Supporting Online Material for

Atomic-Level Characterization of the Structural Dynamics of Proteins

David E. Shaw,* Paul Maragakis, Kresten Lindorff-Larsen, Stefano Piana, Ron O. Dror,
Michael P. Eastwood, Joseph A. Bank, John M. Jumper, John K. Salmon, Yibing Shan,
Willy Wriggers

*To whom correspondence should be addressed. E-mail: David.Shaw@DEShawResearch.com

This PDF file includes:

Materials and Methods

Figs. S1 to S16

Tables S1 to S4

References

Other supporting material includes the following:

PDB files S1 to S5

Materials and Methods

Methods used in the simulation and analysis of reversible folding simulations

Computational setup

The sequences used in the WW and villin simulations were GLY-SER-LYS-LEU-PRO-PRO-GLY-TRP-GLU-LYS-ARG-MET-SER-ARG-ASP-GLY-ARG-VAL-TYR-TYR-PHE-ASN-HIS-ILE-THR-ASN-ALA-SER-GLN-PHE-GLU-ARG-PRO-SER-GLY and LEU-SER-ASP-GLU-ASP-PHE-LYS-ALA-VAL-PHE-GLY-MET-THR-ARG-SER-ALA-PHE-ALA-ASN-LEU-PRO-LEU-TRP-NLE-GLN-GLN-HIS-LEU-NLE-LYS-GLU-LYS-GLY-LEU-PHE, corresponding, respectively, to the sequence of the FiP35 variant of the human Pin1 WW domain (*1*) and to a fast-folding villin variant (*2*). In the villin sequence, NLE refers to norleucine residues. We also simulated six mutants of FiP35, namely Leu4Ala, Trp8Phe, Arg11Ala, Ser13Ala, Tyr19Leu, and Phe21Leu. Finally, we simulated 15 peptide fragments taken from the FiP35 sequence and corresponding to hairpin 1 (residues 7–23), hairpin 2 (residues 18–29), and sequences containing hairpin 1 where residues were added in groups of three to either the N- or C-terminus of the hairpin (or both). This gave rise to the following series of peptides: residues 7–23, 7–26, 7–29, 7–32, 7–35, 4–23, 4–26, 4–29, 4–32, 4–35, 1–23, 1–26, 1–29, and 1–32.

We built all the systems in an extended conformation and modeled all Asp, Glu, Lys and Arg residues and the polypeptide termini in their charged state. In the peptide fragments, the termini were either acetylated or methylated, except for the first and last residues of the FiP35 sequence. The histidine residue in FiP35 was modeled as neutral and protonated in the δ position and the histidine residue in villin was modeled as positively charged (*2*). We solvated the peptides in a cubic box with ~50 Å side length containing ~4000 TIP3P (*3*) water molecules, and added sodium and chloride ions to achieve a ~30 mM ionic concentration. A total of three chlorine ions were added for FiP35 and three chlorine and two sodium ions for villin. We described the physics of the system using the Amber ff99SB-ILDN force field, which is based on the ff99SB force field (*4*) and which includes recently described improved torsion parameters for the side

chains of Ile, Leu, Asp and Asn residues (*5*). Parameters for norleucine were based on those of neutral Lysine up to the Cδ, while the parameters for the methyl group were taken from Isoleucine. The residual charge of −0.00535 was redistributed on the Cδ and Cε atoms. We used the Desmond 2.1.0.1 software (*6*) running on an Intel Xeon cluster to perform equilibration and enhanced sampling runs, and the Anton specialized hardware (*7*) to perform all other simulations.

Simulation protocols

We constrained all bonds involving hydrogen atoms to their equilibrium lengths with the SHAKE algorithm (*8*). We used a cutoff of 9.5 Å for the Lennard-Jones and the short-range electrostatic interactions; for the long-range electrostatic interactions we used the *k*-Gaussian Split Ewald method (*9*) with a 32 × 32 × 32 grid in the Anton simulations and the Particle Mesh Ewald method (*10*) with a 32 × 32 × 32 grid and a fifth-order interpolation scheme in the Desmond simulations. We used a multistep r-RESPA scheme (*11*) for the integration of the equations of motion with time steps of 2.5 fs, 2.5 fs and 5.0 fs for the bonded, short-range non-bonded and long-range non-bonded interactions, respectively.

We first equilibrated the systems using 2.4 ns of MD simulation in the NPT ensemble and chose the frame of this simulation with the volume closest to the average volume as the starting structure for the subsequent long MD simulation on the Anton hardware. We carried out equilibrium MD simulation on Anton in the NVT ensemble using the Nose-Hoover thermostat with a relaxation time of 1.0 ps. Initially, we performed one 88-μs simulation for villin at 300 K and one 58-μs simulation for FiP35 at 337 K. Subsequently, we performed two additional 100-μs simulations for FiP35 at 395 K, slightly below the estimated melting temperature (see below) of 400 K. At the same temperature, we ran four independent simulations of lengths 75, 50, 50, 100, 100 and 100 μs for each of the Leu4Ala, Trp8Phe, Arg11Ala, Ser13Ala, Tyr19Leu and Phe21Leu mutants, respectively, for an aggregate simulation time of 1.9 ms. Finally, we simulated the peptide fragments until the estimated error on the folding rate was less than 20%. This required between 45 and 140 μs of simulation for each system, for an aggregate simulation time of 1.5 ms. In all simulations on Anton, we saved frames every 200 ps.

Molecular visualization of Figs. 1, 2, and S4 was performed using VMD (*12*).

Determination of the melting temperature using replica-exchange metadynamics with a pathlike collective variable

The maximum efficiency in reversible folding simulations (that is, the number of folding/unfolding events per time unit) is achieved when the simulation temperature is close to the melting temperature. Before performing equilibrium reversible folding simulations on Anton, we used a replica-exchange metadynamics approach to estimate the melting temperature of the system in Amber ff99SB-ILDN. First, we used the folding event observed in a 58-μs simulation of FiP35 at 337 K to define a reaction coordinate that was subsequently used in an enhanced sampling protocol. In the 58-μs simulation, we observed that the folding transition had the following sequence of events: first the tip of hairpin 1 formed, followed by formation of the

entire hairpin 1, then by formation of hairpin 2 and finally by collapse of the hydrophobic core to give the native fold. We therefore designed a pathlike reaction coordinate, $q(x)$ (*13, 14*), defined by the $C\alpha$ mean square displacement (MSD) from the native state structure of different parts of the protein sequence:

$$q(x) = \sum_{i=1}^{4} i \cdot e^{-\alpha_i MSD_i} \; .$$

Here, MSD_1 refers to the tip of hairpin 1 (residues 12 to 18), MSD_2 to hairpin 1 (residues 8 to 22), MSD_3 to hairpin 1 and 2 (residues 8 to 30) and MSD_4 to the whole protein (residues 3 to 33). The value of the weighting factors, α_i, were inspired by a scaling law for MSDs (*15*) and given by (in units of Å^{-2}):

$$\alpha_i = \frac{0.4}{\left(\log(N_i) - 1\right)}$$

with N_i being the number of residues included in the MSD calculation. Using this reaction coordinate, we performed eight parallel metadynamics runs (*16*) at temperatures ranging between 345 K and 408 K, coupled in a replica exchange scheme (*17*). We deposited Gaussians of width 0.1 and height 0.05 kcal mol^{-1} every 5 ps of metadynamics simulation, and attempted exchanges between the replicas at different temperatures every picosecond of simulation. Overall, we performed a total of 394 ns of replica exchange metadynamics simulation for each of the eight replicas. These simulations were performed using Desmond (*6*).

From these simulations, we calculated the folding free energy at different temperatures along the chosen reaction coordinate (Fig. S1) and determined the melting temperature as the temperature where the free energy of the folded and unfolded states is equivalent. We estimated the melting temperature to be close to 400 K, and decided to use the more conservative estimate of 395 K for our long equilibrium simulations. This will ensure that a substantial fraction of the simulation will be spent in the native state, in case the estimated value was slightly lower than the real melting temperature of the force field. We find the population of the native state to be ~60% in the equilibrium runs, in excellent agreement with the populations estimated by the replica exchange metadynamics. The experimental melting temperature is ~40 K lower than the computationally-derived value. This discrepancy—as well as the long folding time of villin relative to the folding rate extrapolated from experimental data (*2*; Fig. 1)—may be attributable to the slight tendency of the Amber ff99SB force field to overestimate the stability of the β-sheet region of the Ramachandran map relative to α- and polyproline-II-structures (*18, 19*). Other potential explanations include minor biases in the non-bonded forces that may, for example, lead to incorrect temperature dependencies of secondary-structure propensities (*18*).

Determination of a folding reaction coordinate from equilibrium simulations

Based on the equilibrium folding simulations of FiP35, we determined the transition state ensemble (TSE) for folding. As a first step, we obtained a reaction coordinate for describing the folding transition from the two 100-μs equilibrium-reversible folding simulations of FiP35 at 395 K using the method developed by Hummer (*20*) and applied to protein folding by Best and Hummer (*21*). This method requires an equilibrium distribution and multiple folding and unfolding transitions, and is thus ideally suited to the analysis of equilibrium-reversible folding simulations. Briefly, the method attempts to find a reaction coordinate, $r(x)$, which maximizes the probability that a trajectory visiting the transition state region is a reactive trajectory (*22*). In practice, such a reaction coordinate can be determined using a variational approach in which $r(x)$ is parameterized by a set of "weights," w_i, that are varied to find an optimal reaction coordinate (*21*). A reaction coordinate obtained with this approach is optimal in the sense that equilibrium fluctuations from the folded and unfolded basins are maximally separated from reactive trajectories that result in barrier-crossing events.

We defined the reaction coordinate as a linear combination of several elements with "weights" w_i that are taken as parameters in the optimization procedure (we allow the parameters w_i to be negative during our Monte Carlo based optimization). In the original implementation (*21*) these were the elements of the contact matrix. In our application, we use a function of the mean square displacement (MSD) from the native structure of all consecutive five-residue fragments:

$$r(x) = \sum_{i=3}^{33} w_i e^{-\frac{1}{2}MSD(i-2,i+2)}$$

Here i runs over residues 3 to 33 and $MSD(i-2,i+2)$ is the backbone MSD from the native structure calculated on residues ranging from $i-2$ to $i+2$. We take as initial state for the optimization a reaction coordinate with uniform weights. This reaction coordinate already discriminates between folded and unfolded states and we used it to define transition paths as the parts of the trajectory connecting a frame with $r(x) < 0.1$ to a frame with $r(x) > 0.7$. This definition identifies a total of 15 transitions and selects 27748 frames out of a total of 10^6 to belong to the transition paths. By definition, the maximum value of the probability $p(\text{TP}|r)$ (*21*) is observed at the transition state, and we find it to be 0.37 initially and 0.52 after optimization of the coefficients w_i. The observed maximal value of $p(\text{TP}|r)$, $p(\text{TP}|r = 0.284) = 0.52$, is consistent with the theoretical value of 0.5 expected for diffusion over a free energy barrier (*21*).

P_{fold} analysis of the transition state

Based on the analysis described above, we assigned all of the trajectory frames within 0.2 units of $r(x)$ from the peak of p(TP|r) to the transition state ensemble (TSE). We validated the TSE by performing a P_{fold} analysis (*23*) on 101 frames randomly taken from the TSE. Of these 101 frames, 46 came from actual transition paths in the equilibrium simulation, with the remaining 55 originating from unproductive folding or unfolding attempts that entered and exited the transition state region of $r(x)$ from the same (folded or unfolded) state. Starting from each frame, we

performed four simulations with different initial velocities and ran the simulations until either the folded ($r(x) < 0.1$) or unfolded ($r(x) > 0.55$) states were reached. This required between 0.004 and 3.3 μs of simulation, with an average of 0.37 μs. We calculated P_{fold} for each frame as the ratio between the trajectories that folded and the total number of runs (4 in this case).

φ-value calculation

Contact φ-values (*24, 25, 26*) are ideally suited to quantify the degree of "nativeness" of each residue of the protein in the TSE. With this purpose, we calculated side-chain φ-values for each residue, assuming that the contribution of each residue to the folding free energy is proportional to the fraction of the native side-chain contacts N_i:

$$\varphi_i = \frac{N_i^{TSE} - N_i^{U}}{N_i^{F} - N_i^{U}}$$

Here we considered in the analysis only contacts that are formed more than 80% of the time in the native state part of the simulation. We counted a contact every time two heavy atoms more than one residue apart were closer than 6 Å (*27*) and performed this calculation for every frame of the simulation, assigning each frame to the folded (F), unfolded (U) or transition (TSE) state, with state assignments based on the optimized reaction coordinate $r(x)$.

Additionally, we calculated side-chain φ-values for six selected mutants of FiP35 (Leu4Ala, Trp8Phe, Arg11Ala, Ser13Ala, Tyr19Leu, and Phe21Leu) directly from comparison of the folding rates and stabilities. Errors in the calculated φ-values were estimated using a bootstrapping procedure. For each set of simulations (of either the wild-type FiP35 or a mutant protein) we resampled (with replacement) folding and unfolding events so that the artificial resampled trajectories had the same aggregate length as the original trajectories. This procedure was repeated 10,000 times, and we calculated confidence intervals using the resulting distribution of φ-values. The 68% confidence intervals obtained using this procedure are: L4A ($-1.3 - 0.4$), W8F ($-0.2 - 0.2$), R11A ($0.1 - 0.4$) , S13A ($0.1 - 1.6$) , Y19L ($0.1 - 0.5$) , F21L ($0.3 - 0.5$).

