doi:10.1006/jmbi.1999.3534 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2000) **296**, 1183–1188

JMB

COMMUNICATION

Identifying the Protein Folding Nucleus Using Molecular Dynamics

Nikolay V. Dokholyan[1,2]*, Sergey V. Buldyrev[1], H. Eugene Stanley[1] and Eugene I. Shakhnovich[2]

[1]*Center for Polymer Studies Physics Department, Boston University, Boston, MA 02215, USA*

[2]*Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street Cambridge, MA 02138, USA*

**Corresponding author*

Molecular dynamics simulations of folding in an off-lattice protein model reveal a nucleation scenario, in which a few well-defined contacts are formed with high probability in the transition state ensemble of conformations. Their appearance determines folding cooperativity and drives the model protein into its folded conformation. Amino acid residues participating in those contacts may serve as "accelerator pedals" used by molecular evolution to control protein folding rate.

Keywords: protein folding; kinetics; molecular dynamics; nucleus; transition state

Thermodynamically, the folding transition in small proteins is analogous to a first-order transition whereby two thermodynamic states (folded and unfolded; Makhatadze & Privalov, 1995; Karplus & Shakhnovich, 1992; Jackson, 1998) are free energy minima while intermediate states are unstable. The kinetic mechanism of transitions from the unfolded state to the folded state is nucleation (Karpov & Oxtoby, 1996; Lifshits & Pitaevskii, 1981; Shakhnovich, 1997; Fersht, 1997; Pande *et al.*, 1998). Folding nuclei can be defined as the minimal stable element of structure whose existence results in subsequent rapid assembly of the native state. This definition corresponds to a "post-critical nucleus" related to the first stable structures that appear immediately after the transition state is overcome (Abkevich *et al.*, 1994). The thermal probability of a transition state conformation is low compared to the folded and unfolded states, which are both accessible at the folding transition temperature T_f (see Figure 1(a)).

Lattice model studies showed that passing through the transition state with subsequent rapid assembly of the native conformation requires the formation of some (small) number of specific obligatory contacts (protein-folding nucleus) (Shakhnovich, 1997; Pande *et al.*, 1998; Abkevich *et al.*, 1994; Shakhnovich, 1998). Further studies (Mirny *et al.*, 1998; Pande *et al.*, 1998; Klimov & Thirumalai, 1998) suggested that the folding nucleus location may depend more on the topology of the native structure than on a particular sequence that folds into that structure. This view also received support from experimental analysis (Martinez *et al.*, 1998; Grantchanova *et al.*, 1998).

The dominance of geometrical/topological factors in the determination of the folding nucleus is a remarkable property that has evolutionary implications (see below). It is important to understand the physical origin of this property of folding proteins and assess its generality. To this end, it is important to study other than lattice models and other than Monte-Carlo dynamic algorithms. Here, we employ the discrete molecular dynamics simulation technique, and the Gō model (Taketomi *et al.*, 1975; Gō & Abe 1981; Abe & Gō, 1981; Micheletti *et al.*, 1999) with the square-well potential of the inter-residue interaction, to search for the nucleus (Zhou *et al.*, 1997; Zhou & Karplus, 1997; Dokholyan *et al.*, 1998).

Our proposed method to search for a folding nucleus is based on the observation (Abkevich *et al.*, 1994) that equilibrium fluctuations around the native conformation can be separated into "local" unfolding (followed by immediate refolding) and "global" unfolding that leads to a transition into an unfolded state and requires more time to refold. Local unfolding fluctuations are the ones that do not reach the top of the free energy barrier and, hence, are committed to moving quickly back to the native state. In contrast, global unfolding fluctuations are the ones that overcome the barrier and are committed to descend further

