Supporting Information

Structural characterization of the early events in the nucleation-condensation mechanism in a protein folding process

Predrag Kukic[1], Yulia Pustovalova[2], Carlo Camilloni[1], Stefano Gianni[3], Dmitry M. Korzhnev[2] and Michele Vendruscolo[1,*]

[1]Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, UK
[2]Department of Molecular Biology and Biophysics, University of Connecticut Health Center, Farmington, CT 06030, USA
[3]Istituto Pasteur - Fondazione Cenci Bolognetti and Istituto di Biologia e Patologia Molecolari del CNR, Dipartimento di Scienze Biochimiche "A. Rossi Fanelli", Sapienza Università di Roma, 00185, Rome, Italy
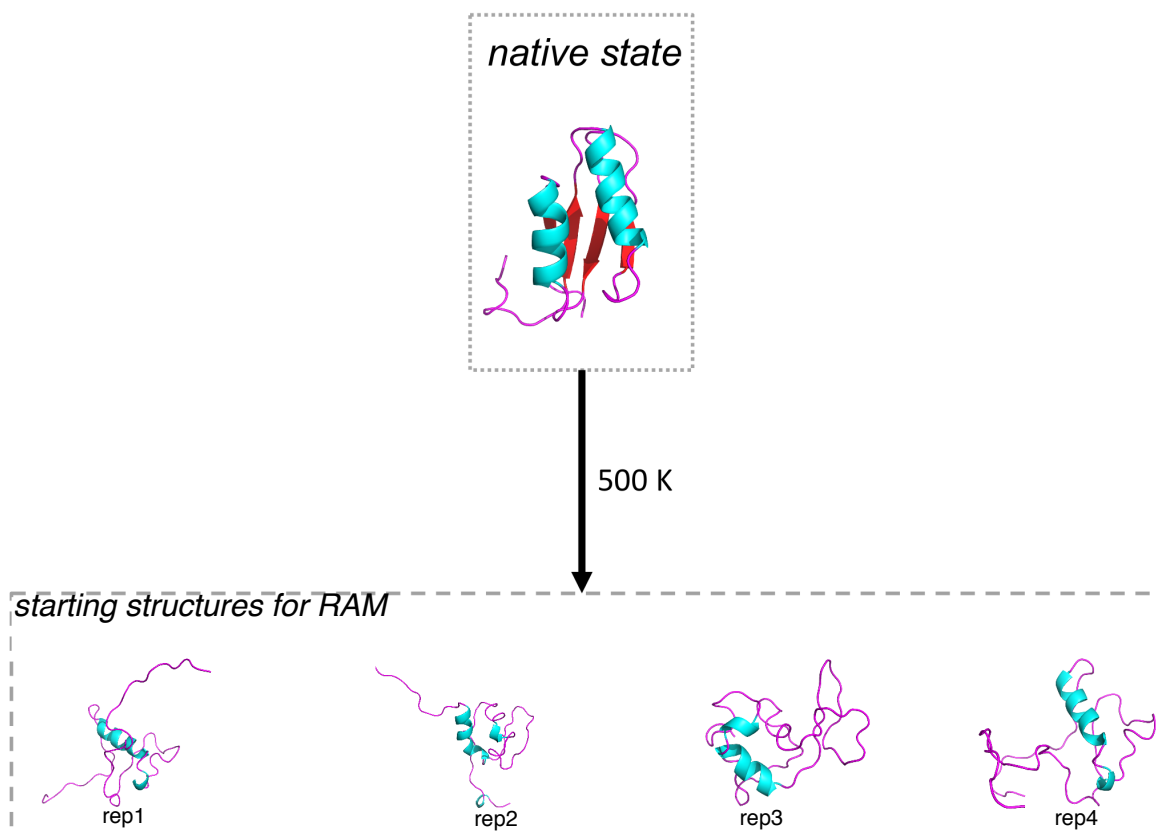
Figure S1. Procedure of selection of the four starting conformations for the RAM simulations of D_{phys}. The starting conformations were obtained by partially unfolding the native state structure of ADA2h I71V. The structures of the converged D_{phys} ensemble showed on average a low topological similarity to the starting conformations (*TM_factor* < 0.4).
