A Lattice Protein with an Amyloidogenic Latent State: Stability and Folding Kinetics

ARTICLE *in* THE JOURNAL OF PHYSICAL CHEMISTRY B · APRIL 2007

Impact Factor: 3.3 · DOI: 10.1021/jp067027a · Source: PubMed

4 AUTHORS, INCLUDING:

Andrey Palyanov
A.P. Ershov Institute of Informatics Systems
16 PUBLICATIONS **68** CITATIONS
SEE PROFILE

Sergei F Chekmarev
Russian Academy of Sciences
55 PUBLICATIONS **271** CITATIONS
SEE PROFILE

A Lattice Protein with an Amyloidogenic Latent State: Stability and Folding Kinetics

Andrey Yu. Palyanov,[†] Sergei V. Krivov,[‡] Martin Karplus,*[,‡,§] and Sergei F. Chekmarev*[,†]

Institute of Thermophysics, 630090 Novosibirsk, Russia, Laboratoire de Chimie Biophysique, ISIS Université Louis Pasteur, 67000 Strasbourg, France, and Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts 02138

Received: October 26, 2006; In Final Form: December 3, 2006

We have designed a model lattice protein that has two stable folded states, the lower free energy native state and a latent state of somewhat higher energy. The two states have a sizable part of their structures in common (two "α-helices") and differ in the content of "α-helices" and "β-strands" in the rest of their structures; i.e. for the native state, this part is α-helical, and for the latent state it is composed of β-strands. Thus, the lattice protein free energy surface mimics that of amyloidogenic proteins that form well organized fibrils under appropriate conditions. A Gō-like potential was used and the folding process was simulated with a Monte Carlo method. To gain insight into the equilibrium free energy surface and the folding kinetics, we have combined standard approaches (reduced free energy surfaces, contact maps, time-dependent populations of the characteristic states, and folding time distributions) with a new approach. The latter is based on a principal coordinate analysis of the entire set of contacts, which makes possible the introduction of unbiased reaction coordinates and the construction of a kinetic network for the folding process. The system is found to have four characteristic basins, namely a semicompact globule, an on-pathway intermediate (the bifurcation basin), and the native and latent states. The bifurcation basin is shallow and consists of the structure common to the native and latent states, with the rest disorganized. On the basis of the simulation results, a simple kinetic model describing the transitions between the characteristic states was developed, and the rate constants for the essential transitions were estimated. During the folding process the system dwells in the bifurcation basin for a relatively short time before it proceeds to the native or latent state. We suggest that such a bifurcation may occur generally for proteins in which native and latent states have a sizable part of their structures in common. Moreover, there is the possibility of introducing changes in the system (e.g., mutations), which guide the system toward the native or misfolded state.

I. Introduction

Scenarios of protein folding vary widely from a simple two-state process to considerably more complex behavior, including the presence of on- and off-pathway intermediates as well as parallel pathways.[1] A specific class of off-pathway intermediates, which may compete in stability and/or kinetic accessibility with the native state, are so-called "latent" states. They occur in a variety of proteins and in some cases are part of the normal folding process and contribute to the function of the protein. Examples are serpins, of which the plasminogen activator inhibitor 1 (PAI-1) folds into an active structure with a subsequent slow conversion into a more stable low-activity state.[2,3] Other systems are α-lytic protease, which folds into an inactive, partially folded state in the absence of the pro-region,[4] and the unphosphorylated signaling protein NtrC, which exists in active and inactive conformations.[5,6] More generally, there are many proteins with more than one conformation, such as GroEL and motor proteins like myosin, which have several stable states but essentially the same fold with somewhat different conformations around hinge regions. For proteins in which the two states have different folds, a common feature is that the latent state is an alternative target structure for folding. It has recently been suggested that essentially all proteins have such latent states, which are less stable in the isolated proteins but become more stable on aggregation.[7] In such systems, which form relatively well characterized amyloid fibrils,[8] it appears that part of the protein has its native fold and part (that incorporated into the aggregate) changes from its native structure to a β-structure with parallel or antiparallel β-sheets. As is well-known, such aggregates lead to a wide range of diseases, including Alzheimer's disease and spongiform encephalopathies. If such latent states exist, then they can give rise to "kinetic partitioning",[9−11] so the folding trajectories are separated in two classes: fast trajectories, which go to the native state directly, and slow trajectories, which visit the latent state. In this respect, the latent state acts as a long-lived off-pathway intermediate, like those observed in some model lattice protein simulations.[12−14] That a sizable part of the structure is kept nearly unchanged during the interconversion between the native and the latent states is likely to be an important element in determining the free energy surface and the nature of the folding process. It is of interest, therefore, to perform folding simulations of a protein model of this type. For this purpose, we make use of a lattice model because it is possible to calculate a large enough set of folding trajectories to obtain statistically meaningful results.[13−15] To mimic the structures thought to be implicated in amyloid fibril formation, we designed a 70 monomer lattice protein in which

* Authors to whom correspondence should be addressed. Phone: (617) 495-4018 (M.K.); 7(383)3391048 (S.F.C.). Fax: (617) 496-32047 (M.K.); 7(383)3308480 (S.F.C.). E-mail: marci@tammy.harvard.edu (M.K.); chekmarev@itp.nsc.ru (S.F.C.).

[†] Institute of Thermophysics.

[‡] ISIS Universit Louis Pasteur.

[§] Harvard University.

10.1021/jp067027a CCC: $37.00
Published on Web 02/22/2007

the native and latent states have a common part consisting of two α-helices; the latent state is different from the native state in that the latter has an additional α-helix, which are β-strands in the former. Thus, the transition between the two states require major structural changes in the form of unwinding/rewinding of the α-helix. Previous studies used native and latent structures that were geometrically completely different states,[10,16] in contrast to our model with the native and latent states having in common a sizable part of the structure. The protein was modeled by a chain of beads (monomers), representing the protein residues, which were restricted to move on the vertices of a cubic lattice. To specify interactions between the monomers, a Gō-like model[17] was employed. The application of this model implies that we focus on the folding process rather than the ability of the proteins to aggregate. A number of model systems (e.g., with hydrophobic−polar interactions[18] or a renormalized Miyazawa−Jernigan potential[19]) have been used to study aggregation explicitly (refs 20−23 and 24 and 25, respectively), which is outside the scope of the present paper.

The paper is organized as follows. Section II describes the lattice model of the protein with a latent state and the methodology of the Monte Carlo simulations. Section III presents the results of equilibrium simulations (free energy surfaces, contact maps, and free energy profiles). Section IV describes the kinetics of folding (the probabilities to reach the native and latent states, the process of achieving equilibrium in the system, and first-passage time distributions). In this section we also present a kinetic model for the folding process and the rate constants for the essential channels of transitions that are determined from a comparison of the theoretical and simulation results. Section V summarizes the results and presents some concluding remarks.

II. Specification of the System and Methodology

The model lattice protein consists of 70 monomers (Figure 1) and mimics the conformational difference between the native state of the protein and the aggregation-prone state, as discussed in the Introduction. The native state has the lower layer composed of two "α-helices"(the red monomers) and the upper layer, which has one "α-helix" and two short "β-strands" (the blue monomers). The latent structure has the same "α-helices" in the lower layer, but the upper layer has four "β-strands", which form a mixed "β-sheet". In the "designed" native state $N_{tot} = 77$, $N_{nat} = 77$, $N_{lat} = 42$, and $N_{com} = 42$, and in the "designed" latent state $N_{tot} = 78$, $N_{nat} = 42$, $N_{lat} = 75$, and $N_{com} = 42$, where N_{tot}, N_{nat}, N_{lat}, and N_{com} are the numbers of total, native, latent, and common contacts between the native and latent states, respectively. The common native and latent contacts mostly involve the two lower "α-helices".

To specify the interactions between the monomers, an extended Gō-type model[17] was employed. The contact energy matrix includes both the native and the latent contacts so as to introduce minima for both states. The monomers that are in contact only in the designed native structure have an interaction energy of $u_{nat} = -1.025$, those only in the designed latent structure have an energy of $u_{lat} = -0.975$, and the monomers common to both structures have an energy of $u_{com} = -1.0$. The difference in the contact energies between u_{nat} and u_{lat} was introduced to provide a bias to the native state. All other pairs of the monomers interact with energy $u_{other} = 0$; i.e., if such monomers are in contact, then they do not contribute to the energy of the system. The total number of contacts (N_{tot}) may thus be larger than the total number of effective contacts, i.e., the contacts with nonzero energy (N_{eff}). The latter is written as $N_{eff} = N_{nat} + N_{lat} - N_{com}$, and the energy of a structure is $U = (N_{nat} - N_{com})u_{nat} + (N_{lat} - N_{com})u_{lat} + N_{com}u_{com}$. The energies of the designed native and latent structures are $U_{nat} = -77.875$ and $U_{lat} = -74.175$.

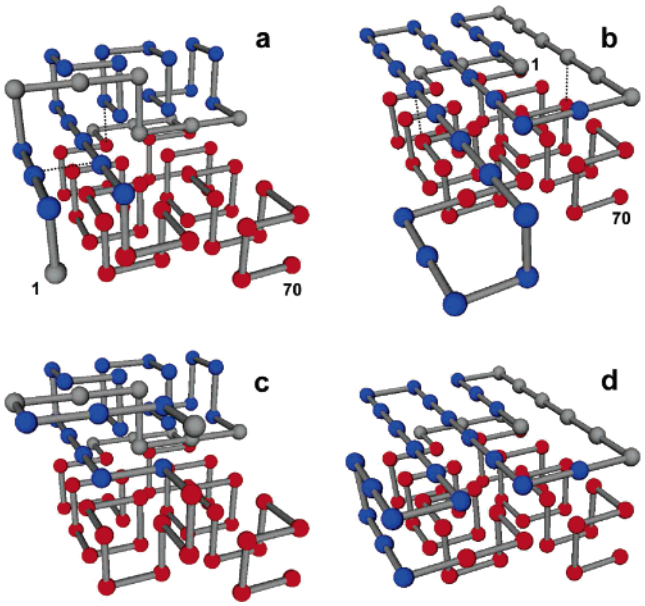

Figure 1. (a and c) Native and (b and d) latent structures. The red monomers represent two "α-helices" common to the native and latent states, and the blue monomers represent the "α-helix" and two short "β-strands" specific to the native state. The numbers 1 and 70 on the structures in the upper row label the two end monomers of the chain. The upper row shows the designed structures for the (a) native and (b) latent states; for the native structure $N_{tot} = 77$, $N_{nat} = 77$, $N_{lat} = 42$, $N_{com} = 42$, and $U = -77.875$, and for the latent structure $N_{tot} = 78$, $N_{nat} = 42$, $N_{lat} = 75$, $N_{com} = 42$, and $U = -74.175$. The lower row shows examples of (c) nativelike and (d) latentlike structures (see text); for them $N_{tot} = 79$, $N_{nat} = 75$, $N_{lat} = 43$, $N_{com} = 41$, and $U = -77.8$ and $N_{tot} = 81$, $N_{nat} = 45$, $N_{lat} = 75$, $N_{com} = 42$, and $U = -77.25$, respectively. The dashed lines connecting the monomers (parts a and b) show the contacts that were neutralized to see how mutations can affect the probability of folding into the latent state (see the text, section V).

