Electronic Supplementary Material (ESI) for Physical Chemistry Chemical Physics.
This journal is © the Owner Societies 2023

SUPPORTING INFORMATION

Molecular insights into the Stereospecificity of Arginine in RNA tetraloop folding

Amal Vijay and Arnab Mukherjee*

arnab.mukherjee@iiserpune.ac.in

Department of Chemistry, Indian Institute of Science Education and Research, Pune-411008, India.

I. GAAA 12-mer RNA Tetraloop simulations in the absence of amino acids

In the absence of amino acids, we have conducted several well-tempered metadynamics simulations of 12-mer GAAA RNA tetraloop system by choosing different combinations of collective variables (CVs) to identify the optimum collective variable (CV) combinations and also to get a reasonable free energy estimate of folding–unfolding events. We have encountered problems related to the existence of false minima states, lack of convergence and CV-dependent evolution of states in these simulations. The details regarding a few of these biased simulations are explained below.

a) Choice of Collective Variables. In the initial simulations of 500 ns, we took Radius of gyration of RNA (Rg) and the total number of hydrogens bonds in the stem region of RNA as CVs. The time evolution of hydrogen bonds (HB) and radius of gyration (Rg) of the RNA motif is shown in the Fig. S1. We have found from the trajectory that the system was unable to explore the complete unfolded state from the starting native folded state. Figure S1a-b show the changes of CVs with time. The corresponding free energy surface is shown in Fig. S1c.

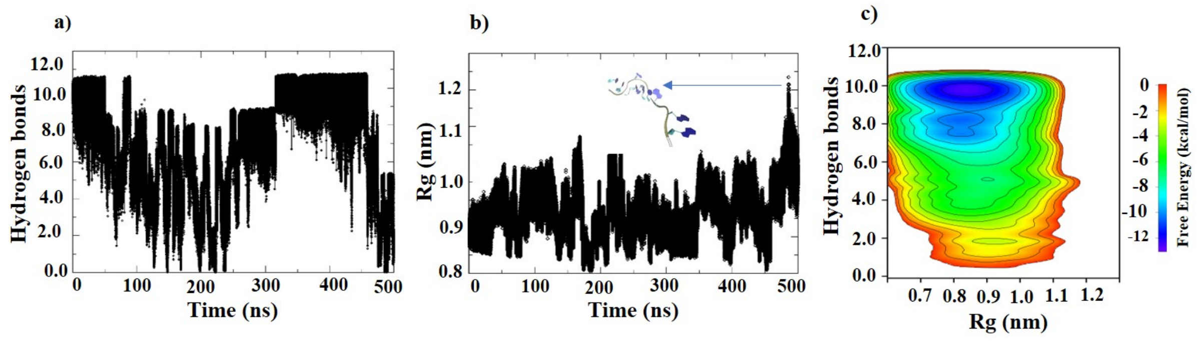

Figure S1. a) Evolution of hydrogen bonds with respect to time in the biased simulations using hydrogen bonds and Rg as CV's b) Evolution of Rg with respect to time in the biased simulations using hydrogen bonds and Rg as CV's c) Free energy plot corresponds to 500 ns simulation with respect to biased variables (Hydrogen bonds and Rg)

To understand the dynamics of RNA in the folded state, we have clustered the trajectory by which the native state is visited by the system in accordance with the ranges of values in biased CV's ($9.7 <$ Hydrogen bonds <10 and $0.89 <$ Rg < 0.93). We have obtained 4 relevant clusters by which the system visited the initial state. The time periods by which the system visited the native state is shown in the Fig. S2a and Fig. S2b by labels A, B, C and D. To understand the clusters in more detail, we have calculated the torsional parameters χ, ζ, ε and δ (Fig. S3) for RNA structures specifically in the residues of stem and loop region of RNA in these clusters. The distributions of these torsional parameters averaged over residues, specifically in the stem region and loop region are shown in the Fig. S4a and Fig. S4b. We found significant changes in the clusters of the loop region in the clusters indicating the need to modify the given description of CV's.

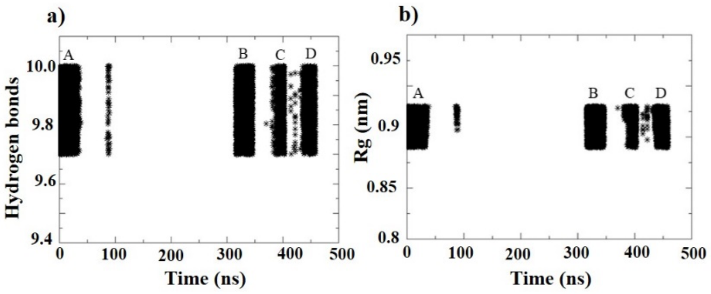

Figure S2. a) Time periods of clusters of structures having equilibrium confirmation of Hydrogen bonds in native folded state for the simulation using hydrogen bonds and Rg as CV's. b) Time periods of clusters of structures having equilibrium confirmation of Rg in native folded state for the simulation using hydrogen bonds and Rg as CV's. Four clusters are indicated by A, B, Cand D.

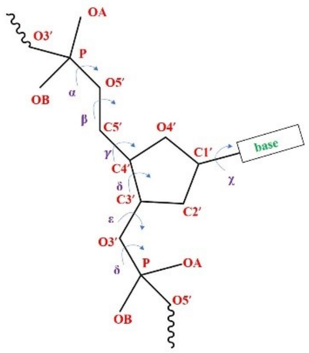

Figure S3. Structural representation of various dihedral parameters in a nucleotide unit.

