Supporting Information

Characterizing the Protonation State of Cytosine in Transient G•C Hoogsteen Base Pairs in Duplex DNA

Evgenia N. Nikolova, Garrett B. Goh, Charles L. Brooks III*, and Hashim M. Al-Hashimi*,

Department of Chemistry and Biophysics, University of Michigan, 930 North University Avenue, Ann Arbor, MI 48109, USA

Supplementary Methods

NMR spectroscopy

NMR samples. DNA samples were prepared as previously described[1,2] in NMR buffer containing 15 mM Sodium phosphate (pH 6.8), 25 mM NaCl, 0.1 mM EDTA, 10% D_2O. For pH titrations of unlabeled A_6-DNA^{1mG10}, the buffer was initially exchanged to NMR buffer (pH 5.2) by washing 3X with Amicon Ultra-4 centrifugal filter. The pH was subsequently adjusted by titrating dilute NaOH solution directly into the NMR sample and monitoring with a pH meter. For the labeled A_6-DNA sample (pH 6.8), pH was varied by either buffer exchange (pH 5.4 to 7.6) or titration of dilute HCl (pH 5.2 to 4.3), as described above.

NMR pH titration and pK_a calculations. All NMR experiments were performed on a Bruker Avance 600 MHz NMR spectrometer equipped with a 5mm triple-resonance cryogenic probe. Standard 2D $^1H,^1H$ NOESY experiments at 26 ^oC and pH 5.2 or 9.2 were used as outlined before[1] to assign and examine the conformational features of A_6-DNA^{1mG10} at variable pH. 2D $^1H,^{13}C$ HSQC correlation spectra for base C2H2, C5H5, C6H6, and C8H8 and sugar C1'H1' of A_6-DNA^{1mG10} were collected as a function of pH (only pH 5.2 and 9.2 for C1'H1'). The average chemical shift perturbation (CSP) was monitored, where CSP is defined as $\sqrt{\Delta\delta_{1H}^2 + (0.25\Delta\delta_{13C})^2}$ (1) $\Delta\delta$ represents the ^1H or ^{13}C chemical shift change relative to the lowest pH 5.2. A pseudo 2-state equilibrium (Scheme 1) was assumed for the deprotonation of HG^+ in fitting of the resultant CSP values.

Scheme S1: 2-state model

$$HG^+ \underset{}{\overset{K_{obs}}{\rightleftharpoons}} WC^* + H^+$$

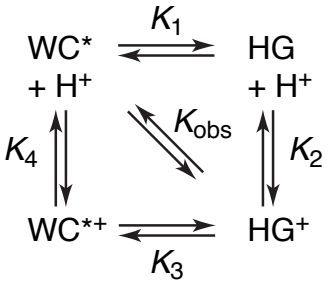
Individual and global data fits of the CSP values as a function of pH to the modified Henderson-Hasselbalch equation were performed using Origin 8.5 (Table S1):

$$\mathrm{CSP} = \left((\delta_A - \delta_B 10^{n(pK_a - pH)}) / (1 + 10^{n(pK_a - pH)}) \right), \quad (2)$$

where δ_A and δ_B are the CSP limits at low and high pH, respectively, and *n* (Hill's coefficient) is set to 1.

The following 4-state mechanism (Scheme S2) was proposed for the pH dependent transition of the 1mG•C^+ HG base pair because of the large conformational changes suggesting formation of a stable distorted WC* base pair at high pH:

Scheme S2: 4-state model



In the context of Scheme S2, the observed pK_a can be represented as a macroscopic protonation constant ($pK_{a,obs}$) as previously described:[3]

