Supporting Information

To

Computer Folding of RNA Tetraloops? What Is Wrong with the Force Fields?

Petra Kührová,[§] *Robert B. Best,*[‡] *Giovanni Bussi,*[†] *Sandro Botaro,*[†] *Jiří Šponer,*[§£]* *Michal Otyepka*[§£] *and Pavel Banáš,*[§£]*

[§]Regional Centre of Advanced Technologies and Materials, Department of Physical Chemistry, Faculty of Science, Palacky University Olomouc, 17. listopadu 12, 771 46 Olomouc, Czech Republic e-mail: pavel.banas@upol.cz

[‡]Laboratory of Chemical Physics, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD 20892-0520

[†]Scuola Internazionale Superiore di Studi Avanzati, Via Bonomea 265, 34136 Trieste, Italy

[£]Institute of Biophysics, Academy of Sciences of the Czech Republic, Kralovopolska 135, 612 65 Brno, Czech Republic

*corresponding authors

Overview of used methods

Replica exchange molecular dynamics (REMD) simulation algorithm runs multiple independent MD simulations in parallel over a range of increasing temperatures. The exchanges between neighboring temperatures are taken at the fixed time intervals. The probability of acceptance of exchanges is determined by Metropolis probability criterion

$$P(i \leftrightarrow j) = \min\left(1, e^{-(\beta_j - \beta_i)(U_i - U_j)}\right) \quad (1)$$

where P is the probability of an exchange between two neighboring replicas i and j; $\beta_i = 1/k_B T_i$ and $\beta_j = 1/k_B T_j$, where k_B is Boltzmann constant, T_i and T_j are temperatures, and U_i and U_j are potential energies of replicas i and j, respectively.

Replica exchange with solute scaling (in a REST2 variant) is a method based on modification of Hamiltonian of the solute atoms.[1] In this method, only the solute atoms are effectively heated up while the solvent molecules remain cold in higher temperature replicas, i.e., the solute energy is scaled whereas the solvent energy is left unchanged. The Hamiltonian is scaled by a factor λ, which is equivalent to a scaling of the temperature by a factor λ^{-1}. The solute-solute interactions are scaled with a factor $\lambda < 1$, solvent-solvent interactions are not scaled, and solute-solvent interactions are scaled by scaling factor between λ and 1. All replicas are run at the same temperature T_0 and the potential energy for replica m is scaled

$$U_m(X) = \lambda U_{ss}(X) + \sqrt{\lambda} U_{sw}(X) + U_{ww}(X) \quad (2)$$

where X represents configuration of the whole system; U_{ss} is the solute intra-molecular energy; U_{sw} is the interaction energy between solute and water; U_{ww} is the interaction energy between water molecules; λ is the scaling factor defined as $\lambda = T_0/T_m$, here, T_m means the effective temperature of the solute with the unscaled potential energy. The exchanges between replicas are attempted at specific intervals.

Table S1: List of native hydrogen bonds used in REMD simulation of GAGA TL with HBfix. For numbering of bases see Figure 1 in main text.

	Bonding partners		
Residue	Name of atom	Name of atom	Residue
G_{S-2}	O6	N4	C_{S+2}
	N1	N3	
	N2	O2	
C_{S-1}	N4	O6	G_{S+1}
	N3	N1	
	O2	N2	
G_{L1}	N2	OP2	A_{L4}
	N2	N7	
	O2’	N7	G_{L3}

Metadynamics is a well-established simulation method used to improve sampling via overcoming energy barriers. This method was used to calculation of free energy surface of GAGA TL. The biasing potential was calculated according to the WT-MetaD scheme using the following formula:

$$V(s,t) = \sum_{t'=0,\tau_G,2\tau_G,\ldots}^{t'<t} \omega\tau_G e^{-V\left(s\left(q(t'),t'\right)\right)/\Delta T} e^{-\sum_{i=1}^{2}\frac{\left(s_i(q)-s_i\left(q(t')\right)\right)^2}{2\sigma_i^2}} \qquad (3)$$

where the deposition rate, ω, and deposition stripe, τ_G, of the Gaussian hills was set to 0.478 kcal/mol·ps (2.0 kJ/mol·ps) and 1 ps, respectively. The bias factor $(T+\Delta T)/T$ was set to 15 and the final FES was calculated as follows:

$$F(s,t) = -\frac{T+\Delta T}{\Delta T}\big(V(s,t) - C(t)\big) \quad (4)$$

Thus, the CVs were sampled at a fictitious temperature $T+\Delta T$ of 3000 K. Two CVs, s_{i}, were employed: H_{core}, which included all native hydrogen bonds of GAGA TL, and *RMSD* of the loop region of GAGA TL. The Gaussian widths were set to 0.2 arbitrary unit, and 0.5 Å for H_{core}, and RMSD respectively.

