Supporting Information

Energetic and topological determinants of a phosphorylation-induced disorder-to-order protein conformational switch

Soundhararajan Gopi,[a] Nandakumar Rajasekaran,[a] Animesh Singh,[b] Sayan Ranu[b] & Athi N. Naganathan[a]*

[a]Department of Biotechnology, Bhupat & Jyoti Mehta School of Biosciences, Indian Institute of Technology Madras, Chennai 600036, India.

[b]Department of Computer Science and Engineering, Indian Institute of Technology Madras, Chennai 600036, India.

*e-mail: athi@iitm.ac.in

Wako-Saitô-Muñoz-Eaton (WSME) Model Parameters The WSME model[1, 2] is parameterized by solely adjusting the mean-field van der Waals interactions energy (ξ) to achieve a minimum native state population of 80% at the experimental temperature of 293 K. We used this criterion as no thermal or chemical unfolding curves of the WT protein or the variants have been reported. The structure corresponding to the phosphorylated variant 4E-BP2 is used as the reference (PDB id: 2MX4). The rest of the parameters are identical to that obtained from a fit to the heat capacity profile of CspB (see below).[3]

Native heavy-atom contacts are identified employing a vdW contact cut-off of 6 Å excluding nearest neighbors together with an all-to-all electrostatic treatment (i.e. without a distance cut-off). The charges are assigned as per the pH 7.0 protonation state. The phosphate groups are assigned an effective charge of -2. The ionic strength value is fixed to 0.16 M in all calculations following the experiments.[4] The final parameters are: ΔS_{conf} = -11.16 J mol^{-1} K^{-1} per residue for residues identified as helix or strand from the Stride server[5] and ΔS_{conf} = -23.13 J mol^{-1} K^{-1} per residue for residues identified as coil, mean-field vdW interaction energy ξ = -74 J mol^{-1} per native contact and the heat capacity change per native contact ΔC_p^{cont} = -0.36 J mol^{-1} K^{-1}. The convergence temperature of 385 K is used as the reference temperature T_{ref}.[6] The difference between the entropy costs above is equivalent to the differences in the backbone conformational entropy costs between Gly and other amino acids as identified by Freire and co-workers[7] - this approach has also been successful in capturing the varying thermodynamic and kinetic behaviors of SH3 and gpW.[8] The effective dielectric constant (ε_{eff}) in the Debye-Hückel electrostatic term is fixed to 29, the magnitude of which is sufficient to capture differences in stability of single point mutations of charged residues and in accurately reproducing the differences in stability between mesophilic and thermophilic/hyper-thermophilic proteins.[3, 9]

It is important to note that all of the parameters are identical between the doubly phosphorylated and singly phosphorylated variants and the mutant AXXXXAΦ. The calculation of total partition function, residue probabilities, residue-contact

probabilities and 1D free-energy profiles are discussed in great detail in previous publications[3, 9] and in the original work by Wako and Saitô.[1] a

Monte-Carlo (MC) Simulation on the SSA Landscape The Single-Sequence Approximation (SSA) landscape is constructed by algorithmically enumerating the statistical weights of all possible single stretches of folded residues (for the different variants) using parameters identical to the exact-solution calculation discussed above. The free-energies of the resulting 1035 microstates (N*(N+1)/2 microstates, where N=45 here) is projected onto two variables - the starting residue (m) and the number of structured residues (n) –resulting in the SSA landscape. This approach has been very adept in identifying partially structured states or intermediates in agreement with experiments in multiple proteins – bovine lactalbumin,[3] IκBα,[10] RNase H,[11] Barstar,[12] bACBP.[13]

The connectivity between different microstates in a SSA landscape is illustrated below employing a five-residue sequence. The fully folded state is 11111 (top row) while microstates with a single structured residue are in the bottom row, and the rest are partially structured (different combinations of 0s and 1s, but all occurring within a single unbroken stretch). In the MC simulation, we start from an unfolded-like microstate and choose another microstate randomly (but following the indicated connectivity) and then evaluate the possibility of transition using the Metropolis criterion. It is important to note that each state is connected to its possible neighbor only through a single bond-flip (i.e., 1 ➔ 0 or 0 ➔1).

