Electronic Supplementary Material (ESI) for Physical Chemistry Chemical Physics.

Supplementary Information
Globular-Disorder Transition in Proteins: A Compromise between Hydrophobic and Electrostatic Interactions?

Anupaul Baruah[a] and Parbati Biswas*[a]

[a] *Address, University of Delhi, Delhi, India. Tel: +911127666646; E-mail: pbiswas@chemistry.du.ac.in*

The *NCαCO* Model

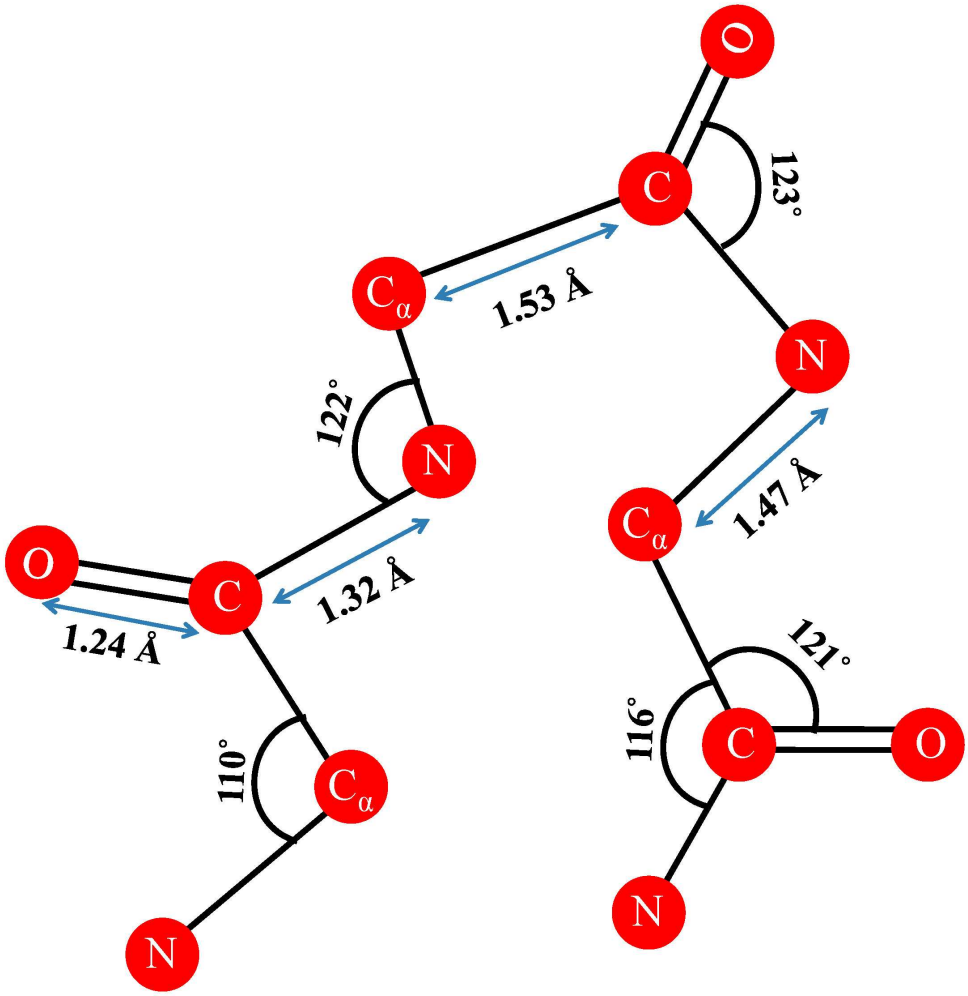

FIG. S1. Schematic representation of the *NCαCO* model used in the manuscript. The *mean angles* and the *mean bond lengths* are shown. The angles are restricted to *mean angle* $\pm 20^{\circ}$ while the bond lengths are restricted to *mean bond length* ± 0.1 Å.

