Oxidized Derivatives of 5-Methylcytosine Alter the Stability and Dehybridization Dynamics of Duplex DNA

Paul J. Sanstead[†‡§], Brennan Ashwood[†‡§], Qing Dai[†‡], Chuan He[†‡‖⊥], Andrei Tokmakoff*[†‡§]

[†]Department of Chemistry, [‡]Institute for Biophysical Dynamics, [§]James Franck Institute, [‖]Department of Biochemistry and Molecular Biology, and [⊥]Howard Hughes Medical Institute, The University of Chicago, Chicago, Illinois 60637, United States

Abstract

The naturally occurring nucleobase 5-methylcytosine (mC) and its oxidized derivatives 5-hydroxymethylcytosine (hmC), 5-formylcytosine (fC), and 5-carboxylcytosine (caC) play important roles in epigenetic regulation and, along with cytosine (C), represent nucleobases currently implicated in the active cytosine demethylation pathway. Despite considerable interest in these modified bases, their impact on the thermodynamic stability of double-stranded DNA (dsDNA) remains ambiguous and their influence on hybridization kinetics and dynamics is even less well-understood. To address these unknowns, we employ steady-state and time-resolved infrared spectroscopy to measure the influence of cytosine modification on the thermodynamics and kinetics of hybridization by assessing the impact on local base pairing dynamics, shifts in the stability of the duplex state, and changes to the hybridization transition state. Modification with mC leads to more tightly bound base pairing below the melting transition and stabilizes the duplex relative to canonical DNA, but the free energy barrier to dehybridization at physiological temperature is nevertheless reduced slightly. Both hmC and fC lead to an increase in local base pair fluctuations, a reduction in the cooperativity of duplex melting, and a lowering of the dissociation barrier, but these effects are most pronounced when the 5-position is formylated. The caC nucleobase demonstrates little impact on dsDNA under neutral conditions, but we find that this modification can dynamically switch between C-like and fC-like behavior depending on the protonation state of the 5-position carboxyl group. Our results provide a consistent thermodynamic and kinetic framework with which to describe the modulation of the physical properties of double-stranded DNA containing these modified nucleobases.

Introduction

The methylation of cytosine (C) at the 5-position to produce 5-methylcytosine (mC) is an essential epigenetic modification in eukaryotes that is associated with transcriptional silencing.[1-2] In mammals this modification occurs mainly at symmetric CpG sites, where it is found in abundance.[3] Although the mechanism and function of DNA methylation have been known for some time,[4-5] the details of the reverse process in which mC is restored to C have come to light only within the last decade. In the active cytosine demethylation pathway mC is sequentially oxidized to 5-hydroxymethylcytosine (hmC), 5-formylcytosine (fC), and finally 5-carboxylcytosine (caC) by the ten-eleven translocation (TET) family of enzymes.[6-7] The bases fC and caC are selectively excised by thymine DNA glycosylase (TDG) and the resulting abasic site is repaired by base excision repair (BER) enzymes, thus closing the demethylation cycle. In contrast to mC, it is currently unknown whether the oxidized derivatives of mC perform unique regulatory functions independent of their role as demethylation intermediates. Given that protein binding partners for these nucleobases have been indentified,[8-10] that at least hmC and fC are known to be genomically stable,[11-12] and that the abundance of these derivatives depends on tissue, developmental stage, and disease,[13-17] it seems likely that some or all of the oxidized derivatives of mC do serve an additional epigenetic function.

Understanding the impact of mC, hmC, fC, and caC on the fundamental properties of CpG sites as well as double-stranded DNA (dsDNA) is a critical step in revealing the molecular recognition and biological utility of modified cytosine derivatives both within the context of active demethylation and as of yet undiscovered epigenetic contexts. To this end, considerable progress has been made characterizing the structural, dynamic, and physical influence of these modifications on dsDNA.[18] For instance, it has been established that cytosine derivatives do not deviate the solution structure of dsDNA significantly away from canonical B-form.[19-22] Evidence suggests that fC and hmC increase while mC decreases the flexibility of dsDNA, and molecular dynamics simulations suggest this effect is accompanied by a corresponding increase or decrease in structural fluctuations about an average B-form structure.[23-26] There is no indication that cytosine modifications shift the tautomeric equilibrium relative to C, as previously suggested to account for the selective activity of TDG,[27] and the keto-amino form predominates in all cases.[28-29] Both fC and caC appear to alter the nucleobase electronic structure as measured through N1 acidity and leaving group ability.[28] It has been proposed that this effect could explain the selective

excision of these bases by TDG as well as the strange features relative to canonical B-DNA observed in the ultraviolet circular dichroism (UV CD) spectrum of oligonucleotides containing these modifications.[19]

Despite this progress, the present knowledge of the impact of cytosine modification on dsDNA remains incomplete and at times contradictory. Most notably, these nucleobases appear to have only a modest effect on duplex stability and no consensus has been reached on the thermodynamic trend with modification. Except for widely reported stabilization due to methylation, different researchers alternately report the same nucleobase to be stabilizing or destabilizing, in some cases even for the same sequence under similar conditions.[21, 29-33] Beyond the ambiguity surrounding trends in thermodynamic stability, understanding the influence of cytosine modifications on the dynamics of base pair opening and the energetic barrier to DNA dissociation is equally if not more important to understanding their biological function, but to our knowledge there is currently a lack of kinetic studies on the impact of these modifications on hybridization.

To date no complete description that can simultaneously account for the influence of the cytosine 5-position substituent on local base pairing dynamics, the thermodynamics of the duplex to single-strand transition, and the kinetics of DNA hybridization has been proposed. Here we seek to address this deficiency using a combination of steady-state and time-resolved temperature jump (T-jump) infrared (IR) spectroscopy. Utilizing the sensitivity of vibrational spectroscopy to pre-melting changes in nucleic acid structure,[34-37] we demonstrate that the standard "all-or-none" description of oligonucleotide dehybridization, in which all base pairs in the duplex are assumed to be fully intact, breaks down to varying degrees for oligonucleotides containing cytosine modifications. We apply an alternative model in which the possibility of a pre-dissociation reduction in base pairing contacts is considered explicitly. This description resolves the ambiguity surrounding the thermodynamic role of cytosine modification and reveals the relationship between increased base pair heterogeneity, duplex stability, and the cooperativity of melting. Temperature jump experiments, including transient two-dimensional infrared (t-2D IR) spectroscopy, characterize the modification-dependent impact on base pair disorder and dehybridization kinetics in oligonucleotide duplexes. A unified description of the thermodynamics and kinetics of hybridization provides a consistent framework with which to discuss the effect of modification on local base pair dynamics, the stability of dsDNA, and the free energy barrier to opening the XpG

step when X is C, mC, hmC, fC, or caC. This insight into the modulation of the properties of DNA by epigenetic cytosine derivatives is an essential step towards a full understanding of the molecular recognition and function of these naturally occurring nucleobases.

Results

Temperature Dependent FTIR and Melting Curves

As an initial assessment of the influence of modified cytosines on dsDNA, temperature-dependent Fourier transform infrared (FTIR) melting experiments on the sequence 5′-TAXGXGXGTA-3′ where X = C, mC, hmC, fC, or caC were conducted between 3-98 °C. The selection of this oligonucleotide is motivated by the many previous studies that have employed this model sequence or closely related sequences with a central triple XpG repeat.[19, 29-30, 33] The 1550-1715 cm^{-1} frequency range of the FTIR spectrum contains in-plane nucleobase ring vibrations and carbonyl stretches that are sensitive to the base stacking and hydrogen bonding interactions that mediate base pairing.[38] As a result the spectrum in this frequency range is reshaped significantly upon DNA hybridization.[39-41]

Figure 1a-e shows the FTIR temperature series for each of the oligonucleotides sampled in ~4.5 °C steps. Inspection of the low temperature spectra plotted in blue reveals that the DNA duplex spectrum is dependent on the presence of modified cytosine bases, but the general peak pattern and intensities observed for the canonical sequence are more-or-less conserved. The fC and caC nucleobases each have additional carbonyl absorptions in this frequency range, but these result in only minor changes to the spectrum, such as a low intensity shoulder at 1675 cm^{-1} when X = fC and additional absorption at 1570 cm^{-1} when X = caC, since the exocyclic carbonyl stretches are strongly mixed with the in-plane nucleobase vibrations.[29] Below 1600 cm^{-1} the spectrum is dominated by peaks corresponding to G ring vibrations whose intensity is suppressed upon the formation of stacked and hydrogen bonded Watson-Crick base pairs.

The frequency range from 1600-1675 cm^{-1} is congested with overlapping absorptions from all four nucleobases, but the intensity growth near 1620 cm^{-1} with increasing temperature is dominated by the loss of T:A base pairing at the termini of the duplex. The feature growing in near 1665 cm^{-1} reflects increasingly unpaired T, X, and G nucleotides as the duplex dehybridizes. This increase in absorption inversely tracks the loss in intensity of the peak centered near 1685 cm^{-1}, which corresponds primarily to a G carbonyl absorption that is shifted 20 cm^{-1} from 1665 cm^{-1} in

the single-strand to 1685 cm^{-1} in dsDNA.[39] The spectra for X = C, mC, hmC and caC reveal a clean isosbestic point between these peaks near 1675 cm^{-1}, which is often considered a sign of two-state behavior. Intensity loss near 1685 cm^{-1} can also reflect the disruption of T:A base pairs since the highest frequency T carbonyl absorption increases when T is engaged in a Watson-Crick pair. Melting curves that reflect the global changes to the IR spectrum in this frequency range were obtained through singular value decomposition (SVD).[36] Fig. 1f shows the normalized melting curve measured for each of the oligonucleotides.

