Large-Scale Context in Protein Folding: Villin Headpiece†

Ariel Fernández,*,‡,§ Min-yi Shen,|| Andrés Colubri,§ Tobin R. Sosnick,‡,⊥ R. Stephen Berry,|| and Karl F. Freed*,||

Institute for Biophysical Dynamics, James Franck Institute and Department of Chemistry, and Department of Biochemistry and Molecular Biology, University of Chicago, 920 East 58th Street, Chicago, Illinois 60637, and Instituto de Matemática, Universidad Nacional del Sur, Consejo Nacional de Investigaciones Científicas y Técnicas, Bahía Blanca 8000, Argentina

Received July 24, 2002; Revised Manuscript Received November 14, 2002

ABSTRACT: The villin headpiece folds autonomously in vitro forming three α-helical regions. Local propensities, however, strongly disfavor the formation of the C-terminal helix because most native residue pairs in that helix are hydrophobic/polar mismatches. Even the N-terminal helix is unfavored according to the AGADIR criterion. Our coarse-grained ab initio simulations reveal three-body correlations in which hydrophobic residues position to protect amide-carbonyl hydrogen bonds from attack by water, thus inducing the growth of the C-terminal helix and guiding the folding process. Similar correlations are also found in all-atom simulations with an implicit solvent model that accurately reproduces the results of simulations with explicit solvent molecules. The correlations establish a large-scale, many-body context that may be probed experimentally by introducing mutations of certain nonobvious residues that reside outside the native hydrophobic core but that are predicted to affect the folding rates and dynamics dramatically.

The problem of protein folding breaks into three parts: (1) the genomic question of relating sequence to structure, (2) the operational question of how structure is related to function, and (3) the kinetic question of what pathways lead to folding and how the system finds them. Here we address the third of these. The experimental data regarding folding is rich with suggestive and coarse-grained results but remains very incomplete from the standpoint of finding folding pathways or, in the vernacular of organic chemistry, mechanisms. Theoretical approaches capable of revealing detailed folding pathways, particularly simulations described in more detail below, have largely been so fine-grained that it has been impractical to carry them to time intervals long enough to compare the computed results with experiments. This report is one of a series based on two approaches to simulation that are designed to provide results for time scales that can be matched in experiments and are time scales typical of the folding processes of real proteins.

Identifying and representing the key interactions among residues adequately has been one impediment to understanding the protein folding process that has to date made it difficult to develop an ab initio approach to describe folding pathways (*1*−*3*). Accounting for local propensities in the primary sequence has been one important step toward making such inferences (*4*). For example, helical regions may be predicted with some degree of accuracy in the case of hierarchical folding schemes, where native local structure formation precedes large-scale organization (*5*), as, for example, in the diffusion-collision scenario (*6*). However, there are clear-cut instances where a large-scale correlated context (involving three- and higher-body correlations) is required to induce and stabilize the formation of local structure (*7*−*9*). Furthermore, the large-scale organization may sometimes overrule propensities one might infer from the local structure (*10*), as in the case of β-lactoglobulin, a quintessential nonhierarchical folder. The preferred folding pathway of this protein has been predicted at the coarse-grained level by our ab initio methods, indicating folding dynamics with an early labile overabundance of α-helical structure that becomes a durable β-strand when the appropriate, stabilizing long-range contacts form (*11*, *12*). These predictions for β-lactoglobulin, consistent with the early studies of this molecule, were soon corroborated experimentally (*13*).

In this paper, we focus on the folding of the villin headpiece, a small 36-amino acid protein that folds autonomously into three helical regions (*14*). These helical regions would be considered as rather unlikely based on local propensities, as revealed by the AGADIR (*4*) plot in Figure 1. A possible folding pathway for this protein has been generated by Duan and Kollman in an all-atom, explicit solvent supercomputer simulation (*15*) for 1 μs, or approximately one-tenth of its estimated folding time (*16*). The duration of this simulation is inadequate to reveal the folding path: the fluctuations in the radius of gyration at the end of the trajectory are of the same magnitude as their overall average. Thus, this 1-μs process is too brief to reach a discernible critical folding stage marked by a dramatic and persistent quenching of structural fluctuations (*17*).

† R.S.B. and T.R.S. acknowledge the support of a grant from the Packard Foundation. R.S.B. likewise acknowledges support from the National Science Foundation. K.F.F. is supported, in part, by Grant GM56678 from the National Institutes of Health (General Medicine).

* Correspondence should be addressed to the following authors. (A.F.) ariel@uchicago.edu. (K.F.F.) k-freed@uchicago.edu.

‡ Institute for Biophysical Dynamics, University of Chicago.

§ Universidad Nacional del Sur.

|| James Franck Institute and Department of Chemistry, University of Chicago.

⊥ Department of Biochemistry and Molecular Biology, University of Chicago.

10.1021/bi026510i CCC: $25.00
Published on Web 01/01/2003

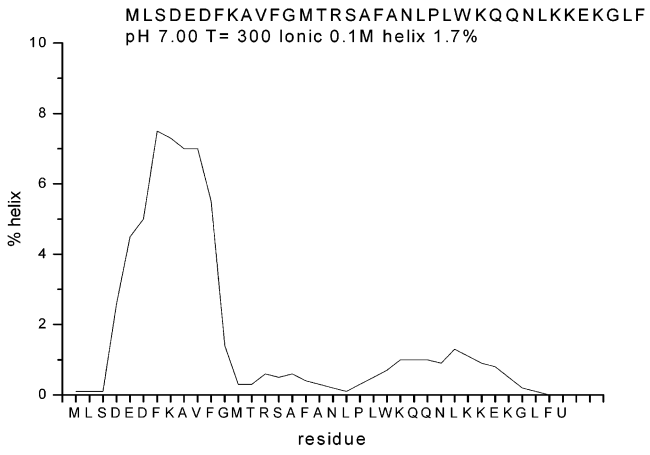

FIGURE 1: AGADIR profile of percentage helix probability based on local propensities for the villin headpiece at $T = 300$ K, pH 7.

Paradoxically, the enormous detail generated by all-atom simulations has not stimulated the examination of an apparent paradox of this process: How is it possible that the villin headpiece forms a C-terminal helix that is mostly mismatched vis-à-vis hydrophobic/polar properties of the (i, $i + 3$) and (i, $i + 4$) residue pairs involved? Heightening this question is the extremely low probability of forming this helix as predicted by AGADIR: below 2% according to well-established local propensity estimates (*4*). We rephrase the question in a more general way: How does large-scale organization (i.e., three- and higher-body correlations), particularly including interactions between residues far from one another in a sequence, prevail over local structural propensities so as to induce the formation of the C-terminal helix?