$$pK_{a,obs} = -\log(f_{HG^+} K_{HG^+} + f_{WC^+} K_{WC^+}) \quad (3)$$

where f_{HG^+} and $f_{WC^{*+}}$ are the fraction of HG^+ and WC^{*+} out of all protonated species, $K_{HG^+} = 1/K_2$ and $K_{WC^{*+}} = 1/K_4$ denote the equilibrium constants for protonation of HG^+ and WC^{*+}, respectively, and $pK_{HG^+} = -\log(K_{HG^+})$ represents the pK_a for cytosine N3 in an HG base pair. Since the conformational exchange between HG^+ and WC^{*+} is at equilibrium, the relative fractions f_{HG^+} and $f_{WC^{*+}}$ can be assumed to be independent of pH. The NMR data strongly suggests that the major protonated species in the HG trapped A_6-DNA^{1mG10} sample is HG^+ and not WC^{*+}. This is based on the observed NOEs, which indicate a *syn* base for 1mG10 (Figure 2A) and yield chemical shifts characteristics of protonated cytosine N3 (Figure S1) at pH 5.2. Therefore, it follows that $f_{HG^+} \rightarrow 1$ and $f_{WC^{*+}} = 1 - f_{HG^+}$ is relatively small (< 0.1). Based on these assumptions, we can approximate the minimum value for pK_{HG^+} to ~ 0.05 units lower than the observed than $pK_{a,obs} \sim 7.24$. We can make further assumptions about the equilibrium constants for the two conformational exchange reactions. For example, the proton affinity of C N3 in a WC* base pair, where it is either solvent exposed or engaged in a weaker H-bond to a nearby G donor (for example, G O6), would in all likelihood be much lower than that in an HG base pair ($pK_{WC^{*+}} < pK_{HG^+}$ or $K_{WC^{*+}} >> K_{HG^+}$) but higher

than that in a canonical WC base pair ($pK_{WC^{*+}}$), where C N3 is deeply buried and forms an H-bond to G N1. Based on this, it is likely that $f_{WC^{*+}}K_{WC^{*+}} >> f_{HG^+}K_{HG^+}$, which would increase the value of pK_{HG^+} above the observed ~ 7.2. Based on the above analysis, $pK_{HG^+} \geq 7.2$.

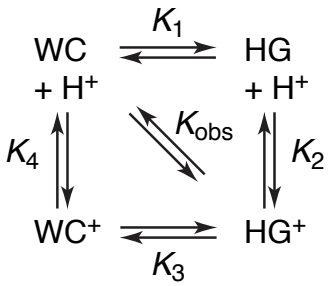
pH and salt dependence for NMR relaxation dispersion. Rotating frame ($R_{1\rho}$) carbon relaxation dispersion profiles for G10 C8' (or C1') and A16 C8 in labeled A_6-DNA were measured and analyzed as previously reported[1] at variable pH from 4.3 to 6.8 for G10 and 4.0 to 7.6 for A16 (Table S2, Figure S3) and variable monovalent (Na^+) and divalent (Mg^{2+}) salt concentration (Figure S4). All relaxation data was analyzed assuming a 2-state exchange process as reported previously.[1] The free energy differences for the forwards and backward transition (Figure S3) were calculated from the corresponding rate constants using Transition State Theory.

As in the analysis of CSPs, the pH dependence of the transient G•C^+ HG base pair population (p_B) was analyzed assuming a reduced 2-state equilibrium, as used for analysis of relaxation dispersion data (Scheme S3), and a more plausible 4-state equilibrium (Scheme S4). In what follows, we derive equations analogous to those used to describe the pH dependence of CSPs that describe the pH dependence of the measured p_B.

Scheme S3: 2-state model

$$WC + H^+ \underset{}{\overset{K_{obs}}{\rightleftharpoons}} HG^+$$

Scheme S4: 4-State model



From Scheme S3, the observed equilibrium constant for the WC to HG^+ transition is given by $K_{obs} = [HG^+]/([WC][H^+]) = 1/K_{a,obs}$, where $K_{a,obs}$ is the observed protonation constant. The fraction or population of WC and HG^+ species can be defined as:

$$p_{WC} = \frac{[WC]}{[WC]+[HG^+]} \quad \text{and} \quad p_{HG^+} = \frac{[HG^+]}{[WC]+[HG^+]} \tag{4}$$

From the above equations, we can derive an expression for the HG^+ population as a function of $[H^+]$:

$$p_{HG^+} = \frac{[HG^+]}{[HG^+]/(K_{obs}[H^+])+[HG^+]} = \frac{K_{obs}[H^+]}{K_{obs}[H^+]+1} \tag{5}$$

Since $pK_{a,obs} = \log(K_{obs})$ and $pH = -\log[H^+]$, the population of HG^+ species obtained from relaxation dispersion measurements (p_B) can be approximated as p_{HG^+} (since WC and HG^+ are the predominant species in this pH range), expressed as a function of pH and $pK_{a,obs}$:

$$p_B = p_{HG^+} = \frac{10^{pK_{a,obs}-pH}}{1+10^{pK_{a,obs}-pH}} \tag{6}$$

Note that Eq. 6 is analogous to Eq. 2 and describes the pH dependence of p_B (rather than CSP).

