Supporting Information

Hydrogen-Bonding Modulation of RNA versus DNA Hairpin Folding Pathways

Yong-Qi Ke[1,*], Li-Qing Tao[1]

[1]School of Chemistry and Chemical Engineering, Beijing Institute of Technology, Beijing 100081, China

Table of Contents

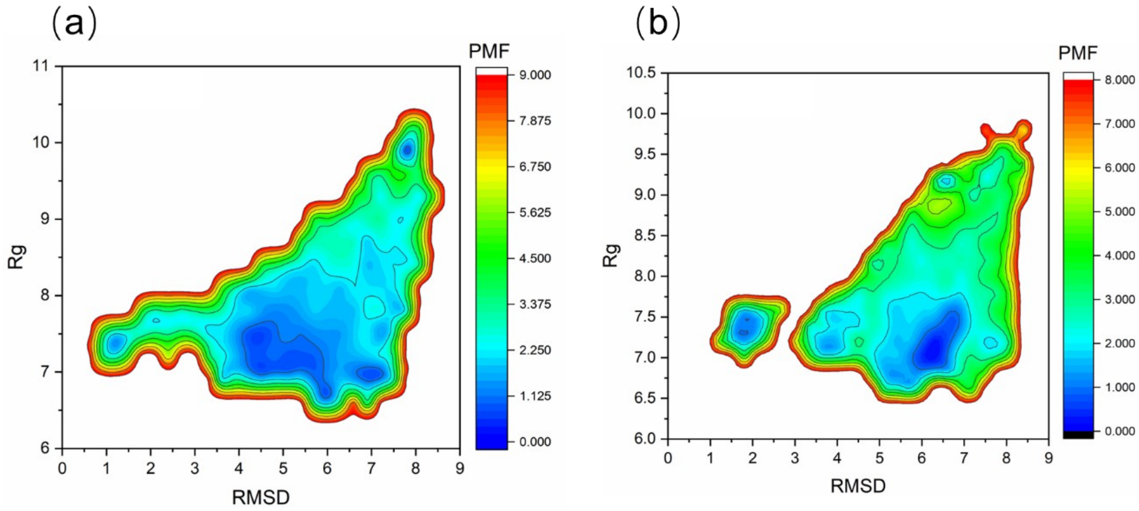


Figure S1. Free energy plot with respect to collective variables RMSD and Rg, (a) RNA UUCG, (b) DNA dUdUdCdG.

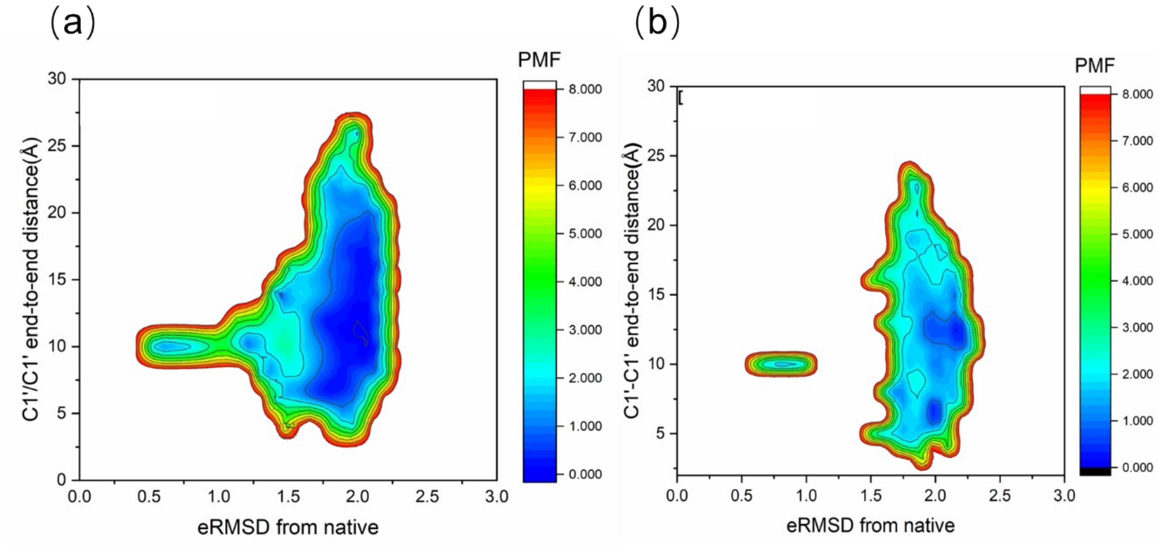
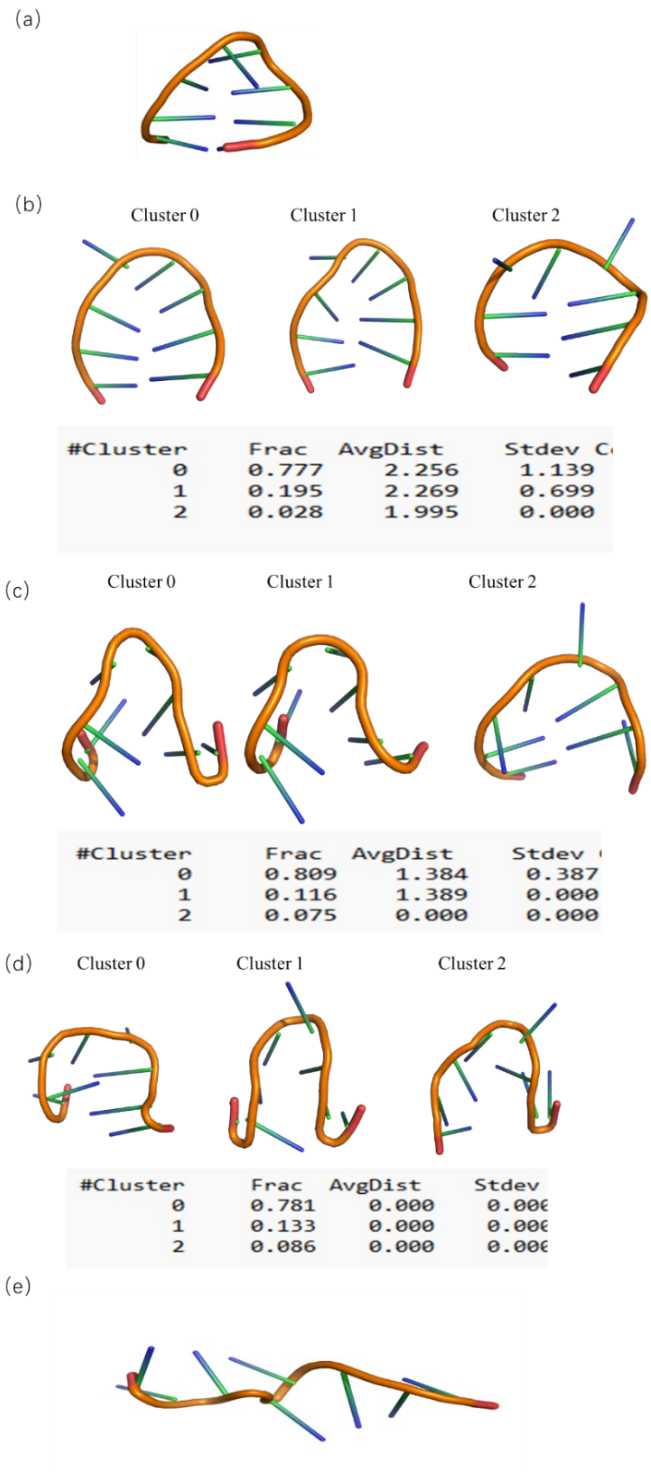


Figure S2. Free energy plot with respect to collective variables eRMSD and the distance between [1]c's C1' and [8]g's C1', (a) RNA UUCG, (b) DNA dUdUdCdG.



```
#Cluster       Frac   AvgDist     Stdev  C
       0      0.777     2.256     1.139
       1      0.195     2.269     0.699
       2      0.028     1.995     0.000
```

```
#Cluster        Frac   AvgDist     Stdev
       0       0.809     1.384     0.387
       1       0.116     1.389     0.000
       2       0.075     0.000     0.000
```

```
#Cluster        Frac   AvgDist     Stdev
       0       0.781     0.000     0.00
       1       0.133     0.000     0.00
       2       0.086     0.000     0.00
```

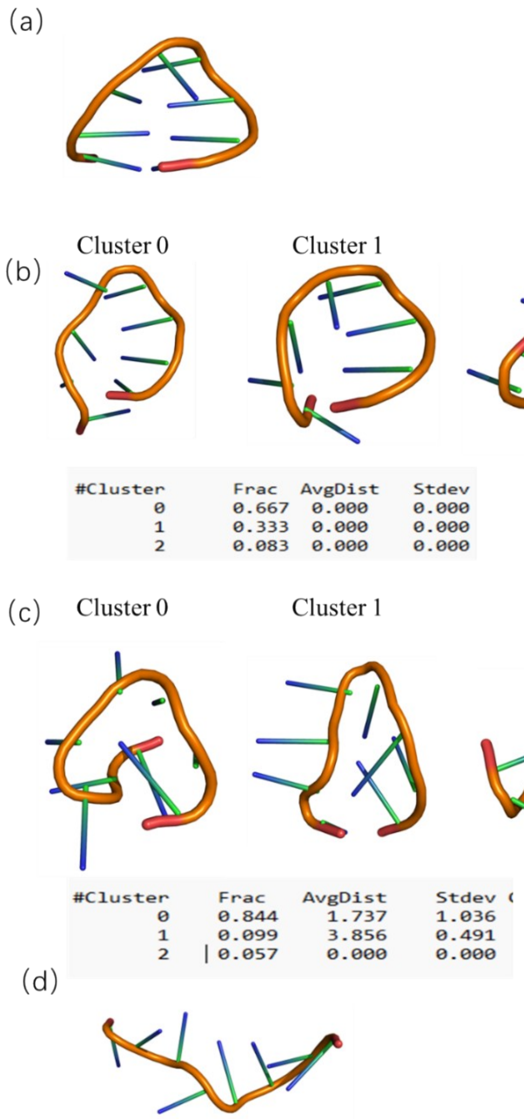
Figure S3. The clustering results of Cluster V-I of the RNA UUCG tetraloop((a)Cluster V, (b)Cluster IV, (c)Cluster III, (d)Cluster II (e)Cluster I). Cluster V only contains one structure (crystalline structure), and the structures within Cluster I are all in the expanded state.

