PCCP

Accepted Manuscript

Volume 1 | Number 1 | Jan 2013 | Pages 1–100
PCCP
Physical Chemistry Chemical Physics
www.rsc.org/pccp
ROYAL SOCIETY OF CHEMISTRY

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/pccp

How determinant is N-terminal to C-terminal coupling for protein folding?

Heinrich Krobath[1], Antonio Rey[2*] and Patrícia FN Faísca[3*]

[1,3]Centro de Física da Matéria Condensada and Departamento de Física, Faculdade de Ciências da Universidade de Lisboa, Portugal

[2]Departamento de Química Física I, Facultad de Ciencias Químicas, Universidad Complutense, Madrid, Spain and Instituto de Biocomputación y Física de Sistemas Complejos (BIFI), Zaragoza, Spain

**Corresponding author*

P.F.N.F, E-mail: pffaisca@fc.ul.pt or A.R, E-mail: areygayo@ucm.es

Abbreviations: Termini interactions, NC; Contact order, CO; Monte Carlo, MC;weighted histogram analysis method, WHAM; transition state ensemble, TSE; circular permutant, CP; alpha-carbon, C_α

Abstract

This work investigates the role of N- to C- termini coupling in the folding transition of small, single domain proteins via extensive Monte Carlo simulations of both lattice and off-lattice models. The reported results provide compelling evidence that the existence of native interactions between the terminal regions of the polypeptide chain (i.e. termini coupling) is a major determinant of the height of the free energy barrier that separates the folded from the denatured state in a two-state folding transition, being therefore a critical modulator of protein folding rates and thermodynamic cooperativity. We further report that termini interactions are able to substantially modify the kinetic behavior dictated by the full set of native interactions. Indeed, a native structure of high contact order with "switched-off" termini-interactions actually folds faster than its circular permutant of lowest CO.

Keywords: Monte Carlo simulation, lattice model, off-lattice model, folding thermodynamics, folding cooperativity, kinetics, folding rate, termini interactions

Introduction

An important landmark in protein folding research was the observation by Plaxco and co-workers, about 15 years ago, that the folding rates of small, single domain proteins that fold with two-state kinetics are well correlated (r~0.9) with a property of the native structure named contact order (CO)[1, 2]. The CO can be viewed as a metric of native geometry in the sense that it measures the average sequence separation of contacting residue pairs in the native structure.

Shortly after this discovery several authors proposed alternative properties (some of which bearing some resemblance with the CO) to quantify the geometry of the native structure (e.g. the long range order[3] and cliquishness[4], just to mention a few examples) that appeared to correlate equally well with the folding rates of small, two-state proteins. Clearly, part of the charm of these reductionist approaches is that they dramatically simplify the solution to the folding puzzle since the protein's primary sequence and all its inherent complexities is no longer at the center stage of the problem. Concretely, Plaxco's 'law' somehow complements Anfinsen's folding principle - that the protein's primary sequence determines the native structure [5] - by stating that it is the native structure itself, through its geometric property CO, that determines the folding rate. This concept had important practical consequences in the folding arena especially in the field of simulations where native-centric Gō models gained considerable popularity throughout the last decade[6, 7].

But if the CO-rate dependence is to become a fundamental principle of protein folding it is necessary to understand its fundamental roots, i.e., to identify the physical mechanism underlying the correlation. Faísca and Ball were the first to explore the CO-rate dependence in the context of Monte Carlo simulations of simple lattice models but they only observed moderately high correlations for long chain lengths and high CO[8]. It appears that when protein energetics is modeled by pair-additive interaction schemes the folding timescales observed in nature (where the fastest, two-state protein folds six orders of magnitude quicker than the slowest one) cannot be recapitulated through simulations[9]. Inspired by the seminal work of Jewett et al.[10], a landmark study by Kaya and Chan[11], also in the context of lattice models, proposed that the CO-rate dependence should rely microscopically on non-additive protein energetics based on a coupling mechanism between local and non-local (i.e. long-range) interactions (i.e. interactions involving pairs of residues that are far away along the sequence). The latter not only renders the folding transition more cooperative (both kinetically and thermodynamically)[9] but also results in highly dispersed folding rates, a necessary condition to observe a strong statistical correlation with the CO. In a recent contribution, also framed on lattice models, we studied different families of circular permutant proteins, i.e., proteins having different chain connectivity (and therefore different CO) but sharing essentially the same native structure[12]. A potential advantage of our approach is that it allows isolating the effect of CO from other protein properties that are known to affect folding rate (e.g. native stability). We proposed that the lattice CO-rate dependence stems from the formation in the transition state ensemble (TSE) of a specific network of native interactions that guarantee that the TSE has essentially the same CO as the native structure in consequence of being distinctively long-ranged. Somehow this result should not be viewed as surprising because if the CO is to correlate with folding rate in proteins whose folding transition is well-modeled by transition state theory, then a close approximation to the native topology must be realized in the TSE, which is the rate-limiting step[13]. In line with this finding, previous simulation studies by Lindorff-Larsen and co-workers[14], who investigated the TSE of three SH3 folding domains (which are all high-CO proteins) with off-lattice models, concluded that the folding transition states exhibit indeed the native topology.

An array of studies collected in the last decade, both theoretical (including simulations) and experimental, somehow contributed to strengthen the idea that the CO is a major drive of two-state folding kinetics[10, 11, 15-21]. However, despite these important insights the physical principles underlying the CO-rate dependence remain elusive and a widely accepted physical theory for the CO-rate dependence is still missing. Furthermore, it is known that the CO alone is not able to predict the folding rate of larger, multi-domain proteins[22,23]. Additionally, when recent experimental kinetic data have been carefully selected for single domain proteins, taking care to eliminate temperature effects, the correlation between folding rates and CO does not seem to be so relevant[24]. Consequently, Muñoz and co-workers have claimed that the folding rate is mainly determined by an even more fundamental protein property, which is that of chain length[25]. Thus, a natural and logical question arises which is to know “how strong a determinant of protein folding rate is chain length alone” [24, 25] A body of work developed by the Muñoz laboratory, combining theoretical predictions with experimental data, put forward the idea that folding rate is fundamentally determined by chain length. In particular, chain length is not only the major determinant of folding rate, as it is also the major determinant of protein stability, which, in turn, also affects folding rate[24-26]. However, the correlation between the folding rates and chain length improves when the type of structural class (all-α, all-β, or α/β) is taken into account[26]. While this result highlights the importance of structural features of the native structure, the possibility to reduce it to a single parameter like the CO is left aside.

With these ideas in mind, the present work revisits the importance of native structure as a determinant of the thermodynamics and kinetics of the folding transition in small, single-domain proteins. However, instead of adopting a reductionist view of the native fold, we explore a specific trait of the native structure, which concerns the conformational preferences of protein termini in small, single-domain proteins. The reason is three-fold: 1) a few years ago, Krishna and Englander performed a statistical survey of the protein data bank (PDB)[27] and established a relation between the proximity of the termini and the type of folding transition[28]. More precisely, they observed that proteins that fold with two-state kinetics (and therefore cooperatively) have their termini close together in the protein surface, whereas proteins that fold in a non-two-state kinetics have their termini separated. 2) In a recent account addressing the investigation of co-translational folding in the context of lattice models, we came to realize that if the native structure exhibits proximate (and interacting) termini, then the fully synthesized chain is necessary for the protein to fold to the native structure[29]. More precisely, we found that the addition of the last two amino acids and the interactions they establish are determinant drivers of the two-state, cooperative transition exhibited by the full-length protein. Finally, 3) in an unrelated research project, framed on full atomistic off-lattice simulations, we studied specific mutations that impair folding cooperativity of small proteins by triggering the formation of aggregation–prone intermediate states[30-32]. The latter are characterized by the existence of unstructured (instead of aligned) termini (as in the wild-type form).

Therefore, these independent results appear to point out to the importance of termini coupling (i.e. interacting termini) as a driver of protein folding cooperativity.

In the present work, in order to access the importance of termini coupling as a determinant of two-state folding transition in small, single domain proteins, we take a two-stage approach: first we use Monte Carlo simulations of simple lattice models as a platform for obtaining theoretical predictions and then we ‘test’ the lattice predictions in the context of off-lattice Monte Carlo simulations of a coarse-grained C_α model[33] . In both cases protein energetics is modeled with a Gō potential[34].

Monte Carlo simulations of coarse-grained models have a long history in the folding literature where they have been used to explore many aspects of protein folding[20, 35-47], and, more recently, of protein aggregation[48-54]. A particularly relevant contribution of lattice models was the prediction of the nucleation-condensation mechanism for small, two-state proteins[55-58]. An important advantage of these models is that their computational feasibility allows accessing very long timescales resulting into accurate measures of thermodynamics and kinetics of folding and aggregation. Our rationale in adopting the outlined approach is as follows: lattice models embed the most fundamental protein properties (compactness of the native structure, chain connectivity, excluded volume interactions, etc.) through the simplest protein representation; therefore, they should succeed in capturing fundamental aspects of protein folding such as those related with chain termini. Off–lattice coarse-grained models, on the other hand, are not affected by the strong geometric constraints imposed by the lattice while allowing the study of specific proteins in the PDB. These models can therefore be used as a first platform for testing the lattice predictions. It is with this goal in mind that they are used in the present study.

