SUPPORTING INFORMATION

for

Efficient Construction of Mesostate Networks from Molecular Dynamics Trajectories

Andreas Vitalis,[1,*] and Amedeo Caflisch[1,*]

[1]*Department of Biochemistry*

University of Zurich

Winterthurerstrasse 190, CH-8057 Zurich, Switzerland

*: To whom correspondence should be addressed:

Amedeo Caflisch: Tel: +41446355521, E-mail: caflisch@bioc.uzh.ch

Andreas Vitalis: Tel: +41446355597, E-mail: a.vitalis@bioc.uzh.ch

SUPPORTING METHODS

Derivation of Efficient Formulas for Computing Intrinsic and Relative Cluster Properties

Mean Distance between Clusters for Periodic Variables: If two mesostates A and B (note that mesostates are addressed by a slightly different syntax in the main text using c_A instead of A) are far away from one another, the individual pairwise distances may or may not be subject to periodic wraparounds. We cannot treat this exactly, but handle it at the centroid level by using another heuristic. The noise introduced by this effect grows with D. The heuristic we use involves precomputing a periodic shift component as follows:

$$\begin{aligned}\vec{\zeta}_\phi &= 2\pi\left[\mathrm{H}\left(\vec{\mathrm{LS}}_A/N_A-\vec{\mathrm{LS}}_B/N_B-\pi\right)+\mathrm{H}\left(\vec{\mathrm{LS}}_A/N_A-\vec{\mathrm{LS}}_B/N_B+\pi\right)-1\right]\\ \lambda &= N_A N_B \vec{\zeta}_\phi\cdot\vec{\zeta}_\phi+2\left[N_B\vec{\mathrm{LS}}_A\cdot\vec{\zeta}_\phi-N_A\vec{\mathrm{LS}}_B\cdot\vec{\zeta}_\phi\right]\end{aligned} \tag{S1}$$

In equation S1, H(x) denotes the Heaviside function, and the remaining variables are parts of the CF vectors of both mesostates as introduced in equation 1 in the main text. The determined offset λ is then simply included in the computation of d_{IC}:

$$d_{IC}^2(A,B)=\frac{\left(\lambda+N_B SS_A+N_A SS_B-2\vec{\mathrm{LS}}_A\cdot\vec{\mathrm{LS}}_B\right)}{N_A N_B D} \tag{S2}$$

Note that equation S2 (unlike equation 2 in the main text) represents the dimensionality-normalized formula that is also generalized to the case where both N_A and N_B may be larger than unity.

Fluctuating Weights: Cluster diameter d is normally defined as the mean distance between snapshots belonging to a cluster. When using weights on the individual dimensions, a normalization by the total weight is the most natural choice to render the "distance" conceptually independent of dimensionality. Now let us consider a mesostate A of size N_A. We will write the sum redundantly (all N_A^2 terms), noting that distances for $i=j$ do not contribute to the double sum, and can simply be removed by normalizing with $N_A(N_A\text{-}1)$:

$$\begin{aligned}N_A(N_A-1)d_c^2(A) &= \sum_{k\in A}^{N_A}\sum_{l\in A}^{N_A} W_{kl}^{-1}\sum_i^D\left(I_k^i+I_l^i\right)\left(X_k^i-X_l^i\right)^2 \quad \text{with} \quad W_{kl}=\sum_i^D\left(I_k^i+I_l^i\right)\\ N_A(N_A-1)d_c^2(A) &\approx \left(W_A+W_A\right)^{-1}\left[\sum_{k\in A}^{N_A}\sum_{l\in A}^{N_A}\sum_i^D\left(I_k^i+I_l^i\right)\left(X_k^i-X_l^i\right)^2\right] \quad \text{with} \quad W_A=N_A^{-1}\sum_k^{N_A}\sum_i^D I_k^i\end{aligned} \tag{S3}$$

The approximation introduced in equation S3 regarding the normalizer is needed to be able to represent the other terms in compact form. If the mesostate is relatively tight, it is reasonable to assume that the error introduced is small (see Table 1 in the main text). Next, we rearrange the terms in the triple sum to be expressed as products and sums of vectors that can be incremented in an extended CF vector.

