doi: 10.1016/S0022-2836(02)00136-5 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2002) **318**, 863–876

JMB

# Thermodynamics and Folding Kinetics Analysis of the SH3 Domain from Discrete Molecular Dynamics

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

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

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

**Corresponding author*

We perform a detailed analysis of the thermodynamics and folding kinetics of the SH3 domain fold with discrete molecular dynamic simulations. We propose a protein model that reproduces some of the experimentally observed thermodynamic and folding kinetic properties of proteins. Specifically, we use our model to study the transition state ensemble of the SH3 fold family of proteins, a set of unstable conformations that fold to the protein native state with probability 1/2. We analyze the participation of each secondary structure element formed at the transition state ensemble. We also identify the folding nucleus of the SH3 fold and test extensively its importance for folding kinetics. We predict that a set of amino acid contacts between the RT-loop and the distal hairpin are the critical folding nucleus of the SH3 fold and propose a hypothesis that explains this result.

*Keywords:* protein folding; SH3; transition state; folding nucleus; molecular dynamics

## Introduction

Studying the protein folding kinetics is a challenging task because it involves the identification of the transition state ensemble (TSE), a set of unstable conformations that form at the top of the free energy barrier separating the unique folded state from the misfolded and unfolded states.[1] The TSE has been the subject of numerous experiments[2–18] and theoretical studies[19–25] of globular proteins. The TSE is defined as the set of protein conformations with a probability to fold, $p_{\mathrm{FOLD}}$, equal to 1/2.[1] Protein engineering experiments[8–18] suggest that in two-state proteins, there is a specific set of amino acid residues that determines the folding properties. Theoretical studies[19–25] of the TSE support the hypothesis of a specific folding nucleus scenario. Passing through the TSE with the subsequent rapid assembly of the native conformation requires the formation of a set of specific obligatory contacts, which are called the protein folding nucleus.[1]

Lattice simulations[19,23] suggest that the folding nucleus location is identical for two different protein sequences, designed with various potentials to fold into the same structure. In protein engineering experiments[26] on the Src SH3 domain[14,15,27–33] and the α-spectrin SH3 domain,[11,13,34] authors find the same structural characteristics for the TSE of the two homologous proteins. These studies suggest that the location of the TSE and the folding nucleus in the native structure of the protein depends more on the topology of the native structure than on the specific protein amino acid sequence folding into that structure. Several groups[35–38] attempted to identify the TSE of some well characterized globular proteins and obtained a significant correlation with experimental kinetic data. However, these models assume the number of ordered residues as an approximation to the reaction coordinate for the folding process. The main difficulty with these studies is that the number of ordered residues does not characterize the TSE. Ding *et al.* (unpublished results) studied the TSE of the Src SH3 domain and found that different conformations with the same number of native contacts as TSE conformations had drastically different probabilities to fold. Another simplification is the assumption that the folding process occurs through the meeting and adoption of native structure of only two fragments of the protein. In contrast, our method to select TSE conformations does not depend on the number of ordered

Abbreviations used: TSE, transition state ensemble; rmsd, root mean square displacement.

E-mail address of the corresponding author: jmborr@bu.edu

0022-2836/02/$ - see front matter © 2002 Elsevier Science Ltd. All rights reserved

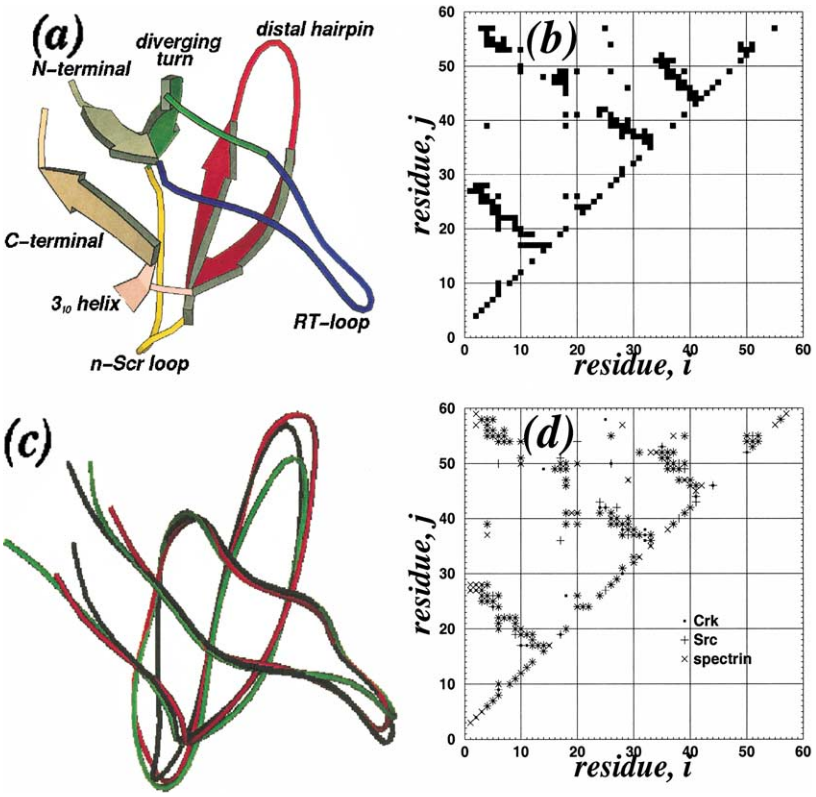

**Figure 1**. (a) Ribbon diagram of the c-Crk SH3 domain. The SH3 fold is a β-sandwich with two nearly orthogonal β-sheets. Anti-parallel oriented strands $\beta_1$, $\beta_2$ and $\beta_5$ make up the first β-sheet, while anti-parallel $\beta_2$, $\beta_3$ and $\beta_4$ build the second one. The c-Crk SH3 domain also contains a short $3_{10}$ helix near the C terminus. (b) Contact map of the c-Crk SH3 domain using coordinates of $C^\beta$ as the force center and cut-off distance $D = 7.5$ Å. Only half of the map is shown, because the plot is diagonal with respect to the diagonal. (c) Structural alignment with FSSP database[42] of c-Crk (black), Src (green) and α-spectrin (red). rmsd between c-Crk and Src is 1.4 Å, and 2.3 Å between c-Crk and α-spectrin. (d) Superposition of related contact maps, again using coordinates of $C^\beta$ as the force center and cut-off distance $D = 7.5$ Å.

residues, and we employ discrete molecular dynamic simulations that do not constrain the number of native elements of structure.

We select the c-Crk SH3 domain[39–41] (Figure 1(a)), as the representative of the SH3 fold family (from the FSSP database[42]). Currently, there are no experimental studies on the folding process of the c-Crk SH3 domain. We aim to understand the folding kinetics of this protein by performing a detailed analysis of the TSE and identifying its folding nucleus. Due to the homology[43] and high structural similarity of c-Crk SH3 domain to Src SH3 (sequence similarity 33%, rmsd = 1.4) and α-spectrin SH3 (sequence similarity 34%, rmsd = 2.3) domains,[9–14,40] our model for c-Crk is virtually identical to that of Src and α-spectrin (Figure 1(c) and (d)). This similarity allows us to compare our predictions for c-Crk with the extensive experimental data of Src and α-spectrin.

During the folding process, a protein conformation is part of the TSE only for a small fraction of the required time for folding, because the free energy of such a conformation is maximal. Thus, the probability of a protein being found in the TSE is minimal. Many folding transitions are needed for a thorough investigation of the TSE, which makes the study of the TSE time-consuming for direct computational approaches. Simplified lattice models[20,44–52] became popular due to their ability to reproduce a significant amount of folding transitions in a reasonable computational time. However, the role of topology in determining the folding nucleus requires study beyond lattice models, which impose unphysical constraints on dihedral angles. All-atom models are the best candidates to address the issue of topology, but these models are computationally difficult to treat because of the large protein conformational space. Simplified off-lattice models[25,53–57] are a compromise between lattice and all-atom models. Here, we determine the TSE and the folding nucleus of the SH3 fold by performing kinetic studies of a protein model. Specifically, we: (a) develop a new, simplified off-lattice model that reproduces the experimentally observed thermodynamic properties of globular proteins, (b) implement the discrete molecular dynamics algorithm (DMD)[25,53,58–62] to rapidly test the folding properties of the model; (c) apply the local fluctuations method (first employed by Dokholyan *et al.*[25] in the kinetic studies of a protein-like chain) to determine the TSE and the folding nucleus of the SH3 fold.

In Results we present the thermodynamic properties of the folding transition. We also present the structural characteristics of the identified TSE in terms of formation of the various secondary structure elements. Finally, we find the folding nucleus and discuss its kinetic properties with a cross-linking simulation. We describe in detail the model and the local fluctuation analysis in Methods.

