The nucleation mechanism of protein folding: a survey of computer simulation studies

TOPICAL REVIEW

The nucleation mechanism of protein folding: a survey of computer simulation studies

Patrícia F N Faísca

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

E-mail: patnev@cii.fc.ul.pt

Abstract
The nucleation mechanism of protein folding, originally proposed by Baldwin in the early 1970s, was firstly observed by Shakhnovich and co-workers two decades later in the context of Monte Carlo simulations of a simple lattice model. At about the same time the extensive use of ϕ-value analysis provided the first experimental evidence that the folding of Chymotrypsin-inhibitor 2, a small single-domain protein, which folds with two-state kinetics, is also driven by a nucleation mechanism. Since then, the nucleation mechanism is generally considered the most common form of folding mechanism amongst two-state proteins. However, recent experimental data has put forward the idea that this may not necessarily be so, since the accuracy of the experimentally determined ϕ values, which are used to identify the critical (i.e. nucleating) residues, is typically poor. Here, we provide a survey of *in silico* results on the nucleation mechanism, ranging from simple lattice Monte Carlo to more sophisticated off-lattice molecular dynamics simulations, and discuss them in light of experimental data.

(Some figures in this article are in colour only in the electronic version)

Contents

1. Introduction

Understanding protein folding, the self-assembly process according to which a linear chain of amino acids acquires

its three-dimensional native (i.e. biologically functional) structure, remains a challenging problem despite more than 70 years of dedicated research [1, 2].

In the 1960s Anfinsen and co-workers performed a series of *in vitro* experiments which showed that some chemically denatured proteins are able to spontaneously refold to their respective native conformations. This observation led to the formulation of the so-called thermodynamic hypothesis of protein folding, according to which the native state is the global minimum of the Gibbs free energy [3]. But how do proteins find this native state? In other words, what is the mechanism of protein folding?

As pointed out by Levinthal, in what was to become widely known as the 'Levinthal paradox', a random search of the entire conformational space is not compatible with the timescale of protein folding [4]; therefore, a smarter mechanism must guide an unfolded polypeptide chain in the search for its native conformation. Solving the mechanism of protein folding is—since the late 1960s—a problem of paramount importance in the field of protein science, but unravelling its solution has been far from straightforward. In part, this difficulty is due to the fact that proteins do not appear to fold by means of a unique mechanism, and over the years several phenomenological models have been proposed for protein folding [4, 5, 7, 13, 16, 17]. Levinthal himself suggested that protein folding would be speeded up by what he termed 'nucleation points', local interactions between the amino acids that form rapidly and in a stable manner, allowing the subsequent formation of larger structural elements that eventually undergo further assembly to yield the native structure [4]. Implicit in Levinthal's suggestion was the core idea of what later became known as the framework model, which envisions protein folding as an hierarchical process, where the formation of hydrogen-bonded secondary structural elements (e.g. α-helices and β-sheets) precedes the formation of the tertiary structure [5, 6]. The diffusion–collision model— also proposed in the 1970s—pictures a different scenario for folding by assuming that an essential part of the process are the diffusive encounters (i.e. collisions) between metastable regions of the structure which result in more stable coalescence intermediates [7].

In the early 1990s Jackson and Fersht provided experimental evidence that the folding kinetics of small (<100 amino acids), single-domain proteins—epitomized by the 64-residue protein Chymotrypsin-inhibitor 2 (CI2)—is remarkably well described by a two-state model [8, 9]. This observation suggests that the only relevant milestones along the folding reaction are the native state (N) and the denatured (D) ensemble, separated by a free energy barrier on the top of which lies the transition state (TS). The absence of significantly populated intermediate species is compatible with the hypothesis that a kinetic mechanism akin to the nucleation-growth mechanism of first-order phase transitions in infinite systems [10, 11] is at play in the folding of small proteins.

Although a model for folding as being *limited* by nucleation had been originally proposed by Baldwin and co-workers in the early 1970s [12] (instead of being *initiated* as in the models of Levinthal [4] and Wetlaufer [5]), it was only

in the 1990s that Shakhnovich and co-workers reported the first detailed microscopic study, developed in the framework of Monte Carlo simulations of a simple lattice model, which supports the hypothesis of a nucleation mechanism being at the heart of the folding process [13]. However, contrary to the earlier models where nucleating events initiate folding, the MC investigation revealed a picture of the folding process that is instead *limited* by nucleation, a scenario that was originally proposed by Baldwin and co-workers [6]. Indeed, Shakhnovich and co-workers observed that in the folding of the lattice polymer, the rate limiting step is the formation of a *specific* set of native contacts (which are predominantly long-ranged), termed folding nucleus (FN), after which the native fold is achieved promptly and reproducibly. From this point onwards, the study of the folding mechanism became inextricably linked with that of the transition state.

Shortly after Shakhnovich's discovery, the extensive use of a protein engineering method termed 'ϕ-value analysis' provided the first microscopic characterization of the structure of the TS of CI2 [18, 19]. Briefly, the ϕ value is obtained by measuring the effect of a single-site mutation on the folding rate and stability, namely, $\phi = -RT \ln(k_{\mathrm{mut}}/k_{\mathrm{WT}})/\Delta\Delta G_{\mathrm{N-D}}$, where k_{mut} and k_{WT} are the folding rates of the mutant and wild-type (WT) proteins, respectively, and $\Delta\Delta G_{\mathrm{N-D}}$ is the change in the free energy of folding upon mutation. For a non-disruptive mutation, which is intended to cause a small perturbation, $-RT \ln(k_{\mathrm{mut}}/k_{\mathrm{WT}})$ can be approximated by the change in the activation energy of folding, $\Delta\Delta G_{\mathrm{TS-D}}$, and therefore $\phi = \Delta\Delta G_{\mathrm{TS-D}}/\Delta\Delta G_{\mathrm{N-D}}$. Likewise, for two-state folding proteins, a ϕ value near unity means that the TS is energetically perturbed upon mutation as much as the native state is perturbed, which has been typically interpreted as if the mutated residue is fully native (i.e. has all its native interactions established) in the TS. On the other hand, a ϕ value near zero is taken as evidence that the residue is as unstructured in the TS as it is in the denatured ensemble. The traditional interpretation of fractional ϕ values is, however, not straightforward as they might indicate the existence of multiple folding pathways or a unique transition state ensemble (TSE) with genuinely weakened interactions [18]. Moreover, the interpretation of the so-called nonclassical ϕ values ($\phi > 1$ and $\phi < 0$) is not straightforward as well and alternative models for ϕ have been recently proposed [20].

According to Fersht and co-workers, the picture of the TS that emerges from the ϕ-value analysis is compatible with CI2 folding via a nucleation mechanism similar to that reported for lattice proteins. The lack of tertiary structure in the TS of CI2 was taken as evidence that secondary and tertiary structures form concomitantly in a process that is triggered by the formation of the FN, a set of local interactions stabilized by a few long-range interactions which are mainly associated with the residues displaying the highest ϕ values. Such a process was coined the nucleation–condensation mechanism of protein folding [21]. Subsequent studies, focusing on other target proteins, have provided further evidence that the nucleation mechanism is common amongst small, two-state proteins [22–24].

A few years ago a couple of studies that investigated the relationship between ϕ-value reliability and the change in the

Figure 1. Protein models used in simulations of protein folding in order of increasing complexity representation. The simple (cubic) lattice model (left) is a generic protein representation displaying the fundamental features of the protein backbone (chain connectivity, excluded-volume, etc). Each bead represents one of the 20 existing amino acids that are connected by sticks representing the peptide bond. The C-α model (centre) is the simplest off-lattice representation. As the lattice model it is also a coarse-grained description of the protein that reduces each amino acid to a sphere catered in the position of each C_α carbon. However, it is a more realistic representation of the protein that not only takes into account the polymeric nature of the protein backbone but also features the specific three-dimensional native structure of the protein. Finally, in the full atomistic off-lattice representation (right) all the heavy atoms of the protein are explicitly taken into account. Figures drawn with *Mathematica* (left) and PyMOL [86].

free energy of folding upon mutation reported that the accuracy of the experimentally determined ϕ values is poor unless $\Delta\Delta G_{N-D} > 7$ kJ mol^{-1} [25]. A subsequent investigation *in silico* established a baseline for $\Delta\Delta G_{N-D}$ of 6 kJ mol^{-1} [26], while the most recent account of this issue reported a smaller baseline of 5 kJ mol^{-1} [27]. In a related study, Raleigh and Plaxco pointed out that only three out of the 125 more accurately determined ϕ values reported in the literature lie above 0.8, and that about 85% of the mutations characterized for single-domain proteins show ϕ values below 0.6 [28]. Overall, these findings have triggered some debate regarding the existence of specific nucleation sites in real proteins, and likewise on the existence of a nucleation mechanism of protein folding.

Since Shakhnovich's pioneering study on the nucleation mechanism, the continuous increase of computing power has been allowing researchers to simulate folding with more sophisticated and more realistic protein representations, which led to new views, interpretations and to a deeper understanding of the nature of nucleation phenomena in protein folding. The purpose of the present review is that of making an assessment of those investigations in light of related experimental data.

We start by making a brief overview of the models and computational methodologies used to simulate folding in the computer. We then review and discuss a selection of computational results. We start with Monte Carlo simulations of lattice models, which deal with the fundamental principles of the mechanism of folding, and we proceed by discussing the nucleation mechanism of specific real-world proteins, for which more sophisticated off-lattice models have been employed. Finally, we draw some concluding remarks.

2. Protein models: a primer

In this section we provide a brief description of the most relevant protein models used in simulations of protein folding. Extensive accounts of the selected models and simulational methods can be found in [29–34].

