Multidimensional Replica Exchange Molecular Dynamics Yields a Converged Ensemble of an RNA Tetranucleotide

AUTHORS:CHRISTINA BERGONZO, NIEL M. HENRIKSEN, DANIEL R. ROE, JASON M. SWAILS, ADRIAN E. ROITBERG, AND THOMAS E. CHEATHAM, III.

PRESENTED BY HARRY MUTH

Overview

Introduction

Methods

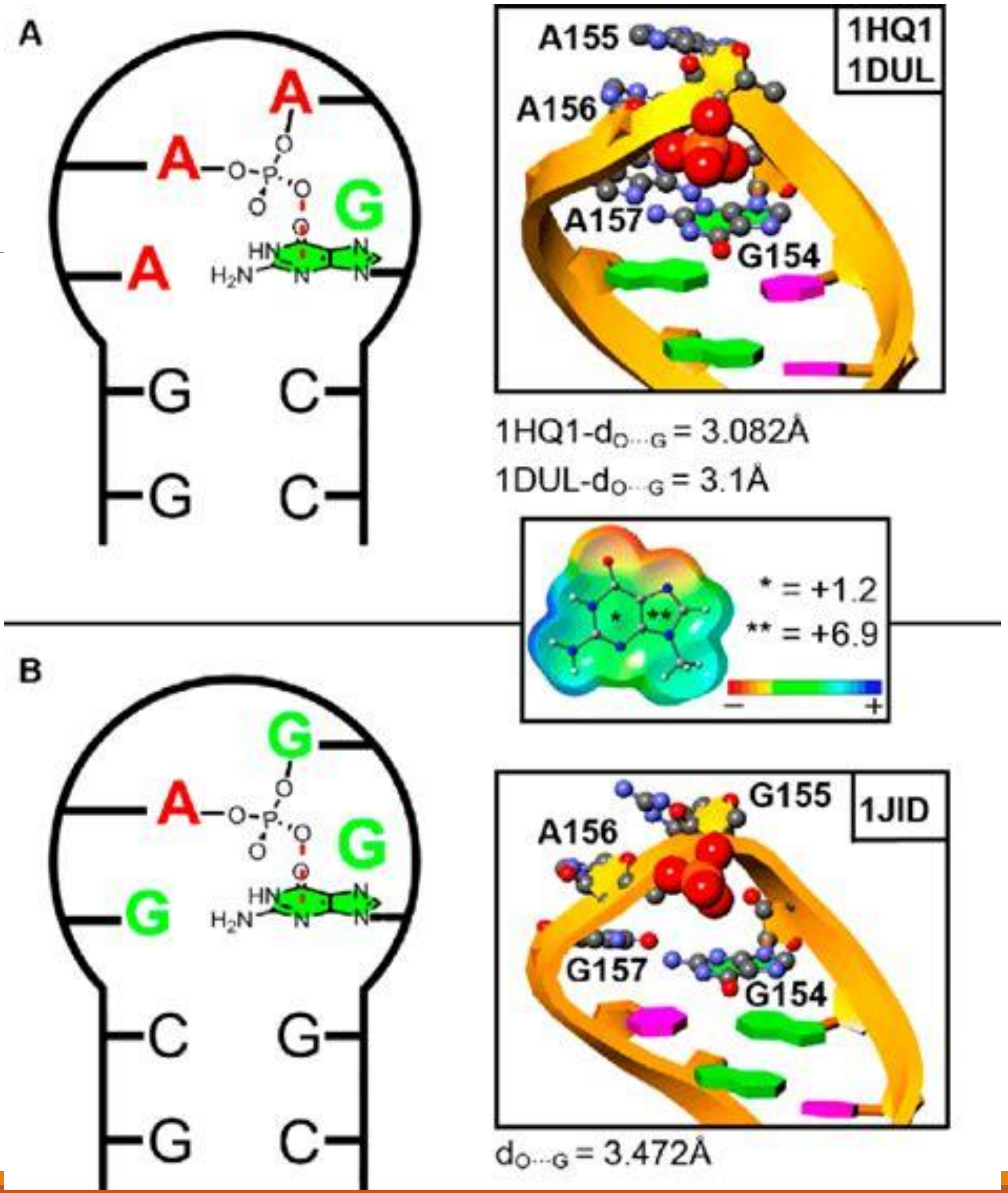
Results (with Methods along the way)

Conclusions

Questions

Introduction

- The role of RNA in a range of biological processes has developed over recent years. Rather than being just restricted to protein synthesis (mRNA and tRNA), RNA has been shown to have multiple other functions.
- Critical to its function in these roles is its structure and due to the dynamic nature of RNA, interchange between multiple conformations.
- Nucleic Acids are highly charged, RNA has many conformations due to the flexibility of the single stranded backbone. These factors have led to very few nucleic acid specific force fields being used for molecular dynamics of nucleic acids.

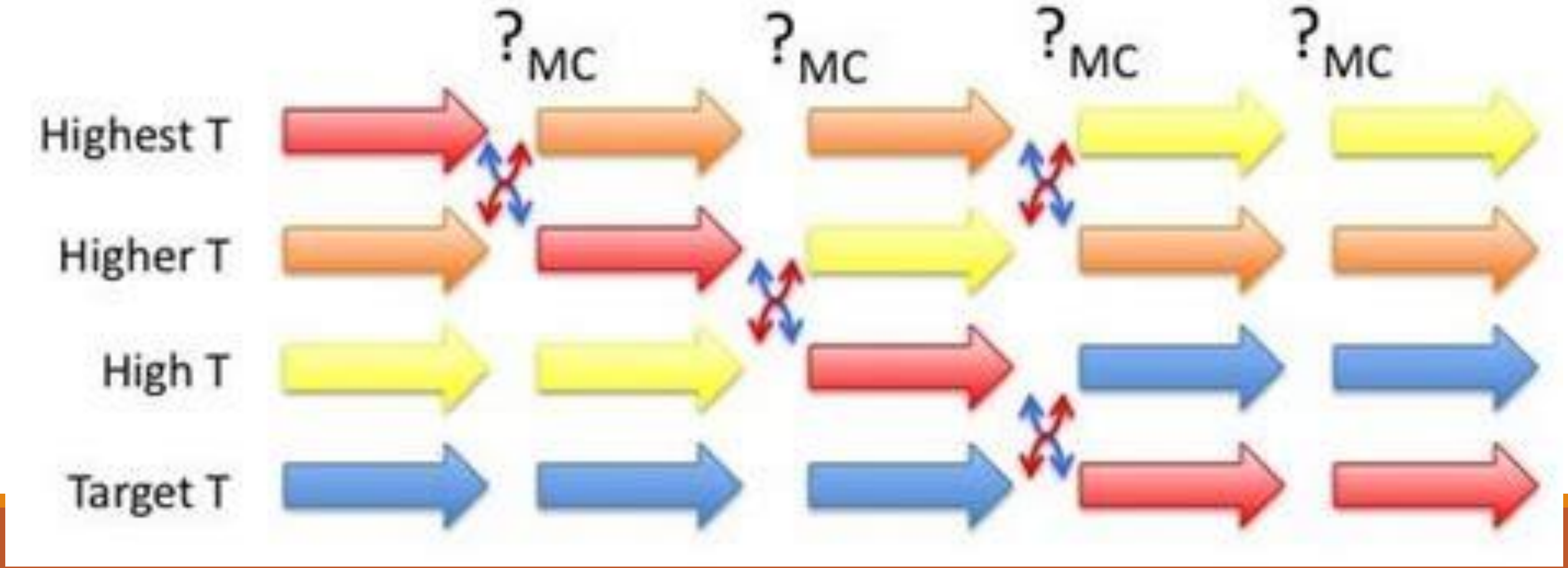


Molecular Dynamics(MD) Introduction

- MD is a tool that can be used to simulate the movement of atoms in molecules and predict the conformational changes that the structures can switch between.
- MD can be utilized with RNA to depict the structure of RNA and interactions of RNA over the course of pico seconds or even microseconds.
- The accuracy of a simulation is heavily reliant on how well the force field reflects the real events on the atomic level.
- Force fields are not typically developed for nucleic acids, focusing instead on proteins.
- This makes RNA simulations difficult to perform and difficult to show validity.

