

# The folding of knotted proteins: insights from lattice simulations

Patrícia F N Faisca<sup>1</sup>, Rui D M Travasso<sup>2</sup>, Tiago Charters<sup>3</sup>, Ana Nunes<sup>1,4</sup>  
and Marek Cieplak<sup>5</sup>

<sup>1</sup> Centro de Física da Matéria Condensada, Universidade de Lisboa, Av. Prof. Gama Pinto 2, 1649-003 Lisboa, Portugal

<sup>2</sup> Centro de Física Computacional, Departamento de Física, Universidade de Coimbra, 3004-516 Coimbra, Portugal

<sup>3</sup> Departamento de Engenharia Mecânica, Área Científica de Matemática, Instituto Superior de Engenharia de Lisboa, Rua Conselheiro Emídio Navarro 1, 1949-014 Lisboa, Portugal

<sup>4</sup> Departamento de Física, Universidade de Lisboa, Av. Prof. Gama Pinto 2, 1649-003 Lisboa, Portugal

<sup>5</sup> Institute of Physics, Polish Academy of Sciences, Aleja Lotnikow, 02-668 Warsaw, Poland

E-mail: [patnev@cii.fc.ul.pt](mailto:patnev@cii.fc.ul.pt) and [mc@ifpan.edu.pl](mailto:mc@ifpan.edu.pl)

Received 17 September 2009

Accepted for publication 4 January 2010

Published 3 February 2010

Online at [stacks.iop.org/PhysBio/7/016009](http://stacks.iop.org/PhysBio/7/016009)

## Abstract

We carry out systematic Monte Carlo simulations of G $\bar{0}$  lattice proteins to investigate and compare the folding processes of two model proteins whose native structures differ from each other due to the presence of a trefoil knot located near the terminus of one of the protein chains. We show that the folding time of the knotted fold is larger than that of the unknotted protein and that this difference in folding time is particularly striking in the temperature region below the optimal folding temperature. Both proteins display similar folding transition temperatures, which is indicative of similar thermal stabilities. By using the folding probability reaction coordinate as an estimator of folding progression we have found out that the formation of the knot is mainly a late folding event in our shallow knot system.

## 1. Introduction

Proteins are linear polymeric chains of amino acids folded into specific 3D structures. While the linearity of polypeptide chains precludes them from forming knots (i.e. closed loops) in a formal topological sense, their spatial structure has been shown to form knotted conformations by two different methods developed for this purpose. One of these methodologies relies on the fact that both the N- and C-termini of open proteins are typically accessible from the surface and can be connected unambiguously to form a closed loop which can then be analysed using standard knot invariants [1]. Another method is based on the Koniaris–Muthukumar–Taylor (KMT) algorithm, a chain smoothing and chain reduction procedure developed independently by Koniaris and Muthukumar [2] and by Taylor [3] to try to deal with a general conformation.

When it was originally applied to a selection of 3440 Protein Data Bank (PDB) entries, the KMT algorithm revealed a figure-eight, or  $4_1$  knot, embedded deep inside the native conformation of a plant protein [3]. A more recent survey,

which considered all the 32 853 PDB entries that contained proteins, identified 273 knotted conformations [4]. One such conformation was shown to contain a rather complex knot—the  $5_2$  knot—in a human protein. Despite its complexity, however, this particular  $5_2$  knot is classified as being a shallow knot because the deletion of 11 amino acids from the N-terminus is sufficient to unknot the fold. Nevertheless, most of the conformations identified as knotted in the PDB correspond to simple, deeply embedded  $3_1$  (also known as trefoil) knots. For the vast majority of the  $3_1$  knots reported so far, the number of residues that must be removed from one of the protein's termini to eliminate the knot lies between 25 and 92 [4].

An interesting variation amongst knotted conformations, which was recently identified in a few proteins, is that of the slipknot [5]. A slipknot is a structure that when examined in its complete form is unknotted in the mathematical sense, but becomes knotted by deletion of a suitable terminal segment, because, just like in a shoelace, it is the arrangement of the terminal segment that unties the knot as the chain folds back upon itself close to one of its ends. Altogether these

recent discoveries have triggered a renewed interest in knotted proteins [6], and studies oriented at revealing how they might fold have just begun [7].

The timing and mechanisms of knot formation in the folding process have been investigated in different experimental and simulational studies on protein YibK, a 160 residue-long chain that contains a deep trefoil knot in its native fold. A recent study based on the use of  $\phi$ -value analysis probed the knotted region of the YibK protein and showed that the native structure in this area develops in a slow manner and very late in folding [8]. A previous investigation by the same laboratory reported evidence that the threading of the polypeptide chain into a loosely knotted conformation occurs early in folding, in a denatured-like state [9]. Thus, contrary to earlier views [10], these findings show that the threading of the polypeptide chain is not the rate-limiting step, at least in the folding reaction of YibK. Other experimental investigations on protein YibK revealed a complex mechanism where two different folding intermediates, formed in parallel pathways, lead to a third intermediate that is en-route to the native structure [11]. Recent molecular simulation studies for the same protein reported as well the existence of two parallel folding pathways, distinguished by the early or late formation of the knot [12]. Interestingly, this investigation points out that chain threading in the folding of YibK seems to involve the formation of a slipknot conformation, created when the C-terminal part of the protein forms a hairpin-like shape, which is then threaded through a loop [12]. According to a very recent study such slipknot conformations play indeed a pivotal role in the folding of knotted proteins by reducing topological bottlenecks [13]. Also noteworthy is the observation by Wallin *et al* [12] that specific non-native interactions are necessary to observe folding of YibK on a biological timescale. Subsequent molecular simulation studies have shown that a native-centric potential can also successfully fold a knotted structure although in a strikingly less efficient manner [13, 14]. Indeed, while Sulkowska *et al* [13] reported folding of protein YibK under the G $\ddot{o}$  potential in 1–2% of the attempted simulations, Wallin and co-workers were able to fold the same protein with 100% efficiency by considering the explicit contribution of proper non-native interactions into the protein energetics [12].

Another largely open issue is the role of native fold knots in protein function. While knots in proteins are rare [15], they are conserved amongst structural homologues (e.g. the trefoil knot in carbonic anhydrase can be found in isozymes ranging from bacteria and algae to humans [4]), which suggests that their existence is not accidental and rather that they would have been preserved throughout evolution because they play a critical role in protein function. Given that all knotted proteins are enzymes and that knots are typically located in the catalytic domain—sometimes even in the highly flexible active site—it was suggested that their presence may provide some added stability to the structure, which, in turn, would enhance catalysis [12]. A very recent simulation effort by Sulkowska *et al* [14] showed that the thermal stability of a trefoil knotted protein is indeed larger than that of an unknotted homologue with a very similar structure (see also [16]).

In this paper we carry further the idea of comparing knotted and unknotted proteins and address these questions by studying the folding processes of two model proteins whose native structures differ from each other due to the presence or absence of a knot but have an otherwise similar native structure. The control system S is ‘simple’, i.e. it is unknotted, and the knotted system K has a trefoil knot. In contrast with other computational studies of proteins with knots where Molecular Dynamics simulations were used to explore their properties and folding dynamics [12–14, 16, 17], here we use systematic Monte Carlo simulations of G $\ddot{o}$  lattice proteins. In this framework, the identification of distinct conformations poses no problem and, more importantly, it is possible to generate large statistical samples of the whole folding process of both model systems. These two data sets are used to measure the effects of knots on stability and folding kinetics and to explore other properties of knotted proteins, adding to the relatively few results of simulations available so far. Our results are also complementary to these, in the sense that in contrast with YibK, the model system K has a shallow knot. Interestingly, we found out that a native-centric potential can fold efficiently the knotted system K considered here, and that the knot forms during the late stage of folding. Moreover, we also found similar thermal stabilities for both knotted and unknotted systems, which suggests that the presence of a knot *per se* is not enough to guarantee an enhancement of protein thermal stability. However, it may be also possible that the stability comparisons are not universal.

