# Quantitative comparison of villin headpiece subdomain simulations and triplet–triplet energy transfer experiments

Kyle A. Beauchamp[a], Daniel L. Ensign[b], Rhiju Das[a,c,d,1], and Vijay S. Pande[a,e,1]

[a]Biophysics Program, [c]Biochemistry, [d]Physics, and [e]Chemistry Departments, Stanford University, Stanford CA 94305; and [b]Department of Chemistry and Biochemistry, The University of Texas at Austin, Austin, TX 78712

Edited* by Robert Baldwin, Stanford University, Stanford, CA, and approved June 22, 2011 (received for review July 24, 2010)

**As the fastest folding protein, the villin headpiece (HP35) serves as an important bridge between simulation and experimental studies of protein folding. Despite the simplicity of this system, experiments continue to reveal a number of surprises, including structure in the unfolded state and complex equilibrium dynamics near the native state. Using 2.5 ms of molecular dynamics and Markov state models, we connect to current experimental results in three ways. First, we present and validate a novel method for the quantitative prediction of triplet–triplet energy transfer experiments. Second, we construct a many-state model for HP35 that is consistent with previous experiments. Finally, we predict contact-formation time traces for all 1,225 possible triplet–triplet energy transfer experiments on HP35.**

protein dynamics | near-native dynamics

A detailed understanding of how proteins fold remains a grand challenge in biophysics, whose solution would yield both satisfying physical insight and useful biological applications. To approach this problem, many have carried out detailed studies of model systems such as the 35-residue villin headpiece protein (HP35) (1). Sometimes called "the hydrogen atom of protein folding," this small protein folds in only microseconds (2), yet retains the clear secondary structure and well-packed core characteristic of larger proteins. These features make HP35 an ideal model; in particular, the fast-folding timescale allows for a detailed connection between experiments and molecular dynamics simulations. The allure of in silico protein folding has led to several computational studies of this small protein. Early landmark simulations showed partial folding (3). Later simulations achieved a small number of complete folding trajectories (4) and suggested residue-level interactions involved in the rate limiting steps of folding. Today's computers permit the study of multiple folding trajectories, permitting a statistically sound approach to questions of states, rates, and pathways (5, 6). With myriad computational and experimental studies of this tiny system, one may expect the story of HP35 to be complete.

However, recent experimental research highlights a number of questions and even a few surprises. Laser temperature jump studies showed deviations of observed relaxations from single-exponential behavior (7). Further studies suggested that such relaxations might be fit using several exponentials, indicating the possibility of high barriers within the unfolded or partially folded states (8, 9). NMR and IR measurements revealed residual structure in the unfolded state (10), particularly in the two N-terminal helices (11). Finally, triplet–triplet energy transfer (TTET) experiments have observed fluctuations between native and several nonnative conformations on multiple timescales (12). This flurry of results suggests that our understanding of this "simple" system remains incomplete.

Recent advances in experimental methods can resolve ever-finer details in the protein folding process (12, 13). Likewise, molecular mechanics force fields have steadily improved (14–16), with accompanying enhancements in computer performance (17, 18). Interpreting molecular dynamics simulations is facilitated by analysis methods such as Markov state models (MSMs) (19–22), which provide a probabilistic framework for describing the conformational transitions observed in simulations. Progress on these fronts suggests that simulation and experiment should be more tightly coupled than ever before, with computation giving specific, falsifiable predictions and experiments quickly testing them (4, 9, 23).

Here, we use molecular dynamics simulations and MSMs to investigate the dynamics of the fast-folding, double-norleucine HP35 mutant (7). By modeling TTET in an MSM framework, we show near-quantitative agreement with recent experiments. Our analysis identifies states consistent with the observations in ref. 12 and elucidates these states in atomic detail, suggesting a set of specific nonnative contacts that frequently occur in the N terminus and a fluctuation in the C terminus that leaves a helix partly but not completely unraveled. These phenomena are strikingly apparent in a comprehensive set of simulated TTET traces, whose fine structure is surprisingly robust to modeling and force field errors. We present these calculations as a challenge to the experimental community. Further validation of these predictions will critically test our understanding of HP35 and demarcate the limitations of state-of-the-art simulation.

## Results

**Comparing TTET Experiments to Simulations.** The recent application of TTET experiments to submicrosecond fluctuations of HP35 (12) raised the possibility of testing simulations at an unprecedented level of detail. We therefore sought to reproduce four experiments that probed the rate of contact formation between a xanthone donor and napthylalanine acceptor attached at residue pairs (0, 23), (7, 23), (35, 23), and (0, 35) (Fig. 1). (The experiments of ref. 12 attached 9-oxoxanthene-2-carboxylic acid to the N terminus via a peptide bond and called this position 0.) To make direct comparisons to TTET measurements, we developed a method for simulating contact-formation time traces that expands the MSM formalism to simultaneously model both conformational and triplet state dynamics (see *Methods*). We created MSMs based on molecular dynamics simulations of a fast-folding double-norleucine HP35 mutant (7), extending prior calculations (5) to include more than 1 ms of simulation and requiring more

Author contributions: K.A.B., R.D., and V.S.P. designed research; K.A.B. and D.L.E. performed research; K.A.B. and D.L.E. contributed new reagents/analytic tools; K.A.B. analyzed data; and K.A.B., R.D., and V.S.P. wrote the paper.

The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

Data deposition: Contact formation time traces are provided at the National Institutes of Health Simbios Center for Biomedical Computing, https://simtk.org/home/hp35-datasets/.

[1]To whom correspondence may be addressed. E-mail: pande@stanford.edu or rhiju@stanford.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1010880108/-/DCSupplemental.

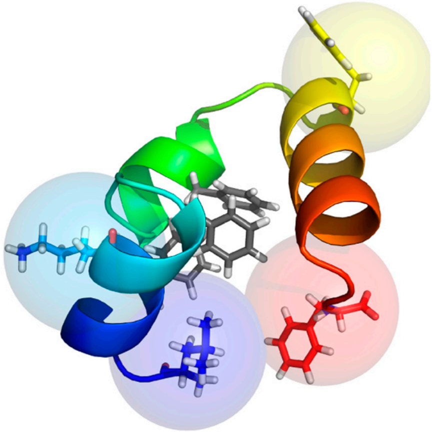


**Fig. 1.** Probe locations For TTET experiments. The TTET experiments in ref. 12 monitored contact formation between a (donor, acceptor) pair inserted at positions (7, 23), (0, 23), (35, 23), and (0, 35). Residues 1 (blue), 7 (cyan), 23 (yellow), and 35 (red) and the tightly packed phenylalanine core (gray) are shown as sticks (PDB ID code 2F4K). The N terminus (residue 1) is located on the left and is colored in blue. TTET interaction radii are shown as 5-Å spheres on beta carbons.

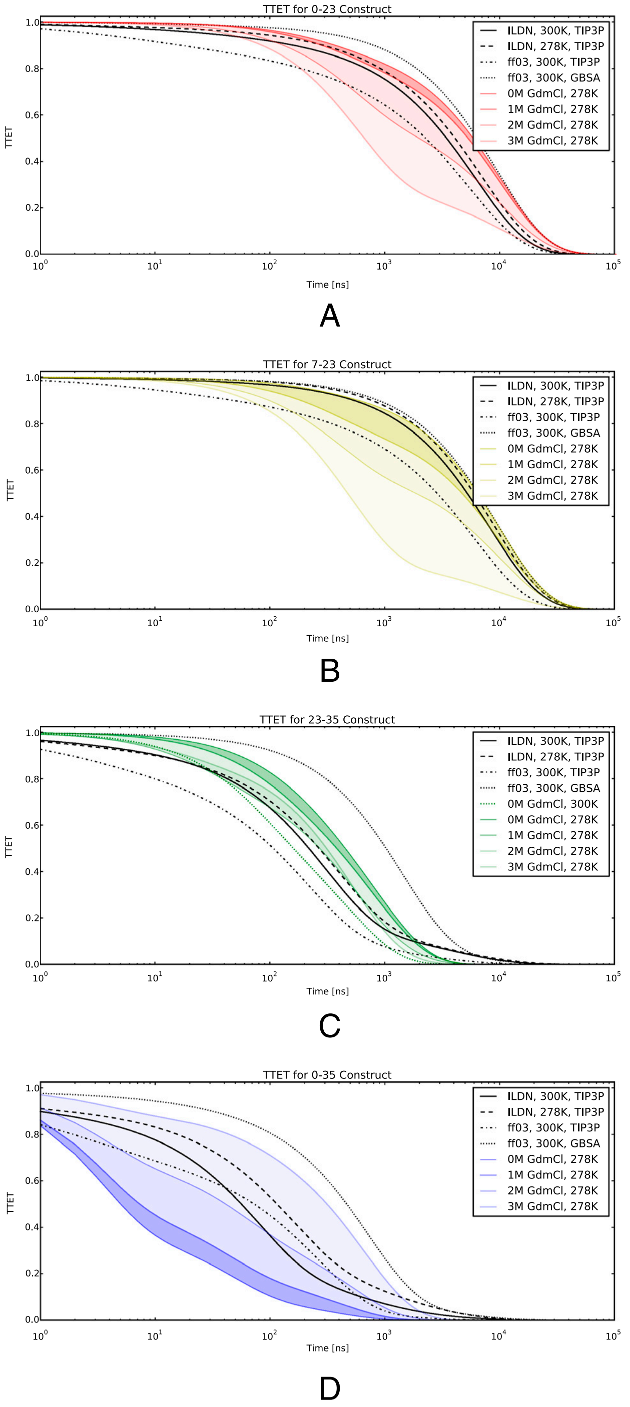


**Fig. 2.** Comparing simulated TTET to experiment. Simulated TTET traces are compared to experiments. The experimental data are calculated using the exponential fits in ref. 12. GdmCl concentrations are shown as color gradients, with the darkest representing 0 M. Simulated TTET is shown for all four datasets. *A* displays the TTET between residues (0, 23). *B* shows (7, 23). *C* shows (23, 35). *D* shows (0, 35). *C* also contains experimental data at (0 M GdmCl, 300 K). Simulations assume 10-μs xanthone decay.

than 10 million CPU hours of computation (see *Methods* and *SI Methods*).

Systematic error in comparing simulations to TTET experiments results from six primary sources: MSM error, uncertainty in modeling the TTET observable, differences between simulated and experimental constructs, differences in solution conditions between simulation and experiment, simulation force field error, and sampling limitations. First, we estimate the systematic error from MSM construction by examining several MSMs for a given simulation dataset (see Fig. S1). Second, simulated TTET depends on the relationship between TTET rate and donor-acceptor distance; the error associated with this relationship is minimized by fitting a single parameter (see *Methods*, *SI Methods*, and Fig. S2). Third, the Lys24Nle/Lys29Nle construct used in molecular dynamics simulations is different from the Met12Nle/double-fluorophore constructs used in the experiments of ref. 12. To address this error, we analyzed short simulations of a fluorophore-containing construct (see *SI Methods* and Fig. S2). Fourth, experiments were performed primarily at 278 K, but fixed charge force fields are optimized for 300 K; we performed most calculations at 300 K but repeated one set at 278 K for comparison. Fifth, we control the error due to force field uncertainty by using datasets simulated with different force field/solvation model combinations. In total, we analyzed four datasets (ff03-TIP3P-300K, ff03-GBSA-300K, ff99SB-ILDN-TIP3P-300K, and ff99SB-ILDN-TIP3P-278K). Finally, limited sampling introduces uncertainties in folding stability; this destabilizing effect can be mitigated by examining experimental data under denaturing conditions, as is discussed below (see Table S1). In the following, we report data from the ff99SB-ILDN-TIP3P-300K dataset, with errors bracketed by the differences between the four datasets, which dominate most other sources of error.

