Fishing for folding nuclei in lattice models and proteins

D Thirumalai[1,2] and DK Klimov[2]

Systematic studies of kinetics using minimal protein models reveal multiple folding nuclei for sequences that reach the native state in a single step. The diversity of the folding nuclei depends on sequence and topology.

Addresses: [1]Institute for Advanced Studies, Hebrew University of Jerusalem, Givat Ram, Jerusalem 91904, Israel. [2]Institute for Physical Science and Technology, University of Maryland, College Park, MD 20742, USA.

Correspondence: D Thirumalai
E-mail: thirum@glue.umd.edu

Folding & Design 15 December 1998, **3**:R112–R118
http://biomednet.com/elecref/13590278003R0112

Introduction

The mechanisms by which proteins fold to their native states (*in vitro*) are beginning to be elucidated thanks to simultaneous advances in theoretical methodology [1] and experimental innovations [2,3]. These developments have been used to unravel the processes by which relatively small proteins, many of which apparently fold rapidly in a single kinetic step, reach their native states. The efficient and rapid access of the native state has been rationalized in terms of a nucleation-collapse (NC) mechanism. The essential idea that nucleation processes may explain rapid folding of proteins was advanced 25 years ago by Wetlaufer [4]. Experiments [5], theoretical arguments [6–8], and computer simulations using lattice models [9,10] have sharpened the questions concerning the nature of the NC mechanism in proteins. It is the purpose of this commentary to assess the current status of our understanding of the NC mechanism with special emphasis on the controversies associated with the lattice model description of the transition states (TSs) in the NC process. In particular, we show the following.

Firstly, in precisely defined lattice models, the analysis of nucleation sites reveals that there is a distribution of native contacts in the folding nuclei with some occurring at a higher probability than others [10]. The data do not support the idea of a specific nucleus [9]. Secondly, although precise measurements of the distribution of TS structures are difficult, we argue that the experiments are broadly consistent with the idea of a distribution of folding nuclei. The difference between α-helical and certain β-sheet proteins seems to be in the width of the transition region, which is a measure of how plastic the TS structures are.

Nucleation mechanism in two-state folders

We have shown [11–13] that, in general, candidate sequences that fold entirely by the NC mechanism are characterized by small values of σ where $\sigma = (T_\theta - T_F)/T_\theta$. T_θ is the temperature at which there is a transition from extended states to a compact phase, and T_F is the folding transition temperature below which the native basin of attraction (NBA) is predominantly populated. When σ is small, the processes of collapse and the acquisition of the native state occur almost simultaneously. As such, the formation of secondary structures, backbone formation, and sidechain packing cannot easily be distinguished. Thus, for two-state proteins which follow the NC mechanism, the major questions are: What are the characteristics of the folding nuclei which are the transition states in the single-step folding reaction $U \rightleftharpoons N$? And what is the width of the associated transition region? The answers to these questions would be simple if one had a reasonable assessment of the underlying reaction coordinate, but the large dimensionality accompanying the folding process makes it difficult to unambiguously determine a simple folding coordinate. In the absence of a suitable reaction coordinate, the answers to the questions posed above have been explored using kinetic simulations of simple models. Three models have been proposed in connection with the nature of the folding nuclei and hence the associated characteristics of transition states.

The specific folding nucleus (SFN) model. Abkevich, Gutin and Shakhnovich (AGS) have stated that “Formation of a specific nucleus, which is a particular pattern of contacts, is shown to be a necessary and sufficient condition for subsequent rapid folding to the native state. The nucleus represents a transition state of folding to the molten globule conformation” [9]. AGS define a nucleus as a “set of contacts that satisfies the following two conditions: (i) Formation of a nucleus is a sufficient condition for folding: that is, after a set of contacts that constitutes the nucleus is formed, the subsequent folding is guaranteed and is very fast (in our search for a nucleus we required that folding should take place in less than 50000 MC steps after the nucleus is formed). We are therefore looking for postcritical nuclei. (ii) Formation of a nucleus is a necessary condition for folding; that is, 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” [9].

The multiple folding nuclei (MFN) model. We (Klimov and Thirumalai; KT), have stated that “The theoretical and experimental studies give a coherent picture of the NC mechanism in which there is a distribution of folding nuclei with some more probable than others" [10]. We (KT) define a folding nucleus to be "a set of native contacts, which (i) consists of a minimal number of stable contacts (stability condition), and (ii) results in rapid assembly of the native conformation (kinetic condition). By 'minimal' and 'stable' we imply that the nucleus consists of the smallest number of contacts, which survive until reaching the native state. By 'rapid assembly' we mean that, when the nucleus is formed, the native state must be reached within the time of less than $\delta \times \tau_{1i}$, where τ_{1i} is the first passage time for the trajectory *i*" [10].

