# Supporting information

## Comparison of clustering methods

**Figures S1-S5** show results from the analyses (Ramachandran plots, secondary structure WebLogos and dissimilarity RMSD matrices) performed on all five clustering methods we examined. Letters H, E and L in the WebLogos represent right-handed helices (α, π, $3_{10}$), extended conformations (β-parallel, β-antiparallel, $P_{II}$) and $\alpha_L$-helices respectively. Helices and extended conformations regions have been defined according to *Best et al.*[23]. The RMSD matrices provide a way to estimate the pairwise dissimilarity of peptides, both internally (within the same cluster) and between distinct clusters. RMSDs between structures that belong to same cluster correspond to the dark blue (low RMSD) squares lying on the diagonal of the matrix. RMSDs between structures that belong to different clusters correspond to the off-diagonal parallelograms. Ideally, a useful clustering method would segregate the structures in such a way as to simultaneously have (a) low RMSDs for all structures that belong to the same cluster, and, (b) high RMSDs (warm colors in these figures) for all pairs of structures that belong to different clusters. The following paragraphs summarize our findings for the five clustering methods.

*Cartesian clustering with automatic cluster number estimation **(Fig S1)***

The clusters produced by this method are quite heterogeneous with relatively high intra-cluster RMSDs. The Ramachandran plots show residues in three regions, a finding which indicates the presence of different transitions within the same cluster. This is in agreement with the WebLogo plots which clearly indicate the presence of a mixture of secondary structure assignments within the individual clusters (most easily seen in the case of the cluster marked as “1” in **Fig S1**).

*Dihedral clustering with automatic cluster number estimation **(Fig S2)***

This method produces only two clusters, and completely fails to differentiate between different transitions. The WebLogos are heterogenous, and the intra-cluster RMSDs are rather high (green and light blue on the RMSD colour scale), indicating high internal diversity in the clusters.

*Dihedral Principal Component Analysis (**Fig S3**)*

The results of the dPCA seem quite interesting and accurate on first sight, however, when studied further, some deficiencies are observed. Although β-sheets are grouped tightly, there are clusters that are almost identical to each other. Some examples are clusters 1 and 4 or clusters 2 and 3. Clusters 2 and 3 are not even distinguishable on the RMSD matrix. α-helices occupy two clusters, one unmixed, and one mixed with $\alpha$-$\alpha_L$-$\alpha$-$\alpha_L$-$\alpha$ transitions. Furthermore the method assigned only 10.101 structures out of the 12.545 total, which is a significant loss of information.

*Cartesian clustering with preset (1.59Å) RMSD cut-off (**Fig S4**)*

The results here are rather organized, in comparison with the previous three methods. The clusters appear to be compact and homogeneous, although some noise is observed, especially in clusters 3, 6 and to a lesser extent, cluster 2. The similarity comparison using the RMSD matrix shows that the clusters are structurally dissimilar, while the similarity is preserved inside every cluster. Two observations arise by these results: one is that $\alpha$-$\alpha_L$-$\alpha$-$\alpha_L$-$\alpha$ and $\beta$-$\alpha_L$-$\beta$-$\alpha_L$-$\beta$ transitions are revealed (although there is some mixture with other transitions); the second, is the differentiation of two β-sheet conformations in clusters 4 and 5, where a short-scale transition motif is identified, as shown in the Ramachandran plots. This was quite unexpected, and the distinction into two clusters may indicate two different transitions motifs between the parallel and anti-parallel β-sheet sub-regions. A similar situation is observed in cluster 3, which demonstrates a transition between the $\alpha$ and $3_{10}$ helix sub-regions.

*Dihedral clustering with a preset (1.44rad) RMSD cut-off (**Fig S5**)*

The final method we tried produced a very large number of clusters, many of whom are nearly identical to each other in structural terms. Although the Ramachandran plots indicate the presence of a detectable variation between these structurally similar clusters, the differences in terms of the actual atomic coordinates are rather minor and unconvincing.

The comparison of these results led us to choose the Cartesian clustering with a preset RMSD cutoff of 1.59Å since it produces clusters which are balanced both in terms of structural diversity and differentiation between distinct transitions in Ramachandran space.

**Figures S1-S5** show the results of the five clustering methods. The population of each cluster is shown in the parentheses. Letters H, E and L in the secondary structure WebLogos represent right-handed helices ($\alpha$, $\pi$, $3_{10}$), extended conformations ($\beta$-parallel, $\beta$-antiparallel, $P_{II}$) and $\alpha_L$-helices respectively, according to *Best et al.* assignments. Colder and warmer colours in the dissimilarity RMSD matrices indicate low and high RMSD values respectively.

