Biophysical Journal, Volume *116*

Supplemental Information

Conformations of an RNA Helix-Junction-Helix Construct Revealed by SAXS Refinement of MD Simulations

Yen-Lin Chen, Tongsik Lee, Ron Elber, and Lois Pollack

Clustering HJH Geometries

It is essential to cluster the HJH conformations into groups with similar structures, as many of the MD generated conformations are highly degenerate. A typical EOM ensemble contains approximately 30,000 to 45,000 conformations depending on the number of cycles of genetic algorithm. It is naive to assume that the representative structures are those that are selected more than a certain number of times. Many of the selected conformers are degenerate and may, in combination, outnumber the above-threshold states and confound the results. An alternate approach is to cluster structurally similar conformations into separate groups. Of the many clustering algorithms that exist for biomolecules, some are RMSD-based[1–3] while others are based on overlapping volumes of two states[4]. These approaches are suitable to describe the relative positions of the two helices in the HJH molecule, but place little emphasis on the junction conformation. We seek a method that accounts for both helix position and junction geometry with similar weighting.

To cluster the structures in a way that considers both the helices and junction, we must first identify a way to classify the different models (in a data matrix M(0) where '0' indicates the step 0), which provides details about both, yet remains computationally reasonable. We considered several possible parameterizations. The four different parameterizations we considered are in Fig. S1. The "backbone" matrix contains the coordinates of all the phosphorus atoms in the backbone and contributes 159 features to M(0). The "backbone+5U bases" matrix contains all the backbone data as well as the coordinates of 5 uracil rings in the junction: 90 extra features. The "Backbone+5U geometry" matrix contains the backbone data, center of mass and the normal vector of the uracil rings. Due to different types of data in the matrix, each feature is normalized before clustering:

$$f'_i = \frac{f_i - \langle f_i \rangle}{\sigma(f_i)} \quad \#(S1)$$

The f'_i and f_i are the normalized and raw data of the i[th] features and $\langle f_i \rangle$ and $\sigma(f_i)$ denote the mean and standard deviation of the raw data respectively. Finally, the "All Atom" matrix is the normal all-atom RMSD method where 5061 features are clustered. The RMSD values increase as the clustering algorithm proceeds and the later clusters (n > 0.8K) usually have worse mutual structural similarity within one cluster.

For the "Backbone" and "All Atom" matrix, the junction geometry features are overshadowed by the two helices therefore the results fail to take the junction base arrangement into account. In addition, the "All Atom" matrix is computationally expensive to cluster. Similarly, the "Backbone+5U geometry" also loses track of the junction because the geometry information is scarce compared to the backbone of the duplexes. The best parameterization appears to be the "Backbone+5U bases" matrix which balances the contribution of the helices and the junction geometry by adding all the atoms in the uracil rings. It also does not take too much computational time.

To carry out the clustering using the "Backbone+5U bases" set, we extracted the geometries of the backbones of H1 and H2 as 48 sets of coordinates. We also extracted the backbone and base coordinates of the nucleotides in the junction, as sets of ring (6) and backbone (1) coordinates of the 5 uracils. These numbers, 144 (48×3) and 105 ($7 \times 5 \times 3$) features for the helices and junction respectively, are placed in M(0). We apply the following algorithm using the K-means clustering[5] of the built-in MATLAB function, *kmeans*, on M(0). There is also no restriction on the number of clusters K. The procedure is as follows:

1. Run K-means on M(0) of K clusters for all the conformations with 249 features.

2. Find the best cluster B(1) among K clusters based on the RMSD value of 249 features.
3. Remove elements in B(1) from M(0) as M(1).
4. Repeat step 1 through 3 using K-i clusters on M(i) for K times.

In contrast to the standard K-means clustering, the matrix M(0) doesn't need to be normalized to avoid possible error-prone weighting because each feature in M(0) is a 3D coordinate. Finally, we end up with K clusters. In general, the structures in later clusters have less mutual similarity than those in the earlier ones. The last (K^{th}) cluster only contains the residues of the previous steps. The reason for the recursive K-means is to improve the accuracy of the clustering while trading off computation time. The mutual similarity within one cluster is good until the n^{th} cluster, where $n \sim 0.8K$. With clustering, we combine all the degenerate states and their corresponding frequencies for further analysis.

The structures of selected clusters are shown in Fig. S2. The earlier clusters have good similarity until cluster #80 beyond which the clusters look more chaotic and cannot be considered as a single representative structure. The last cluster, cluster #100, contains the leftover structures.

Parameters for GAJOE

We used several sets of parameters in Table. S1. for the genetic algorithm program, GAJOE, to match the size of the pool.

	Subpool – 200ns	Subpool – 600ns	All-salt pool
# of structures in the pool	~ 2,980*	~ 13,000	22,540
# of generations	1,000	5,000	10,000
# of ensembles	50	50	50
Ensemble size fixed?	no	no	no
Max # of curves	20	20	20
Min # of curves	5	5	5
Curve repetition	yes	yes	yes
Constant subtraction	yes	yes	yes
# of genetic algorithms	150	150	150

Table S1. GAJOE parameters used for different pools. *The subpool for [KCl] = 500mM is also applied by the same set of parameter despite of fewer structures.

