Electronic Supplementary Material (ESI) for RSC Advances.

Supporting Information

Methylation on CpG Repeats Modulates Hydroxymethylcytosine Induced Duplex Destabilization

Qiong Wu, Jiun Ru Wong, Penny Liu Qing Yeo, Dawei Zhang, and Fangwei Shao*

Division of Chemistry and Biological Chemistry, Nanyang Technological University, 21 Nanyang Link, Singapore 637371, Singapore. fwshao@ntu.edu.sg, phone: +65-6592-2511.

1 General method and materials

Chemicals used were purchased either from Aldrich or Alfa Aesar. Water used for buffer was obtained from Milli-Q filtration system (18.2 MΩ.cm). Reagents for DNA and RNA syntheses, 5-mdC amidite were purchased from Glen Research. Standard CPG supports (1 μmol) are purchased from BioAutomation. All chemicals used as received without further purification. Mermade 4 automated DNA synthesizer was used for the synthesis of all the DNA and RNA strands. Shimadzu reverse phase HPLC (LC-20AD) with column (Microsorb 100-5 C18 Dynamax, 250 × 10.0 mm) was used to purify the DNA and RNA strands. CD spectrums were recorded on a Jasco J-810 CD Spectro-polarimeter. UV-Melting temperature of duplex was analyzed on Shimadzu UV-2550 UV-VIS spectrophotometer equipped with TMSPC-8 peltier controller.

For UV-melting analysis in buffer solution, A-form or B-form DNA duplexes (1.5 μM in 20 mM sodium phosphate pH=7) were prepared by heating the samples from 15 °C to 90 °C at a rate of 0.5 °C/min, then cooling down to 15 °C at a rate of 0.5 °C/min. UV melting data was collected by heating annealed duplex sample from 15 °C to 90 °C at a rate of 0.5 °C/min with measurement interval of 1 °C. The T_m curves of three repeats were analyzed with sigmoidal fitting in OriginPro 8.5.1. Average value of three repeats was used as T_m of the duplex.

For UV-melting analysis in 16% Ficoll solution, A-form or B-form DNA duplexes in sodium phosphate buffer were prepared by heating the samples at 90 °C for 5 minutes then cooling down to room temperature over 3 hours. To the annealed DNA sample, Ficoll (48% as stock solution) was added and mixed thoroughly. The final concentration of DNA, sodium phosphate buffer and Ficoll were 1.5 μM, 20 mM, and 16%, respectively. UV melting data was collected by heating annealed duplex sample from 15 °C to 90 °C at a rate of 0.5 °C/min with measurement interval of 1 °C. The T_m curves of three repeats were analyzed with sigmoidal fitting in OriginPro 8.5.1. Average value of three repeats was used as T_m of the duplex.

2 DNA and RNA sequences preparation and characterization

CHC and **MHM** were synthesized according to previous report.[1] The rest oligonucleotides were synthesized on a MerMade 4 DNA Synthesizer according to manufacturer's recommended procedures. Upon the completion of synthesis, DNA sequence was cleaved from the CpG support by treating with 28% NH_4OH aqueous solution. The collected NH_4OH solution was incubated at 60 °C for 15 hours to remove protecting group of nucleosides. DNA with DMT group was purified by reverse phase HPLC. The DMT group on DNA was later removed by incubating the sample with 80% acetic acid solution (200 μL) at room temperature for 15 min. DNA without DMT group was collected by precipitation with ethanol (800 μL) and purified by HPLC. RNA (*g*) was prepared as the recommended procedure provided by manufacturer. As shown in Table S1, all of the purified oligonucleotides were confirmed by ESI-mass spectroscopy (Sangon Ltd. Shanghai, China).