---

E-mail address of the corresponding author: eugene@belok.harvard.edu

0022-2836/00/051183–6 $35.00/0

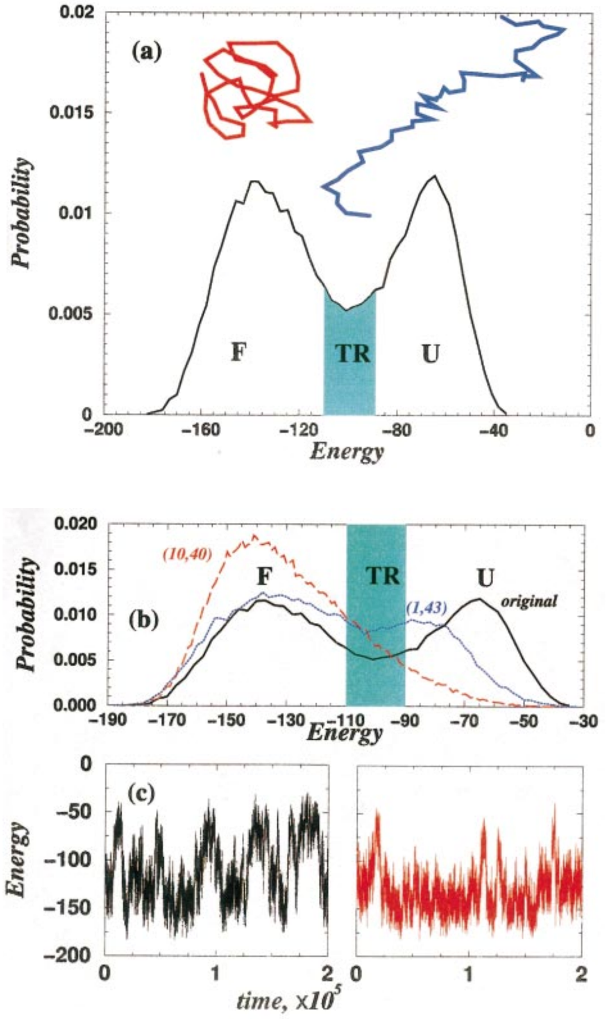

Figure 1. (a) Probability distribution of the energy states E of the 46-mer maintained at the folding transition temperature $T_f = 1.44$. The bimodal distribution indicates the presence of two dominant states: the folded (region F) and the unfolded (region U) states. The transition state ensemble belongs to region TR of the histogram $\{-110 < E < -90\}$. The insets show typical conformations in the folded and unfolded regions. (b) Probability distribution of the energy states E of the: (1) original 46-mer (at $T_f = 1.44$); (2) 46-mer (at $T = 1.46$) with a fixed contact belonging to the protein-folding nucleus, (10, 40); and (3) 46-mer (at $T = 1.46$) with fixed randomly chosen control contact (1, 43), which does not belong to the protein folding nucleus. Note that the probability of the unfolded state of the 46-mer with a fixed contact belonging to the protein folding nucleus, is suppressed compared to that of the original 46-mer. (c) Time evolution of the energy E of (1) original (left) and (2) fixed (10, 40) contact (right). Case (3) fixed (1, 43) contact is similar to (1), so we do not show it. For case (1), the fluctuations are mostly between two extreme values of energy, corresponding to the folded and unfolded states. In contrast, for case (2), the fluctuations are mostly around one energy value, corresponding to the folded state.

to the unfolded state. Similarly, the fluctuations from the unfolded state can be separated into those that descend back to the unfolded state and those that result in productive folding. The difference between the two modes of fluctuation is whether or not the major free energy barrier is overcome. This means that the nucleation contacts (i.e. the ones that are formed on the "top" of the free energy barrier is the chain passes it upon folding) should be identified as contacts that are present in the "maximally locally unfolded" conformations but are lost in the globally unfolded conformations of comparable energy.

Thus, in order to identify the folding nucleus, we study the conformations of the 46-mer that appear in various kinds of folding $\rightleftharpoons$ unfolding fluctuations. The transition state conformations belong to the transition region TR from the folded state to the unfolded state that lies in the energy range $\{-110 < E < -90\}$ (see Figure 1(a)). Region TR corresponds to the minimum of the histogram of the energy distribution. If we know the past and the future of a certain conformation that belongs to the TR, we can distinguish four types of such conformations (see Figure 2): (1) UU conformations that originate in and return to the unfolded region without descending to the folded region; (2) FF conformations that originate in and return to the folded region without ascending to the unfolded region; (3) UF conformations that originate in the unfolded region and descend to the folded region; and (4) FU conformations that originate in the folded region and ascend to the unfolded region. There are $\sim 10^3$ UF, FU, FF, and UU conformations in one simulation run at T_f.

