Supporting Information for:

Protein Dynamics and the Folding Degree.

Vladimir Sladek,[*,†,‡] Ryuhei Harada,[*,¶] and Yasuteru Shigeta[¶]

†*Institute of Chemistry - Centre for Glycomics, Dubravska cesta 9, 84538 Bratislava, Slovakia*

‡*Agency for Medical Research and Development (AMED), Chiyoda-ku, Japan*

¶*Center for Computational Sciences, University of Tsukuba, Tennodai 1-1-1, Tsukuba, Ibaraki 305-8577, Japan*

E-mail: sladek.vladimir@savba.sk; ryuhei@ccs.tsukuba.ac.jp

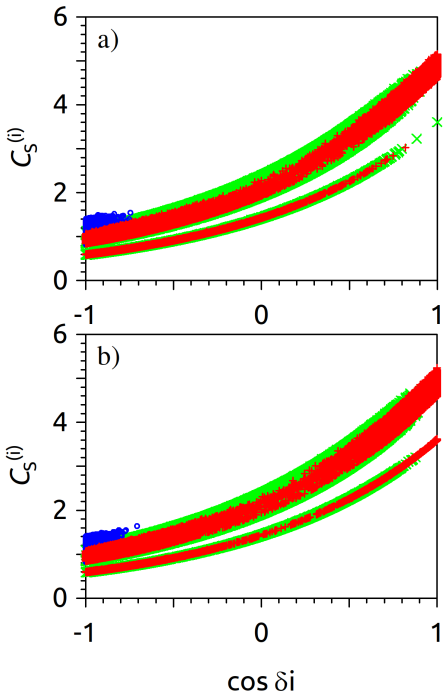

Figure S 1: Node subgraph centrality $C_S^{(i)}$ as a dependence of $\cos\delta_i$, with δ_i being the backbone dihedrals ψ (red), ω (blue), φ (green) for a) chignolin and b) deca-alanine in 1 μs MD simulation of folding.

The apparent gap in the $C_S^{(i)}$ *vs.* $\cos\delta_i$ dependence may seem strange. To explain it we take a deeper look at its origin. One can re-draw the dependence in Fig. S1 as a $C_S^{(i)}$ *vs.* δ_i dependence, however, the gap is still there. These dependencies depict the node subgraph centrality that was acquired in the statistical sample gathered during the 1 μs long folding process (some 10^5 samples for each angle). We see, that certain regions are not well sampled. Apart from the expected narrow sampling window for the peptide bond dihedral angle ω, also φ is sparsely represented for $\cos\varphi \to 1$, i.e. $\varphi \approx 0$. One ought to remember that the values of φ and ψ are not fully arbitrary. Therefore it may be more conclusive to evaluate the residue contribution to the total folding degree (details in the main text). By doing so, we can plot the $RC_S^{(k)}$, k now labelling the residues, as a two dimensional plot of φ and ψ, see Fig. S2. Now it becomes much more apparent, that only some subspaces of φ and ψ are

sampled. It also explains the wide spread of the $C_S^{(i)}$ *vs.* $\cos\delta_i$ dependence. Namely, that e.g. $C_S^{(i)}(\varphi)$ depends also on values of ψ and *vice versa*. The lack of sampling density in certain φ and ψ subspaces is the reason for the gap in the $C_S^{(i)}$ data. Moreover, from the definition of $C_S^{(i)}$ one sees that it depends on the eigenvalues and the associated eigenvectors of the weighted adjacency matrix, hence virtually all dihedral angles play a role in the $C_S^{(i)}$ value for a particular ith node (dihedral). On the other hand, the more or less thin bell-shaped two dimensional surface of $RC_S^{(k)}$ *vs.* φ, ψ seems to suggest that long-range dependence is not a significant contributing factor to $RC_S^{(k)}$ and it is determined predominantly by φ, ψ of the kth residue. By "thin surface" we mean there is not a significant spread along the vertical $RC_S^{(k)}$ axis for a given φ, ψ point. To confirm this assumption, we examine possible correlation between randomly picked $RC_S^{(k)}$ in both deca-alanine and chignolin, see Fig. S3. We cannot detect any such correlation. Hence it seems that the conformation of other residues does not significantly influence the value of $RC_S^{(k)}$ at the given kth residuum. In other words, it seems that the residual contribution to the total folding degree is a (more or less) local property.