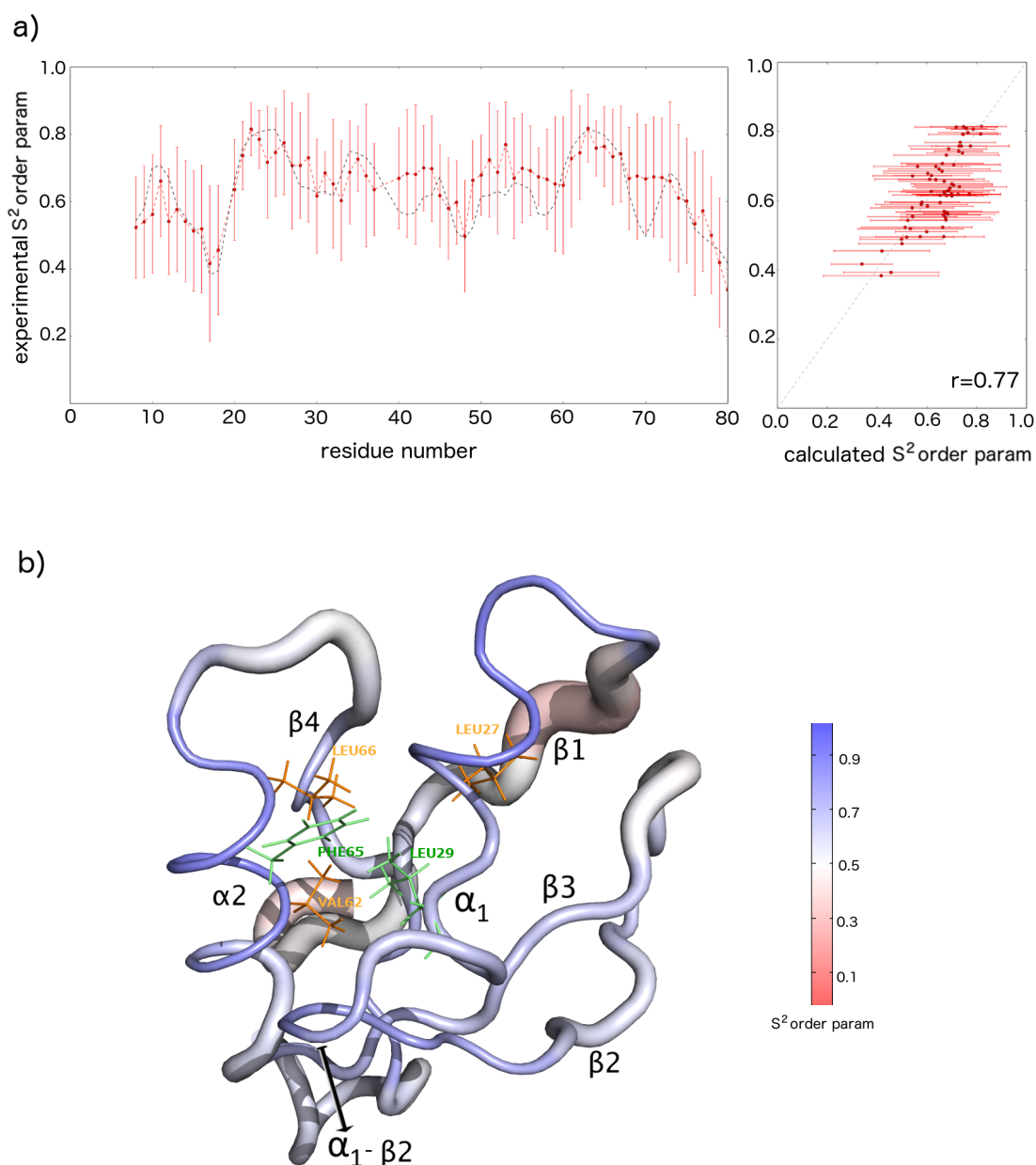
Figure S2. Analysis of the S² order parameters in the D_{phys} ensemble. (a) Correlation between S^2 order parameters calculated in Ref. [1] from chemical shifts using the RCI approach and back-calculated values from the D_{phys} ensemble (b) Structure of ADA2h I71V color-coded according to the backbone amide S^2 order parameters with ribbon width proportional to $(1 − S^2)$. The regions with restricted mobility include α-helix 1, α-helix 2 and loop between α-helix 1 and β-strand 2. In contrast, all four β-strands show increased dynamics.

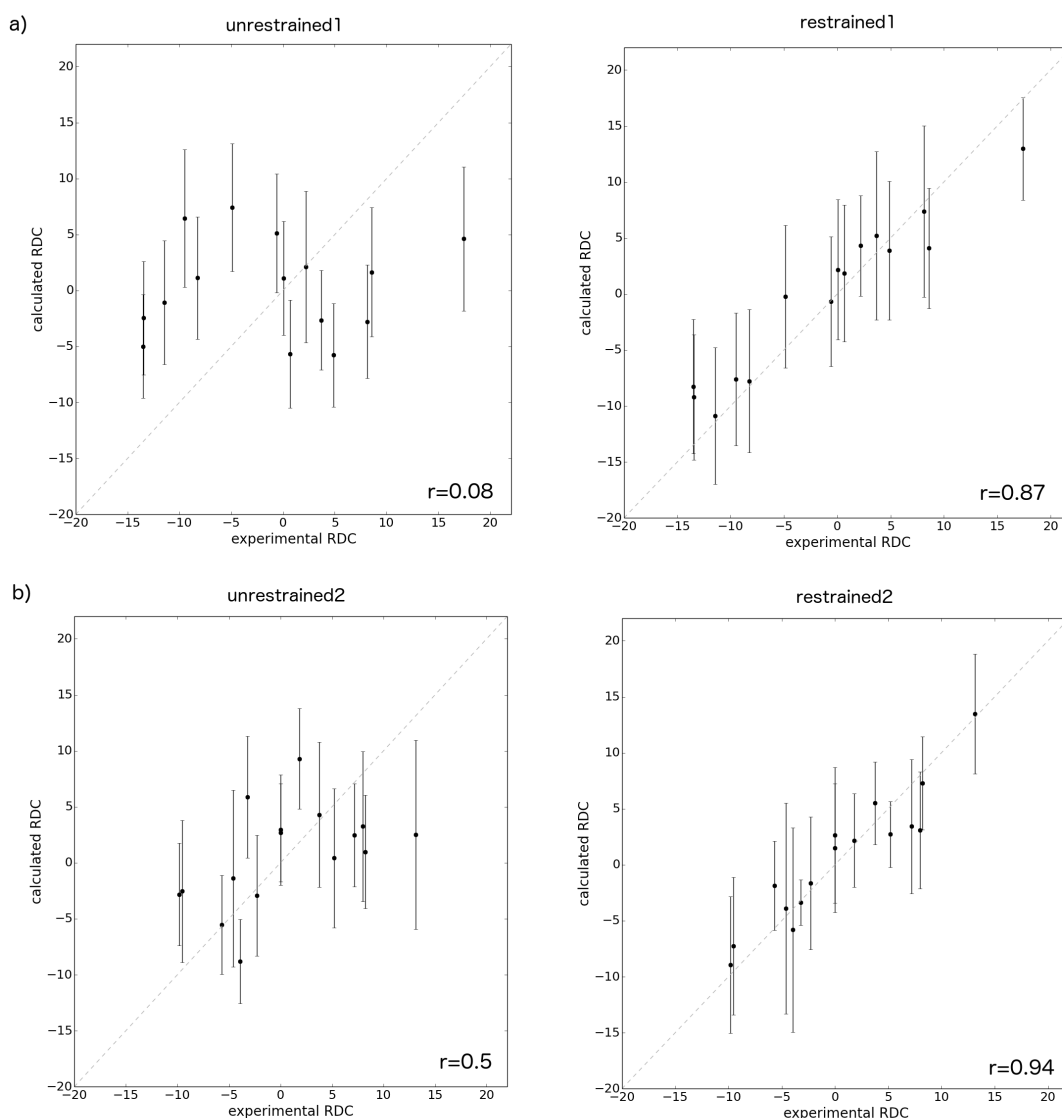**Figure S3. Validation of the D_{phys} ensemble using the correlation between experimental and back-calculated RDC values.