Supporting Information for:
Weighted distance functions improve analysis of high-dimensional data: application to molecular dynamics simulations

Nicolas Blöchliger, Amedeo Caflisch, and Andreas Vitalis*

University of Zurich
Department of Biochemistry
Winterthurerstrasse 190, CH-8057 Zurich

* To whom correspondence should be addressed: a.vitalis@bioc.uzh.ch

Supporting Methods

SAPPHIRE plot for Beta3S (Figure 6 in the main text). We represented the peptide by the sine and cosine values of 99 nonsymmetric dihedral angles. We used a stochastic, approximate algorithm[1] to generate the progress indices for the SAPPHIRE plots in Figure 6. The stochastic algorithm is scalable to large data sets because of the preorganization of the data via tree-based, hierarchical clustering.[2] The upper threshold radius and the tree height for the clustering were set to 1 and 8. The lower threshold radius was set to 0.487, 0.433, and 0.449 for the SAPPHIRE plots based on the UW (eq 1), GW (eq2), and LAW (eq 4) measures, respectively. These settings were chosen to have roughly 100000 clusters at the leaf-level. All the SAPPHIRE plots use snapshot 468441 as the starting snapshot. The number of guesses to find near neighbors[1] was set to 4000. We made use of two recent improvements to the algorithm for generating the approximate progress index (Vitalis, manuscript submitted). First, after the initial clustering of the data, we cluster the data on the three levels of finest resolution again. This improves the homogeneity in the clustering on these levels. The algorithm for generating the approximate progress index requires the computation of near neighbors for the individual snapshots, and the hierarchical clustering is used to focus the search-space. Here, we allow to enlarge this search space if the number of 4000 guesses can otherwise not be satisfied. This is controlled via the CAMPARI keyword “FMCSC_CPROGRDEPTH,” which was set to 3.

SAPPHIRE plot for BPTI (Figure 8 in the main text and Figure S3). For the SAPPHIRE plots shown in Figures 8a-c, 10, and Figure S3, we represented the protein by 271 nonsymmetric dihedral angles. These include χ_2 and χ_3 angles of cysteines. We used an exact algorithm to compute the progress index for these SAPPHIRE plots. Conversely, for Figure 8d, the approximate algorithm[1] was used with the root-mean square deviation (RMSD) of the positions of 699 nonsymmetric atoms as the distance function. Parameter settings for the auxiliary clustering were taken from previous work.[3] In particular, the upper and lower threshold radii and the tree height for the clustering were set to 3.6Å, 2.5Å, and 4, respectively. The number of guesses to find near neighbors[1] was set to 1000, and we made use of the recent improvements to the algorithm as described above for Beta3S in the context of Figure 6.

All the SAPPHIRE plots for BPTI shown in Figures 8, 10, and Figure S3 use snapshot 20521 as the starting snapshot and the value of CAMPARI keyword FMCSC_CPROGMSTFOLD was 1. For the annotation with dihedral angles we used binning into up to three bins with boundaries chosen as follows: Cys14 χ_1 (-120°, -5°, 120°), Cys14 χ_2 (-140°, 0°, 130°), Cys14-Cys38 χ_3 (0°, 150°), Cys38 χ_2 (-155°, -105°, 120°), Cys38 χ_1 (-120°, 0°, 140°), Cys38 ψ (-120°, 80°), Arg42 ψ (-100°, 75°), Cys30-Cys51 χ_3 (0°, 150°), Pro9 ψ (-115°, 70°), Thr11 ψ (-120°, 95°), Asp3 φ (0°, 100°). These boundaries were obtained from direct inspection of the individual histograms for each angle.

