Article No. jmbi.1999.2949 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (1999) 291, 215–225

JMB

Hierarchy of Structure Loss in MD Simulations of src SH3 Domain Unfolding

Jerry Tsai[1], Michael Levitt[2] and David Baker[1]*

[1]*Department of Biochemistry University of Washington Seattle, WA 19195, USA*

[2]*Department of Structural Biology, Stanford University Stanford, CA 94305-5400, USA*

To complement experimental studies of the src SH3 domain folding, we studied 30 independent, high-temperature, molecular dynamics simulations of src SH3 domain unfolding. These trajectories were observed to differ widely from each other. Thus, rather than analyzing individual trajectories, we sought to identify the recurrent features of the high-temperature unfolding process. The conformations from all simulations were combined and then divided into groups based on the number of native contacts. Average occupancies of each side-chain hydrophobic contact and hydrogen bond in the protein were then determined. In the symmetric funnel limit, the occupancies of all contacts should decrease in concert with the loss in total number of native contacts. If there is a lack of symmetry or hierarchy to the unfolding process, the occupancies of some contacts should decrease more slowly, and others more rapidly. Despite the heterogeneity of the individual trajectories, the ensemble averaging revealed an order to the unfolding process: contacts between the N and C-terminal strands are the first to disappear, whereas contacts within the distal β-hairpin and a hydrogen-bonding network involving the distal loop β-turn and the diverging turn persist well after the majority of the native contacts are lost. This hierarchy of events resembles but is somewhat less pronounced than that observed in our experimental studies of the folding of src SH3 domain.

Keywords: molecular dynamics simulation; hierarchy of protein unfolding; src SH3 domain

**Corresponding author*

Introduction

Molecular dynamics (MD) simulations provide a means to obtain atomic level views of protein unfolding not accessible to direct experimental measurement (Brooks, 1998). Despite inaccuracies in the potentials used in the simulations and alterations to folding energetics caused by the artificial conditions used to bring about unfolding, this detail can potentially provide considerable insight into the folding process. Simulations are particularly useful for proteins for which there is extensive experimental data, as illustrated by the very successful collaboration between the Fersht and Daggett groups (Bond *et al.*, 1997; Daggett *et al.*, 1996; Ladurner *et al.*, 1998). These two groups compared experimental data on the folding mechanism, primarily phi values (Fersht, 1985), to analogous quantities computed from unfolding simulations.

Having thoroughly characterized the folding of the src SH3 domain using biophysical and mutational studies (Grantcharova & Baker, 1997; Grantcharova *et al.*, 1998; Riddle *et al.*, 1997; Yi *et al.*, 1998), we turned to MD simulations to obtain an atomic level description of the unfolding process to complement the experimental data. The src SH3 domain is a small (57 residue) predominantly β-sheet protein, which lacks disulfide bonds (Xu *et al.*, 1997; Yu *et al.*, 1993). Work from a number of laboratories has made the src SH3 domain one of the most experimentally characterized models for β-sheet folding. A very large number (>400) of naturally occurring SH3 domain sequences and a phage display selection for simplified SH3 domain sequences that retain the ability to fold (Riddle *et al.*, 1997) provide an almost unparalleled database of information about the sequence determinants of this simple fold. The kinetics and thermodynamics of folding are well described by a two-state model (Grantcharova & Baker, 1997),

Abbreviations used: MD, molecular dynamics; TS, transition state.

E-mail address of the corresponding author: jotter@felix.bchem.washington.edu

0022-2836/99/310215–11 $30.00/0

and the folding transition state ensembles of two SH3 domains of very different sequences have been characterized (Grantcharova *et al.*, 1998; Martinez *et al.*, 1998). This protein has also been previously studied using MD simulation (van Aalten *et al.*, 1996), but the study primarily investigated the protein's stability using various solvent models.

Because experiments measure the average properties of very large numbers of independent folding/unfolding events, our focus here is not on the properties of individual unfolding trajectories, but on features common to most of them. To identify such features, 30 independent unfolding simulations were carried out using the wild-type src SH3 domain. The recurrent properties of the unfolding process were characterized using ''disappearance plots'' which compare the extent to which the different native hydrophobic contacts and hydrogen bonds in the protein are disrupted as the protein structure comes apart. Our approach is similar in spirit to previous studies of the statistical properties of large ensembles of configurations from independent simulations (Bozkco & Brooks, 1995; Lazaridis & Karplus, 1997).

Results

Description of src SH3 domain structure and plots

Because contact maps play an important role in our analysis, we briefly summarize how the major structural features of the src SH3 domain are represented by such plots (Figure 1). In this work, side-chain hydrophobic contacts are defined using the Voronoi construction (Voronoi, 1908), and hydrogen bonds are defined using simple geometric criteria (see Methods). Shaded areas above the diagonal in the contact map indicate hydrophobic contacts between residue pairs, while shaded areas below the diagonal indicate all hydrogen bonds (main-chain to main-chain, side-chain to main-chain, and side-chain to side-chain). The src SH3 domain consists of seven strands, two loops, one hairpin, one turn, and a 3_{10}-helix. Because certain strands (2, 3, and 4) do not make classic β-sheet main-chain hydrogen bonds, they are not shown as β-strands in Figure 1; however, for clarity and simplicity these sequences will be considered strands in this study. Following conventions from previous work, we number the residues from 9 to 65 (Yu *et al.*, 1993). In a classic antiparallel β-sheet fashion, the first strand (strand 1, residues 9 to 13) and last strand (strand 7, residues 61 to 65) interact *via* hydrophobic contacts (upper left corner) and hydrogen bonds (lower right corner). The hydrophobic contacts and a single hydrogen bond close to the diagonal in the upper right-hand corner of the plot reflect the 3_{10}-helix (residues 57 to 61) just before strand 7. The next three strand interactions are highlighted in Figure 1. The irregular line of interactions perpendicular to the diagonal in the lower left corner corresponds to the interactions between strand 2 (residues 14 to 19) and strand 3 (residues 23 to 28), which form the relatively disordered RT loop. Extending from the approximate center of the diagonal, the hydrophobic contacts (going up) and hydrogen bonds (going down) between strand 4 (residues 32 to 38) and strand 5 (residues 42 to 47) correspond to the n-src loop. The hydrophobic interactions and hydrogen bonds between strand 5 (residues 42 to 47) and strand 6 (residues 50 to 55) correspond to the distal β-hairpin.