The weak form of the specific folding nucleus (WFSFN) model. In the accompanying commentary [14], it is stated "In particular, a specific folding nucleus (SFN) scenario was found whereby passing through the transition state with subsequent rapid assembly of the native conformation requires formation of some (small) number of specific obligatory contacts (specific nucleus). According to the SFN scenario, assembly of those obligatory contacts results in rapid folding to the native state." And "Clearly each nucleation conformation (i.e. a conformation that contains a folding nucleus) always also features other, optional native contacts, besides nucleation contacts. Specific nucleus contacts appear simultaneously in nucleation conformations with high probability, however, whereas each optional contact occurs with low probability and their number and location in structure may vary between nucleation conformations".

Even the casual reader will notice the great similarity between the MFN and WFSFN models. Both these models assert that in the folding nuclei, certain contacts are "more probable than others" or "appear in nucleation conformations with high probability". The MFN model suggests that, depending on the sequence, topology, external conditions, or potentials employed in simulations, different degrees of diversity among folding nuclei may be observed. We have arrived at similar conclusions by analyzing nucleation trajectories in off-lattice models [8]. Thus, generically, we expect that there is diversity in the folding nuclei that gives rise to a heterogeneous distribution of contacts in the TS ensemble. There are cases, though, in which a smaller degree of diversity is observed [15,16]. These models are to be contrasted with the SFN model, which asserts that the formation of a specific, small set of contacts has to occur with probability 1 (unity) in each and every molecule in the ensemble of denatured states in order for folding to the native state to take place. It is this stringent requirement (the necessary condition) that conceptually distinguishes SFN and MFN, and this is by no means a merely semantic difference as suggested in [14].

It is worth noting that in the initial proposal of SNF [9], AGS suggested that their model applies only to the formation of the molten globule conformation. Apparently, this idea was generalized (without additional theoretical analysis) to include the formation of the native sidechain packing (see the footnote on page 284 of [5]).

Searching for folding nuclei in lattice models

Seeking SFN in the AGS [9] sequences

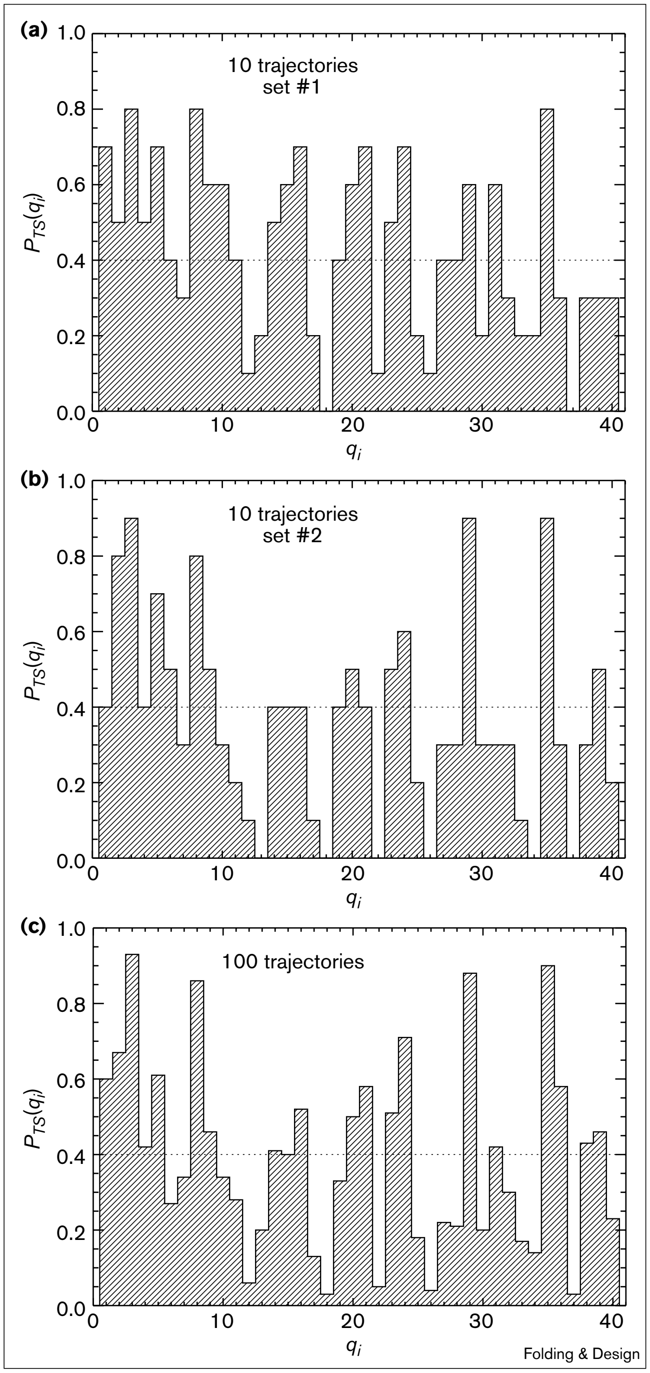
Before addressing the technical points raised in [14] we would like to highlight a number of issues concerning the determination of SFN in [9].