# Results

## Thermodynamics

We perform DMD simulations and compute the equilibrium properties of a model of the SH3 fold

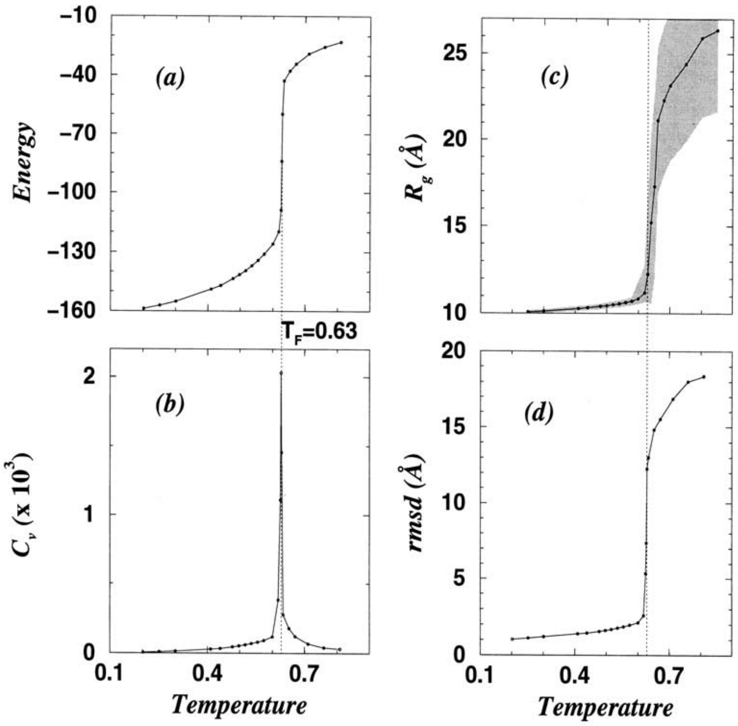

**Figure 2**. The thermodynamic averages of the macroscopic quantities point out the two-state transition of the SH3 fold model when plotted *versus* temperature. (a) Potential energy, (b) specific heat, (c) radius of gyration and (d) root mean square deviation with respect to the native state. The gray shadowed region in (c) inscribes 68% of the total range of values $R_G$ takes for each temperature. At $T_F$, the specific heat is maximal and all extensive thermodynamic quantities show an abrupt change in value.

at temperatures above, at, and below the folding transition temperature, $T_F$. The dependence of average potential energy and related specific heat *versus* temperature is shown in Figure 2(a) and (b), respectively. There is a pronounced increase in the potential energy and a strong peak in the specific heat at $T_F$. We define $T_F = 0.63$ as the midpoint in the abrupt increase of the potential energy *versus* temperature. This temperature coincides with the temperature for which the specific heat is maximal (Figure 2(b)). Below $T_F$, the structure of the globule is the same as that of the native state. The average radius of gyration,[63] $R_G$, does not exceed 10% of the value of $R_G$ in the native state (10 Å), marking the stability of the globular shape (Figure 2(c)). At each studied temperature below $T_F$, fluctuations of $R_G$ around the average value do not exceed 3%, showing that the globule is never disrupted. The root mean square displacement[64] (rmsd) with respect to the native state is never greater than 2 Å, showing that the structure of the globule does not deviate from the structure of the native state (Figure 2(d)). Above $T_F$, the globule is fully unfolded. $R_G$ doubles in magnitude, indicating that the average distance between any two pair of non-bonded amino acid residues is doubled with respect to their distance in the native state. Fluctuations of $R_G$ amount to 20% of the average value, indicating the flexibility of the chain and the lack of any definite structure. Moreover, the rmsd exceeds 15 Å at all times, demonstrating the loss of any structural similarity with the native state.

The histogram of potential energies of protein conformations at $T_F$ is bimodal (Figure 3), a characteristic of the first order-like transition.[1] At $T_F$, the protein exists in two states with equal probability. The folded state corresponds to the left peak of the histogram and the unfolded state corresponds to the right peak of the histogram. The potential energy difference between the maxima of the two peaks indicates the existence of a free energy barrier separating the folded and unfolded states.

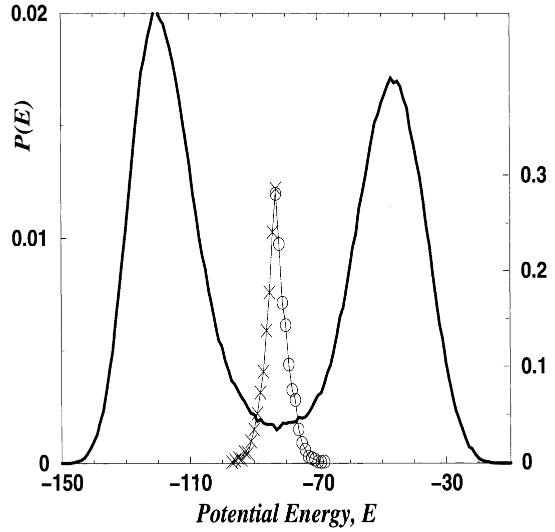

**Figure 3**. Normalized histogram of the potential energies at $T_F$ (thick line, left scale). The bimodality of the distribution is characteristic of a first order-like transition, indicating the strong cooperativity of the folding transition. On the right scale we show the histogram of potential energies of the FF ensemble (circled line) and the UU ensemble (crossed line). The last two distributions are non-zero only within a narrow energy window centered in the minimum of the energy histogram.

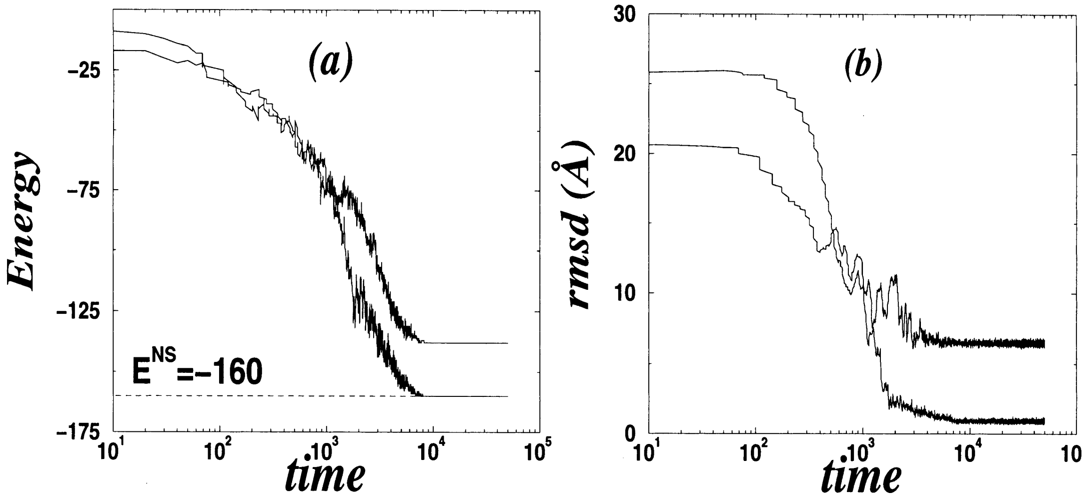

**Figure 4**. Two representatives of a set of five quenches for our model of the SH3 fold. One of the representatives reaches the native state while the other goes to a kinetic trap. The potential energy (a) and rmsd (b) evolution for the two representatives start to differ for times in the range $10^3 <$ time $< 10^4$, which are typical times needed for the folding transition of the model.

## Kinetics

We test the folding properties of our model of the SH3 fold by quenching five different, fully unfolded conformations from temperature $T = 2 \gg T_F$ to $T = 0.1 \ll T_F$. Three conformations out of five fold to the native state, and the remaining two conformations fold into kinetic traps (Figure 4). The average folding time during quenches is $10^4$.

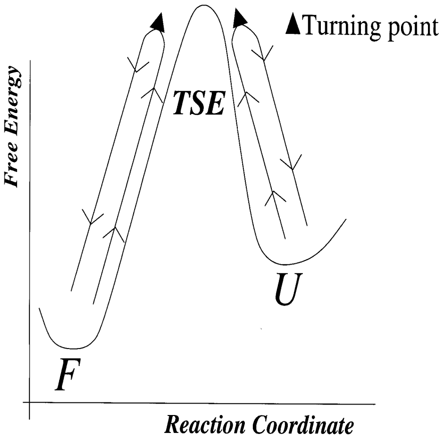

**Figure 5**. Schematic diagram for the interpretation of turning point conformations. The turning point of an FF event is the closest conformation to the TSE, but without crossing the free energy barrier to the unfolded state and with no unfolding of the protein. Analogously, the turning point of a UU event is the closest conformation to the TSE, but without crossing the free energy barrier to the folded state and with no folding of the protein. If close enough to the TSE along a hypothetical reaction coordinate, both FF and UU turning point conformations share structural similarities with protein conformations belonging to the TSE. In addition, differences between FF and UU turning point conformations determine the folding nucleus contacts, defined as the set of contacts that are formed at the top of the free energy barrier in a successful folding event.

To find the TSE and the folding nucleus, we employ a modified local fluctuations method.[20,25] While the protein is at thermal equilibrium, we differentiate between two types of fluctuations that start in the native conformation: (i) "local" disruption of the native structure followed by immediate refolding (FF event) and (ii) successful unfolding of the protein (FU event). We also differentiate between two types of fluctuations that start in the unfolded state: successful (UF) and unsuccessful (UU) folding events. In FF events, the protein does not cross the free energy barrier, but is committed to rapidly descend back to the native state, which is one minimum of the free energy. Similarly, in UU events the protein does not cross the free energy barrier and is committed to rapidly descend back to the unfolded state, which is the other minimum of the free energy.

In the framework of the nucleation scenario for protein folding, the nucleus contacts form at the top of the free energy barrier in the successful folding events. If the nucleus forms, the protein folds into the native state with high probability. If the nucleus does not form, the protein unfolds with high probability. Therefore, we assume the presence of the nucleus at the turning point of the FF trajectory, corresponding to the conformation with the maximal potential energy (Figure 5). Similarly, we assume that the nucleation contacts are not present in the unsuccessful unfolding (UU) events (Figure 5). On the other hand, we conjecture that the turning points of both FF and UU events are protein conformations with a strong structural similarity to the conformations of the TSE ensemble.

