Common Structural Transitions in Explicit-Solvent Simulations of Villin Headpiece Folding

Peter L. Freddolino and Klaus Schulten*
Center for Biophysics and Computational Biology, University of Illinois at Urbana-Champaign, Urbana, Illinois

ABSTRACT   Molecular dynamics simulations of protein folding can provide very high-resolution data on the folding process; however, due to computational challenges most studies of protein folding have been limited to small peptides, or made use of approximations such as Gō potentials or implicit solvent models. We have performed a set of molecular dynamics simulations totaling >50 μs on the villin headpiece subdomain, one of the most stable and fastest-folding naturally occurring proteins, in explicit solvent. We find that the wild-type villin headpiece reliably folds to a native conformation on timescales similar to experimentally observed folding, but that a fast folding double-norleucine mutant shows significantly more heterogeneous behavior. Along with other recent simulation studies, we note the occurrence of nonnative structures intermediates, which may yield a nativelike signal in the fluorescence measurements typically used to study villin folding. Based on the wild-type simulations, we propose alternative approaches to measure the formation of the native state.

INTRODUCTION

Many recent experimental studies of protein folding have focused on the characterization of fast-folding peptides, which approach the hypothesized protein-folding speed limit (1,2). In the process, a variety of small proteins folding on microsecond or even submicrosecond timescales has been identified and characterized (3–6). Atomistic molecular dynamics (MD) simulations could substantially assist in understanding the folding of small proteins by providing highly detailed data on the intermediate structures in the folding funnel, and on the transitions between them. Simulations of protein folding were generally forced until recently to choose between the use of an implicit solvent model, which may lead to incorrect properties for folding intermediates (7,8), or the use of an explicit solvent model, which requires an unfeasible amount of computing power to simulate a complete, microsecond-timescale folding trajectory. Over the past few years, several studies have made use of explicit solvent replica-exchange simulations to analyze the thermodynamics of folding of several small peptides such as Trpcage (9,10), Met-enkephalin (11), and the C-terminal β-hairpin of protein G (12). More recently, it has become possible to obtain trajectories on the folding timescales of fast-folding proteins using explicit solvent MD in a few months of wallclock time (e.g., (13)), now allowing direct comparison of MD and experimental results.

The chicken villin headpiece subdomain is one of the best-characterized fast-folding proteins; most experimental and computational work has focused on a 35-residue fragment (HP-35), which is made up of three α-helices (14). HP-35 has the distinction of being a small, naturally occurring protein, which nevertheless folds independently to a native

state (14), hence its usefulness as a model system. HP-35 has been shown to have bi-exponential folding kinetics with a slow phase (presumably corresponding to folding) characterized by a (4.3 μs)$^{-1}$ rate (3) at 300 K, or (7.4 μs)$^{-1}$ (15) at 323 K, depending on the method used to measure folding. In addition, the related HP-36 (equivalent to HP-35 with an N-terminal methionine residue) was estimated by nuclear magnetic resonance (NMR) lineshape analysis to have a folding rate of (45.5 μs)$^{-1}$ at 306 K (16), albeit with a large uncertainty due to the methods used, and a variant with an N-terminal cysteine shown through measurements of a tryptophan triplet state lifetime (17) to have a folding rate of (3.16 μs)$^{-1}$ at 313 K (17). (Note that although temperatures are noted for the preceding folding rates, the folding rate of villin changes very little with respect to temperature in the 300–350 K range (3).) Importantly, a fast-folding mutant with two lysine residues replaced with norleucine (Nle) shows similar bi-exponential kinetics with a folding rate of (0.7 μs)$^{-1}$ at 300 K (4). It is hypothesized that the accelerated folding, and increased stability, of the K65Nle/K70Nle mutant (henceforth NLE) are due to reduced frustration from the removal of unsolvated charged groups and repulsive charge-charge interactions (4).

Because of its size, simple structure, and rapid folding, HP-35 has been the subject of a wide variety of atomistic in silico folding studies, using both implicit (18–23) and explicit (24–27) solvent. Although several of these studies did indeed produce trajectories folding the villin headpiece to the native state, multiple different folding mechanisms have been proposed in the process. Among recent representatives, Lei and Duan (20) and Lei et al. (21) propose that the primary free energy barrier during folding arises during formation of helices II and III, which occurs before helix I formation. Yang et al., in contrast, propose that the transition state ensemble contains structures with nativelike

conformations of helices I and II (22), and Zagrovic et al. (25) observed a majority of (short) trajectories trapped in an intermediate after secondary structure formation where F76 forms incorrect contacts with the phenylalanine residues of the hydrophobic core, an observation also made in an early 1-μs villin trajectory (24).

With the exception of an impressive series of simulations by Ensign et al. on NLE (26) and a massively parallel study with trajectories only a few tens of nanoseconds in length (27), all MD simulations of the villin headpiece which have reached native states have used implicit solvent models, and as noted above, thus run the risk of altered relative stabilities in intermediate states. We report here a series of three ~7-μs folding simulations of the N68H mutant of HP-35 (henceforth WT) and an additional six simulations of NLE, lasting a total of >50 μs. All three WT simulations reach a native state after 5.6–8.2 μs of simulation; two of the three remain stable in this state for at least one additional μs, while the third instead enters a stable near-native state (average C_α-RMSD to the crystal structure under 2.5 Å). In addition, all three trajectories show very similar pathways at the final stages of folding, although there is substantial variability in the path followed up to that point, in agreement with energy landscape theory (28). In all three cases, we observed early secondary structure formation and hydrophobic collapse, followed by the presence of a varied array of metastable, interconverting conformations, including one (referred to as the flipped state) with correct secondary structure but incorrect relative orientations of the helices. The final transition to the native state occurs after the helices in the flipped state dissociate from each other to yield an open intermediate with a very high radius of gyration; after closure of this intermediate to a nativelike structure, the protein rapidly reaches a native state. More variation is observed for the NLE trajectories, where two trajectories fold to native or near-native states and the others become trapped in misfolded states. Based on the novel folding pathway observed in our simulations, we propose experimental observables, which could be used to test our proposed folding mechanism and better monitor the folding process of the villin headpiece subdomain; we also propose mutations that would be expected to accelerate folding by reducing energetic frustration due to nonnative interactions.