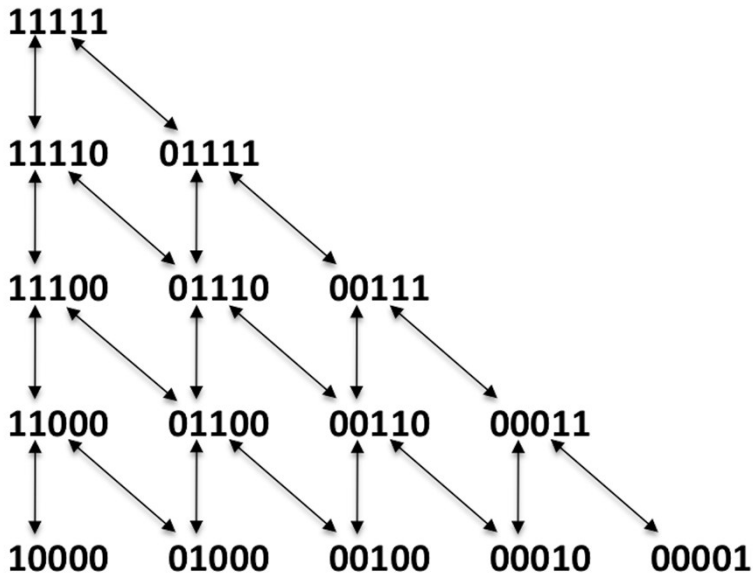

Figure S1. Connectivity in an SSA landscape illustrated with a 5-residue sequence.