2.1. The lattice model

The lattice model is one of the most commonly used protein representations in simulations of protein folding (figure 1, left). In the lattice model the three-dimensional space is discretized by embedding the protein in a lattice (two- or three-dimensional) and a coarse-grained description of the molecule is considered, which is often referred to as the 'bead & stick' representation. Indeed, in its simplest form, the lattice model reduces the amino acids to beads of uniform size and the peptide bond, which covalently connects the amino acids along the polypeptide chain, is represented by sticks of uniform length, corresponding to the lattice spacing. Despite their simplicity lattice models take into account two fundamental traits of protein molecules, namely chain connectivity and excluded-volume interactions. Protein energetics is modelled via the so-called contact Hamiltonian, which defines the energy of each conformation (i.e. the two- or three-dimensional representation of the protein that is defined by the set of bead coordinates) as the sum over all the pairs of amino acids that make a contact (i.e. that are separated by a lattice spacing but are not covalently linked) in the considered conformation. The energy interaction parameters are typically drawn either from the Gō potential [35] or from the Miyazawa–Jernigan (MJ) potential [36]. The Gō potential is based on the idea that the native fold is very well optimized energetically. Likewise, the only contacts that contribute to the system's energy are those present in the native conformation. Thus the Gō potential is ultraspecific; it is defined by the native structure, and because of that it is particularly adequate to simulate the folding of proteins with minimal energetic frustration (i.e. smooth energy landscapes). The MJ potential, on the other hand, is a sequence-specific potential as it considers the 20 naturally occurring amino acids and establishes 20×20 energy parameters for the interactions between them. In this case the native structure does not uniquely determine the energy of a conformation and a protein sequence, with a fixed chemical composition, must be designed in order to have the native structure as the global energy minimum [37]. Moreover, when

the MJ parameters are used to model protein energetics the non-native contacts also contribute to the system's total energy.

Another possibility, which is most often used in combination with the two-dimensional lattice model, is that of the hydrophobic-polar (HP) potential [38]. The HP model is intended to capture a major driving force of protein folding which is the hydrophobic effect. Thus, only two species of amino acids—hydrophobic and hydrophilic (or polar)—are considered and the only stabilizing contacts are those between the hydrophobic residues. Finally, sometimes, the energy interaction parameters are taken as random values with a Gaussian distribution [39].

In simulations of protein folding using the lattice model the exploration of conformational space is done with Monte Carlo (MC) methods. Most often the classical Metropolis algorithm [40] is employed, and the movement of the protein is mimicked with the Verdier–Stockmayer move set [41], including end- and corner-flip moves, or the kink-jump move set which also includes 'crank-shaft' moves that displace two beads simultaneously [42].

2.2. The off-lattice C_α model

The C_α model is the simplest continuous (i.e. off-lattice) protein representation that is constructed on the basis of the protein's crystal structure, generally downloadable from the protein databank (PDB) (figure 1, centre). The C_α model is—like the lattice model—a coarse-grained representation of the protein backbone that reduces each amino acid to a single bead of uniform radius centred in the position of its C_α carbon. The potential energy of a conformation contains bond interactions and angle interactions, routinely used in molecular dynamics (MD) simulations of biopolymers, which guarantee the rigidity of the protein's backbone. The interaction between residues (the so-called non-bonded terms) is typically modelled with a Gō-type interaction energy, based on a van der Waals potential [43, 99, 44], although other energy potentials have been employed [45]. Therefore, attractive interactions are assigned only to the native contacts, and hard-core repulsive (excluded-volume) interactions are assigned to the non-native ones. Contrary to what happens in the lattice representation, where a contact between two beads is accurately defined by the lattice spacing, in the C_α representation a contact between a pair of beads is always defined with a certain degree of arbitrariness, which naturally represents a shortcoming of the model. Indeed, simulation results are sensitive to the choice of the contact cutoff distance employed [43, 45]. In the C_α model the solvent is often modelled by means of an implicit (continuum) approach [43, 44]. Most generally, a random force, which balances energy dissipation through a Langevin noise term, is used to mimic water–protein collisions. Dynamic information on the folding process is obtained by integrating the equations of motion for each C_α atom.

2.3. Full atomistic models

In a full atomistic model all of the protein's heavy atoms are taken into account (figure 1, right). The solvent representation can be implemented either implicitly or explicitly. In the full atomistic representation the potential energy of the protein

(the so-called force field in the jargon of MD) comprises the standard bonded terms together with electrostatic, van der Waals, hydrogen bonding and other intermolecular forces considered at the atomic level. Since the MD approach can entail a level of detail capable of describing the fully atomistic structure of a solvated protein it is an appropriate method to reproduce 'in silico' the protein folding 'reaction'. However, due to limitations of timescale and force field accuracy it remains a challenging task to simulate protein folding with MD. Nevertheless, the development in recent years of alternative simulation algorithms (e.g. ensemble dynamics) combined with world-wide distributed computing has allowed simulators to access timescales comparable to those observed experimentally, making it possible to compare in an absolute manner experimental and simulated data [48, 46, 47].

An alternative methodology was proposed a few years ago by Shakhnovich and co-workers that combines a coarse-grained description of motion and energetics with a full atomistic description of the protein (where all non-hydrogen heavy atoms are represented by spheres of different radii) [54, 50]. Conformational space is explored with Metropolis MC by means of a local move set that mimics side-chain and backbone torsions while maintaining chain connectivity, planar peptide bonds and excluded-volume. Protein energetics is described by means of a square well Gō-potential. This hybrid all-atom approach has the advantage of allowing the exploration of a statistically significant number of folding trajectories with standard computational resources.

3. The nucleation mechanism: general principles from lattice models

3.1. The specific nucleus model

In their seminal work, Shakhnovich and co-workers explored the existence of nucleation phenomena in simple lattice models (36- and 80-mers on the cubic lattice with interaction energies drawn from the MJ potential). The starting point of their investigation is the concept of a folding nucleus (FN) defined as '*a set of contacts that satisfies the following two conditions: (i) formation of a nucleus is a sufficient condition for folding; i.e. after a set of contacts that constitutes the nucleus is formed, the subsequent folding is guaranteed and is very fast. (ii) Formation of a nucleus is a necessary condition for folding; i.e. the pattern of contacts corresponding to the nucleus is always present in prefolding conformations when the number of native contacts is relatively small, but subsequent folding is very fast*' [13]. Implicit in this definition is the idea of a 'postcritical nucleus', i.e. the first stable structures that appear immediately after the transition state is overcome [56]. However, as pointed out by the authors, such a postcritical nucleus will only differ from the one forming in the TS by a few $k_B T$. In practice, the nucleus thus defined was identified as being the set of native contacts which are common to *all* conformations with a fraction of native contacts Q smaller than 0.6 (i.e. which have less than 60% of its native contacts formed), and that fold very fast, in about 1% of the total folding time (i.e. the number of MC steps needed to achieve the
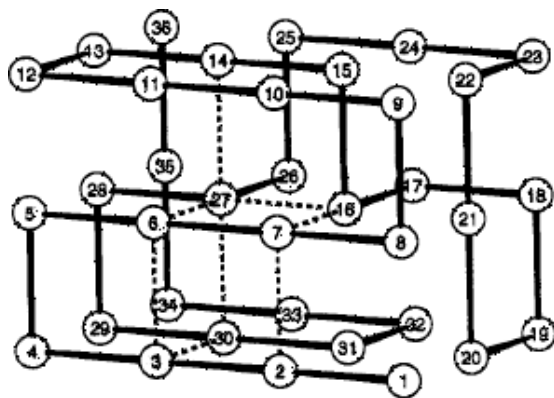
Figure 2. The specific set of native contacts forming the FN of a lattice model system. Reprinted (in part) with permission from [13]. Copyright 1994 American Chemical Society.

native state starting from a randomly unfolded conformation). These contacts form a spatially localized substructure of the native state whose size is about 20% of the total number of native contacts (figure 2). More interesting, however, is the observation that the FN is predominantly formed by non-local contacts (i.e. contacts between beads that are far away from each other in the sequence), which partly explains the cooperative character [14] of the folding transition [15, 51–53]. Of note is the observation that the position of the nucleus (i.e. the location of its residues along the protein sequence) was found to be the same for all the three non-homologous sequences studied (that were nevertheless designed to fold to the same native structure), which suggests a major role played by the native structure in the determination of the FN.

3.2. The multiple nuclei model

The multiple nuclei model was proposed by Thirumalai and Klimov in 1998, based on extensive MC folding simulations of 27- and 36-mers on the cubic lattice. The interaction potential used to model the interactions between beads was sequence-specific with the energy parameters were drawn from a Gaussian distribution [55]. In their study, the FN is defined as the minimal set of native contacts which (i) are stable—meaning that once they form they stay formed until the native state is reached and (ii) results in rapid assembly of the native conformation. Condition (ii) implies that the formation of the nucleus is rate-limiting, which results in the nucleus being identified with the folding TS. Condition (i), however, leads to a completely different nucleation scenario from that described previously. Indeed, by systematically exploring the dynamics of native contact formation during folding Klimov and Thirumalai found out that their adopted definition does not specify a unique/single FN for a given protein sequence. Instead, depending on the initial conformation, many nuclei—differing in size and composition—can be identified, suggesting that the TS does not correspond to a unique conformation but that it contains many conformations forming a transition state ensemble (TSE). Furthermore, different protein sequences display different folding nuclei. As in the single nucleus model, the folding nuclei are formed by local and non-local contacts, but in the multiple nuclei model the local contacts are predominant in the FN.

3.3. Nucleation and transition state heterogeneity

The two computational studies briefly summarized reveal sharply different views of the nucleation mechanism and of the folding TS. In principle, this is not surprising because they are based on significantly different definitions of FN. But of the two proposed views, which is the one that most correctly describes the folding mechanism of real-world proteins? The answer to this question is that both models may be correct. Indeed, *in vitro* investigations on the structure of the TS, based on the use of ψ-value analysis—a variant of ϕ-value analysis developed by Sosnick and colleagues [58]—have shown that essentially two classes of TS heterogeneity can be identified. As discussed in [59, 58] folding may occur via a single TS nucleus (as in the case of protein Ub), or via a TS ensemble (TSE) which contains structurally disjoint members, corresponding to distinct and multiple folding nuclei (as in the case of the dimeric GCN4 coiled coil and titin I27). Furthermore, two variations of the single-nucleus model can be distinguished: either there is a group of contacts which are absolutely required in a given nucleus (single-nucleus model) or there is a group of contacts which are critical for the FN but different groups of structures may as well exist at the TS (single nucleus with microscopic heterogeneity) (figure 3).