Overview of the simulations

MD simulations of src SH3 domain unfolding were carried out using ENCAD (Levitt *et al.*, 1995) under conditions summarized in Table 1. Ten independent simulations were run at 298 K, and 30 independent simulations were run at 498 K. Each run lasted for 1 ns or 500,000 2 fs time-steps. The only difference between the various simulations at a given temperature was that a different random seed was used in the initial temperature equilibration. All properties discussed here are averages over all runs at one of the two simulation temperatures. As evidenced by the small root-mean-squared deviation of the α carbon atoms (C^{α} rmsd) from the native structure at the end of 1 ns (Figure 2), the low temperature simulations produced a population of near-native protein simulations. The slight increase in rms is due to fluctuations within the n-src and RT loops and between the N and C termini (compare Figures 1 and 3). In contrast, at the end of the high-temperature simulations, the C^{α} rmsd was around 8 Å indicating that the structures have largely unfolded. The contact maps in Figure 3 show the average occupancies of side-chain hydrophobic contacts and all hydrogen bonds from the high and low temperature simulations. At low-temperatures, the structure is largely intact with high average occupancies for most native interactions, while at high temperatures, much of the structure is disrupted.

How similar are the different unfolding trajectories?

As shown in Figure 4, the occupancies of different contacts as a function of time differed drastically between the various trajectories at 498 K. Rather than analyzing individual runs, we focused on analyzing properties of the ensemble of structures created by combining all the trajectories together. The thick, bold lines in Figure 4 indicate the mean occupancy of individual contacts averaged over all conformations from all structures within the indicated time interval. Despite the large variations among the 30 trajectories, there are noticeable differences in the average rate of interaction loss between certain regions of the SH3 structure.

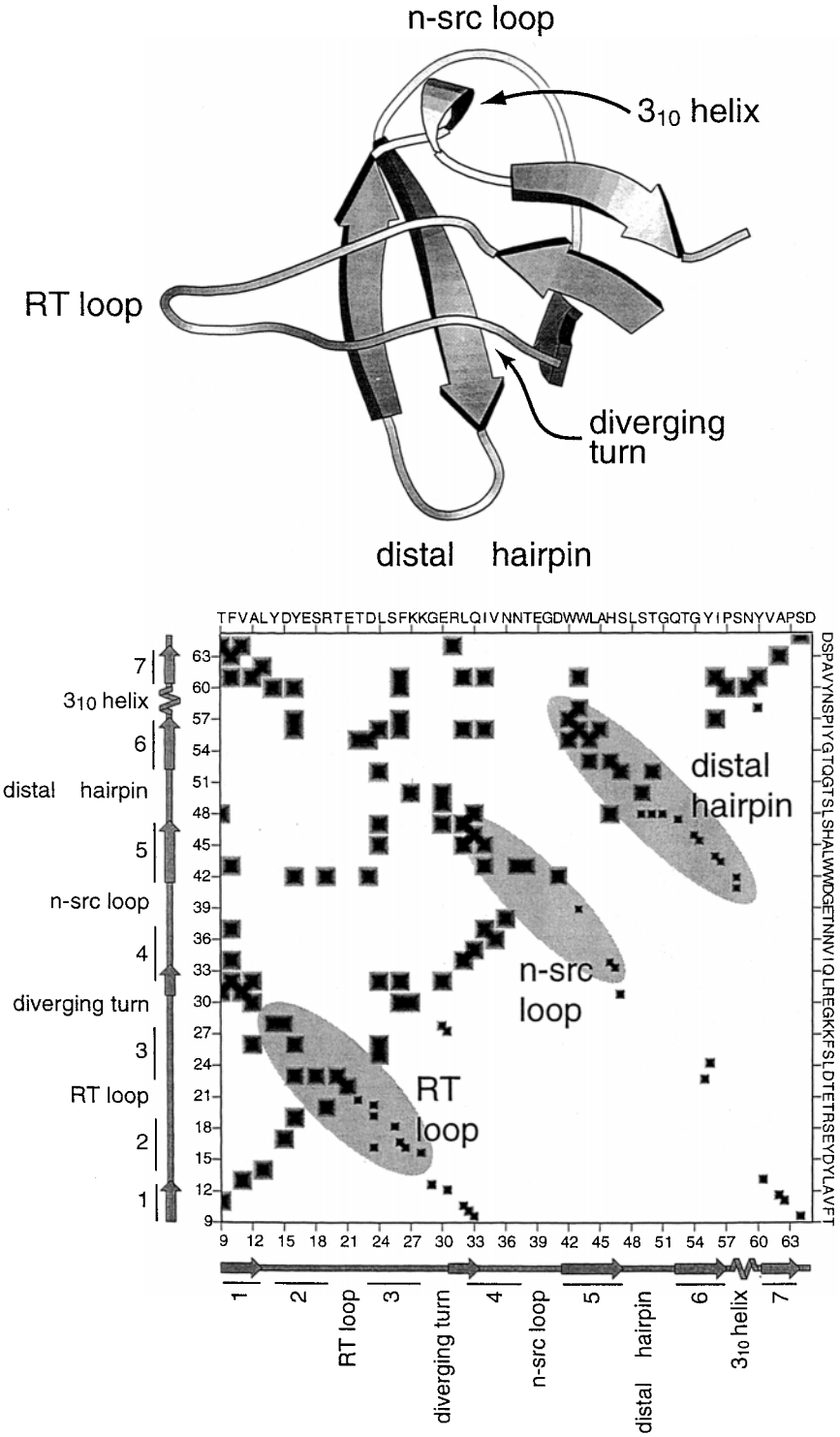

Figure 1. Description of the src SH3 domain. The structure of the src SH3 domain (Xu *et al.*, 1997) is shown in cartoon style. Secondary structure is shown only for those regions making appropriate hydrogen bonds. Important regions of the protein are labeled. Below the structure is a contact map showing the native side-chain hydrophobic contacts and hydrogen bonds. The side-chain hydrophobic contact map is shown above the diagonal, and the map of all hydrogen bonds is shown below. The positions of the secondary structures are indicated along the axes, as well as the definitions of regions described in the text. Opposite each axis is the sequence for reference. Contacts and hydrogen bonds near the diagonal are made between residues close in sequence. The sets of contacts in each of the three hairpin loops in the protein are shaded.

Table 1. Description of simulations

Description	Temperature (K)	Box size (Å^3)	Number[a]
Stable	298	75,783	10
Unfolding	498	89,573	30

Besides the differences noted above, the two sets were derived from simulations of the SH3 domain of src tyrosine kinase (residues 9 to 65) surrounded by 2261 water molecules. Each simulatios used a 2 fs time-step and trajectories were sampled every ps.

[a] The number of 1 ns simulations run under those conditions.