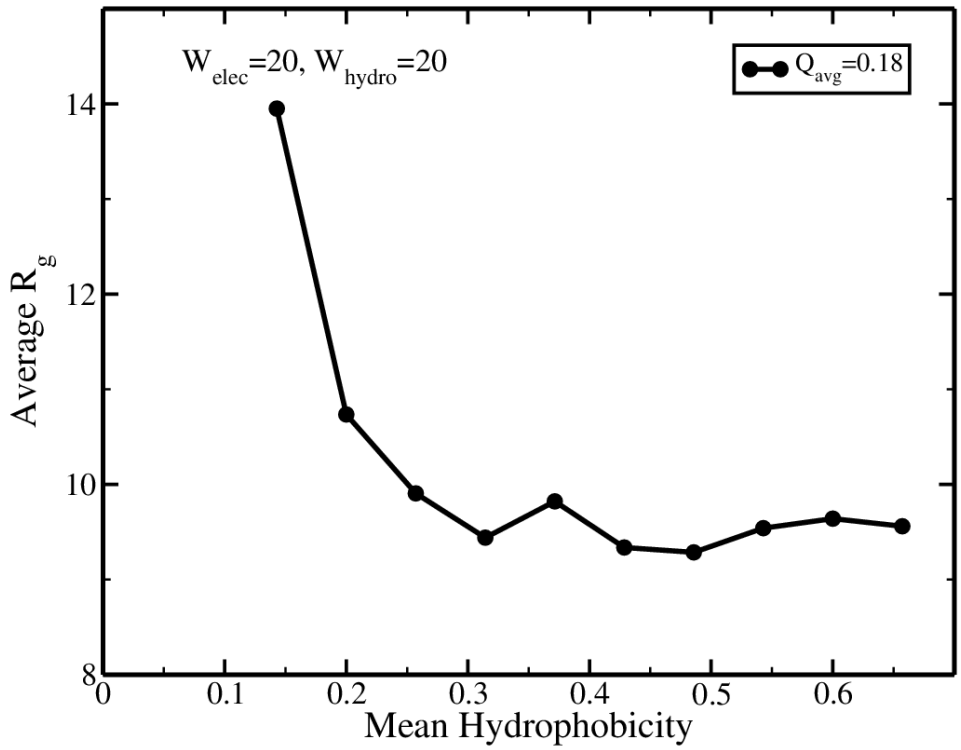

FIG. S2. Average R_g plotted against H_{avg} for $W_{hydro} = 20$ and $W_{elec} = 20$. The average R_g shows an increase at the point where $W_{elec} \times Q_{avg} > W_{hydro} \times H_{avg}$. This implies at the charge hydrophobicity boundary between ordered and disordered region a transtion between the dominant interactions occur. In the disordered region the electrostatic interactions dominate over hydrophobic interactions.

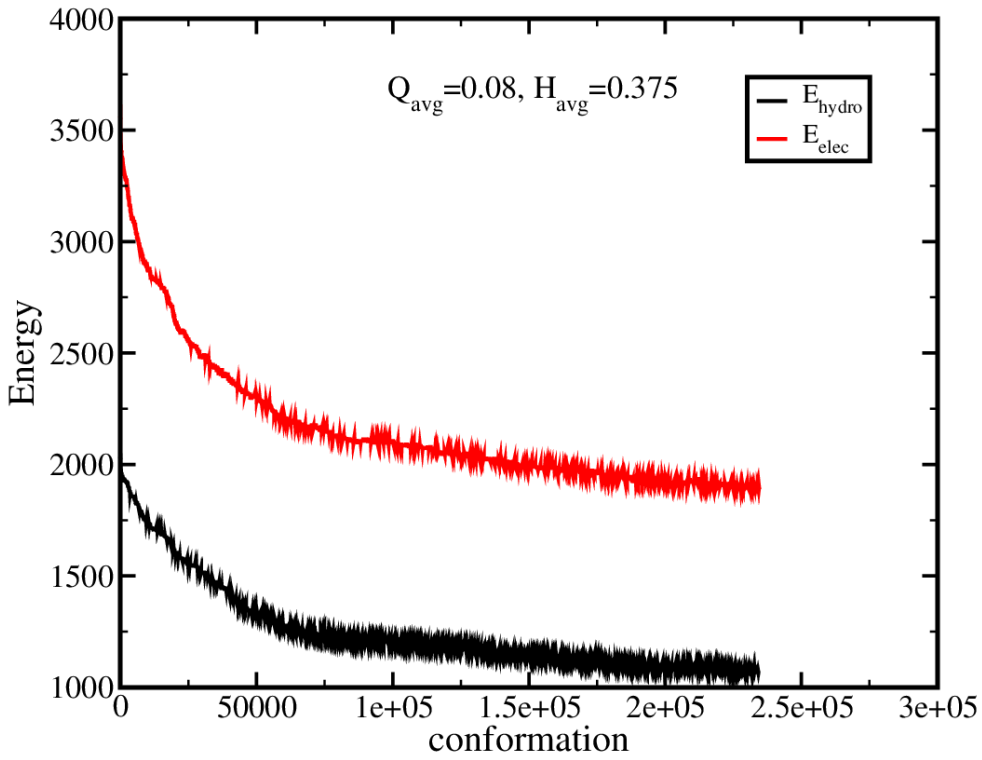

FIG. S3. The electrostatic and hydrophobic energy contributions are plotted against the MC steps. In this sequence the electrostatic interactions are dominant and hence electrostatic interaction are being satisfied (energy lowered to larger extent compared to the hydrophobic ineteraction) to achieve the minimally frustrated state (a dynamic ensemble of expanded conformations).

