

Pathways to folding, nucleation events, and native geometry

Rui D. M. Travasso^{a)}

Centro de Física Teórica e Computacional, Faculdade de Ciências, Universidade de Lisboa, Avenida Professor Gama Pinto 2, 1649-003 Lisboa, Portugal

Margarida M. Telo da Gama^{b)}

Centro de Física Teórica e Computacional, Faculdade de Ciências, Universidade de Lisboa, Avenida Professor Gama Pinto 2, 1649-003 Lisboa, Portugal and Departamento de Física, Faculdade de Ciências, Universidade de Lisboa, Campo Grande, Edifício C8, 1749-016 Lisboa, Portugal

Patrícia F. N. Faísca^{c)}

Instituto de Tecnologia Química e Biológica, Universidade Nova de Lisboa, Avenida da República, EAN 2785-572 Oeiras, Portugal

(Received 6 July 2007; accepted 6 August 2007; published online 12 October 2007)

We perform extensive Monte Carlo simulations of a lattice model and the Gō potential [N. Gō and H. Taketomi, Proc. Natl. Acad. Sci. U.S.A. **75**, 559563 (1978)] to investigate the existence of folding pathways at the level of contact cluster formation for two native structures with markedly different geometries. Our analysis of folding pathways revealed a common underlying folding mechanism, based on nucleation phenomena, for both protein models. However, folding to the more complex geometry (i.e., that with more nonlocal contacts) is driven by a folding nucleus whose geometric traits more closely resemble those of the native fold. For this geometry folding is clearly a more cooperative process. © 2007 American Institute of Physics.

[DOI: [10.1063/1.2777150](https://doi.org/10.1063/1.2777150)]

I. INTRODUCTION

Protein folding is the process by which a linear chain of amino acids *spontaneously* acquires a specific three dimensional native structure.¹ As pointed out by Levinthal in the late 1960s, a random search of the conformational space for the global minimum of the free energy (i.e., for the unique native fold) is not compatible with the biological time frame of folding.² This raised the hypothesis that folding might have to occur through an ordered sequence of events (i.e., an ordered sequence of conformational changes) for the protein to reach rapidly its native conformation when starting from the unfolded state. In other words, kinetic pathways of folding, comprising or not specific intermediates, were envisaged to explain the time scale of protein folding.^{2,3}

The discovery in the early 1990s that the 64-residue protein chymotrypsin inhibitor 2 (CI2) folds rapidly with single-exponential (two-state) kinetics⁴ showed that the existence of discrete folding intermediates is not a prerequisite to fold fast. Indeed, the vast majority of small (with less than 120 amino acids), single domain proteins are, like CI2, rapid two-state folders.⁵ Another “simplifying” feature of small proteins is their topology-dependent folding kinetics; the contact order,⁶ (CO) measuring the average sequence separation of contacting residues in the native fold, and other related metrics of native geometry^{7,8} are strongly correlated with folding rates, suggesting that native topology plays a key role in determining the folding mechanism.

A protein engineering method termed ϕ -value analysis⁹ revealed that CI2 folds via nucleation condensation (NC),¹⁰ a mechanism that was first observed by Abkevich *et al.* in the context of simulations of lattice proteins.¹¹ In the NC mechanism the formation of a small set of local native bonds, stabilized by a few nonlocal native interactions, the so-called folding nucleus, triggers the rapid emergence of the native fold. Subsequent studies suggested that NC is possibly the most common folding mechanism amongst single domain proteins.¹²

The problem of identifying folding pathways along with the formation of folding nuclei is therefore of the utmost importance in protein chemistry and has been investigated within different frameworks.^{15,22} Computer simulations of protein folding and unfolding, both on- and off-lattice, have proven particularly useful in exploring protein folding pathways and mechanisms at different levels of structural detail.^{13–26} For example, at the microstructural level of contact formation, it was shown that folding is dominated by a well-defined sequence of events¹⁵ and that the sequencing of events depends primarily on the native geometry as defined by the CO.¹⁴ On the other hand, a more recent study revealed that the unfolding process of CI2 happens in a highly parallel fashion.²⁶ At a coarser level of structure defined by contact clusters (i.e., secondary structure elements), sequential folding events have been reported within different simulational frameworks.^{14,17,20,21,26}

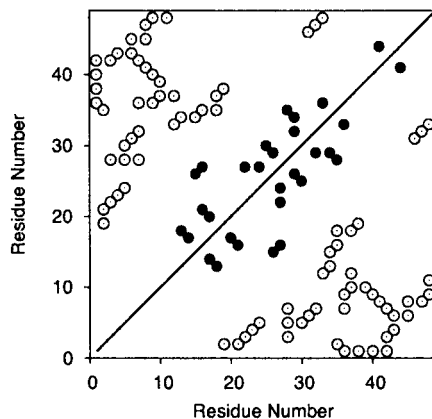
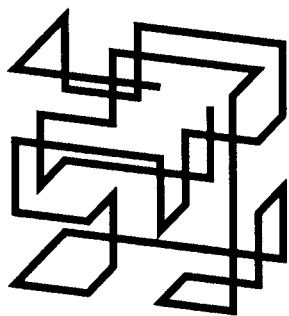
Here we use a lattice model and the Gō potential to explore in some detail the folding pathways leading to different native geometries. In particular, we determine the order according to which different sections of the native fold become structured as folding progresses toward the native

^{a)}Electronic mail: rui@cii.fc.ul.pt

^{b)}Electronic mail: margadrid@cii.fc.ul.pt

^{c)}Author to whom correspondence should be addressed. Electronic mail: patnev@cii.fc.ul.pt

Geometry 1



Geometry 2

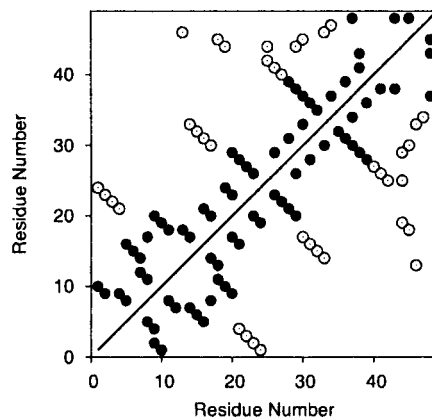
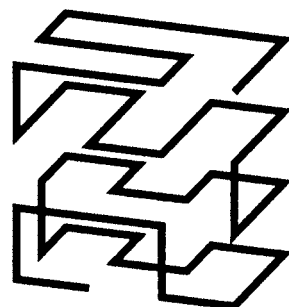


FIG. 1. Three dimensional representations of geometry 1 (top, left) and geometry 2 (bottom, left) and their respective contact maps (right). In the contact maps each circle represents a native contact. Nonlocal LR contacts are shown in white.

state. For both geometries there is one section that exhibits a distinctively different folding pattern. Moreover, the timely formation of this particular section determines the most probable folding pathways. By comparison with previous studies, based on specific strategies to identify the folding nucleus, we have confirmed that this unique section, identified through the analysis of the folding pathways, does contain the critical contacts forming the folding nucleus.

