TASK QUARTERLY 18 No 3, 265–279

DETERMINING CRITICAL AMINO ACID CONTACTS FOR KNOTTED PROTEIN FOLDING

PAWEL DABROWSKI-TUMANSKI[1,2], SZYMON NIEWIECZERZAL[1] AND JOANNA I. SULKOWSKA[1,2]

[1] *Departament of Chemistry, University of Warsaw*
Pasteura 1, 02-093 Warsaw, Poland

[2] *Centre of New Technologies, University of Warsaw*
Banacha 2c, 02-097 Warsaw, Poland

(Paper presented at the CBSB14 Conference, May 25–27, 2014, Gdansk, Poland)

Abstract: Proteins with a non-trivial topological structure are currently well recognized, while a knotted protein chain represents a new motif in protein three dimensional folds. Recent comprehensive analysis of the Protein Data Base shows that knotted proteins represent 1.5% of known protein structures. Determination of a free energy landscape of knotted proteins, and its understanding provides a serious challenge for both experimentalists and theoreticians. Moreover the role of a knot for biological activity of protein still remains elusive. In this work we study the smallest knotted proteins (PDB code 2efv) to understand/investigate their free energy landscape, by means of extensive molecular dynamics simulations. We explore the dependence of the thermodynamics, kinetics and protein folding pathways on the native-likes contact maps and on the length of the chain. We analyze two sets of native-like contacts, which differ by a number of long range interactions, and we consider the 2efv protein with two different lengths of its C-terminus end. We identify the subset of native contacts sufficient to explore the entire free energy landscape. Then, we analyze the influence of the remaining set of native contacts – we show that the set of additional contacts may enhance folding kinetics, and that it has an influence on folding pathways.

Keywords: knotted protein, slipknot, 2efv folding, contact map, native, non-native

1. Introduction

Existence of proteins with a non-trivial topological structure is currently well accepted, and knots are well recognized as new motifs in protein three-dimensional folds [1–4]. The KnotProt database [5] that contains information about all the knotted and slipknotted proteins with a known three-dimensional structure indicates that currently 1184 proteins with non-trivial topology are known: 784 structures form knots, and the rest involves slipknots (we recall that

slipknots arise from threading one loop across a twisted loop, and slipknots can be always reduced to knots after eliminating some number of atoms from one terminal). This new, non-trivial fold is surprisingly well conserved in proteins with a very low sequence similarity and in organisms separated by billions of years [4]. For example, the biggest knotted family with a deep trefoil knot (at least 20 amino acids have to be removed to untie those proteins) is represented by methyltransferase, and those proteins exist in all three branches of the evolutional tree and have sequence similarity as low as 19%. Even stronger conservation of topology is observed for slipknotted membrane proteins – for example proteins with the same topological motif 5_2, 3_1, 3_1 have only 7% of sequence similarity. Those data strongly indicate that knotted proteins withstand evolutional pressure. Nonetheless, it is still not known what is the role of knotted structures in the biological activity of its hosting organism.

One of the problems on the way to understand the biological role of knots in proteins is an unexpected difficulty in untying a protein chain during thermal and chemical denaturation in vitro [6]. Currently, it is possible to observe spontaneous knotting of proteins (YibK and YbeA) when a newly translated chain from ribosome is used [7]. The unknown experimental component of the free energy landscape is the folding pathway, however, the knotting process can be explored in many ways by means of computational approaches. The folding time of known knotted proteins is around 10–20 minutes in vitro [7, 8]. This time scale implies that with the current computational speed numerical approaches have to be restricted to biased methods, or to the analysis of only a part of a protein to minimize the available conformational space. Biased methods, which promote the formation of native contacts based on the protein geometry in the native state (so called structure based models, SBM), have been successful in the analysis of proteins with trivial topology [9], however, some exceptions from its application and a perfect funnel landscape are known [10, 11]. Application of structure based models to proteins with deep knot [12–15] has shown that the main rate-limiting step in folding those proteins arises from the knot formation. Investigation of the thermodynamic behavior of those proteins is still, in general, beyond the capabilities of structure based models. However, the exotic landscape associated with a knot can be entirely explored for the smallest knotted protein, MJ0366 (PDB ID code: 2efv), from Methanocaldococcus jannaschii [3].

MJ0336 is composed of 92 amino acids, however, the structure deposited in the PDB possesses only 82 amino acids (five amino acids at each terminus are missing). The crystallized structure forms a rather shallow trefoil knot (3_1) with negative chirality [3]; 10 and 6 amino acids have to be removed from the N-terminus and the C-terminus, respectively, to untie this protein. Since the discovery of this structure much attention has been dedicated to characterizing its folding mechanism with computational methods [16–18]. In [16], authors, by means of an all-atom structure based model, determined the free energy landscape $F(Q)$, whose shape suggests a three-state system behavior. The third state, apart

from unfolded and native-like knotted states, is characterized by the native-like formed loop. The folding rate-limiting the free energy barrier is of a topological nature and is traversed in two parallel knot-forming pathways: (a) through the slipknot conformation, or (b) by threading as through a needle (a plug motion); the pathway (a) is significantly more populated than (b). The results of an all-atom unbiased explicit-solvent molecular dynamics simulations of 2efv are presented in [17]. The authors have shown that, starting from the slipknotted configuration, one can reach the native-knotted state, and that the same contacts are involved during the threading process as in an SBM model [16]. In the study by Beccara and co-workers [18], the authors have shown that the proportion between the slipknotted and needle-like threading events can be modified by taking non-native contacts into account. Although this protein has already been explored to a large extent, some fundamental questions that could lead to the identification of a proper model to explore free energy landscape of proteins with a deep knots have still not been raised.