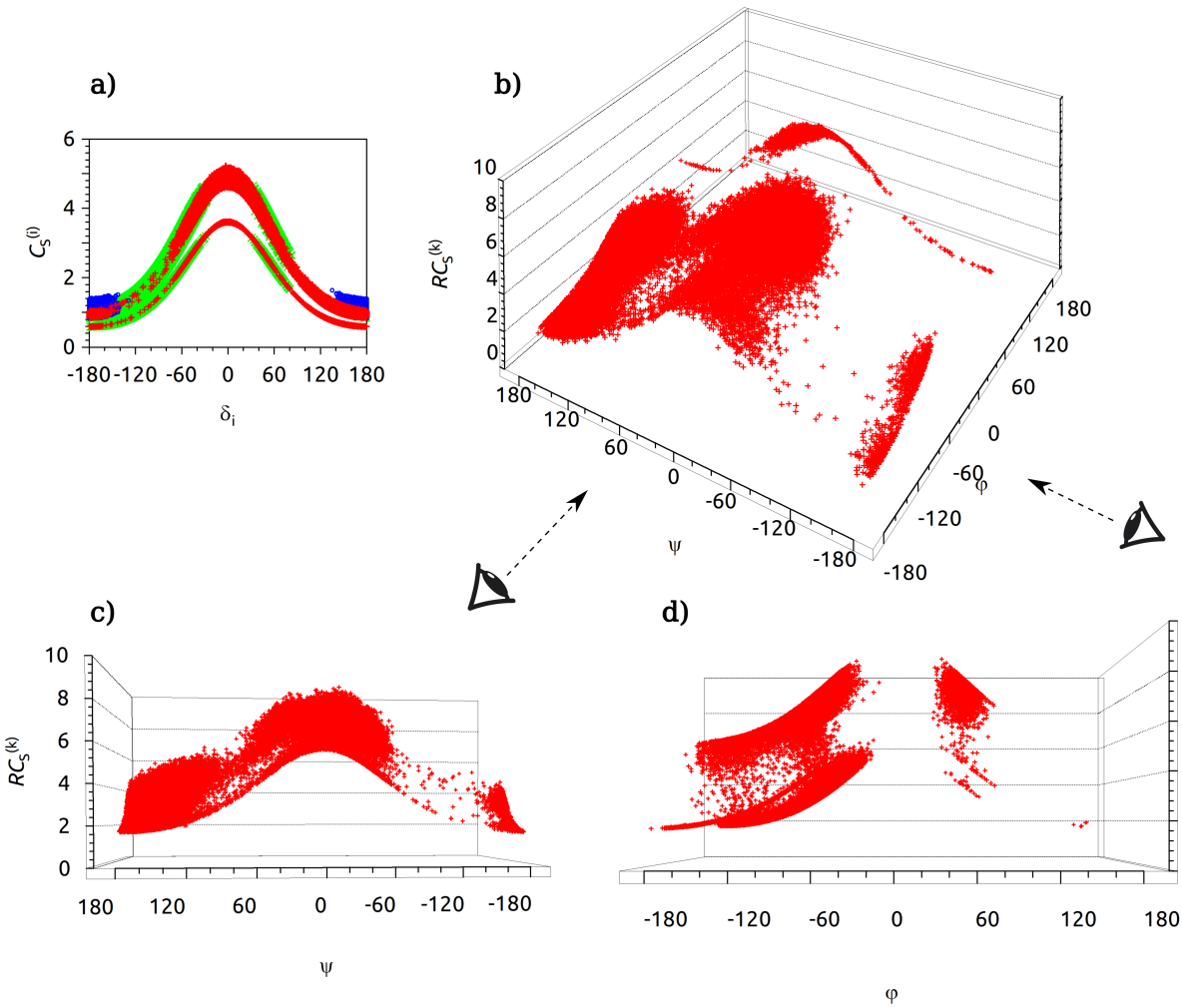

Figure S 2: Node and residual contributions to folding degree of deca-alanine during 1 μs MD simulation of folding: a) node subgraph centralities $C_S^{(i)}$ *vs.* δ_i with δ_i being the backbone dihedrals ψ (red), ω (blue), φ (green) b) residual contribution of non-terminal residues $RC_S^{(k)}$ as a function of ψ, φ c) view along the φ axis depicting the dependence on ψ d) view along the ψ axis depicting the dependence on φ. Note that the sampling gaps in the two dimensional $RC_S^{(k)}$ dependence explain the gap in the one dimensional $C_S^{(i)}$ dependence. The analysis used the 8×10^4 sampled conformations of non-terminal residues.