$$2\,d_c^2(A) \approx M_A^{-1}\left[\sum_{k\in A}^{N_A}\sum_{l\in A}^{N_A}\sum_i^D I_k^i (X_k^i)^2 + I_k^i (X_l^i)^2 + I_l^i (X_k^i)^2 + I_l^i (X_l^i)^2 - 2\left(I_k^i X_k^i X_l^i + I_l^i X_k^i X_l^i\right)\right]$$
$$d_c^2(A) \approx M_A^{-1}\left[N_A \cdot \sum_i^D \sum_{k\in A}^{N_A} I_k^i (X_k^i)^2 + \sum_i^D \sum_{k\in A}^{N_A} I_k^i \sum_{l\in A}^{N_A} (X_l^i)^2 - \sum_i^D \sum_{k\in A}^{N_A}\sum_{l\in A}^{N_A}\left(I_k^i X_k^i X_l^i + I_l^i X_k^i X_l^i\right)\right] \quad \text{(S4)}$$
$$\text{with } M_A = W_A N_A (N_A - 1) = (N_A - 1)\sum_k^{N_A}\sum_i^D I_k^i$$

Hence:

$$d_c^2(A) \approx M_A^{-1}\left[N_A\left(\sum_i^D WQ_A^i\right) + \vec{\mathrm{LI}}_A \cdot \vec{\mathrm{SS}}_A - 2\left(\vec{\mathrm{LS}}_A \cdot \vec{\mathrm{WS}}_A\right)\right]$$
$$d_c^2(A) \approx \frac{N_A\left(\sum_i^D WQ_A^i\right) + \vec{\mathrm{LI}}_A \cdot \vec{\mathrm{SS}}_A - 2\left(\vec{\mathrm{LS}}_A \cdot \vec{\mathrm{WS}}_A\right)}{(N_A - 1)\sum_i^D LI_A^i} \text{ with} \quad \text{(S5)}$$
$$\vec{\mathrm{WS}}_A = \sum_{k\in A}^{N_A} \vec{\mathrm{I}}_k \circ \vec{\mathrm{X}}_k,\ \vec{\mathrm{WQ}}_A = \sum_{k\in A}^{N_A} \vec{\mathrm{I}}_k \circ \vec{\mathrm{X}}_k \circ \vec{\mathrm{X}}_k,\ \vec{\mathrm{SS}}_A = \sum_{k\in A}^{N_A} \vec{\mathrm{X}}_k \circ \vec{\mathrm{X}}_k, \text{ and } \vec{\mathrm{LI}}_A = \sum_{k\in A}^{N_A} \vec{\mathrm{I}}_k$$

The circle denotes the element-by-element (Hadamard) product. Equation S5 implies five different vectors that all need to be accumulated for every mesostate. This is in contrast to the one vector and one scalar that is collected in the CF vector for systems with fixed weights. Importantly, however, the added cost remains independent of mesostate size (still $O(D)$). By assuming a fixed weight of 1 for each dimension, it is easily seen that the above equation relaxes to the dimensionality-normalized variant of the solution in equation 1 in the main text.

The mean inter-mesosate distance is defined as the average distance between snapshots from the two mesostates.

$$N_A N_B d_{IC}^2(A,B) = \sum_{k\in A}\sum_{l\in B} W_{kl}^{-1} \sum_i^D \left(I_k^i + I_l^i\right)\left(X_k^i - X_l^i\right)^2 \text{ with } W_{kl} = \sum_i^D \left(I_k^i + I_l^i\right) \quad \text{(S6)}$$

A largely analogous calculation yields:

$$d_{IC}^2(A,B) \approx M_{AB}^{-1}\left[\sum_{k\in A}^{N_A}\sum_{l\in B}^{N_B}\sum_i^D I_k^i (X_k^i)^2 + I_k^i (X_l^i)^2 + I_l^i (X_k^i)^2 + I_l^i (X_l^i)^2 - 2\left(I_k^i X_k^i X_l^i + I_l^i X_k^i X_l^i\right)\right]$$
$$d_{IC}^2(A,B) \approx M_{AB}^{-1}\left[N_A \cdot \left(\sum_i^D WQ_B^i\right) + N_B \cdot \left(\sum_i^D WQ_A^i\right) + \vec{\mathrm{LI}}_A \cdot \vec{\mathrm{SS}}_B + \vec{\mathrm{LI}}_B \cdot \vec{\mathrm{SS}}_A - 2\left(\vec{\mathrm{LS}}_A \cdot \vec{\mathrm{WS}}_B + \vec{\mathrm{LS}}_B \cdot \vec{\mathrm{WS}}_A\right)\right]$$
$$\text{with } M_{AB} = N_A N_B\left(N_A^{-1}\sum_k^{N_A}\sum_i^D I_k^i + N_B^{-1}\sum_l^{N_B}\sum_i^D I_l^i\right) \quad \text{(S7)}$$
$$d_{IC}^2(A,B) \approx \frac{N_A\left(\sum_i^D WQ_B^i\right) + N_B\left(\sum_i^D WQ_A^i\right) + \vec{\mathrm{LI}}_A \cdot \vec{\mathrm{SS}}_B + \vec{\mathrm{LI}}_B \cdot \vec{\mathrm{SS}}_A - 2\left(\vec{\mathrm{LS}}_A \cdot \vec{\mathrm{WS}}_B + \vec{\mathrm{LS}}_B \cdot \vec{\mathrm{WS}}_A\right)}{N_B\left(\sum_i^D LI_A^i\right) + N_A\left(\sum_i^D LI_B^i\right)}$$