Table S1. Average RMSD values (in Å) with respect to the crystal structure and Rg (in Å) for RNA UUCG.

Cluster Number	Rg	RMSD
I	9.12±0.14	8.03±0.88
II	7.40±0.79	5.24±0.95
III	7.17±0.55	6.28±0.38
IV	7.66±0.47	4.29±0.80
V	7.77±0.98	2.11±0.72

Table S2. Hydrogen bonds and π-π stacking interactions of the misfolded structures of RNA UUCG along the Minimum Free Energy Pathway.
(Hydrogen bonds were identified using a donor–acceptor distance cutoff of 3.5 Å and a D–H···A angle larger than 120°. Base stacking interactions were defined based on a base–base centroid distance ≤ 4.0 Å and an interplanar angle ≤ 30°.)

Cluster Number	Hydrogen bonding	π-π stacking
II	^{2}c-O2’-H---^{3}U-OP1(42%) ^{2}c-N4-H---^{6}g-O4’(35%) ^{7}g-O6---^{6}g-N2-H(28%) ^{7}g-N2-H---^{2}c-OP1(31%) ^{7}g-O6---^{6}G-N2-H(27%) ^{8}g-O2’H▪▪▪ ^{7}g-OP2(46%) ^{4}U-O2---^{1}c-O5’-H(39%)	^{2}c-^{7}g-^{8}g ^{1}C-^{3}U
III	^{1}c-O2’-H---^{2}c-OP1(48%) ^{3}U-O2’-H---^{6}G-O4(41%) ^{3}U-O2-H---^{6}G-O2’(44%) ^{4}U-O4’---^{5}C-O2’-H(36%) ^{8}g-O2’-H---^{7}g-OP2(63%) ^{7}g-O2’-H---^{8}g-OP1(58%) ^{2}C-O2'-H---^{3}U-OP1(40%) ^{1}c-O5’-H---^{3}U-O2’(34%) ^{1}c-N4-H---^{5}c-O2(29%)	^{1}c-^{4}U-^{6}G-^{7}g-^{8}g ^{2}C -^{3}U
IV	standard Watson-Crick hydrogen bonds(^{1}c-^{8}g, ^{2}c-^{7}g)(70-75%) ^{3}U-O2'-H---^{4}U-O5'(52%) ^{4}U-O2'-H---^{5}C-OP2(47%) ^{6}G-O2'-H---^{7}G- O4'(38%)	^{3}U-^{5}C-^{6}G



```
#Cluster      Frac   AvgDist     Stdev
       0     0.667   0.000      0.000
       1     0.333   0.000      0.000
       2     0.083   0.000      0.000
```

```
#Cluster     Frac     AvgDist      Stdev (
       0     0.844      1.737      1.036
       1     0.099      3.856      0.491
       2   | 0.057      0.000      0.000
```

Figure S4. Clustering results of Cluster IV-I for the DNA dUdUdCdG tetraloop((a)Cluster IV, (b)Cluster III, (c)Cluster II, (d)Cluster I). Cluster IV only contains one structure (the folded structure), and the structures within Cluster I are all in the expanded state.

Table S3. Average RMSD values (in Å) with respect to the crystal structure and Rg (in Å) for DNA dUdUdCdG.

Cluster Number	Rg	RMSD
I	9.66±0.57	8.53±0.66
II	6.97±0.89	6.09±0.79
III	7.31±0.62	4.82±0.42
IV	7.26±0.59	2.07±0.33

Table S4. Hydrogen bonds and π-π stacking interactions of the misfolded structures of DNA dUdUdCdG along the Minimum Free Energy Pathway.
(Hydrogen bonds were identified using a donor–acceptor distance cutoff of 3.5 Å and a D–H···A angle larger than 120°. Base stacking interactions were defined based on a base–base centroid distance ≤ 4.0 Å and an interplanar angle ≤ 30°.)

Cluster Number	Hydrogen bonding	π-π stacking
II	^{1}dc-O5'-H---^{6}dG-O6(22%) ^{1}dc-O4'---^{7}dg-N2-H(18%) ^{1}dc-N4-H---^{8}dg-OP2(28%) ^{5}dG-O2---^{6}dG-N2-H(24%) ^{8}dg-N1-H---^{5}dC-OP1(20%) ^{8}dg-N2-H---^{5}dC-OP2(26%) ^{8}dg-N3-H---^{5}dU-O5'(17%)	^{2}dc-^{4}dU ^{6}dG-^{7}dg ^{1}dc-^{6}dG
III	^{2}dc-OP1---^{8}dg-O3'-H(31%) ^{2}dc-O4'---^{8}dg-N2-H(27%) ^{3}dU-O2---^{7}dg-N1-H(23%)	^{1}dc-^{5}dc-^{6}dG-^{7}dg-^{8}dg

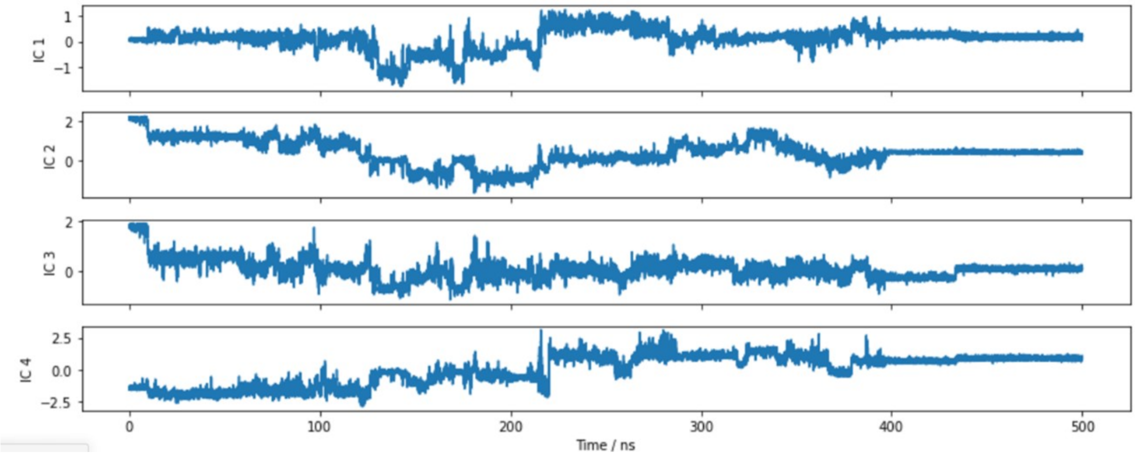


Figure S5. The values for independent components (IC) 1-4 for RNA UUCG.

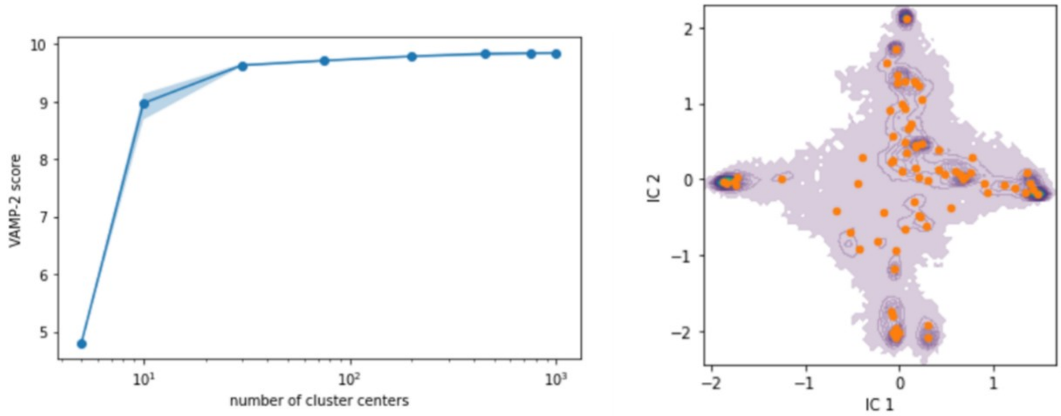
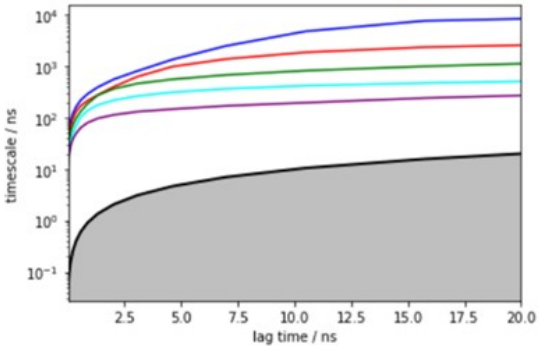


Figure S6. VAMP-2 scores for estimated Markov models with different number of clusters using the k-means clustering method for RNA UUCG.



