Supporting Information

Rao et al. 10.1073/pnas.0915087107

SI Text

MD Simulation. ***The GS10 model.*** The system investigated is a 10-residue peptide (arbitrarily called GS10 throughout the text) defined by the sequence AAAGSAAA with N-acetylated (ACE) and C-amidated (CBX) blocking groups. The GS motif ensures increased stability of the β-hairpin structure because the G and S residues favor a turn (S1). Molecular dynamics (MD) simulations, using the Langevin algorithm and friction coefficient equal to 50 ps^{-1}, were run using the CHARMM program (S2, S3) with polar hydrogens (PARAM19) and SHAKE, so that an integration step of 2 fs could be used. A simple implicit model based on the solvent-accessible surface was used to describe the main effects of the aqueous solvent on the solute. In this model, the solvation free energy is given by $G_{\mathrm{solv}} = \sum_{i=1}^{M} \sigma_i A_i(\mathbf{r})$ for a molecule having M heavy atoms with Cartesian coordinates $\mathbf{r} = (r_1, \ldots r_M)$. $A_i(\mathbf{r})$ is the solvent-accessible surface computed by an approximate analytical expression (S4) using a 1.4-Å probe radius. The model contains only two surface-tension-like parameters: one for carbon and sulfur atoms ($\sigma_{\mathrm{C,S}} = 0.012$ kcal/mol $Å^2$), and one for nitrogen and oxygen atoms ($\sigma_{\mathrm{N,O}} = -0.060$ kcal/mol $Å^2$) (S5).

Inherent Structures. The minimization in the present study was performed by the steepest descent method followed by application of the adopted basis Newton-Raphson (ABNR) algorithm (S2, S3). The former is used to quench the system to the closest potential energy minimum, while the latter is required for convergence of the minimization. Structurally different inherent structures (IS) (i.e., conformations belonging to different regions of the energy surface) were found have very similar energies; for example, snapshots characterized by energy differences within 0.001 kcal/mol have an average all-atom rmsd of 1.83 Å, a value close to the overall average rmsd pairwise difference. This value of the energy difference threshold is the limiting value that can be achieved with the present potential because of the nature of the solvation model. To overcome this problem, the set of IS is defined as the clusters obtained by the application of the leader clustering algorithm (S6, S7) with a very small cutoff of 0.15 Å to the time series of the minimized snapshots. This procedure is robust for different values of the cutoff, provided that the latter is small enough. Clustering with values of the cutoff as small as 0.05 Å give quantitatively similar results (see Fig. S7).

The leader clusterization algorithm. The leader algorithm (S6, S7) is among the fastest and least memory-demanding approaches usable for rmsd-based structural clustering. The algorithm works as follows: The first snapshot of the trajectory is taken as the center of the first cluster. The second snapshot is then compared with the center of the first cluster, if the distance between the two (in this case the rmsd) is smaller than the user-defined cutoff then the second snapshot is assigned to the first cluster; otherwise, a new cluster is created and the snapshot becomes the center of it. Comparisons are always performed on the oldest created clusters first and against cluster centers only. This procedure is iterated till the last snapshot of the trajectory is reached.

Conformation Space Networks. Each microstate (IS) represents a node of the conformation space network (CSN) and a link is present if a *direct* transition between two microstates has been observed during the time series in a time step of a given size S_t (S8). The value of S_t can be as small as the integration time step but can also be larger; for example, considering one snapshot in the time series every 10 or 100 integration steps results in network links describing transitions that occur in that time interval. In the limit of values of S_t much longer than the slowest transition of the system, the CSN will be fully connected (i.e., every node is connected to all the others).

Mincut-Based Free-Energy Profile. In this work, the mean first passage time (m) version of the algorithm was employed (S9). Given the original CSN, m_i is the solution of the equation $m_i = \Delta t + \sum p_{ij} \times m_j$ with boundary condition $m_R = 0$; here p_{ij} is the transition probability between i and j given by the CSN. The time step Δt corresponds to the saving frequency S_t used to build the CSN. Microstates are sorted by increasing values of m_i. Each microstate is labeled according to the cumulative value of the partition function (following the ranking) normalized by the value of the total partition function Z_{tot}, i.e., $Z_i = \sum_{j \le i} z_j / Z_{\mathrm{tot}}$, where z_j is the population of microstate j. The free-energy profile is obtained for each value of the cumulative partition function Z_i by calculating $\Delta F = -k_B T \log(W_i / Z_{\mathrm{tot}})$, where W_i is the number of transitions calculated from the CSN between the group of microstates (i.e., network nodes) with cumulative partition function smaller than Z_i and the rest. The cut-based free-energy profile (CFEP) analysis was done using the program WORDOM (S7, S9).