If the nucleus exists, then the UF, FU, FF, and UU conformations must have different properties depending on their history. One difference between the FF conformations and UU conformations is that the protein folding nucleus is more

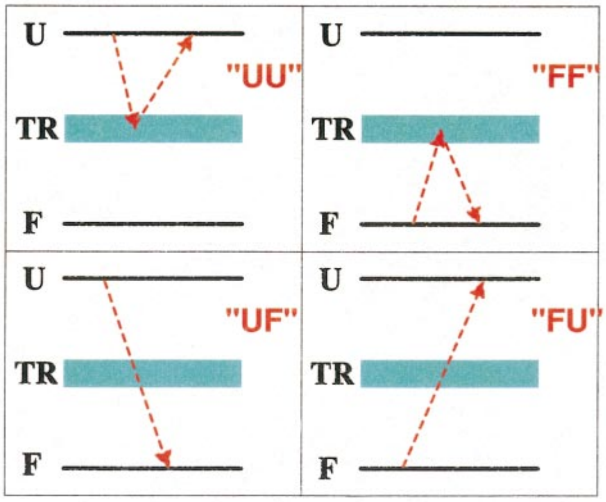

Figure 2. Schematic definition of the four types of conformations: FF, UU, UF, and FU.

likely to be retained in the FF conformations than in the UU conformations. The contacts belonging to the critical nucleus (''nucleation contacts'') start appearing in the UF conformations, and start disappearing in the FU conformations, so that the frequencies of nucleation contacts in UF and FU conformations should be in between FF and UU.

Our goal is to select the contacts that are crucial for the folding $\rightleftharpoons$ unfolding transition. To this end we select the contacts that appear much more often in the FF conformations than in the UU conformations. We calculate the frequencies of all contacts in FF conformations, f_{FF}, and in UU conformations, f_{UU}. We plot the histogram of the differences in frequencies, $f_{\mathrm{FF}} - f_{\mathrm{UU}}$, for all possible (native and non-native) contacts of the 46-mer (see Figure 3(a)). We find that there is a peak at $f_{\mathrm{FF}} - f_{\mathrm{UU}} = 0.2$, that is located over seven standard deviations from the average value of $\langle f_{\mathrm{FF}} - f_{\mathrm{UU}} \rangle = 0 \cdot 008$. We discover that there are only five contacts that belong to this peak: (residue 11, residue 39), (10, 40), (11, 40), (10, 41), and (11, 41) (see Figure 3). These contacts can serve as a putative protein folding nucleus in the folding $\rightleftharpoons$ unfolding transition in our model.

Next, we demonstrate that these five selected contacts indeed belong to the protein folding nucleus. Suppose we fix just one of them, e.g. (10, 40), i.e. we impose a covalent (''permanent'') link between residue 10 and residue 40. If this contact belongs to the protein folding nucleus, its fixation by a covalent bond would eliminate the barrier between the folded and unfolded states, i.e. only the native basin of attraction will remain. Hence, we hypothesize that the cooperative transition between the unfolded and folded state will be eliminated and the energy histogram (Figure 1(a)) should change qualitatively from bimodal to unimodal. Our molecular dynamics simulations support this hypothesis (Figure 1(b) and (c)): fixation of only one nucleation contact, (10, 40), gives rise to a qualitative change in the energy distribution from bimodal to unimodal. Indeed, the probability to find an unfolded state of the 46-mer with a fixed link, (10, 40), which belongs to the protein folding nucleus, is drastically reduced compared to the probability of the unfolded state of the original 46-mer, indicating the importance of the selected contact (10, 40). We also impose a link between the remaining four contacts. We find (results are not shown) that plots for energy distributions of the 46-mer with fixed links between residues (11, 39), (10, 40), (11, 40), (10, 41), and (11, 41) are almost identical.