1. AGS performed Monte Carlo (MC) simulations on 36-mer lattice proteins using what appears to be a variant of the contact interaction energies derived from Miyazawa and Jernigan (MJ) potentials [17]. Unfortunately, the exact method of getting these energies is not clearly explained in [9]. In particular, it is not clarified which table of MJ potentials has been used. AGS used only 10 MC trajectories and searched for nucleation contacts in the final 50000 MC steps. The apparent mean first passage time for folding is ~10^6 MC steps, which implies that δ (see KT and definition of MFN given above) is roughly 0.05. In [14] it is nevertheless asserted that a value of $\delta \approx 0.02$ was used. Because the reaction coordinate is unknown and the relaxation time to the native state after crossing the putative barrier is not calculable, it is necessary to vary δ, as was done by KT, to check the robustness of the conclusions to altering the kinetic requirement (see the definition of MFN) for nucleation. Thus, the fixing of δ to 0.02 to 0.05 is not sacrosanct.

2. It is curious that the specific nucleus identified by AGS (in Figure 1a of [9]) consists mostly of charged or hydrophilic residues. In particular, at least two of the residues in the 'core' of the native structure, namely D and K, are charged. Whether the presence of such 'non-physical' contacts (which do occur in the interior of some proteins but not with the preponderance that AGS observe) in the core is due to the modification of the MJ potentials [17] introduced by AGS is not clear. The eigenvalue decomposition of the MJ potentials clearly shows that they are dominated by two large eigenvalues with little discrimination between the various residues (R Elber, personal communication). In this sense, the MJ potentials [17] are only marginally different from the HP model [1]. Hence, it is not obvious whether the non-physical set of residues constituting the 'specific' nucleus in the AGS sequence is a consequence of the MJ interaction potentials themselves. It is for this reason that KT used various interaction potentials to check the robustness of the MFN.

3. Most importantly, inferring any conclusions from 10 runs, as was done by AGS, is inadequate. When only 10 trajectories are used to identify the folding nuclei, it is quite conceivable that occasionally few contacts would appear with nearly 100% probability in folding nuclei.

Figure 1



Histograms of the probabilities of finding contact q_i in folding nuclei $P_{TS}(q_i)$ for a 36-mer with $\delta = 0.02$, averaging over **(a,b)** two different randomly picked sets of 10 trajectories and **(c)** averaging over 100 trajectories. The probability of concurrent occurrence of the most probable contacts 3, 8, 29, 35 varies between 0.5 (a) to 1.0 (in certain ‘handpicked’ sets of 10 trajectories). The ‘true’ probability calculated from the analysis of 100 trajectories is 0.8 (c).

However, meaningful results may be obtained only when substantially larger sampling of trajectories is performed. The importance of sampling more than 10 MC trajectories is most vividly illustrated in Figure 1, in which we plot the probability of occurrence of various native contacts for the 36-mer using $\delta = 0.02$ (the preferred value in [14]). Figure 1a,b shows $P_{TS}(q_i)$ for two sets of 10 randomly selected trajectories, whereas Figure 1c gives the distribution of contacts for 100 trajectories, which we think is barely representative of the ensemble of random coils. The dramatic differences shown in Figure 1 demonstrate clearly the inadequacy of sampling in the simulations of AGS. The insufficient sampling of the ensemble of denatured conformations as well as lack of verification that the results do not significantly depend on the interaction potentials lead us to conclude that the calculations reported by AGS are technically flawed.

Dissecting the nucleation-collapse mechanism in KT 27-mers

It appears (see the definition of SFN and MFN given above) that the criteria used by AGS and KT to identify the contacts that are part of the folding nuclei are similar. The criteria for folding nuclei together with the pattern recognition algorithm were used to identify folding nuclei in 27-mers and the 36-mer. The findings for 27-mers have been dismissed in [14]. There were basically four reasons given:

1. AGS apparently demand that, in addition to the criteria satisfied (as per the definition given in SFN), the following, most crucial kinetic condition should also be met. If folding is initiated from an ensemble of molecules in which the identified nuclei are constrained but the rest of the chain is denatured (i.e. the putative nucleus conformation (PNC) is preformed) then extremely rapid folding should ensue. We address this criterion below when discussing the 36-mer.