The H_{core} CV was calculated using the switching function as follows:

$$H_{\mathrm{core}} = \sum_i \frac{1-\left(\frac{r_i}{r_0}\right)^n}{1-\left(\frac{r_i}{r_0}\right)^m} \quad (5)$$

where r_0 was set to 2.5 Å, the n and m parameters were set to 6 and 12, respectively, the index i corresponded to the nine hydrogen bonds involved in base pairing of the three GC base pairs of the stem region and additional three hydrogen bonds of the loop region, and r_i was the distance between the hydrogen acceptor and hydrogen atom bound to the hydrogen donor of the abovementioned hydrogen bonds (see Table S2 for the list of atoms involved in definition of H_{core}).

Table S2: List of native hydrogen bonds used in the definition of the $\mathrm{H}_{\mathrm{core}}$ CV in both WT-MetaD simulations.

	Bonding partners		
Residue	Name of atom	Name of atom	Residue
$\mathrm{C}_{\mathrm{S-3}}$	N4	O6	$\mathrm{G}_{\mathrm{S+3}}$
	N3	N1	
	O2	N2	
$\mathrm{G}_{\mathrm{S-2}}$	O6	N4	$\mathrm{C}_{\mathrm{S+2}}$
	N1	N3	

	N2	O2	
C_{S-1}	N4	O6	G_{S+1}
	N3	N1	
	O2	N2	
G_{L1}	N2	OP2	A_{L4}
	N2	N7	
	O2’	N7	G_{L3}

Simulation protocol of classical MD simulation

The molecular dynamics simulation was carried out using pmemd.cuda (AMBER 12.0 package[2]). The solvated system was minimized optimizing the waters and ions, while the position of RNA molecule remained constrained. Subsequently, all RNA atoms were frozen and the solvent molecules with counter-ions were allowed to move during a 500-ps long MD run under NpT conditions (p= 1 atm., T=298.16 K) in order to relax the total density. After this, the RNA molecule were relaxed by several minimization runs, with decreasing force constant applied to the suger-phopshate backbone atoms. After the relaxation, the system was heated in two steps: the first step involved heating under NVT conditions for 100 ps, whereas the second step involved density equilibration under NpT conditions for and additional 100 ps. The particle-mesh Ewald (PME) method for treating electrostatic interactions was used, and the simulation was performed under periodic boundary conditions in the [NpT] ensemble at 298.16 K using weak-coupling Berendsen thermostat[3] with coupling time of a 1 ps. The SHAKE algorithm, with a tolerance of 10^{-5} Å, was used to fix the positions of all hydrogen atoms, and a 10.0 Å cut-off was applied to non-bonding interactions to allow a 2 fs integration step. For REMD simulation, we have chosen 64 structures. The chosen structures corresponded to restart files every 10 ns. In order to unify the box sizes before starting REMD/REST2 simulations that were slightly diversified during NpT simulation, we manually rewritten box size information in all replicas, and subsequently, all structures were re-equilibrated in short 100 ps NVT simulation.

Setting of a HBfix additional potential supporting hydrogen bonds

In order to support hydrogen bonding, which is inherently understabilized in the contemporary force fields due to lack of true electronic structure redistributions, we introduced locally acting HBfix potential. The potential is applied to heavy-atom distances of the supported hydrogen bonds and is defined by following equation:

$$E(r) = \left[\begin{array}{ll} 0 & r < r_{beg} \\ \dfrac{2\eta\left(r - r_{beg}\right)^2}{(1-c)c\left(r_{end} - r_{beg}\right)^2}(1-c) & r_{beg} \le r < r_{beg} + c\left(r_{beg} - r_{beg}\right) \\ \dfrac{2\eta}{(1-c)\left(r_{end} - r_{beg}\right)^2}\left[\left(r - r_{beg}\right)^2 + 2\left(r_{end} - r_{beg}\right)\left(r - r_{beg}\right) - c\left(r_{end} - r_{beg}\right)^2\right] & r_{beg} + c\left(r_{beg} - r_{beg}\right) \le r < r_{end} \\ \varepsilon & r > r_{end} \end{array} \right.$$