METHODS

To address these questions, we need to cover time scales significantly longer than those accessible to all-atom explicit-solvent simulations and do so in a reproducible manner that enables us to make predictions of statistical relevance. We have used two ab initio approaches to attack this problem, one a very coarse-grained method and the other (based on Langevin dynamics), a much finer-grained, all-atom approach. Both avenues of attack treat the solvent implicitly, but they differ in the level of structural detail used to represent the protein chain. The consequence of the implicit treatment of the solvent is that we lose the simple pairwise additivity that characterizes the all-atom potential as every interaction is now dependent on the solvent environment dictated by the large-scale organization of the chain. In the more detailed method, based on Langevin dynamics, the all-atom representation of the protein is used. The more drastically simplifying approach, named the folding machine (FM) (*12*, *17*), sacrifices some chain-structure resolution with the compensation that it can produce enough and long enough trajectories to generate statistically significant information.

Folding Machine. This method is based on a simplified view of backbone torsional dynamics in which the time-dependent variables are not the torsional coordinates themselves but the time-dependent occupancies of the basins of the Ramachandran maps of the individual residues. The algorithm is, in effect, a pattern-biased Monte Carlo method, with transitions made among the available Ramachandran basins of the Φ and Ψ dihedral angles. This device incorporates the constraints to which torsional coordinates are subject as a result of the local steric clashes between backbone and side chain. Structures are then inferred from calculations based on a force field and on the type of structure consistent with the pattern of basin occupancies at each stage of folding. Hence, as in any Monte Carlo method, every structure is physically plausible, but structures occurring in a sequence of steps need not be directly and mechanically accessible as they would in a molecular dynamics simulation. Structure determination is required to find the extent of energetic stabilization of the residues that would occur in a change from their preexisting pattern of interactions. This change of energy, together with the entropic change that would accompany a change of basin occupancies, in turn dictates the probability for the residue to change its Ramachandran basin; the less structurally involved, the more prone is a residue to change its Ramachandran basin. The energetic cost of a virtual move that would disrupt preexisting interactions is not evaluated merely on the basis of the pairwise in-bulk interactions, which are regarded as zeroth-order approximations, but includes a rescaling of the potential that depends on the local context arising from the large-scale organization of the chain (*7*, *9*, *17*). The rescaling depends specifically on the extent of burial of interacting residues in desolvation shells (i.e., on the number of hydrophobic residues surrounding each α-carbon). Because desolvation changes the effective dielectric, a fully desolvated hydrogen bond has roughly 10 times the stability of a fully exposed one. This ansatz is at the crux of cooperativity.

While the nonmechanical FM algorithm makes no attempt to reproduce dynamic-mechanical behavior, whether of a folder or a nonfolder, and overlooks the mechanical restrictions that might otherwise prevent a move forbidden by excluded volume, all the structures it finds are acceptable real structures. Like other random or guided random search algorithms, it has some advantages: (a) It is 6−7 orders of magnitude faster than an all-atom calculation based on mechanics, making realistic folding time scales accessible (*1*−*12*, *17*); (b) it reasonably captures the cooperativity of the folding process insofar as the potential depends on a crude evaluation of the conformation-dependent environments generated by the chain as it folds; (c) it reproducibly generates (on average once every seven runs) stable folds within 5 Å root-mean-square deviation (rmsd) from native for about 61% of sampled autonomous folders of all moderate sizes ($33 < N < 80$) and even allowed us to predict the tertiary structure of non-native on- and off-pathway intermediates for larger folders such as β-lactoglobulin (*12*); (d) it provides a clear-cut way of deciding which runs are unsuccessful (i.e., those that fail to produce a stable fold and whose long-time basin-hopping behavior does not produce a reproducible dramatic quenching of structural fluctuations); and (e) it uses no a priori information on target folds (*1*, *16*, *18*−*22*) or energetic biases, apart from the dependence on stepwise changes in free energies. By dealing with the evolution of constraints (i.e., Ramachandran basins) rather than coordinates (*23*−*29*), the dynamics is judiciously simplified. So far, the rate of success of the FM is widely variable: while no run has properly folded *lambda repressor* (PDB 1lmb), over 85% of runs successfully fold the hyperthermophile variant of protein G (PDB 1gb4).

The algorithm consists of a stochastic assignment of the coarsely resolved dynamics (*12*, *17*, *28*), simplified to the level of time-evolving basin assignments. An operational premise is that steric restrictions imposed by the side-chains on the backbone may be subsumed into the basin-hopping dynamics. The side-chain constraints define low-energy regions in each residue's Ramachandran map that can be explored to obtain an optimized pattern of nonbonded interactions.

The basin location of each residue defines the polypeptide's conformation. This string of basin locations, termed the local topology matrix or LTM(t)[1], is a coarse representation but one that satisfies the inherent geometrical constraints of a real polypeptide chain. The precise coordinates of the chain (i.e., the physical realization of a LTM) are defined by explicit Φ, Ψ angles. To maintain structural continuity during a folding trajectory, the explicit dihedral angles are retained for each residue from one time step to the next until that residue's basin is explicitly changed, as determined by the criteria noted below.

To make moves, a structure is generated with a set of explicit Φ, Ψ angles compatible with the coarse description. At present, these structures are found by the use of PROCHECK. This explicit realization, determined by a minimization of the energy with respect to the Φ, Ψ dihedrals, determines the extent of structural involvement of each residue. The degree of energetic stabilization is quantified with a semiempirical potential (*17*, *28*, *29*). This potential is used to determine which residues change their Ramachandran basin in the next step. Upon a basin transition, the new structure is again energetically minimized by varying the Φ, Ψ angles within the newly chosen basins, but only for the residues that just moved (*11*, *17*, *29*), leaving the rest of the backbone frame fixed.

The basin-hopping probability is dependent on the extent of structural engagement of the residue—quantifiable by the free energy cost associated with the virtual move of the changing basin. The probability of reaching a target basin depends on its lake area whose logarithm is the microcanonical entropy (*11*, *17*). To fit experimental folding measurements (*11*, *17*, *23*−*25*, *28*−*29*), a free residue is arbitrarily assigned a mean basin-hopping rate fixed at $f \sim 10^8$ Hz. From this ansatz, we can calculate the mean basin-hopping time for a free residue: We know from experiment the mean nucleation time for formation of a helix of length n, and we know the mean first-passage time, $\tau(n)$, that it takes to get all n residues in the correct Ramachandran basin needed to generate the helix. The quantity $\tau(n)$ is parametrically dependent on f and is independently known from experiment (*17*). In this way, we have realistically obtained f, and thereby, estimated the time it takes to make a transition to any particular word of Ramachandran basin occupancies.

