JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

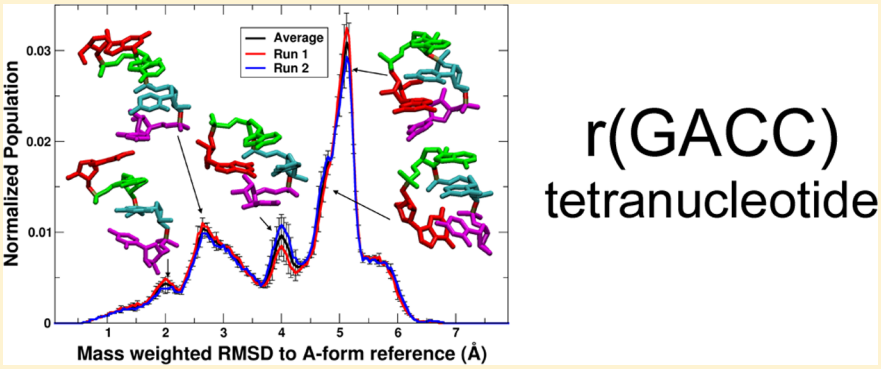
Multidimensional Replica Exchange Molecular Dynamics Yields a Converged Ensemble of an RNA Tetranucleotide

Christina Bergonzo,[†] Niel M. Henriksen,[†] Daniel R. Roe,[†] Jason M. Swails,[‡] Adrian E. Roitberg,[‡] and Thomas E. Cheatham, III*[,†]

[†]Department of Medicinal Chemistry, College of Pharmacy, University of Utah, Salt Lake City, Utah 84112, United States

[‡]Department of Chemistry, University of Florida, Gainesville, Florida 32611, United States

Supporting Information



ABSTRACT: A necessary step to properly assess and validate the performance of force fields for biomolecules is to exhaustively sample the accessible conformational space, which is challenging for large RNA structures. Given questions regarding the reliability of modeling RNA structure and dynamics with current methods, we have begun to use RNA tetranucleotides to evaluate force fields. These systems, though small, display considerable conformational variability and complete sampling with standard simulation methods remains challenging. Here we compare and discuss the performance of known variations of replica exchange molecular dynamics (REMD) methods, specifically temperature REMD (T-REMD), Hamiltonian REMD (H-REMD), and multidimensional REMD (M-REMD) methods, which have been implemented in Amber's accelerated GPU code. Using two independent simulations, we show that M-REMD not only makes very efficient use of emerging large-scale GPU clusters, like Blue Waters at the University of Illinois, but also is critically important in generating the converged ensemble more efficiently than either T-REMD or H-REMD. With 57.6 μs aggregate sampling of a conformational ensemble with M-REMD methods, the populations can be compared to NMR data to evaluate force field reliability and further understand how putative changes to the force field may alter populations to be in more consistent agreement with experiment.

INTRODUCTION

Knowledge of the role RNA plays in biological processes is ever increasing, and critical to RNA function are its structure, folding, dynamics, and interactions with other molecules.[1] Biomolecular simulation methods, including molecular dynamics (MD), are promising tools that augment experimental approaches by providing a detailed depiction of RNA structure and interactions on time scales ranging from pico- to microseconds. Crucial to the usefulness of MD methods are the reliability of the force fields. Due to the lag in the development of nucleic acid force fields compared to protein force fields, many question the accuracy and validity of the current RNA force fields.[1−3] This lag can be attributed to the high charge state of nucleic acids, the subtle balance between interactions within the RNA and with the solvent environment, the high conformational variability in the RNA backbone, and the lack of adequately converged sampling data.[4−6] Current "state of the art" conventional MD simulations cannot be trusted to parametrize force fields because convergence errors could be as large as or equal to those in the force field, leading to corrections for misdiagnosed problems that could instead simply be the result of incomplete sampling. Assessing force fields by standard MD simulation of RNA motifs like tetraloops has proven challenging because efficient sampling is limited by their large size.[7−9] Therefore, smaller tetranucleotides are being used as test systems for force field development and assessment. Though generating the full conformational ensemble is more tractable than it is for tetraloops,[10] tetranucleotides display significant conformational variability.[11−13] The complexity of elucidating the full conformational ensemble of tetranucleotide models is demonstrated by the lack of convergence of conventional MD simulations, even after microseconds of simulation time.[10,12]

Received: October 1, 2013
Published: November 15, 2013

Convergence of population distributions is more readily achieved by using enhanced sampling or an ensemble method such as REMD, which has been widely applied to study protein dynamics, and, to a lesser extent, nucleic acid dynamics.[9,14−19] This is beginning to change, as recent studies by Chen and Garcia[18] and Kührová et al.[20] have used REMD in combination with their respective force field modifications to fold hairpins and tetraloops to their native structure. Traditional REMD involves multiple independent simulations run at different temperatures (T-REMD) which periodically attempt an exchange in temperature space.[21,22] Replicas at low temperature have an opportunity to exchange to a higher temperature where energy barriers may be more easily crossed; in this manner the effective sampling space of each replica is enhanced. However, this method has its drawbacks: T-REMD may not substantially enhance sampling if the system is too large and the number of replicas needed to bridge a temperature range is prohibitive, or if increased temperature does not facilitate the conformational transition of interest. An alternative is Hamiltonian replica exchange (H-REMD)[23] where part of the Hamiltonian is scaled or altered, for example, with a weighted biasing potential with the opposite sign of the dihedral term (targeting protein backbone dihedrals)[24] or with various sets of umbrella sampling-like restraints for different elements of the ensemble.[25] A benefit is that one replica may be simulated with the original Hamiltonian which, given sufficient sampling and exchange, provides the complete unbiased ensemble.