To simulate a folding trajectory for the protein on the lattice, the Metropolis Monte Carlo (MC) method was employed. As in previous work[14] three types of moves were allowed; they are end flips, corner flips, and two-bead crankshaft rotations. The step (time) counter is advanced whether or not the current move is accepted; this means that the number of MC steps coincides with the number of MC trials. Three different types of simulations were performed, depending on the specific goal: (i) an equilibrium sampling of the conformation space of the system involving many folding/unfolding transitions, (ii) time-dependent populations of the characteristic states of the system, with the trajectories started in one of the states and continued until equilibrium in the system is achieved, and (iii) the first-passage (folding) time distributions. More detailed information is given in the corresponding sections.

The temperature, T, measured in the units of the energy (i.e., with the Boltzmann constant $k_B = 1$), ranged from 0.65 to 0.695. This temperature interval was sufficient to allow us to observe two limiting situations, when the latent state is not in equilibrium with the unfolded state ($T = 0.65$) and in equilibrium with it ($T = 0.695$). Within this interval the mean folding time exhibited

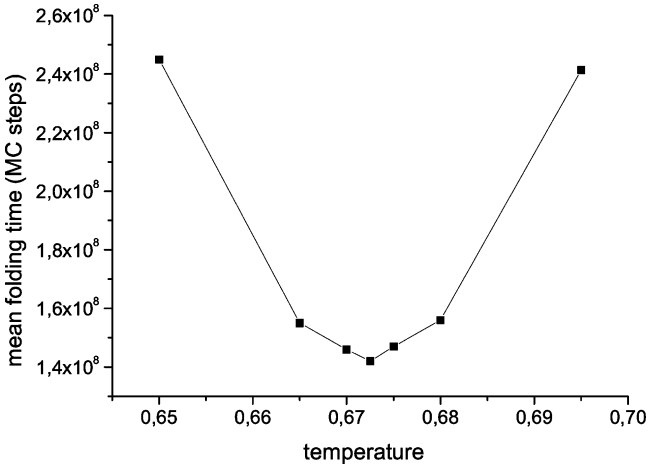

Figure 2. Mean folding time (MC steps) vs temperature.

typical U-shape behavior,[26,27] with the minimum time at $T_f = 0.6725$ (Figure 2).

III. View of Folding from Equilibrium Simulations

A. Free Energy Surfaces. To perform the equilibrium sampling of the conformation space, the trajectory was started at an arbitrary point and continued for 10^{11} MC steps, allowing the system to visit repeatedly all characteristic regions of the space, i.e., the unfolded, native, and latent basins. At each step, the total number of contacts (N), and the numbers of native (N_{nat}) and latent (N_{lat}) contacts were calculated. Collecting these data, we determined the probability for the system to be in a state with specific numbers of contacts $p(N_{tot}, N_{nat}, N_{lat})$. To allow a two-dimensional representation, this function was reduced to $P(N_{nat}, N_{lat}) = \sum_{N_{tot}} p(N_{tot}, N_{nat}, N_{lat})$, which discriminates between the native and the latent states explicitly. Having this function, we calculated the free energy of the system as

$$F(N_{nat}, N_{lat}) = -T \ln P(N_{nat}, N_{lat}) + C \qquad (1)$$

where the arbitrary constant C was determined by the condition $F(77, 42) = U_{nat}$; i.e., the entropy is equal to zero because the designed native state represents a unique conformation for the given numbers of native and latent contacts. The designed latent state does not possess this property: A manifold of conformations is possible for $N_{nat} = 42$ and $N_{lat} = 75$, which differ from each other by the shape of the loop connecting the two lower "α-helices" with the "β-sheet" (Figure 1).

Figures 3a−c show the equilibrium free energy surfaces (FESs) thus obtained at the three characteristic temperatures: $T = T_f = 0.6725$ (part b), at which the mean folding time to the native state is minimal (Figure 2), and two other temperatures, one below and one above this temperature, i.e., (a) $T = 0.65$ and (c) $T = 0.695$, respectively. Three basins of attraction, expected from earlier work, are evident in the figure. These are associated with a semicompact globule state and the native and latent states; these basins are indicated in Figure 3b by labels a, f, and e, respectively, at the bottoms of the basins. Also, interestingly, there is an additional basin of attraction (b), which connects the basin for the semicompact globule with those for the native and latent states. Folding trajectories bifurcate in this basin and proceed to either the native or the latent state. The presence of such a "bifurcation" basin has not been observed in previous studies of "kinetic partitioning",[9−11] perhaps because the native and latent states did not have a sizable part of their

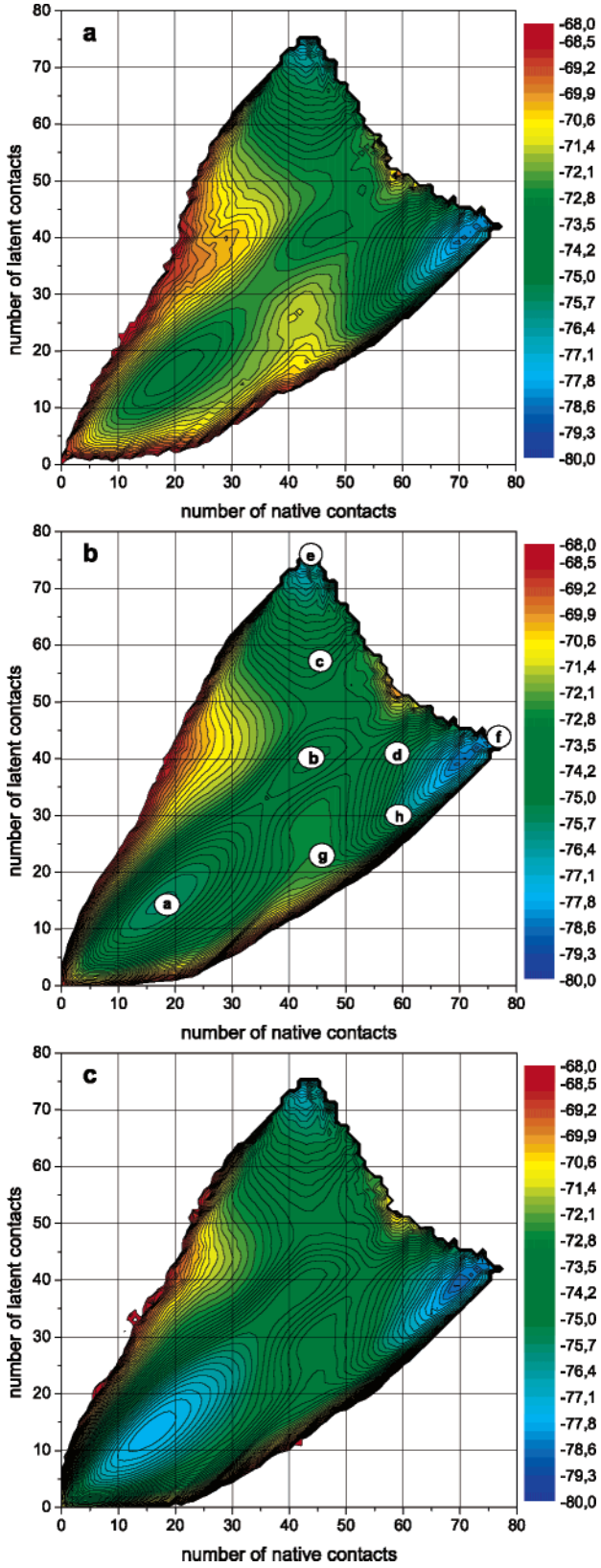

Figure 3. Equilibrium free energy surfaces: (a) $T = 0.65$, (b) $T = 0.6725$, and (c) $T = 0.695$.

structures in common. As the temperature increases, from 0.65 to 0.695, the barriers between the bifurcation basin and the basins of the native and latent states increase, while the barrier to the semicompact globule basin decreases, practically disappearing at the highest temperature (Figure 3c). These regions represent characteristic states of the system and, correspondingly,

the essential stages of folding of the system from the unfolded state, which corresponds to small values of N_{nat} and N_{lat} (Figure 3). At $T = 0.6725$ (Figure 3b), the free energy at the bottom of the bifurcation basin is equal to -74.06 (44, 40), at the transition states to the native and latent basins to -73.43 (52, 40) and -73.68 (48, 51), respectively, and the minimum free energies in the native and latent basins are equal to -78.7 (75, 43) and -77.65 (42, 75); the numbers in the brackets indicate the coordinates of the points of the (N_{nat}, N_{lat}) plane, for which the free energy was calculated. The corresponding values of the mean potential energy at these points are $U = -46.57$, $U = -54.32$, $U = -59.30$, $U = -77.28$, and $U = -74.175$; the difference between the free energy and the potential energy of the latent basin is due to the entropic contribution from the flexible loop connecting the two lower α-helices with the upper β-sheet (Figure 1b). In contrast to the free energies, the potential energy decreases monotonically from the bifurcation basin toward the native and latent states; i.e. because the energy in a Gō-type model decreases monotonically with the number of contacts, the barriers have an entropic origin.