where r_{beg} and r_{end} delimit the region, in which HBfix potential is not constant, c defines the position of the inflex point, so that its position is $r_{beg} + c(r_{end} - r_{beg})$. The η parameter defines total energy support for the hydrogen bonds. This potential might be easily implemented into AMBER simulations. Namely, it can be composed by two flat-bottom restraints with following settings:

```
 &rst iat=first_atom_id, second_atom_id,
      iresid=0,nstep1=0,nstep2=0,
      irstyp=0,ifvari=0,ninc=0,imult=0,ir6=0,ifntyp=0,
      r1=0.0, r2=r_beg, r3=r_beg, r4=r_end, rk2=0.0000, rk3=-
2*eta/((1-c)*(R_end-R_beg)^2),
/
 &rst iat=first_atom_id, second_atom_id,
      iresid=0,nstep1=0,nstep2=0,
      irstyp=0,ifvari=0,ninc=0,imult=0,ir6=0,ifntyp=0,
     r1=0.0, r2=r_beg, r3=r_beg, r4=r_beg+c(r_end - r_beg),
rk2=0.0000, rk3= 2*eta/(c*(1-c)*(R_end-R_beg)^2),
/
```

so for 1 kcal/mol restraint between atoms 1 and 4 it reads as follows

```
 &rst iat=1, 4,
      iresid=0,nstep1=0,nstep2=0,
      irstyp=0,ifvari=0,ninc=0,imult=0,ir6=0,ifntyp=0,
      r1=0.0, r2=3.0, r3=3.0, r4=4.0, rk2=0.0000, rk3=-10.0000,
/
 &rst iat=1, 4,
      iresid=0,nstep1=0,nstep2=0,
      irstyp=0,ifvari=0,ninc=0,imult=0,ir6=0,ifntyp=0,
     r1=0.0, r2=3.0, r3=3.0, r4=3.8, rk2=0.0000, rk3=12.5000,
/
```

Detail of the customization of the clustering algorithm by Laio et al.

The clustering algorithm developed by Laio et al. is based on identification of the cluster centre using so-called decision plot. For each data point (structure) i, a local density ρ_i and a minimal distance δ_i from other points of higher density is calculated. The local density ρ_i of data point i was calculated using Gaussian kernel as follows:

$$\varrho_i = \sum_j e^{\left(\frac{d_{ij}}{d_c}\right)^2} \quad (6)$$

where d_{ij} corresponds to the distance between i-th and j-th datapoint, i.e., εRMSD distance between given structures in our particular case, and d_c is the cutoff equal to 0.35. We used such cutoff value as the Gaussian kernel became insignificant at $2d_c$ distance, which corresponds to εRMSD of 0.7, i.e., the higher expected distance between similar structures. In other words, the ρ_i is related to the number of points that are structurally similar to the point i. δ_i is defined as the minimum distance between point i and any other point with higher density

$$\delta_i = \min_{j:\varrho_j > \varrho_i} \left(d_{ij}\right) \quad (7)$$

The δ_i and ρ_i values and subsequently used to create decision plot, while the potential cluster centres are points with high δ_i and high ρ_i values, i.e., with high value of $\gamma_i = \rho_i \delta_i$. In original algorithm the cluster data points and separated from the cluster hull represented noise based on the density cutoff defined by density of the saddle points between clusters.

In our particular case, such definition of cluster points and cluster hull is problematic as the εRMSD metric is defined by Euclidian distance in vector space of very high dimension, where however, the components are defined in small range compared to distance cutoff $d_{c.}$ As a consequence, clusters in such space are rather isolated and saddle points cannot be often even well defined. Therefore we decided to customize the original algorithm as follows. We create the decision plot as originally suggested. Subsequently, we take a point of highest γ_i and assign the other points to this cluster based on two rules: (i) its nearest neighbour of higher density is also member of this cluster, and (ii) its nearest member of higher density is closer than 0.7 εRMSD. Using this definition, it is guaranteed that any other potential cluster center cannot be assigned to another cluster as it is assumed to have δ_i higher than 0.7 and thus can never meet the second requirement. It should be note that such modification is equivalent with the original algorithm without cluster/hull recognition and selecting all points with δ_i above 0.7 as potential cluster centres. In this way, we typically obtain high number of cluster, so we calculated size of the clusters and ignore those representing less than 1% of entire dataset.