This paper is organized as follows. In section 2 we describe the models and computational methodologies used in the simulations. In section 3 we present and compare the results on the folding time, the folding transition temperature, the thermodynamic cooperativity, the folding scenarios, the structure of the transition states obtained for the knotted system K and for the control system S, as well as results on the timing and mechanism of knot formation for K. In section 4 we discuss the results and draw some concluding remarks.

## 2. Models and methods

### 2.1. The G $\ddot{o}$ model and simulation details

We consider a simple three-dimensional lattice model of a protein molecule with chain length  $N$ . In such a minimalist model amino acids, represented by beads of uniform size, occupy the lattice vertices and the peptide bond that covalently connects amino acids along the polypeptide chain is represented by sticks with uniform (unit) length corresponding to the lattice spacing.

To mimic protein energetics we use the G $\ddot{o}$  model [18]. In the G $\ddot{o}$  model the energy of a conformation, defined by the set of bead coordinates  $\{\vec{r}_i\}$ , is given by the contact Hamiltonian

$$H(\{\vec{r}_i\}) = \sum_{i>j}^N \epsilon \Delta(\vec{r}_i - \vec{r}_j), \quad (1)$$

where  $\epsilon$  is the (uniform) interaction energy parameter and the contact function  $\Delta(\vec{r}_i - \vec{r}_j)$  is unity only if beads  $i$  and  $j$  form a non-covalent native contact, i.e. a contact between a pair

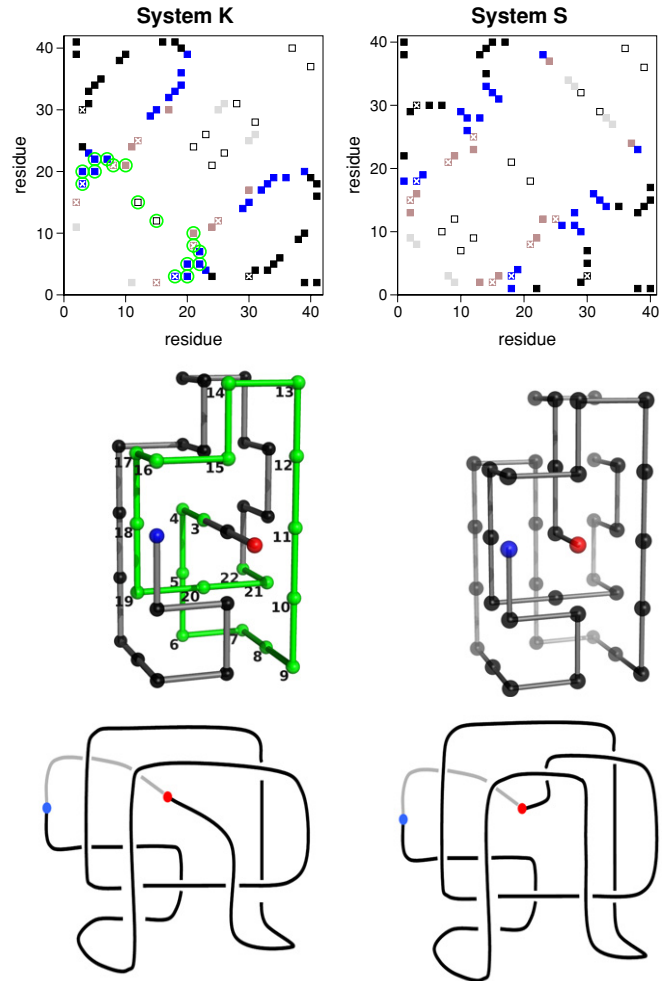
of beads that is present in the native structure, and is zero otherwise.

In order to mimic the protein's relaxation towards the native state we use the Metropolis Monte Carlo (MC) algorithm [19] together with a local move set that includes corner-flips and end-moves (i.e. displacements of one single bead) and the crankshaft move (which involves the displacement of two beads at the same time) [20–22]. At each MC step, the probability of applying the Metropolis criteria to a particular chain displacement is  $0.2/(N + 6)$  if the displacement involves moving only one bead or  $0.8/(2N - 3)$  if it involves the simultaneous movement of two beads. Attempted random moves which are discarded by excluded volume constraints or fail the Metropolis rule are counted as MC steps. A MC simulation starts from a randomly generated unfolded conformation and the folding dynamics is monitored by following the evolution of the fraction of the established native contacts  $Q$ . The number of MC steps required to fold to the native state (i.e. to achieve  $Q = 1.0$ ) is the first passage time (FPT) and the folding time is computed as the median FPT of 500 simulations. Except otherwise stated, folding is studied at the so-called optimal folding temperature  $T_{\text{opt}}$ , the temperature that minimizes the folding time [23–26]. Throughout the paper the temperature ( $T$ ) is measured in units of  $\epsilon/k_B$ , where  $k_B$  is the Boltzmann constant.

## 2.2. Model systems

Two model systems are considered in this study. Their contact maps and three-dimensional structures are shown in figure 1. Both models are geometrically complex as denoted by the high values of contact order (CO) [27]. The backbone of one of these models, termed system K (figure 1, left column), is designed in the form of a trefoil knot; the other, despite forming a geometrically intricate fold, does not contain a knot; this is termed system S (figure 1, right column). System S was constructed from system K by suitably manipulating the arrangement of the backbone within the central core of the knot's structure. As a result there is a very high overlap of 90% between the two structures (only four backbone segments do not coincide) when their backbones are optimally superimposed, indicating that the two native folds are very similar to each other. The chain length of the knotted fold ( $N = 41$ ) is one unit larger than that of the unknotted one. This difference in size allows the extension of one of system's K chain termini above the cuboids' surface so as to guarantee that both termini can be connected unambiguously to form a closed loop (i.e. to form a topological knot).

The two native folds are represented schematically in figure 1 (bottom row), where the conformations have been smoothed and the conventions of planar knot diagrams were used to choose a projection and characterize the crossings. The portion of the line depicted in grey is the connection between the two proteins' termini that is added to the chains in order to close the loop. It is easy to check using a few Reidemeister moves [28] on these planar diagrams that system K with this loop closure is indeed the trefoil knot, while system S can be transformed into a circle.



**Figure 1.** Contact map (top), three-dimensional representation (middle), and planar diagrammatic representation (bottom) of the two model systems considered in this study. Each square in the contact map represents a native contact. There are 40 native contacts in both model proteins. The knotted (left column) and unknotted system (right column) have five native contacts in common which are marked with white crosses (namely, contacts between beads 2–15, 8–21, 12–25, 3–18 and 3–30). The white squares in the contact maps represent local contacts. A contact between two beads  $i$  and  $j$  is deemed local if their sequence separation is smaller than 5 units of backbone distance (i.e.  $|i - j| < 5$ ). Non-local contacts between beads are represented by different colours according to their backbone separation (grey if  $5 \leq |i - j| < 10$ , brown if  $10 \leq |i - j| < 15$ , blue if  $15 \leq |i - j| < 20$  and black if  $|i - j| \leq 20$ ). The high number of non-local contacts translates into high values of the (absolute) contact order parameter (17.22 and 16 for system K and system S, respectively) [27]. The green circles identify the contacts between the beads of the minimal segment of system K's backbone (which is highlighted in green in the 3D representation) that contains the knot  $3_1$ ; these are the contacts between beads 3–18, 3–20, 5–20, 5–22, 7–22, 8–21, 10–21 and 12–15. In the planar diagrammatic representation the portion of the line depicted in grey is the connection between the two termini that is added to the chains in order to close the loop. The first bead is coloured red while the last one is shown in blue.

The deletion of either 20 beads from one of system K chain's terminus or of just three beads from the other terminus is enough to eliminate the knot, which is therefore classified as a shallow knot. There are eight native contacts established

between the beads of the minimal segment of the backbone that contains the knot (i.e. between bead 3 and bead 22); these are marked with green circles in the corresponding contact map (figure 1, top left). For the sake of simplicity we shall refer to this set of contacts as the knot's contacts.