The first comparison (Fig. 2*A*) measures contact formation between residues 0 and 23. This process is thought to require disrupting the entire core of the protein, and is experimentally measured to be slower than 8 μs—TTET measurement is limited by the 10-μs xanthone lifetime (12). The simulation also finds a slow timescale (5.5 ± 2.0 μs; $1/e$ time) for this contact formation. The second comparison (Fig. 2*B*) involves another slow (7.8 ± 2.0 μs) unfolding process that is required for contact between 7 (the N terminus) and 23. The multiexponential MSM traces and the measurements overlay within systematic uncertainties. Furthermore, unlike experimental measurements of TTET, the simulation-based method can observe timescales beyond 10 μs by disabling the intrinsic xanthone decay (Fig. S3). Simulations predict 20–100-μs timescales for these processes, consistent with experimentally measured unfolding under mildly denaturing conditions (7).

A third comparison involves excitation transfer between probes at residues 23 and 35 and is sensitive to fluctuations of the C-terminal helix (Fig. 2*C*). Simulated and experimental TTET traces agree, suggesting a timescale of 400 (±300) and 380 ns, respectively—an order of magnitude faster than global unfolding. For this probe location, experimental data were available at multiple temperatures. With the exception of the implicit solvent data, simulation predictions rest between the experimental values for 278 and 300 K. The experimental observation of a C-terminal unraveled state is not limited to TTET studies; previous NMR work has observed a strong temperature dependence of slow motion order parameters in this helix (24).

The fourth comparison (Fig. 2*D*) measures contact formation between residues 0 and 35. Fits of experimental TTET traces yielded very fast kinetics: a dominant nanosecond (<50 ns) phase and an additional submicrosecond (330 ns) phase. Simulations predict somewhat slower TTET (150 ± 100 ns) than is observed experimentally, which may be attributed to two factors. First, the nonpolar interaction between fluorophores in the experimental construct may accelerate the kinetics of 0–35 contact formation. Control simulations including fluorophores suggest rate enhancement by a factor of three (see Fig. S2); similarly, CD experiments found the 0–35 construct up to 0.4 kcal/mol more stable than the other three constructs. Second, the present simulations (at 300 K) may favor the near-native state rather than the closely packed native state probed by experiments at 278 K, a point further discussed below. Indeed, all four TTET predictions agree with experiments at moderately denaturing (1–3 M GdmCl) conditions.

**Simulation of Denaturant Dependence.** As a final comparison to TTET experiments, we carried out studies of GdmCl denaturation. Additional molecular dynamics simulations at 300 K using available GdmCl parameters (25) gave insignificant destabilization of HP35 by denaturant; the available parameterization may not be sufficiently transferable for use with the ff99SB-ILDN force field. We achieved a more robust and predictive analysis by combining MSM perturbation theory (*SI Methods*) with two separate models for GdmCl effects: Myer's solvent accessible surface area (SASA) *m*-value correlation (26) and a group transfer free energy (GTFE) model (27). In terms of thermodynamics, these calculations yield *m*-values of $0.53 \pm 0.2$ kcal · mol$^{-1}$ · M$^{-1}$ (SASA) and $0.46 \pm 0.2$ kcal · mol$^{-1}$ · M$^{-1}$ (GTFE), in agreement with the measured value of $0.69 \pm 0.1$ kcal · mol$^{-1}$ · M$^{-1}$ (12) (Fig. S4*A*). In terms of kinetics, simulated TTET traces show weakly decreasing or constant rates of 0–23, 7–23, and 23–35 contact formation with increasing GdmCl, and strongly increasing rates of 0–35 contact formation (Fig. S4 *B*–*E*). The 0 M GdmCl calculations are in quantitative agreement with experimental TTET measurements at 2 M GdCl (Fig. S4*B*), further confirming that simulations give an accurate picture of HP35 dynamics in moderately denaturing conditions.

**Comparison to Prior Models.** The quantitative comparison between simulation and experiment described above requires a high-resolution MSM with 10,000 states or more. In contrast, prior models for HP35 folding typically use a small number of states to concisely explain existing experimental data. Thus, a key question is to what extent do simulations capture the behavior of previous empirical models.

Simulation gives reasonable agreement with prior experiment-based models of HP35's structure and stability. Grouping states into folded and unfolded macrostates based on the similarity (rmsd) to the crystal structure [Protein Data Bank (PDB) ID code 2F4K] suggests an apparent two-state $\Delta G$ of $-0.5 \pm 0.5$ kcal/mol (Table S1). Circular dichroism studies of the same mutant report a value of −4.0 kcal/mol (7). This thermodynamic comparison suggests that the native state is understabilized in the simulation, and is consistent with the agreement of simulations and TTET experiments in moderate concentrations of chemical denaturant.

To compare our model with previous experiments and to gain a qualitative picture of folding dynamics, we applied the Perron Cluster Cluster Analysis (28) model reduction procedure to the ff99SB-ILDN-TIP3P-300K dataset (see *SI Methods*). Reduced MSMs permit comparison to the recent four-state model based on TTET data (12). The four-state model included native (N), unlocked (N′), and intermediate (I) states to account for the experimentally observed near-native fluctuations, as well as a globally unfolded state U (U not shown in Fig. 3). In the previous model, the native state N was shown to be associated with close contact between the terminal residues 0 and 35. The near-native state (N′) was shown to be native-like, but lacking close contact between the terminal residues. Finally, the intermediate I was characterized by unfolding of the C-terminal helix, as measured by TTET between residues 23 and 35. The simulation-based reduced MSM is consistent with this picture (Fig. 3, Fig. S5, Fig. S6, and Table S2).

From the 1,000-state reduced MSM, we focus on the ten most populated states, which are labeled **1** through **10** in order of population, with **1** being the most populated state (state properties are given in Table S2). We identify N with states **1** and **5**; both of these states bring together the N and C termini, permitting 0–35 TTET transfer. State 1 is also the most similar to the crystal structure (PDB ID code 2F4K), as indicated by its 2.2-Å rmsd. The states **2**, **6**, **7**, **8**, and **9** are characterized by native-like secondary structure and topology, but with structural differences as compared to the crystallographic structure. We identify this ensemble of near-native states with the unlocked state (N′). States **3**, **4**, and **10** show visible unraveling of the C-terminal helix, suggesting a connection to the intermediate I. Indeed, state **3** puts residues 23 and 35 in contact, providing an explanation for the submicrosecond 23–35 TTET previously assigned to I. The relative populations of the states (N′, N, and I in decreasing population) in the (300 K) simulation are in qualitative agreement with experimental estimates at (278 K, 2 M GdmCl). Simulations thus capture

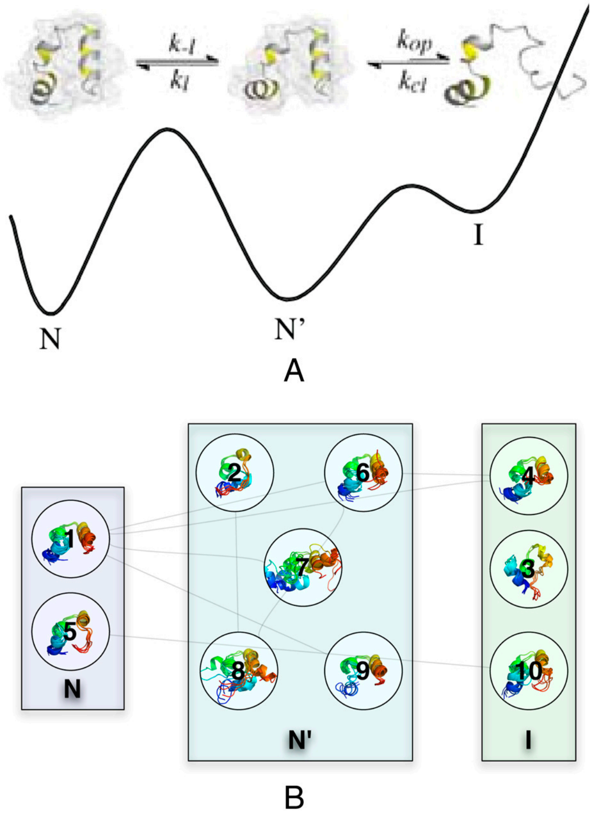


**Fig. 3.** Two models for HP35 dynamics. The model from ref. 12 is shown in *A* and has three distinct states: native (N), near-native globule (N′), and C-terminal unraveled intermediate (I). (Reproduced with permission from ref. 12.) The ten most populated states (*B*) from a 1,000-state reduced MSM show clear similarity to the previous model for HP35.

HP35 dynamics under moderately denaturing conditions in which the native state (N) is sampled but the unlocked (N′) state is more frequently populated.

**Structural Predictions from Simulations.** Despite the similarities above, the reduced MSM is distinct from the previous four-state model (12). We describe two differences here.

The previous model for the observed intermediate (I in ref. 12; states **3**, **4**, and **10** herein) proposed complete unfolding of the 11-residue C-terminal helix (residues 22–32) to allow contact formation between residues 23 and 35. Simulations suggest that this state involves only partial unraveling of the C-terminal helix, rather than a complete disordering. In particular, the segment 22–25 is helical; the next turn (residues 26–29) is partly unstructured, and the C-terminal residues (30–35) are fully disordered. Although the differences between the MSM description and the previous description are subtle, hydrogen/deuterium exchange measurements (1) support the molecular dynamics picture in finding a monotonic decrease in protection factors from residues 22–35 in the C-terminal helix; this monotonic decrease is in contrast to the approximately constant protection factors one would expect of a cooperatively unfolded helix. Additional experiments could further discriminate between these models, including measurements that monitor the position dependence of the rate of contact formation, as described below.

Second, simulations provide a structural model for a near-native, compact state (N′ in ref. 12; states **2**, **6**, **7**, **8**, and **9** herein) that accounts for rapid fluctuations measured in end-to-end contacts (12). This state, previously called an “unlocked” or “dry molten globule” ensemble, was hypothesized to display large fluctuations in the C-terminal helix, making it a natural stepping stone to the more disordered intermediate I (see above). In contrast, molecular dynamics simulations display strong structural heterogeneity in both the N-terminal and C-terminal helices (Figs. S6 and S7) (29). The MSM predicts not just fluctuations but also specific nonnative contacts that have not been previously described. Most notably, residues Asp3 and Ser15, which are well separated in the crystallographic model, are predicted to hydrogen bond via the Ser hydroxyl and Asp carboxylate groups at a frequency of up to 50% (Fig. S7). This contact should be testable experimentally, through NMR nuclear Overhauser effect measurements (of protons on Cβ).

**A Comprehensive Set of TTET Predictions.** Rigorous validation of simulations requires not only consistency checks but also nontrivial predictions that can be quantitatively compared to experiments. We therefore extended our contact-formation analysis from four residue pairs to all pairwise combinations.

We present these predictions as a 35 × 35 matrix of timescales ($1/e$ times in Fig. 4; full time-traces in *SI Methods*). Although the uncertainties in predicted contact timescales are in the range of 2–4 fold, these errors are much smaller than the overall span of timescales, which cover a $10^5$-fold range. The fastest rates are subnanosecond and involve residue pairs that are in close contact in the most populated states. The slowest timescales extend beyond 100 μs and involve the proline-containing turn at residues 18–25. There are small differences in datasets with different force fields, solvation models, and temperatures, but, strikingly, the same overall pattern is observed in all four datasets (Fig. 4). On average, the 278 K data is 1.5 times slower than its 300 K counterpart; however, temperature comparisons at individual probe positions are limited by systematic uncertainty.

The most interesting features of the contact timescale matrix are sharp variations as the probe positions are shifted by single residues. For example, consider contact formation between residue 23 and residues 34 and 35. A previous model (12) suggested that contact formation for these pairs involves full disordering of the C-terminal helix (state I); under this model, the residue pair (23, 34) would have slightly faster timescales than the (23, 35) pair. In contrast, simulations give a different picture of I, with decreasing disorder from residue 35 to lower numbered residues. This picture gives a distinct prediction that TTET timescales for the (23, 34) pair will be nearly 10-fold slower than for (23, 35), although the magnitude of the change is near the limit of the systematic errors of the modeling/experiment comparison.