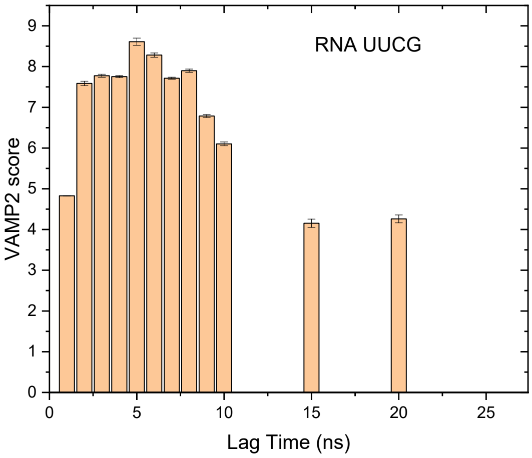


Figure S7. The top implied timescales as function of lag time, derived from the transition probability matrix with all 75 states of RNA UUCG and the VAMP2 for the choice of lag time.

In the Markov State Model (MSM), if the implied timescale stabilizes with increasing lag time, it suggests model convergence. That is, when implied timescales remain unchanged, MSM parameter estimation is considered stable. For RNA UUCG, extending the lag time (20 ns, 60 ns) showed no significant increase in implied timescales, confirming convergence (Figure S7). Additionally, the VAMP2 score, used to assess model quality, peaked at a 5 ns lag time, indicating optimal performance. Thus, 5 ns was chosen for subsequent analysis.

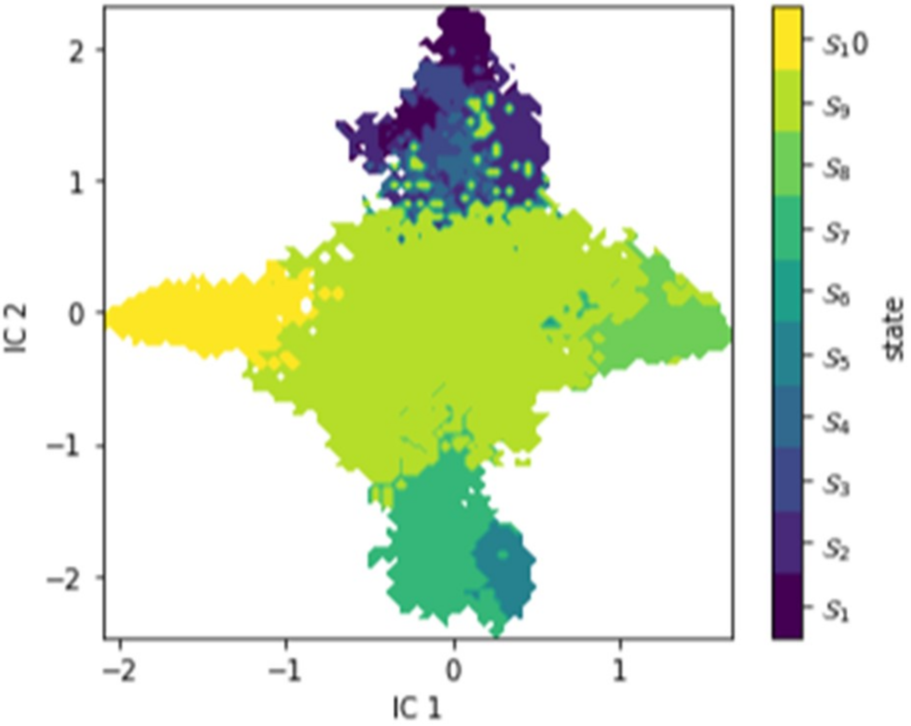


Figure S8. Distribution of the ten metastable states S1-S10 of RNA UUCG.

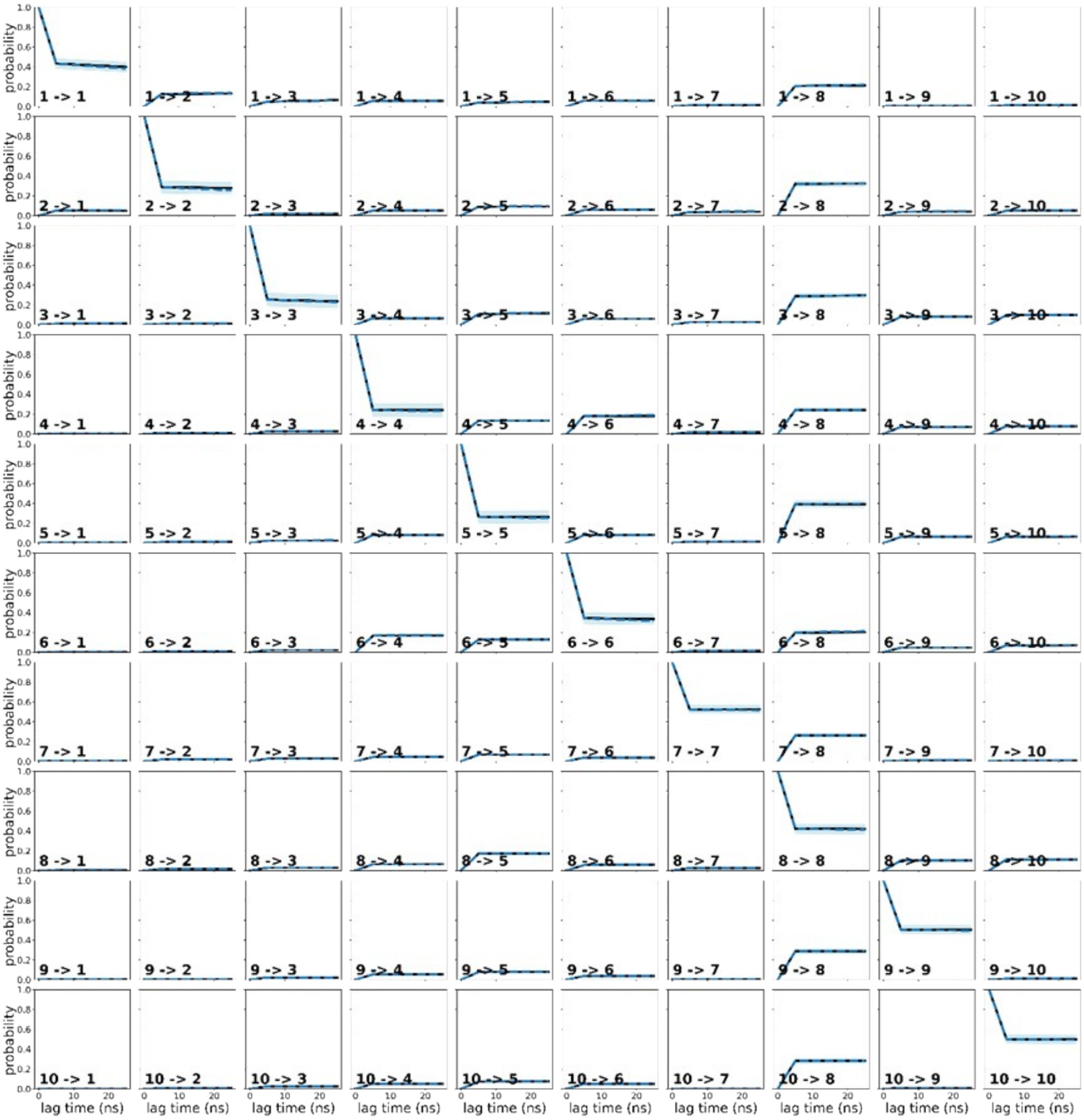


Figure S9. Chapman-Kolmorgov test for transitions among the ten metastable states of RNA UUCG.

The Chapman-Kolmogorov Test (CK Test) is commonly used to verify whether the constructed model satisfies the Markovian property, i.e., whether the long-timescale transition behavior can be derived from short-timescale transitions. In a Chapman-Kolmogorov plot, we typically compare Ppred(kτ) (represented by the solid line), which is derived from the short-time transition matrix, with the true computed P(kτ) (represented by the dashed line). If the scatter points align well with the solid line, it indicates that the MSM approximately satisfies the Markovian property. Conversely, if there is a significant deviation between the two, the model may not satisfy the Markov assumption and requires further adjustment. From Figure S9, it can be observed that for both RNA UUCG, the dashed and solid lines overlap well, indicating that the system satisfies the Markovian property.