When it became clear that averaging over the different trajectories would be necessary, we explored several different measures for grouping conformations. Averages within the time regime correspond to what is found experimentally, but any hierarchy in unfolding may be obscured because the protein unfolds at different times in different simulations, as portrayed by Figure 4. Grouping according to the C^{α} rmsd produced more homogeneous populations of unfolded structures and a clearer view of the unfolding progression (data not shown). However, the best results were obtained with a measure that is better correlated to the loss of native structure: the percentage of native side-chain hydrophobic contacts and hydrogen bonds (Figure 5). This measure is analogous to the reaction coordinate Q often used in studies of lattice models of protein folding (Shakhnovich *et al.*, 1991; Socci *et al.*, 1996). A similar choice of reaction coordinate has been used in previous MD simulation studies (Guo *et al.*, 1997; Lazaridis & Karplus, 1997).

Hierarchy of unfolding

Figure 5 shows the data for the 498 K set of simulations divided into four equal intervals according to the percentage of native interactions. For the sake of simplicity, each interval will be referred to by its central value, i.e. the 100 to 75 % interval will be referred to as 87.5 %, and so on. The average occupancy of each hydrophobic contact and hydrogen bond was calculated for the structures within each interval. Averages over a group of native contacts from a specific region of the src SH3 domain structure are shown in Figure 6. These two Figures clearly show the hierarchical unfolding of the src SH3 domain. The structural elements in the native protein can be classified into three classes according to the rate at which they are lost: unstable (lost between the 87.5 % and the 67.5 % native contact intervals), intermediate (lost between 67.5 % and 37.5 % intervals), and persistent (lost between 37.5 % and 12.5 % intervals). To aid in the discussion, the structural elements in each class are highlighted in Figure 5 according to the last interval that they appeared in.

Unstable

The first interactions lost are primarily non-local. Pairing between the first and last β-strands (residues 9 to 13 and residues 61 to 65) is clearly disrupted in structures in the 67.5 % native interaction interval. Native side-chain hydrophobic contacts in

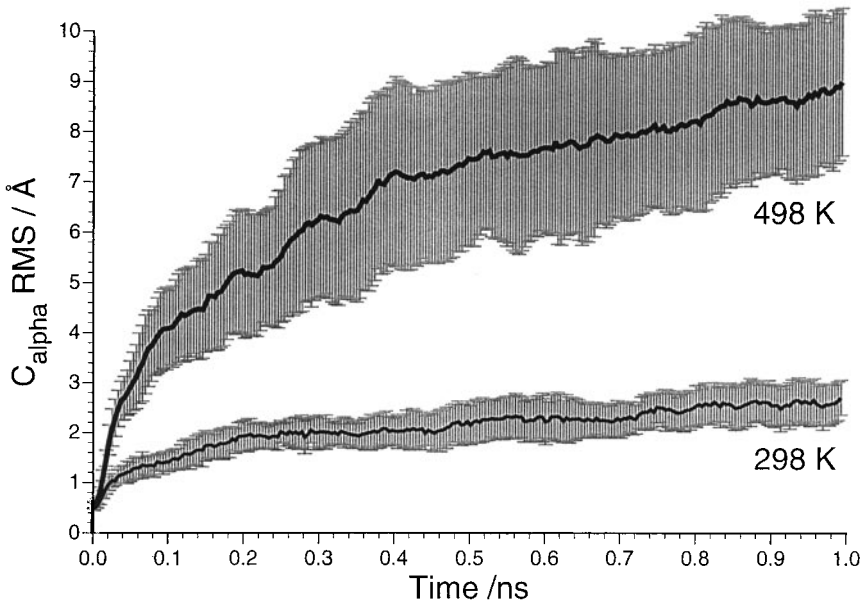

Figure 2. C^{α} rmsd *versus* time. Thick line, 498 K; thin line, 298 K. The error bars indicate the standard deviation of the C^{α} rmsd (this is the standard deviation of each run rather than the standard deviation of the mean which is $1/\sqrt{n}$ smaller, i.e. divide by $\sqrt{10}$ or $\sqrt{30}$ at 298 K or 498 K, respectively).

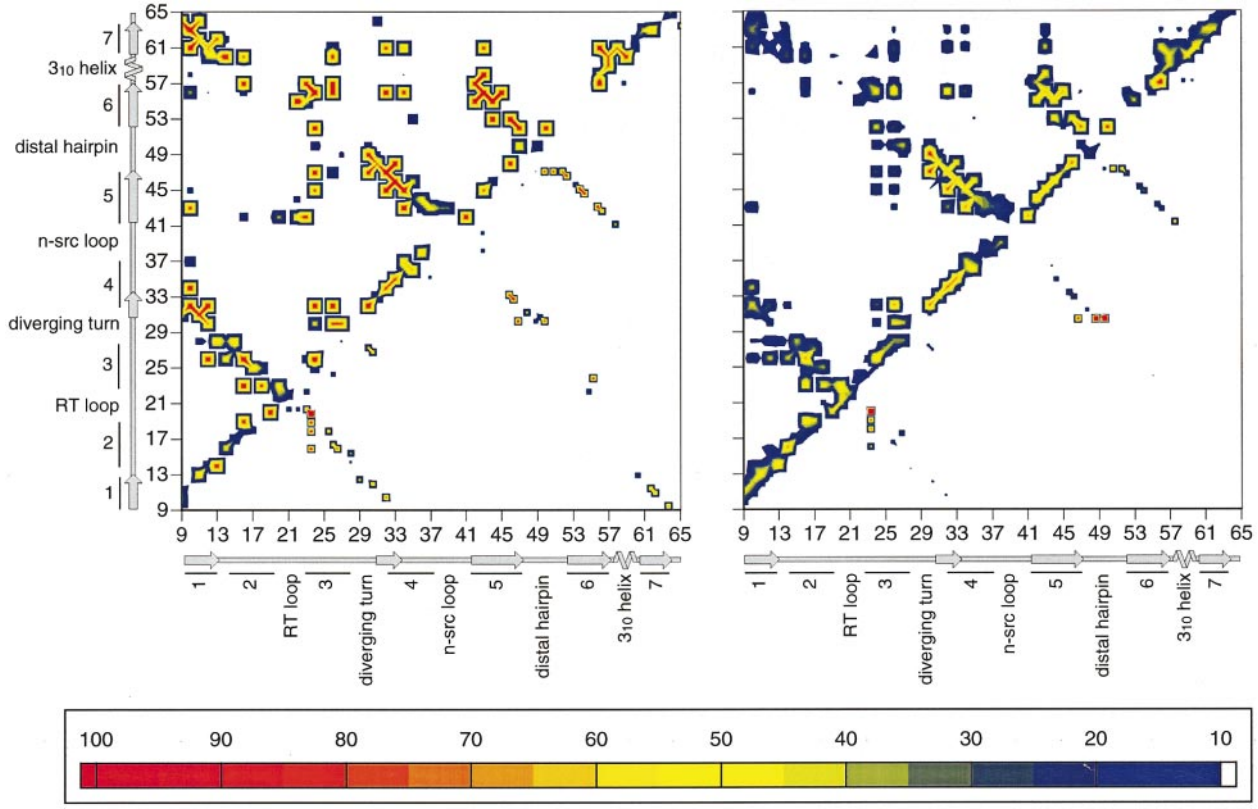

Figure 3. Average interaction occupancies in the simulated structures. The contact map on the left shows the average occupancies from the 298 K set, and on the right, the 498 K set. Above the diagonal in each is shown data for the side-chain hydrophobic contacts, and below the diagonal is shown data for all hydrogen bonds.