The native and latent state basins are rather broad, occupying a considerable part of the conformation space, in contrast to what is found for lattice proteins with the contact energies selected from Gaussian distributions.[14,15] Each basin, native and latent, contains a manifold of conformations, nativelike and latentlike, respectively. Being distributed over the (N_{nat}, N_{lat}) plane, these conformations lead to slightly rugged but relatively flat free energy surfaces at the bottoms of the basins. Two examples of such conformations are shown in Figures 1d and 1d, for the native and latent basins, respectively.

The conformation in Figure 1c is related to the point (75, 43) of the (N_{nat}, N_{lat}) plane, where the free energy in the native basin is minimal (-78.7 at $T = 0.6725$). This conformation differs from the designed structure (a) in that it loses one native and one common contact and acquires two latent contacts; as a result, the energy increases by $1.025 + 1.0 + 2 \times (-0.975) = 0.075$ and becomes equal to -77.8. The decrease in the free energy, -78.7 to -77.875 for the structure in Figure 1a, is due to the fact that in contrast with the point (77,42), where the designed native structure is a unique conformation, a manifold of conformations is observed at the point (75,43). First, along with the previously mentioned conformation, there exists another, which is obtained from it by the crankshaft rotation of monomers 4 and 5 and has the same energy. Further, there belongs a manifold of conformations with energy $U = -76.8$ to this point. In comparison with the conformations with $U = -77.8$, they lose one native and one latent contact and acquire one common contact, which increases their energy by $1.025 + 0.975 - 1.0 = 1.0$. Among themselves, these conformations differ by the shape of the head of the chain (monomers 1−6), which affects neither the number of the native and latent contacts nor the energy. The presence of many different conformations, even with a higher energy, makes the free energy at the point (75, 43) lower (approximately by 0.825) than that for the designed native structure, which is unique at the point (77, 42).

Figure 1d shows the structure that has the lowest (potential) energy in the latent basin. It differs from the corresponding designed structure (b) in that the loop adheres to the body of the polymer, which adds three native contacts and decreases the energy by $3 \times (-1.025) = -3.075$ to -77.25.

B. Contact Maps. To establish the relation between the free energy surface and the conformations of the system, we built the contact probability maps at the points of the (N_{nat}, N_{lat}) plane that represent characteristic regions of the free energy surface. These maps are shown in Figure 4, with panels labeled according to Figure 3b. The maps were obtained in the course of equilibrium sampling at $T = 0.6725$ by counting the number of times the corresponding contacts were present. The sequences of points that are parallel and close to the diagonal of the contact plane correspond to "α-helices", and those parallel but shifted from the diagonal correspond to the pair of parallel "β-strands". The sequences of the points perpendicular to the diagonal represent the pairs of antiparallel "β-strands". Scattered collections of points represent contacts between the secondary structures and the loops that connect them (Figure 1). The probability that a secondary structure ("α-helix" or "β-structure") is formed is equal to the product of the probabilities of forming all contacts relevant to this structure.

Figures 4f and 4e show the contact maps for the native and latent structures, respectively. In accord with Figure 1, the native structure consists of the sequence of secondary structures β-(1−4)-α(11−26)-β(27−30)-α(32−47)-α(51−70), where β stands for a "β-strand", α for a "α-helix", and the figures in brackets indicate the numbers of monomers constituting these structures, counted as shown in Figure 1. Helices α(11−26), α(32−47), and α(51−70) are labeled as α1, α2, and α3, respectively, and the antiparallel β-sheet, formed by strands β(1−4) and β(27−30), as β. Scattered sequences of points correspond to contacts between these secondary structures. The latent structure (Figure 4e) is different from the native one (Figure 1) in that the part of the native structure β(1−4)-α(11−26)-β(27−30) transforms into the sequence of the strands β(1−4)-β(5−10)-β(12−17)-β(18−25), which are connected to the helices α2 and α3, common to both structures, by a loop consisting of monomers 26−to 29. These "β-strands" form two antiparallel "β-hairpins" and one two-stranded parallel "β-sheet"; they are labeled as β1 and β3, and β2, respectively.

Other panels of Figure 4 represent different stages of formation of the native and latent structures. Figure 4a corresponds to the semicompact globule basin, labeled as a in Figure 3b. It is seen that in this basin all contacts, which are necessary to form both the native and the latent structures, are present but with a relatively low probability (between ∼0.1 and ∼0.4). Moreover, these contacts are not formed at the same time; otherwise N_{nat} and N_{lat} would be larger. In the bifurcation basin b, adjoining basin a, the formation of the common part of the native and latent structures (i.e., helices α2 and α3) is practically completed. Figure 5 shows two typical conformations observed in the bifurcation basin. All other contacts observed in basin a are also present in this basin but still with a low probability, so the possibility of formation of both the native and the latent structures (Figures 4f and 4e) is preserved. The progress to these structures is illustrated by Figures 4d and 4c, which show the contact probability maps for points d and c in Figure 3b that are chosen to represent the transition states leading to the native and latent states, respectively. To see the tendency more explicitly, the points are slightly shifted toward these states. As one can see, at these points the probability of formation of the corresponding contacts (i.e., the native and latent contacts, respectively) increases, while that of the competitive contacts (i.e., the latent and native contacts, respectively) decreases.

Figure 4 also shows that first the contacts close to the diagonal form, which represent secondary structures (i.e., the "α-helices" and "β-strands"), and then the contacts between these structures.

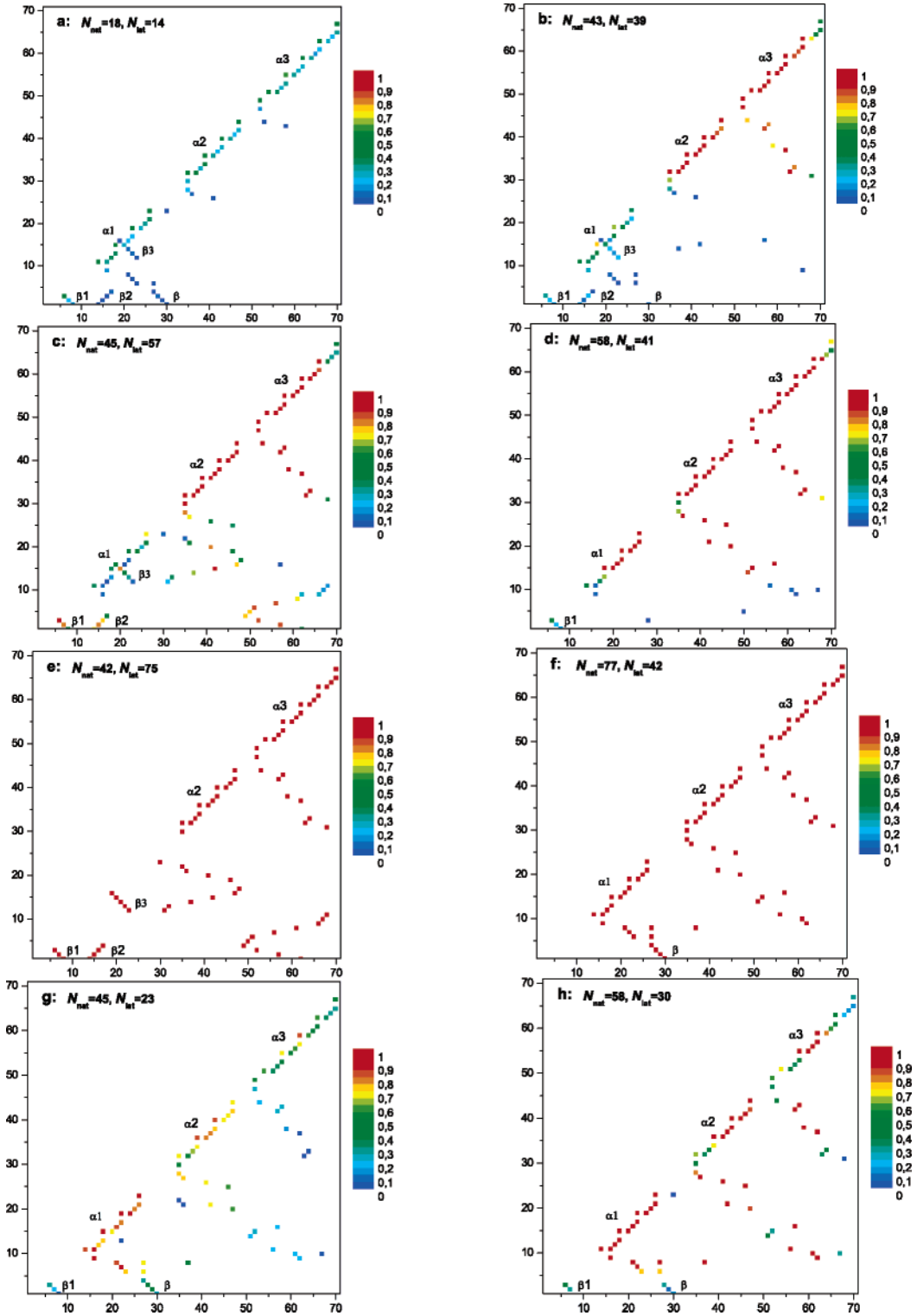

Figure 4. Contact probability maps, $T = 0.6725$: Parts a−h are for the points of the (N_{nat}, N_{lat}) plane that are indicated in Figure 3 by the corresponding labels. Parts e and f show the maps for the latent and native structures, respectively. Figures on the abscissa and ordinate are the monomer numbers; see Figure 1.