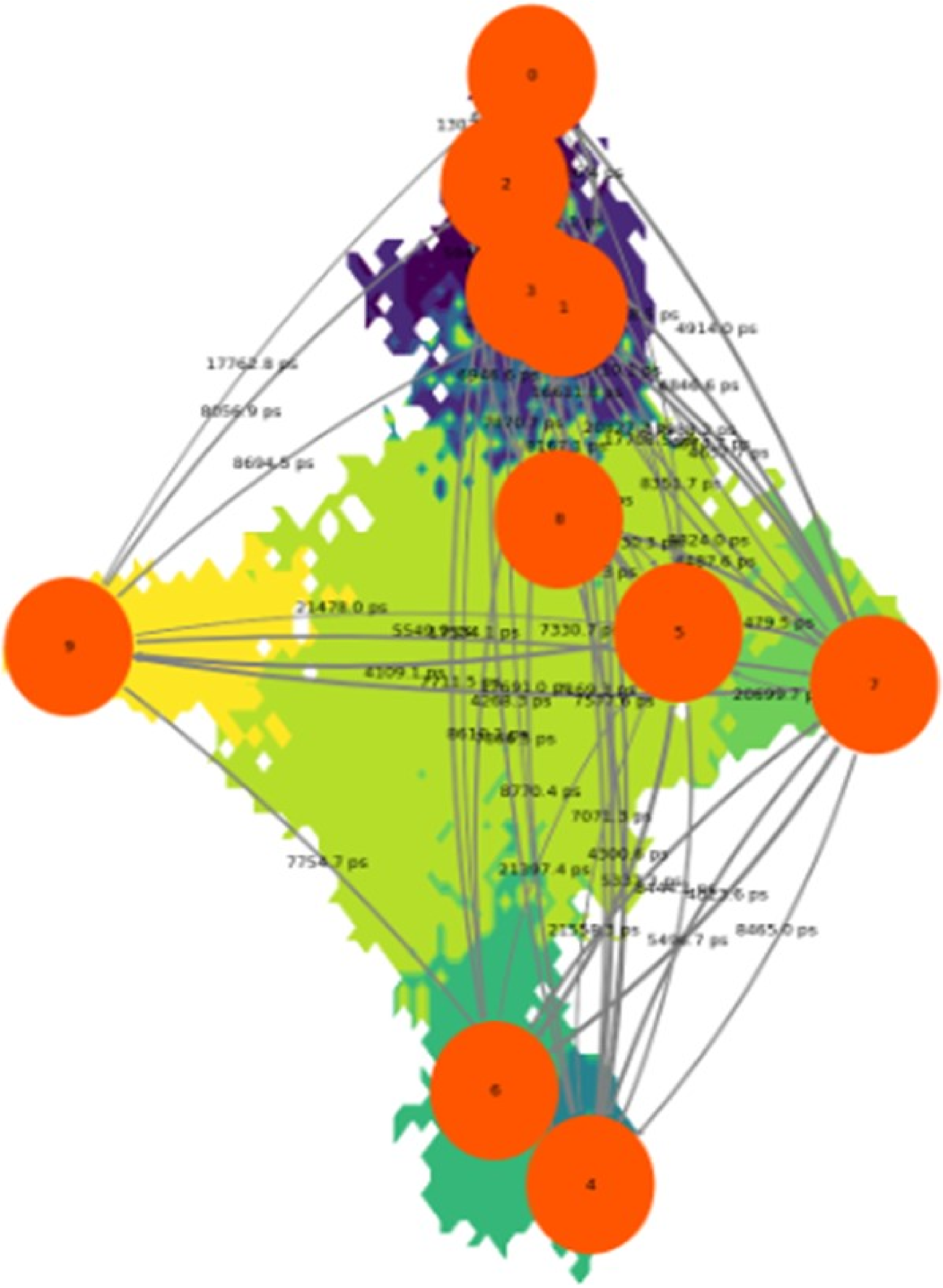
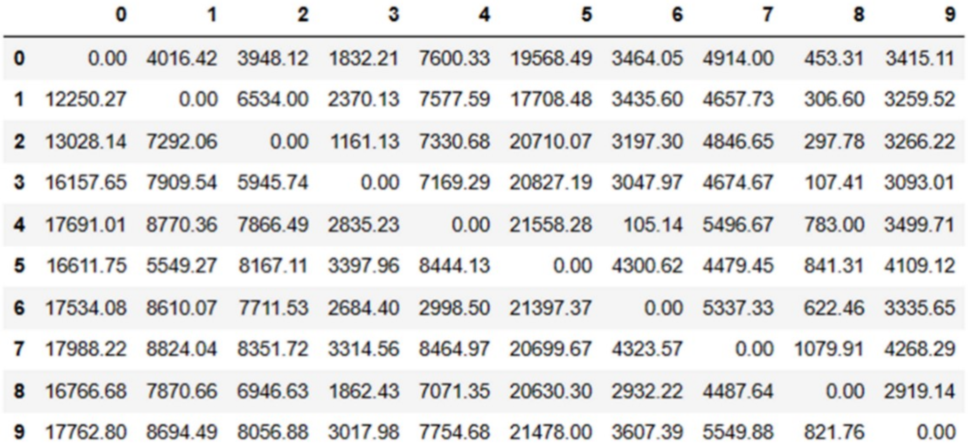


Figure S10. Ten-state Markov state modeling with committor probabilities projected on the first two ICs for RNA UUCG.



	0	1	2	3	4	5	6	7	8	9
0	0.00	4016.42	3948.12	1832.21	7600.33	19568.49	3464.05	4914.00	453.31	3415.11
1	12250.27	0.00	6534.00	2370.13	7577.59	17708.48	3435.60	4657.73	306.60	3259.52
2	13028.14	7292.06	0.00	1161.13	7330.68	20710.07	3197.30	4846.65	297.78	3266.22
3	16157.65	7909.54	5945.74	0.00	7169.29	20827.19	3047.97	4674.67	107.41	3093.01
4	17691.01	8770.36	7866.49	2835.23	0.00	21558.28	105.14	5496.67	783.00	3499.71
5	16611.75	5549.27	8167.11	3397.96	8444.13	0.00	4300.62	4479.45	841.31	4109.12
6	17534.08	8610.07	7711.53	2684.40	2998.50	21397.37	0.00	5337.33	622.46	3335.65
7	17988.22	8824.04	8351.72	3314.56	8464.97	20699.67	4323.57	0.00	1079.91	4268.29
8	16766.68	7870.66	6946.63	1862.43	7071.35	20630.30	2932.22	4487.64	0.00	2919.14
9	17762.80	8694.49	8056.88	3017.98	7754.68	21478.00	3607.39	5549.88	821.76	0.00

Figure S11. The Mean First Passage Time among ten states of RNA UUCG.

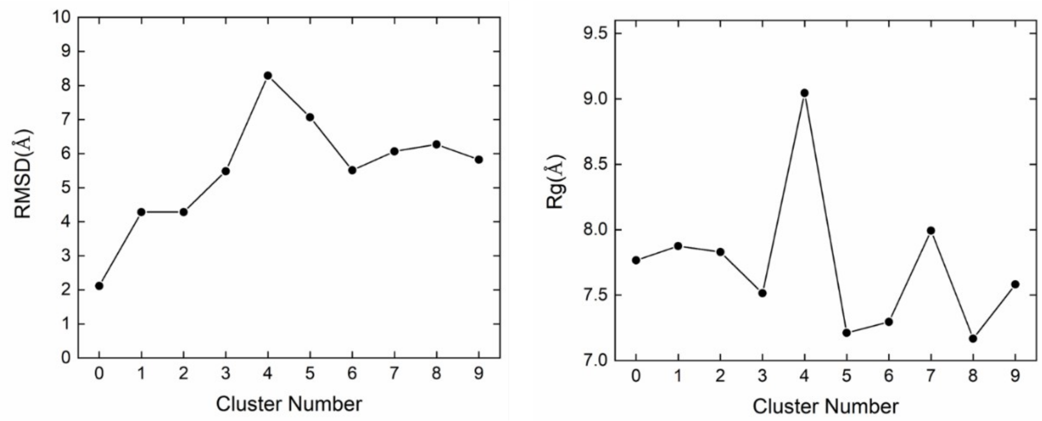


Figure S12. The all-atom RMSD and Rg of RNA UUCG are shown to indicate structural deviation from the native NMR starting structure.

Table S5. The distances and dihedrals for each cluster of RNA UUCG.

	Distanc(Å)	Dihedral(°)
state-0	1.96±0.04	-58.08±1.04
state-1	2.11±0.13	-1.45±0.08
state-2	7.18±0.78	-60.22±1.54
state-3	4.16±0.24	47.37±2.16
state-4	19.29±1.03	177.68±3.16
state-5	11.03±0.94	-120.79±2.04
state-6	14.04±0.87	29.34±3.14
state-7	15.33±1.32	-63.39±2.36
state-8	10.09±0.84	2.84±0.04
state-9	6.24±0.44	22.53±1.22

Table S6. Path fluxes for unfolded and folded states of RNA UUCG.

