Direct Molecular Dynamics Observation of Protein Folding Transition State Ensemble

Feng Ding,* Nikolay V. Dokholyan,† Sergey V. Buldyrev,* H. Eugene Stanley,* and Eugene I. Shakhnovich†
*Center for Polymer Studies, Department of Physics, Boston University, Boston, Massachusetts 02215; and †Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts 02138 USA

ABSTRACT The concept of the protein transition state ensemble (TSE), a collection of the conformations that have 50% probability to convert rapidly to the folded state and 50% chance to rapidly unfold, constitutes the basis of the modern interpretation of protein engineering experiments. It has been conjectured that conformations constituting the TSE in many proteins are the expanded and distorted forms of the native state built around a specific folding nucleus. This view has been supported by a number of on-lattice and off-lattice simulations. Here we report a direct observation and characterization of the TSE by molecular dynamic folding simulations of the C-Src SH3 domain, a small protein that has been extensively studied experimentally. Our analysis reveals a set of key interactions between residues, conserved by evolution, that must be formed to enter the kinetic basin of attraction of the native state.

INTRODUCTION

It has been widely accepted that the physical mechanism underlying folding kinetics of most two-state proteins is nucleation (Abkevich et al., 1994; Fersht, 1997). For the protein to fold into its native state, the overall free energy barrier separating the folded and unfolded states must be overcome. The conformations corresponding to the transition barrier are denoted as transition state ensemble (TSE). Fersht et al. (Matouschek et al., 1989,1990) proposed ϕ-value analysis to quantify the participation of each amino acid in the transition states via the protein engineering experiments. Identification of the transition state ensemble is crucial for the interpretation of experimental results and understanding of protein folding mechanics, which has attracted interests in the past decade (Li and Daggett, 1994,1998; Munoz and Eaton, 1999; Galzitskaya and Finkelstein, 1999; Alm and Baker, 1999; Guerois and Serrano, 2000; Nymeyer et al., 2000; Clementi et al., 2000). Temperature-induced unfolding (Li and Daggett, 1994,1998) in all-atom molecular dynamics simulations with explicit solvent molecules has been applied to study the transition states. However, the limitation of computational ability for traditional molecular dynamics only enables one to sample over several unfolding trajectories from the folded state. Thus, this technique can only capture one or a few transition state conformations instead of a statistically significant ensemble. Moreover, derivation of *folding* transition state ensemble from high-temperature unfolding may be problematic in some cases due to possible significant differences between high-temperature free energy landscape and the free energy landscape on which folding occurs at physiological temperatures (Finkelstein, 1997; Dinner and Karplus, 1999).

Some other theoretical approaches (Munoz and Eaton, 1999; Galzitskaya and Finkelstein, 1999; Alm and Baker, 1999; Guerois and Serrano, 2000) have been proposed to predict the transition states in protein folding and obtained significant correlations with experimental ϕ-values for several proteins. However, each of these models involves drastic assumptions. For example, each amino acid can only adopt two states, native or denatured, and the ability to be in the native state was considered to be independent of other residues. Such an assumption is normal for one-dimensional systems, but may be inappropriate for three-dimensional proteins, because the native state of a residue depends on its contacts with its neighbors. Moreover, the dynamics is only derived from thermodynamics in these works.

The principal difficulty to select TSE conformations is the identification of the reaction coordinate for protein folding. The fraction of native contacts Q (Nymeyer et al., 2000; Clementi et al., 2000) has been proposed as the reaction coordinate to study the TSE. However, the reaction coordinate for folding is not well defined (Abkevich et al., 1994; Du et al., 1998; Klimov and Thirumalai, 2001), so in principle it is difficult to determine the folding TSE from *equilibrium* sampling. The probability for the protein conformation to fold into the native states p_{fold} (Du et al., 1998) is proposed as the robust criterion of TSE. Thus, the TSE can be determined from the *kinetic* simulations as the set of conformations representing the kinetic separatrix between native and unfolded basins of attraction (Du et al., 1998; Klimov and Thirumalai, 2001).

Here we propose an approach to identify TSE from molecular dynamics simulations. Our approach unifies a number of concepts that have been developed in the protein folding community (Gō and Abe, 1981; Sali et al., 1994; Zhou and Karplus, 1999; Dokholyan et al., 2000; Abkevich

Submitted January 28, 2002 and accepted for publication July 31, 2002.

Address reprint requests to Feng Ding, Center for Polymer Studies, Dept. of Physics, Boston University, Boston, MA 02215. Tel.: 617-353-3891; Fax: 617-353-3783; E-mail: fding@polymer.bu.edu.

Dr. Dokholyan's present address is Dept. of Biochemistry and Biophysics, Univ. of North Carolina at Chapel Hill, School of Medicine, Chapel Hill, NC 27599. E-mail: dokh@med.unc.edu.