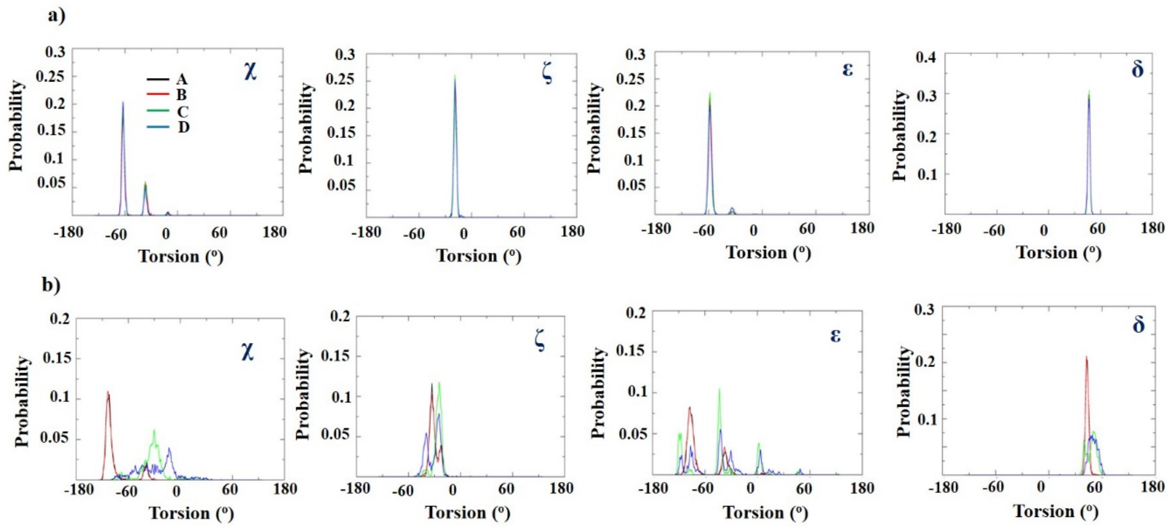

Figure S4. a) Torsional parameters (χ, ζ, ε and δ) in the clusters A, B, C and D averaged over the residues in the stem region of RNA. b) Torsional parameters (χ, ζ, ε and δ) in the clusters A, B, C and D averaged over the residues in the loop region of RNA.

Later, we found that ermsd parameter (discussed below) gave a better description of the loop motion; as the free energy profiles reweighted against ermd captured discrete minima. In our HB collective variable, the hydrogen bonds in the loop region are also considered. The reweighted free energy surface with modified HB definition, Rg, and ermsd parameter is shown in the Fig. S5. The new description of CV's were able to distinguish two important minima states M1 and M2.

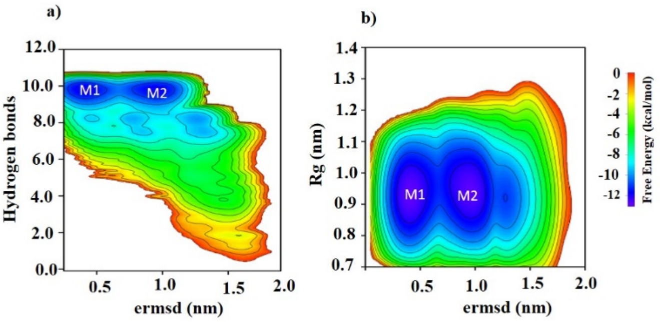

Figure S5. Reweighted energy surface with respect to hydrogen bonds (modified), Rg and ermsd in the well-tempered simulations using hydrogen bonds and Rg as CV's. Two minima states are shown by M1 and M2.

b) ermsd as a collective variable

In the well-tempered metadynamics simulations biased with respect to ermsd, we found that the system tries to stay in the native state and never tried to push the system towards unfolded states. The time evolution of ermsd, HB, and Rg in the trajectory is shown in the Fig. S6. An unusual structure formed by RNA in the simulation is shown in the inset of Fig. S6c.

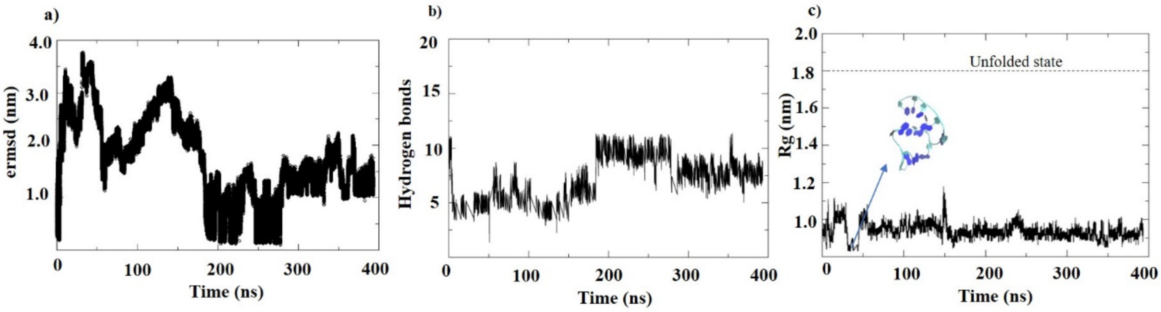

Figure S6. a, b & c) Evolution of ermsd, hydrogen bonds and Rg with respect to time in the biased simulations using ermsd as biased coordinate in the well-tempered metadynamics simulations.

c) End-to-end distance as a collective variable

We have also used end-to-end distance as one of our collective variables to get a completely extended configuration. However, with this, we found that the system fails to come back to the native folded state, once it started unfolding. We also witnessed some complex structures of RNA that could be regarded as false minima, responsible for preventing the folding of RNA in our timescale of simulations. The time evolution of the end-to-end distance and the HB is shown in the Fig. S7a and Fig. S7b, respectively. Some representative structures in the false minima at ~75 ns is shown in the inset of Fig. S7b.