To derive an expression for the pH dependence of p_B assuming Scheme S4, we define the following equilibrium constants for the four steps:

$$K_1 = \frac{[HG]}{[WC]},\ K_2 = \frac{[HG^+]}{[HG][H^+]},\ K_3 = \frac{[HG^+]}{[WC^+]}, \text{ and } K_4 = \frac{[WC^+]}{[WC][H^+]}, \tag{7}$$

where $K_2 = 1/K_{HG^+}$ and $K_4 = 1/K_{WC^+}$ are the inverse of the protonation constants for HG and WC species. The population of HG^+ base pairs can be expressed as:

$$p_{HG^+} = \frac{[HG^+]}{[WC]+[WC^+]+[HG]+[HG^+]} \tag{8}$$

Analogous to the derivation for Scheme S2, we can use Eq. 7 and Eq. 8 to derive an expression for p_B in terms of pH and the pK_a for HG ($pK_{HG^+} = \log K_2$) and WC ($pK_{WC^+} = \log K_4$) base pairs given by:

$$p_B = p_{HG^+} \frac{K_1 10^{pK_{HG+}-pH}}{1+K_1+10^{pK_{WC+}-pH}+K_1 10^{pK_{HG+}-pH}} \tag{9}$$

The pH dependence of the best-fit p_B values for G10 C8/C1' in A_6-DNA obtained from NMR relaxation dispersion was fit to the 2-state (Eq. 6) and 4-state (Eq. 9) model to extract the equilibrium constants. Because of the large number of unknowns in Eq. 9 and small number of data points, we were unable to find a unique solution for all equilibrium constants. However, the data could be well described by the 4-state model (Scheme 4) if we assumed values for pK_{HG^+} and pK_{WC^+} identical to those predicted by constant pH MD simulations (7.1 and 2.4, respectively). As a result, we obtained an estimate for K_1 and K_3, which corresponded to free energy difference of ΔG_1= 5.3 kcal/mol between the neutral HG and WC states, and ΔG_3 = -1.2 kcal/mol between the protonated HG^+ and WC^+ states, which is in line with our expectation that the neutral HG base pair is less stable than its protonated counterpart.

Quantitative analysis of the chemical exchange rate constants for the forward (k_A) and reverse (k_B) reaction was not performed here due to their complex definition under the 4-state model as "net" rate constants and potential complications from additional protonation equilbria, as discussed in the main text.

Molecular Dynamics (MD) Simulations

MD Simulation Setup. The input structure for the DNA duplex was obtained from Nikolova *et. al.*[1] Hydrogen atoms were added using the *HBUILD* facility in CHARMM.[4] The DNA structures were solvated in a cubic volume of explicit TIP3P[5] water molecules of length ~60 Å, with counterions (Na^+, Cl^-) added to simulate an ionic strength of 70 mM NaCl using the convpdb.pl tool from the MMTSB toolset.[6] The terminal ends of the DNA structures were hydroxylated using CHARMM's *5TER* and *3TER* keywords. Patches for the protonated forms cytosine was constructed as previously reported.[7,8] A patch for the N1-methylated guanine residue was also constructed for the simulation of the methylated DNA variant (see Table S3). MD simulations were performed within the CHARMM macromolecular modeling program (version c36a6) using the CHARMM36 all-atom force field for RNA[9] and TIP3P water.[5] Hydrogen-heavy atom bond lengths were constrained using the SHAKE algorithm,[10] and the Leapfrog Verlet integrator was used to propagate the system using an integration time step of 2 fs. A non-bonded cutoff of 12 Å was used with an electrostatic force shifting function and a van der Waals switching function between 10 Å and 12 Å.