**Figure S1 :** Cartesian clustering with automatic cluster number estimation

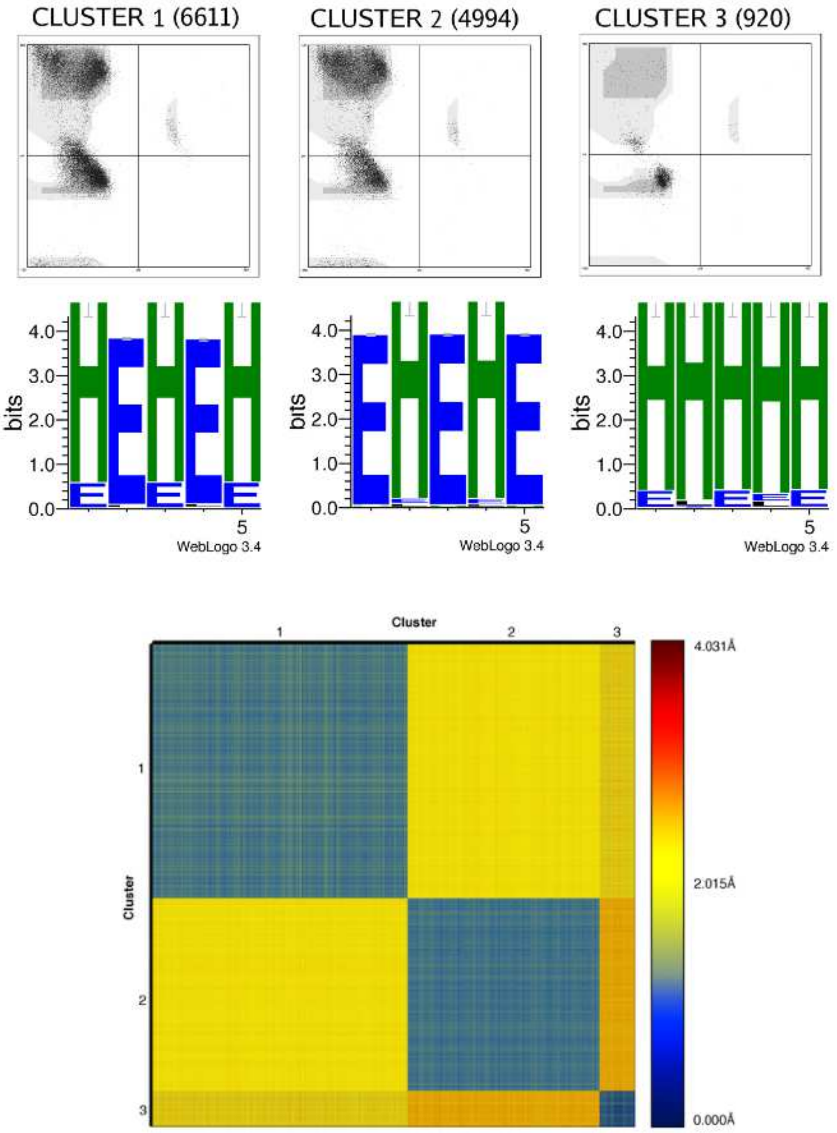


**Figure S2 :** Dihedral clustering with automatic cluster number estimation

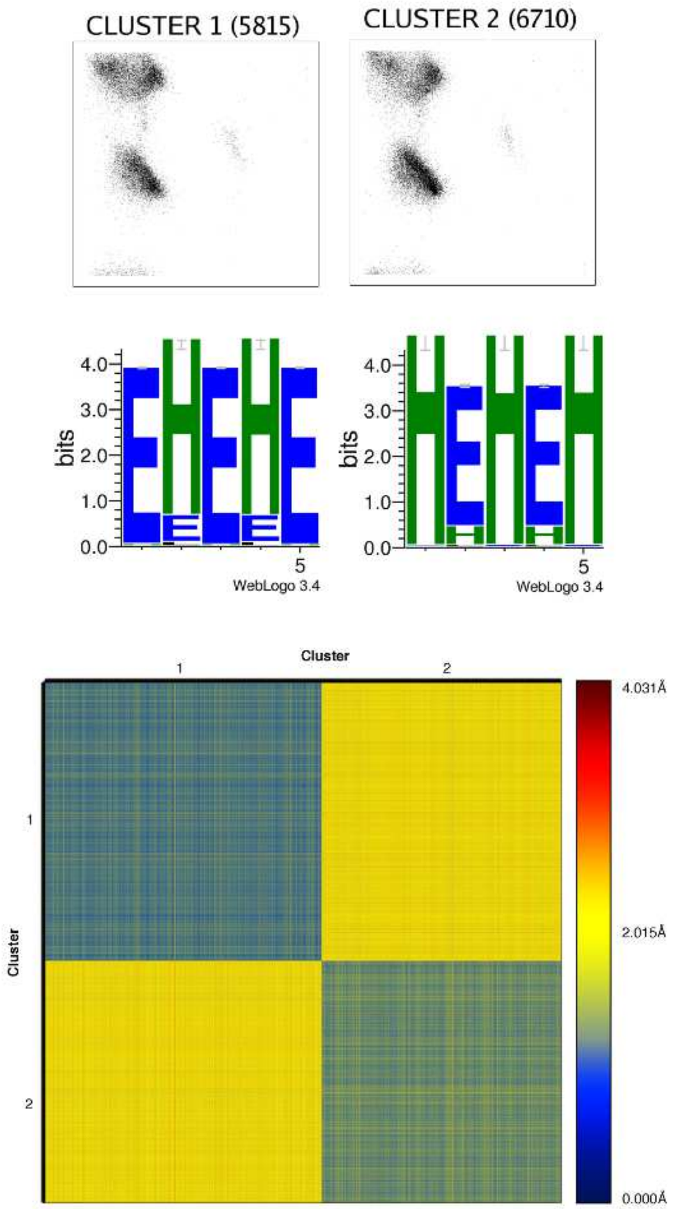


**Figure S3 :** Dihedral Principal Component Analysis

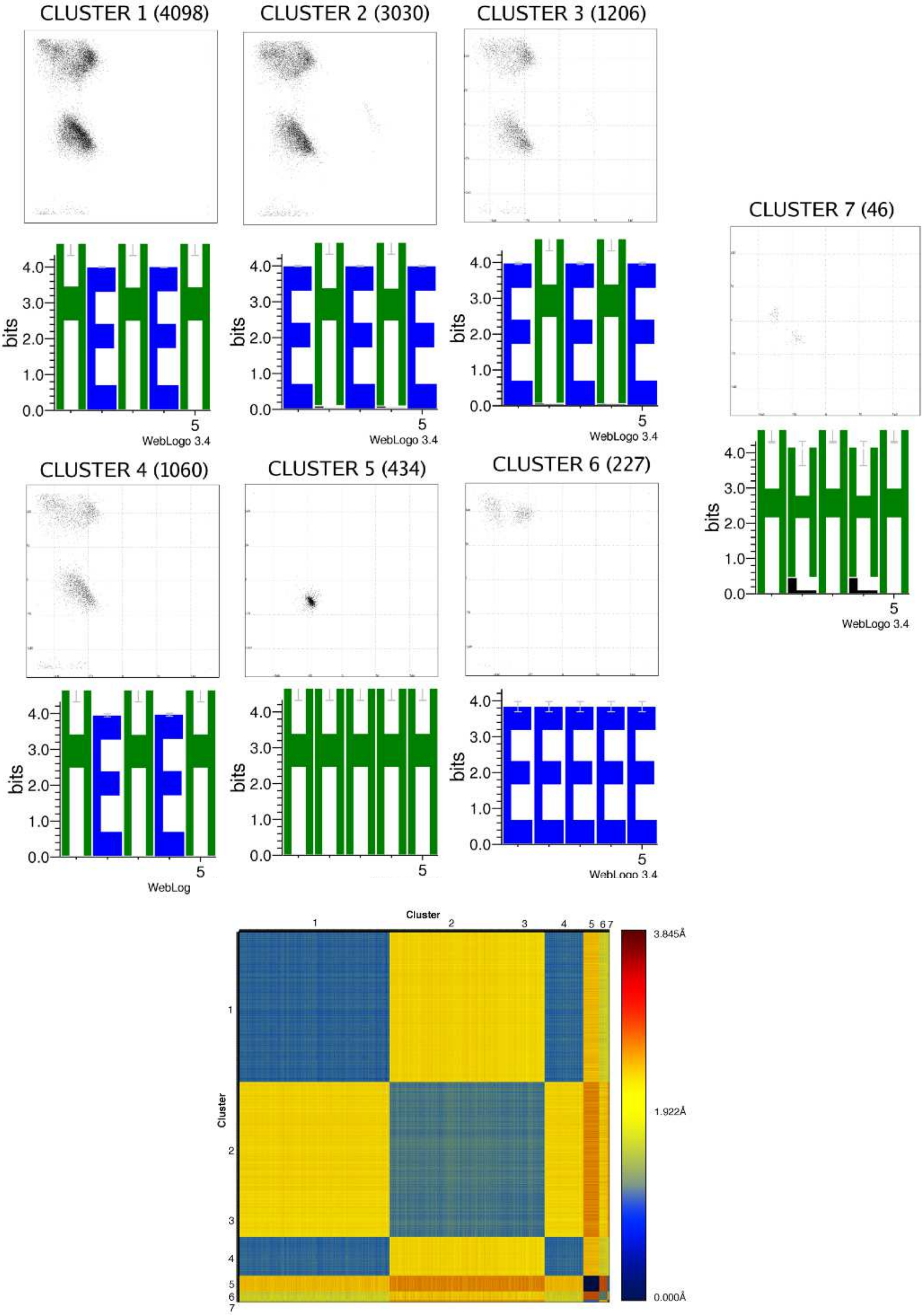