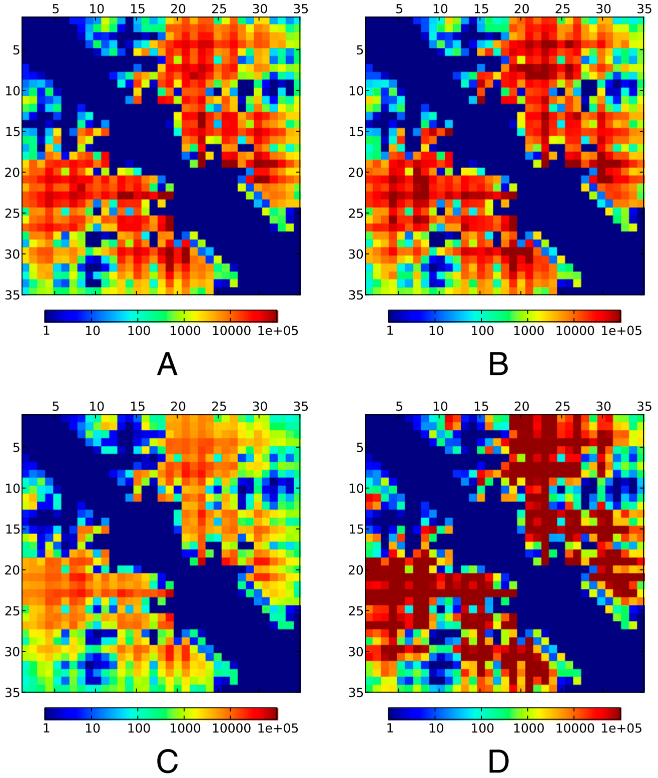


**Fig. 4.** Simulated TTET for all residue pairs provide a single-residue resolution map of contact-formation timescales. The heat map shows the time required for TTET to drop to $1/e$ of its starting amplitude; minimum and maximum represent 1 ns and 100 μs, respectively. (*A*) Using the ff99SB-ILDN-TIP3P-300K dataset. (*B*) Using the ff99SB-ILDN-TIP3P-278K dataset. (*C*) Using the ff03-TIP3P-300K dataset. (*D*) Using the ff03-GBSA-300K dataset. No xanthone decay is assumed.

We further predict that TTET experiments probing contact formation between residue 1 and each of the other residues in HP35 should discover a finely varying spectrum of timescales (Fig. 5). The heterogeneous N terminus of the near-native state N′ places residues 1 and 15 in close contact with a rapid timescale of 20 ± 10 ns, despite the separation of these residues in the crystallographic conformation N. These nonnative contacts are quite specific. For example, residue 12 is also in this N-terminal region but is predicted to interact with residue 1 on a dramatically longer timescale (800 ± 200 ns) due to anchoring by the well-packed phenylalanine core. Further fine structure is apparent in contact-formation rates between residue 1 and residues in the remainder of the protein, including two sites (23 and 35) that have already been probed experimentally (12).

## Discussion

**Challenges of Direct Comparison of Simulation and Experiment.** Our investigation of HP35 is enabled by recent experimental methods that directly probe contact between specific molecular locations. Prior to these advances, comparisons of simulation and experiment were often limited by systematic error. Ideally, one would directly simulate the experimental observable, but it remains challenging to derive circular dichroism spectra or fluorescence changes (30) from structural models. Comparisons are further complicated by

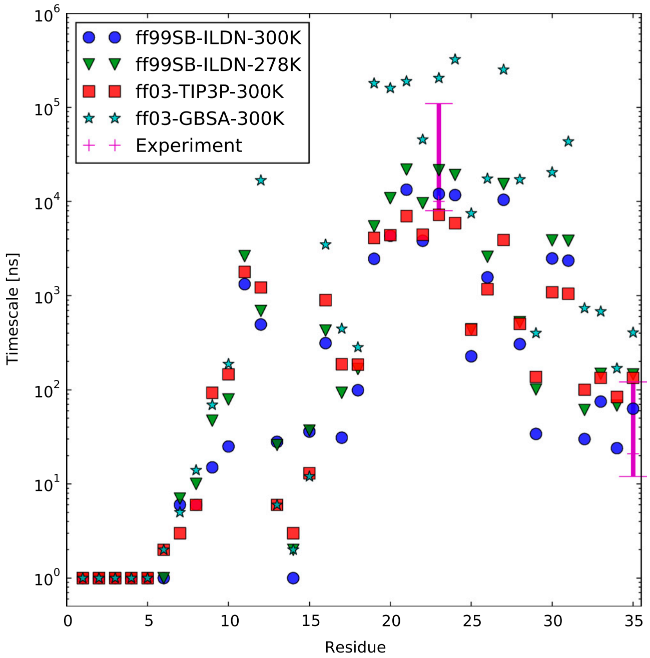


**Fig. 5.** Timescales of contact formation involving residue 1. Experimental values for the 0–35 TTET process are plotted at 0, 1, and 2 M GdmCl. For the 0–23 TTET process, experimental data is shown as the interval [6 μs , 100 μs] to represent uncertainty due to the limited xanthone lifetime. No xanthone decay is assumed in the simulation values.

the presence of multiple structurally distinct states rather than two clearly separated folded and unfolded states. Kinetic measurements introduce additional difficulties, including the difficulty of performing measurements under native conditions and the extraction of relaxation rates rather than microscopic rates. In contrast, contact formation is easily compared to simulation and generally does not require assuming a small number of states.

These challenges appeared several times during our analysis. For example, the 2–4 kcal/mol difference in $\Delta G_{\mathrm{fold}}$ between the circular dichroism model and our own two-state approximation suggests an understabilized native state. However, the lack of a quantitative bridge between the spectroscopic measurement and simulation led to uncertainty in the two-state comparison. The TTET comparisons helped confirm the understabilization of simulated HP35, with the best agreement of traces and denaturant dependences to experiments under moderately denaturing conditions. Comparisons to structural models also highlighted the difficulty of interpreting experimental data. Simulations suggested heterogeneity at the terminal helices of HP35. Early experimental work made no mention of this, although crystallographic models exhibit high temperature factors in the N-terminal helix (Fig. S7) and NMR analysis produced few long-range restraints between this helix and the rest of the molecule (31). Contact-formation experiments identified a near-native state characterized by terminal fluctuations; its TTET signature permits quantitative comparison to simulation. Probes with nanosecond time resolution and residue-level structural resolution allow comparisons that may be unattainable using lower-resolution techniques.

Recent studies have pointed out other challenges in comparing simulation to experiment. Long simulations of HP35 in multiple force fields (32) have suggested that some mechanistic details, such as order of helix folding, may be sensitive to force field differences. This supports the approach used here, where independent datasets using two force fields (ff03 and ff99SB-ILDN) and two water models (TIP3P and GBSA) help evaluate the robustness of simulation-based predictions. Another recent study (33) points out that early simulation-based models of HP35 measured equilibrium constants that were too unfolded. The equilibrium estimates in the current work help ameliorate this deficiency by a more advanced scheme for MSM construction and improved datasets that contain more adequate sampling of the folded state. Even with advanced force field and analysis methods, some deviations between simulation and experiment are expected; for that reason, we stress that the current predictions appear most valid under mildly denaturing conditions.

**Future Challenges.** In the future, faster simulations, well-calibrated denaturant models, and improved force fields should allow the simulation of HP35 across the wide range of solvent and temperature conditions accessible to experiment. Likewise, the present TTET model neglects the effect of fluorophores, except for the analysis presented in Fig. S2. Faster simulation methodologies, increasing computational power, and the facile incorporation of arbitrary chemical modifications will allow simulations to be performed with all choices of donor and acceptor locations. Finally, scaling to larger and more complex proteins may reveal near-native dynamics on a wider spectrum of timescales.

**Conclusion.** Although direct, quantitative agreement between simulation and experiment remains challenging, we have demonstrated a straightforward method to simulate TTET in silico. Using this method, we have predicted 1,225 contact-formation timescales and shown them to be consistent with the four TTET-pairs experimentally probed so far. Simulations lead to a model consistent with existing TTET experiments but with structural differences compared to previous models. Beyond this comparison, atomic-detail simulations predict an ensemble of near-native states and a labyrinth of fine structure in HP35 contact-formation timescales, which will be testable with further TTET experiments. With an aggregate amount of information rivaling high-resolution structural experiments, contact formation provides a way to probe the kinetic, thermodynamic, and structural details of models at high temporal and spatial resolution. HP35 is not yet a "solved" problem, but the validation or refutation of our current predictions will be a powerful step toward that goal.

## Methods

**Theory: A Model For Simulating TTET.** TTET experiments involve stimulating (via laser) a triplet state in a donor (xanthone). Through the short-range

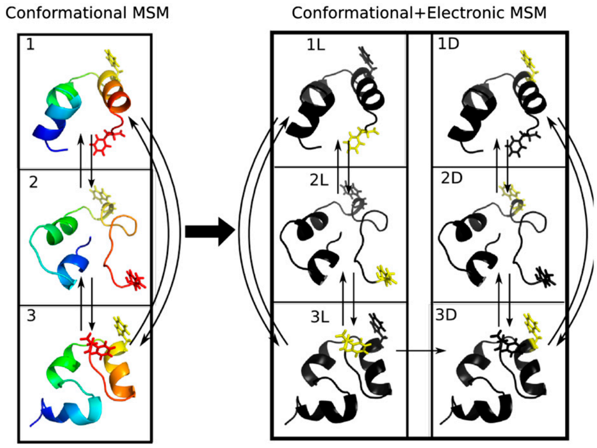


**Fig. 6.** A schematic shows the splitting of each conformational state (*Left*) into two distinct electronic states (*Right*). The two electronic states represent two different possible locations for the excited triplet state (yellow). Experiments that monitor a single wavelength will detect only those molecules where the triplet is on the observable fluorophore. Thus, states with the excitation on the donor we call "light" (L) and the other states we call "dark" (D). Allowed transitions are marked by black arrows. Only conformational state 3 has the donor and acceptor in close contact; thus, the transition L → D is allowed only in state 3.

Dexter mechanism, this triplet excitation is transferred to an acceptor (napthylalanine). The total population of the donor triplet state is recorded by monitoring the wavelength at which the donor triplet state decays. To simulate this experimental setup, we add an additional degree of freedom: a triplet state that starts on the donor and can transfer to the acceptor upon chromophore contact. Given an $n$-state MSM for the conformational dynamics of the protein, we model this effect by splitting each conformational state into two substates: one for each possible location of the triplet (Fig. 6). Because an experiment will typically monitor the wavelength of either the donor or acceptor, we call the two substates "light" and "dark." To parameterize this $2n$-state model requires $4n^2$ transition probabilities, which we can define with three assumptions. First, triplet transfer is irreversible. Second, excitation transfer only occurs within a given conformational state, because transitions between conformations modeled here occur on the 10-ns (or slower) timescale compared to the picosecond-timescale excitation transfer (12). Finally, the rate of triplet transfer within each conformational state is estimated using the distance between the donor and acceptor residues, quantified by transfer coefficients $f_i$. The transition matrix for the joint conformational-triplet dynamics is thus given by the following equations, where $P_0$is the conformational transition matrix and $\delta_{i,j}$ is the Kronecker delta:

$$P(i \to j,\text{dark} \to \text{dark}) = P_0(i \to j) \qquad P(i \to j,\text{dark} \to \text{light}) = 0$$

$$P(i \to j,\text{light} \to \text{light}) = (1 - f_i)P_0(i \to j)$$

$$P(i \to j,\text{light} \to \text{dark}) = \delta_{i,j}f_i.$$

The transfer coefficients $f_i$ for each state were estimated using a 3-state model similar to the one found in ref. 12—see *SI Methods*. To simulate each raw TTET trace, the system is initially placed in its conformational equilibrium with all triplets on the donor (light). The total population with the triplet state on the donor was then monitored as a function of time, leading to the decaying traces in Fig. 2. In Fig. 2 and Figs. S1 and S4, plotted TTET amplitudes were reduced to account for the intrinsic xanthone decay (see *SI Methods*), which occurs on a 10-μs timescale. We note previous work using molecular dynamics to simulate contact formation (34).