Backbone φ-values were calculated using a free energy perturbation approach (*28*) to model the experimental free energy changes due to amide-to-ester substitutions. For each substitution and for each frame in the two long FiP35 simulations, we calculated the change in enthalpy upon a virtual substitution of the backbone amide with an ester, and these were exponentially averaged over the native, transition and unfolded states to calculate φ-values. Errors on the backbone φ-values were estimated from block averaging. In Fig. S3, we show the calculated φ-values (for those mutations giving rise to $\Delta\Delta G_{N-U} > 1$ kcal mol^{-1}) together with the experimental values from the Pin1 WW domain.

Single-stage free-energy perturbations may potentially provide inaccurate results, as relevant configurations of the ester variants may not be sufficiently sampled in simulations of the (amide) wild type. To test whether this effect influences the accuracy of our free energy calculations, we

compared the value from a one-sided free energy perturbation with that obtained from direct simulation of reversible folding in both amide and ester variants. In particular, we calculated free energies of folding from simulations of reversible folding of the isolated hairpin 1 (HP-1 [7–23]; Table S1) and a variant in which we substituted the peptide bond between residues R14 and D15 with an ester. From these two simulations, we calculated that hairpin 1 becomes more stable by 0.8 ± 0.2 kcal mol^{-1} upon substitution of the amide. We also calculated the corresponding value using a single-stage free energy perturbation using only the simulation of the amide variant. The resulting value (0.8 ± 0.4 kcal mol^{-1}) is in good agreement with the value calculated using simulations of reversible folding, suggesting that our free energy approach can be used to calculate reliable free energy changes upon amide-to-ester substitutions.

Simulation of a T-jump experiment

The equilibrium folding simulations presented here contain enough kinetic and thermodynamic information to allow comparison of the simulation data with T-jump experimental data. While calculation of a T-jump relaxation directly from an equilibrium trajectory is possible in principle, in practice the simulation needs to be about two orders of magnitude longer than the longest timescale of interest to give converged results. For this reason, we instead simulated temperature-jump experiments by performing Langevin simulations on a model free energy landscape extracted from the analysis of the two 100-μs FiP35 simulations. We projected the free energy landscape on the optimized reaction coordinate $r(x)$ (Fig. S5, panel a) and determined a position-dependent diffusion coefficient (Fig. S5, panel b) using the method of Hummer (*22*). The free energy profile contains two dominant minima (corresponding to the unfolded and folded basins) separated by a broad and relatively flat free energy barrier. The position-dependent diffusion coefficient drops roughly linearly from a value of ~0.2 μs^{-1} in the unfolded state to ~0.03 μs^{-1} in the folded state. (We note that $r(x)$ does not have units and that the diffusion constant depends on the chosen reaction coordinate and does not correspond to any easily measurable experimental diffusion coefficient.) The folding and unfolding times obtained using Langevin dynamics on the free energy profile are 5.3 μs and 9.1 μs, respectively, and are only a factor of two faster than those observed in the equilibrium runs. This suggests that the model mostly captures the observed dynamics with the remaining difference explained, for example, by a small additional barrier of ~0.6 kcal mol^{-1} not captured in the reaction coordinate.

To simulate a temperature jump experiment, we first estimated the populations at different temperatures, assuming that the position-dependent enthalpy (Fig. S5, panel c) remained constant in the temperature interval considered. The calculated position-dependent enthalpy decreases smoothly from the unfolded to the folded state, consistent with the funnel picture of a folding free energy landscape, and shows how the small free energy barrier arises from large and opposing contributions from the folding entropy and enthalpy. We used the solvent accessible surface area (SASA) of the tryptophan side chain, calculated with a probe of 1.4-Å radius, as a proxy for the fluorescence intensity (Fig. S5, panel d). We performed a total of 15,000 Langevin simulations at 395 K starting from conformations selected at random from a 375 K or 385 K Boltzmann ensemble, and monitored the time evolution of the ensemble-averaged SASA (Fig. S6) mimicking the fluorescence intensity signal observed in the T-jump experiments (*1*). A single-exponential fit to the observed relaxation kinetics resulted in a time-constant of 3 μs, as

expected from the folding and unfolding rates in the model. Such a fit, however, revealed discrepancies at shorter timescales, suggesting the presence of at least one additional relaxation process. This faster "molecular" phase can be fit by an additional exponential term, although the value obtained depends on both the size of the temperature jump and the spectroscopic probe used (Fig. S6). As an alternative method to link the structures seen in our MD simulations to Trp fluorescence measurements, we also calculated the interaction between the Tryptophan and Tyrosine side chains (Fig. S5, panel e). In particular, we calculated the SASA of the single Trp residue, taking only burial by Tyr side chains into account. The Trp indole ring is not in close contact with the phenolic groups in the folded state, and so the SASA in the folded state is only slightly lower than in the unfolded state. In the transition state region, however, there are some transiently formed non-native contacts with the phenolic rings that cause the Trp SASA to be lower than in either the native or unfolded states. When we use the Trp (Tyr-burial only) SASA as a spectroscopic probe, we also observe two kinetic phases. The amplitudes of the two phases in this case differ in their sign, however, so that—after the temperature jump—the SASA first decreases and then increases again (Fig. S6). This behavior is a direct consequence of the non-vanishing contribution of the barrier region to the observed signal, and shows that temperature-jump experiments such as these may, in favorable cases, be used to probe the spectroscopic signal of the transition state region.

Methods used in the simulation and analysis of the BPTI trajectory

Molecular dynamics simulation of BPTI

We prepared the simulation of BPTI using commodity computers and our in-house MD simulation software, Desmond (*6*). We selected the first alternative conformation of the joint neutron/X-ray refined structure with PDB ID 5PTI (*29*) and changed the deuterium atoms to hydrogen atoms. We solvated the resulting protein in a cubic box with sides measuring 52 Å containing 6 chloride ions and 4215 water molecules, using the software Maestro (*30*), we modeled the ions and protein using a variant of the AMBER ff99SB force field (*4*), with the only modification with respect to that force field being a correction to the parameters describing isoleucine (*5*). We modeled the water with the 4-particle TIP4P-Ew force field (*31*)**,** which was previously shown (*32*) to better describe the rotational motion of proteins than the related 3-particle water model, TIP3P (*3*), which is commonly used in simulations with the AMBER ff99SB protein force field. We applied SHAKE to all bonds involving hydrogen atoms (*8*). We minimized the resulting structure to remove any clashes. We applied harmonic positional restraints of strength 10 kcal/mol/Å^2 to the protein backbone atoms, kept the pressure at 1 atm and ramped up the temperature from 10 K to 300 K as a linear function of time over the course of 1.2 ns, using Berendsen temperature and pressure control algorithms with relaxation times of 0.5 picoseconds for both the barostat and the thermostat (*33*). We removed the restraints and performed a 6-ns simulation at constant isotropic pressure of 1 atm and temperature of 300 K, using the MTK algorithm (*34*) with 10-ps relaxation time as implemented in the software Desmond (*35*). We used a 10-Å cutoff radius for range-limited interactions, with Particle Mesh Ewald electrostatics evaluated every third step on a 48 × 48 × 48 mesh, using a sigma of 2.256 and fifth-order spline interpolation. We extracted the configuration that had the volume closest

to the average volume of this short simulation, which resulted in a cubic box with side lengths of 51.26 Å.

The production simulation of BPTI on Anton used a 10.4-Å cutoff radius for range-limited interactions. The long-range electrostatics were calculated with the Gaussian Split Ewald method (*9*) using a 7.1-Å cutoff for charge spreading and force interpolation, and a $32 \times 32 \times 32$ FFT mesh. The simulation box size was kept constant, and the simulation temperature was controlled by a Berendsen thermostat with a 100-ps relaxation time (*33*). The simulation time step was 2.5 fs, with long-range electrostatics evaluated every other time step. We saved snapshots of the trajectory (positions and velocities of all atoms) at 250-ps intervals and a checkpoint of the trajectory (a snapshot with additional information that allows a bit-level accurate continuation of the simulation) at 2.5-ns intervals. The analysis of the BPTI trajectory was performed with scripts developed in house, described below, some of which used the HiMach framework (*36*) that integrates with VMD (*12*) to facilitate seamless parallel access and reduction of large MD trajectories.

For the analysis of the fast relaxation (see section "Analysis of the dynamical content of the BPTI trajectory"), we resampled the trajectory at two additional time grids starting at every integer microsecond between simulated time 20 μs and simulated time 1000 μs. One grid covered 981 segments of 1 ns with samples at 250-fs intervals, and the second grid covered 981 segments of 10 picoseconds with samples at 5-fs intervals. We confirmed that the end of each of the resampled 1-ns segments was a snapshot that is bit-level identical to the corresponding snapshot of the original trajectory.

Molecular visualization of Figs. 4 and S9 was done using OpenStructure (http://www.openstructure.org/).

Kinetic clustering of the BPTI trajectory

In what follows, we describe the details our new kinetic clustering scheme, which simplifies the analysis of the long-time relaxation in the MD simulation trajectory. This clustering scheme was used to color the time series of Fig. 4a in the main text. In short, the clustering method works by assigning conformations to clusters so that the long-timescale behavior in cluster-space mimics the corresponding long-timescale behavior observed in the MD simulation as closely as possible. More precisely, we calculate autocorrelation functions of the time series of a large number of atomic distances, and aim to match the long-timescale tail of these correlation functions with corresponding correlation functions calculated over dynamics in cluster space.

We define $m(t)$, a clustering of the molecular dynamics simulation trajectory onto M clusters, as a map from a sequence of times (or time intervals), indexed by $t \in \{1, N\}$ onto the set of integers in $\{1, M\}$. Typically the number of clusters M is significantly smaller than the number of samples N.

Let $x_k(t)$ be the time series of distance k. We define the cluster-average value $A(x_k, m)$ for cluster m:

$$A(x_k, m) = \frac{\sum_{t=1}^{N} \delta_{m,m(t)} x_k(t)}{\sum_{t=1}^{N} \delta_{m,m(t)}},$$

with δ the Kronecker delta. The cluster-average value $A(x_k, m)$ gives the average value of the distance x_k for samples that belong to cluster m. We define the cluster-mean sequence of the distance x_k as:

$$\tilde{x}_k(t) = A(x_k, m(t)).$$

We consider the non-normalized correlation functions of the distance,

$$C_{x_k}(\tau) = \left\langle x_k(t) x_k(t+\tau) \right\rangle_t,$$

and of the cluster-mean sequence of the distance,

$$C_{\tilde{x}_k}(\tau) = \left\langle \tilde{x}_k(t) \tilde{x}_k(t+\tau) \right\rangle_t,$$

with the brackets denoting averaging over all times for which the arguments of the functions inside the brackets are defined.

We define a quality measure of the clustering, called cluster energy E, for a collection of K time series that are sampled at the same sequence of times:

$$E = \sum_{k=1}^{K} \sum_{\tau=\tau_1}^{\tau_2} \left(C_{\tilde{x}_k}(\tau) - C_{x_k}(\tau) \right)^2.$$

If we pick the time interval $[\tau_1, \tau_2]$ to span the long-time tail of the correlation functions of interest, the cluster energy will be small when this long-time decay can be approximated from the hops in the simplified time series $\tilde{x}(t)$.

We implemented a simulated annealing protocol to optimize the clustering $m(t)$ for a set of time series x_k. This annealing protocol performs a Metropolis Monte Carlo (MC) simulation at a temperature T that is lowered from T_2 to T_1 with decrements that are equidistantly spaced on a logarithmic grid. Each MC move is either of a "rewriting" move or a "grow" move, with 50% probability for each. The rewriting move chooses a random interval with length up to L samples and assigns the samples in that interval to a random cluster. The grow move considers contiguous samples of the time series that share a cluster assignment. The grow move chooses one such sequence at random and randomly decides to grow one of its two ends by up to L samples.

For the analysis of the BPTI trajectory, we randomly selected 1000 pairs of non-hydrogen atoms of BPTI and measured the Euclidean distance of each pair at 250-ps intervals. (We excluded the symmetric aromatic carbons with names CD1, CD2, CE1, CE2, the Val carbons CG1, CG2, the Leu carbons CD1, CD2, and the Arg nitrogens NH1 NH2; the relaxation of certain symmetric groups is extremely slow and does not lead to changes in structure and potential energy.) We coarse-grained the resulting time series by applying a window mean over non-overlapping 250-ns long windows, resulting in a total of 4,000 time intervals for each of the 1,000 time series. We then optimized the cluster energy using the limits $[\tau_1, \tau_2] = [1 \ \mu s, 15 \ \mu s]$ with 3 independent repetitions of 6 million steps of the simulated annealing protocol described above with $L = 20$, $T_2 = 10^4$, and $T_1 = 10^{-6}$. Fig. S7 shows the resulting cluster energy with the number of clusters ranging from 3 to 8.

Analysis of the structural differences between conformational states of BPTI

Our simulation of BPTI transitioned reversibly among a small number of structurally distinct long-lived states. Among other things, these states differed in their RMSD from the crystal structure (Fig. 4a), the chirality of the disulfide bond between Cys14 and Cys38 (Fig. S8a), and the mobility of Tyr10 (Fig. S8b).

In order to identify which residues are most likely to contribute to stabilizing the different conformers of BPTI, we performed analyses of the detailed structural differences among the five conformational states we observed in our simulation.