This article is organized in the following way. In the next section we describe the model and simulational methods employed, then we present and discuss the results of the simulations, and finally we draw some concluding remarks.

II. MODEL AND METHODS

A. $G_{\bar{0}}$ model and simulation details

We consider a simple three dimensional lattice model of a protein molecule with chain length $N=48$. 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 $G_{\bar{0}}$ model.²⁷ In

the $G_{\bar{0}}$ model the energy of a conformation, defined by the set of bead coordinates $\{\mathbf{r}_i\}$, is given by the contact Hamiltonian,

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

where the contact function $\Delta(\mathbf{r}_i - \mathbf{r}_j)$ is unity only if beads i and j form a noncovalent native contact, i.e., a contact between a pair of beads that is present in the native structure and is zero otherwise. The $G_{\bar{0}}$ potential is based on the idea that the native fold is very well optimized energetically. Accordingly, it ascribes equal stabilizing energies (e.g., $\epsilon = -1.0$) to all the native contacts and neutral energies ($\epsilon = 0$) to all non-native contacts. As the $G_{\bar{0}}$ model has a uniform distribution of contact energies, the folding dynamics driven by the $G_{\bar{0}}$ potential is essentially determined by the structural features of the native fold.

In order to mimic the protein's relaxation toward the native state, we use a Metropolis Monte Carlo (MC) algorithm^{28,30,31} together with the kink-jump move set.²⁹ To guarantee that the detailed balance condition is satisfied, the probability of a certain conformational change must be independent of the conformation adopted by the chain.^{30,31} Therefore, at each MC step, the probability of applying the

TABLE I. Absolute contact order (ACO), fraction of long-range (LR) contacts, optimal folding temperature T_{opt} , and folding time for geometries 1 and 2.

Geometry	ACO	Fraction LR	T_{opt}	Folding time ($\times 10^6$ MCS)
1	21.4	0.74	0.65	8.1 ± 0.5
2	10.0	0.33	0.66	2.3 ± 0.1

Metropolis criteria to a particular chain displacement is $0.2/(N+6)$ if the displacement involves moving one single bead or $0.8/(N-3)$ if it involves the simultaneous movement of two beads. A MC simulation starts from a randomly generated unfolded conformation and the folding dynamics is monitored by following the evolution of the fraction of native contacts, $Q=q/L$, where L is the number of contacts in the native fold and q is the number of native contacts at each MC step. 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 mean first passage time of 300 simulations. Folding is studied at the so-called optimal folding temperature T_{opt} , the temperature that minimizes the folding time.^{32–35}

B. Native geometries

In order to explore how native geometry alone drives the folding process, two native folds (Fig. 1), which are amongst the most complex (geometry 1) and the simplest (geometry 2) cuboid geometries found through lattice simulations of homopolymer relaxation,³⁶ were considered in this study.

For structures that like ours are maximally compact cuboids with $N=48$ residues, there are 57 native contacts. A nonlocal contact between two residues i and j is considered long range (LR) if its sequence separation is at least 12 units, i.e., $|i-j| \geq 12$.⁷ Geometry 1 is characterized by a large number of LR contacts, while in geometry 2 the native bonds are predominantly local (Fig. 1 and Table I). The larger number of LR contacts in geometry 1 translates into a large absolute contact order (ACO).⁶

C. Probability to fold, P_{fold}

The folding probability $P_{\text{fold}}(\Gamma)$ for a conformation Γ is defined as the fraction of MC runs which, starting from Γ , fold before they unfold.³⁷ To compute P_{fold} we use an ensemble of 500 MC runs divided into bins of 100 runs. P_{fold} is firstly computed for each bin and the values thus found are subsequently averaged and the respective standard deviation evaluated. Each MC run stops when either the native fold or some unfolded conformation is reached. A conformation is deemed unfolded when its total fraction of native contacts, Q , is smaller than some cutoff, Q_U . In order to estimate Q_U , we compute the probability of finding some fraction of native contacts Q as a function of Q in 200 MC folding runs (Fig. 2). Considerably small Q must necessarily identify unfolded (or denatured) conformations. Indeed, a high-probability peak, centered around the fraction of native contacts $Q_U=0.1$, is readily apparent in the graph reported for geometry 1 (Fig. 2, left). Similarly, the highest probability peak appears around $Q_U=0.2$ for geometry 2 (Fig. 2, right). These fractions of native contacts are relatively low and therefore identify states with minimal residual structure. In this work we use these values of Q to define each geometry's cutoff value Q_U .

III. EXPLORING THE HIDDEN “ARCHITECTURE” WITHIN A LATTICE PROTEIN

In real globular proteins native contacts are *clustered* into the so-called secondary structure elements (α -helices, β -sheets, etc.), which have no direct analog on the lattice. Therefore, in this coarse-grained representation, there are no well-defined clusters of contacts associated with the secondary structural elements. Nevertheless, it is possible to identify well-defined clusters of contacts in lattice proteins that form well-defined sections of the native fold. We have developed a method (based on interresidue contact correlation analysis) that groups native contacts into distinct protein sections according to whether their presence is strongly correlated.

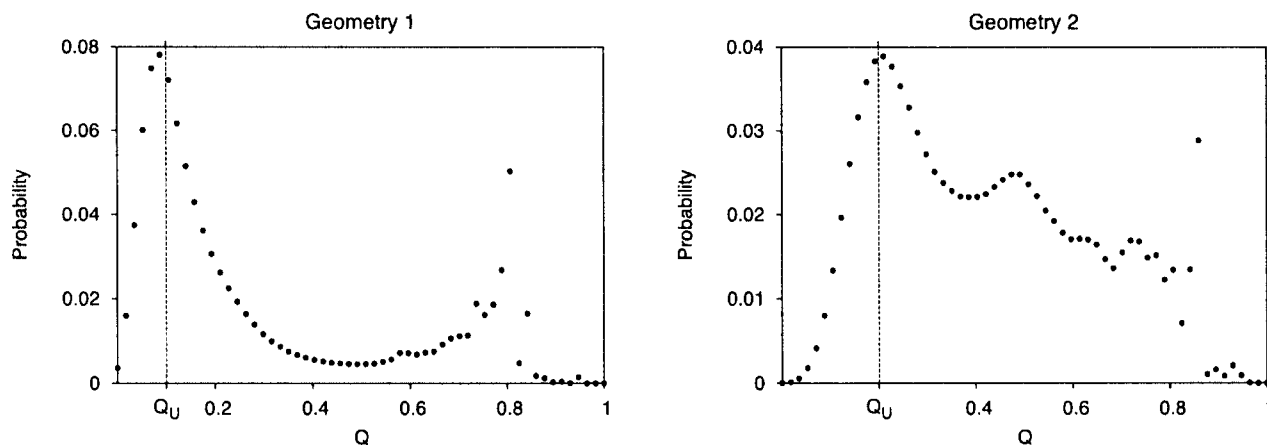


FIG. 2. Probability distribution for the fraction of native contacts, Q , for geometry 1 (left) and geometry 2 (right) as a function of Q . A conformation is considered unfolded when $Q < Q_U$.