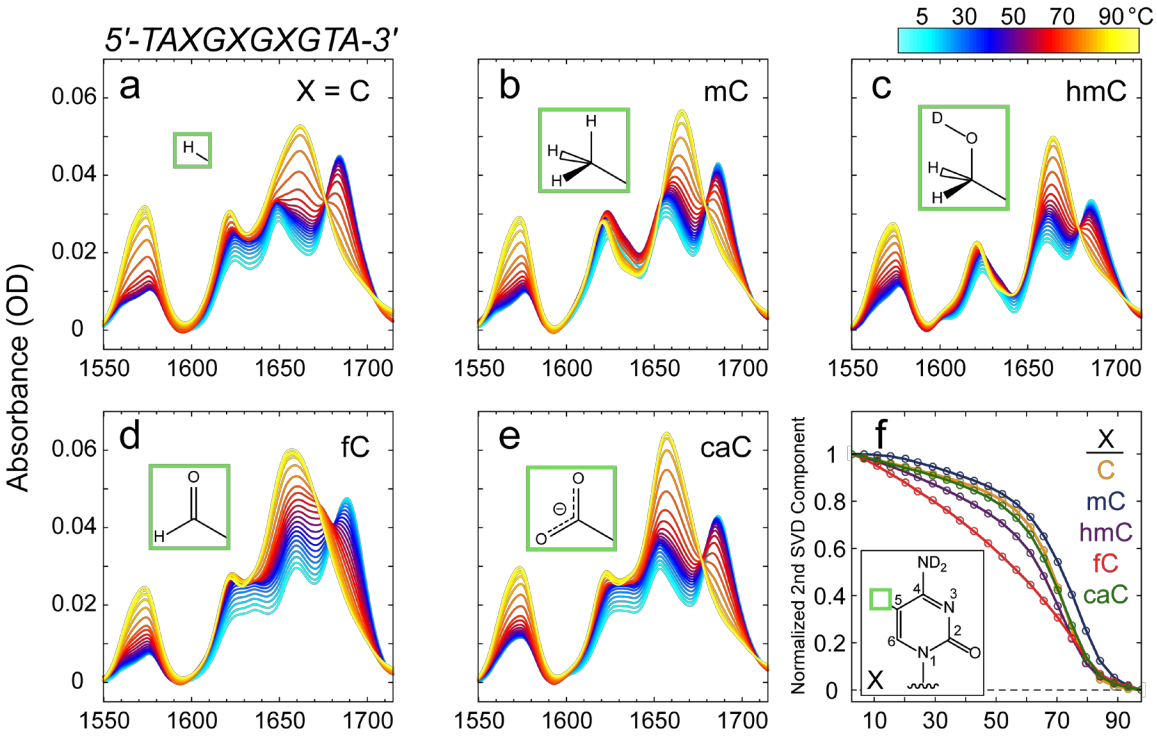


Figure 1: (a-e) FTIR temperature series from 3-98 °C for 5′-TAXGXGXGTA-3′ where X = C, mC, hmC, fC, or caC as indicated in each panel. Sample conditions were 1 mM oligonucleotide in 20 mM sodium phosphate buffer (pD 7.2) plus 16 mM NaCl. (f) The normalized second SVD temperature component corresponding to each of the set of spectra in panels a-e. The inset shows the structure of cytosine with the 5-position highlighted by a green box. The substituent associated with each modification is indicated in the appropriate panel. Labile protons are drawn H-D exchanged to reflect the experimental conditions.

In contrast to the most common method for measuring oligonucleotide melting curves in which the UV hyperchromicity near 260 nm is assumed to track the fraction of duplexed DNA, the IR spectrum relays additional structural insight relative to the comparatively broad and featureless UV spectrum.[29] It has long been known that vibrational spectroscopy is sensitive to pre-melting changes in duplex structure,[34-35] and we have used IR spectroscopy to characterize sequence-dependent disruption of base pairing contacts that precede duplex dissociation in canonical DNA oligonucleotides.[36-37] The origin of the modification-dependent deviations in the shapes of the melting curves in Fig. 1f can be assessed through the FTIR spectra measured along the low temperature baseline. Features in the spectrum indicative of a reduction in the extent of Watson-Crick base pairing, such as increased ring mode intensity and a loss in absorption at 1685 cm^{-1} correlated with a gain at 1665 cm^{-1}, are apparent in the spectrum measured along the baseline for some of the sequences and the extent to which these changes are observed at low temperature is proportional to the degree of melting curve asymmetry.

These observations provide direct evidence that base pairing within the duplex is disrupted even at low temperatures in a modification-dependent manner. A pre-dissociation reduction in the extent of base pairing in the duplex demonstrates that the standard all-or-none model of oligonucleotide melting cannot accurately describe the dehybridization of these sequences. Such considerations provide insight into the lack of consensus surrounding the impact of cytosine modification on duplex stability when assessed through routine melting temperature (T_m) analysis. To demonstrate the shortcomings of the all-or-none model when applied to modified XpG DNA, such an analysis and further discussion are included in the SI as well as additional details of the spectroscopic assignments along the low temperature baseline.

Modification-Dependent Dehybridization Kinetics

To evaluate the pre-melting effects suggested by FTIR experiments and to characterize equally important biophysical properties potentially shaped by cytosine modification, such as local base pairing structural dynamics, the barrier to duplex dissociation, and the nature of the dehybridization transition state, we characterized the impact of X on the dehybridization kinetics of the model oligonucleotide sequences using transient T-jump nonlinear IR spectroscopy. The instrument and its application to the study of DNA oligonucleotide dehybridization have been described in detail previously.[36-37, 42] In brief, a near-IR laser pulse induces a 15 °C temperature

rise within 5 nanoseconds and an electronically synchronized two-dimensional infrared (2D IR) spectrometer probes the response of the oligonucleotide ensemble following a delay τ that is varied from 5 ns to 50 ms after the T-jump.

Figure 2 presents the design of the T-jump experiments and the resulting modification-dependent dehybridization kinetics. The most highly contrasting X = C and fC sequences are shown as illustrative examples, but the discussion is applicable to all of the modified oligonucleotides studied. Fig. 2a,d illustrate the series of T-jumps selected to span the entire melting transition for the X = C and fC sequences as well as the corresponding change in total base pairing contacts between the initial (T_i) and final (T_f) temperature equilibria ($\Delta\theta$). The signal changes in T-jump experiments (ΔS) are expected to be proportional to $\Delta\theta$ if the system is allowed sufficient time to equilibrate at T_f. Therefore the largest ΔS is expected near the melting inflection point.

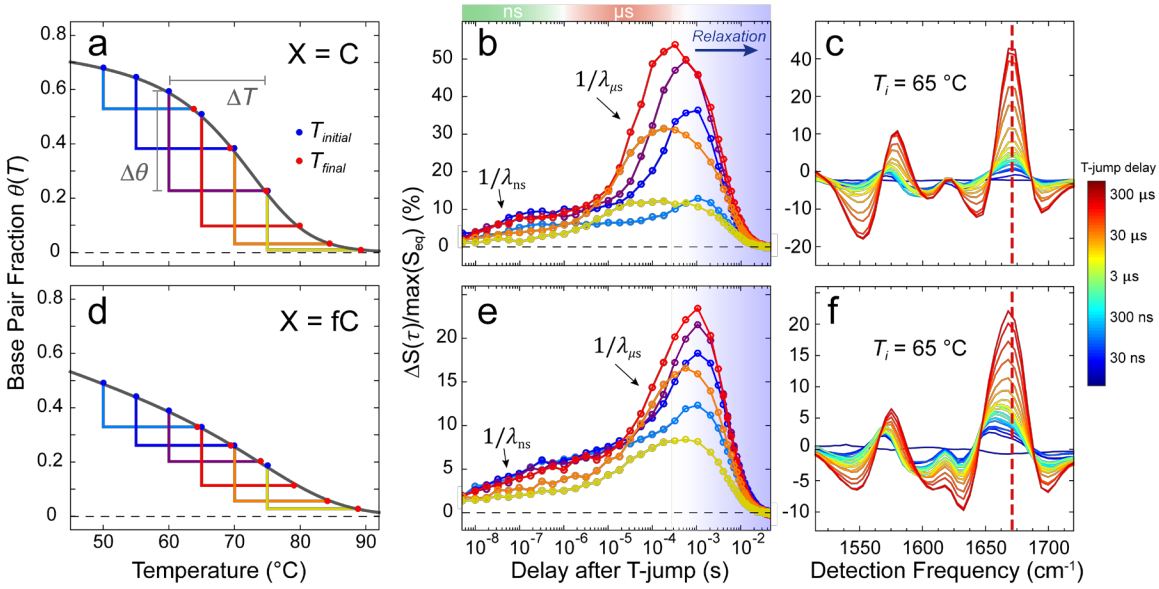


Figure 2: (a) Temperature ranges for the T-jump experiments measured and the corresponding changes in base pairing contacts, $\Delta\theta$ illustrated with color coded bars on the X = C melting curve. (b) Kinetic traces corresponding to each T-jump range indicted in panel (a) tracked at the maximum response of the t-HDVE spectrum at 1673 cm^{-1}. (c) The t-HDVE spectra measured between τ = -5 ns to 320 μs for the X = C sequence where T_i = 65 °C and T_f = 80 °C. The T-jump delay, τ, corresponds to the time between the arrival of the T-jump pulse and the probing mid-IR pulse sequence. (d-f) Corresponding quantities for the X = fC sequence.

Figures 2b,e present the T-jump kinetic traces measured across seven orders of magnitude in time for the temperature ranges indicated in Fig. 2a,d. Traces correspond to the transient heterodyne-detected dispersed vibrational echo (t-HDVE)[43] intensity tracked at 1673 cm^{-1}, the frequency with the largest amplitude signal change. Temperature jump data are presented as the fractional change observed after a delay τ relative to the maximum of the equilibrium spectrum at T_i, $\Delta S(\tau)/\max(S_{eq})$. The time-resolved t-HDVE spectra from which the kinetic traces are derived are illustrated in Fig. 2c,f for the X = C and fC sequences from τ = -5 to 320 μs for the case T_i = 65 °C.

Inspection of the kinetic traces measured for the X = C sequence in Fig. 2b reveals three distinct processes. The signal initially rises in 10-100's of nanoseconds (with an observed rate constant λ_{ns}), followed by a larger amplitude 10-100's of microsecond rise ($\lambda_{\mu s}$), after which the signal relaxes to equilibrium at T_i on a several millisecond timescale. These timescales are indicated respectively by green, red, and blue shading at the top of Fig. 2b. As the initial temperature is increased, the rate of the μs process speeds up noticeably, as expected for an activated process, and the amplitude rises through a maximum as T_i is stepped across the inflection point of the melting curve. In contrast, the amplitude associated with λ_{ns} is most pronounced relative to the total response at low temperature and diminishes as T_i is increased. These observations are consistent with previous dehybridization measurements on canonical oligonucleotides in which $\lambda_{\mu s}$ arose from duplex dissociation, and λ_{ns} was associated with a pre-dissociation reduction in base pairing, such as fraying of the helical termini.[36-37]

The corresponding set of kinetic traces for the X = fC sequence are shown in Fig. 2e. In this case the amplitude associated with $\lambda_{\mu s}$ follows a more gradual intensity trend with increasing T_i, consistent with the shallower profile of the X = fC melting curve. Furthermore, the timescale of the μs response does not increase as markedly with increasing temperature relative to the canonical sequence.

To accurately assign the structural changes occurring within the oligonucleotide ensemble in response to the T-jump, we rely on the sensitivity of the IR spectrum to changes in DNA base pairing contacts discussed above. The full transient 2D IR surface (t-2D IR) measured at a subset of strategically selected T-jump delays reveals cross peaks and 2D line shapes that provide more detailed insight than the t-HDVE spectra. For example, Fig. 3d,e shows the t-2D IR spectra collected at τ = 180 ns and 320 μs for the X = C sequence with T_i = 55 °C and ΔT = 15 °C. These

delays were selected as representative of the λ_{ns} and $\lambda_{\mu s}$ processes observed in the kinetic traces in Fig. 2b,e. Inspection of the t-2D IR spectrum collected at each delay reveals a distinct response associated with each process.