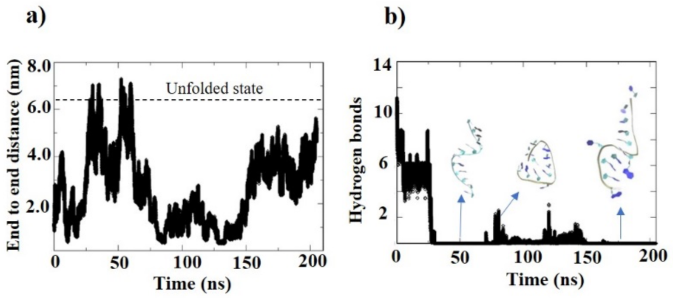

Figure S7. Evolution of end to end distance and hydrogen bonds with respect to time in the biased simulations using end to end distance as biased coordinate in the well-tempered metadynamics simulations. Some of the non-native structures formed by the RNA after 25 ns of the simulations is shown in the insets of fig. S6b.

II. Equations for controlled sampling approach

In well-tempered metadynamics[1] the potential experienced by the system in the collective variable space at a given time is given by

$$V^b(s,t) = \sum_{\tau < t} w(\tau) exp\left[-\frac{\{(s - s(\tau)\}^2}{2(\delta s)^2}\right], \qquad (1)$$

where $w(\tau)$ and δs represents the gaussian height and width parameters to define the gaussian potential at a time interval τ.

In the parallel bias version of metadynamics (PBMetaD)[2], the bias contribution is defined by several one-dimensional bias potentials, each acting on a different CV s_i. The PBMetaD potential is defined by

$$V^{pb}(s_1, \dots, s_n, t) = -\frac{1}{\beta}\ln\sum_{i}^{n} \exp[-\beta V_i^b\ (s_i, t)], \qquad (2)$$

where β = $(k_b T)^{-1}$ and V_i^b is given by Equation (1). The height of the Gaussian bias added along a given dimension is calculated based on the feedback received from other dimensions, as

$$w_i(\tau) = w_i(0) \exp\left[-\frac{V_i^b\ (s_i, \tau)}{k_B \Delta T}\right] P_i(s_i), \qquad (3)$$

where,

$$P_i(s_i) = \frac{exp\ [-\beta V_i^b(s_i, t)]}{\sum_j^n exp\ [-\beta V_j^b(s_j, t)]}, and\ i = 1, \dots., n \qquad (4)$$

$P_i s(i)$ is a feedback function that modulates the Gaussian height based on the bias added along other CVs.

While in the umbrella sampling[3], the harmonic bias is added to the system in the form;

$$w_h^b(s) = \frac{1}{2}\,\kappa_h(s - s_h)^2\,, \qquad h = 1, \ldots M \qquad (5)$$

where κ_h and s_h defines the force constant and mean position of added harmonic bias potential at defined windows h.

In our controlled sampling approach by a combination of umbrella sampling and parallel bias metadynamics, the collective variable space is defined by $s = (s_1, s_2, s_3, s_4, s_5)$. This approach is inspired by the WS-MTD[4] and PBTASS[5] methods. Here, Umbrella sampling potential is experienced by the umbrella sampling coordinate (s_1) at each of the defined windows, along with bias contribution from parallel bias metadynamics along the subsidiary collective variables $(s_2, s_3, s_4, s_5\,)$ in each of these windows (Please see methods for the defined umbrella sampling coordinate and subsidiary collective variables).

So, the overall bias contribution to the system is given by

$$W_h^b(s_1) + V_h^{pb}(s_2, s_3, s_4, s_5, t), \qquad h = 1, \ldots, 20 \qquad (6)$$

where h represents the number of windows taken in consideration for the controlled sampling approach. $W_h^b(s_1)$ and $V_h^{pb}(s_2, s_3, s_4, s_5, t)$ are given by the Equations 2 and 5, respectively.

Thus the modified Hamiltonian in the presence of added bias potentials can be written as

$$H_h(R, P) = H^0(R, P) + W_h^b(s_1) + V_h^{pb}(s_2, s_3, s_4, s_5, t), \qquad h = 1, \ldots, 20 \qquad (7)$$

Where $H^0(R, P)$ defines the underlying potential energy of the system in the absence of any bias.

To obtain the unbiased probability distribution, the reweighting procedure is undertaken as follows. We have concatenated trajectories from each individual windows to have a global trajectory to calculate the bias experienced by each frame in the trajectory. We have reweighted the metadynamics bias potential (contributed by subsidiary collective variables) felt in each of these independent runs above each of the frames in the a global trajectory by the final bias-based reweighting approach for parallel bias metadynamics implemented in PLUMED.[2, 6, 7] This was later combined with the umbrella bias potential, and weight per frame was calculated by self-consistently solving the WHAM equations[8]. Finally, the free energy surface is constructed with respect to defined collective variables using $F(s) = -k_b T\, lnP(s)$.