Again, the accuracy of the approximations underlying equation S7 is documented in Table 1 in the main text. Note that it is also possible to combine the two sets of heuristics for fluctuating weights and periodic quantities. This is largely straightforward and gives rise to dataset "ω,ϕ,ψ/ I" in Table 1.

Implementation of Other Clustering Algorithms

Leader Algorithm: The dataset is scanned from the last entry to the first. For each snapshot, the list of existing mesostates is scanned backwards. Snapshot j is added to the first mesostate c_A, for which $d(j,i_A)$ is less than the threshold criterion. Here, i_A is the first snapshot ever added to c_A (assumed center). If no such mesostate is found, the list of mesostates is appended with a new mesostate containing only snapshot j. Results based on such an algorithm are presented in Figs. 2, 4-7, and S1-S5. The forward Leader algorithm that appears in the analysis of *n*-butane in the main text (Figs. 5 and S2) switches both reading directions, *i.e.*, processes data from the first entry to the last, and scans the list of existing mesostates in the order in which they were created. Differences between the two variants of the Leader algorithm serve to illustrate the sensitivity of different classes of results toward data input order.

Agglomerative Algorithm (Hierarchical Clustering): The distance matrix for the dataset is computed, and an ordered list of snapshot-snapshot distances $d(i,j)$ is computed from it. From the smallest distance onward, the two corresponding snapshots are joined if they are both not part of a mesostate. If one (i) of them is already a member of a mesostate (c_A), the distance $d_{CC}(j,c_A)$ is evaluated. If it is less than the distance threshold criterion, t, c_A is appended with snapshot j. If both are assigned already, the centroid-to-centroid distance $d_{CC}(c_B,c_A)$ is computed, and the two corresponding mesostates are joined if $d_{CC}(c_B,c_A)$ is less than t. The procedure is stopped as soon as the considered distance exceeds twice the threshold criterion. Using centroid-based distances corresponds to a mean linkage criterion. Results from the agglomerative algorithm are shown in Figs. 2, 5, S1, and S2.

Moments of inertia as fluctuating weights

Moments of inertia are computed by standard means as $I_\Phi = \sum_i m_i r_{i,\Phi}^2$ where m_i is the mass of atom i and $r_{i,\Phi}$ is its distance from the axis of rotation. The sum is restricted to those atoms that lie on the shorter of the two chain ends, *i.e.*, it is assumed that there is a specific building direction for each dihedral angle in the system, and that the longer chain ends remain fixed in space upon changes to the value of said dihedral angles.

SUPPLEMENTARY TABLES

Table S1: Statistics for the data in Fig. 7 in the main text.

Dataset	# Mesostates (total)	$\langle r_c \rangle$ for $N_c \geq 2$	$\langle r_c \rangle$ for $N_c \geq 1$	# Microstates in largest mesostate
$H = 16$, $t_1 = 0.27$	184685	0.1836	0.1563	17613
$H = 16$, $t_1 = 0.28$	157147	0.1880	0.1649	24626
$H = 16$, $t_1 = 0.29$	134431	0.1923	0.1728	32894
$H = 16$, $t_1 = 0.30$	116769	0.1956	0.1789	51040
$H = 16$, $t_1 = 0.32$	88367	0.2031	0.1907	59937
$H = 16$, $t_1 = 0.34$	66653	0.2109	0.2018	117009
$H = 16$, $t_1 = 0.36$	45499	0.2193	0.2133	185616
$H = 16$, $t_1 = 0.40$	21586	0.2344	0.2317	195294
Leader, $t_1 = 0.27$	185822	0.1733	0.1564	3722
Leader, $t_1 = 0.28$	158532	0.1789	0.1653	5147
Leader, $t_1 = 0.29$	134514	0.1841	0.1734	8116
Leader, $t_1 = 0.30$	114090	0.1895	0.1812	10178
Leader, $t_1 = 0.32$	80757	0.2005	0.1956	15268
Leader, $t_1 = 0.34$	56420	0.2113	0.2086	15106
Leader, $t_1 = 0.36$	38575	0.2220	0.2206	38637
Leader, $t_1 = 0.40$	17155	0.2435	0.2433	22751