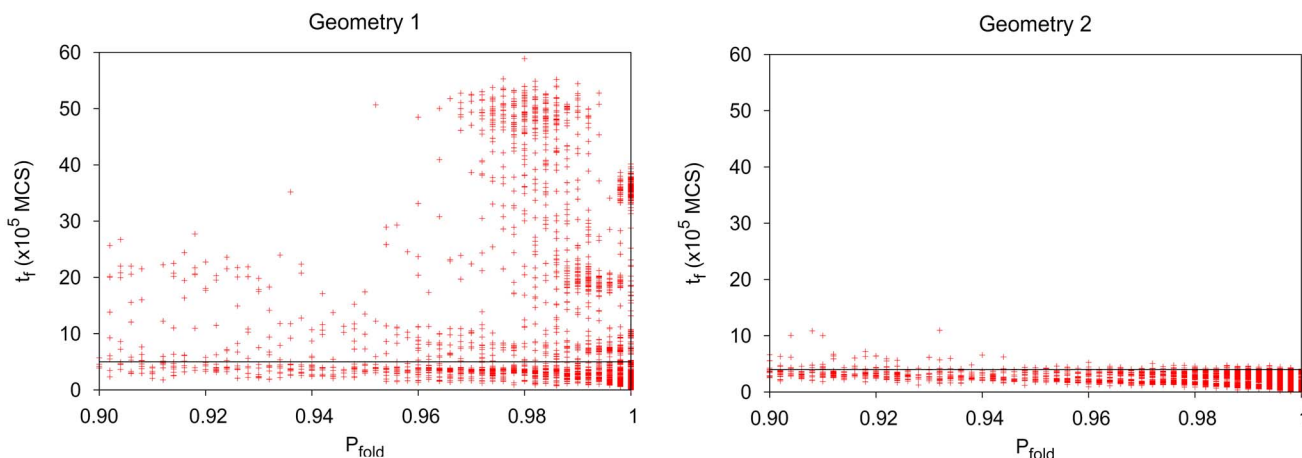


FIG. 3. (Color online) Time to fold t_f as a function of the reaction coordinate P_{fold} . Long lived trapped states are observed in geometry 1 (left) at very high P_{fold} , but are absent in geometry 2 (right). To measure t_f of each conformation we considered only folding events in which the protein folded before unfolding. t_f is the mean time-to-fold averaged over these folding events. The horizontal black lines indicate the cutoff times below which a conformation is committed to fold. For geometries 1 and 2 there are, respectively, 4724 and 4162 conformers with $P_{\text{fold}} \geq 0.9$.

A. Target conformations

The first step in the proposed procedure is that of selecting an ensemble of appropriate target conformations. These must be considerably nativelike and, most importantly, committed to fold. In order to find such productive conformers, we ran 8000 MC simulations for each model geometry and sampled a conformation from each independent MC run when folding was near completion (i.e., at a time close to the run's FPT). Conformations thus selected are dynamically uncorrelated and provide a sample of statistically independent elements. For every conformation we have computed P_{fold} , along with its standard deviation. The time to fold, t_f , was then measured for conformations with $P_{\text{fold}} \geq 0.9$. To compute t_f we have only considered MC runs where the proteins fold before they unfold. *A priori* one would expect such high- P_{fold} conformations to be kinetically very close to the native state. However, for geometry 1, a plot of the dependence of t_f on the folding probability reveals the existence of many conformations with $P_{\text{fold}} > 0.98$ that find the native state in a time frame comparable with that observed in simulations starting from random coil-type conformations (Fig. 3, Table I).

These are trapped states (i.e., off pathway to folding) and are eliminated from the initial sample. Indeed, the vast majority (i.e., 73%) of high P_{fold} conformers find the native state in less than 6% of the folding time. These conformations are on the folding pathway and can be used to cluster native bonds and shed light into the existence of putative protein sections. The mean fraction of native contacts in these conformations is $\langle Q \rangle = 0.73$ and, on average, they differ by 10.31 native contacts.

For geometry 2, 95% of conformations with high $P_{\text{fold}} > 0.9$ rapidly find the native state in less than 17% of the folding time. On average they differ by 9.18 contacts and, like in geometry 1, their mean fraction of native contacts is $\langle Q \rangle = 0.73$.

B. Interresidue contact correlation analysis reveals distinct protein sections

In conformations with high P_{fold} that are committed to fold, one expects that protein sections, comprising groups of correlated native bonds, will be formed with considerably high probability. We say that two native contacts α and β are correlated, i.e., that they belong to the same section, if (i) they have similar probabilities of being present when an arbitrary third contact γ is not, and (ii) the probability of contact γ being present if contact α is not is similar to the probability of γ being present if contact β is absent. Formally, conditions (i) and (ii) may be quantified by correlation between α and β , $C_{\alpha\beta}$, defined as

$$C_{\alpha\beta} = \frac{\sum_{\gamma \neq \alpha, \beta} n_{\gamma} (p_{\gamma\alpha} - p_{\gamma\beta})^2 + (n_{\alpha} + n_{\beta})/2 \sum_{\gamma \neq \alpha, \beta} (p_{\alpha\gamma} - p_{\beta\gamma})^2}{(L-2)(n_{\alpha} + n_{\beta})/2 + \sum_{\gamma \neq \alpha, \beta} n_{\gamma}} \quad (2)$$

being $\ll 1$. In the expression above $p_{\alpha\gamma}$ is the conditional probability of finding contact γ if contact α is not present, n_{α} is the number of conformations in the sample where contact α is not present, and $L=57$ is the total number of native contacts. The error associated with $p_{\alpha\gamma}$ is of the order of $1/\sqrt{n_{\alpha}}$. Therefore the weight of each averaged term in Eq. (2) of either n_{γ} or $(n_{\alpha} + n_{\beta})/2$ implies that $C_{\alpha\beta}$ is determined by the terms which are measured with the highest accuracy (this is an important point since the measurement error associated with the difference between probabilities $p_{\gamma\alpha}$ and $p_{\gamma\beta}$ increases as n_{γ} decreases). Using Eq. (2) the correlation between pairs of native contacts α and β is computed in the ensembles of target conformations selected for geometries 1 and 2, and native contacts are ordered according to their relative correlations in the following way: starting with an arbitrary contact, say, contact 0, contact 1 is the one with the lowest C_{01} , i.e., the contact that is the most strongly correlated with contact 0, contact 2 is that with the lowest C_{12} , and so on. This ordering method sheds light on existing protein