We estimated the correlation of structural properties of simulation snapshots with the membership of these snapshots to a particular cluster. Let $c_n(t)$ in $\{0,1\}$ be the characteristic function that defines membership of a structure that was visited in the simulation at time t to cluster n, with n in $\{0, 1, 2, 3, 4\}$. The value of $c_n(t)$ is 1 when the system is in cluster n at time t and is 0 otherwise. Let $d_{ij}(t)$ denote the time series of the distance between residue i and residue j (for simplicity, we used the distance between the representative side-chain atom of each residue, as defined in [37]). For each cluster n, we construct the positive symmetric matrix C_{ij}^n of absolute Pearson correlation coefficients between the time series $d_{ij}(t)$ and $c_n(t)$. This matrix describes the linear statistical relationship of every residue pair (i,j) with the membership in cluster n. We found that for all clusters, the corresponding matrix C_{ij}^n displays a column-wise striped structure with certain columns having large values for all rows; since the matrix is symmetric with respect to exchange of i with j, the matrix also has a corresponding row-wise striped structure. These strips of uniformly large values in the matrix C_{ij}^n are due to the global nature of the statistical relationship between the change of cluster and the concomitant change of distances from a particular residue to multiple parts of the BPTI structure. The particular structure of this matrix implies that we can compress the information in it by averaging the columns of C_{ij}^n to their mean values C_i^n. We call C_i^n the stabilization coefficient of residue i in cluster n, because we expect that mutations of residues with large values of this quantity will mostly likely influence the stability of cluster n. Table S2 and Fig. S9 show that the resulting dominant values of the stabilization coefficient are highly localized for each cluster, thus providing a prediction for the residues that, upon mutation, are most likely to stabilize one of the alternative conformations of BPTI.

Analysis of the dynamical content of the BPTI trajectory

In this section we detail the method used to estimate the average internal relaxation of the direction of the bond vectors of BPTI, discussed in relation to Fig. 4b of the main paper. This analysis provides a global view of the simulated dynamics of BPTI over a wide range of timescales.

We selected a set of 304 bonds in BPTI. Our selection included commonly-used probes of NMR relaxation, including all bonds between C_α–C, C_α–H, N–H (both the backbone amide and the side-chain bonds), methyl C–C, and methyl C–H atoms. Using the procedure detailed below, the analysis of the relaxation coming from the complete set of 304 bonds resulted in the orange curve in Figs. S10 and 4b; the relaxation coming from the subset of 175 backbone bonds resulted in the green curve in Fig. 4b; the relaxation coming from the subset of 129 side chain bonds resulted in the blue curve in Fig. 4b. The analysis of the relaxation coming from the subsets of bonds involving C_α, C_β, C_γ and C_δ atoms is shown in Fig. S11.

The time-autocorrelation function of the direction of each bond in a molecular frame of reference was calculated in a three-step procedure using a parallel code implementation (*36*) of the procedure we previously described (*38*). First, the internal correlation function was calculated between lag times 250 ps and 50 μs, using the time series resulting from the linearly sampled time grid of the BPTI trajectory. Second, the internal correlation function was calculated between lag times of 250 fs and 250 ps for each of the 981 resampled 1-ns intervals; the resulting 981 correlation functions were averaged. Third, the internal correlation function was calculated between lag times of 5 fs and 1 ps for each of the 981 resampled 10-ps intervals; the resulting 981 correlation functions were averaged. The resulting internal time-autocorrelation functions were averaged over the set of 304 bonds and the resulting mean internal correlation function is plotted in Fig. S10.

We define the *dynamical content* of a time-autocorrelation function to be the negative of its derivative with respect to the logarithm of lag time. Assuming the correlation function decays as a function of lag time, it follows that the dynamical content is a positive quantity. Furthermore, the integral of the dynamical content with respect to the logarithm of the lag time (the area under the orange curve of Fig. 4b) is equal to the decay of the value of the correlation function at the boundaries of the integration interval. It thus follows that the dynamical content is a probability distribution that measures the amount of relaxation that occurs at a certain timescale.

The *average dynamical content* over all correlation functions is also a probability distribution and is given by the negative derivative of the average internal correlation function (since the average and derivative operations commute). To effectively eliminate the noise associated with numerical differentiation of sampled data, we estimated the derivatives on analytical models that we fitted to the data.

We fitted the average internal correlation function between 1-ps and 50-μs lag times on a sequence of multi-exponential models using an in-house code that calls the function NonlinearModelFit of Mathematica 7.0 (*39*). The models with N+1 parameters, S[i], had the

form (in Mathematica notation): `C[t_] = S[0] + Sum[ S[i] Exp[-t/t[i, N]],` `{i, 1, N}]`, with N the number of timescales, ranging from 10 to 32, and with the timescales `t[i, N]` equidistantly spaced on a logarithmic grid according to: `t[i, N] =` `10^( Log[10, tStart] + DT[N]/2 + i * DT[N] )`, with `DT[N]=(Log[10,tStop]-Log[10,tStart])/N`, and `tStart=1 ps, tStop=50` μs. The Bayesian Information Criterion (BIC) (*40*) of these models displays a first minimum as a function of the number of timescales for N=15 (see Fig. S12).

Fig. S13 shows the dynamical content of the fitted models for N=14–23, using the same color coding as used in the plot of the BIC in Fig. S12. Visual inspection of the dynamical content of the models with $N > 20$ reveals minor differences in the large-time tail; however, the differences between models up to lag times of several microseconds are very small compared to the features in Fig. 4b and our results are robust to the choice of model.

Analysis of the transition path time of BPTI

We estimated the transition path duration between selected pairs of states in BPTI using a procedure that is somewhat different from the procedure we used to estimate the transition path time in protein folding. Since obtaining the P_{fold} reaction coordinate for BPTI's conformational changes would require more instances of transitions between pairs of states than even the millisecond simulation provides, we resorted to using an ad hoc one-dimensional coordinate in its place (see below). We expect that the resulting transition path estimation procedure will provide a lower bound for the duration of transition paths in BPTI. The additional orthogonal coordinates that would need to be incorporated to our ad hoc coordinate when describing the P_{fold} reaction coordinate would tend to increase the estimate of the path's duration, since changes in these additional degrees of freedom could be leading or lagging with respect to changes in our one-dimensional coordinate.

We selected the backbone RMSD from the crystal structure (shown in Fig. 4a) as a reaction coordinate. These RMSD fluctuations provide some separation between states 0 and 2, as well as between states 1 and 2 (Fig. S14). Borrowing an idea from the transition-based assignment of states (*41*), we eliminated spurious transitions due to the remaining overlap in the RMSD between these pairs of states by only selecting transition paths connecting the distant parts of the RMSD distributions of each state. The distant parts were defined as those corresponding to the top and bottom 5% of the distributions (shaded areas in Fig. S14). The calculated transition path time is relatively insensitive to the precise choice of the distant part cutoff (see below). We defined the transition path as the segment of trajectory from the time the RMSD time series last exited a distant part of one state's distribution while being in that state until the time when it first entered the distant part of a new state's distribution while being in that new state.

By analysis of the jumps between states in the clustering from our kinetic clustering scheme, we identified a total of 54 transitions between states, including 10 transitions from state 0 to state 2, 9 transitions from state 2 to state 0, 2 transitions from state 1 to state 2, and 2 transitions from state 2 to state 1. For each of the 23 transitions between states 0 and 2 or states 1 and 2 we obtained a transition path segment using the procedure described above. The average transition

path duration, aggregate over all 23 transitions, is 281 ± 104 ns (changing the distant part cutoff criteria to 10% changes the average transition path duration to 233 ± 67 ns).

Analysis of the rotations of the aromatic rings of BPTI

Tables S3, S4 and Fig. S8b report the rate of aromatic ring rotations. The symmetric atoms of the rings of Tyr and Phe residues are distinctly labeled in the simulation; thus, we can monitor their rotations by analysis of the time series of the χ_2 dihedral angles of the Tyr and Phe residues. The value of each χ_2 angle ranges from -180 to 180 degrees and a rotation of a ring shifts its equilibrium fluctuations to a basin whose center is offset by 180 degrees prior to the rotation. We extracted the time series of these dihedral angles from the BPTI trajectory sampled at 250-ps intervals. We post-processed each dihedral angle times series as follows. We assigned the center of one basin at the peak of a histogram of the χ_2 angles with 5-degree bins; we assigned the center of the second basin in $[-180,180)$ differing by 180 degrees from the first. We eliminated spurious recrossing events with the transition-based assignment of states (*41*): we assigned a width of 25 degrees to each basin and counted a rotation event only when the time series entered a new basin after having most recently exited from a different basin. The reported rotation rate is the number of such rotation events per unit of time. Table S4 reports the analysis, for five aromatic rings, of the rotation events that occur while BPTI is visiting a particular state.

In order to better understand the observed rates of rotation of the aromatic rings of BPTI, we estimated the probability density functions of the aromatic rings' χ_2 angles. For each of the eight aromatic rings of BPTI, we first shifted the simulation's samples of the χ_2 dihedral angle to the interval $[0, 180)$ by adding 180 degrees to samples with negative values of the χ_2 angle. Then we used the kernel density estimation procedure implemented by the function *density* of the programming language R (version 2.11.0) to estimate the probability density function of the shifted samples in $[0,180)$. The resulting density was fit over a regular grid of at least 512 points; we then fitted a normal distribution to these points using the R function *nls*, starting with a guess of 90 degrees for the center and 10 degrees for the standard deviation of the normal distribution. The resulting densities, together with their fits, are shown in Fig. S15.

The probability density functions of the χ_2 dihedral angles in the quickly rotating rings differ most noticeably from those in the more slowly rotating ones by displaying pronounced shoulders. It is worth noting that the observed rate of rotation of the aromatics (or the ranking of the aromatic residues according to their rotation rate) does not correlate with the standard deviations of these distributions (listed in the center of each panel in Fig. S15). The observed rate of rotation of the aromatics also does not correlate with the average atomic density of protein atoms around each aromatic ring, as estimated by counting the number of heavy atoms within 4 Å from the CD1, CD2, CE1, CE2 atoms of each ring. The latter densities correlate well with the standard deviations of the χ_2 angle distributions (the Pearson correlation coefficient between the fitted standard deviations and the average atomic density is -0.75, as estimated using the R function *cor*).

In order to further clarify the origin of the observed differences in the kinetics between the fast and the slow rings, we decomposed the sampled χ_2 angles onto contributions coming from each

of BPTI's states. For each ring, we restricted the total set of available samples to only those that were collected while BPTI was visiting a particular state in the simulation. We then obtained the resulting probability density using the R function *density*. The resulting densities for each state are overlaid in each of the eight panels of Fig. S16. Importantly, the shapes of the shoulders of the quickly rotating rings depend substantially on BPTI's state because of the changed local environment the aromatic rings experience in the different states.

Estimation of the distribution of water escape times from the BPTI trajectory

We estimated the lifetime distributions of the internal waters, shown in Fig. 4d of the paper, with the Stable States Picture (SSP) algorithm (*42*). Below, we list the VMD selection command (12) for the SSP "reactant" and "product" surfaces of the internal waters (the parameters of the selection command are typeset in bold, and their values for the reactant and product surfaces are shown in the two rightmost columns):

Water	VMD Selection Command	Reactant	Product
W111	"water and (same residue as (pbwithin **cut** of (protein and resid 7 and name OE1 OE2)) and same residue as (pbwithin **cut** of (protein and resid 8 and name O)) and same residue as (pbwithin **cut** of (protein and resid 43 and name O)) and same residue as (pbwithin **cut2** of (protein and resid 10 and name O)))"	**cut**: 6.0 **cut2**: 7.0	**cut**: 8.5 **cut2**: 9.0
W112	"water and (same residue as (pbwithin **cut** of (protein and resid 43 and name O)) and same residue as (pbwithin **cut** of (protein and resid 10 and name H)))"	**cut**: 3.2	**cut**: 8.0
W113	"water and (same residue as (pbwithin **cut** of (protein and resid 10 and name O)) and same residue as (pbwithin **cut** of (protein and resid 44 and name ND2)) and same residue as (pbwithin **cut** of (protein and resid 41 and name H)))"	**cut**: 4.0	**cut**: 8.0
W122	"water and (same residue as (pbwithin **cut** of (protein and resid 11 and name O)) and same residue as (pbwithin **cut** of (protein and resid 38 and name O H)) and same residue as (pbwithin **cut** of (protein and resid 14 and name H)))"	**cut**: 4.0	**cut**: 8.0

Supplementary snapshots of the BPTI trajectory

We include a series of five representative PDB structures (also shown in Fig. 4c of the paper, superimposed using the Theseus software [*43*]), one for each of the major clusters colored in Fig. 4a of the paper. The representative PDB structures are the snapshots in the middle of the largest

trajectory segment of each cluster. The comments section of the PDB files documents the simulation time of the snapshot.
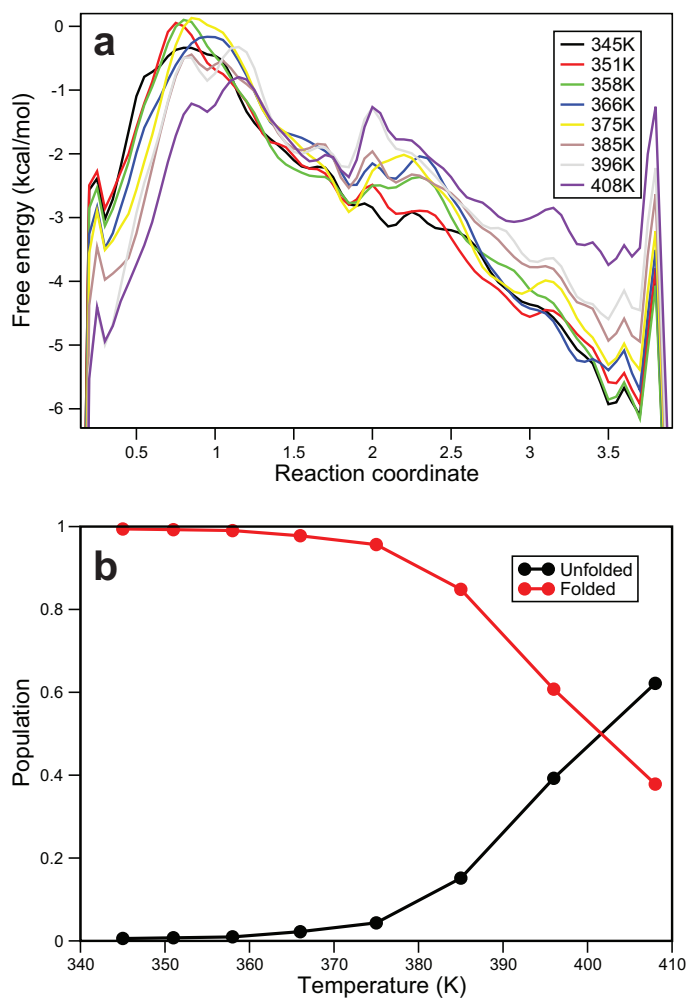
Supporting Figures