Mapping of the folding trajectories onto the free energy surface shows that along with the typical scenario of folding, when the system passed through the bifurcation basin b, there exists an atypical one, in which the trajectories go to the native state directly. Points g and h in Figure 3b and the corresponding contact probability maps of Figure 4 present characteristic stages of such direct pathways. Point g corresponds approximately to the saddle region between basins a and f (which is seen more clearly in Figure 3a), and point h illustrates the progress to the native state. Comparing Figures 4g and 4h with Figure 4b, one can see that the atypical folding scenario is characterized by inverse order of formation of secondary structures; i.e., at first the part specific for the native state (α1) is formed and then the part common for the native and latent states ("helices" α2 and α3). However, such trajectories are very rare (∼1% of the folding trajectories starting from an unfolded chain) and thus do not make a significant contribution to the folding process.

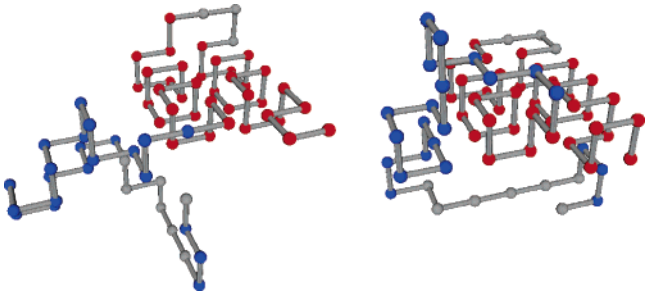

Figure 5. Representative conformations of the protein in the bifurcation basin.

C. Free Energy Profiles. As shown previously,[28] the projection on simple reaction coordinates (e.g., such as the numbers of native and latent contacts that we have chosen) can hide the complexity of the free energy surface. A way to analyze the free energy surface in an unbiased way is to follow the approach presented in ref 28. It is based on root-mean-square deviation (rmsd) clustering of a long equilibrium trajectory, which allows one to determine the equilibrium kinetic network (EKN) and then, by analyzing it, the free energy profile and simplified network.[28,29] However, since configurational space is high dimensional, to have a meaningful connectivity in the EKN (i.e., to have a large enough probability associated with each cluster), the rmsd threshold for the clustering has to be large; in the present case one would have to use 6. Such a large rmsd threshold may not allow local changes in the structure to be seen. To overcome this problem (which, in fact, is general for the rmsd clustering of large systems), we initially reduced the dimensionality of the space using a principal component analysis (PCA), with the entire set of possible local contacts (bond-PCA) as a basis. This allowed us to project the original space onto a reasonable number of principal components. After that we performed clustering in the reduced space. By varying the dimensionality of the reduced space, (i.e., the number of the bond-PCA components included), we were able to control how well the analysis reproduced the essential features of the original free energy surface. One advantage of such bond-based PCA coordinates, in comparison with conventional approaches that select reaction coordinates a priori (e.g., the radius of gyration, number of native contacts), is that the coordinates are introduced in an unbiased way. Also, one can introduce additional bond-PCA coordinates if it turns out to be necessary for the analysis.

The set of local contacts forms a vector of contacts with the dimension $C = n(n-1)/2$, where n is the number of monomers in the protein chain. For simplicity we consider all contacts, including those along the chain, and do not exclude contacts that are impossible to form by geometry. At each point along the trajectory, the components of the vector were taken to be equal to 0 (if a contact is absent) or to 1 (if the contact is present). In applications of the method to continuous off-lattice models the 0/1 indicator is replaced by a function of the contact distance.[30] Following the standard PCA procedure,[31] the covariance matrix $C \times C$ was averaged over the (equilibrium) trajectory of 5×10^9 MC steps; the contacts were calculated every 10^3 step. The resulting matrix was diagonalized to obtain the principal components. The first principal components of the matrix provide the best projection of the multidimensional configurational space onto a reduced low-dimensional space. The bond-PCA coordinates can be used to build both a free energy profile (FEP) and an EKN, similar to what was done for a β-hairpin and 38-atom Lennard-Jones cluster;[29] in those cases rmsd clustering and clustering by local minimization were used and gave meaningful results. To construct the EKN, hypercube clustering was employed. For this, the trajectory was projected on the bond-PCA coordinates (x_i), where $i = 1, ..., N_{PCA}$ and N_{PCA} is the number of the PCA components included. All points of the trajectory for which x_i is between $k_i\,dx$ and $k_i\,dx + dx$ ($i = 1, ..., N_{PCA}$) are associated with the cluster $k_1, k_2, ..., k_{NPCA}$, where dx is the measure of the clustering; in the present case, $dx = 0.1$ was used.

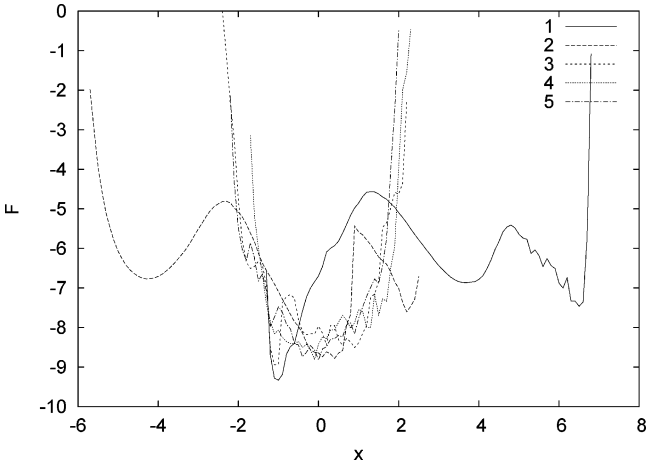

Figure 6. Free energy profiles [$F = -kT \ln(n)$] along the first five PCA components; $T = 0.6725$. x measures the distance along the PCA eigenvector in the bond space.

Figure 6 shows the one-dimensional FEPs along each of the five most essential bond-PCA components. As one can see, the first three components show several minima separated by barriers, while the fourth and fifth components have a single minimum with steep increases in free energy on either side. This indicates that the nature of the FEP is determined primarily by the three lowest PCA components; the other two (and higher components), which show no structure, mainly uniformly decrease the overall connectivity of the network and thereby increase the heights of the barriers.

Figure 7 shows the FEPs obtained by first projecting the free energy surface on the increasing number of PCA components for $N_{PCA} = 1, 2, 3, 4$ and then projecting it on the one-dimensional reaction coordinate. As the reaction coordinate we took the relative partition function of the "reactant" region (Figure 7a).[29] The reactant region is defined as a region around the native structure, separated from the rest of the configurational space by a putative cut.[29] The reaction coordinate is among the simplest and most flexible coordinates that increase monotonically as the system goes from the native to the denatured state. In the neighborhood of the most visited cluster of the native state the coordinate is close to 0, and it increases as the cut moves toward the denatured state; it becomes equal to 1 when the whole free energy surface is included. If there are several well-defined pathways, then this reaction coordinate will adapt its shape to them and will progress mainly along the pathways. The idea behind this choice of the reaction coordinate, as described in ref 29, is that the reaction coordinate is able to include any and all pathways from the initial to the final state without any prejudgment as to the geometric coordinates involved.

Figure 7 shows that the first PCA component allows one to identify three basins: native, denatured, and latent; they correspond to $0 < Z_A/Z < 0.853$, $0.853 < Z_A/Z < 0.92$, and $0.92 < Z < 1$, respectively. Inclusion of the second PCA component also shows the existence of the bifurcation basin, making a total four basins; the bifurcation basin is at $0.853 <$

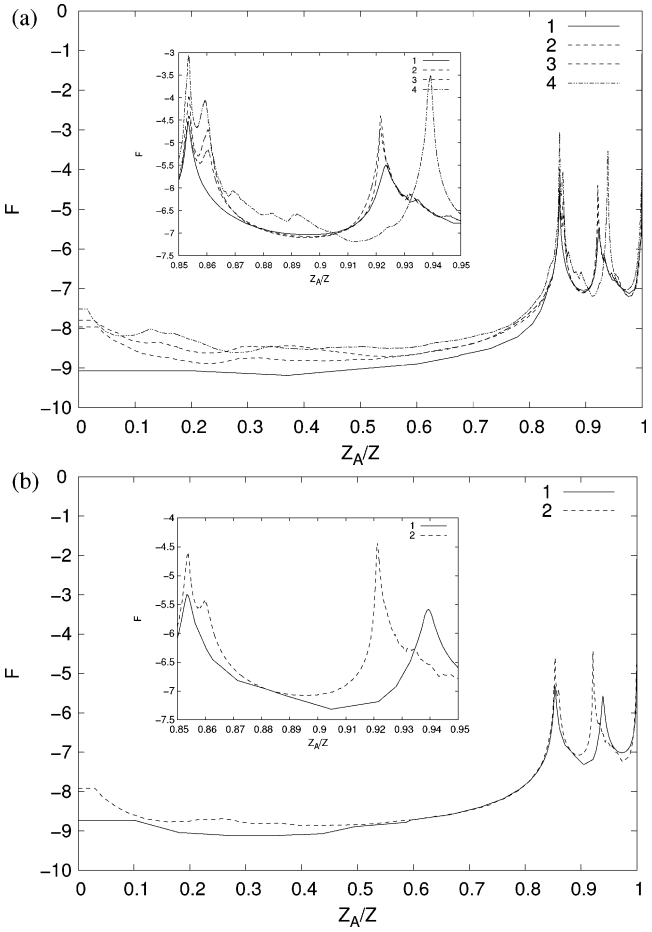

Figure 7. (a) Free energy profiles for the equilibrium kinetic networks obtained by clustering over different numbers of PCA components, from one to four. (b) Free energy profiles for the equilibrium kinetic networks obtained by clustering over the number of (1) native and (2) native and latent contacts. As the reaction coordinate the relative partition function was used (see text); $T = 0.6725$.

$Z_A/Z < 0.86$. The third and fourth PCA components mainly increase the height of the barriers. The positions of the barriers stay almost fixed with respect to the number of components employed since they are essentially defined by the relative partition functions of the basins, and consideration of more PCA components changes mainly the transition state regions, which contribute little to the partition function of a basin. The second barrier shifts to the left for $N_{PCA} = 4$ on Figure 7a due to movement of small basins from the right of the barrier to the left, while each basin has its partition function unchanged, as can be seen from the figure.