Models and methods

The simple lattice model and the simple off-lattice C_α model

As briefly outlined in the previous section, this work explores protein folding in the context of Monte Carlo (MC) simulations of two simple coarse-grained models. We start our investigation by using the simple lattice representation in which amino acids are reduced to beads of uniform size placed on the vertices of a regular three-dimensional cubic lattice. The peptide bond (that covalently connects the amino acids along the polypeptide chain) is represented by uniform sticks with size equal to the lattice spacing. In order to satisfy the excluded volume constraint, only one bead is allowed per lattice site. To generate appropriate lattice model proteins one runs thousands of MC homopolymer relaxation simulations whereby the polymer chain collapses in a non-specific manner to maximally compact cuboids. We then select from these native-like conformations those exhibiting specific traits (e.g. interacting termini), which will be used as native structures in our lattice model studies.

The off-lattice model used in this work classifies as simple because it also reduces the geometry of the polypeptide chain. In particular, every amino acid is modeled by a hard-sphere of uniform size centered at the amino acids' C_α atom. Likewise, connecting beads are separated by 3.8 Å, which is the characteristic size of a *trans* peptide bond.

To model protein energetics we use a Gō potential in both cases[59]. A Gō potential is a native-centric interaction potential. This means that it is determined by the protein's native structure. The nature of the protein representation used in each model implies that the evaluation of the native contact map (and therefore of the intramolecular potential) is model dependent. In the following we briefly explain how the Gō potential is implemented on- and off-lattice.

Protein energetics on-lattice

In the lattice representation with protein energetics modeled by the Gō potential the energy of a conformation, defined by the set of bead coordinates $\{\mathbf{r}_i\}$, is given by the so-called contact Hamiltonian

$$E(\{\mathbf{r}_i\}) = \varepsilon \sum_{i,j>i+2}^{L} \Delta(\mathbf{r}_i - \mathbf{r}_j), \qquad (1)$$

where L is the chain length (i.e. number of beads in the protein), ε is the (uniform) interaction energy parameter (taken as -1 in this study) and the contact function $\Delta(\mathbf{r}_i - \mathbf{r}_j)$ is unity only if beads i and j form a native contact (i.e. a contact that is present in the native structure) and is zero otherwise. The fraction of native contacts, Q, is defined as the number of native contacts found in a specific conformation normalized to the total number of native contacts.

Protein energetics off-lattice

A crucial difference between the lattice and off-lattice representations is that the discrete nature of the lattice automatically defines the contact distance, which is provided by the lattice spacing. The same does not apply off-lattice, for which a cut-off distance for the native interactions must be established *a priori*.

In the context of the off-lattice model used in this work the energy of a conformation is defined according to the following expression

$$E(r_{ij}) = \sum_{all \text{ pairs}}^{L} \varepsilon_{ij}(r_{ij}), \qquad (2)$$

where the interaction energy parameters ε_{ij} for every pair of residues i and j separated by distance r_{ij} are modeled by a harmonic well centered at the native distance d_{ij}^{nat} according to the following expression[33]:

$$\varepsilon_{ij}(r_{ij}) = \begin{cases} \Delta_{ij}\left[(r_{ij} - d_{ij}^{nat})^2 - a^2\right] & if \ (d_{ij}^{nat} - a) < r_{ij} < (d_{ij}^{nat} + a) \\ 0 \quad otherwise \end{cases} \qquad (3)$$

In this expression, a = 0.6 Å establishes the width of the potential well[33]. Depending on the sequence separation between i and j three types of interactions are considered. When $j = i+2$ or $j = i+3$, $\Delta_{ij} = 1/a^2$ in all the cases, so that the corresponding term ε_{ij} = -1 (the energy unit for the model) at the native distance. Moreover, when $j = i+3$, i and j establish a virtual torsion angle interaction. By assigning to the distance between residues the sign of the scalar product of the three vectors defining the virtual torsion angle, local chirality is introduced in the definition of these interactions. When $j \geq i+4$, attractive Gō interactions will occur if, and only if, i and j form a native contact. Two residues i and j are considered to be in contact in the native structure if the shortest distance from all the possible pairs among heavy atoms belonging to both residues is smaller than or equal to 4.5 Å (this cut-off is used because it is slightly larger than the van der Waals radii for heavy atoms in proteins). In these cases, Δ_{ij} = $1/a^2$ again, recovering a unit energy interaction. For all the other pairs, $\Delta_{ij} = 0$.

Monte Carlo folding simulation

In order to mimic the protein's conformational behavior as a function of temperature we use the Metropolis MC algorithm[60] both on- and off-lattice. Conformational sampling on the lattice is achieved through a local move set that includes corner-flips and end-moves (these types of moves displace one bead at a time; the end-moves are exclusively used to move the chain's termini and the corner-flip is used to displace all the other beads in the chain) and the crankshaft move (which involves the simultaneous displacement of two beads except termini beads). In the off-lattice model the exploration of the conformational space proceeds via similar end and spike moves, involving one single bead, and displacement moves which shift a part of the chain (from a randomly selected bead to the chain end) a maximum distance of a virtual bond.

In both cases a MC simulation starts from a randomly generated unfolded conformation and the ensemble characteristics are monitored through several properties (e.g. the fraction of the established native contacts, Q). Further details on the adopted simulation algorithm can be found elsewhere for the lattice[61] and off-lattice models[33, 44].

Computing folding thermodynamics

In order to explore the thermodynamics of the folding transition and compute equilibrium properties we have conducted long replica-exchange (RE) MC simulations at up to 40 (60) different temperatures on-lattice (off-lattice), depending on the considered protein. Each MC trajectory consists of - at least – 10^8 MC-steps per residue after equilibration. We swap replicas every 10^6 MC steps, which is about one order of magnitude larger than the largest auto-correlation time for the energy recorded in simulations at fixed temperature, which allows the replicas to equilibrate between two consecutive RE attempts. The acceptance ratio for the RE is high (>80%) and each replica reliably and repeatedly visits all the temperatures in the grid with cycle time of approximately 40 RE moves. A single total simulation comprises at least 25 full cycles. The calibration of the temperature grid adopted in the RE simulations guarantees good convergence of the data.

The heat capacity C_v is evaluated from the mean squared fluctuations in energy at each temperature considered in the RE simulations in accordance with the definition $C_v = (<E^2> - <E>^2)/T^2$ (where both energy and temperature are used in reduced units along this work). Therefore, the data from the RE simulations allows to directly compute the heat capacity as a function of temperature.

The melting temperature T_m (also known as folding or transition temperature) is defined as the temperature at which the unfolded and native states are equally populated at equilibrium in a two-state transition. Here, as well as in experiments *in vitro*, T_m is estimated as the temperature at which the heat capacity C_v attains its maximum value.

To evaluate the free energy as a function of one (or more) reaction coordinates we use the weighted histogram analysis method (WHAM)[62]. The results presented here correspond to average values from five independent RE simulations. The overlap of the results from the independent trajectories has been checked prior to averaging, as it represents an important proof of correct data sampling.

Computing folding kinetics

To obtain kinetic properties such as the folding rate, we carried out fixed temperature MC lattice simulations at T_m. To get statistically significant kinetic measurements, we computed 2000 independent MC folding runs. The corresponding folding times (i.e. first passage times) allow evaluating the distribution of proteins which remain unfolded as a function of MC 'time' (i.e. MC steps). The folding rate is given by the slope of the linear fitting of this distribution to a single-exponential decay[12, 46].

Results

Lattice model systems

We investigate three lattice model systems, named CP0 (**Fig. 1a**), CP1 (**Fig. 1b**) and CP2 (**Fig. 1c**). The three conformations are related by circular permutation (CP). Consequently the native structure is essentially preserved across the three CPs, but different chain connectivity renders the three CPs with high but different contact order (CO) that varies from 0.45 (CP0) to 0.40 (CP1 and CP2). We note that the reported COs are among the highest values exhibited by lattice proteins. In a previous report we found that within the family of CPs to which our model systems belong the correlation between the CO and the logarithmic folding rate is 0.8[12]. A common structural feature exhibited by these lattice proteins is that they all have the first and the last bead in interaction with each other. We consider terminal regions composed of five residues, which corresponds to 10% of the full protein chain length (L = 48), and define termini interactions as those that establish between the C-terminal and N-terminal residues. We use the acronym NC to refer to termini interactions. An important feature of NC interactions is that they are necessarily long-ranged. Indeed, they are the most long-ranged interactions within the native structure. The three target systems exhibit different conformational arrangements of the termini and, as a consequence, the number of NC interactions is also different. These NC interactions are responsible for coupling the C-terminal and the N-terminal in the considered model systems. In CP0 the chain termini are structurally aligned through four NC interactions (1-48, 2-47, 2-45, 3-44) (**Fig. 1d**) CP1 has the strongest coupling of the termini (1-48, 3-48, 3-46, 4-45, 5-44) (**Fig. 1e**) and CP2 has the weakest termini coupling based on only two NC interactions (1-48, 4-47) (**Fig. 1f**). The NC interactions represent 8%, 10% and 5% respectively, of the CO of CP0, CP1 and CP2. In addition, we consider one control system c1 (**Supplementary Information: Fig. 1**), which is a circular permutant of CP0, CP1, and CP2 with lowest CO.