Coarse Configuration Space Network. A coarse version of the CSN (CCSN) is built by considering only a subset of statistically relevant microstates identified by the CFEP analysis. The essential idea is to keep only a single microstate near the bottom of each valley (e.g., the lowest energy microstate) and determine the transitions between this selected states. The CCSN is obtained from a modified time series in which only the subset of snapshots is kept and all the others are deleted. In this way one excludes, in particular, the many IS in the neighborhood of the top of the barrier (see Fig. S9) where multiple recrossings occur and thereby gets a meaningful measure of the intermicrostate transitions. If there are intermediates that play a role, they have to be explicitly included in the selected subset. A test for the correct description of the intervalley transitions by the CCSN is provided by the agreement between the first passage time distributions from a Markov model of the CCSN and the MD simulation; see Fig. S10.

MD Simulation and Microstate Definition for the PDZ2 Domain. The PDZ2 signaling domain [Protein Data Bank (PDB) ID code 3PDZ] was simulated for 1.6 ns with the CHARMM program (S2, S3) using polar hydrogens (PARAM19) and SHAKE, so that an integration step of 2 fs could be used. The Langevin algorithm with a friction coefficient equal to 1 ps^{-1} was applied. The effect of the water molecules is approximated with an implicit solvation model (FACTS) based on the generalized born approximation (S10). Trajectory conformations were saved every 20 fs for a total of 80,000 snapshots. The trajectory was then minimized using the same procedure applied for the GS10 peptide. Due to the very large number of degrees of freedom, approximate IS are obtained by running the leader algorithm on the minimized trajectory considering the heavy atoms of the backbone and the C_β atoms with a cutoff of 0.2 Å. A total of 3,029 microstates were found. Application of the same clusterization algorithm on the unminimized trajectory with a cutoff of 0.6 Å resulted in 1,464 microstates.

1. Ferrara P, Caflisch A (2000) Folding simulations of a three-stranded antiparallel beta-sheet peptide. *Proc Natl Acad Sci USA* 97:10780–10785.
2. Brooks BR, et al. (1983) Charmm: A program for macromolecular energy, minimization, and dynamics calculations. *J Comput Chem* 4:187–217.
3. Brooks BR, et al. (2009) CHARMM: The biomolecular simulation program. *J Comput Chem* 30:1545–1614.
4. Hasel W, Hendrickson T, Still W (1988) Tetrahedron Comput. *Tetrahedron Comput Methodol* 1:103–116.
5. Ferrara P, Apostolakis J, Caflisch A (2002) Evaluation of a fast implicit solvent model for molecular dynamics simulations. *Proteins: Struct Funct Genet* 46:24–33.
6. Hartigan J (1975) *Clustering Algorithms* (New York: Wiley).
7. Seeber M, Cecchini M, Rao F, Settanni G, Caflisch A (2007) Wordom: A program for efficient analysis of molecular dynamics simulations. *Bioinformatics* 23:2625–2627.
8. Gfeller D, de Lachapelle DM, De Los Rios P, Caldarelli G, Rao F (2007) Uncovering the topology of configuration space networks. *Phys Rev E* 76:026113.
9. Krivov SV, Muff S, Caflisch A, Karplus M (2008) One-dimensional barrier-preserving free-energy projections of a beta-sheet miniprotein: New insights into the folding process. *J Phys Chem B* 112:8701–8714.
10. Haberthur U, Caflisch A (2008) FACTS: Fast analytical continuum treatment of solvation. *J Comput Chem* 29:701–715.
11. Fischer S, Karplus M (1992) Conjugate peak refinement: An algorithm for finding reaction paths and accurate transition states in systems with many degrees of freedom. *Chem Phys Lett* 194:252–261.

