Topological reaction coordinate captures the folding transition state ensemble in a pierced-lasso protein

Jeffrey K. Noel[†] and Ellinor Haglund[*,‡,¶]

[†]*Structural Biology, Max Delbrück Center for Molecular Medicine, Berlin, Germany*

[‡]*Department of Chemistry, University of Hawaii, Manoa, Honolulu, USA*

[¶]*To whom correspondence should be addressed: ellinorh@hawaii.edu*

E-mail: ellinorh@hawaii.edu

Abstract

Proteins with a pierced-lasso topology (PLTs) have a covalent loop created by a disulfide bond, and the backbone circles back to thread the loop. This threaded topology has unique features compared to knotted topologies; notably the topology is controlled by the chemical environment and the covalent loop remains intact even when denatured. In this work, we use the hormone leptin as our model PLT system and study its folding using molecular dynamics simulations that employ a structure-based (Gō-like) model. We find that the reduced protein has a two-state folding mechanism with a transition state ensemble (TSE) that can be characterized by the reaction coordinate Q, the fraction of native contacts formed. In contrast, the oxidized protein, which must thread part of the polypeptide chain through a covalent loop, has a folding process that is poorly characterized by Q. Instead, we find that a topological coordinate that monitors the residue crossing the loop is able identify the TSE of oxidized leptin. By precisely identifying the predicted TSE, one may now reliably calculate theoretical phi-values for the PLT protein, thereby enabling comparison with experimental measurements. We find the loop-threading constraint leads to non-canonical phi-values that are uniformly small because this PLT protein has a flat energy landscape through the TSE.

Introduction

It is now well established that a subset of proteins are able to tie their backbones into structures containing non-trivial topologies such as knots and slipknots.[1–3] Knotted proteins have been observed to fold spontaneously, although chaperones can speed up the process.[4] These proteins can reach their native structures only by threading one of the termini through a twisted loop. However, knots are only a subset of the topological complexities found within protein structure.[2] Proteins containing covalently-tied loops, through which part of the backbone is threaded, is a relatively unexplored, though even larger, class of topologically-

complex proteins.[3,5,6]

Threaded topologies in proteins were described already in the early 80s, when Connolly *et al.*. described topological features of entanglement in folded protein backbones.[7] They showed that in some cases the protein backbone may circle back around itself and form a loop, closed by a contact, which encircles part of the chain. When this contact is anchored by a disulfide, as in the Pierced Lasso Topology (PLT),[5,8–11] the topological constraint can remain even in the denatured ensemble as has been described for knotted proteins.[12] In PLTs, a covalent loop forms a large macrocycle where part of the backbone is threaded through. The most common linker is a disulfide bond where the chemical conditions determine the threaded topology, i.e., the non-trivial topology is promoted under oxidizing conditions while reducing conditions will break the disulfide bond.

PLTs are found in over 600 PDB structures, and in all kingdoms of life, corresponding to 18% of a non-redundant set of disulfide-containing proteins.[5,11] They are mainly found in the extracellular matrix and organelles with an oxidizing environment. While clear connections between disulfide bonds and protein function/stability[13,14] have been established, the impact of threaded, covalent loops in proteins, given their prevalence, remains unexplored.

Here, we use the hormone leptin as a model system to study the impact of PLT constraints on the folding free energy landscape. The PLT in leptin is created by a disulfide bond between C96 and the C-terminal residue C146. In the native state, the covalent loop is threaded by the N-terminal part of the protein, crossing the loop at L51 (Figure 1A). Under reducing conditions, leptin can fold without threading a loop. However, under oxidizing conditions, the folding is complicated by the requirement of threading. This switchability makes leptin an ideal model system for careful study of loop-threading processes. Notably, the interpretation of results in the leptin system is not complicated by the low sequence identity encountered when comparing structurally-homologous knotted/unknotted proteins.[15] Whether studying knotted or PLT proteins, the topological state has not yet been directly probed experimentally without folding mechanism altering techniques, e.g. single-

molecule pulling.[16,17] Therefore molecular dynamics simulations are extremely valuable for the interpretation of results.

Leptin has been previously explored using coarse-grained folding models based in energy landscape theory. These models have been shown to capture the folding mechanisms of many proteins, particularly well for small two-state folding proteins.[18] More recently, these models have been employed in the study of knotted and slipknotted proteins.[19–23] When used to study leptin,[9] this model suggested that oxidized leptin can thread through parallel routes, plugging or slipknotting, but dominated by slipknotting.[9] Perhaps as expected, the oxidized protein folded more slowly than the reduced protein. However, the folding free energy barrier according to the standard reaction coordinate Q, the fraction of native contacts formed, was lower for the oxidized protein. This contradiction, which has yet to be resolved, suggests that Q is not able to capture the folding transition state ensemble (TSE) of oxidized leptin and that the topological constraint significantly alters the folding landscape.

In this study, we show that while Q adequately identifies the TSE for the reduced protein, Q is unable to capture the TSE for the oxidized protein. We find that the crossing reaction coordinate C, which monitors the sequence position of the residue crossing the covalent loop, is able to identify the TSE.[11] We utilize a combination of Q and C to characterize the folding of leptin. Identification of the TSE allowed us to compare phi-values between the reduced and oxidized protein. The oxidized phi-values are predicted to be uniformly low, in striking contrast to the reduced protein. A major challenge with experimental techniques is to determine if the denatured state is unthreaded.[12] Our results suggest phi-value analysis can be used to interrogate threading in the PLT protein Leptin.