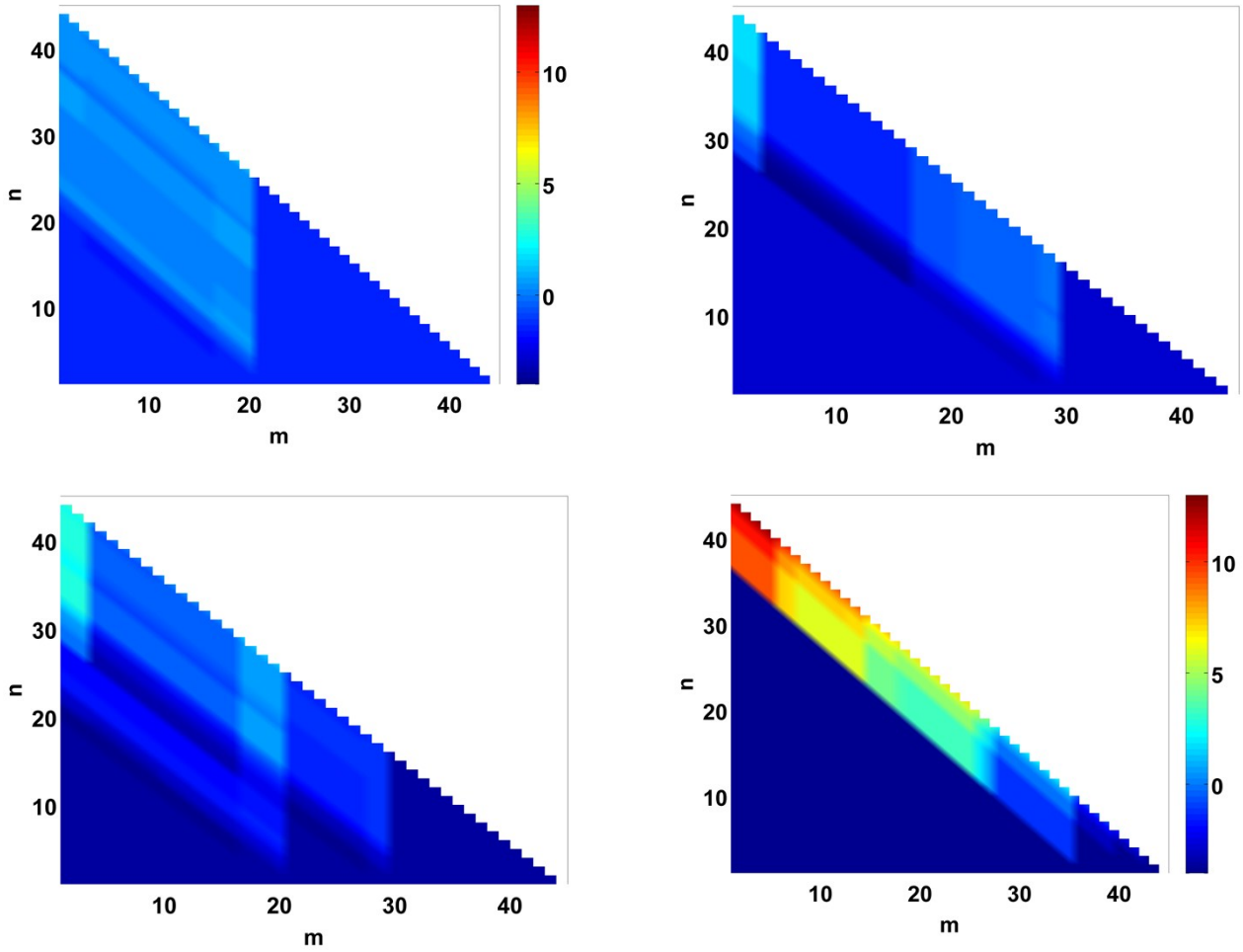

Figure S2. Difference SSA free energy (FE) landscapes of different variants of 4E-BP2 at 293 K with respect to the fully phosphorylated state, i.e. FE (variant) – FE (WT): pT29 (top left), pT20 (top right), pT20pT29 (bottom left) and the doubly phosphorylated mutant AXXXXAΦ (bottom right). The color-coding range (color bar in kJ/mol) is identical between all the variants. The depletion of the native state population and the increased population of the partially structured states (in terms of free energy) can be clearly seen by comparison.