We note that the two model systems studied here are not maximally compact cuboids. This is due to the fact that we were not able to fold a trefoil knot arranged in a maximally compact cuboid. Indeed, we have actually observed that the system gets trapped in conformations which are highly compact and from which it cannot escape due to excluded volume constraints. This might be an indication that alternative chain moves should be developed to efficiently fold systems with topological entanglements and other structural intricacies [29]. Nevertheless the backbones of both system K and system S are arranged in the form of highly compact structures, having 40 native contacts each. Furthermore, the contact maps displayed in figure 1 (top) uniquely define the corresponding three-dimensional folds representing system K and system S (figure 1, middle).

### 2.3. Knot detection

The chain smoothing and chain reduction KMT algorithm [2] is used to detect whether a given backbone conformation is knotted. Given a PDB conformation, its purpose is to produce a reduced and 'topologically equivalent' representation in which the knotted region is sufficiently far from both chain ends for the knot type to be well defined. The algorithm operates along the backbone by sequential repositioning of each amino acid as long as the backbone does not intersect neither of the two triangles set by the amino acid in its initial and final position and by the positions of its two nearest backbone neighbours. The final positions are chosen so as to smooth progressively the angles and curves of the backbone conformation, and the criterion for a move to be accepted should preserve the chain's 'topological type'. Moreover, whenever three consecutive backbone amino acids are collinear (within a certain tolerance), the amino acid in the middle is removed. When iterative application of the algorithm successfully 'unknots' the initial configuration, the final chain has length 2, providing a simple, computationally efficient method for knot detection.

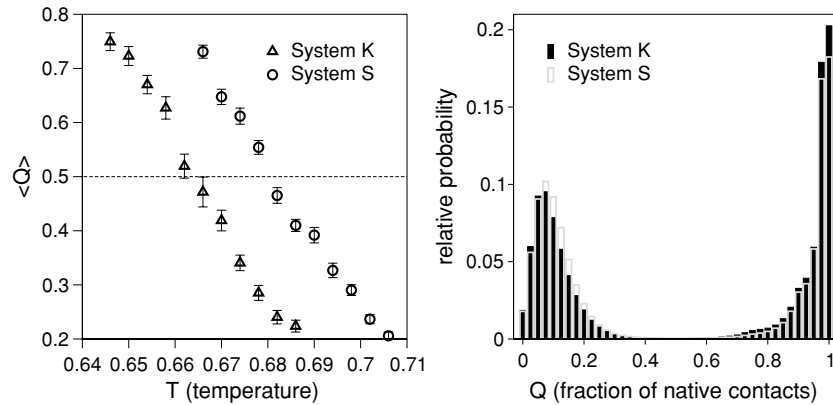
In order to ascertain whether a given conformation is knotted, a version of the KMT algorithm was used in which the topological trivial motifs of on-lattice conformation such as 'U-turns' were previously reduced. However, unless a probabilistic definition of 'knottiness' is adopted [30], determining the knot type of an open chain, whether off- or on-lattice, is of course not a well-defined mathematical problem. In particular, it is known that the KMT algorithm can classify the same conformation as knotted or unknotted depending only on the choice of the chain end that is taken to be the initial terminus [30]. Other strategies [4, 31], more sophisticated than the simple KMT approach, also suffer from some degree of ambiguity that is inherent to the ill-posedness of the problem. Nevertheless, in practical applications, the ambiguous cases have been shown to be rare [5, 7].

Bearing this in mind, in order to determine if a conformation is knotted, we ran the adapted KMT algorithm with different tolerances and in the two different orders for the sequential operations along the chain. All the conformations for which the results depended on the choice of these parameters were discarded. This amounted to at most 2% of the total number of conformations analysed. We stress that the KTM algorithm does not determine the knot type; it only indicates if the knot is present or absent. In order to actually determine the knot type, one should carry out a more detailed analysis, based on knot invariants [28]. So, in principle, other (non-native) trefoil knots, or even knots of higher complexity ( $4_1$ ,  $5_2$ ), could form during the folding process of the knotted model system (or indeed of any other target lattice model). However, this would be highly unlikely for two main reasons. First, since we are using the G $\ddot{o}$  potential to model protein energetics, the knotted protein system is strongly biased to form the knot that is present in the native structure. Secondly, according to a recent estimate by Lua and co-workers [32] the probability of finding non-trivial knots ( $3_1$ ,  $4_1$ ,  $5_2$ ) in random compact lattice loops of the size considered here is extremely small ( $<0.02$ ). Here, we have inspected one-by-one all the relevant conformations identified as knotted by the KTM algorithm to be sure that no spurious knots were formed in the folding of system K.

### 2.4. Folding probability calculation

The folding probability,  $P_{\text{fold}}(\Gamma)$ , of a conformation  $\Gamma$  is defined as the fraction of MC runs which, starting from  $\Gamma$ , fold before they unfold [33]. Because a  $P_{\text{fold}}$  calculation amounts to a Bernoulli trial, the relative error resulting from using  $M$  runs scales as  $M^{1/2}$  [34]. Thus, in order to accurately compute  $P_{\text{fold}}$ , we consider 500 MC runs equally divided into five sets of 100 folding simulations. The average value of  $P_{\text{fold}}$  is computed for each set, and the mean of all five sets, together with its standard deviation, is evaluated. Each MC run stops when either the native fold ( $Q = 1.0$ ) or an unfolded conformation is reached. A conformation is deemed unfolded when its fraction of native contacts  $Q$  is smaller than  $Q_U$ . In order to estimate the cut-off  $Q_U$ , we compute the probability of finding some fraction of native contacts  $Q$  in 200 MC folding runs [35, 36]. A high-probability peak centred around the fraction of native contacts  $Q \approx 0.125$ , appears for system K, while for system S the frequency peaks at  $Q \approx 0.1$  (data not shown). These fractions of native contacts are considerably low and therefore identify states with minimal residual structure. In this work we use these fractions of native bonds to establish  $Q_U$  for each model protein.

Two ensembles, containing 15 000 and 14 000 conformations were collected from independent MC folding runs for system K and system S, respectively. Each conformation was sampled from the run's last  $5 \times 10^6$  MCS. The folding probability of each conformation was measured as outlined above and conformations were partitioned into two ensembles with  $P_{\text{fold}} = 0.0$  and  $P_{\text{fold}} = 1.0$  and into five ensembles with  $P_{\text{fold}}$  in the intervals (0, 0.2), (0.2, 0.4), (0.4, 0.6), (0.6, 0.8) and (0.8, 1).



**Figure 2.** Estimation of the folding transition temperature  $T_f^*$  as the temperature at which the average fraction of native contacts  $\langle Q \rangle$  is 0.5 [38] (left). In order to compute  $T_f^*$  we averaged the fraction of native contacts  $Q$ , after collapse to the native state, over 50 MC simulations lasting  $\sim 10^9$  MC steps each. The error bars indicate the error of the mean value of  $\langle Q \rangle$ . At  $T_m^*$  the conformational distributions of our model systems are equally strongly bimodal indicating similar thermodynamic cooperativity [40, 41] (right).