Folding transition on lattice

We start by assessing the importance of termini coupling in the folding process of system CP0. The free energy profile (i.e. the projection of the free energy onto the fraction of native contacts Q) evaluated at T_m shows a pronounced and well-defined transition state region separating the native state from the denatured basin (**Fig. 2a**). The folding transition at T_m is therefore unambiguously thermodynamically two-state. Up to which extent is this behavior determined by the NC interactions? In order to answer this question we suppressed (i.e. “switched-off”, ε_{ij} = 0) the four NC interactions (which amounts to increase the native energy by 7%) and measured the impact of this energetic perturbation on the folding transition. First, and perhaps not surprisingly, the heat capacity peaks at a lower temperature, indicating a lower thermal stability, and the decrease in T_m is linearly proportional to the number of suppressed native interactions. No significant changes are observed in the shape of the heat capacity curves (data not shown). However, and most importantly, there is a remarkable lowering of the free energy barrier upon "switching-off" the NC interactions. Furthermore, the overall shape of the free energy profile undergoes substantial changes with the TS region becoming considerably broader, which indicates that the perturbed system folds through different pathways. Interestingly, the inspection of the histogram for the number of conformations with fraction of native contacts Q reveals that the number of conformations with energy intermediate between the native and denatured states at T_m is 3.5 times higher with deactivated termini coupling (i.e. when the NC interactions are "switched-off") (**SI: Fig 2**), suggesting that termini coupling contributes to increasing protein folding cooperativity. Furthermore, the projection of the free energy onto reaction coordinates energy, E, and radius of gyration, R_g, reveals that the free energy surface becomes substantially more rugged upon termini deactivation (**Fig. 2b**, **Fig. 2c** and **SI: Fig. 3**). These observations are in line with results reported in Refs.[10, 11] where an increase in protein folding cooperativity was shown to decrease the folding rate. These findings thus indicate that coupling of the termini residues is an important determinant of the folding transition and, in particular, a major contributor to the folding free energy barrier as well as a modulator of protein folding cooperativity.

At this point a question arises which is that of determining if the observed effects are specific of the NC termini interactions. In order to investigate this issue we performed two control experiments. In the first experiment, four native contacts were randomly selected across the native structure of CP0 and the corresponding interactions "switched-off". For this control experiment, five different sets (r1-r5) of four native interactions were tested (**SI: Fig 4a**). The second control experiment takes into account the fact that the set of NC interactions are located close together in the native structure (**SI: Fig 4b**). To evaluate the importance of this structural feature we considered clusters of four native interactions (there are only two such clusters, which we term l1 and l2, in the native structure of CP0). Our results show unambiguously that the strongest stabilization of the TS is obtained upon "switching-off" the NC interactions, thus confirming that coupling of the termini is a critical modulator of the folding transition. However, there are two sets of native interactions (r4 and r5) that imprint an important (although not as striking) impact on the folding transition (**SI: Fig 4a**). In order to understand why this is so we started by identifying the residues involved in those interactions. Interestingly, we found one termini interaction (4-43) in r5 and one long-ranged interaction involving one terminal residue (6-43) in r4. There are no NC or interactions of similar range in sets of native contacts r1-r3. Therefore, the control experiments also underscore the idea that N- to C-termini coupling is critical for cooperative folding.

Finally, since model system CP0 is a high-CO structure it is important to evaluate the role of termini coupling in a low-CO structure. Thus, we performed a last control experiment with model system c1 (**SI: Fig 1**). The lowering of the free energy barrier is non-negligible (10%) although it is not as striking as in CP0 (30%) (**SI: Fig 5**). This stems most likely from the number of NC interactions in c1 (which is only one, that between residues one and 48, against the four NC interactions in CP0).

Having performed these control experiments, we moved on in our investigation by exploring model systems CP1 and CP2. Since CP1 has a stronger coupling of the termini (driven by five NC interactions), one expects to observe a stronger stabilization of the transition state upon "switching-off" the corresponding interactions. Model system CP2, on the other hand, should display an opposite behavior because termini coupling is weaker in this case resting upon on only two NC intramolecular interactions. Results reported in **Fig. 2a** confirm our expectations, with the free energy barrier of CP1 decreasing 37% (against the 30% decrease in CP0), and that of CP2 decreasing only 17%.

Transition state structure on lattice

The results reported in the previous section indicate that energetic coupling between termini residues is an important determinant of the folding TS, contributing significantly to establish the height of the free energy barrier. Therefore, it is likely that the structure of the TSE should somehow reflect this observation. Here, in order to address this issue, we investigate the structure of the TSE when the NC interactions are either "switched-on" or "switched-off". In particular, in each case, we constructed ensembles of 2000 conformations with fraction of native contacts $0.4 < Q < 0.6$, representative of the TSE. These conformations were collected from many independent MC folding runs (at T_m) in order to guarantee that they are statistically uncorrelated. To get insight into the structure of the TSE we evaluated probability contact maps, which indicate how likely is the formation of each native interaction in the TSE. Results reported in **Figure 3 (a-c)** highlight the fact that when the full set of termini interactions is "switched-on" there is a moderate probability (ranging from $p \sim 0.2$ to $p \sim 0.5$) to have the termini interactions established in the TSEs of the three model systems. These probabilities decrease to negligible values when the termini interactions are "switched-off" (**Fig. 3d-f**); in this case the star-shaped cluster of native interactions establishing around interaction 9-34 become more prominent. This analysis indicates that the entropic cost of loop closure associated with the establishment of very long-ranged termini interactions in the TSE is a major contributor the height of the free energy barrier thus consolidating the importance of termini coupling in two-state protein folding.

Folding kinetics on lattice

Here we analyze up the impact of termini coupling in the folding kinetics of model systems CP0, CP1 and CP2. The folding rate of the four model systems is reported in **Figure 4**. Despite having very similar CO (~0.40 - 0.45), the three CPs exhibit different folding rates, with CP2 (CO = 0.40) being the slowest (**Fig. 4b**) folding lattice protein and CP0 (CO = 0.45) the fastest (**Fig. 4a**). The qualitative behavior for the folding rates is in agreement with the height, h, of the free energy barrier ($h_{CP2} > h_{CP1} > h_{CP0}$) reported in **Figure 2a**. In **Figure 4a** we further report the folding rate of c1, the control system of low-CO. We note that the CPs with the "switched-off" NC interactions have an effective CO that is slightly lower ranging from 0.39 (CP2) to 0.42 (CP0), but still much higher than that of c1 (CO = 0.23). The increase in folding speed for the three model systems is, however, remarkable with CP0 and CP1 actually folding faster than the low-CO control system (**Fig. 4a**). This result shows that despite contributing with only 5-10% to the overall CO, termini interactions are critical modulators of the folding speed. Indeed, a high-CO native structure with "switched-off" termini-interactions actually becomes a faster folder than a low-CO protein with essentially the same native structure and coupled termini.

Rationale for protein selection in the off-lattice simulations

In order to confirm the importance of termini coupling as a major determinant of the folding transition as predicted from lattice simulations, we extended our investigation to small, single domain real-world proteins whose folding process was explored with off-lattice Monte Carlo simulations.

In the off-lattice framework we can no longer take advantage of a fundamental feature of lattice models, which is the possibility to design native structures with pre-defined structural traits (e.g. coupled termini). Instead, we are limited to proteins whose native structures are available from the PDB[27]. For the purposes of the present study, the selected proteins must fulfill the following requirements: 1) they must have the C-terminal and N-terminal ends of their polypeptide chains close enough in space in their native structures so that termini contacts appear in the native contact map (we recall that in the context of a Gō model the contact map determines the interaction potential; therefore if there are termini contacts there will be termini interactions); 2) they must be of similar chain length to minimize size effects; and 3) they must show a two-state folding transition, allowing for the possibility of a large free energy barrier between the folded and denatured states whose traits may be perturbed by "switching-off" the termini interactions. This last constraint rules out the possibility to select all-α proteins of small size, which tend to exhibit very small or inexistent folding barriers[45]. The first requirement mentioned above deserves an additional comment. We are not restricting our selection to proteins with interactions between the terminal *residues*, as we have done in the lattice simulations. Instead we are looking for folded structures whose terminal *regions* are in contact. This usually means a certain degree of alignment of the N- and C-terminal ends, which have to be spatially close to be useful in our analysis, even though the very terminal residues at both ends may not necessarily be in contact. Indeed, the focus of Krishna and Englander's work[28], which partially inspired the present study, was the existence of native contacts between the N- and C-terminal secondary structure elements, and not between the terminal residues themselves.