**Simulation Details.** The fast-folding double-norleucine mutant of HP35 (7) was simulated using Gromacs (35). Four independent datasets were used (ff03-TIP3P-300K, ff03-GBSA-300K, ff99SB-ILDN-TIP3P-300K, and ff99SB-ILDN-TIP3P-278K). Further details are provided in *SI Methods*.

**MSM Details.** Simulation data were sampled at 1-ns intervals and clustered using the $k$-centers algorithm (21). Clustering rmsd calculations included all nonequivalent heavy atoms. The number of states was chosen by terminating $k$-centers when cluster radii reached 1.75 Å. Computations were carried out using version 2.0 of MSMBuilder (21, 36).

**ACKNOWLEDGMENTS.** We thank Greg Bowman for discussions on MSMs. Folding@Home donors provided computer resources. MSM construction was performed on the BioX2 cluster (National Science Foundation Award CNS-0619926). R.D. is funded by a Burroughs–Wellcome Career Award at the Scientific Interface. K.B. is funded by a Stanford Graduate Fellowship. We thank the National Science Foundation for support through Frontiers in Integrative Biological Research Grants EF-0623664, MCB-0954714, and DMS-0900700; the National Institutes of Health for support through Grants R01-GM062868 and Simbios U54-GM072970; and the National Science Foundation Award CNS-0619926 for computer resources. We thank Thomas Kiefhaber for discussing his results prior to publication.

1. McKnight JC, Doering DS, Matsudaira PT, Kim PS (1996) A thermostable 35-residue subdomain within villin headpiece. *J Mol Biol* 260:126–134.
2. Kubelka J, Eaton WA, Hofrichter J (2003) Experimental tests of villin subdomain folding simulations. *J Mol Biol* 329:625–630.
3. Duan Y, Kollman PA (1998) Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science* 282:740–744.
4. Zagrovic B, Snow CD, Shirts MR, Pande VS (2002) Simulation of folding of a small alpha-helical protein in atomistic detail using worldwide-distributed computing. *J Mol Biol* 323:927–937.
5. Ensign DL, Kasson PM, Pande VS (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.
6. Freddolino PL, Schulten K (2009) Common structural transitions in explicit-solvent simulations of villin headpiece folding. *Biophys J* 97:2338–2347.
7. Kubelka J, Chiu TK, Davies DR, Eaton WA, Hofrichter J (2006) Sub-microsecond protein folding. *J Mol Biol* 359:546–553.
8. Bowman GR, Pande VS (2010) Protein folded states are kinetic hubs. *Proc Natl Acad Sci USA* 107:10890–10895.
9. Buscaglia M, Kubelka J, Eaton WA, Hofrichter J (2005) Determination of ultrafast protein folding rates from loop formation dynamics. *J Mol Biol* 347:657–664.
10. Brewer SH, et al. (2005) Effect of modulating unfolded state structure on the folding kinetics of the villin headpiece subdomain. *Proc Natl Acad Sci USA* 102:16662–16667.
11. Meng W, Shan B, Tang Y, Raleigh DP (2009) Native like structure in the unfolded state of the villin headpiece helical subdomain, an ultrafast folding protein. *Protein Sci* 18:1692–1701.
12. Reiner A, Henklein P, Kiefhaber T (2010) An unlocking/relocking barrier in conformational fluctuations of villin headpiece subdomain. *Proc Natl Acad Sci USA* 107: 4955–4960.
13. Lapidus LJ, Eaton WA, Hofrichter J (2000) Measuring the rate of intramolecular contact formation in polypeptides. *Proc Natl Acad Sci USA* 97:7220–7225.
14. Lindorff-Larsen K, et al. (2010) Improved side-chain torsion potentials for the amber ff99sb protein force field. *Proteins* 78:1950–1958.
15. Best RB, Hummer G (2009) Optimized molecular dynamics force fields applied to the helix-coil transition of polypeptides. *J Phys Chem B* 113:9004–9015.
16. Duan Y, et al. (2003) A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J Comput Chem* 24:1999–2012.
17. Friedrichs MS, et al. (2009) Accelerating molecular dynamic simulation on graphics processing units. *J Comput Chem* 30:864–872.
18. Shaw DE, et al. (2007) Anton, a special-purpose machine for molecular dynamics simulation. *ACM SIGARCH Computer Architecture News* 35:1–12.
19. Chodera JD, Singhal N, Pande VS, Dill KA, Swope WC (2007) Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics. *J Chem Phys* 126:155101.
20. Noé F, Fischer S (2008) Transition networks for modeling the kinetics of conformational change in macromolecules. *Curr Opin Struct Biol* 18:154–162.
21. Bowman GR, Beauchamp KA, Boxer G, Pande VS (2009) Progress and challenges in the automated construction of Markov state models for full protein systems. *J Chem Phys* 131:124101.
22. Noé F, Schütte C, Vanden-Eijnden E, Reich L, Weikl TR (2009) Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. *Proc Natl Acad Sci USA* 106:19011–19016.
23. Snow CD, Nguyen H, Pande VS, Gruebele M (2002) Absolute comparison of simulated and experimental protein-folding dynamics. *Nature* 420:102–106.
24. Vugmeyster L, McKnight CJ (2008) Slow motions in chicken villin headpiece subdomain probed by cross-correlated NMR relaxation of amide NH bonds in successive residues. *Biophys J* 95:5941–5950.
25. Camilloni C, et al. (2008) Urea and guanidinium chloride denature protein l in different ways in molecular dynamics simulations. *Biophys J* 94:4654–4661.
26. Myers JK, Pace CN, Scholtz JM (1995) Denaturant m values and heat capacity changes: Relation to changes in accessible surface areas of protein unfolding. *Protein Sci* 4:2138–2148.
27. Auton M, Holthauzen LMF, Bolen DW (2007) Anatomy of energetic changes accompanying urea-induced protein denaturation. *Proc Natl Acad Sci USA* 104:15317–15322.
28. Deuflhard P, Huisinga W, Fischer A, Schütte C (2000) Identification of almost invariant aggregates in reversible nearly uncoupled Markov chains. *Linear Algebra Appl* 315:39–59.
29. Lei H, Wu C, Liu H, Duan Y (2007) Folding free-energy landscape of villin headpiece subdomain from molecular dynamics simulations. *Proc Natl Acad Sci USA* 104: 4925–4930.
30. Rogers DM, Hirst JD (2004) First-principles calculations of protein circular dichroism in the near ultraviolet. *Biochemistry* 43:11092–11102.
31. McKnight CJ, Matsudaira PT, Kim PS (1997) NMR structure of the 35-residue villin headpiece subdomain. *Nat Struct Biol* 4:180–183.
32. Piana S, Lindorff-Larsen K, Shaw DE (2011) How robust are protein folding simulations with respect to force field parameterization? *Biophys J* 100:L47–L49.
33. Cellmer T, Buscaglia M, Henry ER, Hofrichter J, Eaton WA (2011) Making connections between ultrafast protein folding kinetics and molecular dynamics simulations. *Proc Natl Acad Sci USA* 108:6103–6108.
34. Yeh IC, Hummer G (2002) Peptide loop-closure kinetics from microsecond molecular dynamics simulations in explicit solvent. *J Am Chem Soc* 124:6563–6568.
35. Lindahl E, Hess B, van der Spoel D (2001) GROMACS 3.0: A package for molecular simulation and trajectory analysis. *J Mol Model* 7:306–317.
36. Bowman GR, Huang X, Pande VS (2009) Using generalized ensemble simulations and Markov state models to identify conformational states. *Methods* 49:197–201.

# SUPPORTING INFORMATION

Beauchamp et al. 10.1073/pnas.1010880108

## SI Methods

**Estimating Transfer Coefficients.** To estimate the transfer coefficients $f_i$, we modify the reaction scheme used in ref. 1 (scheme 1 and equations S4–S6 therein). Within each conformational state $S$, we assume that there are three conformational substates. State $S_i$ consists of those conformations that are triplet-triplet energy transfer (TTET)-inactive. State $S_a$ consists of those conformations that are TTET-active. Finally, state $S_d$ consists of the "dark" state that has already undergone triplet transfer from donor to acceptor. Assuming irreversible TTET, this model has three rate parameters: $k_a$ is the rate of finding the TTET-active substate, $k_i$ is the rate of finding the inactive substate, and $k_t$ is the TTET rate:

$$S_i \underset{k_i}{\overset{k_a}{\rightleftarrows}} S_a \xrightarrow{k_t} S_d.$$

In the reaction scheme above, the second step is irreversible and fast ($k_t \gg k_i, k_a$), occurring with a timescale of approximately 2 ps (1). Solving this three-state reaction provides the fraction of $S_d$ as a function of time:

$$S_d(t) = 1 + A_1 \exp(-\lambda_1 t) + A_2 \exp(-\lambda_2 t) \qquad \lambda_1 = k_t \qquad \lambda_2 = k_a$$

$$A_1 = \frac{k_a - \rho k_t}{k_t - k_a} \qquad A_2 = \frac{-(1-\rho)k_t}{k_t - k_a}.$$

In deriving this equation, we have also assumed that the initial conditions $[S_i](0)$ and $[S_a](0)$ are in relative equilibrium: $[S_a](0) = \rho = k_a/(k_a + k_i)$ and $[S_i](0) = (1-\rho) = k_i/(k_a + k_i)$. We are left with parameterizing the various rate and equilibrium constants. We fix $k_t = (2\ \mathrm{ps})^{-1}$ (1). From our simulation dataset, it is simple to estimate the equilibrium $\rho = k_a/(k_a + k_i)$ by counting the fraction of TTET-active conformations in each conformational state. Next, note that $k_a = \rho(k_a + k_i)$. We estimate $(k_a + k_i)^{-1}$ to be the lagtime of our model (1 ns). With these in place, we have now calculated the fraction of "dark" as a function of time. We assume that $S_d$ (1 ns) $= f$—that is, the transfer coefficient ($f$) is just the fraction of population that has gone dark after 1 lagtime (1 ns).

The remaining step is to classify conformations as TTET-active or TTET-inactive. Because our simulated constructs do not contain the fluorophore groups, we must use approximations (but see Fig. S2). We call a given conformation TTET-active if the donor and acceptor attachment points are located within 10 Å of one another. For the plots in Fig. 2, donor and acceptor positions are defined by Cβ positions, except for positions 0 and 7, which are defined by the N-terminal amino and the Lys7 amino, respectively. These choices were motivated by the chemical attachments used in ref. 1. To ensure a self-consistent set of predictions in Fig. 4, there we defined donor and acceptor positions using β-carbon locations (α-carbon distances for glycine residues). TTET rates are somewhat sensitive to the cutoff distance used; however, rates are not sensitive to the exact functional form used to calculate $f_i$. See Fig. S2 for further discussion and estimates of systematic error. We note that previous experimental work (2) estimated a reactive TTET boundary of 4.4 Å, not far from the value of 5 Å used in the present work.

**Modeling Intrinsic Xanthone Decay.** The model presented thus far does not consider the intrinsic xanthone decay present in real TTET experiments. However, a simple calculation can introduce this effect. First, assume that the xanthone-free TTET model was used to estimate the TTET amplitude $x(t)$. Let $y(t)$ be the amplitude with xanthone decay, and let $d(t)$ be the cumulative amount of xanthone decay. It follows that $y(t) = x(t) - d(t)$. The rate of xanthone decay will be proportional to the remaining undecayed population; thus, $d'(t) = ky(t)$. The resulting ordinary differential equation can be solved via integrating factors, leading to the result

$$d(t) = k \exp(-kt) \int_0^t x(t') \exp(kt') dt.$$

Because $x(t)$ is bounded between 0 and 1, a useful approximation is to assume that $x(t') = x(t)$ within the integral. This simplification leads to

$$d(t) = x(t)(1 - \exp(-kt)).$$

Inserting this equation into the equation for $y(t)$ yields

$$y(t) = x(t) \exp(-kt).$$

This approach was used in Fig. 2 and Figs. S1, S2, and S4. The xanthone lifetime was assumed to be 10 μs (1).