** Structural ensembles were generated by using RAM simulations with the following restraints: (a) NMR chemical shifts and a first subset of 16 RDCs (right plot); (b) NMR chemical shifts and a second subset of 15 RDCs (right plot). Validations were carried out using those RDCs not used in the RAM simulations. Plots on the left side are obtained from an unrestrained RAM simulation, and they are used as controls.
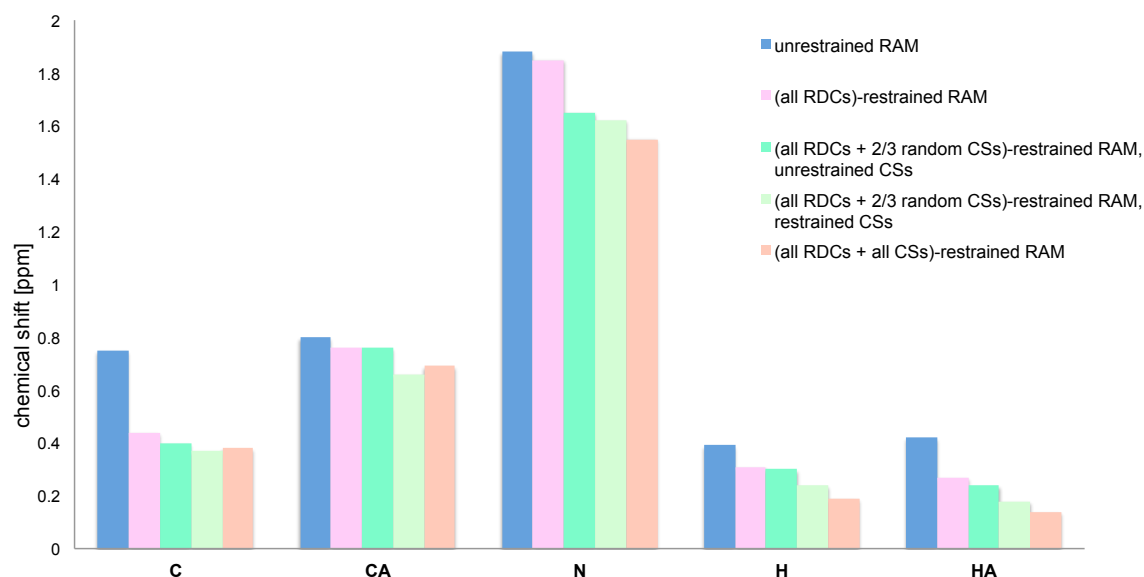
Figure S4. Validation of the D_{phys} ensemble using the RMSD between experimental and back-calculated chemical shifts. Structural ensembles were generated by using unrestrained simulations (blue) and RAM simulations with the following restraints: the complete set of the measured RDCs (pink); the complete set of RDCs and two thirds of the randomly selected NMR chemical shifts (greens); the complete set of RDCs and the complete set of NMR chemical shifts (brown).
