Published in final edited form as:

J Mol Biol. 2007 May 4; 368(3): 884–893. doi:10.1016/j.jmb.2007.02.035.

The folding mechanics of a knotted protein

Stefan Wallin, Konstantin B Zeldovich, and Eugene I Shakhnovich

Department of Chemistry & Chemical Biology, Harvard University, 12 Oxford Street, Cambridge MA 02138, USA

Abstract

An increasing number of proteins are being discovered with a remarkable and somewhat surprising feature, a knot in their native structures. How the polypeptide chain is able to “knot” itself during the folding process to form these highly intricate protein topologies is not known. Here we perform a computational study on the 160-amino acid homodimeric protein YibK which, like other proteins in the SpoU family of MTases, contains a deep trefoil knot in its C-terminal region. In this study, we use a coarse-grained C_{α} -chain representation and Langevin dynamics to study folding kinetics. We find that *specific*, attractive nonnative interactions are critical for knot formation. In the absence of these interactions, i.e. in an energetics driven entirely by native interactions, knot formation is exceedingly unlikely. Further, we find, in concert with recent experimental data on YibK, two parallel folding pathways which we attribute to an early and a late formation of the trefoil knot, respectively. For both pathways, knot formation occurs before dimerization. A bioinformatics analysis of the SpoU family of proteins reveals further that the critical nonnative interactions may originate from evolutionary conserved hydrophobic segments around the knotted region.

Introduction

Are there knots in proteins? This question was first posed in 1994 by Mansfield.¹ At the time, not more than about 400 protein structures were available for analysis and only two potential cases of very shallow knots could be identified. For such borderline cases, classifying proteins as knotted is nontrivial as knots are only rigorously defined for closed loops (how should the ends be connected?).² Since then, the issue has been revisited several times and additional examples of more deeply knotted proteins have surfaced,^{3;4; 5; 6} and we now undoubtedly have to conclude that knots do indeed occur in proteins. In the most recent review,⁶ 273 proteins with some type of knotted conformation were found in the PDB. Still, knots in proteins are relatively rare and most examples of knots are simple trefoils, i.e. they have 3 projected crossings. The known set of knotted proteins is dominated by two classes:⁶ bacterial and viral rRNA methyltransferases, which make up the α/β -knot superfamily,⁷ and various isozymes of carbonic anhydrase. In addition, a few other insular examples of knots in proteins have been found, one of which is ubiquitin hydrolase (UCH-L3), the most complicated knot in a human protein found to date, with five projected crossings.⁶

Here we address the question: What are the mechanics of knot formation in proteins during the folding process? To our knowledge, this issue has not been previously addressed. One likely reason for this is the difficulty in designing suitable experiments that can directly probe knot

Corresponding author: Eugene I Shakhnovich.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

formation as it is a subtle conformational transition. We thus turn to computer simulations in which, at least in principle, every conformation from the unfolded to the native state is available for analysis. To monitor knot formation in our simulations we make use of KMT-reduction,⁸ a test procedure where atoms are sequentially removed if it does not alter the “topology” of the chain. If the test procedure reduces the chain to its first and last atoms the chain is considered the unknot (atom j is removed if the triangle defined by atoms $j - 1$, j , and $j + 1$ is not intersected by any other bond vector in the chain). Despite the lack of mathematical rigor, we find the KMT reduction method reliable and very useful as an operational definition of knot occurrences in simulations.

We focus here on the homodimeric protein YibK (PDB id: 1J85) which belongs to the SpoU family of MTases and is well-suited for a computational study. It contains 160 amino acids and has a deep trefoil knot at approximately 40 residues from its C-terminal,⁹ as depicted in Figure 1. In recent works, Mallam and Jackson extensively characterized^{10; 11} the kinetics and thermodynamics of this protein, although details of knot formation could not be addressed. It was shown that YibK folds reversibly *in vitro* indicating that chaperon activity is not a prerequisite for folding, and it was revealed that YibK has a multi-phasic folding kinetics, with the slowest phase corresponding to dimerization. Moreover, the experimental data suggested that monomeric forms of the protein dominate at low pH and low protein concentrations. For this, and for computational reasons, we focus here mainly on the monomer folding of YibK. We start by straightforwardly applying a well-studied and commonly used Gō-type model¹² to the folding of YibK. We find that for YibK, this standard Gō-type model, in which only native interactions are treated as attractive, fails at producing even a plausible folding scenario as knot formation does not occur. Only by modifying the energetics of the system through the addition of certain *specific* nonnative, attractive interactions, can we achieve successful folding, highlighting the importance of these types of forces during the folding of this protein. In addition to our computer simulations, we perform a bioinformatics analysis of the SpoU family of MTases in an attempt to find evidence for conserved sequence characteristics that relates to knot formation. As knots in proteins are usually conserved within protein families,⁶ we expect most members of the SpoU family to contain similar types of trefoil knots as in the YibK protein. Indeed, several recently determined structures within this family have been found to have deep trefoil knots.^{13; 14; 15; 16} Through an analysis where we compare conserved sequence properties of the SpoU family with those obtained from structurally similar but unknotted protein structures, we find evidence for some unusual sequence properties in the SpoU family close to the trefoil knot, which may be linked to critical nonnative interactions during knot formation.