The total number of mesostates (including those of size 1) for each algorithm is given in column 2. The dimensionality-normalized, Euclidean snapshot-centroid distance (r_c) averaged over all mesostates with at least two microstates is provided in column 3 (unitless). When single microstate clusters are included (column 4) in the computation of the mean, the values for both algorithms converge exactly at low t_1. Large differences are observed for algorithms in the numbers of microstates contained in the respective largest mesostates (column 5). For instance, at $t_1 = 0.40$, the proposed algorithm collects almost 20% of the entire dataset into a single mesostate, whereas for the Leader scheme it is only ~2%. This explains the inability of the first barrier in Fig. 7 to be preserved for the Leader algorithm in combination with large values of t_1.

SUPPLEMENTARY FIGURE CAPTIONS

Figure S1: Comparison of clustering patterns and mesostate overlap between algorithms. The same data are used as in Fig. 2. For different algorithms (**A**: proposed algorithm with $H = 4$; **B**: proposed algorithm with $H = 24$; **C**: simple Leader algorithm; **D**: rigorous agglomerative algorithm with mean linkage criterion), the plots show as colored dots the top largest mesostates that together encompass at least 10000 snapshots. The threshold criterion (or equivalent, see above) was 5 Å in each case. Black lines indicate the chords obtained from circle-circle intersections. Circles are generated using the centroid of each mesostate as the center and a 5 Å radius. The number of lines correlates with the possible volume overlap of mesostates. Actual mesostate overlap is seen as dots of different colors occupying the same area.

Figure S2: Comparison of mean-first passage times (mfpt or τ_{mfp}) from a reference mesostate. The reference is always chosen such that it is the largest mesostate that is part of the basin corresponding to coarse state *ag*$^-$*a*. The data are obtained using the agglomerative scheme and dihedral angles as the measure of similarity as reference. Resultant values using different algorithms and both dihedral angles and RMSD values are plotted against the reference set. Data were obtained by considering each mesostate in each case, assigning it to a coarse state, and constructing the probability-weighted average mfpt for each coarse state with respect to the reference mesostate. The 216 coarse states were obtained by partitioning the three dihedral angles into the three basins (staggered conformations) of size 110º each (centered at -60º, 60º, and 180º) and three high-energy regions (eclipsed conformations) of size 10º each ($6^3 = 216$). The relationships appear uniformly linear with approximately unit slope, and the average slope for the datasets shown was 0.98. Linearity is visually illustrated by the two dotted lines that correspond to $\text{mfpt} = \text{mfpt}_{\text{ref}}$ and $\text{mfpt} = \text{mfpt}_{\text{ref}} + 14.4$ ps, respectively. The actual offsets appear to correlate qualitatively with the increases in numbers of mesostates reported in Table 2 in the main text. All values are in ps.

Figure S3: The same as Fig. 6 in the main text except for using different DSSP maps. Here, the DSSP string is not based on the microstate that either spawned the mesostate (Leader) or is nearest the centroid of the mesostate (proposed algorithm), but was calculated instead as follows. For each of the 7500 largest mesostates plotted, DSSP histograms for each position in the peptide were created. The letter assignment occurring most frequently in a given position was used in the plotted DSSP string. A maximum likelihood estimate constructed this way treats residue positions independently, which is reasonable as long as there is a family of closely related strings that dominate the distribution. For the

data in Figs. S3 and S5, it therefore appeared unnecessary to construct a correct maximum likelihood estimate in 20-letter string space.

Figure S4: Cut-based free energy profiles for beta3S (*DS5*) as a function of H. Data are clustered based on RMSD values of backbone nitrogen and oxygen atoms over residues 3-18 ($D = 96$), and the threshold settings used were $t_1 = 1.8$ Å with $t_H = 10.0$ Å as coarsest criteria for the proposed algorithm. The bottom half shows cFEPs analogously to Fig. 6 in the main text. The top half shows a color trace corresponding to the DSSP letter assignment for the particular case of the proposed algorithm with $H = 16$ (see caption to Fig. 6 in the main text for details). The small differences observed as a function of H do not appear systematic. Lines have been added to aid in the visual identification of small enthalpic basins. Only the largest 10000 clusters are actually plotted to keep the number of represented objects tractable (including more mesostates does not change the plot noticeably at typical resolution/enlargement).