**Simulation Details.** The present work used four independent sets of simulations:

Dataset 1 is a subset of a previously described dataset (3) and was performed using the ff03 force field and the TIP3P water model. Electrostatics were performed using the reaction field formalism, and temperature coupling (300 K) was achieved through the Nose–Hoover thermostat. Simulations started either from the crystal structure [Protein Data Bank (PDB) ID code 2F4K] or from nine unfolded structures. The unfolded structures were selected from 373 K unfolding simulations begun at the crystal structure. Three hundred eighty-six simulations of minimum length 500 ns were used (total aggregate sampling of 300 μs); snapshots were stored every 1 ns. With a 1.75-Å termination diameter, $k$-centers identified 178,744 states from this data.

Dataset 2 used the ff03 force field, GBSA implicit solvent, and stochastic dynamics with water-like viscosity. One hundred simulations were begun at the crystal structure; each simulation was 5–20 μs long (total aggregate sampling of 1,000 μs). With a 1.75-Å termination diameter, $k$-centers identified 39,714 states from this data.

Dataset 3 used the ff99SB-ILDN (4) force field and TIP3P water. Electrostatics were calculated using particle mesh Ewald (5). Temperature coupling (300 K) was achieved via the recent velocity rescaling algorithm (6) with a coupling time of 100 fs. Eight hundred sixty-one folded and unfolded conformations were used as starting conformations; each simulation was at least 500 ns. Snapshots were stored every 1 ns. An aggregate of 300 μs was used. Because short simulations were started from a large number of conformations, not all trajectories intersect one another in conformation space; isolated trajectories were discarded. Such trajectories tend to lie in the unfolded state and have little effect on the near-native dynamics probed in this work. Similar trimming procedures were used in ref. 7. With a 1.75-Å termination diameter, $k$-centers identified 124,068 states from this data.

Dataset 4 was similar to Dataset 3, except with the temperature set to 278 K. Nine hundred sixty-five trajectories of minimum length 500 ns provided an aggregate 300 μs of data. With a 1.75-Å termination diameter, $k$-centers identified 107,580 states from this data.

For all four datasets, microscopic reversibility was assumed by symmetrizing the estimated count matrices (8). After trimming, states with poor statistics (<3 observations) were merged with the state immediately preceding in time. For consistency, the same trimming procedure was used on all four datasets.

**Model Reduction with Perron Cluster Cluster Analysis (PCCA).** To produce models with a reduced number of states, we estimated transition matrices using a 10-ns lagtime and used the PCCA procedure to construct models with 1,000 states. This number of states was chosen to reduce the number of states without creating states with too much internal heterogeneity. From the distribution of eigenvalues (e.g., timescales), we estimate that the 1,000 state models have lumped all dynamics occurring on timescales faster than 20 ns.

**Estimating the Effect of Fluorophores.** Neglecting the effect of the fluorophores contributes uncertainty both from stability differences and from limitations in the geometric model for TTET. To better understand these errors, we ran four short simulations (GBSA, ff99SB-ILDN-300K) that included N-linked 9-oxoxanthone and napthylalanine at position 35 (Fig. S2*A*). Force field parameters were calculated using generalized amber force field (GAFF) and antechamber. Note that by starting all simulations near the crystal structure with the TTET pair in close proximity, these estimates may be biased toward fast TTET as compared to the comprehensive datasets used in the text. With an aggregate of approximately 400 ns of simulation, it is possible to estimate the near-native (0–35) TTET traces. For simulations with the dyes attached, we compared three models for estimating TTET. The first (Cβ) is the model used in main text Fig. 4, and defines TTET in terms of Cβ distance between the residues where the donor and acceptor are to be attached. The second (Amino) is the model used in main text Fig. 2, and defines TTET in terms of distance between the N-terminal amino and the C-terminal Cβ. The third [van der Waals (VDW)] model is motivated by the Dexter TTET mechanism; in this model, TTET activity is judged by whether a conformation has donor and acceptor in VDW contact. VDW contact is defined by a 6-Å cutoff between any heavy atom on the xanthone or napthylalanine rings (e.g., the minimum ring-to-ring distance). As a control, we analyzed four dye-free trajectories (GBSA, ff99SB-ILDN-300K) started from the same conformation.

**Estimating the Thermodynamics of GdmCl Denaturation.** The *m*-value equation leads to the following formula for the free energy difference between two Markov state model (MSM) states, where $c$ is the denaturant concentration:

$$G_i(c) - G_j(c) = \Delta G_{ij}(c) = \Delta G(0) - cm_{ij}.$$

The observed correlation between *m*-value and solvent accessible surface area (SASA) (9) provides an avenue for modeling denaturant effects. Further, this model assumes that the *m*-values are proportional to surface area differences between states:

$$m_{ij} = b\Delta A_{ij},$$

where $\Delta A_{ij}$ is the SASA difference between states $i$ and $j$. Inserting the equation for *m*-value, we have that $\Delta G_{ij}(c) = \Delta G_{ij}(0) - c(b\Delta A_{ij})$ Thus, denoting the equilibrium populations of each state by $\pi_i(c)$, we have:

$$\pi_i(c) = (1/Z(c))\pi_i(0)\exp(-bcA_i).$$

With this formula, one can estimate the denaturant dependence of each state's equilibrium population. From ref. 9, $b = 0.22\ \mathrm{kcal}/(\mathrm{mol}^2\ \mathrm{nm}^2)$. To estimate the SASA of each MSM state, we used the gromacs g_sas tool to calculate the SASA of every conformation in the simulation dataset, averaging the results to find the mean SASA of each conformational state.

The simple *m*-value SASA correlation is a widely used model, but it may not capture all the relevant details of protein denaturation. Thus, we repeated calculations using the more sophisticated group transfer free energy (GTFE) model (10). In that model, the SASA of each backbone and side-chain group is considered separately, with group transfer free energy parameters that were fitted to experiments on model peptides. The literature is currently missing certain parameters necessary for calculations on GdmCl; to avoid this limitation, we used urea parameters with magnitudes multiplied by two, derived from the relative strength of GdmCl and urea as quantified by the slopes of the *m*-value correlations in ref. 9. Once again, the gromacs g_sas tool was used to calculate the backbone and side-chain SASA for each protein residue for each conformation in our datasets. Reference SASAs for glycine peptides were taken from ref. 11.

**A Perturbation Theory for MSM Kinetics.** To model the dependence of contact formation timescale on chemical denaturant, we carried out perturbation analysis of the starting MSM dynamics. The perturbation is controlled by a single parameter $c$, here representing the denaturant concentration. It is necessary to estimate $n^2$ rates from limited information. The known quantities are the rates $K$ (derived from the transition matrix T; see *Estimating Instantaneous Rate Matrices* below) when $c = 0$ and the equilibrium populations $\pi'$ for all values of $c$.

Using the language of transition state theory (12), rates can be expressed in terms of free energy barriers:

$$K'_{ij} = K_{ij}\exp\left(-\frac{\Delta\Delta G^*_{ij}}{kT}\right).$$

Here, $\Delta\Delta G^*_{ij}$ is the difference in barrier heights between state $i$ and $j$ relative to the unperturbed barrier height. The challenge is to estimate the barrier height differences from limited data. The assumed knowledge derives from the unperturbed rate matrix and the known equilibrium populations $\pi'_i = \pi_i\exp(-\frac{\Delta\Delta G_i}{kT})$ in the perturbed ensemble, where $\Delta\Delta G_i$ is the free energy of state $i$ in the perturbed ensemble, relative to the unperturbed ensemble. With $n^2$ unknown parameters to estimate from only $n$ equilibrium populations, modeling the perturbed dynamics requires some approximation. Here, we estimate the error in this approximation by applying three plausible models for relating energy barriers to thermodynamic changes. By construction, all three models maintain detailed balance and the correct equilibrium populations in the perturbed ensemble. The first model ("even") assumes that the barrier heights are changed additively by the free energy difference of the initial and final states:

$$\Delta\Delta G^*_{ij} = -\frac{1}{2}(\Delta\Delta G_j - \Delta\Delta G_i).$$

The second model ("unfolded-like") assumes that transition states approximate the higher energy state of the initial and final states. Thus,

$$\Delta\Delta G^*_{ij} = \begin{cases} \Delta\Delta G_j - \Delta\Delta G_i & \text{if } \pi'_i < \pi'_j \\ 0 & \text{else} \end{cases}.$$

The third model ("folded-like") assumes that transition states approximate the lower energy state of the initial and final states. Thus,

$$\Delta\Delta G^*_{ij} = \begin{cases} 0 & \text{if } \pi'_i < \pi'_j \\ \Delta\Delta G_j - \Delta\Delta G_i & \text{else} \end{cases}.$$

Once the perturbed rate matrix $K'$ is known, it is straightforward to construct a transition matrix via $T' = \exp(-t_l K')$. The three functional forms above show agreement in their predictions for contact formation dynamics (see Fig. S4).

**Estimating Instantaneous Rate Matrices.** For the MSM perturbation theory of the previous section, we require an instantaneous rate matrix. Estimating the rate matrix $K$ is a challenge, as the logarithm is defined implicitly via the inverse relation $T = \exp(-tK)$ (13, 14). Approaches to this problem involving matrix polynomials are not applicable here because of convergence issues and constraints on valid rate matrices. We instead use the following approach. First, we calculate the left and right eigenvectors $\varphi_i$ and $\psi_i$. The eigenvectors of a transition matrix are not orthogonal but rather $D$-orthogonal (e.g., ${\varphi_i}^T D \psi_j = \delta_{ij}$, where $D$ is the diagonal matrix of inverse populations). Using this fact, we express MSM dynamics as an expansion over eigenvalues $\lambda_k$:

$$Tx = \sum_k \lambda_k \phi {\psi_k}^T Dx.$$

This yields the following expression for $T$:

$$T = \sum_k \lambda_k \phi_k {\psi_k}^T D.$$

Because the $T$ and $K$ have the same eigenvectors, we now have a formula for the rate matrix:

$$K = \sum_k -\frac{\log(\lambda_k)}{t_l} \phi_k {\psi_k}^T D.$$

Because of discrete sampling, some eigenvalues were negative; for those, we took their absolute value to ensure no complex values in the final rate matrix. After calculation of the rate matrix, all nondiagonal negative entries were set to zero, and the diagonal entries were then set to $K_{ii} = -\sum_{k \neq i} K_{ij}$. Finally, for calculations that perturbed rate matrix elements ($K$ to $K'$; see *A Perturbation Theory for MSM Kinetics* above), we further regularized the effect of shot noise on the perturbed transition matrix $T' = \exp(-tK')$. We calculated the symmetric count matrix $X' = T' D^{-1}$, where $D$ is a diagonal matrix of inverse populations. We then truncated the count matrix $X'$, deleting all entries less than $10^{-6}$; the threshold was chosen to account for the $5 \times 10^5$ observations in the dataset. After truncation, the count matrix was symmetrized and converted into the final transition matrix $T'_{\mathrm{regularized}}$. In practice, the regularization produced little change. As a control, we calculated the difference between the original $T$ and recalculated $T'_{\mathrm{regularized}}$ in a mock calculation with no rate perturbations ($K' = K$), and we confirmed that transition probabilities changed by an average of 1%.

**Availability of TTET Traces.** Raw TTET contact formation traces for all residue positions are provided at (https://simtk.org/home/hp35-datasets/). Traces are included for the models constructed from each of the four datasets. TTET transfer coefficients were calculated using a 10-Å Cβ distance cutoff (Cα for glycine residues). Traces were continued until the TTET amplitude dropped below 25%.