Fig. S1. **Estimating the folding free energy of FiP35 using replica exchange metadynamics.**
(**a**) Free energy profile along the metadynamics folding reaction coordinate, $q(x)$, at eight
different temperatures. The reaction coordinate goes from values of 0 (unfolded) to 4 (folded).
(**b**) Population of folded ($q > 1$) and unfolded state ($q < 1$) as a function of temperature estimated
using the free energy profile in part a.
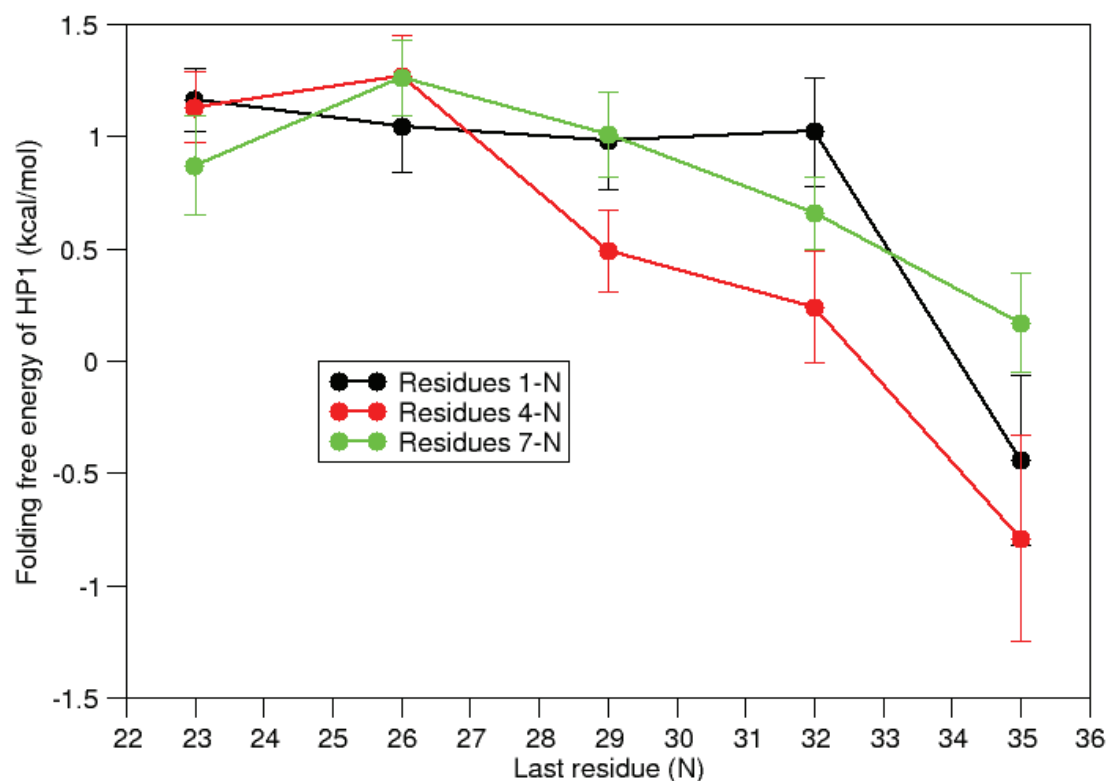
Fig. S2. Hairpin 1 stability in the context of fragments of the FiP35 WW domain. The folding free energy, calculated from reversible folding simulations of the peptide fragments, is reported as a function of the fragment length for peptide fragments starting at residue 1 (black), residue 4 (red) or residue 7 (green). The pronounced drop in stability observed when the last three residues are added can be attributed to the formation of the hydrophobic core around Trp8. The drop is less evident in peptides truncated at residue 7, as they lack the N-terminal residues that contribute to the hydrophobic core structure. The results show that the secondary structure elements in FiP35 are not fully stable on their own and that the formation of the hydrophobic core is required to obtain a stable protein, in line with experimental observations on CI2 and Barnase (*44*, *45*).
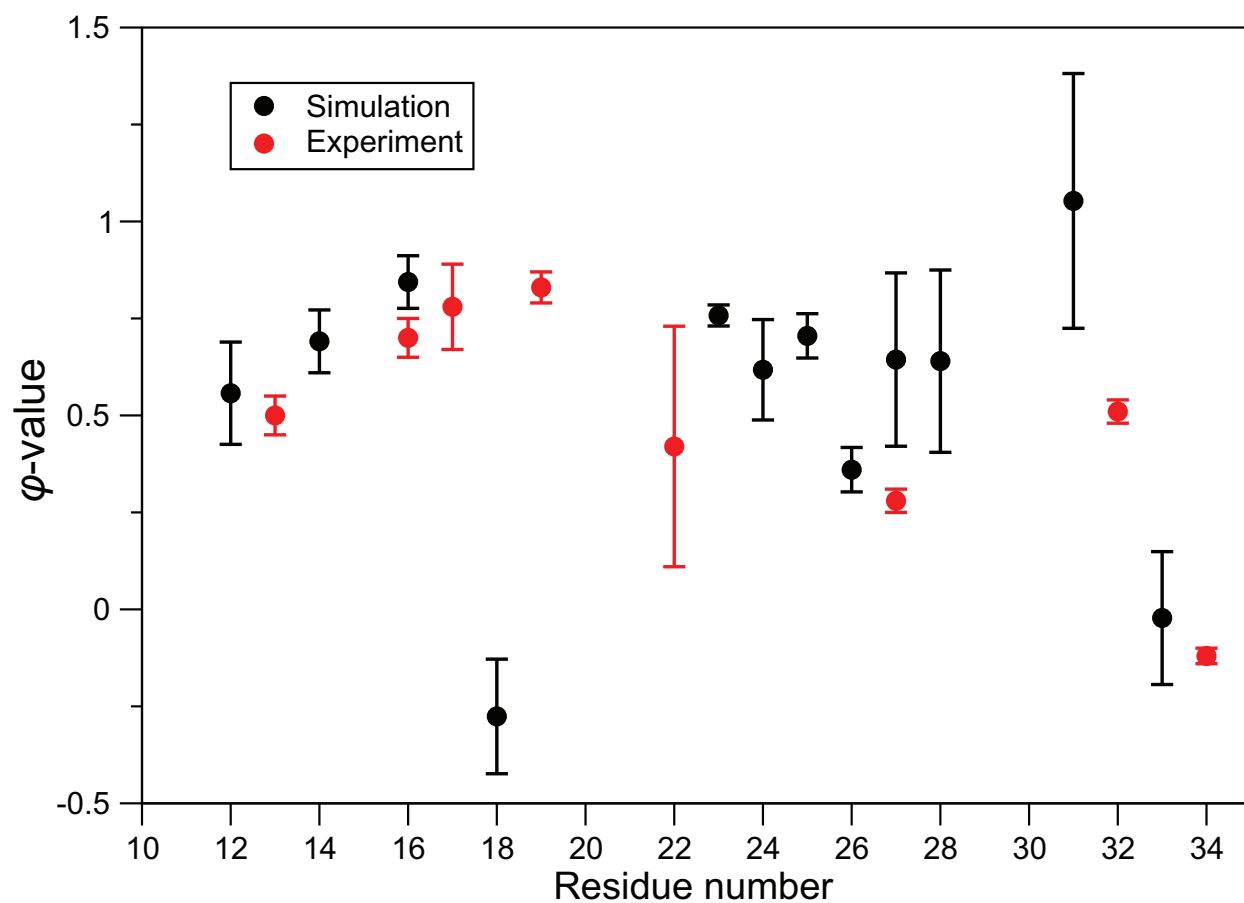
Fig. S3. Comparison of experimental and calculated φ-values for backbone amide to ester substitutions in FiP35. The plot shows the experimental φ-values (*46*) (red) obtained in the hPin1 WW domain together with the values calculated here for FiP35. Calculated values were obtained as described in the text.
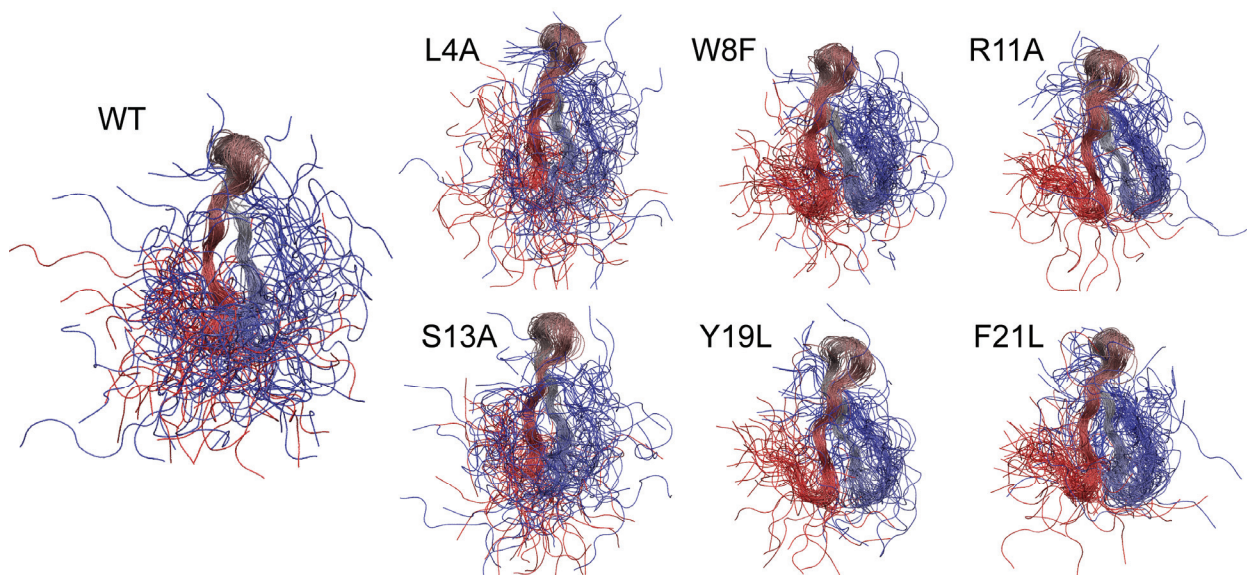
Fig. S4. Transition state ensemble structures of six FiP35 mutants. The transition state ensemble structures were extracted from equilibrium reversible folding simulations of each of the six mutants (L4A, W8F, R11A, S13A, Y19L and F21L) with the procedure described in above (optimizing a new reaction coordinate for each mutant) and aligned with the Theseus software. The transition state ensemble for the original FiP35 sequence (labeled WT) is also pictured for comparison.
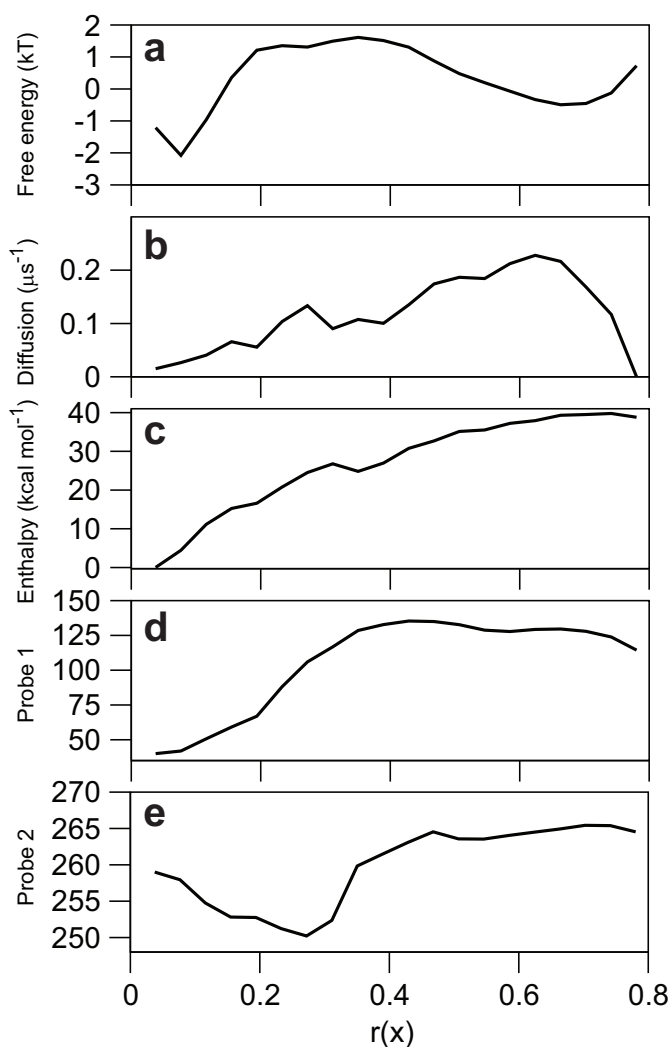