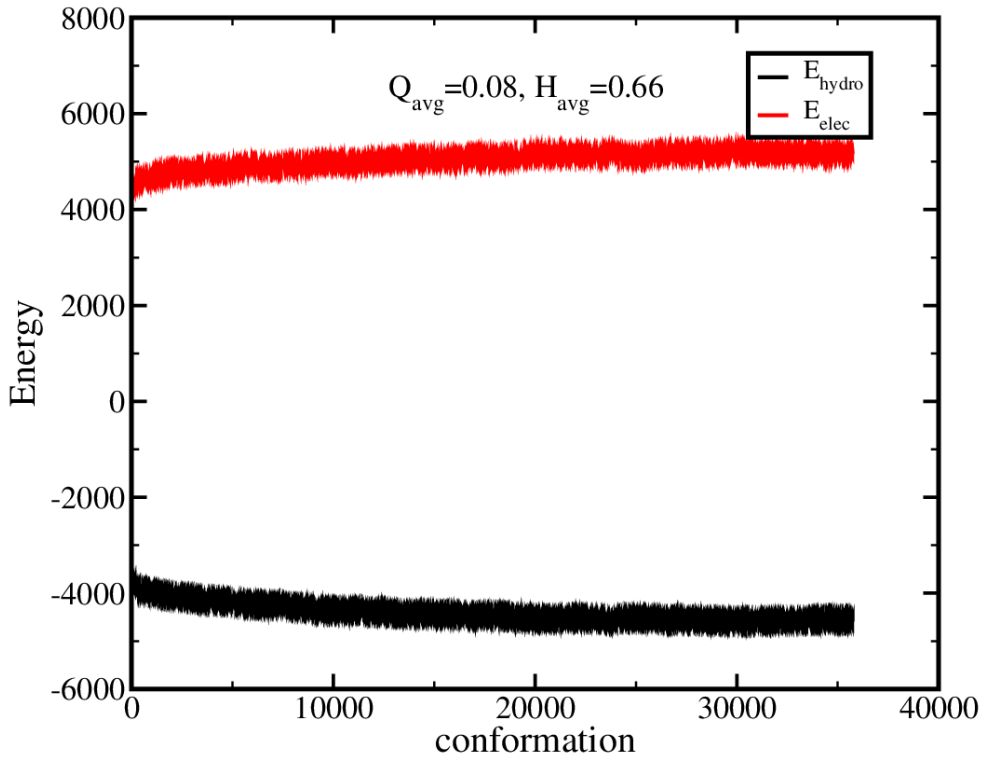

FIG. S4. The electrostatic and hydrophobic energy contributions are plotted against the MC steps. In this sequence the hydrophobic interactions are dominant and hence hydrophobic interactions are being satisfied (energy lowered) to achieve the minimally frustrated state.

Monte Carlo Simulation of the Real Disordered Sequences

For the real disordered sequences native structure is not known. Therefore for each real disordered sequence (DisProt id: $DP00039, DP00116, DP00083, DP00421$) an extended conformation is generated using PyMOL. This conformation is compacted using a molecular dynamics (MD) simulation using AMBER ff99SB force field in vacuum. From the MD trajectory, different conformations with varied Rg are selected as inputs for the MC simulation trial runs. A stable Rg range is identified for each real disordered sequence. Now, any conformation of this Rg range is selected as an input structure for MC simulation with simulated annealing to sample the Rg range effectively. The thermal energy for the simulated annealing is varied from $k_BT = 10000$ to $k_BT = 1$.

Monte Carlo Simulation of the Real Globular Sequences

To check whether the potential defined by eqns. 7, 8, 9, 10, 11, and 12 of the main manuscript can collapse the globular protein sequences MC simulations are performed on the expanded conformations of selected globular proteins (PDB id: $1ERD$, $1DV5$, $1PMR$, $2PKO$ and $2I8L$) The expanded conformations are randomly selected from the molecular dynamics (MD) simulations of the selected globular proteins in explicit solvent (TIP3P water model[2]) with ff99SB force field[3] using AMBER 12[4]. Each protein is energy minimized twice; the energy minimization of the solvent was followed by the energy minimization of the solvated protein. The energy minimized system was equilibrated in a NVT ensemble for 100 ps by gradually raising the temperature from 100 K to 400 K at constant volume. This was followed by equilibration in NPT ensemble for 200 ps at constant temperature of 400 K. The equilibrated system was then again subjected to NVT ensemble for 100 ps to gradually increase the temperature form 400 K to 500 K (600 K for $2I8L$). Finally, NPT production run of 2 ns with a time step of 1 fs is performed for each protein. From the MD trajectory of each protein, an expanded conformation is randomly selected as an input for the subsequent MC simulations using the proposed coarse-grained potential (eqns. 7, 8, 9, 10, 11, and 12 of the main manuscript). Each MC simulation showed sharp decrease in the R_g confirming that the defined potential collapses the globular protein sequences. The extent of collapse is shown from the variation in R_g in Fig. S6.