Not only is convergence crucial for force field development, but it is also even more important for accuracy and comparison with experiment. Generating a formal definition of convergence greatly depends on which system properties one wishes to converge in a simulation. Many authors have determined metrics of convergence for replica exchange simulations, and most, if not all, replica exchange studies discuss the extent to which their simulations are converged. Abraham and Gready proposed that within replica exchange simulations there is interplay between thermodynamic efficiency of sampling the ensemble properties, and mixing efficiency of replicas traversing the temperature or Hamiltonian range of interest.[26] Several studies analyze peptide melting curves, RMSD to known experimental structures, radius of gyration, and principal component projections as ensemble properties, ideally between independent sets of REMD simulations which start from different conformations.[27−30] Quantitative measures of convergence in biomolecular simulation include autocorrelation functions of potential energy,[31] a "decorrelation time" defined by Lyman and Zuckerman,[32] and Kullback−Leibler divergence analysis.[33] Previously, we illustrated the importance of generating a converged ensemble with the r(GACC) tetranucleotide; however, in 2 μs of T-REMD per replica with 24 replicas spanning 277−396 K, the ensemble was not yet completely converged.[10] Even with longer simulation, now at 3.8 μs of T-REMD per replica or an aggregate sampling of over 90 μs, a well-converged ensemble is elusive (Supporting Information Figures S1 and S2). Differences in the populations of conformations in the ensemble are significant and warrant further exploration of sampling methods, including H-REMD and its extension to multidimensional replica exchange (M-REMD) to see if complete sampling can be obtained more quickly.

COMPUTATIONAL METHODS

Amber Implementation. In T-REMD and H-REMD, exchanges occur in only one dimension; i.e., only one part of the Hamiltonian is being altered between replicas. However, two or more exchange dimensions can be coupled; this approach is referred to as multidimensional REMD (M-REMD). This leads to enhanced conformational sampling. Not only are there more replicas in the simulation simultaneously sampling conformational space, but also different conformational properties of the system can be enhanced. M-REMD has been implemented in the MD engines SANDER and PMEMD in a development version of Amber (to be released with Amber 14) and will be briefly described here. The current implementation allows for the use of any number of Hamiltonian dimensions (i.e., changes in the system topology that do not involve changing the number of atoms, as well as changes to input parameters) and/or temperature dimensions. In addition, the code has been designed to be extensible to facilitate the addition of other dimension types. A schematic of the M-REMD simulation setup used in this work is shown in Figure 1.

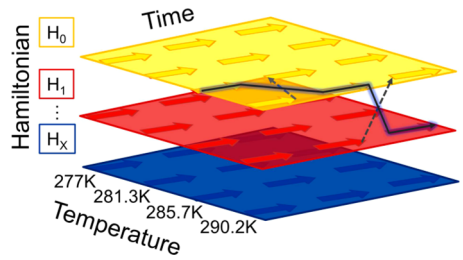


Figure 1. Schematic diagram of the M-REMD simulation. Each plane represents a different Hamiltonian, and each arrow represents a different temperature. The unbiased Hamiltonian, shown in yellow, can be recovered at the Temperature of interest. Exchange in temperature dimension between 285.7 and 290.2 K is followed by an exchange in Hamiltonian space between H_0 and H_1. This figure represents four out of 24 temperatures and three out of eight Hamiltonian dimensions used in this work.

M-REMD in SANDER/PMEMD makes use of the existing T-REMD and H-REMD exchange subroutines. Exchanges are attempted in each dimension in turn, so that the first exchange is attempted in the first dimension, the second exchange is attempted in the second dimension, and so on. This is illustrated in the following pseudocode:

```
multid_exchange(n_exchange) {
  my_dim = n_exchange % remd_dimension;
  if ( remd_types[my_dim] == TEMPERATURE )
    temperature_exchange();
  else if ( remd_types[my_dim] == HAMILTONIAN )
    hamiltonian_exchange();
}
```

Here "%" denotes the modulo operation, n_exchange is the number of the exchange being attempted, remd_dimension is the total number of replica dimensions, and remd_types[ ] is an array of size n_dimension containing an index corresponding to the type of exchange to be performed in that dimension.

As with T-REMD or H-REMD, individual replicas are defined in a file called the "group file", with each line defining the input file, input coordinates, topology file, etc. for each replica. In M-REMD there is an additional file called the

dx.doi.org/10.1021/ct400862k | J. Chem. Theory Comput. 2014, 10, 492−499

"dimension file", which defines both the replica dimensions and which replicas are allowed to exchange within each dimension. This file follows the same Fortran namelist format that Amber MD input files use; there is a title followed by a "&multirem" namelist for each replica dimension with format:

```
Dimension X Title
 &multirem
  exch_type = <Exchange Type>,
  group(1,:) = <Group 1 Replica List>,
  …
  group(N,:) = <Group N Replica List>
  desc = 'Dimension X description'
 /
```

Exchange type (exch_type) is currently restricted at present to either "TEMPERATURE" or "HAMILTONIAN". A "Replica List" consists of a comma-separated list of integers corresponding to positions in the group file, where replica 1 is the first entry in the group file; this defines which replicas are allowed to exchange within that dimension. M-REMD is enabled in a simulation by specifying "-remd-file <dimension file>" on the command line.

Ensemble Postprocessing Using CPPTRAJ. As with the current implementations of T-REMD and H-REMD in Amber, exchanges in temperature space are accomplished by swapping temperatures, and exchanges in Hamiltonian space are accomplished by swapping coordinates. This means that if any dimension is temperature, trajectories will have to be sorted if data are to be processed at a single temperature. This requires that every trajectory file contain information describing the overall dimensionality of the run (i.e., how many dimensions there are and of what type) and each trajectory frame contain information on where it is located in each dimension. Similarly, the restart files used to checkpoint the simulation must contain the same information. Because NetCDF files are by their nature extensible, adding additional information to the format does not break existing parsers that have not yet been set up to recognize it; the additional information is not used. Primarily because of this, it was decided that only the Amber NetCDF trajectory and restart formats would be updated for M-REMD, and thus all M-REMD runs require the use of the Amber NetCDF trajectory and restart formats. Only three additional pieces of data were required to be added to the NetCDF trajectory/restart formats: an integer containing the number of replica dimensions (N), an integer array (of dimension N) containing the type of each dimension, and an array (of dimension $F \times N$, where F is the total number of frames) containing the indices of a replica in each dimension for each frame. CPPTRAJ from AmberTools 13.0 has been modified to process the new information from these trajectories.

