Supporting Information

Sułkowska *et al.* 10.1073/pnas.0811147106

SI Text

In this supporting information (SI) Text we present more details concerning the geometry and properties of knotted proteins (PDB ID codes 1j85, 1o6d, and 1vho) and proteins with slipknots (PDB ID codes 2j6b and 1p6x). We present the Reidemeister moves necessary for a description of knot creation and describe attached video clips (Movies S1 and S2). Furthermore, we provide quantitative details of the fraction of native contacts, as well as the "optimal" energetic map that we found. Finally we describe in detail four folding trajectories of 1p6x.

Proteins with Knots and Slipknots. ***Knotted proteins.*** Monomeric subunits of *Haemophilus influenzae* YibK (PDB ID code 1j85) and *Escherichia coli* YbeA (PDB ID codes 1o6d and 1vho) are shown in Fig. 1, together with additional tails attached to their ends. Another member of the methyltransferase YibK, the protein sr145 from *Bacillus subtilis* (1PDB ID code 1to0), is not suitable for our analysis because its chain is broken in many places.

Methyltransferase (Mtase) from *H. influenzae* YibK is described in detail in refs. 1 and 2. Homodimeric α/β-knot methyltransferase (Mtase) from *E. coli* YbeA (1o6d), consists of 5 α-helices that surround the β-sheet built of 6 β-strands. The secondary structural elements are referred to as follows: 1, β-strand (amino acids 2–8); 2, α-helix (12–27); 3, β-strand (31–37); 4, α-helix (44–58); 5, β-strand (65–69); 6, β-strand (74–75); 7, α-helix (77–92); 8, β-strand (95–99); 9, α-helix (106–113); 10, β-strand (115–118); 11, α-helix (125–145). The structure is knotted between β-strands 5 and 10. The second protein from the YbeA family (1vho) has almost the same structure as 1o6d, but it is longer by 10 aa. It consists of 5 α-helices that surround the β-sheet built of 6 β-strands.

Proteins with slipknots. We analyze two proteins with slipknots: thymidine kinase (3) (PDB ID code 1p6x) and highly conserved protein from crenarchaeal viruses AFV3–109 (4) (PBD ID code 2j6b).

Diagrammatic representation of the protein 2j6b (5) is shown in Fig. 2 *Left*, together with a schematic structure of a slipknot, as well as a knot obtained after cutting 9 aa from the terminal C. 2j6b consists of $n = 109$ aa and belongs to the α/β-class. It is built of 5 β-strands that form the sheet surrounded by a loop from one side and the helix 3–5 on the other side. The secondary structure notation that we use is as follow: 1, β-strand (amino acid 2–5); 2, β-strand (19–25); 3, α-helix (26–37); 4, β-strand (39–41); 5, α-helix (45–57); 6, β-strand (74–80); 7, α-helix (92–100); 8, β-strand (101–109). The function of AFV3–109 is not known, but it was suggested that it could interact with nucleic acids (4).

A thymidine kinase (PDB ID code 1p6x) belongs to the α/β-class protein and is shown in Fig. 2 *Right*. It can be divided into 11 parts denoted by numbers from 1 to 11 in the main text, which are specified as follows: 1, N terminus with the β-strand (amino acid 20–34); 2, α-helix with turn (35–45); 3, β-hairpin and turn (55–60); 4, big long loop created of 3 helices form one side and closed from the other side by loop 4* (120–132) all together (61–132); 5, β-strand (133–139); 6, hairpin built of 2 helices (141–176); 7, β-strand (177–184); 8, bundle of 5 helices (185–299); 9, β-strand (300–306); 10, helix and β-hairpin (306–323); 11, the chain crossing the loop 4* (324–333); 12, helix on the second side of the loop 4 (334–352). The strands 1, 3, and 5 create a characteristic β-sheet. By using this notation for the elements of the secondary structure we identified the following contacts: 1–2, 1–4, 1–4*, 1–5, 1–7, 1–10, 1–11b, 2–3, 2–10, 3–5, 3–10, 3–12, 4, 4–5, 4–6, 4–8, 4–10, 4–11, 4–12, 5–12, 6, 6–8, 7–9, 7–10, 8, 9–10. This list is used to analyze the folding routes and the backtracking between various groups of contacts. Our analysis is mostly concentrated on the loop 4* through which β-strand 1, β-strand 10, and helix 11 have to pass.