2. “It is possible that the 27-mers studied by KT did not fold via a nucleation mechanism. An indication of this is that the distribution of contact frequencies in PNCs reported by KT for 27-mers did not depend on the parameter δ” [14]. This assertion is wrong. Remember that the smaller δ is, the closer a given molecule is to the native conformation. In Table 1, we present the probability of simultaneous occurrence of the most probable contacts in all 100 trajectories as a function of δ for the wild-type (WT) 27-mer and the 36-mer [10]. Clearly there is a strong dependence on δ for the 27-mer as well as the 36-mer. In fact, the rate of convergence of this probability to unity as $\delta \to 0$ is initially larger for the 27-mer than for the 36-mer.

3. “Also the distribution of contact frequencies in the PNCs for 27-mers was quite broad” [14]. Assuming $\delta = 0.02$, the ratio of the dispersion in the nucleus size (measured by the number of contacts) to the mean nucleus size is 0.39 for the WT 27-mer of KT, whereas for the 36-mer this ratio is 0.66. Thus, the dispersion (measuring the

width of the TS) is larger for the 36-mer than the 27-mer — in contrast to the expectation in [14].

This not withstanding, we should emphasize that the width of the transition region is a function of topology and the external conditions. Theoretical arguments [18,19] and computations by KT and others [20] suggest that, in general, the transition region is broad. It is only within the SFN model that the width of the transition region has to be narrow so that one is, for all practical purposes, dealing with a unique TS. A broad distribution of nucleus size in the TS does not imply that the NC mechanism is not obeyed. Rather, it implies that the SFN model is not valid.

4. It is alleged that perhaps the 27-mers studied by KT follow three-stage kinetics. The three-stage multipathway mechanism (TSMM) was first proposed by Camacho and Thirumalai using explicit kinetic simulations [13]. A very direct consequence of TSMM (and presumably the three-stage random search mechanism as well) is that folding occurs in stages so that the time-dependent decay of the fraction of unfolded molecules $P_u(t)$ is best fit by a sum of at least two exponentials. But this is not what is found for the 27-mers studied by us. We find that a single exponential function fits $P_u(t)$ for all values of t (see Figure 3 in [10]), which is consistent with the nucleation mechanism.

Formation of the most probable nucleation contacts in the denatured states does not greatly accelerate folding rates

According to [14], the *sino quo non* of the nucleation mechanism is that if the most probable native contacts (remember that in the SFN model such contacts must occur with probability 1) are preformed in the ensemble of denatured states then the polypeptide chain will fold extremely rapidly to the native state. In the 10 runs done by AGS, such preformed nucleus conformations reach the native state in less than about 3% of the mean folding time (see [9]). There are a couple of comments that should be made about this requirement. Firstly, this third criterion (in addition to the two given in the definition of SFN and MFN), referred to as the ‘kinetic condition’ in [14], is not mandated in the classical theory of nucleation from melt in real materials. For example, in the familiar case of crystallization of, say, argon the procedure for identifying the putative nuclei does not involve subjecting them to the ‘kinetic condition’ test [21,22]. In fact, the virtue of classical nucleation theories is that much can be predicted from thermodynamic considerations alone just as for proteins [18,19]. Secondly, general arguments [23,24] suggest that preformed nuclei in the denatured states are not beneficial and may indeed lead to increased folding times. According to Fersht and coworkers [5] “First, is that the nucleation site does not need to be extensively preformed in the denatured state. A theoretical analysis of the optimization of the rates of protein folding suggests that the less it is formed in the denatured state, the better [23].”

Table 1

Probabilities of concurrent occurrence of four most probable contacts as a function of δ* for 27-mers and 36-mers.

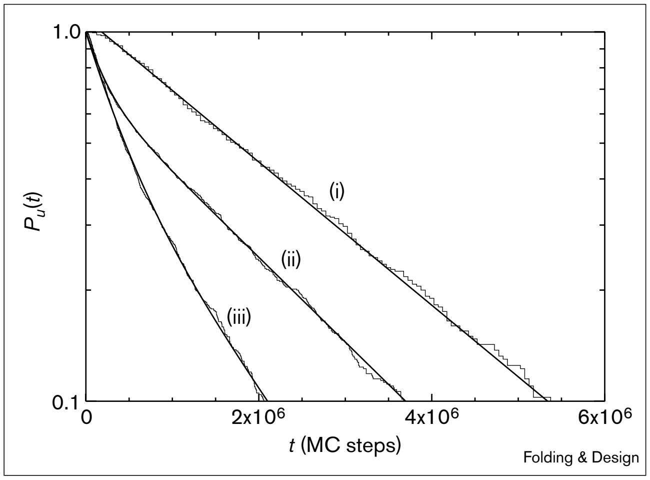
δ	27-mer $P_{5,12,17,22}^{\dagger}$	36-mer $P_{3,8,29,35}^{\ddagger}$
0.2	0.14	0.42
0.05	0.50	0.64
0.02	0.62	0.80
0.01	0.70	0.91
0.00	1.00	1.00