Methods

Structure-based protein folding simulations

The folding process was simulated using a C_α structure-based model.[24] The protein structure was coarse-grained to a single bead per residue and the energy function was based on the leptin crystal structure.[25] Attractive interactions are introduced between "native contacts," residues that are neighboring in the crystal structure,[26] while all other pairs interact only through excluded volume. The backbone structure was biased toward the native conformation. This general strategy is commonly used[27] to study protein folding transition-state ensembles since it is computationally tractable and explicitly encodes a simple realization of a funneled and minimimally-frustrated protein folding energy landscape.[28–30]

Specifically, the protein potential energy is given by,

$$\begin{aligned} V &= \sum_{ij\in\text{bonds}} \epsilon_\text{b}(r^{ij}-r_0^{ij})^2 + \sum_{ijk\in\text{angles}} \epsilon_\theta \left(\theta^{ijk}-\theta_0^{ijk}\right)^2 + \sum_{ijkl\in\text{dihedrals}} \epsilon_\phi F_\text{D}\left(\phi^{ijkl}-\phi_0^{ijkl}\right) \\ &\quad \sum_{ij\in\text{contacts}} \epsilon C_{ij}(r^{ij}-r_0^{ij}) + \sum_{ij\notin\text{contacts}} \epsilon R_{ij}(r^{ij}-r_0^{ij}) \end{aligned} \tag{1}$$

Subscript zeros denote the value in the native state. F_D is a typical dihedral-like potential

$$F_\text{D}(x) = (1-\cos x) + 0.5(1-\cos 3x)$$

and C_{ij} is a native contact interaction consisting of a Gaussian well G_{ij} coupled a r^{-12} repulsion R_{ij} in such a way that fixes the minimum at (r_0,ϵ),[31] homogenizing the excluded volume for all residue pairs.

$$C_{ij} = [(1+R_{ij})(1+G_{ij}-1]$$

with

$$G_{ij} = -e^{(r^{ij}-r_0^{ij})^2/2\sigma^2}$$

and

$$R_{ij} = (r_0/r^{ij})^{12},$$

where $r_0 = 4$ Å sets the bead excluded volume. The values of the energetic parameters are $\epsilon_{\mathrm{b}} = 20000\epsilon$, $\epsilon_\theta = 40\epsilon$, $\epsilon_\phi = \epsilon$, $\epsilon_{\mathrm{N}} = \epsilon$, where ϵ is the reduced energy unit. The time step is 0.0005 reduced time units.

Simulations were performed using GROMACS 4.5.3[32] containing code edits implementing Gaussian contact interactions (available at http://smog-server.org). The necessary input files for GROMACS were generated using the SMOG2 software[33] using the forcefield "SBM_CAgauss". The input structure was residues 3-146 from PDB ID 1AX8.[25]

The computationally slowest folding protein studied was wild-type oxidized leptin. Its trajectory of 5.2e11 time steps contained 163 folding or unfolding transitions.

Analysis of folding simulations

Reaction coordinates Q and C

The number of native contacts formed in a simulation snapshot was calculated as in ref[24]

$$Q = \sum_{ij \in \mathrm{contacts}} \theta(1.2r_0^{ij} - r^{ij}), \tag{2}$$

where θ is the Heaviside function.

The detection of the lasso topology and determining the depth of the threading coordinate C follows the method employed by the LassoProt web server[6,34] which is detailed in ref.[11] Briefly, it consists of four steps: (i) defining the covalent loop; (ii) spanning a minimal surface on this loop; (iii) detection of surface piercings; (iv) reduction of artificial piercings. The covalent loop in leptin is given by residues 96-146 (Figure 1). The minimal area spanning the closed loop is a triangulated surface, following techniques in computer graphics that are optimized for proteins. Notably, this technique takes into account surface orientation

(to detect supercoiling). The piercings arise as an intersection of the vector joining two neighboring C_α atoms (which points along the direction of increasing residue index) with the oriented triangles. The crossing residue is then defined as the spatially closest of the two neighboring C_α atoms to the surface. LassoProt additionally reports the orientation of the crossing (the sign of the dot product between the crossing vector and the normal to the triangle). leptin^{ox} crosses the loop in a positive direction. All snapshots with multiple crossings or with negative crossings were removed before creating the plots shown in this paper for clarity. The rare negative crossings corresponded to shallow plugs or slipknots that were topological traps and could not progress to the native state. Multiple crossings were due to an N-terminal helix that crossed the loop while lying nearly parallel to the loop surface. C becomes ambiguous in this case, but these events were rare ($<$0.01%). Inclusion of these multiply-crossed snapshots in the analysis, by taking the value of C as either the largest or smallest crossing into the statistics, showed no visible change to the histograms or TSE metric.

Identifying the TSE

The transition state ensemble is defined by the ensemble of structures that are at a multidimensional separatrix, those equally likely to fold or unfold. This structural ensemble therefore also represents the ensemble of states at the rate limiting step to the folding reaction. The structures found at a peak in the potential of mean force ($-\ln(P(Q))$) along a reaction coordinate is not necessarily indicative of the TSE or "folding barrier." Further characterization is required to ensure that the peak contains TSE structures.