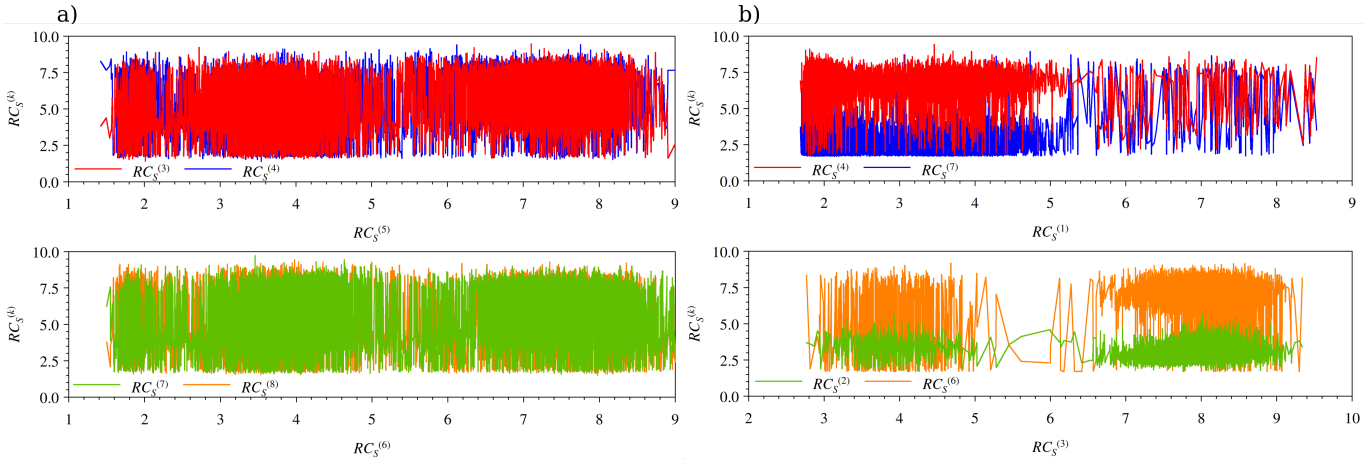

Figure S 3: The (non-existent) correlation of $RC_S^{(k)}$ for randomly selected residues for deca-alanine (columns a) and chignolin (column b) during 1 μs MD simulation of folding.