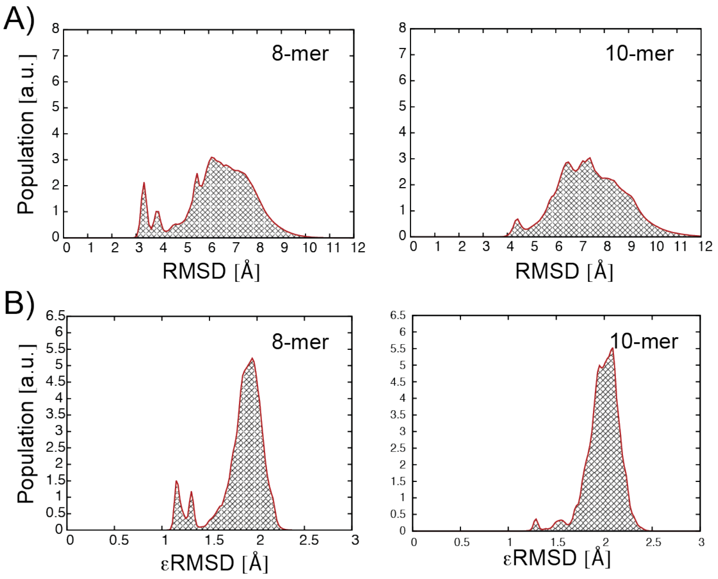

Figure S1. (A) All atoms RMSD and (B) εRMSD histogram profiles calculated over all replicas of T-REMD simulation of 8-mer and 10-mer under net-neutral conditions (i.e., χ_{OL3}-neut. and χ_{OL3}-neut.-10mer, see Table 1 in the main text), respectively.

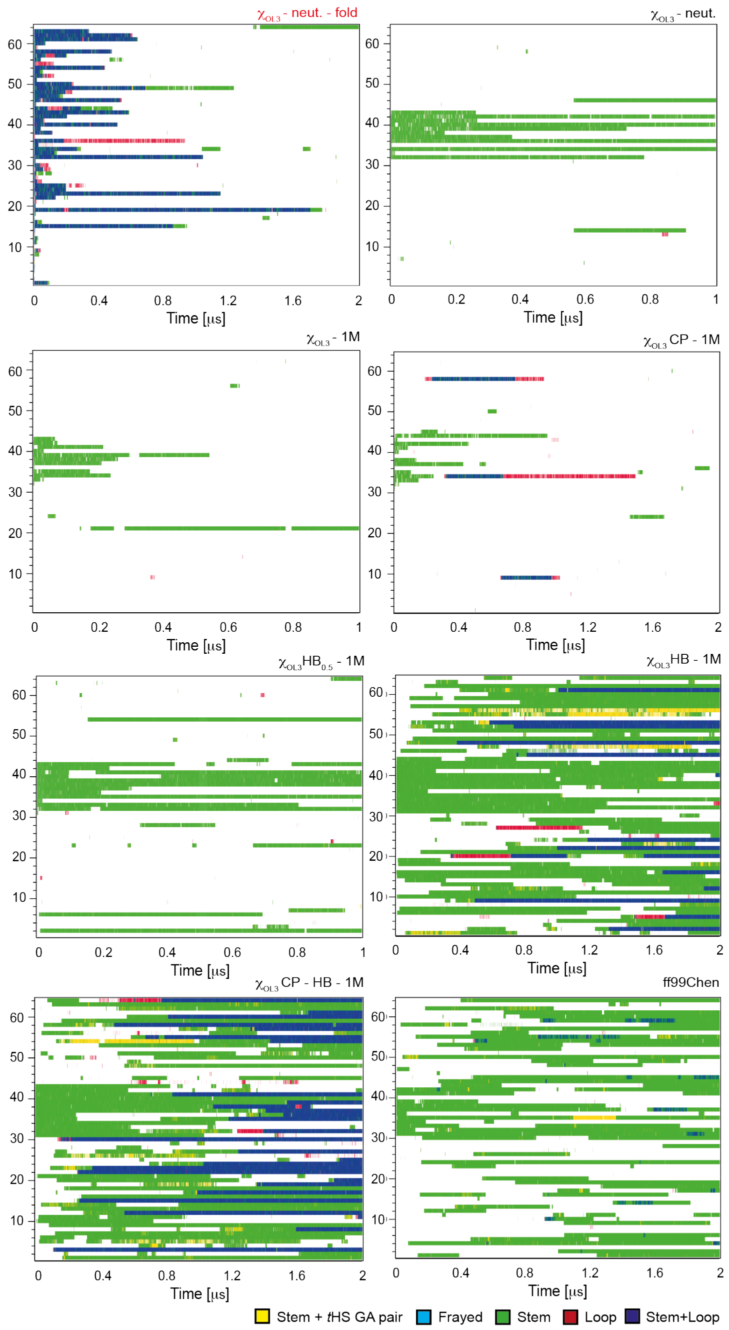

Figure S2. The time evolution of structures with folded stem, folded stem with frayed terminal base pair, stem with *t*HS GA base pair, loop and loop with stem calculated over all T-REMD 8-mer simulations. Each of the horizontal stripes corresponds to one to one of 64 unique replica simulations. The colours of the stripes correspond to the given state.