One computationally-efficient method to analyze a reaction coordinate's ability to capture the TSE is a Bayesian analysis of transition paths.[35] A transition path (TP) comprises the portion of the trajectory corresponding to the last time the protein was in the folded (unfolded) ensemble and the first time the protein is in the unfolded (folded) ensemble. The denatured ensemble was defined by $Q < 100$ and C unthreaded, and the folded ensemble was

defined by $Q > 345$ and $45 < C < 55$. To determine the extent to which a coordinate X (or combination of coordinates) could uniquely identify the TSE, the conditional probability of being on a transition path is calculated as function of coordinate, $p(\mathrm{TP}|X)$. This quantity peaks at 0.5 for a coordinate that perfectly captures the TSE of a two-state folding process because a TSE configuration is equally likely to either complete the transition path or reverse and any other configuration is even more likely to reverse.

Phi-value calculation

The phi-value of residue i is defined as:

$$\phi_i = \frac{\Delta\Delta G^{\mathrm{TSE}\to\mathrm{D}}}{\Delta\Delta G^{\mathrm{F}\to\mathrm{D}}}, \tag{3}$$

where $\Delta\Delta G$ refers to the change in free energy upon substitution of residue i when comparing the transition state to the denatured state (TSE $\to$ D) or the folded state to the denatured state (F $\to$ D). Given a definition of the three ensembles TSE, N, D, this can be readily estimated from folding simulations by calculating Q_i, the fraction of native contacts formed by residue i. Namely,

$$\phi_i \approx \frac{Q_i^{\mathrm{TSE}} - Q_i^{\mathrm{D}}}{Q_i^{\mathrm{F}} - Q_i^{\mathrm{D}}} \tag{4}$$

provides the estimate, since $Q_i^X \propto \langle \Delta G_i \rangle^X$, where $\langle \Delta G_i \rangle^X$ is the stabilizing free energy provided by residue i within ensemble X. Note that all contacts are equally stabilizing in the employed simulation model (see Eq. 1).

Results and Discussion

Wild-type leptin contains a disulfide bond between C96 and C146, forming a 50-residue-long covalent loop under oxidizing conditions (Figure 1). This loop is threaded by the N-terminal portion of the protein to reach the native state. In a reducing environment, leptin is able to

fold without a topological constraint. The folding of the reduced protein (leptin^{red}) and the oxidized protein (leptin^{ox}) was simulated at folding temperature using an unbiased coarse-grained structure-based model until approximately 100 folding events were captured. The potential of mean force (PMF) with respect to the reaction coordinate Q is shown in Figure 2A. Both proteins appear to exhibit a typical two-state folding free energy landscape, with an unfolded basin at low Q, a folded basin at high Q, and a barrier in between. The shoulder in leptin^{ox} is due to the transient fraying of the N-terminal helix[8] (Figure S1).

A topological coordinate captures the TSE for leptin^{ox}

Notably, there is a contradiction when the dynamics is described in terms of diffusion along the collective coordinate Q:[36] leptin^{red} folds 4.0 times faster despite having a 1.5 kT higher free energy barrier. The discrepancy between barrier height and folding rate suggests that Q is not sufficient to capture the TSE. Whether Q is capturing the TSE can be checked by using the conditional probability $p(\mathrm{TP}|Q)$, the probability of being on a transition path given a value Q of the reaction coordinate.[35] For two-state folding proteins, a peak in $p(\mathrm{TP}|X)$ near 0.5 indicates that the reaction coordinate X is uniquely identifying the TSE. The peak in $p(\mathrm{TP}|Q)$ above 0.3 shows that Q is adequately capturing the TSE for leptin^{red} (Figure 2A). However, the peak below 0.1 for leptin^{ox} indicates that Q fails to classify the TSE, and therefore, under-estimates the folding barrier in leptin^{ox} .

Since leptin^{ox} folding has the added complexity of threading the loop, one can additionally probe the topological state of leptin^{ox} . A natural coordinate to monitor the threading progress is the crossing residue C, namely the sequence position of the residue that is crossing the covalent loop[34] (see Methods for details). The PMF for C is shown in Figure 2B. The native basin has $C \approx 50$. Remarkably, $p(\mathrm{TP}|C)$ peaks near 0.5, meaning that C identifies the TSE. There are two peaks in $p(\mathrm{TP}|C)$, at $C \approx 70$ and $C \approx 25$. A unified representation of the folding along Q and C clarifies why Q is unable to distinguish the TSE in leptin^{ox} (Figure 2C,D). Q is between 0.25 and 0.3 for the conformations populating the peaks in $p(\mathrm{TP}|C)$.

Therefore, the TSE in leptin^{ox} is mapped to the denatured ensemble in the one dimensional Q representation (Figure 2A), and Q alone does not classify TSE structures differently from the denatured ensemble.

To our knowledge this study is the first to report that a topological coordinate is identifying a protein folding TSE. Topological coordinates have previously been used in the analysis of equilibrium folding studies of knotted proteins. For example, studies of the smallest knotted protein showed that the transition from unknotted to knotted coincided either with the peak in PMF(Q)[20] or at even higher Q.[37] Additionally, a coordinate related to C, the depth of slipknotting, was shown to be downhill towards the native ensemble and interpreted as post-TSE.[38] The coordinate C used here has some differences when compared to the typical collective coordinates commonly used to describe protein folding, e.g., Q or root mean squared deviation from the native state. First, it is not a collective coordinate that can identify the native ensemble. Second, C is undefined when there is no crossing of the covalent loop, meaning that all uncrossed conformations are absent from Figure 2B,C,D. We note that no transitions to high Q were observed in leptin^{ox} in the absence of threading (Figure S1). Third, C is discontinuous in the sense that uncrossed conformations transition to crossed at both high and low C. This last feature enables C to automatically differentiate slipknotting and plugging routes as discussed in the next section.