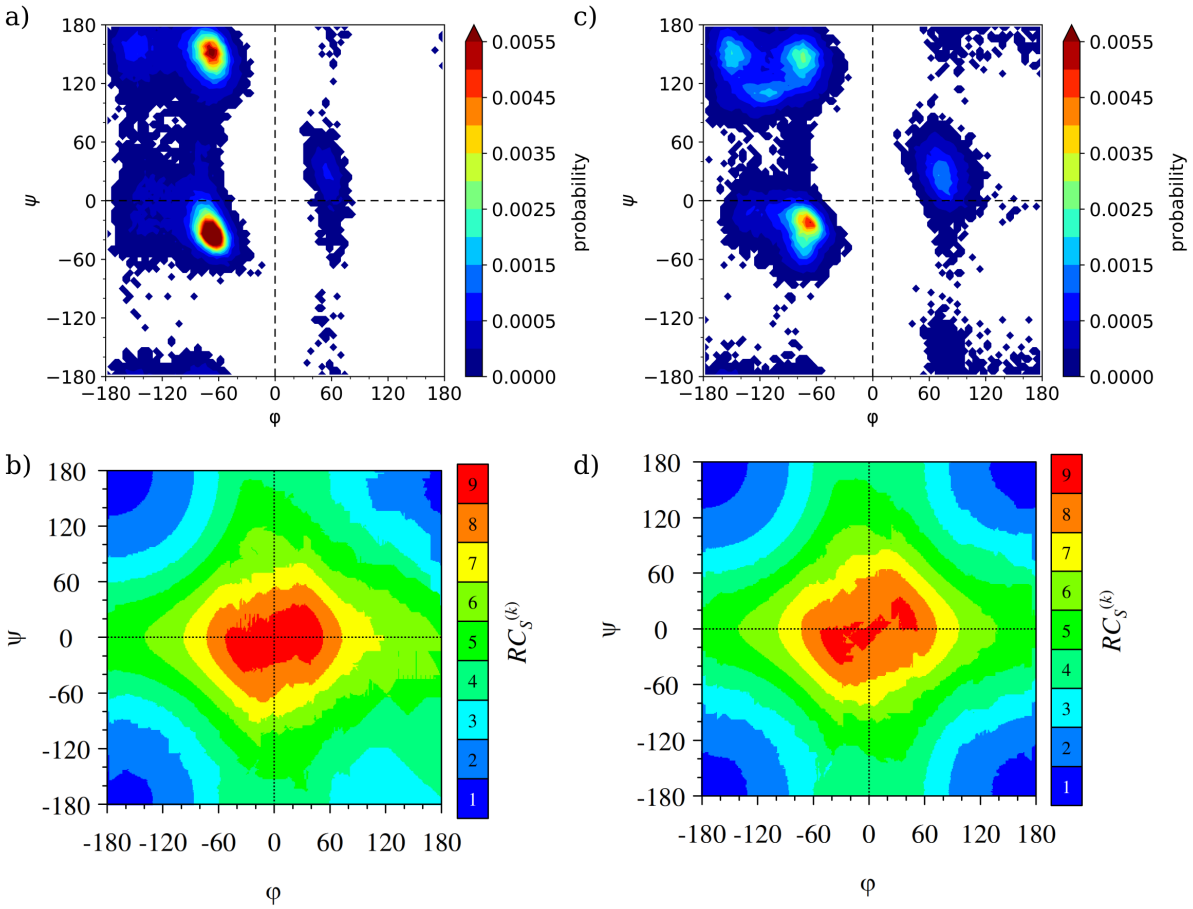

Figure S 4: Ramachandran-like histogram plots showing frequency of accessed combinations of dihedral angles ψ, φ of non-terminal residues (top) and $RC_S^{(k)}$ as a function of ψ, φ (bottom) for deca-alanine (left) and chignolin (right) during 1 μs MD simulation of folding. The histograms (a, c) are normalised such that sum of values in all bins yields unity. There are in total 8×10^4 values of φ, ψ combinations and the associated $RC_S^{(k)}$ values as we sampled 10^4 snapshots and there are eight non-terminal residues.