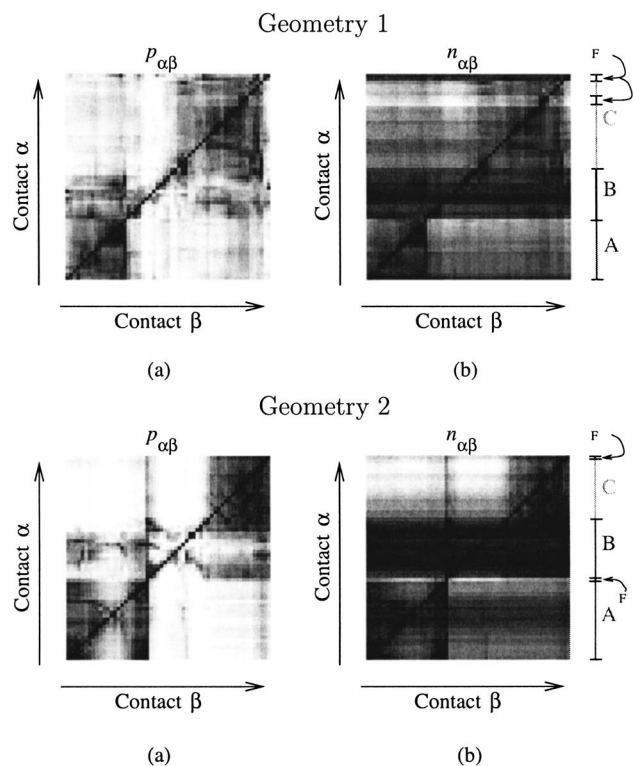


FIG. 4. Density plots of the probability (left column) and fraction of conformations (right column), where contact β is present and α is not for geometry 1 (top) and geometry 2 (bottom). Native contacts are ordered according to their relative values of $C_{\alpha\beta}$ (the order is the same for the $p_{\alpha\beta}$ and $n_{\alpha\beta}$ plots). The groups of contacts forming sections A, B, and C are identified. Contacts that were not assigned to any section ("free" contacts) are identified by the letter F. The range of $p_{\alpha\beta}$ lies between 0 (black) and 1 (white), while $n_{\alpha\beta}$ varies between 0 (black) and 0.54 (white) in geometry 1 and between 0 (black) and 0.64 (white) in geometry 2.

sections since it block diagonalizes the contact matrix C . Indeed, density plots for the probability that contact α is present if contact β is not, $p_{\alpha\beta}$, and for the fraction of conformations in the sample satisfying the same condition, $n_{\alpha\beta}$ reveals the existence of three protein sections, namely, sections A, B, and C in the two model proteins (Fig. 4).

In both geometries, contacts within sections A and C are strongly correlated. This is shown by the low probability (i.e., dark) squared spots located along the diagonal in the $p_{\alpha\beta}$ and in the $n_{\alpha\beta}$ ordered matrices. In the $p_{\alpha\beta}$ plot, such well-defined regions indicate that when a contact belonging to A (or C) is *not* formed, any other contact in A (or C) has a considerably low probability of being formed [Fig. 4(a)]. Correspondingly, the darker squares identifying sections A and C in the $n_{\alpha\beta}$ matrices show that for any pair of contacts within those sections, there is a small number of conformations in which one of the contacts in the pair is formed while the other is not [Fig. 4(b)].

In the $p_{\alpha\beta}$ and $n_{\alpha\beta}$ density plots the brighter spots located in the matrices' off diagonal indicate that contacts belonging to C (or A) can be formed with a relatively high probability, when a contact in A (or C) is missing. Hence, we conclude that the target conformations have either A or C formed.

Contacts in section B behave differently from those in sections A and C as they are *always* present with high prob-

ability. This is shown by the existence of the white vertical bar in the $p_{\alpha\beta}$ density plot [Fig. 4(a)] and the dark (and homogeneous) horizontal band that spans the vertical axis in the $n_{\alpha\beta}$ matrices [Fig. 4(b)]. The white spots on the diagonal in the $p_{\alpha\beta}$ matrices indicate that, by contrast to contacts in sections A and C, when one contact within B is missing, other contacts within B may still be formed with high probability.

Some contacts are located at the boundaries of the identified sections. The correlation between their presence and other contacts' presence does not fit the correlation patterns found for sections A, B, or C. For this reason we decide not to assign them to any section and denote them by *free* contacts. There are five free bonds in geometry 1 (namely, 4–23, 5–24, 12–33, 13–34, and 25–30) and two free bonds (2–9 and 13–46) in geometry 2.

C. Section's geometric traits

The protein sections thus identified as clusters of strongly correlated native bonds form well-defined, separate parts in the native fold (Fig. 5). Indeed, clusters of strongly correlated bonds are grouped together in the protein's three dimensional representation. The structural characterization of each individual section is reported in Table II.

In geometry 1 the three sections are geometrically different. In section A, all contacts but one are long range and link residues located in opposite ends of the chain. On the other hand, about 50% of the native bonds in section C are local. They connect residues in the middle of the chain (between residues 17 and 34). Contacts forming section B link residues located in the middle of the chain to residues located in either end of the chain. Interestingly, the geometric features of section B reflect those of the overall native fold. In geometry 2, on the other hand, the three sections are highly geometrically similar, being formed essentially by local bonds.

IV. FOLDING PATHWAYS

A folding pathway is an ordered sequence of events (i.e., of conformational changes) observed along the time coordinate. In this section we investigate if the previously identified protein sections become structured by following some preferential order, and how such ordering preferences depends on native geometry. In other words, we investigate the existence of folding pathways at the macrostructural level of section formation and how the latter depend on the native fold geometry. In order to do so, the fraction of native contacts in each section, Q_S , is monitored during each folding event. A section is considered folded from the time when its fraction of native contacts Q_S reaches 1.0 until it decreases below a certain threshold Q_S^U . Accordingly, the time at which a section folds is the smallest time t_S such that $Q_S=1.0$ at time t_S and $Q_S \geq Q_S^U$ at times larger than t_S .