We generate two ensembles of turning points, taken from thermal fluctuations of the protein

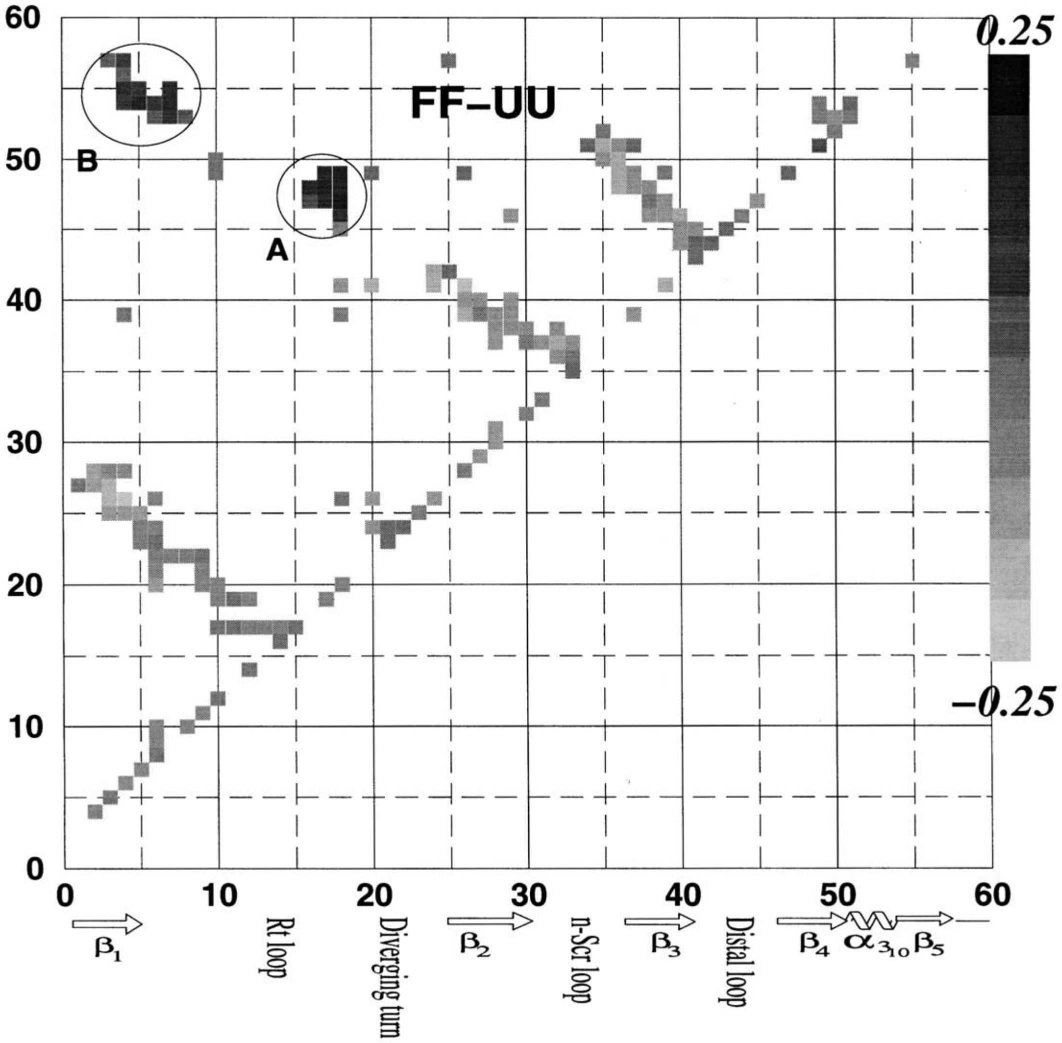

**Figure 6**. Difference contact probability map ($f_{ij}^{FF} - f_{ij}^{UU}$). Regions *A* and *B*within the ellipses have positive values, while all remaining contacts have negative values. We assign the putative folding nucleus of the SH3 fold to the cluster of contacts between the RT-loop and the distal hairpin (cluster A). We do not assign the cluster of contacts between the two termini (cluster B) as part of the folding nucleus because these contacts are not likely to form in the TSE.

near $T_F$: one ensemble for FF events and the other for UU events. If a native contact between amino acids *i* and *j* has a high contact probability, $f_{ij}$, in the FF ($f_{ij}^{FF}$) and UU ($f_{ij}^{UU}$) ensembles, the contact is a candidate to be present in the TSE. We assume that contacts with a positive difference of ($f_{ij}^{FF} - f_{ij}^{UU}$) form at the top of the free energy barrier as the protein folds, and thus may be identified as the folding nucleus. We test this assumption extensively in Methods.

We identify two clusters of amino acid contacts with maximal difference ($f_{ij}^{FF} - f_{ij}^{UU}$) (Figure 6). As discussed above, both clusters may contain the folding nucleus of the SH3 fold. One cluster is formed by contacts between amino acids of the N and C termini. The other cluster is formed by contacts between the RT-loop and the distal hairpin. We find that contacts between the N and C termini have low probabilities ($f_{ij} < 0.5$) in both FF and UU turning point ensembles. These contacts are not relevant to the folding process because they are unlikely to form as the protein crosses the TSE during folding. In contrast, a cluster of contacts between the RT-loop and the diverging turn has high contact probability values ($0.7 < f_{ij} < 1$) in the FF ensemble and moderate values ($0.5 < f_{ij} < 0.7$) in the UU ensemble. Thus, these contacts form when the protein crosses the TSE during folding and we assign them to be the folding nucleus of the SH3 fold. Our findings are in agreement with simulation studies by Ding *et al*. (unpublished results) of the TSE of the Src SH3 domain with a different model of the SH3 fold.

We argue that the cluster of contacts between the RT-loop and the distal hairpin may be stabilized by a hydrogen bond network. We find two hydrogen bonds in the native state of the c-Crk SH3 domain between amino acid residues E16 and L18 in the RT-loop, with amino acid M48 in the distal hairpin (E16–M48 and L18–M48). Both hydrogen bonds

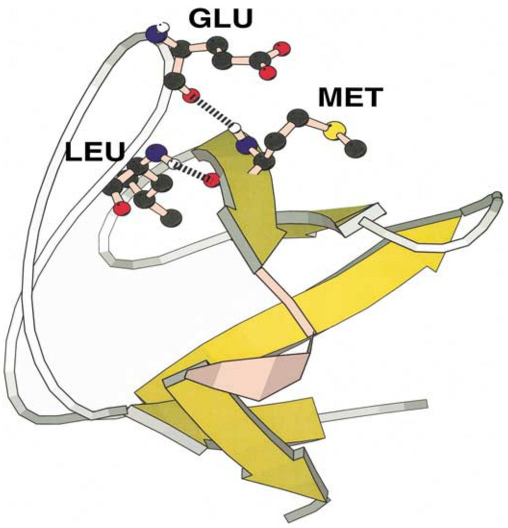

**Figure 7**. Putative folding nucleus of the c-Crk SH3 domain. Backbone hydrogen bonds E16-M48 and L18-M48 bring together the RT-loop and the distal hairpin. We also find backbone hydrogen bonds in the respective native states of the α-spectrin SH3 domain[13] (R21–F52 and V23–F52) and the Src SH3 domain (T22–Y55 and L24–Y55) at structurally equivalent positions.

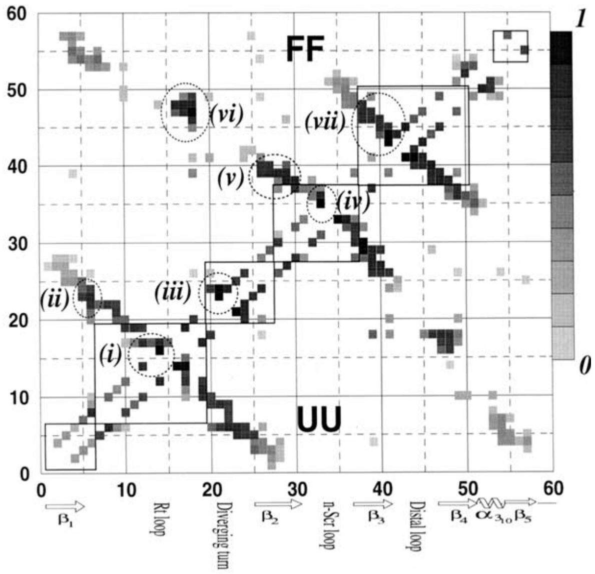

**Figure 8**. Contact probability maps for the FF (upper half) and the UU (lower half) ensembles of turning point conformations. Contacts within squares are local contacts for each of the secondary structure elements of the SH3 fold. Encircled contacts have high contact probability values ($0.7 < f_{ij} < 1$) and represent the structure of the TSE, namely: (i) turn of the RT-loop; (ii) contacts between N terminus and diverging turn; (iii) diverging turn; (iv) turn of the n-Src loop; (v) contacts between the n-Src loop and the distal hairpin; (vi) putative folding nucleus contacts between the RT-loop and the distal hairpin; (vii) distal hairpin.

are interactions of the backbone, while the relative orientation of the respective side-chains suggest no interaction between residues (Figure 7). We find hydrogen bonds in α-spectrin (R21–F52 and V23–F52)[13] and Src (T22–Y55 and L24–Y55) at structurally equivalent positions.

Next, we present our results for the TSE in terms of formation of secondary structure (Figure 1) of c-Crk.

### N terminus

Residues 1 to 6; AEYVRA: The first six amino acid residues form the first β-strand of c-Crk, which forms along with the C terminus a β-sheet in the native state. Experimental studies of α-spectrin and Src[11,14,33] show that the β-sheet is not formed in the TSE. In our simulations, we find that the 13 native contacts of our model within the β-sheet have a low contact probability ($0 < f_{ij} < 0.5$) to be present in both FF and UU turning point ensembles (Figure 8). We also find that these contacts have higher probability values in FF turning point conformations than in UU turning point conformations (Figure 6). This result is not sufficient to assign these contacts as part of the folding nucleus, since the contacts are unlikely to form in the TSE due to their low contact probabilities.

The N terminus and the diverging turn make 16 contacts in the native state of c-Crk, of which nine contacts have low probabilities ($0 < f_{ij} < 0.5$), five contacts have moderate probabilities, and two contacts (A6–K22 and A6–G23) have high probabilities ($0.7 < f_{ij} < 1$). Thus, the N terminus is not likely to contact the diverging turn in the TSE. Residue A6 is located in the c-Crk sequence at the end of the N terminus, before the beginning of the RT-loop, and residues K22 and G23 are located at the beginning of the diverging turn, before the end of the RT-loop. Thus, contacts A6–K22 and A6–G23 form in the TSE the structured base of an elongated RT-loop.