**Figure S4 :** Cartesian clustering with set RMSD cut-off  (1.593Å)

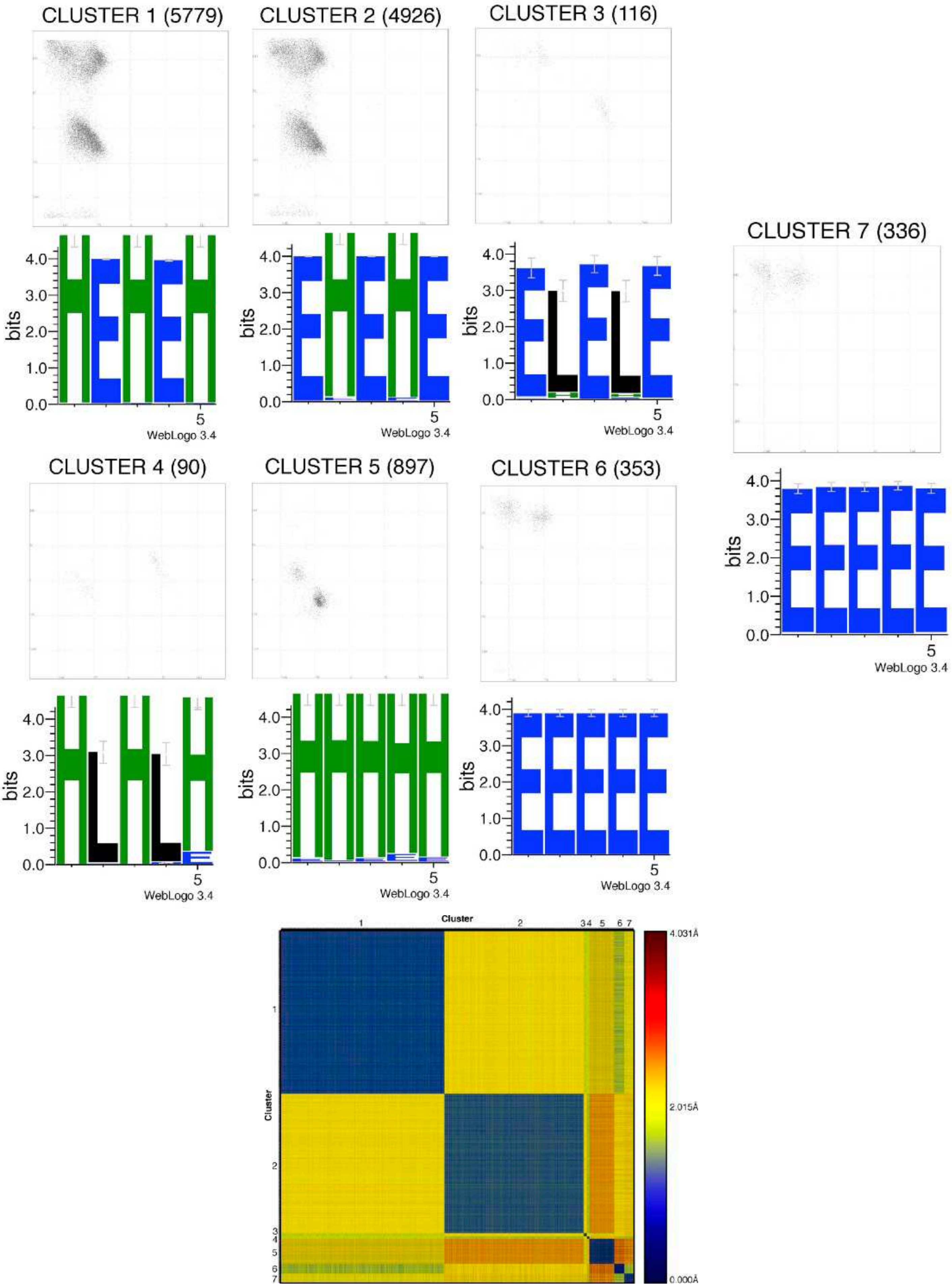
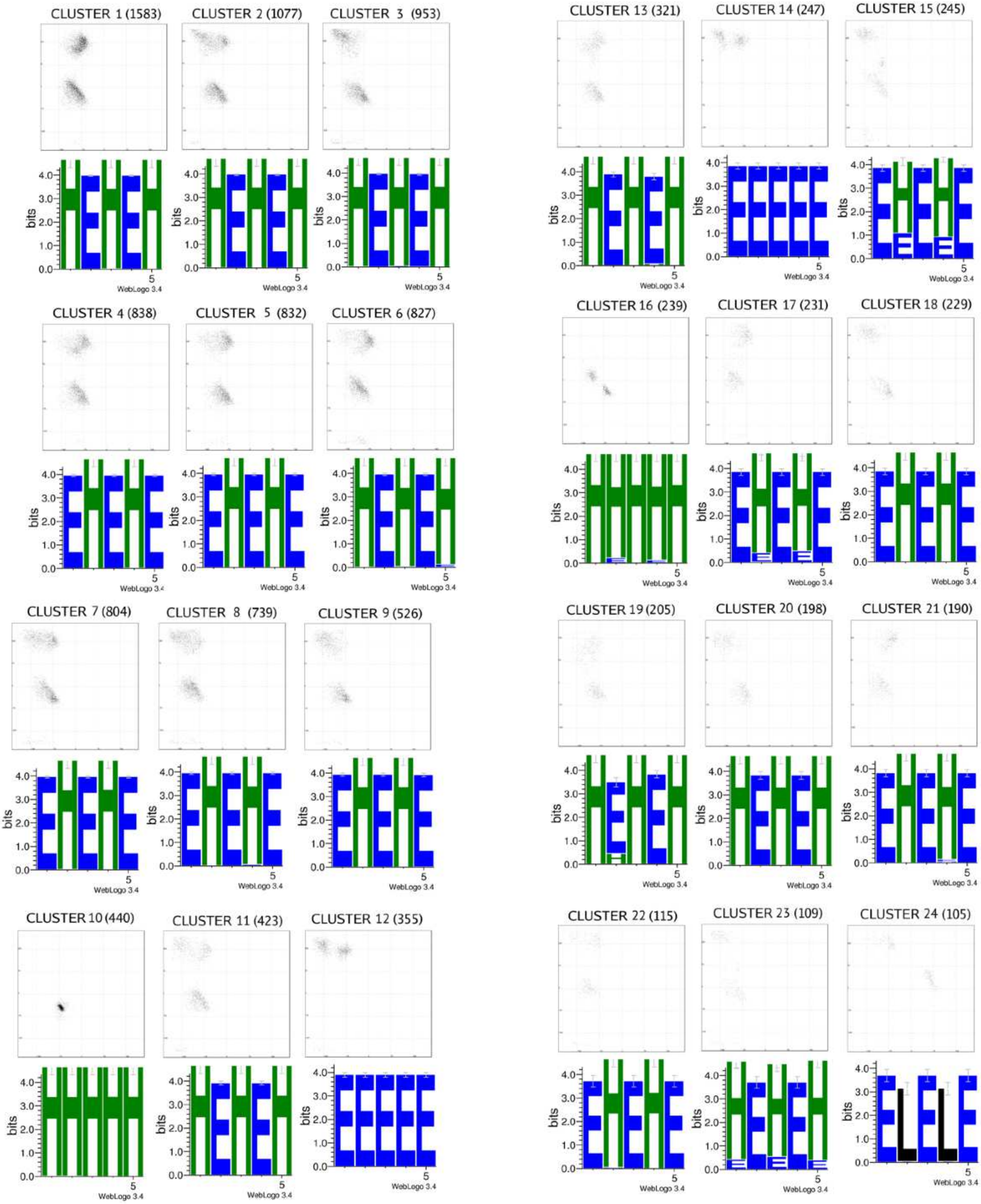
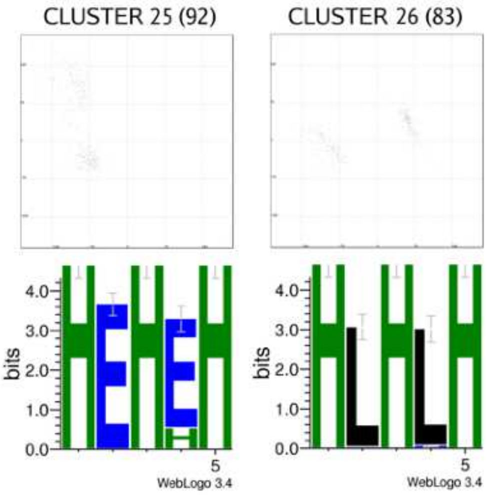


**Figure S5 :** Dihedral Clustering with set RMSD cut-off (1.44rad)




CLUSTER 25 (92)
CLUSTER 26 (83)
4.0
3.0
2.0
1.0
0.0
bits
5
WebLogo 3.4
4.0
3.0
2.0
1.0
0.0
bits
5
WebLogo 3.4