Constant pH MD Simulations. The CPHMD simulations performed is an extended Hamiltonian approach, where the protonation state of the residue is described by a continuous variable, λ, which is propagated simultaneously with the spatial coordinates at a specified external pH via multi-site λ-dynamics (MSλD).[11] Additional patches were constructed to represent the protonated forms of adenine and cytosine. All of the associated bonds, angles and dihedrals were explicitly defined in the patch. Each titratable residue was simulated as a hybrid model that explicitly included atomic components of both the protonated and unprotonated forms. The titratable fragment included the nitrogen atom that is protonated, the protonated hydrogen and adjacent atoms whose partial charge differed according to the protonation state as reported previously by Goh *et. al.*[7,8] The environment atoms were defined as all atoms that were not included in the titratable fragments. λ dynamics was performed within the BLOCK facility using the MSλD framework (*MSLD*) and selecting the λ^{Nexp} functional form for λ (*FNEX*). Linear scaling by λ was applied to all energy terms except bond, angle and dihedral terms, which were treated at full strength regardless of λ value to retain physically reasonable geometries. Each θ_α was assigned a fictitious mass of 12 amu•$Å^2$ and λ values were saved every 10 steps. Variable biases (F_{var}) were added to the hybrid potential energy function to enhance transition rates between the two protonation states, and the associated force constant (k_{bias}) used were identical to the optimized values reported by Goh *et. al.*[7,8] The temperature was maintained at 298 K by coupling to a Langevin heatbath using a frictional coefficient of 10 ps^{-1}.

Only the cytosine in the G•C base pair was titrated in our simulations, with the other residues kept at their reference protonation state. $CPHMD^{MS\lambda D}$ simulations were augmented by a recently developed pH-based

replica exchange protocol, where attempts to exchange the pH of the simulation between pairs of adjacent replicas are attempted every 1 ps.[12] Prior to the simulation, each system was minimized using 300 steps of steepest descent (SD), followed by 200 steps of adopted basis Newton-Raphson (ABNR). After an initial heating of 200 ps and equilibration of 800ps), a production run 1 ns was performed.

Calculation of pK_a values. CPHMD$^{MS\lambda D}$ simulations were repeated for 3 independent runs that used different initial seeds. The populations of unprotonated (N^{unprot}) and protonated (N^{prot}) states from the simulation were used to derive the unprotonated fraction (S^{unprot}):

$$S^{unprot}(\mathrm{pH}) = \frac{N^{unprot}(\mathrm{pH})}{N^{unprot}(\mathrm{pH}) + N^{prot}(\mathrm{pH})} \tag{10}$$

The S^{unprot} values were computed at each pH condition simulated, which was 5.0 to 9.0 for the HG G•C base pair and 1.0 to 5.0 for the WC G•C base pair. The S^{unprot} values were then fitted to a generalized version of the Hendenson-Hasselbach equation to obtain a single pK_a value:

$$S^{unprot}(\mathrm{pH}) = \frac{1}{1+10^{-n(pH-pKa)}} \tag{11}$$

This procedure was repeated to obtain 3 sets of pK_a values, and the reported pK_a value and its error correspond to the mean and standard deviation respectively (Figure 3 in main text).

Table S1. Best-fit pK_a values for A_6-DNA^{1mG10}.

Residue (spins)	pK_a
Individual fits	
T9 (C6H6)	7.10 ± 0.10
G10 (C8H8)	6.65 ± 0.06*
C15 (C5H5)	7.15 ± 0.10
C15 (C6H6)	7.36 ± 0.18*
A16 (C8H8)	7.07 ± 0.12
A16 (C2H2)	7.13 ± 0.15
A17 (C2H2)	7.18 ± 0.10
Global fit	
All residues	7.24 ± 0.06

* profile for G10 C8H8 and C15 C6H6 were incomplete due to chemical exchange broadening at intermediate pH values

Table S2. Best-fit parameters obtained from global fits of $R_{1\rho}$ relaxation dispersion data, where the carbon C8 chemical shift difference between WC and HG base pairs ($\Delta\omega_{AB}$) was kept constant across the pH range.