One of the current challenges in the field of knotted proteins is to identify the set of native-like/non-native contacts, which implemented in the structure based model will provide a driving force that will allow us to effectively fold deeply knotted proteins. The results of [12, 19–21] show that it is possible to add non-native interactions to accelerate kinetics of knotting. However, neither the thermodynamics of the investigated proteins from the viewpoint of the contact map nor a comparison between these two contact maps have been studied previously.

Here, based on the MJ0336 protein, we explore the robustness of thermodynamics, kinetics and folding pathways from the perspective of native-like contacts. We analyze two sets of native-like contacts (using two different methods to identify native-like contacts, Shadow [22] and Tsai [23]), and two lengths of proteins, 82 and 87 amino acids. Firstly, we identify a subset of native-like contacts which are sufficient to explore the entire free energy landscape of knotted proteins. Secondly, we present a subset of non-native contacts or native-like contacts if a different definition of establishing a native contact is taken into account, which dramatically enhances folding kinetics. Additionally we show that the choice of native-like contacts also has an influence on the folding pathways.

Our results raise naturally new questions. What is the smallest possible set of contacts essential for the protein to tie? What is the role of native-like contacts detected by a less strict criterion to identify native-like interactions in proteins around the knotted core of the protein. Can those contacts be used as a hydrophobic long range driving force? How does the folding pathway depend on the length of the chain? Are these considerations general, or are they applicable only to the 2efv protein? These questions could be the starting point for future analysis.

2. Methods and Models

Protein. In this work we use two structures of the 2efv protein: one of the length of 82 residues that is available in the PDB database (denoted as 2efv-82), and a structure with a complete C-terminal end, denoted as 2efv-87. The additional five residues which were not crystallized (GLU, GLY, GLU, ARG, ALA) were reconstructed as an extended helix using the CHARMM package in NAMD 2.6. In the crystal structure the knot begins at Tyr11 and ends at Cys76, hence, $K_N = 10$ and $K_C = 6$ describe the lengths of the N-terminal and C-terminal tails of the knot. For the model with the complete C-terminus $K_C = 11$. The knot covers 66 residues.

Coarse-Grain Model and simulation method. We use the standard Cα coarse-grained structure based model as presented in detail in [24]. Native amino acid contacts are mimicked by the Gaussian potential [25] with the standard parameters proposed by the SMOG server [26]. All simulations were conducted using Gromacs v4.5.4 with the Gaussian potential as implemented in [26]. A leap-frog stochastic dynamics integrator with inverse friction constant equal to 1.0 was used. The time step was equal to 0.0005τ. The number of steps varied between $0.5\times10^9 - 1.5\times10^9$ depending on a mean number of folding events in a trajectory (see Section 3.4).

Contact Maps. We consider two different native-like contact maps. The first one is the Shadow map, described previously in [22]. In this map two beads form a native contact, if they are separated by no more than 6 Å in the native structure, and unless there is another bead between them. The second one, the so called Tsai map follows from the procedure described in [23], based on the criterion of overlaps of enlarged atoms. A native contact is formed between two residues, if there exists at least one pair of heavy atoms, each from each residue, for which the distance between them is not greater than the sum of their van der Waals radii multiplied by the factor of 1.244 (the inflection point in the Lennard-Jones potential) [27]. A comparison of the considered maps for 2efv is presented in Figure 1. There are 212 native contacts in the Shadow map for 2efv-82, while for 2efv-87 there are 11 more (6 within the C-terminal helix, and another 5 with other parts of the protein). The Tsai map of the 2efv-82 consists of 153 native contacts, all of which exist also in the Shadow map. 46 out of the 55 contacts that do not exist in the Tsai map represent long range contacts. The significant fraction of those contacts is formed by amino acids from the N-terminal fragment of 2efv, in particular within the β-sheet motif. The detailed map of native contacts for the Shadow map is presented in Figure 1.

Reaction coordinates. Q is defined as a fraction of native contacts formed. A contact is considered as formed if the distance between the Cα atoms is less than 1.2 times their native distance [28]. Cα atoms are considered to be in non-native contact, if the distance between them is less than 6 Å but they do not form a native contact.