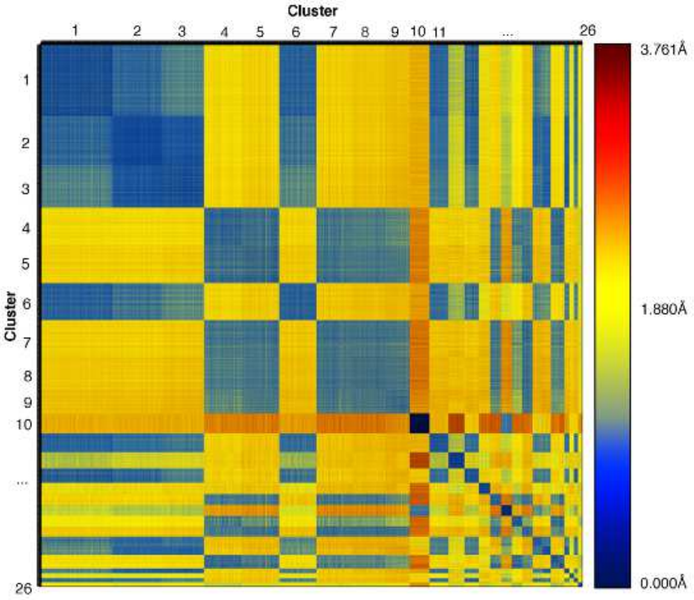
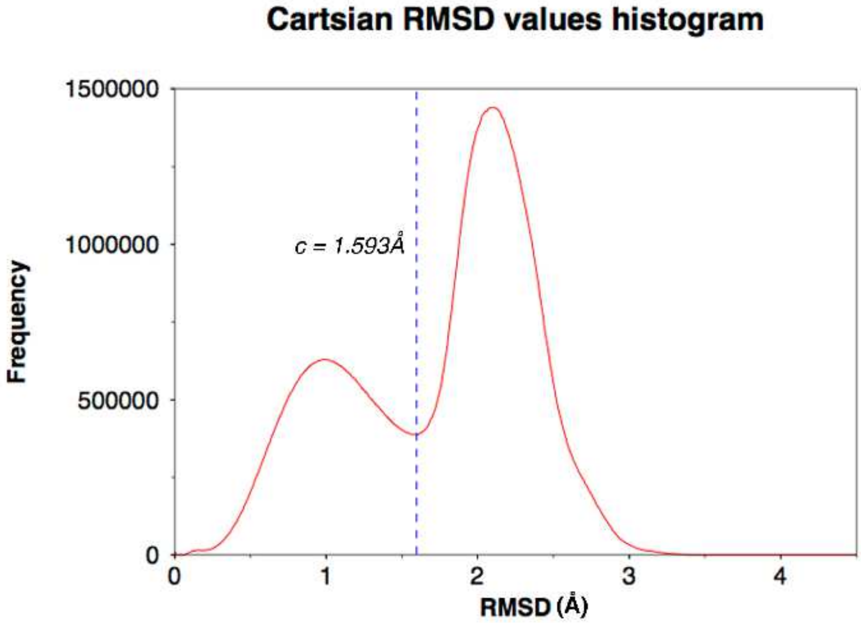
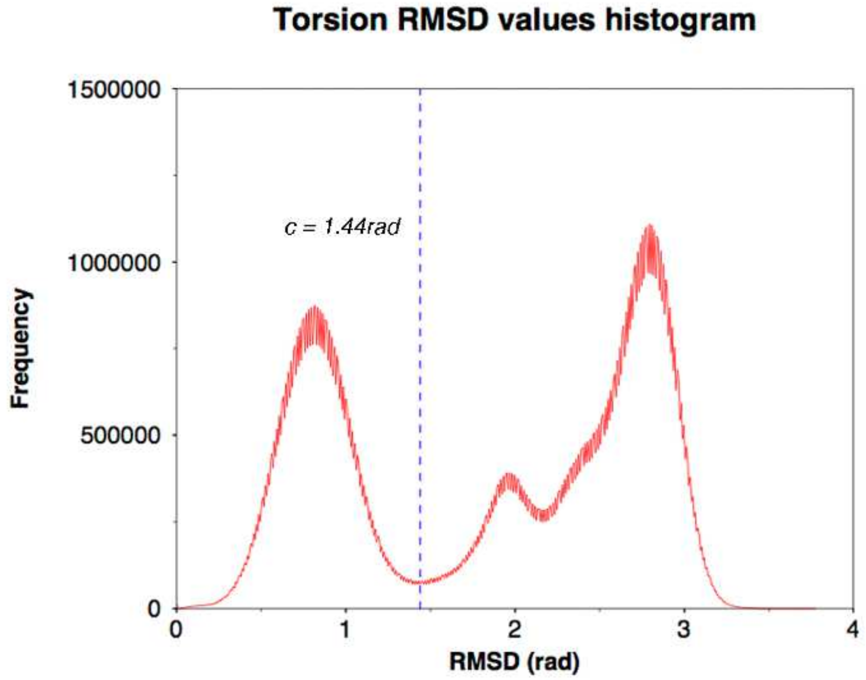

Cluster
1
2
3
4
5
6
7
8
9
10
11
...
26
Cluster
1
2
3
4
5
6
7
8
9
10
...
26
3.761Å
1.880Å
0.000Å

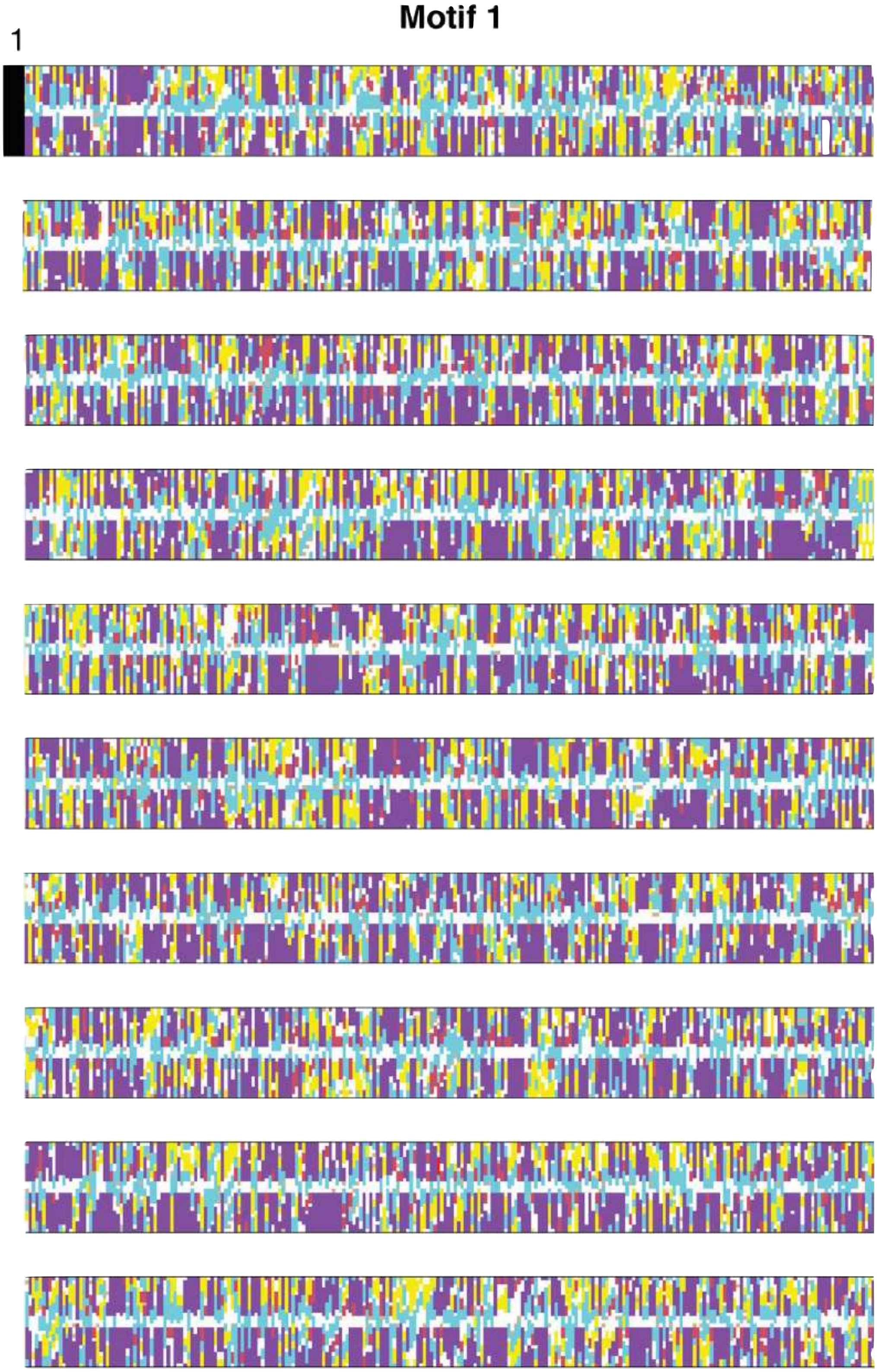
**Figure S6 :** The two graphs are the histograms of RMSD values in the non-automatic cartesian and dihedral clustering. The vertical blue dashed lines in each histogram indicate the local minima in the valleys between the major peaks. Structures with RMSD higher than these cut-offs are considered structurally similar.