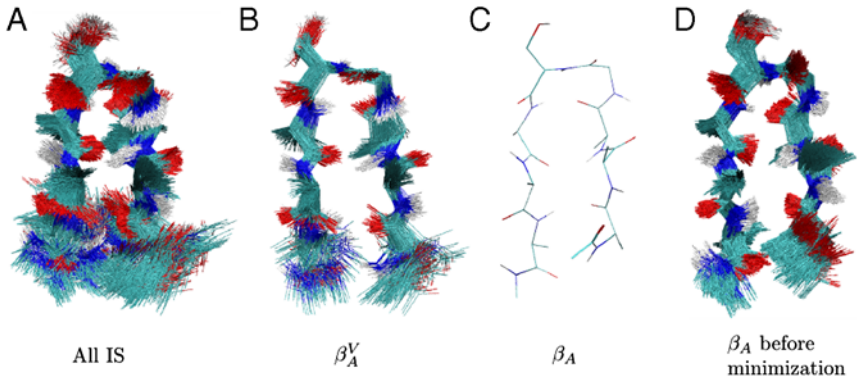

Fig. S1. Superposition of GS10 structures. (*A*) Superposition of the 1,561 IS, (*B*) the 263 IS belonging to the β_A^V valley, (*C*) the minimized snapshots belonging to the β_A IS, and (*D*) a sample of the unminimized snapshots belonging to β_A.

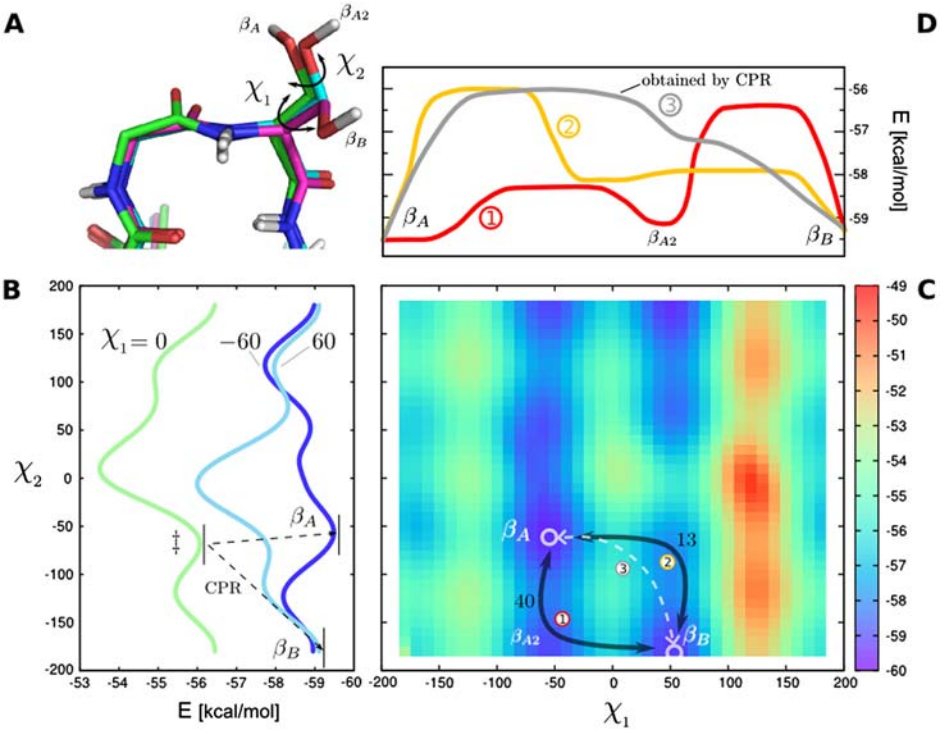

Fig. S2. Predicted and observed pathways energetics. (*A*) SER side chain orientations for β_A, β_{A2}, and β_B IS. The three IS differ in the values of χ_1 and χ_2 dihedral angles. (*B*) Potential energy calculated from β_A upon the variation of the χ_2 dihedral angle for different values of χ_1. The minimum energy pathway from β_A to β_B predicted by the conjugate peak refinement (CPR) algorithm (S11) is shown as a dashed line. (*C*) Two-dimensional potential energy calculated from β_A as a function of χ_1 and χ_2. The CPR pathway is shown as a dashed line, whereas the observed pathways from the simulation, with their respective populations, are shown in black. The latter pathways are obtained by the analysis of the time series of the IS and show the difference between the minimum energy pathway and the free-energy based pathway. (*D*) Potential energy profile along the three pathways shown in Fig. S1*C* (the evolution is in arbitrary units).