Results and Discussion

Native interactions and the folding of YibK

Our initial goal is to investigate the role of native interactions for the folding dynamics of YibK. We therefore apply a well-studied coarse-grained Gō-type C_α model^{12; 17; 18; 19; 20; 21} to the YibK monomer, in which only interactions present in the native structure are attractive, and explore the kinetic behavior using Langevin dynamics (see Models and Methods). This type of “native-centric” model has been used widely in the past several years to study folding kinetics and thermodynamics because of computational limitations on more realistic sequence-based all-atom protein force fields.^{22; 23; 24} As nonnative interactions are ignored, the native-centric approach to protein modeling usually guarantees folding to a proteins’ native state, and some justification for its use exists.²⁵ In particular, it has been found useful when tested against certain experimental data.^{18; 26; 27}

We find, however, that an entirely native-centric approach can not explain the folding of YibK. Nonetheless, it is illuminating to explore the kinetic behavior obtained for the coarse-grained

C_α model. Figure 2(a) shows the time evolution of the fraction of native contacts, Q , in a typical simulation for the YibK monomer started from a random configuration, at $T = 0.975$, which is slightly below the melting temperature, T_m . A reversible transition to a state characterized by $Q \approx 0.7$ – 0.8 occurs. This state is compact and quite native-like in terms of the cRMSD similarity measure (data not shown), but the chain remains unknotted and is thus unable to reach its native state. It should be pointed out that, by construction, the energy-minimum state of this model is very close to the native structure, which, of course, in the case of YibK includes a trefoil knot. It could therefore be suggested from Figure 2(a) that folding does not occur because the native state is kinetically inaccessible from the unfolded state. This is not the case, however. To show this, we perform several additional simulations started from the native structure at a slightly elevated temperature $T = 1.04 > T_m$. In all cases, unfolding is readily observed. A typical unfolding trajectory can be seen in Figure 2(b). Hence, we conclude that the unfolded and native states are kinetically connected in our model, but in a dynamic process dictated by native interactions alone the probability of reaching the native state is exceedingly low. Clearly, important interactions to the folding of YibK are absent from this standard Gō-model.

Attractive nonnative interactions are critical for knot formation

In order to achieve a successful folding behavior for our knotted protein, it is necessary to include attractive, nonnative interactions in our model. Moreover, we find that these forces must be included in a rather specific manner. We tested a number of different ways to add interactions to the Gō-forces of the native-centric C_α model (see Supplemental Information). Of these, only one exhibits a reliable folding behavior, consistent with the reversible folding observed for YibK.^{10; 11} This improved model is characterized by, in addition to the original Gō-forces, nonnative attractions between two separate regions of the protein chain, the chain “end” from the native trefoil knot to the C-terminus (positions 122–147) and a region roughly in the middle of the knot (positions 86–93), i.e. nonnative attractive forces are applied to amino acids pairs ij in the set $\tilde{C}_{\text{opt}} = \{i = 86 \dots 93; j = 122 \dots 147\}$ as described in Models and Methods. These two regions are highlighted in Figure 1. We emphasize that different ways of including nonnative attractions yield significantly less reliable folding behavior. For example, the *nonspecific* prescription of including nonnative interactions used in a recent study²⁸ of the present C_α model, which was found to generally increase folding rates at moderate interaction strengths, does not provide any improvement of the knot formation ability vis-à-vis the original Gō-model. In studying the YibK folding kinetics, we will therefore focus on our improved “nonnative” C_α model, which reliably and reproducibly folds YibK into its native state.

Before turning to a more detailed investigation of the folding kinetics of the knotted protein, it is useful to consider more generally the role of nonnative interactions for protein folding. It is obvious that in misfolding and aggregation events, such as those occurring for prion proteins²⁹ and Alzheimer’s A β peptide,³⁰ nonnative interactions play a detrimental role. In ordinary folding processes, however, moderately strong nonnative interactions can in fact play a quite different role. Several theoretical studies^{28; 31; 32; 33; 34} have found that the addition of nonnative interactions increase the rate of folding. This somewhat counter-intuitive effect results from the fact that nonnative attractions can help stabilize the transition state, which effectively lowers the free-energy barrier for folding. Moreover, for proteins with disulfide bonds, it has been found³¹ that nonnative intermediates, i.e. states containing SS bonds not present in the native state, can mediate the formations of the native structure. The partial stabilization of nonnative conformations in the transition state has also been observed in experiments.^{35; 36} For proteins with a deep knot, such as YibK, we find here that nonnative stabilization of the transition state is not simply “beneficial” but may indeed be critical to obtain proper folding.