METHODS

Molecular dynamics simulations were performed with a development version of NAMD 2.7 (29). Simulations were performed using Langevin dynamics with a damping constant of 0.1 ps^{-1}; constant pressure simulations also used a Nosé-Hoover Langevin piston barostat (29) with a pressure of 1.0 atm, period of 200.0 fs, and decay rate of 100.0 fs, with isotropic cell scaling. Multiple time stepping was employed, with an integration timestep of 2.0 fs, short-range forces evaluated every timestep, and long-range electrostatics evaluated every three timesteps. Cutoffs for short-range interactions are shown in Table 1; long-range electrostatics were calculated using particle-mesh Ewald with 64 grid points along each axis. All bonds involving hydrogen in the protein were constrained using RATTLE (30), and water geometries were maintained using SETTLE (31). A temperature

TABLE 1   Summary of simulations performed, including notation for each simulation that will be used throughout the text

Name	Variant	Duration (ns)	Conformation	Cutoff (Å)	Ensemble	Min. C_α-RMSD (Å)
WT-NAT	WT	280	Crystal	7.0/8.0	NVT	0.46
NLE-NAT	NLE	250	Crystal	7.0/8.0	NVT	0.47
WT-FOLD1	WT	7010	Extended	7.0/8.0	NVT	0.46
WT-FOLD2	WT	8892	Extended	10.0/12.0	NPT	0.81
WT-FOLD3	WT	7584	Extended	7.0/8.0	NVT	0.45
NLE-FOLD1	NLE	3000	Extended	7.0/8.0	NVT	0.49
NLE-FOLD2	NLE	3000	Extended	7.0/8.0	NVT	2.83
NLE-FOLD3	NLE	8126	Extended	7.0/8.0	NVT	1.22
NLE-FOLD4	NLE	4330	Denatured	7.0/8.0	NVT	3.46
NLE-FOLD5	NLE	5410	Denatured	7.0/8.0	NVT	3.16
NLE-FOLD6	NLE	7496	Denatured	7.0/8.0	NVT	3.63
Total		55.4 μs				

The "Cutoff" column gives the switching distance and cutoff used.

of 300 K was assumed unless otherwise noted. The CHARMM22 force field with CMAP corrections (32) was used for the protein and ions. Coordinates were written once every 6.0 ps.

The initial HP-35 structure was taken from PDB code 1YRF (33), with arbitrarily chosen rotamers for side chains with multiple positions. This structure is referred to as the crystal structure throughout. The initial structure was placed in a box of 9607 TIP3P water molecules and neutralized with 200 mM NaCl using VMD (34). The structure was then subjected to 6000 steps of conjugate gradient minimization, and equilibrated for 200 ps with harmonic restraints applied to all protein heavy atoms and for 2.0 ns with no restraints; these equilibrations occurred at constant pressure and with a Langevin damping constant of 5.0 ps^{-1} and involved ~30,000 atoms. The resulting equilibrated structure is referred to as "Crystal" in Table 1, and its periodic cell size (66.25 Å along each axis) was assumed for all NVT WT simulations. A fully extended starting structure for folding simulations was generated from the WT villin sequence by setting all backbone (ϕ, ψ) angles to (−135,135) and then equilibrating the resulting structure for 100 ps.

The reference structure of the K65Nle/K70Nle (4) mutant (NLE) was generated by appropriately modifying the same structure as used for the WT simulations. Standard CHARMM22 parameters for aliphatic groups were used for the terminal portion of the norleucine side chain. The NLE structure was then equilibrated using the same protocol as the WT crystal structure, and the resulting periodic cell size (66.30 Å) was assumed for all constant volume NLE simulations. For several runs using the double norleucine mutant, a thermally denatured conformation was used instead (see Table 1). This conformation was obtained by beginning simulation from the crystal structure and running for 100 ns at 450 K and 60 ns at 500 K. The denatured structure has no secondary structure (as calculated by STRIDE (35)), and has a backbone atom RMSD of 12.6 Å to the crystal structure, and Q_{res} (36) of 0.1439.

Clustering analysis was performed using the g_cluster program of GROMACS 3.3 (37) with the GROMOS clustering method (38). Frame-frame RMSDs were calculated using all heavy atoms except those that are chemically identical to another atom in the same residue.

RESULTS

To confirm the stability of the villin headpiece crystal structure with the simulation parameters employed, simulations of 280 ns and 250 ns were performed starting from the WT and NLE crystal structures, respectively (see Table 1 for all

simulation notation). In both cases, the crystal structures were stable over the simulated timescales, with average ($\pm$ standard deviation) C_α-RMSD values to the crystal structure of 1.54 ± 0.37 and 1.61 ± 0.44 Å for the WT and NLE cases, respectively.

WT villin folding simulations

We performed three folding simulations of WT villin headpiece starting from a fully extended conformation (see Table 1); all three trajectories reach folded conformations within 5.6–8.2 μs. Here we define a single frame as folded (or in the native state) if its C_α-RMSD to the crystal structure is <2.0 Å, and treat the trajectory as having reached a folded state if its distributions of C_α-RMSD and Q_{res} (36) values reach those of the crystal structure simulation. These values are shown along with the secondary structure for each trajectory in Fig. 1; a set of other quantities of interest (such as radius of gyration) are shown in Fig. S1 in the Supporting Material. In each case, the simulation was continued for 0.5–1.0 μs after the folding event to assess the stability of the folded structure that was reached. WT-FOLD1 and WT-FOLD3 remain folded once they reach a native state, with the exception of a short return to a near-native state in WT-FOLD1 at ~6.6 μs. WT-FOLD2 instead briefly reaches the native state at ~8.2 μs and then settles in a near-native state for the duration of the simulation (although even during the latter period many individual conformations satisfy our native state criteria). The deviations between the near-native state in WT-FOLD2 and the crystal structure are mostly in the N- and C-terminal residues; exclusion of the first and last residues from C_α-RMSD calculations (as in (20)) yields an average RMSD of ~1.65 Å and minimum RMSD of 0.55 Å for the near-native state.

All three WT folding trajectories show a similar order of events along the folding pathway: the protein undergoes a rapid collapse over the first 1–2 μs, although the solvent-exposed hydrophobic surface area does not reach nativelike values (see Fig. S1); in addition, during this collapse, helices I and III form (accompanied, in some cases, by helix II). The protein then has a long dwell time (~5 μs) in a series of interconverting nonnative states. To identify significant intermediate states, clustering analysis was performed on each trajectory (see Methods). The highly occupied clusters and plots of the cluster occupied throughout the WT simulations are shown in Fig. S2. The complete folding trajectories of WT-FOLD1, WT-FOLD2, and WT-FOLD3 are shown in three respective movies supplied in the Supporting Material.