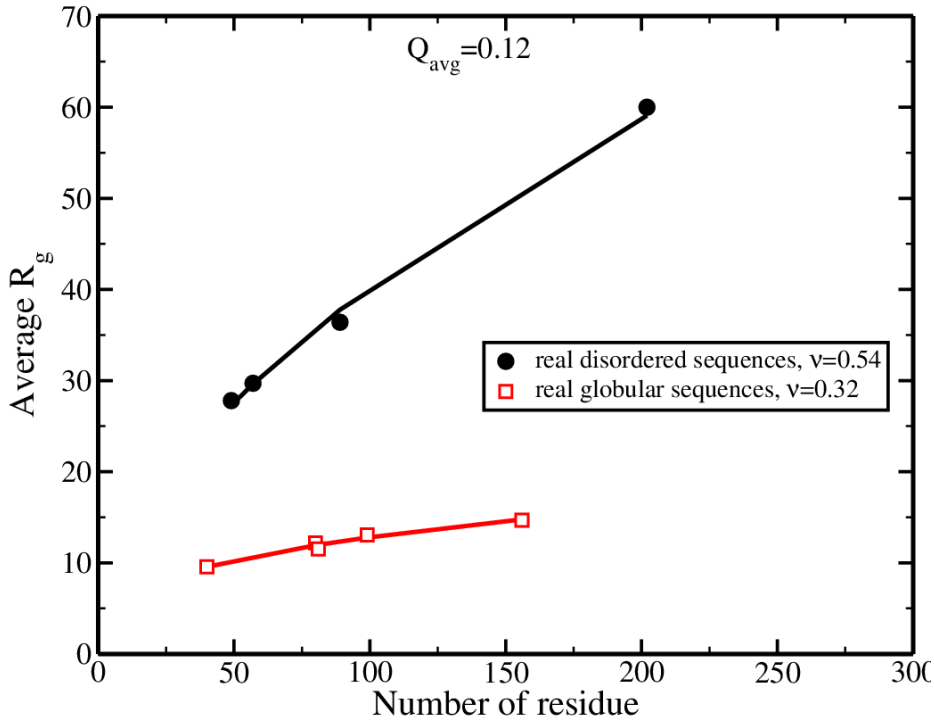

FIG. S5. R_g values plotted against the respective sequence lengths. The sequence length ranges between 40 and 156 residues for selected globular proteins and between 49 and 202 residues for disordered proteins. The power law fitting gives the Flory exponent $\nu = 0.54$ for disordered sequences with $Q_{avg} \sim 0.12$ and $\nu = 0.32$ for globular sequences with similar mean net charge.