Table S1. Calculated Mass and ESI-Mass of DNA and RNA sequences

	DNA	Sequence (5'-3')	**Calculated Mass *m/z***	**ESI-Mass found *m/z***
ODN1	**CCC**	TTC CAC G**C**G CGT TCC TGA CTG ACT C	7544.9	7544.3
	CHC	TTC CAC G**hmC**G CGT TCC TGA CTG ACT C	7574.9	7577.8
	CMC	TTC CAC G**mC**G CGT TCC TGA CTG ACT C	7558.9	7558.3
ODN2	**MCM**	TTC CA**mC** G**C**G **mC**GT TCC TGA CTG ACT C	7573.0	7572.2
	MHM	TTC CA**mC** G**hmC**G **mC**GT TCC TGA CTG ACT C	7603.0	7608.2
	MMM	TTC CA**mC** G**mC**G **mC**GT TCC TGA CTG ACT C	7587.0	7587.4
ODN3	**G**	GAG TCA GTC AGG AAC GC***G*** CGT GGA A	7781.1	7781.2
ODN4	**A**	GAG TCA GTC AGG AAC GC***A*** CGT GGA A	7765.1	7765.5
ODN5	**C**	GAG TCA GTC AGG AAC GC***C*** CGT GGA A	7741.1	7740.2
ODN6	**T**	GAG TCA GTC AGG AAC GC***T*** CGT GGA A	7756.1	7755.4
RNA1	***g***	gag tca gtc agg aac gcg cgt gga a	8139	8140.5

3 CD spectrums of A-form and B-form duplexes in buffer and crowding condition

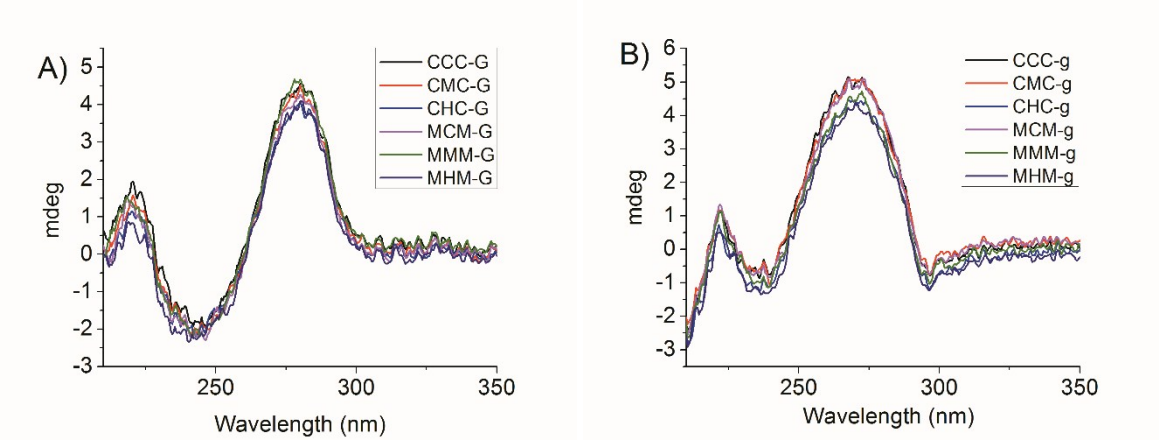

Figure S1. CD spectrums of A) **CXC**/**MXM**-**G** duplexes and B) **CXC**/**MXM**-**g** duplexes in sodium phosphate buffer (20 mM, pH=7), **X**=**C**/**M**/**H**.