**Figures S7-S10** are the per-sequence secondary structure diagrams of the context of each motif. Coloured vertical lines represent different secondary structure assignments, as produced by STRIDE (purple : α-helix, yellow : extended, magenta : $3_{10}$-helix, cyan : turn, white : coil). x axis represents the 25 residues (10 + 5 +10),  and y axis represents all the cluster members.

**Figure S7**:



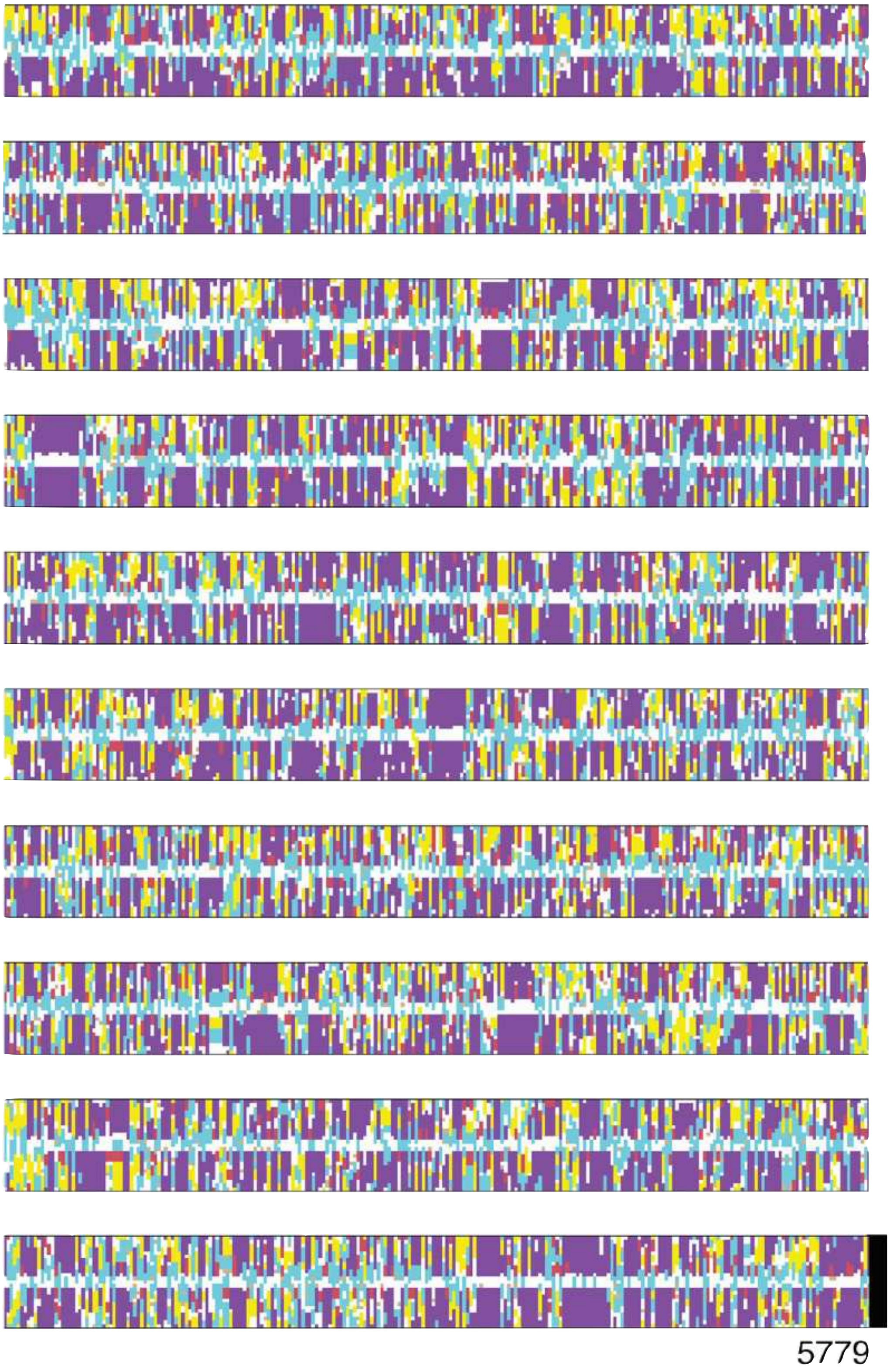
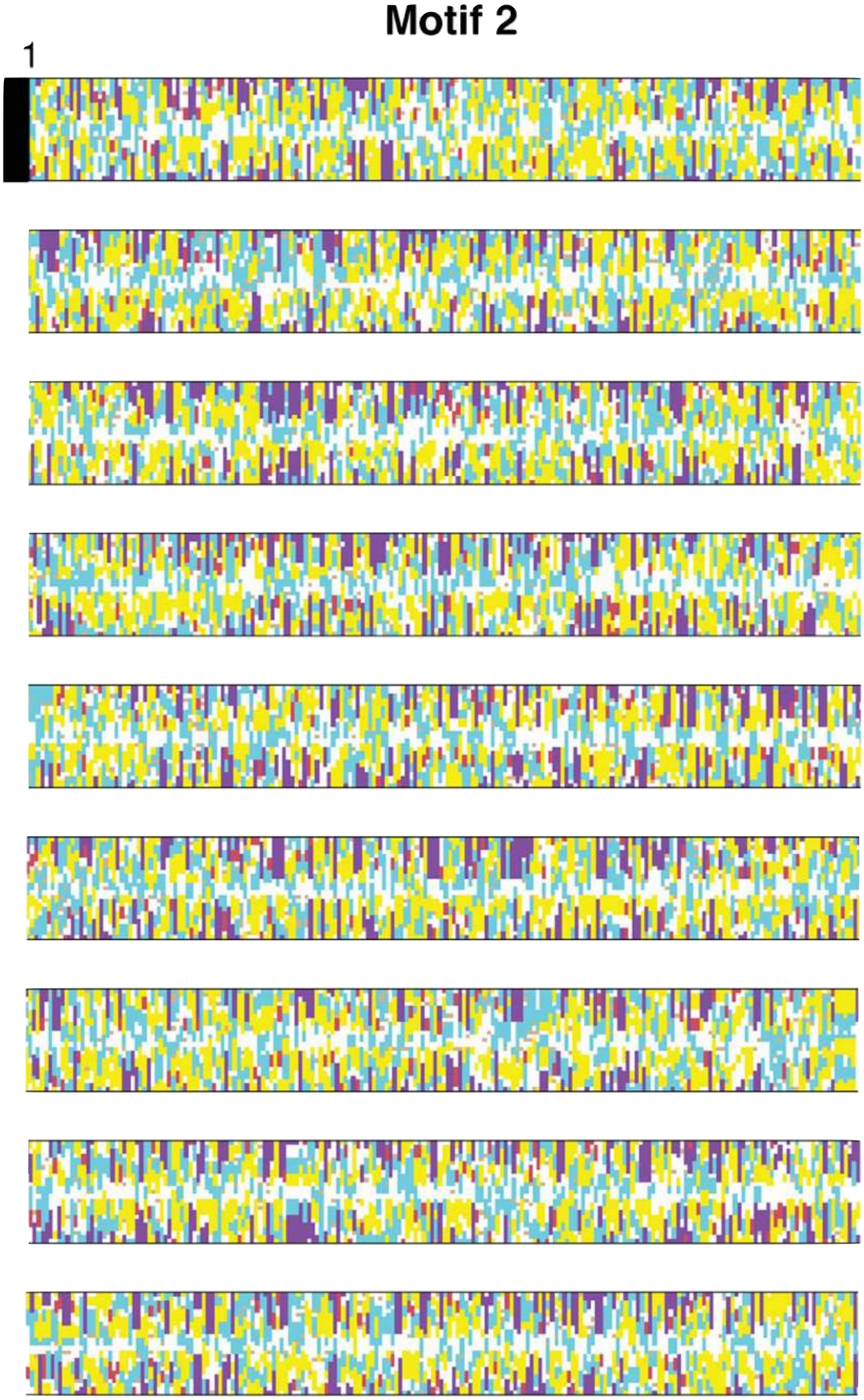