As was anticipated, the fourth component mostly contributes to the heights of the barriers and not to the overall pattern of the FEP, which illustrates the robustness of the approach. Moreover, as one can see, the heights of the barriers change approximately equally. This means that reaction rates between basins are changed by the same multiplication factor, which can be associated with an overall diffusion coefficient for the motion of the chain. Note that it cannot be said that the model kinetics on the obtained FEP would correspond to the original system kinetics in contrast to the profiles obtained with the approach suggested in ref 32. Note, however, that in ref 32 the profile was specifically designed to match the system kinetics (using the FEP along the p_{fold} reaction coordinate as a criterion and assuming a constant diffusion coefficient), while here we calculate the profile by direct means and comparison with system kinetics is not a tautology in this case. The work on

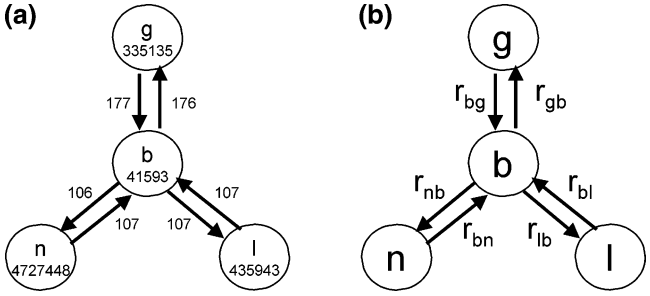

Figure 8. (a) Equilibrium kinetic networks; $T = 0.6725$. Four PCA components are taken into account, and $dx = 0.1$. The numbers in the circles show the numbers of times that the system was found in the basins, and those beside the arrows show the corresponding numbers of transitions between the basins. (b) Kinetic scheme of the folding process. The basin labels g, b, n, and l correspond to the globule, bifurcation basin, native, and latent states, respectively.

representing the obtained profile and diffusion coefficient in a way to make it possible to reproduce the system kinetics is in progress.

If the numbers of native and latent contacts are used for clustering, then the FEP (Figure 7b) is very similar to that obtained with the first two bond-PCA components (Figure 7a). Moreover, Figure 7b also reveals a very shallow, but noticeable, bifurcation basin. The relatively accurate representation of the FES by the conventional set of native and latent contacts is likely to be due to the Gō potential employed in the model.

Figure 8a presents a simplified EKN obtained via the mincut procedure.[28] In accord with the free energy surfaces (Figure 3) and the FEP, the network has four free energy basins, with a bifurcation basin in the middle of the network. The EKN also supports Figure 3 in that the bifurcation basin is very shallow, which is seen both from its partition function and from the FEPs (Figure 7). We thus infer that in the process of folding from the denatured state the system dwells for a short time in the intermediate (bifurcation) basin, where the trajectory bifurcates and then goes either to the native or the latent basin.

Figure 9 shows the projection of the four basins individually (Figures 9a−d) and together (Figure 9e) on the (N_{nat}, N_{lat}) plane. The combined picture (Figure 9e), in which the basins are superimposed, is very similar to Figure 3b. However, the heights of the barriers that separate the bifurcation basin from the others are reduced, which is a result of a partial overlap of the basins; the barriers are increased with inclusion of more PCA components, as already mentioned (Figure 7).

IV. Folding Kinetics

A. Probabilities to Reach the Native and Latent States. With two alternative folded structures, native and latent, it is of interest to determine the probabilities with which the system first reaches these structures. For this, at each characteristic temperature ($T = 0.65$, 0.6725, and 0.695), 10^3 trajectories were calculated starting from randomly chosen fully unfolded conformations. The latter were generated by adding successive beads at one end of the protein chain, each of the beads being randomly oriented with respect to the preceding bead in one of the three positive directions. We found that the probability to fold to the native state varied with temperature as 0.45 at $T = 0.65$, 0.44 at $T = 0.6725$, and 0.5 at $T = 0.695$. (Correspondingly, the probabilities to fold into the latent state are 0.55, 0.56, and 0.5, respectively.) Thus, over the temperature range

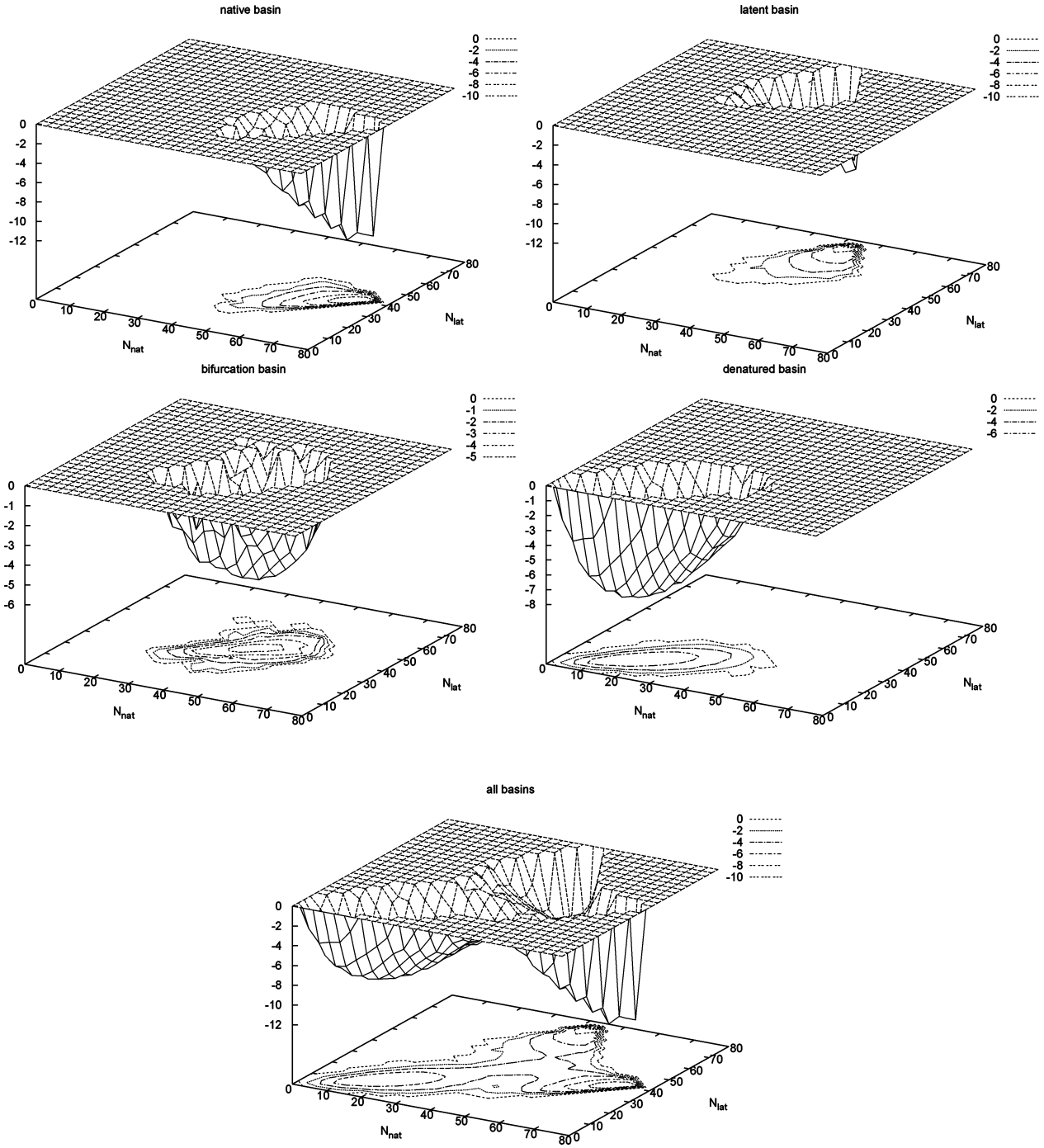

Figure 9. Projection of the equilibrium kinetic network of Figure 8a on the (N_{nat}, N_{lat}) plane; $T = 0.6725$.

TABLE 1: Fitted Values of the Rate Constants

T	r_{bg}	r_{gb}	r_{nb}	r_{bn}	r_{lb}	r_{bl}
0.65	2.8×10^{-8}	7.3×10^{-8}	2.3×10^{-7}	4.1×10^{-10}	2.7×10^{-7}	5.0×10^{-9}
0.6725	2.5×10^{-8}	2.5×10^{-7}	2.2×10^{-7}	1.2×10^{-9}	2.2×10^{-7}	1.4×10^{-8}
0.695	2.0×10^{-8}	7.1×10^{-7}	2.5×10^{-7}	4.9×10^{-9}	1.8×10^{-7}	3.8×10^{-8}

considered, the two states are almost equally accessible. However, the residence times of the system in these states, which determine their stability, are drastically different; i.e., the system remains in the native state much longer than in the latent state. Specifically, at $T = 0.65$, $T = 0.6725$, and $T = 0.695$ the residence times in the native state are, respectively, 12.2, 11.7, and 7.8 times longer than those in the latent state (see Table 1 for the rate constants of transitions from the native and latent states to the bifurcation basin).

B. Simulation Results for Achieving Equilibrium. On the basis of Figures 3, 4, and 8a, the exploration of the conformation space of the system can be considered as a series of successive transitions between characteristic basins of attraction on the free energy surface, with the possibility of return to a previously

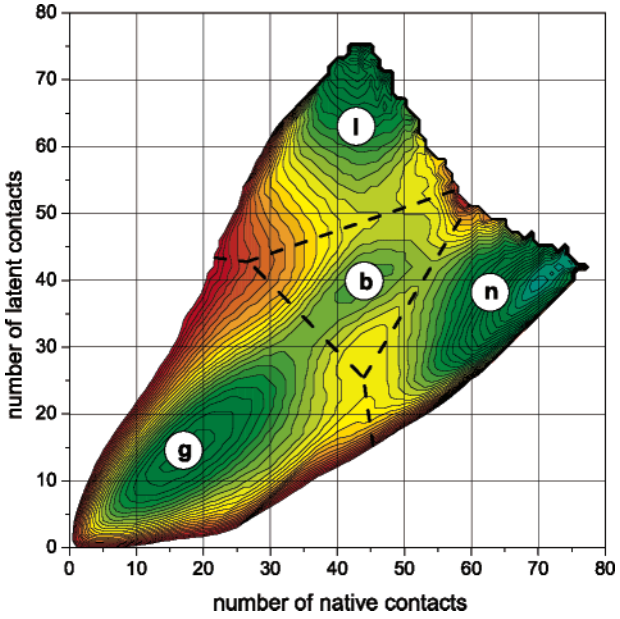

Figure 10. Separation of the reduced conformation space into the basins of attractions: (g) globule, (b) bifurcation basin, (n) native, and (l) latent states.