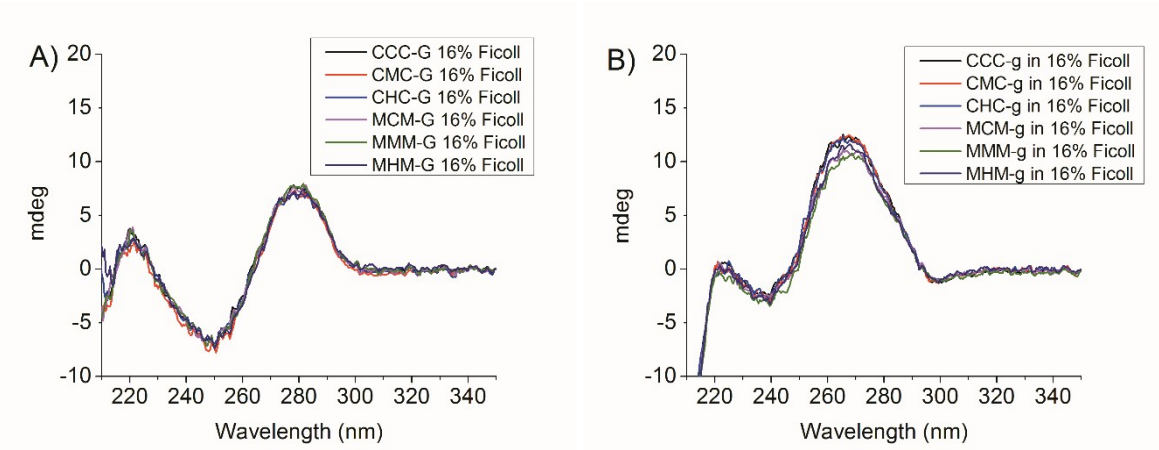

Figure S2. CD spectrums of A) **CXC**/**MXM**-**G** duplexes and B) **CXC**/**MXM**-**g** duplexes in sodium phosphate buffer (20 mM, pH=7) mixed with 16% of Ficoll, **X**=**C**/**M**/**H**.

4 Validation of T_m analysis method

Table S2. Three times individually repeats of T_m °C analysis for **CCC**-**G** in buffer condition

T_m	**CCC**-**G**
1	62.09
2	62.01
3	62.04

The UV melting of **CCC**-**G** in sodium phosphate buffer (20 mM, pH=7) was collected for three times. For each time, T_m was calculated as an average value from three simultaneous repeats, as show in Table S2. Sample was annealed by heating the sample from 15 °C to 90 °C in a rate of 0.5 °C/min, then cool down to 15 °C in a rate of 0.5 °C/min. UV melting data was collected by heating annealed duplex sample from 15 °C to 90 °C in a rate of 0.5 °C/min with measurement interval of 1 °C. The average value of three times individually repeats is 62.05 °C with standard deviation of 0.05 °C. Therefore, data acquired by using current method has good repeatability.

5 Melting Temperatures of A-form and B-form duplexes in buffer and crowding condition

Table S3. T_m (°$\mathbb{C}$) of A-form and B-form duplexes in buffer and crowding condition

		CCC	**CMC**	**CHC**	**MCM**	**MMM**	**MHM**
Buffer	**G (ODN3)**	62.05(5)	62.4(2)	61.0(1)	62.6(2)	63.0(2)	62.6(3)
	g (RNA1)	62.63(5)	63.4(3)	62.0(1)	63.65(2)	64.6(4)	64.3(4)
16% Ficoll	**G (ODN3)**	65.3(1)	65.6(2)	64.66(7)	65.6(3)	66.1(2)	66.1(3)
	g (RNA1)	66.02(3)	66.8(2)	65.00(6)	66.9(1)	67.7(2)	67.1(1)

* Standard deviation on last digit is in parentheses.

6 Melting Temperatures of mismatched duplexes in buffer condition

Table S4. T_m (°C) of mismatched duplexes in buffer condition

	CCC	**CMC**	**CHC**	**MCM**	**MMM**	**MHM**
A (ODN4)	57.04(7)	57.6(2)	56.03(5)	57.37(6)	58.1(2)	57.6(3)
C (ODN5)	54.73(4)	55.2(2)	53.23(2)	55.4(3)	55.7(2)	54.9(3)
T (ODN6)	54.2(2)	54.4(3)	52.77(3)	54.47(8)	55.2(3)	54.2(2)

* Standard deviation on last digit is in parentheses.