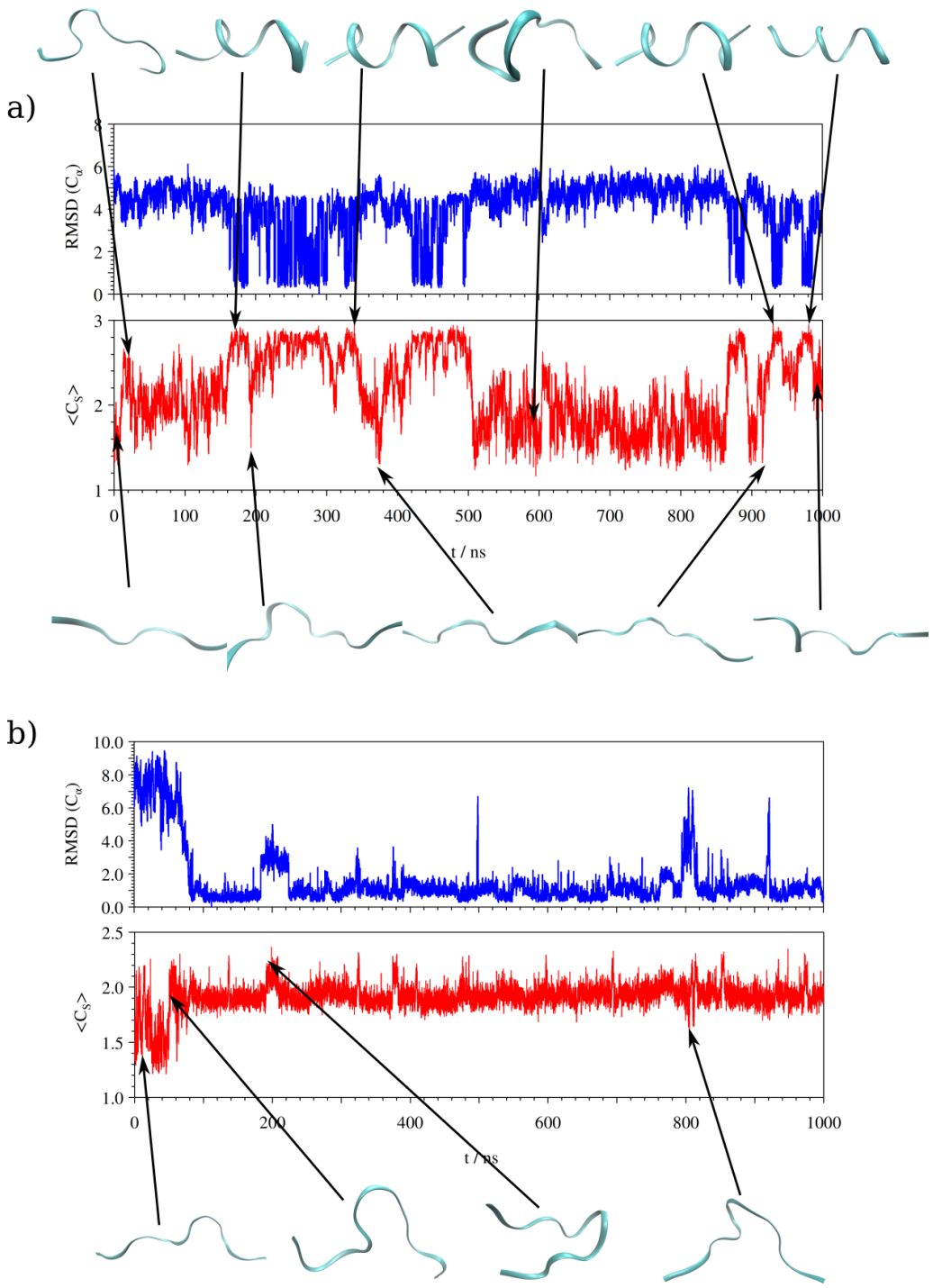

Figure S 5: Total folding degree (red) and RMSD of the alpha carbons (blue) as a time series for a) deca-alanine b) chignolin during 1 μs MD simulation of folding (time step for the recorded data points is 100 ps). Some representative structures of different folding stages are depicted and associated with the $\langle C_S \rangle$ value. The RMSD was calculated with respect to the structure corresponding to the minima on the $\langle C_S \rangle$ based PMFs, see Fig. S9.

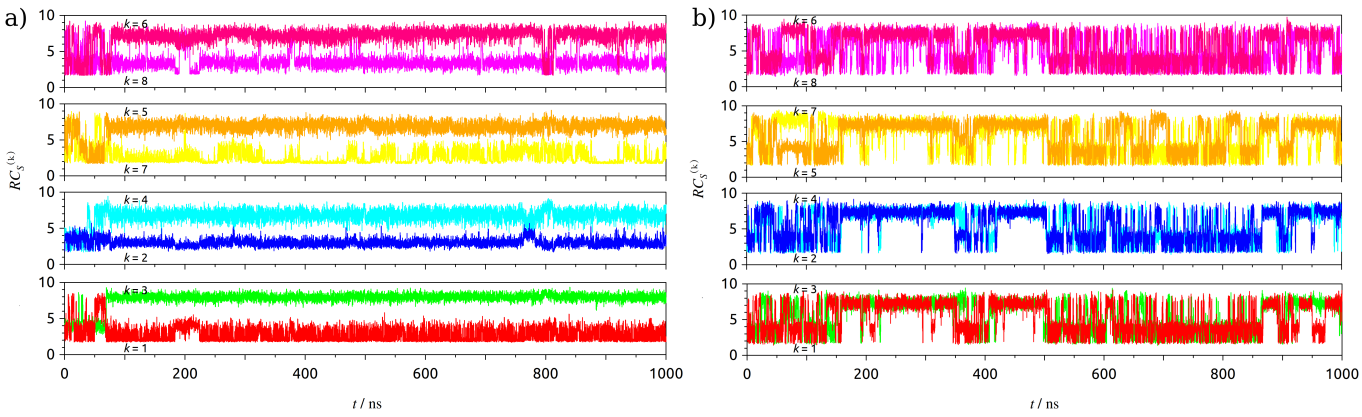

Figure S 6: Residue folding degree $RC_S^{(k)}$ as a time series for a) chignolin and b) deca-alanine during 1 μs MD simulation of folding (time step for the recorded data points is 100 ps). Terminal residues not shown.