Reidemeister Moves. Three Reidemeister moves I, II, and III are shown in Fig. 3. They describe, after a projection on a plane, basic geometric transformations that do not change a topology of a knot. In each move the relative location of the ends of all of the strings involved is unchanged. Move I is a transformation of a straight piece of a string into a loop. Move II corresponds to transforming 2 parallel strings into a configuration with 2 crossings. Move III involves 3 strings, and corresponds to a shift of one of them over the crossing made by the other two. These Reidemeister moves are very useful in a simplified description of proteins with nontrivial topology.

Video Clips of Folding Trajectories for Knots and Slipknots. We enclose video presentations of the mechanism of knot formation, generated by using our implementation of the Go-like model. The first animation presents protein 1o6d. The process of the knot formation in this animation corresponds to Fig. 2 in the main text. The knotted regions of protein is shown by green color. Furthermore, we also enclose a video presentation of the mechanism of slipknot formation in our model for 1p6x.

Intermediate Slipknot and the Fraction of Native Contacts for 1o6d. In Fig. 4 the fraction of native contacts Q is shown at the moment when the slipknot is being created, i.e., when its hook-part starts to be threaded through the loop 66–96. In the main trajectory the structure elements located closer to the N terminus of the protein fold in last stages of the folding process. These trajectories are represented by red dots and in this case $Q \simeq 0.8$. In the second type of trajectories, represented by blue dots, the N-terminal folds in initial stages of the folding, and $Q \simeq 0.6$. One additional point in the figure represents a very rare knot formation, corresponding to threading the C terminus through the loop 66–96 without a slipknot intermediate. The data shown in this figure correspond to the optimal model considered in the main text and denoted in bold in the Table S1. However, there is no qualitative difference in the values of Q between this model and the results obtained in the basic model (with the uniform energetic map).

Apart from the protein 1o6d we also checked the folding ability of another member of the YbeA family, i.e., protein with PDB ID code 1vho. Similarly to 1o6d, this protein can fold to the native state in the model only based on the native contacts. The folding route includes the slipknot transition state, similarly as for 1o6d and 1j85. In this case, we also observed one trajectory where the C terminus crosses the loop not in a hairpin-like configuration. The probability of creation of a knot for this protein in a simple model is $\approx$0.1%.

New Energetic Map for Folding Knotted Protein 1o6d. From the analysis of the behavior (in particular, involving the backtracking) of various elements of the knot structure in 1o6d during folding, shown in Fig. 3 in the main text, we constructed a series of models with new energetic maps. These maps are based only on native contacts and differ in strengths of interactions inside the knotted region of a protein. The energetic maps that we

considered are shown in Table S1, where these groups of contact are listed whose strength was changed. We found a few energetic maps for which the folding ability is highest—they are shown in rows 15, 17, and 20 in Table S1. For these models we increased twice the sampling set, and under such refined statistics we found that folding ability was the best for the model from row 15, which we call "optimal." It is also possible to construct energetic maps that prevent knot formation completely, for example, by increasing the strength of interaction between contacts for which backtracking is observed (such as β-strands 5, 6, and 9).

Folding Trajectories for Slipknots in 1p6x. We present now, in more detail, four main folding routes shown in Fig. S5, which are briefly discussed in the main text.

The first route is represented by 5 trajectories that we found. Formation of the slipknot is one of the last steps during folding, and it is possible because of backtracking between β-strands 1–5, 3–5, and contacts between 3–11 and 4–5. The characteristic property of these trajectories is threading of the N and C termini through the loop 4^*, which takes place after or almost in the same time when the entire loop 4 establishes contacts with segment 8. This event, as well as a correlation between threading almost in the same time elements 1 and 10, makes this trajectory distinct from all others that we observed. From folding scenario for this route we deduce that almost all native contacts are created before threading appears, apart from the contacts between the N terminus and the β-hairpin of the C terminus with loop 4*. At this state the contacts between 1–4, 1–5, 4–5, and 4–6 have to be broken to open the loop 4 and allow the N and C termini to thread through it. Now a hairpin is created close to the N terminus, which subsequently moves across loop 4. By temporary breaking and creating the contacts between the N terminus and hairpin 10, the entire C terminus is threaded across loop 4 also in the hairpin conformation.