Two threading routes lead to two separate TSEs

The minimal spanning surface of the covalent loop has two sides described by the direction of the surface normal.[6] The direction is given by the right hand rule with the fingers curving along the direction of increasing residue index. The two sides of the loop create two threading routes, termed slipknotting and plugging (Figure 1D), in analogy to the threading of twisted loops in knotted proteins.[20] During folding, the N-terminal region crosses the loop surface from the negative to the positive side during slipknotting, whereas it crosses from the positive to the negative side during plugging (Figure 1D). During unfolding the process is reversed.

From the perspective of the crossing residue C, a folding (unfolding) slipknotting route starts (ends) with C near 90, and a plugging route starts (ends) with C near 1.

We find that both routes occur in leptin^{ox} simulations (Figure 3), however the slipknotting route dominates 15:1, consistent with previous work.[9] Interestingly, the two routes have disparate TSEs, identified by the peaks in $P(\mathrm{TP}|C)$ (Figure 2B). The slipknotting route climbs a barrier until reaching its TSE at $C \approx 70$, whereas the plugging route is nearly barrierless through its TSE at $C \approx 25$. The dominance of slipknotting comes from the greater sampling of shallow slipknots as compared to shallow plugs. This is likely explained entropically: the residues participating in the shallow slipknot are directly next to the loop, while in contrast, the N-terminus is separated by a loop length of 90 residues. For the remainder of the paper, leptin^{ox} TSE refers to the more-likely and well-sampled slipknotting TSE and defined by $68 \leq C \leq 71$.

The leptin^{ox} TSE contains a slipknotting αC and an extended N-terminal portion (Figure S2). A detailed look at the average contact map in the TSE shows that the strongest native contacts are found at the location of the disulfide. Relative to the unthreaded configurations, the TSE has lesser formation of secondary structure and αB-αC contacts, due to the necessity to thread sequentially αC and then αB through the loop. As the overall Q is on average constant during threading, we find that some αC-αD native contacts are more likely formed, which makes sense as αC is threading the covalent loop in the TSE.

Topology dramatically alters phi-value distribution

With the TSEs identified, calculating phi-values from the folding simulations becomes possible. Phi-value analysis is an experimental technique that characterizes the degree to which a residue is participating in the TSE by observing the changes in protein stability and folding kinetics upon mutation. Here, simulated phi-values are obtained through a ratio of the contact formation in the TSE relative to the native basin (see Methods). The simulated phi-values are shown in Figure 4. leptin^{red} has a folding nucleus containing αB, α4, and

αD. The distribution of phi-values is dramatically altered for leptin^{ox} , the phi-values are uniformly low. This result could have been expected because Q in the leptin^{ox} TSE is in line with the denatured ensemble. Since the phi-value experiment measures only the relative change between the denatured and folded ensembles upon mutation, the contact formation in the denatured ensemble is subtracted out during the phi-value computation.

The striking difference in phi-value distribution is important because it provides an assay to detect whether *in vitro* folding experiments are capturing leptin^{ox} threading. Obtaining threaded leptin is trivial, one simply moves leptin^{red} to oxidizing conditions. Whether leptin^{ox} in fact unthreads during kinetic folding experiments upon the addition of denaturant and then re-threads as denaturant is washed away is difficult to prove. It has been shown that deeply knotted proteins can maintain their knotted topology even in the denatured state.[12] PLTs differ from knotted proteins in that the PLT loop cannot tighten, suggesting that unthreading should be faster in PLTs. Nonetheless, the unthreaded state is difficult to identify without single-molecule techniques that disrupt the folding pathway[16,17,39] and unthreading must be verified for proper interpretation of experimental results. If the kinetic refolding experiments are capturing full threading, then the phi-value distribution should be similar to Figure 4B. However, if the kinetic folding experiments are rather capturing refolding from a deeply threaded configuration, the phi-value distribution is predicted to be similar to leptin^{red} , but still sufficiently different to be able to differentiate it from a reduced disulfide bond (Figure S3).

Conclusions

The pierced-lasso topology (PLT) protein, whose topology is tunable based on the environmental conditions, offers fascinating possibilities both in nature and in the laboratory. In the current study, we have mapped the folding free energy landscape of leptin, the best studied PLT, using coarse-grained protein folding simulations. We showed that a topological coor-

dinate C, which monitors the residue crossing the covalent loop, captures the TSE. Precise identification of the TSE allowed us to calculate theoretical phi-values, and we found that the phi-values are uniformly small. These non-canonical phi-values are not shared by the reduced protein and are a consequence of the topological constraint imposed by the covalent loop during folding.