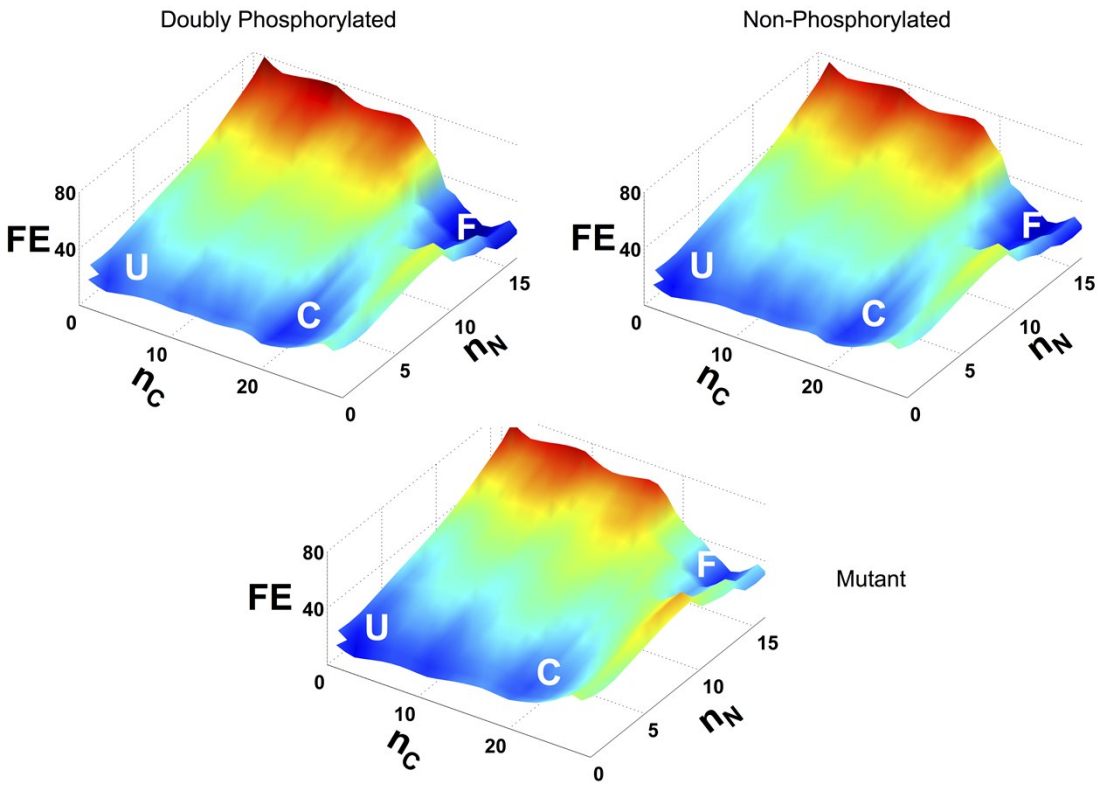

Figure S3. Free energy landscapes constructed by accumulating the statistical weights of 9,531,031 microstates at 293 K (by combining single-, double- and triple-sequence approximations of the WSME model). They are projected on to two coordinates, the number of C-terminal structured residues (n_C; residues 18-45) and the number of N-terminal structured residues (n_N; residues 1-17). The partially structured state C, the unfolded state U and the fully folded state F are highlighted. It is clear from the representations that the N-terminal part of the structure (that includes strand 1 and the long loop) folds the last in agreement with the SSA landscape (Figure 2a of the main text). The landscape can also be seen to progressively tilt towards the states U and C in going from the phosphorylated variant to the mutant. However, microstate connectivity is lost in this representation.