*By definition, when δ = 0 all native contacts occur with unit probability. †$P_{5,12,17,22}$ is the probability that four most probable contacts 5,12,17 and 22 occur simultaneously in the folding nucleus in all the trajectories. ‡$P_{3,8,29,35}$ is the probability that four most probable contacts 3,8,29 and 35 occur simultaneously in the folding nucleus in all the trajectories.

Despite these general observations we feel obliged to examine the ‘kinetic condition’ requirement as demanded in [14] for the 36-mer. This test was already done in two different ways for the 27-mer by KT in [10]. We first created an ensemble of conformations in which the most probable contacts in the folding nuclei were preformed to be in the native states. With these as initial conformations, simulations were performed at temperatures at which the native state is the most stable. In one experiment the most probable contacts remained fixed throughout the folding simulations, while in the other case the constraints were released after the temperature was lowered. The results showed that constraining the contacts helps folding at best marginally, if at all.

Here, we report results for similar calculations for the 36-mer in which the most probable contacts (3, 8, 29 and 35) in the TS, which were identified by averaging over 100 trajectories (Figure 1c), were constrained to be in the native state in the denatured ensemble. The time dependence of $P_u(t)$ for three distinct cases is shown in Figure 2. Curve (ii) corresponds to the case in which the constraints were released upon temperature quench. The folding time, obtained by averaging over 800 trajectories, decreases by a factor of about 2 compared to unconstrained folding. When the most probable native contacts are always retained, which appears to be advocated in [14], we find that the folding time for this sequence decreases by a factor of 3 compared with unconstrained folding. The decrease in folding time, with preformed most probable contacts in the denatured conformations, is marginal and not as dramatic as expected from the suggestions made by AGS and in [14]. More importantly, the nature of folding kinetics with constraints has been altered. In the absence of preformed contacts the decay of the fraction of molecules that have not folded at time t, $P_u(t)$, is quantitatively fit using a single exponential (curve (i) in Figure 2). In the presence of preformed putative nucleation contacts in the denatured ensemble, however, $P_u(t)$ is best fit by a sum of

Figure 2



Fraction of unfolded molecules $P_u(t)$ as a function of time t (semilog plots) for the 36-mer studied in [10]. Curve (i) is the single exponential fit to $P_u(t)$ for unconstrained spontaneous folding. The mean folding time is $\tau_f = 2.4 \times 10^6$ MC steps. Curve (ii) is the bioexponential fit to $P_u(t)$ for the case in which the four most probable contacts are preformed in the denatured (high temperature) conformations but the constraints are released upon temperature quench. The mean folding time is $\tau_f = 1.4 \times 10^6$ MC steps, but the folding kinetics (as measured by $P_u(t)$) are quantitatively fit by a sum of two exponentials. The amplitude of the fast phase is 0.30. Curve (iii) shows a bioexponential fit to $P_u(t)$ for the case in which four most probable contacts are preformed to be in the native state in the denatured conformations and remain so until the native state is reached. The mean folding time in this case is $\tau_f = 0.8 \times 10^6$ MC steps. The decay of $P_u(t)$ is best fit by a sum of two exponential with the amplitude of the fast phase being 0.51.

two exponentials (see Figure 2 for details). This exercise, together with the results of KT, shows that, depending on the sequence and the resulting topology, there may be some decrease in the overall folding time by having most probable contacts preformed in the denatured conformations. But in all instances examined, folding starting from preformed nucleus does not occur extremely rapidly as is expected in [14]. Thus, the presence of preformed nucleation contacts in the denatured conformations does not lead to great enhancement in the folding rates in contrast to the expectation in [14].