this region are decreased by over one-half (upper left-hand corner of interval 67.5 % in Figures 5 and 6(a)) with hydrogen bonding all but disappearing (lower right-hand corner of interval 67.5 % in Figures 5 and 6(b)). At lower per cent native contacts, the average occupancies for these interactions exhibit very low values. The next most unstable set of interactions are those between the src SH3 domain's two orthogonal β-sheet structures, which as indicated in Figure 5, make up the src SH3 domain's hydrophobic core. One sheet consists of the terminal β-strands (between residues 9 to 13 and residues 61 to 65) and the RT loop (residues 14 to 28). The other sheet consists of the n-src loop (residues 32 to 47) and distal β-hairpin (residues 42 to 55).

Intermediate

Interactions in the 3_{10}-helix and the RT loop are lost in going from the 67.5 % native contact interval to the 37.5 % interval. The only helix in the structure, the 3_{10}-helix (residues 57 to 61) is the last piece of secondary structure before the chain returns to the pairing made between the termini of β-strands 1 and 7. Native side-chain hydrophobic contacts in the 3_{10}-helix are lost at a rate close to the average for all contacts (Figure 6(b)), but the single native main-chain hydrogen bond in the 3_{10}-helix (between the proline 57 O and tyrosine 60 H) is lost more rapidly (Figure 6(b)).

In contrast to the 3_{10}-helix, the RT loop (residues 14 to 28) exhibits hydrogen bonds that are more stable than average (Figure 6(b)) and side-chain hydrophobic contacts that are less so (Figure 6(a)). As the population of structures become less native, this region's hydrogen bonds and side-chain hydrophobic contacts occur mostly around interactions made by the side-chain of aspartate 23. The local hydrogen-bond network around aspartate 23 is also responsible for the residual structure seen near the diagonal in the 12.5 % native interaction interval.

Persistent

The last part of the structure to be disrupted is the three-stranded β-sheet made by the n-src loop and distal loop β-hairpin. The n-src loop (residues 32 to 47) is involved in the recognition and binding functions of the src SH3 domain. The second part of the three-stranded sheet, the distal β-hairpin (residues 42 to 55), displays classic β-sheet side-chain hydrophobic contacts and main-chain hydrogen bonds, which are well preserved into populations of non-native structures. Figure 6 shows that both types of interactions behave similarly for the n-src loop and distal β-hairpin.

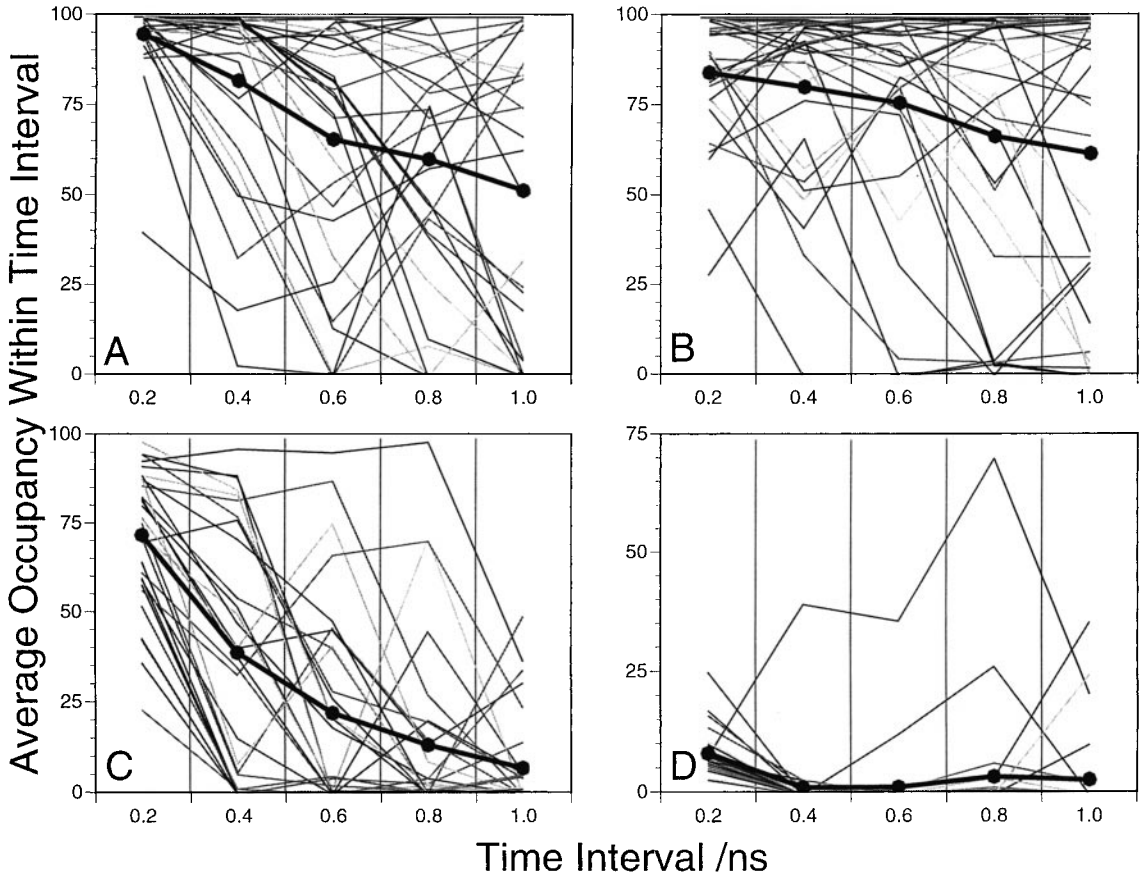

Figure 4. Variability in the 498 K trajectories. Plot of average occupancy within a 0.2 nanosecond time interval for four native side-chain hydrophobic contacts. Light lines are data from each of 30 simulations run at 498 K. Dark lines running through filled circles are averages over the high-temperature simulations. (a) Ile34 with Trp43 (interactions between strand 4 to strand 5). (b) Glu30 with Ser49 (diverging turn to distal hairpin). (c) Ile34 to Ile56 (core contacts between strand 4 and strand 6). (d) Arg31 to Ser64 (diverging turn to C terminus).