7 Optimization of C-G, hmC-G, mC-G base pairs and calculation of their hydrogen bonding energies

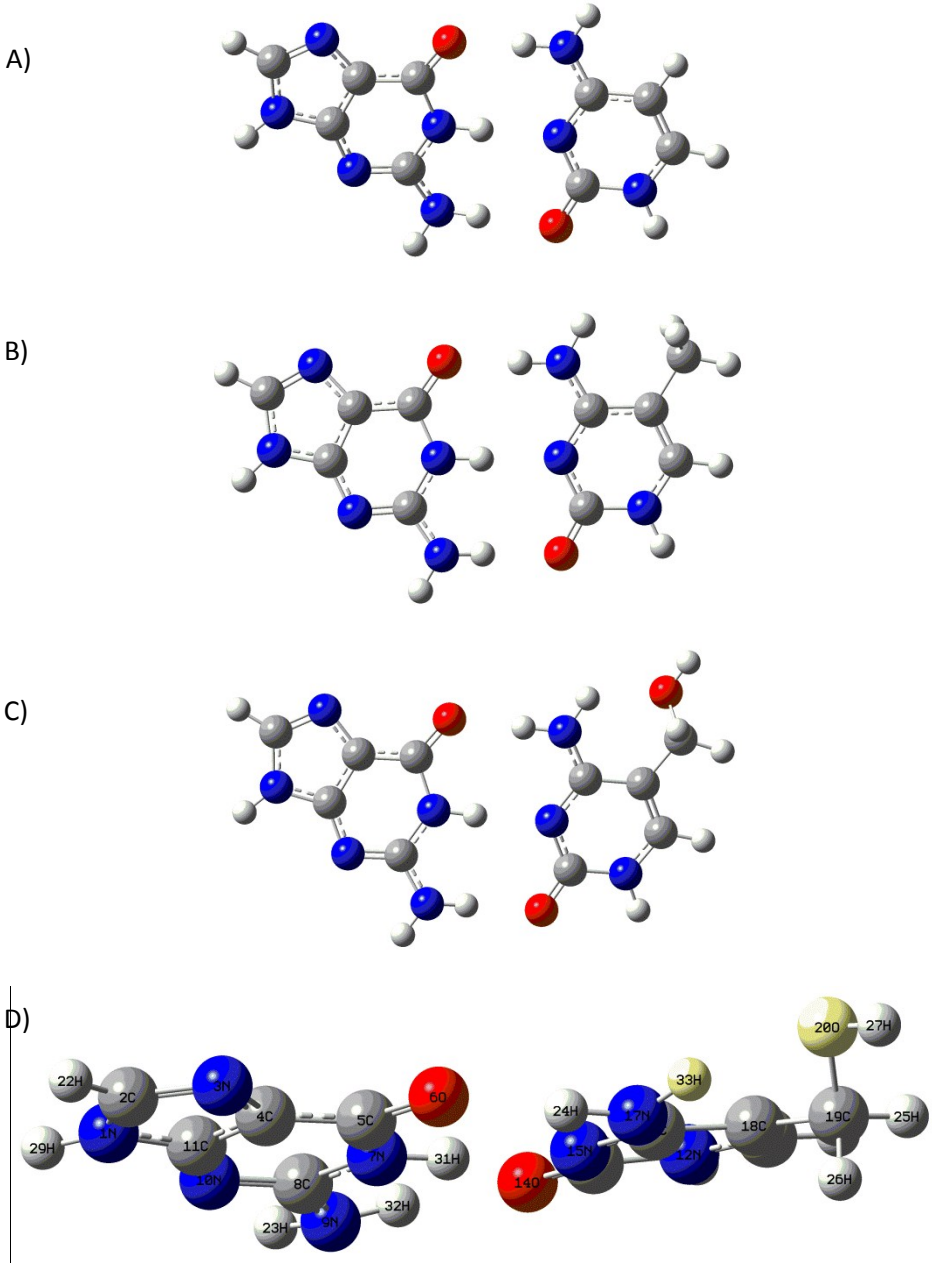

Figure S3. Optimized C-G, hmC-G, mC-G base pairs. The initial structures of C-G, mC-G, hmC-G base pairs were extracted from reported crystal structures. C-G and hmC-G were extracted from **PDB 4GLC**, mC-G was from **PDB 4GJU**.[2] The Simulation was performed on the platform of Gaussian 09 at MP2/6-31G(d) level. A) front view of C-G base pair; B) front view of mC-G base pair; C) front view of hmC-G base pair; D) side view of hmC-G base pair.

Table S5. Total energies of base pairs (E_{tot}) and hydrogen bonding interaction energies (E_{int}) of base pairs ($E_{int} = E_{tot} - E_{Guanine} - E_{Cytosine}$). Energy calculations for base pairs and individual bases were performed at MP2/aug-cc-pvdz level with Gaussian 09.

	E_{tot} hartree	$E_{Cytosine}$ hartree	$E_{Guanine}$ hartree	E_{int} hartree	E_{int} kcal mol^{-1}
C-G	-935.1347879	-393.910505	-541.1764782	-0.047805	-29.9981
hmC-G	-1049.390317	-508.167464	-541.1764781	-0.0463746	-29.1005
mC-G	-974.3285951	-433.103321	-541.1764782	-0.0487961	-30.62

Table S6. Total energies of base pairs (E_{tot}) and hydrogen bonding interaction energies (E_{int}) of base pairs ($E_{int} = E_{tot} - E_{Guanine} - E_{Cytosine}$). Structure optimizations and energy calculations for base pairs and bases were performed at M06-2X/6-31+G(d,p) level by using Gaussian 09.