pH	p_B (%)	k_{ex} (Hz)	$\Delta\omega_{AB}$ (ppm)	p_B (%)	k_{ex} (Hz)	$\Delta\omega_{AB}$ (ppm)
		G10 C8			**A16 C8**	
7.6				0.188 ± 0.017	3797 ± 500	3.37 ± 0.04
6.8	0.050 ± 0.036	604 ± 584	3.13 ± 0.01	0.214 ± 0.023	3786 ± 552	
6.0	0.166 ± 0.013	885 ± 119		0.254 ± 0.020	3723 ± 414	
5.4	0.679 ± 0.027	606 ± 33		0.301 ± 0.011	3192 ± 190	
5.2	1.081 ± 0.033	675 ± 30				
4.9	2.045 ± 0.036	804 ± 24		0.741 ± 0.026	3715 ± 185	
4.6	3.770 ± 0.041	1005 ± 19		1.178 ± 0.030	3953 ± 120	
4.3	8.382 ± 0.074*	1795 ± 29*	3.76 ± 0.02*	2.122 ± 0.049	3595 ± 100	
4.0				3.684 ± 0.079	4875 ± 114	

*data was collected for G10 C1' due to extensive line broadening at C8

Table S3. Parameters for N1-methylguanine (1mG) residue.

Atom Name	Atom Type	Partial Charge
C1A	CN8	-0.27
H1A1	HN8	0.09
H1A2	HN8	0.09
H1A3	HN8	0.09