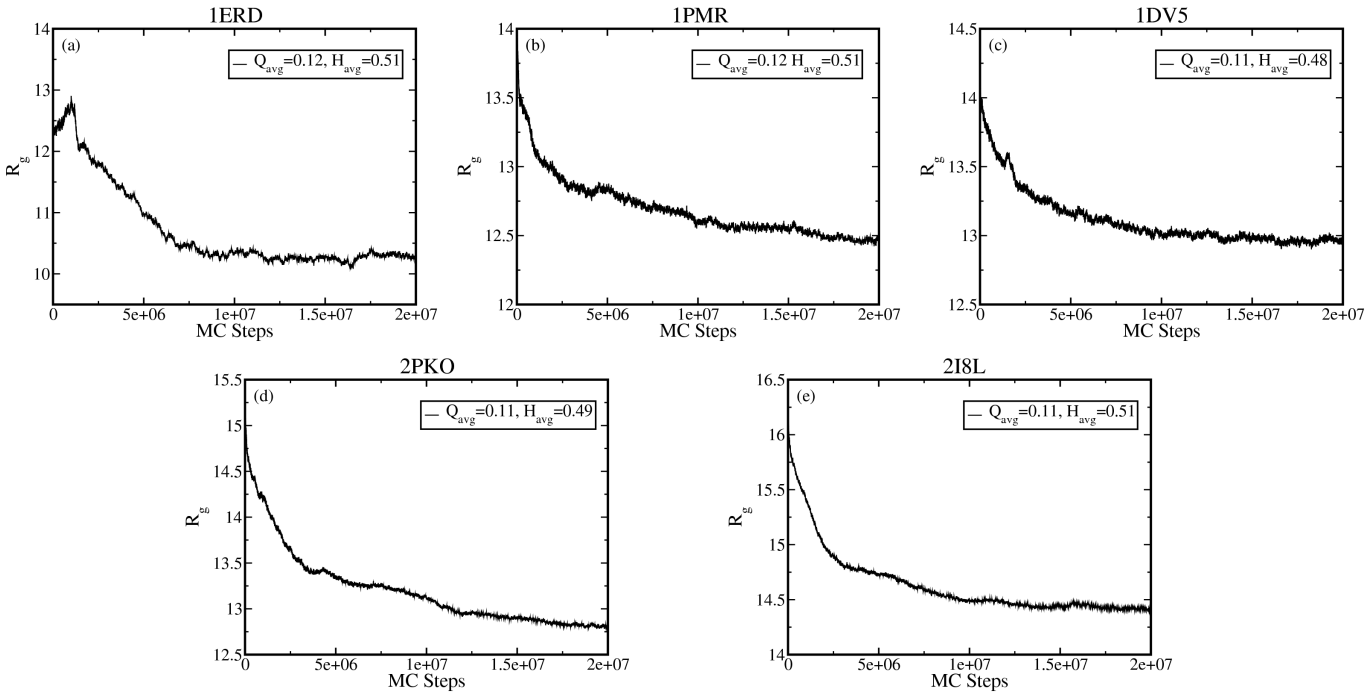

FIG. S6. The R_g values are plotted against the MC steps for a) 1ERD, b) 1PMR, c)1DV5, d)2PKO and, e) 2I8L. The MC simulations are performed using the proposed coarse-grained potential. In each simulation the R_g shows a sharp decrease confirming that the potential induces collapse in globular protein sequences.

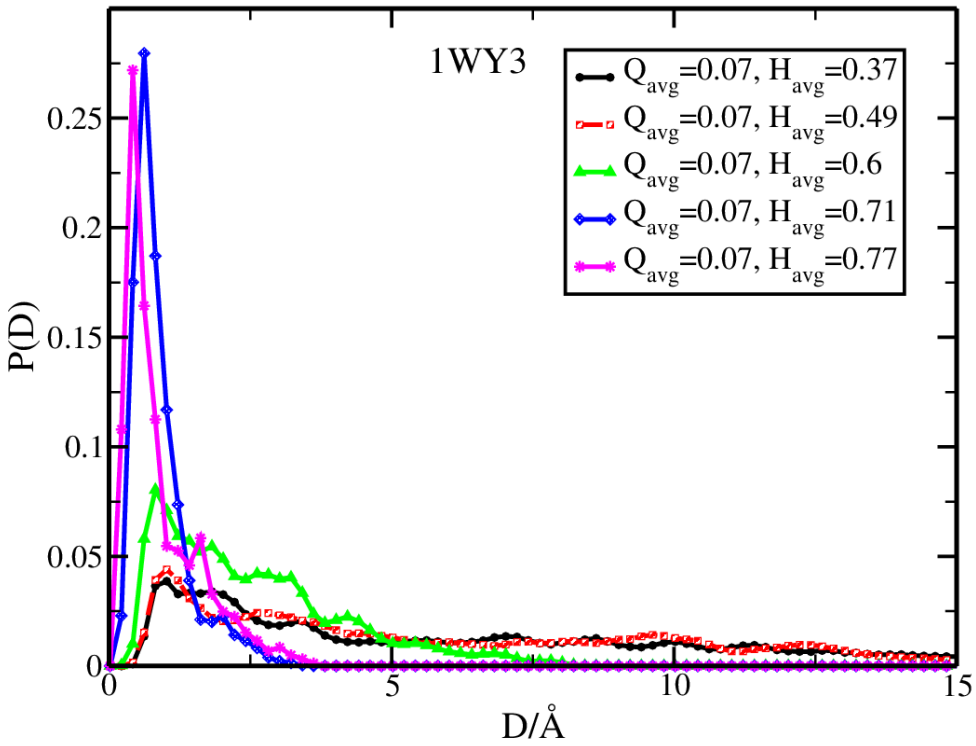

FIG. S7. Probability of conformational heterogeneity plotted against the conformational heterogeneity for designed sequences for the target conformation 1WY3 with Q_{avg}=0.07.