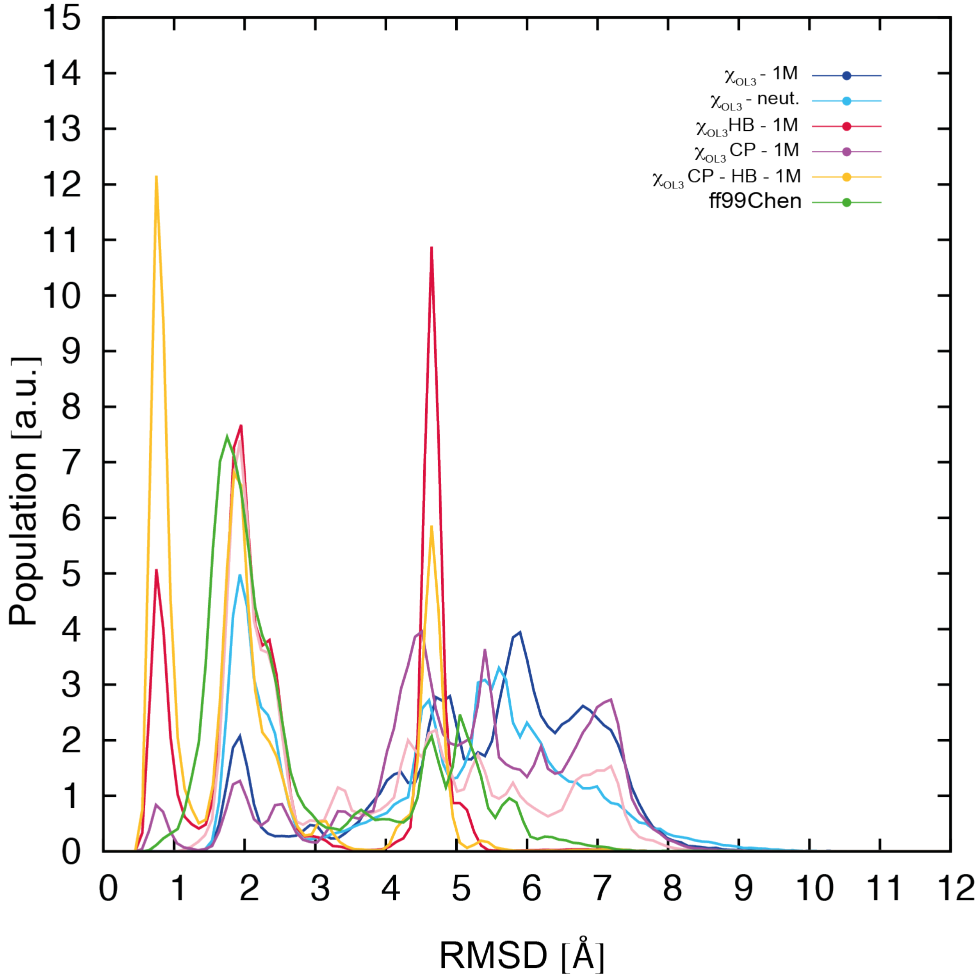

Figure S3. RMSD histogram profiles calculated over all replicas of T-REMD 8-mer simulation. RMSD calculation included all non-hydrogen atoms in the two C-G base pairs of stem, and the backbone atoms of the loop segment.