EOM Fit of All [KCl]

The EOM fitting to the experimental SAXS profiles with χ^2 listed is shown in Fig. S3. At high [KCl] = 500mM, the fitting is improved by the inclusion of extended models generated in lower [KCl] simulations.

End-to-end Distance Distribution

The distributions of end-to-end distance (d) of the pool (red), "all-cycle" ensemble (blue) and "best-cycle" ensemble (green) are shown in Fig. S4. The end-to-end distance of HJH is calculated using the separation of the phosphorus atoms at the 5' and 3' end. The bimodal distribution also appears at medium salt where [KCl] = 100mM, with d ~ 60 and 80□. These two states split further as [KCl] increases and the base-stacking comes into play at high salt, resulting in the extended HJH conformation with d = 92□. The subpools at [KCl] = 500mM are also missing the extended conformations as seen in the Rg distribution.

Different Thresholding Values

The HJH solution ensemble with different threshold values and junction conformations are shown in Fig. S5 – Fig. S9 for [KCl] series. The thresholds are set to be the mean ± one standard deviation.

Supporting References

1. Betancourt, M. R. & Skolnick, J. Universal similarity measure for comparing protein structures. *Biopolymers* **59,** 305–309 (2001).

2. Zemla, A. LGA: A method for finding 3D similarities in protein structures. *Nucleic Acids Res.* **31,** 3370–3374 (2003).

3. Petoukhov, M. V. *et al.* New developments in the ATSAS program package for small-angle scattering data analysis. *J. Appl. Crystallogr.* **45,** 342–350 (2012).

4. Rodrigues, J. P. G. L. M. *et al.* Clustering biomolecular complexes by residue contacts similarity. *Proteins Struct. Funct. Bioinforma.* **80,** 1810–1817 (2012).

5. MacQueen, J. *Some methods for classification and analysis of multivariate observations. In Fifth Berkeley Symposium on Mathematical Statistics and Probability.* (University of California Press, Berkeley, CA. 666, 1967).

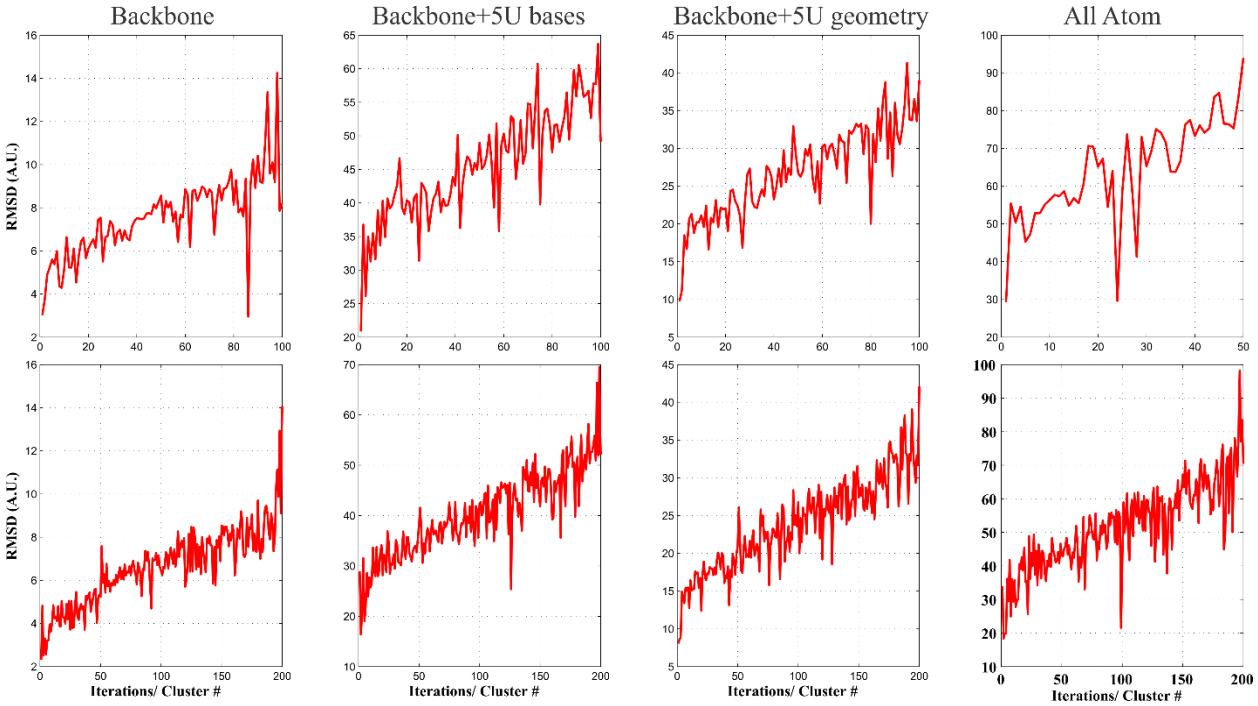

Figure S1. The best-RMSD traces of clustering using different data matrix M for different K value of 100 and 200. Backbone: coordinates of all phosphorus atoms in HJH, features = 159; Backbone+5U bases: backbone coordinates and the 6 atoms of uracil rings, features = 249; Backbone+5U geometry: backbone coordinates and the COM and normal vectors of uracil rings, features = 189; All atom: all atoms of HJH, features = 5,061.