Fig. S5. **Simulating a temperature-jump experiment on a one-dimensional folding free energy landscape of FiP35.** (**a**) Free energy profile along the reaction coordinate obtained from analysis of equilibrium simulations. (**b**) Position-dependent diffusion constant. (**c**) Average enthalpy of structures in each region of $r(x)$. We used two different quantities to link the structures seen in MD simulations to the experimental fluorescence measurements. Probe 1 (**d**) is the solvent accessible surface area (SASA) of the Trp indole ring (measured in $Å^2$). Probe 2 (**e**) monitors more specifically the surface contact s between the Trp indole ring and the phenolic groups of the two Tyr residues. Similar to probe 1, this interaction was quantified as the SASA of the indole ring, but for probe 2 we only included burial by the phenolic groups when calculating the SASA (i.e., low values of Probe 2 correspond to increased contacts).
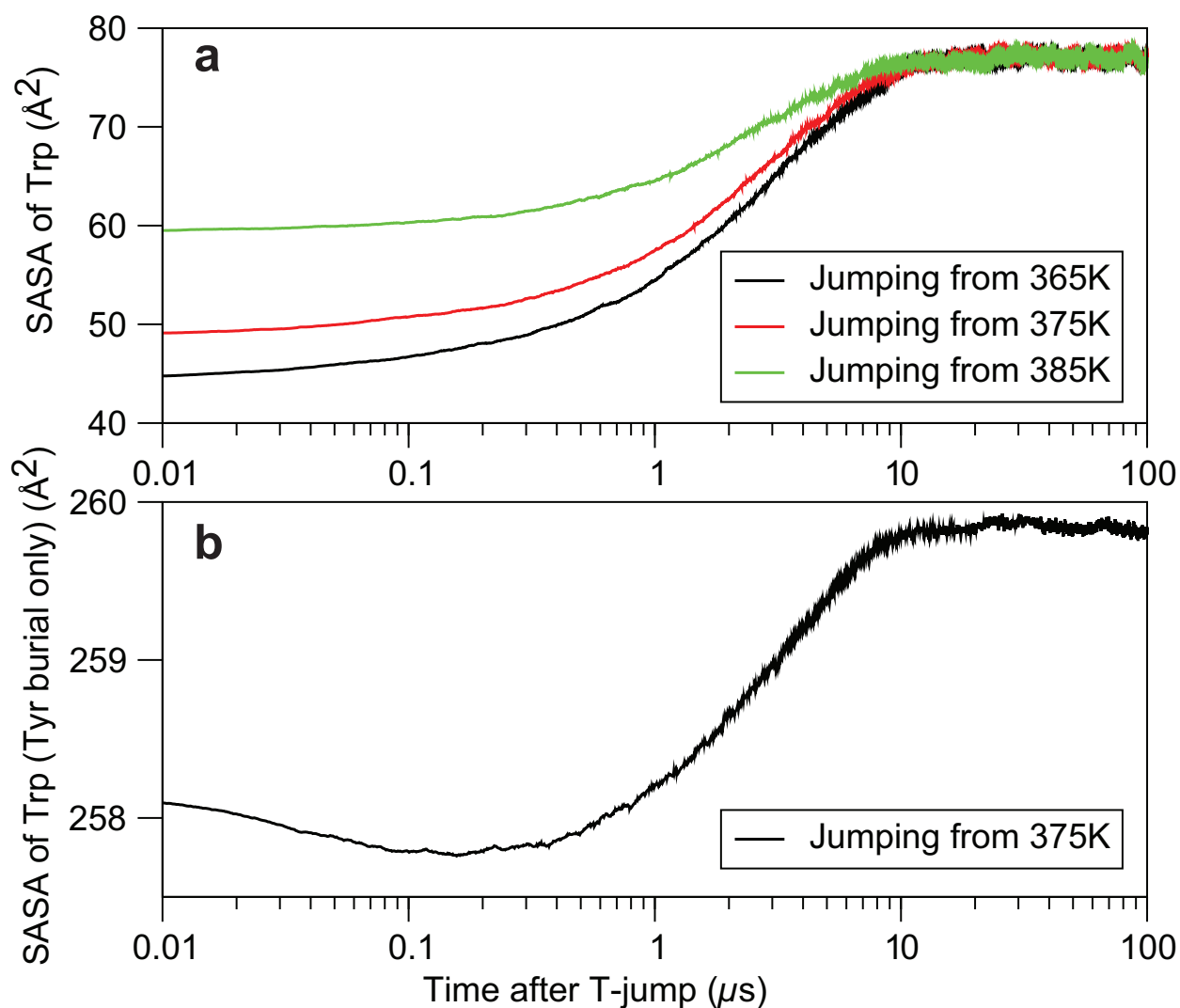
Fig. S6. **Effect of temperature and probe on simulated T-jump experiment probing folding of FiP35**. (**a**) Results of three T-jump experiments from 365 K (black), 375 K (red), and 385 K (green), all to a final temperature of 395 K. The plots show the time evolution of the SASA of the Trp indole ring (probe 1 in Fig. S2). (**b**) T-jump from 375 K to 395 K. The plot shows the time evolution of the interaction between the Trp indole ring and Tyr side chains (probe 2 in Fig. S2). All plots were obtained from averaging over a large number of Langevin simulations as described in the text.
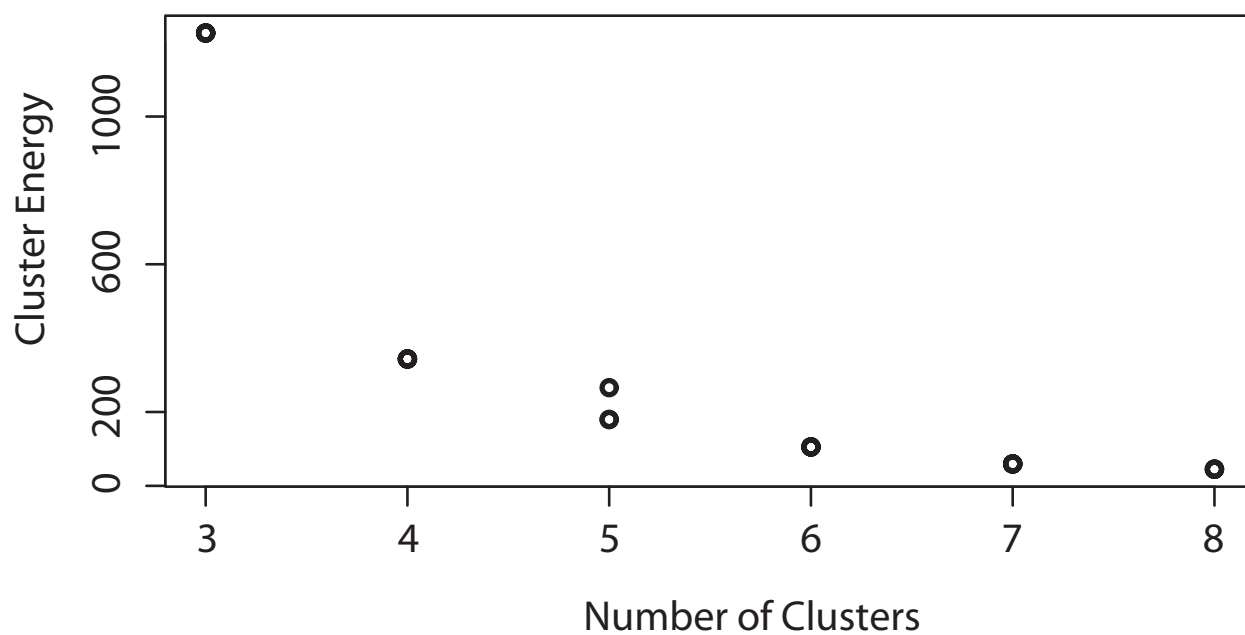
Fig. S7. The cluster energy as a function of the number of clusters in the trajectory of folded BPTI. The figure shows the results of three simulated annealing runs for each choice of the number of clusters. (Fewer than three points are visible for each cluster, because the cluster energies for different runs are very similar.)
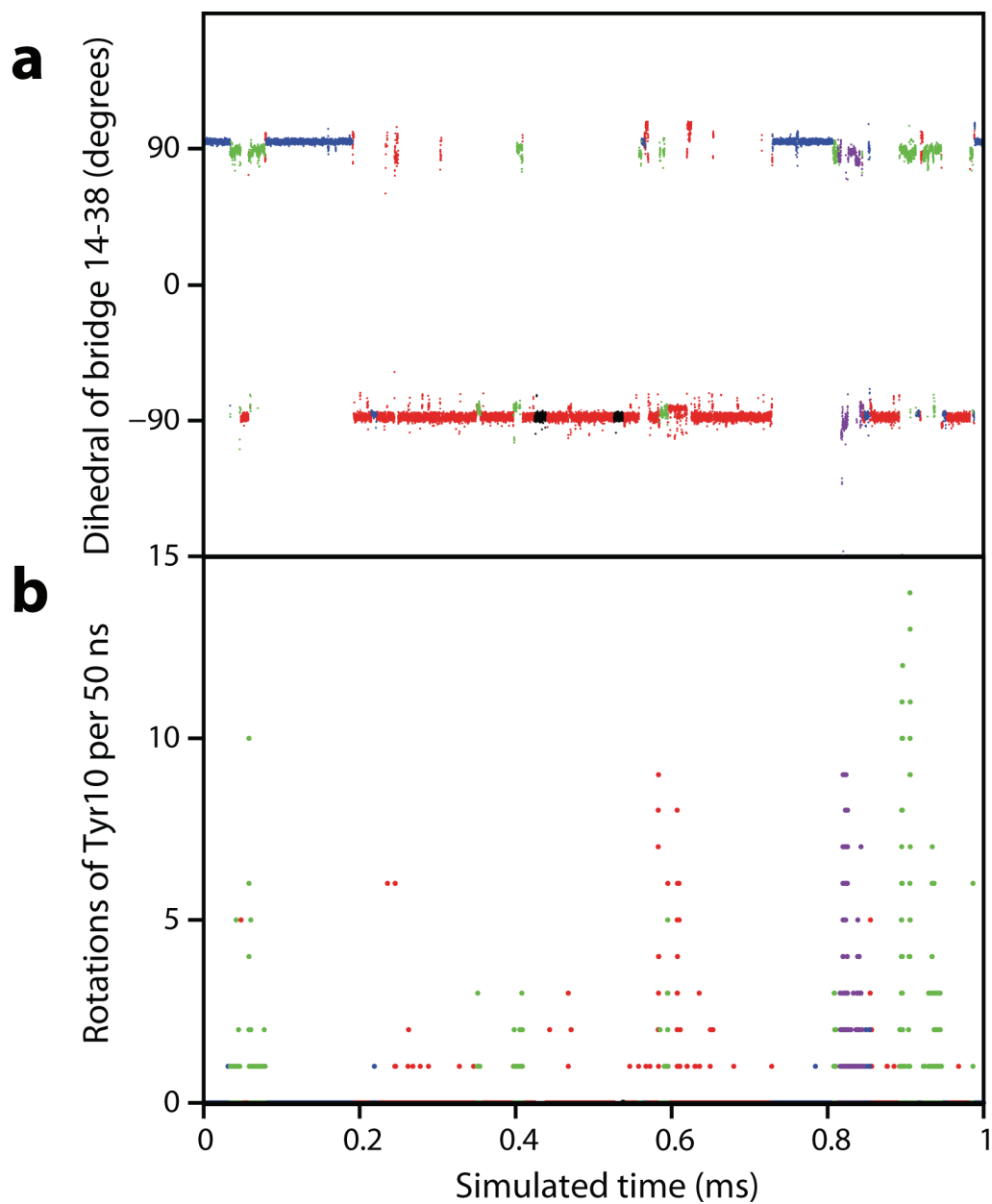
Fig. S8. Transitions among distinct conformational states of folded BPTI. Panel (**a**) shows the median values of the dihedral angle of the disulphide bond between residues Cys14 and Cys38 in windows of 50 ns. Panel (**b**) shows the number of flips per 50 ns of the aromatic ring of Tyr10. The color of the data points denotes cluster membership.
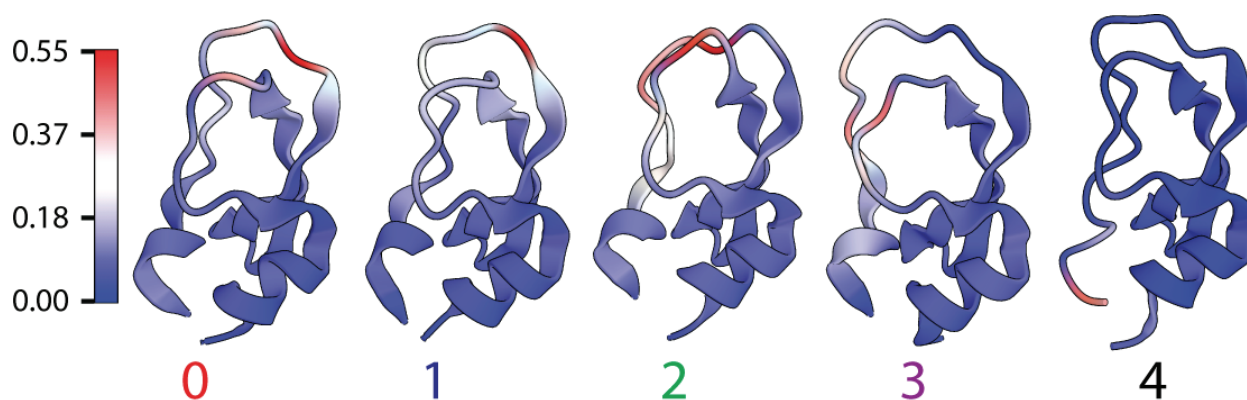
Fig. S9. The most discriminative amino acids for each state of folded BPTI. The figure shows a representative structure for each cluster (see supplementary snapshots of the BPTI trajectory). The color coding of each structure shows the value of the mean correlation coefficient listed in table S2. The entries in table S2 correspond to the residues that are colored red in this figure. The figure uses a non-photo-realistic rendering style for clarity: diffuse and specular lighting is replaced with a hemispheric lighting model, black outlines emphasize the overall shape, and the inside of the helices have a touch of gray to improve contrast.