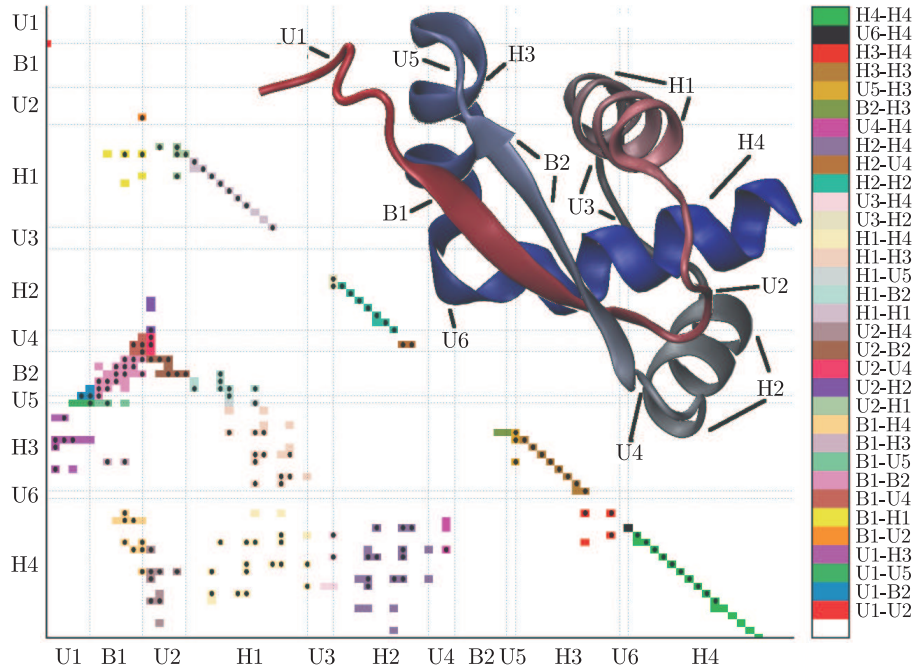

Figure 1. Detailed contact map for 2efv-82; the protein sequence is divided into fragments based on its secondary structure. H1, H2, H3, H4 denote helical fragments along the sequence, B1 and B2 – fragments of the β-sheet, and U1, U2, U3, U4, U5, U6 denote unstructured fragments (*e.g.* turns); native contacts between every pair of these fragments present in the Shadow map are marked by squares of different color, as defined in the legend on the right-hand side; black dots mark the native contacts present in the Tsai map

Thermodynamics and visualization. Thermodynamics data were obtained by constant-temperature simulations and histograms from different temperatures were combined using the Weighted Histogram Analysis Method [29]. Structures were visualized using VMD [30].

Identification of the knot in the protein, KMT algorithm. The presence of the knot during a simulation was determined in each snapshot by using the Koniaris-Muthukumar-Taylor (KMT) algorithm [31] following the procedure described in [32]. To determine the chirality of a knot the HOMFLY-PT polynomial [33, 34] was computed using the Ewing-Millet program [35] implemented in [4].

3. Results

Two main goals of our analysis are as follows. First, we wish to identify a set of native-like contacts that will improve the thermodynamic sampling of knotted proteins in structure based models analysis. We investigate two contact maps, Tsai and Shadow, respectively, with 153 and 212 native-like contacts. Contacts which are not included in the Tsai map represent mostly long range (along the sequence) contacts, $j-i>10$. Our second goal is to elucidate how the knotting

mechanism depends on long range native-like contacts and on an extension of the C-terminal end of the protein by 5 amino acids (based on the known sequence of amino acids). For all the three studied models we observed reversible folding and unfolding transitions, as it was expected at least for the model with the Shadow map [16]. An extensive number of transitions which we obtained for each model allowed us to investigate in detail the free energy landscape of proteins. Differences in the analyzed maps and lengths of protein lead to distinct thermodynamics and kinetics, which we discuss in the following sections.

3.1. Thermodynamic Description of the Folding Energy Landscape

To characterize changes in the stability of the proteins and cooperativity of folding transitions in different models we analyzed the heat capacity curves as a function of the temperature, Figure 2. In general, it is expected that a larger number of native-like contacts should, on the one hand, enhance the thermal stability, and on the other hand, it should speed up the kinetics. The position of heat capacity peaks, which define thermodynamics temperature T^* (in our model) for both lengths of the protein with the Shadow map, are distinctly shifted to higher temperatures in comparison to the model with the Tsai map. This suggests that the latter model needs much lower temperatures to be present dominantly in the unfolded state. This effect arises from the fact, that the number of native contacts, which have to be broken to destabilize the protein, is significantly smaller in the Tsai map. The ratio of native contacts per residue for the Tsai map is equal to 1.87, while for the Shadow map it is 2.58 for 2efv-82 and 2.56 for 2efv-87, respectively. Moreover, a significant part of the native contacts present in the Shadow map and not included in the Tsai map is formed by the N-terminal fragment (U1,B1, and U2; symbols as introduced in Figure 1) with other parts of the polypeptide chain. These contacts, especially with residues from the second β-strand (B2 and U4), are responsible for the stabilization of the loop, which is indispensable for the knot formation. An increase in the heat capacity due to a larger number of native-like contacts for the same protein was observed, for example in [27].

The temperature shift between the models of 2efv-82 and 2efv-87 with the Shadow map is of a topological nature, since the ratios of native contacts per residue are the same, and only 5 out of 11 additional contacts for 2efv-87 are formed not within the C-terminal helix. In case of the protein with the reconstructed C-terminal end, the folding-unfolding process has probably a higher energy barrier to overcome before reaching the native state, due to a necessity of threading (or slipknotting) of the longer tail through the loop. It is worth mentioning that experimental studies suggested that self tying of proteins with a deep knot could take up to even 20 minutes [7]. Details of the folding mechanism are still not available in the experiment, however, independent theoretical simulations have shown that an unexpectedly long time to fold has a dual nature. Firstly, the rate limiting step to fold those proteins is