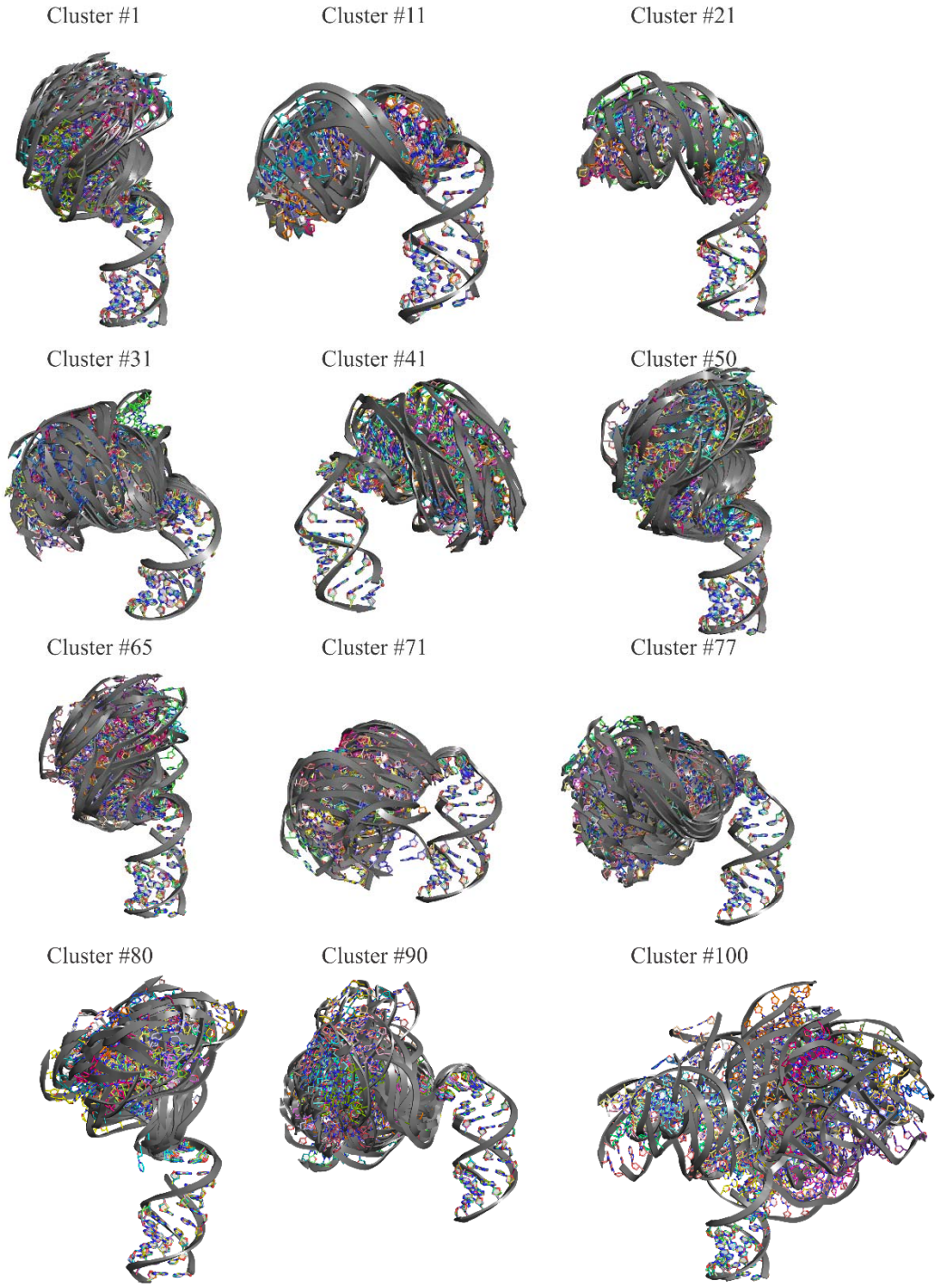

Figure S2. The structures in one cluster using the "Backbone+5U bases" and K=100. The performance is good until cluster # n, where n~0.8K=80 in this case. For cluster #80, the best-RMSD increases to more than twice of cluster #1. Prior to cluster #80, the structures within certain cluster are similar to each other. Notice that the use of atoms in the uracil rings accounts for junction geometries.