A priori, the threshold Q_S^U could be section specific. However, different values of Q_S^U were tested for both proteins and the results reported hereafter are robust to changes in the exact value of this threshold. Therefore, and for the sake of simplicity, each section's Q_S^U was set equal to the

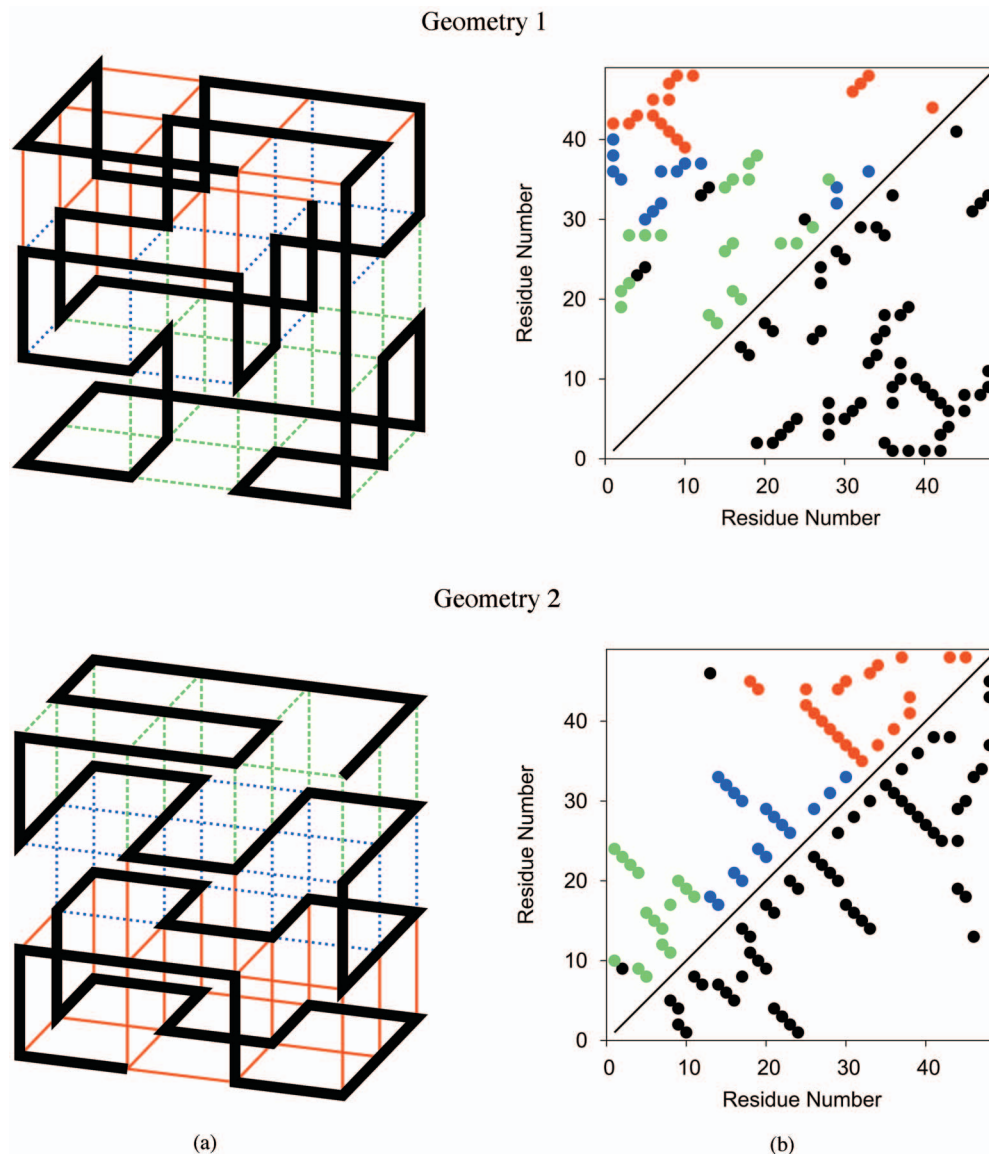


FIG. 5. (Color) Protein sections identified for geometry 1 (top row) and geometry 2 (bottom row). Native contacts forming sections A, B, and C are, respectively, colored red, blue, and green, in the three dimensional representations (left) and contact maps (right). Note that the protein sections identified as groups of correlated native bonds are grouped together in the protein's three dimensional native structure.

cutoff Q_U used previously to determine the protein's unfolded state.

We consider 5000 folding events. For each one the times t_S at which each section folds are recorded and the corresponding folding pathway is identified. The probability of observing specific pathways is then computed (Table III).

TABLE II. Number of native bonds forming each protein section, absolute contact order (ACO), and fraction of long-range (LR) contacts of each protein section.

Name	Number of contacts	Fraction LR	ACO	
Geometry 1	Section A	17	0.94	30.8
	Section B	14	0.79	24.1
	Section C	21	0.52	13.0
Geometry 2	Section A	22	0.45	10.9
	Section B	17	0.24	7.1
	Section C	16	0.25	10.5

Interestingly the most probable folding pathways are those in which section B is the second to fold. Structurally, this preference translates into folding starting either at the top or at the bottom of the native structure followed by the consolidation of the structure's middle "layer" (Fig. 5). The next most probable pathways are those where B folds first and, for both geometries, the probability that B folds last is vanishingly small. These observations suggest that in either case it is the folding of section B that determines the probability of a folding pathway. We disregarded the folding events in which two sections fold simultaneously (i.e., in the same MC step) as they result from the discretization of time and space imposed by the lattice.

For the most probable folding pathways, we measured the time elapsing between the formation of the first section and the emergence of the native structure. For both geometries the shorter time intervals are observed when section B folds first. However, these time intervals are systematically

TABLE III. Folding pathways at the macrostructural level of section formation (showing the first, second, and third sections to fold) and their relative probabilities of occurrence. The probabilities do not add to one, since there are some events in which two sections fold simultaneously. The average time elapsing between the formation of the first section and the formation of the last section in each pathway is given in units of 100 000 MCS.

Geometry 1					Geometry 2				
First	Second	Third	Prob.	Time	First	Second	Third	Prob.	Time
A	B	C	0.28	2.2 ± 0.3	A	B	C	0.40	4.0 ± 0.2
C	B	A	0.26	1.8 ± 0.2	C	B	A	0.31	5.2 ± 0.2
B	A	C	0.16	1.7 ± 0.1	B	A	C	0.13	3.4 ± 0.1
B	C	A	0.11	0.9 ± 0.2	B	C	A	0.12	2.7 ± 0.1
A	C	B	0.04		A	C	B	0.00	
C	A	B	0.00		C	A	B	0.00	

larger in the folding of geometry 2. Here, and once the first section is completely formed, the protein takes on average 25% of the folding time to achieve the native state if it follows the slowest pathway. For geometry 1 the equivalent time interval is just 2.5% of the overall folding time. This feature is particularly interesting because geometry 2 folds faster than geometry 1 (Table I).

V. SECTION FORMATION AS A FUNCTION OF THE FOLDING PROBABILITY

Here we analyze the folding progression of individual sections as a function of the probability to fold, P_{fold} . In other words, we investigate how the different sections of the protein become structured, i.e., how their fraction of native bonds, Q_S , evolves along the folding reaction. In order to do so, two ensembles, each comprising 8000 conformations,

were considered for each native geometry and the folding probability of each conformation evaluated (Sec. III A). (The standard deviation σP_{fold} was also measured. Hence, the probability for a conformation Γ to have some P_{fold} is considered to be given by the Gaussian distribution with average $P_{\text{fold}}(\Gamma)$ and standard deviation $\sigma_{P_{\text{fold}}(\Gamma)}$. These Gaussian distributions are used as weighting terms for calculating the probabilities of having a section with fraction of native bonds Q_S as a function of P_{fold} .) The probabilities of having a section with fraction of native bonds Q_S as a function of P_{fold} are shown as density plots in Fig. 6.