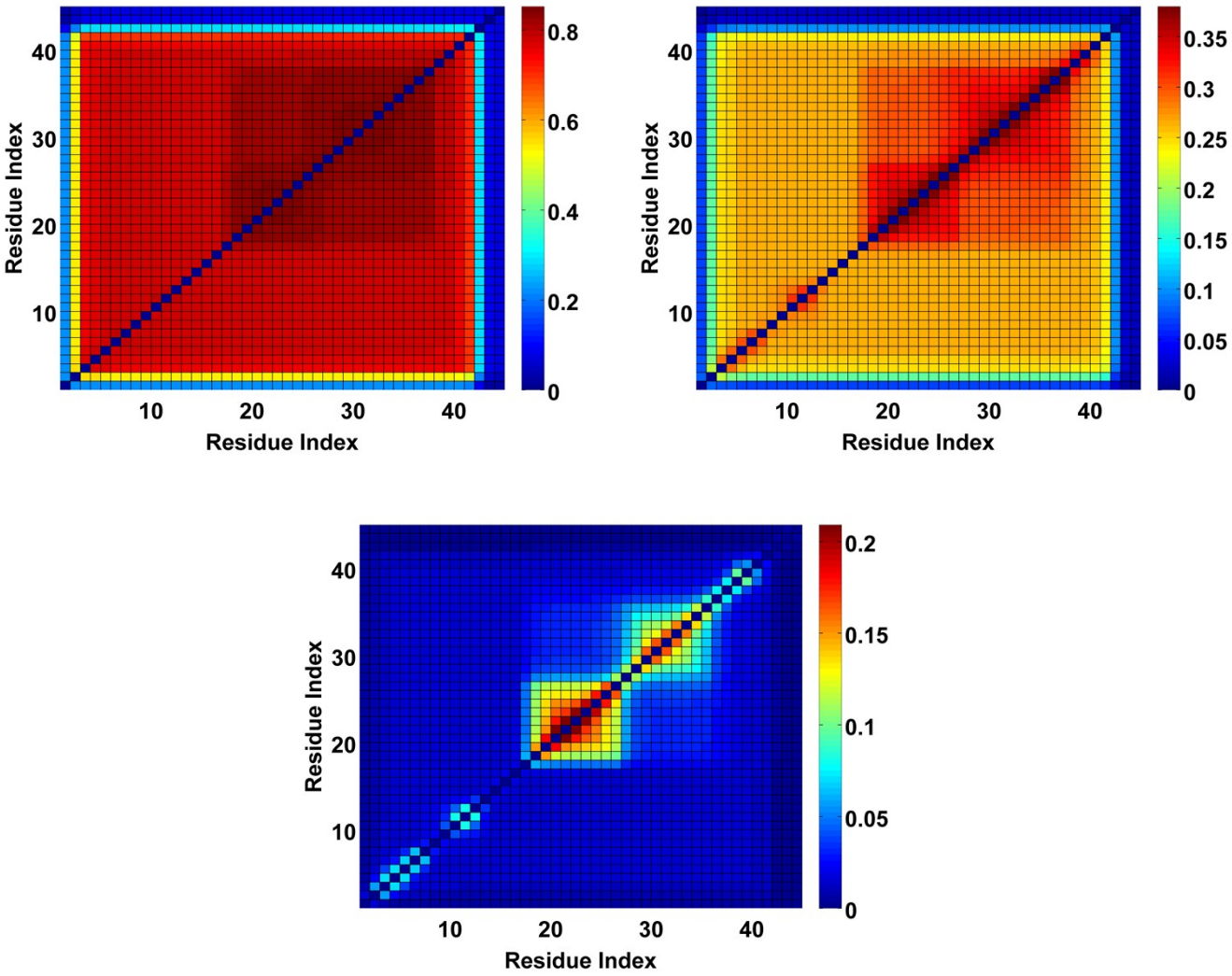

Figure S4. The probability of interaction between any two residues at 293 K for the different variants of 4E-BP2: (Top Left) Doubly phosphorylated, (Top Right) Non-phosphorylated and (Bottom) Mutant AXXXXAΦ. Note that the range of the color-bar is different for the different panels. The location of the different secondary structure elements is as follows: 2-7 (Strand 1), 18-20 (Strand 2), 24-28 (Strand 3) and 34-39 (Strand 4). In the doubly phosphorylated variant all inter-residue interactions have high probability as the protein is ~80% folded. For the non-phosphorylated variant, it is clear that there is a large probability of interaction between residues 18-39 (State C in Figure 2 of the main text). In the mutant AXXXXAΦ, the local difference in stabilities are even more vivid with the residues 18-28 being the most folded and hence corresponding to the State A; the next most populated state is one with residues 27-37 corresponding to State B. These calculations are in good agreement with the nature of partially structured states generated from the SSA landscape.