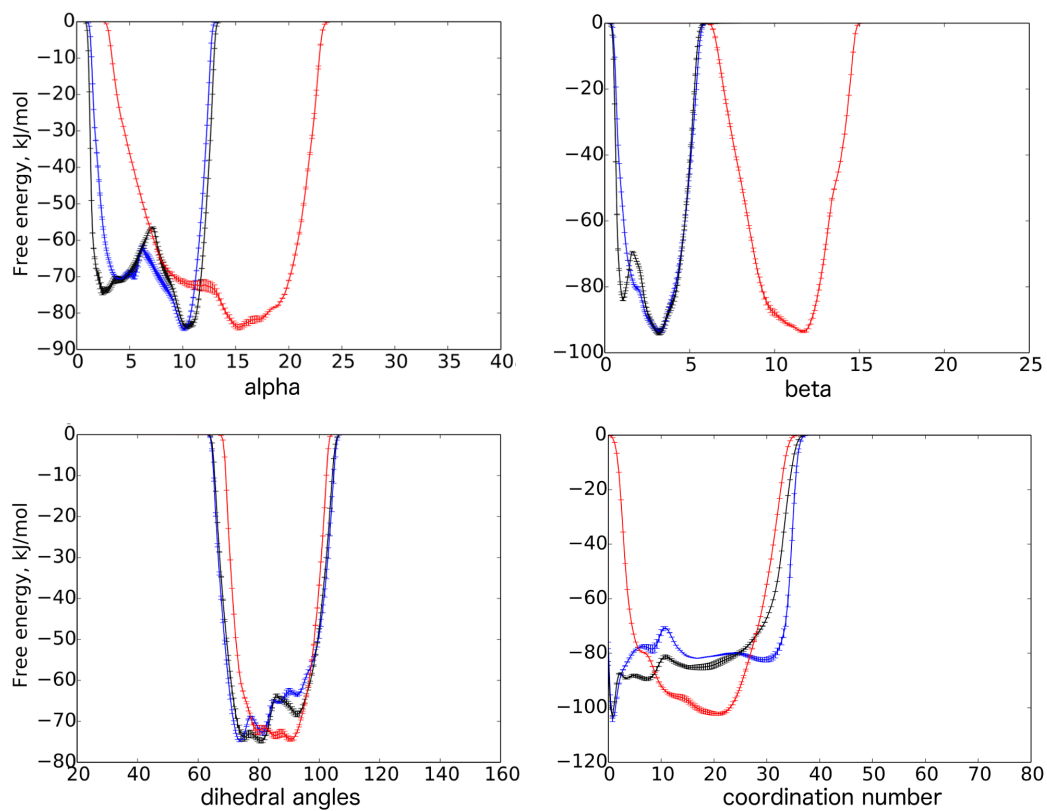
Figure S5. Comparison of the free energy profiles of the unrestrained (TIP3P water, red), restrained (TIP3P water, blue), and restrained (TIP4P/2005 water, black) ensembles of ADA2h I71V. The free energy profiles are calculated along the four CVs (Materials and Methods). They are averaged after convergence over the last 80 ns of the RAM simulations and their standard deviations are reported as error bars.
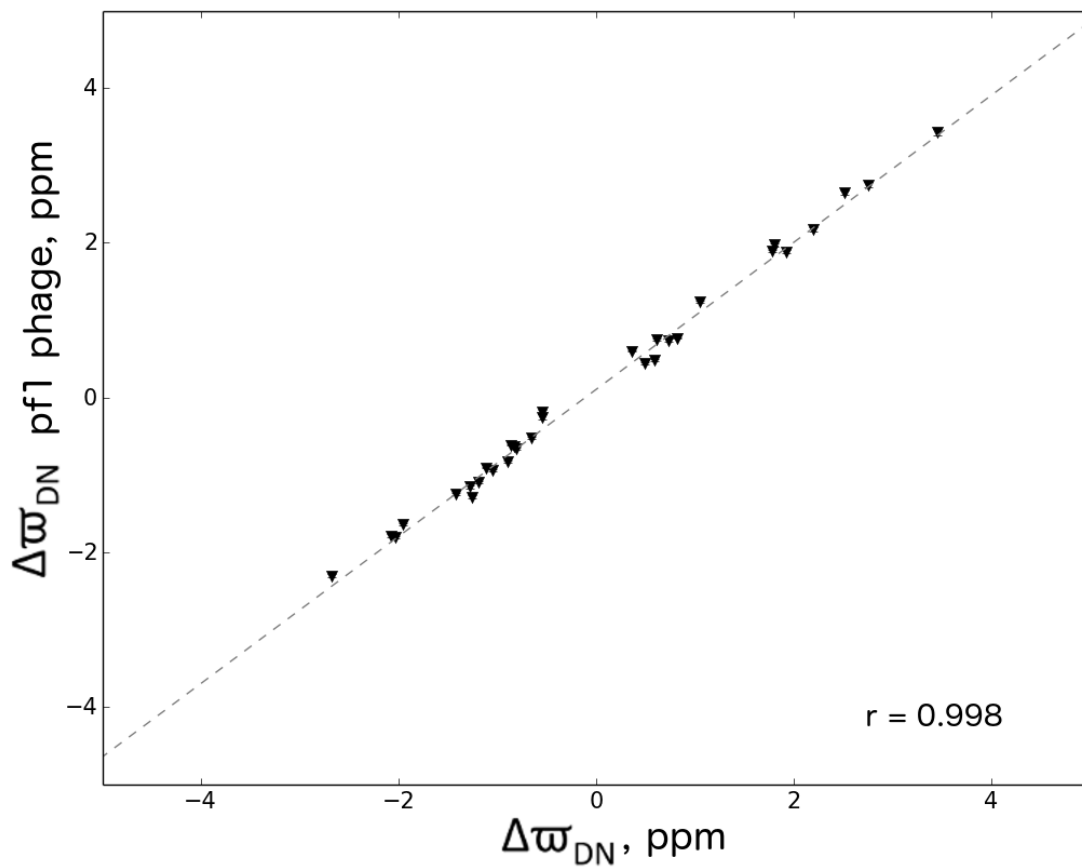
Figure S6. **Correlation of ^{15}N chemical shift differences, $\Delta\varpi_{DN}$, between the D_{phys} and N states in the presence and absence of weakly aligned Pf1 phages.** Since the chemical shift differences, $\Delta\varpi_{DN}= \Delta\varpi_D\text{-}\Delta\varpi_N$, are obtained from relaxation dispersion measurements they are independent from the D_{phys} population. Overall, the correlation is high, as indicated by a Pearson's coefficient of correlation of r=0.998.