The concept of protein folding funnels[29,40] explains why proteins, given their enormous configuration space, do not require a precise kinetic pathway to reach their native states. Protein folding should be considered as a progressive organization of partially folded structures guided by the gradient in solvent-averaged free energy. A protein is a polypeptide that evolution has molded to make native structures on average more stabilizing than non-native ones. The structure-based model employed in this study is a so-called "perfect funnel model," where only native contacts contribute stabilizing energy and non-native contacts are neglected. In this model Q inversely correlates with the energy of the system, and is its natural reaction coordinate. Leptin$^{\mathrm{red}}$ behaves as expected, progression of folding through the rate limiting TSE takes the protein down the funnel or higher in Q. Leptin$^{\mathrm{ox}}$, however, behaves unexpectedly, progress along Q is delayed until after the TSE, despite the perfectly funneled model. It appears that this arises due to location of the covalent loop relative to the folding nucleus, which in leptin$^{\mathrm{red}}$ is comprised by αB-αD contacts. These contacts cannot be formed simultaneously with threading since αB fills the covalent loop. Thus, leptin$^{\mathrm{ox}}$ embarks on an unbiased, stochastic search for a threaded state, whereupon it can begin forming native contacts.

The real leptin protein is not perfectly funneled, but instead is minimally frustrated. This frustration can greatly lengthen unbiased search times. To overcome this, leptin may have evolved some non-native contacts that aid in threading or act as gatekeepers preventing long-lived unthreaded traps. Inclusion of these effects into the simulation energy model may tilt the landscape to have an energetic bias toward the TSE. Of course, in nature, leptin may fold only in a reducing environment, obviating the need for evolution to optimize

threading. Although, since leptin is a rather unstable protein and is a long-lived signalling hormone, it is more likely that leptin commonly finds itself both unfolded and oxidized. Thus, understanding the energy landscape of leptin and how leptin functions is contingent on new experiments, guided by our findings, unraveling whether and how it threads and folds.

Acknowledgements

The research reported in this manuscript is supported by the National Science Foundation award number CHE2145906 and the Hawaii Community Foundation award number HCF40846 (013357-00002). JKN would like to thank Oliver Daumke for the use of his compute resources.

References

(1) Taylor, W. R. A deeply knotted protein structure and how it might fold. *Nature* **2000**, *406*, 916–919.

(2) Sulkowska, J. I. On folding of entangled proteins: knots, lassos, links and -curves. *Current Opinion in Structural Biology* **2020**, *60*, 131–141.

(3) Dabrowski-Tumanski, P.; Rubach, P.; Goundaroulis, D.; Dorier, J.; Sułkowski, P.; Millett, K. C.; Rawdon, E. J.; Stasiak, A.; Sulkowska, J. I. KnotProt 2.0: a database of proteins with knots and other entangled structures. *Nucleic Acids Research* **2018**, *47*, D367–D375.

(4) Lim, N. C.; Jackson, S. E. Mechanistic Insights into the Folding of Knotted Proteins In Vitro and In Vivo. *Journal of Molecular Biology* **2015**, *427*, 248–258.

(5) M., S. J.; E., H. Topological Twists in Nature. *Trends Biochem. Sci.* **2021**, *46*, 461–471.

(6) Dabrowski-Tumanski, P.; Niemyska, W.; Pasznik, P.; Sulkowska, J. I. LassoProt: server to analyze biopolymers with lassos. *Nucleic Acids Res.* **2016**, *44*, W383–9.

(7) Connolly, M. L.; Kuntz, I. D.; Crippen, G. M. Linked and threaded loops in proteins . *Biopolymers* **1980**, *19*, 1167–1182.

(8) Haglund, E.; Sułkowska, J. I.; He, Z.; Feng, G.-S.; Jennings, P. A.; Onuchic, J. N. The unique cysteine knot regulates the pleotropic hormone leptin. *PLoS ONE* **2012**, *7*, e45654.

(9) Haglund, E.; Sulkowska, J. I.; Noel, J. K.; Lammert, H.; Onuchic, J. N.; Jennings, P. A. Pierced Lasso Bundles are a new class of knot-like motifs. *PLOS Comput. Biol.* **2014**, *10*, e1003613.

(10) Haglund, E. Engineering covalent loops in proteins can serve as an on/off switch to regulate threaded topologies. *J. Phys.: Condens. Matter* **2015**, *27*, 354107.

(11) Niemyska, W.; Dabrowski-Tumanski, P.; Kadlof, M.; Haglund, E.; Sułkowski, P.; Sulkowska, J. I. Complex lasso: new entangled motifs in proteins. *Sci. Rep.* **2016**, *6*, 36895–36810.

(12) Mallam, A. L.; Rogers, J. M.; Jackson, S. E. Experimental detection of knotted conformations in denatured proteins. *Proc. Nat. Acad. Sci. USA* **2010**, *107*, 8189–8194.

(13) Dombkowski, A. A.; Sultana, K. Z.; Craig, D. B. Protein disulfide engineering. *FEBS Letters* **2014**, *588*, 206–212, Protein Engineering.

(14) Fass, D. Disulfide Bonding in Protein Biophysics. *Annual Review of Biophysics* **2012**, *41*, 63–79, PMID: 22224600.

(15) Sriramoju, M. K.; Yang, T.-J.; Hsu, S.-T. D. Comparative folding analyses of unknotted versus trefoil-knotted ornithine transcarbamylases suggest stabilizing effects of protein knots. *Biochemical and Biophysical Research Communications* **2018**, *503*, 822–829.