Replica Exchange Molecular Dynamics

- To simulate RNA, tetranucleotides are used. However, these highly flexible molecules can have many different conformations. Conventional MD cannot run long enough to elucidate these. Instead, Replica Exchange MD (REMD) can be used.
- Running an ensemble of simulations that have various different temperatures (T-REMD) at the same time. Conformations in these replicas can exchange to different temperatures, if the probability that the structure can remain at the lower temperature is high enough.



Convergence and T-REMD

- Convergence for a REMD simulation is a key to showing that the simulation is effective. This means that multiple conformations that occur in the structure can be repeatedly shown by the simulation, even with different starting points.
- Among the conformations will be lower energy, meaning higher prevalence, and higher energy, less prevalent. The most prevalent structures should be repeatedly shown if the system is reaching convergence.
- The authors previously used T-REMD on the r(GACC) tetranucleotide RNA. However, even after 3.8 microseconds per 24 replicates, convergence was not reached.

T-REMD Data

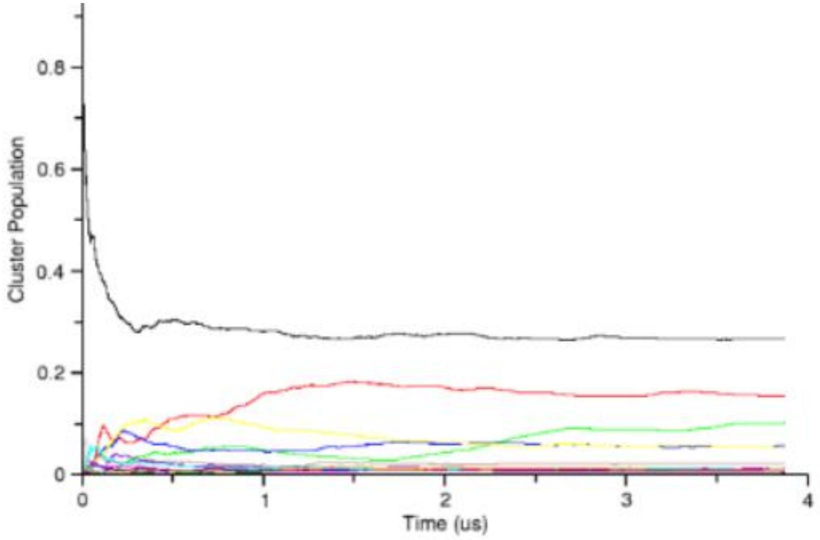


Figure S1: Cluster populations or occupancies as a function of time. The plot shows the percent occupancy for each of the clusters from the r(GACC) T-REMD simulation with 24 replicas as a function of time (extended to ~3.8 µs per replica) at 300 K. DBscan clustering, as described in the previous section, was performed on the RNA structures in the 300 K (temperature-sorted) trajectory. Each color represents a different cluster/conformation. Note that the DBscan clustering (using an epsilon of 0.9 Å and 25 minimum points) placed both the A-form-minor and A-form-major conformations into the first cluster (black).

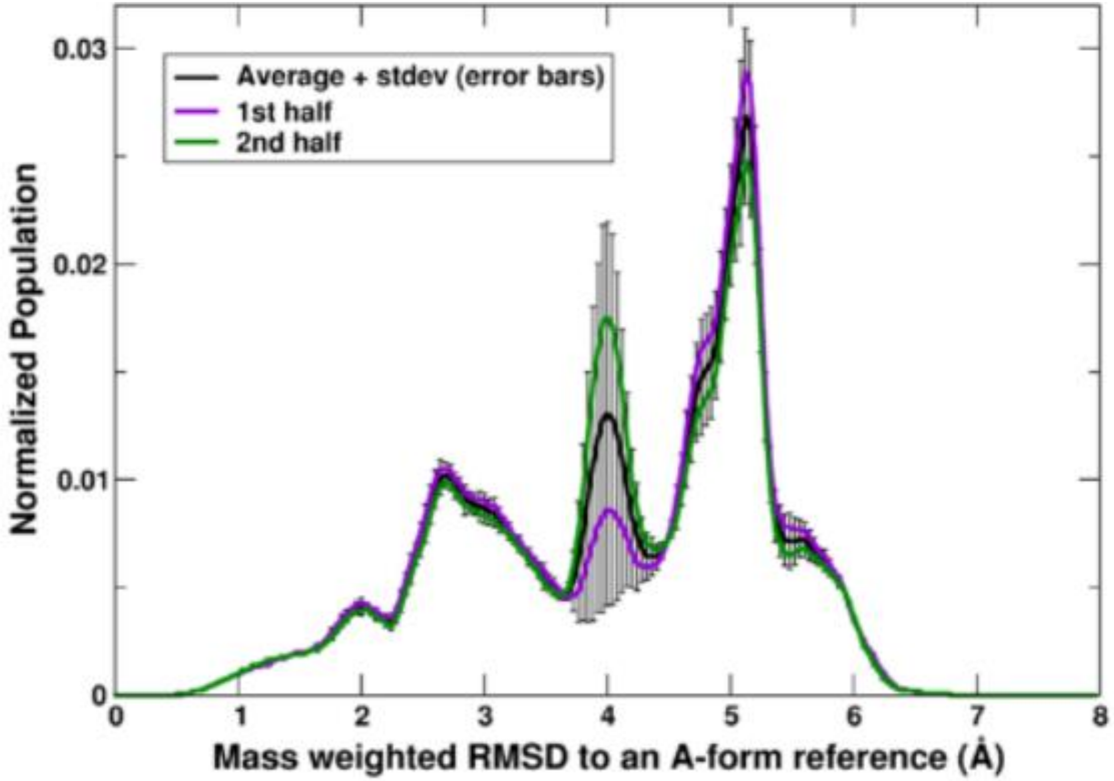


Figure S2: Estimating convergence of GACC T-REMD simulations from RMSD population histograms. Shown are the normalized populations (y-axis) of mass weighted RMSD values (x-axis) of all atoms in residues 1-4 to a reference structure (A-form RNA). The black line is the average of the first half and second half of the simulation, and the error bars are standard deviation. The first half histogram is shown in purple, and the second half histogram is shown in green.