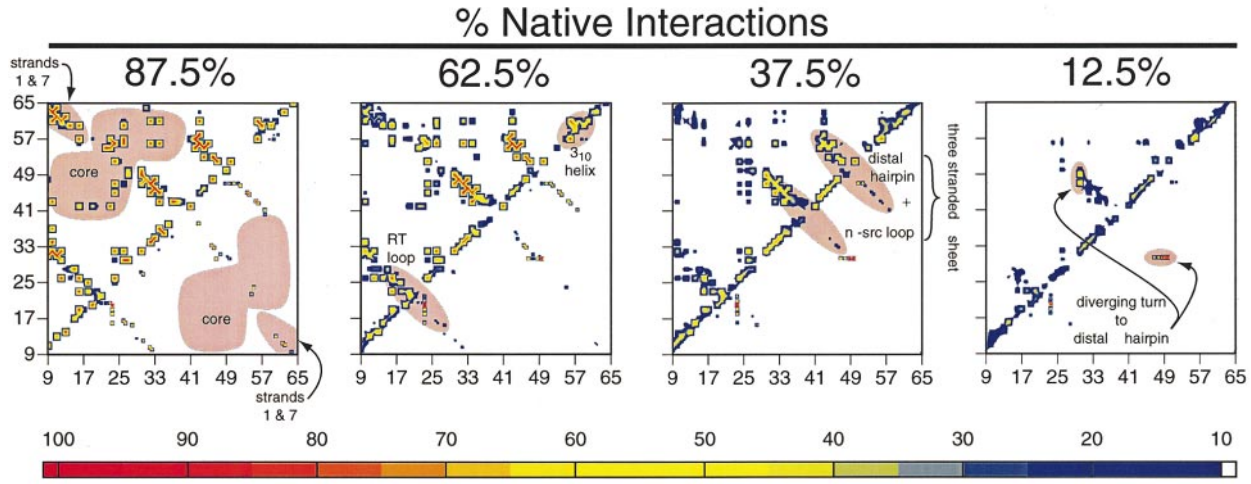

Figure 5. Disappearance plots at 498 K. Individual structures from all 30 trajectories were binned according to the percentage of native contacts, as described in the text. The average occupancy of each contact is indicated by color according to the gradient on the bottom of the Figure. In each plot, side-chain hydrophobic contact results are plotted above the diagonal and results from all hydrogen bonds are plotted below. As an aid in the Discussion, certain regions are highlighted in pink in the interval where they last appear.

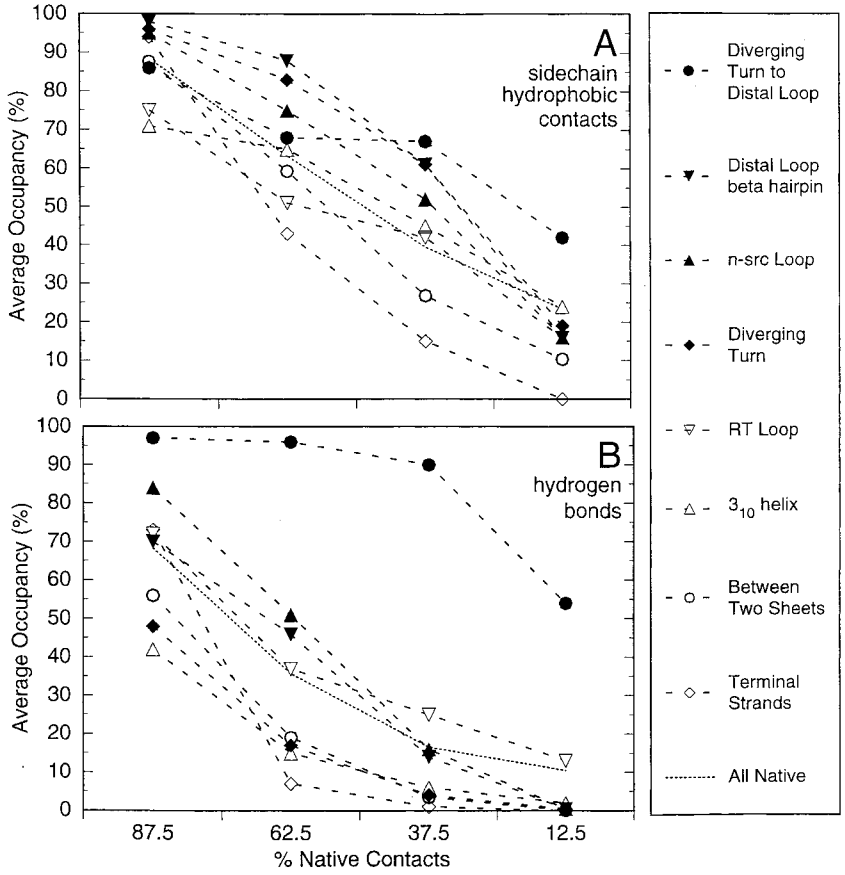

Figure 6. Loss of interactions within and between selected substructures. For each native contact interval, average occupancies were calculated for native interactions within the substructures indicated. (a) Loss of side-chain hydrophobic contacts. (b) Loss of hydrogen bonds.

Because the sequence in the diverging turn has been conserved across SH3 domains (Feng *et al.*, 1995; Guruprasad *et al.*, 1995) and shown to be important by a number of experiments (Grantcharova *et al.*, 1998; Riddle *et al.*, 1997; Yi *et al.*, 1998), we discuss its interactions as a separate substructure. The diverging turn (residues 28 to 32) connects strand 3 to strand 4 in a 90° kink. The diverging turn's persistence is due to particularly stable native side-chain hydrophobic contacts and hydrogen bonds (Figure 6).

Interactions between the diverging turn (residues 28 to 32) and the distal hairpin (residues 47 to 50) are preserved throughout the simulations. Specifically, the side-chain of glutamate 30 in the diverging turn makes many interactions with the distal hairpin. Both Figures 5 and 6 clearly illustrate the strong stability of this region over all others, especially in structures that are less than 25% native (in the 12.5% interval). While not always the most populated, the side-chain hydrophobic contacts between these regions exhibit the most consistent occupancies over all intervals (Figure 6(a)), and the hydrogen bonds display the highest occupancies within every interval (Figure 6(b)). Consistent with this finding, recent double mutant experiments suggests that the hydrogen bond network is largely formed in the transition state ensemble (V.P. Grantcharova & D.B., unpublished results).