In each WT trajectory, a long-lived intermediate state eventually forms where native or nearly native secondary structure has formed, but helix I is flipped relative to the rest of the protein, such that the N-terminus is directed away from the C-terminus, and the hydrophobic face of helix I is rotated away from the protein's center (clusters 1, 3, and 6 in WT-FOLD1; 2 and 14 in WT-FOLD2; and 3 and 16 in
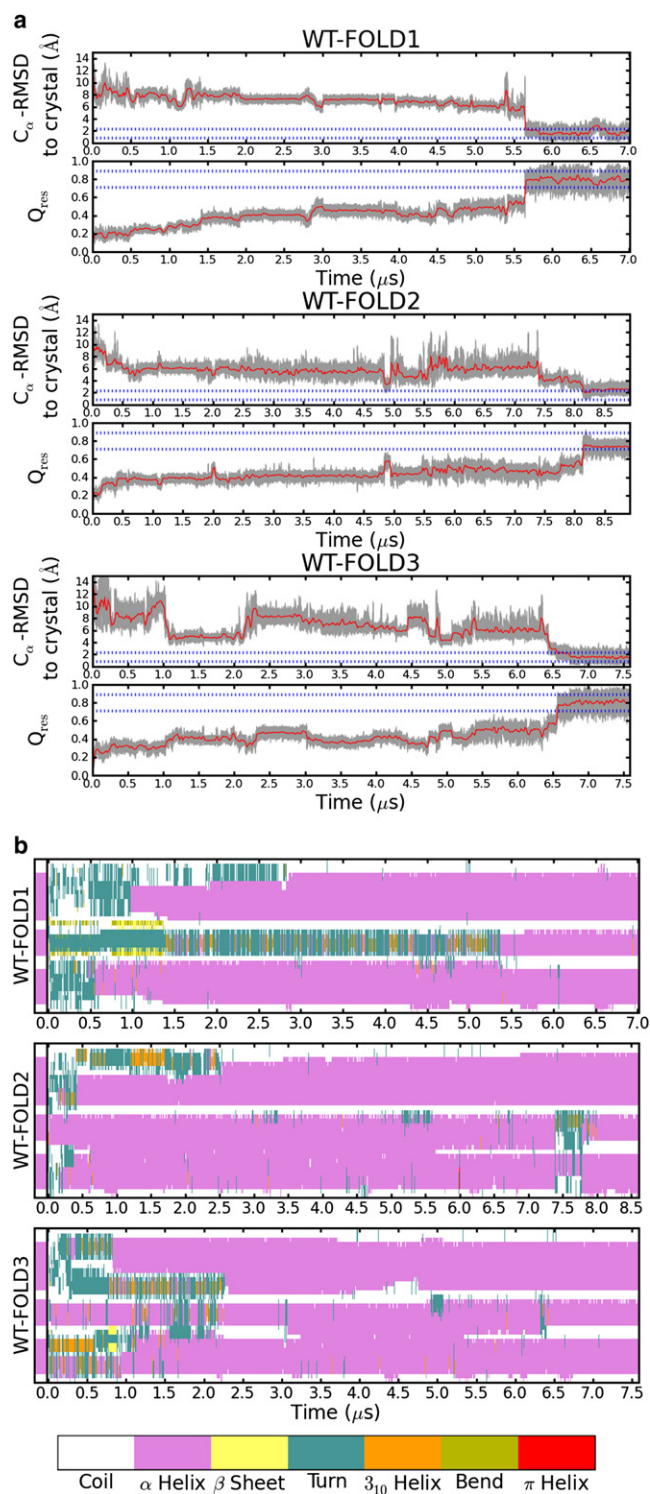

FIGURE 1 Progress of the WT folding simulations. (*a*) C_α-RMSD and Q_{res} relative to the 1YRF crystal structure; running averages over 30 ns are shown in red, and the range defined by the mean $\pm$ two standard deviations from simulation WT-NAT as blue bars. (*b*) Secondary structure throughout the WT folding trajectories. The secondary structure of the crystal structure is shown at left.

WT-FOLD3). This conformation is discussed in more detail below under Common Interactions in Folding Intermediates, and could represent either an obligatory intermediate occurring near the final stages of most folding trajectories, or a frequently encountered kinetic trap resulting from residual frustration in the protein. The nature of the long-lived intermediate varies somewhat between simulations: The intermediate state in WT-FOLD1 generally contains a well-packed hydrophobic core, whereas cluster 2 of WT-FOLD2 and cluster 3 of WT-FOLD3 are mostly open, although they interconvert rapidly with clusters (14 and 16, respectively, for WT-FOLD2 and WT-FOLD3), which have significantly more helix I-helix III contacts. A similar, shorter-lived semi-open state occurs in simulation WT-FOLD1 (clusters 8 and 13). Despite the presence of similar long-lived intermediates in all three cases, the paths followed by the three trajectories before the flipped state (and nature of other transient intermediates) are quite different. WT-FOLD1 shows mostly disordered (and short-lived) structures until the flipped intermediate is reached; WT-FOLD2 fluctuates rapidly between several conformations containing a single joint helix in place of helices I and II; and WT-FOLD3 progresses through a series of intermediate states with lifetimes on the order of hundreds of nanoseconds and a variety of two-helix and three-helix architectures (including several transient occurrences of the flipped conformation).

The eventual transition to a folded structure occurs immediately after an event in which the secondary structure elements lose the majority of their contacts with one another, and the protein is observed to take on an extended conformation (albeit with mostly intact secondary structure, with some transient losses of helical character in helix II and, in one case, the N-terminal half of helix I). In the process, a large number of nonnative contacts that had been present throughout the trajectory are replaced with native contacts due to correct relative arrangement of the helices, allowing rapid subsequent formation of the native state. The path followed by each trajectory during the transition is shown in Fig. S3. The appearance of such similar conformational transitions immediately before formation of the native state, and the microsecond or longer dwell time in a flipped conformation in all WT trajectories, strongly suggest that either the dissociation or reassociation event is the rate-limiting step in folding, corresponding to the presence of one long relaxation time in temperature-jump experiments (3).

NLE mutant folding simulations

Significantly more variability was observed in the case of the NLE mutant; while one trajectory (NLE-FOLD1) folded to a native state in ~2.5 μs, only one other trajectory, NLE-FOLD3, even reached a near-native state. NLE-FOLD3 showed near-native conformations after <1.3 μs, and again after 7.1 μs, but we do not consider the protein to have reached the native state since the distribution of folding observables does not match that from the crystal structure simulation NLE-NAT. C_α-RMSD values throughout the NLE trajectories are shown in Fig. 2, and a number of other properties are shown in Fig. S4 and Fig. S5. Clustering analysis was performed as for the WT simulations, with highly populated clusters presented in Fig. S6.