TABLE 2: Comparison of the Equilibrium Populations of the Basins from the Equilibrium Kinetic Network of Figure 8a with Those from Figure 11; $T = 0.6725$

	g	b	n	l
Figure 11	4.7×10^{-2}	4.4×10^{-3}	8.7×10^{-1}	7.9×10^{-2}
Figure 8a	6.0×10^{-2}	7.5×10^{-3}	8.5×10^{-1}	7.9×10^{-2}

visited basin (Figure 8a). It is, therefore, of interest to calculate the time-dependent populations of the states, i.e., $n_g(t)$, $n_b(t)$, $n_n(t)$, and $n_l(t)$, with the trajectories initiated in one of the states. Since we intend to determine the rate constants from these results and then to use them to calculate the first-passage time distributions theoretically (section IV.C), we do not use the globule as the initial state. Otherwise, the sequence of transitions would be similar to the process of folding from the unfolded sate, and thus the resulting rate constants could be biased. Specifically, the latent state was chosen as the initial one. To obtain the populations, we divided the (N_{nat}, N_{lat}) space into the basins of attraction, following the highest lines on the free energy surface at $T = 0.6725$ (Figure 3b), as is shown in Figure 10. Although the positions of the boundaries might vary slightly with the temperature (cf. Figures 3a−c), we assumed them to be the same at the three temperatures under consideration. As is evident by comparing Figure 3b and Figure 9, the separation into basins from this method and the FEP is very similar. Having the boundaries, we generated an ensemble of 1.5×10^3 trajectories initiated in the latent state. Given these trajectories, we counted the number of times, i.e., the number of the MC steps, that the system was found in each of the basins. At each temperature, the trajectories were continued until equilibrium in the system was attained. The resulting distributions are shown in Figures 11a−c. The nonmonotonic behavior of the globule and bifurcation basin populations, in contrast to that of the native state, shows that the process of populating the latter passes through the former states; i.e., the system, starting in the latent state, generally goes through the bifurcation and globule basins to reach the native state. At the lower temperature, however, visits to globule basin are rare, because the height of the barrier between the bifurcation basin and the globule is higher than

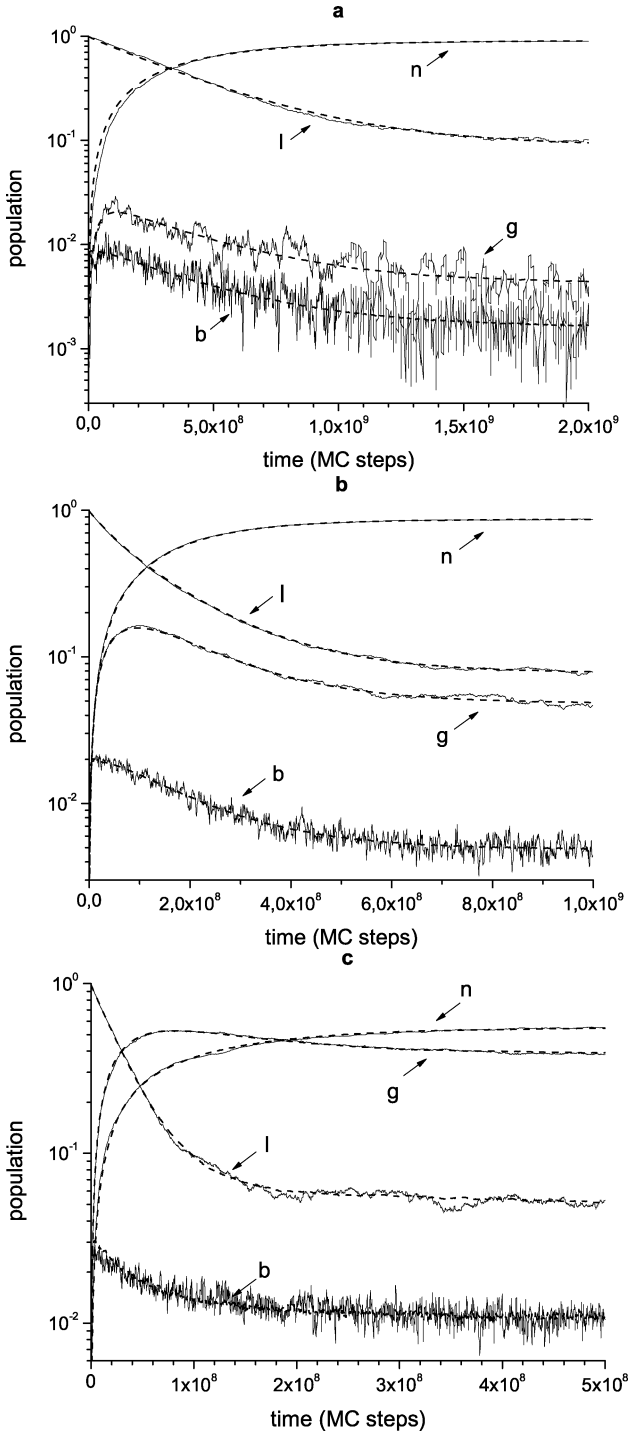

Figure 11. Time-dependent populations of the basins: (a) $T = 0.65$, (b) $T = 0.6725$, and (c) $T = 0.695$. The label g stands for the globule, b for the bifurcation basin, and n and l for the native and latent states, respectively.

those between the bifurcation basin and the native and latent states (Figure 3a). In contrast, at the higher temperature, at which the bifurcation basin is merged with the globule basin (Figure 3c), the visits to the globule basin are frequent. These observations are in accord with the values of the rate constants presented in Table 1 below. Also, it can be seen that the equilibrium population of the globule basin increases monotonically with temperature, which is typical for protein folding and explains the increase of the mean folding time with temperature (e.g.,

ref 14). In the present case, this increase is mostly at the expense of the population of the latent state, which monotonically decreases.

C. Kinetic Model. The time dependence of the populations in the various basins can be described by a system of coupled linear equations, similar to that used to describe folding kinetics of the 27-bead lattice protein[14] and ubiquitin.[33] In the present case, the system can be written as follows

$$\frac{dn_g}{dt} = r_{gb}n_b - r_{bg}n_g \quad (2)$$

$$\frac{dn_b}{dt} = r_{bg}n_g + r_{bn}n_n + r_{bl}n_{nl} - (r_{gb} + r_{nb} + r_{lb})n_b \quad (3)$$

$$\frac{dn_l}{dt} = r_{lb}n_b - r_{bl}n_l \quad (4)$$

$$\frac{dn_n}{dt} = r_{nb}n_b - r_{bn}n_n \quad (5)$$

where $n_i = n_i(t)$ is the probability of finding the system in state i at MC step t, r_{ji} is the rate constant for transitions from state i to j, and the indices g, b, l, and n stand for the semicompact globule basin, bifurcation basin, and the basins for the native and latent states, respectively; the kinetic scheme is shown in Figure 8b. Since the collapse of the unfolded chain into the globule is very fast (typically, it takes less than 1×10^5 MC steps) and the backward process (complete unfolding of the globule) is highly improbable at the temperatures considered, we simplify the system of equations by neglecting these processes. Also, we do not present the analytical solution of the system of eqs 2−5, even though it is readily written as a sum of four exponential terms. Such a solution is of little use here, because the roots of the characteristic equation, determining the decay times, are too complicated to reveal their dependence on the rate constants.

Given the system of equations, we can estimate the rate constants for the transitions between the characteristic states of the system. For this, the system of equations was solved numerically with the initial conditions $n_g = n_b = n_n = 0$ and $n_l = 1$ for the time-dependent populations of the states, i.e., $n_g(t)$, $n_b(t)$, $n_n(t)$, and $n_l(t)$. The rate constants appearing in the theoretical solution were then fitted to the simulation results; the maximum likelihood approach with an "entropic" functional[14] was employed. The resulting rate constants are presented in Table 1, and the fitted theoretical distributions are shown in Figures 11a−c. It is seen from the figure that the fitted distributions are in excellent agreement with the distributions from the MC simulations.

Tables 2 and 3 compare the equilibrium populations of the basins and the rate constants, which are calculated from the EKN of Figure 8a, with those from Figure 11 and Table 1, respectively. The populations were calculated as $n_i = Z_i/\sum_i Z_i = N_i/\sum_i N_i$, where N_i is the number of times the system was found in basin i, and the rate constants were calculated as $r_{ji} = Z_{ji}/Z_i = N_{ji}/(N_i \times 10^3)$, where N_{ji} is the number of times the system left basin i for basin j (i, j = g, b, n, l), and 10^3 is the length of the checking interval (the number of MC steps); i.e., transition state theory is used with the transmission coefficient equal to one. The populations agree reasonably well, and the rate constants differ by an order of magnitude throughout (a factor of 10 to 37), with the values from the EKN always larger. Given the very different approaches used, the agreement is reasonable. The difference arises from several sources. First,