1. Reiner A, Henklein P, Kiefhaber T (2010) An unlocking/relocking barrier in conformational fluctuations of villin headpiece subdomain. *Proc Natl Acad Sci USA* 107:4955–4960.
2. Möglich A, Joder K, Kiefhaber T (2006) End-to-end distance distributions and intrachain diffusion constants in unfolded polypeptide chains indicate intramolecular hydrogen bond formation. *Proc Natl Acad Sci USA* 103:12394–12399.
3. Ensign DL, Kasson PM, Pande VS (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.
4. Lindorff-Larsen K, et al. (2010) Improved side-chain torsion potentials for the amber ff99sb protein force field. *Proteins* 78:1950–1958.
5. Essmann U, et al. (1995) A smooth particle mesh Ewald method. *J Chem Phys* 103:8577–8593.
6. Bussi G, Donadio D, Parrinello M (2007) Canonical sampling through velocity rescaling. *J Chem Phys* 126:014101.
7. Noé F, Schütte C, Vanden-Eijnden E, Reich L, & Weikl TR (2009) Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. *Proc Natl Acad Sci USA* 106:19011–19016.
8. Bowman GR, Beauchamp KA, Boxer G, Pande VS (2009) Progress and challenges in the automated construction of Markov state models for full protein systems. *J Chem Phys* 131:124101.
9. Myers JK, Pace CN, Scholtz JM (1995) Denaturant m values and heat capacity changes: Relation to changes in accessible surface areas of protein unfolding. *Protein Sci* 4:2138–2148.
10. Auton M, Holthauzen LMF, Bolen DW (2007) Anatomy of energetic changes accompanying urea-induced protein denaturation. *Proc Natl Acad Sci USA* 104:15317–15322.
11. O'Brien EP, Ziv G, Haran G, Brooks BR, Thirumalai D (2008) Effects of denaturants and osmolytes on proteins are accurately predicted by the molecular transfer model. *Proc Natl Acad Sci USA* 105(36):13403–13408.
12. Truhlar DG, Garrett BC, Klippenstein SJ (1996) Current status of transition-state theory. *J Phys Chem* 100:12771–12800.
13. Higham NJ (2001) Evaluating Padé approximants of the matrix logarithm. *SIAM J Matrix Analysis Applications* 22:1126–1135.
14. Moler C, Van Loan C (1978) Nineteen dubious ways to compute the exponential of a matrix. *SIAM Rev* 20:801–836.
15. Shell MS, Ritterson R, Dill KA (2008) A test on peptide stability of amber force fields with implicit solvation. *J Phys Chem B* 112:6878–6886.

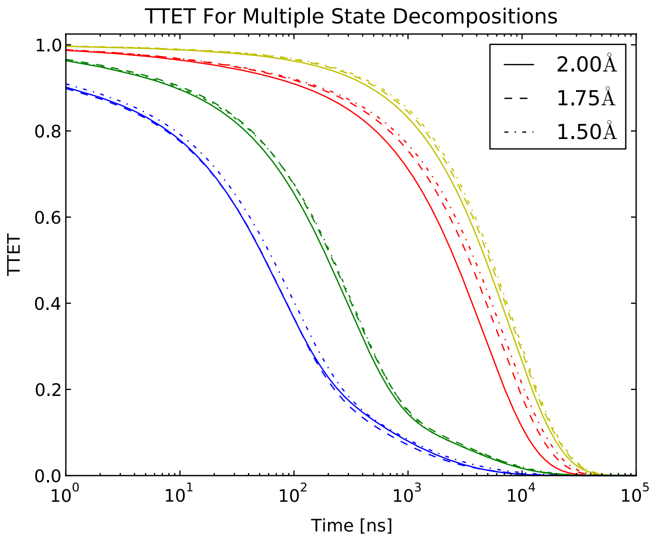


**Fig. S1.** Uncertainty due to MSM construction. TTET traces calculated with three MSMs show the uncertainty associated with MSM construction and sampling. The three MSMs were constructed using *k*-centers termination diameters of 1.5, 1.75, and 2.0 Å. Calculations were performed on the ff99SB-ILDN-TIP3P-300K dataset. The differences in timescale (1/*e*) between the different MSMs are approximately 20–50% and are smaller than the other uncertainties involved in TTET calculation. Calculations using the other datasets lead to similar uncertainties. For the faster timescale TTET processes, sampling uncertainty is generally small; for the slowest processes, sampling uncertainty is larger due to the rarity of observing contact formation events.

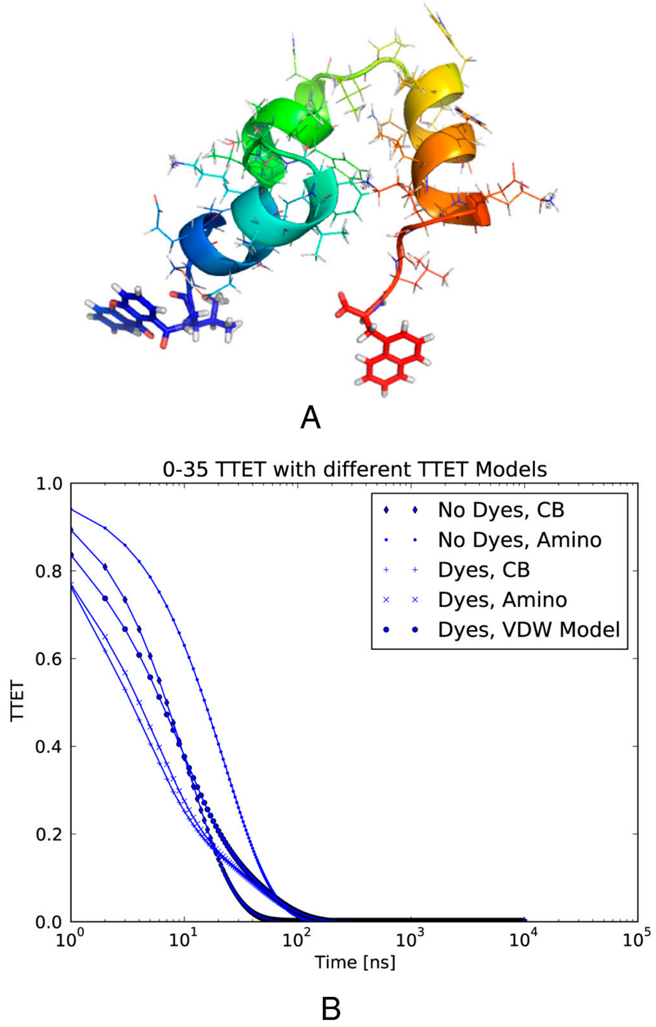


**Fig. S2.** Effect of fluorophore group on 0–35 TTET. (*A*) Structure of HP35 with fluorophores attached at position 0–35. (*B*) Native-state (0–35) TTET with and without fluorophores. Comparing the Cβ, amino, and VDW models suggests that the error associated with the simpler models (Cβ, amino) is approximately a factor of 2.0. Simulations with the dyes attached showed faster TTET than the simulations of the dye-free construct, due to hydrophobic interactions between the dyes. Estimates using the Cβ model suggest that the presence of dye accelerates TTET by twofold. Similarly, estimates using the amino model suggest that the presence of dye accelerates TTET by threefold. Circular dichroism experiments find that the 0–35 TTET construct is up to 0.5 kcal/mol more stable than the other three constructs, so the observed rate enhancement is expected. Overall, neglecting the dyes and using the simpler models for TTET lead to predictions that are up to 2× slower than the fluorophore-containing simulations with the VDW model. We emphasize that this error is a worst-case scenario, because the near-native (0–35) TTET involves the precise packing of the terminally attached fluorophores, whereas the other three TTET processes involve greater distances. Other functional forms for the TTET-distance relationship led to similar results. The TTET in this figure is faster than that in Fig. 2 because of limited sampling—with only 400 ns of aggregate sampling, these control simulations explore a limited subset of conformational space.

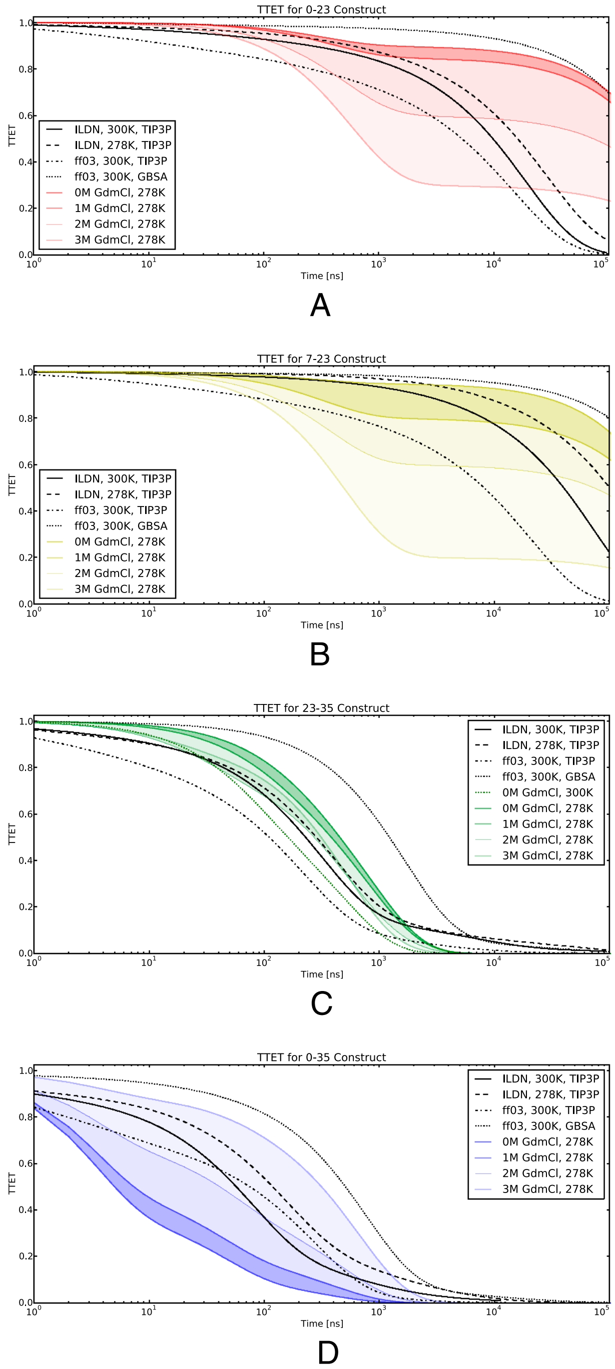


**Fig. S3.** TTET without xanthone decay. This figure is similar to Fig. 2, except here the lifetime of xanthone has been set to infinity. Because of the 10-µs xanthone lifetime, experimental TTET studies are typically limited to monitoring nanosecond to microsecond fluctuations. Simulations, however, are not limited by this. Here, we perform an "idealized" TTET experiment where the donor has infinite lifetime. Unhindered by the xanthone lifetime, the slowest timescales in HP35 range into the hundreds of microseconds, comparable to the unfolding rates measured by temperature jump studies.