3.4. Non-native interactions and the nucleation mechanism

Understanding the role played by non-native interactions in the energetics and dynamics of protein folding is a major issue in protein folding research, and has motivated several computational investigations during the last decade [60, 61, 67–70].

The observation that for many small proteins folding is cooperative and two-state does not preclude the establishment of non-native interactions between the amino acids, especially when the protein is still only partially folded [70]. Thus, it is important to understand the way(s) in which non-native interactions affect the whole process, and in particular the folding TS.

The relation between non-native interactions and the nucleation mechanism was firstly explored by Shakhnovich and co-workers within the scope of a cubic lattice model with side chains (the explicit representation of side chains is intended to capture packing effects) [62]. In this study the FN is defined as the set of contacts forming with probability, p, larger than 0.5 in the ensemble of conformations folding in 2% of the total folding time, and with the fraction of native contacts $Q = 0.41$ [57]. Clearly, this definition of FN underpins the model of a single nucleus with microscopic heterogeneity. Indeed, only 20% of the contacts identified as being part of the FN form with very high probability $p \sim 0.8$. More interesting, however, is the observation that 45% of the nucleus contacts are non-native and predominantly non-local. When these contacts were simultaneously mutated (through the destabilization of their interaction energies) a threefold deceleration was observed in the folding rate, with no change in the stability of the native state (as indicated by similar values of the melting temperatures for both mutant and WT protein). This finding shows that specific non-native interactions may have an important effect in stabilizing the folding TS, without changing the stability of the native structure, which may explain the origin of nonclassical ϕ values.
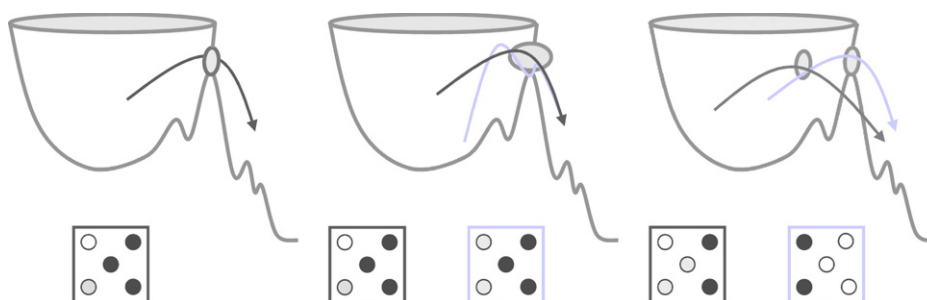
5

Figure 3. Classes of TS heterogeneity. Folding may occur through a single TS nucleus (left) where some residues are absolutely required for folding (dark grey beads), some residues are completely unfolded (white boxes) and others have partially formed interactions in the TS (light grey beads). Folding may also occur via a structurally heterogeneous TS (middle). In this case the ensemble of conformations forming the TS has different contacts formed but they share a group of conserved residues that is the FN (single nucleus with microscopic heterogeneity). Finally, folding may also occur through distinct folding nuclei (multiple nuclei model). Adapted from figure 12 in [58].

3.5. The folding probability as a TSE locator

The concept of folding probability, P_{fold}, was introduced by Du *et al* [63], and further explored by Snow and Pande [64], as a way to understand and describe protein folding kinetics. It is equivalent to the standard transmission coefficient in Eyring's TS theory of chemical kinetics [49]. In operational terms the P_{fold} of a given conformation is defined as the probability that a given conformation folds before it unfolds. Thus, while conformations with $P_{fold} = 1$ define the native state, conformations with $P_{fold} \sim 0$ are representative of the unfolded ensemble. Also, from its definition it follows that P_{fold} is a measure of the kinetic distance between a given conformation and the folded state or the unfolded ensemble. Clearly, any state with $P_{fold} > 0.5$ is more likely to fold first than to unfold first and is therefore kinetically closer to the folded state. A similar argument holds for $P_{fold} < 0.5$. The case of $P_{fold} = 0.5$ is more interesting. Since a conformation with $P_{fold} = 0.5$ has an equal probability to either fold or unfold it is reasonable to define the ensemble of conformations with $P_{fold} = 0.5$ as the TSE (figure 4).

Since the evaluation of P_{fold} amounts to a Bernoulli trial, the relative error resulting from using M runs in the calculation of P_{fold} scales as $M^{-1/2}$. Therefore, to obtain an accurate estimate of P_{fold} a very large number of simulation runs should be considered. While this requirement generally poses no challenge in the case of simple MC lattice simulations (for which a substantially large number of folding simulations can be obtained in a relatively short amount of time), some difficulty arises when off-lattice models are used to simulate folding. Thus, and in an attempt to bypass this problem, the suitability of alternative measures of folding progression has been investigated. In particular, it was recently reported that for proteins that fold by a simple two-state mechanism the ϕ values of the TSE predicted by structural reaction coordinates, such as the fraction of native contacts Q, are almost identical to those of the TSE based on the use of P_{fold} [65]. These results are thus suggestive that, for proteins with smooth energy landscapes, the fraction of native contacts can provide a good approximation to P_{fold} and therefore it can also be used to locate the TS and to probe the progress of the folding reaction. We stress, however, that a recent study that investigated the
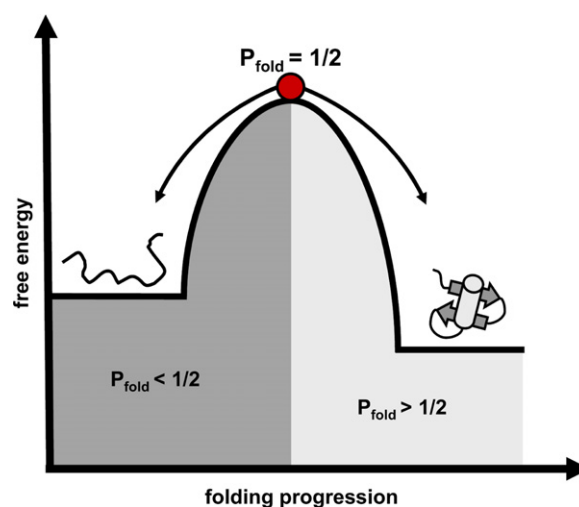

Figure 4. The free energy profile of a protein whose folding kinetics is two-state. The native state is separated from the unfolded ensemble by a free energy barrier on the top of which stays the TS. By definition of folding probability, conformations on the top of the free energy barrier have an equal probability to fold and unfold. On the other hand, pre-TS conformations have $P_{fold} < 0.5$ while for post-TS conformations $P_{fold} > 0.5$.

energy landscape of an exactly solvable model of a small β-hairpin with 12 residues has refuted the suitability of P_{fold} and Q as reaction coordinates for folding and, in particular, their suitability as TS locators [66].

3.6. The importance of native geometry in determining the nucleation mechanism

In the late 1990s Plaxco and Baker provided empirical evidence for a major role played by native geometry in protein folding kinetics [71]. Indeed, a strong correlation was found between a parameter of native geometry named contact order—measuring the average sequence separation between all pairs of residues in contact in the native structure relative to the total length of the protein—and the folding rates of small, two-state proteins. The following studies using related metrics of native geometry further strengthened the idea that native geometry plays a key role in folding [72].
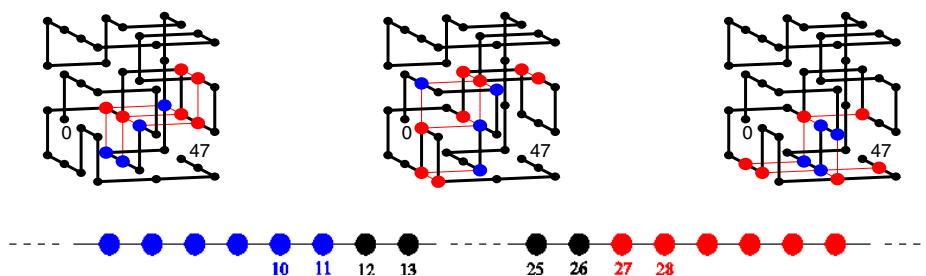
Figure 5. The FN for the Gō model (left), for sequence 1 (centre) and for sequence 2 (right), is the set of nine, ten and eight contacts, respectively. Beads pertaining to the FN whose number along the sequence is less than 12 are coloured in blue (dark grey) and those whose number along the sequence is larger than 26 are coloured in red (light grey) (bottom). Adapted from figure 6 in [79].

A circular permutant is an engineered form of a protein which results from linking the C- and N-termini after disrupting the protein backbone at some selected peptide bond. While this rather radical mutation procedure typically leads to minimal changes of the WT native structure and energetics, it can alter chain connectivity and contact order up to a large extent. It is therefore a clever procedure to evaluate the roles played by native geometry in folding kinetics and mechanisms. Based on this premise, Li and Shakhnovich [73] constructed two circular permutants of the lattice protein with side chains described previously. One of them cuts the backbone at the FN, while the other cuts the backbone in a region unrelated to the TS. It was found that the latter retains the FN, but in the former a new nucleus is formed that results in a slower folding kinetics. This observation led to the conclusion that the native structure of a protein does not completely determine the FN, and that a significant change in backbone connectivity can move the FN from one region to another.

The change in FN upon circular permutation has been observed in investigations with real-world proteins (src-SH3 and protein S6) [74–76], although it is not generic (e.g. CI2 was shown to retain the FN upon circular permutation [77]). These observations have been rationalized on the basis of the relative amount of local and long-ranged native contacts in the unperturbed TS. More precisely if—just like in CI2—the nucleating contacts are mostly local and uniformly distributed over the native fold, significant changes are not expected upon shifting of the sequence over the structure. If, in contrast, the formation of the TS involves the establishment of specific non-local, long-ranged contacts, as in complex native geometries like src-SH3, then changing the chain connectivity in those regions is expected to have a much more dramatic impact in the formation of the TSE [78].