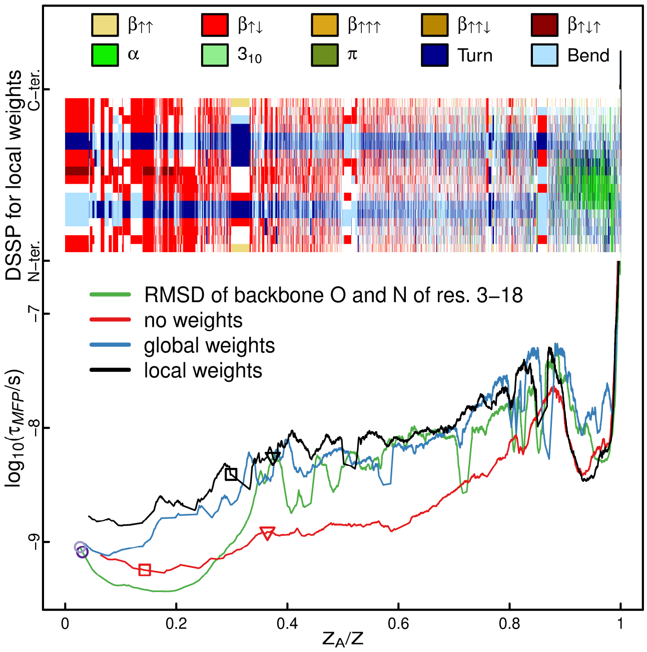

Figure S1 | Cut-based free energy profiles (cfep)[4] for Beta3S. The snapshots were clustered using a recent tree-based, hierarchical clustering algorithm[2] with four different distance functions. All the four cfeps used the largest cluster as their reference cluster. The coordinate root mean square deviation (RMSD) computed over backbone oxygen and nitrogen atoms of residues 3-18 after pairwise alignment serves as reference to illustrate manual feature selection (green curve). The clustering used coarse and fine thresholds of 10 and 1.5 Å, respectively, and yielded 161778 clusters with a tree height of 16. Note that the same parameters were used as in Figure 6A in prior work.[2] The resulting cfep displays the native state ($Z_A/Z \leq 0.38$) as a homogeneous basin without any internal structure. The second distance function we employed is the Euclidean distance of the sine and cosine values of 99 nonsymmetric dihedral angles (red curve, UW measure). The clustering used a tree height of 8, and 1 and 0.473 as coarse and fine thresholds, respectively. These settings produced 162039 clusters. The resulting cfep has a smaller folding barrier than the one based on RMSD, and the other metastable states are not as pronounced. The third and the fourth distance functions used are the GW (eq 2 in the main text, $\tau = 2$ ns) and LAW (eq 4 in the main text, $\Delta = 2$ ns, $\alpha = 1$) measures, respectively. We again chose the sine and cosine values of the same dihedral angles to represent the data (blue and black curves, respectively). Here the clustering used a tree height of 8, and a value of 1.0 as coarse threshold. The fine thresholds were set to 0.399 (GW) and 0.422 (LAW), which gave 160574 and 159754 clusters, respectively. In both cases, the folding barrier near $Z_A/Z = 0.4$ is slightly higher than for the RMSD-based profile and substructure can be found within the native basin ($Z_A/Z \leq 0.4$). The cfep based on local weights is annotated by the secondary structure[5] of the centroids of the individual clusters (legend on top). The positions of the clusters from the native basin and from the folding barrier that were extracted for Figure 3 are indicated with squares and triangles, respectively. The purple circles highlight the positions of the clusters containing the two snapshots shown in Figure 6d in the main text in the RMSD-based profile (green curve).