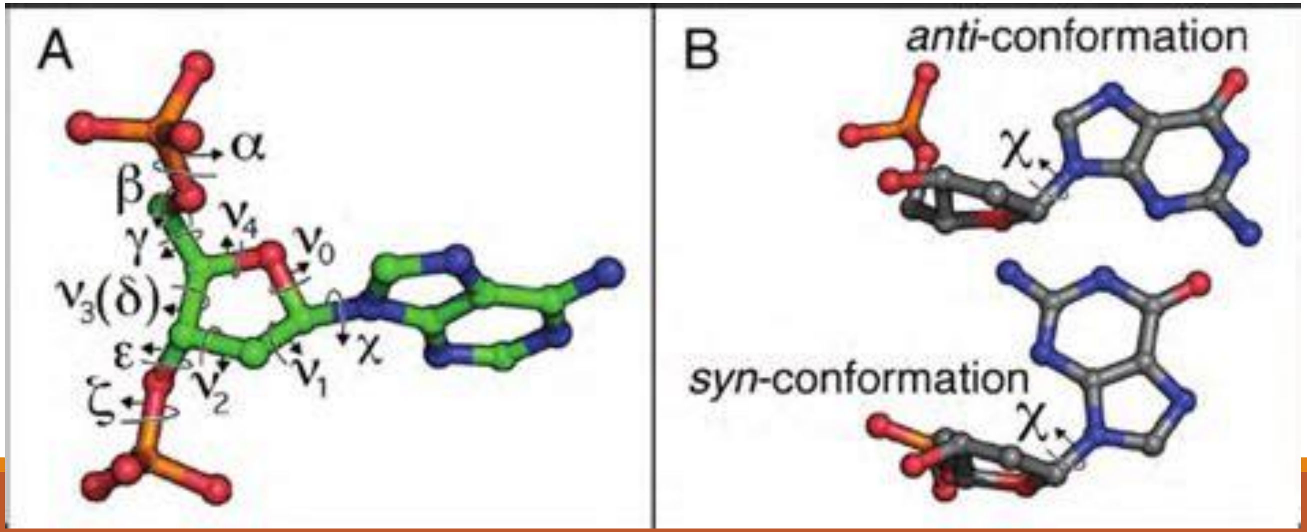
Changing the Force Field Parameters

Similar to changing temperature, force field parameters can be adjusted to allow for increased conformation changes.

In this paper, the dihedral force constant(DFC) was lowered, leading to a lower torsional energy barrier.

The Hamiltonian REMD (H-REMD) in this case means that the DFC was scaled from 1.0 to 0.3.

When this torsion barrier is decreased, interconversion between two states is easier as there is a lower barrier to overcome.

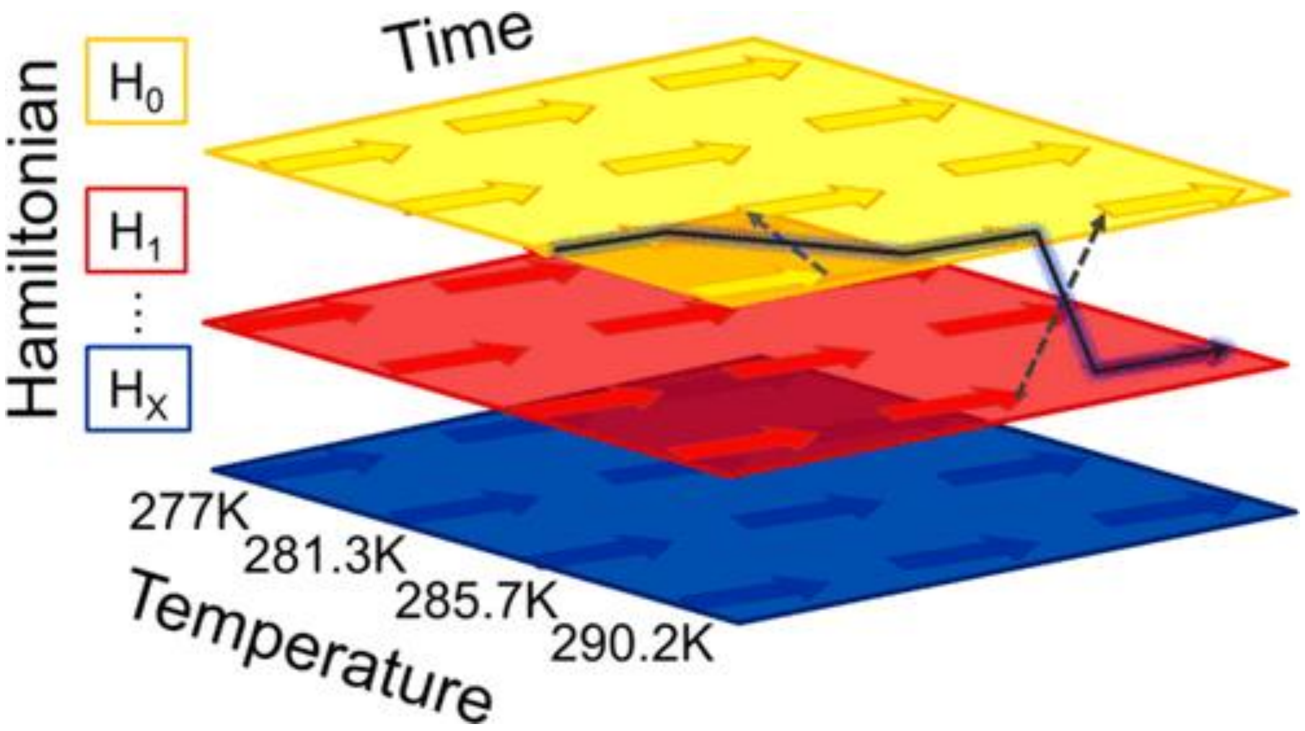


Multidimensional Replica Exchange

- To make the REMD more effective, the replicas can be exchanged in two dimensions. Meaning that the structure can be tested to see if it will remain in the lower temperature or in a higher dihedral constraint. Alternatively, exchange can occur if a structure needs to be in a higher energy state to change conformation.
- With this M-REMD, convergence can take place with much shorter run times. At the expense of running many more replicas.
- In this paper, the temperature and the Hamiltonian exchange via DFC were paired for the M-REMD.

Methods:

- The M-REMD used 8 different Hamiltonians, ranging 0.3 to 1.0 for the DFC, and 24 different temperatures, ranging 277K to 396K. This combination led to an ensemble of 192 replicas.
- This was compared to H-REMD where only the Hamiltonian exchange occurred. (Dihedral constant reduced to reduce torsional energy barrier).
- The H-REMD also had Dihedral constant adjusted on a scale to give 192 replicas.
- Used development version of AMBER12 (will be incorporated into AMBER14).



The DFC Scalars for the H-REMD

- The DFC scalars used were chosen based on the test with scalars ranging from the 1.0, unbiased, to 0.1, highly biased.
- After 0.3 there was little to no overlap, meaning exchange was no longer occurring.