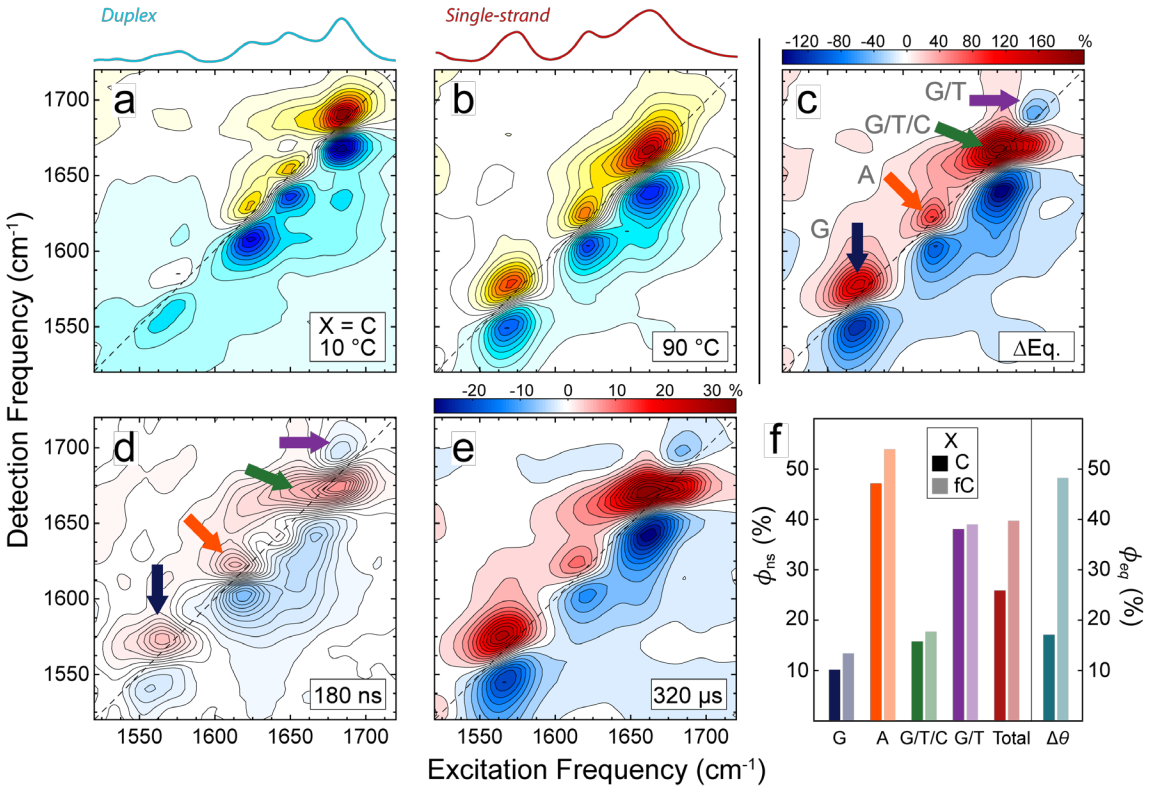


Figure 3: Equilibrium 2D IR spectra for the X = C sequence collected at (a) 10 °C where the oligonucleotide ensemble is highly duplexed and (b) 90 °C where single-stranded DNA predominates. The corresponding linear spectrum is plotted above each 2D IR surface for reference. (c) The equilibrium difference spectrum between the 90 °C and 10 °C surfaces. Arrows highlighting features of interest are labelled with the nucleobases which most contribute to each difference feature. (d) The τ = 180 ns t-2D IR surface collected for the X = C sequence. (e) The τ = 320 μs t-2D IR surface collected for the X = C sequence, exhibiting agreement with the equilibrium difference spectrum in panel (c). The waiting time was fixed at τ_2 = 150 fs and the polarization was set to ZZZZ for all measurements. For the t-2D IR measurements, T_i = 55 °C and ΔT = 15 °C. (f) Percentage of signal change present by 180 ns relative to 320 μs (ϕ_{ns}) for each feature indicated in panel (c) as well as the total signal change across the entire t-2D IR surface. The total change is compared against the change in the internal base pairing fraction, $\Delta\theta_{int}$ relative to the overall change in the melting curve, $\Delta\theta$ over the same temperature range (ϕ_{eq}).

Figure 3c shows the equilibrium difference spectrum between the 90 °C (Fig. 3a) and 10 °C (Fig. 3b) 2D IR surfaces measured for the X = C sequence, illustrating the spectral changes expected in response to duplex dissociation. Equilibrium difference and t-2D IR spectra are plotted with a blue-white-red color gradient to distinguish these data from the 2D IR spectra in Fig. 3a,b. Positive features are red while negative features are blue. Since 2D IR resonances appear as a vertically displaced doublet along the diagonal axis, a gain on the t-2D IR surface corresponds to a red-over-blue doublet and a loss appears as blue-over-red.

The equilibrium difference spectrum in Fig. 3c is dominated by intensity gain along the diagonal, such as the increase near 1575 cm^{-1} (highlighted by the blue arrow) corresponding to G ring mode absorption and the increase near 1620 cm^{-1} (orange arrow) corresponding primarily to A ring mode absorption. The gain feature near 1665 cm^{-1} marked by the green arrow and the loss feature near 1685 cm^{-1} indicated by the purple arrow correspond to the 20 cm^{-1} shift of the G carbonyl absorption upon loss of C:G base pairing discussed above. The reduction in intensity of the highest frequency T carbonyl absorption upon the loss of T:A base pairing also contributes to the loss observed near 1685 cm^{-1}.

The equilibrium difference spectrum in Fig. 3c and the 320 μs t-2D IR spectrum in Fig. 3e are in good agreement, indicating that the $\lambda_{\mu s}$ process corresponds to the duplex to single-strand dissociation process. This assignment of the μs timescale is consistent across the set of modified oligonucleotides and is in agreement with many past reports of DNA dehybridization.[37, 44-45]

To assign the processes occurring prior to duplex dissociation (λ_{ns}), we turn to the 180 ns t-2D IR surface for the X = C sequence in Fig. 3d. The four prominent features identified in the equilibrium difference and 320 μs t-2D IR spectra in Fig. 3c,e are apparent at 180 ns as well and are consistent with changes in T:A and C:G base pairing, but the overall variation in signal intensity is reduced and the difference in line shapes relative to the equilibrium spectrum is distinct.

To quantify the degree to which base pairing is disrupted prior to duplex dissociation, we calculate the percentage of the total signal change that is already present by 180 ns for each base-specific resonance: $\phi_{ns} = \Delta S_{ns}/\Delta S_{us}$, where ΔS_{ns} and ΔS_{us} reflect the change in T-jump signal amplitude at 180 ns and 320 μs, respectively. This ratio is shown in Fig. 3f for each of the four prominent difference features along the diagonal as well as the integrated intensity across the entire transient spectrum for both the X = C and fC sequences. The details of this analysis as well as further discussion of the t-2D IR surfaces measured at each delay are included in the SI.

We observe that ~50% of the growth in the A ring mode absorption (orange arrow) is already present by 180 ns, indicating that much of the T:A base pairing at the termini of the duplex is disrupted before dissociation. In contrast only ~10% of the growth of the G ring mode absorptions (blue arrow) is observed on this timescale, suggesting that the response from the central X:G base pairs is suppressed at early times. At higher frequency, overlapping contributions from multiple nucleobase absorptions complicate assignment, but the correspondence of ϕ_{ns} at 1685 cm^{-1} (purple) to the large ϕ_{ns} observed for the A ring mode suggest that both largely correspond to shifts in T:A base pairing. Similarly, the growth at 1665 cm^{-1} follows the behavior observed at the G ring modes and primarily reflects a disruption in X:G base pairing. The ns timescales characteristic of these changes suggest rapid interconversion between the available base pairing conformations at thermal equilibrium, consistent with enhanced base pair mobility and increased structural fluctuations within the duplex. Regardless of modification, the T:A termini of the double helix show the largest early-time response, consistent with past reports of duplex fraying.[36-37, 46]

To quantify the total overall signal change we determine the ratio of the integrated intensity across the entire t-2D IR surface measured at 180 ns versus 320 μs. This quantity can be compared against the percent change in pre-dissociation base pairing relative to the global change in the melting curve over the same temperature range as the T-jump: $\phi_{eq} = \Delta\theta_{int}/\Delta\theta$. The determination of θ_{int} and θ from the melting curve is discussed in the next section. Across the entire t-2D IR surface, a greater signal change occurs within 180 ns for the X = fC sequence compared to the X = C sequence, indicating an enhanced ns response that reflects increased base pair fluctuations when X = fC. This trend is also observed in the greater contribution from disrupted base pairing contacts within the X = fC duplex as quantified by ϕ_{eq} plotted at right in Fig. 3f.

Self-Consistent Modeling of Modification-Dependent Thermodynamics and Kinetics

Having measured the thermodynamic and kinetic impact of cytosine modification through a combination of FTIR and T-jump nonlinear IR spectroscopy, we propose a unified description that can account for all of the experimental observations and provide a consistent framework for discussing the results. An interpretation of the melting curves that allows for the possibility of a pre-dissociation reduction in base pairing is clearly required. A statistical thermodynamic

description in which the many possible base pair conformations involved in dehybridization are considered explicitly would be the most complete. In this context the IR melting curve is assumed to reflect the fraction of intact base pairs in the sample $\theta(T)$ such that the melting curve is proportional to the product of an internal (θ_{int}) and external (θ_{ext}) base pair fraction, where θ_{int} is the ensemble average fraction of intact base pairs among dsDNA and θ_{ext} is the fraction of duplexed DNA strands among all DNA strands.[47] For example, Fig. S1b in the SI shows the melting curve for the canonical sequence simulated using such an approach that employs nearest-neighbor (NN) parameters[48] and simple polymer physics to model the oligonucleotide dimer ensemble.[49] Application of this model to the canonical sequence provides further insight into the origin of the shape of the experimental melting curve, but existing statistical models parameterized for the four canonical DNA nucleobases cannot be applied directly to modified sequences, especially since the thermodynamic impact of the cytosine derivatives remains poorly understood. We therefore propose a simplified description in which the duplex [D] to single-strand [S] transition is modeled as a two-state process as in the all-or-none model, but within the duplex state the opening and closing of base pairs is allowed. Fig. 4 presents a diagram of this reaction scheme.

A detailed description of the model is included in the SI. In brief, the product $\theta_{int}\theta_{ext}$ is assumed to describe the melting curve and the temperature dependence of the internal and external base pair fractions is given by an internal and external equilibrium constant,

$$K_i(T) = \exp\left[-\Delta G_i^\circ / RT\right] = \exp\left[-\Delta H_i^\circ / RT\right]\exp\left[\Delta S_i^\circ / R\right] \qquad (i = int,ext) \qquad (1)$$

where R is the ideal gas constant and the standard enthalpy ΔH_i° and entropy ΔS_i° are assumed to be temperature independent. Here we assume that a single value of ΔH_{int}° and ΔS_{int}° can be used to describe base pair opening for each base pair, such that the internal thermodynamics of the oligonucleotide are treated in an average way across the duplex.