III. Convergence of free energy profiles

The free energy surface for the folding-unfolding equilibria of RNA tetraloop with respect to all biased collective variables in the presence of L-arginine and D-arginine is shown in the Fig. S8.

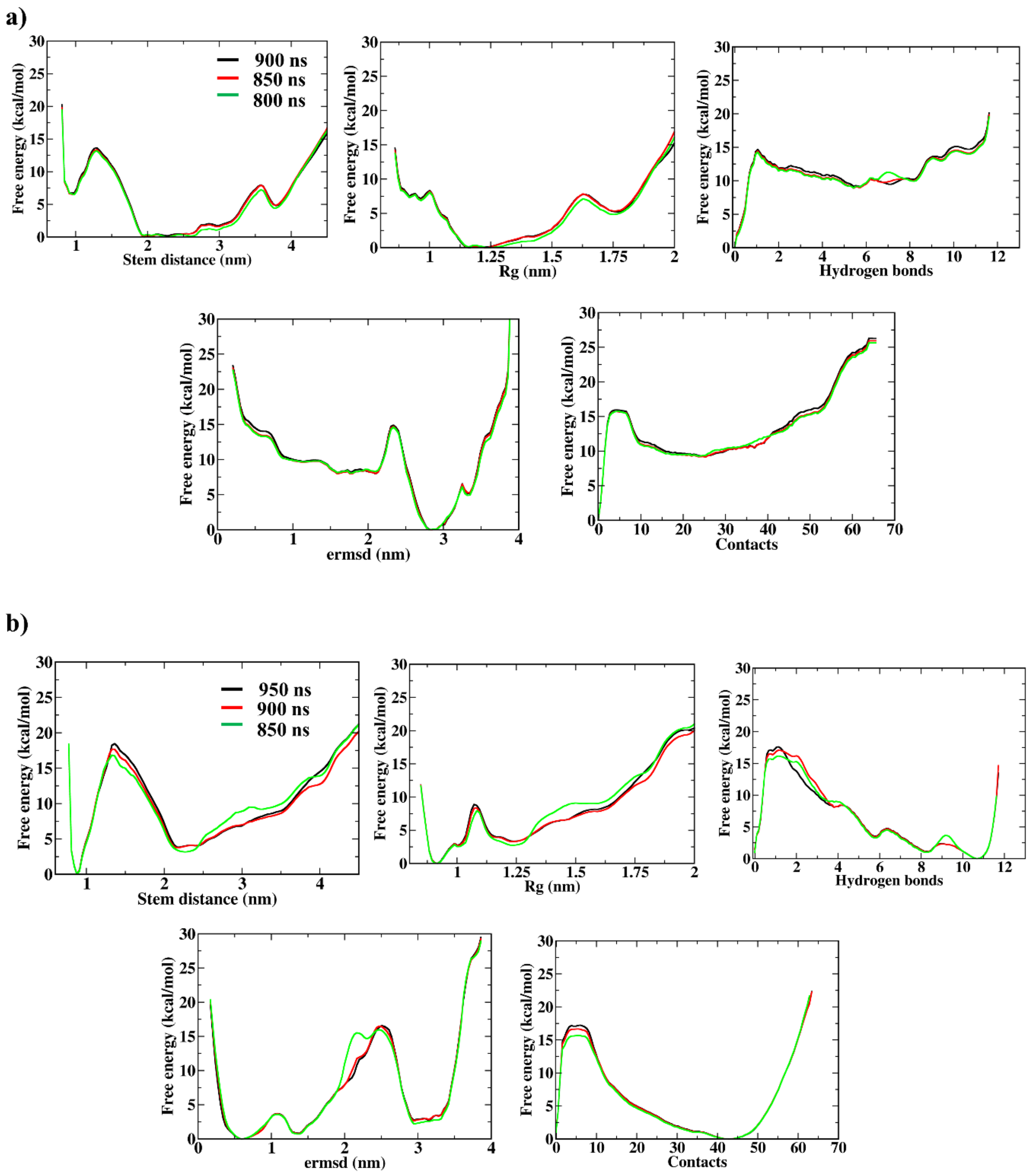

Figure S8. Free energy profiles at different timescales corresponding to RNA folding-unfolding equilibria projected with respect to all biased collective variables in the controlled sampling approach, a) in the presence of L-arginine, and b) in the presence of D-arginine.