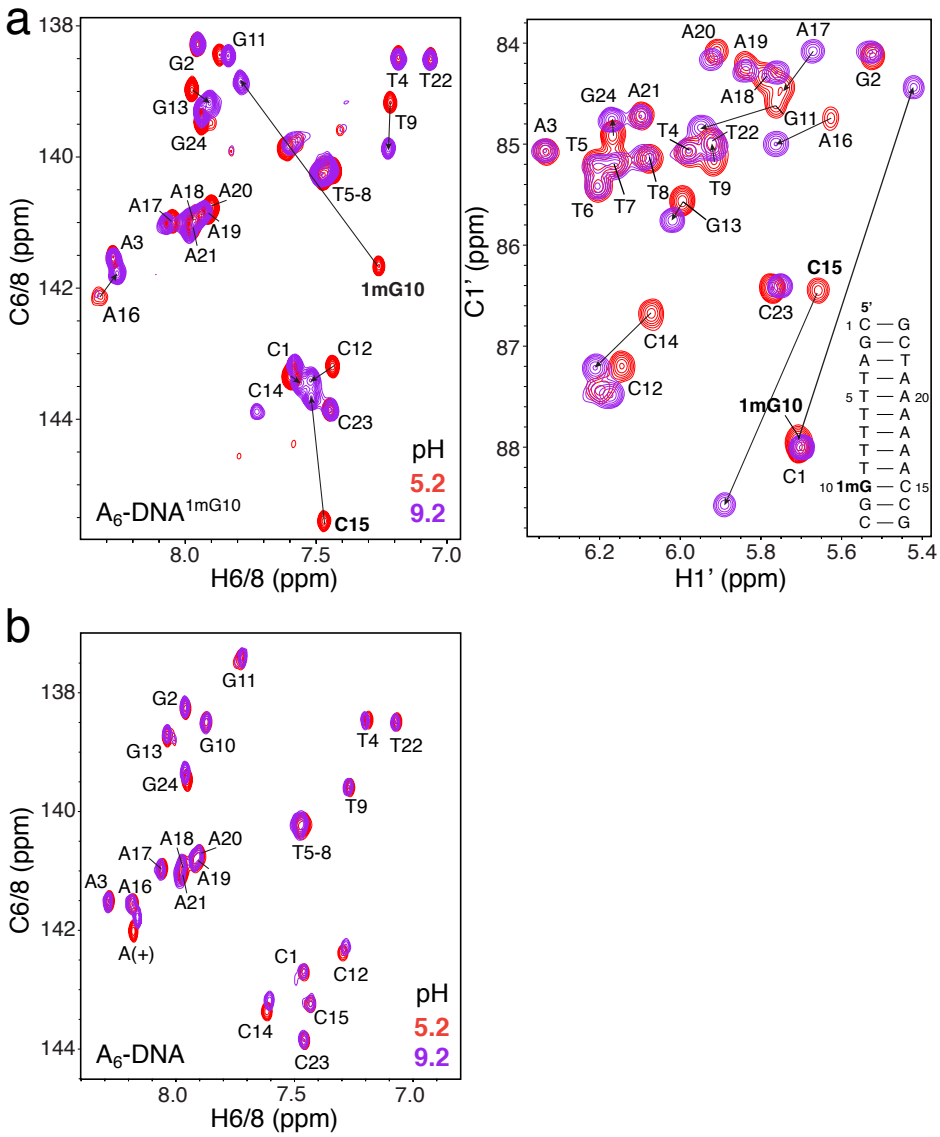


Figure S1. pH dependent conformational change at the HG trapped 1mG•C base pair in A_6-DNA^{1mG10}. (a) 2D $^{1}H,^{13}C$ HSQC overlays of base C6H6/C8H8 and sugar C1'H1' region of A_6-DNA^{1mG10} at pH 5.2 (red) and 9.2 (purple) showing large conformation changes at the 1mG10•C15 and the two neighboring base pairs (A16•T9 and G11•C14) that support transition from a protonated HG base pair at low pH to a WC-like base pair at high pH. (b) Control 2D $^{1}H,^{13}C$ HSQC overlays of unmodified A_6-DNA showing no significant change between pH 5.2 (red) and 9.2 (purple).

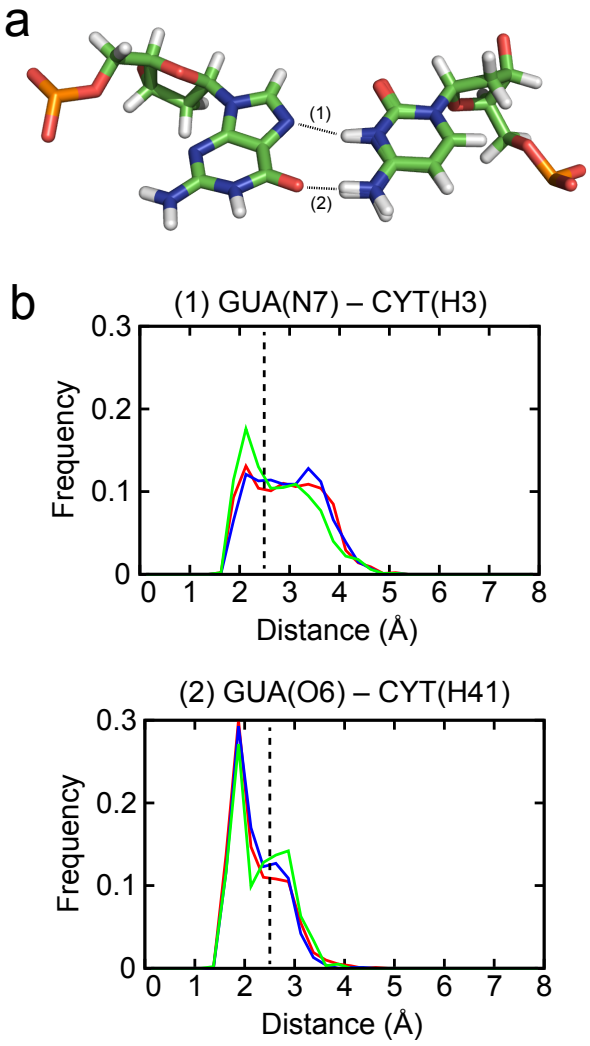


Figure S2. HG base pairs form stable hydrogen bonds during constant pH MD simulations. (a) Ideal geometry of a protonated HG G•C^+ base pair with 2 hydrogen bonds formed between (1) GUA(N7) – CYT(H3) and (2) GUA(O6) – CYT(H41). (b) Histogram of hydrogen bonding distances at pH 7 indicates the formation of a HG-like G•C^+ base pair, with a strong hydrogen bond between GUA(O6) and CYT(H41) and a moderate hydrogen bond between GUA(N7) and CYT(H3) due to incomplete protonation at pH 7.