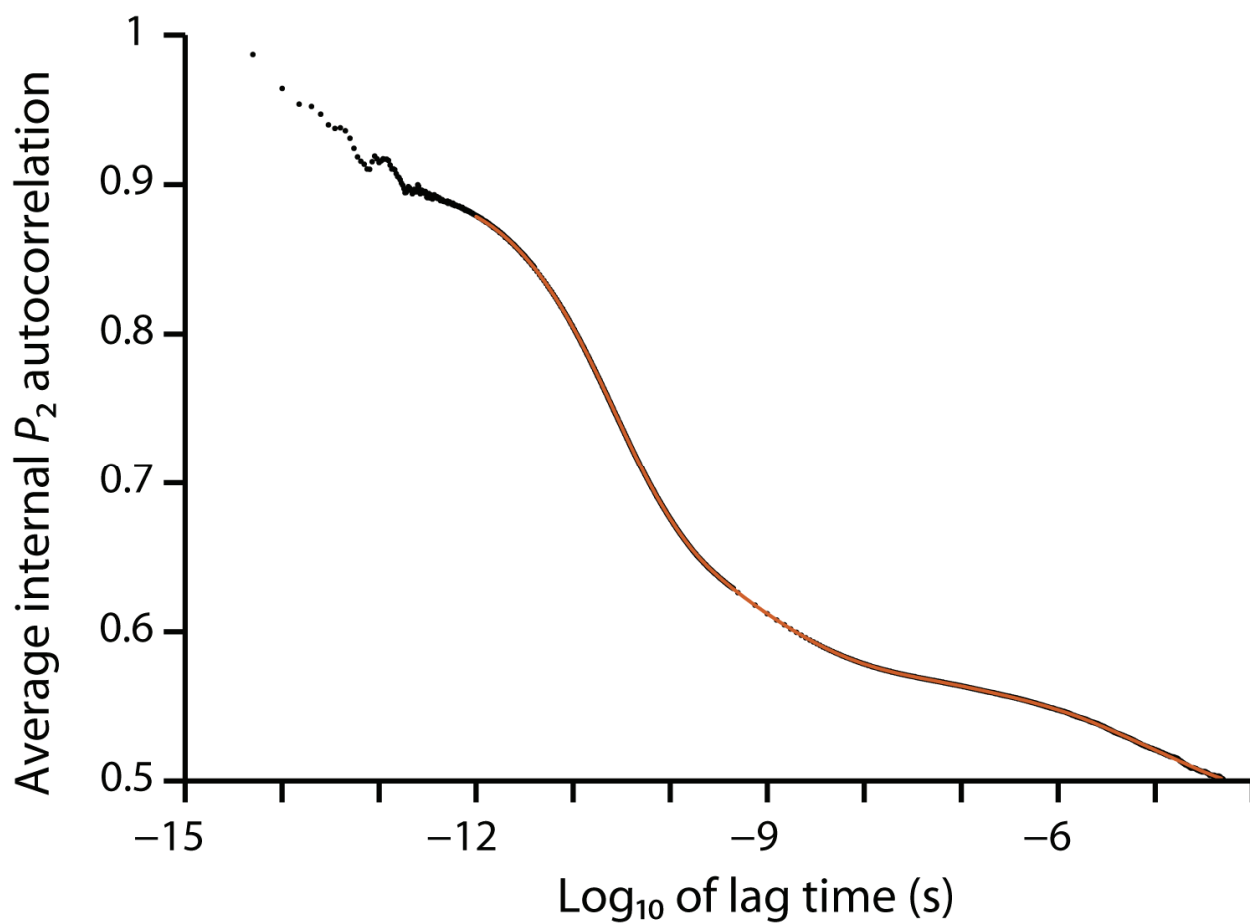
Fig. S10. Average relaxation of bond directions of folded BPTI. The figure shows the average (over 304 bonds) of the autocorrelation of the bond directions in the molecular frame of reference (internal P_2 correlation function; black) and its model fit (orange). The peaks at short lag times correspond to the quasi-periodic fluctuations of the high-frequency angles.
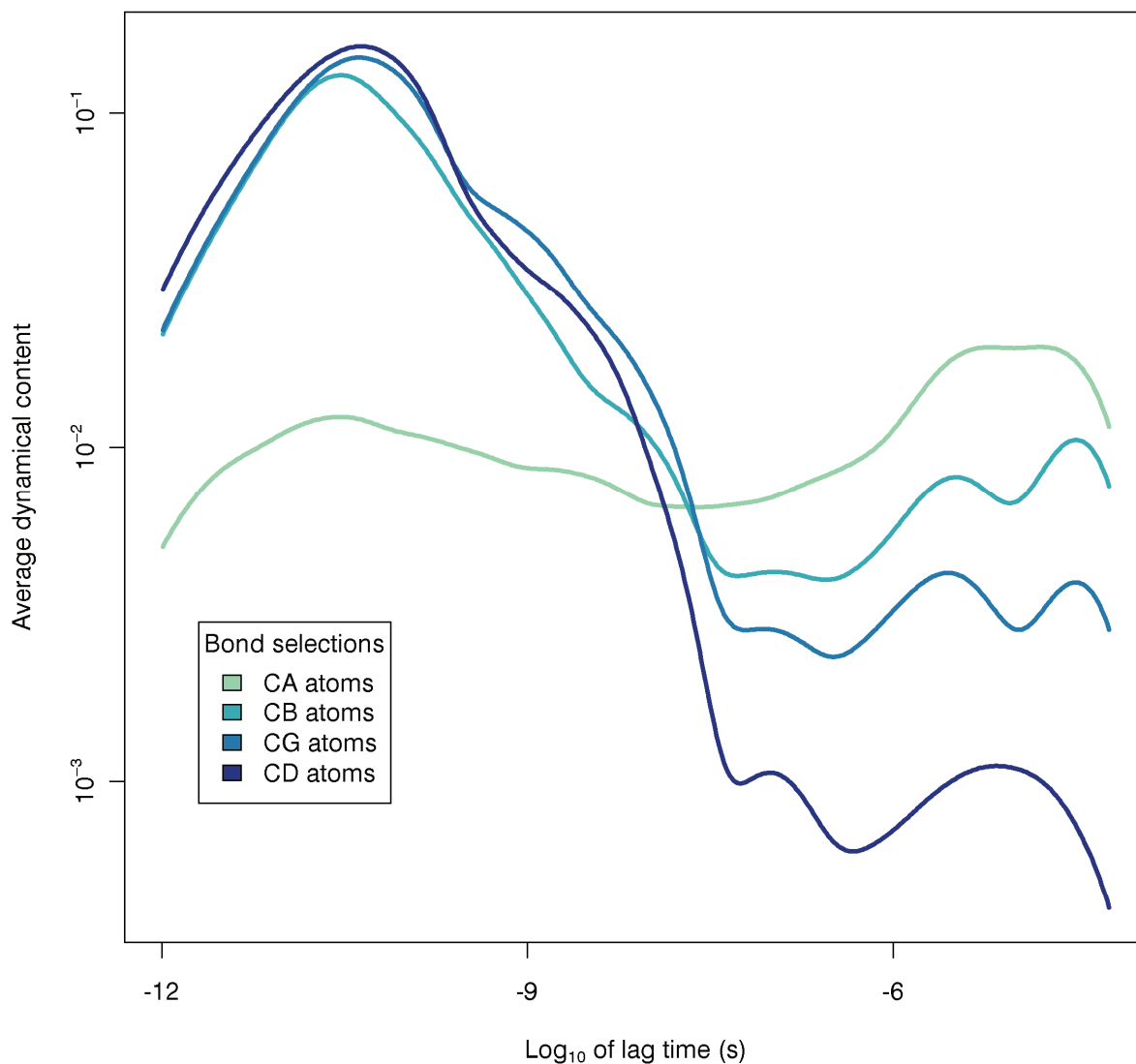
Fig. S11. The dynamical content of bonds to selected carbons of folded BPTI. The figure shows the average dynamical content coming from fluctuations of the subset of bonds that contain the atoms C_α (CA), C_β (CB), C_γ (CG), and C_δ (CD). The contribution to the relaxation at slow timescales decreases with the distance of the bonds from the backbone, which explains that the small contribution of the side chains to the long-timescale relaxation was due primarily to the side-chain atoms nearest the backbone.
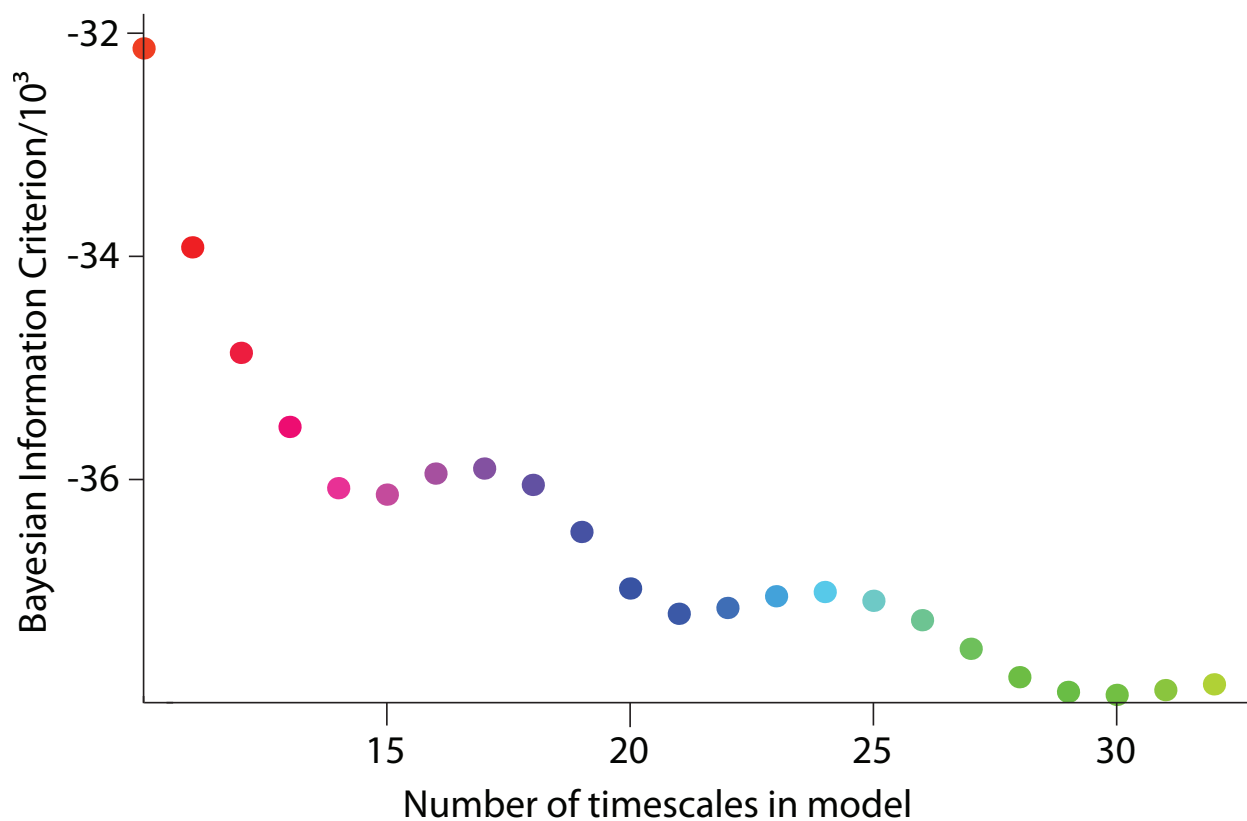
Fig. S12. The Bayesian Information Criterion of fits to the average correlation function of folded BPTI. The Bayesian Information Criterion of the fit to the average internal correlation function as a function of the number of timescales included in the Mathematica model.
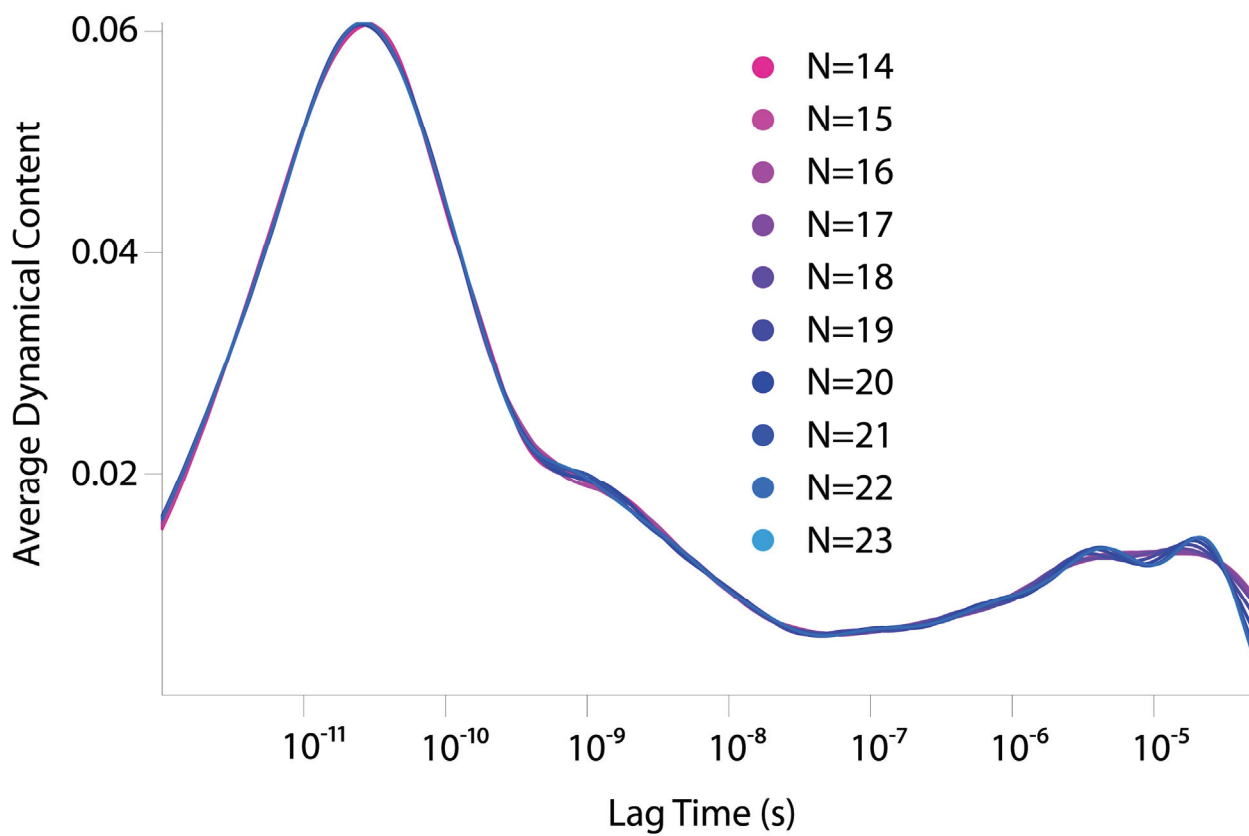
Fig. S13. Models of the average dynamical content of the BPTI simulation. The average dynamical content of the BPTI simulation, as estimated from a sequence of models with the number exponentials, *N*, used to represent the correlation function ranging from 14 to 23.
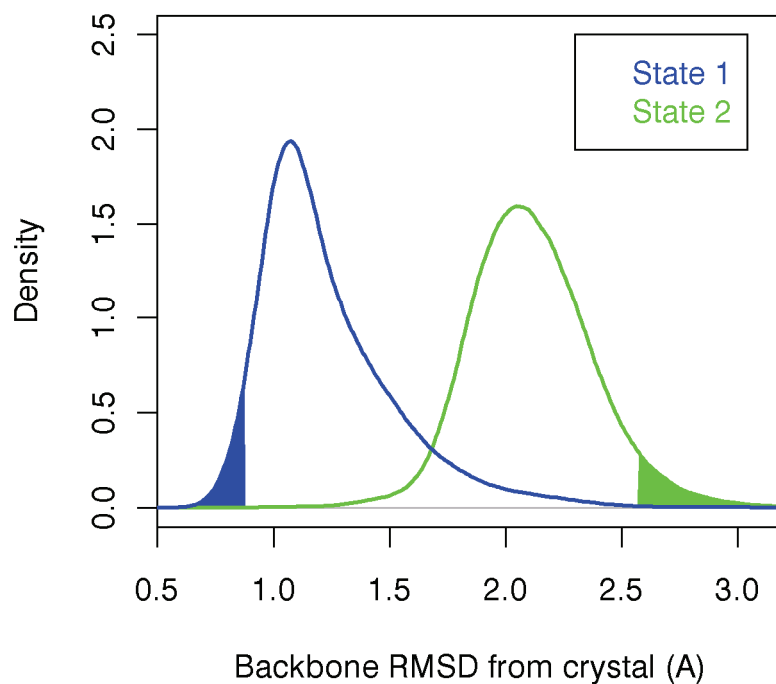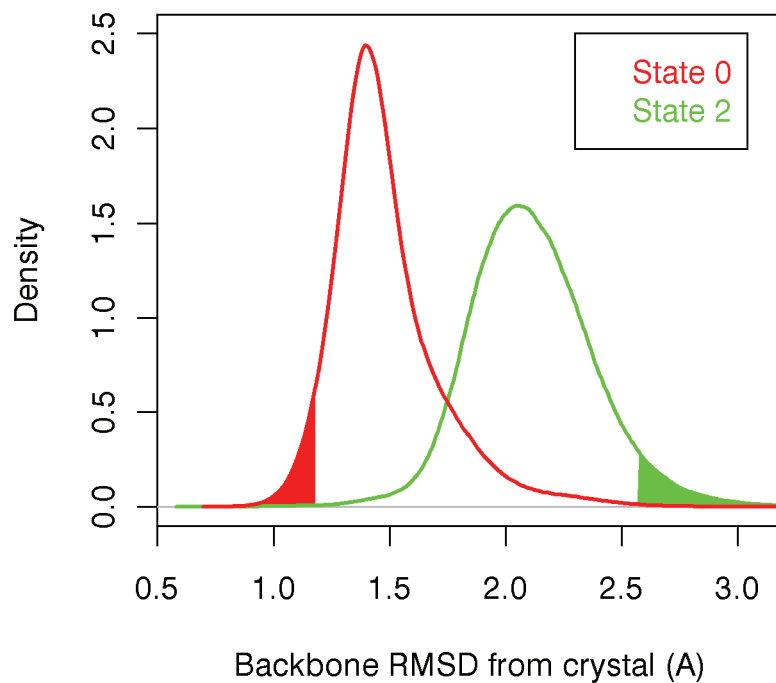
Fig. S14. RMSD distributions for selected states of folded BPTI. The figure shows the distribution of backbone RMSD from the crystal structure for selected states. It also highlights their 5% and 95% quantiles, which we used to define the transition paths.
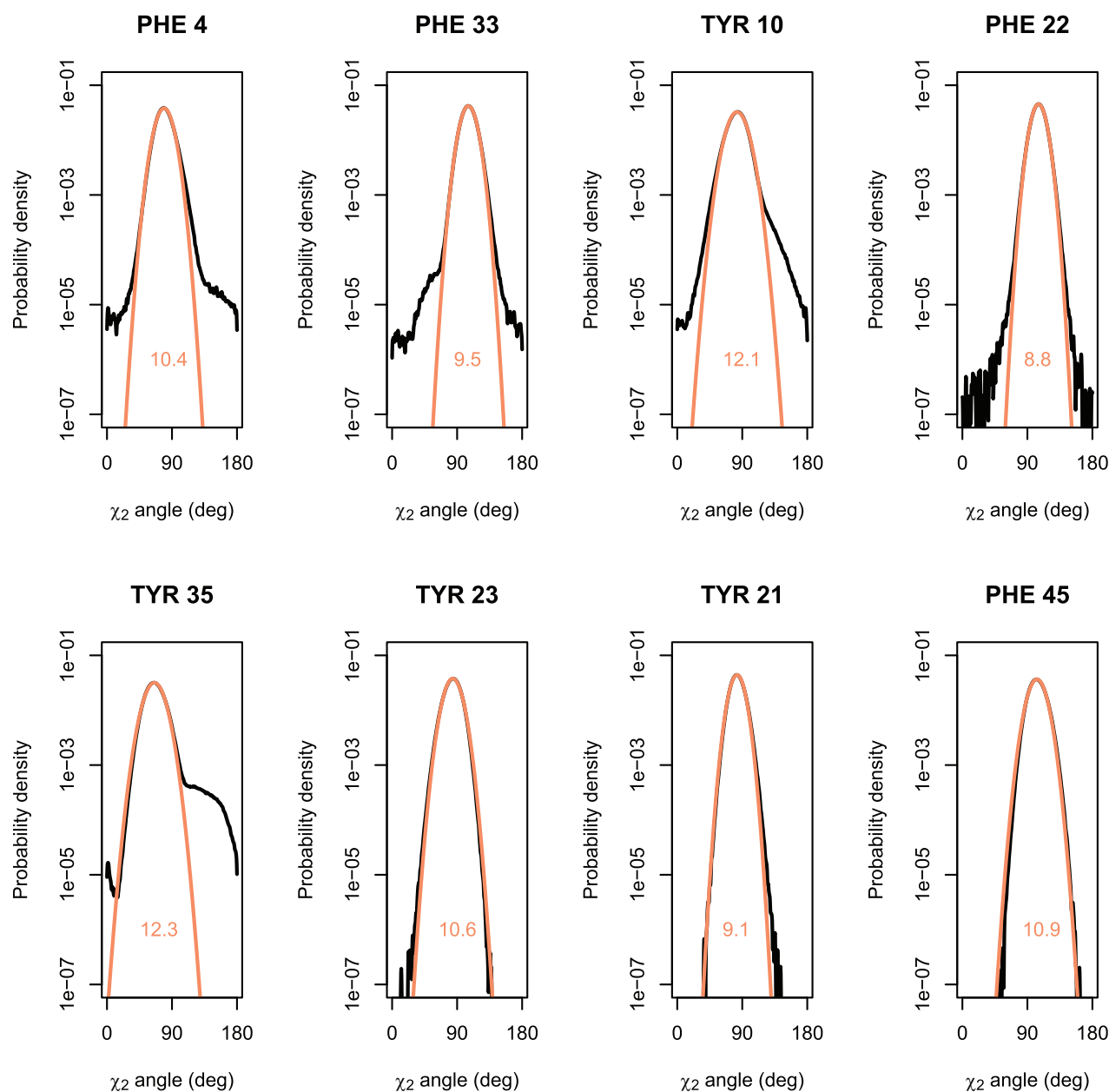
Fig. S15. Distributions of χ_2 dihedrals for each aromatic ring of folded BPTI. The panels show the probability density functions of the aromatic rings' χ_2 dihedral angle for each of the eight aromatic rings of BPTI. The panels are ordered, from left to right, top to bottom, according to the number of flips of the corresponding ring during the course of the millisecond simulation of BPTI. The orange curves show the fits of the density to a normal distribution; the orange number is the standard deviation, in degrees, of the corresponding normal distribution.