Figure S5: The same as Fig. 7 in the main text except for using different DSSP maps. This is the analogous variant of Fig. 7 that Fig. S3 is for Fig. 6. See captions to Figs. 7 and S3 for details.

Figure S1:

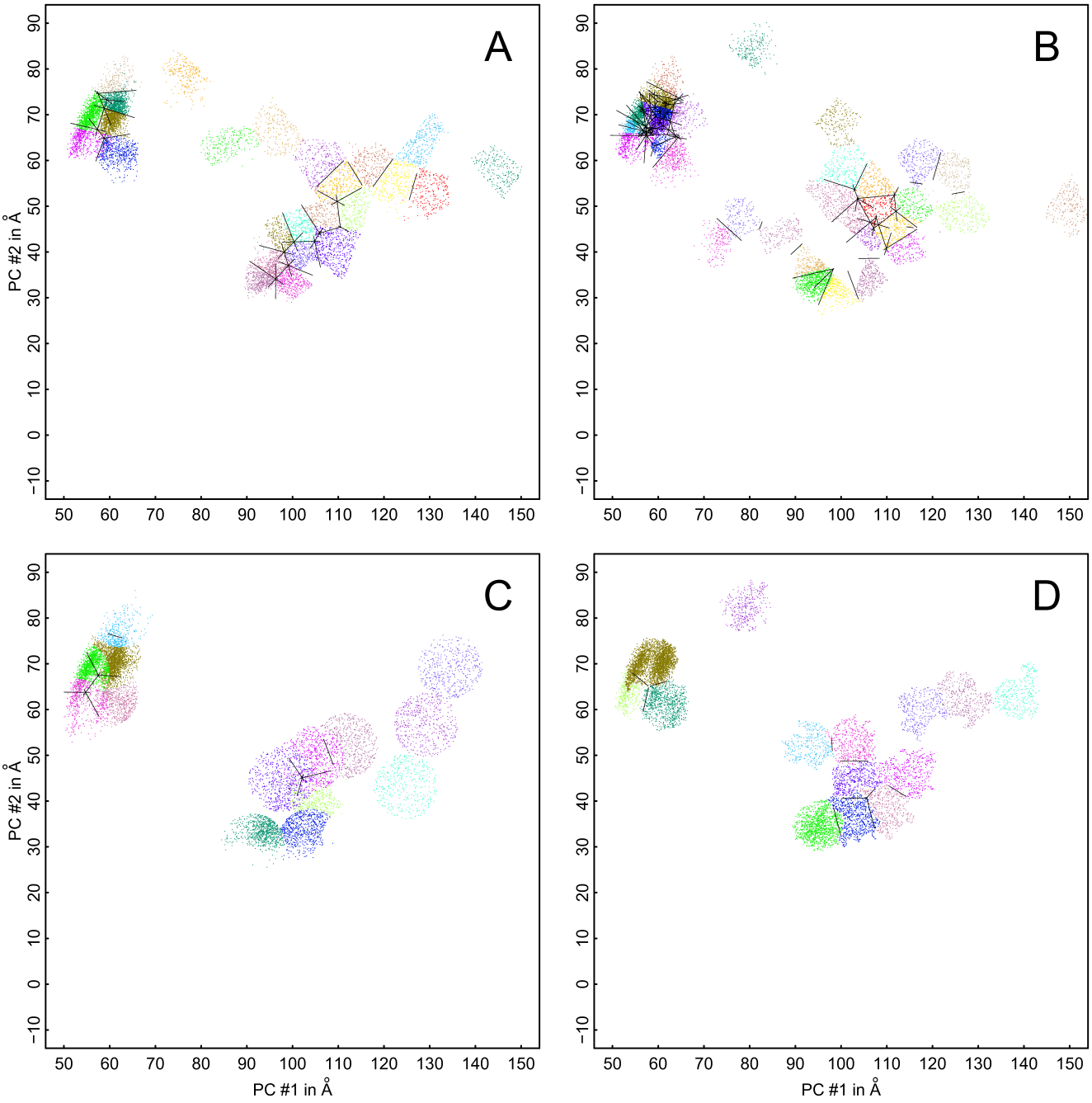

A
B
C
D
PC #2 in Å
PC #1 in Å
−10
0
10
20
30
40
50
60
70
80
90
50
60
70
80
90
100
110
120
130
140
150

Figure S2:

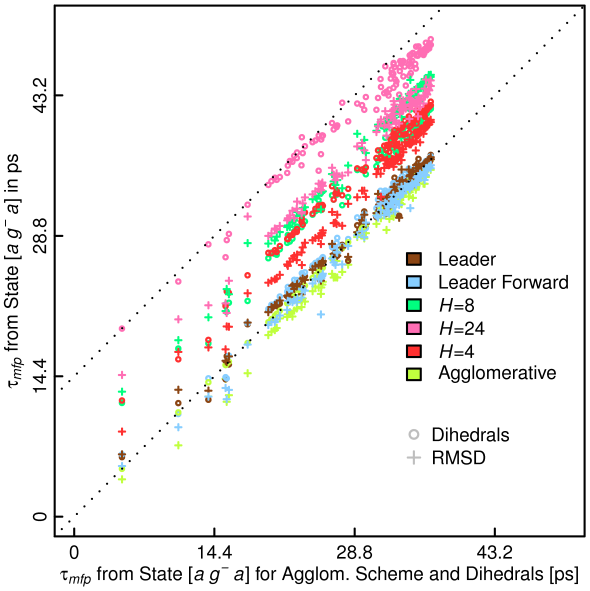

43.2
28.8
14.4
0
τ_{mfp} from State [$a\ g^-\ a$] in ps
0
14.4
28.8
43.2
τ_{mfp} from State [$a\ g^-\ a$] for Agglom. Scheme and Dihedrals [ps]
Leader
Leader Forward
H=8
H=24
H=4
Agglomerative
Dihedrals
RMSD

Figure S3:

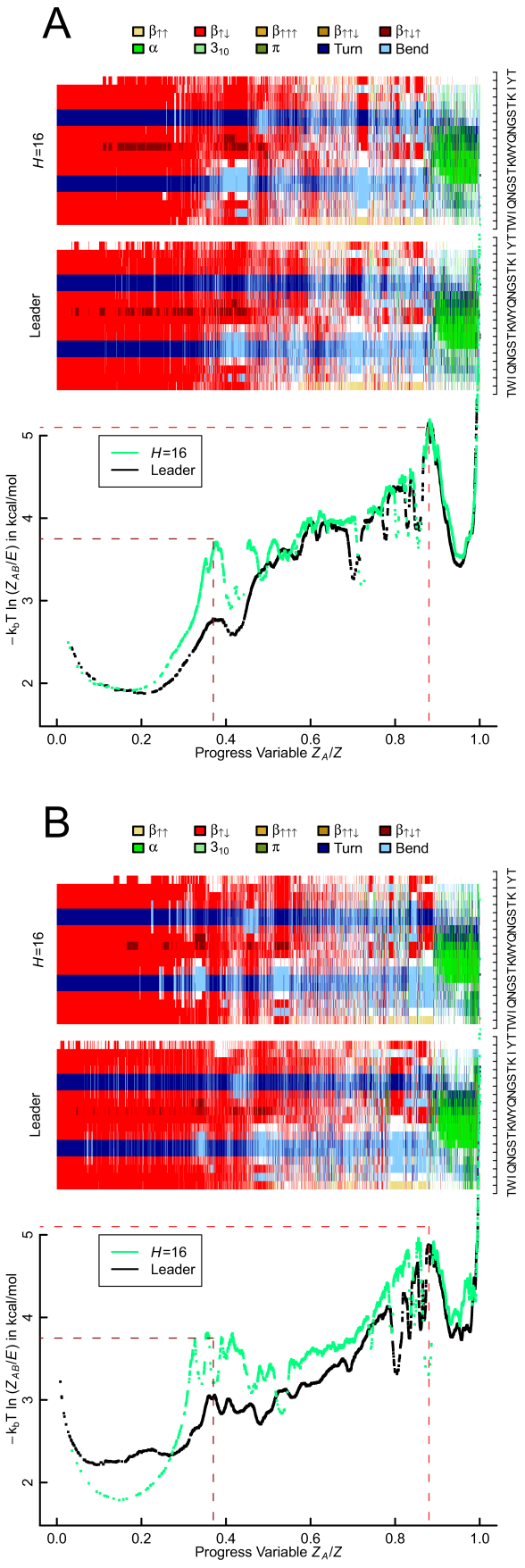

A
β↑↑
β↑↓
β↑↑↑
β↑↑↓
β↑↓↑
α
3₁₀
π
Turn
Bend
H=16
Leader
TWI QNGSTKWYQNGSTK I YTTWI QNGSTKWYQNGSTK I YT
H=16
Leader
−k_b T ln (Z_AB/E) in kcal/mol
Progress Variable Z_A/Z
B
β↑↑
β↑↓
β↑↑↑
β↑↑↓
β↑↓↑
α
3₁₀
π
Turn
Bend
H=16
Leader
TWI QNGSTKWYQNGSTK I YTTWI QNGSTKWYQNGSTK I YT
H=16
Leader
−k_b T ln (Z_AB/E) in kcal/mol
Progress Variable Z_A/Z