We start with the analysis of geometry 1. Here, section A is essentially unfolded for the most part of the folding reaction. Indeed, up to $P_{\text{fold}} \sim 0.8$, the most probable conformations are those with fraction of native bonds $Q_A \sim 0.1$, and it is only when folding is near completion that the probability

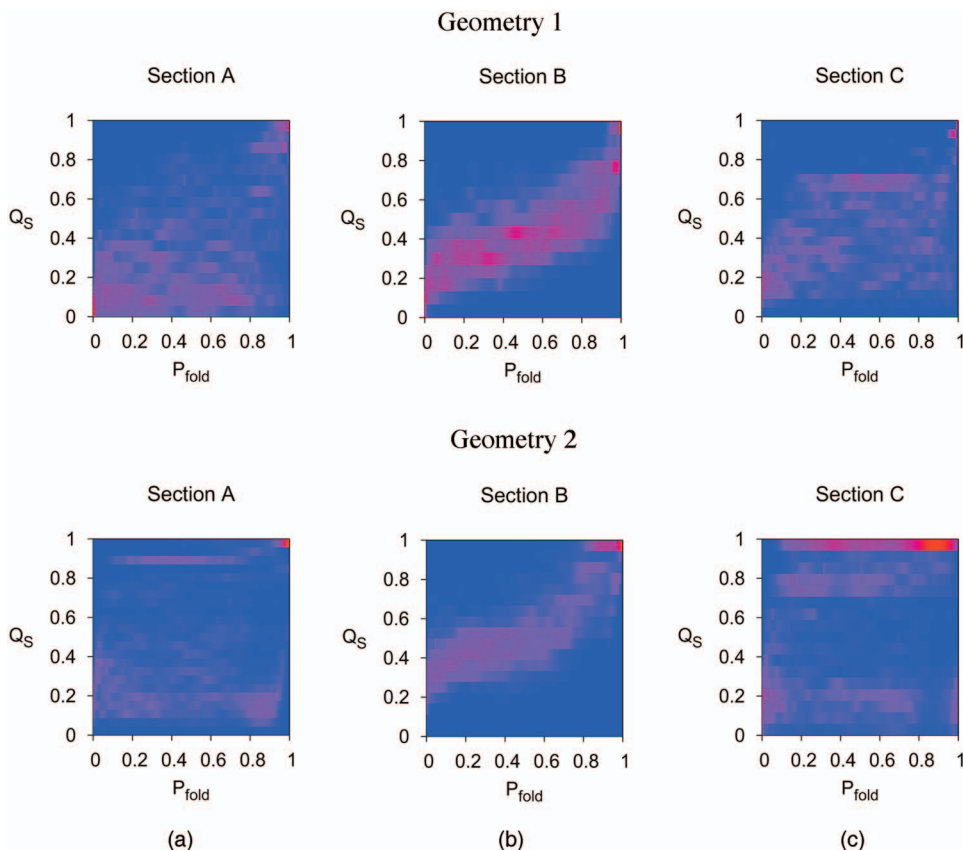


FIG. 6. (Color) Density plots of the probability for having a certain Q_S as a function of P_{fold} for the sections A, B, and C in geometry 1 (top) and geometry 2 (bottom).

to find A folded or close to folded (i.e., with $Q_A > 0.9$) is nonzero. Due to its local nature, bonds in section C can break and form more easily than in other sections where nonlocal bonds abound. It is perhaps for this reason that Q_C distributes rather uniformly in the range of $0.1 < Q_C < 0.75$ up to late folding stages (i.e., up to $P_{\text{fold}} \sim 0.8$). It is only when $P_{\text{fold}} > 0.9$ that there is a significant group of conformations with more than 90% of section C folded.

While there is no correspondence between the behavior of P_{fold} and that of the fraction of native contacts formed in A and C—in the sense that higher (lower) P_{fold} does not necessarily imply higher (lower) Q_S —for section B, on the other hand, at high P_{fold} , Q_B is on average high, while early on in folding (at low P_{fold}) section B is essentially unfolded. Therefore, an increase in P_{fold} typically leads to an increase in Q_B , suggesting that the folding of section B acts as a driver for the folding of the whole protein.

For geometry 2 the folding scenarios of sections A and C are rather distinct from those found in the more complex geometry 1. Indeed, for geometry 2, the probability of finding sections A and C with fraction of native bonds Q_S is strongly bimodal for any P_{fold} . This means that at any stage of the folding reaction, it is possible to find conformations with either A or C almost folded (peak at high Q_S) and others where A and C are very little structured (peak at low Q_S). This observation agrees with our previous findings regarding the most probable folding pathways, where folding initiates at A (and C folds last) or, conversely, it starts at C (and A folds last). However, as with geometry 1, the fraction of native bonds of section B increases with P_{fold} , and when it achieves some critical value, it becomes large enough to prompt folding of section A or C.

To gain further insight into the folding reactions of both model proteins, we have determined how the average fraction of native bonds in each section, $\langle Q_S \rangle$, changes with P_{fold} (Fig. 7).

The average fraction of native bonds in sections A and C decreases considerably when folding of geometry 2 is near completion [Fig. 7(b)], which is suggestive of existing unfolding events at large P_{fold} . This presumably happens due to a partial or complete folding of both sections A and C prior to the complete folding of section B. Such unfolding events, which are required to ensure that folding follows the right pathway, do not occur to such an extent in geometry 1 where A and B cannot fold at low P_{fold} [Fig. 7(a)], perhaps due to topological constraints. A comparison of the data reported on Figs. 7(a) and 7(b) with that shown in Fig. 7(c) indicates that the folding of the whole protein follows the folding of section B, in agreement with the idea that section B drives the folding of the whole native structure.