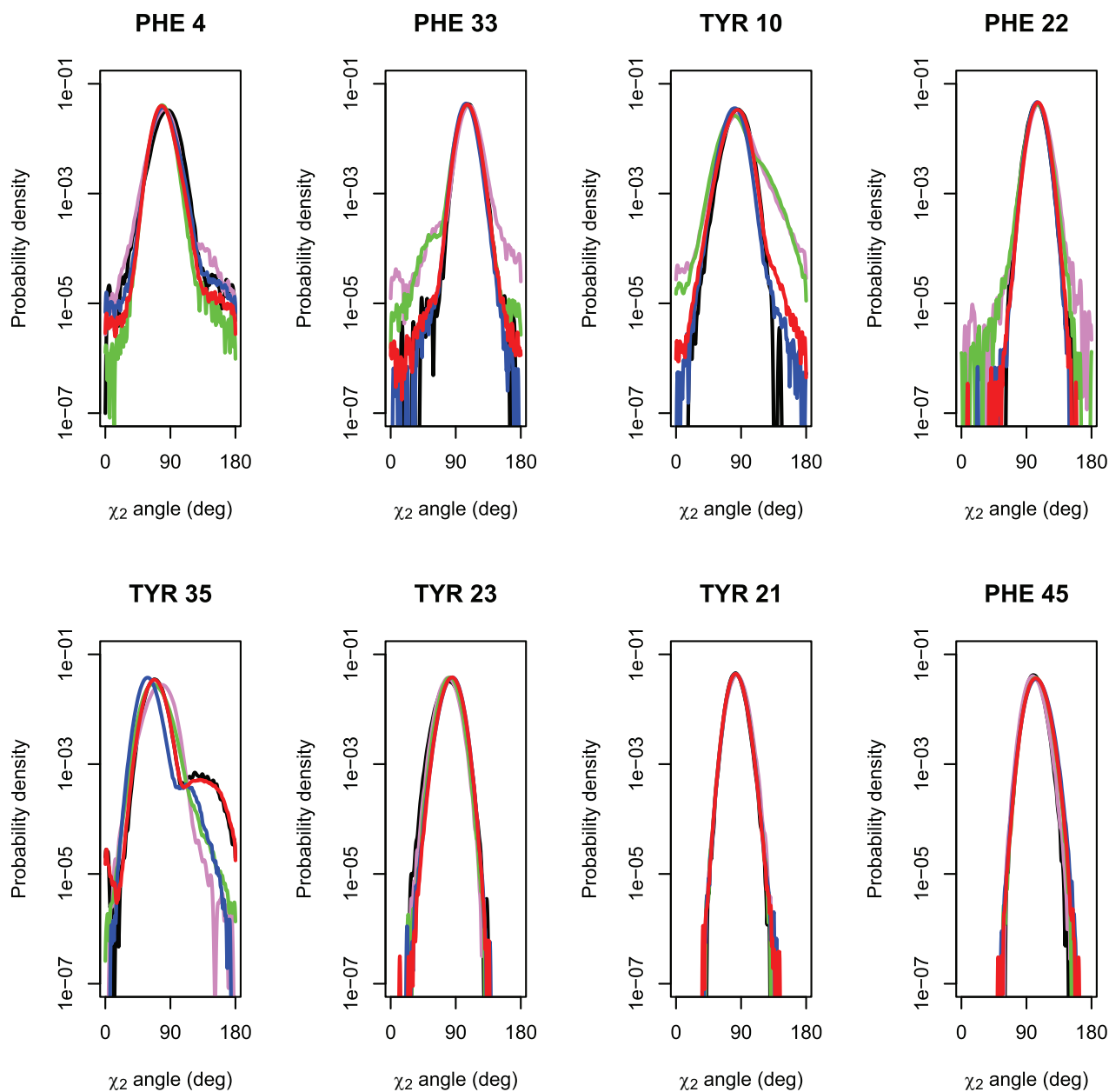**Fig. S16. Decomposition of distributions of χ_2 dihedrals on each state of folded BPTI.** The panels show the probability density function of the χ_2 dihedral angle for each of the eight aromatic rings of BPTI, confined to state 0 (red), 1 (blue), 2 (green), 3 (violet), 4 (black). The order of the panels is that of Fig. S14.