3.7. The importance of protein sequence in determining the nucleation mechanism

In a recent study we have determined the folding nuclei displayed by three different protein sequences that fold to the same native structure on the cubic lattice [79]. In one of the sequences the interactions between beads were modelled with the Gō potential, while the MJ interaction parameters were used for the other two model sequences which display specific amino acid contents. Therefore, in one case the FN is exclusively ascribable to the native geometry while in the other

two cases the interplay between native geometry and stability (i.e. energetics) establishes and drives a nucleation pattern.

We define the FN as the set of most probable native contacts formed in the ensemble of conformations which are separated from the native conformation as far away as possible in time, and yet fold with high folding probability, $P_{fold} \geqslant 0.9$, and rapidly (in less than 5% of the total folding time). Based on this definition three different folding nuclei were identified, one for each considered model sequence (figure 5, top). This diversity in folding nuclei results from using the MJ potential which biases the nucleation pattern towards the lowest energy (i.e. most stable) interactions. Nevertheless, the MJ nuclei share up to 33% of their native contacts with the Gō nucleus, and these common contacts are mostly determined by the native geometry. Indeed, their average energy is 25% higher than the average energy of the remaining contacts in the nuclei but they form with equally very high probability. Furthermore, we have observed that independently of protein sequence the beads forming the three folding nuclei are distributed along the protein chain in a very similar way (i.e. they occupy similar regions of the chain) (figure 5, bottom). As a consequence, the nucleation mechanism comprises the coalescence of two separated parts of the protein chain, which happens through the establishment of the long-range interactions corresponding to the non-local contacts that are common to all the three nuclei. These results are suggestive that the native geometry determines the distribution of the FN along the protein chain, but the specific location of nucleating residues is modulated by protein sequence.

3.8. Folding nucleus and the structure of the TS

In their influential work on the folding mechanism of CI2, Fersht and co-workers considered the FN as '*the best formed part of the native structure in the TS*' [19], suggesting a relation between the kinetic relevance of a residue and its degree of nativeness (i.e. the extent to which the residue is in its native environment) in the TS. We have recently explored this relation in the context of MC simulations of simple Gō model systems with different native geometries [80]. The adopted strategy was as follows: (i) determine the set of residues that lead to the largest increase in folding time upon mutation via a simulational proxy of the ϕ-value analysis, and assume that these residues (and associated contacts) constitute the FN, (ii) determine the TS as the ensemble of fast folding

conformations with $P_{fold} = 1/2$ and investigate the degree of nativeness of the putative nucleating residues in the TSE and (iii) determine the FN via the P_{fold} reaction coordinate as the set of native contacts (and associated residues) that exhibit the most dramatic changes between pre- ($P_{fold} = 0.05$) and true-TS conformations [83]. Results from (i) and (iii) show that the identification of the FN via the P_{fold} reaction coordinate agrees with the identification of the FN carried out with the mutational analysis. The overlap observed between the folding nuclei determined by both methodologies is particularly stronger for the native geometry dominated by non-local, long-range contacts. This is possibly a direct consequence of its smaller conformational plasticity (i.e. small number of alternative folding pathways accessible to the mutant), which makes it a more suitable target for ϕ-value analysis [81]. However, we have observed a very frail relation between the kinetic relevance of a residue and its degree of nativeness in the TSE. Indeed, the vast majority of the nucleating residues have all its native contacts formed in the TSE with a very small probability. As in [84] these findings suggest that the ϕ value correlates with the acceleration/deceleration of folding induced by mutation, rather than with the degree of nativeness of the TS.

4. The nucleation mechanism of specific proteins explored with off-lattice models

It is clear from the previous section that simple lattice models are extremely valuable tools to explore protein folding at a fundamental level. However, they are of no use in studies focused on specific proteins, which necessarily require more realistic representations. In this case off-lattice models, using the crystal structure of the protein as their basic input, are usually employed to explore folding. Also, since the lattice representation is a natural 'environment' to observe nucleation phenomena (e.g. formation of specific contacts), it is important to evaluate the universality of the nucleation mechanism by exploring the folding process beyond the lattice model. More precisely, it is important to investigate if a mechanism based on the formation of specific contacts is exclusive of the lattice representation, or if it is also observed off-lattice, where other mechanisms can *a priori* drive folding. The goal of this section is that of trying to provide an answer to this question by making an analysis of off-lattice computational studies addressing the nucleation mechanism. We stress, however, that this is not intended to be a comprehensive account. Rather, the proteins selected are preferentially those that have been studied by more that one research group, and by means of different models and methodologies, and also for which experimental data has been reported.

4.1. Acylphosphatase

A few years ago Vendruscolo and co-workers proposed a method to probe TS conformations that uses experimentally determined ϕ values as input data in off-lattice simulations of protein unfolding [85]. Since then it has been extensively used to investigate the TS of several target proteins [87, 88, 100, 93–95, 91, 97, 106].

The method's basic assumption is that the ϕ value of a residue can be interpreted as the fraction of native contacts it forms in a particular conformation of the TSE. The method's rationale is that by simultaneously constraining each simulational ϕ value, ϕ^{sim}, to its experimentally determined counterpart, ϕ^{exp}, an ensemble of conformations can be generated that corresponds to the folding TSE. In practice this idea is implemented by adding an extra energy term, which has typically the form of an harmonic restraint, $(\phi^{exp} - \phi^{sim})^2$, to the protein's energy. Briefly, the sampling procedure consists in driving the protein from its native conformation to one in which the restraints are satisfied (an extensive and very detailed account of the method, emphasizing its advantages and limitations, can be found in [87]).

The 'restraints method' was originally applied to study the TS of protein acylphosphatase (AcP) together with MC unfolding simulations of a coarse-grained C_α Gō model. AcP is a 98-residue protein, with an α/β architecture, which folds slowly with two-state kinetics. By using the whole set of experimentally determined ϕ values [89] as restraints, as well as many different subsets of the latter, it was found that only three out of the 24 ϕ^{exp}s are sufficient to determine the TS of AcP. Indeed, a very high correlation ($R = 0.86$) between ϕ^{exp} and ϕ^{sim} was obtained when only the largest ϕ^{exp} (Y11, P54 and F94) were used are restraints in the unfolding simulations (figure 6, top). Interestingly, this very result was subsequently confirmed in the context of MD simulations of a full atomistic model with an implicit solvent representation [90]. These investigations suggested a new definition for FN, namely that *the FN is the minimal set of experimentally determined ϕ values required for finding the TS*. In AcP the residues forming the nucleus establish a large number of native interactions in the TS—which are mainly long-ranged—and are enough to determine the overall fold of the protein. In other words, the members of the TSE have the same overall topology than the native structure, and the latter is determined by the FN. Subsequent studies by the same group have confirmed this observation for other target proteins [91, 92]. In particular, the correlation between their folding rates and the average contact order of their TSE was similar to that found for the contact order of the native state [92], which provides a mechanistic insight into the empirical observation that native topology is a key determinant of protein folding kinetics.

4.2. Chymotrypsin-inhibitor 2

Chymotrypsin-inhibitor 2, CI2, is a small protein with 64 amino acids arranged in an α-helix packed against six β-strands, which form an hydrophobic core (figure 7). It represents a paradigm for two-state protein folding kinetics, and it is a classic example of the nucleation–condensation (NC) mechanism of folding [21, 23].

By performing extensive ϕ-value analysis Itzhaki *et al* [19] found a broad distribution of ϕ values (ranging from 0 to 1) for CI2. The vast majority of amino acids in CI2 display ϕ values smaller than 0.5, and those for which $\phi \geqslant 0.5$ are exclusively localized in the α-helix. The β-strands 3 and 4
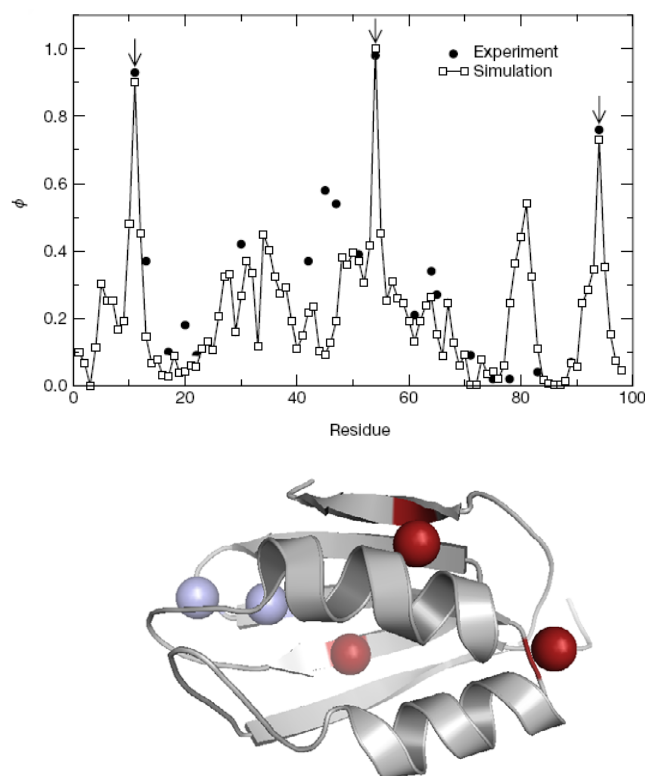
Figure 6. Comparison of ϕ^{exp} (circles) and ϕ^{sim} (squares) when only three experimentally determined ϕ values (marked with arrows) are used as restraints in MC unfolding simulations of protein AcP (top). The high correlation $R = 0.86$ between ϕ^{exp} and ϕ^{sim} suggests that the three highest ϕ values are sufficient to determine the overall structure of the TS. The native structure of AcP (1APS.pdb) where the beads indicate the C_α carbons of residues with $\phi^{\text{exp}} > 0.5$ (bottom). The C_α carbons of residues Y11, P54 and F94 are indicated as red (dark grey) beads. These residues and the extensive network of 28 long-range interactions they establish in the TS form the FN of AcP. The top figure is reprinted by permission from Macmillan Publishers Ltd: Nature [85], copyright (2001). The bottom figure was drawn with PyMOL [86].

display the higher ϕ values after the α-helix. The amino acid A16 ($\phi^{\text{exp}} = 1.1$), which establishes long-range interactions with L49 ($\phi^{\text{exp}} = 0.5$) and I57 ($\phi^{\text{exp}} = 0.1$), was considered of key importance in the folding of CI2.