Analysis Using the DBScan Clustering Algorithm. Postprocessing of all REMD simulations was performed using PTRAJ and specially modified development versions of CPPTRAJ.[34] Cluster analysis was performed using two methods: the average-linkage hierarchical agglomerative and DBscan clustering methods.[35,36] For both algorithms, coordinate RMSD was used as the distance metric. The average-linkage agglomerative algorithm clustered on heavy atoms of residues 1−4 used a critical distance ε value of 2.3 Å and a variable sieve value to ensure a 5000 frame initial pass through the trajectory. DBscan clustering was performed on a subset of atoms described using the AMBER mask syntax (:1@N2,O6,C1′,P,:2@H2,N6,C1′,P,:3@O2,H5,C1′,P,:4@O2,H5,C1′,P), used a distance cutoff ε of 0.9 Å between clusters and a minimum of 25 points required to form a cluster. Again, a variable sieve was employed to ensure at least a 5000 frame first pass through the trajectory.

Principal Component Analysis. Principal component analysis was performed using CPPTRAJ. To compare principal components obtained from two independent runs, the covariance matrix was calculated from the combined trajectories of the two independent simulations for a given REMD type; each frame was first RMS-fit to an overall average structure to remove global translational and rotational motion. The eigenvectors and eigenvalues were then obtained from diagonalization of the combined covariance matrix, after which coordinates from each independent trajectory were projected along eigenvectors of interest to obtain projection values for given modes.

Measuring Convergence using Kullback−Leibler Divergence Analysis. The Kullback−Leibler divergence analysis[33] over time was performed using a development version of CPPTRAJ. At each frame **t**, a histogram for each data set being compared (**Pt** and **Qt**) was constructed from data at all previous frames ($0 \rightarrow \mathbf{t}$) using a Gaussian kernel density estimator with 400 bins and a bandwidth estimated from the normal distribution approximation for the entire data set. The Kullback−Leibler divergence at frame t ($KL(t)$) was then calculated:

$$\mathrm{KL}(t) = \sum_i Pt(i) \ln\left(\frac{Pt(i)}{Qt(i)}\right) \quad (1)$$

To ensure that the Kullback−Leibler divergence was properly defined, the sum over each histogram was normalized to 1.0 and frames in which a bin had density in one histogram but no density in the other histogram were ignored.

Simulation Details. The GACC RNA, r(GACC), system setup, building, and equilibration protocol used here was previously described in Henriksen et al.[10] The Amber ff12SB force field parameters were used that include the base AMBER ff99 force field[37] and the parmbsc0 corrections for α/γ nucleic acid backbone torsions and χ_{OL3} torsion corrections for RNA.[4,5] The r(GACC) structures used in temperature replica-exchange (T-REMD) simulations were directly taken from the previous work and were solvated with 2497 TIP3P waters[38] and 3 Na^+ ions.[39] Equilibration was performed at each of the 24 replicas' target temperature (temperatures ranged from 277 to 396 K). Replicas were spaced to yield at least a ∼20% acceptance rate. The previous T-REMD runs were extended from ∼2 μs per replica to ∼3.8 μs per replica.

All (nonreplica exchange) molecular dynamics (MD) simulations were carried out with the *pmemd.cuda.MPI* module of the Amber12 suite of programs.[40,41] All replica exchange molecular dynamics (MD) simulations were carried out with the development version of the *pmemd.cuda.MPI* and *pmemd.MPI* modules based on the Amber12 suite of programs; this code will eventually be released as part of Amber14. Production dynamics for each replica were carried out in the *NVT* ensemble. Temperature was regulated using the Langevin thermostat with a collision frequency of 2 ps^{-1} using the "ig=-1" option to randomly set the random number seeds at each restart, avoiding synchronization effects.[42,43] An integration time step of 2 fs was used. The exchange attempt time interval was set to 1 ps. The nonbonded direct space cutoff was set to 8.0 Å, and default Amber12 particle mesh Ewald settings were used for reciprocal space calculations. SHAKE was used to

constrain bonds to hydrogen.[44] Simulations and analysis were performed both locally at the University of Utah's Center for High Performance Computing and also on the NCSA Blue Waters and XSEDE Keeneland and Stampede computational resources.

H-REMD simulations were performed with eight replicas; this drastically cuts back on the number of replicas in the simulation compared to T-REMD, which was possible because decent exchange probabilities between the Hamiltonians could be obtained. For the H-REMD, we uniformly scaled the dihedral force constant (DFC) of all dihedrals by a constant, from 1.0 (full force field) to 0.3 (lowest torsion barriers in this work) by 0.1 intervals: (1.0, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3) leading to eight replicas. Support for such parameter/topology file modifications will be included in the ParmEd and CPPTRAJ tools distributed with AmberTools 14. When the torsion barrier heights are decreased, interconversion between minima and overall sampling of structural populations is enhanced. On the basis of our system benchmarks, scaling from full DFC to 30% DFC by 10% intervals gave acceptable potential energy overlap between replicas and led to an exchange acceptance of 57% to 26% (Supporting Information Figure S3). Simulations were performed using the development version of Amber14 and AmberTools14 suite of programs in which H-REMD and M-REMD are implemented.[45−47] For the initial H-REMD simulations, the equilibrated structure at 300 K was used as a starting structure for all eight replicas. To generate starting structures for the subsequent and independent H-REMD simulation, the restart structures from the original H-REMD simulation after 510 ns were assigned new velocities and simulated for 1 ns in the NVT ensemble.