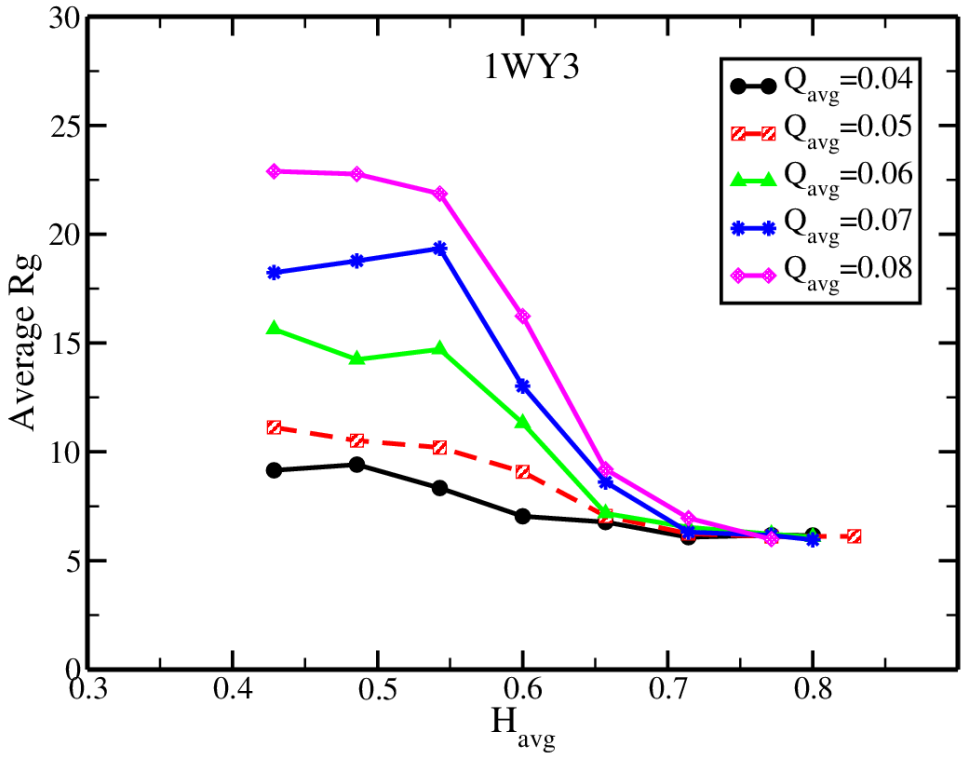

FIG. S8. Average radius of gyration (Å) for each MC simulation is plotted as a function of the mean hydrophobicity of the corresponding designed sequences for the target structure 1WY3.

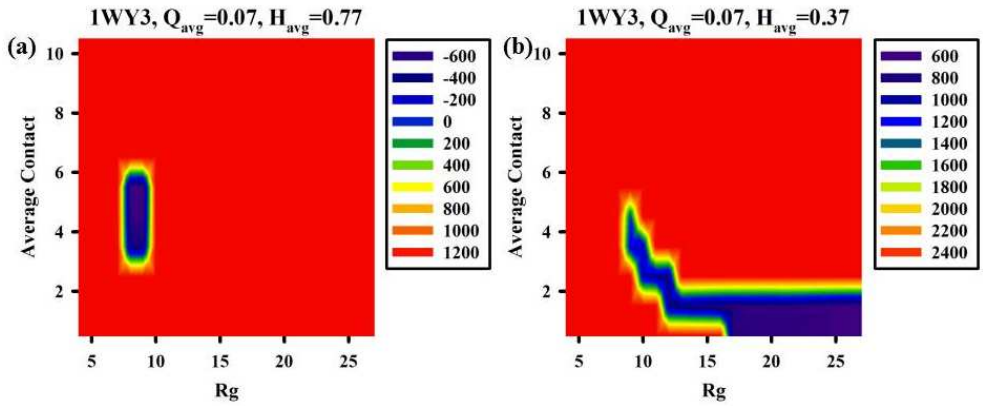

FIG. S9. The contour map of energy landscape plotted against the R_g (Å) and average contact for the target structure 1WY3.