On the possible bimodality in the distribution of native contacts in TS

Consider the distribution of contacts in the TS such as, for example, the ones computed by KT (see Figure 13 in [10]). If such a distribution is strictly bimodal with some set of contacts occurring with high probability (as would be predicted by the MFN model) then these contacts could be kinetically important. If such contacts occur with unit probability then these kinetically important contacts would constitute a 'specific' nucleus. It is stated in [14] that such a bimodal distribution is observed in the 36-mer of KT (see Figure 3a). Notice that in Figure 13 of [10] there are four contacts that occur on average with a probability greater than 0.4. This already supports the MFN model, and is also in accord with the WFSFN model of [14]. Since it is stated in [14] that the probability of finding all four most probable contacts should depend strongly on δ it is incumbent upon us to examine how the distribution displayed in Figure 3a varies with δ. In Figure 3b,c the distributions for $\delta = 0.05$ and 0.02 (the value advocated in [14]) are presented. The apparent bimodality seen in Figure 3a vanishes as δ is varied and in particular when δ is 0.02 (Figure 3c).

Lessons from experiments

At the present time, the most direct glimpse of the TS structures of two-state folders is obtained using the protein engineering method and ϕ value analysis pioneered by Fersht and coworkers [3,5]. Insightful as these experiments are, they only enable us to infer certain average characteristics of TS structures. Recently, the protein engineering method has been used to infer the extent to which the TS structures are heterogeneous in the SH3 domain family of proteins [15,16]. In both the Src SH3 domain and the α-spectrin SH3 domain it has been suggested that the distal loop hairpin formed by a tight connection between two β strands is nearly fully formed in the TS. A similar proposal could be put forward for the folding of certain cold-shock proteins [25]. Baker and coworkers [15] have further characterized partial structure formation in the diverging type II β-turn as well as in the hydrophobic core, both of which have residues that interact favorably with the distal loop. Serrano and coworkers [16] have investigated the effect of mutations on D48G α-spectrin SH3 domain and found ϕ values similar to that in the wild-type protein. The largest ϕ values are found in the distal loop. The findings of Baker and Serrano are taken to imply that the formation of the tight distal loop is an obligatory step in ensuring efficient folding.

It is also worth pointing out that in an earlier study, Serrano and coworkers [26] showed that the patterns of ϕ values for α-spectrin SH3 domain and its circular permutants (which have the same topology) are quite distinct. This suggested that a particular set of contacts in the TS ensemble are not necessary for efficient folding, that is, the putative nucleus is not specific.

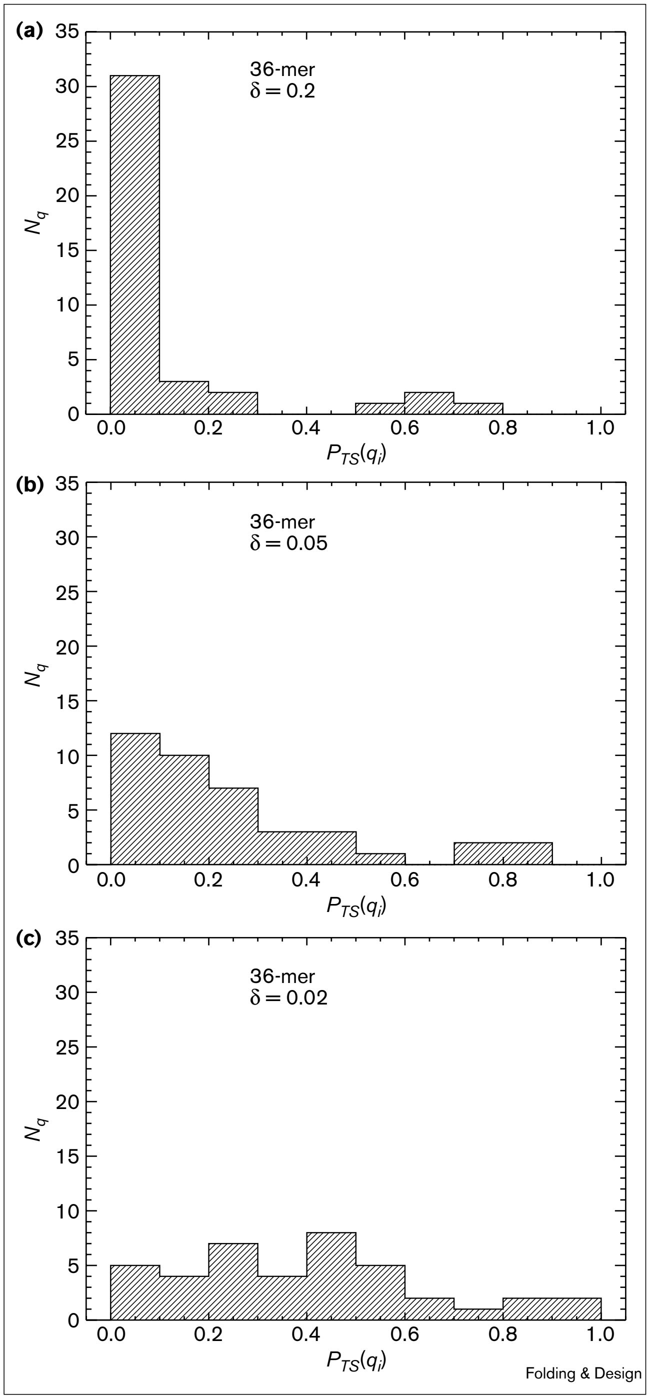
The idea that nucleation events occur with larger probability near turn regions in proteins with predominantly β-sheet topology was predicted using theoretical analysis [6,8]. In the SH3 family of proteins the formation of structure in the distal loop is mandated by its intrinsic rigidity, that is, it is a 'mechanic nucleus' [16]. This suggests that the dispersion in the structures of the TS ensemble in the SH3 family is less than is apparently the case for chymotrypsin inhibitor 2 and λ repressor. In fact, Baker has argued that it is the width of the transition region (see also [27]) that is decreased in the Src SH3 domain (due to a tight hairpin stabilized by hydrogen bonding). According to the SFN model the dispersion in the transition region is always extremely small, independent of the protein (due to the expected strict bimodality in the distribution of contacts in the TS ensemble). In the MFN model (or for that matter in the WFSFN model), by contrast, the width is determined by topology and energetic considerations. Thus, the structuring of a part of the protein with higher probability than other parts (polarized transition states [15]) does not violate the underlying scenario of the NC mechanism of the MFN model. It is possible that stiff β-turns (large persistence length) might decrease the heterogeneity of the folding nuclei compared with α-helical proteins.