Folding kinetics of YibK

Figure 3(a) shows an example of a folding trajectory obtained using the nonnative C_α -model, at $T = 0.96$. In sharp contrast to the original Gō-model, the trefoil knot is readily formed and the native state can be reached. Interestingly, despite the very different folding behavior between the original C_α model and the nonnative model, the energetic contribution of the nonnative interactions is small compared to the native interactions, as can be seen in Figure 3 (b). The effect of the nonnative interactions on the overall thermodynamic properties is therefore likely very limited. Instead, the improved “success rate” for folding must originate from favorable changes in the folding dynamics leading to an increased kinetic accessibility of the native state. How this occurs can be understood in the following way: the native trefoil knot in YibK is located approximately 40 residues from the C-terminus and 70 residues from the N-terminus. Assuming therefore that the knot is formed at the C-terminal region, the C-terminus must enter a loop formed by a preceding chain segment (which may or may not be in its native conformation) in order to form the native trefoil knot. In our model, this “loop entrance” by the C-terminus is effectively mediated by the added attractive nonnative interactions. The effect of these transient nonnative interactions during knot formation can be seen in greater detail in Figure 4, which shows snapshots of a folding event including the knot formation. (A movie of this folding trajectory can be seen at <http://www-shakh.harvard.edu/movies/yibk.html>.)

In order to examine more quantitatively the folding process of the YibK monomer, we construct an ensemble of structures in the following way: From 1000 folding simulations we select, for each successful folding trajectory, the first occurring conformation which contains a knot. Note that a conformation is selected only if the chain proceeds to fold directly without any further unknotting/knotting of the chain, so that only one conformation is selected per folding event. Hence, this “knot transition state ensemble” represents structures at the transition between unknotted to knotted states during folding. Figure 5 shows the probability distribution of Q^* , where Q^* is the fraction of native contacts for this knot transition state. The bimodal shape of the distribution shows that folding occurs through one of two parallel “pathways,” distinguished by a late ($Q^* \approx 0.8$) or an early ($Q^* \approx 0.2$) formation of the trefoil knot. In fact, the folding trajectories in Figures 3 and 4, respectively, are representative examples of these two folding pathways. As can be seen from Figure 3(a), the “late-knot” pathway is associated with an (unknotted) intermediate state during folding. This intermediate is quite native-like with approximately 70% of native contacts formed. In the $Q^* = 0.2$ folding pathway, the chain proceeds quickly to the native state once the trefoil knot has been formed (data not shown). Hence, the corresponding intermediate, which contains a knot but otherwise closely resemble conformations in the unfolded state, is short-lived in our model. Early and late intermediates corresponding to nonnative-like and native-like states, respectively, have been observed previously³¹ in lattice simulations, in the context of proteins with disulfide bonds.

Another interesting question in terms of the knot transition state is the size and location of these initial knots. The first and last amino acids of a knot can be determined by repeatedly applying the KMT-reduction scheme, noting the first instance at which no knot is detected as amino acids are sequentially removed from one chain end. In the native structure of YibK, for example, this procedure leads to a knot at positions 75–121. Figure 5(inset) shows the probability distributions of the first and last residues of the knots in our set of transition conformations. These distributions show that, in most case, these initial knots stretch approximately positions 75 to 147, meaning that the knots are usually formed by the threading of the C-terminus. In some of the cases, however, the last residue is found at around position 121, close to the native position. Visual inspection of these structures reveals that in these cases the knot is formed by threading not the C-terminus directly, but a preceding part of the chain folded in a hairpin-like manner. In the majority of the cases, however, the knot is formed by

the threading of the C-terminus followed by a tightening from the C-terminal end to form the fully native trefoil knot stretching positions 75 to 121.

Finally, to address the dimerization of YibK, we perform 100 folding simulations with two interacting chains in a cube with periodic boundary conditions, using our nonnative C_α model at $T = 0.96$. Successful folding to the native dimer occurs in 45 of these trajectories. In the remaining trajectories, binding occurs with one, or both, chains in the unknotted intermediate state resulting in a trapped “misfolded” dimerized state. None of these proceed to form the correct dimer form. In all cases with successful folding and binding, knot formation occurs before the dimerization step, as illustrated in Figure 6. A likely reason for the observed misfolding events is the high chain density used in our 2-chain simulations, which was chosen for computational efficiency. At a lower chain density, each monomer would have a greater probability of forming its knot before contacting another chain and thus dimerizing successfully. Hence, based on the consistent behavior of the successful dimerization events, we find that our model suggests that knot formation occurs before dimerization, i.e. in the first phase of the folding process of the YibK protein.

The obtained folding behavior for our model can be compared to experimental data on the kinetics of YibK obtained by Mallam and Jackson.^{10; 11} They found their data consistent with two parallel folding pathways, with structurally different intermediates I_1 and I_2 , leading to a single monomer state which subsequently dimerizes in a slow, rate-limiting step (see Figure 8 (c) in Ref. 11). This folding scenario fits remarkably well with our simulation results and allows us to discuss various aspects of the proposed folding process of YibK as it relates to knot formation. Based on our simulation results, we identify the two parallel pathways with an “early” and a “late” formation of the trefoil knot. The two corresponding intermediate states are found to be structurally quite distinct in our model: The “knot-late” pathway is associated with an *unknotted*, quite compact intermediate state characterized by approximately 60–80% native contacts. By contrast, the “knot-early” intermediate is a *knotted* but otherwise relatively poorly structured state with approximately 20% native contacts formed. Some experimental support exist for a structural distinction between the two intermediate states. m_{k_u} and m_{k_f} -values obtained for the two fastest kinetic phases of YibK (which corresponds to the formation of I_1 and I_2) are slightly different,¹¹ suggesting that the two intermediates have different solvent accessible surface area. This may, in fact, indicate a difference in compactness. However, a quantitative assessment of structural features such as compactness and nativeness of the two intermediates obtained by our model will require further experimental data. We note in this context also that knot formation is predicted by our model to occur before dimerization, regardless of folding pathway.