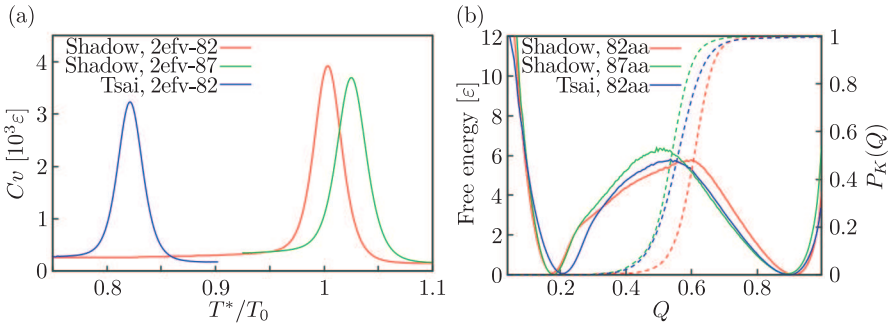

Figure 2. (a) Heat capacity curves as a function of temperature for 2efv-82 in the Tsai map (blue line), the Shadow map (red line), and for 2efv-87 in the Shadow map (green line); (b) one-dimensional free energy profiles as a function of native contacts fraction Q represented by the solid line for 2efv-82 in the Shadow map (red), 2efv-82 in the Tsai map (blue), and 2efv-87 in the Shadow map (green) along with the corresponding knot probability (dashed lines)

due to the tying of a knot (a threading tail across a twisted loop); secondly, some conformations of a protein may act as kinetic traps forming topological barriers [13, 36].

Those results and our analysis of the system suggest that the increase in T^* that we observed for 2efv-87 model facilitates escaping from the traps and increasing the time when the loop is in the open conformation, which is needed to thread a longer tail from the simple kinetics argument. Therefore, the higher T^* for the 2efv-87 model is of a topological nature. However, the observed increase in T^* due to the extended C-terminal end is relatively small in comparison to the decrease in T* due to the decrease in the number of native-like contacts. A comparison of sharp maxima in the heat capacity curves suggests that all the models share the same cooperative folding pattern, however, the folding time is smaller for models with a higher number of native-like contacts, as we will present in what follows.

3.2. Free energy landscape

The one-dimensional free energy profile for the 2efv-82 model with the Tsai map at the folding temperature T_F, presented in Figure 2 (b), has a typical two-state behavior, with no apparent intermediate states. On the other hand, for both models with the Shadow map, the free energy curve bends at Q near 0.24, which suggests the existence of additional substates, that may be hidden at T_F. The point of bending correlates with the position of the intermediate state reported in [16] for the all-atom structure based model. The bending is much less clear in the 2efv-87 model with the Shadow map and it does not appear in the 2efv-82 model with the Tsai map. The heights of energy barriers for all three models are comparable. The free energy barrier for 2efv-87 with the Shadow map is slightly higher than for the other two models. It might suggest that the kinetics of folding and unfolding processes is similar for all three models. Nevertheless, the kinetics

of the folding process of 2efv-82 with the Shadow map is about 3 times faster than for the other models (see Section 3.4). It implies that the fraction of native contacts, Q, in this case is not a good reaction coordinate for this protein [17, 18]. In order to monitor knot formation along the reaction coordinate, we determine a probability $P_K(Q)$ of the knot existence for a given Q. The comparison of $F(Q)$ with $P_K(Q)$ is presented in Figure 2 (b). For all three models the knot formation is strongly correlated with the positions of the top of the barrier in the free energy curves. For Q lower than 0.3 $P_K(Q)$ is zero, implying that no random knots are formed during early stages of the folding process. This behavior agrees with previous theoretical results obtained from different models [12, 14, 15].

Furthermore, numerical simulations based on polymers show that longer chains have a higher probability of tying [37], however, such behavior is not observed for different lengths of the 2efv protein. To understand the thermodynamics of proteins with topological barrier better, we also analyzed a two-dimensional landscape presented in Figure 3, where the second coordinate (other than Q) is the native structure similarity measure, RMSD (Root Mean Square Deviation). For the 2efv-82 model with the Tsai map (Figure 3 a) we observed two broad global minima for denaturated and folded states, while these minima were much more compact for the Shadow map models, especially the native state wells. The larger extent of the minimum of the unfolded state for smaller contact maps arises from two effects. On one hand, at very low $Q = 0.25$, which is on the border of the unfolded basin, we observed that helices were much more structured in the Tsai map than in the Shadow map, see Figure 4. On the other hand, at the same Q in the case of the Shadow map we already observed a formation of native long range contacts. Those contacts are responsible for the loop formation. 2efv-82 with the Shadow map (Figure 3 b) has clearly the smoothest slope connecting the unfolded state minimum with the transition state, which was also captured in the one-dimensional free energy profile. The free energy surface for 2efv-87 provides a much more narrow passage between both global minima (Figure 3 c), significantly slowing down the kinetics of this model in comparison with 2efv-82 with the Shadow map.

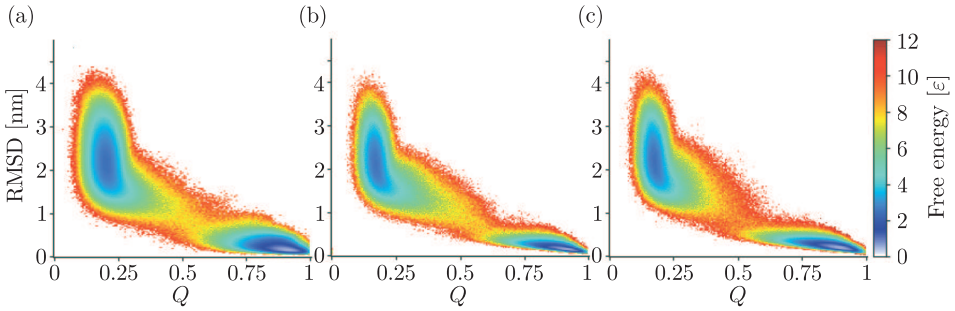

Figure 3. Two-dimensional free energy landscape of the knotted protein measured by the fraction of native contacts formed Q and RMSD for (a) the Tsai map 2efv-82, (b) the Shadow map, 2efv-82, and (c) the Shadow map, 2efv-87