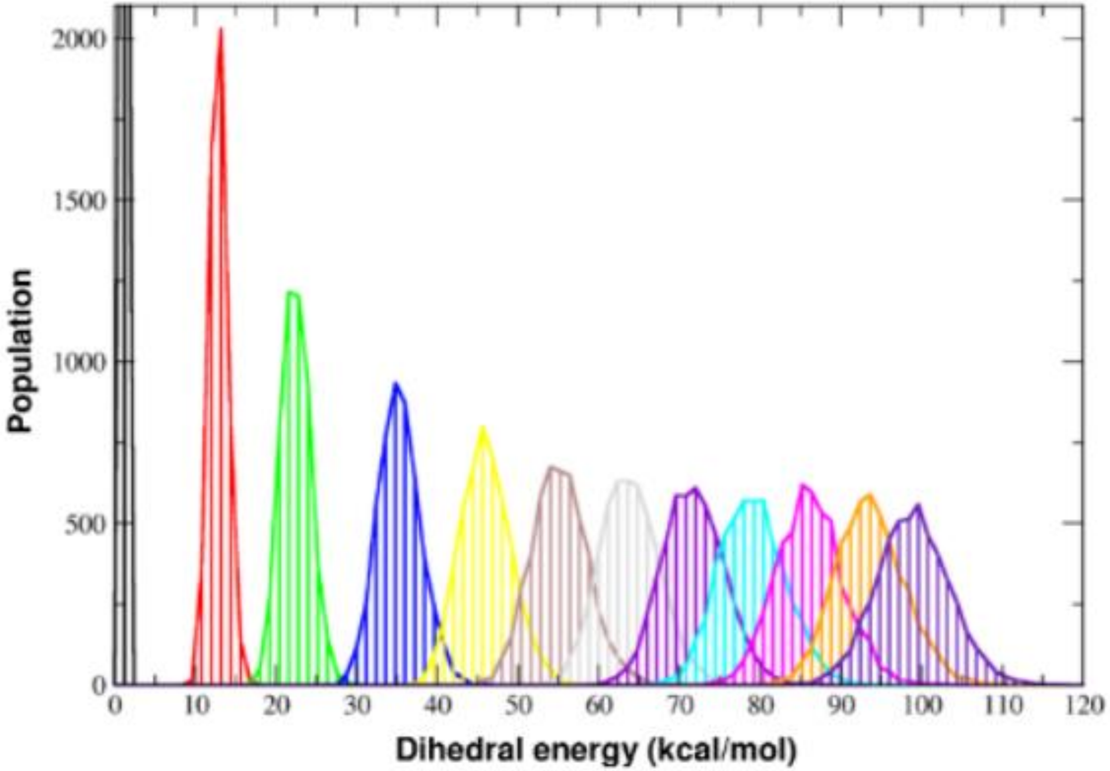
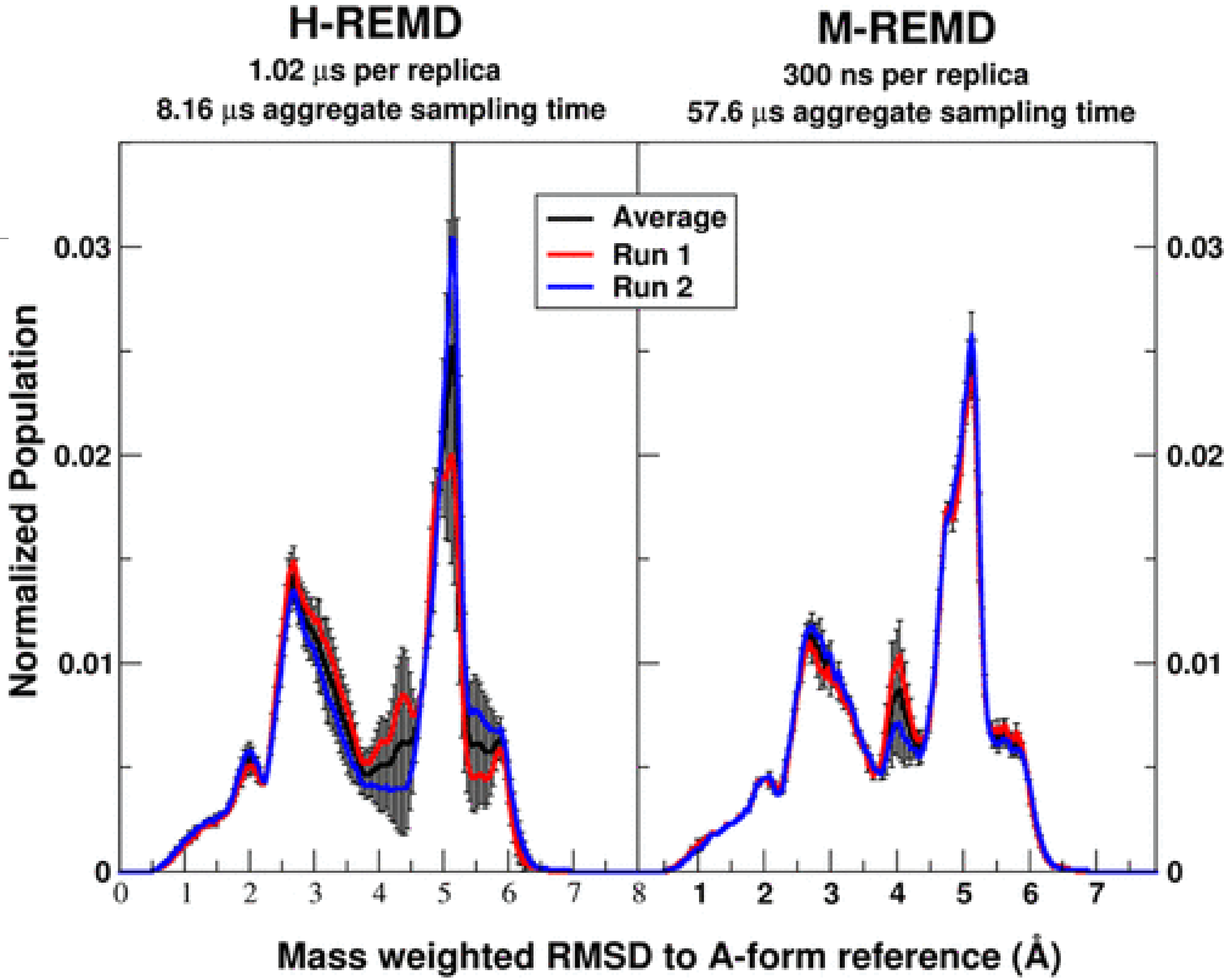


Figure 2: Overlap of RMSD.

- The RMSD between each individual H-REMD and M-REMD to a r(GACC) RNA A-form reference structure (found using NMR) was calculated.
- Overlap between the two runs shows how well converged the two runs are, meaning they sample similar RMSD space. Each run started at a different structure set.
- M-REMD has much more overlap than H-REMD.



Cluster Analysis

- Using an algorithm to group a set of objects that are more similar to each other (in the same group) than another groups (clusters).
- In this study, the most prevalent conformations that the r(GACC) occupied in the simulations were clustered together.
- To do this, two algorithms were used: average-linkage hierarchical agglomerative and DBscan clustering.
- This was done with the 300 K trajectories using CPPTRAJ to characterize the populations.

Figure 3: Cluster Analysis Structures

- The percentages of each structure is displayed.
- Red is guanine, green is adenine, blue and purple are cytosine.
- These are the average of two independent simulations.