0006-3495/02/12/3525/08 $2.00

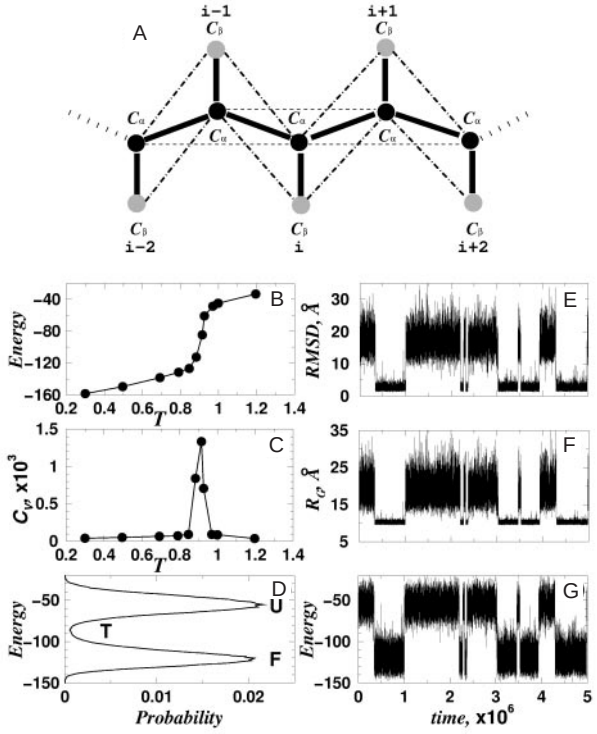

FIGURE 1 (*A*) Schematic diagram of the protein model. Gray spheres represent α carbons, black ones represent β carbons (for Gly, α and β carbons are the same). In the present model only the interactions between side chains are counted, so that the interaction only exists between β carbons, and the α carbon only plays the role of the backbone. (*B*) The average potential energy and (*C*) the specific heat dependence on temperature, where the solid dots indicate the actual temperatures at which simulations were performed. There is a sharp transition at the folding temperature $T_f = 0.91$. (*D*) The probability distribution of the potential energy at T_f. It is bimodal, with a low probability between the peaks corresponding to folded (F) and unfolded (U) states, which corresponds to the putative TSE (T). (*E*) The radius of gyration, (*F*) *RMSD,* and (*G*) potential energy of the protein at folding temperature T_f, respectively. A typical run is shown and the unit of time is 10^6 time units. The folded and unfolded states are stable states in that the minimum time (see the transitions around 2.2 and 3.5) that the model protein is present in the folded or unfolded states is of the order of 10^5 time units, while the time of temporary fluctuations is of the order of 10^2 time units. In folded states, the *RMSD* is around 2 Å. The energy difference between the folded state and unfolded state is ≈70 energy units.

and Shakhnovich, 2000). We test this approach on the folding kinetics of the C-Src SH3 domain (PDB access code 1NLO), within the Gō model approximation for the amino acid interactions (Gō and Abe, 1981; Zhou and Karplus, 1999). We introduce a coarse-grained representation of the C-Src SH3 domain, which includes the C_α and C_β atoms and a set of additional specific constraints that allow us to mimic protein flexibility (see Methods and Fig. 1 *A*).

THERMODYNAMICS AND KINETICS

To test whether the model faithfully reproduces the experimentally observed (Jackson, 1998; Grantcharova et al., 1998) thermodynamic and kinetic properties of the C-Src SH3 domain, we first perform the discrete molecular dynamics simulations of the model C-Src SH3 domain at various temperatures (Fig. 1 *B*). At each temperature we calculate the potential energy E, the radius of gyration R_g, the rms deviation from the native-state *RMSD,* and the specific heat $C_v(T)$. At low temperatures, the average potential energy $\langle E \rangle$ increases slowly with temperature, and the *RMSD* remains below 3 Å. Near the transition temperature $T_f = 0.91$, the quantities E, R_g, and *RMSD* fluctuate between values characterizing two states, folded and unfolded, yielding bimodal distribution of potential energy (Fig. 1). Potential energy fluctuations at T_f give rise to a sharp peak in $C_v(T)$ (see Fig. 1 *C*), which is characteristic of a first-order phase transition for a finite system. Our findings are consistent with experimental observations for the C-Src SH3 domain (Jackson, 1998; Grantcharova et al., 1998).

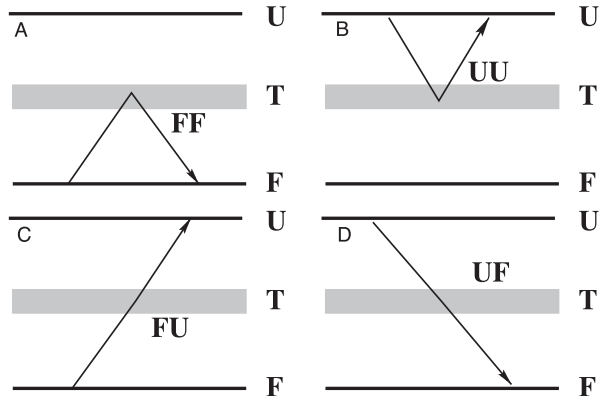

FIGURE 2 Schematic diagram of four types of fluctuations that bring the protein to the putative transition state region. FF (*A*), UU (*B*), FU (*C*), and UF (*D*). The upper line corresponds to the average energy of unfold states (U), E_U, and the lower line corresponds to the average energy of folded states (F), E_F. The shaded region indicates the putative transition state (T) energy range $\{E_{TS}\}$, $-91 < E < -80$. All the fluctuations are selected along the trajectory and are partitioned according their history and their future. Starting from a fluctuation in the putative transition state region $\{E_{TS}\}$ and tracing backward or forward along the trajectory, if the energy reaches E_F/E_U we denote that the fluctuation starts from or ends at the unfolded (U)/folded (F) state. To not mistakenly count a temporary fluctuation as the unfolded state (e.g., see the big fluctuations around 0.5 in Fig. 1 *G*), we set $E_U = -50$, which is slightly higher than the average energy of unfolded states.