3.3. Folding Routes on the Energy Landscape

In the previous section the influence of the contact map on the free energy landscape of protein folding and on the heat capacity was investigated. Now, we look closer at the dependence of protein folding on the map used. In order to elucidate details of the 2efv protein folding process, protein conformations from trajectories obtained at T_F were clustered according to the native contacts fraction, Q. Normalized occurrences of formed native-like contacts (which we call later probabilities of contact formation) for several values of Q are presented in Figure 4, below the diagonal to see the influence of topological constrains. In addition, we also considered non-native interactions, which were counted when two interacting beads were within the distance of 6.0 Å. Although the choice of the cutoff distance is arbitrary, such choice is consistent with the definition of the cutoff distance for native contacts. Moreover, such a cutoff distance for non-native contacts allowed us to investigate the relative distance of the beads in the neighborhood of the loop. Data for non-native contacts are presented above the diagonal. The detailed description of events along the folding process for secondary structure elements is presented in Table 1.

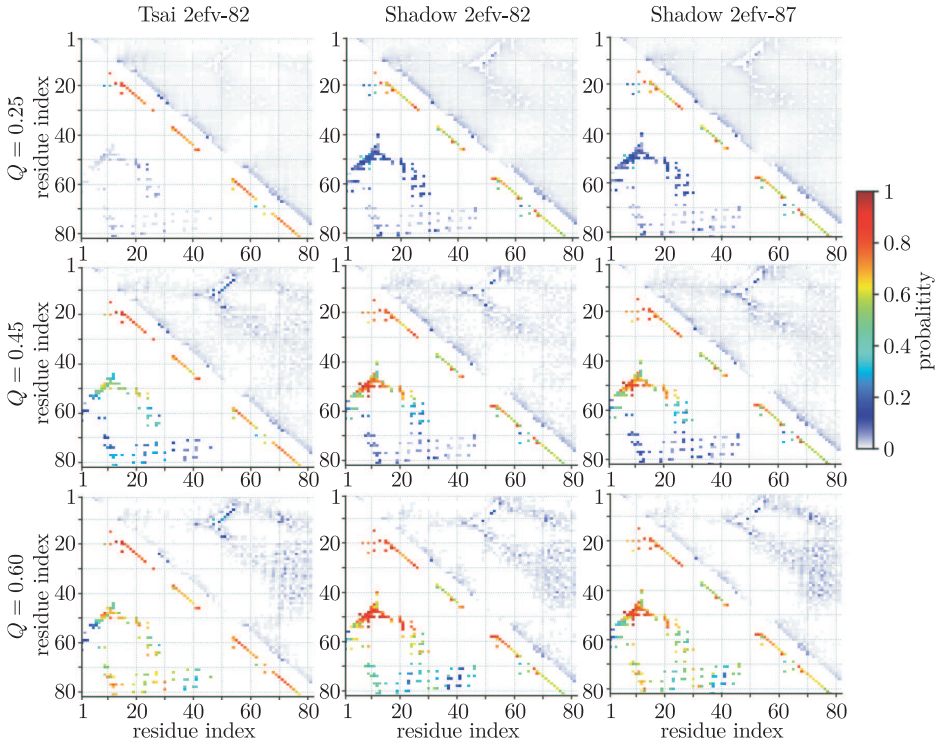

Figure 4. Probability of the presence of interaction in a fraction of conformations with Q=0.25, 0.45 and 0.60 for 2efv-82 with the Tsai map, 2efv-82 with the Shadow map and 2efv-87 with the Shadow map; the native contacts with nonzero probability are marked under the diagonal, and the non-native contacts present in the selected fraction of conformations are denoted above the diagonal

Table 1. The order of native contacts formation in various models; contacts forming the loop are the contacts between B1-B2, B1-U5, B1-U4, U2-B2, U2-B4, U2-H2; for $Q = 0.4$ and for $Q = 0.45$ we observe strengthening the same set of contacts; $*$ means all contacts, *e.g.* H1-$*$ means all contacts formed by amino acids in H1 helix

Q	Tsai 2efv-82	Shadow 2efv-82	Shadow 2efv-87
0.25	formation of H1-4	formation of loop, H1-4, H1-$*$ contacts	formation of loop, H1-4, H1-$*$ contacts
0.30	formation of B1-B2, B1-U4, B2-U2, H1-$*$ and H4-$*$ contacts	loop tightens	loop tightens
0.35	B1-B2, B1-U4, B2-U2, H1-$*$, H4-$*$ contacts tighten	loop tightens	loop tightens
0.45	loop tightens	loop and H1-$*$ contacts tighten	loop and H1-$*$ contacts tighten
0.50	H1-$*$, H4-$*$ contacts tighten	loop and H1-$*$ contacts tighten	loop, H1-$*$, H3-$*$ and H4-$*$ contacts tighten
0.55	H1-$*$, H3-$*$, H4-$*$ contacts tighten	H1-$*$, H3-$*$, H4-$*$ contacts tighten	H1-$*$, H3-$*$, H4-$*$ contacts tighten, loop relaxes
0.60	H1-$*$, H3-$*$, H4-$*$ contacts tighten, loop relaxes	H1-$*$, H3-$*$, H4-$*$ contacts tighten, loop relaxes	H1-$*$, H3-$*$, H4-$*$ contacts tighten, loop relaxes
0.65	stabilization of all contacts	H1-$*$, H3-$*$, H4-$*$ contacts tighten, loop relaxes	stabilization of all contacts
> 0.70	stabilization of all contacts	stabilization of all contacts	stabilization of all contacts