With fewer replicas, convergence of the H-REMD distribution requires long simulations per replica, or long wall-clock time, to aggregate a similar amount of sampling. To accumulate sampling on the same order as the M-REMD and T-REMD, the DFC was scaled from a factor of 1.0× (full DFC) to 0.3× over 192 replicas using an interval of 0.0036. Starting structures for the 192 replica H-REMD were identical to the 8 replica H-REMD run 2 structures and were each used 24 times. An exchange acceptance rate from 96% to 99% was observed.

To further enhance sampling, the temperature and Hamiltonian dimensions were combined into a two-dimensional M-REMD run. By combining the temperature and Hamiltonian dimensions,[48] we can reach the converged ensemble for a solvated r(GACC) tetranucleotide much faster than using either T-REMD or H-REMD alone. For the M-REMD simulations, the temperature range used was the same as in the previously published T-REMD simulation, totaling 24 T-REMD replicas from 277 to 396 K, which, when coupled with the 8 H-REMD replicas, leads to a total of 192 replicas. Each equilibrated starting structure from T-REMD was copied eight times, and each copy was assigned a different Hamiltonian. To generate the starting structures for an independent M-REMD simulation, the initial r(GACC) structures equilibrated at their target temperatures were assigned new velocities and simulated for 1 ns.

RESULTS

Figure 2 shows histograms of the mass-weighted heavy atom RMSD of each individual H-REMD and M-REMD simulation to an A-form reference structure. Overlapping histograms indicate that separate simulations are sampling the same RMSD space, which is a necessary but not sufficient condition for

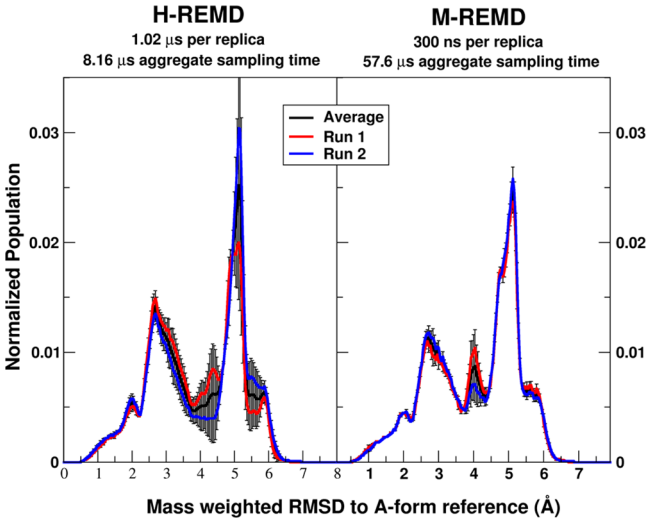


Figure 2. Population analysis showing the number of structures at specific RMSD values from an A-form reference structure. Mass weighted RMSD histograms of the unbiased replicas at 300 K are averaged (shown in black) between two runs (shown in red and blue). Error bars represent standard deviation between two runs. Independent sets of starting structures were used for each simulation.

convergence. The large error bars between the two independent H-REMD runs in Figure 2 indicates the simulations have sampled somewhat different populations. Error bars between the two M-REMD simulations are smaller, and indicate that the simulations, which start from different structure sets, ultimately sample more similar RMSD space than the H-REMD simulations, as well as the T-REMD simulation (shown in Supporting Information Figure S2). To characterize the populations, cluster analysis on the 300 K trajectories was performed using CPPTRAJ (sample scripts are available in the Supporting Information).[34] Representative structures and cluster populations are shown in Figure 3 and Supporting Information Table 1 and Figure S4. The two H-REMD simulations disagree about the populations of the major peak at 5.0 Å (the intercalated-anti and inverted-syn structures in Figure 3), as well as the structures at 4.0 (1_3-basepair) and 6.0 Å (intercalated-syn). In contrast, the smaller error bars between independent runs of M-REMD indicate much better convergence of the populations between the two simulations (shown more clearly in Supporting Information Table 1 and Figures S5, S6, and S7). This also indicates that the higher temperature replicas are critical to resolve structures in the unconverged regions of the H-REMD simulation.

Poorer apparent convergence between the H-REMD runs is supported by the lower correlation between cluster populations of the two independent runs compared to M-REMD, shown in Figure 4. Independent simulations are clustered together, and the cluster population from each independent run is reported. If the simulations are sampling the same conformational space despite the difference in their starting structure conditions, they are better converged, and the cluster populations from the two independent runs will be the same. The cluster populations from the H-REMD simulations have a correlation coefficient of 0.93, whereas the populations from the M-REMD simulations have a correlation coefficient of 0.99. To test the robustness of the correlation, the 95% confidence intervals were calculated for the slope (α) and intercept (β) of the linear fits for the H-REMD and M-REMD simulations, shown as the shaded area in