(16) Ziegler, F.; Lim, N. C.; Mandal, S. S.; Pelz, B.; Ng, W. P.; Schlierf, M.; Jackson, S. E.; Rief, M. Knotting and unknotting of a protein in single molecule experiments. *Proceedings of the National Academy of Sciences of the United States of America* **2016**, *113*.

(17) Rivera, M.; Hao, Y.; Maillard, R. A.; Baez, M. Mechanical unfolding of a knotted protein unveils the kinetic and thermodynamic consequences of threading a polypeptide chain. *Scientific Reports* **2020**, *10*, 9562.

(18) Whitford, P. C.; Sanbonmatsu, K. Y.; Onuchic, J. N. Biomolecular dynamics: order-disorder transitions and energy landscapes. *Reports on progress in physics Physical Society (Great Britain)* **2012**, *75*, 076601.

(19) Sulkowska, J.; Sułkowski, P.; Onuchic, J. Dodging the crisis of folding proteins with knots. *Proc. Nat. Acad. Sci. USA* **2009**, *106*, 3119–3124.

(20) Noel, J. K.; Sulkowska, J. I.; Onuchic, J. N. Slipknotting upon native-like loop formation in a trefoil knot protein. *Proc. Nat. Acad. Sci. USA* **2010**, *107*, 15403–15408.

(21) Bölinger, D.; Sulkowska, J.; Hsu, H.-P.; Mirny, L. A.; Kardar, M.; Onuchic, J. N.; Virnau, P. A Stevedore's protein knot. *PLOS Comput. Biol.* **2010**, *6*, e1000731.

(22) Sułkowska, J. I.; Noel, J. K.; Onuchic, J. N. Energy landscape of knotted protein folding. *Proc. Nat. Acad. Sci. USA* **2012**, *109*, 17783–17788.

(23) Sulkowska, J. I. On folding of entangled proteins: knots, lassos, links and θ-curves. *Curr. Opin. Struct. Biol.* **2020**, *60*, 131–141.

(24) Clementi, C.; Nymeyer, H.; Onuchic, J. N. Topological and energetic factors: what determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? an investigation for small globular proteins. *J. Mol. Biol.* **2000**, *298*, 937–953.

(25) Zhang, F.; Basinski, M. B.; Beals, J. M.; Briggs, S. L.; Churgay, L. M.; Clawson, D. K.; DiMarchi, R. D.; Furman, T. C.; Hale, J. E.; Hsiung, H. M. et al. Crystal structure of the obese protein Ieptin-E100. *Nature* **1997**, *387*, 206–209.

(26) Noel, J. K.; Whitford, P. C.; Onuchic, J. N. The Shadow Map: A General Contact Definition for Capturing the Dynamics of Biomolecular Folding and Function. *J. Phys. Chem. B* **2012**, *116*, 8692–8702.

(27) Noel, J. K.; Onuchic, J. N. The Many Faces of Structure-Based Potentials: From Protein Folding Landscapes to Structural Characterization of Complex Biomolecules. **2012**, 31–54.

(28) Bryngelson, J. D.; Wolynes, P. G. Intermediates and barrier crossing in a random energy model (with applications to protein folding). *J. Phys. Chem.* **1989**, *93*, 6902–6915.

(29) Leopold, P. E.; Montal, M.; Onuchic, J. N. Protein folding funnels: a kinetic approach to the sequence-structure relationship. *Proc. Nat. Acad. Sci. USA* **1992**, *89*, 8721–8725.

(30) Onuchic, J. N.; Wolynes, P. G. Theory of protein folding. *Curr. Opin. Struct. Biol.* **2004**, *14*, 70–75.

(31) Lammert, H.; Schug, A.; Onuchic, J. N. Robustness and generalization of structure-based models for protein folding and function. *Proteins: Struct., Funct., Bioinf.* **2009**, *77*, 881–891.

(32) Pronk, S.; Páll, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; van der Spoel, D. et al. GROMACS 4.5: a high-throughput

and highly parallel open source molecular simulation toolkit. *Bioinformatics* **2013**, *29*, 845–854.

(33) Noel, J. K.; Levi, M.; Raghunathan, M.; Lammert, H.; Hayes, R. L.; Onuchic, J. N.; Whitford, P. C. SMOG 2: A Versatile Software Package for Generating Structure-Based Models. *PLOS Comput. Biol.* **2016**, *12*, e1004794.

(34) Gierut, A. M.; Niemyska, W.; Dabrowski-Tumanski, P.; Sułkowski, P.; Sulkowska, J. I. PyLasso: a PyMOL plugin to identify lassos. *Bioinformatics* **2017**, *33*, 3819–3821.

(35) Best, R. B.; Hummer, G. Reaction coordinates and rates from transition paths. *Proc. Nat. Acad. Sci. USA* **2005**, *102*, 6732–6737.

(36) Socci, N. D.; Onuchic, J. N.; Wolynes, P. G. Diffusive dynamics of the reaction coordinate for protein folding funnels. *Journal Of Chemical Physics* **1996**, *104*, 5860–5868.

(37) a Beccara, S.; Skrbić, T.; Covino, R.; Micheletti, C.; Faccioli, P. Folding pathways of a knotted protein with a realistic atomistic force field. *PLOS Comput. Biol.* **2013**, *9*, e1003002.