In Ref. 11, it was concluded that the parallel folding pathways originate from a heterogeneity in the proline isomer population in the denatured state; 1 of the 10 prolines in YibK is natively in its intrinsically unfavorable *cis* conformation. Our C_α coarse-grained chain model is, of course, unable to address folding in such detail. It could therefore be suggested that in the absence of such details in our model, we should observe only one of the folding pathways, namely I_2 , as it corresponds to the folding pathway initiated from the unfolded state where all proline isomers are in native-like conformations. However, as the non-native prolyl-peptidyl species in the denatured state cause folding through a structurally different folding intermediate I_1 , and not by a slower parallel reaction to the same species, which has been observed for a number of other proteins,¹¹ it is possible that two intrinsic parallel pathways exist for the YibK monomer, which we observe here despite the coarse-grained nature of our model.

Robustness of computational approach

The pattern of nonnative interactions included in our C_α model results in a highly reliable folding behavior for the YibK monomer, as we have seen. It is nonetheless of value to study

the robustness of our computational approach, i.e. to investigate the impact of small variations in the nonnative interaction pattern around the optimal choice, $\tilde{C}_{\text{opt}} = \{i = 86 \dots 93; j = 122 \dots 147\}$. To quantify the folding “success rate” of a particular $\tilde{C}$, we calculate the fraction r of 100 independent folding simulation at $T = 0.96$ which reach the native state N (including, of course, the trefoil knot), within 10^8 MD steps. For $\tilde{C}_{\text{opt}}$, we obtain in $r = 1$, i.e. all 100 test trajectories reach N . Indeed, this is how $\tilde{C}_{\text{opt}}$ was selected. Achieving $r \approx 1$ is in fact crucial as experimental data suggests that no off-pathway intermediates are significantly populated during the folding of YibK.^{10; 11} We find several other choices of $\tilde{C}$ for which $r > 0$, but no other $\tilde{C}$ (significantly deviating from $\tilde{C}_{\text{opt}}$) achieve a fraction r close to 1. Of the tested alternative nonnative interaction patterns (a complete list is given in Supplementary Information), the two most successful are $\tilde{C} = \{i = 74 \dots 81; j = 122 \dots 147\}$ ($r = 0.62$) and $\tilde{C} = \{i = 92 \dots 99; j = 122 \dots 147\}$ ($r = 0.59$), which interestingly involve the β_4 and β_5 regions, respectively.

Sequence characteristics of the SpoU family of proteins

Our chain modeling results suggest that certain specific nonnative interactions are critical to the knot formation of YibK. An important question is then, of course, if signatures of such interactions can be discerned from the amino acid sequence properties of YibK. A natural candidate for these nonnative attractive forces is the hydrophobic effect, as it is strongly attractive and one of the main driving forces for folding. Before turning to this question, however, it must be pointed out that our simulations results are applicable to any protein with a similar native structure regardless of the particular amino acid sequence, as for any native-centric-type model. Hence, the validity of the proposed nonnative interactions is more accurately discussed in terms of the sequence properties of the SpoU family of proteins rather than the YibK sequence alone, as we expect most of these proteins to have native topologies similar to YibK (including the deep trefoil knot).³⁷

In our quest for sequence signals responsible for knot formation, we apply the following logic. It is known that sequence properties can partially be attributed to local structural propensities, i.e. a significant correlation between local structure and sequence properties exists.³⁸ Knot formation, on the other hand, is a global event as it involves specific non-local (along the sequence) amino acid pairs being brought into close proximity. Therefore, we seek sequence patterns located around the knotted region which can not be attributed to local structural propensities. To carry out this program, we construct first a multiple sequence alignment³⁹ of the approximately 1500 sequences available⁴⁰ to date for the SpoU family of proteins. The average hydrophobicity profile of this SpoU sequence alignment is shown in Figure 7(a) (red curve). Second, to find the role of local interactions in this profile, we construct a reference alignment (green curve) based on the *local* structure of the YibK protein, as described in Models and Methods. At each position, this reference alignment represents the typical amino acid composition for small fragments with roughly the same local structure as in YibK. We note in particular two regions which are strongly hydrophobic in the SpoU alignment. These are located in the middle of the native trefoil knot (β_5) and in the C-terminal end region (positions 139–142 of α_5). It is not unreasonable to expect that attractions between these two regions might have an effect similar to the nonnative attractive forces introduced in our C_α model. Further, we note that these two “hydrophobic signals” deviate clearly from the reference alignment, indicating that they can not be attributed simply to the conservation of local structural propensities. The origin of such deviations might arise from several factors, such as protein function or stability, but it is also consistent with the conservation of the type of global non-native attractions suggested by our simulation to be critical to the formation of the trefoil knot. We perform a similar analysis for another parameter, namely the β -sheet propensity.⁴¹ As can be seen from Figure 7(b), the largest deviation between the SpoU alignment and the reference alignment is again in the middle of helix α_5 , around positions 139–142 (note that in

this propensity scale,⁴¹ low values mean a higher propensity for β -sheet structure). It should be pointed out that the hydrophobicity and β -sheet propensity scales are not uncorrelated; strongly hydrophobic residues tend to favor β -sheet structure. Nonetheless, it is interesting that in this region (139–142), which is natively α -helical, the propensity for β -sheet structure is comparable to that of the natively β -sheet structured regions, β_4 , β_5 , and β_6 .