The second folding route is represented by 2 trajectories. In this case the chain strongly folds from N terminus by a few different nuclei. This causes creation of a β- sheet from β-strands 1, 3, 5, entire loop 4, and its contacts with 6 and 8. In such a case (no other native structure of protein is yet established) a small backtracking between elements 1–4 and 1–5 allows easily for crossing the N terminus in the bend shape through loop 4*, which creates a knot. After that all other elements of the protein fold. The final event is threading the C terminus across loop 4 as in the first folding route described above.

The third folding route is most distinct from the others. A part of loop 4, loop 6 (made of 2 helices) and bundle 8 fold in the same time when contacts between 9 and 10 are established, but all nuclei still exist separately. In the next step big nuclei start to rearrange, which results in a creation of a slipknot with segments 9, 10, and 11 located inside loop 4. In contrast to all other trajectories, after this step the N terminus starts to fold and makes half of the β-sheet clamping the C terminus in loop 4. Now, only the end of the N terminus has to be threaded by loop 4* (as in route 1), whereas the C terminus has to go out of the clamp (as in route 2). These moves are possible because of backtracking between segments 1–4, 1–5, 4–5, 1–10, and 3–10.

The fourth trajectory is schematically described in the main text, but now we include more details. We recall that in this route almost the entire native structure, apart from contacts between the loop 4 and elements 6 and 8, is created before creation of a slipknot. This conformation corresponds almost entirely to the β-sheet created (involving the elements 1, 3, and 5) with strands 1 and 10 located close to their native positions (as in the knot or slipknot configuration). The final stages of knotting are accompanied by a backtracking, as shown in Fig. 6 in the main text. We discuss it starting from the moment when the loop 4 is still above the sheet 1-3-5. From this moment the contacts between β-strands 3–5, helix 2, and strand 3, and a few others, perform amazing backtracking that leads to the rotation of loop 4 by almost 360°. During this rotation the values of rmsd and RG steadily grow. The contacts inside loop 4 break temporarily to provide enough space to accommodate both N and C termini, which are subsequently threaded through the loop.

Some deviation of the second folding route is represented by two other trajectories, for which the critical role is also played by closing loop 4. In this case, the β-sheet 1-3-5 is created after establishing the entire remaining native structure and before closing loop 4 (as also described above). This allows for reopening the contacts in β-sheet and rotation of loop 4 to make the knot, despite a somewhat awkward position of the C terminus. In consequence, closing of loop 4 clamps the C terminus between this loop and the β-sheet for a long time.

1. Mallam AL, Jackson SE (2007) Comparison of the folding of two knotted proteins: YbeA and Yibk. *J Mol Biol* 366:650–665.
2. Mallam AL, Jackson SE (2007) The dimerization of an α/beta-knotted protein is essential for structure and function. *Structure* 15:111–122.
3. Gardberg A, Shuvalova L, Monnerjahn C, Konrad M, Lavie A (2003) Structural basis for the dual thymidine and thymidylate kinase activity of herpes thymidine kinases. *Structure* 11:1256–1277.
4. Keller J, *et al.* (2007) Crystal structure of AFV3–109 a highly conserved protein from crenarchaeal viruses. *Virol J* 4:12.
5. Yeates T, Todd O, Norcross S, King NP (2007) Knotted and topologically complex proteins as models for studying folding and stability. *Curr Opin Chem Biol* 11:595–603.