Next we determine for the C-Src SH3 domain the folding TSE, a set of conformations with p_{fold} equal to ½. It is computationally impossible to find p_{fold} for every single conformation of a protein. Thus, following Dokholyan et al. (2000), we limit the search for TSE conformations to the energy range $\{E_{TS}\}$, defined to be $-91 < E < -80$, corresponding to the unstable region with the lowest probability in the potential energy histogram at T_f (Fig. 1 *D*). Not all conformations from $\{E_{TS}\}$ belong to the TSE, so we partition these conformations into four kinds of fluctuations that bring the protein to the unstable state within the range of $\{E_{TS}\}$ (see Fig. 2): 1) FF, when the folded protein unfolds

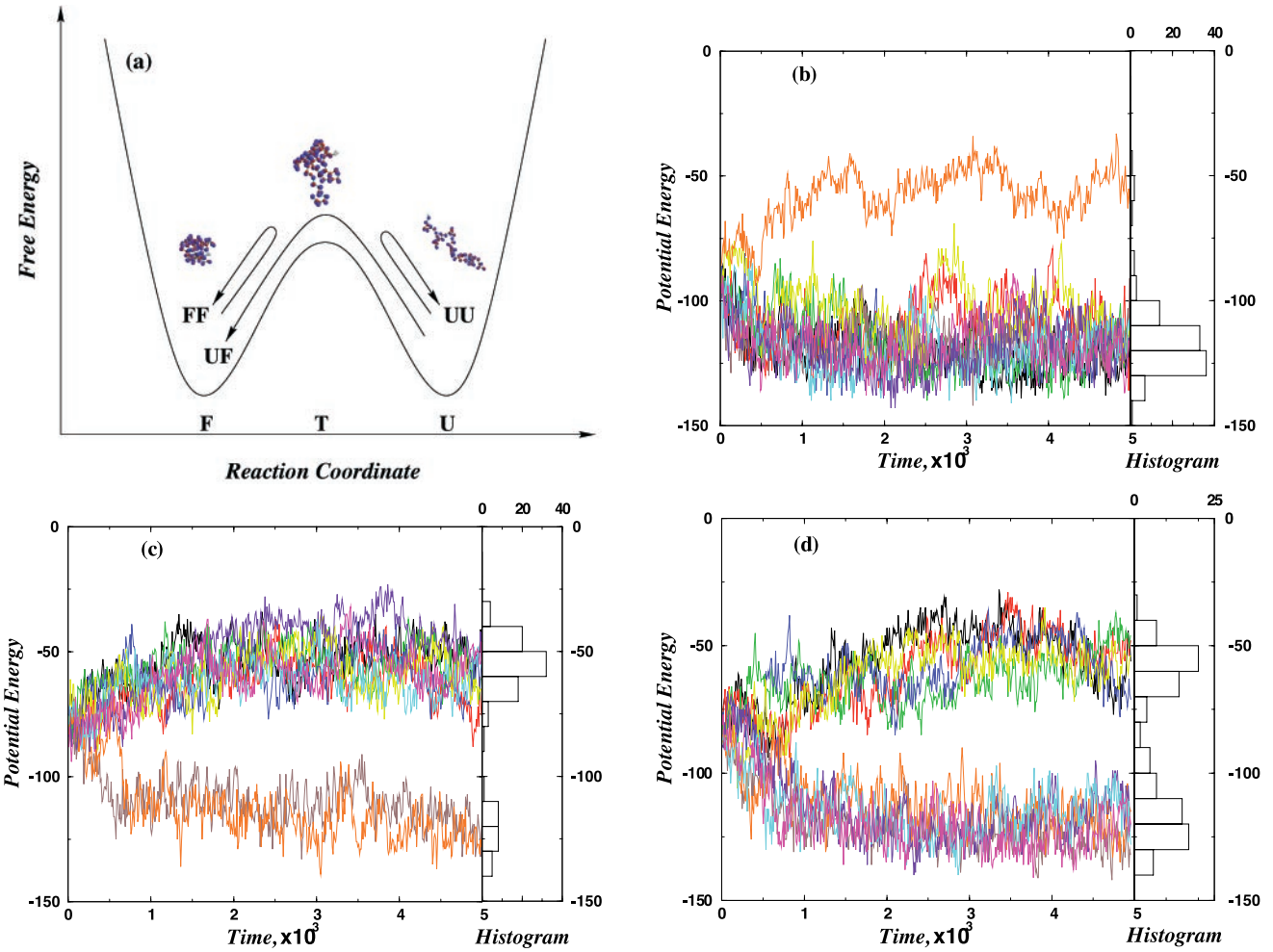

FIGURE 3 (*a*) A schematic representation of TSE conformations. TSE conformations belong to the top of the free energy barrier between folded and unfolded states, and have 50% probability to descend to the folded state and 50% probability to descend to unfolded states. (*b*) The evolution of potential energy for simulations starting from a conformation from the native state basin of attraction (FF conformations). Most simulations fold (see histogram). (*c*) The evolution of potential energy of the protein for simulations starting from a conformation that belongs to an unfolded basin of attraction (UU conformations). Most simulations unfold (see histogram). (*d*) The fluctuations of potential energy starting at time zero from a conformation belonging to the TSE: there is ≈50% probability to fold, and ≈50% to unfold. All three classes of fluctuations shown in (*b*)–(*d*) start from conformations of the same potential energy, and only 10 of the 100 energy trajectories are shown.

to $\{E_{TS}\}$ and then rapidly refolds to its native state; 2) UU, when the unfolded protein partly folds into $\{E_{TS}\}$ and then rapidly unfolds; 3) FU, when the folded protein unfolds to $\{E_{TS}\}$, and then proceeds unfolding further; and 4) UF, when the unfolded protein traverses the energy range $\{E_{TS}\}$ on its way to folded conformations.

We determine p_{fold} from 100 simulation runs for conformations from these four UU, FF, FU, and UF ensembles. For each ensemble, we randomly select 10 conformations to calculate the corresponding p_{fold} values. In each run we reassign the initial velocities of each residue, keeping the temperature unchanged at T_f. Because the initial state is unstable, it rapidly evolves to a stable folded or unfolded state. Indeed, p_{fold} varies greatly between starting conformations, despite the fact that their energies are similar: FF (Fig. 3 b) conformations have $p_{fold} \approx 1$, while UU (Fig. 3 *c*) conformations have low p_{fold}, UF (Fig. 3 *d*) and FU (data not shown) conformations exhibit $p_{fold} \approx ½$, and thus belong to the TSE. UU and FF conformations represent basins of attraction of unfolded and native states, respectively, so the energy and also the fraction of native contacts Q, which is related to energy in the Gō models, are *not* appropriate reaction coordinates for folding. For simplicity, we construct our TSE only of UF conformations from the energy window $\{E_{TS}\}$, i.e., conformations that are collected only along trajectories that traverse this energy range on the way from the unfolded state to the folded state. We analyze 200 independent folding transitions to create the TSE.