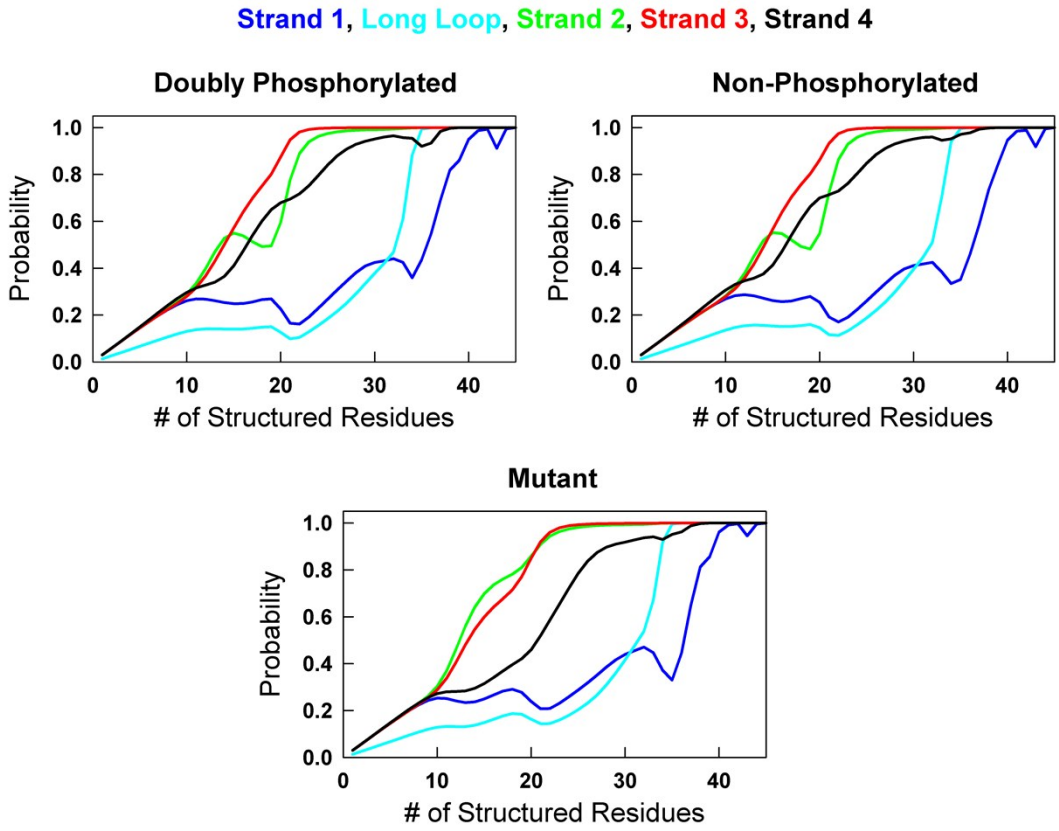

Figure S5. The average probability of forming specific secondary structure elements including the long loop at near the melting temperature as a function of the reaction coordinate, the number of structured residues. In all cases, it can be seen that the strand 1 forms the last just after the condensation of the long loop. The doubly-phosphorylated and non-phosphorylated variants follow a similar folding mechanism with an early formation of strand 3, followed by strand 2 and then strand 4 before the main folding barrier located at ~30 structured residues. However, the mutant exhibits a very different folding mechanism during the early stages of folding – the strand 3 forms earlier than strand 2, and strand 4 forms even later than the variants resulting in the conformational heterogeneity highlighted in Figure 2f of the main text.