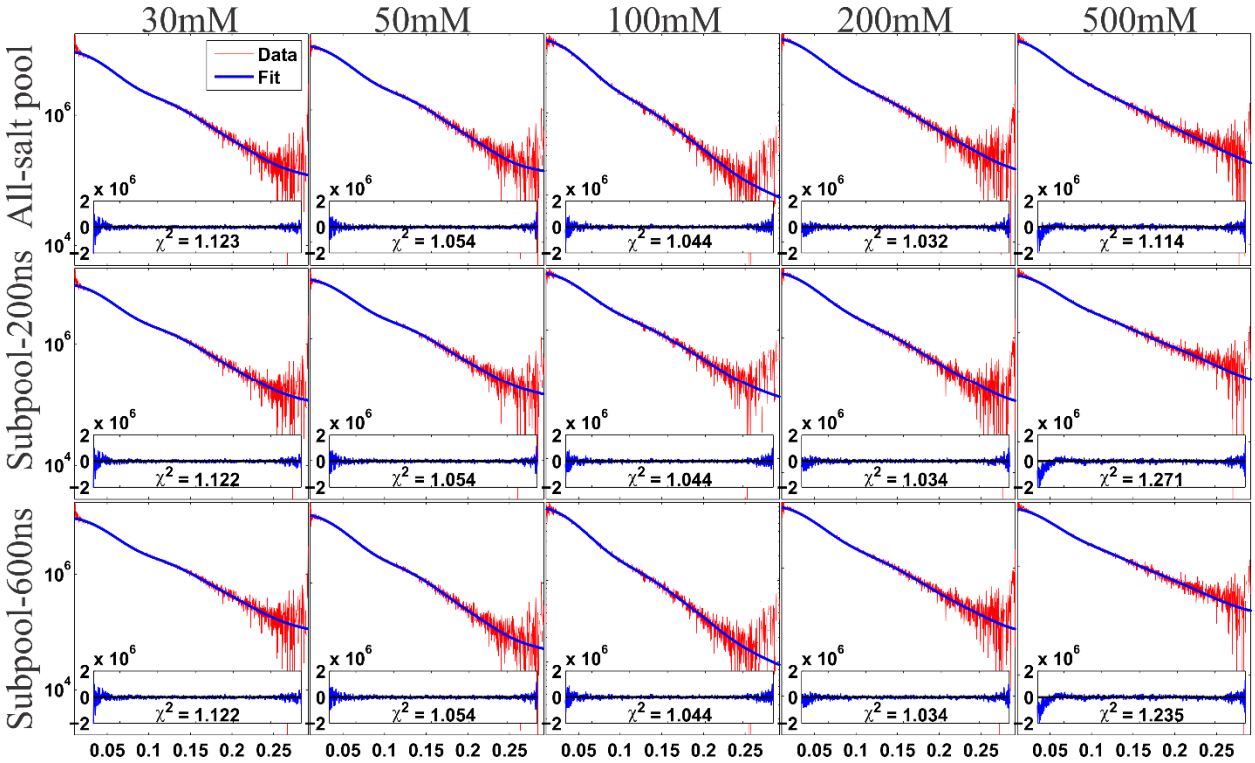

Figure S3. The EOM fitting to all the SAXS profiles using different search pools.

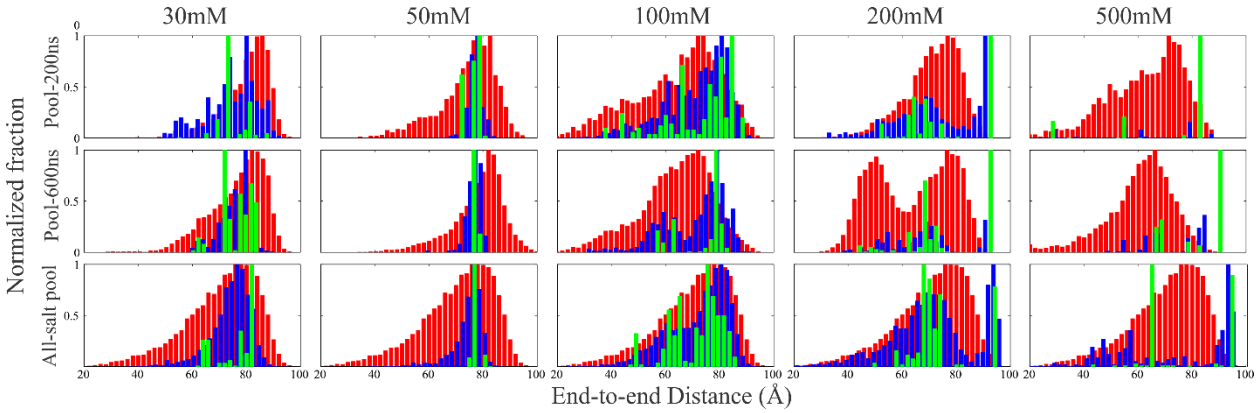

Figure S4. Distribution of end-to-end distance at different [KCl] using different search pools. The red bars show the distribution of the pool while the blue and green ones are the distributions from the "all-cycle" and "best-cycle" analysis respectively. Each distribution is normalized by the maximum number of counts for presentation purposes.