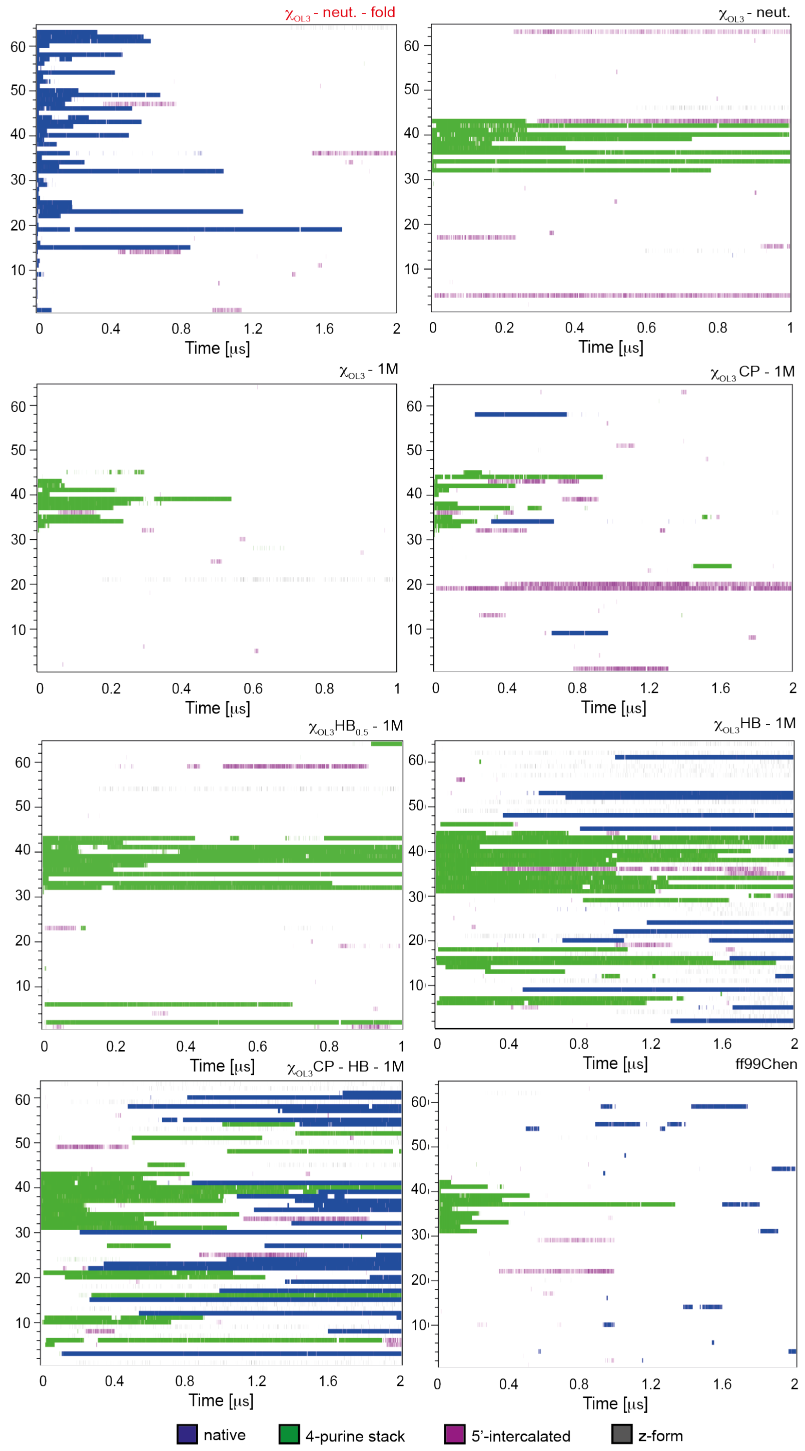

Figure S4. Cluster's development calculated over all T-REMD 8-mer simulations. Each of the horizontal stripes corresponds to one to one of 64 unique replica simulations. The colours of the stripes correspond to the given cluster.

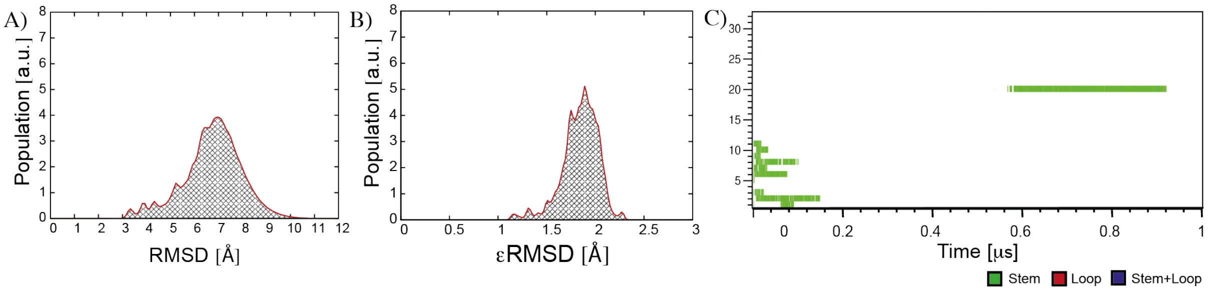

Figure S5. (A) All atoms RMSD and (B) εRMSD histogram profiles calculated over all replicas of REST2 8-mer simulation. (C) The time evolution of structures with folded stem, loop and loop with stem. Each of the horizontal stripes corresponds to one to one of 32 unique replica simulations. The colours of the stripes correspond to the given state.