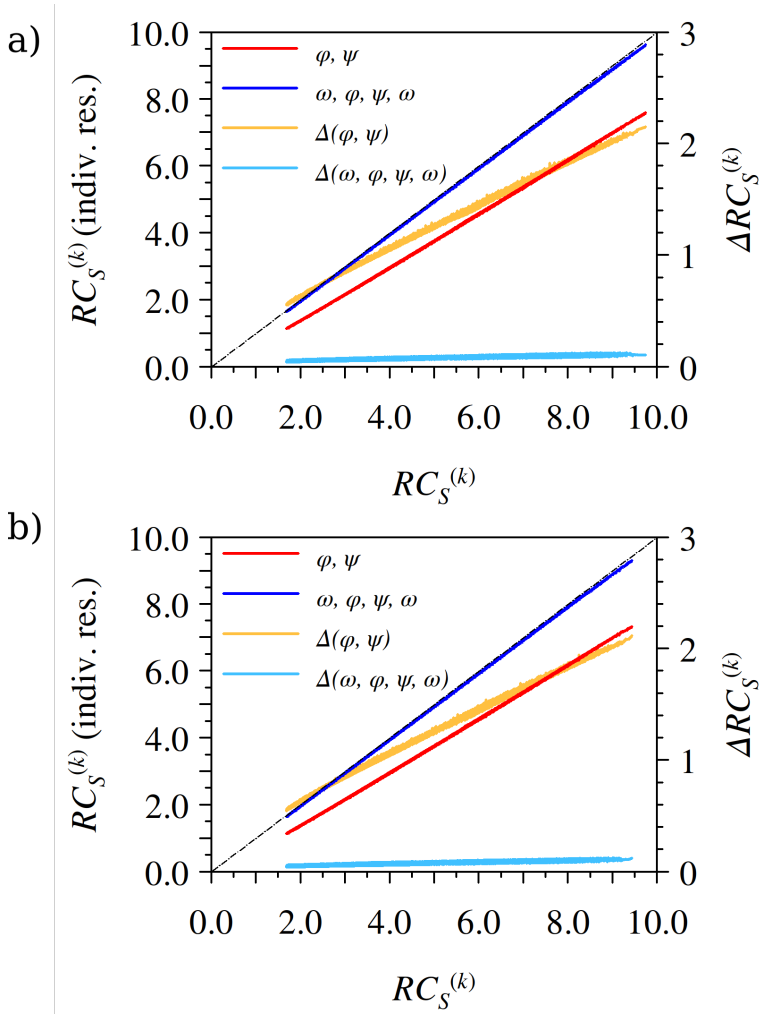

Figure S 7: $RC_S^{(k)}$ values for a) deca-alanine, b) chignolin calculated form sub-matrices of the weighted adjacency matrix **A** plotted against $RC_S^{(k)}$ values calculated form the full matrix **A** (on horizontal axis). Red lines correspond to 2×2 sub-matrices (φ, ψ) and blue to the 4×4 sub-matrices $(\omega, \varphi, \psi, \omega)$. The differences from the $RC_S^{(k)}$ values calculated form the full weighted adjacency matrix **A** are in orange $(\Delta(\varphi, \psi))$ and cyan $(\Delta(\omega, \varphi, \psi, \omega))$, respectively. Correlation coefficient is 1 in all cases, but the slopes do differ; deca-alanine: 0.791 for the 2×2 and 0.991 for the 4×4 sub-matrix and chignolin: 0.799 and 0.991 for the 2×2 and 4×4 sub-matrices, respectively. The average differences are 1.370 and 0.085 for deca-alanine and 1.250 and 0.078 for chignolin and the 2×2 and 4×4 sub-matrices, respectively. The analysis used the 8×10^4 sampled conformations of non-terminal residues.