Next, we determine the ϕ values for each residue using the "virtual screening" method as described in the Methods

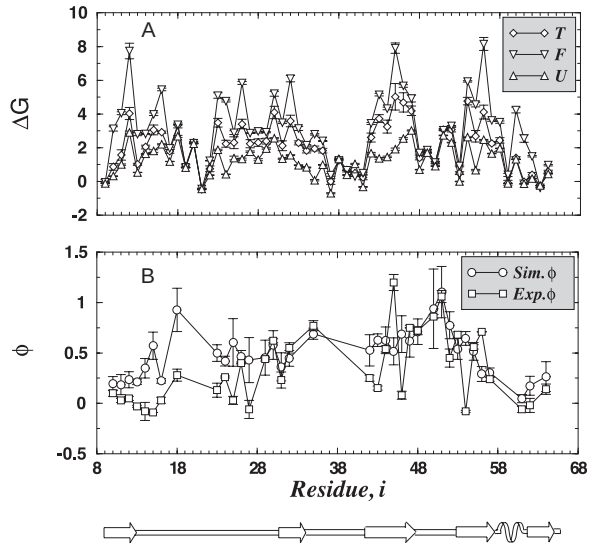

FIGURE 4 (*A*) The values of ΔG for folded (F), transitional (T), and unfolded (U) conformations determined in simulations at T_c. (*B*) ϕ-Values determined from simulations by the virtual screening method (○) and by experiment (□) (Grantcharova et al., 1998; Riddle et al., 1999). Only residues for which experimental ϕ-values are known are shown. The statistical errors of ΔG values are estimated as the standard deviations. The errors of ϕ-values are derived from that of ΔG by the error propagation. Below the *x* axis, the linear structure of C-Src SH3 domain is shown. The arrows denote the β strands, the spaces between the arrows are RT loop, n-src loop, and distal β hairpin, respectively, and the short spiral denotes the 3_{10} helix.

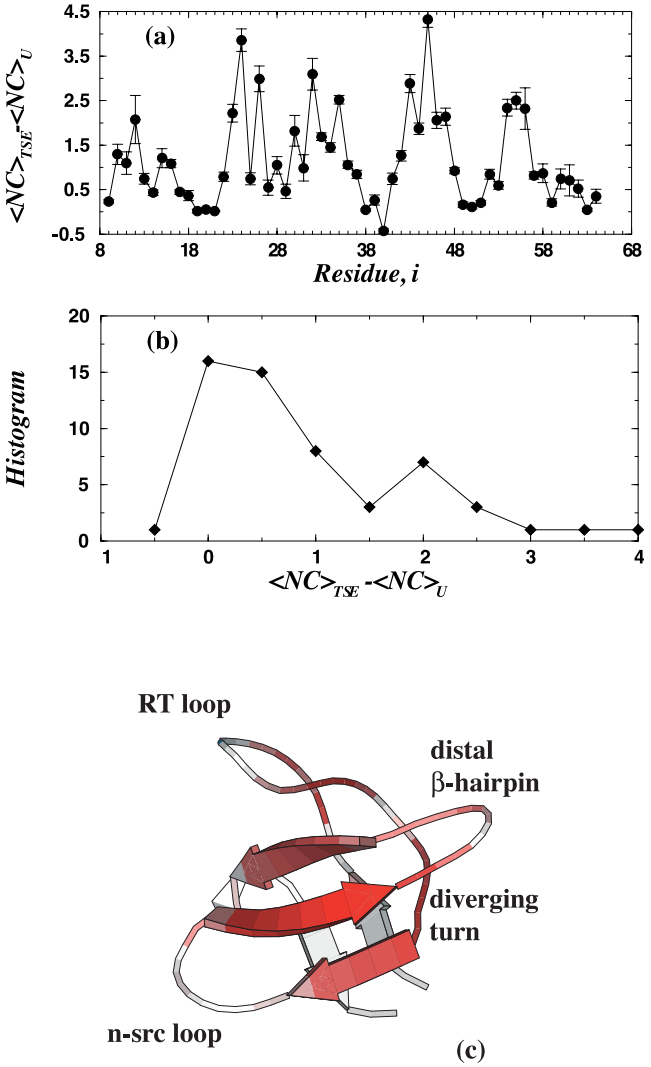

FIGURE 5 (*a*) The number of extra contacts that each residue forms in the TSE, compared to the unfolded ensemble. (*b*) The histogram of the extra contact numbers for each amino acid. There is a second peak at contact number 2. We set the cutoff as 2 for the selection of amino acids that contribute most to the folding TSE. (*c*) Structure of the native state of the C-Src SH3 domain. The color code (from *red* to *white*) represents the relative contribution of individual amino acids to the TSE. The brighter colors of red represent the kinetically most important structures.