In the first stage of folding we observe a formation of helices, which are present in the free energy minimum corresponding to the unfolded state. For higher values of Q, folding scenarios differ between models. For $Q = 0.25$ (which is right after the bending point of the free energy curve for the 2efv-82 model with the Shadow map), the probability of presence of β-sheet contacts oscillates around 0.2, while in case of the 2efv-82 model with the Tsai map for the same value of Q, contacts within the β-sheet structure are hardly probable. The presence of the β-sheet in the 2efv-82 model with the Shadow map at this stage of folding agrees with previous simulations [16], where the intermediate state observed at $Q = 0.24$ was reported to have the loop already formed. While the probability of β-sheet forming is lower for the 2efv-82 with the Tsai map, the probability of forming helices is much higher. For $Q = 0.25$ the probability of each native contact corresponding to helices in the 2efv-82 model with the Tsai map is at least equal to 0.7, while for the same Q in the model with the Shadow map it oscillates around 0.6, rarely reaching values greater than 0.8. It means that during folding the protein with the Tsai map is much stiffer than those with the Shadow map, in which helices can deform more easily. The difference in the folding pathway at an early stage of folding is probably responsible for a different shape of $F(Q)$ in the range of $0.2 < Q < 0.23$, which shows a higher gradient for a model with a larger number of non-local contacts. The first derivative of $F(Q)$ in the range of $0.19 < Q < 0.22$ is larger for models with the Shadow map. In general, the

entropic cost of forming non-local contacts and a twisted loop, is bigger than for local contacts as in the case of the Tsai map. It can be understood as a result of a restriction on the number of conformations available to the intervening segment. The amount of configurational entropy lost before groups of native-like long range contacts are formed depends on the geometry of the native states, as it was shown in [38]. In our case both proteins have the same geometry and topology, however the Shadow map favors the non-trivial topology of proteins by a larger number of non-local contacts.

At higher values of Q, up to $Q = 0.4$, for models with the Shadow map, we observed a significant increase in the probability of the presence of contacts corresponding to the β-sheet structure. On the other hand, in the model with the Tsai map there are no special regions for which the increase in the probability is significantly larger.

For Q between 0.4 and 0.55 we observed the same sequence of the following events for all three models analyzed. First, the number of loop forming contacts and the probability of their occurrence increases. In the meantime, helices H1 and H3 establish interactions with each other. This brings the H4 helix with the C-terminal end closer to the loop. For Q close to 0.55, the probability of the loop forming contacts decreases, which can be interpreted as a relaxation of the loop needed to let the C-terminal end go through it. The effect of loop relaxation and threading is also visible in the non-native contacts. Comparing maps for $Q = 0.45$ and $Q = 0.6$ it can be noticed, that many marks representing non-native contacts between H1, U3, H2, U4 and B1 vanish.

It should be noted, that the decrease in the probabilities of the loop forming contacts (especially β-sheet contacts) is strongly correlated with the position of the free energy profiles maxima. This means that along with threading the C-terminal end through the loop, relaxing the loop in order to let the C-end go through is one of the reasons for the appearance of the free energy barrier. However, relaxation of the loop cannot be the main reason for the height of the free energy barrier. Although more contacts should be relaxed in the Shadow map than in the Tsai map, the height of the free energy barrier is almost equal for the 2efv-82 models. On the other hand, the free energy barrier is the highest for the 2efv-87 model with the Shadow map. This means that the barrier comes not only from relaxation of the loop, but also from pushing of the C-terminal end helix through it, and the longer the end to be pushed, the higher the energy cost to be paid. For higher Q, helices H3 and H4 establish native contacts with other parts of the protein in all the three models analyzed. The probabilities of later occurrence for each contact increase, leading to the stabilization of the folded structure.

In summary, in each model a general folding mechanism looks similarly. A twisted loop through which one of the termini should thread is required to tie a trefoil knot. In our simulations helices are formed first and then the loop formation begins, and then the C-terminal end is moved towards the loop by H1-H3 contacts. Subsequently, other contacts with H3 and H4 helices are established,

followed by relaxation of the loop and pushing the C-terminal end through it. The last step is the stabilization of the structure. Both maps differ in stiffness of the helices (helices are stiffer in the Tsai map). Moreover, the first important, distinguishable event in the Shadow map is formation of long range contacts – the loop that can be related to the bending point on the free energy curve. However, in the Tsai map there is no such distinguishable event: at the beginning all contacts seem to form in parallel, and in effect there is no evident bending point in the free energy curve corresponding to the Tsai map. The difference between 2efv-82 and 2efv-87 models with the Shadow map is less evident. For the 2efv-87 model each event happens at a lower value of Q. But considering the number of native contacts instead of Q, it turns out that the relaxation of the loop takes place when the number of formed native contacts is around 127. It can be interpreted that there is some set of native contacts, which after being formed, push the C-terminal end through the loop. These contacts probably also determine the geometry of the C-terminal end during threading it through the loop, by slipknot topology in 45% and 22%, for the Tsai and Shadow maps, respectively.