(38) Noel, J. K.; Onuchic, J. N.; Sulkowska, J. I. Knotting a protein in Explicit Solvent. *The journal of physical chemistry letters* **2013**, *4*, 3570–3573.

(39) Sułkowska, J. I.; Sułkowski, P.; Szymczak, P.; Cieplak, M. Untying knots in proteins. *J. Am. Chem. Soc.* **2010**, *132*, 13954–13956.

(40) Bryngelson, J. D.; Wolynes, P. G. Spin Glasses and the Statistical Mechanics of Protein Folding. *Proc. Nat. Acad. Sci. USA* **1987**, *84*, 7524.

Figures

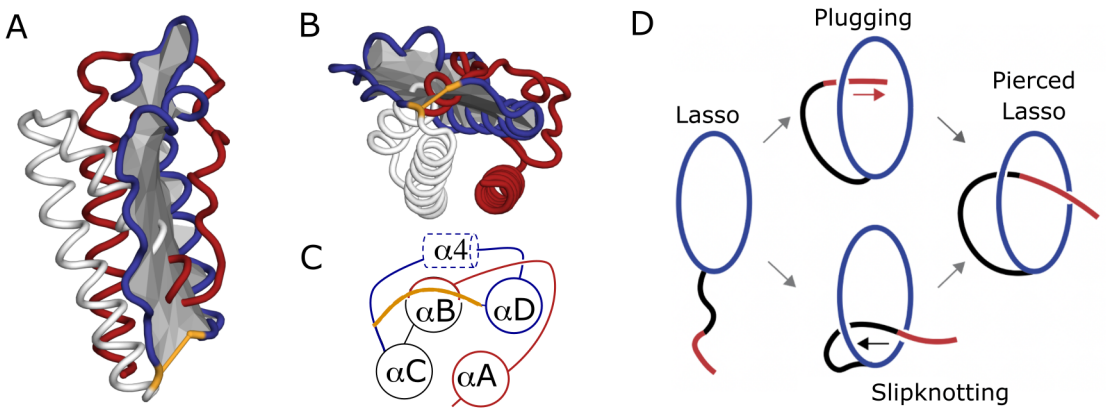

Figure 1: The Pierced Lasso Topology (PLT) in Leptin. A) A PyLasso representation[34] of leptin, where the covalent loop (blue), fixed by one disulfide bond (yellow), is threaded by the N-terminal part of the protein. The minimal surface spanning the loop is colored gray. In the native state, the loop crossing roughly divides the N-terminal chain in two, with 50 amino acids on one side (white) and 50 on the other (red). In yellow is the native disulfide bond (solid line). B) A rotated view with the native disulfide bond on top and C) a cartoon of (B) with labeled helices. D) The two threading routes, plugging versus slipknotting.