section. The correlation coefficient between experimental (Grantcharova et al., 1998; Riddle et al., 1999) and simulated ϕ-values is 0.58 (Fig. 4 *B*). In Fig. 4 *b* there are regions where our determined ϕ-values apparently mismatch the experimental ones, such as residues 10–20, residues around 24, residues 43–46, and residue 54. One of the main reasons is that the mutations cannot probe all the surrounding interactions, especially the backbone interactions, as what we do by the "virtual screening" method. For example, residues 10–20 belong to the N-terminal strand of the RT-loop, which is mostly stabilized by backbone interactions and is persistent (Grantcharova and Baker, 1997) in the partially unfolded states, and thus the mutations in this region produce low ϕ-values. For residues 43–46 we predict all intermediated ϕ-values around 0.6 so that the corresponding β strand adopts the native-like structure in TSE, which is consistent with that fact that residue A45 has the highest experimental ϕ-value. The reason why we cannot capture the fluctuations of experimental values is because our simplified Gō model does not consider the specific nature of different amino acids. In addition, the mutation on the same site to different amino acids may yield different ϕ-values, while the "virtual screening" method does not consider the specificity of mutations. Residues L24 and G54, which we found to be crucial for the folding kinetics, will be discussed later in this paper. For comparison, we have also calculated the ϕ-values by using potential energy as the reaction coordinate, i.e., selecting all the UU, FF, FU, and UF conformations as transition states. The correlation coefficient between thus-calculated and experimental ϕ-values reduces to 0.49 (data not shown). Our results allow us to directly evaluate the relative importance of various interactions in the TSE — an insight difficult to obtain solely from experiments, which report on the structure of TSE only implicitly, via ϕ-values, the interpretation of which is too complex in some cases (Itzhaki et al., 1995; Abkevich et al., 1998).

By comparing the number of contacts N_C that an amino acid makes in the TSE with that number in the unfolded state (Fig. 5 a), we select amino acids that are most impor-

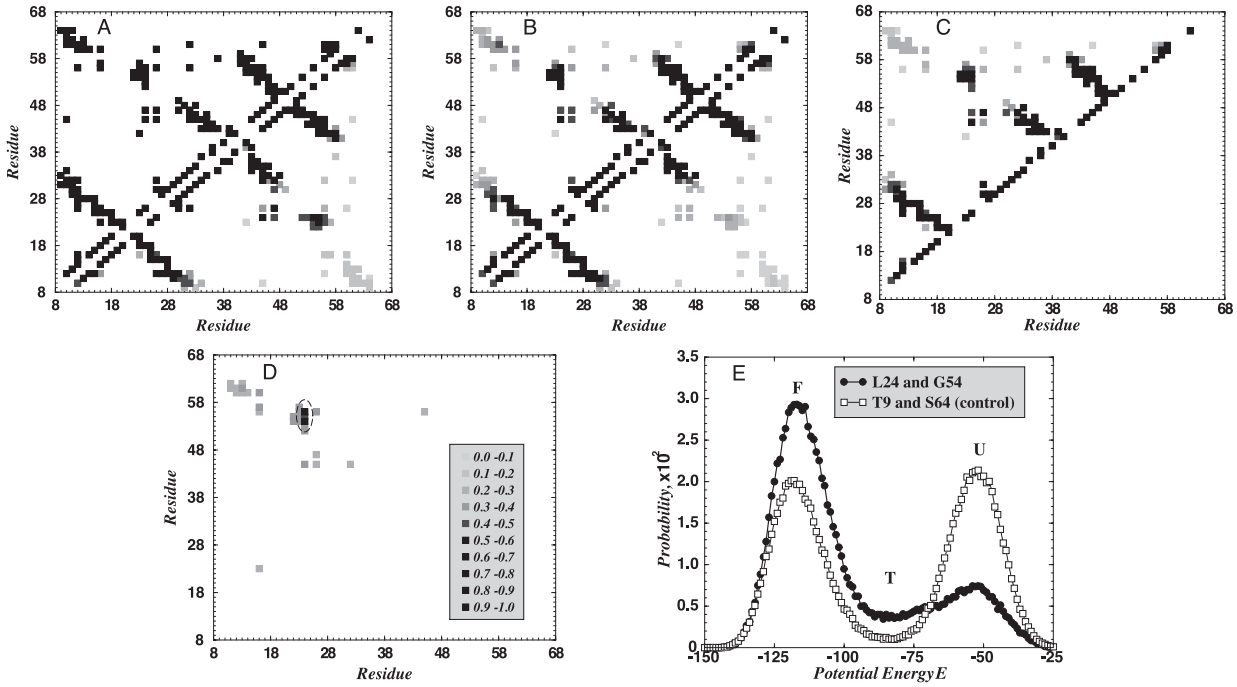

FIGURE 6 (*A*) Above the diagonal is the contact map of the native C-Src SH3 conformation, while below the diagonal is the map of frequencies of contacts between residues obtained from the averaging over 200 conformations of TSE. (*B*) Above the diagonal is the map of frequencies of contacts between residues obtained from the averaging over 200 conformations of FF conformations, while below the diagonal is the map of frequencies of contacts between residues obtained from the averaging over 200 conformations of UU conformations. (*C*) The contact map of putative TSE is calculated by using fraction of native contacts *Q* to select the TSE conformations (Nymeyer et al., 2000; Clementi et al. 2000). (*D*) The difference of the frequency maps for FF and UU conformations shows that the key contacts distinguishing FF and UU basins are between L24 and G54–I56 (*dashed ellipse*). A long range contact L24–G54 occurs with high probability in all conformations that belong to the basin of attraction of the native state. The gray scale represents the frequency scale. (*E*) The probability distribution of potential energy *E* for cross-linked L24 and G54 shows suppressed bimodality. The distribution for NC cross-linked protein (T9 and S64) is as bimodal as for the wild type of Fig. 1 *D*.

tant for the formation of the TSE by setting the cutoff as 2 (Fig. 5 *b*): A12, N23, L24, F26, L32, V35, W43, A45, H47, G54, Y55, and I56. These amino acids have high calculated ϕ-values except for A12. The low calculated and experimentally derived ϕ-value for A12 indicates that it still need to form many more contacts to be native-like by noticing that it forms 13 contacts in the native states (Fig. 6 *A*). In general, the majority of the residues from that list also have high experimental ϕ-values; remarkably, residue A45, which has the highest number of contacts in the TSE with respect to the unfolded states, has the highest experimental ϕ-value (1.2). Notable exceptions are N23, L24, W43, and G54, which have ϕ-values that are either small or negative, as in the case of G54. For residue G54, mutation destabilizes the protein while accelerating folding, strongly suggesting that it indeed participates in the TSE (Itzhaki et al., 1995; Ozkan et al., 2001).