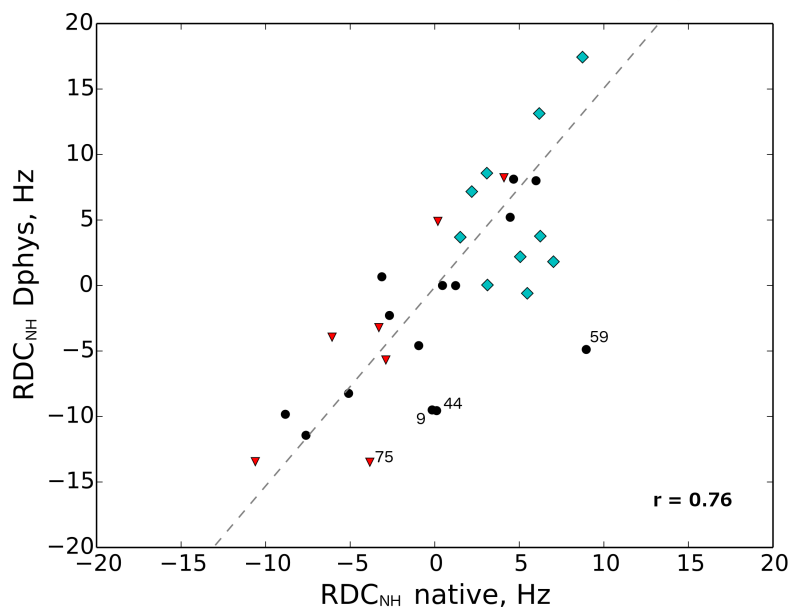
Figure S7. Correlation between ${}^{1}H^{N}$-${}^{15}N$ RDCs of the _N_ and _D${}_{phys}$_ states of I71V ADA2h. The values measured for the _N_ state are plotted along the x-axis, whereas the values determined for the _D${}_{phys}$_ state are plotted along the y-axis. Overall, a high correlation was obtained for all secondary structure elements (with a Pearson's coefficient of correlation, r=0.76). The highest correlation was obtained for the β-sheet region (red, r=0.84), while lower correlations were found for the loop region (black, r=0.64) and the α-helical region (blue, r=0.4). The outliers belong to the disordered N-terminal region (Gly9), β2 strand (Thr44), α2 helix (Val59) and β4 strand (Ile75). Note that all β-sheet points apart from the outlier Ile75 belong to β1 and β3 strands.
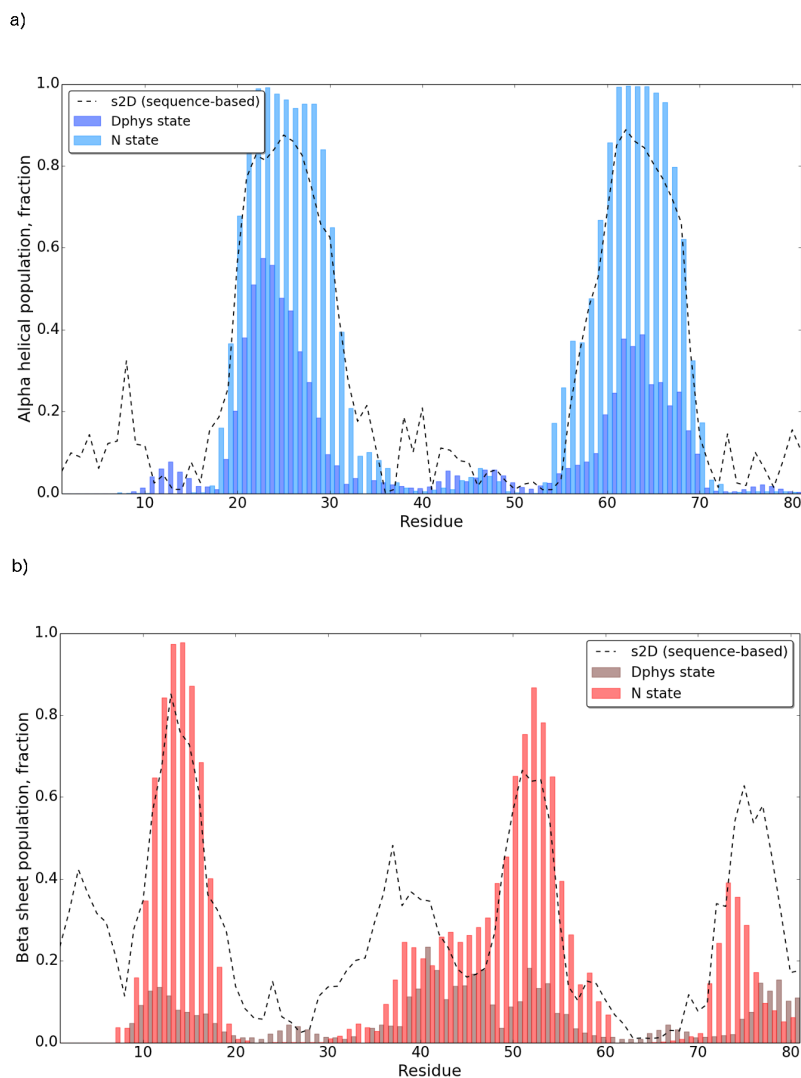
a)

b)

Figure S8. Correlations between secondary structure populations in the D_{phys} and N states and sequence-based secondary structure propensities of ADA2h to form (a) α-helix (b) β-sheet. Secondary structure propensities from the amino acid sequence of ADA2h are calculated using the s2D method[2] and represented by black dotted lines. Secondary structure populations in the D_{phys} and N states are determined from the CPMG-derived backbone ^{15}N, $^{1}H^{N}$, $^{13}C^{α}$, $^{1}H^{α}$, $^{13}C'$ chemical shifts using the δ2D