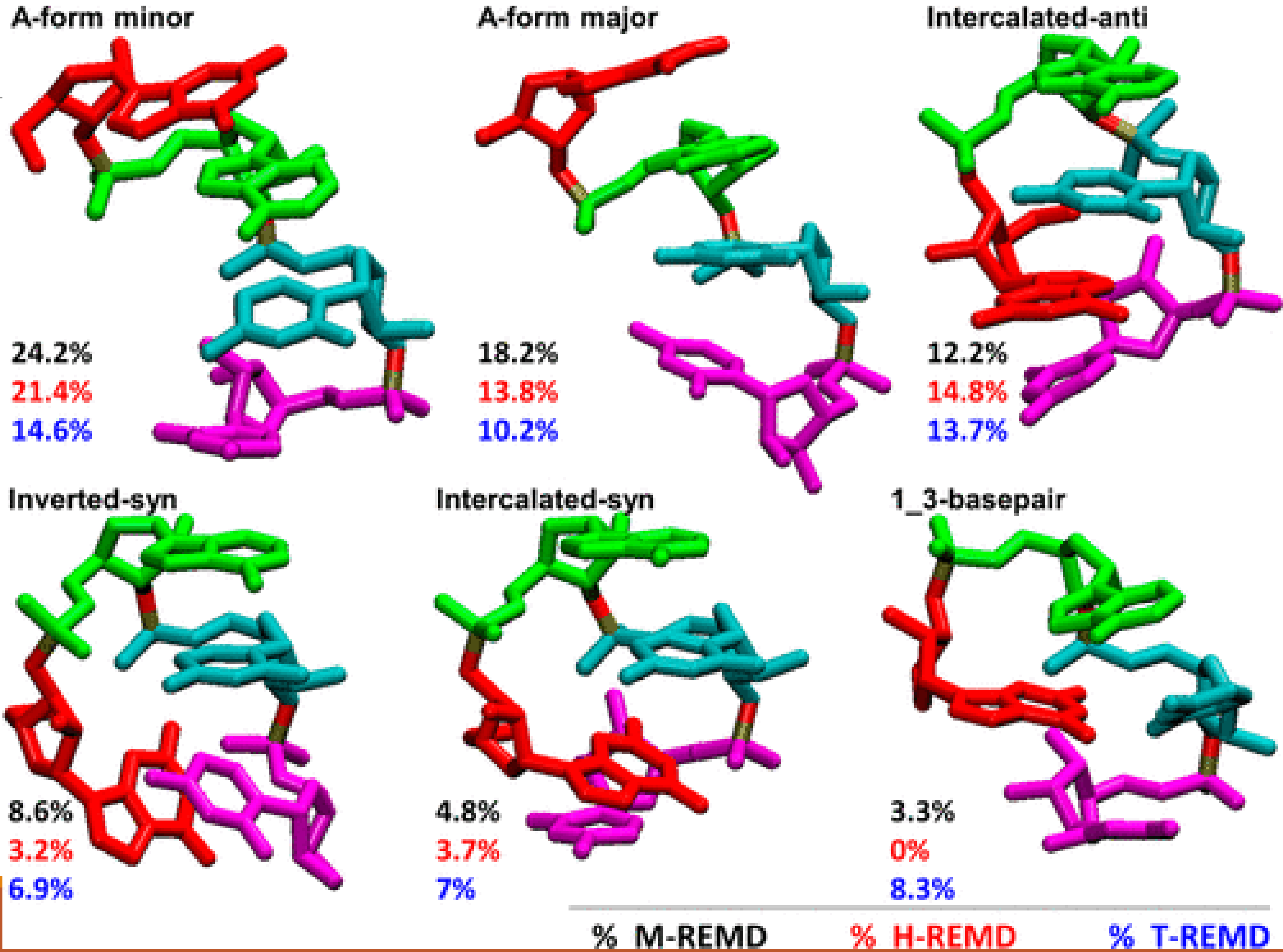
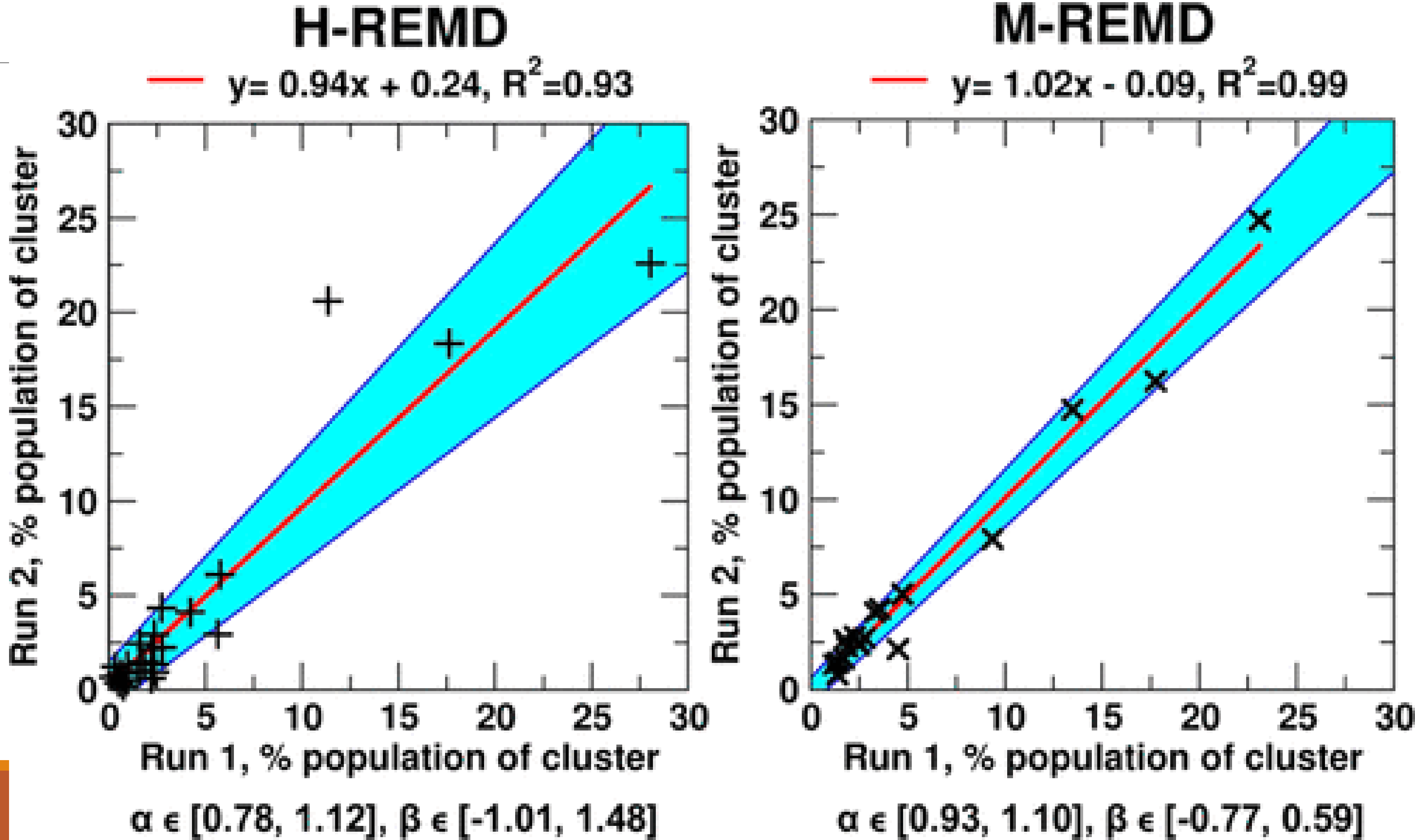


Figure 4: Cluster Population Correlations

- The difference between the two cluster percentages was plotted.
- The 95% confidence interval is the blue shaded region.
- The R^2 indicates how well the trendline fit the difference between the two runs.



Principal Component Analysis

- Principal Component Analysis (PCA) is a statistical technique used to reduce the dimensionality of a dataset so that large datasets with many dimensions can be more easily interpreted.
- The key to this is to transform the data into a new coordinate set without removing the variance but allowing it to be seen in fewer dimensions.
- In this experiment, the covariance matrix was calculated using a combination of two simulations for a REMD type. The eigenvectors from these were used to project coordinates.
- This was used to assess the overall dynamics of the system. The dynamics of two independent runs can be tested for convergence using the principal component projections.