Conclusions

Firstly, systematic and extensive analyses of the NC mechanism using several interaction potentials and sequence lengths are possible with the use of minimal lattice models. The major advantage (perhaps the only one) of such models is that very precise answers to precisely formulated questions can be given. Extensive computations on lattice models as well as off-lattice models [8,19] do not support the original purist version of the SFN [9] model. Rather they indeed show the validity of the MFN model or the WFSFN model stated in [14]. The severe restrictions of the lattice models, such as a lack of realistic topology or secondary structure, prevent them from providing insights into experiments such as those reported by Baker and Serrano. Secondly, some of the conclusions based on protein engineering analysis are still relatively fluid. The conclusion that efficient folding of two-state folders clearly follows the NC mechanism is very robust. These experiments suggest that, depending on topology, different scenarios for nucleation processes in protein can arise. The major difference between these scenarios is associated with the degree of heterogeneity in the TS ensemble. (It is worth remembering (Anfinsen hypothesis) that because protein folding is a self-assembly process, topology itself is a consequence of a sequence (chain connectivity) and interaction energies.) Many more experiments on a variety of different proteins with differing topologies are needed before a fuller understanding of the nature of the folding nuclei and the associated characteristics of the TS ensemble can emerge.

If we take the existing computational and experimental evidence into account, it is logical to conclude that (we quote from [20]) “The extreme picture of a small single specific nucleus determining entirely the rate is at best a convenient oversimplification of the experimental data. A more generally useful picture is that of a delocalized structure [7] or a set of many smaller nuclei [8].” Even if we discount the fact that the strict SFN has not been demonstrated in lattice models, we agree that if what is meant by SFN is WFSFN [14], then the distinction between MFN and WFSFN is indeed semantic.

Figure 3



Distribution of probabilities $P_{TS}(q_i)$ that a contact q_i is found in folding nuclei. This distribution shows the number of native contacts N_q that have a particular probability to participate in folding nuclei. Calculation for **(a)** is with $\delta = 0.2$, and for **(b,c)** with $\delta = 0.05$ and 0.02, respectively. The smaller the value of δ, the closer a given molecule is to the native state.

Acknowledgements

We have greatly benefited from numerous insightful conversations with R Elber and JE Straub. They and J Skolnick made many useful comments on an earlier version of this report.