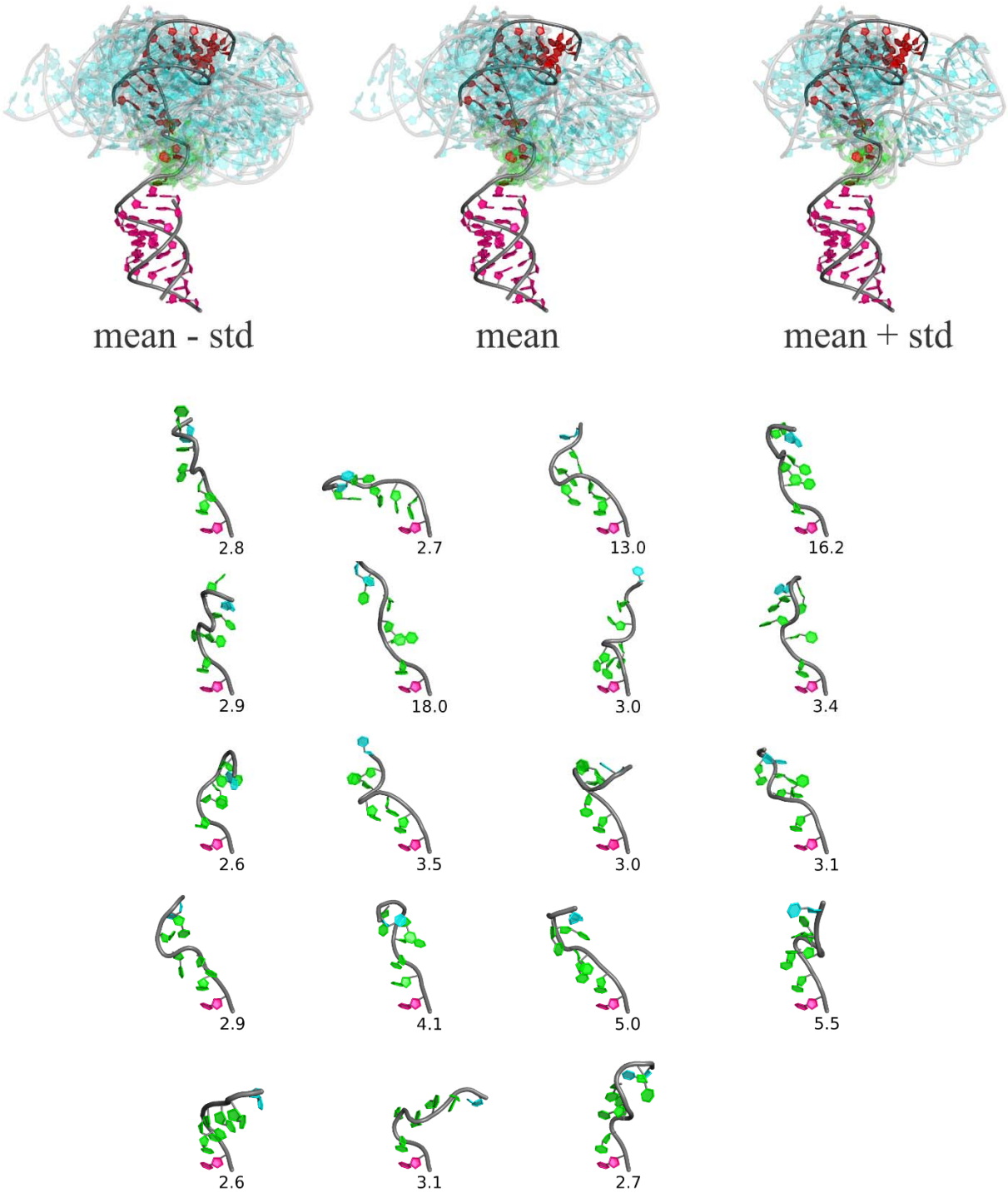

Figure S5. The ensemble and junction conformations at [KCl] = 30mM.