3.4. Kinetics

Having analyzed the influence of the number of native contacts on the free energy landscape and on the folding pathway, we extended our investigation of 2efv toward the kinetics of protein folding. In order to compare kinetics we counted the mean number of free energy barrier crossing (either way, towards folded or unfolded state) during simulation at T_F. The barrier crossing was defined as a sudden increase or decrease in the fraction of the native contacts formed. As the simulations were conducted at T_F, the probability of crossing the barrier towards a folded or unfolded state should be comparable.

In the Tsai map we observe approximately 1 crossing on 375×10^6 steps, while one transition happens approximately per 120×10^6 steps in the Shadow map for 2efv-82. It means that although additional contacts in the Shadow map are not necessary for the protein to fold, they increase the probability of crossing the barrier almost three times (see Figure 5). As most of the additional contacts are

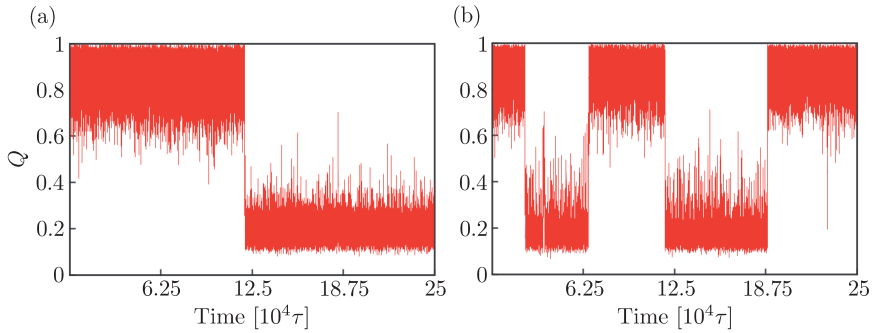

Figure 5. A comparison of the fraction of native contacts for two maps in equally long simulations; (b) 2efv-82 in the Tsai map (one barrier crossing), (b) 2efv-82 in the Shadow map (4 barrier crossings); each simulation consists of 500×10^6 steps

located in the loop region it stresses the importance of the loop formation during protein folding and contribution in the threading mechanism. The contribution of long-range native-like contacts in the threading mechanism was confirmed by explicit solvent simulations without any bias toward the native state [17]. As the free energy barrier is higher for 2efv-87, the elongation of the chain results in the slowdown of the kinetics. As a consequence, the frequency of crossing the barrier for 2efv-87 in the Shadow map is comparable with the frequency for 2efv-82 in the Tsai map. A similar frequency for those models arises just from a coincidence.

4. Discussion

In this study we analyzed the influence of native contact selections (focusing on long range contacts) and the depth of a knot on the thermodynamics and kinetics of the smallest knotted protein, PDB code 2efv. The larger contact map, the Shadow map, with a significant number of additional native-like contacts in the region of the β-sheet, in both 2efv-82 and 2efv-87 models leads to a nonlinear slope curvature in $F(Q)$. We found that this effect was correlated with sudden growth of contacts formed in the β-sheet at Q around 0.24. In the case of the 2efv-82 model with the Tsai map that lacks 46 long range interactions, we showed that the more restricted criterion of the definition of native-like contacts was still sufficient to lead protein across the free energy landscape. However, we observed a rather monotonous increase in the contacts in the β-sheet region, and no bending of the $F(Q)$ curve as it is expected based on all the atom structure based model. We showed that the smaller contact map and longer C-terminal end resulted in a slowdown in the kinetics of folding-unfolding processes. The protein in the 2efv-82 model with the Shadow map folds and unfolds 3–4 times faster than in the Tsai map. Those results indicate that long-range native-like contacts, which have the meaning of non native interactions from the point of view of the Tsai map, facilitate the folding process (instead of introducing noise to the system). An increase in the foldability due to the introduction of specific sets of non-native interactions was observed in other systems. To understand better these results, we additionally compared randomly chosen 10 sets of non-native contacts used to fold the 1j85 protein in [12] and we found that contacts identified by the Shadow map overlapped best with the set of non-native contacts giving the highest probability to fold a protein. This suggests that the Shadow map could be used to identify critical contacts which are treated as non-native interactions by different maps.

Along with what is known for other knotted proteins to date, these results suggest the fundamental physics underlying the folding of proteins with non-trivial topology and the role of native-like contacts. It is possible to find a subset of native contacts that is sufficient to self-tie a protein chain at least for 2efv. This subset of native contacts has to be formed in a specific order to allow self tying. Additionally it is possible to identify, by a different method, a set of non-native or native-like contacts that dramatically enhance the kinetics of a knotting process. A comparison between sets of native-like contacts shows that the knotting process

is surprisingly robust. Contacts identified by us represent an ideal set of amino acids which could be tested experimentally by the phi value analysis [11] to find the driving force to tie a protein.

Our study represents the starting point for analysis of at least the three following problems. The first problem is a designation of the smallest contact map sufficient to self-tie a 2efv protein. It should be possible to generalize our method and to obtain such minimal maps for proteins with deep knots, which are still out of our range of folding simulations. Secondly, it should be possible to identify critical contacts (which will accelerate the folding of proteins with deep knots) by subtracting long range contacts from the less strict criterion that identifies native-like contacts. The third problem is to understand the dependence of kinetics and thermodynamics of a knotted protein on the depth of the knot.