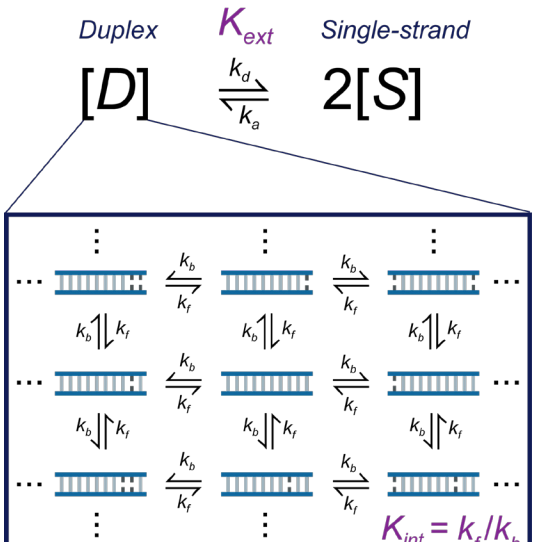


Figure 4: Schematic of the model applied to describe the modification-dependent effects on the thermodynamics and kinetics of hybridization. [*D*] and [*S*] are the duplex and single-strand concentrations determined by the external equilibrium constant, K_{ext}, which can be expressed as the ratio of the rate constants for duplex dissociation and association, k_d/k_a. Within the duplex state, the possibility of disrupted base pairing contacts is modeled by an internal equilibrium constant, K_{int}, that is related to the elementary step of base pair formation/breaking with rate constants k_f and k_b.

The thermodynamic model can be extended to incorporate the measured dehybridization kinetics by expressing the duplex to single-strand equilibrium constant as the ratio of the dissociation and association rate constants, $K_{ext} = k_d/k_a$. Since the μs process measured in the T-jump experiments corresponds to the activated duplex to single-strand transition, the observed rate $\lambda_{\mu s}$ can be related to k_d and k_a by,

$$\lambda_{\mu s} = k_d + 4[S_{eq}]k_a \tag{2}$$

where $[S_{eq}]$ is the single-strand concentration at the T_f equilibrium following the T-jump.[50] The temperature dependence of the rate constants is assumed to follow Kramers equation in the high-friction limit since we seek to describe the hybridization of DNA strands in solution,[51-52]

$$k_j(T) = \frac{C_j^\circ \lambda_{ns}}{\eta(T)} \exp\left[\Delta S_j^\ddagger / R\right] \exp\left[-\Delta H_j^\ddagger / RT\right] \qquad (j = d,a) \qquad (3)$$

where $\eta(T)$ is the known viscosity of the D_2O solvent[53] and a free energy of activation $\Delta G_j^\ddagger = \Delta H_j^\ddagger - T\Delta S_j^\ddagger$ describing duplex dissociation and association ($j = d,a$) is introduced such that $\Delta G_{ext}^\circ = \Delta G_d^\ddagger - \Delta G_a^\ddagger$. The C_j° in the prefactor is a reference parameter set to 1 that carries the units of Pa·s for $j = d$ and Pa·s/M for $j = a$. The observed rate of the ns process is included in the prefactor because our t-2D IR assignment of λ_{ns} above as well as previous characterization of similar rapid shifts in nucleobase contacts within canonical DNA duplexes[37] suggests that the ns response originates from a barrierless reduction in the extent of base pairing within the duplex ensemble. Since these rapid shifts in base pairing are in response to a reshaping free energy surface following the T-jump, λ_{ns} is sensitive to the shape of the duplex well and offers a reasonable estimate for the rate of hybridization in the limit that the barrier height approaches zero.[54-55]

Modification-Dependent Trends in Hybridization Thermodynamics and Kinetics

Applying this model to the series of modified oligonucleotides, Fig. 5a shows fits to the experimentally measured melting curve for each sequence while Fig 5b shows the θ_{int} and θ_{ext} determined from the fit plotted independently. Note that the standard all-or-none assumption that the melting curve ranges between 0 and 1 is no longer applicable since the melting curve in this case does not strictly equal the fraction of duplexed DNA. The offset relative to 1 along the vertical axis for each melting curve at the lowest temperature point is determined from the fit as described in the SI and reflects the value of θ_{int} at 3 °C. Comparing the θ_{ext} in Fig. 5b, the curves for the X = C, hmC, and caC sequences essentially overlay suggesting that the X = hmC and caC modifications do not significantly perturb the overall duplex to single-strand equilibrium relative to the canonical sequence. In contrast the X = mC and fC external fractions are shifted to higher temperature relative to the canonical sequence by several degrees, indicating that these modifications shift K_{ext} in favor of duplex formation. These trends in θ_{ext} are in agreement with the T_m trends suggested by the baseline corrected melting curves in Fig. S1a in the SI, supporting our assignment of asymmetry as arising mainly from θ_{int} and suggesting that the all-or-none treatment of the melting curves is primarily sensitive to the fraction of duplexed DNA, as one would expect.

Comparing θ_{int} for each sequence plotted in dashed lines in Fig. 5b, it is clear that considering trends in θ_{ext} in isolation will lead to erroneous conclusions about the influence of cytosine modification on base pairing within the XpG step. Those sequences with the most asymmetric melting curves correspond to the most sharply descending θ_{int} with increasing temperature. Most notably, the X = fC and hmC sequences exhibit sharply sloping θ_{int} relative to the canonical sequence, suggesting that base pairing contacts within the duplex are measurably disrupted at temperatures well below the inflection point of θ_{ext} where half of all DNA strands are duplexed. This effect is most pronounced when X = fC. When X = caC, θ_{int} is also reduced relative to the X = C sequence, but to a far less significant degree. The internal fraction for the X = mC sequence starts above the canonical sequence at low temperature, but descends more sharply with increasing temperature, dropping below the X = C sequence above 80 °C.

Considering the thermodynamic parameters determined from the model, Fig. 6a shows the trend in the external enthalpy, ΔH°_{ext} as a function of cytosine modification. This quantity reports on the standard enthalpy change associated with the duplex to single-strand transition and is therefore related to the van't Hoff enthalpy determined from the slope of the melting transition at T_m when applying the two-state all-or-none model.[56] The magnitude of ΔH°_{ext} is thus a reflection of the cooperativity of duplex melting, with larger values corresponding to more cooperative dehybridization in which increasing numbers of base pair contacts are lost in concert over a narrower temperature range. The canonical sequence is associated with the largest ΔH°_{ext} and shows the most sharply transitioning θ_{ext}, suggesting that the unmodified CpG domain opens more cooperatively than any of the modified XpG domains. The X = mC sequence shows a relatively modest ~15 kJ/mol drop in ΔH°_{ext} compared to X = C, while the X = caC, hmC, and fC sequences respectively show a ~30, 40, and 70 kJ/mol reduction, suggesting increasingly less cooperative melting and an increasingly gradual duplex to single-strand transition.

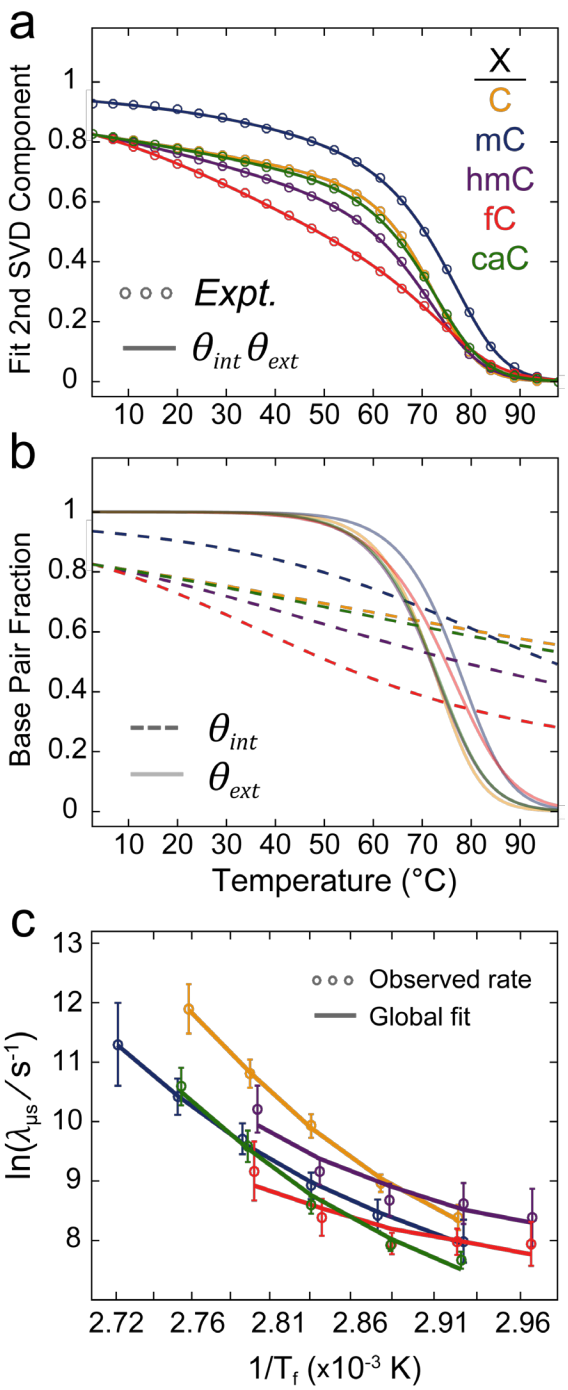


Figure 5: (a) Fitting the experimental melting curves at pD 7.2 (points) to the product $\theta_{int}\theta_{ext}$ (solid line). (b) The internal (dashed line) and external (solid line) base pair fractions determined from the fits in panel (a). (c) Global fit to the observed rate $\lambda_{\mu s}$ and the external fraction θ_{ext} for each sequence. Error bars reflect the amplitude weighted standard deviation of the observed rate as determined from the rate domain representation of the t-HDVE data (SI).

The external entropy, ΔS°_{ext} in Fig. 6a follows a similar trend to ΔH°_{ext} as a function of XpG modification and reflects the standard entropy gain upon duplex dissociation. The interplay between the external enthalpy and entropy dictates the duplex to single-strand equilibrium at a given temperature and determines θ_{ext} according to eq. S5 in the SI. The similar trends in ΔH°_{ext} and ΔS°_{ext} reflect an enthalpy/entropy compensation in which a decreasing positive enthalpy change is offset by a decreasing positive entropy gain upon duplex dissociation. Furthermore, the trend in ΔH°_{ext} and ΔS°_{ext} is correlated with the pattern of decreasing internal fractions seen in Fig. 5b. This observation is consistent with the assignment of asymmetry in the melting curves as arising from an accumulating reduction in base pairing character within the duplex ensemble (reduced θ_{int}) since such pre-dissociation dehybridization reflects non-cooperative melting (reduced ΔH°_{ext}) as well as a reduction in the entropy change between an increasingly disordered duplex state and the single-strand state (reduced ΔS°_{ext}).