References

1. Dill, K.A. & Chan, H.S. (1997). From Levinthal to pathways to funnels. *Nat. Struct. Biol.* **4**, 10-19.
2. Fersht, A.R. (1995). Characterizing transition states in protein folding: an essential step in the puzzle. *Curr. Opin. Struct. Biol.* **5**, 79-84.
3. Fersht, A.R. (1997). Nucleation mechanism of protein folding. *Curr. Opin. Struct. Biol.* **7**, 10-14.
4. Wetlaufer, D. (1973). Nucleation, rapid folding, and globular intrachain regions in proteins. *Proc. Natl Acad. Sci. USA* **70**, 697-701.
5. Itzhaki, L.S., Otzen, D.E. & Fersht, A.R. (1995). The structure of the transition state of chymotrypsin inhibitor 2 analyzed by protein engineering methods: evidence for a nucleation-condensation mechanism for protein folding. *J. Mol. Biol.* **254**, 260-288.
6. Matheson, R.R. & Scheraga, H.A. (1978). A method for predicting nucleation sites for protein folding based on hydrophobic contacts. *Macromolecules* **11**, 814-829.
7. Bryngelson, J.D. & Wolynes, P.G. (1990). A simple statistical field theory of heteropolymer collapse with application to protein folding. *Biopolymers* **30**, 177-188.
8. Guo, Z. & Thirumalai, D. (1995). Kinetics of protein folding: nucleation mechanism, time scales and pathways. *Biopolymers* **36**, 83-103.
9. Abkevich, V.I., Gutin, A.M. & Shakhnovich, E. (1994). Specific nucleus as the transition state for protein folding: evidence from the lattice model. *Biochemistry* **33**, 10026-10036.
10. Klimov, D.K. & Thirumalai, D. (1998). Lattice models for proteins reveal multiple folding nuclei for nucleation-collapse mechanism. *J. Mol. Biol.* **282**, 471-492.
11. Klimov, D.K. & Thirumalai, D. (1996). Factors governing the foldability of proteins. *Proteins* **26**, 411-441.
12. Veitshans, T., Klimov, D.K. & Thirumalai, D. (1996). Protein folding kinetics: timescales, pathways and energy landscapes in terms of sequence-dependent properties. *Fold. Des.* **2**, 1-22.
13. Camacho, C.J. & Thirumalai, D. (1993). Kinetics and thermodynamics of folding in model proteins. *Proc. Natl Acad. Sci. USA* **90**, 6369-6372.
14. Shakhnovich, E.I. (1998). Folding nucleus: specific or multiple? Insights from lattice models and experiments. *Fold. Des.* **3**, R108-R111.
15. Grancharova, V.P., Riddle, D.S., Santiago, J.V. & Baker, D. (1998). Important role of hydrogen bonds in the structurally polarized transition state for folding of the Src SH3 domain. *Nat. Struct. Biol.* **5**, 714-720.
16. Maritinez, J.C., Pisabarro, M.T. & Serrano, L. (1998). Obligatory steps in protein folding and the conformational diversity of the transition state. *Nat. Struct. Biol.* **5**, 721-729.
17. Miyazawa, S. & Jernigan, R.L. (1985). Estimation of effective inter-residue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules* **18**, 534-552.
18. Wolynes, P.G. (1997). Folding funnels and energy landscapes of larger proteins within the capillary approximation. *Proc. Natl Acad. Sci. USA* **94**, 6170-6175.
19. Guo, Z. & Thirumalai, D. (1997). The nucleation-collapse mechanism in protein folding: evidence for the non-uniqueness of the folding nucleus. *Fold. Des.* **2**, 377-391.
20. Onuchic, J.N., Socci, N.D., Luthey-Schulten, Z. & Wolynes, P.G. (1996). Protein folding funnels: the nature of the transition state ensemble. *Fold. Des.* **1**, 441-450.
21. Yang, J., Gould, H., Klein, W. & Mountain, R.D. (1990). Molecular dynamics investigation of deeply quenched liquids. *J. Chem. Phys.* **93**, 711-723.
22. Swope, W.C. & Andersen, H.C. (1990). 10^6 particle molecular dynamics study of homogenous nucleation of crystals in a supercooled atomic liquid. *Phys. Rev. B.* **41**, 7042-7053.
23. Fersht, A.R. (1995). Optimization of rates of protein folding: the nucleation-condensation mechanism and its implications. *Proc. Natl Acad. Sci. USA* **92**, 10869-10873.
24. Camacho, C.J. & Thirumalai, D. (1995). Modeling the role of disulfide bonds in protein folding: entropic barriers and pathways. *Proteins* **22**, 27-40.
25. Perl, D., Welker, C., Schindler, T., Schroder, K., Marahiel, M.A., Jaenicke, R. & Schmid, F.X. (1998). Conservation of rapid two-state folding in mesophilic, thermophilic, and hyperthermophilic cold shock proteins. *Nat. Struct. Biol.* **3**, 229-235.
26. Viguera, A.R., Wilmanns, M. & Serrano, L. (1996). Different folding transition states result in the same native structure. *Nat. Struct. Biol.* **3**, 874-880.
27. Gruebele, M. & Wolynes, P.G. (1998). Satisfying turns in folding transitions. *Nat. Struct. Biol.* **5**, 682-685.