	E_{tot} hartree	$E_{Cytosine}$ hartree	$E_{Guanine}$ hartree	E_{int} hartree	E_{int} kcal mol^{-1}
C-G	-937.2220926	-394.79668	-542.3790456	-0.0455348	-28.5735
hmC-G	-1051.713941	-509.289872	-542.3798775	-0.0441912	-27.7304
mC-G	-976.5224447	-434.096208	-542.3798774	-0.0463592	-29.0908

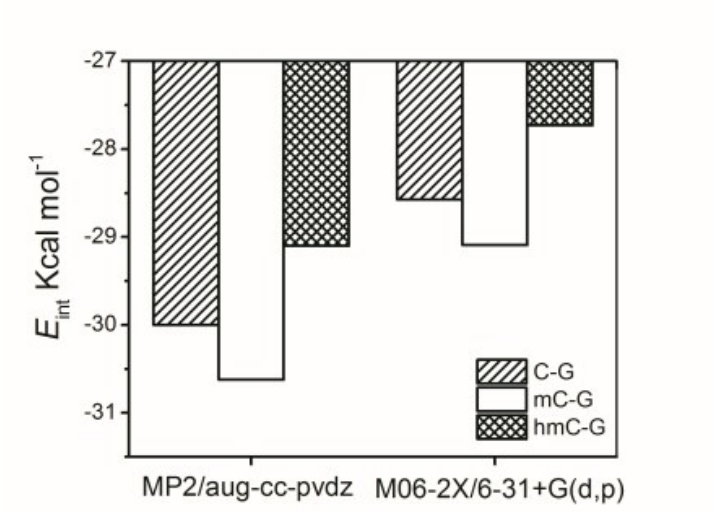

Figure S4. Similar trends of E_{int} for the three base pairs can be obtained by applying both MP2/aug-cc-pvdz method and M06-2X/6-31+G(d,p) method. MP2 (second-order Moeller-Plesset) theory, is considered as a standard method for introducing dispersion energy in theoretical studies. The aug-cc-pVDZ basis set has been proved to gives a reliable estimate of the interaction energy between mC-G base pair.[3] Beside MP2 method, density functional theory (DFT) is another widely used method in the computational chemistry community. M06-2X, designed by Truhlar and co-workers, predicts accurate valence and is an excellent functional for the study of noncovalent interactions.[4] These calculation

results showed consistent trend that hmC-G base pair maintained the weakest H-bonding interaction comparing to C-G and mC-G base pairs.