To provide a ''control'' for the purposes of illustrating that a specific contact plays such a dramatic role in changing the character of the energy landscape, we fix a randomly chosen contact, (1, 43), which is not predicted by our analysis, to belong to the critical nucleus. Our hypothesis predicts no qualitative change in the energy distribution histogram, since the barrier should not change dramatically for this control. Figure 1(b) and (c) shows that this is indeed the case. In addition, we impose a link between four non-nucleic contacts, other than (1, 43): (19, 37), (18, 39), (22, 46), and (29, 45). We find (results are not shown) that plots for energy distributions of the 46-mer with these fixed control links are bimodal, similar to that of the control with fixed (1, 43) contact.

To further demonstrate the dramatic role of the five nucleic contacts we compute a parameter which is 1 or 0 depending on whether there exists at least one contact out of five selected. Then, we average the value of this parameter in the window of approximately 200 time units. For comparison, we compute the similar parameter for the five control contacts. The results are shown in Figure 4. From the top plots shown in Figure 4 it is clear that at least one nucleic contact is present in the folded states and none of them exist in the unfolded state and the transition between these two states is sharp. On the contrary, the parameter reflecting the appearance of control non-nucleus contacts fluctuates in a wide range even in the folded state without apparent correlation with folding-unfolding transition.

An interesting point is that the contacts, such as (12, 39), that have slightly smaller values of $f_{\mathrm{FF}} - f_{\mathrm{UU}}$ than nucleic ones, that belong to the next peak of the histogram in Figure 3(a) at $f_{\mathrm{FF}} - f_{\mathrm{UU}} = 0.12$, do not behave as nucleic. Actually, contact (12, 39) is specifically interesting, since its appearance is correlated with the nucleic contacts, residue 12 is close to both residues 10 and 11, while residue 39 participates in the nucleic contacts itself. Fixation of the (12, 39) contact has two implications. First, it apparently facilitates the formation of the nucleic contact (11, 39), resulting in rapid transition to the native state. Second, we found that in many cases contact (12, 39) leads the 46-mer to a misfolded state (trap). This can be seen from the energy trajectory or the rms displacement histogram, where one can identify the second peak corresponding to the misfolded conformations (data not shown).

Our analysis shows that there is a well-defined set of contacts that is responsible for the rapid assembly of the native state of the 46-mer. The example of the contact (12, 39) indicates that even contacts that are located in the vicinity of nucleic contacts may cause trapping of the 46-mer in the misfolded conformation.

Another interesting point is that the contacts that have negative $f_{\mathrm{FF}} - f_{\mathrm{UU}}$ values (see Figure 3(a)) persist in the unfolded conformations of the 46-mer, while they are less frequent in the folding conformations. These contacts maybe responsible for the kinetic traps during folding of the 46-mer. A more detailed study of these contacts is underway.

Our main conclusion is that a few (≈five) structure-specific contacts play a major role in determining the free energy landscape of a protein. This is well illustrated by our results that show that fixation of even one nucleation contact can eliminate the free-energy barrier between folded and unfolded states. These contacts are most frequently