The basic tenets of the FM approach are as follows:

(a) The Ramachandran basin areas, given by local steric constraints, do not change during the folding process (*27*). This invariance arises because local contributions to the potential are incorporated as constraints on Φ, Ψ angles, while nonbonded interactions are treated explicitly within the effective potential.

(b) Interbasin hopping is slower than intrabasin exploration. This justifies limiting the search to the backbone (Φ, Ψ) exploration of Ramachandran basins, readily represented by the evolving LTM(t).

(c) Side-chain torsional exploration occurs on a faster time scale than backbone LTM dynamics. This adiabatic approximation justifies the averaging of side-chain torsional motions in the stages of folding that precede the final side-chain packing. This simplification is adequate to represent early stages of compaction and hydrogen-bond protection.

(d) The effective enhancement of dielectric-dependent two-body interactions can be based on the extent of surrounding desolvation. This translates in our model as the rescaling of the zeroth-order (in-bulk) pairwise contributions depending on the number of neighboring hydrophobic residues. The weakening of hydrophobic attraction depending on the extent of hydrophobic burial is treated in a similar manner (*11*).

We model the long-range nonbonded interactions between the residues by including the following terms in our effective potential:

$$U_{nb} = U_{LJ} + U_{solv} + U_{coul} + U_{dip} + U_{Hbond} \qquad (1)$$

where U_{LJ} represents a Lennard-Jones contribution that accounts for excluded volume, U_{solv} is the effective solvophobic term accounting for the attraction between hydrophobic residues and the repulsion between hydrophobic and polar residues, U_{coul} represents the ionic energy between charged side-chains, U_{dip} models the backbone dipole−dipole interactions, and U_{Hbond} corresponds to the backbone hydrogen bonding.

Local environment strongly influences the pairwise interactions (*7*−*9*, *17*, *26*, *28*, *29*). The stabilities of hydrogen bonds are extremely context-sensitive. Solvent ordering around hydrophobic side chains inhibits backbone solvation and enhances intraprotein hydrogen bonding between two residues so bound. The algorithm used in this work treats the solvent implicitly in a way that three-body correlations affect the intramolecular potential. If a hydrophobic residue lies close enough that its interaction energy with a component of a hydrogen bond is greater than 2 kT, then that residue is considered in the desolvation sphere of the hydrogen bond. The energy of that bond is then reevaluated, by multiplying the interaction energy between the H-bonded residues i and j, $u^{(0)}_{hb}(i,j)$, by a factor $M(i,j,t) = [b(i,t)b(j,t)]^{1/2}$ where $b(i,t)$ is the number of hydrophobic residues desolvating H-bonded residue i, including that residue itself. Refs *11*, *12*, *17*, *28*, and *29* give a detailed description of the algorithm.

Langevin Dynamics Simulations with Implicit Solvent. The molecular simulations are generated with an algorithm that uses Langevin dynamics rather than deterministic Hamiltonian equations of motion. The solvent is treated implicitly, although comparison with calculations based on explicit solvent indicates that the method is reliable in this regard (*40*). The implicit solvent simulations use a modified version of the TINKER 3.7 molecular design package, as recently applied to met-enkephalin (*30*), and with the AMBER95 (parm94) force field as applied to the villin headpiece (*31*). The Ooi-Scheraga solvent-accessible surface area (*32*) solvation potential is added to calculate the solute portion of the total free energy. A sigmoidal-type distance-dependent dielectric coefficient (*33*)

[1] Abbreviations: LTM, local topology matrix.

$$\epsilon(r) = D - \frac{D-1}{2}(S^2r^2 + 2Sr + 2)e^{-Sr} \qquad (2)$$

(with $D = 78.0$ and $S = 0.3$) is used throughout the simulation for mimicking the differences between the electrostatic screening inside the protein and in bulk water. Nonbonding interactions are truncated at 8.0 Å. The numerical integration of the Langevin equation employs the velocity Verlet algorithm (*34*) with step sizes of 2.0 fs for 65 ns of the folding trajectory and 2.5 fs for the remainder. We use the Pastor−Karplus (*35*) scheme for calculating the atomic friction coefficients along with the experimental water viscosity of $h = 0.84$ cp. A 50-ns simulation with constant bulk water dielectric constant fails to preserve the native structure and finally produces a loose globular structure without helices, while a similar simulation with the dielectric function in eq 2 preserves the native structure (*36*). The simulation temperature is controlled at 300 K with a Berendsen-type (*37*) thermal bath coupling. All trajectories are stored at 10 ps time intervals. The rmsd between structures is calculated by matching the backbone C_α point sets by rigid translations and rotations using the conjugated gradient minimization algorithm (*38*). The secondary structures are assigned by the method suggested by Kabsch and Sander using the program DSSP (*39*). The method for estimating the speed-up gained by the implicit solvent method is explained in our previous work (*40*), where the dependence of the computational cost on N (the number of residues) is linear for the implicit water simulation but grows as $N^{1.5}$ for the explicit water case. We obtain a 200-fold speed-up for a 5-residue chain (*41*), so a 540-fold speed-up is estimated for a 36-residue peptide.

RESULTS

Coarsely and Finely Resolved Folding Pathways. An important distinction can be made between the N- and C-terminal helices of the villin headpiece. The N-terminal helix would be expected to have a mild helical propensity based solely on its pairwise according to AGADIR (*4*). The change upon folding in the pairwise-additive potential energy from the folding machine simulations is −11.0 kcal/mol, while many-body correlations lower this value to −25.6 kcal/mol. The corresponding free energy changes are, respectively, −0.7 and −2.2 kcal/mol according to the parametrization given in ref *17*. On the other hand, because of the pairwise mismatches, the native C-terminal helix is computed using the same parametrization as unstable with respect to the random coil by +8.2 kcal/mol, while its full correlated energy in the native fold lies 1.9 kcal/mol below that of the random coil.