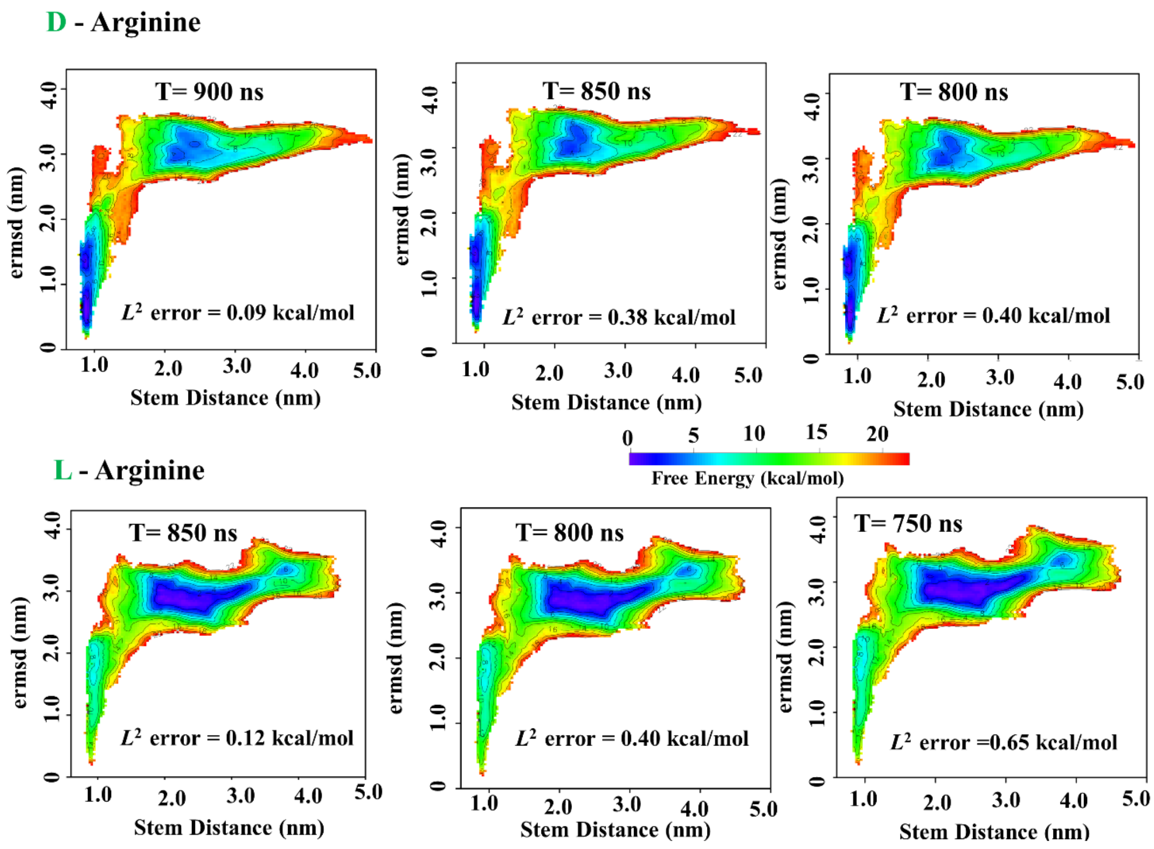

Figure S9. 2D Free energy profiles of folding-unfolding equilibria in the presence of D-arginine and L-arginine. The L^2 error errors corresponding to the final free energy surfaces (after 950 ns for D-arginine and after 900 ns for L-arginine) are shown in the inset of the free energy surfaces.

For L^2 error calculations, we have considered reweighted 2D free energy surfaces at 50 ns intervals in the last 150 ns and compared to the final free energy surfaces (at T=950 ns for D-arginine and T=900 ns for L-arginine) shown in Fig. 3a and Fig. 3b of the manuscript. The L^2error is calculated using the following equation;

$$L^2 = \left[\iint |\, f_t(x) - F(x)|^2\, dx\right]^{\frac{1}{2}},$$

where $f_t(x)$ is the t-th function in a series of functions and F(x) is the ideal function (referred with respect to final free energy surface) being compared to. For calculations, we have used only the regions in CV space where the free energy is lesser than 5 kcal/mol in the final free energy surface. This is because we are mostly interested in the convergence of the free energy minima states corresponding to folded, misfolded, and unfolded states described in the manuscript. The free energy surfaces and the corresponding L^2 error are shown in Fig. S9.

IV. Reweighted free energy plots with respect to subsidiary collective variables in the presence of amino acids.

The 2D reweighted free energy surface for folding-unfolding equilibria in the presence of D-arginine and L-arginine with respect to Stem distance (umbrella sampling coordinate) and other subsidiary collective variables (discussed in the methods section) is shown in the Fig. S10. The native folded state is found to have maximum number of hydrogen bonds and possible contacts between the stem regions with minimum Rg and end to end distance values. While, the misfolded and unfolded states found to form lowest possible contacts and hydrogen bonds with gradual increase in the values for Rg and end to end distance. The global minima states is found to be folded state (Stem distance < 1.8 nm) in the presence of D-arginine and misfolded state (Stem distance between 1.8 nm and 4.0 nm) for L-arginine.