8 Cartesian Coordinates for All Species

Table S7. C-G base pair optimized coordinates, at MP2/6-31G(d) level

Tag	Symbol	X	Y	Z
1	N	-4.673683	0.483845	-0.101785
2	C	-4.974102	-0.857475	-0.171017
3	N	-3.901302	-1.630353	-0.134514
4	C	-2.85624	-0.733571	-0.03412
5	C	-1.443658	-0.96399	0.033671
6	O	-0.834826	-2.048172	0.019309
7	N	-0.746251	0.250592	0.128104
8	C	-1.301951	1.509452	0.147919
9	N	-0.433629	2.54367	0.321055
10	N	-2.601392	1.741314	0.075916
11	C	-3.308474	0.586495	-0.01156
12	N	4.335494	0.916886	-0.169074
13	C	2.940704	1.068886	-0.118965
14	O	2.466446	2.211341	-0.198339
15	N	2.198003	-0.069737	0.013744
16	C	2.787131	-1.268221	0.073986
17	N	1.994958	-2.339908	0.193554
18	C	4.217446	-1.4256	0.027081
19	C	4.955623	-0.293231	-0.096989
20	H	-5.99455	-1.210747	-0.247387
21	H	0.55672	2.422856	0.095308
22	H	2.389408	-3.269507	0.194106
23	H	4.68804	-2.399852	0.08424
24	H	6.040023	-0.28875	-0.144917
25	H	-5.31837	1.265245	-0.106399
26	H	4.862381	1.7783	-0.265129
27	H	0.285584	0.166204	0.141575
28	H	-0.834902	3.453666	0.137455
29	H	0.968455	-2.233403	0.154359

Table S8. mC-G base pair optimized coordinates, at MP2/6-31G(d) level

Tag	Symbol	X	Y	Z
1	N	-3.934564	1.461901	-0.162253
2	C	-2.535527	1.483119	-0.131378
3	N	-1.907164	0.275688	-0.020597
4	C	-2.608281	-0.860353	0.039815
5	C	-4.054497	-0.896442	0.016962
6	C	-4.815679	-2.185816	0.099424
7	C	-4.666546	0.314136	-0.088428
8	O	-1.952012	2.575472	-0.206868
9	N	-1.914562	-2.002109	0.139656
10	N	4.988208	0.18367	-0.070215
11	C	5.163258	-1.179184	-0.148909
12	N	4.022784	-1.848966	-0.132195
13	C	3.064613	-0.859445	-0.035364
14	C	1.635987	-0.957587	0.012705
15	O	0.929411	-1.980548	-0.019528
16	N	1.053166	0.315409	0.110415
17	C	1.723407	1.516841	0.150238
18	N	0.953526	2.625466	0.323916
19	N	3.039904	1.627145	0.095645
20	C	3.637414	0.41245	0.004797
21	H	-5.892218	-1.997477	0.077026
22	H	-4.590867	-2.723996	1.02691
23	H	-4.577354	-2.847545	-0.741021
24	H	-5.747003	0.426793	-0.118898
25	H	-0.883298	-1.98514	0.082117
26	H	6.147261	-1.625381	-0.216775
27	H	1.43958	3.496226	0.155924
28	H	-4.381741	2.368905	-0.242346
29	H	5.702781	0.901624	-0.060385
30	H	0.017519	0.327626	0.112559
31	H	-2.385093	-2.894656	0.134973
32	H	-0.042747	2.599886	0.091315