To examine the dynamics of formation of the context-dependent helix, a typical FM trajectory of 2×10^5 iterations for the villin headpiece is shown in Figure 2. The following features are observed and are reproducible in 55 of 81 runs (the other runs are irreproducible in that they do not lead to a stable fold and fail to quench structural fluctuations): (a) There is a fairly reproducible critical time $t^* \sim 12 \pm 1$ μs at which structural fluctuations are ultimately quenched (Figure 2F−H). (b) A periodicity of ~3 μs is evident in the broad cycles of compaction and expansion (see Figure 2G) in all successful runs. (c) The initial burst phase of sudden compaction occurs during the first 600 ns (see Figure 2G and ref *32*); a similar collapse occurs even for a random sequence of the same amino acids. In this sense, this burst phase is an artifact—not a true folding—inasmuch as the initial configuration has no interresidue contacts. (d) Typical intermediate structures and the stable final structure, found by the FM, are displayed in Figure 2A−E. The rmsd of the stationary folded structures is 4.5−5.5 Å from the native structure. The structures found by the explicit solvent molecular dynamics method are similar. At the end of the 1-μs MD run of Duan and Kollman, the radius of gyration still fluctuates in the range of 16−8.7 Å, indicating that the

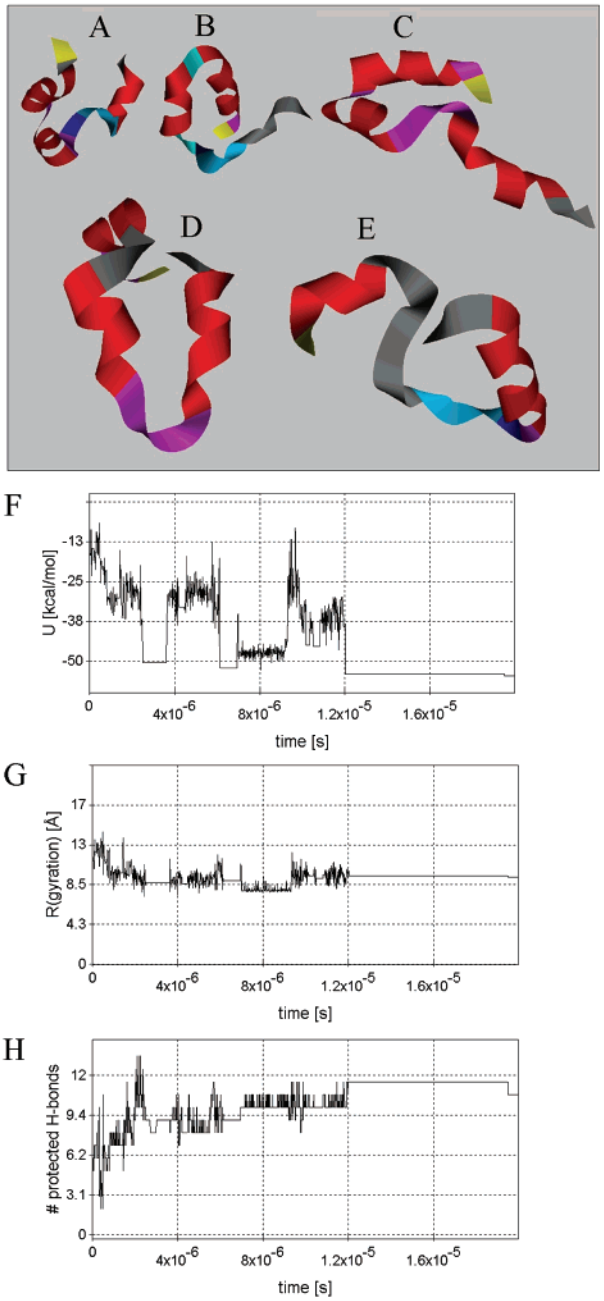

FIGURE 2: Typical FM run for the villin headpiece involving 2×10^5 iterations. The five schematic structure snapshots (A−E) are taken at 0.8, 3.5, 6.5, 10, and 16 μs, respectively. The plots in panels F−H represent, respectively, the time-dependence of the intramolecular correlated internal energy, the radius of gyration, and the number of protected hydrogen bonds. To qualify as a protected hydrogen bond, the residue pair must be surrounded by at least three hydrophobic residues, which are thus obviously engaged in three-body correlations.

ground-state had not been found (*15*). On the other hand, the final stationary value of 9.4 Å in the FM simulations agrees well with the native structure (9.33 Å) (*14*). The dispersion in this observable over the 55 successful runs is 7% of the mean reported value.

A correlation is detectable in the FM trajectory between the quenching of structural fluctuations (Figure 2F−G) and a build-up of a sustainable population of protected (partially buried) amide-carbonyl hydrogen bonds (Figure 2H). The sudden decrease and stabilization of the structure, apparent in the time-dependent plots of internal energy and radius of gyration (Figure 2F,G), occurs precisely at the time when a sudden and dramatic buildup of highly protected (dehydrated) hydrogen bonds appears. This buildup is followed by a plateau in the number of desolvated hydrogen bonds. To qualify as a protected hydrogen bond, the residue pair must be surrounded by at least three hydrophobic residues, which are obviously engaged in strong three-body correlations.

Two regions of stasis are detectable in the folding machine dynamics as displayed in Figure 2F−H. They represent two different ways to achieve intramolecular protection of the C-terminal helix, ways that exchange with different transition times for different runs (varying almost continuously with the run from 2.2 to 4 μs). The conformations in the regions of stasis are not significantly populated in the average over the ensemble of runs, as stasis occurs at different times for different runs. Only when the N-terminus stabilizes and protects the fragile C-terminal helical structure (otherwise it hardly deserves to be called a real helix) does the villin headpiece achieve its native structure.

In consonance with the all-atom trajectory (*15*), the FM reveals an initial burst phase marked by a sudden rise in helicity (up to 55% of the native level) within the first 600 ns (Figure 2A,F−H). This helicity appears mainly because local propensities tend to favor the formation of the N-terminal helix. The correlation between the nonbonded energy U (Figure 2F), chain compaction (Figure 2G), and hydrogen-bond burial (Figure 2H) implies that secondary structure formation is concurrent, and its extent might be even proportional to surface burial, in accord with refs *41* and *42*.

The backbone hydrogen bonds in the N-terminal helix are internally protected (i.e., desolvated), mainly through (i + 3, i, i + 4) and (i, i + 1, i + 4) three-body correlations, where the first entry denotes the hydrophobic or desolvating group (*17*, *23*−*26*, *28*, *29*). Within the same time range, the nucleus for the C-terminal helix is being shaped by three-body correlations, especially as the looped residue Ala19 approaches the turn of the amide-carbonyl hydrogen bond Gln26−Lys30. The Ala19 and Leu21 exchange roles as desolvators of the Gln26−Lys30 hydrogen-bonded pair on a time scale of about 800 ns in all 55 successful runs with a dispersion in the cycle length of less than 20 ns. We conclude that this exchange is responsible for the fluctuations in the folding process revealed by the temporal behavior of the radius of gyration, native helicity, and internal energy within this time window (*15*).