The trajectories that do reach native or near-native states, NLE-FOLD1 and NLE-FOLD3, do so through different mechanisms than seen for WT HP-35. Most notably, in neither case does a flipped conformation similar to the pre-folded state from the WT simulations form. Instead, the initial hydrophobic collapse and secondary structure formation occur with the formation of the correct relative positioning of secondary structure elements, likely because of the formation of a nativelike hydrophobic core much earlier in NLE-FOLD1 and NLE-FOLD3 than in the WT trajectories. Aside from the native conformation (cluster 1) observed in simulation NLE-FOLD1, a commonly observed near-native state occurred which differs only from the native state in that helix I is rotated further away from helix III (clusters 3, 5, and 10 in NLE-FOLD1, and 1 and 3 in NLE-FOLD3). This near-native conformation appears to be stabilized by hydrophobic effects, as it allows the norleucine residues to be more closely associated with the hydrophobic core than they are in the native state (see Fig. S7 *b*). The increased burial of the Nle residues is apparent in Fig. S4, which shows that the near-native conformation taken by NLE-FOLD3 has significantly more burial of the hydrophobic Nle residues than occurs in the crystal structure simulation (NLE-NAT).

Of the NLE trajectories that fail to even reach a near-native state, three (NLE-FOLD2, NLE-FOLD4, and NLE-FOLD5) of the four trajectories are dominated by conformations in which helix I and helix II are merged to form a single long helix (see Fig. S7 *c*), reminiscent of the dominant early intermediate from WT-FOLD2. The Nle mutations again appear to play a key role in stabilizing the favored conformations, as they show lower levels of exposed hydrophobic surface area than those observed in the crystal structure simulation (NLE-NAT). The remaining trajectory, NLE-FOLD6, forms nativelike secondary structure in helix I and helix III rather than an extended helix I/II, but still does not reach a native or near-native state, instead settling into a series of conformations where Nle65/Nle70 are surrounded by the residues at the center of the crystal structure hydrophobic core (F47, F51, F58, L61).

Folding rates and experimental observables

Temperature-jump experiments probing the quenching of W64 fluorescence by H68 have yielded folding rates at 300 K of $(4.3 \pm 0.6 \mu s)^{-1}$ for the WT (N68H) variant and $(0.73 \pm 0.03 \mu s)^{-1}$ for the NLE variant of HP-35 (3,4); on the other hand, recent experiments instead monitoring absorption at 1632 cm^{-1} yielded a folding rate of $(7.4 \mu s)^{-1}$
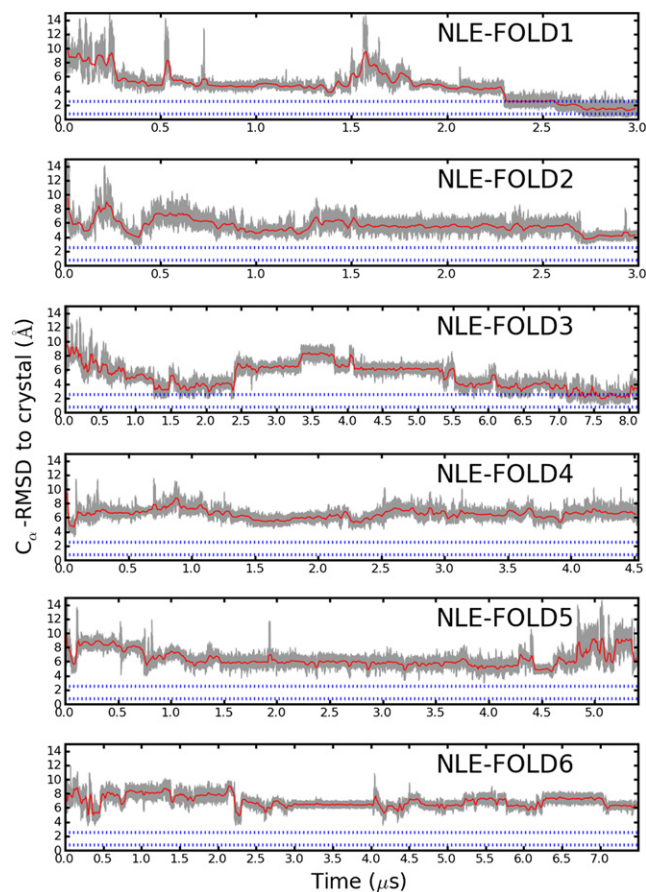
FIGURE 2 C_α-RMSD to the crystal structure throughout the NLE folding trajectories. The mean $\pm$ 2 standard deviations of values from simulation NLE-NAT are shown as dotted lines, and an average over 30 ns is shown as a red line.

at ~333 K (15). Variations between simulated and experimental folding rates can be attributed to factors including inaccuracies in the MD potential used, characteristics of the chosen starting structure for the simulations (26), the low effective viscosity of MD water models (39,40), and possible discrepancies between the relaxation time of experimental observables and the actual folding process (26). The use of a thermostat (and, in the case of WT-FOLD2, a barostat) could also alter the simulated folding kinetics, but a thermostat was necessary both to allow the use of multiple time-stepping (29) and to avoid temperature changes during the (nonequilibrium) protein folding events observed in the simulations. Langevin thermostats can, in some cases, slow such relaxation processes as protein folding due to the frictional term present in the Langevin equation (studied in the implicit solvent case in (41)). In this study, the kinetic effects of the thermostat were minimized by using a very low damping constant of 0.1 ps^{-1}, which did not perturb the dynamics of a TIP3P water box (data not shown); however, minor effects on folding kinetics still cannot be ruled out.

The folding times exhibited by WT villin in our simulations (5.6, 8.2, and 6.6 μs) are all longer than the experimentally

characterized folding time measured based on W64-H68 quenching; given that the folding times for individual molecules should be roughly exponentially distributed, the significance of the observed difference is unclear. The direction of the discrepancy is opposed to what would be expected from arguments based solely on water viscosity, and thus other factors must be investigated. One possible explanation is the starting conformation used in the folding simulations: all three WT trajectories began from a fully extended conformation, whereas the denatured state of villin is expected on the basis of NMR results to have helical secondary structure in the region of helices I and II (42,43), although it is not known whether the native helices are formed. We used an extended starting structure to avoid introducing a bias toward folding given that the nature of the unfolded structural ensemble is unknown; as noted above, 1–2 μs are required for the initial formation of secondary structure elements in the WT folding simulations. Correction of our observed folding times by ~2 μs yields folding times in good agreement with the experimentally observed range for the WT simulation (neglecting the effects of solvent viscosity), but it is unclear how appropriate such an ad hoc correction is. More information on the denatured state in villin folding kinetics experiments would be required to determine an appropriate time correction (and, perhaps more importantly, to provide a more realistic starting structure for further simulations). On the other hand, the folding times from simulation are in much better agreement with a rate of (7.4 μs)$^{-1}$ obtained through global measurement of amide I peak relaxation (15); the latter rate was obtained at a higher temperature than the simulations, but the villin folding rate appears fairly insensitive to temperature in the vicinity of 300–350 K (3,15).