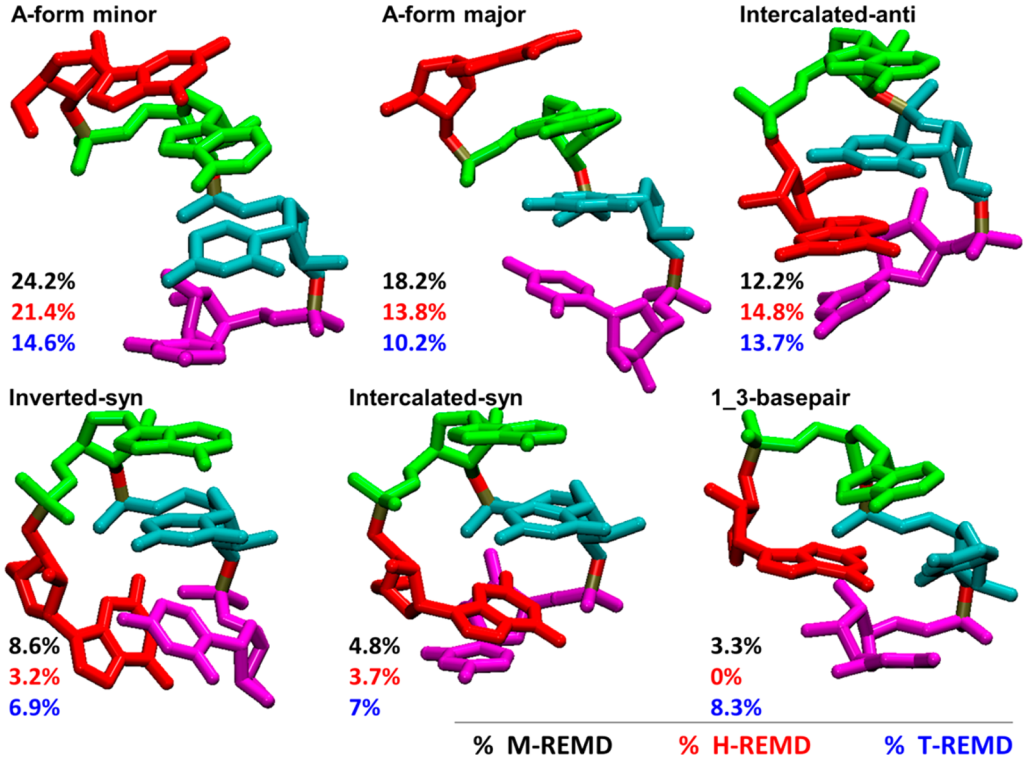


Figure 3. Representative structures from DBscan cluster analysis at 300 K. The top six most populated clusters in most REMD simulations are shown above. For the r(GACC) sequence the coloring is G1 (red), A2 (green), C3 (cyan), and C4 (magenta). Cluster populations for M-REMD are shown in black, H-REMD in red, and T-REMD in blue. M-REMD and H-REMD cluster populations represent the average of two independent simulations. Cluster names were assigned on the basis of analysis of the 277 K T-REMD trajectory.

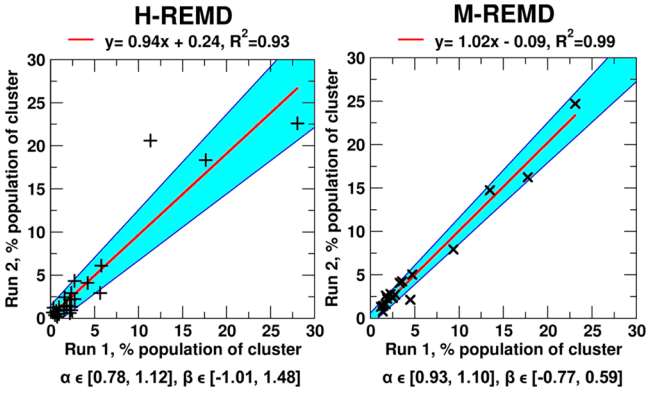


Figure 4. Correlation of cluster populations between independent H-REMD runs and independent M-REMD runs. Linear fit and correlation coefficients for H-REMD and M-REMD are shown, where red fits all clusters. The shaded area represents a 95% confidence interval where the slope and *y*-intercept bounds are denoted by α and β, respectively.

Figure 4. The correlation was also calculated with the most populated cluster left out (Supporting Information Figure S8). The difference between the two sets is more apparent in the H-REMD simulations, again indicating that the independent simulations remain unconverged.

Principal components, which describe the overall dynamics of the system, are less easily converged than other metrics, particularly those components corresponding to the lowest frequency motions. To ascertain whether the dynamics of two independent runs appear to be converged, we looked at the overlap of histograms of principal component projections calculated from Cartesian coordinates obtained from each simulation (see Computational Methods: Principal Component Analysis for more details). If the replicas have sufficiently diffused through temperature and Hamiltonian dimensions, the different initial structure sets should sample the same conformational space resulting in a close principal component overlap. A script that can be used to perform this calculation with CPPTRAJ can be found in the Supporting Information. Figure 5 shows the overlap in the first five principal components for the H-REMD and M-REMD simulations. The independent H-REMD simulations diverge, particularly in the first low frequency mode (black solid line vs black dashed line), which is representative of the majority of fluctuations in the system. Conversely, the M-REMD principal component histograms show excellent overlap, indicating similarity between the overall dynamics of each independent M-REMD run.

To quantitatively examine rates of convergence, we performed Kullback–Leibler divergence analysis (KL divergence) over time on histograms of principal component (PC) projections for the unbiased 300 K trajectories from H-REMD and M-REMD, as well as for the histograms of mass weighted RMSD to an A-form reference structure.[33] The results are a measure of the difference between probability distributions sampled from each simulation as a function of simulation length, yielding a time-dependent measure of ensemble differences. Low divergence values indicate the ensemble at time X is not significantly different from the final ensemble, whereas high divergence values indicate the ensemble at time X is very different from the final, or most converged ensemble. Figure 6 shows the KL divergence of the first three principal components for the H-REMD and M-REMD simulations. M-

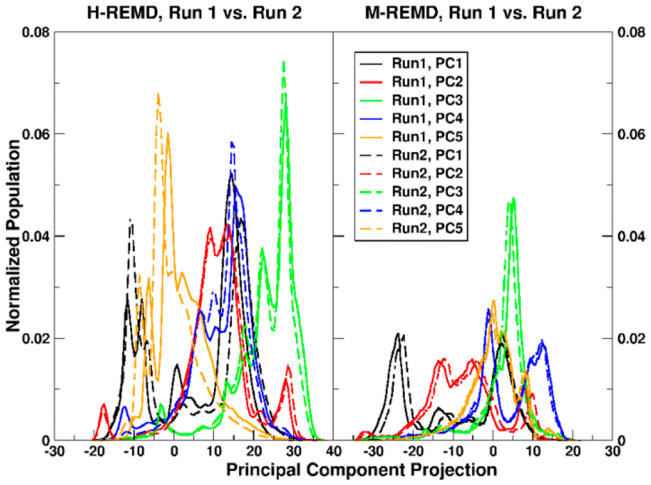


Figure 5. Principal component projection of top five modes onto (left) H-REMD run 1 (solid lines) and H-REMD run 2 (dashed lines) show little overlap in the low frequency modes, whereas (right) M-REMD run 1 and M-REMD run 2 show the overlap in the low frequency modes.