No.	State-4 as final unfolded structure and State-0 as initial native-like structure
1	4-6-8-1-0
2	4-6-8-2-0
3	4-6-8-3-2-0
4	4-6-8-0
5	4-6-3-0
6	4-6-8-3-0
7	4-6-8-5-1-0
8	4-6-2-0

9	4-6-8-9-1-0
10	4-6-8-7-1-0

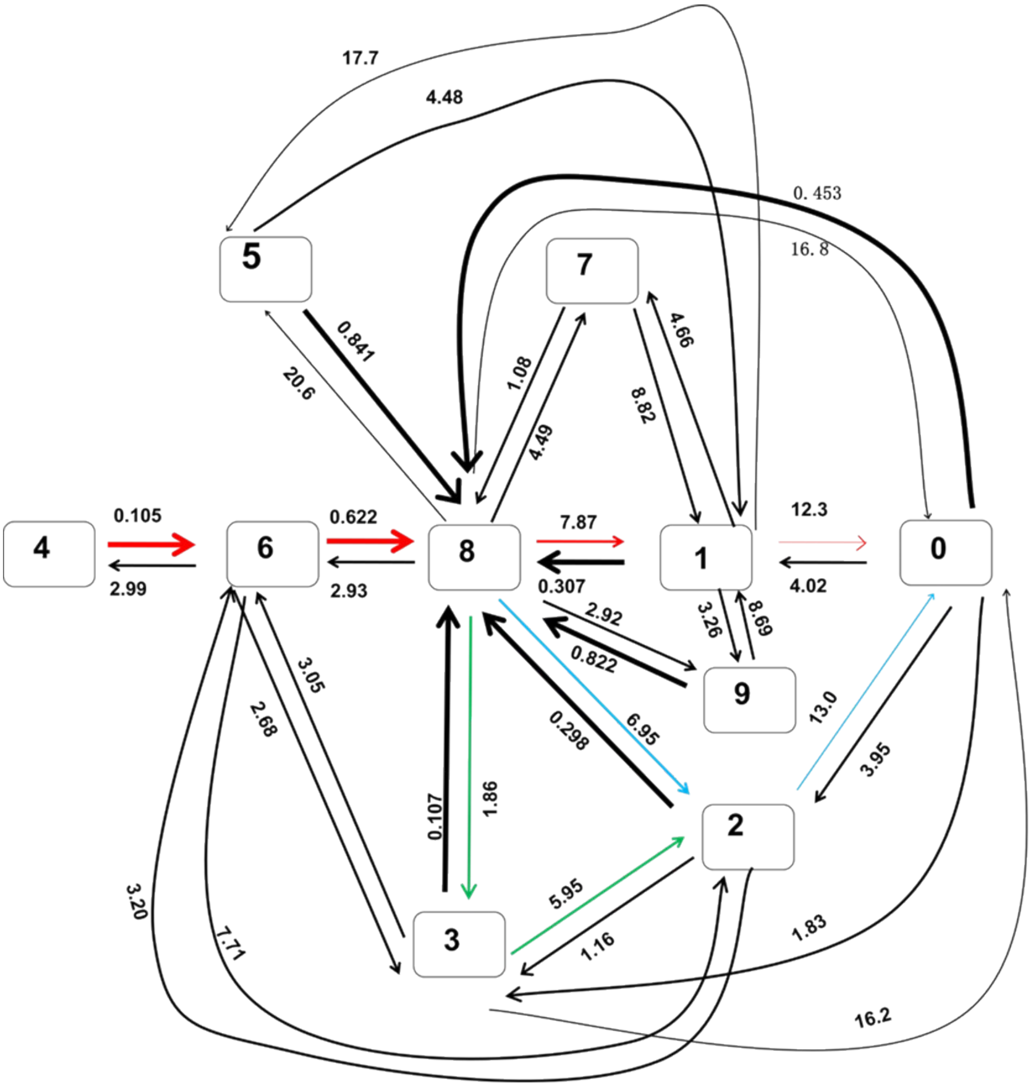


Figure S13. Folding pathways of RNA UUCG from RNA state 4 (unfolded state) to RNA state 0 (folded state). Arrow colors denote different pathways and the weight of arrow indicates the transition probability between states. Transition times (in μs) between two states are derived from Mean First Passage Time (MFPT) calculations. The red, blue and green curves represent the first, second and third folding pathway, respectively.

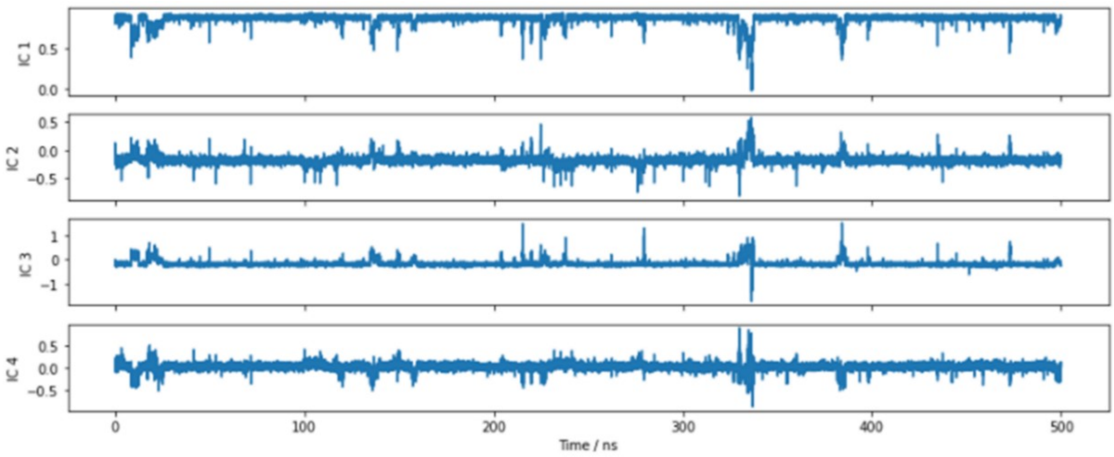


Figure S14. The values for the independent components (IC) 1-4 of DNA dUdUdCdG.

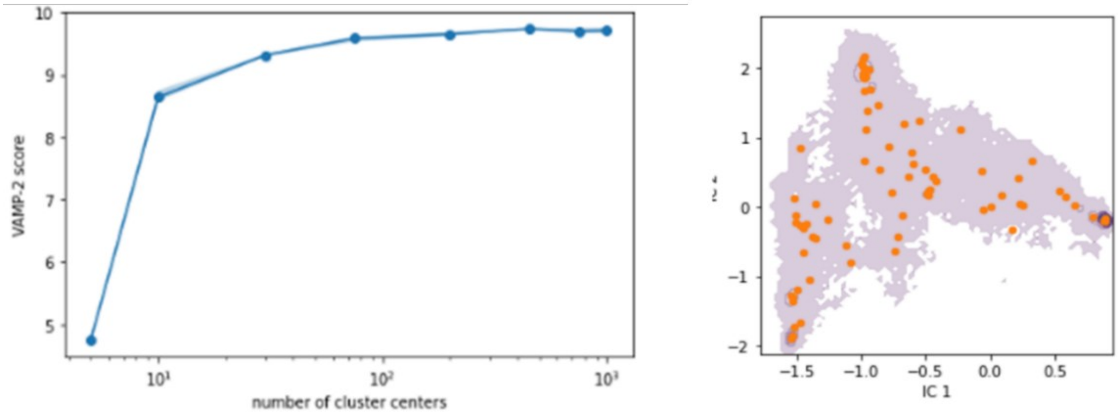
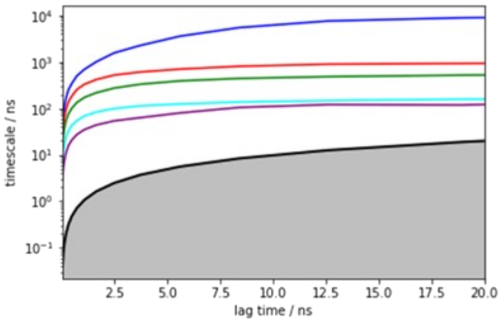


Figure S15. VAMP-2 scores for estimated Markov models with different number of clusters using the k-means clustering method for DNA dUdUdCdG.



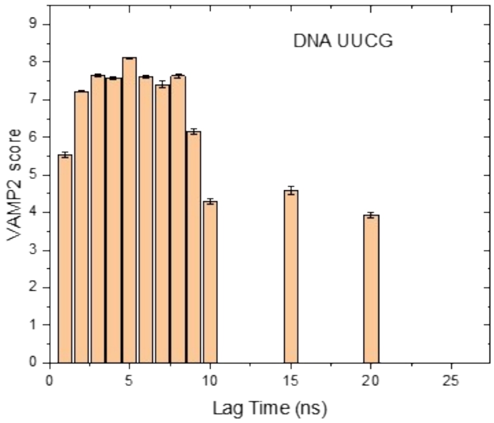


Figure S16. The top implied timescales as function of lag time, derived from the transition probability matrix with all 75 states of DNA dUdUdCdG, and the VAMP2 for the choice of lag time.

In the Markov State Model (MSM), if the implied timescale stabilizes with increasing lag time, it suggests model convergence. That is, when implied timescales remain unchanged, MSM parameter estimation is considered stable. For DNA dUdUdCdG, extending the lag time (20 ns, 60 ns) showed no significant increase in implied timescales, confirming convergence (Figure S16). Additionally, the VAMP2 score, used to assess model quality, peaked at a 5 ns lag time, indicating optimal performance. Thus, 5 ns was chosen for subsequent analysis.