software (bars)[3]. These results indicate that the α-helical secondary structure elements in the D_{phys} state match closely the intrinsic (i.e. sequence-based) propensities of ADA2h, as well as the populations in the N state. By contrast, the β-structure elements in the D_{phys} state are less populated and poorly correlated with those in the N state and with the intrinsic β-structure propensities of this protein.
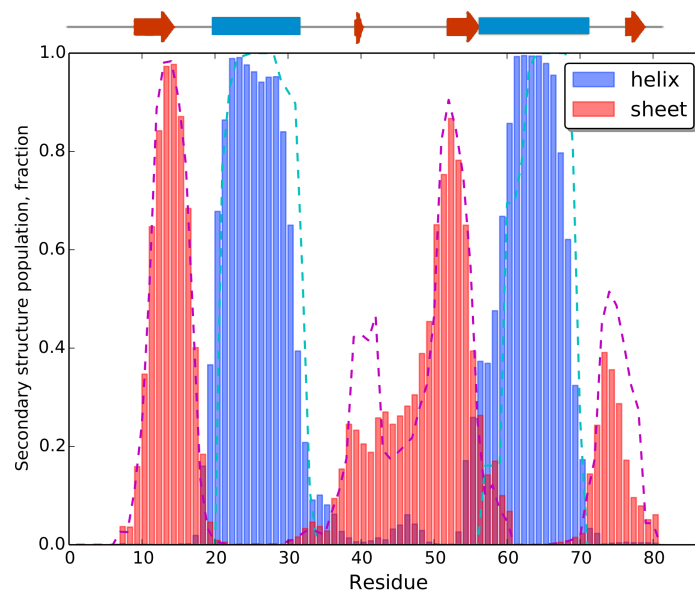
Figure S9. Agreement between chemical shift- and ensemble-derived secondary structure populations in the *N* state. Secondary structure populations of the *N* state of ADA2h I71V obtained from the CPMG-derived backbone ^{15}N, ^{1}H^{N}, ^{13}C$^{\alpha}$, ^{1}H$^{\alpha}$, ^{13}C' chemical shifts using the δ2D software (bars)[3]. Secondary structure population back-calculated from the *N* ensemble is depicted by the dotted blue and red lines.
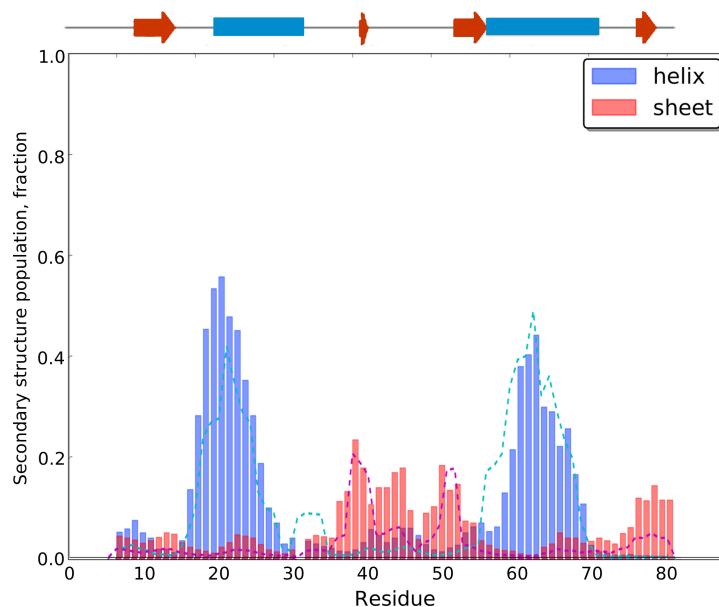
Figure S10. Agreement between chemical shift- and ensemble-derived secondary structure populations in the D_{phys} state. Secondary structure populations of the D_{phys} state of ADA2h I71V obtained from the CPMG-derived backbone ^{15}N, $^{1}H^{N}$, $^{13}C^{\alpha}$, $^{1}H^{\alpha}$, $^{13}C'$ chemical shifts using the δ2D software (bars)[3]. Secondary structure population back-calculated from the D_{phys} ensemble is depicted by the dotted blue and red lines.