In contrast to the WT trajectories, only one NLE simulation reached a native state, after ~2.5 μs; a second trajectory also reached a near-native state after 1.3 μs, but subsequently unfolded and again reached a near-native state after 7.1 μs. While more rapid folding events were observed for the NLE mutant than the WT protein, given the experimentally observed folding rate of (0.7 μs)$^{-1}$ and the duration of the NLE simulations most of the trajectories should have folded, even if a 1–2 μs offset was allowed for formation of a realistic denatured state.

Aside from the obvious possibility of force-field inaccuracies overstabilizing nonnative states for the NLE mutant (as in (13,44)), it is also useful in light of recent simulations of the NLE mutant by Ensign et al. (26) to ponder the experimental observable being used to measure folding rates. For both the WT and NLE variants of HP-35, folding rates have been measured based on the quenching of W64 fluorescence by protonated H68 (3,4). Since W64 and H68 reside one turn apart on helix III, it is possible that folding intermediates in which the first turn of helix III is formed are spectroscopically indistinguishable from the folded state (26). Indeed, in a large set of simulations of the NLE mutant from a variety of thermally denatured structures, Ensign et al. observed that estimating the folding time based on the population of nativelike

W64-H68 interactions systematically yielded shorter folding times than do structural observables, because of the presence of nonnative states with nativelike local structure around the fluorescent probe (26). The possibility that measurements of W64-H68 quenching underestimate the folding time of villin is indirectly supported by recent measurements of a slower folding rate after the relaxation of a nonspecific, global observable (15). It is also unclear why results from NMR line shape analysis (16) yield folding kinetics an order of magnitude slower than other methods, although the authors of that report note that their estimate is based on an extrapolation from experiments with high denaturant concentrations and thus carries a large uncertainty (16), and thus the disagreement may not be significant.

The distance between the centers of mass of the W64 and H68 ring systems throughout the simulations in this study are shown in Fig. S8. As with the results of Ensign et al. (26), we find that the W64-H68 distance reaches nativelike values faster than the protein actually folds. Formation of a native-like distance distribution generally coincides with formation of the first turn of helix III, and is present in the common flipped structure of the WT trajectories and in most highly occupied misfolded states from the NLE simulations. The W64-H68 distance thus takes on a nativelike distribution in <2.5 μs in all simulations except NLE-FOLD5.

Common interactions in folding intermediates

The identification of alternative experimental observables for the villin folding process based on MD simulations would allow testing of both the validity of our proposed folding mechanism and the appropriateness of quenching of W64 by H68 as an experimental surrogate for folding. Aside from fluorescence measurements, time-resolved infrared (IR) spectroscopy has been successfully applied to track protein folding (e.g., in (45)) and conformational equilibria, in some cases using proteins with ^{13}C labeled backbones to isolate the contributions of specific residues (46). In principle, contributions of other specific interactions such as salt bridges ought to be observable by monitoring peaks such as the asymmetric C-O stretch of isotope-labeled carboxylate-containing residues (47).

We thus sought to identify salt bridges that formed only either before or after the large opening transition between the flipped and folded states in the WT simulations. In all three WT folding trajectories, the prefolded state has helix I not only flipped relative to the rest of the protein but also rotated, such that its hydrophobic face is directed toward solvent or helix II and its hydrophilic face is closer to helix III (see Fig. 3 a), leading to the potential for consistent differences in salt bridging between the two states. Several such interactions could be identified that were formed only in the flipped structure of one or two trajectories, including D44-K65 and K48-E72; one such interaction, which was consistently formed in the flipped structure and other pre-folded conformations, but not the folded state, in all three simulations, involves D44 and K48 (see Fig. 3 b).

We thus propose measurement of the breakage of the D44-K48 salt bridge via isotope-labeled IR spectroscopy as an alternate probe of HP-35 folding kinetics. Based on the simulations presented here, this salt bridge should be significantly populated in the denatured state and in most major folding intermediates, but not in the native state. We would thus expect relaxation of the corresponding IR peak to occur at the same rate as, or more slowly than, other folding observables such as W64 fluorescence and 1632 cm^{-1} absorbance during temperature jump experiments. We also note that, in our simulations, the D44-R55 salt bridge only forms in the presence of native structure in the segment encompassing helices I and II. Thus, measuring the rate of its formation relative to other folding observables might allow the determination of whether or not native structure in helices I and II forms significantly before or in tandem with completion of folding.

In addition to the D44-K65 interaction noted above, which arises in ~50% of timesteps during the flipped state of simulation WT-FOLD1, a number of other salt bridges form transiently between residues K65, K70, and K73 and the acidic residues on helix I; some combination of D44, E45, and D46 interacting with these lysine residues occurs in the flipped state in all WT trajectories. Given that the apparent rate-limiting step in the wild-type folding trajectories involves dissociation of the flipped state, and that the flipped state appears to be stabilized by salt bridges between helix I and helix III, the mutation of D46, and possibly E45 to polar or basic residues, would be expected to accelerate folding. The removal of the frustration caused by these salt bridges may also be partly responsible for the faster folding of NLE, although the increased hydrophobicity of the Nle residues also likely encourages earlier formation of a correct hydrophobic core. We also note that D44 would be a poor choice for mutation because it forms a salt bridge with R55 in the native state.

Given that the flipped state also lacks nativelike interactions in the hydrophobic core, one expects that, as suggested by Ensign et al., stabilization of the F47-F58 contact would also accelerate folding (26). Indeed, we observe that involvement of F58 in the hydrophobic core is observed only in the native state, with the exception of a period during the early stages of WT-FOLD2, as seen in Fig. 3 c. This observation suggests that villin folding could also be monitored by substituting F58 with cyanophenylalanine, which can be used to measure the hydrophobicity of its environment (48).