The internal standard enthalpy, ΔH°_{int} and entropy, ΔS°_{int} can be interpreted as the standard thermodynamic quantities per base pair averaged across the disruption of contacts that contribute to the observed asymmetry in the melting curves. As such, ΔH°_{int} and ΔS°_{int} for the canonical sequence should be less than or equal to the empirically determined dinucleotide NN parameters from the nearest-neighbor model since these parameters are the standard enthalpy and entropy changes upon the complete dehybridization of a specific dinucleotide step within a DNA sequence.[48] We can estimate this upper bound for the X = C sequence by calculating the average of the dinucleotide step parameters per base pair in the 5′-TACGCGCGTA-3′ sequence, resulting in a mean standard enthalpy of 19 kJ/mol and a mean standard entropy of 50 J/mol·K. These values are a factor of five greater than the ΔH°_{int} = 3.7 kJ/mol and ΔS°_{int} = 10 J/mol·K determined from fitting the melting curve of the X = C sequence, suggesting that the internal thermodynamic parameters report on a reduction in base pairing contacts within the duplex but do not correspond to the enthalpy/entropy changes anticipated for a complete loss of all hydrogen bonding and base stacking within a base pair.

The internal standard enthalpy and entropy in Fig. 6a show a distinct pattern relative to the external thermodynamic quantities, but the trends observed in ΔH°_{int} and ΔS°_{int} are similar to each

other, as seen for ΔH°_{ext} and ΔS°_{ext} above. This result indicates an enthalpy/entropy compensation similar to that observed for the external parameters, but at the level of individual base pairing contacts. The X = mC, fC, and to a lesser extent hmC sequences each show an increased average internal enthalpy relative to the canonical sequence, but this is also accompanied by an increase in the average internal entropy upon the disruption of a base pair. The X = caC sequence shows a modest increase of a few kJ/mol in ΔH°_{int} and a few tens of J/mol·K in ΔS°_{int}. As one would expect from their similar θ_{int}, the internal thermodynamics of the X = caC sequence are the most comparable to the X = C sequence at physiological pH.

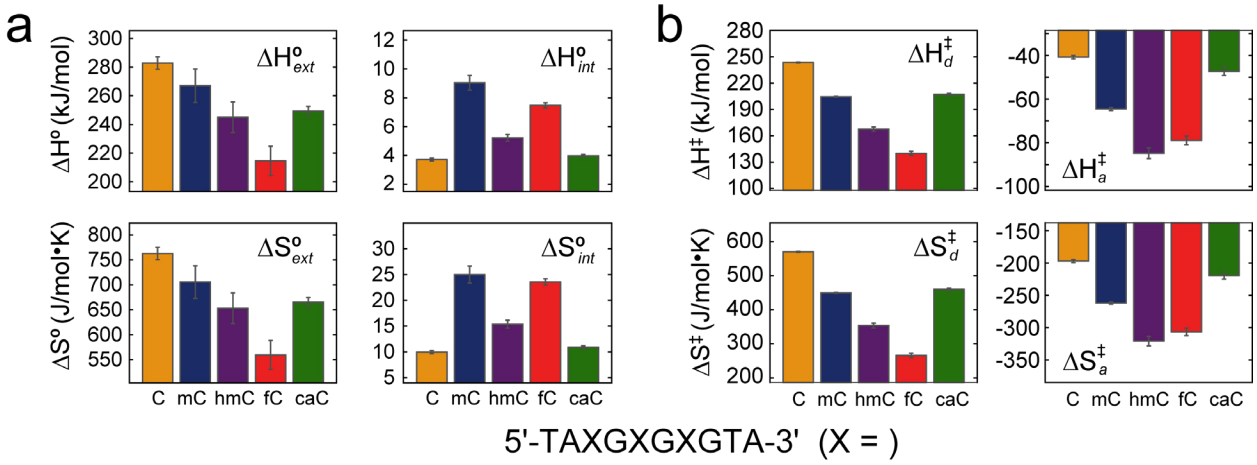


Figure 6: (a) The modification-dependent trends in external enthalpy ΔH°_{ext}, internal enthalpy ΔH°_{int}, external entropy ΔS°_{ext}, and internal entropy ΔS°_{int} determined from the fits presented in Fig. 5. (b) Modification dependent trends in the activation enthalpy and entropy for DNA oligonucleotide dissociation and association, which represent the fit parameters of the global model. Error bars in panels (a,b) correspond to 95% confidence intervals from the fit.

To self-consistently describe modification dependent trends in DNA oligonucleotide thermodynamics and kinetics, we globally fit the observed rate constant $\lambda_{\mu s}$ and the external fraction θ_{ext} from the melting curve with respect to the four activation parameters $\Delta H^{\ddagger}_{d}$, $\Delta H^{\ddagger}_{a}$, $\Delta S^{\ddagger}_{d}$, and $\Delta S^{\ddagger}_{a}$ for each sequence. In practice, the observed rate constants λ_{ns} and $\lambda_{\mu s}$ are determined from the T-jump kinetics by transforming the time-domain data into a rate-domain representation as described previously.[37, 57] An example of the rate domain representation of the

T-jump measurements outlined in Fig. 2a is included in the SI. This approach allows the observed rates to be read off directly without resorting to fitting the kinetic traces to an assumed functional form. The observed rate constant is related to the dissociation and association rate constants by eq. 2. To model the external (or duplex) fraction with respect to the activation parameters, the temperature dependence of the rate constants given by eq. 3 is used to describe the temperature dependence of K_{ext}, which in turn can be related to θ_{ext} by eq. S5. Fig. 5c shows the global fit to the temperature dependence of the observed rate $\lambda_{\mu s}$ for each sequence. The lowest amplitude T-jump data set is excluded from the fit for all sequences, which corresponds to T_i = 50 °C for the X = C, mC, and caC sequences and T_i = 75 °C for the X = hmC and fC sequences. Exclusion of these points improves the quality of the fit but does not alter the trends in activation parameters. The external fractions obtained from global fitting overlay with the θ_{ext} plotted in Fig. 5b.

As can be seen in Fig. 5, the data are simultaneously well described suggesting that the thermodynamics and kinetics of hybridization for each sequence are reasonably captured by the global model. Furthermore, the results are consistent irrespective of whether the free parameters are defined relative to activation or thermodynamic quantities. Fig. 6b shows the modification dependent activation enthalpy and entropy of duplex dissociation and association. The parameters corresponding to duplex dissociation are positively signed and mirror the trends in ΔH°_{ext} and ΔS°_{ext} in Fig. 6a, suggesting that the effects of modification X on the duplex state discussed above are also imparted on the dehybridization transition state. Conversely, the association parameters are negatively valued irrespective of X, consistent with the exothermic formation of base pairing contacts and the reduction in translational and configurational entropy expected for the capture of a single-strand by its complement in the formation of the hybridization transition state. The magnitude of $\Delta H^{\ddagger}_{a}$ is inversely related to that of $\Delta H^{\ddagger}_{d}$, with the X = C sequence returning the smallest magnitude at -40 kJ/mol and the X = hmC and fC sequences exhibiting the largest at -85 and -79 kJ/mol. Similar to the trends observed for $\Delta H^{\ddagger}_{d}$, X = mC and caC represent intermediate cases with $\Delta H^{\ddagger}_{a}$ = -65 and -47 kJ/mol, respectively. The activation entropy of association follows a similar trend with modification, ranging between -197 J/mol·K for the X = C sequence and -321 J/mol·K for the X = hmC sequence.

Evaluating the pH Dependence of Hybridization when X = caC

At physiological pH, the modification that behaves most similarly to canonical C with respect to its influence on both hybridization thermodynamics and kinetics is X = caC. However, the TDG excision activity towards caC and its impact on the physical properties of the nucleobase are known to be pH-dependent.[28, 58] The pK_a of the exocyclic carboxyl group of the caC nucleobase is around 4.5, meaning that this group is predominantly deprotonated at physiological pH, but it has been proposed that protonation of this moiety within the active site of TDG or even transiently within dsDNA could contribute to the selective excision or recognition of caC in the genome.[28-29, 58-59] To evaluate the influence of the protonated caC nucleobase on the thermodynamics and kinetics of hybridization, we conducted a series of pH-dependent FTIR and T-jump nonlinear IR measurements on the X = caC sequence following the same approach as above. Since the pH-dependent experiments are conducted in D_2O, we report pH* values that correspond to the reading from a standard glass electrode pH meter in deuterated solution.

A detailed characterization of the distinct pH-dependence of dsDNA in response to N3 protonation (pK_a = 4.3) when X = C versus protonation of the exocyclic carboxyl group when X = caC is the focus of a complementary study published previously.[60] In brief, we find that protonation at N3 completely disrupts the ability of the canonical sequence to form a duplex and significantly perturbs the duplex to single-strand equilibrium. In contrast, protonation of the -COO^- group is less perturbative, leading to increased base pair fluctuations and a reduction in the activation energy of dehybridization without preventing Watson-Crick base pairing. The pertinent pH-dependent results from Ref. 60 are summarized below in order to facilitate comparison of the anionic and protonated forms of the X = caC duplex against the X = C, mC, fC, and hmC oligonucleotides studied here. Fig. 7a shows a series of FTIR melting curves for the X = caC sequence acquired between pH* 3.5-6.8, corresponding to a pD range of 3.9-7.2.[61] The melting curve measured at each pH* is fit to the product of $\theta_{int}\theta_{ext}$. The internal (dashed line) and external (solid line) base pair fraction determined from the fit are plotted separately in Fig. 7b. From pH* 4.5-6.8, there is little impact on θ_{ext}, but θ_{int} is observed to descend more sharply with rising temperature as the pH* of the solution is lowered. This result suggests that, as carboxyl sites within the major groove of the duplex begin to protonate, local base pair fluctuations increase and the overall average fraction of base pair character within duplexes decreases. However, down to pH* 4.5 these changes do not appear to significantly perturb the duplex to single-strand

equilibrium, as indicated by the insensitivity of θ_{ext}. Below pH* 4.5 the internal fraction continues to drop with increasing protonation, but the inflection point of the external fraction also begins to shift towards lower temperature. This change suggests that as the pH of the solution descends below the pK_a of the carboxyl group and the majority of caC nucleotides become protonated, K_{ext} begins to shift increasingly in favor of single-strands.