### RT-loop

Residues 7 to 19; LFDFNGNDEEDLP: Several experiments[11,14,33] show that the RT-loop has no native structure in the TSE. There are 18 local native contacts in the RT-loop of our model, the majority found with moderate ($0.5 < f_{ij} < 0.7$) contact probabilities in the TSE. Contacts in the turn of the loop (G12–D14, N13–E17, D14–E16 and D14–E17) have high contact probabilities in both FF and UU ensembles, thus we conclude that the turn of the RT-loop is structured to the same degree in the TSE as in the native structure. In realizations at temperatures higher than $T_F$, the turn is still formed while the two strands forming the loop are unordered, in accordance with the experiments[65–68] showing that the rate-limiting step in the formation of a long loop is the formation of contacts between the two strands,[69] while the turn is usually formed. Residues D17

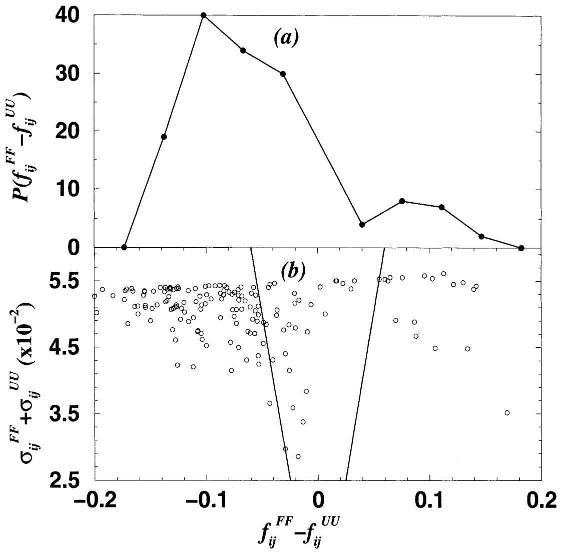

**Figure 9**. (a) The histogram of $(f_{ij}^{FF}-f_{ij}^{UU})$ values has two peaks, the smaller for contacts with positive $(f_{ij}^{FF}-f_{ij}^{UU})$ values. Most of the contacts have negative $(f_{ij}^{FF}-f_{ij}^{UU})$ values because of the specific sampling algorithm we use for FF and UU turning point conformations. (b) Estimation of the error in the value $(f_{ij}^{FF}-f_{ij}^{UU})$ for each contact ij. The two lines $y=x$ and $y=-x$ separate the insufficient sampling region (central region) from the two statistically significant regions. The error $(\sigma_{ij}^{FF}+\sigma_{ij}^{UU})$ is smaller than the absolute value of $(f_{ij}^{FF}-f_{ij}^{UU})$ in the two statistically significant regions.

and P19, located in the middle of one of the strands, have a high probability to contact with the amino acid residues across the loop in the other strand, G12 ($f_{ij}^{FF}=0.8$) and N11 ($f_{ij}^{FF}=0.81$). However, D17 and P19 have only moderate ($0.5<f_{ij}<0.7$) contact probability with residues adjacent in sequence to G12 and N11. We conclude that the structure of the RT-loop in the TSE is not stable and is flexible.

The number of non-local native contacts between the RT-loop and the rest of the protein amounts to 28 contacts. Of these 28 contacts, the diverging turn participates with six contacts and the distal hairpin participates with 15 contacts. We find moderate ($0.5<f_{ij}<0.7$) contact probabilities between residues L7, F8, D9 and F10 in the beginning of the RT-loop with residues F20 and K22 in the diverging turn. We find contact D9–K21 with a high probability ($f_{ij}^{FF}=0.71$). The 15 contacts between the RT-loop and the distal hairpin span a range of all values. Only a cluster of six contacts (out of the 15) has high contact probability values ($0.7<f_{ij}<1$) in the FF ensemble and moderate values ($0.5<f_{ij}<0.7$) in the UU ensemble (contacts E16–M48, D17–G47, D17–M48, L18–R46, L18–G47 and L18–M48). Contacts in this cluster are more likely to form in the FF ensemble than in the UU ensemble (Figures 6 and 8). Thus, the contacts form preferentially as the protein crosses the TSE during folding and we assign them to be the folding nucleus (Figure 7).

### Diverging turn

Residues 20 to 27; FKKGDILR: The turn (F20–D24) and the β-strand following the turn (I25–R27) are highly structured in the TSE of both α-spectrin and Src experiments.[11,14,33] There are five local native contacts in our model, three of them with high contact probabilities and two with moderate contact probabilities. Non-local contacts between the β-strand (I25–R27) and strand $\beta_3$ of the distal hairpin have high contact probability for UU turning points and moderate probability for FF turning points (contacts I26–V39, L27–V39, I26–E40 and L27–E40). Thus, we conclude that the diverging turn is structured in the TSE as it is in the native state.

### n-Scr loop

Residues 28 to 37; IRDKPEEQWW: In our model there are 12 native local contacts and only E33–Q35 and E33–W36, located at the turn, have high contact probabilities ($f_{ij}^{FF}=0.92, 0.73$). Residue W37, located at the end of the loop, has contacts with the first half of the loop (I28–E33), which is the part with no secondary structure in the native state. We find these contacts to have moderate contact probability values ($0.5<f_{ij}<0.7$) in the TSE. We conclude that excluding the structured turn, the loop is only partially structured in the TSE. The n-Zrc loop has non-local native contacts with the distal hairpin, the $3_{10}$ helix following the distal hairpin and the N terminus. We distinguish two different clusters of contacts between the n-Src loop and the distal hairpin: (i) the beginning of the n-Src loop forms contacts with the beginning of the distal hairpin (I28–N38, I28–A39, R29–A39, D30–N38, D30–A39) and (ii) the end of the n-Src loop forms contacts with the end of the distal hairpin (W36–M48, W36–I49, W36–P50, W37–M48, W37–I49). The first cluster is made of contacts with high probability values ($0.7<f_{ij}<1$), and is adjacent to the cluster of contacts between the diverging turn and the distal hairpin, which is also made of contacts with high probability values. The second cluster is made of contacts with low to moderate contact probability values ($f_{ij}<0.7$) and is therefore not stable in the TSE. The remaining non-local contacts with the $3_{10}$ helix and the N terminus have low contact probabilities and are therefore not likely to form in the TSE.

### Distal hairpin

Residues 38 to 50; NAEDSEGKRGMIP: According to experiments in α-spectrin and Src,[11,14,33] the distal hairpin is the most stable structure in the TSE of the SH3 fold. These experiments find that the turn is fully formed and the two β-strands of the hairpin are brought together, forming part of the hydrophobic core of the protein. The previous experiments also identify stable non-local interactions between the distal hairpin and the

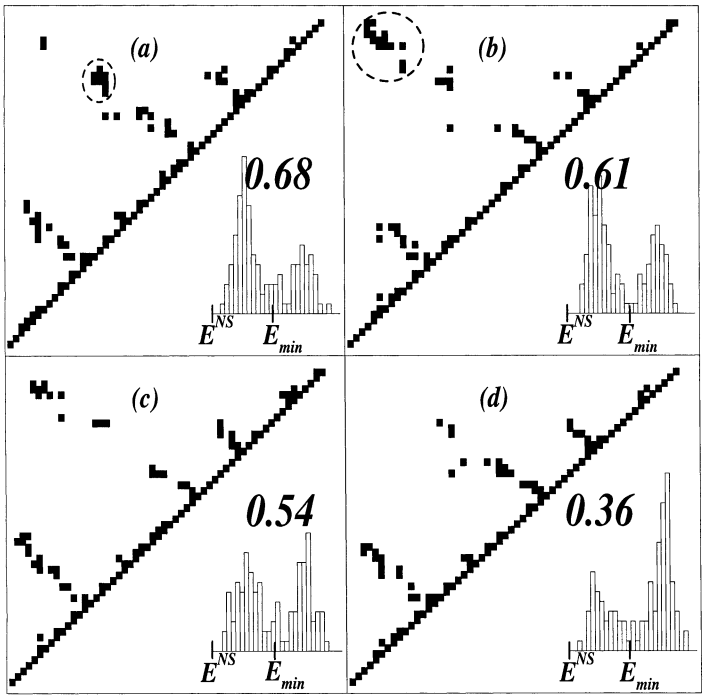

**Figure 10.** $p_{\mathrm{FOLD}}$ analysis of four different FF turning point conformations. Each box contains the contact map of the initial turning point conformation, from which 120 simulations with identical initial amino acid coordinates, but with different initial velocities, evolve for $10^4$ time steps. We record the potential energies of the last five measurements for each of the 120 simulations, totaling 600 energy values gathered in the histogram plot next to the contact map. From the normalized histogram, we calculate $p_{\mathrm{FOLD}}$ (the number shown above the histogram) as the area of the histogram in the potential energy range ($E^{\mathrm{NS}} = -160$, $E_{\mathrm{min}} = -82$). The potential energy gap between the two peaks of the histograms, averaged over all the histograms, is 70 units. The gap cannot be accounted for only by the appearance/disappearance of the long range clusters (broken ellipses in (a) and (b)), but because of the folding/unfolding of the protein. We show in (a) the folding nucleus within the broken ellipse. Conformations (a)–(c) have $p_{\mathrm{FOLD}} > 1/2$ and correlate with the number of present nucleus contacts. Conformation (d) has $p_{\mathrm{FOLD}} < 1/2$, but it has only three contacts belonging to the folding nucleus.

diverging turn and partially stable interactions with the n-Src loop. From our simulations, out of a total of 18 local native contacts, we find 12 contacts with high contact probability values ($0.7 < f_{\mathrm{ij}} < 1$) and four contacts with moderate values ($0.5 < f_{\mathrm{ij}} < 0.7$). The turn (S42, E43 and G44) has the highest values, making the turn the most stable fragment of the hairpin. Thus, we find that the hairpin is the most stable fragment of the secondary structure in the TSE of the SH3 fold. Non-local native contacts with the diverging turn (11 contacts), the n-Scr loop (13 contacts) and the RT-loop (13 contacts) have already been discussed in the previous respective subsections.

### *C terminus*

Residues 54 to 57; VEKY: The C terminus makes contacts only with the N terminus, already discussed in the N terminus subsection.