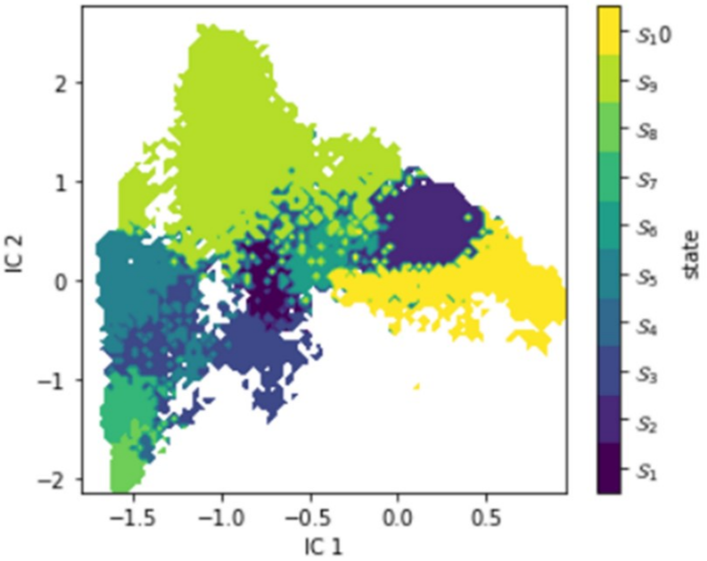


Figure S17. Distributions of the ten metastable states S1-S10 of DNA dUdUdCdG.

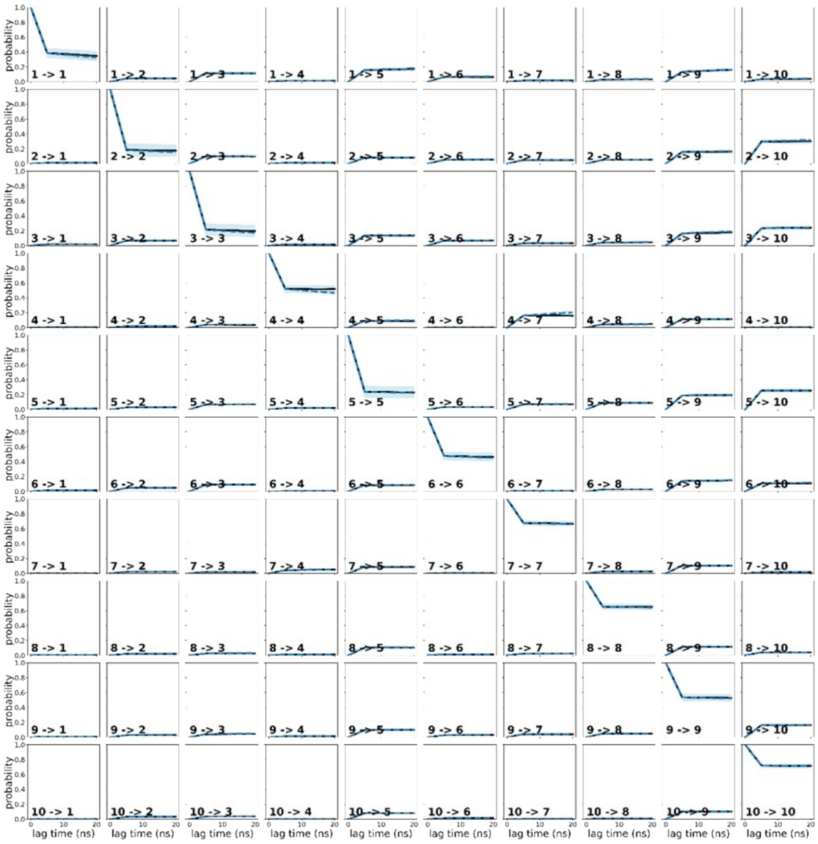


Figure S18. Chapman-Kolmorgov test for transition among the ten metastable states of DNA dUdUdCdG.

The Chapman-Kolmogorov Test (CK Test) is commonly used to verify whether the constructed model satisfies the Markovian property, i.e., whether the long-timescale transition behavior can be derived from short-timescale transitions. In a Chapman-Kolmogorov plot, we typically compare Ppred(kτ) (represented by the solid line), which is derived from the short-time transition matrix, with the true computed P(kτ) (represented by the dashed line). If the scatter points align well with the solid line, it indicates that the MSM approximately satisfies the Markovian property. Conversely, if there is a significant deviation between the two, the model may not satisfy the Markov assumption and requires further adjustment. From Figure S18, it can be observed that for both DNA dUdUdCdG, the dashed and solid lines overlap well, indicating that the system satisfies the Markovian property.

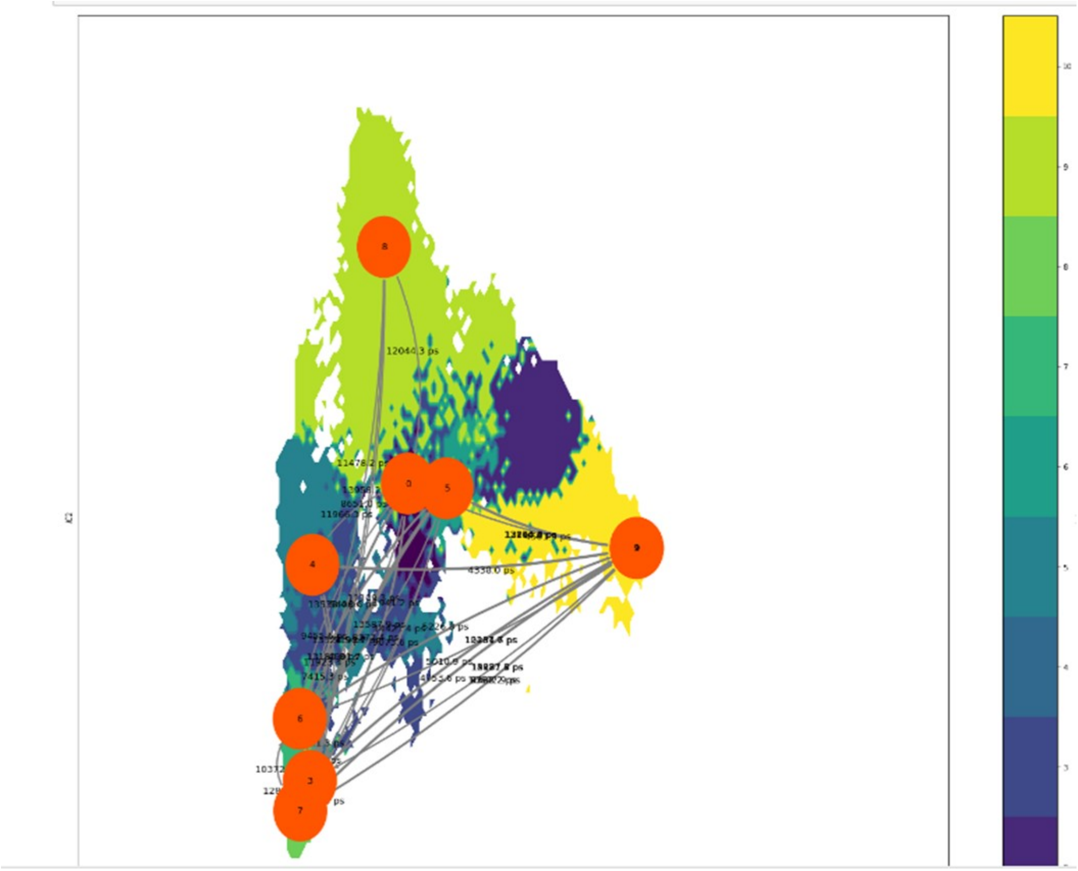
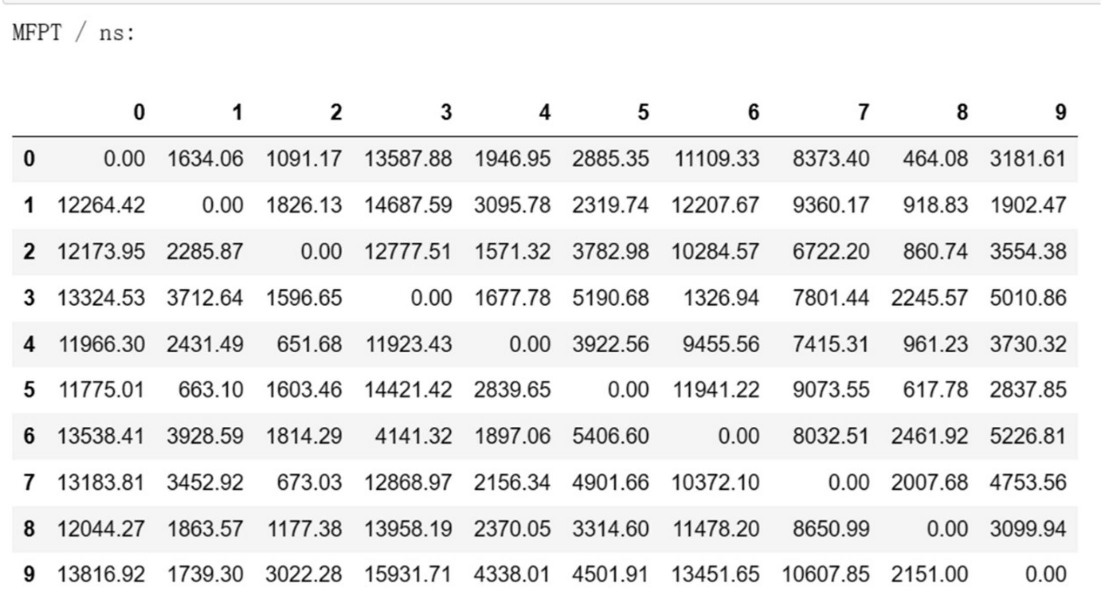


Figure S19. Ten-state Markov state modeling with committor probability projected on the first two ICs for DNA dUdUdCdG.