Proteins studied off-lattice

With these criteria in mind, we selected from the PDB three well-studied two-state folding proteins, with high CO, whose native structures and contact maps are reported in **Figure 5**. 2GB1 is the PDB code of the immunoglobulin binding domain of streptococcal protein G[63]. It is a protein domain of 56 residues, with α/β structure, whose termini form the central strands of a four membered β-sheet; its CO is 0.34. As seen in both the native structure and in the contact map (**Fig. 5a**), the terminal β-strands are packed against each other in a parallel manner so that the N- and C-terminal residues are not in direct contact, but the terminal secondary structure elements are. With the adopted cut-off distance for the native interactions, there are 108 native contacts in 2GB1 (we note that native contacts corresponding to virtual bonds, virtual bond angles and virtual torsion angles are not taken into account in this number). In particular, the chain ends interact through 20 native contacts (18.5% of the total number of contacts that represent 32% of the protein's CO) that are encircled within the red ellipse in the contact map (**Fig. 5d**). As in the lattice simulations, we designate by NC the termini contacts and corresponding interactions.

The code 1SHG[64] corresponds to the spc-SH3 folding domain. Although the PDB structure contains the atomic coordinates of residues 6 to 61, we have only considered residues 9 to 60. We did so because the curtailed parts are floppy and their elimination results into a folded structure with well-defined termini contacts (**Fig. 5b)**. Therefore, the investigated structure has 52 residues, 116 native contacts and its CO is 0.38. As before, we marked the NC contacts between the termini with a red ellipse (**Fig. 5e**). The corresponding 14 interactions (12.1% of the total number of native interactions that represent 18% of the total CO) establish directly between the terminal residues located in an anti-parallel arrangement of β-strands.

Finally, we have also selected the structure with PDB code 2CI2[65], which corresponds to chymotrypsin inhibitor 2, CI-2, a well-studied two-state folding protein (**Fig. 5c**). The solved structure has 65 residues. However, since three of its end residues do not establish native interactions with other parts of the protein we have also curtailed them in order to obtain a native structure with well-defined termini contacts. The investigated native structure, with 62 residues and 127 native contacts, is only slightly larger than those of the other proteins considered in this work. Its CO is 0.36. Furthermore, as one may appreciate from the analysis of the contact map (**Fig. 5f**), this protein has the largest number of NC contacts among those considered in this work; these 27 native contacts (21.3% of the total number of native contacts that represent 36% of the protein's CO) are encircled within the red ellipse in the contact map.

Setting-up the off-lattice control simulations

In contrast with lattice proteins, real proteins exhibit well-defined levels of structural organization. In particular, the interactions between the secondary structural elements form clusters of native contacts that can be easily spotted in the native contact maps (**Fig. 5d-f**). Therefore, for the construction of the control simulations, instead of randomly "switching-off" native interactions, we have taken into account the characteristics of the folded structure. In the context of off-lattice simulations the sets of interactions that are switched-off in the control experiments are named control interactions.

In the case of 2GB1 the four β-strands provide three contacting regions as shown in the contact map (**Fig. 5d**). One of these regions corresponds to the already mentioned NC contacts between terminal ends. Therefore, for the control simulations we have considered the remaining two regions in the contact map. The first one comprises 23 native contacts defining the N-terminal β-hairpin (enclosed within the green ellipse), while the other encapsulates 16 native contacts in the C-terminal β-hairpin (enclosed within the blue ellipse). The number of suppressed native interactions is different in each control simulation but the differences are relatively small. Therefore, instead of artificially selecting an identical number of native contacts we decided to keep the differences while letting the native structure dictate natural choices for the control simulations.

In the case of 1SHG (**Fig. 5b**) there are two β-hairpins in the structure with respectively 16 native contacts (green ellipse) and 21 native contacts (blue ellipse). The 18 native contacts inside the pink ellipse correspond to tertiary interactions that stabilize the native fold. The 30 native contacts involving the protein's N-terminus are marked with a brown ellipse (**Fig. 5e**). Although the native structure motivates a control simulation based on this set of native contacts, the number of interactions that must be “switched-off” is excessively high and much larger than in the other controls. Indeed, upon “switching-off” these native interactions, we observe a three-state folding transition with the N-terminus of the chain folding at a temperature lower than the remainder of the structure (data not shown). This situation precludes a proper comparison with the other simulations carried out for this protein. Likewise, the ensemble of interactions at the N-terminal region is not a reasonable control system for the purposes of this work and will not be considered further.

Finally, for 2CI2 we considered the 23 contacts inside the green ellipse that correspond to native interactions between the two parallel β-strands, and the 11 contacts at the C-terminus of the chain, encircled into the brown ellipse (**Fig. 5f**). Interestingly, since the NC contacts are clustered into two different subsets we have separately considered each one of these clusters for the purposes of additional control simulations: the most terminal region inside the blue ellipse with 11 contacts, and the region inside the pink ellipse with 16 native contacts, which involves the residues next in the sequence to the chain ends.

Folding transition off-lattice: Heat capacity

As in the lattice model, we studied the folding transition when protein energetics is driven by the full set of native interactions and when the NC (and control) interactions are suppressed. The heat capacity curves, which provide the transition temperatures T_m, and the free energy profiles computed at T_m by applying the WHAM method to the full set of temperatures, are reported in **Figure 6**.

In the off-lattice model, in order to decide if a native contact is formed in a given conformation sampled along the MC simulation one evaluates the distances between all pairs of alpha-carbons that are in contact in the native structure. A native contact is considered formed if the corresponding evaluated distance is within 10% of the corresponding native distance. Given this somehow arbitrary cut-off distance, we used the energy as reaction coordinate. We note, however, that in a Gō model, the energy is equivalent to the fraction of native contacts Q that was used in the lattice simulations. In particular, the highest value of Q, which corresponds to the native state, has the lowest value of energy.

The analysis of the heat capacity curves shows, as expected, and in line with lattice simulations, that the suppression of native interactions leads to the thermal destabilization of the native state, as indicated by smaller values of T_m (**Fig. 6a-c**). However, contrary to what happens in the lattice simulations, the decrease in T_m is not linearly proportional to the number of suppressed native interactions. For example, in 2GB1 the effect of suppressing the 20 NC interactions is remarkably more striking than that of suppressing 23 control interactions at the N-terminal hairpin (red and green curves in **Fig. 6a**). To rationalize this result we should recall that in a two-state transition T_m is determined by the energetic and entropic components of both the native and the denatured states, and cannot be reduced to the energy of the native state alone. This is an important fact, which, in addition, provides an interesting ground for comparing the behavior of the lattice and off-lattice models. In the lattice model (**Fig. 2a**, **Fig. SI 4** and **Fig. SI 5**), the native state is conserved across the different simulations (i.e. the minimum of the free energy is always found for Q =1), and the unfolded state is also fairly conserved. In the off-lattice model, however, thermal fluctuations have an effective impact on the ensemble of protein conformations that form the folded state sampled at T_m, and the corresponding free energy minima shifts to lower or higher energies depending on the amplitude of these fluctuations (**Fig. 6d-f**). This is valid for the denatured state as well. The limitation of the lattice model in capturing the effects of conformational entropy stems most likely from the strong geometric constraints imposed by the lattice that impede the occurrence of most conformational changes that are driven by thermal fluctuations.

Folding transition off-lattice: Free energy profiles

The shape of the heat capacity curves (**Fig. 6a-c**) already indicates significant changes in the folding transition upon suppressing native interactions. Indeed, the heat capacity curves become broader (and therefore lower), which implies a less cooperative folding transition. In some cases, the curves show a very smooth low temperature tail, which indicates that the main heat capacity peak does not always signal a two-state transition between the native and the denatured states but may involve intermediate states or even underlie a continuous downhill transition[45]. This can be clearly appreciated in the free energy profiles (**Fig. 6d-f**). When the full set of native contacts contributes to protein energetics (black curves) the free energy profiles always show two well-defined free energy minima, and a large free energy barrier between them. In 2GB1 and 2CI2, the suppression of NC interactions leads to a marginal, almost inexistent barrier (**Figs. 6d** and **6f**, respectively), while in 1SGH the free energy barrier at T_m is reduced to approximately 33% of its original value (**Fig. 6e**). Therefore, the off-lattice results support the lattice predictions that termini coupling plays a major role in determining the characteristics of the folding transition, and, in particular, the height of the free energy barrier.

In the three proteins, the reduction in the barrier height is related to the amplitude of thermal fluctuations, which distinctly perturb the denatured state, the folded state, or both. In 2GB1, the absence of a free energy barrier at T_m upon suppressing NC interactions implies the occurrence of a 'continuous' ensemble of conformations going from the fully folded to the denatured states (data not shown). The situation for 2CI2 is more complex, though, since the lowest energy conformations sampled at T_m are not actually folded upon suppressing the NC interactions. That this is so can be confirmed by analyzing the ensemble-averaged root mean square deviation (RMSD) measured with respect to the corresponding PDB native structure at different temperatures (**SI Fig. 6**). When the full set of native interactions contributes to protein energetics the three proteins exhibit an abrupt two-state transition at T_m, with high RMSD above T_m falling close to 1 Å below T_m. The same happens with most of the NC and control simulations. However, the NC case for 2CI2 is an exception to this general behavior (**SI Fig. 6c**). This happens because the number of native interactions "switched-off" in this case is quite large and clearly reduces the stability of the native state. As a matter of fact, the ensemble of lowest energy conformations sampled at T_m (and, indeed, slightly below T_m as well) form a compact but distorted state that is similar to the native structure with loose ends (especially the N-terminus), which only become native at lower temperatures (**SI movie**). The breakdown of the folding transition is also observed for the control interactions "mid1" (**Fig. 6f** and **SI Fig. 6c**).