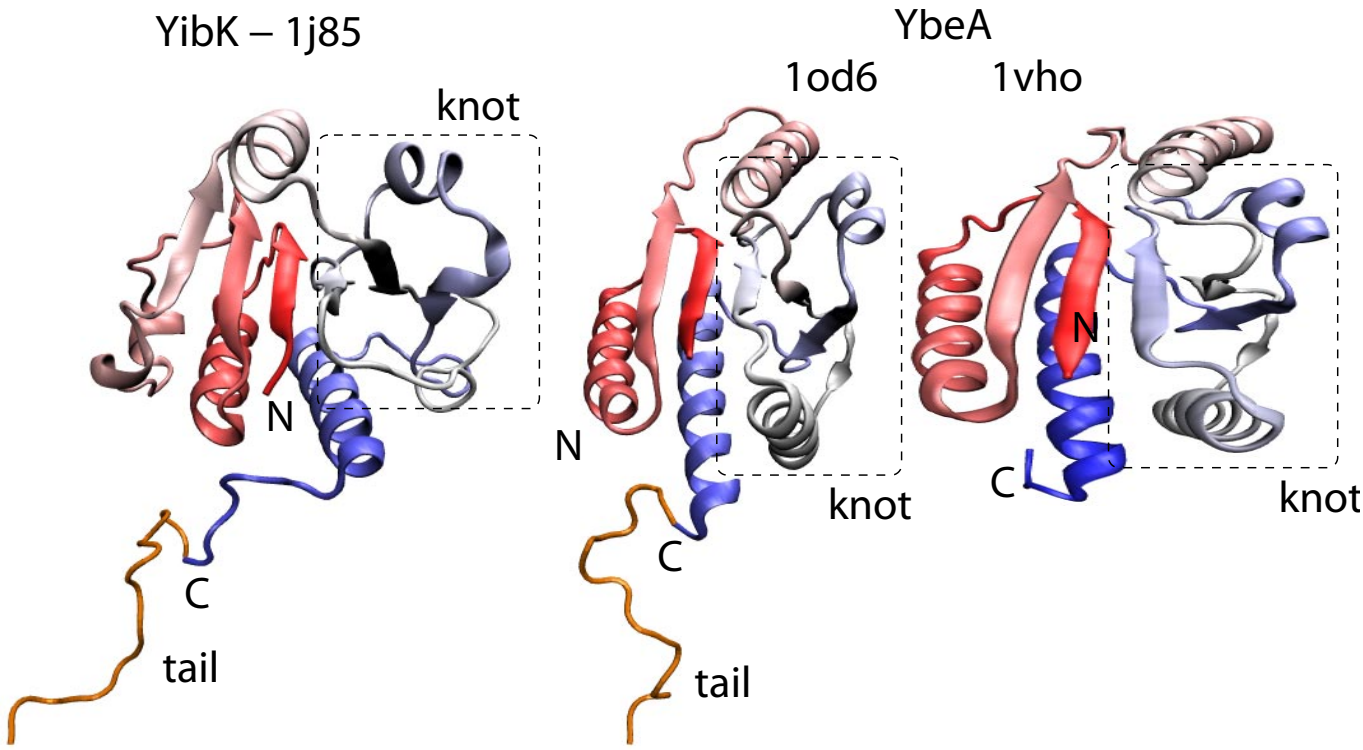


Fig. S1. Structure of knotted proteins from families YibK and YbeA. (*Left*) Protein 1j85 with attached designed flexible tail. (*Center*) Protein 1o6d with attached designed flexible tail. (*Right*) 1vho. For each protein a region where the knot is localized is denoted by a dashed rectangle.