Supporting Tables

Fragment	Simulation time (μs)	Hairpin folding time (μs)	Stability (kcal mol^{-1})
HP 2 (19-30)	144	0.86 (0.06)	2.47 (0.2)
HP 1 (7-23)	131	1.2 (0.1)	0.87 (0.22)
7–26	44	1.5 (0.3)	1.26 (0.17)
7–29	69	1.8 (0.3)	1.01 (0.19)
7–32	117	2.3 (0.4)	0.66 (0.16)
7–35	138	2.5 (0.5)	0.17 (0.22)
4–23	122	2.3 (0.4)	1.13 (0.16)
4–26	120	2.9 (0.5)	1.27 (0.18)
4–29	90	2.7 (0.5)	0.49 (0.18)
4–32	119	3.6 (0.8)	0.24 (0.25)
4–35	133	4.9 (1.7)	−0.79 (0.46)
1–23	140	2.5 (0.4)	1.16 (0.14)
1–26	119	2.9 (0.5)	1.04 (0.20)
1–29	141	3.9 (0.7)	0.98 (0.22)
1–32	138	4.5 (0.9)	1.02 (0.24)
1–35	200	5. (2)	−0.44 (0.38)

Table S1. Kinetics and thermodynamics of FiP35 peptide fragments. The total simulation time for each fragment is reported together with the folding rates and stabilities of hairpin 1 in the context of the different sequences obtained from reversible folding simulations at 395 K. A dual-cutoff approach was used to determine folding/unfolding transitions: a folding event was

counted every time the hairpin was unfolded and the Cα rmsd of residues 8–22 with respect to the crystal structure become smaller than 1.25 Å, and an unfolding event was counted every time the hairpin was folded and the Cα rmsd become larger than 5.5 Å. The same approach was used for hairpin 2, but the Cα rmsd was computed on residues 19–30 and cutoffs of 1.8 and 4.9 Å were used for folding and unfolding, respectively.

State	Res Name	Res ID	Stabilization coefficient
0	ALA	16	0.541762
0	CYS	38	0.372383
0	CYS	14	0.334574
1	ALA	16	0.508978
2	CYS	14	0.531035
2	CYS	38	0.483888
2	THR	11	0.393154
2	PRO	13	0.381127
2	GLY	37	0.373036
2	GLY	12	0.350927
3	ARG	39	0.418139
3	PRO	9	0.400069
3	THR	11	0.305087
3	LYS	41	0.303865
4	PRO	2	0.423836
4	ARG	1	0.373713

Table S2. The most discriminative amino acids for each state of folded BPTI. The table lists the amino acids (residue name and residue number) with stabilization coefficient larger than 0.3. The last column lists the stabilization coefficient for that amino acid in the state listed in the first column of the table.

Aromatic group	Rotation rate (s^{-1})		
	Simulation (300 K)	Experiment (313 K)	Experiment (309 K)
TYR10	1×10^6	fast	fast
TYR21	1×10^3	fast	fast
TYR23	1×10^3	3×10^2	3×10^2
TYR35	1×10^4	5×10^1	5×10^1
PHE4	7×10^6	fast	fast
PHE22	7×10^4	2×10^3	fast
PHE33	6×10^6	fast	fast
PHE45	0	5×10^1 at 353 K	2×10^3

Table S3. Comparison of experimental and simulated aromatic ring rotation rates. The aromatic ring rotation rates from simulation (this work); the original experimental investigation at 313 K (*47*); and the reinvestigation of the rotation rates at 309 K (*48*).

State	p	τ (ms)	Rate (ms^{-1})				
			PHE4	TYR10	PHE33	PHE22	TYR35
0	0.551	25.8	5331	376	833	4	9
1	0.268	25.1	12792	48	185	0	0
2	0.128	8.8	3287	5280	419	240	16
3	0.029	6.0	12607	14766	206	1131	103
4	0.025	12.8	1684	0	80	0	0

Table S4. The aromatic ring flips in each state of BPTI. The table lists the population p and the mean residence time τ for each of the five states of BPTI in the millisecond MD simulation. The table also lists the number of aromatic ring flips and corresponding rate of rotation in each state, for each of the five rings of BPTI that are colored orange in Fig. 5c. We estimate the population p for state m by taking the fraction of frames from the cluster sequence that equal m. We estimate the mean residence time τ for state m by taking the mean length of consecutive segments from the cluster sequence that equal m, and multiplying the result with the time spacing of the cluster sequence. We calculate the number of aromatic ring flips in state m by counting the ring flip events that occur while BPTI is in state m; we calculate the corresponding rate of rotation in state m by multiplying that state's population p with the number of ring flips in state m.

Supporting References

1. F. Liu *et al,* An experimental survey of the transition between two-state and downhill protein folding scenarios. *Proc. Natl. Acad. Sci. USA* **105**, 2369–2374 (2008).

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

3. W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, M. L. Klein, Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys*. **79**, 926–935 (1983).

4. V. Hornak *et al*, Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins Struct. Funct. Bioinf.* **65**, 712–725 (2006).

5. K. Lindorff-Larsen *et al*, Improved side-chain torsion potentials for the Amber99SB protein force field. *Proteins Struct. Funct. Bioinf*. **78**, 1950–1958 (2010).

6. K. J. Bowers *et al*, Scalable algorithms for molecular dynamics simulations on commodity clusters. *Proc. ACM/IEEE Conf. on Supercomputing (SC06)*, Tampa, FL, November 11–17, 2006.

7. D. E. Shaw *et al*, Anton: A special-purpose machine for molecular dynamics simulation. *Proc. 34th Annl. Intnl. Symp. on Computer Architecture (ISCA '07)*, San Diego, CA, June 9–13, 2007.

8. R. A. Lippert *et al*, A common, avoidable source of error in molecular dynamics integrators. *J. Chem. Phys*. **126**, 046101 (2007).

9. Y. Shan *et al*, Gaussian split Ewald: A fast Ewald mesh method for molecular simulation. *J. Chem. Phys*. **122**, 054101 (2005).

10. U. Essmann *et al*, A smooth particle mesh Ewald method. *J. Chem. Phys*. **103**, 8577–8593 (1995).

11. M. Tuckerman, B. J. Berne, G. J. Martyna, Reversible multiple time scale molecular dynamics. *J. Chem. Phys*. **97**, 1990–2001 (1992).

12. W. Humphrey, A. Dalke, K. Schulten, VMD—Visual Molecular Dynamics. *J. Molec. Graphics* **14**, 33–38 (1996).

13. D. Branduardi, F. L. Gervasio, M. Parrinello, From A to B in free energy space. *J. Chem. Phys*. **126**, 054103 (2007).

14. A. Berteotti *et al*, Protein conformational transitions: the closure mechanism of a kinase explored by atomistic simulations. *J. Am. Chem. Soc*. **131**, 244–250 (2009).

15. O. Carugo, S. Pongor, A normalized root-mean-square distance for comparing protein three-dimensional structures. *Protein Sci.* **10**, 1470–1473 (2001).

16. A. Laio, M. Parrinello, Escaping free-energy minima. *Proc. Nat. Acad. Sci. USA* **99**, 12562–12566 (2002).

17. G. Bussi, F. L. Gervasio, A. Laio, M. Parrinello, Free-energy landscape for beta-hairpin folding from combined parallel tempering and metadynamics. *J. Am. Chem. Soc*. **128**, 13435–13441 (2006).

18. R. B. Best, G. Hummer, Optimized molecular dynamics force fields applied to the helix-coil transition of polypeptides. *J. Phys. Chem. B*. **113**, 9004–15 (2009).

19. R. B. Best, N. V. Buchete, G. Hummer, Are current molecular dynamics force fields too helical? *Biophys. J.* **95**, L07–L09 (2008).

20. G. Hummer, Position-dependent diffusion coefficients and free energies from Bayesian analysis of equilibrium and replica molecular dynamics simulations. *New J. Phys.* **7**, 34 (2005).

21. R. B. Best, G. Hummer, Reaction coordinates and rates from transition paths. *Proc. Natl. Acad. Sci. USA* **102**, 6732–6737 (2005).

22. G. Hummer, From transition paths to transition states and rate coefficients. *J. Chem. Phys*. **120**, 516–523 (2004).

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

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

25. K. Lindorff-Larsen, E. Paci, L. Serrano, C. M. Dobson, M. Vendruscolo, Calculation of mutational free energy changes in transition states for protein folding. *Biophys. J.* **85**, 1207–1214 (2003).

26. X. Salvatella, C. M. Dobson, A. R. Fersht, M. Vendruscolo, Determination of the folding transition states of barnase by using φ_I-value-restrained simulations validated by double mutant φ_{IJ}-values. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 12389–12394 (2005).

27. E. Cota, S. J. Hamill, S. B. Fowler, J. Clarke, Two proteins with the same structure respond very differently to mutation: the role of plasticity in protein stability. *J. Mol. Biol.* **302**, 713–725 (2000).

28. R. W. Zwanzig, High-temperature equation of state by a perturbation method. I. Nonpolar gases. *J. Chem. Phys*. **22**, 1420–1426 (1954).

29. A. Wlodawer, J. Walter, R. Huber, L. Sjölin, Structure of bovine pancreatic trypsin inhibitor. Results of joint neutron and X-ray refinement of crystal form II. *J. Mol. Biol.* **180**, 301–329 (1984).

30. Maestro, Version 8.5.015. Schrödinger, LLC, New York, NY (1999–2007).

31. H. W. Horn *et al*, Development of an improved four-site water model for biomolecular simulations: TIP4P-Ew. *J. Chem. Phys.* **120**, 9665–9678 (2004).

32. V. Wong, D. A. Case, Evaluating rotational diffusion from protein MD simulations. *J. Phys. Chem. B* **112**, 6013–6024 (2008).

33. H. J. C. Berendsen *et al*, Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **81**, 3684–3690 (1984).

34. G. J. Martyna, D. J. Tobias, M. L. Klein, Constant pressure molecular dynamics algorithms. *J. Chem. Phys.* **101**, 4177–4189 (1994).

35. D. E. Shaw Research. Desmond User's Guide. http://www.deshawresearch.com/resources_desmond.html (2009).

36. Tu, T. *et al*, A scalable parallel framework for analyzing terascale molecular dynamics simulation trajectories. *Proc. ACM/IEEE Conf. on Supercomputing (SC08)*, Austin, Texas, November 15–21, 2008.

37. W. Wriggers *et al*, Automated event detection and activity monitoring in long molecular dynamics simulations. *J. Chem. Theory Comput.* **5**, 2595–2605 (2009).

38. P. Maragakis *et al*, Microsecond molecular dynamics simulation shows effect of slow loop dynamics on backbone amide order parameters of proteins. *J. Phys. Chem. B* **112**, 6155–6158 (2008).

39. Mathematica, Version 7.0. Wolfram Research, Inc., Champaign, IL (2008).

40. G. Schwarz, Estimating the dimension of a model. *Annal. Stat.* **6,** 461–464 (1978).

41. N.-V. Buchete, G. Hummer, Peptide folding kinetics from replica exchange molecular dynamics. *Phys. Rev. E* **77**, 030902(R) (2008).

42. R. F. Grote, J. T. Hynes, The stable states picture of chemical reactions. II. Rate constants for condensed and gas phase reaction models. *J. Chem. Phys.* **73**, 2715–2732 (1980).

43. D. L. Theobald, D. S. Wuttke, THESEUS: maximum likelihood superpositioning and analysis of macromolecular structures. *Bioinformatics* **22**:2171–2172 (2006).

44. G. de Prat Gay *et al*, Conformational pathway of the polypeptide chain of chymotrypsin inhibitor-2 growing from its N terminus *in vitro*. Parallels with the protein folding pathway. *J. Mol. Biol*. **254**, 968–979 (1995).

45. J. L. Neira, A. R. Fersht, Acquisition of native-like interactions in C-terminal fragments of barnase. *J. Mol. Biol.* **287**, 421–432 (1999).

46. S. Deechongkit *et al*, Context-dependent contributions of backbone hydrogen bonding to β-sheet energetics. *Nature* **430**, 101–105 (2004).

47. G. Wagner, A. DeMarco, K. Wüthrich, Dynamics of the aromatic amino acid residues in the globular conformation of the basic pancreatic trypsin inhibitor (BPTI). I. 1H NMR studies. *Biophys. Struct. Mech.* **2**, 139–158 (1976).

48. G. Wagner, D. Brühwiler, K. Wüthrich, Reinvestigation of the aromatic side-chains in the basic pancreatic trypsin inhibitor by heteronuclear two-dimensional nuclear magnetic resonance. *J. Mol. Biol.* **196**, 227–231 (1987).