With the exception of 2GB1, the picture conveyed by the off-lattice control simulations is not as clear-cut as in the lattice model. In 1SGH, NC interactions and control interactions "mid3" (involving tertiary contacts) have both a similar impact on the folding transition (**Fig. 6e**), and their impact is smaller than that of control interactions "mid2" (within one of the β-hairpins). However, the number of native interactions suppressed in control simulations “mid2” and “mid3” is larger than the number of NC interactions, which may partly explain the larger free energy reduction recorded for “mid2” and “mid3”. Furthermore, when the number of “switched-off” control interactions is very similar to that of NC interactions (e.g. in the “mid1” control), the impact of suppressing the latter is clearly larger.

In the case of 2CI2 the analysis of control simulations NC1 and NC2 provides important insights (**Fig. 6f**). We recall that NC1 and NC2 are the two sub-clusters of NC with respectively 11 and 16 termini interactions. In the case of NC1, the height of the free energy barrier is reduced to approximately 60% of its original value, in line with predictions from lattice simulations. However, this control experiment renders a thermodynamic behavior that is essentially the same as that observed for the “C” control simulation, where the suppressed native interactions only involve one of the chain ends and are considerably more short-ranged than in the NC1 control (**Fig. 6f**). Moreover, even though the interactions in NC1 are more ‘termini-like’ than those in NC2, we observe that “switching-off” the latter plays a more striking effect in the barrier height. It is true that the difference between 11 and 16 interactions removed in these control simulations is important in relative terms, and this difference can indeed affect the results. But another possibility should be considered when analyzing the results obtained for the different cases simulated for this protein in more detail.

As already pointed out, the native structure of 2CI2 is not populated at T_m when the NC interactions are “switched-off” because the N-terminus never attains its native position at this temperature (**SI movie**). In the NC1 control simulations, on the other hand, there is still a two-state transition between the native and denatured states, although the height of the free energy barrier is smaller. In the NC2 control simulations the transition is downhill, and its free energy profile shows shallow minima for the native and denatured states, and for an additional state with intermediate energy (which corresponds to the lowest energy minimum in the free energy profile of the NC simulation, **Fig. 6f**). This intermediate state is essentially folded except for the unstructured N-terminal end and is also populated when the NC interactions are “switched-off”. Therefore, the characteristics of the folding transition observed upon suppressing NC interactions appear to be majorly driven by the native interactions forming the NC2 subset (**Fig. 5f**). By analyzing in detail the contact map of 2CI2, one notices that the C-terminal residues not only interact with the N-terminal residues, but also interact to a significant extent with residues 41 to 48 (according to our numbering) through the set of native contacts circled in brown (**Fig. 5f**). On the other hand, the N-terminal residues interact mostly with the C-terminus and with the helix through a set of nine scattered native contacts encircled by the orange ellipse (**Fig. 5f**). The results from the control simulations thus show that even if the sub-set of termini interactions NC1 are suppressed, the structural integrity of the native structure is still ensured through the set of alternative native contacts indicated by the brown and orange ellipses. The energetic stability of the native state is reduced, but these sets of native interactions ensure that the native structure is still populated at T_m. On the other hand, if the interactions corresponding to the NC2 contacts are "switched-off", the corresponding residues at the N-terminus (with numbers 7 to 13) are not significantly stabilized via interactions with other regions of the native structure, and the conformational entropy, which plays a critical role in the off-lattice simulations, favors the N-terminal part of the chain to become unstructured.

Transition state structure off-lattice

In the off-lattice proteins the two-state behavior is lost in many of the NC and control simulations. This is the reason why an analysis of the TSE at T_m as the one carried out on-lattice does not make sense in most of the cases simulated off-lattice. However, the NC system in 1SHG still exhibits two-state folding behavior, as indicated by the presence of a significant free energy barrier (**Fig. 6e)**. This allows for the selection of conformational snapshots at T_m corresponding to the TSE. As in the on-lattice analysis, we use probability contact maps to investigate the structural features of the TSE (**SI: Fig. 7)**. When protein energetics is driven by the full set of native interactions in 1SHG there is a significant population (30-40%) of established long-range interactions (including those defined as NC) (**SI: Fig. 7a**). The reported frequencies may be compatible with the ϕ values experimentally reported[66], but the populations resulting from our simple simulation model are comparable to other native contacts of middle range. As in the lattice simulations, the entropic cost involved in bringing together the residues forming these contacts may then be considered as responsible for the larger barrier existing for the protein when the NC contacts are "switched-on". When the NC interactions are "switched-off" the TSE shows highly frequent native contacts relatively close to the main diagonal, i.e., for residues that are not too far away along the sequence, and the NC interactions are no longer present, as expected (**SI: Fig. 7b**). As a result there is an entropy increase resulting from unstructured (and wobbling) termini that decreases the free energy barrier. Since the effect of unstructured termini is not properly captured in the lattice model (due to the strong geometric constraints imposed by the lattice) the decrease in the barrier height is not as significant on-lattice as it is in the off-lattice simulations.

We do not report results for off-lattice folding kinetics because its analysis is redundant. As indicated from the reported and previous lattice results (see, e.g.[12]) there is a direct correlation between the folding rate and the height of the free energy barrier for systems that exhibit a two-state folding behavior. We have checked that the same behavior holds for off-lattice systems as well (data not shown).

Conclusions

A major challenge in molecular biophysics is the establishment of the general principles of protein folding. Because large proteins, formed by several domains, typically fold via intermediate states and require the assistance of molecular chaperones[67], researchers have been focusing their attention toward single domain proteins with chain length between 44 and 100 amino acids. The vast majority of these small proteins fold through a remarkably cooperative process, whereby a large free energy barrier separates the native from the denatured state. Here, we investigated the importance of termini coupling as a major determinant of the thermodynamics and kinetics of the two-state folding transition through extensive Monte Carlo simulations that combine lattice predictions with off-lattice ‘experiments’. The reported results indicate that termini interactions (i.e. interactions established directly between the termini residues or between the N- and C- terminal regions of the polypeptide chain) play a pivotal role in folding thermodynamics by increasing the height of the free energy barrier. Termini interactions are thus responsible for the thermodynamic cooperativity of the folding transition.

A physical rationale for the reported observation is the following. The tails of a polypeptide chain are more prone to gain entropic stabilization by becoming unfolded than fragments of equivalent length located in the middle of the chain[68]. Therefore, in order to overcome the conformational entropy loss associated with the folding transition the chain ends must necessarily establish a sufficient amount of stabilizing interactions. Since many single domain proteins display their termini at the surface[28], it appears just natural that the establishment of those stabilizing interactions will occur precisely between the two chain ends that are brought into contact in the native structure[28]. Therefore, by virtue of their native location, terminal elements of secondary structure and the interactions they establish must necessarily play a leading and *direct* role in setting up the free energy barrier between native and unfolded states in small, single domain proteins. We placed an emphasis on the word direct because there are other physical mechanisms underpinning protein folding cooperativity that result from solvent mediated effects such as desolvation barriers in hydrophobic association[69].

Despite the striking importance of termini coupling for protein folding, one cannot neglect the interactions that termini residues establish with other parts of the native structure. Indeed, as indicated by the results reported for chymotrypsin inhibitor 2, it is likely that in some native structures protein termini are stabilized through the independent establishment of native interactions with other parts of the chain, instead of (or in addition to) with each other. This situation is likely less frequent in single domain proteins[28], because it implies having two independent regions of highly stabilizing native interactions instead of just one, which, at least intuitively, represents a structural constraint more difficult to fulfill.

From a biological standpoint, a cooperative two-state folding process is also justifiable. Indeed, the reasons why such a transition is biologically advantageous is, at least, three-fold: 1) because it eliminates from the folding space intermediate states that may be prone to aggregate; 2) because it provides an enhancement of kinetic stability against transformation into non-functional forms; and 3) because it modulates potentially complicated effects of co-translational folding[70]. It is therefore likely that interacting termini is an evolutionary selected trait achieved through the design of proteins with aligned terminal regions. The results we now report are in line with this view.

Chan and co-workers previously proposed that protein folding cooperativity results from non-trivial energetics based on multi-body interactions[9]. In particular, a mechanism of local-nonlocal coupling whereby the establishment of local interactions is greatly enhanced by the formation of non-local native interactions, was shown to significantly enhance protein folding cooperativity and the CO-rate correlation in coarse grained models[9, 11]. Although the atomistic origins of many-body interactions are not well understood, desolvation effects and side-chain packing are likely contributors[71]. The results reported here add termini coupling to the list of already identified physical ingredients responsible for protein folding cooperativity. Interestingly, since termini coupling involves the establishment of the most long-ranged interactions in the native structure it may be seen as an extreme case of the local-non-local coupling mechanism in that the establishment of termini interactions favors the formation of the remainder of the native fold through a 'clean' folding transition where a large free energy barrier separates native and denatured states.