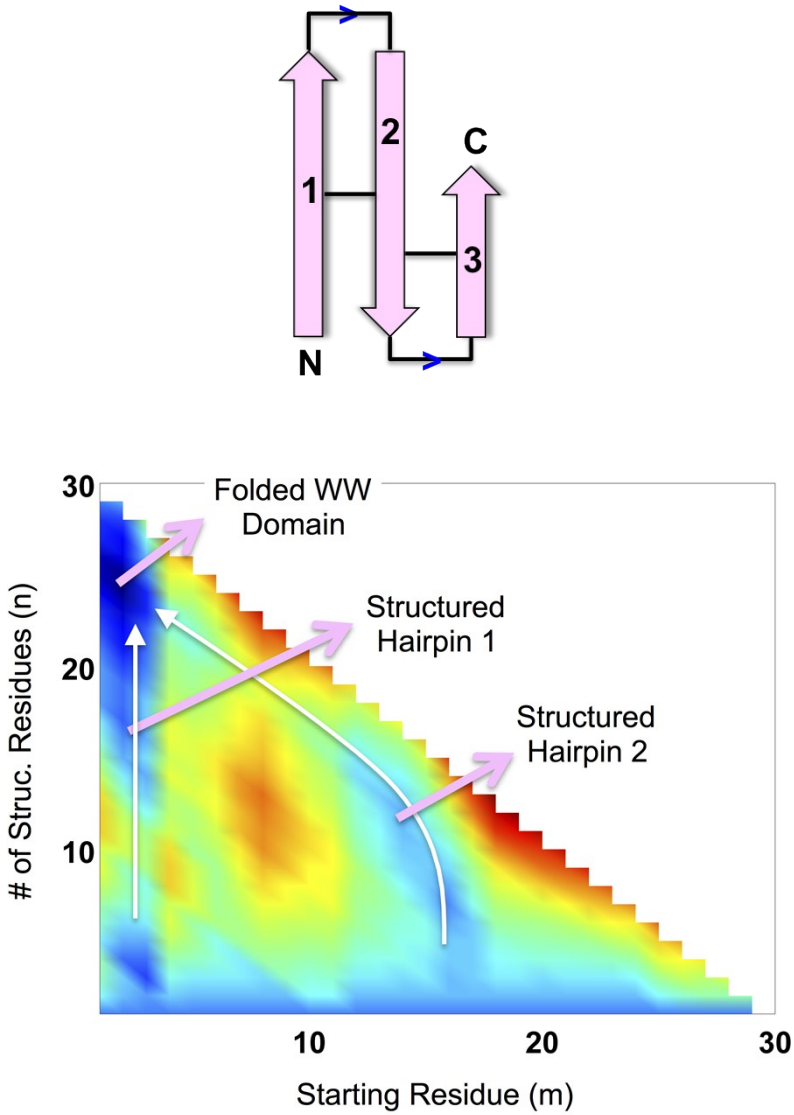

Figure S6. Folding mechanism of Pin WW domain (topology on top) as inferred from the SSA landscape (bottom panel). Spectral color-coding is employed going from low free energy (blue) to high (red). The landscape clearly indicates two paths that involve the formation of either the first or the second hairpin (white arrows) in agreement with experiments and simulations.[14, 15] This agreement further validates the SSA landscape approach for identifying intermediates in simple systems. 4E-BP2, which is built on a WW-domain-like scaffold, displays similar complexity during early stages of folding (Figure 2a of the main text).

Supplementary References

1. H. Wako and N. Saito, *J. Phys. Soc. Japan*, 1978, 44, 1939.
2. V. Muñoz and W. A. Eaton, *Proc. Natl. Acad. Sci. U.S.A.*, 1999, 96, 11311.
3. A. N. Naganathan, *J. Chem. Theory Comput.*, 2012, 8, 4646.
4. A. Bah, R. M. Vernon, Z. Siddiqui, M. Krzeminski, R. Muhandiram, C. Zhao, N. Sonenberg, L. E. Kay and J. D. Forman-Kay, *Nature*, 2015, 519, 106.
5. M. Heinig and D. Frishman, *Nuc. Acids Res.*, 2004, 32, W500.
6. A. D. Robertson and K. P. Murphy, *Chem. Rev.*, 1997, 97, 1251.
7. J. A. Daquino, J. Gomez, V. J. Hilser, K. H. Lee, L. M. Amzel and E. Freire, *Proteins*, 1996, 25, 143.
8. P. Bruscolini and A. N. Naganathan, *J. Am. Chem. Soc.*, 2011, 133, 5372.
9. A. N. Naganathan, *J. Phys. Chem. B*, 2013, 117, 4956.
10. S. Sivanandan and A. N. Naganathan, *PLOS Comp. Biol.*, 2013, 9, e1003403.
11. A. Narayan and A. N. Naganathan, *J. Phys. Chem. B*, 2014, 118, 5050.
12. A. N. Naganathan, J. M. Sanchez-Ruiz, S. Munshi and S. Suresh, *J. Phys. Chem. B*, 2015, 119, 1323.
13. S. Munshi and A. N. Naganathan, *Phys. Chem. Chem. Phys.*, 2015, 17, 11042.
14. K. Lindorff-Larsen, S. Piana, R. O. Dror and D. E. Shaw, *Science*, 2011, 334, 517.
15. A. J. Wirth, Y. Liu, M. B. Prigozhin, K. Schulten and M. Gruebele, *J. Am. Chem. Soc.*, 2015, 137, 7152.