Discussion

We find that there is a definite hierarchy to the loss of structure in the src SH3 domain under high-temperature MD unfolding conditions. Despite considerable differences between individual unfolding trajectories, a systematic analysis of 30 independent simulations shows that as the native structure breaks down, some structural elements in the native protein are, on average, lost earlier than others. The most persistent element of secondary structure is the three-stranded β-sheet made up of the n-src loop and distal β-hairpin. This β-sheet is effectively stabilized by two groups of interactions: hydrophobic contacts and hydrogen bonding within the sheet and a hydrogen-bond network between the diverging turn and distal hairpin. All other elements of the src SH3 domain (the RT loop, the 3_{10}-helix, the pairing between the terminal strands 1 and 7, and interactions in the src SH3 domain's core) are lost earlier in the unfolding process.

The intellectual origins of the method of analyis used here lie in the work of Brooks and co-workers, who reconstructed the free energy landscapes of several small proteins (Boczko & Brooks, 1995; Guo *et al.*, 1997; Sheinerman & Brooks, 1998) using the results of many independent low-temperature molecular dynamics simulations, and that of Lazaridis & Karplus (1997) who identified a statistically preferred unfolding

pathway through analyisis of 24 independent high-temperature unfolding simulations of CI-2 using an implicit solvent model. Both approaches analyzed the statistical properties of large numbers of configurations generated in multiple independent simulations, and explored the population of configurations as a function of the number of native contacts and other order parameters. The ''disappearance plots'' have two principal advantages. First, as pointed out by Lazaridis & Karplus (1997), Q is a more robust reaction coordinate than the simulation time, because unfolding takes place at different times in different simulations. Second, they simultaneously display the changes in the average occupancies of each native side-chain hydrophobic contact and each native hydrogen bond in the protein as the native structure breaks down, providing a comprehensive view of the hierarchy of the unfolding process. The approach should be useful in analyses of the unfolding dynamics of other proteins. The averaging procedure provides a means to connect the behaviors of individual molecules with the ensemble averaged properties measured in experiments on protein folding in solution.

Analysis of the effects of mutations on folding and unfolding kinetics suggested that the diverging turn and distal loop beta hairpin interact in the folding transition state in the src SH3 (Grantcharova *et al*., 1998). The distal loop β-hairpin was also found to be formed in the folding transition state of the spectrin SH3 domain (Martinez *et al*., 1998). It appears that the folding transition state structure of this family of proteins is largely determined by the topology of the src SH3 domain. More recent experiments (D.S. Riddle & D.B., unpublished results), in which the consequences of mutations of almost every residue in the src SH3 domain on the kinetics of folding have been characterized, suggest that the three-stranded sheet composed of the n-src loop and the distal β-hairpin is largely formed in the folding transition state ensemble. This is the same region of the protein observed to fall apart last in the MD unfolding simulations described here.

To facilitate comparison to experimental data, we computed quantities from the MD simulations analogous to the experimentally measured ϕ values. Figure 7 compares these to the experimentally determined ϕ values (Grantcharova *et al*., 1998). The comparison of experimental phi values to quantities computed from simulations was pioneered by Daggett *et al*. (1996). The ϕ values capture the extent to which the interactions made by a residue in the native state are formed at the transition state: a value of one indicates that the interaction is nearly completely made at the transition state, whereas a value of zero indicates that it is largely disrupted. The ϕ value-like quantities were computed for each residue by summing the average occupancies of the native side-chain hydrophobic contacts made by a residue in the partially disrupted structures from the 50-60 % native contact interval. The ϕ values from the simulation (Figure 7(a)) roughly match those observed experimentally (Figure 7(b)), with the highest values in the three-stranded beta sheet formed by the n-src loop and the distal loop beta hairpin, and the lowest values in the N and C-terminal strands. Recent experimental results indicate that the phi values for all residues in the N and C-terminal strands are very close to zero (D. S. Riddle & D.B., unpublished results), which is again consistent with the results from the simulations.

While the overall hierarchy of structure loss is similar in the simulations and in experiment (the first and last β-strands come apart first, and the central β-sheet, last), the differences are more dramatic in the experimental results. This is evident in

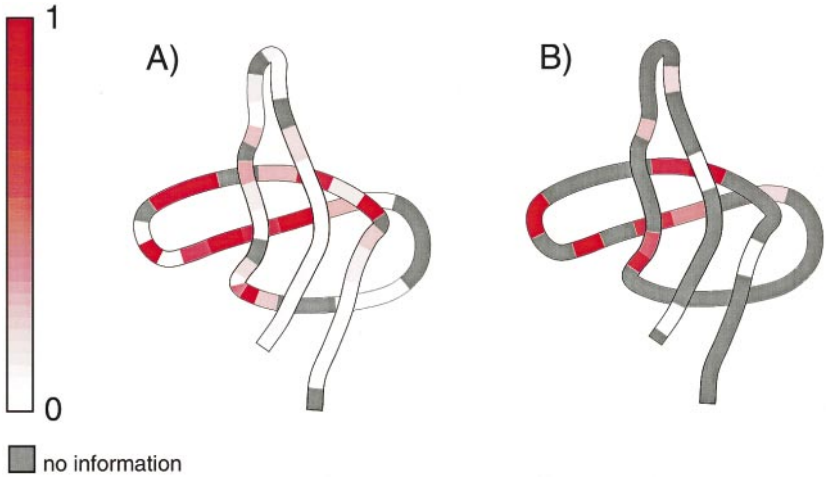

Figure 7. Comparison of ϕ value-like quantities from the MD simulations with ϕ values from experiment. ϕ values are represented from white ($\phi = 0$) to red ($\phi = 1$) as displayed in the legend. (a) The ϕ value-like quantities were computed as described in the text, and the values cubed to accentuate the differences between the various parts of the structure. Residues in gray made no native contacts. (b) The ϕ values from experiment (Grantcharova *et al*., 1998). For those residues in gray, a ϕ value was not measured.

the ϕ values: the experimental ϕ values are close to zero in the first and last strand, and near one in the distal loop beta hairpin (Figure 7(b)), while the values computed from the simulations are on average 0.4 and 0.8, respectively (Figure 7(a)). The reduced structural polarization observed in the simulations could reflect inaccuracies in the potential functions, insufficient sampling, the strongly denaturing conditions under which the simulations were carried out, or simply the fact that the transition state ensemble is not singled out for analysis.