In Fig. 6 we present the contact map of the native state and maps of frequencies of the relative participation of contacts in TSE, FF, and UU conformations and the difference of the frequency maps for FF and UU conformations. For comparison, we also show the contact map of the putative TSE (Fig. 6 *C*) derived from using the fraction of native contacts *Q* as the method to select TSE conformations (Nymeyer et al., 2000; Clementi et al., 2000). We find that the three-strand β sheet (residues 28–56) forms first — it is already present in the majority of UU conformations (Fig. 6 *B*). This is not surprising because the three-strand β sheet is just a combination of distal hairpin (residues 44-56) and the n-src loop (Riddle et al., 1999) (Fig. 5 *c*). It is a substructure of relatively short-range contacts that form rapidly, in accord with general observations (Plaxco et al., 1998) and experimental data on the rate of β hairpin formation (Munoz et al., 1997). However, formation of the three-strand β sheet is necessary, but not sufficient for a conformation to enter the basin of attraction of the native state. Comparison of the FF contact map with the UU contact map in Fig. 6 *d* reveals a crucial structural element that needs to be formed to rapidly fold into the native conformation: specific long-range contact between L24 from the RT loop and/or G54 and/or I56 from the distal hairpin (*dashed oval* in Fig. 6 *D*). Interestingly, using the equilibrium sampling method to select TSE by Q (Nymeyer et al., 2000, Clementi et al., 2000), we do not find any specific role for the contact L24 and G54 (Fig. 6 *C*).

DISCUSSION

Remarkably, L24, and especially G54, are two of the most conserved structural residues in the SH3 fold family (Larson and Davidson, 2000) (in which case, they are residue 18 and residue 48, respectively). Furthermore, Baker and co-workers (Grantcharova et al., 1998) showed that L24 cannot be diversified in phage-selection experiments, along with other kinetically important residues. The fact that ϕ-values of these two particular residues are not close to unity—despite strong evidence of their participation in the folding nucleus—is similar to the case of I76 in chymotrypsin inhibitor 2 (CI2) protein, which also has a low ϕ-value but appears to participate in the folding nucleus (Itzhaki et al., 1995). Such apparent contradiction was explained for CI2 by Fersht and co-workers who showed that the strain in the native structure may account for this anomalous behavior of a residue. This explanation is likely also to be valid for the C-Src SH3 domain given the extremely tight packing of C_α of G54 against C_β of L24 in the native structure of C-Src SH3. The site mutation of G54 destabilizes the native state, but may not destroy the backbone interaction and thus cannot probe the transition states properly. Importantly, all residues that are sequence neighbors of G54 have large ϕ-values, while sequence neighbors of L24 have low ϕ-values, fully consistent with our findings (Figs. 4 *B* and 6).

We further verify the crucial role of contact between L24 from the RT loop and the C-terminal strand of the distal hairpin by "cross-linking" L24 and G54. As shown in Fig. 6 *e*, the cross-linking dramatically changes the cooperativity of the folding transition by essentially eliminating the free energy barrier between folded and unfolded states, and shifting equilibrium toward the manifold of folded states. To determine whether this change can be attributed to nonspecific stabilization due to the entropy reduction of the unfolded state caused by cross-link (Abkevich and Shakhnovich, 2000; Dokholyan et al., 2000), we perform a control simulation with N- and C-termini cross-linked (Grantcharova and Baker, 2001) and rule out this possibility (Fig. 6 *E*). We find that the NC cross-linked protein is indeed more stable (T_f increases) than the wild-type, but the barrier between the native and unfolded states remains intact, in sharp contrast to the L24–G54 cross-linked protein.

Thus we reconstruct a comprehensive picture of the C-Src SH3 folding mechanism derived directly from folding kinetics simulations. The three-stranded β sheet and diverging turn is present in the TSE, in accord with previous analysis. However, while this structural feature is present in the TSE, it is not sufficient for folding. A key long-range contact between L24 and the distal hairpin (residues 54–56) must be formed to enter the basin of attraction of the native state, causing direct and fast descent to the native state. These kinetically relevant amino acid interactions cannot be obtained from the thermodynamic approach (Clementi et al., 2000; Nymeyer et al., 2000) to study the TSE by using some global reaction coordinate (such as Q). We predict that cross-linking these residues (by mutating them to cysteines) (Grantcharova et al., 2000) would dramatically change the free energy landscape, and it would be interesting to test this prediction experimentally. The crucial kinetic roles of these amino acids, especially G54, may contribute to the high conservatism in the SH3 fold family (Larson and Davidson, 2000). This model and discrete molecular dynamics simulations used to analyze it represent a combination of structural and dynamic realism with computational efficiency needed to gain statistically significant insights into structural features of the main milestones along the protein folding pathway.

METHODS

Discrete molecular dynamics simulations

Due to the computational burden of traditional molecular dynamics (Du and Kollman, 1998), simplified simulation methods are needed to study protein folding. The discrete molecular dynamics method (Zhou and Karplus, 1997,1999; Dokholyan et al., 1998,2000) is a compromise between computational simplicity and dynamic realism. This method replaces integration of the dynamic (Newtonian) equations of motion by solving momentum and energy conservation laws at each collision that involves polymer (protein) atoms and "solvent" particles used to thermalize the system. Earlier applications of discrete molecular dynamics simulations to protein folding (Dokholyan et al., 1998,2000; Zhou and Karplus, 1999) used a simple "bead-on-a-string" model, where beads were placed at the positions of C_α atoms and the bonds could rotate freely with respect to each other. The bead-on-a-string off-lattice models usually feature greater flexibility than occurs in real proteins, and for this reason they often exhibit several metastable intermediates that are not observed in experiments. Here we propose a realistic off-lattice model that also includes the C_β atoms.