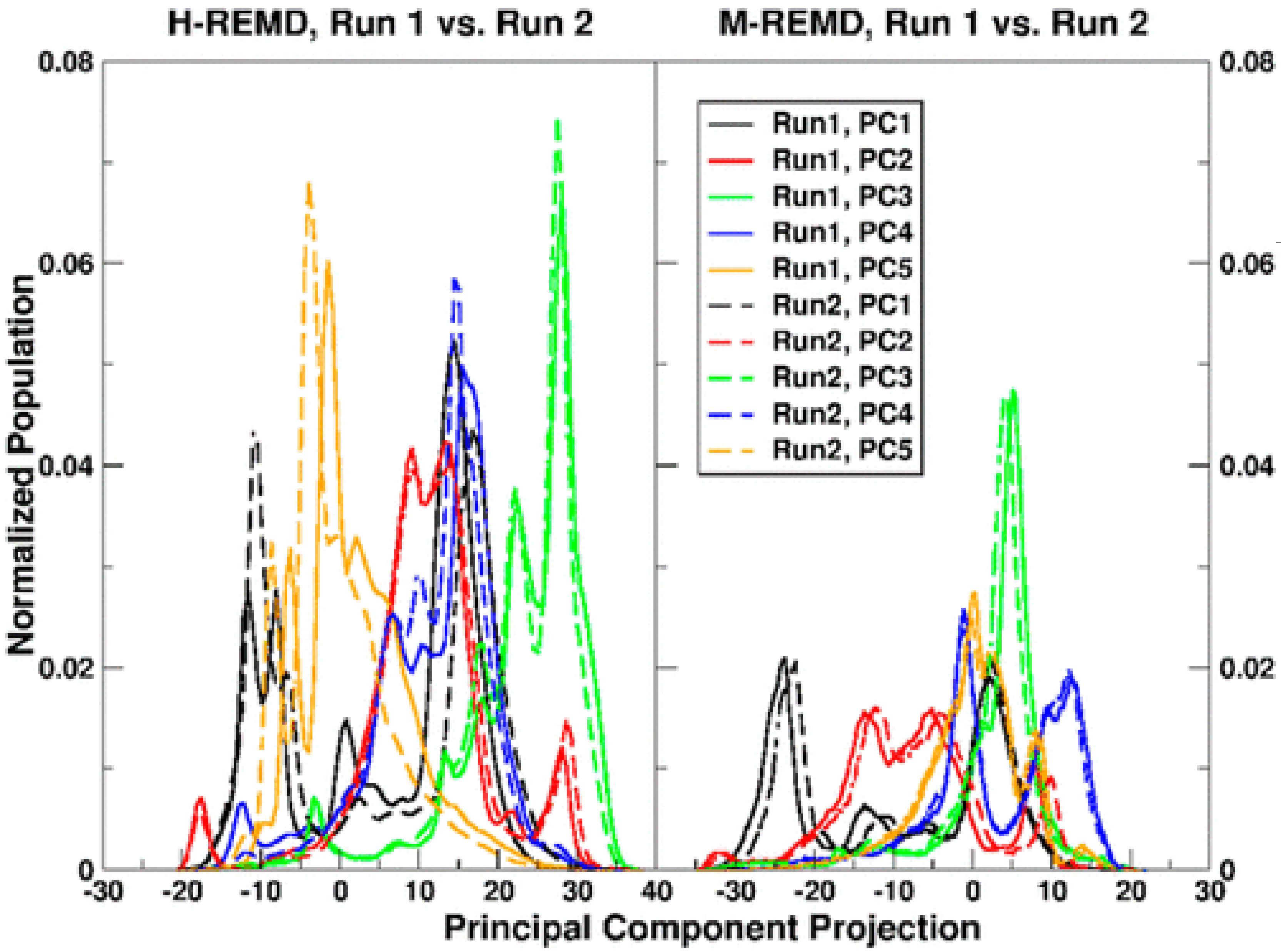


Figure 5:PCA

- The H-REMD two runs diverged, greatly in the low frequency range (black and yellow).
- The M-REMD runs overlap much better. Even the low frequency had very similar dynamics.

Kullback-Leibler Divergence

- To measure convergence in the REMD a KL divergence analysis was used to evaluate the difference between the most converged final ensemble versus the ensemble at various time points.

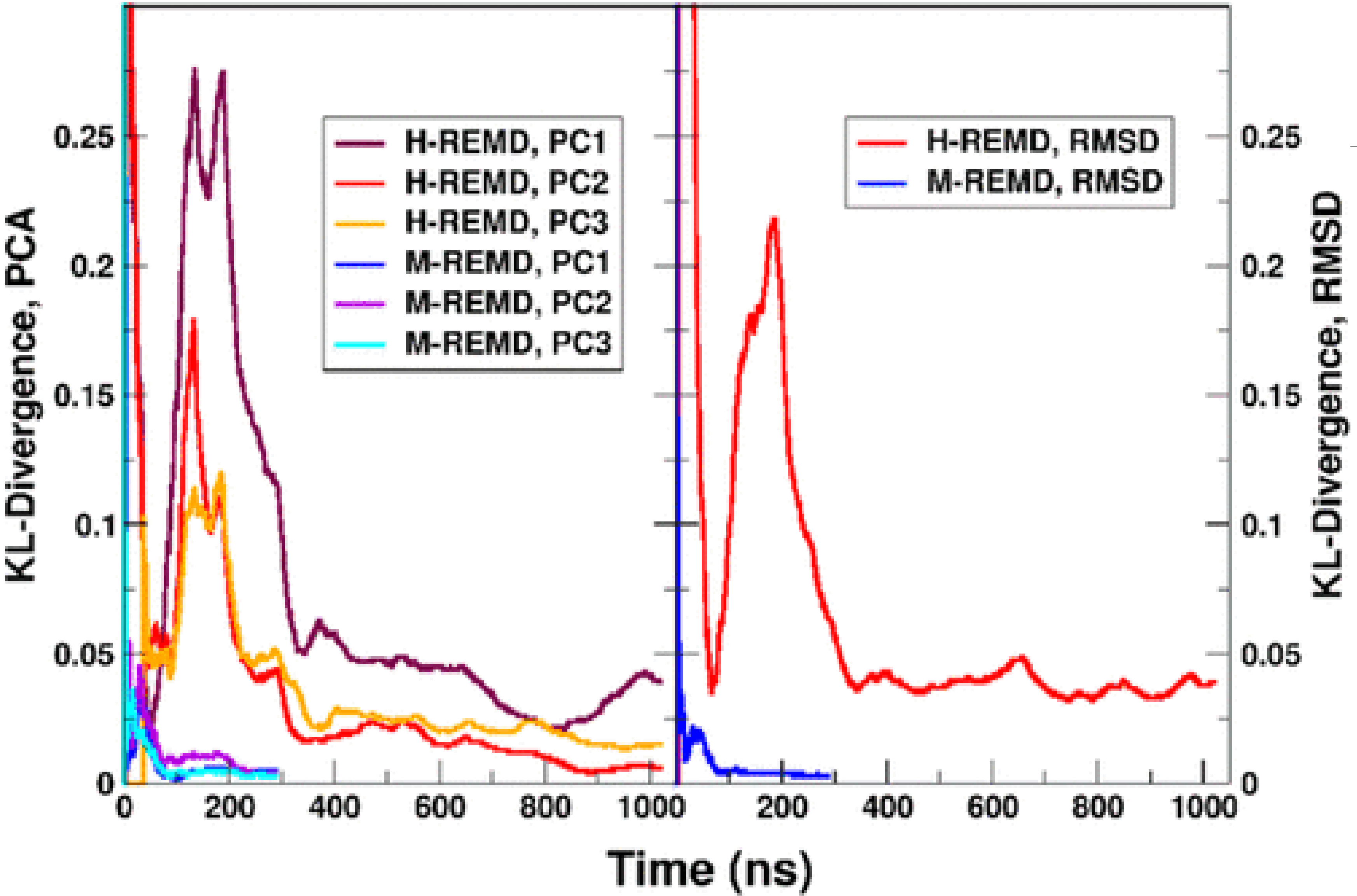
- KL Divergence is a statistical distance that measures the difference between one probability distribution P versus another probability distribution Q (the reference probability).

$$\mathrm{KL}(t) = \sum_{i} Pt(i) \ln\left(\frac{Pt(i)}{Qt(i)}\right)$$

- The principle component projections was the unbiased, 300K trajectories for H-REMD and M-REMD which were able to run to the end (to give the final ensemble).