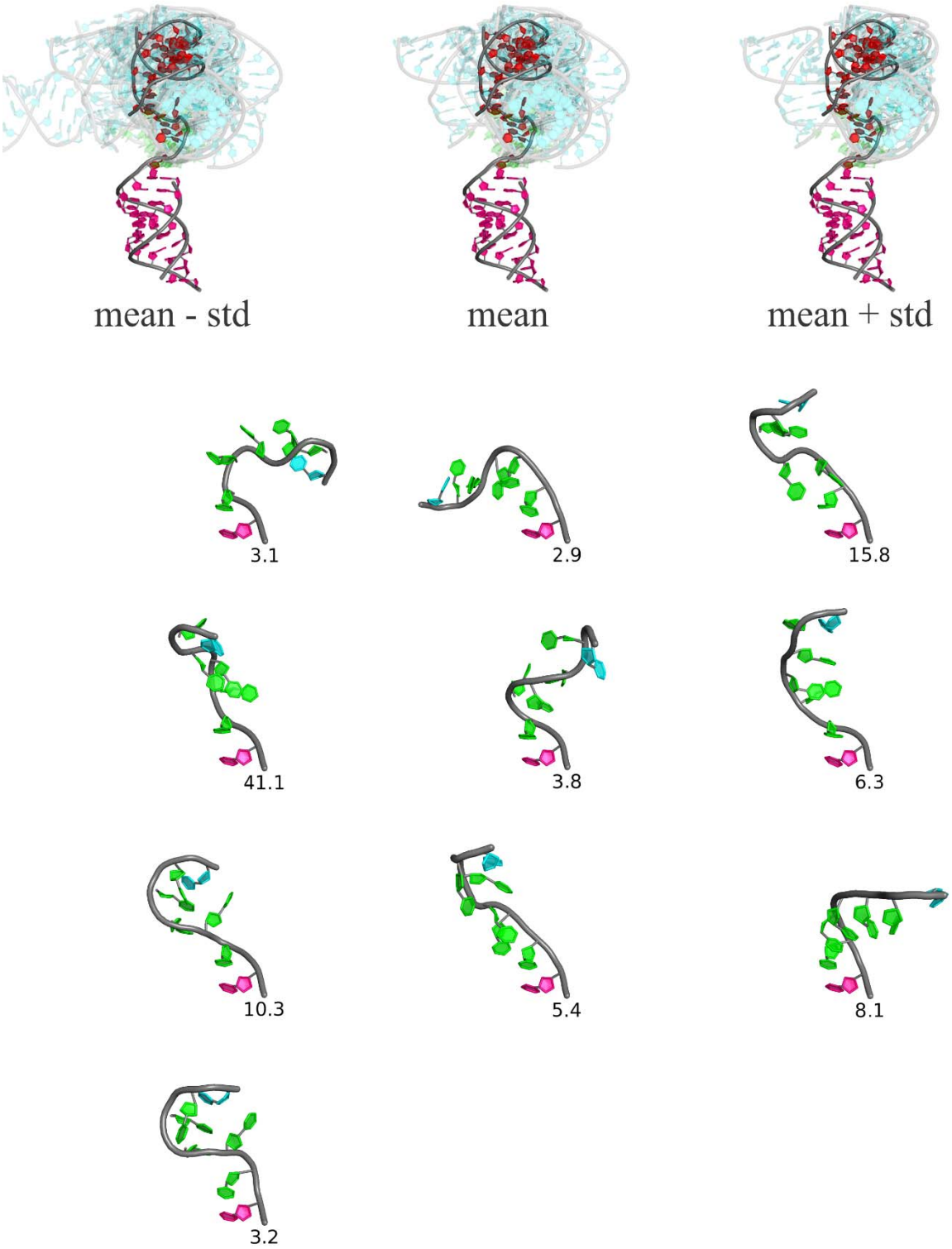

Figure S6. The ensemble and junction conformation at [KCl] = 50mM.