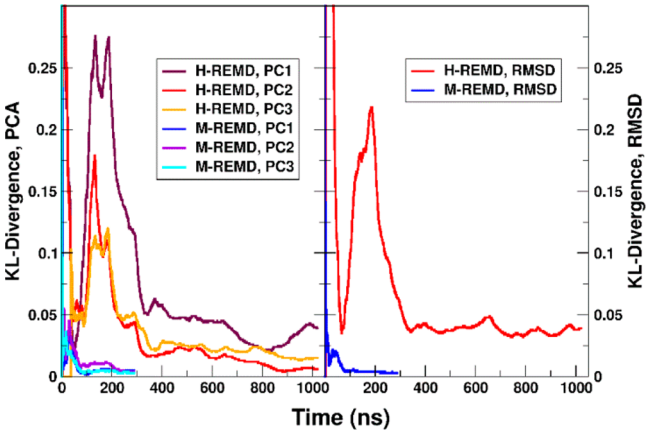


Figure 6. Kullback–Leibler divergence analysis. Time dependent Kullback–Leibler divergence. Left: first three principal components from the H-REMD and M-REMD simulations. H-REMD principal components 1, 2, and 3 are maroon, red, and orange. M-REMD principal components 1, 2, and 3 are blue, purple, and cyan. Right: mass-weighted heavy atom RMSD to an A-form reference from the M-REMD (blue) and H-REMD (red) simulations.

REMD converges to the final ensemble quickly, reaching a divergence of almost zero after 250 ns. In contrast, the H-REMD remains unconverged (particularly in the first PC) even after more than 1 μs per replica. This pattern is echoed in the KL divergence analysis of the mass weighted RMSD to an A-form reference for the M-REMD and H-REMD simulations, where the H-REMD simulations are clearly not converged whereas the M-REMD simulations appear well-converged.

In addition to poorer convergence, even with equivalent aggregate MD simulation times (i.e., using 192 H-REMD replicas to match the sampling of the M-REMD simulation), H-REMD tends to underpopulate the peak at 4.0 Å (the 1_3-basepair structure) and show significant differences between resulting ensembles (Supporting Information Figure S9). For example, the H-REMD simulations yield several low-populated clusters that differ from more populated clusters solely due to glycosidic X flips and shifts to more optimal π-stacking geometries. The T-REMD samples few of these extended structures, whereas the M-REMD cluster structures contain a mix of both T-REMD-like structures and lower population H-REMD-like structures (Supporting Information Table S1 and Figure S4).

DISCUSSION

Sampling of alternate X-flipped/π-stacked structures in the H-REMD simulations is likely a direct result of our choice to bias the dihedral force constant. In knocking down the dihedral barriers, the nonbonded terms of the force field have a greater effect on the structure. At high biasing levels, less energy is required to overcome unfavorable torsional barriers to form favorable nonbonded contacts. Conformations which contain flipped X dihedrals, for example, are seen more often in the H-REMD simulations compared to T-REMD simulations. When the temperature is added into the sampling with M-REMD, the higher temperature replicas are less likely to become trapped because they can more easily break these favorable nonbonded interactions.

The choice of bias and the resulting overemphasis of nonbonded interactions may result in the nonconvergence of the H-REMD simulations. If specific structures (e.g., the X-flipped structures) are favored at high biases, rearranging to a canonical X value is prohibitive in the unbiased replica where the dihedral barriers are at their full heights. Supporting Information Figure S9 illustrates this point; through clustering, a X-flipped structure becomes representative of one of the four major clusters in H-REMD. This phenomenon could possibly be remedied by increasing the number of replicas at low biasing levels or by providing more time at low to no bias for X dihedrals to repopulate canonical values.

A summary of the populations sampled by each REMD method is shown in Figure 7. Here the danger in drawing conclusions about conformational preferences from unconverged data becomes evident. The relative differences in population at the 3.0 and 5.0 Å peak reflect differences in the NMR minor and intercalated-anti conformations, two of the highest populated clusters. If only H-REMD was used to study

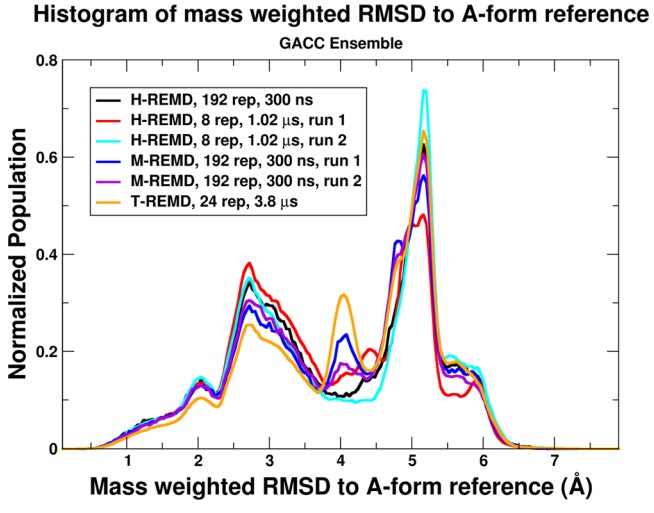


Figure 7. RMSD population histograms at 300 K from the various REMD simulations. Shown are the normalized populations (y-axis) of structures at particular mass-weighted RMSD values (x-axis) of all atoms in residues 1–4 to a reference structure (A-form RNA) for the various REMD simulations.

dx.doi.org/10.1021/ct400862k | J. Chem. Theory Comput. 2014, 10, 492–499

r(GACC), the 1_3-basepair conformation, represented by the peak at 4.0 Å, would most likely not be seen because this structure seems to be the most difficult to converge on the basis of the M-REMD and T-REMD simulations. The efficiency of the M-REMD also becomes obvious; with the same amount of aggregate sampling, the 1_3-basepair conformation is not sampled in H-REMD (black), whereas the two independent M-REMD simulations appear more converged in this region than any other simulations performed here.