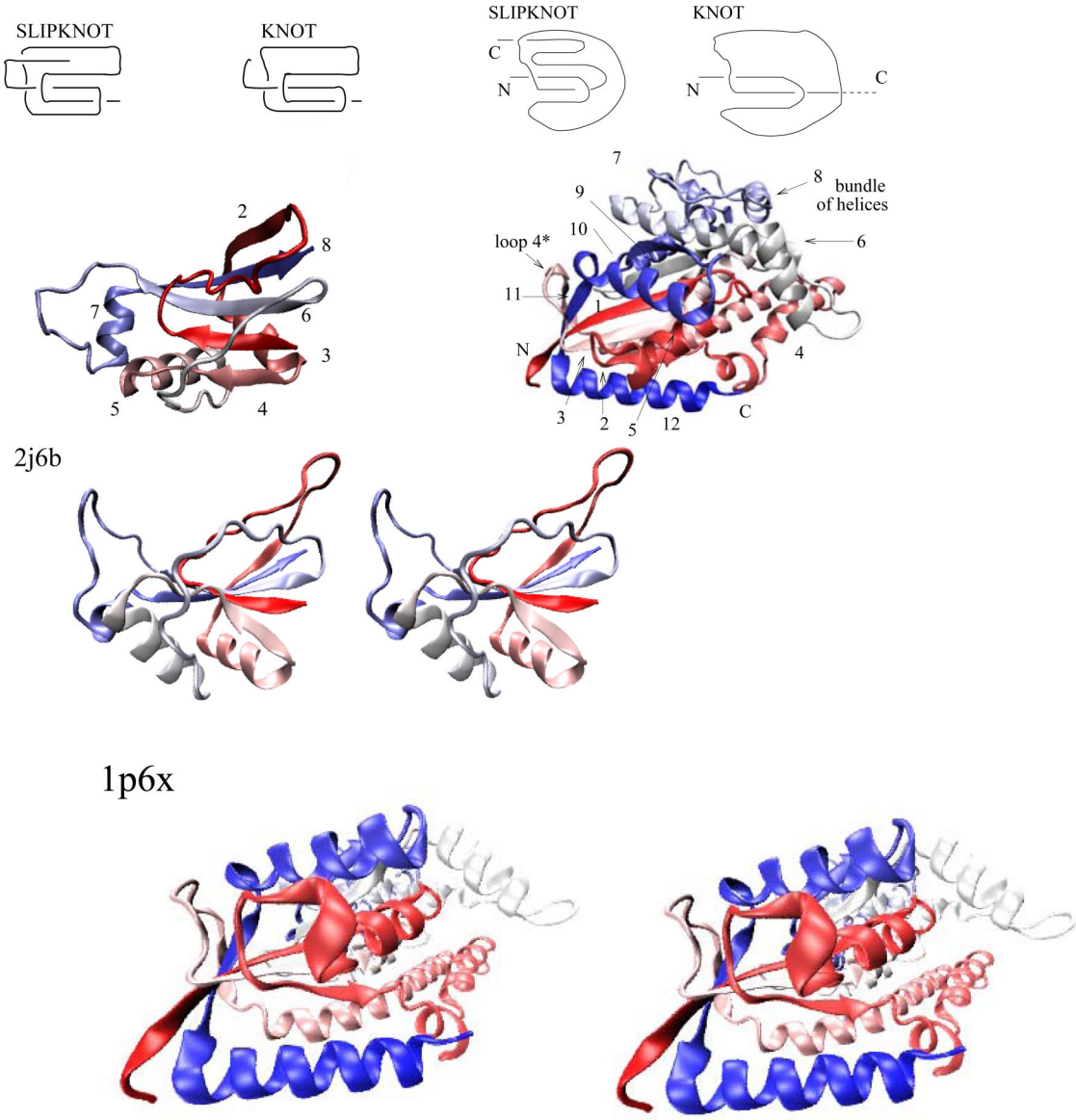


Fig. S2. Structure of a slipknotted protein and a smooth representation of its slipknot, as well as a modification to the knot. (*Upper*) The corresponding schematic structures of a slipknot in these proteins, as well as the knotted conformation 3_1 obtained after cutting, respectively, 20 aa from the N terminus (for 2j6b), or a few amino acids from the C terminus (for 1p6x). (*Lower*) Stereoviews of proteins 2j6b (*Left*) and 1p6x (*Right*).

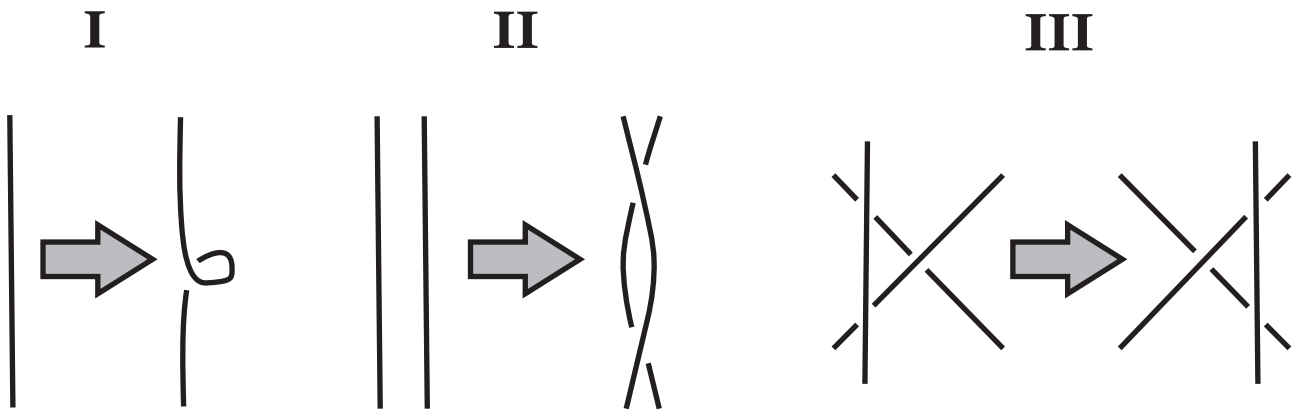


Fig. S3. Reidemeister moves I, II, and III.