## Discussion

### Error analysis

We address the question of whether our numerical results for the different $(f_{\mathrm{ij}}^{\mathrm{FF}} - f_{\mathrm{ij}}^{\mathrm{UU}})$ values are statistically significant or if they are the result of insufficient sampling. First, we compute the histogram of $(f_{\mathrm{ij}}^{\mathrm{FF}} - f_{\mathrm{ij}}^{\mathrm{UU}})$ values, which has two peaks, the smaller one corresponding to the positive values of $(f_{\mathrm{ij}}^{\mathrm{FF}} - f_{\mathrm{ij}}^{\mathrm{UU}})$ (Figure 9(a)). Next, we estimate the expected error for each $(f_{\mathrm{ij}}^{\mathrm{FF}} - f_{\mathrm{ij}}^{\mathrm{UU}})$ value. We assume that the presence of a specific contact *ij* at any time is a random variable, with an output of 0 or 1. We analyze uncorrelated FF and UU events separated by at least an interval of 2500 time units, which is longer than the typical folding transition time. We collect $N^{\mathrm{FF}} = 153$ FF events and $N^{\mathrm{UU}} = 181$ UU events, then estimate the error in $(f_{\mathrm{ij}}^{\mathrm{FF}} - f_{\mathrm{ij}}^{\mathrm{UU}})$ as $\sigma_{\mathrm{ij}}^{\mathrm{FF}} + \sigma_{\mathrm{ij}}^{\mathrm{UU}}$, where $\sigma_{\mathrm{ij}}^{\mathrm{FF(UU)}} = \sqrt{(f_{\mathrm{ij}}^{\mathrm{FF(UU)}}(1 - f_{\mathrm{ij}}^{\mathrm{FF(UU)}})/N^{\mathrm{FF(UU)}})}$. We define $(f_{\mathrm{ij}}^{\mathrm{FF}} - f_{\mathrm{ij}}^{\mathrm{UU}})$ to be statistically significant if $|f_{\mathrm{ij}}^{\mathrm{FF}} - f_{\mathrm{ij}}^{\mathrm{UU}}| > \sigma_{\mathrm{ij}}^{\mathrm{FF}} + \sigma_{\mathrm{ij}}^{\mathrm{UU}}$ and find $(f_{\mathrm{ij}}^{\mathrm{FF}} - f_{\mathrm{ij}}^{\mathrm{UU}}) > 0.06$ to be statistically significant (Figure 9(b)).

### $p_{\mathrm{FOLD}}$ Analysis

We test the hypothesis that the FF and UU ensembles have structural similarities with the TSE in two complimentary tests. In the first test, we assume that the energy is the reaction coordinate for the folding–unfolding transition. This assumption is supported by the bimodality of the potential energy histogram at $T_{\mathrm{F}}$ (Figure 3). We then estimate the TSE as the set of conformations with energies within a narrow energy window centered in the minimum of the histogram, $E_{\mathrm{min}} = -82$. According to this criterion, both FF and UU ensembles are candidates to represent the TSE (Figure 3). Even though we accept FF turning point conformations with energies far above the minimum of the histogram ($E_{\mathrm{th,f}} = -56$), we do

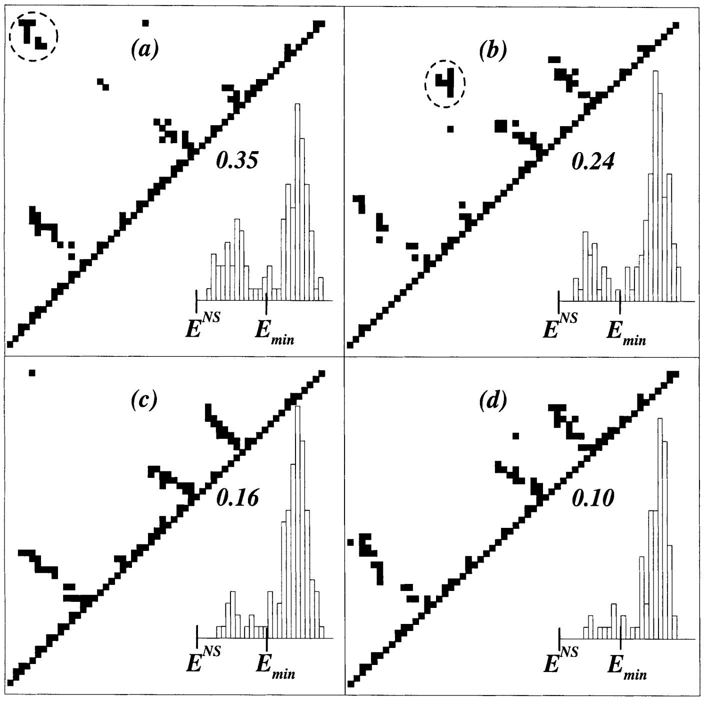

**Figure 11**. $p_{\text{FOLD}}$ analysis of four different UU turning point conformations. (a) The presence of long range contacts between the two termini (broken ellipse) does not guarantee $p_{\text{FOLD}} > 1/2$ because these contacts are not the folding nucleus (broken ellipse in (b)). (b) The contact map represents an entangled conformation. Thus, the protein unfolds first in order to fold to the native state, and $p_{\text{FOLD}} < 1/2$ even though the entangled conformation contains most of the nucleus (broken ellipse). Contact maps in (c) and (d) have most of the contacts that make the TSE, yet $p_{\text{FOLD}}$ is much lower than 0.5 because they lack the putative folding nucleus.

not find any FF turning point with an energy bigger than $E = -68$, out of an ensemble of 2620 FF turning point conformations. We retrieve 95% of the FF ensemble with an energy threshold even closer to $E_{\text{min}}$, $E_{\text{th,f}} = -76$. We also find analogous results for the UU ensemble (Figure 3).

The second test is to apply the $p_{\text{FOLD}}$ criterion according to which a conformation belongs to the TSE if it has a probability $p_{\text{FOLD}} = 1/2$ to evolve either to the folded or unfolded state. If a turning point conformation is structurally similar to the TSE conformations, then $p_{\text{FOLD}}$ is close to 1/2. Moreover, if the turning point conformation belongs to a FF event, the nucleus is likely to be present and we expect $p_{\text{FOLD}}$ to be larger than 1/2. Analogously, if the turning point conformation belongs to a UU event, the nucleus is likely to be absent and we expect $p_{\text{FOLD}}$ to be less than 1/2. Thus, $p_{\text{FOLD}}$ is a measure of the presence or absence of the folding nucleus in a turning point conformation.

The $p_{\text{FOLD}}$ analysis for one turning point conformation has five steps: (i) random selection of the turning point conformation; (ii) change of initial conditions by replacing all amino acids and heat bath particle velocities with random velocity values from a Maxwell velocity distribution corresponding to thermal equilibrium at $T_{\text{F}}$ (we perform the velocity replacement 120 times, creating 120 different initial conditions); (iii) subsequent time evolution for $10^4$ time steps, which is above the typical time of folding or unfolding; (iv) collect the potential energy values of the last five measurements, giving a total of $5 \times 120 = 600$ energy values that we plot in a normalized histogram; (v) compute $p_{\text{FOLD}}$ as the area below the histogram curve in the range of potential energies ($E^{\text{NS}} = -160$, $E_{\text{min}} = -82$).

We compute an estimate of the average $p_{\text{FOLD}}$ for the FF ensemble, $\langle p_{\text{FOLD}} \rangle_{\text{FF}}$, with the analysis of 16 randomly selected FF turning point conformations totaling $16 \times 120 = 1920$ different conformations plus initial conditions. Once simulations are over, we find that 1113 fold and 807 unfold, thus $\langle p_{\text{FOLD}} \rangle_{\text{FF}} = 0.58$. The same analysis on 13 randomly selected UU turning point conformations produces 1560 different conformations plus initial conditions, of which 622 fold and 938 unfold, thus $\langle p_{\text{FOLD}} \rangle_{\text{UU}} = 0.40$. The difference of $\langle p_{\text{FOLD}} \rangle$ with respect to 1/2 states the preference of the FF (UU) ensemble toward the folded (unfolded) state. As expected, we find that both $p_{\text{FOLD}}$ values are close to 1/2, indicating the closeness of both FF and UU ensembles to the TSE.

In addition, $p_{\text{FOLD}}$ analyses quantify the ability of the various structural elements of the TSE to bias the protein toward the folded or the unfolded states. For this study, we randomly select eight FF turning point conformations and eight UU turning point conformations with different contact maps and perform $p_{\text{FOLD}}$ analysis for each of them.

We find seven out of the eight FF turning point conformations with $p_{\text{FOLD}} \geq 1/2$ (Figure 10), and the correlation coefficient between $p_{\text{FOLD}}$ and the number of present folding nucleus contacts is 0.98. The coefficient drops to 0.86 when we include the turning point conformation with $p_{\text{FOLD}} < 1/2$ (Figure 10(d)). A similar analysis of $p_{\text{FOLD}}$ *versus* number of present contacts between the two termini gives a correlation coefficient of 0.0. In the analogous analysis of the UU ensemble, we find two turning point conformations with $p_{\text{FOLD}} > 1/2$.

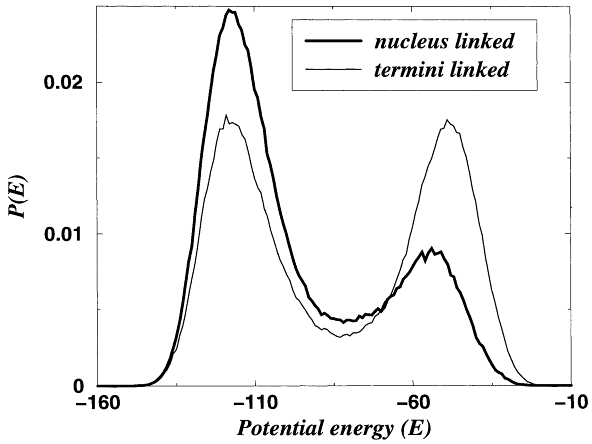

**Figure 12**. Potential energy histogram for two cross-linking simulations at $T = 0.64$. Protein with linked termini has a bimodal histogram with equal probability for the folded and the unfolded states (light line). In contrast, the protein with a linked folding nucleus (thick line) has three times more area under the left peak than under the right peak, thus favoring the folded state more than the unfolded state.