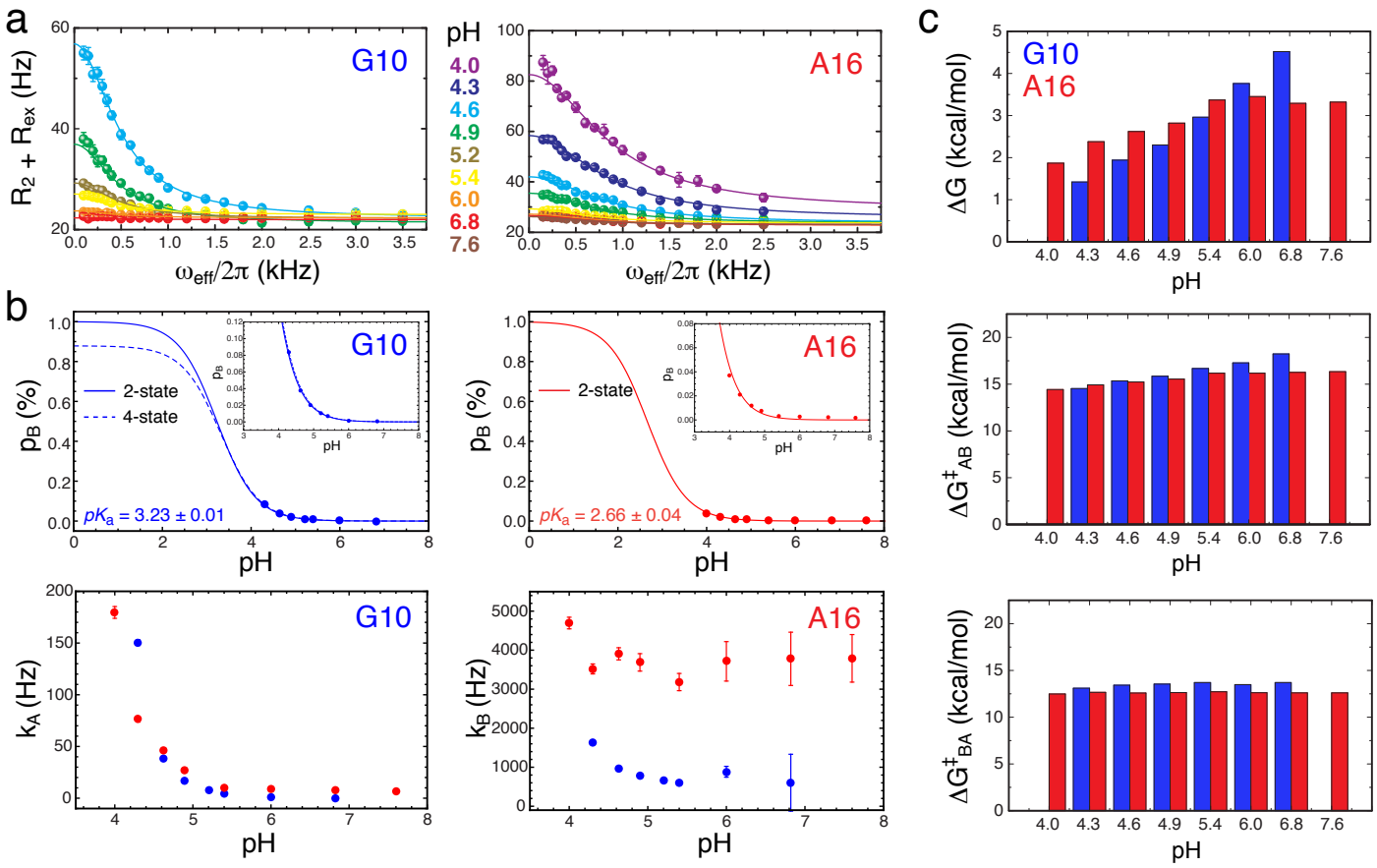


Figure S3. pH dependence of NMR ^{13}C relaxation dispersion data for a G•C and A•T base pairs in A_6-DNA. (a) Representative on-resonance relaxation dispersion profiles for G10 and A16 C8 site as a function of pH (at 26 ^{o}C), where the solid line represents the best fit of the combined on- and off-resonance data to a 2-state equilibrium (the profile for G10 at pH 4.3 was performed on C1' due to extensive line broadening at C8).[1] (b) pH dependence of the transient HG base pair population (p_B) and the forward (WC→HG$^+$, k_A = $p_B k_{ex}$) and reverse (HG$^+$→WC, k_B = (1 - p_B)k_{ex}) rate constants for G10 and A16 derived from the chemical exchange rate constant k_{ex}, showing the best fit to a 2-state equilibrium (Scheme S3, solid line) and a 4-state equilibrium (Scheme S4, dashed line) for the change in p_B. (c) Corresponding calculated free energy differences for the net reaction (ΔG) and forward ($\Delta G^{\ddagger}_{AB}$) and reverse ($\Delta G^{\ddagger}_{AB}$) transitions, showing a coupled decrease in ΔG and $\Delta G^{\ddagger}_{AB}$ with pH and little dependence of $\Delta G^{\ddagger}_{AB}$ on pH ($\Delta G^{\ddagger}$ values were calculated from the transition rate constants using Transition state theory).