Protein model

We model the protein by beads representing C_α and C_β (Fig. 1). There are four types of bonds: *(i)* covalent bonds between $C_{\alpha i}$ and $C_{\beta i}$; *(ii)* peptide bonds between $C_{\alpha i}$ and $C_{\alpha(i \pm 1)}$; *(iii)* effective bonds between $C_{\beta i}$ and $C_{\alpha(i \pm 1)}$; and *(iv)* effective bonds between $C_{\alpha i}$ and $C_{\alpha(i \pm 2)}$. To determine the effective bond length, we calculate the average and the standard deviation of distances between carbon pairs of types *(iii)* and *(iv)* for 10^3 representative globular proteins obtained from the PDB. We find that the average distances are 4.7 and 6.2 Å for type *(iii)* and type *(iv)* bonds, respectively. The ratio σ of the standard deviation over the average for bond types (iii) and *(iv)* are, respectively, 0.036 and 0.101. The standard deviation of bond type *(iv)* is larger than that of bond type *(iii)* because it relates to the angle of two consecutive peptide bonds. Thus, the bond lengths of type *(iv)* fluctuate less than that of type *(iii)*. The effective bonds impose additional constraints on the protein backbone, so our model closely mimics the stiffness of the protein backbone, and can give rise to cooperative folding thermodynamics.

In our simulation, the four types of bonds are realized by assigning infinitely high potential well barriers (Dokholyan et al., 1998):

$$V_{ij}^{bond} = \begin{cases} 0, & D_{ij}(1-\sigma) < |r_i - r_j| < D_{ij}(1+\sigma), \\ +\infty, & \text{otherwise} \end{cases} \quad (1)$$

where D_{ij} is the distance between atoms i and j in the native state, $\sigma = 0.0075$ for a bond of type *(i)*, $\sigma = 0.02$ for a bond of type *(ii)*, $\sigma = 0.036$ for a bond of type *(iii)*, and $\sigma = 0.101$ for a bond of type *(iv)*. The covalent and peptide bonds are given a smaller width and the effective bonds are given a wider width to mimic the protein flexibility. We use a modified Gō model similar to one described in Dokholyan et al., (1998), in which interactions are determined by the native structure of proteins. In our model, only C_β atoms that are not next to each other along the chain interact with each other. The cutoff distance between C_β atoms is chosen to be 7.5 Å.

Despite the drawback of the Gō model, associated with the prerequisite knowledge of the native structure, it has important advantages. It is the simplest model that satisfies the principal thermodynamic requirements for a protein-like model: 1) the unique and stable native state; and 2) a cooperative folding transition resembling a first-order phase transition. Furthermore, it has been widely applied in the past to study various aspects of protein folding thermodynamics and kinetics (Zhou and Karplus, 1999; Alm and Baker, 1999; Dokholyan et al., 2000). In addition, experimental works (Grantcharova et al., 1998; Martinez et al., 1998; Riddle et al., 1999) show that the transition state ensemble of many two-state fast-folding proteins is primarily determined by native states topologies.

"Virtual screening" method

We use a technique similar to experimental ϕ-value analysis to predict the TSE via computer simulations. We assume that the mutation does not give rise to significant variation of the three-dimensional structures of folded-state and transition-state ensembles, the same assumption that is made in protein engineering experiments. In our simulations, the free energy shifts due to mutation can be computed separately in the unfolded, transition, and folded state ensembles:

$$\Delta G_x = -kT \ln\langle \exp(-\Delta E/kT)\rangle_x \qquad (2)$$

Here x denotes a state ensemble: folded, F; unfolded, U; and transition, T; ΔE is the change of potential energy due to the mutation, and the average $\langle \cdots \rangle_x$ is taken over all conformations of unfolded; transition; and folded-state ensembles. We compute

$$\phi \equiv \frac{\Delta G_T - \Delta G_U}{\Delta G_F - \Delta G_U} = \frac{\ln\langle \exp(-\Delta E/kT)\rangle_T - \ln\langle \exp(-\Delta E/kT)\rangle_U}{\ln\langle \exp(-\Delta E/kT)\rangle_F - \ln\langle \exp(-\Delta E/kT)\rangle_U} \qquad (3)$$

The same equation has been applied to calculate ϕ values in Clementi et al. (2000). Interestingly, if one adopts a simplified definition of the ϕ-value used in recent work (Vendruscolo et al., 2001) as proportional to the number of contacts a residue makes in the TSE, the correlation coefficient between theoretical and experimental ϕ-values is reduced to 0.27. An approximation to the ϕ-value, the *difference* between the average number of contact residues formed in the TSE and in unfolded states, $\phi \approx (\langle N_i\rangle_T - \langle N_i\rangle_U)/(\langle N_i\rangle_F - \langle N_i\rangle_U)$, provides a better correlation coefficient between predicted and experimentally observed ϕ-values (0.48) than does the approximation of Vendruscolo et al. (2001). The reason why a thermodynamic definition of the ϕ-value yields better agreement with experiments can be inferred from the ΔG plot (Fig. 4 *a*), which shows that $\Delta G_F - \Delta G_U$ for most of the amino acids is not negligible. Indeed, there are several amino acids that make persistent short-range contacts in the unfolded states.

ACKNOWLEDGMENTS

This work was supported by the Petroleum Research Fund of the American Chemical Society #37237-AC4, by National Institutes of Health Grant GM52126 (to E.I.S.) and National Institutes of Health National Research Service Award Fellowship GM20251 (to N.V.D.).

REFERENCES

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

Abkevich, V. I., A. M. Gutin, and E. I. Shakhnovich. 1998. A protein engineering analysis of the transition state for protein folding: simulation in the lattice model. *Folding & Design.* 3:183–194.

Abkevich, V. I., and E. I. Shakhnovich. 2000. What can disulfide bonds tell us about protein energetics, function and folding?: simulations and bioinformatics analysis. *J. Mol. Biol.* 300:975–985.