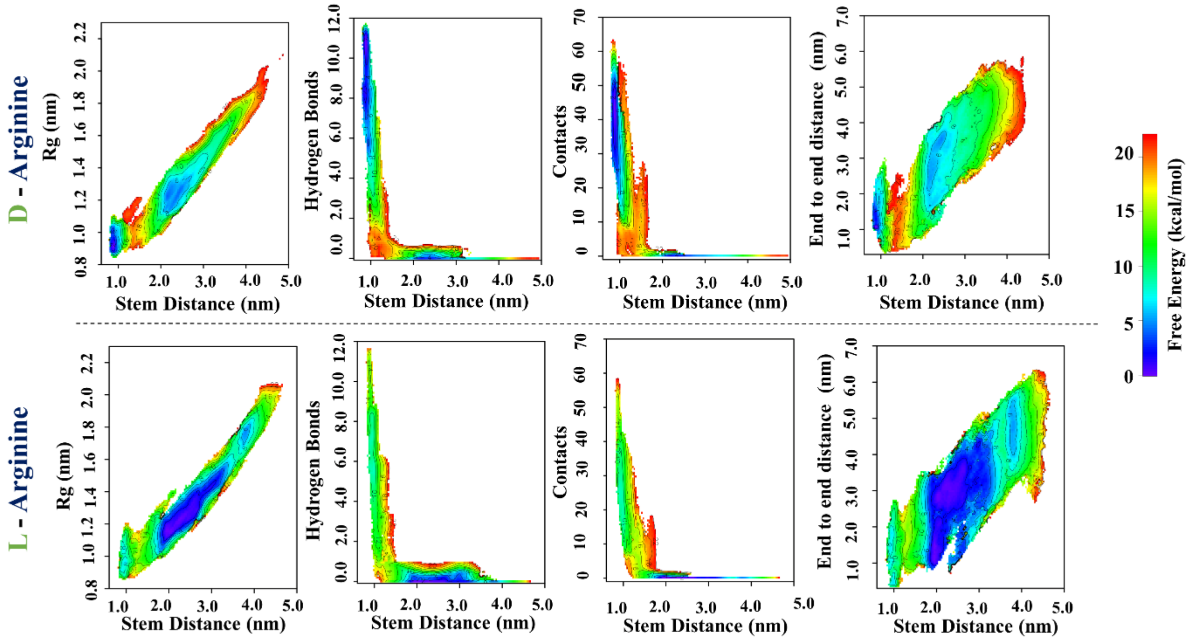

Figure S10. 2D Free energy landscape of folding-unfolding equilibria in the presence of D-arginine and L-arginine reweighted with respect to stem distance (umbrella sampling coordinate) and other subsidiary collective variables.

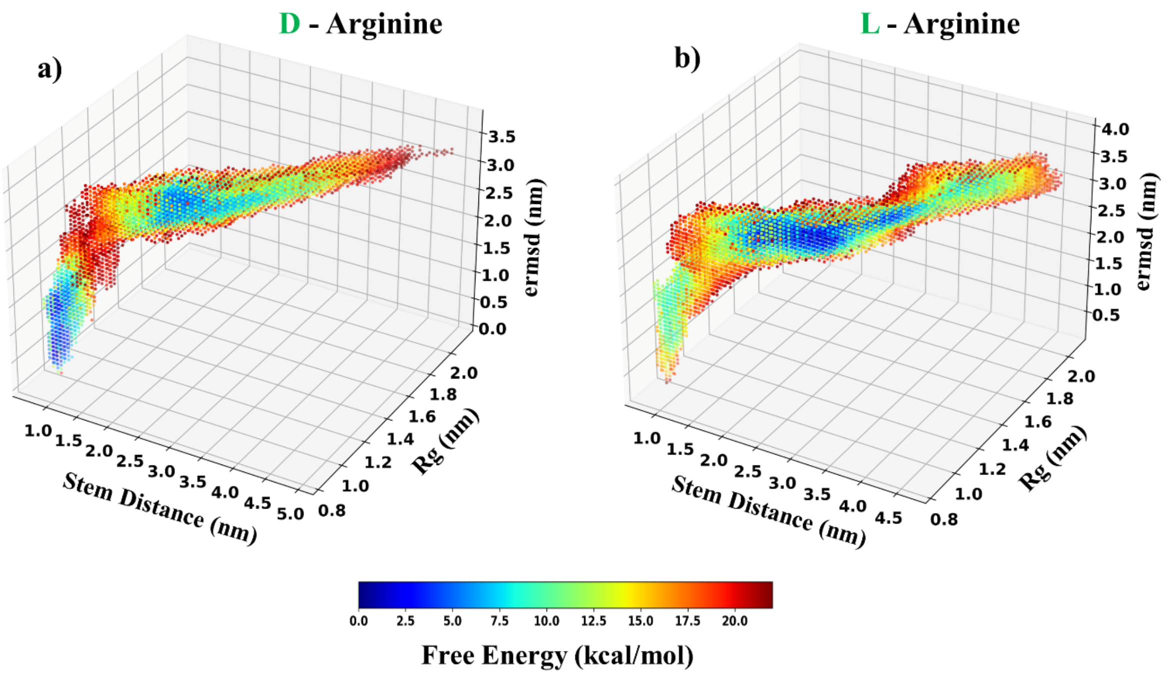

Figure S11. 3D free energy surface of folding-unfolding equilibria in the presence of D-arginine and L-arginine with respect to Stem distance, Rg and ermsd.

The 3D free energy surface with respect to stem distance, Rg and ermsd is shown in Fig. S11. The 3D FES also distinguishes the significant free energy minima states associated with folding-unfolding equilibria similar to the observations discussed in the result section of the manuscript, in accordance with the 2D free energy profiles.

V. Distribution of dihedral parameters in the RNA with and without arginine

We have calculated the distribution of various dihedral parameters determining the secondary structure content of the RNA motif. If there is any distinct change in peak positions of the dihedral distribution, it tells about the conformational change in the structures of RNA. The dihedral parameters are averaged over the residues of RNA along the trajectory of simulation in the folded cluster of RNA. The dihedral distribution of various dihedral parameters is shown in Fig. S12 along with the structural representation of dihedral parameters.