Clementi and co-workers were amongst the first to explore the folding mechanism of CI2 with MD simulations of an off-lattice C_α model with protein energetics modelled by a Gō-type potential [99]. The evaluation of ϕ^{sim} values was made *exclusively* through free energy calculations. Furthermore, instead of considering the mutation of each residue they considered the mutation of each native contact (i.e. the removal/destabilization of each single native interaction), so that a ϕ^{sim} value was evaluated for each native contact. The fraction of native contacts Q was used as the reaction coordinate for folding and the TS was identified as the ensemble of conformations having $Q \sim 0.5$. By measuring the probability of formation of each native interaction in the TSE, it was found that the interactions formed with higher probability were those within the α-helix, between β-strands 4 and 5, and also the interactions between residues 32, 38 and

50. The computation of ϕ^{sim} values revealed three regions with ϕ value higher than 0.6, namely, the α-helix, the mini-core defined by β-strands 3 and 4 (and their connecting loop) and the regions between the C-terminus of β-strand 4 and the N-terminus of β-strand 5. Thus, the *in silico* ϕ-value analysis strengthens the importance of the α-helix, and of β-strands 3 and 4 in the folding of CI2. However, it is only for the α-helix that a clear relation can be established between ϕ^{exp} and the residue's degree of nativeness in the TS.

In a subsequent study, Li and Shakhnovich [100] used the 'restraints method', and 39 experimentally determined ϕ values, in unfolding MC simulations of a full atomistic protein model with interactions modelled by the Gō potential [54]. The goal was, of course, that of constructing an ensemble of conformations representative of the TS of CI2. By calculating the mean value of ϕ^{sim} relative to each element of secondary structure it was concluded that the α-helix is the most structured element of the TS ($\langle\phi^{\text{sim}}\rangle \sim 0.4$), followed by β-strand 3 ($\langle\phi^{\text{sim}}\rangle \sim 0.25$) and β-strand 4 ($\langle\phi^{\text{sim}}\rangle \sim 0.3$). Strands 1 and 6, on the other hand, showed $\langle\phi^{\text{sim}}\rangle < 0.1$, which was taken as evidence that these regions of the native fold are very little structured in the TS. Thus, this picture of the TS essentially corroborates the experimental data and, to some extent, the picture of the TS drawn by Clementi and colleagues.

Nevertheless, a rather interesting and perhaps surprising result that came out of the all-atom investigation was that *the ϕ value of a residue does not necessarily translates the residue's kinetic significance*. Li and Shakhnovich arrived at this conclusion by measuring the folding probability, P_{fold}, of two distinct classes of conformations. One such class was characterized for having the α-helix disrupted, while in the other one it was the β-strands 3 and 4 the structural elements that were simultaneously disrupted. Since the helix has a larger $\langle\phi\rangle$ than the strands it might be expected to find a lower average P_{fold} for the helix-disrupted conformations than for the beta-disrupted ones. However, the result turned out to be the other way round since the beta-disrupted states display a lower folding probability. Thus, it seems that in the folding of CI2 the main nucleation sites are the β-strands 3 and 4, despite their low $\langle\phi^{\text{exp}}\rangle$.

Further evidence supporting a major role played by the $\beta 3$–$\beta 4$ region in the folding of CI2 was recently reported by Kmiecik and Kolinski through MC simulations of the high resolution CABS lattice model [101]. By measuring the frequency of native contact formation it was found that the tertiary residual structure of the $\beta 3 - \beta 4$ hairpin persists at the folding transition temperature, which demonstrates their importance as key elements of TS stabilization.

4.3. Protein G

The B1 domain of streptococcal protein G, commonly referred to as protein G, is a small chain with 56 residues arranged in an α/β structure. It contains one α-helix packed against a four-stranded β-sheet (figure 8). The N-terminal strands and the first β turn ($\beta 1$-turn1-$\beta 2$) is generally termed hairpin 1, while the C-terminal strands together with the second β-turn ($\beta 3$-turn 2-$\beta 4$) is termed hairpin 2.
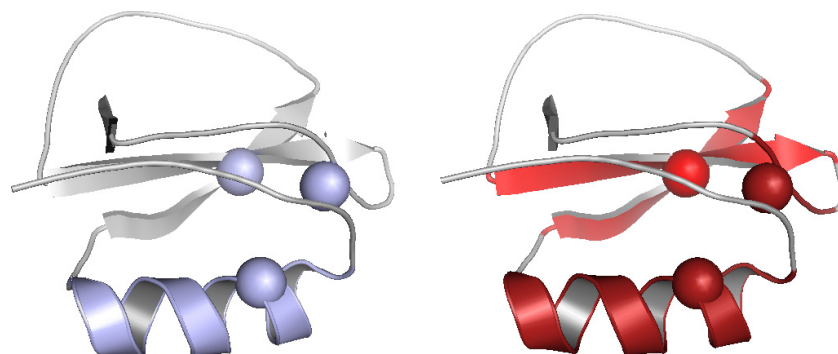
Figure 7. The nucleating regions of CI2 (1YPA.pdb) determined experimentally by ϕ-value analysis (left) and via MC simulation by the 'restraints method' (right). Beads indicate the C_α carbons of residues A16, L49 and I57. Residue A16, located in the α-helix, was considered a key element in nucleation due to its high $\phi^{exp} = 1.1$. Its interactions with residues L49 and I57 form the hydrophobic core and constitute the FN. The role of A16 as a key element of TS stabilization was confirmed through *in silico* investigations [100], which revealed a considerably high number ($\sim$30) of native interactions established by A16 in the TS (50% of these interactions are established with other residues also located in the α-helix and 13% are with L49 or I57). However, the same computational study also showed that β-strands 3 and 4 (highlighted in light red/light grey) have a more determinant kinetic role in folding than the α-helix despite its lower average ϕ^{exp}-value. This figure was drawn with PyMOL [86].
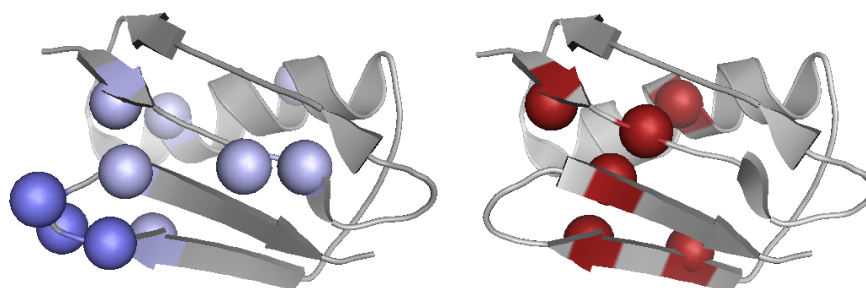

Figure 8. The native structure of protein G (1GBI.pdb) where the C_α carbons of the 10 residues with the $\phi^{exp} > 0.3$ (Y3, I6, L7, A26, Q32, Y45, D46, D47, T49, T51) are represented as beads. The three residues with highest ϕ values, located on the $\beta3$–$\beta4$ turn, are highlighted (left). The FN determined via molecular simulation through a combination of the 'restraints method' with the P_{fold} analysis (right). The six hydrophobic residues that constitute the computational FN are spread over the native fold: in hairpin 1 (Y3 and L5), in the α-helix (F30), and in hairpin 2 (W43, Y45 and F52). Except for Y3 ($\phi^{exp} = 0.38$), the other residues forming the FN have undetermined ϕ^{exp} or $\phi^{exp} \leqslant 0.3$. This figure was drawn with PyMOL [86].

Experimental investigations of the TS of protein G using ϕ-value analysis have pointed out an important role played by the residues located in the second hairpin in the stabilization of the TS [102]. In particular, the ϕ-value analysis highlights the role of residues D46 ($\phi^{exp} = 0.96$), T49 ($\phi^{exp} = 0.84$) and D47 ($\phi^{exp} = 0.67$).