A powerful feature of all atom MD simulations is the detailed view they provide into aspects of the folding reaction, which are only crudely mapped by experiments. In particular, there are a number of mutations in the n-src loop which either speed up both the folding and unfolding rates, or reduce both, suggesting that the residues affected make more interactions in the transition state (TS) ensemble than they do in the native state. To investigate what these residues might do in the TS, conformations with the distal loop β-hairpin largely intact were extracted from the simulations (the distal loop β-hairpin appears to be largely intact in the folding TS in experiment, and thus these conformations are plausible members of the TS ensemble). A superposition of these conformations shows that the n-src loop is displaced below the plane formed by the three strands, and that the diverging turn at the beginning of the n-src loop is more constrained than the tip of the loop (Figure 8). This ensemble of possible transition state conformations should be valuable in interpreting the results of ongoing studies of the effects of mutations in this region on folding kinetics.

How do the changes we observe in the native contacts as unfolding progresses compare to those observed in studies of other proteins? As noted in the abstract, in the symmetric funnel limit (Zwanzig, 1995; Doyle *et al*., 1997), the occupancies of all contacts should decrease in concert with the loss in total number of native contacts. For the src SH3 domain the folding funnel clearly deviates from such symmetry: the first and last β-strands come apart first, and the central β-sheet, last. Deviations from symmetry have also been observed for the other small proteins whose folding free energy landscapes have been characterized using simulation methods. Lazaridis & Karplus (1997) found that the primary unfolding event in their implicit solvent simulations was the disruption of tertiary interactions between the single helix in CI-2 and a two-stranded portion of the beta sheet; similar results were obtained in explicit solvent simulations by Li & Daggett (1998). Onuchic *et al*. (1996) found that the folding transition state ensembles of lattice and full atom models, while quite broad, contained a subset of particularly hot contacts. The free energy landscapes constructed by Brooks and co-workers for several small proteins (Guo *et al*., 1997; Sheinerman & Brooks, 1998), which notably are based on simulations carried out under folding conditions, also suggested a thermodynamic ordering to folding: formation of the third helix in protein A was observed to follow formation of helices I and II, and the N terminus of the helix and the second beta turn of protein G were observed to form before the rest of the structure.

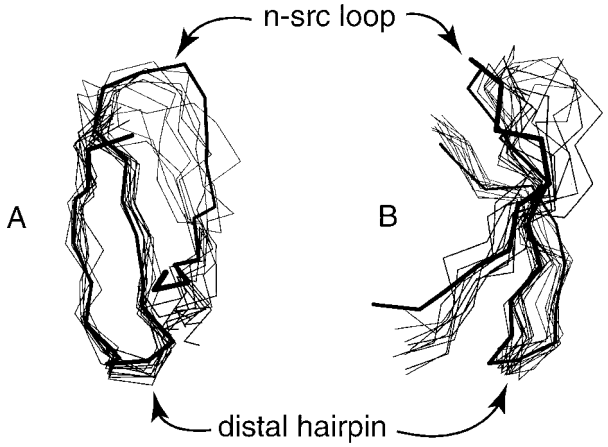

Figure 8. Possible structure of the three-stranded β-sheet in the TS ensemble. Structures that had between 50 and 60% native contacts were clustered into ten groups. The representative closest to all others in the group was chosen. These ten structures were then superimposed based on the residues in the distal loop β-hairpin. Only residues from 29 to 59 are shown. (a) View from the perspective of Figure 1. (b) View across the sheet axis rotated by 90° around the *x* axis.

A central question given the agreement between the experiments and the simulations is the physical origin of the observed hierarchy of structure loss/structure formation. Recent experimental results suggests that folding mechanisms are largely determined by the topology of the native state (Alm & Baker, 1999). Prompted by these observations, our group has developed a simple model for the folding free energy landscape based on native state topology, which takes into account only the entropic cost of chain ordering and the free energy gain associated with hydrophobic burial. For the src SH3 domain, this model predicts that the lowest free energy path to the native state involves the formation of the three-stranded sheet composed of the n-src loop and the distal loop β-hairpin. The simple model completely ignores non-native interactions and uses a simplified treatment of the native interactions, suggesting that the topology of the SH3 domain fold rather than the details of the inter-residue interactions determines the importance of the three-stranded sheet in folding. Far more interactions are made within the three-stranded sheet than within any other SH3 domain segment of of similar length. The molecular mechanics-based potential functions used in MD simulations are known to be far from perfect, and the potential function used in the simple model makes no attempt to accurately represent all the forces involved in folding. The fact that both agree reasonably well with experimental data points to

an underlying robustness of the protein folding process consistent with the observed importance of native state topology in determining folding mechanism.

Methods

The coordinates from the crystal structure (Xu *et al*., 1997) were used as the starting point of the molecular dynamics simulations. To be consistent with previous work, the residue numbering scheme used throughout this is from the NMR study (Yu *et al*., 1993). Simulations were run with ENCAD (Levitt *et al*., 1995) as described (Daggett & Levitt, 1993). Simulation parameters are listed in Table 1. Each simulation ran through 500,000 2 fs time-steps to produce runs 1 ns long. A smooth force shifting truncation of non-bonded interactions at 8.0 Å was used and timesteps were saved every 500 steps (1 ps). Within each set of simulations at a given temperature, the only difference between runs was the random number seed used to initially equlibrated the system.

Side-chain hydrophobic contact analysis was done using the Voronoi procedure as described by Gerstein *et al*. (1995). Two residues were considered to be in contact if they shared a face of a Voronoi polyhedron. This contact was used to calculate occupancy of side-chain hydrophobic contacts. The Voronoi construct allows the proper identification of real contacting neighbors and is well suited to all atom MD simulations. As discussed in previous work (Tsai *et al*., 1997), Voronoi polyhedra are an exact solution to the number of non-bonded contacts. Because the sizes of atoms in proteins is heterogeneous, a distance cutoff usually over or underestimates these interactions.

Hydrogen bonds were defined geometrically: the distance cutoff between a hydrogen and an acceptor atom was 2.6 Å and the angle between the acceptor atom, hydrogen, and the hydrogen's covalently bonded donor atoms had to have been greater than 120°. In order to put data from all hydrogen bonds and side-chain hydrophobic contacts on one graph, hydrogen bonds between O_i and H_j are plotted at (i,j) and hydrogen bonds between H_i and O_j are plotted at $(i + 0.5, j + 0.5)$.

Native interactions were defined as those found in the crystal structure for all hydrogen bonds and side-chain hydrophobic contacts. At each time-step, the percentage of native interactions was calculated and structures were binned accordingly.