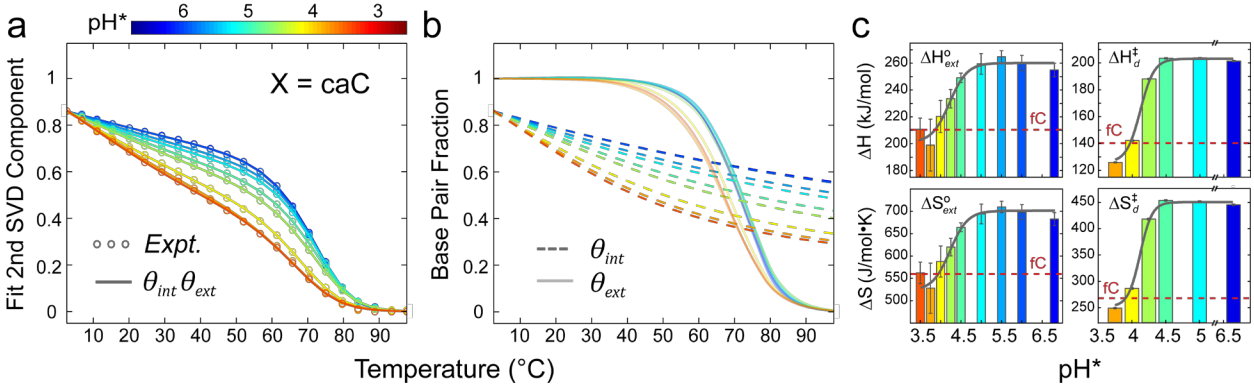


Figure 7: (a) FTIR melting curves measured for the X = caC sequence between pH* 3.5-6.8, corresponding to pD 3.9-7.2. The experimental data is plotted as points while fits to $\theta_{int}\theta_{ext}$ are plotted as solid lines. (b) The internal (dashed line) and external (solid line) base pair fractions determined from the fits in panel (a) (c) The external enthalpy and entropy from fitting the thermodynamic model to the melting curves (left) and the activation enthalpy and entropy from globally fitting θ_{ext} and the temperature dependence of $\lambda_{\mu s}$ as a function of pH* (right). Error bars indicate 95% confidence intervals from the fit. The dashed red line in each subpanel corresponds to the value measured for the X = fC sequence under physiological conditions for reference. Figure is adapted from Figs. 5 and 6 in Ref. 60.

In addition to pH-dependent FTIR melting studies, we also measured the influence of caC protonation on the dehybridization kinetics of the X = caC sequence using T-jump nonlinear IR spectroscopy and by globally fitting to the Kramers model described above. Fig. 7c summarizes the pH-dependent results by showing the trends in ΔH°_{ext} and ΔS°_{ext} from the melting curve and the trends in $\Delta H^\ddagger_d$ and $\Delta S^\ddagger_d$ from the global fit. Both sets of parameters trace out titration profiles centered near the pK_a of the caC carboxyl group. As the pH* is lowered from 6.8 to 3.5 ΔH°_{ext} drops ~60 kJ/mol, indicating that the cooperativity of the duplex to single-strand transition is reduced as caC protonates. This drop-off is accompanied by a 170 J/mol·K reduction in ΔS°_{ext},

consistent with a diminishing entropic difference between the increasingly disordered duplex state and the single-strand state. A similar pattern in the dissociation activation enthalpy and entropy with decreasing pH* suggests a less concerted loss of contacts between the duplex state and dehybridization transition state. Taken together, these results indicate that the X = caC sequence behaves increasingly like the X = fC sequence as the carboxyl groups in the duplex are protonated. For reference, the quantities in Fig. 7c determined for the X = fC sequence under physiological conditions are indicated on each subpanel with a dashed red line, showing that by pH* ~4.0 the external and activation parameters associated with X = caC have approached the values associated with X = fC. This observation suggests that the caC nucleotide can dynamically switch between C-like and fC-like behavior depending on the protonation state of the exocyclic carboxyl group.

Discussion

The Influence of X = mC is more Complicated than Simple Stabilization of the Duplex State

The mC nucleotide lowers the duplex standard free energy relative to the X = C sequence and shifts the duplex to single-strand equilibrium to higher temperature by ~5 °C, consistent with many past reports of modest duplex stabilization.[29-30, 33] However, this shift in the duplex fraction is not the only effect on hybridized DNA. Despite the inflection point lying at higher temperature, the slope of the melting transition is reduced relative to the canonical sequence, reflecting a small reduction in the standard enthalpy of duplex dissociation and a minor decrease in the cooperativity of dehybridization. This effect is also reflected in the internal base pairing fraction for the X = mC sequence, which starts above and roughly parallel to the θ_{int} associated with the X = C sequence at low temperature, but then begins to drop off more sharply above 50 °C, eventually overtaking X = C near 80 °C. Taken together, these results indicate a more nuanced impact on DNA thermodynamics than simple stabilization of the duplex state. It appears that mC stabilizes base pairing within the XpG domain and leads to a reduction in base pair fluctuations at temperatures below the duplex melting transition, but this trend reverses with increasing temperature. The conclusion that mC inhibits base pairing motions at low temperature is consistent with reports of increased duplex rigidity at 22 °C in helices containing mC as well as molecular dynamics (MD) simulations that suggest a reduction in structural fluctuations in mC:G base pairs at room temperature.[23-25] A combined NMR and MD study found that methylation of CpG steps restricts

base pairing dynamics, increases duplex stiffness, and reduces major groove solvation.[26] Despite the attenuation in local base pairing dynamics and the stabilizing influence on overall duplex thermodynamic stability, the free energy barrier to dehybridization of the XpG domain at physiological temperature is reduced slightly by 1.6 kJ/mol. This result indicates that the more tightly structured mCpG step is nevertheless kinetically destabilized relative to the unmodified CpG step. These changes to the duplex and transition state are likely attributed to a combination of effects due to the 5-position methyl group, such as increased electron density in the pyrimidine ring, expulsion of solvating waters from the major groove, and increased steric constraints upon the mC:G base pair.

Base Pair Fluctuations are Increased and the Barrier to Opening the XpG Domain is Lowered when X = hmC or fC

The hmC nucleotide does not significantly shift the duplex to single-strand equilibrium at 37 °C, but the internal free energy decreases markedly, reflecting an increased degree of structural fluctuation within hmC:G base pairs, consistent with previous reports of moderately increased flexibility and enhanced base pair mobility within hmC-containing dsDNA.[23-24] Most strikingly, the free energy barrier to dehybridization is lowered with respect to the canonical sequence by 8.8 kJ/mol. Taking all of these observations into consideration, modification with X = hmC is disruptive at the level of base pairing and lowers the barrier to opening the XpG domain, but the duplex to single-strand equilibrium is not perturbed significantly relative to the canonical sequence.

At physiological temperature and pH, the modification that most significantly alters the physical properties of dsDNA is X = fC, which is coincidentally the excision target toward which TDG displays the greatest activity under these conditions.[28] Not only is the standard free energy of the duplex state increased relative to the unmodified sequence, but fC:G base pairs are considerably more disordered than C:G base pairs and induce greater structural fluctuation about average B-form DNA, as evidenced by the sharp decline in the internal base pair fraction as well as the enhanced response observed in the 180 ns t-2D IR surface. This result is consistent with past experimental and computational work that established that the fC nucleotide confers even greater flexibility to dsDNA and induces even greater disorder within the duplex than the hmC nucleotide.[23] Both hmC and fC have also been observed to accelerate the rate of imino proton

exchange in dsDNA, interpreted as an increase in the rate of base pair opening.[21] Despite the apparent disruption at the local level of base pairing, the overall duplex to single-strand equilibrium for the X = fC sequence is shifted towards higher temperature by almost as much as the X = mC sequence, as indicated by the ~3 °C shift of the inflection point of the duplex fraction relative to the canonical sequence. However, dehybridization is considerably less cooperative for the X = fC sequence. As a result this duplex is the least stable at physiological temperature and the standard free energy of duplex formation at 37 °C is 5.2 kJ/mol less favorable than the canonical sequence.

The primary impact of X = fC on hybridization thermodynamics is thus to disrupt the cooperativity of opening the XpG step. Rather than dehybridizing as a collective unit over a narrow temperature range, as is typical of short C:G rich oligonucleotides,[45,48] the fC:G base pairs increasingly loosen and disorder as the temperature rises. As a consequence the eventual dissociation into single-strands involves the disruption of fewer base pairing contacts on-average, but dsDNA can nevertheless persist at higher temperatures relative to the X = C sequence due to entropic stabilization of the duplex state. These effects are also reflected in the transition state and the activation free energy of dehybridization, which is lowered with respect to the canonical sequence by 9.1 kJ/mol at 37 °C. Modification with X = fC not only most disorders base pairing within dsDNA, but also most significantly lowers the barrier to opening the XpG step at physiological temperature.

Interestingly, at physiological pH and temperature the X = hmC nucleotide is the most similar to the X = fC nucleotide both in terms of its impact on local base pairing dynamics and on the barrier to opening the XpG step, but the effects observed are less drastic. In contrast to the fC nucleobase, hmC is not excised in the active demethylation pathway. Multiple reports suggest that specific interactions between caC and fC within the TDG enzyme are critical in flipping the target nucleobase into the active site and in subsequent base excision activity.[22, 62-64] Furthermore, it has been determined that the leaving group ability of the nucleobase, as reported through N1 acidity, determines excision activity and that this effect can account for the fact that fC and caC are the only bases removed by TDG.[28] The current results do not contradict this picture, but do suggest that hmC could play a unique role in genomic DNA by imparting a degree of fC-like character on the double helix, thereby facilitating enzymatic recognition at hmCpG sites.

Several compounding factors likely account for the unique properties of the fC-containing and, to a milder degree, hmC-containing duplexes. The 5-position formyl group is capable of

accepting a hydrogen bond from the 4-position amino group or solely from solvating waters in the major groove depending on whether the formyl oxygen atom is rotated towards the 4- or 6-positions. Evidence from crystal structures and solution phase NMR studies suggest that the formation of an intramolecular hydrogen bond with the amino group is the preferred orientation, but perhaps transient rotation of the formyl group within the duplex occurs.[19,21,65-66] Crystal structures of hmC-containing dsDNA reveal at least two rotational conformations of the hydroxymethyl group with hydrogen bonds to the adjacent G nucleobase.[21,67] Competing hydrogen bond partners as well as the ability to dynamically shift the hydrophobicity of the major groove environment depending on the orientation of the 5-position substituent could thus contribute to the disordered base pairing and the reduced dehybridization barrier observed for the X = fC and hmC sequences. In addition, we have previously suggested that N3 hydrogen bonding is weakened in fC:G base pairs, which could result in increased hydrogen bond fluctuations, not to mention possible modification-dependent modulation of base stacking affinities.[29, 66, 68] The 460 J/mol drop in ΔG°_{int} at 37 °C relative to X = C is suggestive of such reductions in base pair stability. In reality, the influence of the 5-position substituent on the interactions that mediate base pairing and the impact of the moiety on the shell of hydration surrounding the duplex are inseparable contributions to the free energy since the solvation and structure of nucleic acids are linked intimately.[69-72] It is thus most likely a combination of these effects that determines the observed impact of cytosine modification on the physical properties of dsDNA.