The protective role of Ala19 as a desolvator for the growth of the C-terminal helix can also be observed in the all-atom Langevin computation with implicit solvent that energetically penalizes the exposure of hydrophobic surfaces (See refs *36* and *40* and Methods). The snapshots in Figure 3A−D reveal

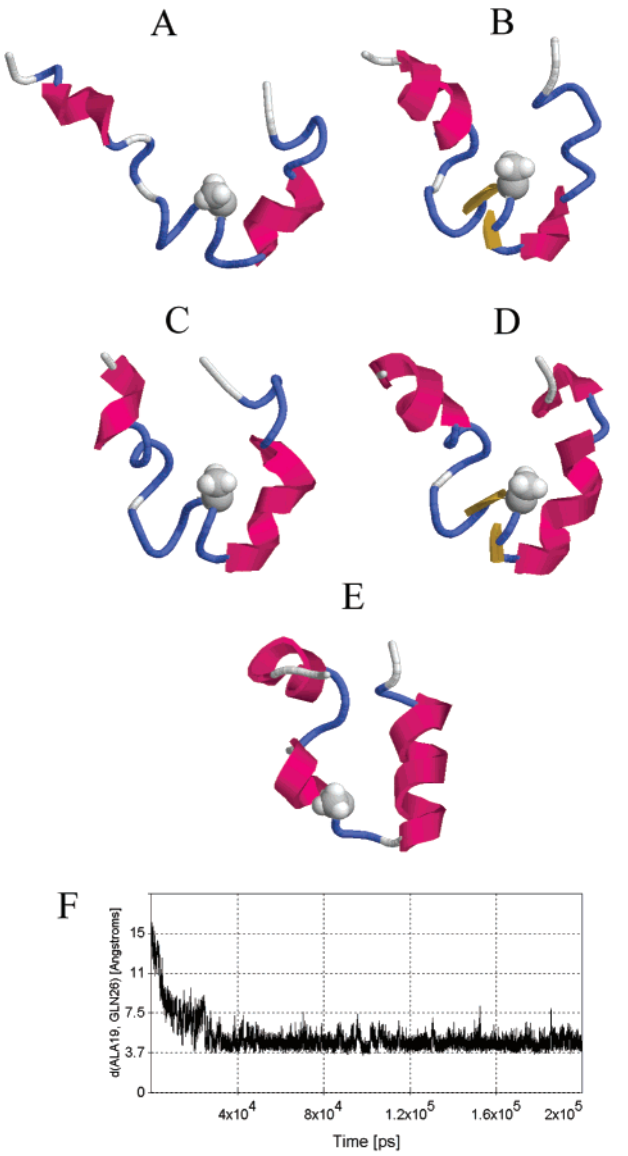

FIGURE 3: Five schematic structure snapshots (A−D) obtained from the 200 ns all-atom implicit solvent trajectory for the villin headpiece. The pictures are obtained respectively at 38, 65, 110, and 164 ns, while the picture at the end represents the native structure (added for comparison). Only Ala19 is shown in a space-filling representation. Panel E displays the native fold. Panel F represents the time-dependence of the α-carbon distance between Ala19 and Gln26.

precisely how the 26−30 helix turn is approached by the hydrophobic Ala19 (in a space filling representation), which thus becomes the protector of the nucleus for the helix formation until the native-core residues Phe11 and Phe18 assume the protective role (Figure 3E). This behavior is corroborated by monitoring the Ala19−Gln26 α-carbon distance obtained from the all-atom simulation with implicit solvent (Figure 3F).

One rationalization of the treatment by the FM of third-body stabilization is the interpretation that the desolvation of a hydrogen bond lowers its enthalpy, compensating for the unfavorable burial of the polar intervening moieties, in the manner of Makhatadze and Privalov (*42*). An alternative approach would be to attribute the stabilization to an increase of the kinetic barrier for solvent attack; this interpretation gives very similar results (*29*). At present, the relative

contributions of the two phenomena are not known.

Well beyond the 1-μs range, the FM reveals a second and more important cycle of compaction−expansion−compaction with a ~5.7-μs period (Figure 2F−H). This cycle may be understood by inspection of Figure 2B−D as corresponding to an exchange of protecting roles that entails a vast structural rearrangement of the molecule (e.g., the Ala19/Leu21 protectors of the C-terminal helix are replaced by the more distant Met1/Leu2 protectors). (The hydrophobic N-terminus is finally exposed to bulk solvent in the native fold (*14*, *15*) at a considerable free energy expense.)

Furthermore, Figure 2 reveals that the high-frequency pattern of compaction−expansion occurs because the exchange of protective roles between the Ala19−Gln26−Lys30 and the Leu21−Gln26−Lys30 correlation accompanies the rapid switching of the Met1−Gln26−Lys30 correlation with the Leu2−Gln26−Lys30 correlation. The high-frequency fluctuations in the energy and radius of gyration are built onto a coarser pattern associated with the large-scale switching back and forth from Ala19/Leu21 to Met1/Leu2 as protectors of the C-terminal helix, a motif that invariably relies on the large-scale context for stability. Although the finer structure of the initial exchange might be detectable at the all-atom level (*15*), the coarser pattern remains to be confirmed by future all-atom simulations or experiments or by site-directed mutations such as Met1Ser and Leu2Ser at the N-terminus. While stabilizing the N-terminus helix by eliminating hydrophobic-polar mismatches, such mutations should prevent the N-terminus helix from providing a hydrophobic template upon which the locally unfavorable C-terminus helix can safely nucleate.

Once the C-terminal helix is completed (Figure 2D,F), it is partially stabilized through the following internal three-body correlations: Leu23−Gln26−Lys30, Leu23−Gln27−Lys30, and especially, Trp24−Gln27−Lys30. These correlations provide an increase in stability that still is insufficient to overcome the entropic advantage of the random coil. The correlations place the C-terminal helix 0.4 kcal/mol above the random coil, according to the parametrization given in ref *17*. It is not until the core residues Phe11 and Phe18 become engaged in three-body correlations that the C-terminal helix becomes stabilized (cf. Figure 2).

The previous discussion indicates that specific hydrophobic residues not belonging to the native core take turns as protectors—desolvators—of the hydrogen bonds that form in the nucleus of the C-terminus helix. This dynamic behavior cannot be inferred solely by inspection of the native structures because the desolvation of the hydrogen bonds in the C-terminus helix ultimately relies only on hydrophobic residues of the native core. Furthermore, it provides a dynamical justification for hydrophobic residues that are partially exposed to the solvent and thereby are seemingly misplaced in the native fold.