While studying the folding mechanism of protein G with the all-atom MC approach [54], Shakhnovich and co-workers have recently pointed out that *interpreting the experimental ϕ value as the fraction of native interactions established by an amino acid in the TS, is unable to uniquely specify the TSE* [97]. This conclusion came out of the observation that conformations selected with the 'restraints method' display a P_{fold} distribution that is strongly bimodal, independently of the number of ϕ^{exp} values used as restraints. In other words, conformations constrained by ϕ^{exp} do not necessarily display $P_{fold} = 0.5$, and therefore are not necessarily *bona fide* members of the TS. Thus, while the 'restraints method' provides a remarkably straightforward process to construct a putative TSE for folding, a more accurate TSE determination will always require the evaluation of the

folding probability, P_{fold}, of each constrained conformation. In doing so, Shakhnovich and co-workers separated the putative TSE of protein G into three ensembles of conformations: the true TSE, containing conformations with $0.45 \leqslant P_{fold} \leqslant 0.55$, the pre-TS ensemble, with conformations having $P_{fold} < 0.4$, and the post-TSE whose conformations have $P_{fold} > 0.6$. The differential probability map (obtained from subtracting the pre-TS from the post-TS probability map) shows that hairpin 2—whose role as TS stabilizer is highlighted in ϕ-value analysis—actually begins forming before the TSE and hairpin 1 only starts developing at the TSE. Thus, at least according to simulations, the formation of hairpin 2 is not the rate-limiting step in folding. Indeed, what distinguishes TSE conformations from pre-TS conformations is an FN formed by six hydrophobic residues, namely L3 ($\phi^{exp} = 0.38$) and Y5 (ϕ^{exp} undetermined), both located in hairpin 1, F30 ($\phi^{exp} = 0.05$), located in the helix, and W43 (ϕ^{exp} undetermined), Y45 ($\phi^{exp} = 0.3$) and F52 ($\phi^{exp} = 0.19$), all pertaining to hairpin 2. Interestingly, this very result for the identification of the FN was recently recapitulated by Kolinski and co-workers, through a multiscale simulation approach based on the CABS model [98].
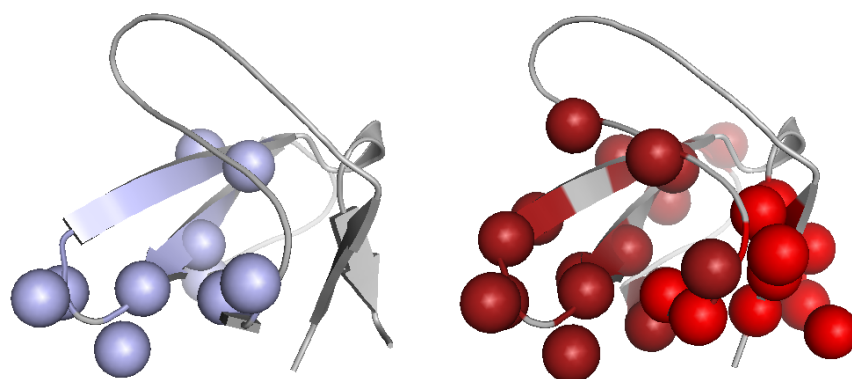
Figure 9. The native structure of the src-SH3 domain (1SRL.pdb) where the distal β-hairpin, a structural region of the protein formed by β-strands 3 and 4, and by the type I β-turn is shown in blue (left). The residues with $\phi^{\text{exp}} > 0.5$ are represented by spheres centred around the respective C_α carbons. The FN determined via computer simulations (left) with the 'restraints method' (represented in dark red/grey) [82] and the FN determined by using the maximum free energy as a criteria for locating the TS [83] (shown in light red/grey). This figure was drawn with PyMOL [86].

All in all, the computational investigations—not only with protein G but also with CI2—clearly suggest that a high ϕ value does not necessarily translate the importance of a residue for the folding TS, and that the recipe for identifying the FN as the set of residues with highest ϕ^{exp} may not always hold. The difficulty in using the experimentally determined ϕ values to conclude about the microscopic structural features of the TS is partly due to the fact that the ϕ value translates the ensemble average of the folding 'reaction' that is not always simple, but may involve several alternative routes to the native state. Indeed, this may actually be the case of protein G for which three folding pathways were recently observed in all-atom MC simulations [96].

4.4. src-SH3 domain

The src-SH3 is an all-β-sheet protein domain with 64 amino acids arranged in five anti-parallel β-sheets, orthogonally packed to form a single hydrophobic core (figure 9).

The distribution of the experimentally determined ϕ values reported by Baker and co-workers is suggestive of a highly polarized TS [103]. Indeed, 9 out of the 12 residues with $\phi^{\text{exp}} \geqslant 0.5$ are located in the distal β-hairpin, a structural region of the protein formed by β-strands 3 and 4, and by the type I β-turn that links them together. This fragment of the polypeptide chain, which extends between residues 43–57, contains four residues with $\phi^{\text{exp}} \sim 1$ (A45, S47, T50 and T51) that are clustered around the β-turn. The ϕ-value analysis suggests that the distal β-hairpin is the most ordered structural element in the TS of src-SH3, followed by the diverging turn together with β-strand 2, which also contains two residues with $\phi^{\text{exp}} \geqslant 0.5$ (figure 9, left). The other regions of the native structure display ϕ^{exp} values compatible with the protein being either unstructured, or having very little amount of native structure formed in the TS.

Folding simulations carried out by the same group, using the *ab initio* folding method ROSETTA, revealed that the most frequent contacts in conformations with fraction of native contacts $Q \sim 0.5$ are those forming the distal hairpin,

which supports the role of this substructural element as a TS stabilizer [103].

A picture of the TS compatible with that obtained via the ϕ-value analysis was reported by Clementi *et al* based on the coarse-grained computational approach employed for CI2 [99]. A following study by Gsponer and Caflisch also confirmed the picture of the TS obtained experimentally [104]. In their investigation Gsponer and Caflisch used the 'restraints method' with 18 experimentally determined ϕ values, representative of all the regions of the native fold. They selected 12 putative TS conformations from MD unfolding simulations of a standard all-atom force field which was combined with an implicit solvent representation. The evaluation of the folding probability of six of these conformations confirmed a $P_{\text{fold}} = 0.5$ for four of them, and $P_{\text{fold}} = 0.6$ and $P_{\text{fold}} = 0.4$ for the other two. This finding was taken as evidence that TS theory— a basic pillar of ϕ-value analysis—is valid, at least in the case of the src-SH3 domain. Interestingly, a finer analysis of the structural features of the 12 selected conformations revealed a significant number of non-native interactions established by residues I34 (unknown ϕ^{exp}), N37 (unknown ϕ^{exp}) and W43 ($\phi^{\text{exp}} = 0.15$). Non-native interactions were also found between the diverging turn, through residues E30 ($\phi^{\text{exp}} = 0.62$), R31 ($\phi^{\text{exp}} = 0.23$) and the distal hairpin, through residues S47 ($\phi^{\text{exp}} = 0.95$), L48($\phi^{\text{exp}} = 0.72$), S49 (unknown ϕ^{exp}) and T50($\phi^{\text{exp}} = 0.86$). By studying non-TS conformations, characterized for having no side-chain interactions in the central β-strand ($\beta2-\beta3-\beta4$), it was found that they were unable to fold fast, showing that in this case— and contrary to CI2—the most kinetically relevant residues are those located in the most structured region of the protein.

More recently, Shakhnovich and co-workers investigated the nucleation mechanism of the src-SH3 domain by means of the all-atom MC approach [82]. A putative TS for src-SH3 was constructed via the 'restraints method' where 10 experimental high ϕ values (mainly from residues localized on the $\beta3-\beta4$ hairpin) were used. The ensemble of putative TS conformations thus found was subsequently refined by calculating the folding probability of each conformation and
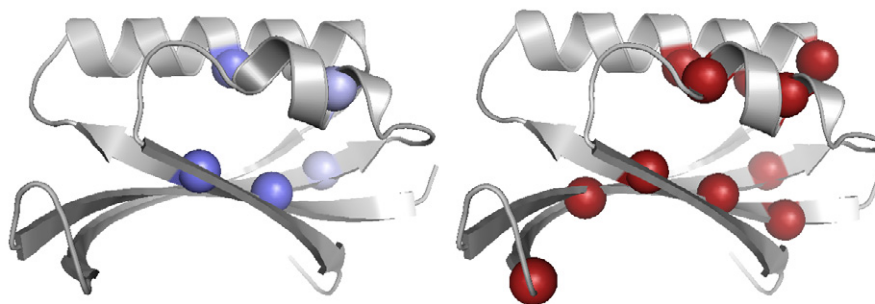
Figure 10. The native structure of protein S6 (1RIS.pdb), where the residues with $\phi^{\exp} > 0.3$ are represented as spheres centred around their C_α carbons. The three residues forming the experimentally determined FN are highlighted (left). The FN determined computationally via the 'restraints method' and P_{fold} analysis (right). This figure was drawn with PyMOL [86].

retaining only those for which $0.4 < P_{\text{fold}} < 0.6$. Interestingly, the average fraction of native contacts formed in these conformations ($Q = 0.17$) as well as their high C^α root mean square deviation (RMSD = 7.1 Å) are both indicative of a TS which occurs early on during folding. The average number of native contacts $\langle N \rangle$ established by each amino acid in the more accurately determined representative of the TS was evaluated, and the FN was defined as the set of amino acids with higher $\langle N \rangle$. Naturally, the FN includes all the residues whose $\phi^{\exp}$ were used as restraints (L44, S47, L48, T50, Q52, T53, Y55 and I56 together with E30 and L32). However, it also includes residues W43 ($\phi^{\exp} = 0.15$), A45 ($\phi^{\exp} = 1.20$) and T51 ($\phi^{\exp} = 1.06$) as well as residues L24 ($\phi^{\exp} = 0.26$) and F26 ($\phi^{\exp} = 0.40$), both from the RT loop. This FN of src-SH3 is particularly interesting in that it is not prototypical with regard to the hydrophobic content. Indeed, the protein's large hydrophobic residues are not localized in the nucleus but rather in regions which are very little structured in the TS. These observations are compatible with a folding mechanism where hydrophobic collapse occurs after the TS is crossed, making it a paradigm of a pure nucleation–condensation mechanism.

Finally, we refer to a study by Ding and co-workers where a multiscale computational approach was used to explore the TS of the src-SH3 domain [83]. In this investigation, instead of using the 'restraints method', a putative TS ensemble was constructed by selecting the highest free energy conformations of a free energy surface generated by importance sampling MD of a full atomistic solvated model. These conformations were subsequently subjected to the P_{fold} analysis in order to separate those with $0.4 < P_{\text{fold}} < 0.6$, which form the TS ensemble, from other spurious elements. However, due to its very high computational cost, the P_{fold} evaluation was carried out by means of discrete MD simulations of a coarse-grained Gō model [105]. The FN was defined as the set of residues involved in the contacts that exhibit '*the most dramatic changes between pre- and true-TS structures*', and it differs significantly from that evaluated with the 'restraints method'. Namely, it is composed of residues F10 ($\phi^{\exp} = 0.1$), V11 ($\phi^{\exp} = 0.03$) and A12 ($\phi^{\exp} = 0.05$), all located in β-strand 1; residues G29 ($\phi^{\exp} = 0.44$), R31 ($\phi^{\exp} = 0.23$) and L32 ($\phi^{\exp} = 0.22$), located in the diverging turn/β-strand 2, and also V61 ($\phi^{\exp} = -0.06$), A62 ($\phi^{\exp} = -0.02$), P63 (unknown $\phi^{\exp}$) and S64 ($\phi^{\exp} = 0.14$), all localized in β-strand 5. Again, this result supports the idea that the importance of a residue in the TS cannot be judged solely by its $\phi^{\exp}$.