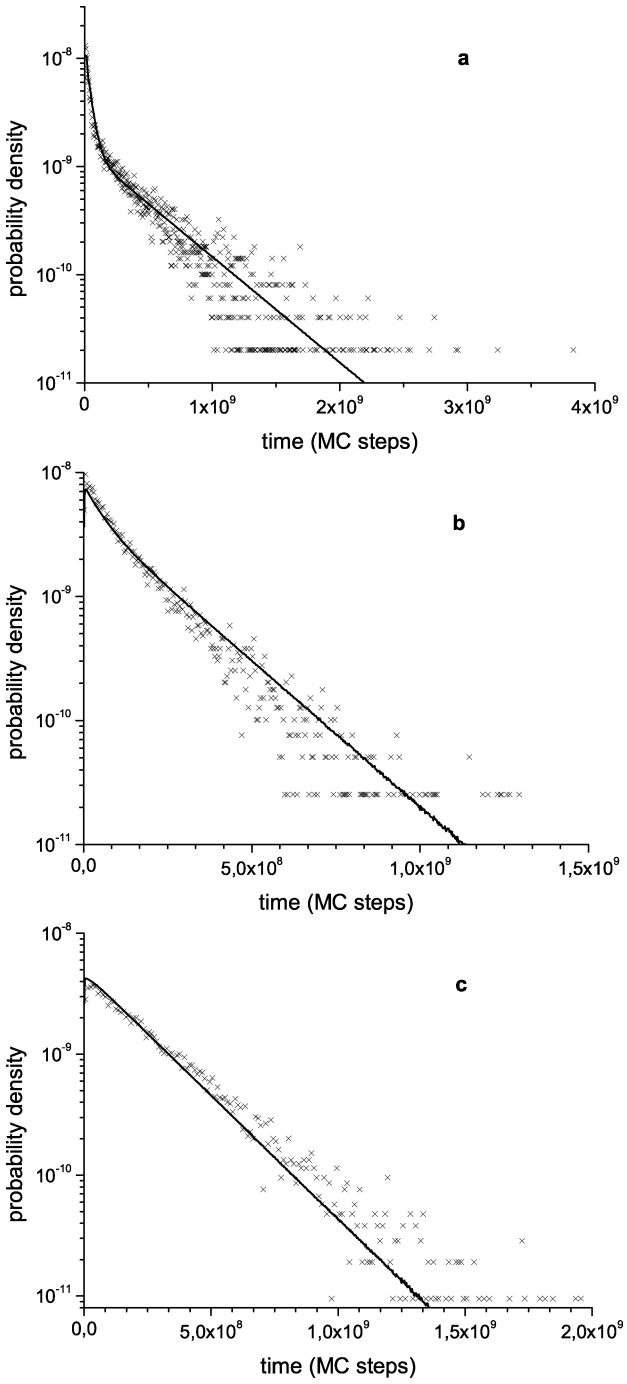

Figure 12. Folding time distributions: (a) $T = 0.65$, (b) $T = 0.6725$, and (c) $T = 0.695$. Crosses are for the simulation results, and the solid lines are for the corresponding theoretical distributions with the rate constants from Table 1.

N_{ji} in the above equation for r_{ji} is such that each time the system leaves a basin, crossing the transition state, it counts as a transition; thus every recrossing event contributes to the number of transitions. Analysis shows that if the trajectory were required to reach the bottom of the basin, then the transmission coefficient would be approximately $1/3$. The transmission coefficient is smaller than 1 due to the diffusive motion across the transition state. This diffusive motion has a component that is inherent in the kinetics and exists in the original multidimensional conformational space and a component due to the fact that additional recrossings occur when the trajectory is projected to a lower-dimensional space. A larger number of the PCA components used for clustering is expected to increase the heights of the

TABLE 3: Comparison of the Rate Constants from the Equilibrium Kinetic Network of Figure 8a with Those from Table 1; $T = 0.6725$

	r_{bg}	r_{gb}	r_{nb}	r_{bn}	r_{lb}	r_{bl}
Table 1	2.5×10^{-8}	2.5×10^{-7}	2.2×10^{-7}	1.2×10^{-9}	2.2×10^{-7}	1.4×10^{-8}
Figure 8a	2.5×10^{-7}	2.6×10^{-6}	2.6×10^{-6}	2.3×10^{-8}	4.3×10^{-6}	5.3×10^{-7}

barriers, so the rate constants from the EKN would be lower and thus closer to those from Table 1. However, to keep the results robust, such an extension of the PCA space dimensionality would require considerably better statistics and, correspondingly, a much longer trajectory.

D. First-Passage Time Distributions. To study folding time distributions,[14] the trajectories were started from unfolded conformations and terminated upon reaching the native state, so the "folding time" corresponds to the first-passage time. The unfolded conformations were generated as previously in section IV.A, when the probabilities to fold into the native and latent states were calculated. For each temperature, 10^4 trajectories were run with the maximum length of the trajectories equal to 10^{10} MC steps; this time was long enough so that all observed trajectories reached the native state. The folding time distributions are shown in Figure 12. It is seen that the distributions change gradually from a double-exponential distribution at the lower temperature (Figure 12a) to a single-exponential distribution at the highest temperature (Figure 12c). The single-exponential character of the distribution suggests two-state kinetics, i.e. that all states that contribute to the unfolded state are in equilibrium, forming an "extended globule".[14] As is seen from Figure 3c, at $T = 0.695$ the bifurcation basin merges with the globule basin, so these basins do form an extended globule, while the latent state is separated from them by a noticeable barrier. With the latent state considered as a dead-end trap, the mean folding time ($\bar{t}_f$) is determined by the equation that we derived in ref 14 (eq 28); in the present case it takes the form

$$\bar{t}_f = \frac{1}{\tilde{r}_{ng}}\left(1 + \frac{\tilde{r}_{lg}}{r_{gl}}\right) \tag{6}$$

where $\tilde{r}_{ng}$ and $\tilde{r}_{lg}$ are the effective rates of transitions from the extended globule to the native and latent states. The values of these constants can be estimated as $\tilde{r}_{ng} = r_{bg}r_{nb}/(r_{gb} + r_{nb} + r_{lb})$ and $\tilde{r}_{lg} = r_{bg}r_{lb}/(r_{gb} + r_{nb} + r_{lb})$, which with the data of Table 1 gives $\tilde{r}_{ng} \approx 4.4 \times 10^{-9}$ and $\tilde{r}_{lg} \approx 3.2 \times 10^{-9}$. Correspondingly, eq 6 yields for the mean folding time $\bar{t}_f \approx 2.5 \times 10^8$, which is in good agreement with its value obtained in the simulations, which is $\sim 2.45 \times 10^8$ (Figure 2). Kinetic MC simulations, i.e., the MC simulations made on the basis of the kinetic equations with given rate constants (see, e.g., ref 14 for how such simulations are performed), show that the latent state joins the extended globule (i.e., comes into equilibrium with the bifurcation basin and globule) rather quickly, at approximately $t = 5 \times 10^7$. However, as eq 6 shows, the contribution of the latent state, which is determined by the second term in the brackets on the right-hand side of the equation, remains small; it is on the order of 10%.

In contrast, at the lower temperature ($T = 0.65$), the latent state plays an essential role. Here the barriers from the bifurcation basin to the native and latent basins are comparable and low, while the globule is separated from the bifurcation basin by a relatively high barrier. The kinetics are far from the two-state kinetics here, and no extended globule exists. The short times in the double-exponential distribution of Figure 12a correspond to the trajectories that pass to the native state without visiting the latent basin, and the long times correspond to the trajectories that visit the latent basin before reaching the native state. Due to the high barrier between the bifurcation and the globule basins, in comparison with those to the native and latent basins, the events of partial unfolding of the system, i.e., passing into the globule basin, are relatively rare (see the corresponding rate constants in Table 1). If we neglect the partial unfolding and use in eq 6 the rate constants r_{nb} and r_{lb} instead of $\tilde{r}_{ng}$ and $\tilde{r}_{lg}$, respectively (i.e., do not include the globule), then the equation gives $\bar{t}_f \approx 2.5 \times 10^8$, which is also in good agreement with the simulation results ($\sim 2.41 \times 10^8$, Figure 2). However, in this case the contribution of the latent state is dominant, i.e., the second term in the brackets on the right-hand side of the equation is as large as ~ 54. The case of $T = 0.6725$ (Figure 12b), which corresponds to the minimum value of the mean folding time (Figure 2), represents an intermediate situation; here the contribution of the latent state is noticeable but not so significant as at $T = 0.65$.

The folding time distributions can also be calculated from kinetic equations with the rate constants of Table 1. For this, the system of eqs 2−5 is modified to exclude the return from the native state to the bifurcation basin since we are dealing with the first-passage time. Equation 3 then reduces to

$$\frac{dn_b}{dt} = r_{bg}n_g + r_{bl}n_l - (r_{gb} + r_{nb} + r_{lb})n_b \tag{7}$$

and eq 5 to

$$\frac{dn_n}{dt} = r_{nb}n_b \tag{8}$$

Equations 2 and 4 are unchanged. The system of equations in eqs 2, 4, 7, and 8 is solved with the initial conditions $n_g = 1$ and $n_b = n_n = n_l = 0$. From $n_n(t)$ (see eq 17 of ref 14), the folding time distribution is calculated as

$$p_n(t) = \frac{dn_n}{dt} \tag{9}$$

Figure 12 compares the theoretical distributions thus obtained with the corresponding distributions from the simulations. The results show good agreement. It is important to note that the sequence of events for the first passage to the native state from the unfolded state (Figure 12) is different from that for the case when the trajectories were started at the latent state (Figure 11). Therefore, the fact that the rate constants determined in the latter process describe the first-passage time distributions of folding from the unfolded state is evidence of the validity of the kinetic scheme presented in Figure 8; i.e. the four characteristic states of the protein (the semicompact globule, bifurcation basin, native, and latent states) are necessary and sufficient for describing the folding kinetics of the model protein under consideration.