Acknowledgements

We thank Professor Andrzej Kolinski and Aleksandra Grzeszczak for the continuing discussion. The National Science Center [2012/07/E/NZ1/01900 to J.S. and Sz.N.]; the European Molecular Biology Organization [Installation Grant #2757/2014 to J.S. and P.D-T.]; the Foundation for Polish Science [SKILLS/Inter/2014 to J.S.].

References

[1] Taylor W R 2000 *Nature* **406** (6798) 916
[2] Takusagawa F and Kamitori S 1996 *Journal of the American Chemical Society* **118** (37) 8945
[3] Bölinger D, Sułkowska J I, Hsu H-P, Mirny L A, Kardar M, Onuchic J N and Virnau P 2010 *PLoS computational biology* **6** (4), e1000731
[4] Sułkowska J I, Rawdon E J, Millett K C, Onuchic J N and Stasiak A 2012 *Proceedings of the National Academy of Sciences* **109** (26), E1715
[5] Jamroz M, Niemyska W, Rawdon E J, Stasiak A, Millett K C, Sułkowski P and Sułkowska J I 2014 *Nucleic Acids Research*, gku1059
[6] Mallam A L, Rogers J M and Jackson S E 2010 *Proceedings of the National Academy of Sciences* **107** (18) 8189
[7] Mallam A L and Jackson S E 2012 *Nature chemical biology* **8** (2) 147
[8] Andrews B T, Capraro D T, Sułkowska J I, Onuchic J N and Jennings P A 2012 *The journal of physical chemistry letters* **4** (1) 180
[9] Chavez L L, Onuchic J N and Clementi C 2004 *Journal of the American Chemical Society* **126** (27) 8426
[10] Ferguson N, Capaldi A P, James R, Kleanthous C and Radford S E 1999 *Journal of molecular biology* **286** (5) 1597
[11] Wensley B G, Batey S, Bone F A, Chan Z M, Tumelty N R, Steward A, Kwa L G, Borgia A and Clarke J 2010 *Nature* **463** (7281) 685
[12] Wallin S, Zeldovich K B and Shakhnovich E I 2007 *Journal of molecular biology* **368** (3) 884
[13] Sułkowska J I, Sułkowski P and Onuchic J N 2009 *Proceedings of the National Academy of Sciences* **106** (9) 3119
[14] Sułkowska J I, Noel J K and Onuchic J N 2012 *Proceedings of the National Academy of Sciences* **109** (44) 17783

[15] Li W, Terakawa T, Wang W and Takada S 2012 *Proceedings of the National Academy of Sciences* **109** (44) 17789
[16] Noel J K, Sułkowska J I and Onuchic J N 2010 *Proceedings of the National Academy of Sciences* **107** (35) 15403
[17] Noel J K, Onuchic J N and Sulkowska J I 2013 *The Journal of Physical Chemistry Letters* **4** (21) 3570
[18] Beccara S a, Škrbić T, Covino R, Micheletti C and Faccioli P 2013 *PLoS computational biology* **9** (3), e1003002
[19] Škrbić T, Micheletti C and Faccioli P 2012 *PLoS computational biology* **8** (6), e1002504
[20] Covino R, Škrbić T, Faccioli P, Micheletti C, *et al.* 2013 *Biomolecules* **4** (1) 1
[21] Godzik A, Skolnick J and Kolinski A 1993 *Protein Engineering* **6** (8) 801
[22] Noel J K, Whitford P C and Onuchic J N 2012 *The Journal of Physical Chemistry B* **116** (29) 8692
[23] Tsai J, Taylor R, Chothia C and Gerstein M 1999 *Journal of molecular biology* **290** (1) 253
[24] Clementi C, Nymeyer H and Onuchic J N 2000 *Journal of molecular biology* **298** (5) 937
[25] Lammert H, Schug A and Onuchic J N 2009 *Proteins: Structure, Function, and Bioinformatics* **77** (4) 881
[26] Noel J K, Whitford P C, Sanbonmatsu K Y and Onuchic J N 2010 *Nucleic acids research* **38** (2), W657
[27] Cieplak M and Hoang T X 2002 *International Journal of Modern Physics C* **13** (09) 1231
[28] Whitford P C, Noel J K, Gosavi S, Schug A, Sanbonmatsu K Y and Onuchic J N 2009 *Proteins: Structure, Function and Bioinformatics* **75** (2) 430
[29] Kumar S, Rosenberg J M, Bouzida D, Swendsen R H and Kollman P A 1992 *Journal of computational chemistry* **13** (8) 1011
[30] Humphrey W, Dalke A and Schulten K 1996 *Journal of molecular graphics* **14** (1) 33
[31] Koniaris K and Muthukumar M 1991 *The Journal of chemical physics* **95** (4) 2873
[32] Sułkowska J I, Sułkowski P, Szymczak P and Cieplak M 2008 *Physical review letters* **100** (5), 058106
[33] Freyd P, Yetter D, Hoste J, Lickorish W R, Millett K and Ocneanu A 1985 *Bulletin of the American Mathematical Society* **12** (2) 239
[34] Przytycki J H and Traczyk P 1988 *Kobe Journal of Mathematics* **4** (2) 115
[35] Ewing B and Millett K C 1997 *Progress in knot theory and related topics* **56** 51
[36] Tuszynska I and Bujnicki J M 2010 *Journal of Biomolecular Structure and Dynamics* **27** (4) 511
[37] Rawdon E, Dobay A, Kern J C, Millett K C, Piatek M, Plunkett P and Stasiak A 2008 *Macromolecules* **41** (12) 4444
[38] Baker D 2000 *Nature* **405** (6782) 39