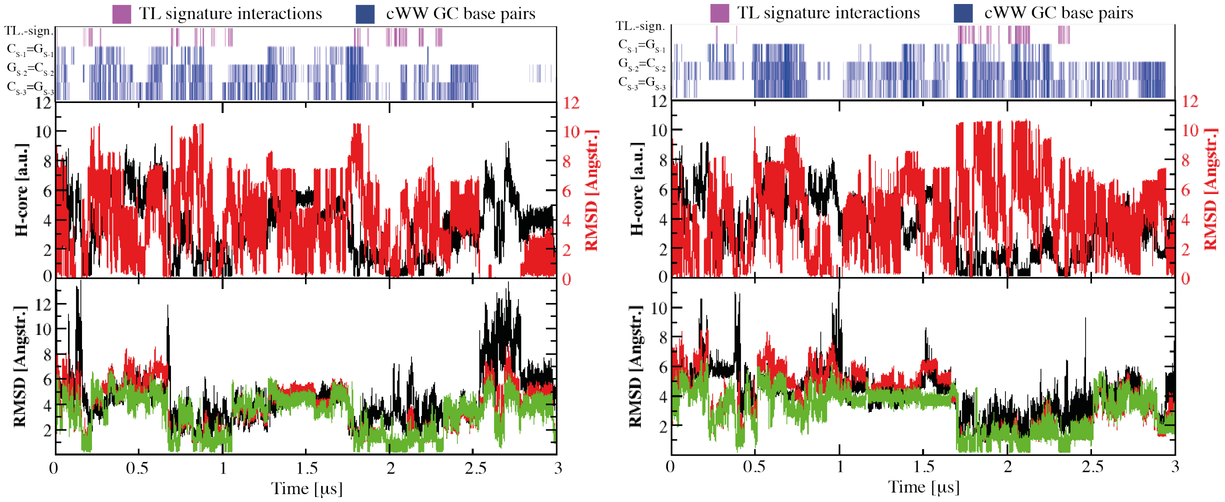

Figure S6. The development of key structural parameters obtained from our WT-MetaD simulations of GAGA TL. Red and blue stripes in the top panels indicate the presence of TL signature interactions and GC base pairs of the stem, respectively, both defined on the basis of hydrogen bonding with 4.0 A° cutoff for heavy-atom distance. The middle panels show the time evolution of the Hcore (black) and RMSD (red) CVs, see Methods in the main text. The lower left panels show evolution of the root-mean-square deviation (RMSD) of the whole RNA hairpin (black), the TL (red), and the tripurine $A_{L2}G_{L3}A_{L4}$ stack (green).

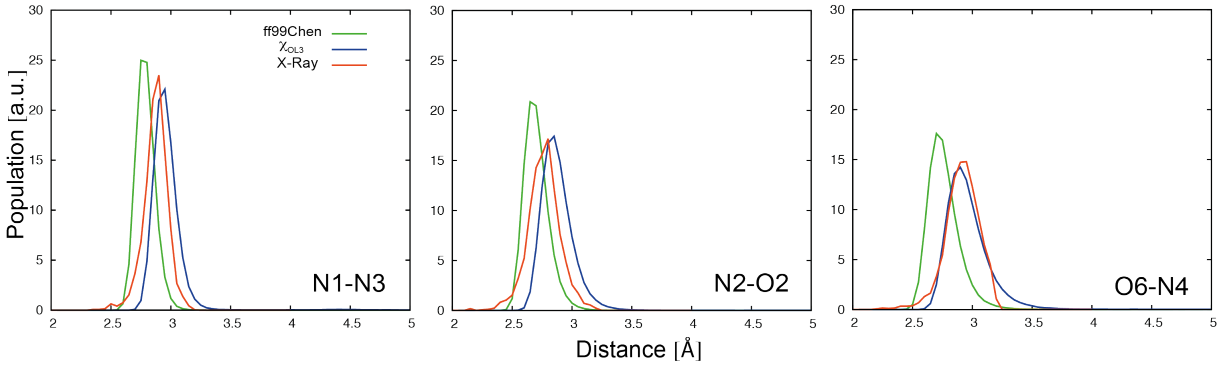

Figure S7. The statistical distribution of the distances between heavy atoms of both hydrogen bonds of the GC base pair as observed in MD simulations of r(CGCGC) A-RNA douplex in χ_{OL3} (blue lines) and ff99Chen (green lines) force fields, and in all X-ray structures of RNAs from the protein data bank having resolution below 2.5 Å.