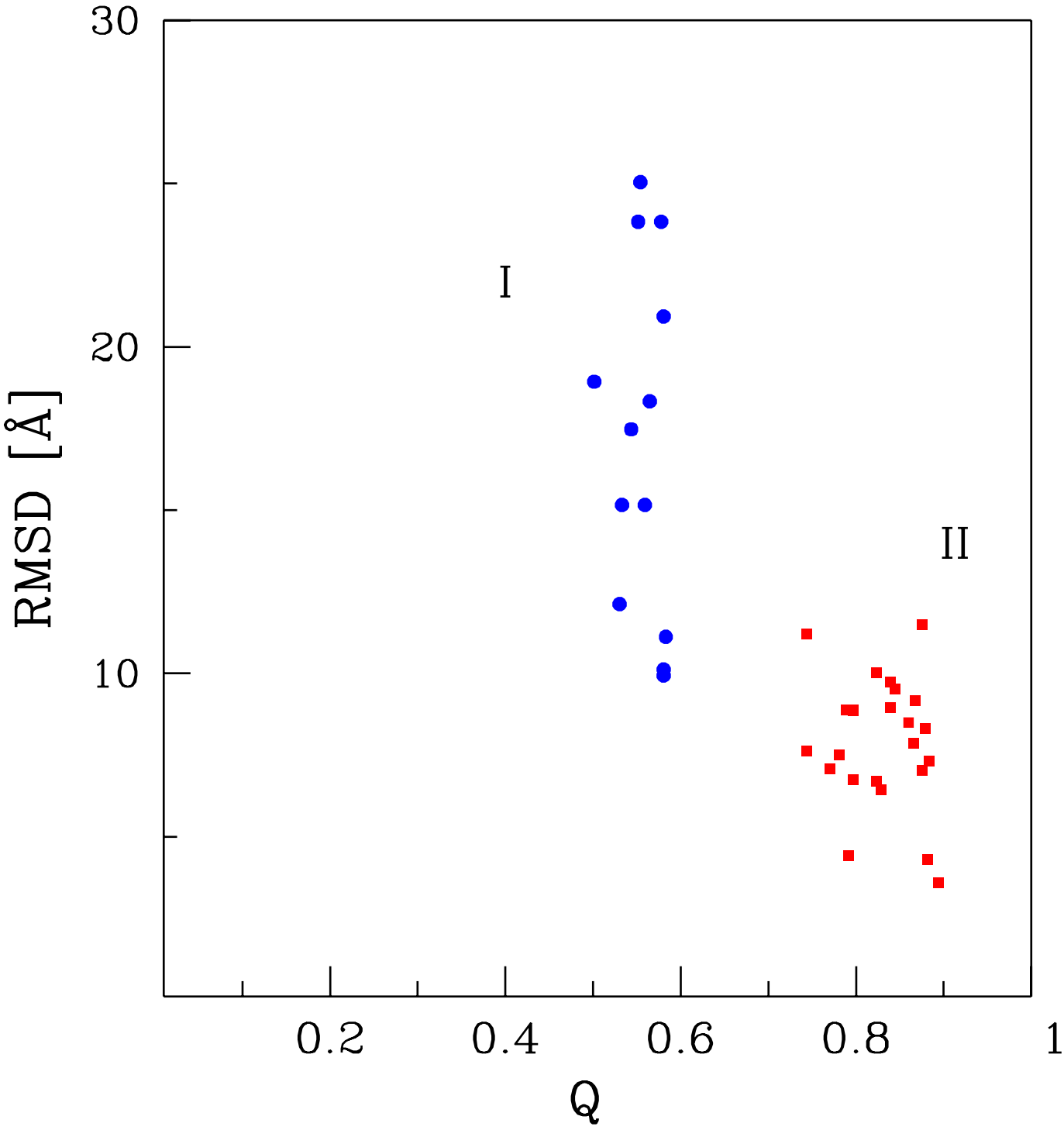


Fig. S4. Distribution of native contacts in protein 1o6d at the moment when slipknot configuration appears for the first time during folding. This moment corresponds to the fourth step in Fig. 2 in the main text. The results are similar both for the model with the optimal native-like energy scale and for the basic model with uniform energy. The symbol I corresponds to folding from the C terminus, and II corresponds to folding from the N terminus.