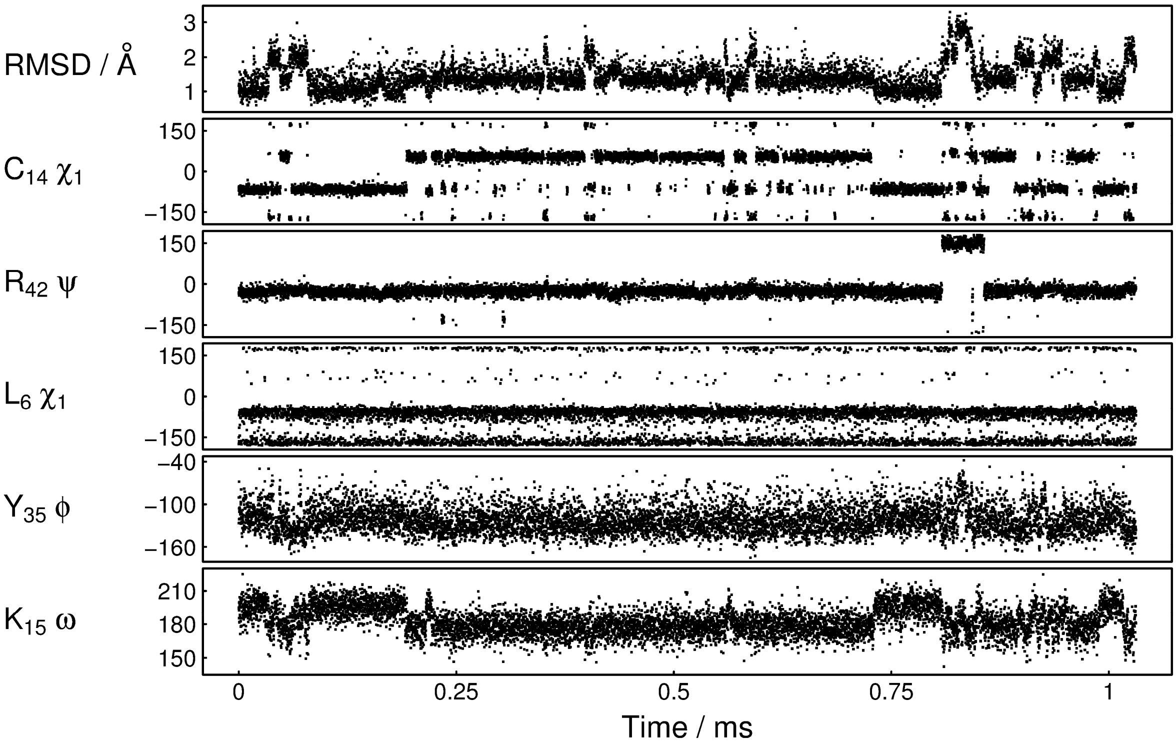

Figure S2 | Time series of the RMSD to a reference conformation and of selected dihedral angles for BPTI. For the slow dihedral angles (Cys14 χ_1, Arg42 ψ) jumps coincide with jumps in the RMSD time series. No such conclusion is obtained for fast dihedral angles (Leu6 χ_1). Other dihedral angles include a recognizable slow component but exhibit overlap among the distributions within different metastable states (Tyr35 φ and the ω angle between Cys14 and Lys15). Note that these observations are reflected in the respective autocorrelation functions (Figure 7a in the main text). Data are plotted for every 5th snapshot, *i.e.*, every 125 ns.

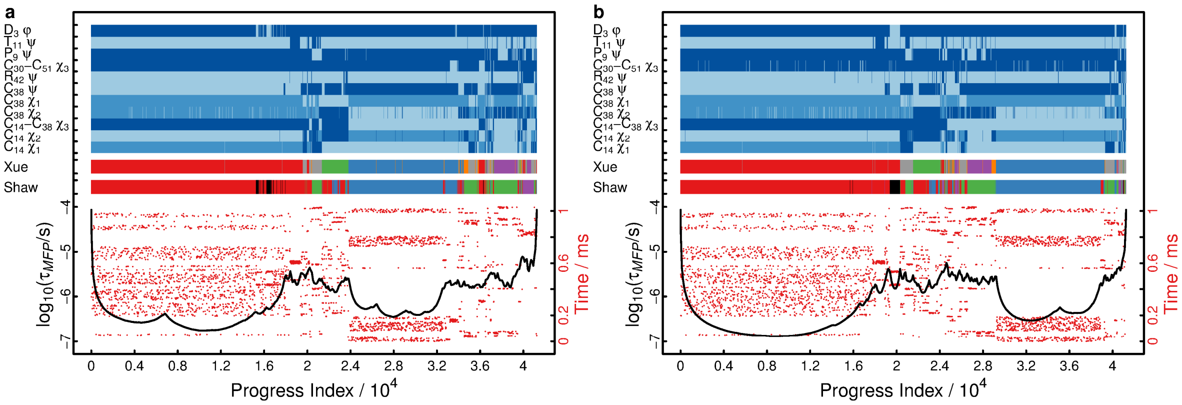

Figure S3 | Influence of the time lag τ for global weights. These SAPPHIRE plots are identical to Figure 8b in the main text with the exception that the employed lag times τ are 25 ns (**a**) and 100 µs (**b**), respectively. The saving frequency of the trajectory equals 25 ns, *i.e.*, τ = 25 ns is the lowest possible time lag. Note that most autocorrelation functions have decayed completely at τ = 100 µs (Figure 7a in the main text), which leads to very noisy estimates for the weights w_i. Nevertheless, the resulting profile highlights the most important features of the conformational landscape of BPTI. Please refer to the caption of Figure 8 in the main text for plotting details and to the Supporting Methods above for further information.

Supporting References

(1) Blöchliger, N.; Vitalis, A.; Caflisch, A., A scalable algorithm to order and annotate continuous observations reveals the metastable states visited by dynamical systems. *Comp. Phys. Comm.* **2013,** *184* (11), 2446-2453.

(2) Vitalis, A.; Caflisch, A., Efficient construction of mesostate networks from molecular dynamics trajectories. *J. Chem. Theor. Comput.* **2012,** *8* (3), 1108-1120.

(3) Blöchliger, N.; Vitalis, A.; Caflisch, A., High-resolution visualisation of the states and pathways sampled in molecular dynamics simulations. *Sci. Rep.* **2014,** *4*, 6264.

(4) Krivov, S. V.; Karplus, M., One-dimensional free-energy profiles of complex systems: Progress variables that preserve the barriers. *J. Phys. Chem. B* **2006,** *110* (25), 12689-12698.

(5) Kabsch, W.; Sander, C., Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983,** *22* (12), 2577-2637.