- The RMSD to an A-form RNA reference structure was also used for KL Divergence.

Figure 6: KL Divergence Analysis



- The H-REMD PCs did not converge.
- The M-REMD converged rapidly for the PCA.
- The H-REMD RMSD also failed to reach convergence, whereas the M-REMD did quickly.

H-REMD Comparisons

- The H-REMD allowed for a glycosidic X-flip to occur more easily. Once this occurred, it became a lower energy conformation and the structure remained stuck in this conformation.

- M-REMD where temperature was also a factor, provided enough energy to overcome this X-flipped conformation. This meant that M-REMD had a mix between the H-REMD and T-REMD.

- This reason may be a factor in why H-REMD failed to reach convergence. To use H-REMD only and reach a converging model, more replicas at the lower biasing levels.

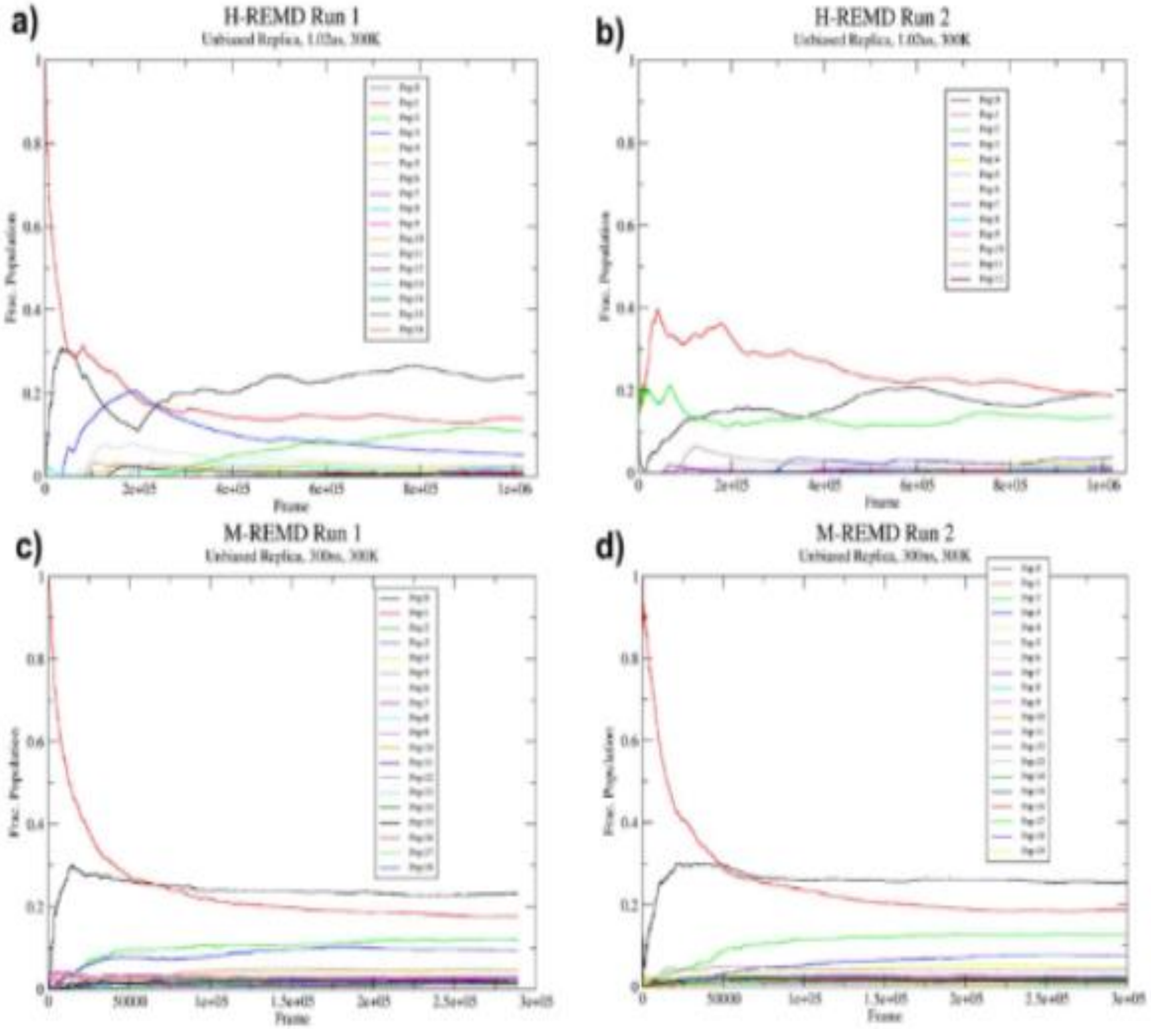
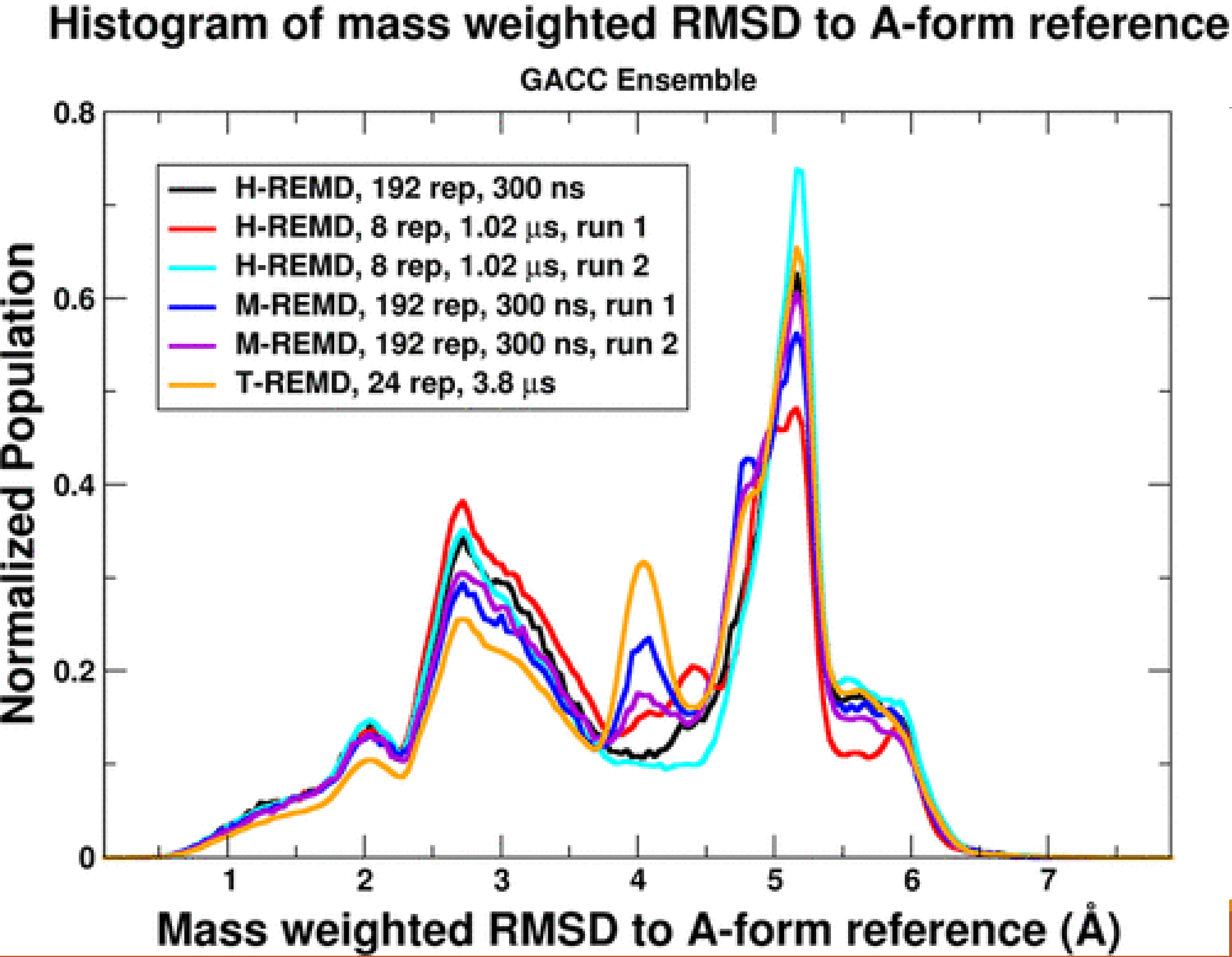