Although we have not reported folding kinetics in the off-lattice simulations the results we obtained in the scope of lattice models indicate that despite contributing with up to 10% to the overall CO, termini interactions are important determinants of the folding rate, being able to reverse the trend driven by the full set of native interactions in proteins of the same size.

The results reported in this work are based on native-centric models. Despite recent evidence, based on atomistic simulations, supporting the view that the protein folding transition is driven by native interactions with non-native interactions playing no significant part in determining the folding mechanism[72] one cannot rule out the possibility that non-native interactions may play a role in the enthalpic stabilization of protein termini in transient conformations en-route to the native state. Therefore, it would be interesting to investigate termini energetics in the context of more realistic off-lattice protein models with sequence-specific protein energetics, and eventually test our results in experiments *in vitro* with real proteins.

Acknowledgments

PFNF and HK thank Fundação para a Ciência e a Tecnologia for financial support through grants PTDC/FIS/113638/2009 (to PFNF) and Pest-OE/FIS/UI0261/2014.

AR acknowledges partial financial support from UCM-Santander Grant GR3/14. PFNF and AR designed the research, HK and AR performed the calculations, all authors analyzed the data and PFNF and AR wrote the paper. The authors would like to thank HS Chan for his critical reading of the manuscript and suggestions to improve it.

References

1. K. W. Plaxco, K. T. Simons and D. Baker, *J. Mol. Biol.*, 1998, **277**, 985-994.
2. K. W. Plaxco, K. T. Simons, I. Ruczinski and B. David, *Biochemistry*, 2000, **39**, 11177-11183.
3. M. M. Gromiha and S. Selvaraj, *J. Mol. Biol.*, 2001, **310**, 27-32.
4. C. Micheletti, *Proteins: Structure, Function, and Bioinformatics*, 2003, **51**, 74-84.
5. C. B. Anfinsen, *Science*, 1973, **181**, 223-230.
6. C. Clementi, *Curr. Opin. Struct. Biol.*, 2008, **18**, 10-15.
7. R. D. Hills and C. L. Brooks, *Int. J. Mol. Sci.*, 2009, **10**, 889-905.
8. P. F. N. Faisca and R. C. Ball, *J. Chem. Phys.*, 2002, **117**, 8587-8591.
9. H. S. Chan, S. Shimizu and H. Kaya, in *Energetics of Biological Macromolecules, Pt E*, eds. J. M. Holt, M. L. Johnson and G. K. Ackers, Elsevier Academic Press Inc, San Diego, 2004, vol. 380, pp. 350-379.
10. A. I. Jewett, V. S. Pande and K. W. Plaxco, *J. Mol. Biol.*, 2003, **326**, 247-253.
11. H. Kaya and H. S. Chan, *Proteins*, 2003, **52**, 524-533.
12. P. F. N. Faísca, R. D. M. Travasso, A. Parisi and A. Rey, *PLoS One*, 2012, **7**, e35599.
13. D. Barrick, *Physical Biology*, 2009, **6**,015001.
14. K. Lindorff-Larsen, M. Vendruscolo, E. Paci and C. M. Dobson, *Nat. Struct. Mol. Biol.*, 2004, **11**, 443-449.
15. A. R. Fersht, *Proc. Natl. Acad. Sci. U. S. A.*, 2000, **97**, 1525-1529.
16. E. Paci, K. Lindorff-Larsen, C. M. Dobson, M. Karplus and M. Vendruscolo, *J. Mol. Biol.*, 2005, **352**, 495-500.
17. D. E. Makarov and K. W. Plaxco, *Protein Sci.*, 2003, **12**, 17-26.
18. F. Chiti, N. Taddei, P. M. White, M. Bucciantini, F. Magherini, M. Stefani and C. M. Dobson, *Nat. Struct. Biol.*, 1999, **6**, 1005-1009.
19. D. S. Riddle, V. P. Grantcharova, J. V. Santiago, E. Alm, I. Ruczinski and D. Baker, *Nat. Struct. Biol.*, 1999, **6**, 1016-1024.
20. P. F. N. Faisca and M. M. T. da Gama, *Biophys. Chem.*, 2005, **115**, 169-175.
21. R. D. M. Travasso, P. F. N. Faisca and M. M. Telo da Gama, *J. Phys.-Condes. Matter*, 2007, **19**, 285212.
22. D. N. Ivankov, S. O. Garbuzynskiy, E. Alm, K. W. Plaxco, D. Baker and A. V. Finkelstein, *Protein Sci.*, 2003, **12**, 2057-2062.
23. O. V. Galzitskaya, S. O. Garbuzynskiy, D. N. Ivankov and A. V. Finkelstein, *Proteins*, 2003, **51**, 162-166.
24. A. N. Naganathan and V. Muñoz, *J. Am. Chem. Soc.*, 2004, **127**, 480-481.
25. D. De Sancho, U. Doshi and V. Munoz, *J. Am. Chem. Soc.*, 2009, **131**, 2074-2075.

26. D. De Sancho and V. Munoz, *Phys. Chem. Chem. Phys.*, 2011, **13**, 17030-17043.
27. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne, *Nucleic Acids Res.*, 2000, **28**, 235-242.
28. M. M. G. Krishna and S. W. Englander, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 1053-1058.
29. H. Krobath, E. I. Shakhnovich and P. F. N. Faisca, *J. Chem. Phys.*, 2013, **138**, 215101.
30. H. Krobath, S. G. Estácio, P. F. N. Faísca and E. I. Shakhnovich, *J. Mol. Biol.*, 2012, **422**, 705-722.
31. S. G. Estácio, H. Krobath, D. Vila-Viçosa, M. Machuqueiro, E. I. Shakhnovich and P. F. N. Faísca, *PLoS Comput Biol*, 2014, **10**, e1003606.
32. S. G. Estacio, C. S. Fernandes, H. Krobath, P. F. N. Faisca and E. I. Shakhnovich,*J. Chem. Phys.*, 2012, **137**, 085102.
33. L. Prieto, D. de Sancho and A. Rey, *J. Chem. Phys.*, 2005, **123**, 154903.
34. N. Go and H. Abe, *Biopolymers*, 1981, **20**, 991-1011.
35. M. A. Soler and P. F. N. Faísca, *PLoS One*, 2012, **7**, e52343.
36. K. A. Dill, S. Bromberg, K. Z. Yue, K. M. Fiebig, D. P. Yee, P. D. Thomas and H. S. Chan, *Protein Sci.*, 1995, **4**, 561-602.
37. H. S. Chan and K. A. Dill, *Proteins*, 1998, **30**, 2-33.
38. A. Sali, E. Shakhnovich and M. Karplus, *Nature*, 1994, **369**, 248-251.
39. A. M. Gutin, V. I. Abkevich and E. I. Shakhnovich, *Phys. Rev. Lett.*, 1996, **77**, 5433-5436.
40. P. F. N. Faisca, A. Nunes, R. D. M. Travasso and E. I. Shakhnovich, *Protein Sci.*, 2010, **19**, 2196-2209.
41. H. Krobath and P. F. N. Faísca, *Phys. Biol.*, 2013, **10**, 016002.
42. M. A. Soler and P. F. N. Faísca, *PLoS One*, 2013, **8**, e74755.
43. R. Perezzan and A. Rey, *J. Chem. Phys.*, 2012, **137**,. 185102.
44. M. Larriva, L. Prieto, P. Bruscolini and A. Rey, *Proteins*, 2010, **78**, 73-82.
45. L. Prieto and A. Rey, *J. Chem. Phys.*, 2007, **127**,175101.
46. P. F. N. Faisca and K. W. Plaxco, *Protein Sci.*, 2006, **15**, 1608-1618.
47. M. A. Soler, A. Nunes and P. F. N. Faísca,*J. Chem. Phys.*, 2014, **141**, 025101.
48. S. Abeln, M. Vendruscolo, C. M. Dobson and D. Frenkel, *PLoS One*, 2014, **9**, e85185.
49. M. S. Li, N. T. Co, G. Reddy, C. K. Hu, J. E. Straub and D. Thirumalai, *Physical Review Letters*, 2010, **105**, 218101.
50. C. Holzgrafe, A. Irback and C. Troein, *J. Chem. Phys.*, 2011, **135**. 195101.
51. A. Irback, S. A. Jonsson, N. Linnemann, B. Linse and S. Wallin, *Phys. Rev. Lett.*, 2013, **110**, 058101.
52. M. Enciso and A. Rey, *J. Chem. Phys.*, 2013, **139**, 115101.
53. M. Enciso and A. Rey, *J. Chem. Phys.*, 2012, **136**,. 215103
54. L. Prieto and A. Rey, *J. Chem. Phys.*, 2009, **130**, 115101.
55. V. I. Abkevich, A. M. Gutin and E. I. Shakhnovich, *Biochemistry*, 1994, **33**, 10026-10036.
56. D. K. Klimov and D. Thirumalai, *J. Mol. Biol.*, 1998, **282**, 471-492.
57. P. F. N. Faisca, *J. Phys.-Condes. Matter*, 2009, **21**,. 373102.
58. L. S. Itzhaki, D. E. Otzen and A. R. Fersht, *J. Mol. Biol.*, 1995, **254**, 260-288.
59. H. Taketomi, Y. Ueda and N. Go, *Int. J. Pept. Protein Res.*, 1975, **7**, 445-459.