An additional concern raised in review and in presentation of these results is the influence of salt on the conformational ensemble and the potential for biasing the population due to the inclusion of only net-neutralizing Na^+ ions. To address this, additional M-REMD simulations equivalent to the previously discussed simulations were performed in 100 mM NaCl and also 100 mM KCl using the Joung/Cheatham ion parameters.[39] As shown in Figure S10 and described in the Supporting Information, the change in salt concentration and ion identity appears to have a minimal effect on the RMSD population histograms suggesting a nearly equivalent conformational ensemble comparing net-neutralizing Na^+ to 100 mM NaCl or 100 mM KCl.

As previously mentioned, rigorous analysis of simulation convergence is necessary to properly validate force fields, so after having apparently converged our simulations, we are now in a unique position to truly interrogate our current parameters. Though we generate a very well-converged ensemble for r(GACC), it is apparent that our current force field does not adequately reproduce the experimental NMR data seen for this tetranucleotide.[13] An A-form major conformation was observed via NMR, as well as a minor conformation characterized by a flipped C4 base. In our converged M-REMD simulations, the populations of the NMR major and minor conformations were 18% and 24%, respectively. The intercalated-anti structure accounted for 12% of the population. This structure has been seen in previous simulations,[10,13] yet there is no experimental evidence showing that it should be this highly populated. The structure is characterized by intercalated stacking and increased number of hydrogen bonds. On the basis of these observations, we are in the process of both comparing to other existing RNA force fields and attempting to refine various force field terms, such as altering charges, dihedral terms, and stacking interactions to help guide revisions to the Amber force fields.[6]

CONCLUSIONS

In summary, we have demonstrated that apparent convergence for very flexible systems can be more quickly achieved by combining temperature and dihedral force constant biasing in M-REMD. By itself, the H-REMD simulation did not converge in an adequate time frame; overcoming the high-bias replica's preference for π stacked and X flipped structures at low or unbiased replicas was prohibitively slow. By addition of the temperature dimension, a well-converged ensemble, determined by various metrics presented here, was achieved within 300 ns of sampling (per the 192 replicas). As M-REMD tends to utilize more replicas, the time to solution in wall-clock hours improves if sufficient access to large-scale, high performance computational resources is available. In the case of AMBER simulations, speed-up is especially impressive with access to large-scale GPU resources such as NCSA Blue Waters and the XSEDE Stampede and Keeneland resources where these simulations were performed. This method will be used to investigate the ensembles of other (or all) tetranucleotide sequences, providing a comprehensive look at the underlying structure preference of current force fields.

ASSOCIATED CONTENT

Supporting Information

A summary of simulation methods, additional convergence analysis, cluster population as a function of time, molecular graphics of r(GACC) conformations, and analysis scripts are provided. This material is available free of charge via the Internet at http://pubs.acs.org

AUTHOR INFORMATION

Corresponding Author

*T. E. Cheatham: e-mail, tec3@utah.edu.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This research was enabled by the Blue Waters sustained-petascale computing project (NSF OCI 07-25070 and PRAC OCI-1036208), the NSF Extreme Science and Engineering Discovery Environment (XSEDE, OCI-1053575) and allocation MCA01S027P, the Center for High Performance Computing at the University of Utah, and substantial friendly user time on Keeneland and Titan. Research funding comes from National Institutes of Health R01-GM098102 (T.E.C.) and National Science Foundation OCI-1147910 (A.E.R.).

ABBREVIATIONS

REMD, replica exchange molecular dynamics; *T*, temperature; *H*, Hamiltonian; M, multidimensional; RNA, ribonucleic acid

REFERENCES

(1) Meister, G. *RNA biology: an introduction*; Wiley-VCH: Weinheim, 2011.
(2) McDowell, S. E.; Spacková, N.; Sponer, J.; Walter, N. G. *Biopolymers* **2007**, *85*, 169−184.
(3) Cheatham, T. E., III; Case, D. A. *Biopolymers* **2013**, *99*, 969−977.
(4) Pérez, A.; Marchán, I.; Svozil, D.; Sponer, J.; Cheatham, T. E., III; Laughton, C. A.; Orozco, M. *Biophys. J.* **2007**, *92*, 3817−3829.
(5) Zgarbová, M.; Otyepka, M.; Sponer, J.; Mládek, A.; Banáš, P.; Cheatham, T. E., III; Jurečka, P. *J. Chem. Theory Comput.* **2011**, *7*, 2886−2902.
(6) Steinbrecher, T.; Latzer, J.; Case, D. A. *J. Chem. Theory Comput.* **2012**, *8*, 4405−4412.
(7) Krepl, M.; Réblová, K.; Koča, J.; Sponer, J. *J. Phys. Chem. B* **2013**, *117*, 5540−5555.
(8) Sorin, E. J.; Engelhardt, M. A.; Herschlag, D.; Pande, V. S. *J. Mol. Biol.* **2002**, *317*, 493−506.
(9) Garcia, A. E.; Paschek, D. *J. Am. Chem. Soc.* **2008**, *130*, 815−817.
(10) Henriksen, N. M.; Roe, D. R.; Cheatham, T. E., III. *J. Phys. Chem. B* **2013**, *117*, 4014−4027.
(11) Tubbs, J. D.; Condon, D. E.; Kennedy, S. D.; Hauser, M.; Bevilacqua, P. C.; Turner, D. H. *Biochemistry* **2013**, *52*, 996−1010.
(12) Yildirim, I.; Kennedy, S. D.; Stern, H. A.; Hart, J. M.; Kierzek, R.; Turner, D. H. *J. Chem. Theory Comput.* **2012**, *8*, 172−181.
(13) Yildirim, I.; Stern, H. A.; Tubbs, J. D.; Kennedy, S. D.; Turner, D. H. *J. Phys. Chem. B* **2011**, *115*, 9261−9270.
(14) Meng, Y.; Dashti, D. S.; Roitberg, A. E. *J. Chem. Theory Comput.* **2011**, *7*, 2721−2727.
(15) Sorin, E. J.; Rhee, Y. M.; Nakatani, B. J.; Pande, V. S. *Biophys. J.* **2003**, *85*, 790−803.
(16) Wolf, M. G.; Jongejan, J. A.; Laman, J. D.; de Leeuw, S. W. *J. Phys. Chem. B* **2008**, *112*, 13493−13498.
(17) Kim, S.; Roitberg, A. E. *J. Phys. Chem. B* **2008**, *112*, 1525−1532.