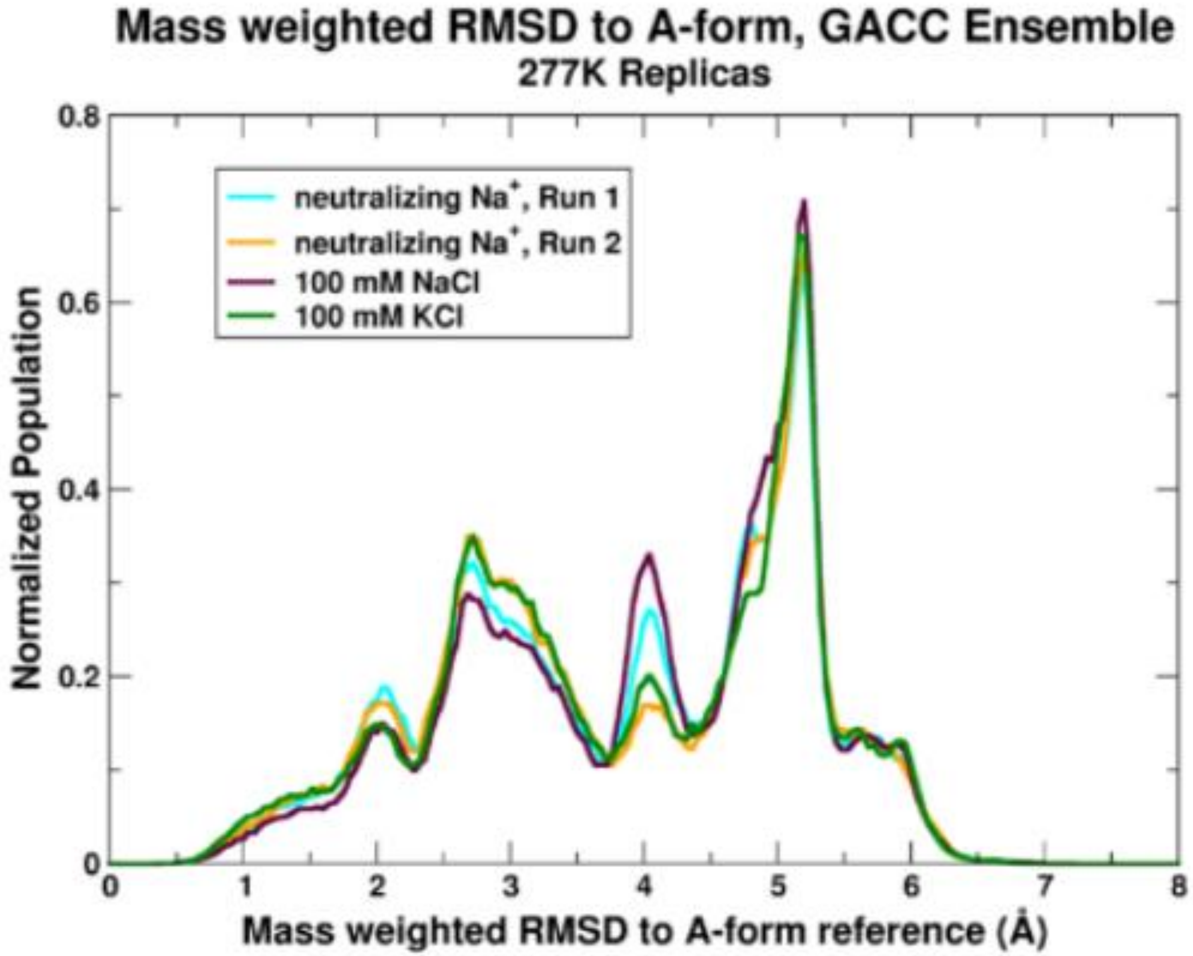
Figure S7: Cluster populations versus time. Shown are cluster populations as a function of simulation frame number (units of picoseconds) for each cluster comparing

Figure 7: Summary of the RMSD for the Populations



- The biggest deviations from the NMR A form RNA occur at 3 and 5 Angstroms. These represent the A-form minor and intercalated-anti conformations respectively.
- The H-REMD failed to show the conformation at 4 Angstroms (1_3 base pair).

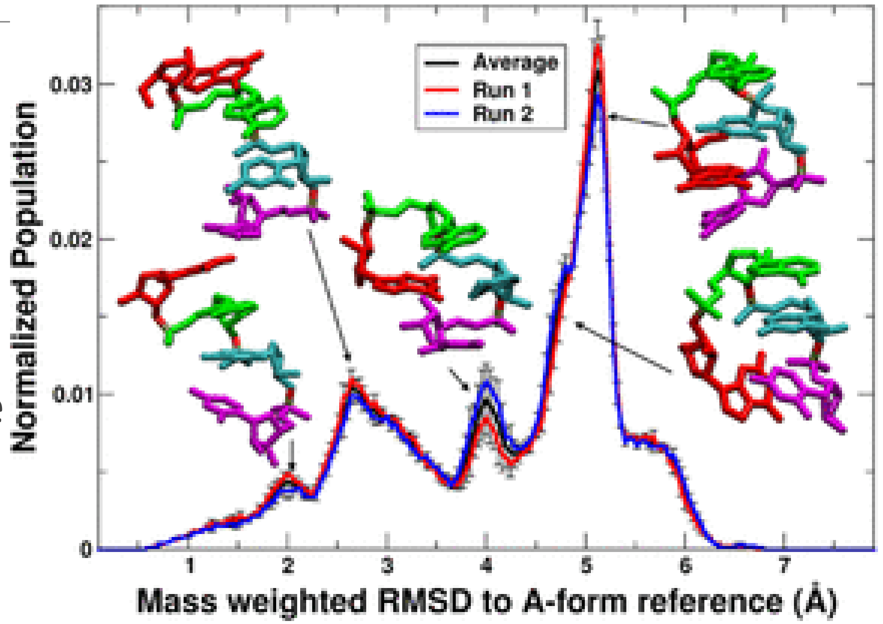
High Salt Content RMSD Histogram



- High Salt content allowed the same results to occur for the M-REMD

Conclusions

- The flexible system of the tetranucleotide RNA was able to reach convergence using M-REMD more quickly than T-REMD or H-REMD on their own.
- However, there are some limitations to this method. The simulations did not replicate the NMR data. The presence of the intercalated-anti structure was much more highly present compared to the NMR data.
- Another limitation is that high GPU count is required to do these ensembles of simulations.



r(GACC)
tetranucleotide

THANK YOU! Questions?