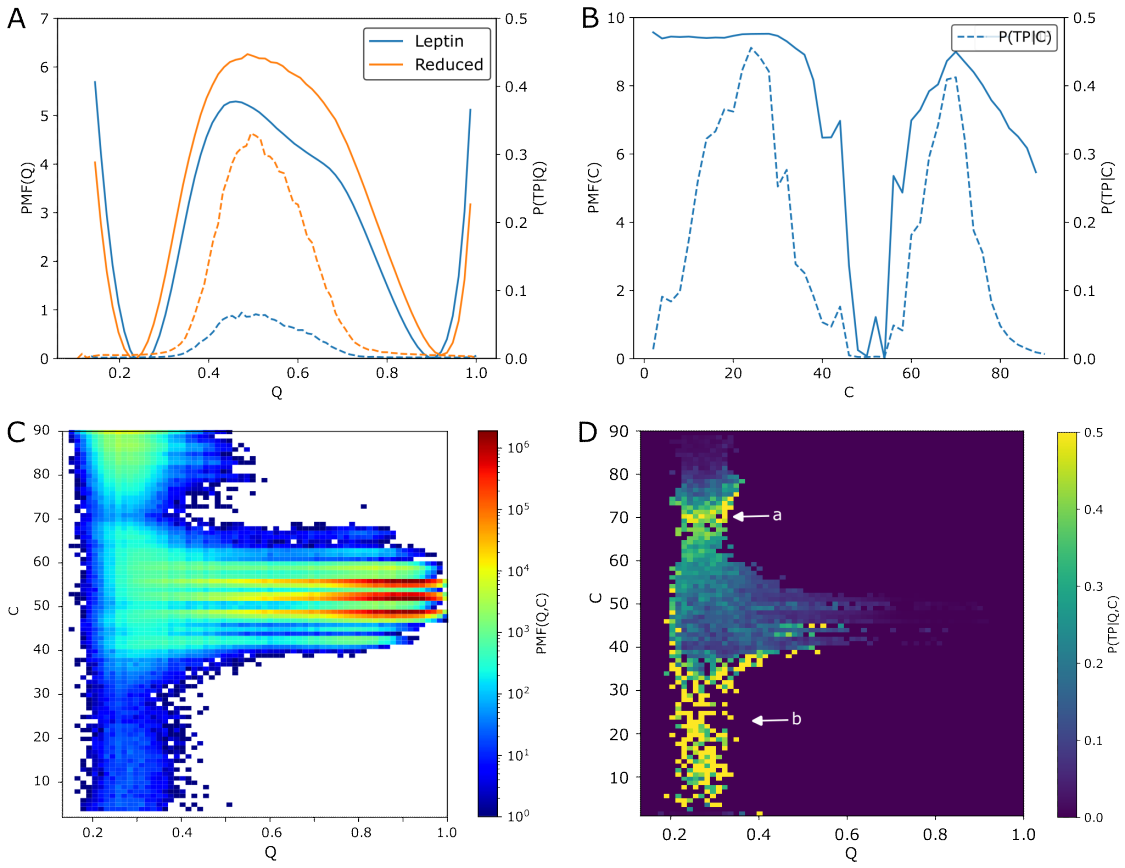

Figure 2: A) (Solid lines) Potential of mean force (PMF) with respect to Q, the fraction of native contacts formed, for leptin$^{\text{ox}}$ (blue) leptin$^{\text{red}}$ (orange). (Dashed lines) Probability to be on a transition path given a particular value of Q ($p(\text{TP}|Q)$). B) (Solid line) PMF with respect to C, the residue crossing the loop. (Dashed line) $p(\text{TP}|C)$. Two peaks correspond to the plugging route at low C and the slipknotting route at high C. C) Two dimensional PMF over Q adn C. Threading, represented moving vertically on the plot, clearly occurs as values of Q corresponding to the unfolded state. D) Probability to be on a transition path given Q and C. C value corresponding to (a) the slipknot peak, or (b) the plugging peak, in panel B. The plugging route is not sufficiently well sampled for the two dimensional representation. Note, a value of 0 is plotted for bins containing fewer than 4 samples.

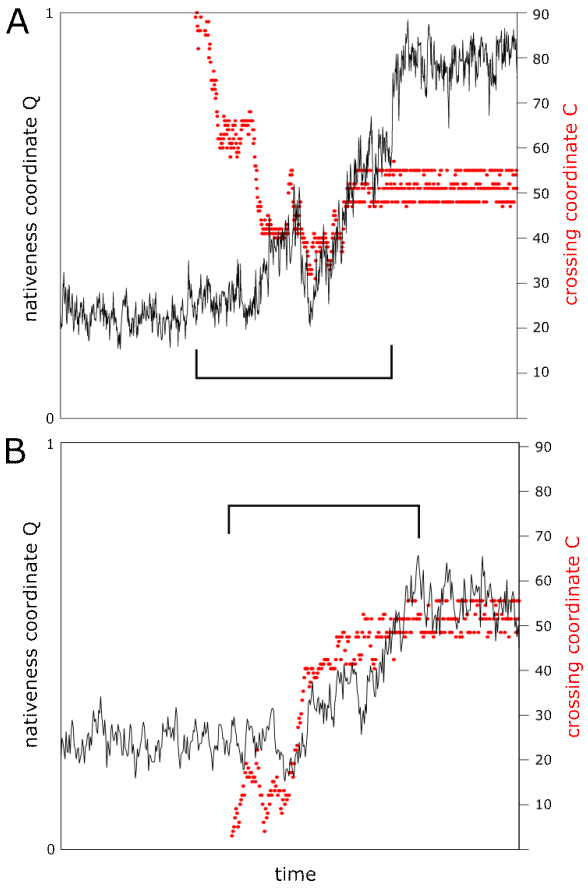

Figure 3: Slipknotting and plugging transition paths. Q (black line) and C (red dots) are shown for two folding transition paths. Bracket shows the portion of the trajectory counting as a transition path. Where there are no red dots plotted, C is undefined, i.e. there is no crossing. Note that in both cases the crossing initially appears at low Q, within the values of Q belonging to the unfolded basin. A) Folding via slipknot. Crossing starts at high C. B) Folding via plug. Crossing starts at low C.

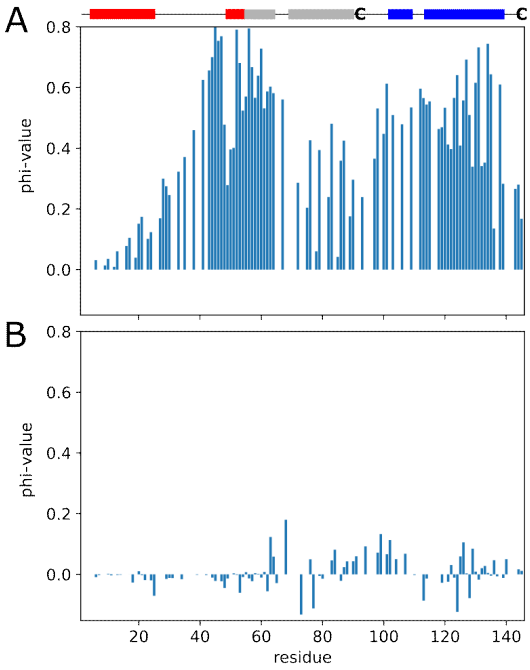

Figure 4: Phi-values calculated from folding simulations. Phi-values are not shown for residues where $\Delta\Delta G^{\mathrm{F}\to\mathrm{U}} < 2\epsilon$. Secondary structure of leptin is shown above the plots. A) reduced leptin$^{\mathrm{red}}$, B) oxidized leptin$^{\mathrm{ox}}$. The topological constraint introduced by closing the loop in leptin$^{\mathrm{ox}}$ creates a strikingly different pattern of phi-values compared to the reduced protein leptin$^{\mathrm{red}}$. For leptin$^{\mathrm{ox}}$, the phi-values are uniformly low.