The combination of results from the bioinformatics analysis performed here indicates the following: (i) the possibility of nonnative attractive interactions between α_5 and the β_5 regions; and (ii) this attraction might transform the natively helical segment 139–142 into a transient β -sheet structure with β_5 during the threading of the C-terminal. We point out, however, that hypothesis (ii) in particular must be tested rigorously by experiment. It is interesting though, that α - to β -structure transitions have been found in simulations to be expediated by a close spatial proximity of a polypeptide chain in a pre-formed β -sheet conformation.²⁹ Moreover, amide NH and CO groups of the protein backbone rarely exist in the hydrophobic core of proteins without participating in hydrogen bonds,⁴² and so the energetics of polypeptide chains may favor the formation of backbone-backbone hydrogen bonds between the “loop” and the “entering” chain of a tight trefoil knot. Indeed, in the native structure of YibK the trefoil knot region is flanked by two β -sheets (see β_4 and β_6 in Figure 1), and the same is true for the other known structures of SpoU methyltransferases with deep trefoil knots in their C-terminal. Hence, it could be valuable for the C-terminal region, including helix α_5 , to be able to transiently form β -sheet structure as the knotted structure is formed. Figure 7 (see Supporting Information) shows more closely the amino acid composition of the α_5 helix region, 133–147, in the SpoU alignment. The two most prevalent amino acid types are Alanine and Valine which, in fact, differ sharply in their secondary structure propensities; Alanine strongly favors α -helix structure and Valine is frequently observed in β -sheet structure, which appears consistent with an amino acid segment with an ability to make both α - and β -type secondary structure.

Conclusion

We have investigated the folding behavior of the deeply knotted YibK protein through simulations of a coarse-grained C_α model, and by a bioinformatics study of related homologous proteins. Taken together, our results suggest that the folding of YibK depends critically on *specific* attractive, nonnative interactions. Although a complete thermodynamic characterization of the folding process is computationally beyond the scope of our study, a picture of the folding kinetics have been obtained through a large number of folding trajectories, which fits remarkably well with previously obtained experimental data on YibK. As proposed by Mallam and Jackson,¹¹ we see two parallel folding pathways in the first step of the folding process. We attribute these different pathways to an early and a late formation of the trefoil knot. Moreover, we find that the corresponding intermediate states are structurally different, with the “knot-late” pathway involving a more native-like intermediate than the “knot-early” pathway. In order for successful dimerization to occur, we find that knots in the individual monomers have to be pre-formed.

The origin of these critical nonnative interactions may be two conserved hydrophobic regions, in the middle of the trefoil knot (β_5) and at C-terminal end of the protein (α_5). Hydrophobic attractions between these regions fit reasonably well with the optimal energetics of the nonnative C_α model. From our folding simulations, we see that the role of the nonnative interactions is mainly to enhance the kinetic accessibility of the topologically intricate native state of YibK; in the absence of these attractions, the probability of knot formation is exceedingly low. Despite this dramatic impact on folding kinetics, their energetic contributions are relatively small and they likely do not affect strongly the overall thermodynamics. An interesting possibility suggested by the bioinformatics analysis is that an unusually high β -sheet propensity in the α_5 helix may prompt the formation of transient β -sheet structure in the

C-terminal region, as this chain segment is threaded to form the trefoil knot. Rigorous experimental tests are needed to examine this hypothesis. It would, in particular, be interesting to see if suitable amino acid mutations in positions 129–132 in helix α_5 prohibit, or significantly slow down, the folding of the YibK protein.

Models and Methods

Native-centric C_α model

This C_α chain model was introduced by Clementi et al in 2000.¹² Here we apply it to the residues 1–147 of the YibK protein, as the last 13 residues are unstructured in the native state. The energy function can be expressed as:

$$E_0 = k_{\text{bond}} \sum_i (b_i - b_i^0)^2 + k_{\text{bend}} \sum_i (\theta_i - \theta_i^0)^2 + \sum_i k_{\text{tors}}^{(1)} [1 - \cos(\phi_i - \phi_i^0)] + k_{\text{tors}}^{(3)} [1 - \cos 3(\phi_i - \phi_i^0)] + k_{\text{cont}} \sum_{ij \in C} \left[5 \left(\frac{r_{ij}^0}{r_{ij}} \right)^{12} - 6 \left(\frac{r_{ij}^0}{r_{ij}} \right)^{10} \right] + k_{\text{ev}} \sum_{i < j - 3} \left(\frac{\sigma_{\text{ev}}}{r_{ij}} \right)^{12} \quad (1)$$