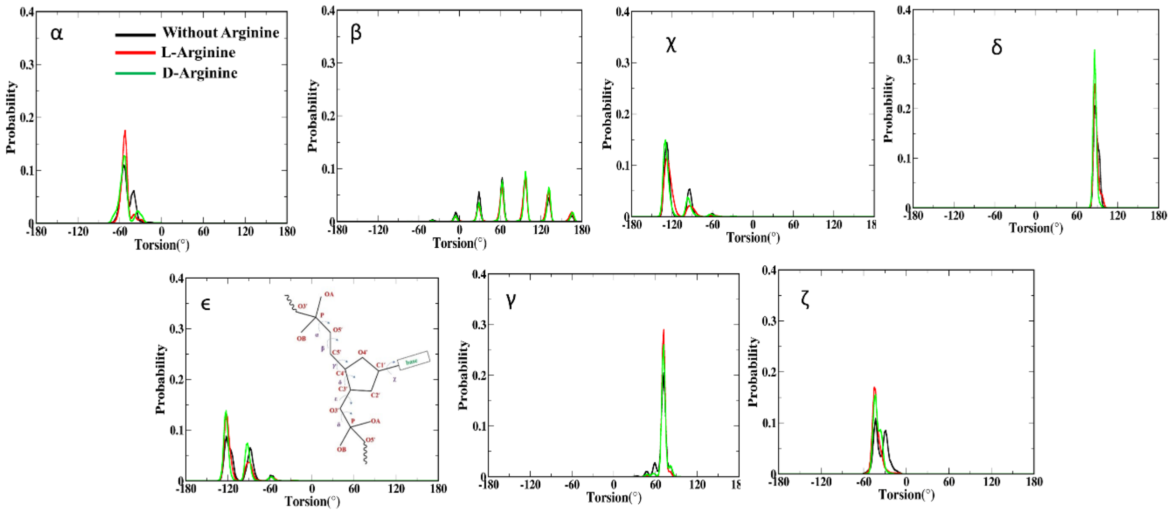

Figure S12. Distribution of structural content in RNA motif in terms of dihedral parameters in the systems of RNA (absence and presence of arginine). The distributions are obtained based on simulation trajectory from controlled sampling approach (for the systems in the presence of arginine) and well-tempered metadynamics simulation trajectory (for the system in the absence of arginine) in the folded cluster of RNA.

We observed that the perks in the probability distribution are found to be in a similar range for various dihedral parameters with moderate deviations in the height of distributions. As we could not observe distinct peaks in each of these observations in the RNA (with and without arginine), it concludes that inversion in geometry with respect to structural changes in RNA is not observed in our simulations.

VI. Representation of edges of interactions in the folded state of RNA

The mapping of RNA bases based on various edges of interaction is shown in the Fig. S13. The three letter code based on Leontis RNA structure classification[9, 10] is shown in Table S1. Triangular representations of some of these relevant interactions with respect to to cis/trans glycosidic rotations is shown in the Fig. S14.

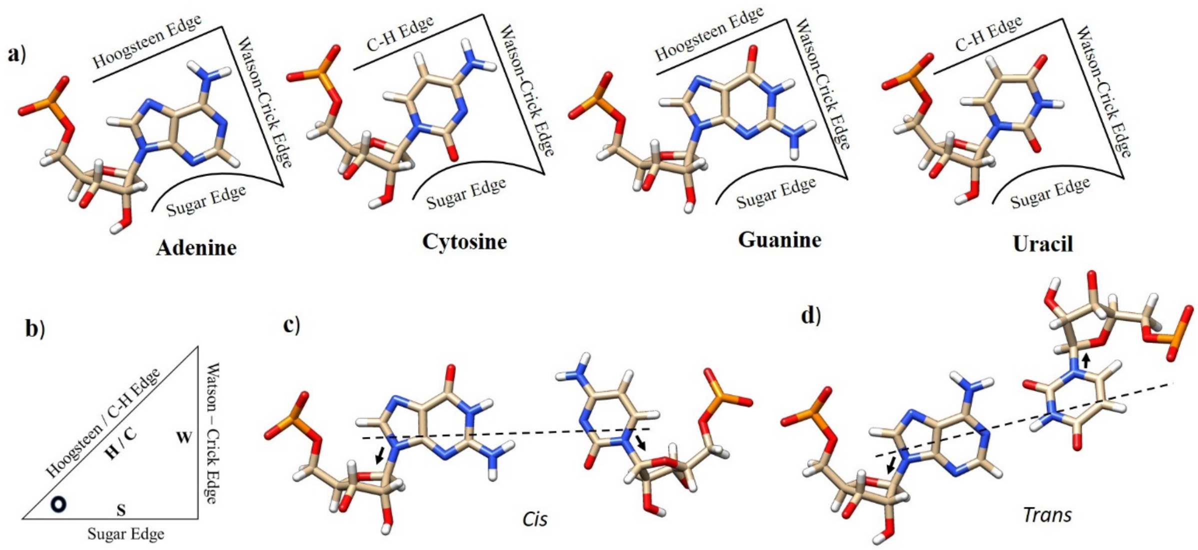

Figure S13. a) Representation of interacting edges in RNA bases (Adenine, Cytosine, Guanine and Uracil). b) Triangle representation of Watson-Crick, Hoogsteen, and C-H edges. The circle near the corner of Hoogsteen/C-H edge and sugar edge indicates the position of ribose unit. c) Representation of cis geometry formed with respect to the glycosidic bonds formed by neighbouring bases. d) Representation of trans geometry formed with respect to the glycosidic bonds formed by neighbouring bases.