DISCUSSION

The most significant finding of the WT simulations presented here is that all three simulations reached the native state in 5.6–8.2 μs, despite following different folding pathways. The early stages of folding for the three WT trajectories
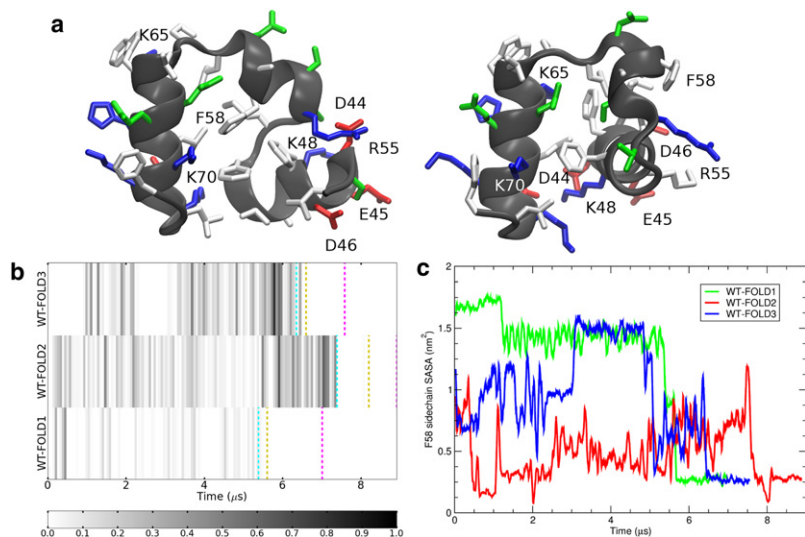
FIGURE 3 Proposed experimental metrics for villin folding. (a) Clusters 1 and 2 from simulation WT-FOLD1. Hydrophobic, polar, basic, and acidic side chains are shown in white, green, blue, and red, respectively. (b) Density plot showing the formation of the D44-K48 salt bridges throughout the WT simulations; for each 6-ns segment, the fraction of frames with a formed salt bridge (<3.5 Å heavy atom distance) is shown. Dashed cyan, yellow, and magenta lines indicate the opening transition, formation of native structure, and final frame, respectively, in each trajectory. (c) Solvent-exposed surface area of the side chain of F58, averaged over 60 ns.

are similar only in that they all show secondary structure formation and hydrophobic collapse (albeit not to nativelike values). In all three, a set of long-lived intermediate states is eventually formed in which native secondary structure is present, but helix I is flipped and rotated relative to helix III, preventing formation of the complete set of hydrophobic interactions present in the native state. All three simulations then proceed to fold after a structural transition in which the three helices dissociate from and then reassociate with each other.

The pattern observed in the WT simulations of independent trajectories showing very different early stages of folding, a series of interconverting folding intermediates, and then a final, consistent transition to the native state, is qualitatively consistent with the folding funnel hypothesis (49) and energy landscape theory (28). Indeed, plotting histograms of the number of clusters present in each trajectory as a function of Q_{res} (Fig. S9) illustrates that except for very small values of Q_{res} (which are poorly sampled because of the rapid early stages of collapse of the protein), the number of distinct accessible conformations drops with increasing Q_{res}, until a single native cluster is observed (or, in the case of WT-FOLD2, two near-native clusters). Energy landscape theory suggests that proteins satisfy the principal of minimal frustration (28,50), and in general we find that the number of nonnative contacts formed during villin folding is significantly less than the number of native contacts formed in the native state (as seen in Fig. S1). Frustration is, however, still present in the villin structure, and appears to give rise to the flipped intermediate state observed in villin folding, given the presence of nonnative helix I-helix III salt bridging in this intermediate. Interestingly, a general folding mechanism of rapid initial collapse, followed by a variety of different paths as the protein converts between a series of intermediates, and a late transition state giving rise to exponential decay, has previously been suggested based on on-lattice simulations of model proteins (51).

The mechanism for folding observed in our simulations does not appear to match any of the hypotheses drawn from other recent simulations of villin folding (see Introduction). Our mechanism most closely resembles the results of implicit solvent Monte Carlo simulations by Yang et al. indicating that the transition state ensemble contains well-formed secondary structure elements but is mostly disordered (22). However, we observe a transient opening event in all three trajectories in which the radius of gyration spikes sharply, whereas the results of Yang et al. (22) indicate that the transition state ensemble is only moderately more extended than the native state.

Although we did observe rapid folding to native or near-native states in two of six simulations of the NLE mutant, the other four trajectories became trapped in nonnative states for periods significantly longer than the expected folding time of the mutant (or even, in several cases, the WT protein). Both effects appear to be largely mediated by the increased hydrophobicity of the Nle residues, as the successful folding trajectories immediately recruit nativelike hydrophobic cores, and the misfolded states contain clusters of hydrophobic residues interacting with both norleucines. In no case is a flipped state resembling the typical folding intermediate from WT simulations observed, suggesting that the free energy surface on which the Nle mutant folds may be substantially different from that of the WT protein due to destabilization of the most favorable set of WT folding intermediates. We note that the presence of several very slow folding pathways for the NLE mutant was also observed by Ensign et al. (26) using a different force field.

The specific folding times observed in the wild-type simulations are ~2 μs slower than the experimentally expected range, although as noted above, the discrepancy may be partially due to a fully extended starting conformation, or may simply be due to the small number of simulations sampled. The possible effects of the starting conformation on folding rates are more apparent in the case of the NLE

simulations: two trajectories starting from the extended structure reached nativelike states, but none starting from the denatured structure did. Neither of the starting structures used in this study are likely to be representative of conformations occurring in the denatured state ensemble in villin folding experiments, but, as noted above, they do offer starting conformations that avoid unduly biasing the simulation toward rapid folding. Protein folding simulations beginning from more realistic denatured states would be a clear step toward even closer linkage of protein folding simulation and experiment, and would remove one possible cause for discrepancies in folding kinetics, but will only be possible through more detailed experimental data on the denatured state or extensive simulations characterizing it.

Despite the wide variety of folding behaviors exhibited by the WT and NLE trajectories in this study, nativelike interaction distances were usually observed between W64 and H68 after the first few microseconds of simulation. This observation indicates that the corresponding fluorescent probe may underestimate folding times or mask heterogeneity in the folding process by measuring only formation of local native structure around the beginning of helix III, in line with the observation of a slower folding rate in recent temperature jump experiments tracking a global observable (15). Because the detailed mechanism of W64 fluorescence quenching by the neighboring histidine is not known, it is possible that some sensitivity to the local protein environment allows the quenching to distinguish between the folded structure and folding intermediates, but comparisons to time-resolved data using other folding observables would be needed to verify this possibility (26). On the other hand, it is also possible that the kinetic discrepancies observed in these simulations (slower than expected folding of WT villin, and very slow folding of the NLE variant) are due to some failure of the MD force field in properly describing the folding process, e.g., due to neglect of atomic polarizability that could affect dipole-dipole interactions along the helical backbone. Certainly, even recent explicit-solvent MD simulations on realistic folding timescales have not always given correct results (e.g., (13)), and even in cases where simulations agree with experiment in terms of structure and kinetics, previous attempts to computationally predict mutations accelerating villin folding have been unsuccessful (3). MD simulations of loop closure in a model peptide using different force fields have also illustrated that even when different force fields agree with experiment, they may yield different mechanistic details (52). Molecular dynamics simulations do, however, benefit from ongoing improvements both in the force fields used and the amount of sampling that can be performed, and thus are expected to provide continually more accurate views of processes such as protein folding. In the case of the villin headpiece subdomain, two sets of explicit solvent-folding simulations of the NLE mutant (this work and that of (26)), using different force fields, have both suggested that a frequently used experi-mental probe for villin folding may overestimate the folding rate of the protein. Our results also suggest a certain overestimate of the folding rate for the WT version of the protein. Importantly, the results provide clear mechanistic detail which can be tested experimentally both to assess the general accuracy of the model for villin folding that we present, and to answer the specific question of whether suggested alternative observables will yield slower folding rates.