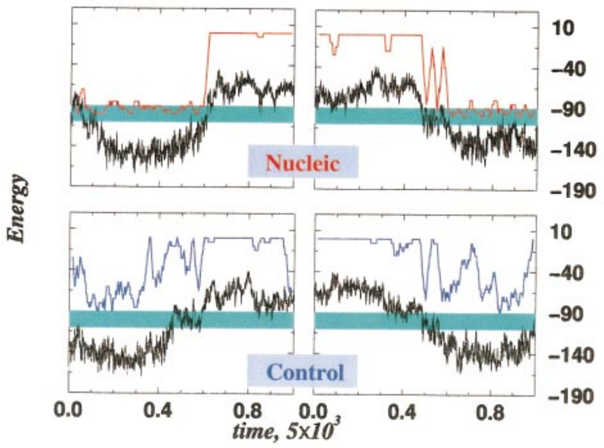

Figure 4. Dependence on time of the parameter, that is defined to be 1 if at least one contact out of five selected is present, or 0 if none of the contacts are present. The values of the parameter are averaged in the window of 200 time units. For demonstration purpose only we multiplied the average values of the parameter by −100. The top two plots are presented for the FU (left) and UF (right) conformations for the set of the nucleic contacts. The bottom two plots are similar to the top ones but for the set of the control contacts. From the top plots it is clear that at least one nucleic contact is present at the folded states and they do not exist at the unfolded state and the transition between these two states is sharp. On the contrary the average parameter values when the 46-mer is in the folded state are strongly fluctuating around −40, which indicates that at least one control contact present only in 40% of the lifetime of the 46-mer in the folded state.