MFPT / ns:

	0	1	2	3	4	5	6	7	8	9
0	0.00	1634.06	1091.17	13587.88	1946.95	2885.35	11109.33	8373.40	464.08	3181.61
1	12264.42	0.00	1826.13	14687.59	3095.78	2319.74	12207.67	9360.17	918.83	1902.47
2	12173.95	2285.87	0.00	12777.51	1571.32	3782.98	10284.57	6722.20	860.74	3554.38
3	13324.53	3712.64	1596.65	0.00	1677.78	5190.68	1326.94	7801.44	2245.57	5010.86
4	11966.30	2431.49	651.68	11923.43	0.00	3922.56	9455.56	7415.31	961.23	3730.32
5	11775.01	663.10	1603.46	14421.42	2839.65	0.00	11941.22	9073.55	617.78	2837.85
6	13538.41	3928.59	1814.29	4141.32	1897.06	5406.60	0.00	8032.51	2461.92	5226.81
7	13183.81	3452.92	673.03	12868.97	2156.34	4901.66	10372.10	0.00	2007.68	4753.56
8	12044.27	1863.57	1177.38	13958.19	2370.05	3314.60	11478.20	8650.99	0.00	3099.94
9	13816.92	1739.30	3022.28	15931.71	4338.01	4501.91	13451.65	10607.85	2151.00	0.00

Figure S20. The Mean First Passage Time among the ten states of DNA dUdUdCdG.

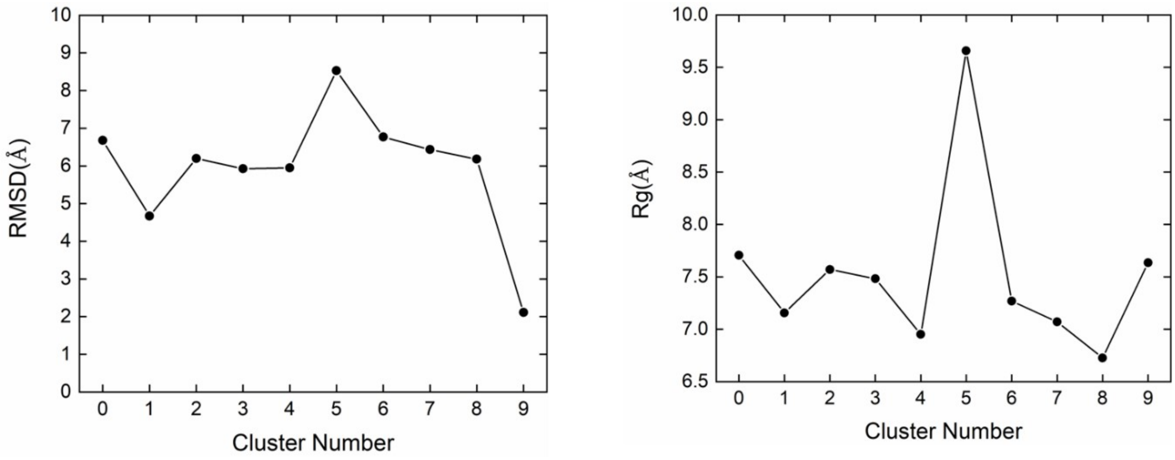


Figure S21. All-atom RMSD and Rg of DNA dUdUdCdG are shown to indicate structural deviation from the native NMR starting structure.

Table S7. The distances and dihedrals for each cluster of DNA dUdUdCdG.

	distance	dihedral
state-0	16.82±0.65	42.09±1.36
state-1	4.43±0.09	-29.66±1.21
state-2	8.19±0.36	-73.49±0.77
state-3	11.96±0.79	58.41±2.42
state-4	3.79±0.06	90.78±2.25
state-5	20.10±0.96	178.99±4.59
state-6	13.69±0.55	85.05±2.38
state-7	7.14±0.21	-1.91±0.09
state-8	7.98±0.19	59.75±1.26
state-9	1.95±0.09	-59.99±1.98

Table S8. Path fluxes for unfolded state and folded state of DNA dUdUdCdG.

No.	State-5 as final unfolded structure and State-9 as initial native-like structure
1	5-1-9
2	5-8-9
3	5-2-9
4	5-8-2-1-9
5	5-8-1-9
6	5-0-4-9

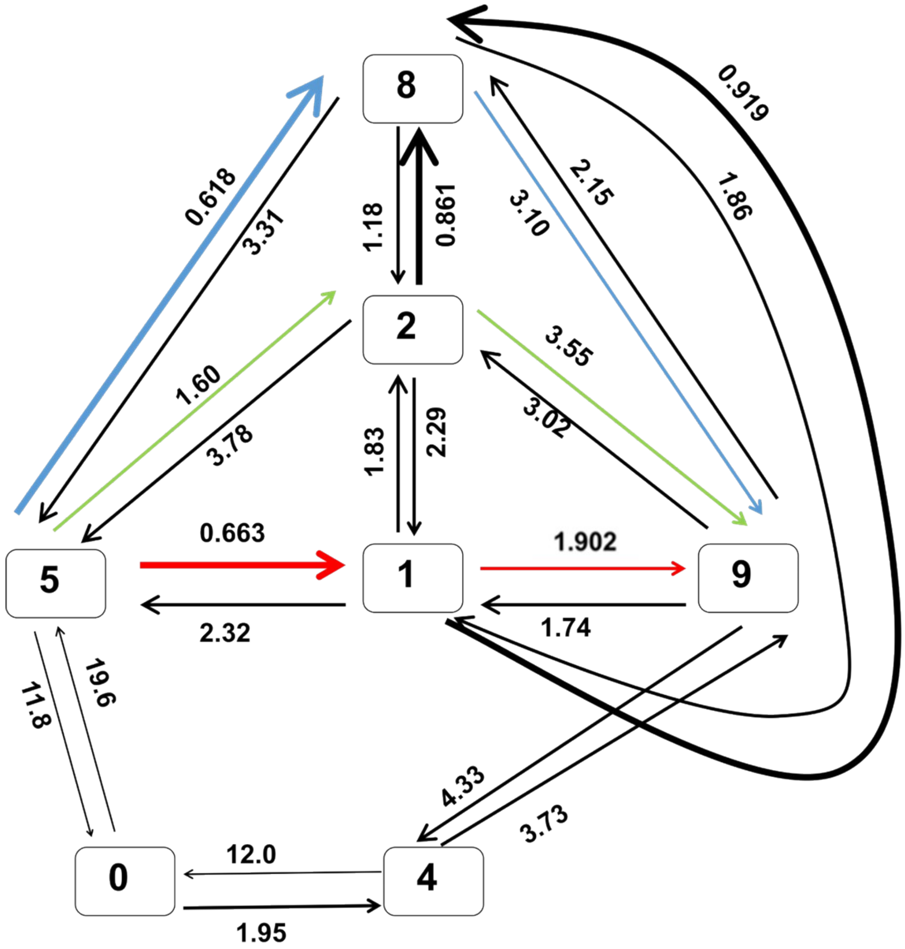


Figure S22. Folding pathways of DNA dUdUdCdG from DNA state 5 (unfolded state) to DNA state 9 (folded state). Arrow colors denote different pathways and the weight of each arrow indicates the transition probability between states. Transition times (in μs) between two states are derived from MFPT calculations. The red, blue and green curves represent the first, second and third folding pathway, respectively.