5779

**Figure S8:**



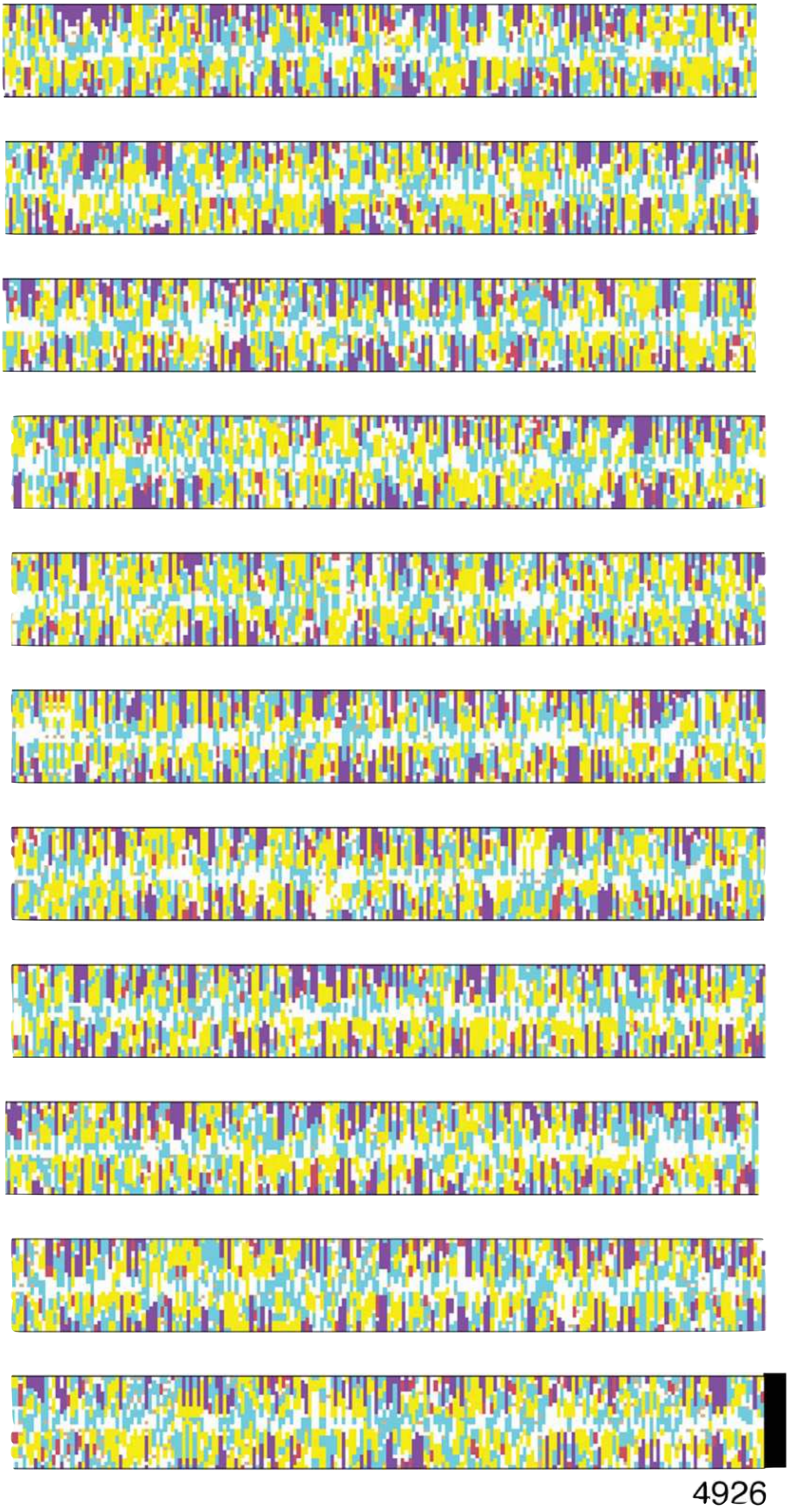

4926

**Figure S9**:

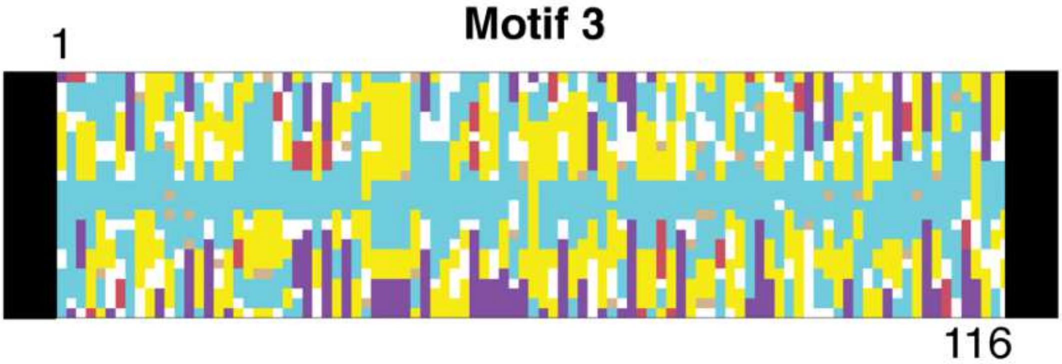


**Figure S10**:

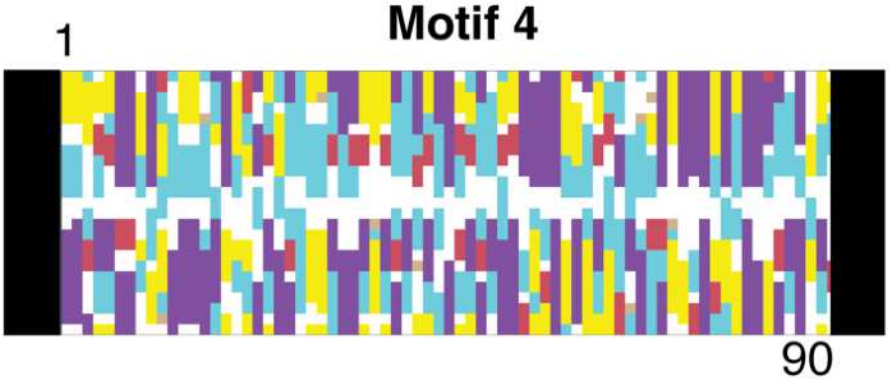


**Figure S11**: The four residue preferences (D,S,N,T) in position 4 of motif 1. These are stick representations of the backbone, with the side chains of the four conserved residues added in position 4 (green colour).