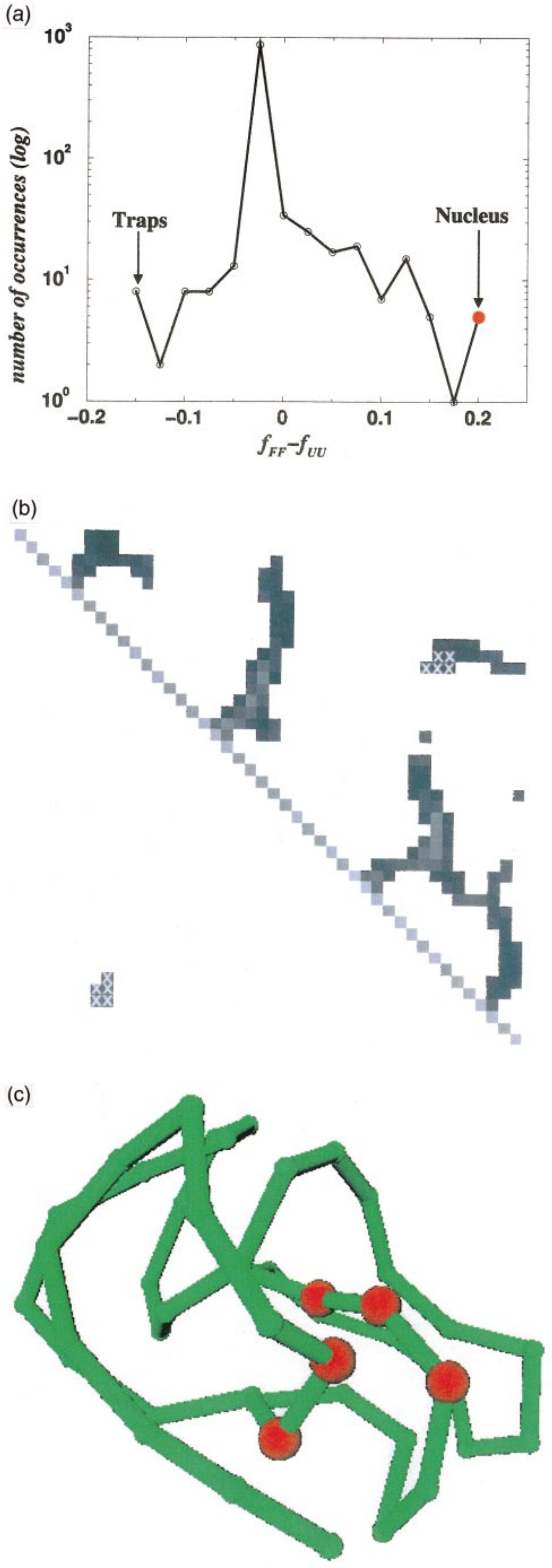

Figure 3. (a) The histogram of the differences in frequencies between FF and UU conformations, $f_{\mathrm{FF}} - f_{\mathrm{UU}}$, for all possible (native and non-native) contacts of the 46-mer. The peak of the histogram at $f_{\mathrm{FF}} - f_{\mathrm{UU}} = 0.2$ is formed by five native contacts (11, 39), (10, 40), (11, 40), (10, 41), and (11, 41). These contact form the nucleus of the 46-mer. The peak corresponding to the negative values of the $f_{\mathrm{FF}} - f_{\mathrm{UU}} = -0.15$ is formed by the contacts that persist in the unfolded conformations of the 46-mer, while less frequent in the folding conformations. These contacts may be responsible for the kinetic traps during folding of the 46-mer. (b) Contact map of the model protein. The darker the shade of grey, the larger is the frequency of a contact. Above the diagonal of the square matrix shows the native contacts (see Dokholyan *et al.*, 1998) of the FF conformations (if the native contact frequency is larger than 0.2). Below the diagonal of the square matrix shows the difference between the frequencies of the native contacts in FF and UU conformations (if this difference is larger than 0.2). Five contacts that persist in the FF conformations, ((11, 39). (10, 40), (11, 40), (10, 41), and (11, 41)), are marked by crosses. The Figure shows that identification of the protein folding nucleus is facilitated by the method used to construct the region of the matrix below the diagonal. (c) The structure of the 46-mer at the native state. Five selected residues, (10, 11, 39, 40, and 41), form nucleic contacts.

formed in the folding transition state hence their kinetic role as folding nucleus.

The Gō model does not discriminate between native contacts based on the energetic properties of these contacts. Nevertheless, a few native contacts turn out to play a key role in determining free energy landscape and folding kinetics even in this model. The only reason for this may be the fact that the topology of the native structure determines a special role for those (nucleic) contacts, i.e. nucleus location may be determined to a great extent by the topology of the native state. This discovery has a direct implications for protein evolution, raising the possibility that proteins that have similar structures but different sequences may have similarly located protein folding nuclei. Recent experiments (Martinez *et al*., 1998; Chiti *et al*., 1999; Clarke *et al*., 1999) and lattice simulations (Abkevich *et al*., 1994; Mirny *et al*., 1998) point out to dominant role of topological factors in determining folding nucleus in two-state proteins.

It should be emphasized that energetic factors play also an important role in nucleation scenario by providing stabilization to nucleus residues *via* selection of proper sequences. In terms of the evolutionary selection of protein sequences, the robustness of the folding nucleus suggests that any additional evolutionary pressure that controls the folding rate may have been applied selectively to nucleus residues, so that nucleation positions may have been under double (stability + kinetics) pressure in all proteins that fold into a given structure. Such additional evolutionary pressure has indeed been found in the analysis of several protein superfamilies (Mirny *et al*., 1998; Ptitsyn, 1998; Ptitsyn & Ting, 1999; Mirny & Shakhnovich, 1999).

Acknowledgments

We thank R. S. Dokholyan for careful reading of the manuscript. N.V.D. is supported by NIH NRSA molecular biophysics predoctoral traineeship (GM08291-09) and by NIH postdoctoral fellowship (GM20251-01). E.I.S. is supported by NIH grant RO1-52126. The Center for Polymer Studies acknowledges the support of the NSF.

References

Abe, H. & Gō, N. (1981). Non-interacting local-structure model of folding and unfolding transition in globular proteins. II. Application to two-dimensional lattice proteins. *Biopolymers*, **20**, 1013-1031.

Abkevich, V. I., Gutin, A. M. & Shakhnovich, E. I. (1994). Specific nucleus as the transition state for protein folding: evidence from the lattice model. *Biochemistry*, **33**, 10026-10036.

Chiti, F., Taddei, N., White, P. M., Bucciantini, M., Magherini, F., Stefani, M. & Dobson, C. M. (1999). Mutational analysis of acylphosphatase suggests the importance of topology and contact order in protein folding. *Nature Struct. Biol.* **6**, 1005-1009.

Clarke, J., Cota, E., Fowler, S. & Hamill, S. J. (1999). Folding studies of immunoglobulin-like β-sandwich proteins suggest that they share a common folding pathway. *Structure Fold. Design*, **7**, 1145-1154.

Dokholyan, N. V., Buldyrev, S. V., Stanley, H. E. & Shakhnovich, E. I. (1998). Molecular dynamics studies of folding of a protein-like model. *Fold. Design*, **3**, 577-587.

Fersht, A. (1997). Nucleation mechanisms in protein folding. *Curr. Opin. Struct. Biol.* **7**, 10-14.