Probing Dynamic Predictions. The critical dynamical role of Ala19 found to arise in the helix-nucleation event in all 55 successful runs for the wild type may be probed experimentally with site-directed mutagenesis. This loop residue has not been identified as belonging to the native core, although it may be shown to be weakly coupled to the core (*14*). In view of the crucial role we ascribe for Ala19 in the folding process, the mutations Ala19Val and Ala19Ser would lead to striking and persuasive results. Because Ala19 is a hydrophobic loop residue, the Ala19Val replacement would seem to be thermodynamically unfavorable, making that site in the 12−23 loop even more hydrophobic, by as much as 2.6−3 kcal/mol. However, site 19 plays a crucial dynamic role in the scaffolding of the C-terminal helix. Our analysis, based on the mutant's in vitro folding by the FM, predicts this mutant to be a far more expedient folder than the wild type. The enhanced scaffolding of the kernel for C-terminal helix initiation in the mutant enhances the kinetics of folding enough to more than compensate for its perturbation of the native state stability. Mutations of this kind potentially yield ϕ-values less than 0 or greater than 1 (see Figure 4), depending on whether the native fold is destabilized (most likely in this instance) or stabilized by the mutation. For more standard mutations, that destabilize both transition state and native fold, a ϕ-value between 0 and 1 is a measure of a residue's energetic involvement in the transition state (*2*, *16*, *19*). Val19, with its larger hydrophobic area, is a better desolvator of the Gln26−Lys30 and Gln27−Lys30 hydrogen-bonded pairs than is Ala19.

In sharp contrast, as illustrated by a typical run displayed in Figure 5, the mutation Ala19Ser is predicted to be

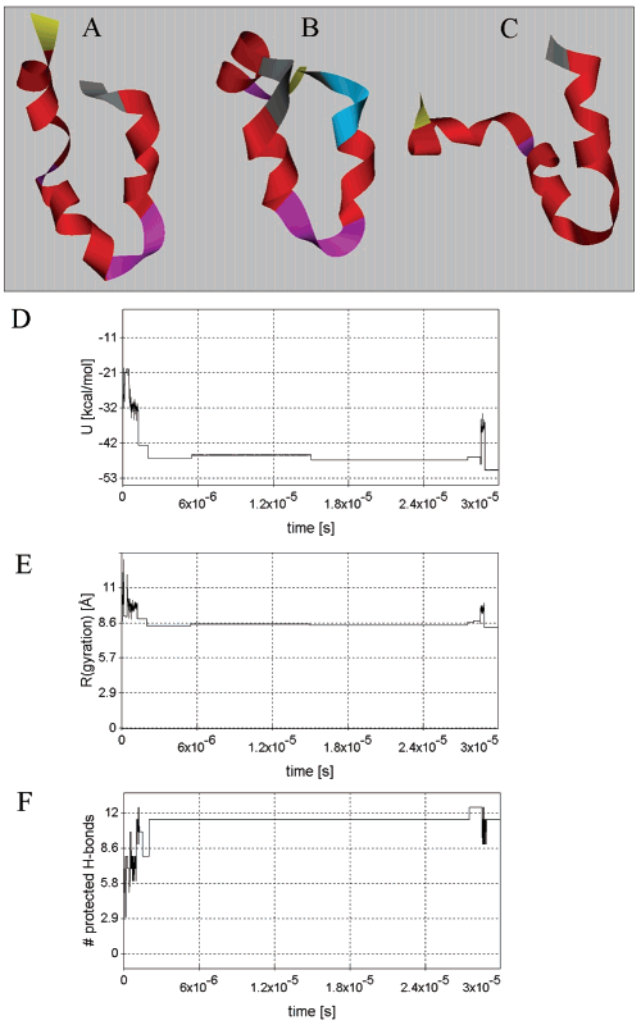

FIGURE 4: Typical FM run for the Ala19Val mutant made up of 2 $\times$ 10^5 iterations. The three schematic structure snapshots (A−C) are taken at 1.5, 4, and 18 μs, respectively. The plots in panels D−F represent, respectively, the time-dependence of the intramolecular correlated internal energy, the radius of gyration, and the number of protected hydrogen bonds.

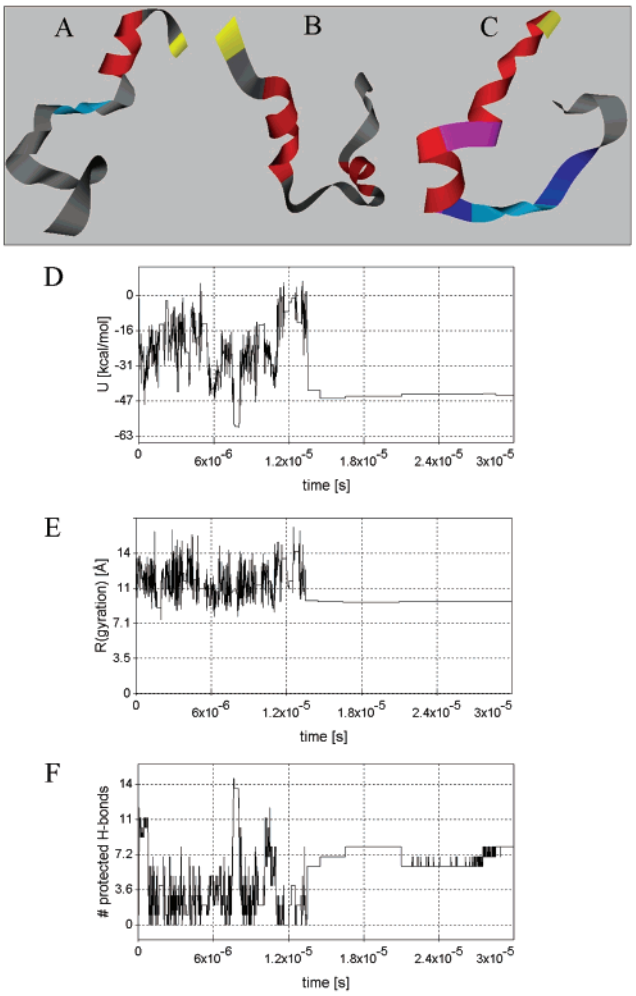

FIGURE 5: Typical FM run for the Ala19Ser mutant involving 2×10^5 iterations. The three schematic structure snapshots (A−C) are taken at 6, 7.5, and 18 μs, respectively. The plots in panels D−F represent, respectively, the time-dependence of the intramolecular correlated internal energy, the radius of gyration, and the number of protected hydrogen bonds.

deleterious for formation of the C-terminal helix because its large-scale context fails to stabilize the unique initiation kernel of this helix: the looped region (*19*−*21*) forms prematurely in simulations for this mutant. This behavior has been detected in all the 78 runs performed with the FM for the Ala19Ser mutant.