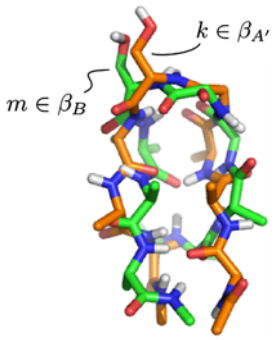

Fig. S3. Example of two microstates with the same energy but different structures; the rmsd is 2.4 Å.

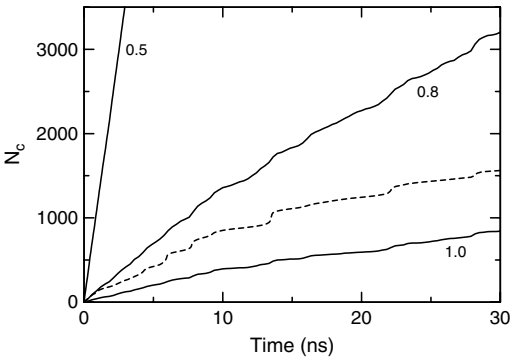

Fig. S4. Number of visited IS and all-atom rmsd microstates as a function of time. IS data are shown in the dashed line and indicate that the simulation is approximately converged at 30 ns. Three values of the cutoff are shown (see main text for details). A fit of the curves with a function of the form $N = \tilde{N}[1 - \exp(-t/\tau)]$ indicated that, for the IS case, the sampling covered 82% of the total number of expected microstates. In the case of the 0.8 and 1.0 Å cutoff realizations, this number is less than 60% for both cases.

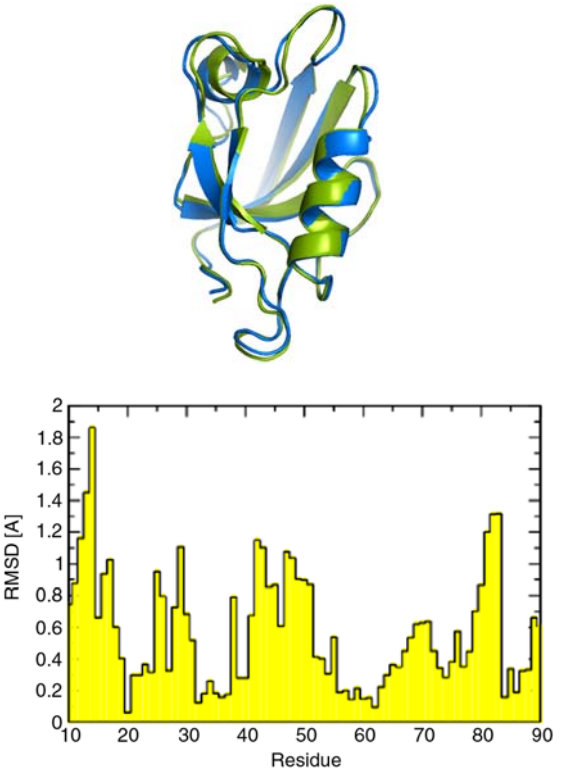

Fig. S5. Comparison of the two most populated microstates of the α and γ valleys of the PDZ2 IS-CFEP. (*Left*) Cartoon representation of the two microstates. (*Right*) Per residue backbone heavy atoms rmsd differences between the two structures.

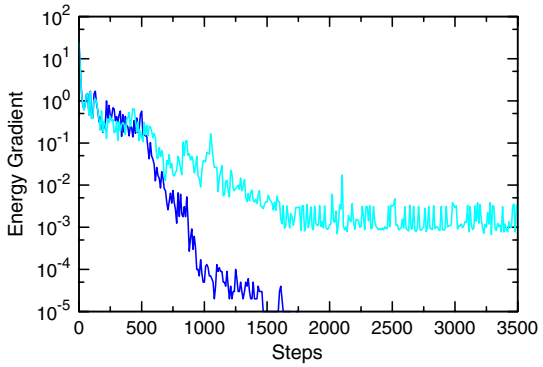

Fig. S6. Energy gradient upon minimization of a peptide snapshot. In this illustration, 100 steps of Steepest Descent are followed by the application of the ABNR algorithm. The light blue and blue lines show the behavior in the presence or absence of the implicit solvation term SASA in the energy function, respectively. The SASA term contains an approximate estimation of the solvent exposed area that is preventing the minimization algorithm to properly converge.