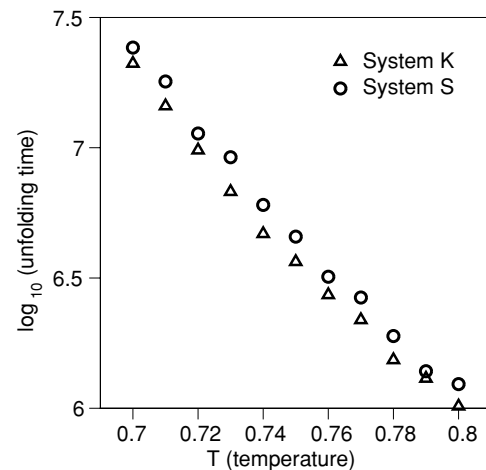
### 2.5. Structural clustering

To investigate the degree of conformational heterogeneity in an ensemble of conformations with some  $P_{\text{fold}}$  we use the structural clustering method of [36, 37]. The method groups conformations according to their degree of structural similarity by comparing every possible pair of conformations based on one (or more) measures of structural similarity. As in [36] the measure of structural similarity between two conformations here used is the parameter  $s$ , which is defined as the ratio between the number of native contacts the two conformations have in common and the number of native contacts formed in the conformation with higher  $Q$ . Two conformations are considered structurally similar (i.e. they belong to the same structural class) if  $s$  is larger than some cut-off  $S$ . In order to establish  $S$  we note that as  $S$  increases from 0 to 1, the number and size of the identified clusters changes. Since we are interested in representative clusters (i.e. with a statistically sound number of conformations) with distinct structural traits, we selected clusters with size  $\geq 5\%$  than that of the starting ensemble, and for which no conformation is similar by more than  $s = 0.9$  to a conformation in another identified cluster (i.e.  $S = 0.9$ ).

## 3. Results

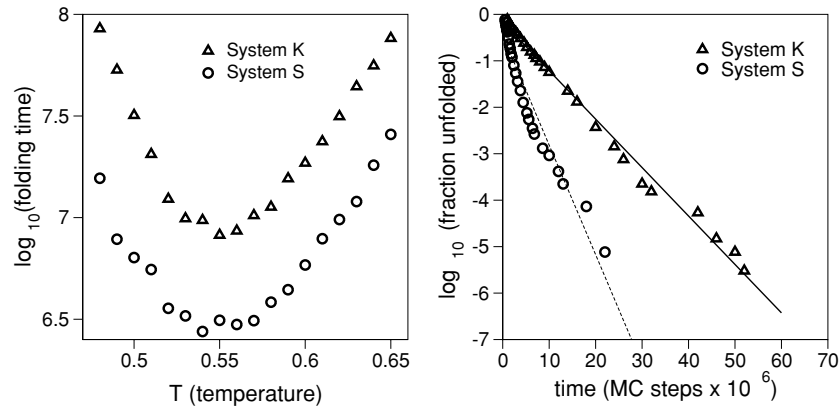
### 3.1. The folding transition temperature

In this subsection we compare the thermal stabilities of systems K and S. The folding transition temperature  $T_f$  (typically denoted as melting temperature  $T_m$  in the experimental literature) is a measure of the protein's thermal stability. By assuming a strict two-state folding transition, experimentalists usually estimate the transition temperature as the temperature at which the denatured and native states are equally populated at equilibrium. An easy way to estimate  $T_f$  in the lattice model, which is similar to that employed in experiments, is from the temperature dependence of a macroscopic quantity, e.g. the average number of native contacts  $\langle Q \rangle$  [38] (figure 2, left). In doing so, we find that the estimated folding



**Figure 3.** Dependence of the logarithm of the unfolding time on the simulation temperature. The unfolding time is computed as the median FPT necessary to achieve fraction of native contacts  $Q = 0$  of 500 MC runs.

transition temperature  $T_f^*$  is 0.662 for model system K, while model system S displays a slightly higher  $T_f^* = 0.682$ . For the lattice model, however,  $T_f$  is defined microscopically as the temperature at which the equilibrium population of the native state is half of the total population (i.e. the probability that  $Q = 1$  is 1/2 at  $T_f$ ) [39]. In other words, at  $T_f$ , the protein spends half of the time in its native conformation. Therefore, in order to determine  $T_f$  we actually count the number of times the native state (i.e.  $Q = 1$ ) appears in very long MC runs. By using the microscopic definition of  $T_f$  we find that the transition temperature of system S is  $T_f = 0.54$  (which is the same as its optimal folding temperature), while for system K it is just slightly larger, namely,  $T_f = 0.58$ . Thus, the folding transition temperatures of the two model systems show marginal and negligible differences which is indicative of similar thermal stabilities. We further confirmed this by exploring the dependence of the unfolding time on the simulation temperature. The proximity between both curves is also indicative of similar thermal stabilities (figure 3).



**Figure 4.** Dependence of the logarithm of the folding time on the simulation temperature. At the optimal folding temperature  $T_{\text{opt}}$  (i.e. the temperature that minimizes the folding time) the folding of system K is three times slower than that of system S (left). The right panel shows evidence for single-exponential relaxation at  $T_{\text{opt}}$  for both knotted and unknotted folds. The single-exponential nature of the folding kinetics is particularly stronger in system K ( $R^2 = 0.99$ ) than in system S ( $R^2 = 0.93$ ).

**Table 1.** Summary of kinetic and thermodynamic properties for model systems K and S.

System	$T_{\text{opt}}$	$\log_{10}$ (folding time) at $T_{\text{opt}}$	$T_f(T_f^*)$	$\log_{10}$ (folding time) at $T_f(T_f^*)$
K	0.55	6.91	0.58 (0.662)	7.05 (8.01)
S	0.54	6.44	0.54 (0.682)	6.44 (7.78)

Table 1 provides a summary of the kinetic and thermodynamic properties of both model systems.

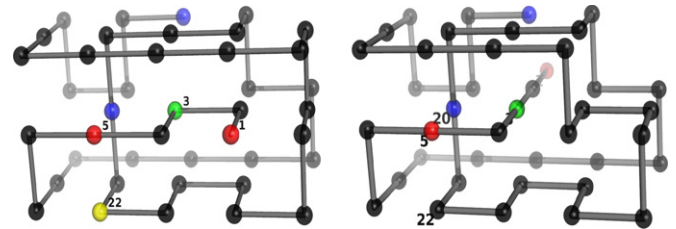
### 3.2. Thermodynamic cooperativity

In this subsection we present evidence for similar thermodynamic cooperativities for both the knotted and the unknotted fold. Since a more thermodynamic cooperative folding transition corresponds to a more pronouncedly bimodal conformational distribution [40], we have analysed the probability distribution at  $T_m^*$  for the fraction of native contacts  $Q$  in order to explore possible differences in thermodynamic cooperative behaviour between system K and system S. Results reported in figure 2 (right) show no differences between both model structures.

### 3.3. Folding times

In this subsection we show that the knotted fold displays a larger folding time at the temperature of fastest folding  $T_{\text{opt}}$ . We started by exploring the dependence of the folding time on the simulation temperature. Results reported in figure 4 (left) show that the folding of both target models is strongly temperature dependent. At  $T_{\text{opt}}$  system K folds three times slower than system S, which shows that the formation of the knot increases indeed the difficulty of finding the native conformation. Furthermore, this difference in folding performance is amplified for  $T < T_{\text{opt}}$  (e.g. at  $T = 0.48$  system S folds seven times more rapidly than system K), showing that the folding of the knotted protein is less robust against temperature change.

At  $T_{\text{opt}}$  the folding of both model systems closely approximates a single-exponential relaxation. Interestingly,

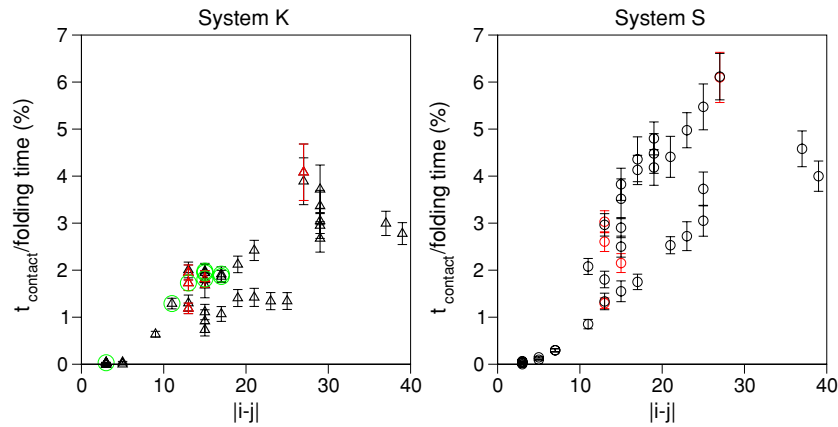


**Figure 5.** Example of a slipknotted conformation (left) found in a folding trajectory to system K (right). The contact between bead 5 (red) and bead 20 (blue) is the first to form, which occurs relatively early in folding. The contact between bead 5 and bead 22 (yellow) forms afterwards, which simultaneously establishes the axis that contains beads 19, 20 and 21. This axis stays formed until the end of the folding process. By then the contact between bead 3 (green) and bead 20 forms. Finally, a small hairpin formed by beads 1, 2, 3 and 4 enters the core of the structure that subsequently extends to lock the knot (right).

the single-exponential behaviour of folding kinetics is clearly stronger in the case of system K (figure 4, right).