Table S9. hmC-G base pair optimized coordinates, at MP2/6-31G(d) level

Tag	Symbol	X	Y	Z
1	N	5.2316	-0.102191	-0.132887
2	C	5.286262	-1.476155	-0.188522
3	N	4.092981	-2.043948	-0.129484
4	C	3.2266	-0.973607	-0.028687
5	C	1.795919	-0.947102	0.05938
6	O	1.002092	-1.903026	0.067636
7	N	1.32942	0.374936	0.145848
8	C	2.101732	1.513345	0.14305
9	N	1.435005	2.689091	0.314363
10	N	3.420493	1.508172	0.05278
11	C	3.907986	0.244137	-0.028343
12	N	-3.54061	1.925654	-0.266382
13	C	-2.145827	1.836302	-0.136556
14	O	-1.473955	2.873126	-0.237651
15	N	-1.625404	0.591793	0.076322
16	C	-2.418774	-0.480045	0.154569
17	N	-1.852314	-1.680302	0.335248
18	C	-3.863296	-0.383429	0.103818
19	C	-4.737796	-1.56066	0.385459
20	O	-4.430787	-2.594352	-0.563113
21	C	-4.371018	0.856523	-0.129503
22	H	6.225879	-2.007261	-0.272281
23	H	1.991345	3.508397	0.108471
24	H	-0.82572	-1.755927	0.265829
25	H	-5.792159	-1.258994	0.314918
26	H	-4.555262	-1.919144	1.409316
27	H	-4.94116	-3.384925	-0.312542
28	H	-5.437146	1.054931	-0.202247
29	H	6.005349	0.551218	-0.155742
30	H	-3.901494	2.857936	-0.43905
31	H	0.299489	0.476564	0.175824
32	H	0.438969	2.744576	0.087841
33	H	-2.415814	-2.49165	0.10605

Table S10. C-G base pair optimized coordinates, at M06-2X/6-31+G(d,p) level

Tag	Symbol	X	Y	Z
1	N	4.664283	0.456713	-0.000147
2	C	4.950682	-0.895978	-0.00051
3	N	3.882153	-1.639117	-0.000548
4	C	2.836527	-0.738639	-0.000188
5	C	1.420157	-0.940959	-0.00006
6	O	0.790377	-2.001579	-0.000227
7	N	0.730754	0.278431	0.000301
8	C	1.302618	1.525823	0.000599
9	N	0.454685	2.57213	0.001304
10	N	2.605716	1.728778	0.000453
11	C	3.303372	0.573288	0.000074
12	N	-4.317906	0.92393	-0.000516
13	C	-2.922234	1.057691	-0.000641
14	O	-2.443881	2.191313	-0.001134
15	N	-2.184192	-0.080398	-0.000236
16	C	-2.770009	-1.27758	0.000289
17	N	-1.985054	-2.353355	0.000669
18	C	-4.207458	-1.424505	0.000448
19	C	-4.936393	-0.285803	0.000024
20	H	5.968977	-1.258397	-0.000753
21	H	-0.559477	2.461923	0.000129
22	H	-2.386914	-3.27624	0.00097
23	H	-4.677847	-2.398336	0.000902
24	H	-6.020518	-0.276333	0.000068
25	H	5.316706	1.225772	0.000074
26	H	-4.839647	1.789513	-0.000925
27	H	-0.29881	0.198231	0.000273
28	H	0.86612	3.490341	0.000597
29	H	-0.957208	-2.24415	0.000384