The two conformations with $p_{\mathrm{FOLD}} > 1/2$ contain the folding nucleus, thus belonging rather to the FF ensemble than to the UU ensemble. We find that conformations with no folding nucleus contacts and with no contacts between termini have the lowest $p_{\mathrm{FOLD}}$ values (Figure 11). It is interesting to observe that one of the eight UU turning point conformations contains most of the folding nucleus contacts, yet $p_{\mathrm{FOLD}}$ equals only 0.24 (Figure 11(b)). Inspection of the structure reveals that this conformation is entangled and it is necessary for the protein to unfold first in order to fold to the native state. In summary, $p_{\mathrm{FOLD}}$ analysis confirms that the putative folding nucleus plays an important role in the folding transition.

## Cross-linking of folding nucleus

We explicitly confirm the effect of a specific folding nucleus contact in the stability of the protein by a cross-linking simulation, in which we impose a permanent bond in the nucleus contact. A permanent bond reduces the entropy of the unfolded state and leaves the entropy of the folded state unchanged, thus increasing the probability of the folded state. At $T_{\mathrm{F}}$, the probability of the folded and unfolded states must be equal and therefore, the folding temperature of the cross-linked protein, $T_{\mathrm{F,new}}$, is large than the original folding temperature, $T_{\mathrm{F}}$. First, we perform a control simulation in which we cross-link the N and C termini by imposing a permanent bond between Y3 in the N terminus and Y57 in the C terminus. We selected this contact because cross-linking of the termini of a homopolymer maximally changes the entropy of the unfolded state. We find a temperature $T_{\mathrm{F,new}} = 0.64$ such that the histogram of potential energies has equal probabilities for the folded and the unfolded states. The new folding temperature is close to the original temperature $T_{\mathrm{F}} = 0.63$. Next, we perform at $T_{\mathrm{F,new}}$ a different simulation in which we replace the putative folding nucleus contact E16–M48 by a permanent bond. We find that the folded state is roughly three times more probable than the unfolded state (Figure 12) and conclude that E16–M48 is a folding nucleus contact that favors the folded state more than the control termini contact Y3–Y57.

## Conclusion

We perform systematic thermodynamic and folding kinetic studies of a model of the SH3 fold with DMD. We find a strong cooperative transition at a specific folding temperature. We find that the structure of the TSE in simulations is similar to the structure emerging from experimental studies of SH3 fold members Src and α-spectrin. We predict the TSE of the SH3 fold to have the distal hairpin and diverging turn structured as in the native state. The n-Src and RT loops are structured but flexible, and do not adopt the native conformation except for their turns, in contrast with previous theoretical predictions[35–37] where the RT-loop was fully ordered in the TSE. The reason for the disparity may be due to the assumption of these studies that each amino acid has only two possible configurations, either ordered as in the native state or disordered. While this assumption may be more appropriate for amino acids that adopt a α-helix, β-strands or tight turn conformation in the native state, it overemphasizes the change of entropy for amino acids with no secondary structure, like the RT-loop. We address the role of flexibility on the change of entropy with our protein model, which reproduces a strong cooperative transition yet does not over-restrict the number of accessible conformations of each amino acid.

We predict that the folding nucleus forms when E16, D17 and L18 in the RT-loop form contacts with R46, G47 and M48 in the distal hairpin. We hypothesize that the backbone hydrogen bond network E16–M48 and L18–M48 is the folding nucleus of c-Crk. Due to the structural similarity between c-Crk and other members of the SH3 fold family (Figure 1(c) and (d)), our results are not only applicable to c-Crk, but to these members as well. Analogous hydrogen bonds are found in the native states of α-spectrin (R21–F52 and V23–F52)[13] and Src (T22–Y55 and L24–Y55) SH3 domain proteins. The variety of amino avids (E, M, L, V, R, F, T, Y) involved in the two backbone hydrogen bonds of c-Crk, Src and α-spectrin altogether, indicates that backbone interactions are resilient to evolutionary pressure and, thus, provide a way to conserve the dynamics of folding. Experimental verification of our prediction may

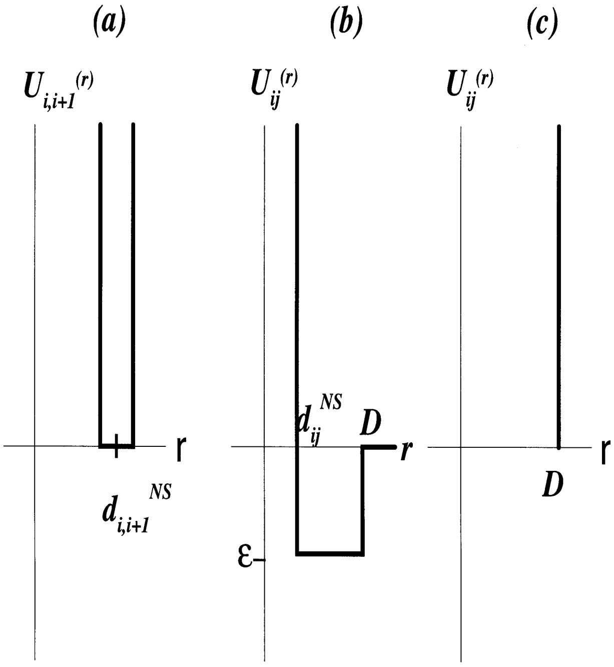

**Figure 13**. Interaction potentials used between each pair of amino acid residues (a) Covalent-like bonding for neighboring amino acid residues along the backbone. $d^{\mathrm{NS}}_{\mathrm{i,i+1}}$ is the native distance between two particular neighbor $C^\beta$s ($C^\alpha$ in case of Gly). (b) Attractive interaction with strength ε for the native contact formed by amino acid *i* and amino acid *j*. We set a repulsive barrier due to excluded volume slightly below their mutual distance in the native state, $d^{\mathrm{NS}}_{\mathrm{i,j}}$ (we take $0.99\times d^{\mathrm{NS}}_{\mathrm{i,j}}$). (c) Repulsive barrier between amino acid *i* and *j* when their native distance is bigger than the predefined cut-off distance $D = 7.5$ Å.

come from an attempt to inhibit formation of hydrogen bonds. According to our prediction, inhibition of formation of the hydrogen bond network E16–M48, D17–M48 in c-Crk, R21–F52, V23–F52 in α-spectrin and T22–Y55, L24–Y55 in Src would reduce the folding rate of these proteins.

## Methods

### Model

As the unfolded protein approaches the TSE, the number of allowed conformations reduces drastically, resulting in an abrupt decrease of entropy.[1] A successful model for protein folding must reproduce this abrupt entropy reduction. Flexibility is the main factor contributing to entropy and thus, it is crucial to define an appropriate set of effective constraints among amino acids. First, we study the folding transition of the SH3 fold with the beads-on-a-string model.[25] We find that the protein folds into the native state by passing through a series of meta-stable intermediates, giving the beads-on-a-string model a flexibility that does not reproduce the cooperative transition experimentally observed in globular proteins. Next, we introduce a set of additional constraints among amino acids in order to make the folding transition strongly cooperative.

We use the Gō model[70–72] of interactions (Figure 1(b)), determined with the knowledge of the native topology of the protein under study. We represent each of the 57 amino acid residue by beads centered in their respective $C^\beta$ coordinates[41] ($C^\alpha$ in case of Gly). We model the peptide bond between to neighboring amino acids *i* and $i+1$ along the sequence by a narrow, infinitely high potential well (Figure 13(a)), so that the bond length can fluctuate within 2% of the distance in the native state, $d^{\mathrm{NS}}_{\mathrm{i,i+1}}$, between the $C^\beta$ coordinates of the two amino acids. We define a contact between two non-bonded amino acids *i* and *j* to be a native contact when the distance between their respective $C^{\beta\prime}$ is in the native state, $d^{\mathrm{NS}}_{\mathrm{ij}}$, is smaller than some predefined cut-off distance *D*. We use $C^\alpha$ in case of Gly. We plot the resulting set of native contacts on the contact map (Figure 1). We model a native contact between amino acids *i* and *j* by an attractive square-well potential[53,70,71] of depth ε with a hard-core distance $0.99\times d^{\mathrm{NS}}_{\mathrm{ij}}$ and an attraction distance *D*

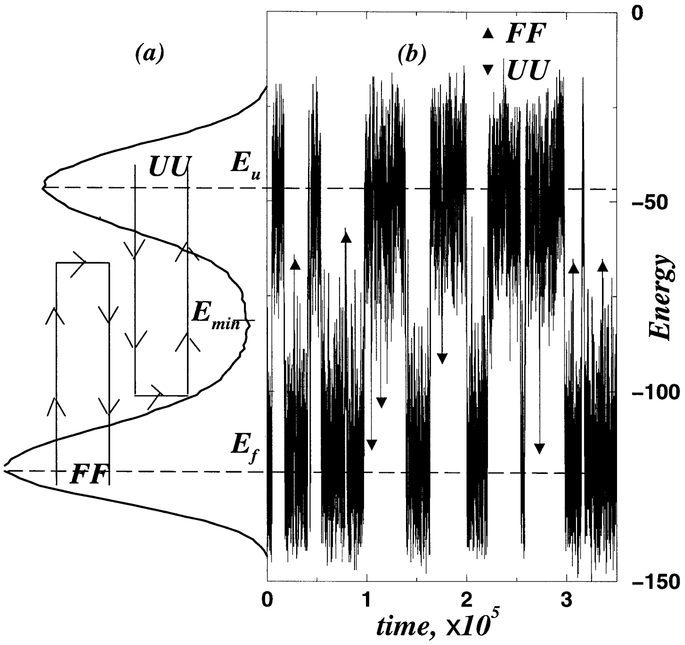

**Figure 14**. (a) A diagram of the histogram of potential energies at $T_F$. The two peaks with respective energies $E_f$ and $E_u$ correspond to the folded and unfolded states. Crossing the peaks are schematic diagrams for FF and UU events. An FF event originates below the $E_f$ level, crosses the minimum of the histogram but not the $E_u$ level and finally returns below $E_f$. Analogously, a UU event originates above the $E_u$ level, crosses the minimum of the histogram but not the $E_f$ level and finally returns above $E_u$. (b) Typical trajectory of the potential energy of the protein at $T_F$. Some of the FF (up triangle) and UU (down triangle) turning point conformations are marked. Note that FF turning point conformations have higher potential energies than UU turning point conformations.