Table S1: Notation for the three letter codes assigned based on the edges of interactions between the bases and glycosidic bond orientation. The notation is followed based on Lenotis represention of RNA base classification

Sl No	Starting Edge	Neighbouring Edge	Orientation	Representation
1	Hoogsteen	Sugar	*cis*	**WSc**
2	Sugar	Watson-Crick	*cis*	**SWC**
3	Hoogsteen	Hoogsteen	*trans*	**HHt**
4	Sugar	Hoogsteen	*trans*	**SHt**
5	Hoogsteen	Watson-Crick	*cis*	**HWc**
6	Watson-Crick	Sugar	*cis*	**WSc**
7	Hoogsteen	Hoogsteen	*cis*	**HHc**
8	Sugar	Sugar	*cis*	**SSc**
9	Watson-Crick	Hoogsteen	*trans*	**WHt**
10	Watson-Crick	Watson-Crick	*trans*	**WWt**
11	Sugar	Hoogsteen	*cis*	**SHc**
12	Watson-Crick	Hoogsteen	*cis*	**WHc**
13	Watson-Crick	C-H	*cis*	**WCc**
14	Hoogsteen	Watson-Crick	*trans*	**HWt**
15	Sugar	Watson-Crick	*trans*	**SWt**
16	Watson-Crick	Watson-Crick	*cis*	**WWc**
17	Hoogsteen	Sugar	*trans*	**HSt**
18	Watson-Crick	Sugar	*trans*	**WSt**
19	Sugar	Sugar	*trans*	**SSt**

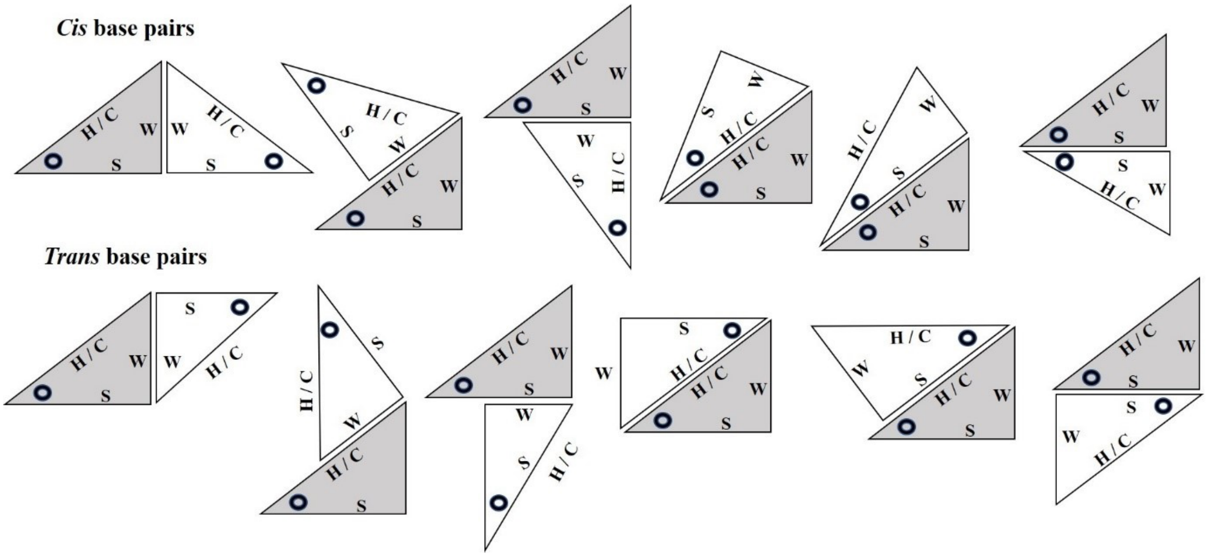

Figure S14. Triangle representation of relevant possible orientations that can be formed by nucleobases in RNA in cis and trans orientation. The interacting edges between the neighbouring bases are separated by the surface of contact between the grey and white coloured triangles representing neighbouring bases.

References

1. A. Barducci, G. Bussi and M. Parrinello, *Phys Rev Lett*, 2008, **100**, 020603.
2. J. Pfaendtner and M. Bonomi, *J Chem Theory Comput*, 2015, **11**, 5062-5067.
3. G. M. Torrie and J. P. Valleau, *Journal of Computational Physics*, 1977, **23**, 187-199.
4. S. Awasthi, V. Kapil and N. N. Nair, *J Comput Chem*, 2016, **37**, 1413-1424.
5. A. Gupta, S. Verma, R. Javed, S. Sudhakar, S. Srivastava and N. N. Nair, *J Comput Chem*, 2022, **43**, 1186-1200.
6. M. Bonomi, A. Barducci and M. Parrinello, *Journal of Computational Chemistry*, 2009, **30**, 1615-1621.
7. G. A. Tribello, M. Bonomi, D. Branduardi, C. Camilloni and G. Bussi, *Computer Physics Communications*, 2014, **185**, 604-613.
8. S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen and P. A. Kollman, *Journal of Computational Chemistry*, 1992, **13**, 1011-1021.
9. N. B. Leontis, J. Stombaugh and E. Westhof, *Nucleic Acids Res*, 2002, **30**, 3497-3531.
10. N. B. Leontis and E. Westhof, *RNA*, 2001, **7**, 499-512.