Figures 4 and 5 display our predictions and designed probes for the role of large-scale context in folding. The mutation Ala19Val converts the loop residue into such a good desolvator of the scaffolding hydrogen bond Gln26−Lys30 that the kernel for growth of the C-terminal helix survives throughout the entire simulation and does not require an exchange of protective roles or a major structural rearrangement to bring into proximity the N-terminus and the initial C-terminal helix turn. Thus, the net result of this mutation is a molecule that folds in FM simulations 1 order of magnitude faster than the wild type (Figure 4). This behavior has been found without exception in all 50 runs performed for the Ala19Val mutant.

On the other hand, Ala19Ser in the simulations folds reproducibly into a less stable structure whose C-terminal helix is absent (Figure 5). In this mutant, given the impossibility of sustaining the kernel for C-terminal helix growth until the N-terminus intervenes as a protector, the mutated protein simply uses the hydrophobic residues in the unstructured C-terminus region as desolvators of the N-terminal helix (Figure 5A−C). The required desolvation takes place by engaging the previously helical (*30*−*34*) sequence LysLysGluLysGly into a loop needed to bring the C-terminus into proximity with the initial helix (Figure 5C). In this mutant, the large-scale context reinforces the local propensity of formation of the N-terminal helix (cf. Figures 1 and 5).

CONCLUSION

Through a comparison between coarse-grained and all-atom simulations, we have identified critical features of the villin headpiece whereby tertiary context determines secondary structure and overrules simplistic local propensities. Such comparisons serve to validate the coarse-grained simulations, and more importantly, provide insights into features of the folding process that are often opaque in the full complexity of all-atom simulation. While the formation of a hydrophobic core is a recognized driving force in protein folding, the establishment of specific three-body correlations provides a more precise, detailed picture at the molecular level of how this collapse stabilizes essential structures within the core (*43*). Thus, in the villin headpiece, we have identified cycles of scaffolding patterns that eventually induce the formation of a helix with very low local propensity. Further, we identify site mutations that should have dramatic kinetic consequences, unexpected on the basis of their extent of participation in the final native core. This result is compatible with recent findings revealing a high resilience of the folding process to point mutations in the native core (*44*).

The results are statistically significant insofar as 55 out of 81 coarse-grained FM simulations of the wild type produced only stable folds within 5 Å rmsd from the native, clearly quenched the structural fluctuations within a 1-μs (∼1/10-folding time) time interval and produced cycles of structure compaction and expansion with period dispersion lower than 10% of the mean value and invariably within 10% rmsd of those encountered in the all-atom explicit-solvent trajectory. That is, the FM, with its artificial identification of a time scale, finds the native structure about 10 times more efficiently, in terms of the molecular time scale, than does an efficient all-atom Langevin dynamics simulation. The latter is, in turn, considerably more efficient than traditional molecular dynamics (*36*, *40*). On the other hand, the predicted behavior for the dynamically crucial mutations Ala19Val (a folder 1 order of magnitude faster) and for Ala19Ser (a nonfolder) were found to be invariably reproducible in all runs performed.

A recently performed equilibrium analysis (*44*) has persuaded us that the villin headpiece subdomain exhibits a core resilience comparable to a larger protein that is able to sustain organization on a larger scale. The mutational perturbations F47L, F51L, and F58L at the core of the villin headpiece prove to be destabilizing as evidenced by a decrease in the thermal unfolding midpoint monitored by CD (*44*). Nevertheless, all three mutants and even the F47, 51L double mutant retain the wild-type fold, as corroborated by CD and 1-D-NMR spectroscopy, although the double mutant fold might be closer to a molten globule, as evidenced by the broadening of downfield NMR resonances.

However, mutational kinetic experiments providing a ϕ-value analysis (*1*, *2*) will be necessary to elucidate whether the core formation is the decisive feature that commits the protein to fold or, rather, whether this commitment stems from the transient protecting roles played by the loop hydrophobic residues in the vicinity of Ala19, as we suggest in this paper. This point is especially important given that more often than not high ϕ-values are not associated with the core residues but rather with residues in crucial turn or loop regions (*44*).