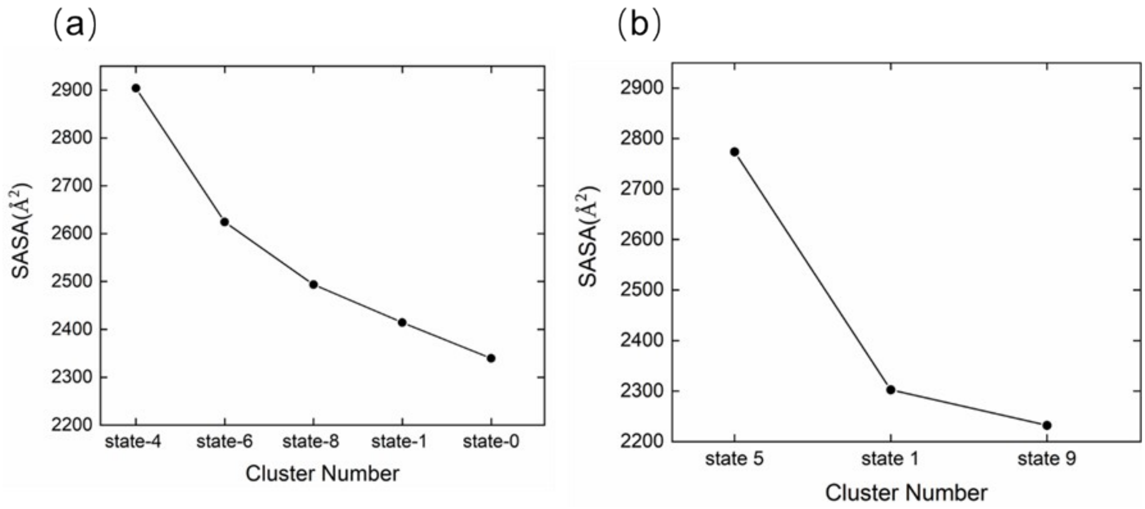
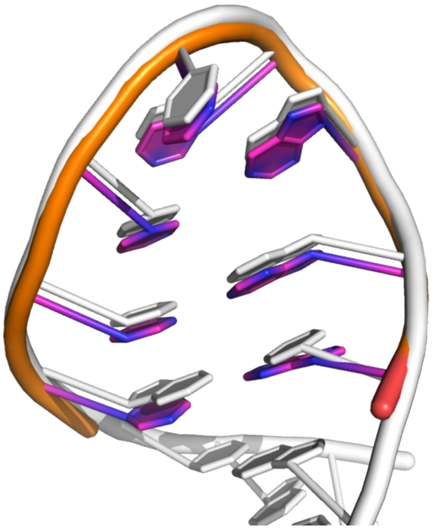
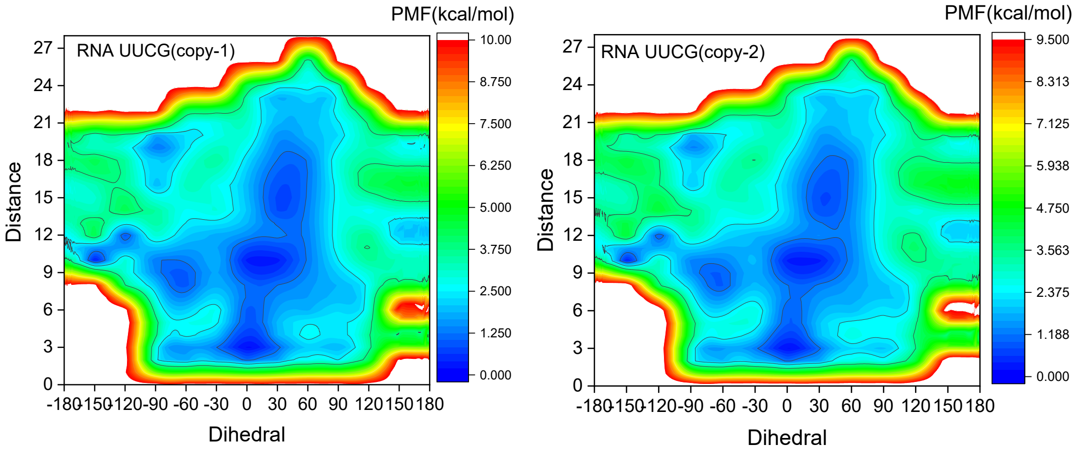


Figure S23. SASA for each cluster in the maximum path flux of MSM analysis, (a) RNA UUCG, (b) DNA dUdUdCdG

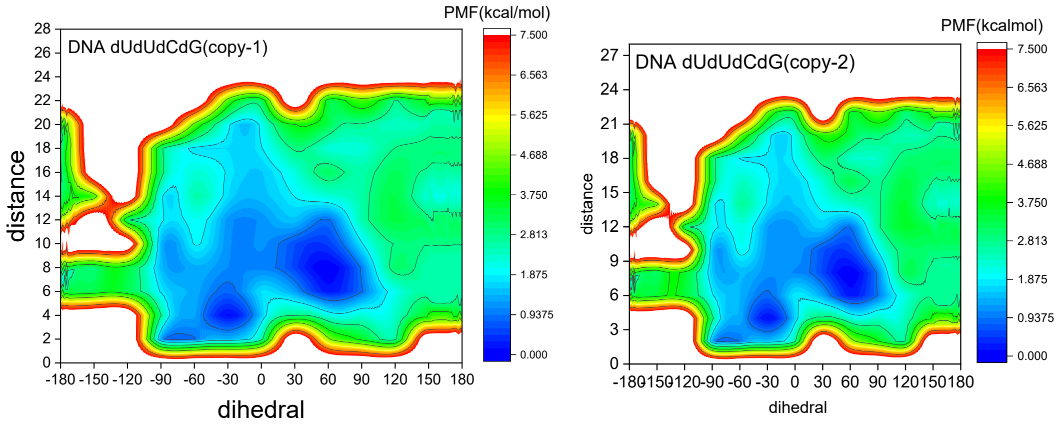
Note S1. **The comparison between the modeling folded state and the referenced NMR (PDB: 2KOC, using MODEL 1) structures. The RMSD of the alignment is 1.2 Å.**



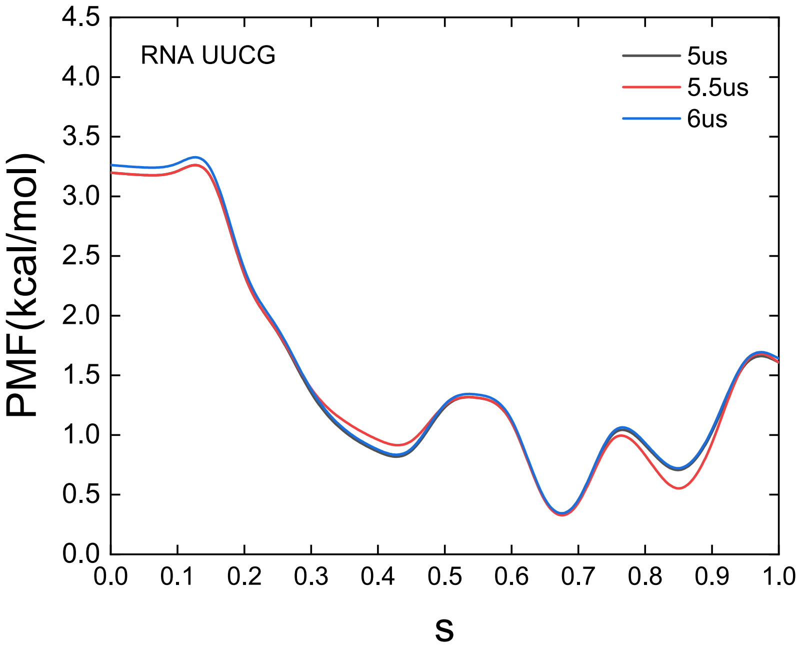
Note S2. Free energy landscape of RNA UUCG folding–unfolding equilibrium projected on the dihedral parameter and the stem distance in the replicas.



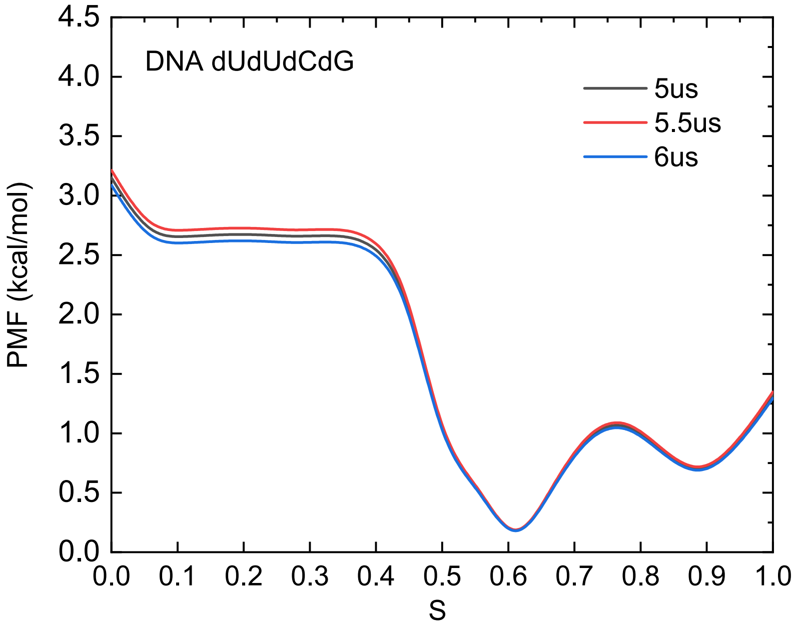
Note S3. Free energy landscape of DNA dUdUdCdG folding–unfolding equilibrium projected on the dihedral parameter and the stem distance in the replicas.



Note S4. Convergence analysis of GaMD sampling for RNA UUCG



Note S5. Convergence analysis of GaMD sampling for DNA dUdUdCdG



Note S6. Average number of hydrogen bonds in unfolded, misfolded, and folded states of RNA UUCG and DNA dUdUdCdG.

System	Average number of hydrogen bonds in the Unfolded state	Average number of hydrogen bonds in the Misfolded state	Average number of hydrogen bonds in the Folded state
RNA	3	9	8
DNA	2	4	7