4.5. Protein S6

The ribosomal protein S6 from *Thermus thermophilus* is a 101-residue long chain arranged in a four-stranded β-sheet packed against two α-helices with a hydrophobic core (figure 10).

The TS of protein S6 was investigated by Oliveberg *et al* [76] through ϕ-value analysis, which revealed a diffuse TS with a uniform distribution of fractional ϕ values (<0.52) and an FN centred around residues V6 ($\phi^{\exp} = 0.52$), I8 ($\phi^{\exp} = 0.46$) and I26 ($\phi^{\exp} = 0.40$). Thus, just like protein CI2 that displays a similarly diffuse nucleation pattern, one could expect protein S6 to be similarly robust against circular permutation. Remarkably, a circular permutant construct, termed P13-14, which cuts the protein backbone between residues 13 and 14 was shown to display a nucleation pattern substantially different from that of the WT protein. Indeed, with the exception of residues V6 and I8, which are common to both folding nuclei, the mutant shows a polarized FN whose core is shifted towards residues L75 ($\phi^{\exp} = 1.55$), V88 ($\phi^{\exp} = 1.10$) and V90 ($\phi^{\exp} = 0.70$), which were considerably unstructured in the TS of the WT protein.

In a recent account Shakhnovich and collaborators have applied the 'restraints method', in combination with the P_{fold} analysis, to explore the TS of protein S6 [106]. In this study the FN was defined as being the subset of long-range contacts common to *all* conformations in the ensemble representative of the TS. Together with residues V6, I8 and I26 this long-ranged FN also includes residues Y4, L30, F60, V65, L75 and L79 (all with $\phi^{\exp} < 0.4$) and residues Y33 and L48 (both with unknown $\phi^{\exp}$).

A subsequent study, involving a collaboration between the Shakhnovich group and the Oliveberg laboratory, investigated the TS and nucleation mechanism of five circular permutants of protein S6 [107], including P13-14. A striking result that came out from this investigation was the finding that the same residues of the WT nucleus (especially Y33, F60 and L75), although located at different positions in the sequence, play a predominant role in the nucleation mechanism of the mutants. What then explains the differences observed

between the several nucleation patterns? A considerably good anticorrelation ($R = -0.78$) was found between the average loop length of the contacts established by a residue and the corresponding change in the residue's ϕ^{exp} value upon circular permutation [108]. Accordingly, increased sequence separation between interacting residues reduces their contribution to the stability of the TS. Based on this premise it was suggested that the WT FN is mainly altered by the extent to which the sequence separation between its residues is changed when linking the N- and C-termini of the protein [107]. However, the change in sequence separation is itself dependent on the mutant's native geometry, and more specifically on its relative content in local and long-ranged contacts. If the mutant's content in local contacts is higher than that of the WT protein (as in P13-14 and P68-69), then it is likely that part of these less entropically costly contacts will be recruited to be part of the mutant's FN, leading to a pronounced change in the nucleation pattern. If, in contrast, the mutant's native geometry is such that its content in long-ranged contacts is equal to or higher than that of the WT protein, it is not at all clear that a substitution (even if partial) of the WT FN will be energetically advantageous [105, 109].

5. Summary and outlook

The folding kinetics of small, single-domain proteins is remarkably well described by a simple single exponential process. This observation suggests that protein folding rates can be explained by TS theory (TST), developed by Eyring in the 1930s for elementary chemical reactions. Accordingly, the native state is separated from the unfolded ensemble by a free energy barrier on the top of which stays the highly energetic TS (TS). The nucleation mechanism of protein folding proposed by Shakhnovich and co-workers in the 1990s has put forward the idea that the folding TS corresponds to the formation of a specific set of native contacts, termed folding nucleus (FN), after which the native structure is achieved promptly. A great many simulations of protein folding on the lattice have such a microscopic definition of FN as the starting point and seek for strategies to identify it. One such strategy is based on the concept of folding probability P_{fold}—which is the equivalent of the transmission coefficient in TST—and on the operational definition of the TS as the ensemble of conformations with $P_{fold} = 0.5$. The picture of nucleation that emerges from lattice investigations is typically centred on the existence of a single FN with microscopic heterogeneity where the non-local contacts, associated with long-range interactions between the amino acids, are dominant. The lattice model approach allows the exploration of fundamental features of the nucleation mechanism, such as the interplay between native geometry and protein sequence in producing a certain nucleation pattern.

The vast majority of *in silico* investigations of nucleation using off-lattice protein representations are based on a rather different strategy to identify the FN, and have been largely framed in studies of TS structure based on the use of ϕ-value analysis. Indeed, the off-lattice simulations typically use the experimental ϕ values as input data, and the classic interpretation of the latter as a measure of the degree of nativeness of a residue in the TS. Interestingly, this approach to the study of the nucleation mechanism was deeply influenced by the idea—originally proposed by Ferhst and co-workers in their seminal study on the structure of the TS of CI2— that the FN is the part of the native structure that is the most well formed in the TS. This idea suggests the existence of a correlation (or at least a relation) between the structure of the TS—which is itself approximated by the ϕ value—and the kinetic relevance of a residue. Such a relation between structure and kinetics does not necessarily hold, as several simulation studies have shown. Furthermore, simulation studies based on the use of P_{fold} have also shown that the classical structural interpretation of ϕ values does not necessarily identify TS conformations, which may be due to the complexity of the folding reaction and to the existence of a large number of pathways leading to the native state.

Despite the difficulties associated with using and interpreting ϕ values, it is clear that a more complete picture of the folding mechanism has been achieved by combining experimental and simulational data. The study of the nucleation mechanism represents indeed a paradigmatic example of the importance of molecular simulations as a research tool that leads to a deeper understanding of a process through the generation of new data complements that obtained via real-world investigations. Nevertheless, it would be interesting to carry out further explorations of the nucleation mechanism of protein folding with off-lattice simulations by using alternative approaches which do not require the use of ϕ^{exp}-value data. For example, it would be interesting to explore up to which extent coarse-grained off-lattice simulations are able to capture the kinetic importance of a residue by performing systematic mutations and measuring the associated folding rate change. As far as we know this was never explored before, at least systematically, and it could give insight into the existence of nucleation phenomena in a unbiased manner. In a step ahead it would also be very interesting to compute the simulational ϕ values by adopting *in silico* a procedure identical to that carried out in investigations with real-world proteins.

Acknowledgments

References

[1] Dill K A, Ozkan B, Weikl T R, Chodera J D and Voelz V A 2007 *Curr. Opin. Struct. Biol.* **17** 342
[2] Service R F 2008 *Science* **321** 784
[3] Anfinsen C 1973 *Science* **181** 223
[4] Levinthal C 1969 *Mossbauer Spectrosc. Biol. Syst. Proc.* **67** 22
[5] Wetlaufer D 1973 *Proc. Natl Acad. Sci. USA* **70** 697
[6] Baldwin R L and Rose G D 1999 *Trends Biochem. Sci.* **24** 26
[7] Karplus M and Weaver D L 1978 *Biopolymers* **18** 1421
[8] Jackson S E and Fersht A R 1991 *Biochemistry* **43** 10428
[9] Jackson S E 1998 *Fold. Des.* **3** R81

[10] Binder K 1987 *Rep. Prog. Phys.* **50** 783

[11] Chen J, Bryngelson J D and Thirumalai D 2008 *J. Phys. Chem.* B **112** 16115

[12] Tsong T Y, Baldwin R, McPhie P and Elson E L 1972 *J. Mol. Biol.* **63** 453

[13] Abkevich V I, Gutin A M and Shakhnovich E I 1994 *Biochemistry* **33** 10026

[14] Chan H S, Shimizu S and Kaya H 2004 *Methods Enzymol.* **380** 350

[15] Hao M H and Scheraga H A 1994 *J. Phys. Chem.* **98** 4940

[16] Makarov D E and Plaxco K W 2003 *Protein Sci.* **12** 17

[17] Ivarsson Y, Travaglini-Allocatelli C, Brunori M and Gianni S 2008 *Eur. Biophys. J.* **37** 721

[18] Fersht A 1998 *Structure and Mechanism in Protein Science: a Guide to Enzyme Catalysis and Protein Folding* 3rd edn (San Francisco, CA: Freeman) pp 560–3

[19] Itzhaki L S, Otzen D E and Fersht A R 1995 *J. Mol. Biol.* **245** 260

[20] Weikl T R and Dill K A 2006 *J. Mol. Biol.* **365** 1578

[21] Fersht A R 1995 *Proc. Natl Acad. Sci. USA* **92** 10869

[22] Nolting B and Andert K 2000 *Proteins* **41** 288

[23] Nolting B and Agard D A 2008 *Proteins* **73** 754

[24] Daggett V and Fersht A 2003 *Nature Rev. Mol. Cell. Biol.* **4** 497

[25] Sánchez I E and Kiefhaber T 2003 *J. Mol. Biol.* **334** 1077

[26] Settanni G, Rao F and Caflisch A 2005 *Proc. Natl Acad. Sci. USA* **102** 628

[27] De Los Rios M A, Muralidhara B K, Wildes D, Sosnick T R, Marqusee S, Wittung-Stafshede P, Plaxco K W and Ruczinski I 2006 *Protein Sci.* **15** 53