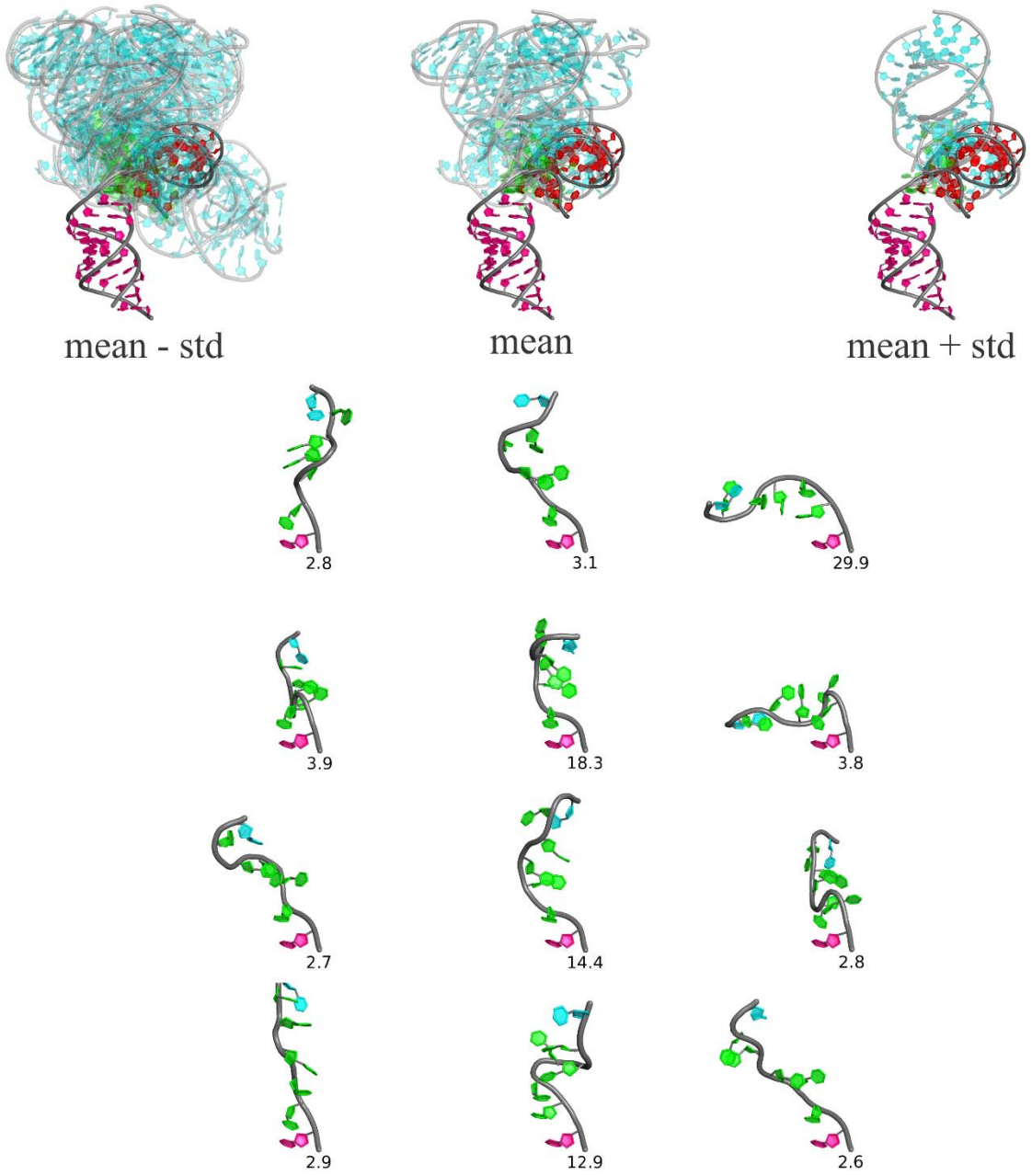

Figure S7. The ensemble and junction conformation at [KCl] = 100mM.