Table S11. mC-G base pair optimized coordinates, at M06-2X/6-31+G(d,p) level

Tag	Symbol	X	Y	Z
1	N	-3.906673	1.474378	-0.005939
2	C	-2.509469	1.470742	-0.00179
3	N	-1.89291	0.263091	0.003394
4	C	-2.596817	-0.867679	0.003533
5	C	-4.049751	-0.887239	-0.0005
6	C	-4.814156	-2.178374	0.000309
7	C	-4.643852	0.329116	-0.005132
8	O	-1.915082	2.549916	-0.002818
9	N	-1.919339	-2.015431	0.008017
10	N	4.974064	0.13784	-0.001621
11	C	5.123582	-1.237088	-0.003945
12	N	3.985901	-1.869291	-0.003999
13	C	3.035892	-0.868338	-0.001704
14	C	1.606029	-0.92792	-0.000707
15	O	0.875519	-1.921751	-0.001633
16	N	1.041543	0.353854	0.001661
17	C	1.736096	1.537776	0.003305
18	N	0.996997	2.663391	0.007197
19	N	3.053117	1.60905	0.002307
20	C	3.631724	0.389893	-0.000134
21	H	-5.889721	-1.988629	-0.00231
22	H	-4.585976	-2.77852	0.888499
23	H	-4.582264	-2.781943	-0.884609
24	H	-5.722421	0.45052	-0.0085
25	H	-0.885803	-2.001199	0.008031
26	H	6.100628	-1.699604	-0.005439
27	H	1.497963	3.536013	0.003528
28	H	-4.343293	2.385782	-0.009444
29	H	5.699636	0.838372	-0.001134
30	H	0.008643	0.375457	0.002846
31	H	-2.40186	-2.898412	0.006751
32	H	-0.023588	2.653337	0.000861

Table S12. hmC-G base pair optimized coordinates, at M06-2X/6-31+G(d,p) level

Tag	Symbol	X	Y	Z
1	N	-5.221939	-0.12509	0.086878
2	C	-5.266053	-1.506133	0.034359
3	N	-4.084086	-2.047394	-0.032109
4	C	-3.213893	-0.976449	-0.023953
5	C	-1.784151	-0.924823	-0.077634
6	O	-0.978631	-1.855222	-0.14656
7	N	-1.320984	0.397895	-0.042878
8	C	-2.103598	1.52281	0.02887
9	N	-1.454712	2.702944	0.042759
10	N	-3.42102	1.490962	0.079089
11	C	-3.903554	0.230603	0.049637
12	N	3.529661	1.93951	0.077871
13	C	2.135293	1.825498	-0.006002
14	O	1.46109	2.854855	0.025126
15	N	1.617802	0.576736	-0.102004
16	C	2.407566	-0.497161	-0.127249
17	N	1.835626	-1.696489	-0.18811
18	C	3.858043	-0.388908	-0.117345
19	C	4.737036	-1.584141	-0.314721
20	O	4.444104	-2.54998	0.689738
21	C	4.355893	0.863445	0.006426
22	H	-6.203865	-2.043364	0.050091
23	H	-2.021224	3.531392	0.115131
24	H	0.806305	-1.760033	-0.162794
25	H	5.787592	-1.270185	-0.266185
26	H	4.554211	-2.015127	-1.31033
27	H	4.989212	-3.332678	0.551064
28	H	5.421682	1.067557	0.039237
29	H	-5.998519	0.51598	0.142097
30	H	3.888217	2.880403	0.164335
31	H	-0.294558	0.503558	-0.076412
32	H	-0.437023	2.775101	0.030564
33	H	2.397535	-2.511792	0.011977

9 References

[1] S. Xuan, Q. Wu, L. Cui, D. Zhang, F. Shao, *Bioorg. Med. Chem. Lett. ,* **2015**, *25*, 1186-1191.

[2] D. Renciuk, O. Blacque, M. Vorlickova, B. Spingler, *Nucleic Acids Res. ,* **2013**, *41*, 9891-9900.

[3] Q. Song, Z. Qiu, H. Wang, Y. Xia, J. Shen, Y. Zhang, *Struct. Chem.,* **2013**, *24*, 55-65.

[4] Y. Zhao, D. G. Truhlar, *Acc. Chem. Res. ,***2008**, *41*, 157-167.