Alm, E., and D. Baker. 1999. Prediction of protein-folding mechanisms from free-energy landscapes derived from native structures. *Proc. Natl. Acad. Sci. USA.* 96:11305–11310.

Clementi, C., H. Nymeyer, and J. N. Onuchic. 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.* 298:937–953.

Dinner, A. R., and M. Karplus. 1999. Is protein unfolding the reverse of protein folding? A lattice simulation analysis. *J. Mol. Biol.* 292: 403–419.

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

Dokholyan, N. V., S. V. Buldyrev, H. E. Stanley, and E. I. Shakhnovich. 2000. Identifying the protein folding nucleus using molecular dynamics. *J. Mol. Biol.* 296:1183–1188.

Du, Y., and P. Kollman. 1998. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science.* 282:740–743.

Du, R., V. S. Pande, A. Y. Grosberg, T. Tanaka, and E. I. Shakhnovich. 1998. On the transition coordinate for protein folding. *J. Chem. Phys.* 108:334–350.

Fersht, A. R. 1997. Nucleation mechanisms in protein folding. *Curr. Opin. Struct. Biol.* 7:3–9.

Finkelstein, V. 1997. Can protein unfolding simulate protein folding? *Protein Eng.* 10:843– 845.

Galzitskaya, O. V., and V. Finkelstein. 1999. A theoretical search for folding/unfolding nuclei in three-dimensional protein structures. *Proc. Natl. Acad. Sci. USA.* 96:11299–11304.

Gō, N., and H. Abe. 1981. Noninteracting local-structure model of folding and unfolding transition in globular proteins. I. Formulation. *Biopolymers.* 20:991–1011.

Grantcharova, V. P., and D. Baker. 1997. Folding dynamics of the src SH3 domain. *Biochemistry.* 36:15685–15692.

Grantcharova, V. P., and D. Baker. 2001. Circularization changes the folding transition state of the src SH3 domain. *J. Mol. Biol.* 306: 555–563.

Grantcharova, V. P., D. S. Riddle, and D. Baker. 2000. Long-range order in the src SH3 folding transition state. *Proc. Natl. Acad. Sci. USA.* 97:7084–7089.

Grantcharova, V. P., D. S. Riddle, J. V. Santiago, and D. Baker. 1998. Important role of hydrogen bonds in the structurally polarized transition state for folding of the src SH3 domain. *Nat. Struct. Biol.* 5:714–720.

Guerois, R., and L. Serrano. 2000. The SH3-fold family: experimental evidence and prediction of variations in the folding pathways. *J. Mol. Biol.* 304:967–982.

Itzhaki, L. S., D. E. Otzen, and A. R. Fersht. 1995. The structure of the transition-state for folding of chymotrypsin inhibitor-2 analyzed by protein engineering methods: evidence for a nucleation-condensation mechanism for protein-folding. *J. Mol. Biol.* 254:260–288.

Jackson, S. E. 1998. How do small single-domain proteins fold? *Folding & Design.* 3:R81–R91.

Klimov, D. K., and D. Thirumalai. 2001. Multiple protein folding nuclei and the transition state ensemble in two-state proteins. *Proteins.* 43: 465–475.

Larson, S., and A. Davidson. 2000. The identification of conserved interactions within the SH3 domain by alignment of sequences and structures. *Protein Sci.* 9:2170–2180.

Li, A., and V. Daggett. 1994. Characterization of the transition state of protein unfolding by use of molecular dynamics: chymotrypsin inhibitor 2. *Proc. Natl. Acad. Sci. USA.* 91:10430–10434.

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

Martinez, J. C., M. T. Pissabarro, and L. Serrano. 1998. Obligatory steps in protein folding and the conformational diversity of the transition state. *Nat. Struct. Biol.* 5:721–729.

Matouschek, A., J. T. Kellis, L. Serrano, M. Bycroft, and A. R. Fersht. 1990. Transient folding intermediates characterized by protein engineering. *Nature.* 346:440–445.

Matouschek, A., J. T. Kellis, L. Serrano, and A. R. Fersht. 1989. Mapping the transition state and pathway of protein by protein engineering. *Nature.* 340:122–126.

Munoz, V., and W. A. Eaton. 1999. A simple model for calculating the kinetics of protein folding from three-dimensional structures. *Proc. Natl. Acad. Sci. USA.* 96:11311–11316.

Munoz, V., P. Thompson, J. Hofrichter, and W. A. Eaton. 1997. Folding dynamics and mechanism of beta-hairpin formation. *Nature.* 390: 196–199.

Nymeyer, H., N. D. Socci, and J. N. Onuchic. 2000. Landscape approaches for determining the ensemble of folding transition states: success and failure hinge on the degree of frustration. *Proc. Natl. Acad. Sci. USA.* 97:634–639.

Ozkan, S. B., I. Bahar, and K. A. Dill. 2001. Transition states and the meaning of Phi-values in protein folding kinetics. *Nat. Struct. Biol.* 8:765–769.

Plaxco, K. W., K. T. Simons, and D. Baker. 1998. Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.* 277:985–994.

Riddle, D. S., V. P. Grantcharova, J. V. Santiago, E. Alm, I. Ruczinski, and D. Baker. 1999. Experiment and theory highlight role of native state topology in SH3 folding. *Nat. Struct. Biol.* 6:987–990.

Sali, A., E. I. Shakhnovich, and M. Karplus. 1994. How does a protein fold? *Nature.* 369:248– 251.

Vendruscolo, M., E. Paci, C. Dobson, and M. Karplus. 2001. Three key residues form a critical contact network in a protein folding transitional state. *Nature.* 409:641–645.

Zhou, Y., and M. Karplus. 1997. Folding thermodynamics of a three-helix-bundle protein. *Proc. Natl. Acad. Sci. USA.* 94:14429–14432.

Zhou, Y., and M. Karplus. 1999. Interpreting the folding kinetics of helical proteins. *Nature.* 401:400–403.