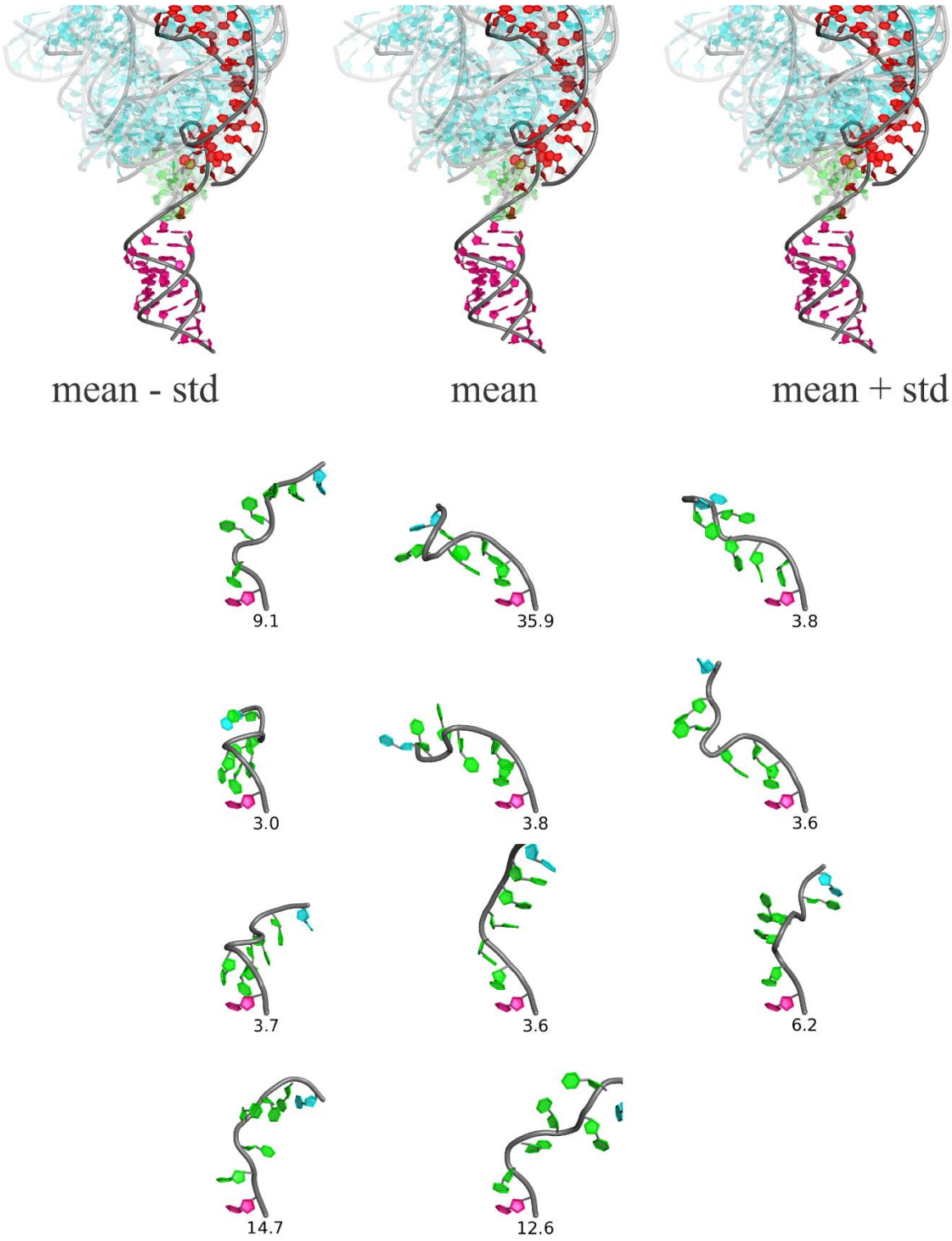

Figure S8. The ensemble and junction conformation at [KCl] = 200mM.

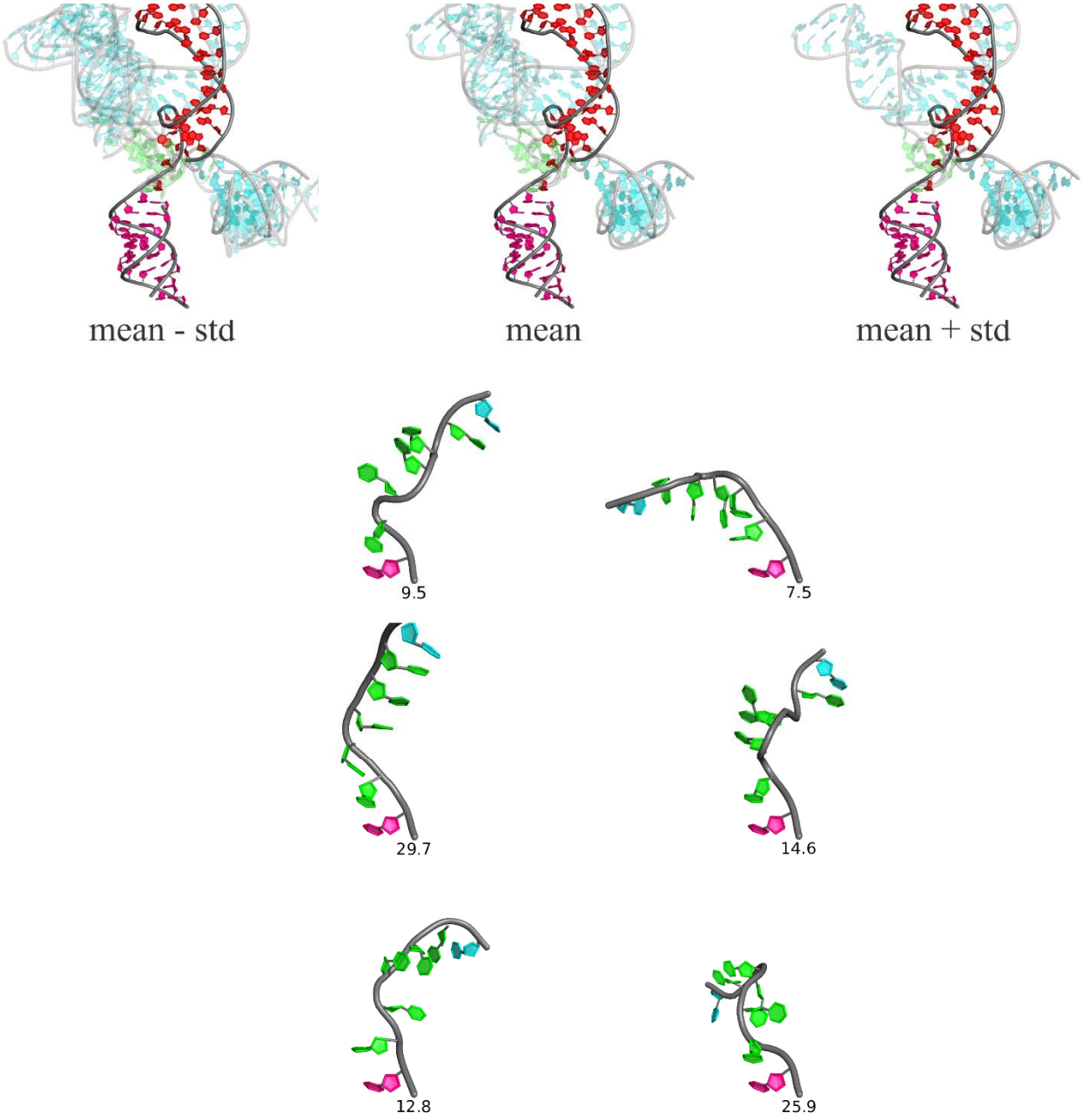

Figure S9. The ensemble and junction conformation at [KCl] = 500mM.