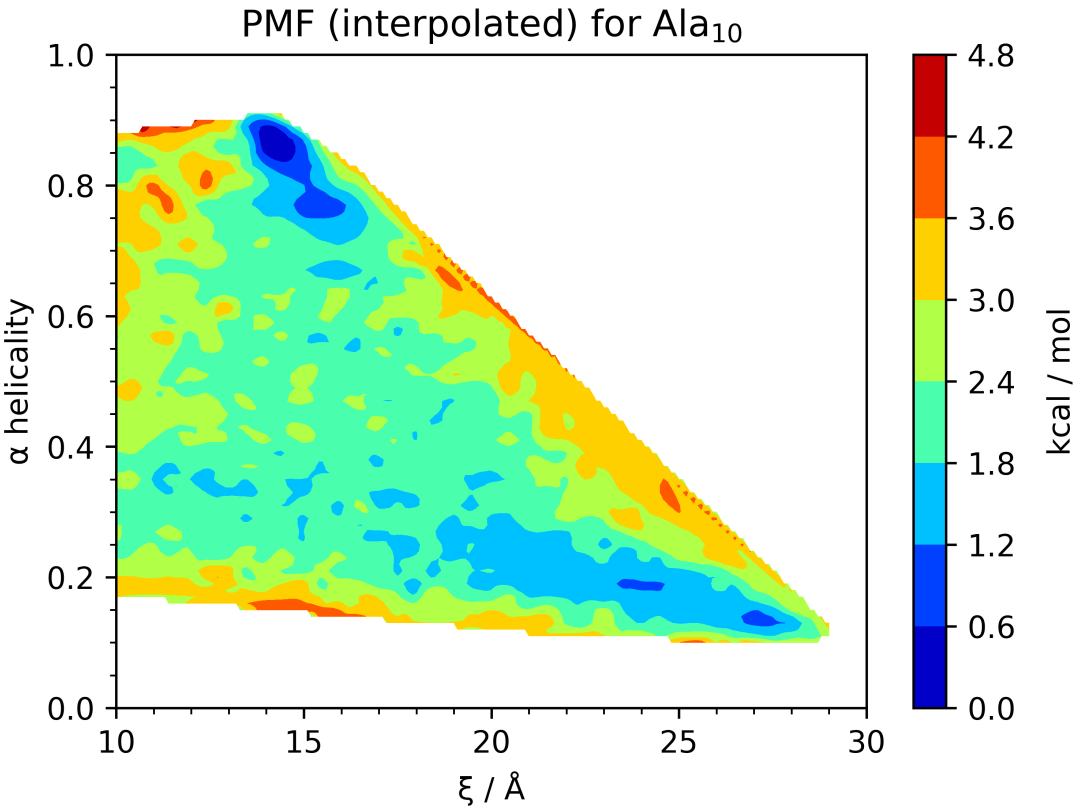

Figure S 8: PMF at 300K for deca-alanine during 1 μs MD simulation of folding.