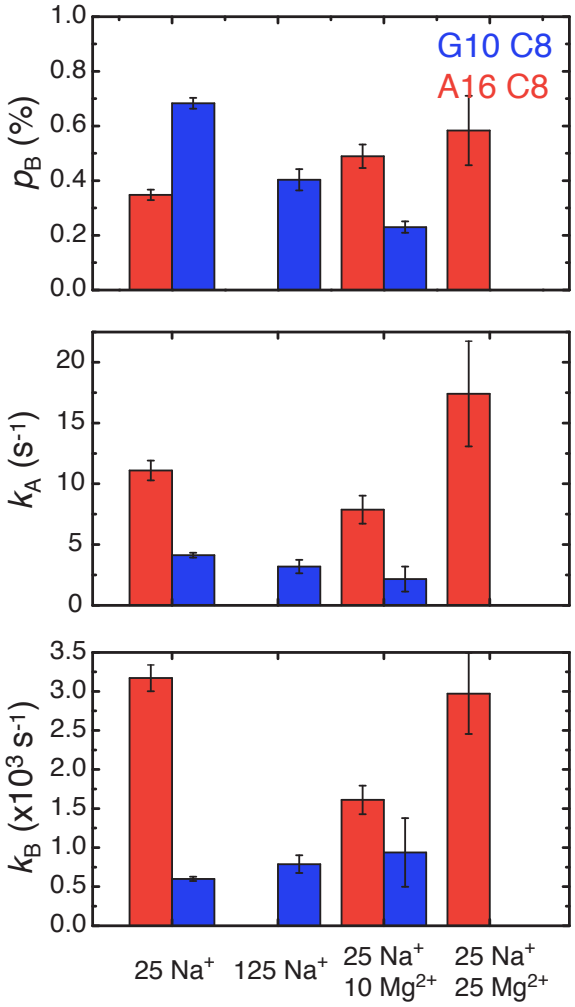


Figure S4. Salt dependence of chemical exchange at G10 and A16 C8 in A_6-DNA showing the variation in transient HG^+ base pair population (p_B), forward (k_A), and reverse (k_B) rate constants as a function of added Na^+ (NaCl) and Mg^{2+} ($MgCl_2$) concentration (mM) at pH 5.4 (A16 C8 measurements performed at pH 6.8/125 mM Na^+ also showed no change relative to 25 mM Na^+).

Supplementary References

(1) Nikolova, E. N.; Kim, E.; Wise, A. A.; O'Brien, P. J.; Andricioaei, I.; Al-Hashimi, H. M. *Nature* **2011**, *470*, 498.
(2) Nikolova, E. N.; Gottardo, F. L.; Al-Hashimi, H. M. *J. Am. Chem. Soc.* **2012**, *134*, 3667.
(3) Shi, C.; Wallace, J. A.; Shen, J. K. *Biophys. J.* **2012**, *102*, 1590.
(4) Brooks, B. R.; Brooks, C. L., III; Mackerell, A. D., Jr.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caflisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M. *J. Comput. Chem.* **2009**, *30*, 1545.
(5) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926.
(6) Feig, M.; Karanicolas, J.; Brooks, C. L., 3rd *J. Mol. Graph. Modell.* **2004**, *22*, 377.
(7) Goh, G. B.; Knight, J. L.; Brooks, C. L., III *J. Chem. Theory Comput.* **2012**, *8*, 36.
(8) Goh, G. B.; Knight, J. L.; Brooks, C. L., III *J. Chem. Theory Comput.* **2013**, *9*, 935.
(9) Denning, E. J.; Priyakumar, U. D.; Nilsson, L.; Mackerell, A. D., Jr. *J. Comput. Chem.* **2011**, *32*, 1929.
(10) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327.
(11) Knight, J. L.; Brooks, C. L., III *J. Comput. Chem.* **2011**, *32*, 3423.
(12) Goh, G. B.; Knight, J. L.; Brooks, C. L., III *J. Phys. Chem. Lett.* **2013**, *4*, 760.