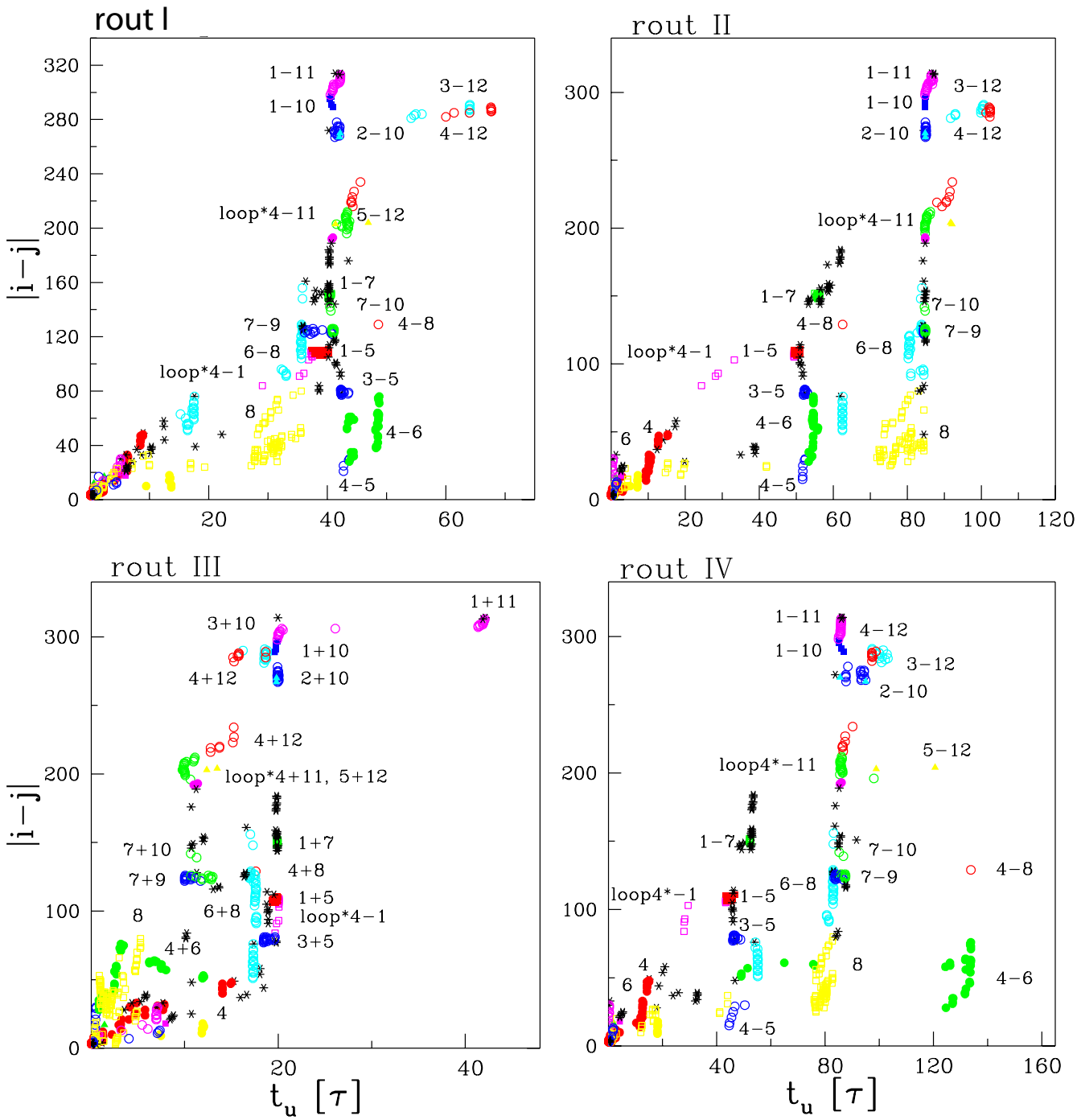
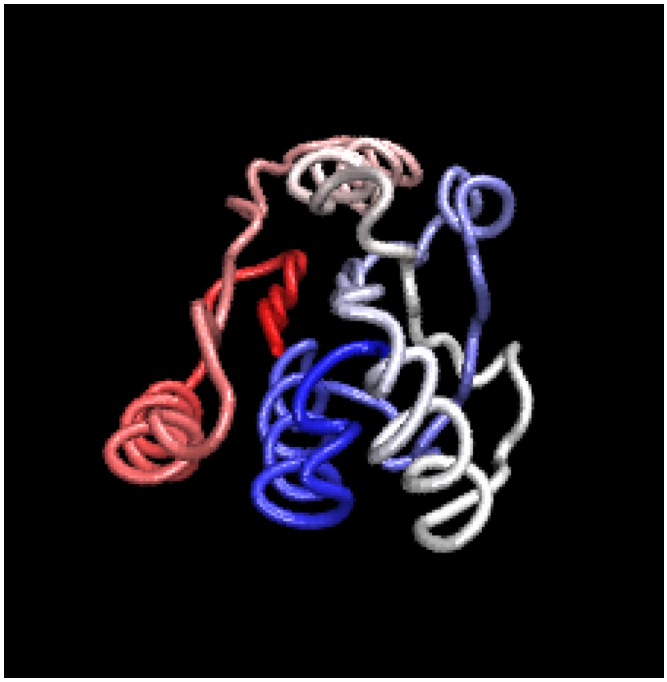
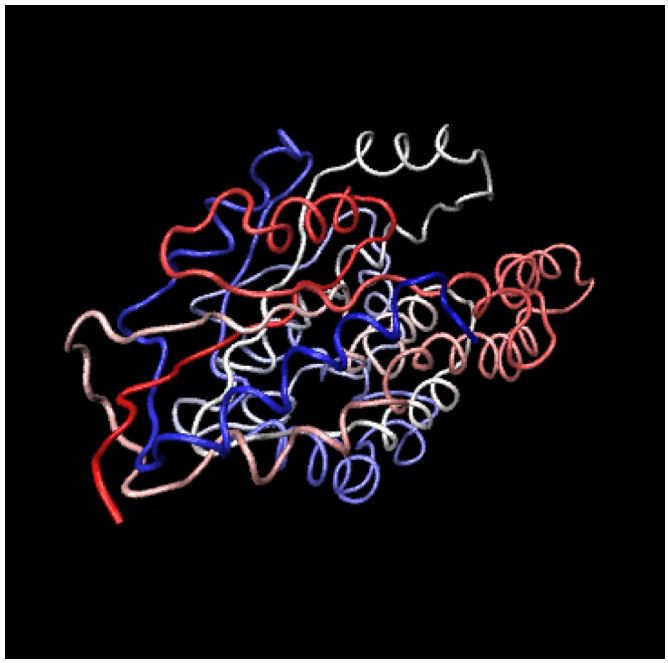