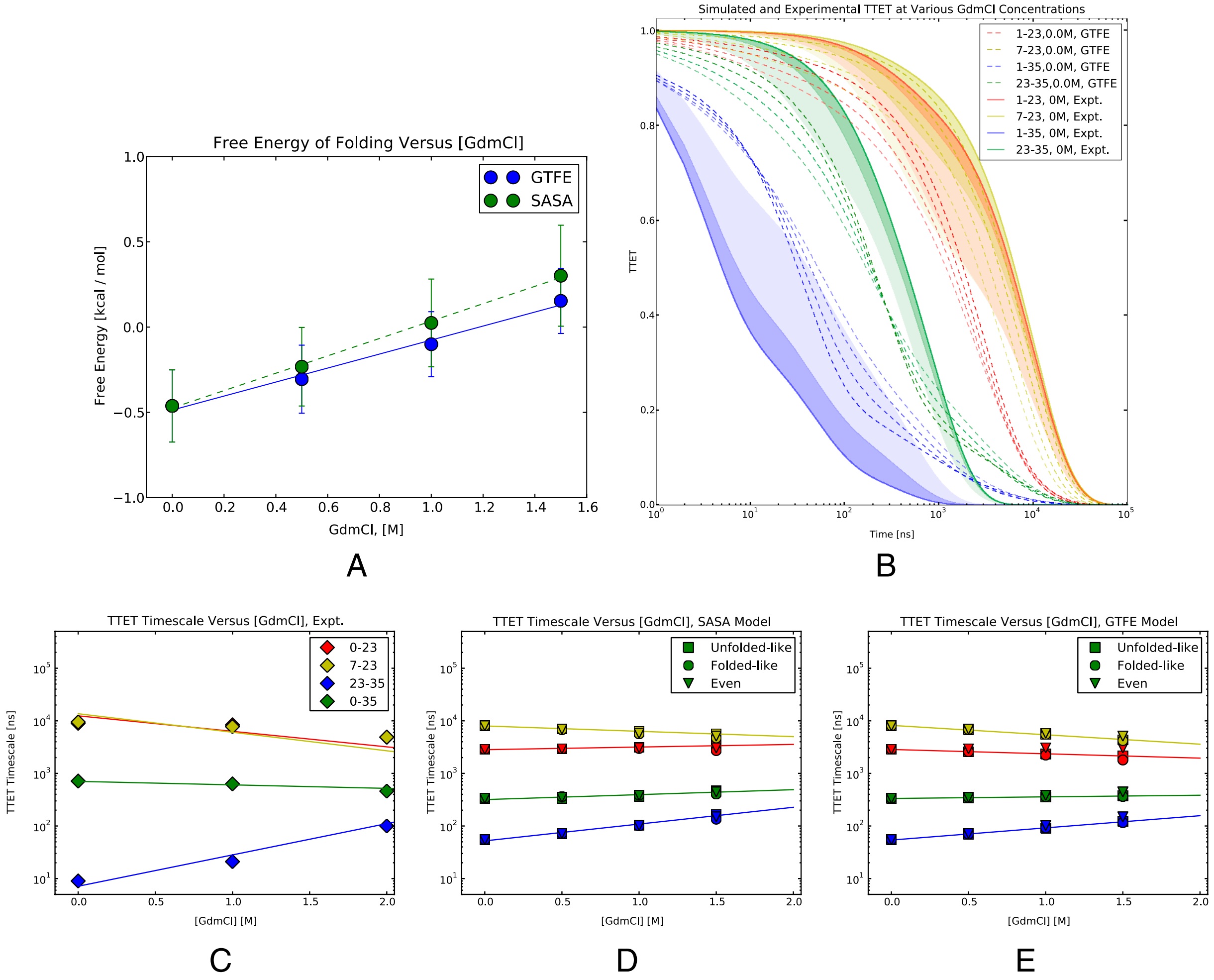
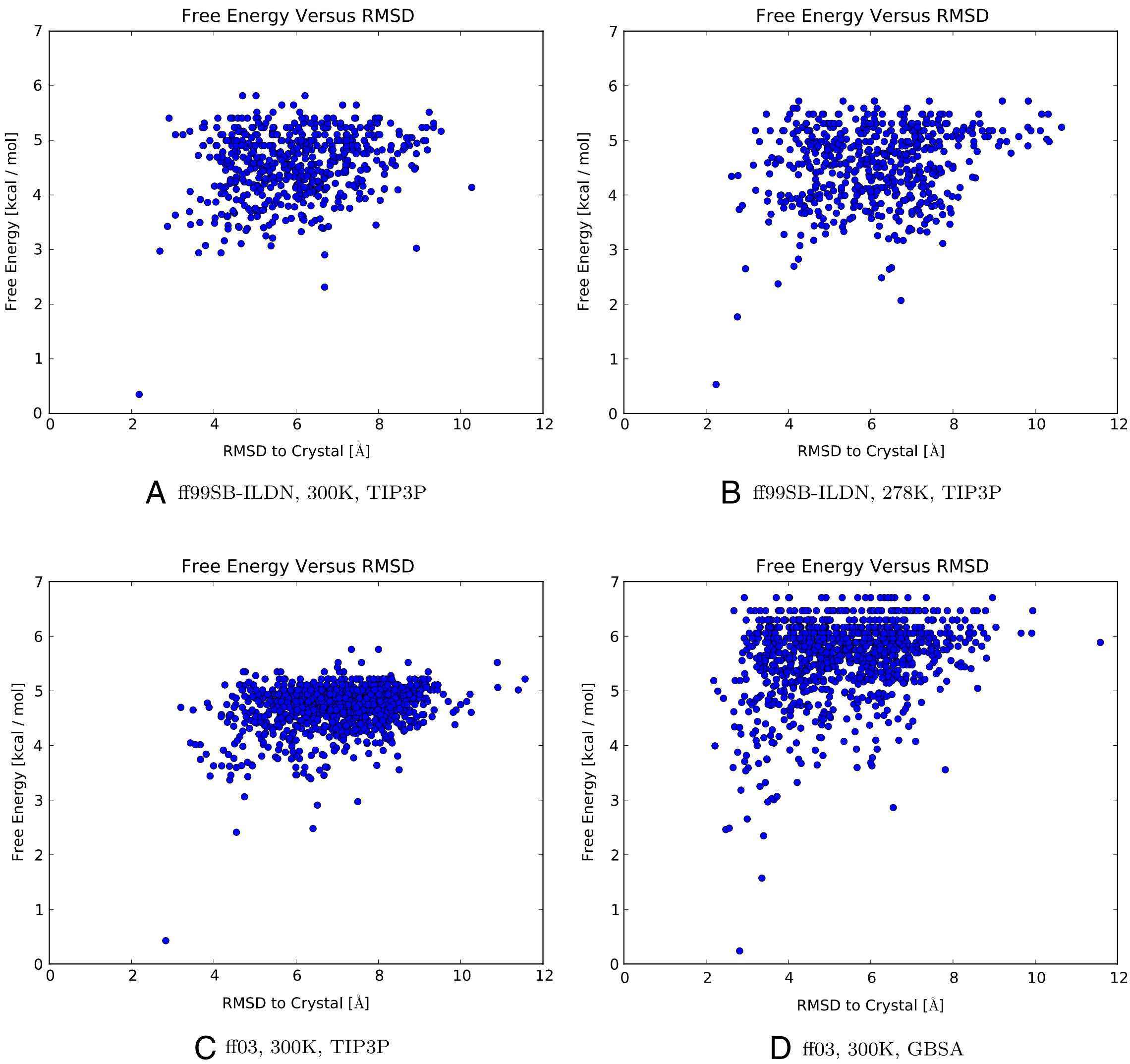


**Fig. S4.** Denaturant dependence of predictions. (*A*) The free energy of folding as a function of [GdmCl] provides a validation for the current model of denaturant. Linear fits suggest an *m*-value of $0.53 \pm 0.2$ kcal $\cdot$ mol$^{-1}$ $\cdot$ M$^{-1}$ for the SASA model and $0.46 \pm 0.2$ kcal $\cdot$ mol$^{-1}$ $\cdot$ M$^{-1}$ for the GTFE model. The folded state is defined here by all conformations with rmsd (to PDB ID code 2F4K) of under some cutoff: The three data series use rmsd cutoffs of 4.0, 5.0, and 6.0 Å, respectively. (*B*) Full TTET traces. The full TTET traces at various values of [GdmCl]. The darkest traces correspond to zero-denaturant condition. Several trends are observed, all consistent with the experimental results. First, the 0–35 TTET decelerates with increasing denaturant, as is seen experimentally (1). The unfolding processes (0–23) and (7–23) become faster with increasing denaturant, again qualitatively consistent with experimental observations (1). For the 23–35 TTET process, both simulation and experiment predict a nonmonotonic behavior in the TTET timescale. For simplicity, simulated TTET traces are shown for only the GTFE denaturant model with the "folded-like" transition state model. (*C*) The experimentally measured timescales ($1/e$) for each of the four TTET processes as a function of [GdmCl]. (*D*) Simulation timescale ($1/e$) results using the SASA model of denaturant. (*E*) Simulation timescales ($1/e$) using the GTFE model of denaturant. Simulation results are shown using all three transition state models.



A ff99SB-ILDN, 300K, TIP3P

B ff99SB-ILDN, 278K, TIP3P

C ff03, 300K, TIP3P

D ff03, 300K, GBSA

**Fig. S5.** Thermodynamics and structure of HP35 from reduced models. The free energy of each state summarizes the equilibrium thermodynamics observed in simulations. The most populated conformations structurally resemble the experimental crystal structure (PDB ID code 2F4K), as measured by Cα rmsd. Data is shown for each of the four datasets.

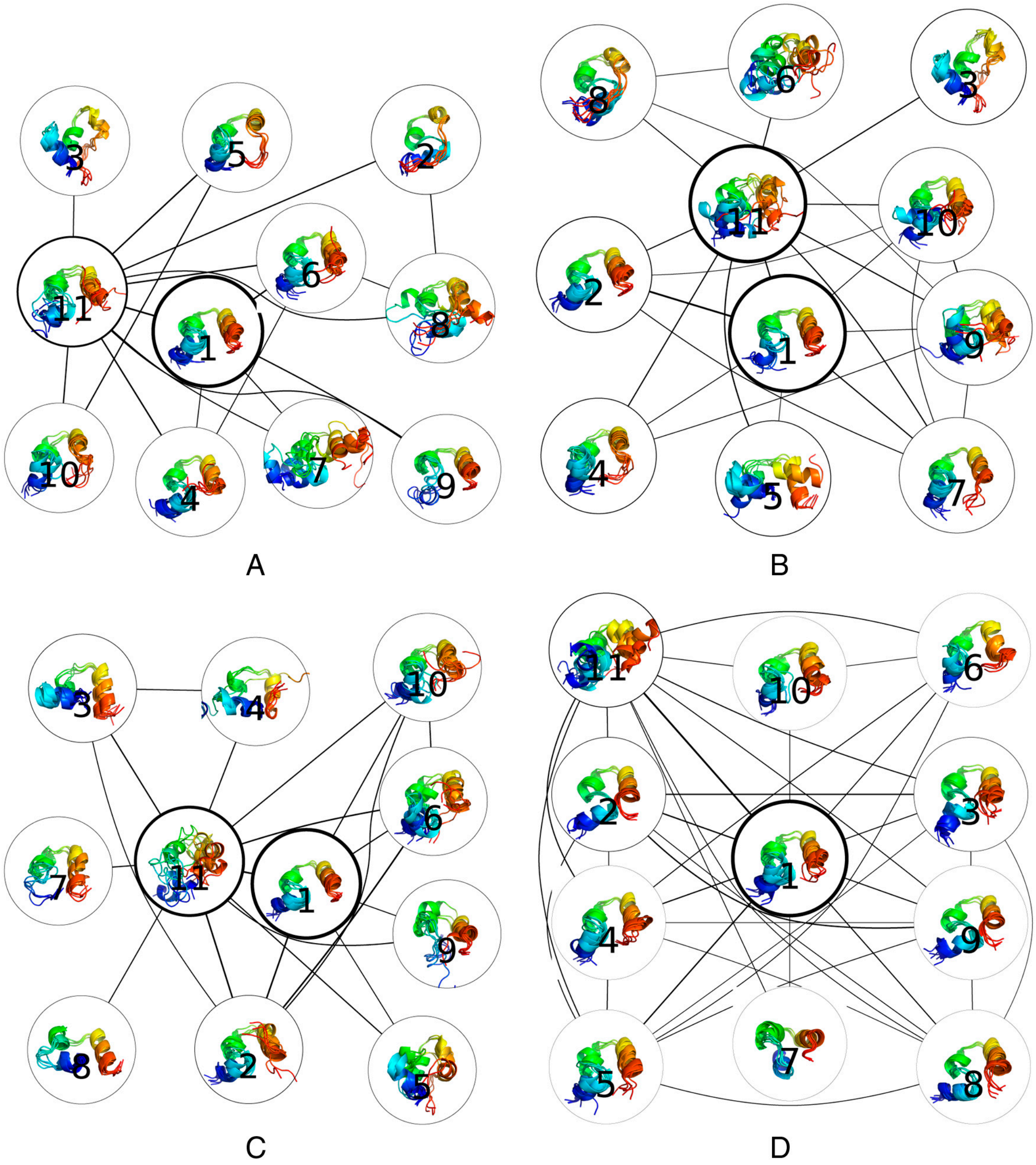


**Fig. S6.** Populated states in reduced models. Five random conformations are displayed from the ten most populated states for each of the four reduced models. The graph edges are the symmetric counts from each of the states. The line widths represent the log of the number of counts. The darkened outline around each state represents the log population of each state. In addition to the 10 most populated states, the remaining conformations in each dataset were grouped into "state" 11; thus, these figures summarize all the conformations in each dataset.