Figure S4:

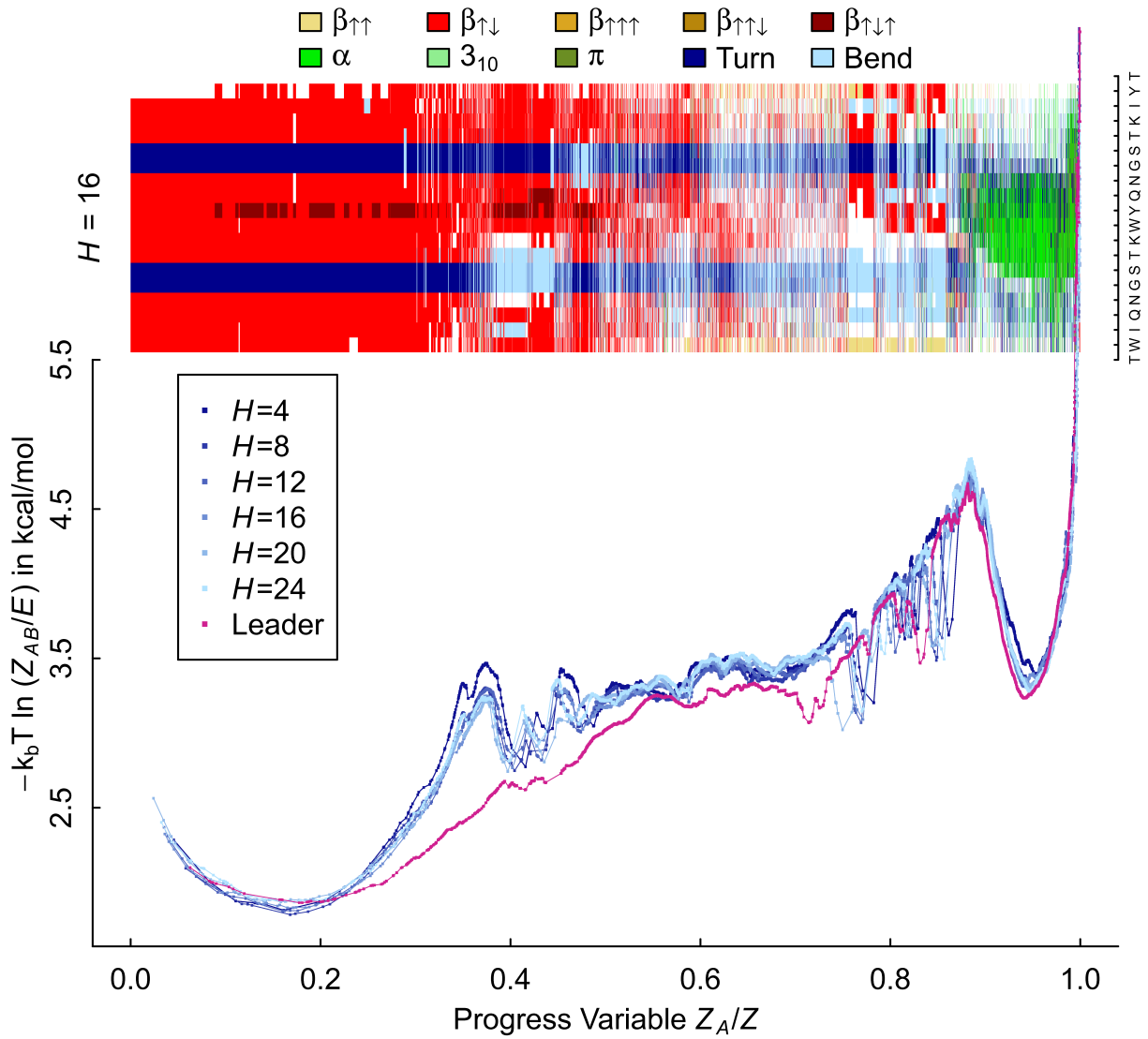
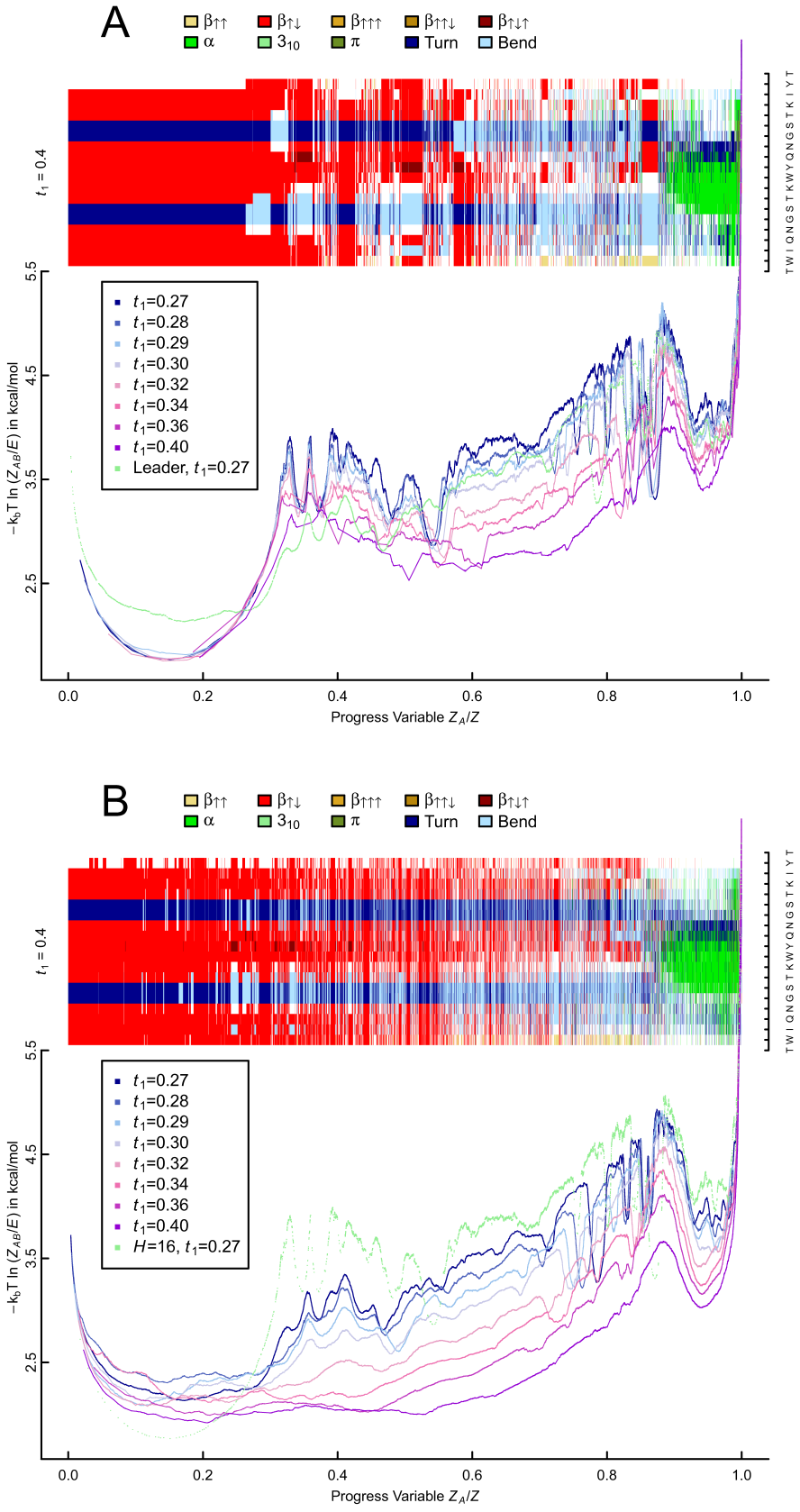

β↑↑
β↑↓
β↑↑↑
β↑↑↓
β↑↓↑
α
3₁₀
π
Turn
Bend
H = 16
T W I Q N G S T K W Y Q N G S T K I Y T
H=4
H=8
H=12
H=16
H=20
H=24
Leader
−k_bT ln (Z_AB/E) in kcal/mol
2.5
3.5
4.5
5.5
0.0
0.2
0.4
0.6
0.8
1.0
Progress Variable Z_A/Z

Figure S5:


A
β↑↑
β↑↓
β↑↑↑
β↑↑↓
β↑↓↑
α
3₁₀
π
Turn
Bend
t_1 = 0.4
T W I Q N G S T K W Y Q N G S T K I Y T
t_1=0.27
t_1=0.28
t_1=0.29
t_1=0.30
t_1=0.32
t_1=0.34
t_1=0.36
t_1=0.40
Leader, t_1=0.27
$-k_bT \ln (Z_{AB}/E)$ in kcal/mol
Progress Variable Z_A/Z
B
H=16, t_1=0.27