where b_i , θ_i , and ϕ_i denotes the (pseudo) bond lengths, bond angles and torsion angles, respectively, of the C_α chain, and r_{ij} denotes the distance between amino acids i and j . Superscript “0” indicates the corresponding values in the native structure. The strengths of the various terms are chosen as in previous studies,^{12; 18} i.e. $k_{\text{bond}} = 100\varepsilon$, $k_{\text{bend}} = 20\varepsilon$, $k_{\text{tors}}^{(1)} = \varepsilon$, and $k_{\text{tors}}^{(3)} = 0.5\varepsilon$, and $k_{\text{cont}} = k_{\text{ev}} = \varepsilon$; ε sets the energy scale of the model and here we use $\varepsilon = 1.0$. C is a set of native contacts, where amino acids ij are considered in contact if any two of its non-H atoms are within 4.5 Å and $|i - j| < 3$. Applied to the YibK structure, this gives 327 and 720 native contacts for the monomer and dimer forms, respectively. In the excluded-volume term, the repulsion between nonnative atom pairs is set to $\sigma_{\text{ev}} = 4.0$ Å. Dimerization is simulated with two interacting chains in a cube with side length 200 Å and with periodic boundary conditions.

Nonnative C_α model

The native-centric model described by Equation 1 is insufficient for our purposes, as it fails to produce the native trefoil knot in YibK. Therefore, we modify its energetics through the addition of attractive, nonnative interactions, i.e. $E = E_0 + E_{\text{nonnat}}$, where E is the new energy function. More precisely, we choose

$$E_{\text{nonnat}} = -k_{\text{nonnat}} \sum_{ij \in \tilde{C}} e^{-(r_{ij} - \sigma_{\text{ev}})^2/2} \quad (2)$$

where $k_{\text{nn}} = 0.8\varepsilon$ is the nonnative interaction strength, slightly weaker than the native interactions. The sum goes over a “nonnative contact set” $\tilde{C}$, i.e. a set of amino acid pairs ij which are not in native contact. Our main focus in the present work is on $\tilde{C}_{\text{opt}} = \{i = 86 \dots 93; j = 122 \dots 147\}$, obtained through a test procedure described in Supplementary Information.

Langevin dynamics

Folding and unfolding kinetics are studied using Langevin dynamics. This means that, for a coordinate x , each C_α atom in our model is described by

$$m\dot{v}(t) = -\frac{\partial E}{\partial x} - m\gamma v(t) + R(t) \quad (3)$$

where m , $v(t)$, $-\frac{\partial E}{\partial x}$, γ , and $R(t)$ are the mass, velocity, conformational force, friction coefficient, and random force, respectively. The random forces R are uncorrelated and drawn from a Gaussian distribution; the variance of this distribution effectively controls the temperature T . To numerically integrate Equation 3, we use the velocity form of the Verlet algorithm (see Ref. 43). All parameters, including friction coefficient γ , mass m , and time-step constant in the numerical integration, are the same as in Refs. 18; 21.

Bioinformatics

The approximately 1500 sequences in the SpoU family of MTases were retrieved from the UniProt database,⁴⁰ and a multiple sequence alignment was built using the ClustalW program with default parameters.³⁹ We compare this SpoU sequence alignment with a reference structural alignment constructed in the following way: Each 9-mer structural segment of YibK is matched against all 9-mer segments in a data set of 1621 nonhomologous protein structures.⁴⁴ The middle (i.e. the 5th) amino acid type of the matched segment is included in the alignment if (i) the C_α cRMSD between the two segments less than 2 Å and (ii) $|K_i - K_j| < 3$, where K_i is the number of intramolecular contacts for the middle C_α atom of segment i (C_α - C_α distance less than 8.5 Å). Each matched segment thus adds one amino acid letter to the reference alignment. Criterion (i) assures that this amino acid type comes from a segment which is locally highly similar to the corresponding segment in the YibK structure. The second criterion (ii) ensures that this segment also has a local environment, in terms of the number of amino acids in close spatial proximity, which is similar to the situation in YibK.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Jason E. Donald for help with standard bioinformatics tools and databases. This work is supported by NIH.

References

1. Mansfield ML. Are there knots in proteins? *Nat Struct Biol* 1994;1:213–214. [PubMed: 7656045]
2. Adams, CC. The knot book: An elementary introduction to the mathematical theory of knots. W. H. Freeman and Company; New York: 1994.
3. Lua RC, Grosberg AY. Statistics of knots, geometry of conformations, and evolution of proteins. *PLoS Comp Biol* 2006;2:350–357.
4. Mansfield ML. Fit to be tied. *Nat Struct Biol* 1997;4:166–167. [PubMed: 9164450]
5. Taylor WR, Lin K. A tangled problem. *Nature* 2003;25:25. [PubMed: 12511935]
6. Virnau P, Mirny L, Kardar M. Intricate knots in proteins: Function and evolution. *PLoS Comp Biol* 2006;2:1074–1079.
7. Murzin AG, Brenner SE, Hubbard T, Clothia C. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 1995;247:536–540. [PubMed: 7723011]
8. Virnau P, Kantor Y, Kardar M. Knots in globule and coil phases of a model polyethylene. *J Am Chem Soc* 2005;127:15102–15106. [PubMed: 16248649]