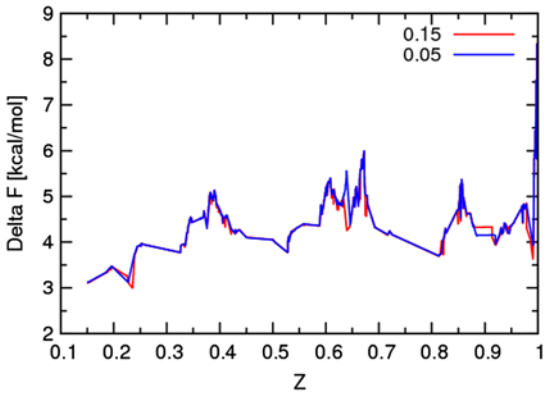

Fig. S7. IS CFEP for two different values of the cutoff to define the IS. The profiles built with 0.15-Å and 0.05-Å cutoffs in the leader algorithm (see main text) are shown in red and blue, respectively. This profile shows that the results are robust with respect to the cutoff in this range.

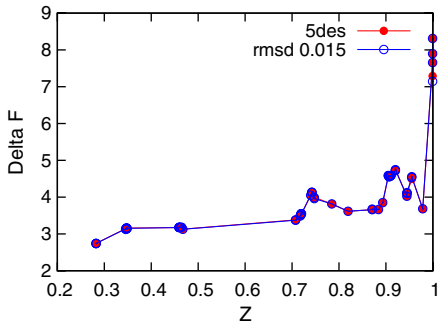

Fig. S8. IS cut-based free-energy profiles when the energy minimizations are carried out without the solvation term (SASA). The results obtained with IS defined using an energy (five decimals) or an rmsd (0.015 Å cutoff) criterion are shown in red and blue, respectively. Identical results are obtained with a 0.15-Å cutoff.

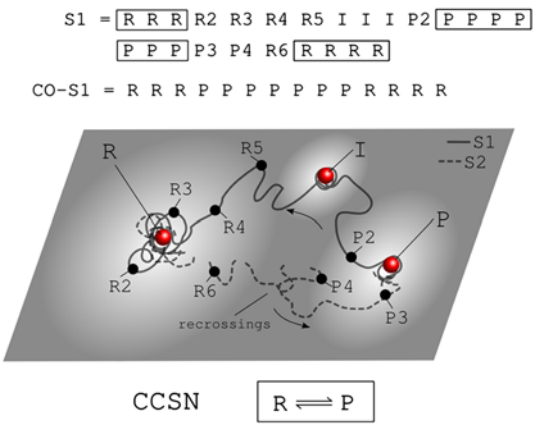

Fig. S9. Meaning of the CCSN. Suppose to have a time series S1 and want to count the number of times the transition between the R and P microstates have been observed. A simple way to do this is to delete from the time series all the microstates except R and P and then build the CSN (CCSN) out of the new coarse trajectory. The deletion results in the creation of a new time series CO-S1. From this new time series the number of times the full transition from R to P (or vice versa) occurs is readily available by counting how many times the event/link R-P (or P-R) is present. In this reduced description, details of the transition, such as recrossings at top of the barrier and intermediate states (I), are omitted. The latter can be introduced into the CCSN by not deleting them from the original time series.

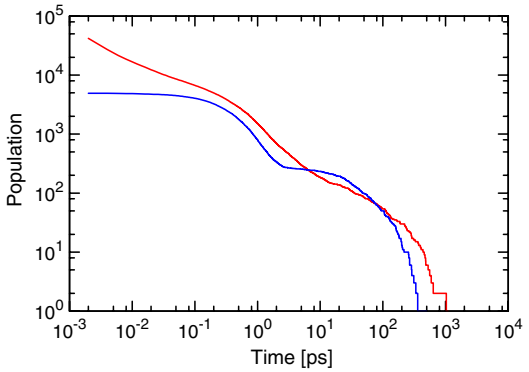

Fig. S10. The first passage time (FPT) distributions to the β_A microstate calculated from the original trajectory and a CCSN model are shown in red and blue, respectively. The double exponential behavior with the correct characteristic times is well reproduced by the CCSN model. The CCSN FPT distribution is calculated from a time series by performing a random walk on the network of 1.5×10^6 steps. The transition matrix for this Markov process is directly obtained from the CCSN network using the link weights presented in Fig. 1*D*. Self transitions are estimated from the CFEP and are taken equal to the population of the corresponding valleys. Each column of the weighted network matrix is normalized to one.