Gō, N. & Abe, H. (1981). Non-interacting local-structure model of folding and unfolding transition in globular proteins. I. Formulation. *Biopolymers*, **20**, 991-1011.

Grantchanova, V., Riddle, D., Santiago, J. & Baker, D. (1998). Important role of hydrogen bonds in the structurally polarized transition state for folding of the src SH3 domain. *Nature Struct. Biol.* **5**, 714-720.

Jackson, S. (1998). How do small single-domain proteins fold? *Fold Design*, **3**, R81-R91.

Karplus, M. & Shakhnovich, E. I. (1992). Theoretical studies of protein folding thermodynamics and dynamics. In *Protein Folding* (Creighton, T., ed.), chapt. 4, pp. 127-196, W.H. Freeman and Company, New York.

Karpov, V. G. & Oxtoby, D. W. (1996). Nucleation in disordered systems. *Phys. Rev. ser. B*, **54**, 9734-9745.

Klimov, D. & Thirumalai, D. (1998). Lattice models for proteins reveal multiple folding nuclei for nucleation-collapse mechanism. *J. Mol. Biol.* **282**, 471-492.

Lifshits, E. M. & Pitaevskii, L. P. (1981). *Physical Kinetics*, Pergamon Press, Oxford and New York.

Makhatadze, G. & Privalov, P. L. (1995). Energetics of protein structure. *Advan. Protein Chem.* **47**, 307-425.

Micheletti, C., Banavar, J. R., Maritan, A. & Seno, F. (1999). Protein structures and optimal folding from a geometrical variational principle. *Phys. Rev. Letters*, **82**, 3372-3375.

Martinez, J., Pissabarro, T. & Serrano, L. (1998). Obligatory steps in protein folding and the conformational diversity of the transition state. *Nature Struct. Biol.* **5**, 721-729.

Mirny, L. A. & Shakhnovich, E. I. (1999). Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. *J. Mol. Biol.* **291**, 177-196.

Mirny, L., Abkevich, V. I. & Shakhnovich, E. I. (1998). How evolution makes proteins fold quickly. *Proc. Natl Acad. Sci. USA*, **95**, 4976-4981.

Pande, V. S., Grosberg, A. Yu, Rokshar, D. & Tanaka, T. (1998). Pathways for protein folding: is a new view needed? *Curr. Opin. Struct. Biol.* **8**, 68-79.

Ptitsyn, O. B. (1998). Protein folding and protein evolution: common folding nucleus in different subfamilies of c-type cytochromes? *J. Mol. Biol.* **278**, 655-666.

Ptitsyn, O. B. & Ting, K.-L. H. (1999). Non-functional conserved residues in globins and their possible role as a folding nucleus. *J. Mol. Biol.* **291**, 671-682.

Shakhnovich, E. I. (1997). Theoretical studies of protein-folding thermodynamics and kinetics. *Curr. Opin. Struct. Biol.* **7**, 29-40.

Shakhnovich, E. I. (1998). Protein design: a perspective from simple tractable models. *Fold. Design,* **3**, R108-R111.

Shakhnovich, E. I., Abkevich, V. I. & Ptitsyn, O. B. (1996). Conserved residues and the mechanism of protein folding. *Nature,* **379**, 96-98.

Taketomi, H., Ueda, Y. & Gō, N. (1975). Studies on protein folding, unfolding and fluctuations by computer simulations. *Int. J. Peptide Protein Res.* **7**, 445-459.

Zhou, Y. & Karplus, M. (1997). Folding thermodynamics of a three-heix-bundle protein. *Proc. Natl Acad. Sci. USA,* **94**, 14429-14432.

Zhou, Y., Karplus, M., Wichert, J. M. & Hall, C. K. (1997). Equilibrium thermodynamics of homopolymers and clusters: molecular dynamics and Monte Carlo simulations of system with square-well interactions. *J. Chem. Phys.* **107**, 10691-10708.

Edited by A. R. Fersht

(Received 23 August 1999; received in revised form 11 November 1999; accepted 19 November 1999)