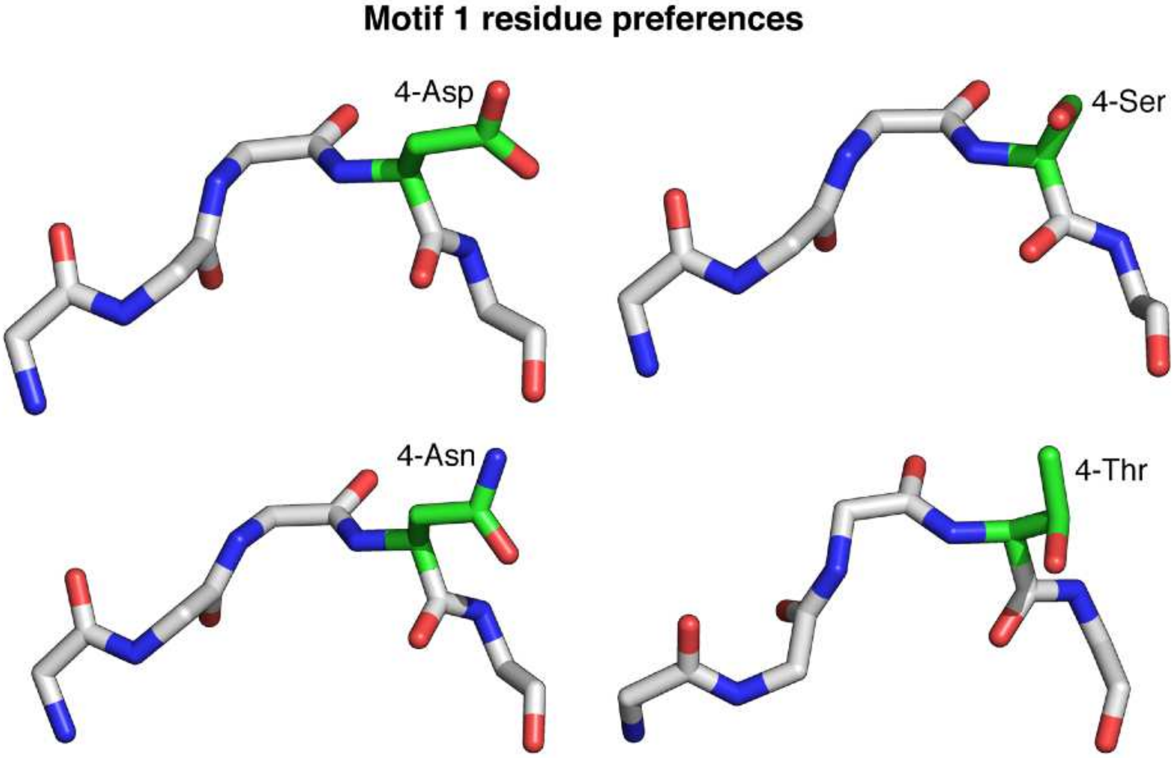
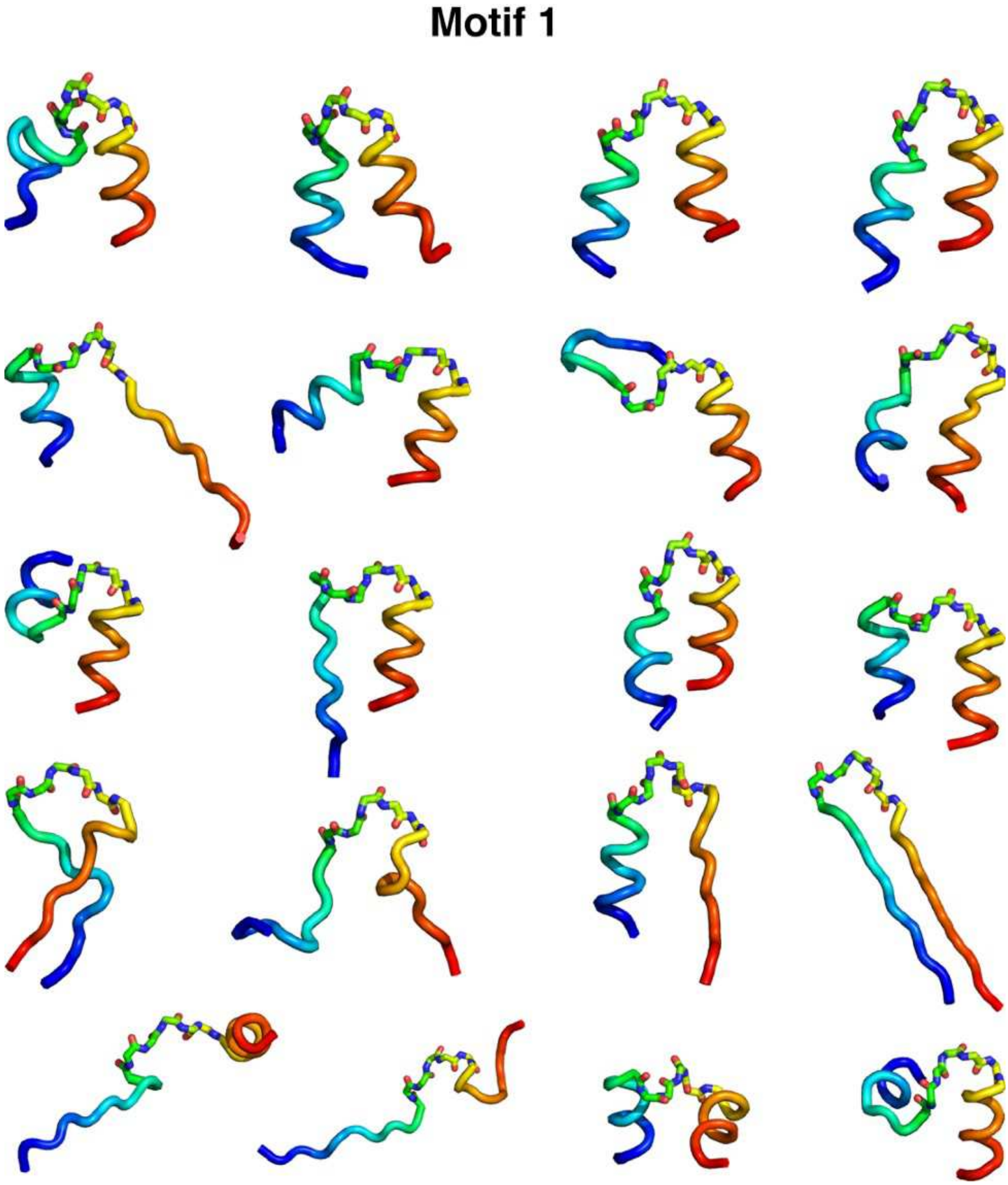


**Figure S12**: Structure collages of 20 selected members from clusters 1 and 2 (motif 1 and 2 respectively). Ten residues were added in the N- and C- termini to show the variability in observed secondary structure elements and their orientations, on each side of the peptides adopting the two motifs.