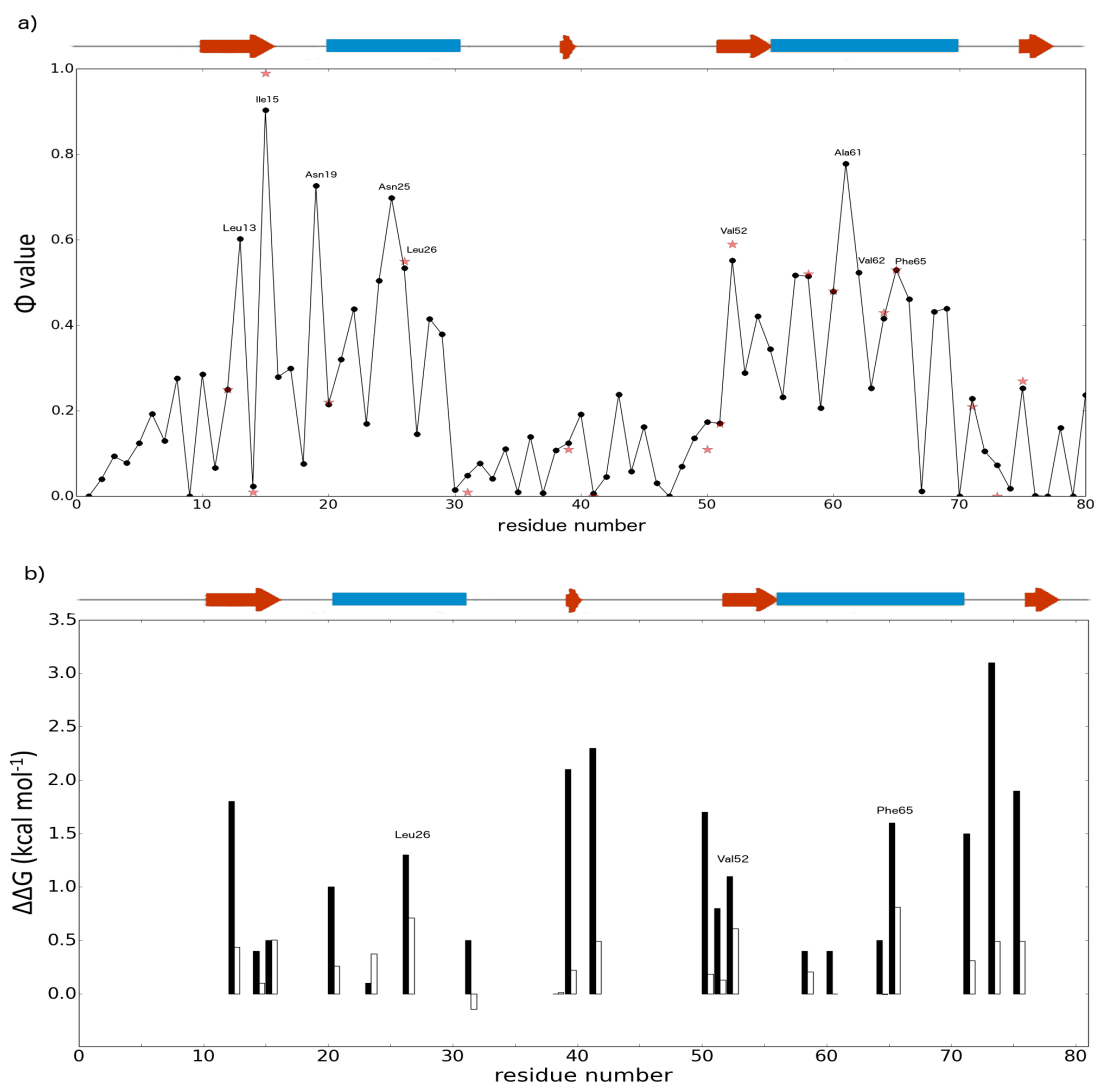
Figure S11. Analysis of residue-specific thermodynamic properties of the TS ensemble. (a) Back-calculated Φ_i-values from the TS ensemble; values denoted by red stars were experimentally measured and used as structural restraints in this work. Residues with high Φ values are located in α-helices α1 and α2, and β-strand β1 and β3. Residues with high Φ values selected as candidates for the folding nucleus are labeled. (b) Comparison of the $\Delta\Delta G_{TS-Dphys}$ (black bars) and $\Delta\Delta G_{N-Dphys}$ (open bars) values for the various mutants. Three residues with the highest folding free energy barriers are Leu26, Val52 and Phe65.