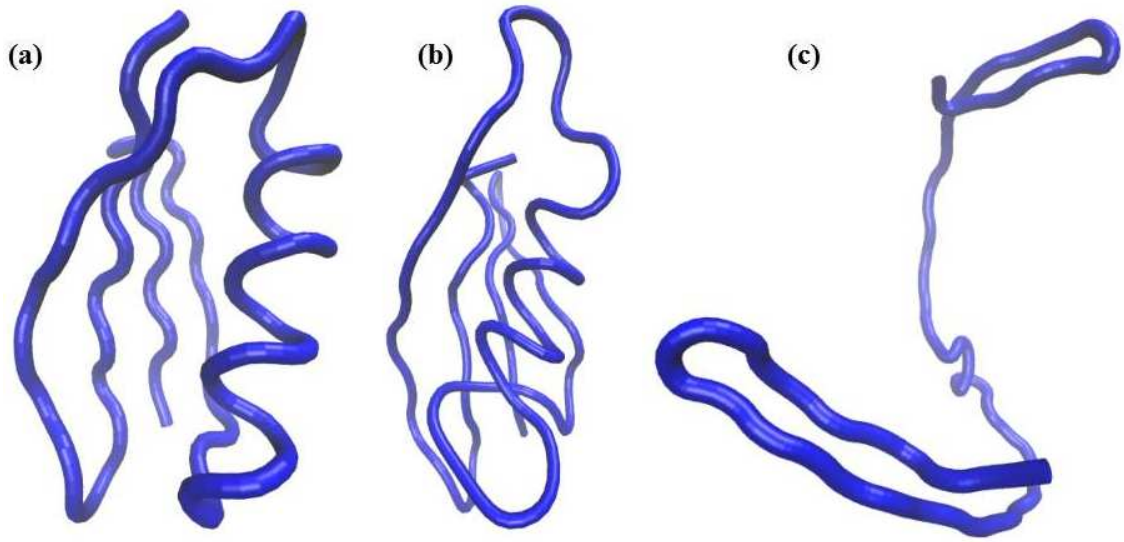

FIG. S10. (a) The 1PGA target structure, (b) random structure from the MC simulation of the ordered sequence and (c) random structure from the MC simulation of the disordered sequence.

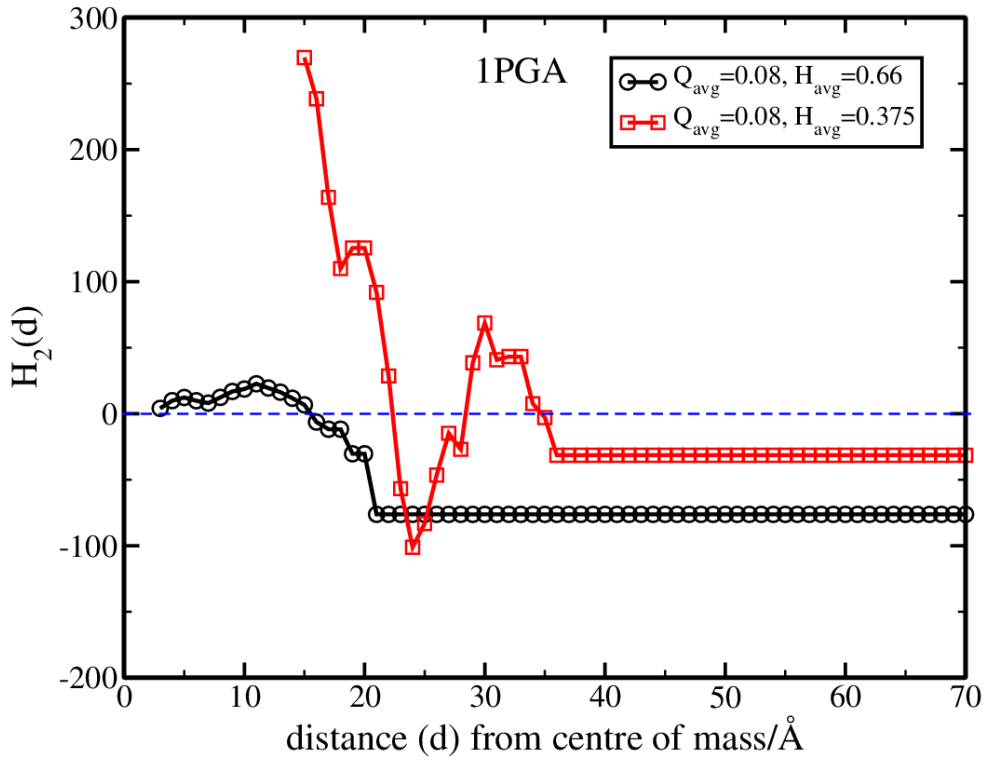

FIG. S11. The second order hydrophobic moment[1] ($H_2(d)$) of a designed ordered sequence ($H_{avg} = 0.66$) and a designed disordered sequence ($H_{avg} = 0.375$) is plotted against the distance (d) from centre of mass. The plot shows that in the ordered sequence the hydrophobic residues form a hydrophobic core while in the disordered sequence the hydrophobic residues fail to form any hydrophobic core.