### 3.4. Slipknots and knot formation

It has been suggested that to generate a knot in its structure, the protein needs to thread a part of its backbone through a loop *via* a slipknot conformation [4]. The observation that the folding of knotted proteins involves an intermediate conformation with a slipknot was recently reported in the context of off-lattice Molecular Dynamics simulations of a simple coarse-grained model of proteins YibK and YbeA [13]. In order to get insight into the mechanism that governs the formation of the knot we have explored in detail a few individual MC runs in order to look for precursors of the knotted conformation. In some cases, we could observe that the formation of the knot proceeds indeed through a slipknot conformation. An example of a slipknotted conformation is shown in figure 5: towards the end of the folding process a small hairpin formed by four terminal residues enters the core of system S, subsequently extending its end branch to knot the backbone.



**Figure 6.** Folding scenarios I: dependence of the (mean) first time formation of each native contact on the contact's range (i.e. the backbone separation between beads  $i$  and  $j$  that establish the contact). During folding, native contacts in system K form earlier than native contacts with similar backbone separation in system S. Bonds represented in red are those common to both structures, and the green circles indicate the knot's contacts, which form for first time in less than 2% of the folding time.

### 3.5. Folding scenarios

Cieplak and co-workers [42–44] have introduced the so-called ‘folding scenario’ representation to characterize the folding process in off-lattice simulations of protein folding. In a folding scenario, the average time to form a native contact for the first time,  $t_{\text{contact}}$ , is plotted against the contact length (i.e. the backbone separation between the two beads that establish the contact). Typically, in off-lattice simulations, the folding scenario shows a mostly monotonic increase of  $t_{\text{contact}}$  with the sequential distance between the two beads forming a contact.

Figure 6 reports the average time needed to form a native contact (normalized to the folding time) for system K (left) and system S (right), where the error bars indicate dispersion in the values of  $t_{\text{contact}}$ ; the average is computed as the mean time to form a contact for the first time in 500 MC runs. Contacts marked in red are those that exist in both structures, while contacts marked with green circles in the plot for system K are the knot's contacts.

Compared with that of system S, the scenario displayed by the knotted system shows a weaker monotonic dependence of  $t_{\text{contact}}$  on the contact length. For example, about 20% of system's K native contacts (with sequence separation ranging between 10 and 25 units of backbone distance) form during the first 1% of the folding time. Also, the vast majority of system's K contacts takes 2% of the folding time to form for the first time. It is interesting to note that in system K the contact that takes more time to form (the non-local contact between beads 3 and 30) does it 20% earlier than the equivalent contact in system S. This observation translates the general trend observed in the folding of system K which is that all of its native contacts form for the first time earlier than in system S. We interpret this as being a consequence of the different time scales of the two folding processes we are comparing, while the exploration of conformational space in early folding and the first occurrence of native contacts takes place on the same time scale for the two systems.

We have also considered an alternative description of the folding process that is based on measuring the ‘fixation time’,

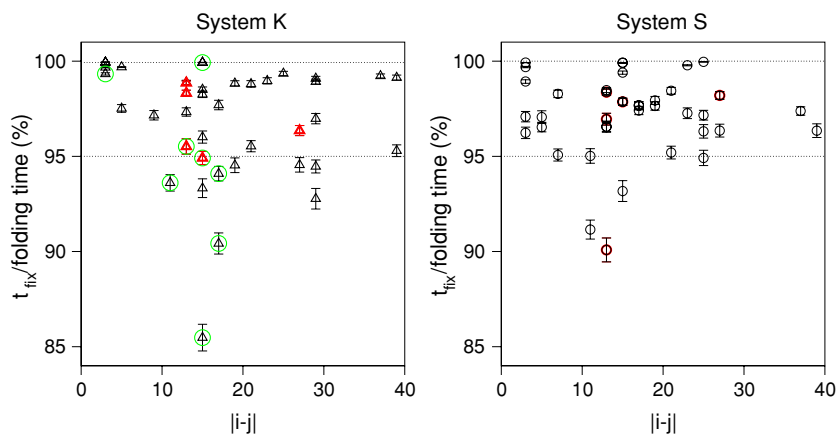
$t_{\text{fix}}$ , of a native contact, a measure similar to that introduced in [45] for the formation of secondary structural elements. By fixation time we mean the (average) time (in number of MC steps) corresponding to the final formation (or fixation) of a native contact: from this ‘instant’ onwards the contact does not break until the protein reaches its (final) native conformation. Results reported in figure 7 show the ‘fixation scenarios’ for system K (left) and for system S (right), where  $t_{\text{fix}}$  is normalized to the run's FPT.

A comparison between both scenarios shows that the number of native contacts that get fixed first (i.e. in less than 95% of the total folding time) is three times larger for system K. Interestingly, 50% of these native contacts are knot's contacts. However, the fixation scenarios are overall much more similar than the scenarios displayed for first time formation of native contacts.

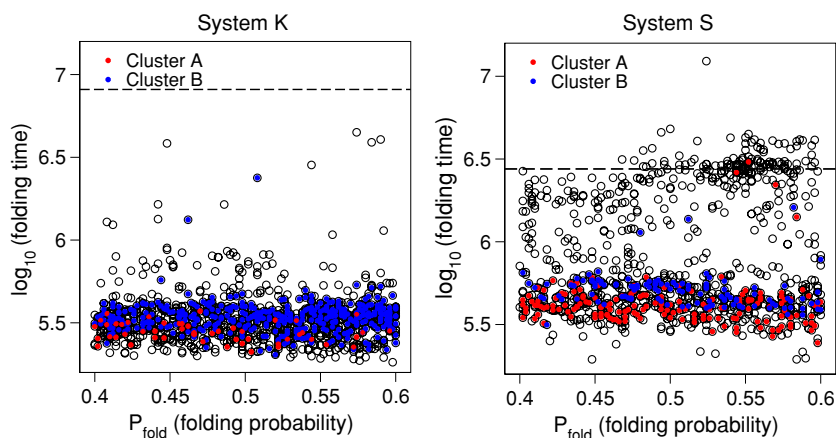
### 3.6. Transition state ensemble

Here we explore the transition state (TS) ensembles of systems S and K by using the folding probability analysis method. It has been reported in studies of lattice models that  $P_{\text{fold}}(\Gamma)$  may serve as a reaction coordinate for the folding process when it corresponds to a two-state transition [33]. In particular, any optimal (i.e. minimal free energy) trajectory connecting unfolded states with the native basin must cross conformations for which  $P_{\text{fold}}(\Gamma)$  is equal to 1/2. It is thus interesting to consider the properties of all conformations that are close to this threshold, specifically those satisfying condition  $0.4 < P_{\text{fold}}(\Gamma) < 0.6$ . It should be noted, however, that not all of these states would actually lie on paths which are truly optimal. The set of fast folding conformations with  $0.4 < P_{\text{fold}}(\Gamma) < 0.6$  will be referred to as the transition state ensemble (TSE).

The conformations with  $0.4 < P_{\text{fold}}(\Gamma) < 0.6$  were selected from the original ensemble of  $\sim 15\,000$  conformations; there are 1008 and 877 such conformations for model systems K and S, respectively. The results reported in figure 8 show the time necessary to reach the naive structure



**Figure 7.** Folding scenarios II: dependence of the (mean) fixation time of each native contact on contact's range. The fixation time of a contact means that from that instant onwards the contact forms and does not break again until the protein gets into the native state. In the folding of system K there are more contacts that get fixed in the first 95% of the folding time. Interestingly, the contacts that fix in first and second place (between beads 3–18 and 3–20, respectively) are the knot's contacts.