[28] Raleigh D P and Plaxco K W 2005 *Protein Pep. Sci.* **12** 117

[29] Chan H S, Kaya H and Shimizu S 2002 Computational methods for protein folding: scaling a hierarchy of complexities *Current Topics in Computational Molecular Biology* ed T Jiang, Y Xu and M Q Zhang (Cambridge, MA: MIT Press) 16 pp 403–447

[30] Caflisch A and Paci E 2004 *Molecular Dynamics Simulations to Study Protein Folding and Unfolding* (Weinheim: Wiley-VCH) chapter 32, pp 1143–69

[31] Tozzini V 2004 *Curr. Opin. Struct. Biol.* **15** 144

[32] Snow C D, Sorin J E, Rhee Y M and Pande V S 2005 *Annu. Rev. Biophys. Biomol. Struct.* **34** 43

[33] Clementi C 2008 *Curr. Opin. Struct. Biol.* **18** 10

[34] Hills R D and Brooks C L 2008 *Int. J. Mol. Sci.* **10** 889

[35] Go N and Taketomi H 1978 *Proc. Natl Acad. Sci. USA* **75** 559

[36] Miyazawa S and Jernigan R L 1985 *Macromolecules* **18** 534

[37] Shakhnovich E and Gutin A 1993 *Proc. Natl Acad. Sci. USA* **90** 7195

[38] Dill K A 1985 *Biochemistry* **24** 1501

[39] Gutin A and Shakhnovich E 1990 *J. Chem. Phys.* **93** 5967

[40] Metropolis N, Rosenbluth A W, Rosenbluth M N, Teller A H and Teller E 1953 *J. Chem. Phys.* **21** 1087

[41] Verdier P H and Stockmayer W H 1962 *J. Chem. Phys.* **36** 227

[42] Landau D P and Binder K 2000 *A Guide to Monte Carlo Simulations in Statistical Physics* (Cambridge: Cambridge University Press) pp 122–3

[43] Cieplak M and Hoang T X 2002 *Int. J. Mod. Phys.* C **13** 1231

[44] Wallin S and Chan H S 2006 *J. Phys.: Condens. Matter* **18** S307

[45] Prieto L, de Sancho D and Rey A 2005 *J. Chem. Phys.* **123** 154903

[46] Pande V S, Baker I, Chapman J, Elmer S P, Khaliq S, Larson S M, Rhee Y M, Shirts M R, Snow C, Sorin E and Zagrovic B 2003 *Biopolymers* **68** 91

[47] Rhee Y M and Pande V S 2003 *Biophys. J.* **84** 775

[48] Snow C D, Nguyen H, Pande V S and Gruebele M 2002 *Nature* **420** 102

[49] Berry S, Rice S A and Ross J 2002 *Physical and Chemical Kinetics* (Oxford: Oxford University Press) chapter 30, p 911

[50] Hubner I A, Deeds E J and Shakhnovich E I 2005 *Proc. Natl Acad. Sci. USA* **102** 18914

[51] Faisca P F N, Gama M M T and Ball R C 2004 *Phys. Rev.* E **69** 051917

[52] Faisca P F N and Plaxco K W 2006 *Protein Sci.* **15** 1608

[53] Prieto L and and Rey A 2007 *J. Chem. Phys.* **127** 175101

[54] Shimada J, Kussell E L and Shakhnovich E I 2001 *J. Mol. Biol.* **308** 79

[55] Klimov D K and Thirumalai D 1998 *J. Mol. Biol.* **282** 471

[56] Shakhnovich E I 1998 *Fold. Des.* **3** 108

[57] Li L, Mirny L A and Shakhnovich E I 2000 *Nat. Struct. Biol.* **7** 336

[58] Sosnick T R, Krantz B A, Dothager R S and Baxa M 2006 *Chem. Rev.* **106** 1862

[59] Krantz B A, Dothager R S and Sosnick T R 2004 *J. Mol. Biol.* **337** 463

[60] Kazmirski S L and Daggett V 1998 *J. Mol. Biol.* **284** 793

[61] Li L, Mirny L A and Shakhnovich E I 2000 *Nat. Struct. Biol.* **7** 336

[62] Klimov D K and Thirumalai D 1998 *Fold. Des.* **3** 127

[63] Du R, Pande V S, Grosberg A Y, Tanaka T and Shakhnovich E I 1998 *J. Chem. Phys.* **108** 334

[64] Snow C and Pande V S 2006 *Biophys. J.* **91** 14

[65] Cho S S, Levy Y and Wolynes P G 2006 *Proc. Natl Acad. Sci. USA* **103** 586

[66] Chang I, Cieplak M, Banavar J R and Maritan A 2004 *Protein Sci.* **13** 244

[67] Pacci E, Vendruscolo M and Karplus M 2002 *Proteins* **47** 379

[68] Treptow W L, Barbosa M A A, Garcia L G and Pereira de Araujo A F 2002 *Proteins* **49** 167

[69] Clementi C and Plotkin S 2004 *Protein Sci.* **13** 1750

[70] Zarrine-Afsar A, Wallin S, Neculai A M, Neudecker P, Howell P L, Davidson A R and Chan H S 2008 *Proc. Natl Acad. Sci. USA* **105** 9999

[71] Plaxco K W, Simmons K T, Ruczinski I and Baker D 2000 *Biochemistry* **39** 11177

[72] Gromiha M M and Selvaraj S 2002 *J. Mol. Biol.* **310** 27

[73] Li L and Shakhnovich E I 2001 *J. Mol. Biol.* **306** 121

[74] Viguera A R, Serrano L and Wilmanns M 1996 *Nat. Struct. Biol.* **3** 874

[75] Grantcharova V P and Baker D 2001 *J. Mol. Biol.* **306** 555

[76] Lindberg M, Tangrot J and Oliveberg M 2002 *Nat. Struct. Biol.* **9** 818

[77] Otzen D E and Fersht A R 1998 *Biochemistry* **37** 8139

[78] Clementi C, Jennings P A and Onuchic J N 2001 *J. Mol. Biol.* **311** 879

[79] Travasso R D M, Faisca P F N and Gama M M T 2007 *J. Phys.: Condens. Matter* **19** 285212

[80] Faisca P F N, Travasso R D M, Ball R C and Shakhnovich E I 2008 *J. Chem. Phys.* **129** 095108

[81] Faísca P F N and Gomes C M 2008 *Biophys. Chem.* **138** 99

[82] Hubner I A, Edmonds K A and Shakhnovich E I 2005 *J. Mol. Biol.* **349** 424

[83] Ding F, Guo W, Dokholyan N V, Shakhnovich E I and Shea J E 2005 *J. Mol. Biol.* **350** 1035

[84] Ozkan S B, Bahar I and Dill K A 2001 *Nat. Struct. Biol.* **8** 765

[85] Vendruscolo M, Paci E, Dobson C M and Karplus M 2001 *Nature* **409** 641

[86] DeLano W L 2002 *The PyMOL Molecular Graphics System* (Palo Alto, CA: DeLano Scientific)

[87] Allen L R and Paci E 2007 *J. Phys.: Condens. Matter* **19** 285211

[88] Gsponer J and Caflisch A 2002 *Proc. Natl Acad. Sci. USA* **99** 6719

[89] Chiti F, Taddei N, White P M, Bucciantini M, Magherini F, Stefani M and Dobson C M 1999 *Nat. Struct. Biol.* **6** 1005

[90] Paci E, Vendruscolo M, Dobson C M and Karplus M 2002 *J. Mol. Biol.* **324** 151

[91] Lindorff-Larsen K, Vendruscolo M, Paci E and Dobson C M 2004 *Nat. Struct. Mol. Biol.* **11** 443

[92] Paci E, Lindorff-Larsen K, Karplus M, Dobson C M and Vendruscolo M 2005 *J. Mol. Biol.* **352** 495

[93] Paci E, Clarke J, Steward A, Vendruscolo M and
     Karplus M 2003 *Proc. Natl Acad. Sci. USA* **100** 394
[94] Lindorff-Larsen K, Paci E, Serrano L, Dobson C M and
     Vendruscolo M 2003 *Biophys. J.* **85** 1207
[95] Paci E, Friel C T, Lindorff-Larsen K, Radford S E,
     Karplus M and Vendruscolo M 2004 *Proteins* **54** 513
[96] Shimada J and Shakhnovich E I 2004 *Proc. Natl Acad. Sci.
     USA* **99** 11175
[97] Hubner I A, Shimada J and Shakhnovich E I 2004 *J. Mol.
     Biol.* **336** 745
[98] Kmiecik S and Kolinski A 2008 *Biophys. J.* **94** 726
[99] Clementi C, Nymeyer H and Onuchic J N 2000 *J. Mol. Biol.*
     298 937
[100] Li L and Shakhnovich E I 2001 *Proc. Natl Acad. Sci. USA*
     98 13014
[101] Kmiecik S and Kolinski A 2007 *Proc. Natl Acad. Sci. USA*
     104 12330

[102] McCallister E L, Alm E and Baker D 2000 *Nat. Struct. Biol.*
     7 669
[103] Riddle D S, Grantcharova V P, Santiago J, Alm E,
     Ruczinski I and Baker D 1999 *Nat. Struct. Biol.* **6** 1016
[104] Gsponer J and Caflisch A 2002 *Proc. Natl Acad. Sci. USA*
     99 6719
[105] Dokholyan N V, Buldyrev S V, Stanley H E and
     Shakhnovich E I 2000 *J. Mol. Biol.* **296** 1183
[106] Hubner I A, Oliveberg M and Shakhnovich E I 2004 *Proc.
     Natl Acad. Sci. USA* **101** 8354
[107] Hubner I A, Lindberg M, Haglund E, Oliveberg M and
     Shakhnovich E I 2006 *J. Mol. Biol.* **359** 1075
[108] Lindberg M O, Haglund E, Hubner I A, Shakhnovich E I and
     Oliveberg M 2006 *Proc. Natl Acad. Sci. USA* **103** 4083
[109] Baker D 2000 *Nature* **405** 39