9. Lim K, Zhang H, Tempczyk A, Krajewski W, Bonander N, Toedt J. Structure of the YibK methyltransferase from *Haemophilus influenzae* (HI0766): a cofactor bound at a site formed by a knot. *Proteins Struct Funct Genet* 2003;51:56–67. [PubMed: 12596263]
10. Mallam AL, Jackson SE. Folding studies on a knotted protein. *J Mol Biol* 2005;346:1409–1421. [PubMed: 15713490]
11. Mallam AL, Jackson SE. Probing nature's knots: The folding pathway of a knotted homodimeric protein. *J Mol Biol* 2006;359:1420–1436. [PubMed: 16787779]
12. Clementi C, Jennings PA, Onuchic JN. How native state topology effects the folding of Dihydrofolate Reductase and Interleukin-1beta. *Proc Natl Acad Sci USA* 2000;97:5871–5876. [PubMed: 10811910]
13. Lim K, Zhang H, Tempczyk A, Krajewski W, Bonander N, Toedt J, Howard A, Eisenstein E, Herzberg O. *Proteins Struct Funct Genet* 2003;51:56–67. [PubMed: 12596263]
14. Michel G, Sauve V, Larocque R, Li Y, Matte A, Cygler M. The Structure of the RlmB 23S rRNA Methyltransferase Reveals a New Methyltransferase Fold with a Unique Knot. *Structure* 2002;10:1303–1315. [PubMed: 12377117]
15. Nureki O, Shirouzu M, Hashimoto K, Ishitani R, Terada T, Tamakoshi M, Oshima T, Chijimatsu M, Takio K, Vassilyev D, Shibata T, Inoue Y, Kuramitsu S, Yokoyama S. An enzyme with a deep trefoil knot for the active-site architecture. *Acta Crystallogr Sect D Biol Crystallogr* 2002;58:1129–1137. [PubMed: 12077432]
16. Zereminski T, Kim Y, Peterson K, Christendat D, Dharamsi A, Arrowsmith C, Edwards A, Joachimiak A. *Proteins Struct Funct Genet* 2003;50:177–183. [PubMed: 12486711]
17. Clementi C, Nymeyer H, Onuchic JN. Topological and energetic factors: What determines the structural details of the transition state ensemble and “on-route” intermediates for protein folding. *J Mol Biol* 2000;298:937–953. [PubMed: 10801360]
18. Kaya H, Chan HS. Solvation effects and driving forces for protein thermodynamics and kinetic cooperativity: How adequate is native-centric topological modeling? *J Mol Biol* 2003;326:911–931. [PubMed: 12581650]Corrigendum: 337, 1069–1070 (2004)
19. Knott M, Chan HS. Criteria for downhill protein folding: Calorimetry, chevron plot, kinetic relaxation, and single-molecule radius of gyration in chain models with subdued degrees of cooperativity. *Proteins Struct Funct Bioinf* 2006;65:373–391.
20. Levy Y, Wolynes PG, Onuchic JN. Protein topology determines binding mechanism. *Proc Natl Acad Sci USA* 2004;101:511–516. [PubMed: 14694192]
21. Wallin S, Chan HS. Conformational entropic barriers in topology-dependent protein folding: Perspectives from a simple native-centric polymer model. *J Phys Condens Matter* 2006;18:S307–S328.
22. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. CHARMM: A program for macromolecular energy minimization, and dynamics calculation. *J Comp Chem* 1983;4:187–217.
23. Case DA, Cheatham TD, Darden T, Gohlke H, Luo R, Merz KM, Onufriev A, Simmerling C, Wang B, Woods R. The Amber biomolecular simulation program. *J Comp Chem* 2005;26:1668–1688. [PubMed: 16200636]
24. Lindahl E, Hess B, Spoel Dvd. GROMACS 3.0: A package for molecular simulation and trajectory analysis. *J Mol Model* 2001;7:306–317.
25. Baker DA. A surprising simplicity to protein folding. *Nature* 2000;405:39–42. [PubMed: 10811210]
26. Dokholyan NV, Li L, Ding F, Shakhnovich EI. Topological determinants of protein folding. *Proc Natl Acad Sci USA* 2002;99:8639–8641.
27. Shimada J, Shakhnovich EI. The ensemble folding kinetics of protein G from an all-atom Monte Carlo simulation. *Proc Natl Acad Sci USA* 2002;99:11175–11180. [PubMed: 12165568]
28. Clementi C, Plotkin SS. The effects of non-native interactions of protein folding rates: Theory and simulation. *Prot Sci* 2004;13:1750–1765.
29. Malolepsza E, Boniecki M, Kolinski A, Pielak L. Theoretical model of prion propagation: A misfolded protein induces misfolding. *Proc Natl Acad Sci USA* 2005;102:7835–7840. [PubMed: 15911770]