The X = caC Sequence Behaves like the X = C or fC Sequence Depending on the Protonation State of the 5-Position Carboxyl Group

At physiological pH and temperature, the X = caC sequence behaves most similarly to the canonical sequence. This result is consistent with past reports that modification with this nucleobase does not alter the flexibility of DNA at 22 °C, does not induce greater base pair motion in MD simulations, and does not accelerate the rate of imino proton exchange at 15 °C in dsDNA.[21,23] A minimal shift in the standard free energy of duplex formation is observed and the free energy barrier to dissociation is only reduced by 2.4 kJ/mol at 37 °C. The similarity of the X = caC sequence and the canonical sequence is perhaps surprising given that the caC nucleobase represents the most highly oxidized step in the active demethylation cycle and is a known excision

target of TDG. However, as discussed above, it is well established that the excision activity towards caC is pH dependent.[28, 58-59] Our study of the impact of caC modification on duplex thermodynamics and kinetics reveals that the X = caC sequence behaves increasingly like the X = fC sequence with decreasing pH. Transient structural variation in the base pairs of the duplex increases as the exocyclic carboxyl groups become increasingly protonated, but up until pH* ~4.5 the overall duplex to single-strand equilibrium is not perturbed significantly and the barrier height to dehybridization remains unchanged. These observations indicate that the primary effect of protonation occurs at the level of base pairing, leading to enhanced base pair mobility without substantially altering duplex stability or the dehybridization transition state as long as the anionic form of caC predominates.

As protonation of the caC nucleobases in the duplex becomes abundant, the impact on the thermodynamics and kinetics of dehybridization begins to increase to the point that the highly protonated X = caC sequence behaves more like the X = fC sequence under neutral conditions, with a measurable impact on both the hybridization barrier height and duplex equilibrium. Most strikingly, the free energy barrier to dissociation is lowered 12.8 kJ/mol at pH* = 4.0 relative to the canonical sequence at neutral pH, representing a reduction several kJ/mol greater than the effect of X = hmC and fC modification. The internal base pair fraction also continues to descend with decreasing pH*, reflecting increased base pair fluctuations induced by protonation of the carboxyl group.[60]

As with the fC and hmC nucleobases, there are several interrelated factors that could account for the effects observed with increasing protonation of caC. The inductive effect of a protonated carboxyl group is similar to that of a formyl group and it has been proposed that caC protonation weakens the stability of the N1 glycosidic bond and N3 hydrogen bond.[28-29] It is possible that this effect leads to increased hydrogen bond fluctuations in the protonated caC:G base pair. Crystal structures of caC-containing dsDNA indicate that the deprotonated carboxyl group adopts a similar arrangement to the formyl group of fC by orienting in the plane of the pyrimidine ring and forming an intramolecular hydrogen bond with the 4-position amino group.[21, 73] In this respect the protonated caC nucleobase is likewise more similar to the fC nucleobase in that the anionic carboxyl group, with its delocalized negative charge, becomes neutral and asymmetric upon protonation. In this case rotation of the 5-position substituent would dynamically shift the solvation environment of the major groove by switching the orientation of the hydrogen

bond donating and accepting sites on the protonated carboxyl group. In addition to variable interactions with solvating waters, the protonated oxygen could also form transient hydrogen bonds to adjacent nucleobases in analogy to such contacts formed by hmC discussed above. On the whole, the pH-dependent results indicate that even a modest degree of protonation has a measurable impact on base pairing, as evidenced by the immediate drop in the internal base pairing fraction of the X = caC sequence as pH* decreases, but a considerable number of protonated caC sites are required before a transition to X = fC-like behavior is observed duplex-wide.

In addition to the thermodynamic and kinetic framework proposed to account for our experimental observations, a discussion of the shape of the time traces measured for each sequence and pH* is included in the SI to emphasize that the modification and pH-dependent degree of disorder in the duplex is apparent in the kinetic data independent of any model.

Conclusions

Our steady-state and transient T-jump results provide a comprehensive characterization of the impact of cytosine modification on several fundamental physical aspects of dsDNA, such as the modulation of local base pairing dynamics, reduction in the cooperativity of XpG melting, and alteration of the hybridization transition state. By applying IR spectroscopy and a thermodynamic framework that allows for the possibility of disrupted base pairing contacts in the duplex, the impact of X = mC on the dehybridization of a model oligonucleotide sequence was found to be more nuanced than the well-established result that this modification is stabilizing. Although mC:G base pairs are more tightly bound at temperatures below the dehybridization transition and the free energy of the duplex state is lowered, the cooperativity of melting and the free energy barrier to dissociation at physiological temperature are nevertheless reduced relative to canonical DNA. Modification with either X = hmC or fC leads to increased base pair fluctuations in the duplex state, possibly due in part to the ability of the 5-position substituents to transiently switch between local hydrogen bond partners and to change the degree of solvation of the major groove depending on their orientation. These changes to the duplex state are also accompanied by a decrease in melting cooperativity and a lowering of the free energy barrier to dissociation for both sequences, but in all instances the effect is greatest when X = fC. At physiological temperature and pH, modification with X = caC has the least impact on dsDNA, with only modest destabilization

observed in the standard free energy of the duplex state and a decrease of a few kJ/mol in the dehybridization barrier. However, the influence of the caC nucleobase is found to be highly dependent upon the protonation state of the exocyclic carboxyl group. As the 5-position substituent becomes increasingly protonated, the X = caC sequence behaves increasingly like the X = fC sequence such that by pH* ~4.0, the impact of modification with X = caC is more perturbative in some respects.

This study focuses on a single heavily modified model oligonucleotide design in order to magnify the modification-dependent effects. The fact that these oligonucleotides still adopt canonical B-form helical structures and demonstrate only minor shifts in the duplex to single-strand equilibrium suggests that this is a reasonable model system for modified dsDNA. The authors of Ref. 33 considered constructs containing both single, multiple, and non-contiguous modified CpG sites, noting that although the magnitude of the shift in stabilization in response to mC and hmC substitution varies according to sequence context, the overall trend that hmC appears to reverse the stabilization conferred by methylation at the 5 position is observed regardless of the sequence. More broadly, such behavior is expected since it is well-established that the thermodynamics of nucleic acid hybridization in both polymers and oligonucleotides is additive and can be decomposed into the contributions of discrete dinucleotide steps.[48] Therefore although we do not discount the importance of local sequence on the precise magnitude of the reported effects, we have no reason to expect that the trends we observe with modification do not apply more generally beyond the sequence studied here.

Regardless of the identity of modification X, the free energy barrier to dehybridization at 37 °C is reduced relative to the canonical sequence. Furthermore, we find that each modification tunes the properties of the XpG step away from the unmodified CpG step, even if in opposing directions. Perhaps the distinct influence of X on local base pairing dynamics, the thermodynamic stability of dsDNA, and the barrier to opening an XpG step is utilized by enzymes to selectively identify, bind, or act on their specific nucleobase targets. For example, recently reported high-resolution crystal structures of a TDG-DNA complex indicate that a proton is likely transferred to the exocyclic carboxyl group when it is flipped into the active site of TDG.[74] Understanding the effect of protonation on the fundamental physical properties of the caC nucleobase is undoubtedly an important aspect of understanding the significance of such events when observed in biology. The focus of the present work is to assess the impact of each naturally occurring 5-position

modification on fundamental biophysical properties at the level of dsDNA since such insight is a necessary step towards understanding the larger questions surrounding these nucleobases.

Materials and Methods

Synthesis and Purification of 5′-TAXGXGXGTA-3′ (X = mC, hmC, fC and caC)

Unmodified, mC, hmC, fC and caC phosphoramidites and other UltraMild reagents for oligonucleotide synthesis were purchased from Glen Research. DNA oligomers were synthesized at 1 μmol scale in several batches. After synthesis, the X = mC beads were treated with concentrated NH_4OH at room temperature for 4 hrs. The X = hmC beads were treated with 0.4 N NaOH in 4:1 MeOH/H_2O at room temperature for 16 hrs, followed by neutralization with AcOH. The X = fC beads were treated with concentrated NH_4OH at room temperature for 4 hrs. After drying by vacuum concentrator, the residue was dissolved in 900 μl H_2O plus 100 μl 3 M NaOAc (pH = 5.3) and incubated at 37 °C for 16 hrs to ensure the acetal protecting groups fully hydrolyzed. The X = caC beads were treated with 0.1 M K_2CO_3 in 1:1 MeOH/H_2O at 42 °C for 16 hrs, then neutralized with AcOH. Maldi TOF MS gave the expected MS: X = mC oligonucleotide, [MH]+ = 3069; X = hmC, [MH]+ = 3117; X = fC, [MH]+ = 3111; X = caC, [MH]+ = 3159. Oligonucleotides were dialyzed at 4 °C in purified water (18 MΩ, Millipore) for at least 48 hours, lyophilized, and dissolved in a deuterated buffer solution in preparation for IR spectroscopy. A second lyophilization step followed by the addition of the appropriate volume of pure deuterium oxide (D_2O, Cambridge Isotopes, 99.9%) was necessary to HD exchange labile protons and ensure isotopically pure solutions. For all experiments at physiological pH, sample conditions were 1 mM oligonucleotide in 20 mM sodium phosphate buffer (pD 7.2) plus 16 mM NaCl. A DCl solution was used to tune the pH* of each solution for the pH-dependent experiments. DNA samples were annealed prior to all experiments by heating the solution to 95 °C and allowing the samples to cool to room temperature.

Equilibrium IR Measurements

Temperature dependent FTIR spectra were acquired using a Bruker Tensor FTIR spectrometer at 4 cm^{-1} resolution averaging 30 scans per temperature point. Samples were held between two 1 mM thick CaF_2 windows with a 50 μm spacer setting the path length. The bath temperature was stepped between 0-105 °C in 5 °C steps using a recirculating chiller

(Ministat 125, Huber). The sample temperature in the home-built brass holder at a given bath set point was calibrated using a thermocouple to measure the temperature at the center of the CaF_2 window. Equilibrium 2D IR spectra were measured in the boxcar geometry using a spectrometer that has been described previously.[75] The waiting time was fixed at $\tau_2 = 150$ fs for all nonlinear measurements. The coherence time was stepped in 4 fs steps from $\tau_1 = -60$ to 2500 fs and -60 to 3000 fs for the nonrephasing and rephasing surfaces, respectively. All spectra were acquired with parallel (ZZZZ) polarization to maximize the total signal intensity. As with the FTIR measurements, the sample temperature was set using a home-built brass sample holder and a recirculating chiller.

Transient T-Jump IR Measurements

The temperature jump spectrometer and experiment have been described in detail previously.[42-43] Six T-jumps were measured for each sequence with the initial temperature (T_i) set between 50-75 °C and spaced 5 °C apart such that the entire temperature range of the duplex to single-strand transition was sampled for each sequence. An additional $T_i = 80$ °C T-jump was measured for the X = mC sequence since the melting curve is shifted to higher temperature for this oligonucleotide. The T_i of the sample was set using a recirculating chiller and the T-jump magnitude (ΔT) was set between 14-15 °C by monitoring the change in transmission of the bend-libration combination band of the D_2O solvent.[76] An undersampling scheme in which the coherence time was stepped in 16 fs steps from -60 to 1250 fs and -60 to 1750 fs for the nonrephasing and rephasing surfaces was employed when collecting the t-2D IR spectra to reduce data collection time.

Author Information

Corresponding Author

*tokmakoff@uchicago.edu

Notes

The authors declare no competing financial interests.