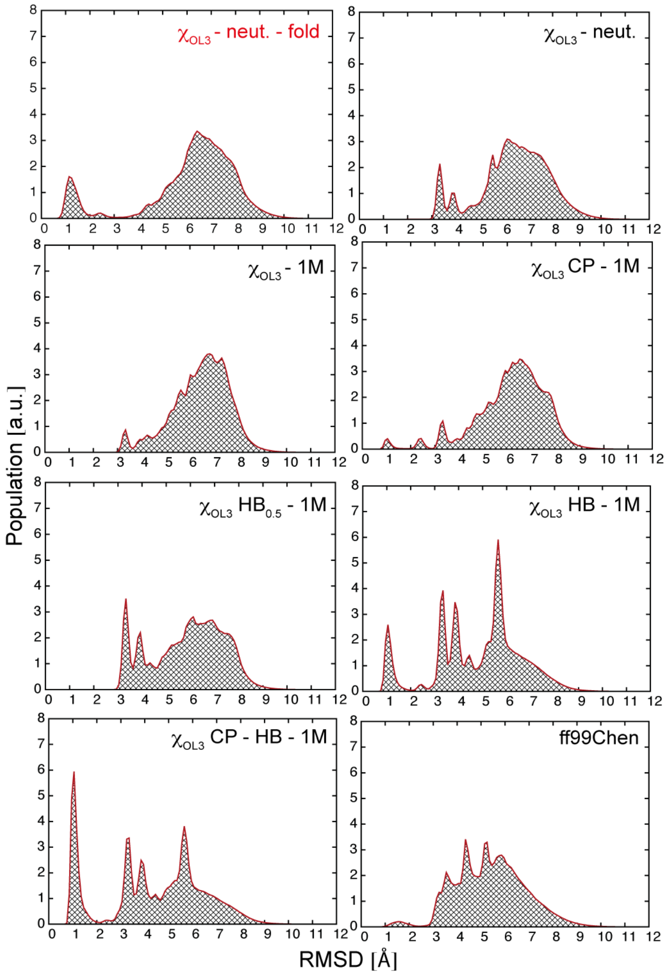

Figure S8. All atoms RMSD histogram profiles calculated over all replicas of T-REMD 8-mer simulations.

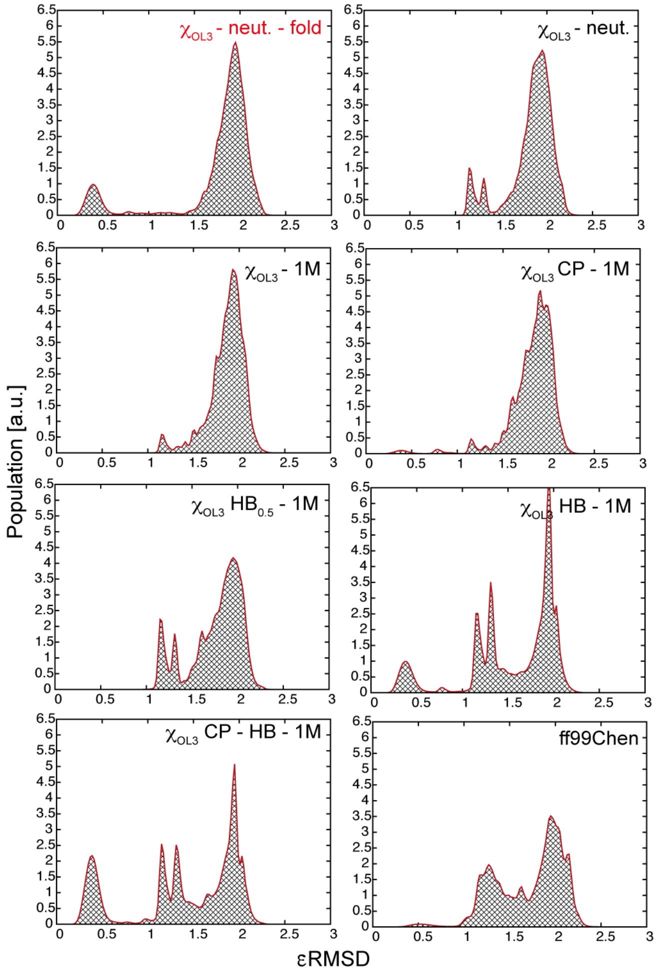

Figure S9. εRMSD histogram profiles calculated over all replicas of T-REMD 8-mer simulations.

1. Wang, L.; Friesner, R. A.; Berne, B. J., Replica Exchange with Solute Scaling: A More Efficient Version of Replica Exchange with Solute Tempering (REST2) (vol 115, pg 9431, 2011). *J Phys Chem B* **2011,** *115* (38), 11305-11305.

2. Salomon-Ferrer, R.; Case, D. A.; Walker, R. C., An overview of the Amber biomolecular simulation package. *Wires Comput Mol Sci* **2013,** *3* (2), 198-210.

3. Berendsen, H. J. C.; Postma, J. P. M.; Vangunsteren, W. F.; Dinola, A.; Haak, J. R., Molecular-Dynamics with Coupling to an External Bath. *J Chem Phys* **1984,** *81* (8), 3684-3690.