(Figure 13(b)). If two amino acids are neither neighbors along the sequence, nor form a native contact, i.e. if $d_{ij}^{NS} > D$, we model their interaction by a hard core potential barrier located at distance $D$ (Figure 13(c)). The latter constraint stiffens the backbone, a property reported to increase the cooperativity of the coil-to-globule transition.[1,73,74] In our model $\varepsilon = -1$, $D = 7.5 Å$[75] and all amino acids have unit mass.

We thermally equilibrate the protein model with a heat bath by elastic collisions between the amino acids and a $10^3$ heat bath particles. The total energy of the system, consisting of the protein chain plus the heat bath particles, is conserved. We set the initial temperature of the system by adjusting the kinetic energy of all the particles and we define the temperature of the simulation, $T$, as 2/3 of the time average kinetic energy per particle in units of $|\varepsilon|/k_B$. We perform simulations for $10^5$ time units in the range of temperatures $0.1 < T < 0.8$. Near $T_F$, we increase the simulation time to $10^6$ time units to make a more precise determination of $T_F$. After each simulation is equilibrated for $10^4$ time units, we compute time averages of the temperature $T$, potential energy $E$ (measured in units of $|\varepsilon|$), specific heat $C_v$ (computed from fluctuations of the potential energy), radius of gyration $R_g$ and root mean square distance rmsd.[64]

### Kinetics

In the narrow temperature range near the folding transition, the protein frequently undergoes reversible folding and unfolding transitions (on average, one transition per $4 \times 10^5$ time units). The distribution of the potential energy at $T_F$ is bimodal, with two well-separated peaks centered at $E_f$ and $E_u$, corresponding to the folded and unfolded states, respectively (Figure 14(a)). At this temperature, we define an FF event as a fluctuation that: (i) begins in the folded state with a potential energy below $E_f$; (ii) crosses above the minimum of the potential energy histogram, $E_{min}$; (iii) reaches a maximum potential energy which must be smaller than a threshold energy $E_{th,f} = (1 - \alpha)E_f + \alpha E_u$, where $0.5 < \alpha < 1$ is a predefined constant and (iv) returns to the folded state with an energy again below $E_f$. We introduce the thresold energy $E_{th,f} < E_u$ in order to discriminate FF events from successfully unfolding (FU) events in which the potential energy exceeds $E_u$. Similarly, we define a UU event as a fluctuation that: (i) begins in the unfolded state with a potential energy above $E_u$; (ii) crosses below the minimum of the energy histogram, $E_{min}$; (iii) reaches a minimum potential energy, that must be larger than a threshold energy $E_{th,u} = (1 - \alpha)E_u + \alpha E_f$ with the same constant $\alpha$ and (iv) returns to the unfolded state with an energy again above $E_u$.

We define the turning point conformation of an FF (or a UU) event to be the confirmation with the highest (or lowest) potential energy reached during such an event. We generate ensembles of FF and UU turning point conformations from a set of ten simulations at $T_F$, each $5 \times 10^6$ time step units (Figure 14(b)). We fix the energy thresholds $E_{th,f}$ and $E_{th,u}$ by choosing the parameter $\alpha = 7/8$. For each native contact, we measure its probability $f_{ij}^{FF}$ to be found in the FF turning point ensemble and its probability $f_{ij}^{UU}$ to be found in the UU turning point ensemble. Due to the specific sampling algorithm of FF and UU events that we use, FF turning point conformations have higher potential energies that UU turning point conformations. Consequently, FF turning point conformations have less number of native contacts than UU turning point conformations. As a consequence, $(f_{ij}^{FF} - f_{ij}^{UU})$ is negative for the majority of native contacts. However, some contacts have positive $(f_{ij}^{FF} - f_{ij}^{UU})$ differences and we select these contacts as candidates for the folding nucleus.

## Acknowledgments

We thank Pollyanna R. G. Chavez, Feng Ding, Chris M. Dobson and Ilike Merey for useful discussions and revision of the manuscript. We acknowledge the support of the Petroleum Research Fund.

## References

1. Shakhnovich, E. I. (1997). Theoretical studies of protein-folding thermodynamics and kinetics. *Curr. Opin. Struct. Biol.* **7**, 29–40.
2. Micheletti, C., Banavar, J. T., Martian, A. & Seno, F. (1998). Protein structures and optimal folding emerging from a geometrical variational principle. *Phys. Rev. Letters*, **82**, 3372–3375.
3. Villegas, V., Martinez, J. C., Aviles, F. X. & Serrano, L. (1998). Structure of the transition state in the folding process of human procarboxypeptidase a2 activation domain. *J. Mol. Biol.* **283**, 1027–1036.
4. Chiti, F., Taddei, N., White, P. M., Bucciantini, M., Megherini, 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.
5. Lorch, M., Mason, J., Clarke, A. & Parker, M. (1999). Effects of core mutations on the folding of a beta-sheet protein: implications for backbone organization in the i-state. *Biochemistry*, **38**, 1377–1385.
6. Kim, D. E., Fisher, C. & Baker, D. (2000). A breakdown of symmetry in the folding transition state of protein L. *J. Mol. Biol.* **298**, 971–984.
7. McCallister, E. L., Alm, E. & Baker, D. (2000). Critical role of β-hairpin formation in protein G folding. *Nature Struct. Biol.* **7**, 669–673.
8. Jackson, S. E. & Fersht, A. R. (1991). Folding of Chymotrypsin inhibitor-2. I. Evidence for a 2-state transition. *Biochemistry*, **30**, 10428–10435.
9. Viguera, A. R., Martinez, J. C., Filimonov, V. V., Mateo, P. L. & Serrano, L. (1994). Thermodynamic and kinetic-analysis of the SH3 domain of Spectrin shows a 2-state folding transition. *Biochemistry*, **33**, 10925–10933.
10. Villegas, V., Azuaga, A., Catasus, L., Reverter, D., Mateo, P. L., Aviles, F. X. & Serrano, L. (1995). Evidence for a 2-state transition in the folding process of the activation domain of human procarboxypeptidase-a2. *Biochemistry*, **46**, 15105–15110.
11. Martinez, J. C., Viguera, A. R., Berisio, R., Wilmanns, M., Mateo, P. L., Filimonov, V. V. & Serrano, L. (1999). Thermodynamic analysis of alpha-spectrin SH3 and two of its circular permutants with different loop lengths: discerning the reasons for rapid folding in proteins. *Biochemistry*, **38**, 549–559.
12. Filimonov, V. V., Azuaga, A. I., Viguera, A. R., Serrano, L. & Mateo, P. L. (1999). A thermodynamic analysis of a family of small globular proteins: SH3 domains. *Biophys. Chem.* **77**, 195–208.