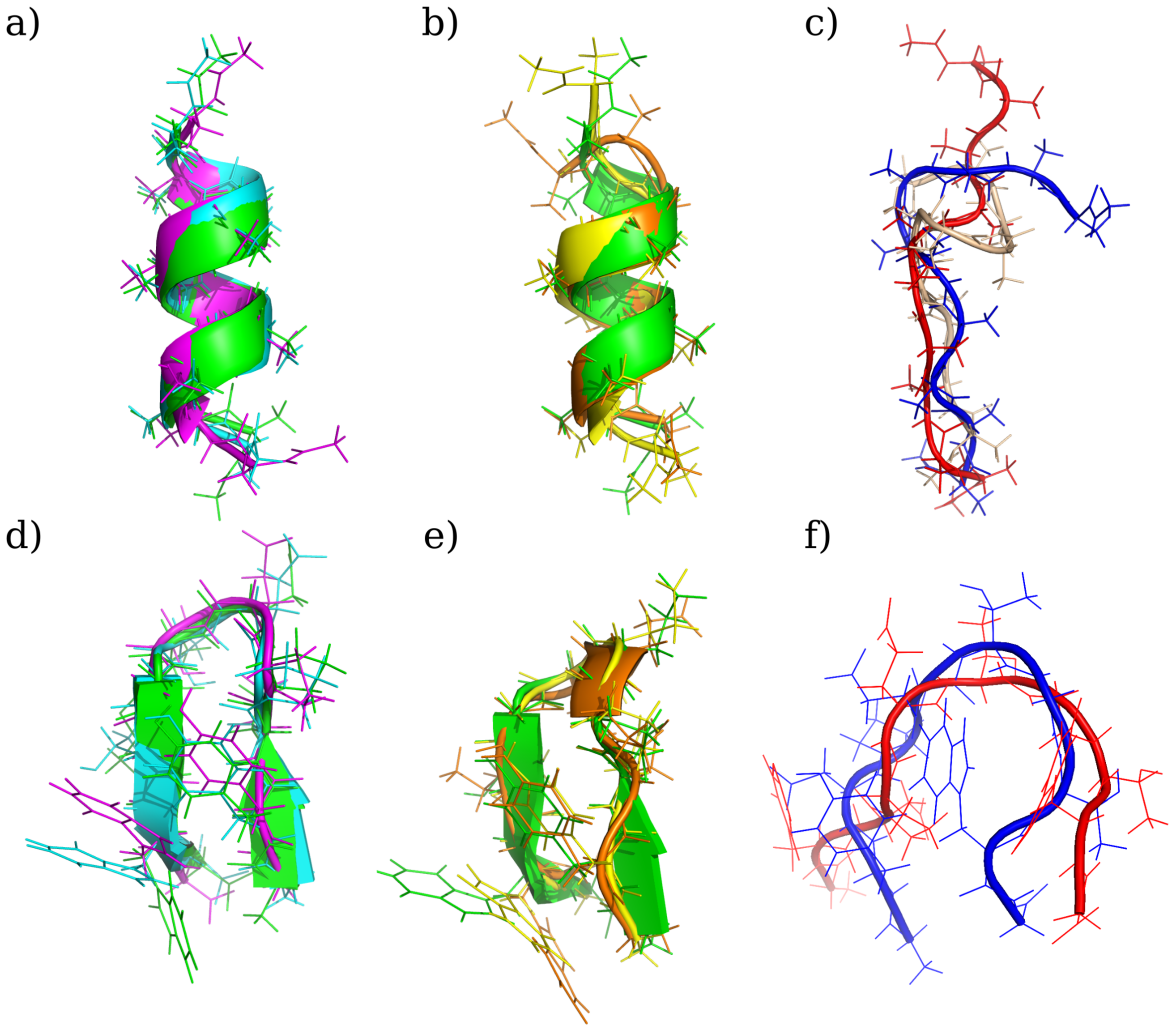

Figure S 9: Structures corresponding to the minima of the $\langle C_S \rangle$ based PMF at 300K. Top row (a-c) for deca-alanine, bottom row (d-f) for chignolin. The green coloured ones in a) & b) are the structures for $\langle C_S \rangle = 2.80$ and $\langle C_S \rangle = 1.89$ in d) & e), the global minima of the PMFs for deca-alanine and chignolin, respectively. Magenta and cyan (a, d) depict structures with $\langle C_S \rangle \pm 0.01$ from the global minima. Yellow and orange (b, e) depict structures with $\langle C_S \rangle \pm 0.1$ from the global minima. Three conformations of deca-alanine with $\langle C_S \rangle = 1.75, 1.85, 1.95$ are shown in c). Two conformations of chignolin from the local minima with $\langle C_S \rangle = 1.42, 1.55$ are shown in f).

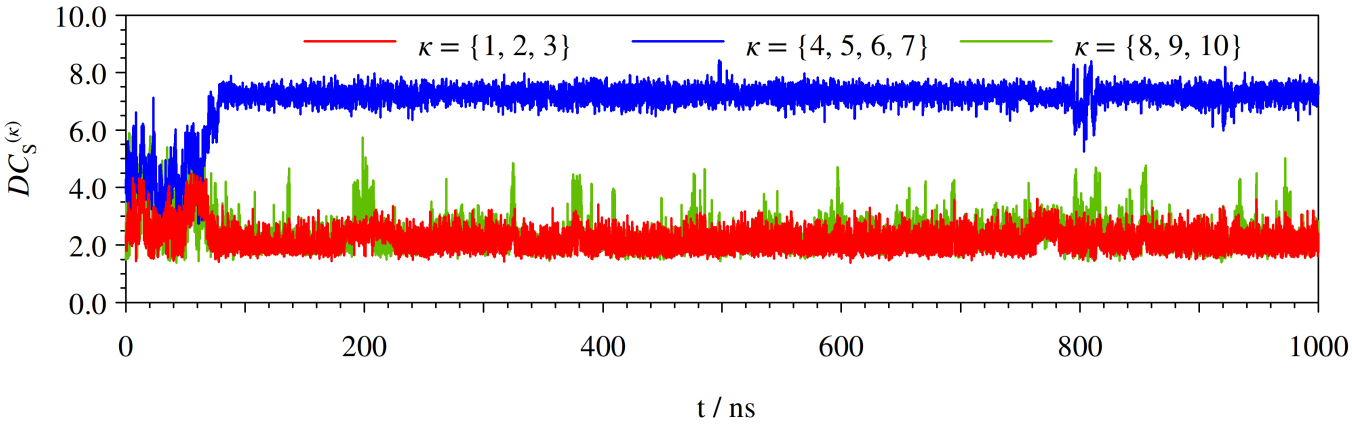

Figure S 10: Domain folding degree as a time series for chignolin during 1 μs MD simulation of folding (time step for the recorded data points is 100 ps).