# Motif 2

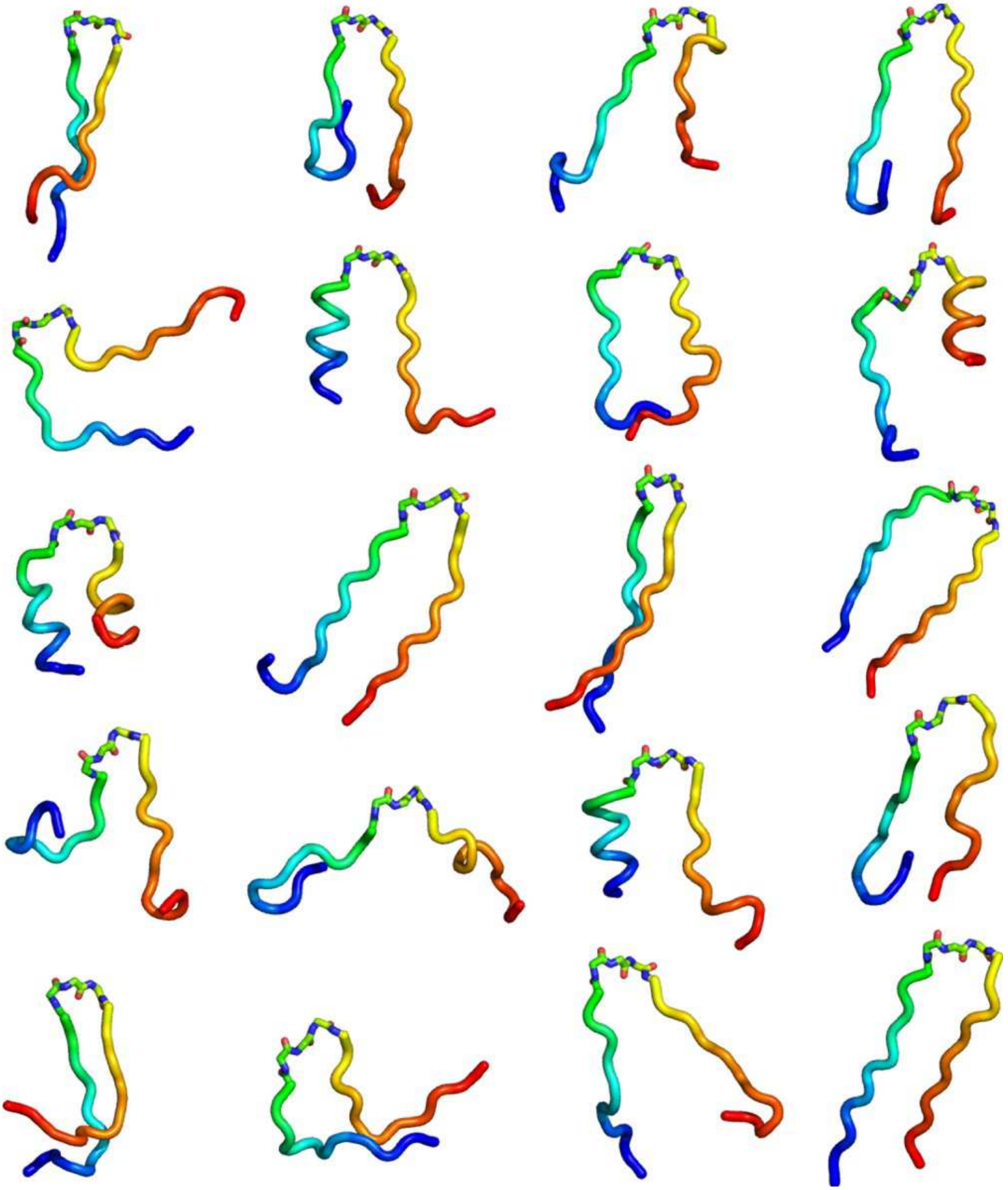
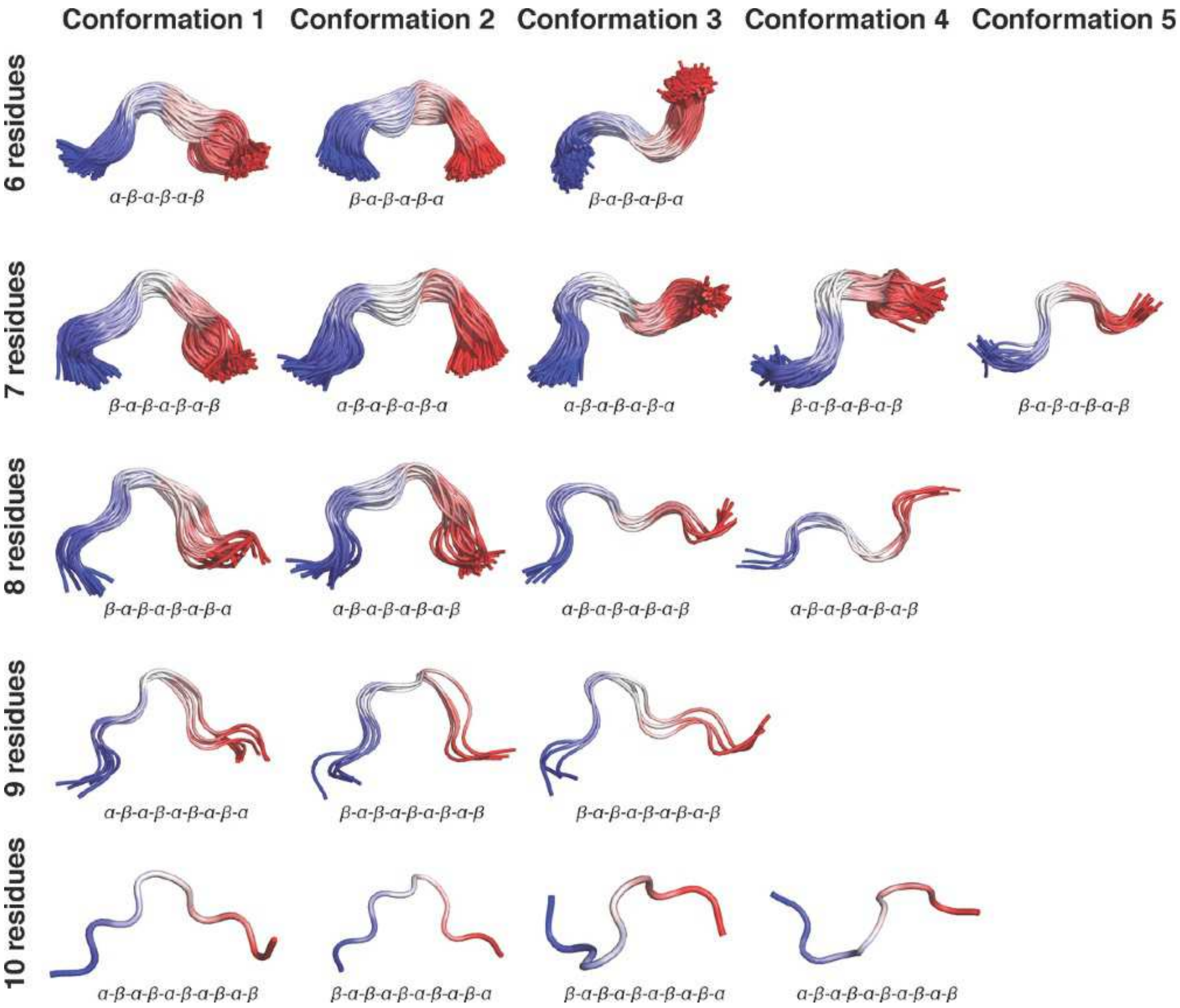


**Figure S13:** Dihedral angle transitions in peptides longer than five residues. Structures for 6,7,8 and 9 residue peptides are shown in superposition for each highly populated cluster. For 10-residue peptides, only some representative structures are show due to insufficient sample size for clustering.


Conformation 1
Conformation 2
Conformation 3
Conformation 4
Conformation 5
6 residues
α-β-α-β-α-β
β-α-β-α-β-α
β-α-β-α-β-α
7 residues
β-α-β-α-β-α-β
α-β-α-β-α-β-α
α-β-α-β-α-β-α
β-α-β-α-β-α-β
β-α-β-α-β-α-β
8 residues
β-α-β-α-β-α-β-α
α-β-α-β-α-β-α-β
α-β-α-β-α-β-α-β
α-β-α-β-α-β-α-β
9 residues
α-β-α-β-α-β-α-β-α
β-α-β-α-β-α-β-α-β
β-α-β-α-β-α-β-α-β
10 residues
α-β-α-β-α-β-α-β-α-β
β-α-β-α-β-α-β-α-β-α
β-α-β-α-β-α-β-α-β-α
α-β-α-β-α-β-α-β-α-β