60. N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller, *J. Chem. Phys.*, 1953, **21**, 1087-1092.
61. P. F. N. Faisca and R. C. Ball, *J. Chem. Phys.*, 2002, **116**, 7231-7237.
62. J. D. Chodera, W. C. Swope, J. W. Pitera, C. Seok and K. A. Dill, *J. Chem. Theory Comput.*, 2007, **3**, 26-41.
63. A. Gronenborn, D. Filpula, N. Essig, A. Achari, M. Whitlow, P. Wingfield and G. Clore, *Science*, 1991, **253**, 657-661.
64. A. Musacchio, M. Noble, R. Pauptit, R. Wierenga and M. Saraste, *Nature*, 1992, **359**, 851-855.
65. C. A. McPhalen and M. N. G. James, *Biochemistry*, 1987, **26**, 261-269.
66. V. P. Grantcharova and D. Baker, *J. Mol. Biol.*, 2001, **306**, 555-563.
67. F. U. Hartl and M. Hayer-Hartl, *Science*, 2002, **295**, 1852-1858.
68. T. R. Weikl and K. A. Dill, *J. Mol. Biol.*, 2003, **329**, 585-598.
69. Z. Liu and H. S. Chan, *J. Mol. Biol.*, 2005, **349**, 872-889.
70. T. Chen and H. S. Chan, *Phys. Chem. Chem. Phys.*, 2014, **16**, 6460-6479.
71. H. S. Chan, Z. Q. Zhang, S. Wallin and Z. R. Liu, *Annu. Rev. Phys.l Chem.*, 2003, **62**, 301-326.
72. R. B. Best, G. Hummer and W. A. Eaton, *Proc. Natl. Acad. Sci. U. S. A.*, 2013, **110**, 17874-17879.
73. W. Humphrey, A. Dalke and K. Schulten, *J. Mol. Graph.*, 1996, **14**, 33-38.

Figures

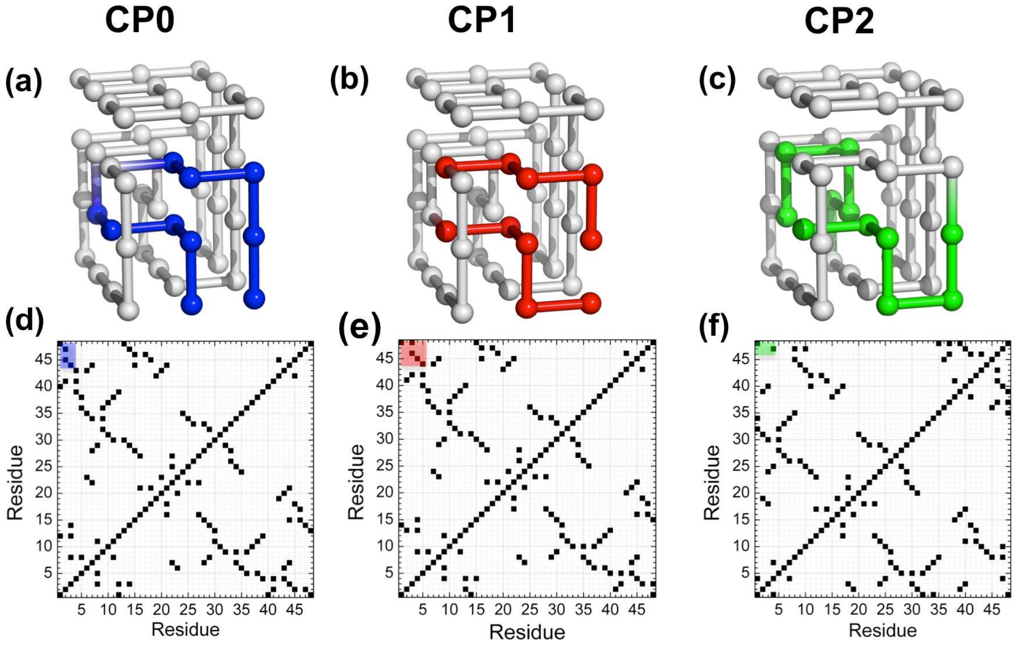

Figure 1. Lattice model systems. Three dimensional native structures (a-c) and native contact maps (d-f) of the lattice model systems CP0, CP1 and CP2 investigated in this study. In the native structures the chain termini are highlighted. The native interactions established by the termini residues, which we term NC, are also highlighted in the contact maps.

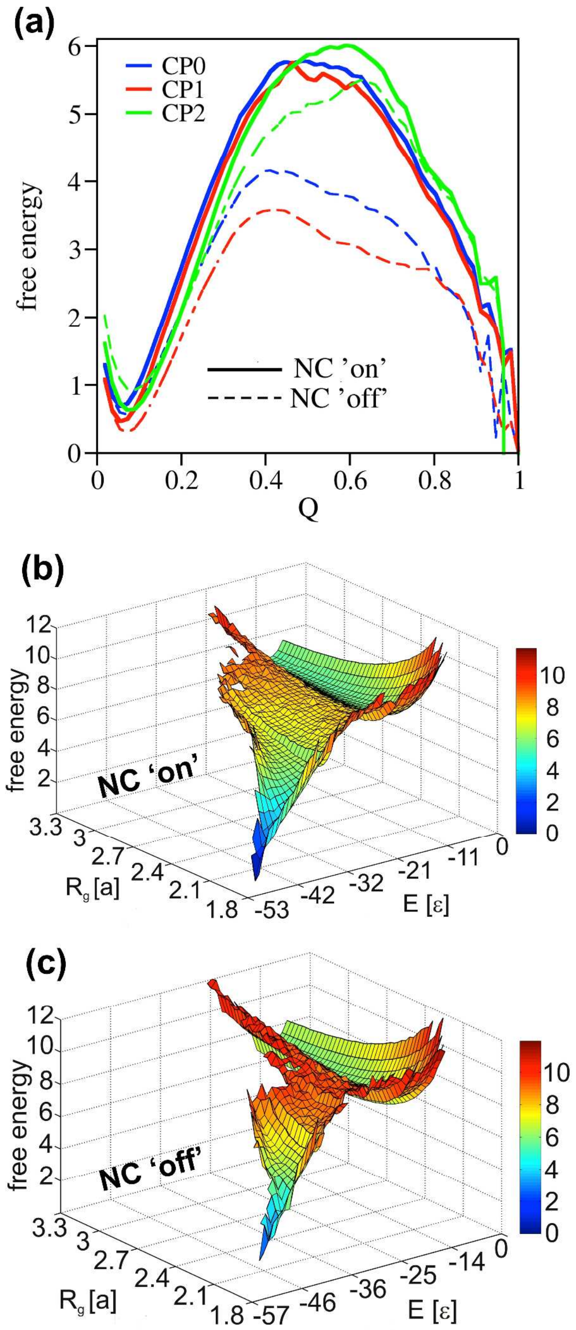

Figure 2. Effect of termini interactions on the thermodynamics of the folding transition on lattice. (a) Free energy profile showing the projection of the free energy on reaction coordinate fraction of native contacts, *Q*, for model system CP0, CP1 and CP2, when the NC interactions are "switched-on", or "switched-off". Free energy surface (i.e. the projection of the free energy on reaction coordinate energy, *E*, and radius of gyration, R_g) for model system CP0, when the NC interactions are 'switched-on' (b) and 'switched-off' (c).