To derive representative structures from the ensemble with 50 to 60 % native contacts, the structures were clustered into ten groups based on the C^α rmsd and used in a multi-linkage algorithm. The structure closest to all others in the groups, again based on the C^α rmsd, was chosen as the representative structure.

Figures of structures were oriented with RasMol (Sayle & Milner-White, 1995) and then rendered using MOLSCRIPT (Kraulis, 1991).

Acknowledgements

We thank the members of the Baker Laboratory for helpful comments on the manuscript. D.B. and J.T. thank the NIH (GM51888) and NSF (VAR9214821/9423347) for support. M.L. and J.T. thank the NIH (GM41455) for support.

References

Alm, E. & Baker, D. (1999). Matching theory and experiment in protein folding. *Curr. Opin. Struct. Biol.* **9**, 189-196.

Bond, C. J., Wong, K. B., Clarke, J., Fersht, A. R. & Daggett, V. (1997). Characterization of residual structure in the thermally denatured state of barnase by simulation and experiment: description of the folding pathway. *Proc. Natl Acad. Sci. USA,* **94**, 13409-13413.

Bozcko, E. M. & Brooks, C. L., III (1995). First-principles calculation of the folding free energy of a three-helix bundle protein. *Science,* **269**, 393-396.

Brooks, C. L., III (1998). Simulations of protein folding and unfolding. *Curr. Opin. Struct. Biol.* **8**, 222-226.

Daggett, V. & Levitt, M. (1993). Realistic simulation of native protein dynamics in solution and beyond. *Annu. Rev. Biophys. Biomol. Struct.* **22**, 353-380.

Daggett, V., Li, A., Itzhaki, L. S., Otzen, D. E. & Fersht, A. R. (1996). Structure of the transition state for folding of a protein derived from experiment and simulation. *J. Mol. Biol.* **257**, 430-440.

Doyle, R., Simons, K., Quian, H. & Baker, D. (1997). Local interactions and the optimization of protein folding. *Proteins: Struct. Funct. Genet.* **29**, 282-291.

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.

Fersht, A. (1985). *Enzyme and Structure Mechanism*, 2nd edit., W. H. Freeman and Company, New York.

Gerstein, M., Tsai, J. & Levitt, J. (1995). The volume of atoms on the protein surface: calculated from simulation, using Voronoi polyhedra. *J. Mol. Biol.* **249**, 955-966.

Grantcharova, V. P. & Baker, D. (1997). Folding dynamics of the src SH3 domain. *Biochemistry,* **36**, 15685-15692.

Grantcharova, 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. *Nature Struct. Biol.* **5**, 714-720.

Guo, Z., Brooks, C. L. R. & Boczko, E. M. (1997). Exploring the folding free energy surface of a three-helix bundle protein. *Proc. Natl Acad. Sci. USA,* **94**, 10161-10166.

Guruprasad, L., Dhanaraj, V., Timm, D., Blundell, T. L., Gout, I. & Waterfield, M. D. (1995). The crystal structure of the N-terminal SH3 domain of Grb2. *J. Mol. Biol.* **248**, 856-866.

Kraulis, P. J. (1991). MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallog.* **24**, 946-950.

Ladurner, A. G., Itzhaki, L. S., Daggett, V. & Fersht, A. R. (1998). Synergy between simulation and experiment in describing the energy landscape of protein folding. *Proc. Natl Acad. Sci. USA,* **95**, 8473-8478.

Lazaridis, T. & Karplus, M. (1997). ''New view'' of protein folding reconciled with the old through multiple unfolding simulations. *Science,* **278**, 1928-1931.

Levitt, M., Hirshberg, M., Sharon, R. & Daggett, V. (1995). Potential energy function and parameters for simulation of the molecular dynamics of proteins and nucleic acids in solution. *Comp. Phys. Commun.* **91**, 215-231.

Li, A. & Daggett, V. (1998). Molecular dynamics simulation of the unfolding of barnase: characterization of the major intermediate. *J. Mol. Biol.* **275**, 677-694.

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

Onuchic, J. N., Socci, N. D., Luthey-Schulten, Z. & Wolynes, P. G. (1996). Protein folding funnels: the nature of the transition state ensemble. *Fold Design,* **1**, 441-450.

Riddle, D. S., Santiago, J. V., Bray-Hill, S. T., Doshi, N., Grantcharova, V. P. & Baker, D. (1997). Functional rapidly folding protein from simplified amino acid sequences. *Nature Struct. Biol.* **4**, 805-809.

Sayle, R. & Milner-White, E. J. (1995). RasMol: biomolecular graphics for all. *Trends Biochem. Sci.* **20**, 374.

Shakhnovich, E. I., Farztdinov, G. M., Gutin, A. M. & Karplus, M. (1991). Protein folding bottlenecks: a lattice Monte-Carlo simulation. *Phys. Rev. Letters,* **67**, 1665-1668.

Sheinerman, F. B. & Brooks, C. L. R. (1998). Calculations on folding of segment B1 of streptococcal protein G. *J. Mol. Biol.* **278**, 439-456.

Socci, N. D., Onuchic, J. N. & Wolynes, P. G. (1996). Diffusive dynamics of the reaction coordinate for protein folding funnels. *J. Chem. Phys.* **104**, 5860-5868.

Tsai, J., Gerstein, M. & Levitt, M. (1997). Simulating the minimum core for hydrophobic collapse in globular proteins. *Protein Sci.* **6**, 2606-2616.

van Aalten, D. M., Amadei, A., Bywater, R., Findlay, J. B., Berendsen, H. J., Sander, C. & Stouten, P. F. (1996). A comparison of structural and dynamic properties of different simulation methods applied to SH3. *Biophys. J.* **70**, 684-692.

Voronoi, G. F. (1908). Nouveles applications des paramétres continus á la théorie de formes quadratiques. *J. Reine Agnew. Math.* **134**, 198-287.

Xu, W., Harrison, S. C. & Eck, M. J. (1997). Three-dimensional structure of the tyrosine kinase c-Src. *Nature,* **385**, 595-602.

Yi, Q., Bystroff, C., Rajagopal, P., Klevit, R. E. & Baker, D. (1998). Prediction and structural characterization of an independently folding substructure in the src SH3 domain. *J. Mol. Biol.* **283**, 293-300.

Yu, H., Rosen, M. K. & Schreiber, S. L. (1993). ^{1}H and ^{13}N assignments and secondary structure of the Src SH3 domain. *FEBS Letters,* **324**, 87-92.

Zwanzig, R. (1995). Simple model of protein folding kinetics. *Proc. Natl Acad. Sci. USA,* **92**, 9801-9804.

Edited by B. Honig

(Received 26 January 1999; received in revised form 3 June 1999; accepted 11 June 1999)