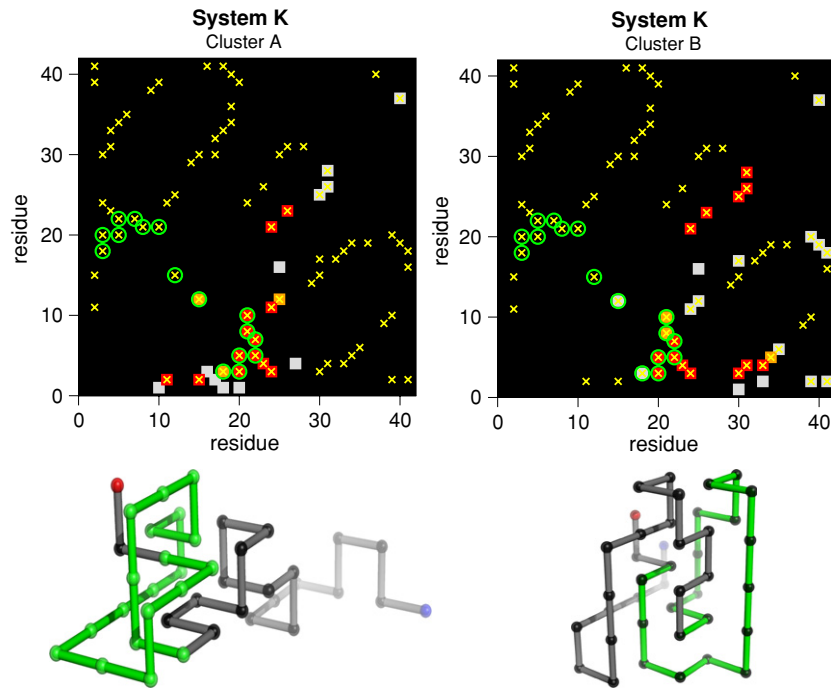


**Figure 8.** Folding time measured as the median FPT of 500 MC runs starting from each conformation within the ensemble of 1008 conformations with  $0.4 < P_{\text{fold}} < 0.6$  for system K (left). For system S (right), an ensemble of 877 conformations within the same  $P_{\text{fold}}$  interval was considered. The dotted line indicates the folding time starting from a random coil-type conformation. The structural clustering method identifies two structural classes, termed cluster A and cluster B. Conformations pertaining to these clusters are highlighted. For model system K the number of conformations in cluster A is 51 and in cluster B is 398. For model system S there are 168 conformations in cluster A and cluster B is formed by 116 conformations.

starting from each conformation with  $0.4 < P_{\text{fold}}(\Gamma) < 0.6$  for system K (left) and system S (right). In the case of system K, the vast majority of these conformations are able to reach the native structure very rapidly. Indeed, they do so one order of magnitude faster than a randomly generated unfolded conformation. In contrast, for system S there is a significantly large number of conformations with  $0.4 < P_{\text{fold}} < 0.6$  that are actually trapped states. Indeed, these conformations need more (or at least as much) time to find the native structure than a random coil type conformation (figure 8, right) and display non-native contacts formed with considerably high probability ( $p > 0.75$ ). This should not be surprising because at the optimal folding temperature,  $T_{\text{opt}}$ , the folding kinetics of system S is markedly less single-exponential than that of system K (figure 4). Fortunately, the application of the structural clustering method separates the kinetic traps from TSE conformations.

The assessment of the degree of conformational heterogeneity in the TSE pinpointed two relevant structural

classes for the knotted model system. As shown in figure 8 (left)—and with the exception of two outliers—the conformations within both classes are typically very fast folders. The fastest folding cluster, which is deemed cluster A, folds on average in 3.4% of the folding time, while the other cluster, termed cluster B folds slightly slower in 4.3% of the folding time. The probability maps reported in figure 9 (top) show the probability of occurrence  $p$  of each contact (native and non-native) in the ensemble of conformations forming cluster A (left) and cluster B (right) (contacts formed with small probability  $p < 0.2$  are discarded). A comparison between the two probability maps shows that both clusters have a very similar number of native contacts formed with high probability ( $p > 0.8$ ). However, cluster A ( $\langle Q \rangle = 0.41$ ) is less consolidated (i.e. it has less native contacts and more non-native contacts formed with probability  $p > 0.2$ ) than cluster B ( $\langle Q \rangle = 0.51$ ). The three-dimensional structure of the conformation that is the more nativelylike (i.e. which has more native contacts formed) illustrates the difference



**Figure 9.** Probability to find a contact (native and non-native) in the ensemble of 51 conformations forming cluster A (left) and in the ensemble of 395 conformations (with  $\log_{10}t < 6$ ) forming cluster B (right). The probability map only shows the contacts forming with probability  $p > 0.2$ . Among the latter we distinguish those that form with probability  $0.5 \leq p < 0.8$  (orange squares) from those that form with higher probability  $p \geq 0.8$  (red squares). Grey squares represent contacts forming with intermediate probability  $0.2 < p < 0.5$ . The yellow crosses represent the native contacts and the green circles indicate the knot's contacts. To highlight the differences between both clusters, the three-dimensional structures of the most nativelike conformation within each cluster ( $Q = 0.53$  for cluster A and  $Q = 0.7$  for cluster B) are also shown (bottom). The conformation representative of cluster A is knotted.

between the identified structural classes (figure 9, bottom). Furthermore, an important but subtle difference between both probability maps concerns the formation of the so-called knot's contacts. All of them are systematically formed with very high probability ( $p \sim 1$ ) in cluster A, and two of them (namely, contacts 8–21 and 10–21) are actually twice more probable in cluster A than in cluster B (figure 9, top). Therefore, one is led to conjecture that the backbone of system K is knotted in cluster A but not necessarily so in cluster B. Indeed, by applying the KMT algorithm to each conformation, we confirmed that the knot is present with a remarkably high probability of 74% in cluster A, but only in 1.3% of the conformations pertaining to cluster B.

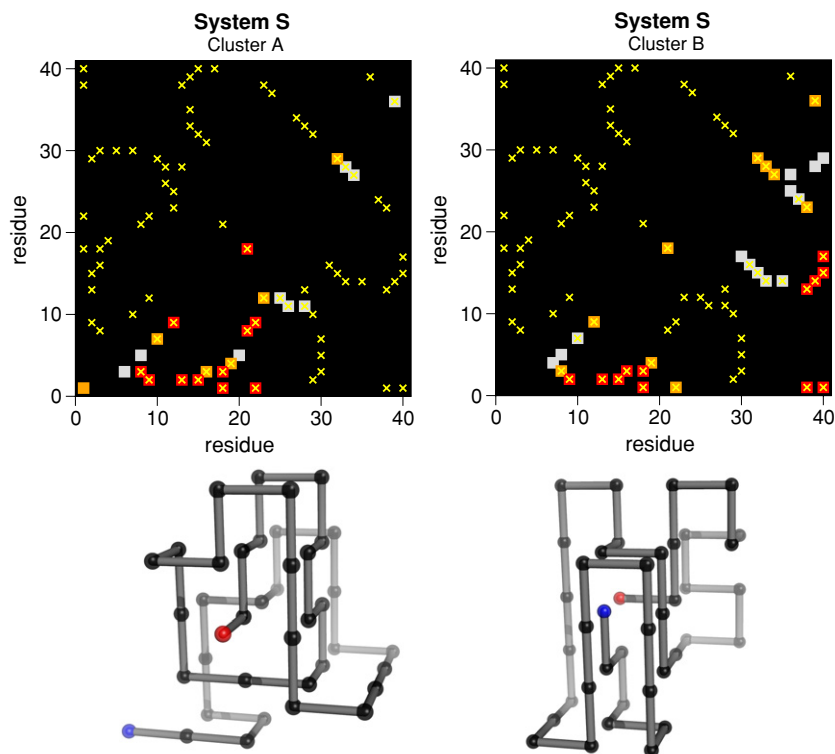
A similar analysis carried out for system S pinpointed as well the existence of two structural classes. There is cluster A (with fraction of native contacts  $\langle Q \rangle = 0.46$ ), whose conformations fold on average in 15% of the total folding time, and cluster B (with fraction of native contacts  $\langle Q \rangle = 0.53$ ), which folds slightly slower in 17% of the folding time (figure 10). Interestingly, the conformations that represent system's K TSE fold up to five times more rapidly than those representing the TSE of system S.