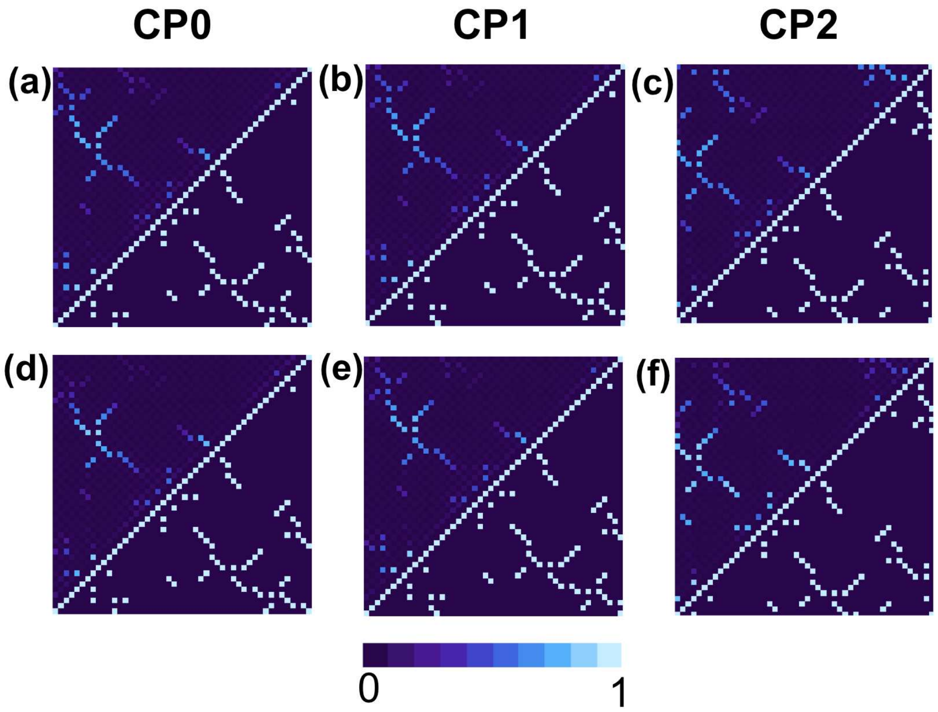

Figure 3. Transition state structure in the lattice model. Probability maps showing the likelihood of formation of each native contact in the transition state ensemble (i.e. ensemble of 2000 conformations with $0.4 < Q < 0.6$) when the NC interactions are 'switched-on' (a-c) and 'switched-off' (d-f). When the termini interactions contribute to stabilizing the native structure they form in the transition state with non-negligible probability in CP0 ($p_{(1\text{-}48)} = 0.26$, $p_{(2\text{-}47)} = 0.31$, $p_{(2\text{-}45)} = 0.44$, $p_{(3\text{-}44)} = 0.46$)), CP1 ($p_{(1\text{-}48)} = 0.24$, $p_{(2\text{-}48)} = 0.30$, $p_{(3\text{-}46)} = 0.42$, $p_{(4\text{-}45)} = 0.47$, $p_{(5\text{-}44)} = 0.53$) and CP2 ($p_{(1\text{-}48)} = 0.48$, $p_{(4\text{-}45)} = 0.56$).

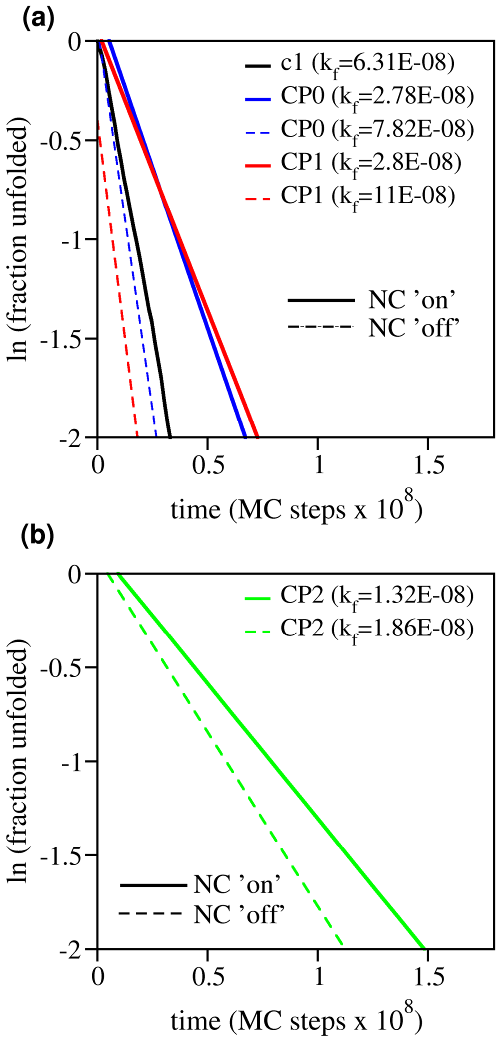

Figure 4. Effect of termini interactions on folding rate in the lattice model. Evaluation of the folding rate at T_m for model systems CP0 (a), CP1 (a) and CP2 (b) when the NC interactions are “switched-on” or “switched-off”. The black line in panel (a) corresponds to control system c1, which is the circular permutant of low contact order.

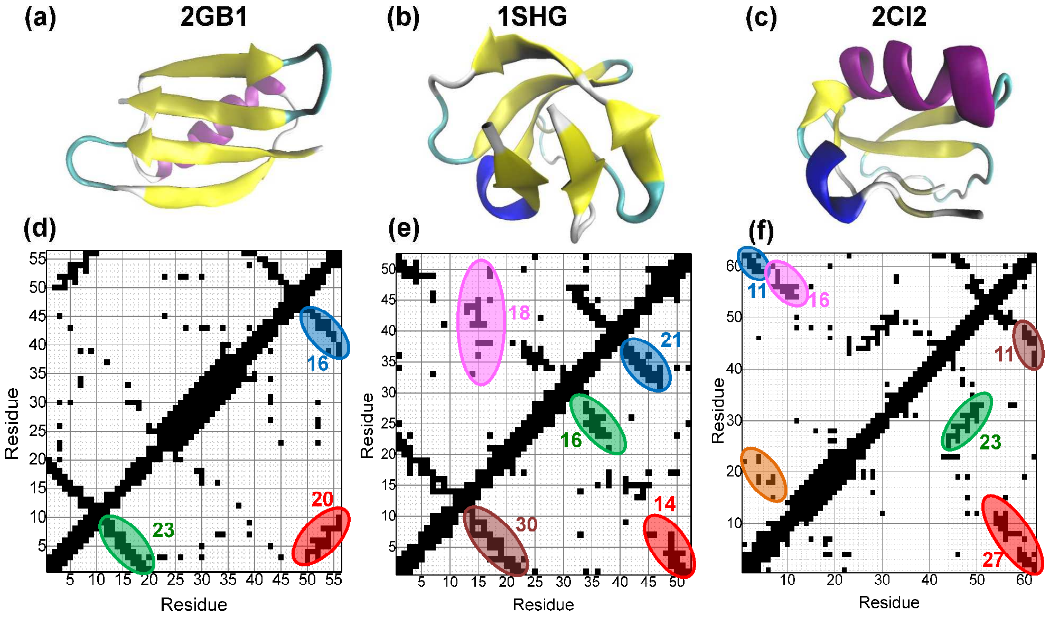

Figure 5. Proteins studied off-lattice. Ribbon diagrams (a-c) and native contact maps (d-f) of proteins with PDB codes 2GB1, 1SHG (residues 9 to 60) and 2CI2. The colored ellipses in the contact maps indicate the sets of interactions that are "switched-off" in the different simulated systems. The nearby numbers indicate the number of native contacts involved in each case. The three-dimensional structures were prepared with VMD[73].

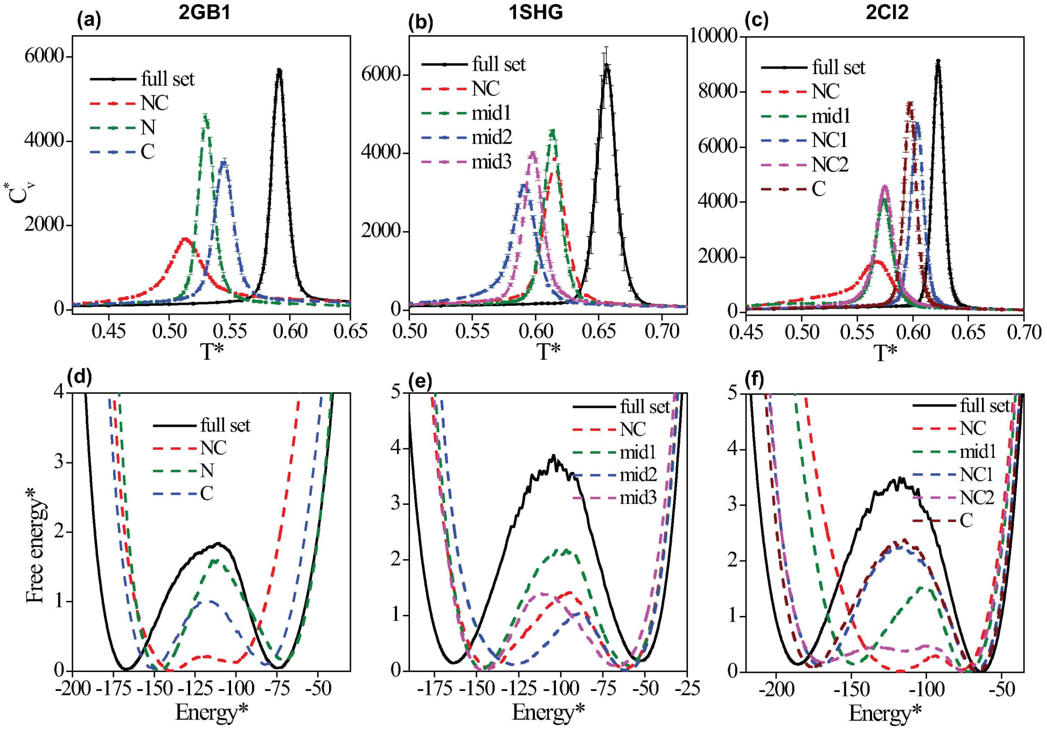

Figure 6. Effect of termini interactions on the thermodynamics of the folding transition off lattice. Heat capacity curves (a-c) and free energy profiles (in reduced units) (d-f) for the folding transition of the three proteins investigated in the scope of off-lattice simulations. The black curves correspond to the simulations where protein energetics is driven by the full set of native interactions. The colored curves report folding thermodynamics exhibited by the three proteins when specific sets of native interactions were “switched-off”, including the NC interactions at the chain termini and several control interactions. The color code is that adopted in Figure 5.