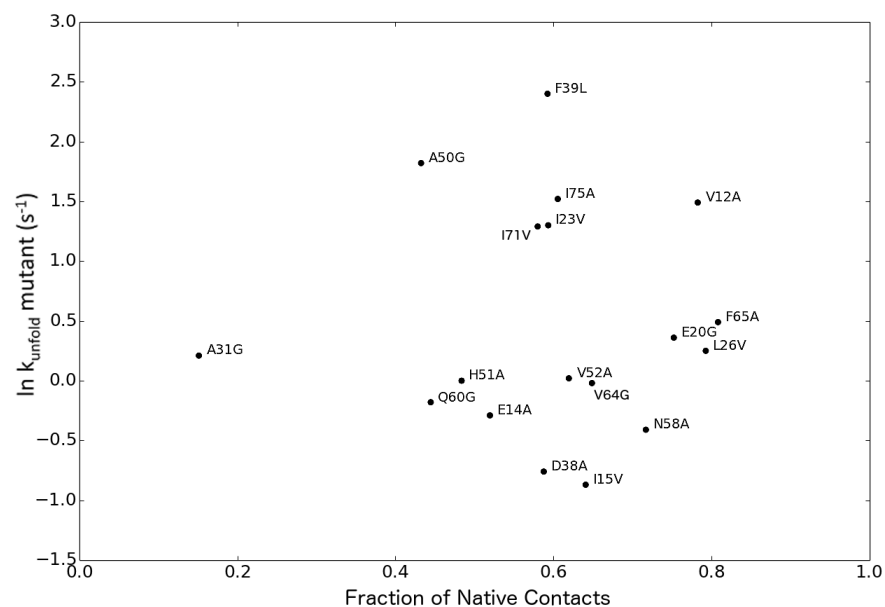
Figure S12. Analysis of the correlation between the fraction of native contacts in *D_{phys}* **and unfolding rates of ADA2h mutants.** Our results indicate a lack of correlation between the fraction of native contacts in *D_{phys}* of the mutated residues and the corresponding unfolding rates, (the coefficient of correlation, 0.05).
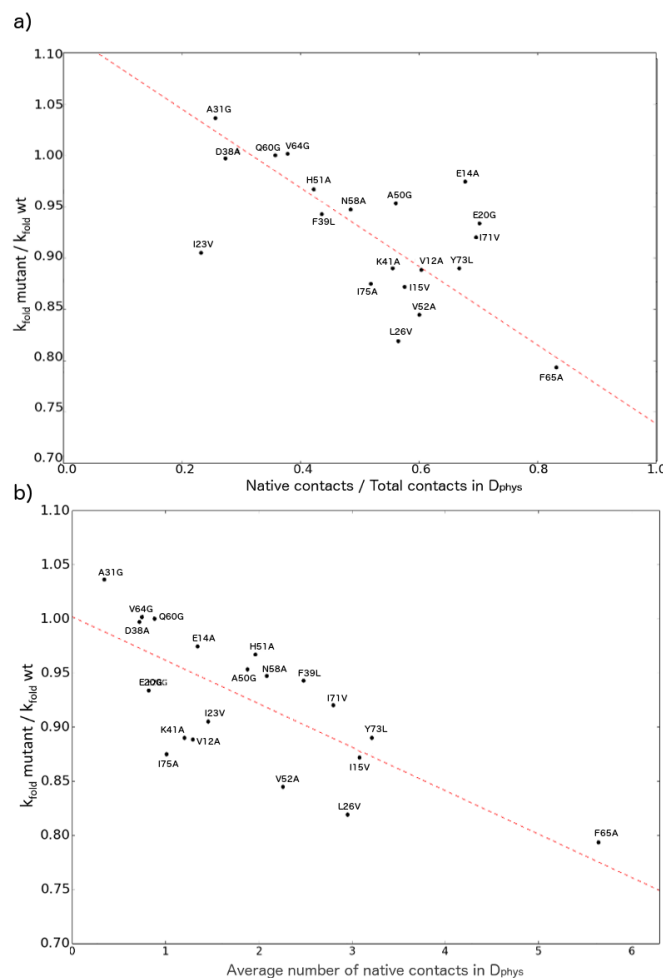
Figure S13. Influence on the folding rate of transiently-formed contacts in the D_{phys} ensemble produced with the TIP4P/2005 water model. We found a statistically significant correlation between the fraction of native contacts in D_{phys} and the folding rate of ADA2h mutants. (a) When we considered the fraction of native contacts relative to the total number of contacts, the Pearson's coefficient of correlation was 0.63. (b) When we considered the average number of native contacts, the coefficient of correlation was 0.72.

Supplementary References

(1)   Pustovalova, Y.; Kukic, P.; Vendruscolo, M.; Korzhnev, D. M. *Biochemistry* 2015, *54*, 4611.

(2)   Sormanni, P.; Camilloni, C.; Fariselli, P.; Vendruscolo, M. *J. Mol. Biol.* 2015, *427*, 982.

(3)   Camilloni, C.; De Simone, A.; Vranken, W. F.; Vendruscolo, M. *Biochemistry* 2012, *51*, 2224.