(18) Chen, A. A.; García, A. E. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, 16820−16825.
(19) Paladino, A.; Zangi, R. *Biophys. Chem. 180−181*, 110−118.
(20) Kührová, P.; Banáš, P.; Best, R. B.; Šponer, J.; Otyepka, M. *J. Chem. Theory Comput.* **2013**, *9*, 2115−2125.
(21) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314*, 141−151.
(22) Hansmann, U. H. E. *Chem. Phys. Lett.* **1997**, *281*, 140−150.
(23) Fukunishi, H.; Watanabe, O.; Takada, S. *J. Chem. Phys.* **2002**, *116*, 9058−9067.
(24) Kannan, S.; Zacharias, M. *Proteins* **2007**, *66*, 697−706.
(25) Kara, M.; Zacharias, M. *Biophys. J.* **2013**, *104*, 1089−1097.
(26) Abraham, M. J.; Gready, J. E. *J. Chem. Theory Comput.* **2008**, *4*, 1119−1128.
(27) Periole, X.; Mark, A. E. *J. Chem. Phys.* **2007**, *126*, 014903−1−014903-11.
(28) Okur, A.; Roe, D. R.; Cui, G.; Hornak, V.; Simmerling, C. *J. Chem. Theory Comput.* **2007**, *3*, 557−568.
(29) Okur, A.; Wickstrom, L.; Layten, M.; Geney, R.; Song, K.; Hornak, V.; Simmerling, C. *J. Chem. Theory Comput.* **2006**, *2*, 420−433.
(30) Sanbonmatsu, K. Y.; García, A. E. *Proteins* **2002**, *46*, 225−234.
(31) Zhang, W.; Wu, C.; Duan, Y. *J. Chem. Phys.* **2005**, *123*, 154105-1−154105-9.
(32) Lyman, E.; Zuckerman, D. M. *J. Phys. Chem. B* **2007**, *111*, 12876−12882.
(33) Kullback, S.; Leibler, R. A. *Ann. Math. Stat.* **1951**, *22*, 79−86.
(34) Roe, D. R.; Cheatham, T. E., III. *J. Chem. Theory Comput.* **2013**, *9*, 3084−3095.
(35) Shao, J.; Tanner, S. W.; Thompson, N.; Cheatham, T. E., III. *J. Chem. Theory Comput.* **2007**, *3*, 2312−2334.
(36) Ester, M.; Kriegel, H. P.; Sander, J.; Xu, X. In *Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining*; Simoudis, E., Han, J., Fayyad, U., Eds.; AAAI Press: Menlo Park, CA, 1996; pp 226−231.
(37) Wang, J.; Cieplak, P.; Kollman, P. A. *J. Comput. Chem.* **2000**, *21*, 1049−1074.
(38) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926−935.
(39) Joung, I. S.; Cheatham, T. E., III. *J. Phys. Chem. B* **2008**, *112*, 9020−9041.
(40) Case, D. A.; Darden, T. A.; Cheatham, T. E., III; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Walker, R. C.; Zhang, W.; Merz, K. M.; Roberts, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Swails, J.; Goetz, A. W.; Kolossváry, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; Wolf, R. M.; Liu, J.; Wu, X.; Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Cai, Q.; Ye, X.; Wang, J.; Hsieh, M. J.; Cui, G.; Roe, D. R.; Mathews, D. H.; Seetin, M. G.; Salomon-Ferrer, R.; Sagui, C.; Babin, V.; Luchko, T.; Gusarov, S.; Kovalenko, A.; Kollman, P. A. *AMBER 12*; University of California: San Francisco, 2012.
(41) Götz, A. W.; Williamson, M. J.; Xu, D.; Poole, D.; Le Grand, S.; Walker, R. C. *J. Chem. Theory Comput.* **2012**, *8*, 1542−1555.
(42) Loncharich, R. J.; Brooks, B. R.; Pastor, R. W. *Biopolymers* **1992**, *32*, 523−535.
(43) Sindhikara, D. J.; Kim, S.; Voter, A. F.; Roitberg, A. E. *J. Chem. Theory Comput.* **2009**, *5*, 1624−1631.
(44) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327−341.
(45) Le Grand, S.; Götz, A. W.; Walker, R. C. *Comput. Phys. Commun.* **2013**, *184*, 374−380.
(46) Salomon-Ferrer, R.; Goetz, A. W.; Poole, D.; Le Grand, S.; Walker, R. C. *J. Chem. Theory Comput.* **2013**, *9*, 3878−3888.
(47) Case, D. A.; Cheatham, T. E., III; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. *J. Comput. Chem.* **2005**, *26*, 1668−1688.
(48) Sugita, Y.; Kitao, A.; Okamoto, Y. *J. Chem. Phys.* **2000**, *113*, 6042−6051.