### 3.7. Timing of knot formation

Here, we investigate when the knot in system K forms during the folding process and show that the formation of the knot is mainly a late folding event. In doing so, we get insight

into the following question: Is the formation of the knot the rate-limiting step in folding? Or are the folding and knotting mechanisms uncoupled from each other? In order to address this question we have evaluated the fraction of knotted conformations in each considered  $P_{\text{fold}}$  ensemble. To detect the presence (or absence) of the knot in some conformation we have applied the KMT algorithm as outlined before. The probability of the knot being present in an ensemble with  $M$  conformations follows the binomial distribution whose error scales with  $\sqrt{M}$ . Since the number of conformations in each  $P_{\text{fold}}$  ensemble is typically large (e.g. it ranges between 1008 conformations in the ensemble with  $0.4 < P_{\text{fold}} < 0.6$  to 3493 conformations in the ensemble with  $0.0 < P_{\text{fold}} < 0.2$ ) the relative error is typically very small. We note that the conformations with  $\log_{10}t \geq 6$  were discarded from the ensemble of conformations with  $0.4 < P_{\text{fold}} < 0.6$  as some of them are clearly trapped states.

Results reported in figure 11 show that the formation of the knot is mainly a late folding event. Indeed, the knot forms with very small probability ( $\sim 0.001$ ) in early folding ( $P_{\text{fold}} < 0.3$ ) but this probability increases sharply when folding is close to completion (e.g. it shows a threefold increase for  $P_{\text{fold}} = 0.7$  and a fivefold increase for  $P_{\text{fold}} = 0.9$ ). The probability of the knot being formed very early in folding is practically null. This observation is consistent with the scenario where the threading of the chain and the formation of the native structure in the knotted region both occur late in folding. Thus, the formation of the shallow knot follows a different process from that of the



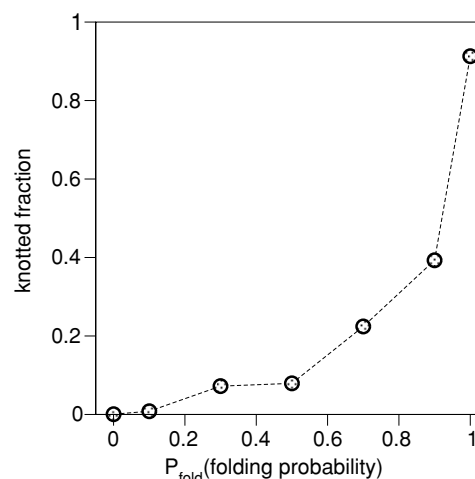
**Figure 10.** Probability to find a contact (native and non-native) in the ensemble of 164 conformations forming cluster A (left) and in the ensemble of 118 conformations forming cluster B (to construct the probability maps conformations with  $\log_{10}t \geq 6$  were discarded). The probability map only shows the contacts forming with probability  $p > 0.2$ . Among the latter we distinguish those that form with probability  $0.5 \leq p < 0.8$  (orange squares) from those that form with higher probability  $p \geq 0.8$  (red squares). Grey squares represent contacts forming with intermediate probability  $0.2 < p < 0.5$ . The three-dimensional structure of the most nativelike conformation within each cluster ( $Q = 0.6$  and  $Q = 0.7$  for clusters A and B, respectively) is also shown (bottom).

knot in protein YibK. In the latter, although the native structure in the knotted region of the protein remains undeveloped until very late in the folding reaction [8], the threading of the chain occurs early on in the folding process [11].

#### 4. Discussion and conclusions

We have performed extensive lattice Monte Carlo simulations to explore the folding process of a model system whose backbone is arranged in the form of a trefoil knot. This knot is rather shallow since it is located near one of the ends of the polypeptide chain. We have further compared the folding process of the knotted structure with that of a decoy structure constructed from the knotted one which—despite sharing an overall similar topology—does not contain the knot (i.e. it is unknotted).

Our results show that, under conditions of optimal temperature, folding to the knotted protein is considerably slower than folding to the unknotted one. However, and perhaps surprisingly, both knotted and unknotted folds have identical thermal stabilities as denoted by very similar transition temperatures. This finding is in contrast with recent reports from computational studies which indicate that a knotted topology enhances thermal stability [14]. However, the observed discrepancy might be due to the location of the lattice knot which is very near to the protein termini, while previous investigations were concerned with knots that



**Figure 11.** Probability to find the knot along the folding process. Folding progression is monitored by means of the folding probability reaction coordinate,  $P_{\text{fold}}$ . The formation of the knot occurs essentially late during folding as denoted by the sharp increase in the fraction of knotted conformations for  $P_{\text{fold}} > 0.6$  thus suggesting that knotting and folding are two decoupled processes. The relative error in evaluating the knot's probability of being present in each  $P_{\text{fold}}$  ensemble ranges between 0.0004 and 0.01; due to its small size the error is not represented in the plot.

are deeply embedded into the native structure. We therefore conclude that the presence of a knot *per se* is not sufficient

to guarantee an enhancement of the protein's thermal stability and might explain why the vast majority of trefoil knots found in real world proteins correspond to deeply embedded knots, instead of shallow knots.

The study of the transition state ensemble revealed that the times needed to achieve the native state starting from TSE conformations are shorter for the model system that contains the knot. Adding up the fact that its folding time (starting from random coil type conformers) is larger than the folding time of the system without the knot, this observation implies that the knotted system spends considerably more time (than the unknotted one) in the exploration of pre-transition state conformational space. This is possibly due to the existence of a larger amount of topological frustration in the form of backtracking. The concept of backtracking was introduced by Onuchic and co-workers [46] to describe a situation where the protein unfolds and refolds again due to the incorrect order along time of formation of the native contacts (the idea is that topology determines the time ordering according to which native contacts should form during folding, and as the topological complexity increases such an ordering process should get more strict).

We note that with regard to the presence of non-native contacts, only marginal differences were found between both TSEs, and potentially allowed non-native contacts appear to be scarce in both. Recent off-lattice simulations [12] have shown that non-native interactions play an important role in driving efficient folding to a native structure having a deeply embedded knot. Here, we have focused on a lattice system displaying a shallow knot that we could fold efficiently by using a native-centric potential. Thus, our results suggest that non-native interactions are not essential to fold proteins with shallow knots. However, it is expected that adding well-chosen non-native contacts may accelerate the folding process and perhaps even allow for folding into knotted maximally compact structures.

By evaluating the probability of knot formation in ensembles of conformations with increasing folding probabilities, we have found that the knotting of the backbone is mainly a late folding event. This shows that in the case of shallow knots, the threading of the polypeptide chain into a knot occurs mostly late in folding. This late process may not be possible in the case of deeply embedded knots for which threading movements leading to knot formation involve a substantially large part of the polypeptide chain. In this case, and as the experiments with YibK suggest [11], threading should occur in a denatured-like conformation, resulting in the formation of a loosely knotted structure that subsequently slowly consolidates its knotted region towards the late stages of folding [8].

## Acknowledgments

PFNF and RDMT thank Fundação para a Ciência e Tecnologia (FCT) for financial support through the Ciência 2007 program. PFNF acknowledges support from the FCT through grant POCI/QUI/58482/2004. MC appreciates many discussions

with J Sulkowska. MC was supported by grant N N202 0852 33 from the Ministry of Science and Higher Education in Poland and by the European Union within European Regional Development Fund, through grant Innovative Economy (POIG.01.01.02-00-008/08).