VI. FROM MACRO- TO MICROSTRUCTURAL FORMATION: EVIDENCE FOR NUCLEATION PHENOMENA

A postcritical folding nucleus (FN) is defined as a set of native bonds which, once formed, prompts rapid and highly probable folding.¹¹ We have recently developed a methodology, based on the concept of folding probability, aimed at identifying critical (i.e., nucleating) bonds in the folding of

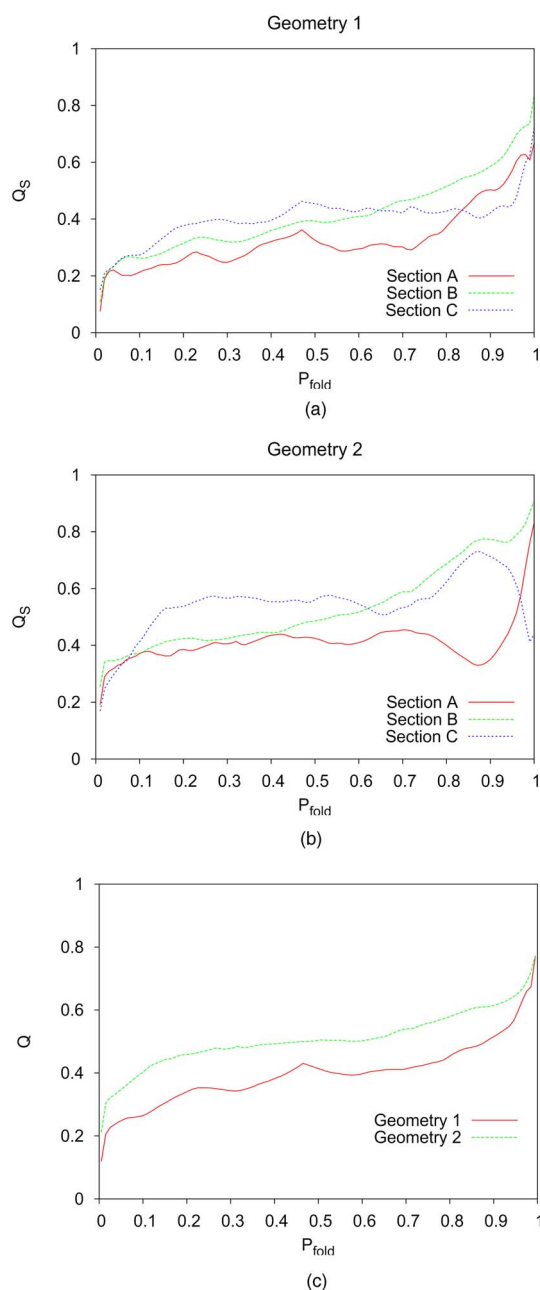


FIG. 7. (Color online) Average fraction of native bonds in each protein section, Q_S , as a function of P_{fold} in geometry 1 (a) and geometry 2 (b). Also shown is the dependence of the protein's average fraction of native bonds on the reaction coordinate, P_{fold} , for both geometries. Note that when folding is near completion at high P_{fold} , there is a sharp increase in the fraction of native contacts for geometry 1.

small lattice proteins.³⁸ In a related effort, a simulational proxy of the phi-value analysis was used to identify nucleating residues in the folding of the two model proteins investigated in the present work.⁴⁰ We have found that the set of residues 6, 33, and 35 in geometry 1 and residues 19, 20, and 29 in geometry 2 lead to the largest increase in folding time upon mutation. Interestingly, the vast majority (i.e., more than 60%) of contacts formed by these residues are present in section B of both proteins.

The conclusion that section B encapsulates the nucleating residues (and therefore the set of native bonds forming a postcritical FN) can actually be drawn from an independent

analysis of the results obtained so far. In addition to shedding light on the existence of protein sections, the contact correlation analysis introduced in Sec. III B shows that the folding of section B is a prerequisite to observe inevitable (i.e., highly probable) folding of the whole protein. Indeed, section B is always folded in the high- P_{fold} conformations that are on pathway to the native state (i.e., that fold fast). Therefore, if these model proteins fold via nucleation, section B must necessarily contain the critical residues forming the FN. In support of this argument we have found an increase in the correlation between Q_B and P_{fold} for both geometries when folding is near completion (i.e., $P_{\text{fold}} > 0.85$), which implies that the folding of section B determines the inevitable folding of the whole protein. For example, in geometry 1, a conformation where section B is folded has folding probability $P_{\text{fold}} > 0.93$. Also illuminating is the fact that the presence, with high probability, of the bonds forming section B is independent of which other bonds are formed in the protein. Moreover, the most probable folding pathways are those in which section B folds early, while those in which it folds last have a vanishingly small probability of occurrence.

VII. COOPERATIVITY AT THE LEVEL OF MACROSTRUCTURAL FORMATION

In protein folding the term cooperativity is generally used in connection with specific thermodynamic and kinetic features exhibited by small, single domain proteins. Indeed, extraordinary experimental traits such as the linear chevron behavior (kinetic cooperativity) and the verification of the van t'Hoff criterion (thermodynamic cooperativity) have been typically ascribed to the existence of highly unusual energetics involving nonadditive multibody interactions.^{39,41}

The results reported in this work are suggestive that geometry 1 folds in a more cooperative manner than geometry 2. This difference in cooperative behavior is particularly evident from the study of section formation along the reaction coordinate (Fig. 6). Here, it is shown that both sections A and C in geometry 2 can have the vast majority of its bonds formed ($Q_S > 0.9$) early on in folding (i.e., at low P_{fold}). In geometry 1, on the other hand, the formation of bonds within one section does not happen in such an independent manner. Indeed, it is only when folding of the overall protein is near completion (i.e., for $P_{\text{fold}} > 0.9$) that the fraction of bonds within each section comes close to unity. Also suggestive of the more cooperative behavior of geometry 1 are the considerably smaller times elapsing between the formation of the first section and the folding of the whole protein (Fig. 3). Indeed, not only these time intervals are considerably smaller in geometry 1 than in geometry 2, as they are (on average) 33% smaller than the cutoff time that was used to select the conformations that fold inevitably fast from other high P_{fold} conformations (Sec. III A). For geometry 2 such time intervals are similar to this cutoff parameter and much larger than the average folding time of conformations on pathway with $P_{\text{fold}} > 0.9$. These times are in line with the finding that the first section to fold can do it relatively early during the process (i.e., $P_{\text{fold}} \ll 0.9$). Finally, the higher cooperativity of ge-

ometry 1 is also evident from the sharper increase in the fraction of native contacts Q that is observed near the very end of its folding process [Fig. 7(c)].

VIII. CONCLUSIONS

In the present work we investigated the existence of folding pathways for two model proteins differing in native geometry at a coarse-grained level of structure formation. To this end we developed and applied a methodology, based on native contact correlation analysis, which identifies protein sections with clusters of highly correlated native bonds. The latter were shown to map onto well-defined structural three dimensional domains within the native fold.