Based on our WT trajectories, we propose that monitoring of the D44-K48 interaction using isotope-labeled IR spectroscopy offers an alternative probe that is sensitive to the final folding transition. If our observed folding mechanism is correct, this interaction should relax more slowly than W64-H68 quenching, and more accurately reflect the transition to the native state.

In summary, we have performed what are, to our knowledge, the first complete folding simulations of N68H villin headpiece subdomain in explicit solvent over physically realistic timescales. The use of unbiased, explicit solvent atomistic simulations of folding provided detailed information on the nature of intermediate structures occurring during folding that might be obscured through approximations such as Gō potentials, coarse-graining, or implicit solvent models. Three separate trajectories of WT villin reached native states through similar mechanisms, suggesting a class of long-lived intermediate states and revealing the approximate nature of the transition state ensemble. Similar simulations of a fast-folding double-norleucine mutant yielded more heterogeneous behavior, and suggest either inaccuracies in the simulation methods used or the presence of a variety of slower folding pathways, which are masked by the spectroscopic observable used experimentally to monitor folding (as in (26)). Based on our results, we are able to propose both alternative experimental observables to monitor folding and a pair of mutations that would be expected to accelerate villin folding, in hopes of allowing experimental testing of the folding mechanism we observed for WT villin. As molecular dynamics simulations promise to reach timescales beyond the 50 μs covered here and force fields become more accurate, we expect that their utility for studying large conformational transitions such as those involved in protein folding will continue to grow.

SUPPORTING MATERIAL

Nine figures and three movies are available at http://www.biophysj.org/biophysj/supplemental/S0006-3495(09)01359-9.

REFERENCES

1. Yang, W. Y., and M. Gruebele. 2003. Folding at the speed limit. *Nature.* 423:193–197.

2. Kubelka, J., J. Hofrichter, and W. A. Eaton. 2004. The protein folding 'speed limit'. *Curr. Opin. Struct. Biol.* 14:76–88.

3. Kubelka, J., W. A. Eaton, and J. Hofrichter. 2003. Experimental tests of villin subdomain folding simulations. *J. Mol. Biol.* 329:625–630.

4. Kubelka, J., T. K. Chiu, D. R. Davies, W. A. Eaton, and J. Hofrichter. 2006. Sub-microsecond protein folding. *J. Mol. Biol.* 359:546–553.

5. Arora, P., T. G. Oas, and J. K. Myers. 2004. Fast and faster: a designed variant of the B-domain of protein A folds in 3 μs. *Protein Sci.* 13:847–853.

6. Wang, T., Y. Zhu, and F. Gai. 2004. Folding of a three-helix bundle at the folding speed limit. *J. Phys. Chem. B.* 108:3694–3697.

7. Zhou, R., and B. J. Berne. 2002. Can a continuum solvent model reproduce the free energy landscape of a β-hairpin folding in water? *Proc. Natl. Acad. Sci. USA.* 99:12777–12782.

8. Zhou, R. 2003. Free energy landscape of protein folding in water: explicit vs. implicit solvent. *Proteins.* 53:148–161.

9. Juraszek, J., and P. G. Bolhuis. 2006. Sampling the multiple folding mechanisms of Trp-cage in explicit solvent. *Proc. Natl. Acad. Sci. USA.* 103:15859–15864.

10. Paschek, D., S. Hempel, and A. E. Garca. 2008. Computing the stability diagram of the Trp-cage miniprotein. *Proc. Natl. Acad. Sci. USA.* 105:17754–17759.

11. Sanbonmatsu, K. Y., and A. E. Garca. 2002. Structure of Met-enkephalin in explicit aqueous solution using replica exchange molecular dynamics. *Proteins.* 46:225–234.

12. Zhou, R., B. J. Berne, and R. Germain. 2001. The free energy landscape for β-hairpin folding in explicit water. *Proc. Natl. Acad. Sci. USA.* 98:14931–14936.

13. Freddolino, P. L., F. Liu, M. Gruebele, and K. Schulten. 2008. Ten-microsecond MD simulation of a fast-folding WW domain. *Biophys. J.* 94:L75–L77.

14. McKnight, C. J., D. S. Doering, P. T. Matsudaira, and P. S. Kim. 1996. A thermostable 35-residue subdomain within villin headpiece. *J. Mol. Biol.* 260:126–134.

15. Bunagan, M. R., J. Gao, J. W. Kelly, and F. Gai. 2009. Probing the folding transition state structure of the villin headpiece subdomain via side chain and backbone mutagenesis. *J. Am. Chem. Soc.* 131:7470–7476.

16. Wang, M., Y. Tang, S. Sato, L. Vugmeyster, C. J. McKnight, et al. 2003. Dynamic NMR line-shape analysis demonstrates that the villin headpiece subdomain folds on the microsecond time scale. *J. Am. Chem. Soc.* 125:6032–6033.

17. Buscaglia, M., J. Kubelka, W. A. Eaton, and J. Hofrichter. 2005. Determination of ultrafast protein folding rates from loop formation dynamics. *J. Mol. Biol.* 347:657–664.

18. Fernandez, A., M. Y. Shen, A. Colubri, T. R. Sosnick, R. S. Berry, et al. 2003. Large-scale context in protein folding: villin headpiece. *Biochemistry.* 42:664–671.

19. Jang, S., E. Kim, S. Shin, and Y. Pak. 2003. Ab initio folding of helix bundle proteins using molecular dynamics simulations. *J. Am. Chem. Soc.* 125:14841–14846.

20. Lei, H., and Y. Duan. 2007. Two-stage folding of HP-35 from ab initio simulations. *J. Mol. Biol.* 370:196–206.

21. Lei, H., C. Wu, H. Liu, and Y. Duan. 2007. Folding free-energy landscape of villin headpiece subdomain from molecular dynamics simulations. *Proc. Natl. Acad. Sci. USA.* 104:4925–4930.