Fig. S5. The folding routes for the 1p6x protein as described by the first time of forming a contact corresponding to the sequence of length $j - i$, respectively, for the first, second, third, and fourth route.



Movie S1. Folding mechanism for protein 1o6d.

Movie S1 (MPG)



Movie S2. Foldingn mechanism for slipknot protein 1p6x.

Movie S2 (GIF)

Table S1. Energetic maps for folding knotted protein 1o6d

Order	Region*ε	Region*ε	Region*ε	Region*ε	Success rate
1	5–8*0.5				3
2	5–8*2				9
3	5–8*4				10
4	5–8*6				9
5	5–8*10				1
6	5–8*2	6–10*2			15
7	5–8*4	6–10*2			9
8	5–8*10	6–10*2			12
9	5–8*2	6–10*4			14
10	5–8*2	6–10*6			8
11	5–8*2.5	6–10*2.5			6
12	5–8*2		(7–10,8–10,9–10)*0.5		2
13	5–8*2		(7–10,8–10,9–10)*1.5		2
14	5–8*2	6–10*2	(7–10,8–10,9–10)*0.5		25
15	5–8*2	6–10*2	(7–10,8–10,9–10)*0.5	without 74–115	34
17	5–8*2.5	6–10*1.5	(7–10,8–10,9–10)*0.5	without 74–115	25
18	5–8*2	6–10*3	(7–10,8–10,9–10)*0.5	without 74–115	27
19	5–8*1	6–10*3	(7–10,8–10,9–10)*0.5	without 74–115	15
20	5–8*2	6–10*2	(7–10,8–10,9–10)*0.5	turns*2	34
21	5–8*2			5–9*2	3
22	5–8*2			6–9*2	2

The *region* corresponds to 2 elements of the secondary structure (or to a particular number of amino acid) for which the strength of interaction was changed by a factor ε. For example, 5–8*2 means that the interaction strength between structure elements 5 and 8 was increased twice. The *success rate* shows the number of correctly knotted structures (with ≈95% of all native contacts) in the set of 2,500 runs.