V. Concluding Remarks

Taking amyloid-fiber-forming proteins as the prototype, we have designed a model lattice protein possessing two low-energy states, native and latent, which have a sizable part of their structures in common (two "α-helices") and differ in the content of "α-helices" and "β-strands" in the rest of the structures; i.e.

for the native state this part is α-helical, and for the latent state it is composed of β-strands. A Gō-like potential was used, with the contact energy matrix including both the native and the latent contacts. This contrasts with the recent work of Okazaki et al.,[34] which combines two separate Gō potentials. By performing Monte Carlo simulations, we have shown that the system can fold into both states and pass from one state to the other. To gain insight into the folding process for such a complex system, we started by mapping the folding trajectories on a reduced free energy surface (depending on the number of native and latent contacts) and contacts maps. We found that a semicompact globule is formed very rapidly. It has contacts that appear in the secondary structural elements in the native and latent states but not simultaneously. From the globule, the system can follow two different folding pathways: In one case, which presents the typical folding scenario, the system passes to a basin, called the bifurcation basin, in which the common part of the native and latent structures is formed (i.e., the two "α-helices" in the lower part), with the upper part of the structure retaining the possibility of forming either the native or the latent structure (i.e., one "α-helix" and two "β-strands", specific to the native state, and four "β-strands", specific to the latent state). Thus, the bifurcation basin represents an on-pathway intermediate, from which the system can pass into either the native or the latent state. If the system does not reach native state directly, because it enters the latent basin, it partly unfolds and repeats the attempt. In the other case, which presents an alternative folding scenario, which occurs with low probability, the system folds into the native state directly; here the part specific to the native state is formed first and then the part that is common to the native and latent structures. It is interesting to note that similar (direct) pathways to the latent state were not observed.

Since projections on simple reaction coordinates, such as the numbers of native and latent contacts used here to determine the free energy, could hide the complexity of the free energy surface,[28] we verified the results with a new approach.[30] On the basis of consideration of the entire set of local contacts (bonds), we reduced the dimensionality of the conformation space by using a principal component analysis (PCA), which allowed us to project the original space on a reasonable number of principal components (bond-PCA coordinates). Knowledge of the bond-PCA coordinates made it possible to construct one-dimensional free energy profiles (FEPs),[28] which include different numbers of the bond-PCA coordinates, as well as the equilibrium kinetic network (EKN). For this purpose, a long equilibrium trajectory was projected on the bond-PCA coordinates, and the points in this reduced bond-PCA space were clustered. Both the FEPs and EKN revealed four basins of attraction, namely, the semicompact globule, bifurcation, native, and latent basins. Moreover, the projection of the basins on the (N_{nat}, N_{lat}) plane gave a picture very close to the reduced free energy surface obtained using the standard approach (i.e., through the calculation of the residence probabilities at the (N_{nat}, N_{lat}) points). The fact that the latter approach turns out to be quite accurate is probably due to the Gō-like potential used, where the energy is determined by the numbers of native and latent contacts. The FEPs also showed that the structure of the EKN is well reproduced by a few (the three first) PCA components, while the higher components contribute mainly to the heights of the barriers between the basins.

For a more detailed understanding of the folding kinetics, we calculated the time-dependent populations of the basins. The free energy surface, based on the numbers of native and latent contacts, was separated into the regions corresponding to these basins. The trajectories were started at the latent state and continued until equilibrium was achieved. The equilibrium populations of the basins are in satisfactory agreement with those from the EKN. One advantage of studying such time-dependent populations is that they allow estimation of the rate constants for the essential channels of transitions. For this, we solved the system of linear kinetic equations describing the transitions between the characteristic states, and compared the solution with the simulated distributions. The rate constants showed that the rates of the transitions from the bifurcation basin to the globule and from the native and latent states to the bifurcation basin are the most sensitive among them to the temperature. All of these rates noticeably increase with temperature, which results in a faster achievement of equilibrium. The rate constants thus estimated are in agreement with those from the EKN, except for a common factor on the order of 10, which can be attributed to an insufficient number of the bond-PCA coordinates used to build the EKN, as well as the relatively small simulation length used to construct the EKN. Also, we simulated first-passage (folding) time distributions and compared them to the corresponding theoretical distributions from the kinetic model with the rate constants obtained from the previously mentioned time-dependent populations of the basins. The results were found to be in very good agreement. We note that the first passage to the native state from the unfolded state involves a different sequence of events from the process of populating the characteristic states of the system when the trajectories are started at the latent state. Therefore, the fact that the rate constants determined in the latter process describe the first-passage time distributions is evidence of the validity of the kinetic scheme; i.e., the four characteristic states of the protein (the semicompact globule, bifurcation basin, native and latent states) are necessary and sufficient to describe the folding kinetics in the model protein under study.

An interesting finding of this work is that the region where the folding trajectories bifurcate (to reach either the native or the latent state) is a basin of attraction, in which the common part of the native and latent structures is formed. This phenomenon may be general for proteins in which native and latent states have a sizable part of the structure in common. The fact that the common part of the structures is formed in the bifurcation basin is of particular interest because it offers the possibility of altering the folding process in the late stages of folding, thereby increasing or decreasing the probability of misfolding. In lysozyme, for example, there are amyloidogenic mutants, whose X-ray structures show an essentially unchanged native state.[35] Decrease of the probability of misfolding can be achieved, for example, by introducing mutations that hinder the formation of the latent state and do not affect the formation of the common part of the native and latent structures. Model simulations show that if the latent contacts between monomers 8 and 61 and 21 and 36 (Figure 1b) are neutralized (i.e., the contacts energies are set zero instead of -0.975) at $T = 0.6725$, the probability of folding into the latent state decreases from $\sim$0.56 in the original protein to $\sim$0.36 in the protein with the two latent contacts neutralized. In contrast, if certain native contacts are neutralized, then the probability of folding into the latent state increases; e.g., if the contact energies between monomers 3 and 28 and 20 and 47 (Figure 1a) are set to zero, the probability of folding into the latent state increases to $\sim$0.75. It will be interesting to see whether the bifurcation basin found here exists and can be trapped in real proteins.

Acknowledgment. This work was supported in part by a grant from the Civilian Research and Development Foundation (Grant No. RUP2-2629-NO-04). The research at Harvard was supported in part by a grant from the National Institutes of Health. A.Y.P. and S.F.C. acknowledge support from the Russian Foundation for Basic Research (Grant No. 06-04-48587).

References and Notes

(1) Dinner, A. R.; Šali, A.; Smith, L. J.; Dobson, C. M.; Karplus, M. *Trends Biochem. Sci.* **2000**, *25*, 331−339.
(2) Wang, Z.; Mottonen, J.; Goldsmith, E. J. *Biochemistry* **1996**, *35*, 16443−16448.
(3) Ye, S.; Goldsmith, E. J. *Curr. Opin. Struct. Biol.* **2001**, *11*, 740−745.
(4) Sohl, J. L.; Jaswal, S. S.; Agard, D. A. *Nature* **1998**, *395*, 817−819.
(5) Volkman, B. F.; Lipson, D.; Wemmer, D. E.; Kern, D. *Science* **2001**, *291*, 2429−2433.
(6) Hu, X.; Wang, Y. *J. Biomol. Struct. Dyn.* **2006**, *23*, 509−517.
(7) Dobson, C. M. *Nature* **2003**, *426*, 884−890.
(8) Jiménez, J. L.; Guijarro, J. I.; Orlova, E.; Zurdo, J.; Dobson, C. M.; Sunde, M.; Saibil, H. R. *EMBO J.* **1999**, *18*, 815−821.
(9) Guo, Z. Y.; Thirumalai, D. *Biopolymers* **1995**, *36*, 83−102.
(10) Abkevich, V. I.; Gutin, A. M.; Shakhnovich, E. I. *Proteins* **1998**, *31*, 335−344.
(11) Nakamura, N. K.; Sasai, M. *Proteins* **2001**, *43*, 280−291.
(12) Abkevich, V. I.; Gutin, A. M.; Shakhnovich, E. I. *J. Chem. Phys.* **1994**, *101*, 6052−6062.
(13) Dinner, A. R.; Karplus, M. *J. Phys Chem. B* **1999**, *103*, 7976−7994.
(14) Chekmarev, S. F.; Krivov, S. V.; Karplus, M. *J. Phys Chem. B* **2005**, *109*, 5312−5330.
(15) Šali, A.; Shakhnovich, E.; Karplus, M. *Nature* **1994**, *369*, 248−251.
(16) Locker, C. R.; Hernandez, R. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 9074−9079.
(17) Gō, N. *Annu. Rev. Biophys. Bioeng.* **1983**, *12*, 183−210.
(18) Dill, K. A. *Biochemistry* **1985**, *24*, 1501−1509.
(19) Leonhard, K.; Prausnitz, J. M.; Radke, C. J. *Phys. Chem. Chem. Phys.* **2003**, *5*, 5291−5299.
(20) Gupta, P.; Hall, C. K.; Voegler, A. C. *Protein Sci.* **1998**, *7*, 2642−2652.
(21) Harrison, P. M.; Chan, H. S.; Prusiner, S. B.; Cohen, F. E. *J. Mol. Biol.* **1999**, *286*, 593−606.
(22) Dima, R. I.; Thirumalai, D. *Protein Sci.* **2002**, *11*, 1036−1049.
(23) Fawzi, N. L.; Chubukov, V.; Clark, L. A.; Brown, S.; Head-Gordon, T. *Protein Sci.* **2005**, *14*, 993−1003.
(24) Cellmer, T.; Bratko, D.; Prausnitz, J. M.; Blanch H. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 11692−11697.
(25) Cellmer, T.; Bratko, D.; Prausnitz, J. M.; Blanch, H. *Biotechnol. Bioeng.* **2005**, *89*, 78−87.
(26) Galzitskaya, O. V.; Finkelstein, A. V. *Protein Eng.* **1995**, *8*, 883−892.
(27) Karplus, M. *Fold. Des.* **1997**, *2*, 569−576.
(28) Krivov, S. V.; Karplus, M. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 14766−14770.
(29) Krivov, S. V.; Karplus, M. *J. Phys. Chem. B* **2006**, *110*, 12689−12698.
(30) Krivov, S. V.; Karplus, M., to be submitted for publication.
(31) Lay, D. *Linear Algebra and Its Applications*, 2nd ed.; Addison-Wesley: New York, 2000.
(32) Rhee, Y.; Pande, V. *J. Phys. Chem. B* **2005**, *109*, 6780−6786.
(33) Chekmarev, S. F.; Krivov, S. V.; Karplus, M. *J. Phys. Chem. B* **2006**, *110*, 8865−8869.
(34) Okazaki, K.; Koga, N.; Takada, S.; Onuchic, J. N.; Wolynes, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 11844−11849.
(35) Dumoulin, M.; Kumita, J. R.; Dobson, C. M. *Acc. Chem. Res.* **2006**, *39*, 603−610.