Three protein sections and four folding pathways, corresponding to different ordering preferences of section formation, were identified for each protein. Interestingly, the analysis of folding pathways at a macrostructural level of structure formation revealed a common underlying folding mechanism, based on nucleation phenomena, for both target geometries. Indeed, our results show that one of the protein sections contains a set of critical bonds that form the folding nucleus. In the most complex geometry this section and the folding nucleus have a topology similar to that of the native fold.^{42,43}

Despite these similarities, a relevant difference was identified between the folding processes related to the different cooperative behaviors of the two proteins. The higher cooperativity observed for the most complex geometry is probably due to the larger number of nonlocal, long-range native bonds of the native fold as well as of the folding nucleus.²⁷ In other words the higher cooperativity of the folding process of the complex geometry is ascribed to the nontrivial order of the native fold, that is mimicked by that of the folding nucleus. Despite the small size of the two model proteins, this structural difference has a marked effect in the dynamics of the folding process and for the complex geometry it resembles the dynamics of first order transitions in the thermodynamic limit. Quantitative measures of cooperativity and, in particular, the size dependence of the nucleation barrier for the different geometries are outside the scope of this work.⁴⁴

We speculate that by introducing chemical specificity in our model proteins the number, or at least the probability of occurrence, of the folding pathways identified here, and that are solely driven by native geometry, will probably change. The use of a sequence-specific model (e.g., using the Miyazawa-Jernigan potential) is, however, out of the scope of the present study and will be investigated in future work.

ACKNOWLEDGMENTS

Two of the authors (R.D.M.T. and P.F.N.F.) thank Fundação para a Ciência e Tecnologia (FCT) for financial support through Grant Nos. SFRH/BPD/27328/2006 and SFRH/BPD/21492/2005, respectively. This work was also supported by FCT through Project Nos. POCI/FIS/55592/2005 and POCTI/ISFL/2/618.

¹C. Anfinsen, *Science* **181**, 223 (1973).

²C. Levinthal, *J. Chim. Phys.* **65**, 44 (1968).

³R. L. Baldwin, *Nat. Struct. Biol.* **6**, 814 (1999).

- ⁴S. E. Jackson and A. R. Fersht, *Biochemistry* **29**, 10428 (1991).
- ⁵S. E. Jackson, *Folding Des.* **3**, R81 (1998).
- ⁶K. W. Plaxco, K. T. Simmons, I. Ruczinski, and D. Baker, *Biochemistry* **39**, 11177 (2000).
- ⁷M. M. Gromiha and S. Selvaraj, *J. Mol. Biol.* **310**, 27 (2001).
- ⁸H. Zhou and Y. Zhou, *Biophys. J.* **82**, 458 (2002).
- ⁹A. Fersht, *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding* (Freeman, San Francisco, 1998).
- ¹⁰L. S. Itzhaki, D. E. Otzen, and A. R. Fersht, *J. Mol. Biol.* **254**, 260 (1995).
- ¹¹V. I. Abkevich, A. M. Gutin, and E. I. Shakhnovich, *Biochemistry* **33**, 10026 (1994).
- ¹²B. Nolting and K. Andert, *Proteins* **41**, 288 (2000).
- ¹³V. S. Pande and D. S. Rokhsar, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 1273 (1999).
- ¹⁴T. X. Hoang and M. Cieplak, *J. Chem. Phys.* **113**, 8319 (2000).
- ¹⁵G. Tiana and R. A. Broglia, *J. Chem. Phys.* **114**, 2503 (2001).
- ¹⁶R. A. Broglia and G. Tiana, *J. Chem. Phys.* **114**, 7267 (2001).
- ¹⁷P. Ferrara and A. Caffisch, *J. Mol. Biol.* **306**, 837 (2001).
- ¹⁸T. X. Hoang, M. Cieplak, and M. O. Robbins, *Proteins* **49**, 114124 (2002).
- ¹⁹M. A. Seeliger, S. E. Breward, and L. S. Itzhaki, *J. Mol. Biol.* **325**, 189 (2005).
- ²⁰M. Cieplak, T. X. Hoang, and M. O. Robbins, *Proteins* **56**, 285 (2004).
- ²¹A. Irback, S. Mitternacht, and S. Mohanty, *Proc. Natl. Acad. Sci. U.S.A.* **102**, 13427 (2005).
- ²²I. A. Hubner, M. Oliveberg, and E. I. Shakhnovich, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 8354 (2006).
- ²³L. Sutto, G. Tiana, and R. Broglia, *Protein Sci.* **15**, 1638 (2006).
- ²⁴L. G. Garcia and A. F. P. Araujo, *Proteins* **62**, 4663 (2006).
- ²⁵T. S. Norcross and T. O. Yeates, *J. Mol. Biol.* **362**, 605 (2006).
- ²⁶L. Reich and T. R. Weikl, *Proteins* **63**, 10521058 (2006).
- ²⁷N. Go and H. Taketomi, *Proc. Natl. Acad. Sci. U.S.A.* **75**, 559563 (1978).
- ²⁸N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, *J. Chem. Phys.* **21**, 1087 (1953).
- ²⁹D. P. Landau and K. Binder, *A Guide to Monte Carlo Simulations in Statistical Physics* (Cambridge University Press, Cambridge, 2000).
- ³⁰H. S. Chan and K. A. Dill, *J. Chem. Phys.* **100**, 9238 (1994).
- ³¹M. Cieplak and T. X. Hoang, *Phys. Rev. E* **58**, 3589 (1998).
- ³²M. Oliveberg, Y. Tan, and A. R. Fersht, *Proc. Natl. Acad. Sci. U.S.A.* **92**, 8926 (1995).
- ³³A. Gutin, A. Sali, V. Abkevich, M. Karplus, and E. I. Shakhnovich, *J. Chem. Phys.* **108**, 6466 (1998).
- ³⁴M. Cieplak, T. X. Hoang, and M. S. Li, *Phys. Rev. Lett.* **83**, 1684 (1999).
- ³⁵P. F. N. Faisca and R. C. Ball, *J. Chem. Phys.* **116**, 7231 (2002).
- ³⁶P. F. N. Faisca and R. C. Ball, *J. Chem. Phys.* **117**, 8587 (2002).
- ³⁷R. Du, V. S. Pande, A. Y. Grosberg, T. Tanaka, and E. I. Shakhnovich, *J. Chem. Phys.* **108**, 334 (1998).
- ³⁸R. D. M. Travasso, P. F. N. Faisca, and M. M. Telo da Gama, *J. Phys.: Condens. Matter* **19**, 215212 (2007).
- ³⁹H. S. Chan, S. Shimizu, and H. Kaya, *Methods Enzymol.* **380**, 350 (2004).
- ⁴⁰P. F. N. Faisca, R. D. M. Travasso, M. M. Telo da Gama, R. C. Ball, and E. I. Shakhnovich (unpublished).
- ⁴¹P. F. N. Faisca and K. W. Plaxco, *Protein Sci.* **15**, 1608 (2006).
- ⁴²A. Fersht, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 15251529 (2000).
- ⁴³E. Paci, K. Lindorff-Larsen, C. M. Dobson, M. Karplus, and M. Vendruscolo, *J. Mol. Biol.* **352**, 495 (2005).
- ⁴⁴W. B. Hu and D. Frenkel, *J. Phys. Chem. B* **110**, 3734 (2006).