22. Yang, J. S., S. Wallin, and E. I. Shakhnovich. 2008. Universality and diversity of folding mechanics for three-helix bundle proteins. *Proc. Natl. Acad. Sci. USA.* 105:895–900.

23. Lei, H., X. Deng, Z. Wang, and Y. Duan. 2008. The fast-folding HP35 double mutant has a substantially reduced primary folding free energy barrier. *J. Chem. Phys.* 129:155104–155107.

24. Duan, Y., and P. A. Kollman. 1998. Pathways to a protein folding intermediate observed in a 1-μs simulation in aqueous solution. *Science.* 282:740–744.

25. Zagrovic, B., C. D. Snow, M. R. Shirts, and V. S. Pande. 2002. Simulation of folding of a small α-helical protein in atomistic detail using worldwide-distributed computing. *J. Mol. Biol.* 323:927–937.

26. Ensign, D. L., P. M. Kasson, and V. S. Pande. 2007. Heterogeneity even at the speed limit of folding: large-scale molecular dynamics study of a fast-folding variant of the villin headpiece. *J. Mol. Biol.* 374:806–816.

27. Jayachandran, G., V. Vishal, and V. S. Pande. 2006. Using massively parallel simulation and Markovian models to study protein folding: examining the dynamics of the villin headpiece. *J. Chem. Phys.* 124:164902.

28. Onuchic, J. N., Z. Luthey-Schulten, and P. G. Wolynes. 1997. Theory of protein folding: the energy landscape perspective. *Annu. Rev. Phys. Chem.* 48:545–600.

29. Phillips, J. C., R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, et al. 2005. Scalable molecular dynamics with NAMD. *J. Comput. Chem.* 26:1781–1802.

30. Andersen, H. C. 1983. RATTLE: a ''velocity'' version of the SHAKE algorithm for molecular dynamics calculations. *J. Chem. Phys.* 52:24–34.

31. Miyamoto, S., and P. A. Kollman. 1993. SETTLE: an analytical version of the SHAKE and RATTLE algorithm for rigid water models. *J. Comput. Chem.* 13:952–962.

32. MacKerell, Jr., A. D., M. Feig, and C. L. Brooks, III. 2004. Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J. Comput. Chem.* 25:1400–1415.

33. Chiu, T. K., J. Kubelka, R. Herbst-Irmer, W. A. Eaton, J. Hofrichter, et al. 2005. High-resolution x-ray crystal structures of the villin headpiece subdomain, an ultrafast folding protein. *Proc. Natl. Acad. Sci. USA.* 102:7517–7522.

34. Humphrey, W., A. Dalke, and K. Schulten. 1996. VMD—visual molecular dynamics. *J. Mol. Graph.* 14:33–38.

35. Frishman, D., and P. Argos. 1995. Knowledge-based secondary structure assignment. *Proteins.* 23:566–579.

36. Eastwood, M. P., C. Hardin, Z. Luthey-Schulten, and P. G. Wolynes. 2001. Evaluating protein structure-prediction schemes using energy landscape theory. *IBM J. Res. Develop.* 45:475–497.

37. van der Spoel, D., E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, et al. 2005. GROMACS: fast, flexible, and free. *J. Comput. Chem.* 26:1701–1718.

38. Daura, X., K. Gademann, B. Jaun, D. Seebach, W. F. van Gunsteren, et al. 1999. Peptide folding: when simulation meets experiment. *Angew. Chem. Int. Ed.* 38:236–240.

39. Shen, M.-y., and K. F. Freed. 2002. Long time dynamics of Met-encephalin: comparison of explicit and implicit solvent models. *Biophys. J.* 82:1791–1808.

40. Feller, S. E., R. W. Pastor, A. Rojnuckarin, S. Bogusz, and B. R. Brooks. 1996. Effect of electrostatic force truncation on interfacial and transport properties of water. *J. Phys. Chem.* 100:17011–17020.

41. Zagrovic, B., and V. Pande. 2003. Solvent viscosity dependence of the folding rate of a small protein: distributed computing study. *J. Comput. Chem.* 24:1432–1436.

42. Tang, Y., D. Rigotti, R. Fairman, and D. Raleigh. 2004. Peptide models provide evidence for significant structure in the denatured state of a rapidly folding protein: the villin headpiece subdomain. *Biochemistry.* 43:3264–3272.

43. Havlin, R. H., and R. Tycko. 2005. Probing site-specific conformational distributions in protein folding with solid-state NMR. *Proc. Natl. Acad. Sci. USA.* 102:3284–3289.

44. Freddolino, P. L., S. Park, B. Roux, and K. Schulten. 2009. Force field bias in protein folding simulations. *Biophys. J.* 96:3772–3780.

45. Mukherjee, S., P. Chowdhury, M. R. Bunagan, and F. Gai. 2008. Folding kinetics of a naturally occurring helical peptide: implication of the folding speed limit of helical proteins. *J. Phys. Chem. B.* 112:9146–9150.

46. Huang, C. Y., Z. Getahun, T. Wang, W. F. DeGrado, and F. Gai. 2001. Time-resolved infrared study of the helix-coil transition using ^{13}C-labeled helical peptides. *J. Am. Chem. Soc.* 123:12111–12112.

47. Holloway, P., and H. Mantsch. 1988. Infrared spectroscopic analysis of salt bridge formation between cytochrome b5 and cytochrome *c*. *Biochemistry.* 27:7991–7993.

48. Getahun, Z., C.-Y. Huang, T. Wang, B. D. Len, W. F. DeGrado, et al. 2003. Using nitrile-derivatized amino acids as infrared probes of local environment. *J. Am. Chem. Soc.* 125:405–411.

49. Onuchic, J. N., P. G. Wolynes, Z. Luthey-Schulten, and N. D. Socci. 1995. Toward an outline of the topography of a realistic protein-folding funnel. *Proc. Natl. Acad. Sci. USA.* 92:3626–3630.

50. Bryngelson, J. D., and P. G. Wolynes. 1987. Spin glasses and the statistical mechanics of protein folding. *Proc. Natl. Acad. Sci. USA.* 84:7524–7528.

51. Camacho, C. J., and D. Thirumalai. 1993. Kinetics and thermodynamics of folding in model proteins. *Proc. Natl. Acad. Sci. USA.* 90:6369–6372.

52. Yeh, I.-C., and G. Hummer. 2002. Peptide loop-closure kinetics from microsecond molecular dynamics simulations in explicit solvent. *J. Am. Chem. Soc.* 124:6563–6568.