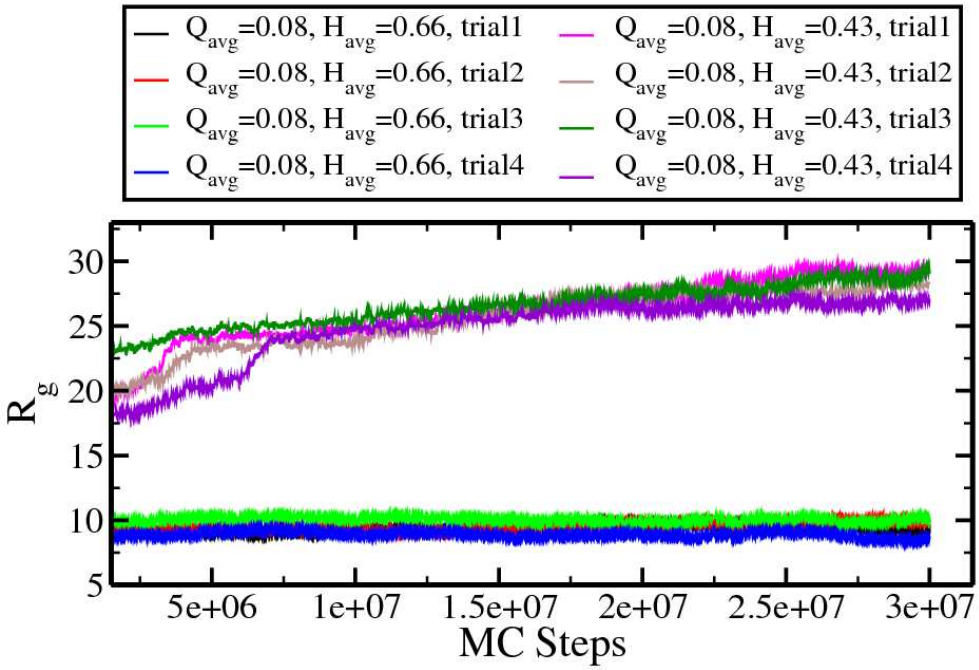

FIG. S12. Plot showing the convergence of the MC simulations.

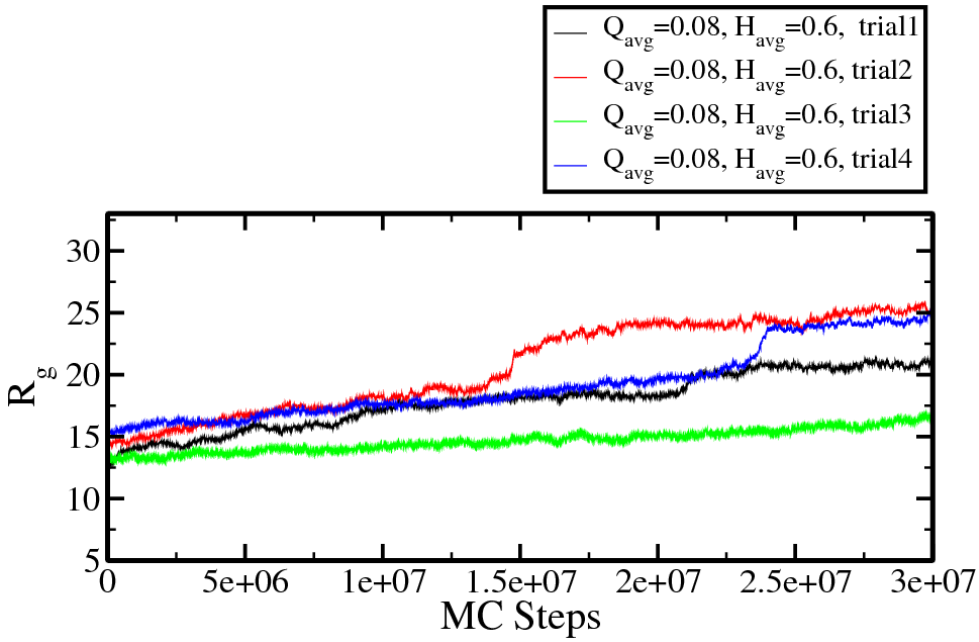

FIG. S13. Plot showing that sequence with mean net charge and mean hydrophobicity in the order-disorder transition region populates conformations in between compact/folded state and noncompact/unfolded state.

References

1 N. Rawat and P. Biswas, *J. Phys. Chem. B*, 2012, **116**, 6326.

2 W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, M. L. Klein, *J. Chem. Phys.*, 1983, **79**, 926.

3 V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, C. Simmerling, *Proteins Struct. Funct. Bioinf.*, 2006, **65**, 712.

4 D. A. Case et al., AMBER 12 *University of California, San Francisco*, 2012, **1**, 3.