13. Viguera, A. R., Serrano, L. & Wilmanns, M. (1996). Different folding transition states may result in the same native structure. *Nature Struct. Biol.* **10**, 874–880.
14. Grantcharova, V. P., Riddle, D. S., Santiago, J. N. & Baker, D. (1998). Important role of hydrogen bonds in the structurally polarized transition state for folding of the src SH3 domain. *Nature Struct. Biol.* **8**, 714–720.
15. Grantcharova, V. P. & Baker, D. (1997). Folding dynamics of the Src SH3 domain. *Biochemistry*, **36**, 15685–15692.
16. Otzen, D. E., Itzhaki, L. S., Elmasry, N. F., Jackson, S. E. & Fersht, A. R. (1994). Structure of the transition-state for the folding/unfolding of the barley Chymotrypsin inhibitor-2 and its implications for mechanisms of protein-folding. *Proc. Natl Acad. Sci. USA*, **22**, 10426–10429.
17. Knapp, S., Mattson, P. T., Christova, P., Berndt, K. D., Karshikoff, A., Vihinen, M. *et al.* (1998). Thermal unfolding of small proteins with SH3 domain folding pattern. *Proteins: Struct. Funct. Genet.* **23**, 309–319.
18. Jackson, S. E., Elmasry, N. & Fersht, A. R. (1993). Structure of the hydrophobic core in the transition-state for folding of Chymotrypsin inhibitor-2: a critical test of the protein engineering method of analysis. *Biochemistry*, **32**, 11270–11278.
19. Shakhnovich, E. I., Abkevich, V. I. & Ptitsyn, O. (1996). Conserved residues and the mechanism of protein folding. *Nature*, **379**, 96–98.
20. 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.
21. Pande, V. S., Grosberg, A. Y., Tanaka, T. & Rokhsar, D. S. (1998). Pathways for protein folding: is a new view needed? *Curr. Opin. Struct. Biol.* **1**, 68–79.
22. Mirny, L. A., Abkevich, V. I. & Shakhnovich, E. I. (1998). How evolution makes proteins fold quickly. *Proc. Natl Acad. Sci. USA*, **95**, 4976–4981.
23. Shakhnovich, E. I. (1998). Folding nucleus: specific or multiple? *Folding Des.* **3**, R108–R111.
24. Gutin, A. M., Abkevich, V. I. & Shakhnovich, E. I. (1998). A protein engineering analysis of the transition state for protein folding: simulation in the lattice model. *Folding Des.* **3**, 459–480.
25. Dokholyan, N. V., Buldyrev, S. V., Stanley, H. E. & Shakhnovich, E. I. (2000). Identifying the protein folding nucleus using molecular dynamics. *J. Mol. Biol.* **296**, 1183–1188.
26. Fersht, A. R. (1995). Characterizing transition-states in protein folding: an essential step in the puzzle. *Curr. Opin. Struct. Biol.* **1**, 79–84.
27. Yu, H., Rosen, M. K., Shin, T. B., Seidel-Dugan, C. & Brugde, J. S. (1992). Solution structure of the SH3 domain of Src and identification of its ligand-binding site. *Science*, **258**, 1665–1668.
28. Yu, H., Chen, J. K., Feng, S., Dalgarno, D. C. & Brauer, A. W. (1994). Structural basis for the binding of proline-rich peptides to SH3 domains. *Cell*, **76**, 933–945.
29. Feng, S., Chen, J. K., Yu, H., Simons, J. A. & Schreiber, S. L. (1994). Two binding orientations for peptides to the Src SH3 domain: development of a general model for SH3–ligand interactions. *Science*, **266**, 1241–1247.
30. Feng, S., Kasahara, C., Rickles, R. J. & Schreiber, S. L. (1995). Specific interactions outside the proline-rich core of two classes of Src homology 3 ligands. *Proc. Natl Acad. Sci. USA*, **92**, 12408–12415.
31. Combs, A. P., Kapoor, T. M., Feng, S. & Chen, J. K. (1996). Protein structure-based combinatorial chemistry: discovery of non-peptide binding elements to Src SH3 domain. *J. Am. Chem. Soc.* **118**, 287–288.
32. Feng, S. B., Kapoor, T. M., Shirai, F., Combs, A. P. & Schreiber, S. L. (1996). Molecular basis for the binding of SH3 ligands with non-peptide elements identified by combinatorial synthesis. *Chem. Biol.* **8**, 661–670.
33. Riddle, D. S., Grantcharova, V. P., Santiago, J. V., Ruczinski, A. E. & Baker, D. (1999). Experiment and theory highlight role of native state topology in SH3 folding. *Nature Struct. Biol.* **6**, 1016–1024.
34. Musacchio, A., Noble, M., Pauptit, R., Wierenga, R. & Saraste, M. (1992). Crystal-structure of a src-homology-3 (SH3) domain. *Nature*, **359**, 851.
35. Munoz, V. & Eaton, W. A. (1999). A simple model for calculating the kinetics of protein folding from three-dimensional structures. *Proc. Natl Acad. Sci. USA*, **96**, 11311–11316.
36. Galzitskaya, O. V. & Finkelstein, A. V. (1999). A theoretical search for folding/unfolding nuclei in three-dimensional protein structures. *Proc. Natl Acad. Sci. USA*, **96**, 11299–11304.
37. Alm, E. & Baker, D. (1999). Prediction of protein-folding mechanisms from free-energy landscapes derived from native structures. *Proc. Natl Acad. Sci. USA*, **96**, 11305–11310.
38. Guerois, R. & Serrano, L. (2000). The SH3-fold family: experimental evidence and prediction of variations in the folding pathways. *J. Mol. Biol.* **304**, 967–982.
39. Wu, X. D., Knudsen, B., Feller, S. M., Sali, J., Cowburn, D. & Hanafusa, H. (1995). Structural basis for the specific interaction of lysine-containing proline-rich peptides with the n-terminal SH3 domain of c-Crk. *Structure*, **2**, 215–226.
40. Branden, C. & Tooze, J. (1999). *Introduction to Protein Structure*, Garland Publishing Inc, New York.
41. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H. *et al.* (2000). The protein data bank. *Nucl. Acids Res.* **28**, 235–242.
42. Holm, L. & Sander, C. (1996). Mapping the protein universe. *Science*, **273**, 595–602.
43. Dokholyan, N. V. & Shakhnovich, E. I. (2001). Understanding hierarchical protein evolution from first principles. *J. Mol. Biol.* **312**, 289–307.
44. Kolinski, A., Galazka, W. & Skolnick, J. (1996). On the origin of the cooperativity of protein folding: implications from model simulations. *Proteins: Struct. Funct. Genet.* **26**, 271–287.
45. Doye, J. P. K. & Wales, D. J. (1996). On potential energy surfaces and relaxation to the global minimum. *J. Chem. Phys.* **105**, 8428–8445.
46. Abkevich, V. I., Gutin, A. M. & Shakhnovich, E. I. (1995). Domains in folding of model proteins. *Protein Sci.* **4**, 1167–1177.
47. Abkevich, V. I., Gutin, A. M. & Shakhnovich, E. I. (1994). Free-energy landscape for protein-folding kinetic: intermediates, traps, and multiple pathways in theory and lattice model simulations. *J. Chem. Phys.* **7**, 6052–6062.
48. Shakhnovich, E. I. (1994). Proteins with selected sequences fold into unique native conformation. *Phys. Rev. Letters*, **72**, 3907–3910.

49. Hao, M. H. & Scheraga, H. A. (1994). Monte Carlo simulation of a first-order transition for protein-folding. *J. Phys. Chem.* **98**, 4940–4948.
50. Camacho, C. J. & Thirumalai, D. (1993). Kinetics and thermodynamics of folding in model proteins. *Proc. Natl Acad. Sci. USA*, **90**, 6369–6372.
51. Skolnick, J., Kolinski, A., Brooks, C. L., Godzik, A. & Rey, A. (1993). A method for predicting protein-structure from sequence. *Curr. Biol.* **3**, 414–423.
52. Sali, A., Shakhnovich, E. & Karplus, M. (1994). How does a protein fold. *Nature*, **6477**, 248–251.
53. Dokholyan, N. V., Buldyrev, S. V., Stanley, H. E. & Shakhnovich, E. I. (1998). Molecular dynamics studies of folding of a protein-like model. *Folding Des.* **3**, 577–587.
54. Clementi, C., Nymeyer, H. & Onuchic, J. N. (2000). Topological and energetic factors: what determines the structural details of the transition state ensemble and en-route intermediates for protein folding? An investigation for small globular proteins. *J. Mol. Biol.* **5**, 937–953.
55. Irback, A. & Schwarze, H. (1995). Sequence dependence of self-interacting random chains. *J. Phys. A: Math. Gen.* **28**, 2121–2132.
56. Berriz, G. F., Gutin, A. M. & Shakhnovich, E. I. (1997). Cooperativity and stability in a Langevin model of protein like folding. *J. Chem. Phys.* **106**, 9276–9285.
57. Guo, Z. & Brooks, C. L. (1997). Thermodynamics of protein folding: a statistical mechanical study of a small all-β protein. *Biopolymers*, **42**, 745–757.
58. Zhou, Y. Q., Karplus, M., Wichert, J. M. & Hall, C. K. (1997). Equilibrium thermodynamics of homopolymers and clusters: molecular dynamics and Monte Carlo simulations of systems with square-well interactions. *J. Chem. Phys.* **24**, 10691–10708.
59. Alder, B. J. & Wainwright, T. E. (1959). Studies in molecular dynamics. I. General method. *J. Chem. Phys.* **31**, 459–466.
60. Allen, M. P. & Tildesley, D. J. (1987). *Computer Simulation of Liquids*, Clarendon Press, Oxford.
61. Rapaport, D. C. (1997). *The Art Of Molecular Dynamics Simulation*, Cambridge University Press, Cambridge.
62. Grosberg, A. Y. & Khokhlov, A. R. (1997). *Giant Molecules*, Academic Press, Boston, MA.
63. Doi, M. (1997). *Introduction to Polymer Physics*, Oxford University Press, New York.
64. Kabsch, W. (1978). A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallog. sect. A*, **34**, 327–380.
65. Matouschek, A., Serrano, L. & Fersht, A. R. (1992). The folding of an enzyme. IV. Structure of an intermediate in the refolding of barnase analyzed by a protein engineering procedure. *J. Mol. Biol.* **224**, 819–835.
66. Matouschek, A., Serrano, L., Meiering, E. M., Bycroft, M. & Fersht, A. R. (1992). The folding on an enzyme. V. H/H-2 exchange nuclear-magnetic-resonance studies on the folding pathway of barnase: complementarity to and agreement with protein engineering studies. *J. Mol. Biol.* **224**, 837–845.
67. Serrano, L., Kellis, J. T., Cann, P., Matouschek, A. & Fersht, A. R. (1992). The folding of an enzyme. II. Substructure of barnase and the contribution of different interactions to protein stability. *J. Mol. Biol.* **224**, 783–804.
68. Serrano, L., Matouschek, A. & Fersht, A. R. (1992). The folding of an enzyme. III. Structure of the transition state for unfolding of barnase analyzed by a protein engineering. *J. Mol. Biol.* **224**, 805–818.
69. Viguera, A. R., Blanco, F. J. & Serrano, L. (1995). The order of secondary structure elements does not determine the structure of a protein but does affect its folding kinetics. *J. Mol. Biol.* **247**, 670–681.
70. Taketomi, H., Ueda, Y. & Gō, N. (1975). Studies on protein folding, unfolding and fluctuations by computer simulations. *Int. J. Pept. Protein Res.* **7**, 445–459.
71. Gō, N. & Abe, H. (1975). Non-interacting local-structure model of folding and unfolding transition in globular proteins. I. Formulation. *Biopolymers*, **20**, 991–1011.
72. Abe, H. & Gō, N. (1981). Noninteracting local-structure model of folding and unfolding transition in globular proteins. ii. Application to two-dimensional lattice proteins. *Biopolymers*, **20**, 1013–1031.
73. Lifshitz, I. M., Grosberg, A. Y. & Khohlov, A. R. (1978). Some problems of statistical physics of polymers with volume interactions. *Rev. Mod. Phys.* **50**, 683–713.
74. Sfatos, C. M., Gutin, A. M. & Shakhnovich, E. I. (1993). Phase diagram of random copolymers. *Phys. Rev. E*, **48**, 465–475.
75. Jernigan, R. L. & Bahar, I. (1996). Structure-derived potentials and protein simulations. *Curr. Opin. Struct. Biol.* **2**, 195–209.

***Edited by A. R. Fersht***

*(Received 5 November 2001; received in revised form 21 February 2002; accepted 21 February 2002)*