REFERENCES

1. Muñoz, V., and Eaton, W. A. (1999) A simple model for calculating the kinetics of protein folding from three-dimensional structures, *Proc. Natl. Acad. Sci. U.S.A. 96*, 11311−6.
2. Fersht, A. (2000) Transition state structure as a unifying basis in protein folding mechanisms: contact order, chain topology, stability and the extended nucleus mechanism, *Proc. Natl. Acad. Sci. U.S.A. 97*, 1525−1529.
3. Nymeyer, H., Socci, N. D., and Onuchic, J. N. (1998) Landscape approaches for determining the ensemble of folding transition states: success and failure hinges on the degree of frustration, *Proc. Natl. Acad. Sci. U.S.A. 95*, 5921−5928.
4. Lacroix, E., Viguera, A. R., and Serrano, L. (1998) Elucidating the folding problem of alpha helices: local motifs, long-range electrostatics, ionic strength dependence and prediction of NMR parameters, *J. Mol. Biol. 284*, 173−181.
5. Baldwin, R., and Rose, G. D. (1999) Is protein folding hierarchic? II. Folding intermediates and transition states, *Trends Biochem. Sci. 24*, 77−83.
6. Bashford, D., Karplus, M., and Weaver, D. (1990) in *Protein folding* (Gierasch, L. M., and King, J., Eds.) pp 283−290, Am. Assoc. Adv. Sci., Washington.
7. Park, K., Vendruscolo, M., and Domany, E. (2000) Towards an energy function for the contact map representation of proteins, *Proteins 40*, 237−248.
8. Minor, D. L., and Kim, P. S. (1996) Context-dependent secondary structure formation of a designed protein sequence, *Nature 380*, 730−734.
9. Vendruscolo, M., and Domany, E. (1998) Efficient dynamics in the space of contact maps, *Folding Des. 3*, 329−336.
10. Forge, V., Hoshino, M., Kuwata, K., Arai, M., Kuwajima, K., Batt, C. A., and Goto, Y. (2000) Is folding of β-lactoglobulin nonhierarchic? Intermediate with nativelike β-sheet and nonnative α-helix, *J. Mol. Biol. 296*, 1039−1051.
11. Fernández, A., Colubri, A., and Berry, R. S. (2001) Topologies to geometries in protein folding: Hierarchical and nonhierarchical scenarios, *J. Chem. Phys. 114*, 5871−5887.
12. Fernández, A., Colubri, A., and Berry, R. S. (2000) Topology to geometry in protein folding: β-lactoglobulin, *Proc. Natl. Acad. Sci. U.S.A. 97*, 14062−14066.
13. Kuwata, K., Shastry, R., Cheng, H., Hoshino, M., Batt, C. A., Goto, Y., and Roder, H. (2001) Structural and kinetic characterization of early folding events in β-lactoglobulin, *Nat. Struct. Biol. 8*, 151−155.
14. McKnight, C. J., Matsudaira, P. T., and Kim, P. S. (1997) NMR structure of the 35-residue villin headpiece subdomain, *Nat. Struct. Biol. 4*, 180−184.
15. Duan, Y., and Kollman, P. A. (1998) Pathways to a protein folding intermediate observed in a 1 microsecond simulation in aqueous solution, *Science 282*, 740−744.
16. Plaxco, K. W., Simons, K. T., and Baker, D. (1998) Contact order, transition state placement and the refolding rates of single-domain proteins, *J. Mol. Biol. 277*, 985−994.
17. Fernández, A. (2001) Conformation-dependent environments in folding proteins, *J. Chem. Phys. 114*, 2489−2502.
18. Daggett, V., Li A., Itzhaki, L. S., Otzen, D. E., and Fersht, A. R. (1996) Structure of the transition state for folding of a protein derived from experiment and simulation, *J. Mol. Biol. 257*, 430−440.
19. Martínez, J. C., Pisabarro, M. T., and Serrano, L. (1998) Obligatory steps in protein folding and the conformational diversity of the transition state, *Nat. Struct. Biol. 5*, 721−729.
20. Alm, E., and Baker, D. (1999) Prediction of protein folding mechanisms from free energy landscapes derived from native structures, *Proc. Natl. Acad. Sci. U.S.A. 96*, 11305−11310.
21. Baldwin, R. L., and Rose, G. D. (1999) Is protein folding hierarchic? I. *Trends Biochem. Sci. 24*, 26−37.
22. Vendruscolo M., Paci, E., Dobson, C. M., and Karplus, M. (2001) Three key residues form a critical contact network in a protein folding transition state, *Nature 409*, 641−645.
23. Fernández, A., Kostov, K., and Berry, R. S. (1999) From residue matching patterns to protein folding topographies: General model and bovine pancreatic trypsin inhibitor, *Proc. Natl. Acad. Sci. U.S.A. 96*, 12991−12996.
24. Fernández, A., Kostov, K., and Berry, R. S. (2000) Coarsely resolved topography along protein folding pathways, *J. Chem. Phys. 112*, 5223−5229.
25. Fernández, A., and Berry, R. S. (2000) Self-organization and mismatch tolerance in protein folding. General theory and an application, *J. Chem. Phys. 112*, 5212−5222.
26. Fernández, A. (2001) Cooperative walks in a cubic lattice: Protein folding as a many-body problem, *J. Chem. Phys. 115*, 7293−7297.
27. Thornton, J. (1992) in *Protein Folding* (Creighton, T. E., Ed.) pp 59−63, Freeman, New York.
28. Fernández, A. (2000) Protein design from in silico dynamic information: the emergence of the "turn-dock-lock" motif, *Protein Eng. 15*, 1−6.
29. Fernández, A., Colubri, A., and Berry, R. S. (2002) Three-body correlations in folding proteins: The origin of cooperativity, *Physica A 307*, 235−259.
30. Ponder, J. W., Rubenstein, S., Kundrot, C., Huston, S., Dudek, M., Kong, Y., Hart, R., Hodsdon, M., Pappu, R., Mooij, W., and Loeffler, G. (1999) TINKER: Software tools for molecular design, version 3.7, Washington University, St. Louis.
31. Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Jr., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W., and Kollman, P. A. (1995) A second generation force field for the simulation of proteins, nucleic acids, and organic molecules, *J. Am. Chem. Soc. 117*, 5179−15197.
32. Ooi, T., Oobatake, M., Nemethy, G., and Scheraga, H. A. (1987) Accessible surface areas as a measure of the thermodynamic parameters of hydration of peptides, *Proc. Natl. Acad. Sci. U.S.A. 84*, 3086−3090.
33. Ramstein, J., and Lavery, R. (1988) Energetic coupling between DNA bending and base pair opening, *Proc. Natl. Acad. Sci. U.S.A. 85*, 7231−7235.
34. Allen, M. P., and Tildesley, D. J. (1987) *Computer simulation of liquids*, Oxford University Press, Oxford.
35. Pastor, R. W., and Karplus, M. (1988) Parametrization of the friction constant for stochastic simulations of polymers, *J. Phys. Chem. 92*, 2636−2641.
36. Shen, M.−y., and Freed, K. F. (2000), All-atom fast protein folding: The villin headpiece, *Proteins 44*, 439−445.
37. Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A., and Haak, J. R. (1984) Molecular dynamics with coupling to an external bath, *J. Chem. Phys. 81*, 3684−3690.
38. Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1992) *Numerical Recipes*, Cambridge University Press, Cambridge.
39. Kabsch, W., and Sander, C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, *Biopolymers 22*, 2577−2637.
40. Shen, M.-y., and Freed, K. F. (2002) Long Time Dynamics of Met-Enkephalin: Comparison of Explicit and Implicit Solvent Models, *Biophys. J. 82*, 1791−1808.
41. Krantz, B., Moran, L. B., Kentsis, A., and Sosnick, T. R. (2000) D/H amide isotope effects reveal when hydrogen bonds form during protein folding, *Nat. Struct. Biol. 7*, 62−71.
42. Makhatadze, G., and Privalov, P. (1995) Energetics of protein structure, *Adv. Protein Chem. 47*, 307−425.
43. Garcia, A. E., and Sanbonmatsu, K. Y. (2001) Exploring the energy landscape of a beta-hairpin in explicit solvent, *Proteins 42*, 345−354.
44. Frank, B. S., Vardar, D., Buckley, D. A., and McKnight, C. J. (2002) The role of aromatic residues in the hydrophobic core of the villin headpiece subdomain, *Protein Sci. 11*, 680−687.

BI026510I