## References

- [1] Mansfield M 1997 Fit to be tied *Nat. Struct. Biol.* **4** 166–7
- [2] Koniaris K and Muthukumar M 1991 Knottedness in ring polymers *Phys. Rev. Lett.* **66** 2211–4
- [3] Taylor W 2000 A deeply knotted structure and how it might fold *Nature* **406** 916–9
- [4] Virnau P, Mirny M A and Kardar M A 2006 Intricate knots in proteins: function and evolution *PLoS Comput. Biol.* **2** 1074–9
- [5] King N P, Yeates E O and Yeates T O 2007 Identification of rare slipknots in proteins and their implications for stability and folding *J. Mol. Biol.* **373** 153–66
- [6] Mansfield M 1994 Are there knots in proteins *Nat. Struct. Biol.* **1** 213–4
- [7] Yeats T O, Norcross T S and King N P 2007 Knotted and topologically complex proteins as models for studying folding and stability *Curr. Opin. Chem. Biol.* **11** 595–603
- [8] Mallam A L, Morris E L and Jackson S E 2008 Exploring knotting mechanisms in protein folding *Proc. Natl Acad. Sci. USA* **105** 18740–5
- [9] Mallam A L, Onuoha S C and Jackson S E 2008 Knotted fusion proteins reveal unexpected possibilities in protein folding *Mol. Cell* **30** 642–8
- [10] Taylor W 2007 Geometry knots and fold complexity: some new twists *Comput. Biol. Chem.* **31** 151–62
- [11] Mallam A L and Jackson S E 2007 Probing nature's knots: the folding pathway of a knotted homodimeric protein *J. Mol. Biol.* **359** 1420–36
- [12] Wallin S, Zeldovich K B and Shakhnovich E I 2007 Folding mechanics of a knotted protein *J. Mol. Biol.* **368** 884–93
- [13] Sulkowska J I, Sulkowski P and Onuchic J 2008 Dodging the crisis of folding proteins with knots *Proc. Natl Acad. Sci. USA* **9** 3119–24
- [14] Sulkowska J, Sulkowski P and Cieplak M 2008 Stabilizing effect of knots on proteins *Proc. Natl Acad. Sci. USA* **105** 19714–9
- [15] Lua R C and Grosberg A Y 2006 Statistics of knots, geometry of conformations and evolution of proteins *PLoS Comput. Biol.* **2** 350–7
- [16] Sulkowska J I, Sulkowski P, Szymczak P and Cieplak M 2008 Tightening of knots in proteins *Phys. Rev. Lett.* **100** 058106
- [17] Dzubiella J 2009 Sequence-specific size, structure, and stability of tight geometry knots *Biophys. J.* **96** 19
- [18] Gō N and Taketomi H 1978 Respective roles of short- and long-range interactions in protein folding *Proc. Natl Acad. Sci.* **75** 559–3
- [19] Metropolis N, Rosenbluth A W, Rosenbluth M N, Teller A H and Teller E 1953 Equation of state calculations by fast computing machines *J. Chem. Phys.* **21** 1087–92
- [20] Landau D P and Binder K 2000 *A Guide to Monte Carlo Simulations in Statistical Physics* (Cambridge: Cambridge University Press)
- [21] Chan H S and Dill K A 1993 Energy landscapes and the collapse dynamics of homopolymers *J. Chem. Phys.* **99** 2116–27
- [22] Chan H S and Dill K A 1994 Transition states and folding dynamics of proteins and heteropolymers *J. Chem. Phys.* **100** 9238–57
- [23] Oliveberg M, Tan Y and Fersht A R 1995 Negative activation enthalpies in the kinetics of protein folding *Proc. Natl Acad. Sci. USA* **92** 8926–9

- [24] Gutin A, Sali A, Abkevich V, Karplus M and Shakhnovich E I 1998 Temperature dependence of the folding rate in a simple protein model: search for a 'glass' transition *J. Chem. Phys.* **108** 6466–83
- [25] Cieplak M, Hoang T X and Li M S 1999 Scaling of folding properties in simple models of proteins *Phys. Rev. Lett.* **83** 1684–7
- [26] Faisca P F N and Ball R C 2002 Thermodynamic control and dynamical regimes in protein folding *J. Chem. Phys.* **116** 7231–8
- [27] Plaxco K W, Simmons K T, Ruczinski I and Baker D 2000 Topology, stability, sequence and length: defining the determinants of two-state protein folding kinetics *Biochemistry* **39** 11177–83
- [28] Adams C 2004 *The Knot Book: An Elementary Introduction to the Mathematical Theory of Knot* (Providence, RI: American Mathematical Society)
- [29] Liu Z and Chan H S 2008 Efficient chain moves for Monte Carlo simulations of a wormlike DNA model: excluded volume, supercoils, site juxtapositions, knots, and comparisons with random-flight and lattice models *J. Chem. Phys.* **128** 145104
- [30] Millett K, Dobay A and Stasiak A 2005 Linear random knots and their scaling behavior *Macromolecules* **38** 601–6
- [31] Kolesov G, Virnau P, Kardar M A and Mirny L A 2007 Protein knot server: detection of knots in protein structures *Nucleic Acids Res.* **35** W425–8
- [32] Lua R, Borovinskiy A L and Grosberg A Yu 2004 Fractal and statistical properties of large compact polymers: a computational study *Polymer* **45** 717–31
- [33] Du R, Pande V S, Grosberg A Y, Tanaka T and Shakhnovich E S 1998 On the transition coordinate for protein folding *J. Chem. Phys.* **108** 334–50
- [34] Hubner I A, Shimada J and Shakhnovich E I 2004 Commitment and nucleation in the protein G transition state *J. Mol. Biol.* **336** 745–61
- [35] Travasso R D M, Faisca P F N and Gama M M T 2007 Nucleation phenomena in protein folding: the modulating role of protein sequence *J. Phys.: Condens. Mater.* **19** 285212
- [36] Faisca P F N, Travasso R D M, Ball R C and Shakhnovich E I 2008 Identifying critical residues in protein folding: insights from  $\phi$ -value and  $P_{\text{fold}}$  analysis *J. Chem. Phys.* **129** 095108
- [37] Hubner I A, Deeds E J and Shakhnovich E I 2006 Understanding ensemble protein folding at atomic detail *Proc. Natl Acad. Sci. USA* **103** 17747–52
- [38] Abkevich V I, Gutin A M and Shakhnovich E I 1995 Impact of local and non-local interactions on thermodynamics and kinetics of protein folding *J. Mol. Biol.* **252** 460–71
- [39] Shakhnovich E and Gutin A 1993 Engineering of stable and fast-folding sequences of model proteins *Proc. Natl Acad. Sci. USA* **90** 7195–9
- [40] Makhatadze G I and Privalov P L 1995 Energetics of protein structure *Adv. Protein Chem.* **47** 307–9
- [41] Faisca P F N and Plaxco K W 2006 Cooperativity and the origins of rapid, single-exponential kinetics *Prot. Sci.* **15** 1608–18
- [42] Cieplak M and Hoang T X 2003 Universality classes in folding times of proteins *Biophys. J.* **84** 475–88
- [43] Cieplak M, Hoang T X and Robbins M O 2002 Thermal folding and mechanical unfolding pathways of protein secondary structures *Proteins: Funct., Struct. Genetics* **49** 104–13
- [44] Cieplak M and Sulkowska J I 2005 Thermal unfolding of proteins *J. Chem. Phys.* **123** 194908
- [45] Kaya H and Chan H S 2005 Explicit-chain model of native-state hydrogen exchange: implications for event ordering and cooperativity in protein folding *Proteins* **58** 31–44
- [46] Gosavi S, Chavez L L, Jennings P A and Onuchic J N 2006 Topological frustration and the folding of interleukin-1 beta *J. Mol. Biol.* **357** 986–96