathways, unfolded states, and differences between the B and E domains. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 133−138.
(32) Berriz, G. F.; Shakhnovich, E. I. Characterization of the folding kinetics of a three-helix bundle protein *via* a minimalist Langevin model. *J. Mol. Biol.* **2001**, *310*, 673−685.
(33) Myers, J. K.; Oas, T. G. Preorganized secondary structure as an important determinant of fast protein folding. *Nat. Struct. Biol.* **2001**, *8*, 552−558.
(34) Ghosh, A.; Elber, R.; Scheraga, H. A. An atomically detailed study of the folding pathways of protein A with the stochastic difference equation. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 10394−10398.
(35) Jang, S.; Kim, E.; Shin, S.; Pak, Y. Ab initio folding of helix bundle proteins using molecular dynamics simulations. *J. Am. Chem. Soc.* **2003**, *125*, 14841−14846.
(36) Garcia, A. E.; Onuchic, J. N. Folding a protein in a computer: An atomic description of the holding/unfolding of protein A. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 13898−13903.
(37) Vila, J. A.; Ripoll, D. R.; Scheraga, H. A. Atomically detailed folding simulation of the B domain of staphylococcal protein A from random structures. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 14812−14816.
(38) Dimitriadis, G.; Drysdale, A.; Myers, J. K.; Arora, P.; Radford, S. E.; Oas, T. G.; Smith, D. A. Microsecond folding dynamics of the F13W G29A mutant of the B domain of staphylococcal protein A by laser-induced temperature jump. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 3809−3814.
(39) Sato, S.; Religa, T. L.; Daggett, V.; Fersht, A. R. Testing protein-folding simulations by experiment: B domain of protein A. *Proc Natl. Acad. Sci. USA* **2004**, *101*, 6952−6956.
(40) Liwo, A.; Khalili, M.; Scheraga, H. A. Ab initio simulations of protein-folding pathways by molecular dynamics with the united-residue model of polypeptide chains. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 2362−2367.
(41) Khalili, M.; Liwo, A.; Jagielska, A.; Scheraga, H. A. Molecular dynamics with the United-Residue model of polypeptide chains. II. Langevin and Berendsen-Bath dynamics and tests on model α-helical systems. *J. Phys. Chem. B* **2005**, *109*, 13798−13810.
(42) Cheng, S.; Yang, Y.; Wang, W.; Liu, H. J. Transition state ensemble for the folding of B domain of protein A: A comparison of distributed molecular dynamics simulation with experiments. *J. Phys. Chem. B* **2005**, *109*, 23645−23654.
(43) Khalili, M.; Liwo, A.; Scheraga, H. A. Kinetic studies of folding of the B-domain of staphylococcal protein A with molecular dynamics and a united-residue (UNRES) model of polypeptide chains. *J. Mol. Biol.* **2006**, *355*, 536−547.
(44) Jagielska, A.; Scheraga, H. A. Influence of temperature, friction, and random forces on folding of the B-domain of staphylococcal protein A: All- atom molecular dynamics in implicit solvent. *J. Comput. Chem.* **2007**, *28*, 1068−1082.
(45) Maisuradze, G. G.; Liwo, A.; Oldziej, S.; Scheraga, H. A. Evidence, from simulations, of a single state with residual native structure at the thermal denaturation midpoint of a small globular protein. *J. Am. Chem. Soc.* **2010**, *132*, 9444−9452.
(46) Berendsen, H. J. C.; Postma, J. P. M; van Gunsteren, W. F.; Dinola, A.; Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **1984**, *81*, 3684−3690.
(47) Gibson, D.; Kleinberg, J. Raghavan, P. *Inferring web communities from link topology*; ACM Press: New York, 1998.

(48) Newman, M. E. J. The structure of scientific collaboration networks. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 404−409.
(49) Barabasi, A. L.; Oltvai, Z. N. Network biology: understanding the cell's functional organizations. *Nat. Rev. Genet.* **2004**, *5*, 101−113.
(50) Girvan, M.; Newman, M. E. J. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 7821−7826.
(51) Clauset, A.; Newman, M. E. J.; Moore, C. Finding community structure in very large networks. *Phys. Rev. E* **2004**, *70*, 066111.
(52) Van Dongen, S. Graph clustering by flow simulation. *PhD thesis*, University of Utrecht: Utrecht, The Netherlands, May 2000.
(53) Enright, A. J.; Van Dongen, S.; Ouzounis, C. A. An efficient algorithm for large scale detection of protein families. *Nucleic Acids Res.* **2002**, *30*, 1575−1584.
(54) Ohgushi, M.; Wada, A. "Molten-globule state": a compact form of globular proteins with mobile side-chains. *FEBS Lett.* **1983**, *164*, 21−24.
(55) Bondy, J. A.; Murty, U. S. R. *Graph Theory with Applications*; Macmillan: London, 1976.
(56) Harary, F. Graph Theory; Perseus: Cambridge, MA, 1995.
(57) Bollobas, B. *Modern Graph Theory*; Springer: New York, 1998.
(58) Richard, B. On a Routing Problem. *Quarterly of Applied Mathematics* **1958**, *16*, 87−90.
(59) Dijkstra, E. W. A note on two problems in connexion with graphs. *Numerische Math.* **1959**, *1*, 269−271.
(60) Floyd, R. W. Algorithm 97: Shortest Path. *Commun. ACM* **1962**, *5*, 345.
(61) Hart, P. E.; Nilsson, N. J.; Raphael, B. Correction to "A Formal Basis for the Heuristic Determination of Minimum Cost Paths". *SIGART Newsletter* **1972**, *37*, 28−29.
(62) Newman, M. E. J. Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Phys. Rev. E* **2001**, *64*, 016132.
(63) Privalov, P. L.; Khechinashvili, N. N. A thermodynamic approach to the problem of stabilization of globular protein structure: a calorimetric study. *J. Mol. Biol.* **1974**, *86*, 665−684.
(64) Eastwood, M. P.; Hardin, C.; Luthey-Schulten, Z.; Wolynes, P. G. Statistical mechanical refinement of protein structure prediction schemes: Cumulant expansion approach. *J. Chem. Phys.* **2002**, *117*, 4602−4615.