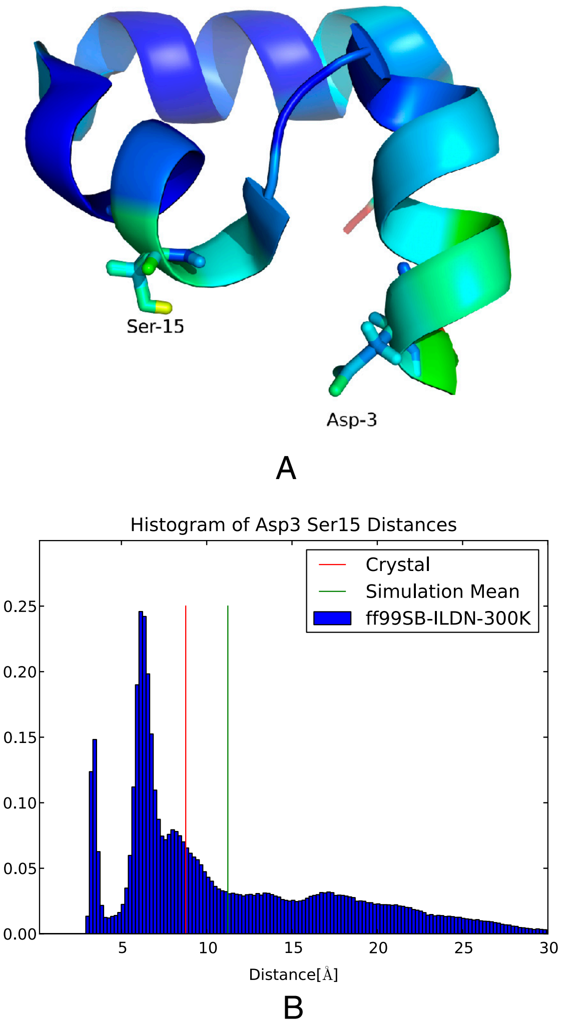


**Fig. S7.** A predicted Asp3 Ser15 contact. (*A*). The experimental model (PDB ID code 2F4K) is displayed in cartoon representation, colored by temperature factor. The locations of high temperature factors (cyan-green-red) correspond to the locations of variation observed in simulations. The N terminus, residues ASP3 and SER15 (shown as sticks), and the C terminus all show elevated temperature factors. (*B*). Simulations suggest that near-native fluctuations of HP35 involve a number of adjustments in the terminal helices. Here, the Asp3-Ser15 distances are plotted in a histogram. The simulation ensemble suggests at least two possible states. It is worth noting that the crystallographic distance and simulation ensemble average distance are similar. However, the actual distribution shows several peaks, suggesting that sensitive experimental techniques might detect multiple conformational states via this probe.

**Table S1. Folding free energies**

| Dataset/rmsd cutoff | ff99SB-ILDN-TIP3P-300K | ff99SB-ILDN-TIP3P-278K | ff03-TIP3P-300K | ff03-GBSA-300K |
|---|---|---|---|---|
| Free energy [300 K, (kcal/mol) 4.0 Å rmsd cutoff] | −0.18 | −0.09 | 0.08 | −1.16 |
| Free energy [300 K, (kcal/mol) 5.0 Å rmsd cutoff] | −0.46 | −0.38 | −0.07 | −1.51 |
| Free energy [300 K, (kcal/mol) 6.0 Å rmsd cutoff]) | −0.70 | −0.65 | −0.28 | −1.80 |

Using the rmsd to crystal structure as an order parameter, we estimate the free energy of folding for the four datasets used in this study. The folded state was defined as all MSM states with mean rmsd to crystal structure less than some cutoff value. Cutoffs of 3, 4, and 5 Å were used. The largest sources of error in these calculations are force field differences and limited sampling; one way to control sampling error is to consider datasets with different starting conformations. The ff99SB-ILDN datasets had approximately 25% of trajectories starting from near-native conformations. The ff03-TIP3P dataset had approximately equal amounts of data started from unfolded and folded conformations. The ff03-GBSA dataset had all trajectories start from the native conformations; because the GBSA trajectories are quite long (>5 μs), they nonetheless have sufficient time to explore the near-native dynamics probed in this work. Because the ff03-GBSA-300K trajectories were all started in the folded state, that estimate of free energy is likely a lower bound. Including additional trajectories started from unfolded states tends to shift the free energy toward the unfolded state. Consistent with the comparison to 0–35 TTET, these datasets describe HP35 in a state that is folded but close to folding midpoint.

## Table S2. State properties of reduced MSM

| Dataset | State | Population | −kT log(P), kcal/mol | Rmsd, Å | 7–23 $f_i$ | 1–23 $f_i$ | 1–35 $f_i$ | 23–35 $f_i$ | SASA, nm$^2$ |
|---|---|---|---|---|---|---|---|---|---|
| ILDN-300K-TIP3P | 1 | 0.555 | 0.349 | 2.185 | 0.000 | 0.000 | 0.718 | 0.177 | 28.338 |
| | 2 | 0.020 | 2.314 | 6.693 | 0.000 | 0.000 | 0.000 | 0.000 | 28.554 |
| | 3 | 0.007 | 2.903 | 6.696 | 0.000 | 0.000 | 0.000 | 0.061 | 29.720 |
| | 4 | 0.007 | 2.940 | 4.175 | 0.000 | 0.000 | 0.000 | 0.000 | 29.678 |
| | 5 | 0.007 | 2.942 | 3.631 | 0.000 | 0.000 | 1.000 | 0.000 | 29.226 |
| | 6 | 0.007 | 2.972 | 2.688 | 0.000 | 0.000 | 0.000 | 0.000 | 29.304 |
| | 7 | 0.006 | 3.025 | 8.925 | 0.000 | 0.000 | 0.000 | 0.000 | 30.078 |
| | 8 | 0.006 | 3.069 | 5.389 | 0.000 | 0.000 | 0.000 | 0.000 | 29.987 |
| | 9 | 0.006 | 3.075 | 3.797 | 0.000 | 0.000 | 0.000 | 0.000 | 29.379 |
| | 10 | 0.005 | 3.108 | 4.663 | 0.000 | 0.000 | 0.000 | 0.000 | 30.059 |
| ILDN-278K-TIP3P | 1 | 0.409 | 0.530 | 2.240 | 0.000 | 0.000 | 0.295 | 0.000 | 28.393 |
| | 2 | 0.051 | 1.768 | 2.760 | 0.000 | 0.000 | 0.257 | 0.000 | 28.765 |
| | 3 | 0.030 | 2.069 | 6.736 | 0.000 | 0.000 | 0.598 | 0.000 | 29.769 |
| | 4 | 0.018 | 2.373 | 3.747 | 0.021 | 0.000 | 0.000 | 0.021 | 29.123 |
| | 5 | 0.015 | 2.484 | 6.263 | 0.000 | 0.000 | 0.000 | 0.000 | 29.489 |
| | 6 | 0.012 | 2.643 | 6.453 | 0.000 | 0.000 | 0.000 | 0.295 | 29.545 |
| | 7 | 0.011 | 2.650 | 2.957 | 0.000 | 0.000 | 0.205 | 0.000 | 29.334 |
| | 8 | 0.011 | 2.668 | 6.508 | 0.000 | 0.000 | 0.000 | 0.000 | 28.981 |
| | 9 | 0.011 | 2.695 | 4.134 | 0.000 | 0.000 | 0.814 | 0.711 | 29.789 |
| | 10 | 0.008 | 2.826 | 4.242 | 0.000 | 0.000 | 0.000 | 0.000 | 29.948 |
| ff03-300K-TIP3P | 1 | 0.487 | 0.427 | 2.830 | 0.000 | 0.000 | 0.000 | 0.000 | 28.858 |
| | 2 | 0.017 | 2.411 | 4.549 | 0.000 | 0.000 | 1.000 | 0.000 | 30.067 |
| | 3 | 0.015 | 2.480 | 6.410 | 0.000 | 0.000 | 0.000 | 0.000 | 29.461 |
| | 4 | 0.007 | 2.908 | 6.516 | 0.000 | 0.000 | 0.000 | 0.000 | 29.612 |
| | 5 | 0.007 | 2.971 | 7.498 | 0.000 | 0.000 | 0.000 | 0.000 | 29.906 |
| | 6 | 0.006 | 3.061 | 4.745 | 0.000 | 0.000 | 0.000 | 0.000 | 30.140 |
| | 7 | 0.003 | 3.369 | 4.387 | 0.000 | 0.000 | 0.000 | 0.000 | 28.577 |
| | 8 | 0.003 | 3.390 | 6.358 | 0.000 | 0.000 | 0.000 | 0.000 | 30.164 |
| | 9 | 0.003 | 3.421 | 6.310 | 0.000 | 0.000 | 0.000 | 0.000 | 28.972 |
| | 10 | 0.003 | 3.430 | 4.824 | 0.000 | 0.000 | 0.000 | 1.000 | 29.969 |
| ff03-300K-GBSA | 1 | 0.668 | 0.239 | 2.813 | 0.000 | 0.000 | 1.000 | 0.000 | 29.707 |
| | 2 | 0.070 | 1.572 | 3.355 | 0.000 | 0.000 | 1.000 | 0.000 | 29.744 |
| | 3 | 0.019 | 2.347 | 3.397 | 0.000 | 0.000 | 0.000 | 0.000 | 30.081 |
| | 4 | 0.016 | 2.461 | 2.478 | 0.000 | 0.000 | 1.000 | 0.000 | 29.369 |
| | 5 | 0.015 | 2.486 | 2.561 | 0.000 | 0.000 | 0.000 | 0.000 | 29.581 |
| | 6 | 0.011 | 2.655 | 2.998 | 0.000 | 0.000 | 0.000 | 0.000 | 28.624 |
| | 7 | 0.008 | 2.863 | 6.548 | 0.000 | 0.000 | 0.000 | 0.000 | 27.763 |
| | 8 | 0.007 | 2.967 | 3.499 | 0.000 | 0.000 | 0.000 | 0.000 | 30.028 |
| | 9 | 0.006 | 3.006 | 3.649 | 0.000 | 0.000 | 0.000 | 0.000 | 29.921 |
| | 10 | 0.006 | 3.026 | 3.594 | 0.000 | 0.000 | 0.000 | 0.000 | 28.649 |

A 1,000-state reduced MSM was constructed from each dataset. Here, we summarize key features of the 10 most populated states in each reduced model. First, the most populated states are characterized by strong similarity to the crystal structures, with the primary difference being slight heterogeneity in the orientation and packing of the terminal residues. In particular, the simulation structures suggest that the crystallographically observed Leu1-Phe35 interaction may be weak at 300 K. A second observation is that the 7–23 and 1–23 TTET transfer coefficients ($f_i$) are almost all zero. This is expected—these slow TTET processes are coupled to unfolding and do not monitor the most populated states of HP35. A third observation is that the 1–35 and 23–35 TTET transfer coefficients ($f_i$) all show activity in the populated states. In particular, the 1–35 $f_i$ are strongest for the most populated one or two states, consistent with the interpretation in ref. 1. The 23–35 TTET $f_i$ are also strong for several near-native states. A key exception is the implicit solvent GBSA data—in the GBSA data, all 10 of the most populated states have zero 23–35 transfer coefficients. This corroborates suggestions (15) that implicit solvent models tend to overstabilize secondary structure.