30. Luhers T, Ritter C, Adrian M, Riek-Loher D, Bohrmann B, Dobeli H, Schubert D, Riek R. 3D structure of Alzheimer's amyloid -beta(1–42) fibrils. *Proc Natl Acad Sci USA* 2005;102:17342–17347. [PubMed: 16293696]
31. Camacho CJ, Thirumalai D. Modeling the role of disulfide bonds in protein folding: entropic barriers and pathways. *Proteins: Structure, Function, and Genetics* 1995;22:27–40.
32. Cieplak M, Hoang TX. The range of the contact interactions and the kinetics of the Go models of proteins. *Int J Mod Phys C* 2002;13:1231–1242.
33. Li L, Mirny LA, Shakhnovich EI. Kinetics, thermodynamics and evolution of non-native interactions in a protein folding nucleus. *Nature* 2000;7:336–342.
34. Treptow WL, Barbosa MAA, Garcia LG, Araujo AFPd. Non-native interactions, effective contact order, and protein folding: A mutational investigation with the energetically frustrated hydrophobic model. *Proteins: Structure, Function, and Genetics* 2002;49:167–180.
35. Nardo AAD, Korzhnev DM, Stogios PJ, Zarrine-Afsar A, Kay LE. Dramatic acceleration of protein folding by stabilization of a nonnative backbone conformation. *Proc Natl Acad Sci USA* 2004;101:7954–7959. [PubMed: 15148398]
36. Viguera AR, Vega C, Serrano L. Unspecific hydrophobic stabilization of folding transition states. *Proc Natl Acad Sci USA* 2002;99:5349–5254. [PubMed: 11959988]
37. Watanabe K, Nureki O, Fukai S, Ishii R, Okamoto H, Yokoyama S, Endo Y, Hori H. Roles of conserved amino acid sequence motifs in the SpoU (TrmH) RNA methyltransferase family. *J Biol Chem* 2005;280:10368–10377. [PubMed: 15637073]
38. Bystroff C, Simons K, Han K, Baker D. Local sequence-structure correlations in proteins. *Curr Opin Biotechnol* 1996;7:417–421. [PubMed: 8768900]
39. Higgins D, Thompson J, Gibson T. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 1994;22:4673–4680. [PubMed: 7984417]
40. Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, Boeckman B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Muzumder R, O'Donovan C, Redaschi N, Suzek B. The universal protein resource (UniProt): An expanding universe of protein information. *Nucleic Acids Res* 2006;34:D187–D191. [PubMed: 16381842]
41. Street AG, Mayo SL. Intrinsic beta-sheet propensities results from the van der Waals interactions between side chains and the local backbone. *Proc Natl Acad Sci USA* 1995;96:9074–9076. [PubMed: 10430897]
42. Lesk, A. *Introduction to Protein Architecture: The Structural Biology of Proteins*. Oxford University Press; USA, New York: 2001.
43. Veitshans T, Klimov D, Thirumalai D. Protein folding kinetics: Timescales, pathways, and energy landscapes in terms of sequence-dependent properties. *Fold Des* 1997;2:1–22. [PubMed: 9080195]
44. Wang FG, Dunbrack RL. PISCES: A protein sequence culling server. *Bioinformatics* 2003;19:1589–1591. [PubMed: 12912846]

Figure 1.

Structure of the YibK monomer (a) and part of its amino acid sequence (b). Residues 122–147 (orange) are threaded through a trefoil knot consisting of residues 75–121 (underlined in b). In our optimal “nonnative” C_α model, knot formation is mediated by nonnative attractions between residues in 86–93 (magenta) and 122–147 (orange). Notation of secondary structure elements (e.g. α_5 and β_4) are taken from Ref. 11.

Figure 2.

Kinetic behavior of the YibK monomer as obtained for the original Gō-type model.¹² Two different trajectories are shown: a “folding” run started from an extended, unstructured conformation at $T = 0.975 < T_m$ (a) and an unfolding run started from the native structure at $T = 1.04 > T_m$, where T_m is the melting temperature. Q is the fraction of native contacts (‘+’), where two amino acids i and j contact each other if $r_{ij} < 1.2 r_{ij}^0$, where r_{ij} is the C_α - C_α distance between i and j and r_{ij}^0 the corresponding native value. The presence of a knot is determined at each MD step by KMT reduction,⁸ and is indicated in the top panels (‘□’). In the unfolding trajectory (b), the knot from the native structure disappears after approximately 60×10^6 MD steps.

Figure 3.

A typical folding trajectory in Q (a) obtained for the optimal nonnative C_α model, at $T = 0.96$. Also shown is the contribution of native and nonnative contact energy terms (see Models and Methods) during folding (b).

Figure 4. Snapshots of a folding event obtained at $T = 0.97$. The time t is given in units of 10^3 MD steps. The trefoil knot is formed at $t \approx 300$.

Figure 5.

Probability distribution of the fraction of native contacts in the knot transition state, Q^* , at $T = 0.96$. Also shown (inset) are the distributions of the first (l_{first}^*) and last (l_{last}^*) amino acids of the knots in this transition state.

Figure 6.

Example of a successful folding and binding trajectory for two interacting YibK chains. The fraction of native contacts, Q , for each individual chain is shown for chains 1 (red) and 2 (green). The fraction of inter-chain native contacts is also shown (blue).

Figure 7. Profiles of the hydrophobicity (a) and β -sheet propensity (b) obtained for a multiple sequence alignment of the SpoU family (red) and a reference structural alignment (green, see text); both curves have been weakly smoothed using a sliding window of 3 amino acid positions, allowing trends along the sequences to be easily seen.