Associated Content

Supporting Information Available. Application of the all-or-none model to the melting curves. Assignment of spectroscopic changes along the low temperature baseline of the melting curves. Details of the thermodynamic/kinetic model. Discussion and analysis of t-2D IR spectra for the X = C and fC sequences. Rate domain representation of t-HDVE spectra and determination of observed rates. Modification and pH-dependent trends in stretched exponential kinetics. This information is available free of charge via the Internet at http://pubs.acs.org.

Acknowledgements

A.T. thanks the National Science Foundation (Grant No. CHE-1856684) and the National Institute of General Medical Sciences of the National Institutes of Health (Award No. R01GM118774) for support of this research. C.H. thanks the National Institutes of Health (Award No. R01HG006827). Q.D. was supported by the National Institutes of Health grant 5K01HG006699. B.A. acknowledges support from the NSF GRFP.

Citations

1. Law, J. A.; Jacobsen, S. E., Establishing, Maintaining and Modifying DNA Methylation Patterns in Plants and Animals. *Nat. Rev. Genet.* **2010,** *11*, 204-220.
2. Schübeler, D., Function and Information Content of DNA Methylation. *Nature* **2015,** *517*, 321-326.
3. Ehrlich, M.; Gama-Sosa, M. A.; Huang, L.-H.; Midgett, R. M.; Kuo, K. C.; McCune, R. A.; Gehrke, C., Amount and Distribution of 5-Methylcytosine in Human DNA from Different Types of Tissues or Cells. *Nucleic Acids Res.* **1982,** *10*, 2709-2721.
4. Bird, A. P., CpG-Rich Islands and the Function of DNA Methylation. *Nature* **1986,** *321*, 209-213.
5. Cedar, H., DNA Methylation and Gene Activity. *Cell* **1988,** *53*, 3-4.
6. He, Y.-F.; Li, B.-Z.; Li, Z.; Liu, P.; Wang, Y.; Tang, Q.; Ding, J.; Jia, Y.; Chen, Z.; Li, L., Tet-Mediated Formation of 5-Carboxylcytosine and its Excision by TDG in Mammalian DNA. *Science* **2011,** *333*, 1303-1307.
7. Ito, S.; Shen, L.; Dai, Q.; Wu, S. C.; Collins, L. B.; Swenberg, J. A.; He, C.; Zhang, Y., Tet Proteins can Convert 5-Methylcytosine to 5-Formylcytosine and 5-Carboxylcytosine. *Science* **2011,** *333*, 1300-1303.
8. Iurlaro, M.; Ficz, G.; Oxley, D.; Raiber, E.-A.; Bachman, M.; Booth, M. J.; Andrews, S.; Balasubramanian, S.; Reik, W., A Screen for Hydroxymethylcytosine and Formylcytosine Binding Proteins Suggests Functions in Transcription and Chromatin Regulation. *Genome Biol.* **2013,** *14*, R119.
9. Song, J.; Pfeifer, G. P., Are there Specific Readers of Oxidized 5-Methylcytosine Bases? *Bioessays* **2016,** *38*, 1038-1047.

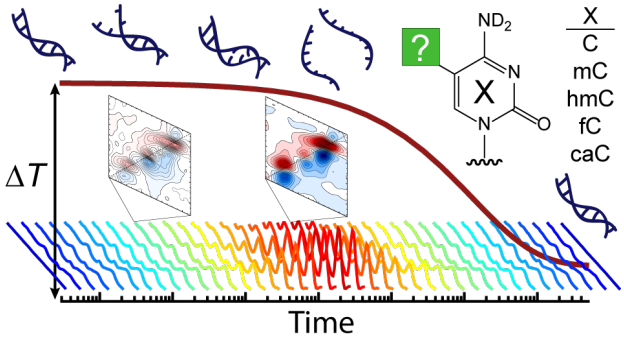
10. Mellén, M.; Ayata, P.; Dewell, S.; Kriaucionis, S.; Heintz, N., MeCP2 Binds to 5hmc Enriched within Active Genes and Accessible Chromatin in the Nervous System. *Cell* **2012,** *151*, 1417-1430.
11. Bachman, M.; Uribe-Lewis, S.; Yang, X.; Williams, M.; Murrell, A.; Balasubramanian, S., 5-Hydroxymethylcytosine is a Predominantly Stable DNA Modification. *Nat. Chem.* **2014,** *6*, 1049-1055.
12. Bachman, M.; Uribe-Lewis, S.; Yang, X.; Burgess, H. E.; Iurlaro, M.; Reik, W.; Murrell, A.; Balasubramanian, S., 5-Formylcytosine can be a Stable DNA Modification in Mammals. *Nat. Chem. Biol.* **2015,** *11*, 555-557.
13. Kriaucionis, S.; Heintz, N., The Nuclear DNA Base 5-Hydroxymethylcytosine is Present in Purkinje Neurons and the Brain. *Science* **2009,** *324*, 929-930.
14. Globisch, D.; Münzel, M.; Müller, M.; Michalakis, S.; Wagner, M.; Koch, S.; Brückl, T.; Biel, M.; Carell, T., Tissue Distribution of 5-Hydroxymethylcytosine and Search for Active Demethylation Intermediates. *PLoS One* **2010,** *5*, e15367.
15. Zhu, C.; Gao, Y.; Guo, H.; Xia, B.; Song, J.; Wu, X.; Zeng, H.; Kee, K.; Tang, F.; Yi, C., Single-Cell 5-Formylcytosine Landscapes of Mammalian Early Embryos and ESCs at Single-Base Resolution. *Cell Stem Cell* **2017,** *20*, 720-731.
16. Eleftheriou, M.; Pascual, A. J.; Wheldon, L. M.; Perry, C.; Abakir, A.; Arora, A.; Johnson, A. D.; Auer, D. T.; Ellis, I. O.; Madhusudan, S., 5-Carboxylcytosine Levels Are Elevated in Human Breast Cancers and Gliomas. *Clin. Epigenetics* **2015,** *7*, 88.
17. Pfeifer, G. P.; Kadam, S.; Jin, S.-G., 5-Hydroxymethylcytosine and its Potential Roles in Development and Cancer. *Epigenetics Chromatin* **2013,** *6*, 10.
18. Hardwick, J. S.; Lane, A. N.; Brown, T., Epigenetic Modifications of Cytosine: Biophysical Properties, Regulation, and Function in Mammalian DNA. *BioEssays* **2018,** *40*, 1700199.
19. Hardwick, J. S.; Ptchelkine, D.; El-Sagheer, A. H.; Tear, I.; Singleton, D.; Phillips, S. E.; Lane, A. N.; Brown, T., 5-Formylcytosine Does Not Change the Global Structure of DNA. *Nat. Struct. Mol. Biol.* **2017,** *24*, 544-552.
20. Renciuk, D.; Blacque, O.; Vorlickova, M.; Spingler, B., Crystal Structures of B-DNA Dodecamer Containing the Epigenetic Modifications 5-Hydroxymethylcytosine or 5-Methylcytosine. *Nucleic Acids Res.* **2013,** *41*, 9891-9900.
21. Szulik, M. W.; Pallan, P. S.; Nocek, B.; Voehler, M.; Banerjee, S.; Brooks, S.; Joachimiak, A.; Egli, M.; Eichman, B. F.; Stone, M. P., Differential Stabilities and Sequence-Dependent Base Pair Opening Dynamics of Watson–Crick Base Pairs with 5-Hydroxymethylcytosine, 5-Formylcytosine, or 5-Carboxylcytosine. *Biochemistry* **2015,** *54*, 1294-1305.
22. Fu, T.; Liu, L.; Yang, Q.-L.; Wang, Y.; Xu, P.; Zhang, L.; Liu, S.; Dai, Q.; Ji, Q.; Xu, G.-L., Thymine DNA Glycosylase Recognizes the Geometry Alteration of Minor Grooves Induced by 5-Formylcytosine and 5-Carboxylcytosine. *Chem. Sci.* **2019**, *10*, 7407-7417.
23. Ngo, T. T.; Yoo, J.; Dai, Q.; Zhang, Q.; He, C.; Aksimentiev, A.; Ha, T., Effects of Cytosine Modifications on DNA Flexibility and Nucleosome Mechanical Stability. *Nat. Commun.* **2016,** *7*, 10813.
24. Wanunu, M.; Cohen-Karni, D.; Johnson, R. R.; Fields, L.; Benner, J.; Peterman, N.; Zheng, Y.; Klein, M. L.; Drndic, M., Discrimination of Methylcytosine from Hydroxymethylcytosine in DNA Molecules. *J. Am. Chem. Soc.* **2010,** *133*, 486-492.

25. Teng, X.; Hwang, W., Effect of Methylation on Local Mechanics and Hydration Structure of DNA. *Biophys. J.* **2018,** *114*, 1791-1803.
26. Derreumaux, S.; Chaoui, M.; Tevanian, G.; Fermandjian, S., Impact of CpG Methylation on Structure, Dynamics and Solvation of Camp DNA Responsive Element. *Nucleic Acids Res.* **2001,** *29*, 2314-2326.
27. Karino, N.; Ueno, Y.; Matsuda, A., Synthesis and Properties of Oligonucleotides Containing 5-Formyl-2′-Deoxycytidine: In Vitro DNA Polymerase Reactions on DNA Templates Containing 5-Formyl-2′-Deoxycytidine. *Nucleic Acids Res.* **2001,** *29*, 2456-2463.
28. Maiti, A.; Michelson, A. Z.; Armwood, C. J.; Lee, J. K.; Drohat, A. C., Divergent Mechanisms for Enzymatic Excision of 5-Formylcytosine and 5-Carboxylcytosine from DNA. *J. Am. Chem. Soc.* **2013,** *135*, 15813-15822.
29. Dai, Q.; Sanstead, P. J.; Peng, C. S.; Han, D.; He, C.; Tokmakoff, A., Weakened N3 Hydrogen Bonding by 5-Formylcytosine and 5-Carboxylcytosine Reduces their Base-Pairing Stability. *ACS Chem. Biol.* **2015,** *11*, 470-477.
30. Raiber, E.-A.; Murat, P.; Chirgadze, D. Y.; Beraldi, D.; Luisi, B. F.; Balasubramanian, S., 5-Formylcytosine Alters the Structure of the DNA Double Helix. *Nat. Struct. Mol. Biol.* **2015,** *22*, 44-49.
31. Sumino, M.; Ohkubo, A.; Taguchi, H.; Seio, K.; Sekine, M., Synthesis and Properties of Oligodeoxynucleotides Containing 5-Carboxy-2′-Deoxycytidines. *Bioorg. Med. Chem. Lett.* **2008,** *18*, 274-277.
32. Münzel, M.; Lischke, U.; Stathis, D.; Pfaffeneder, T.; Gnerlich, F. A.; Deiml, C. A.; Koch, S. C.; Karaghiosoff, K.; Carell, T., Improved Synthesis and Mutagenicity of Oligonucleotides Containing 5-Hydroxymethylcytosine, 5-Formylcytosine and 5-Carboxylcytosine. *Chem.: Euro. J.* **2011,** *17*, 13782-13788.
33. Thalhammer, A.; Hansen, A. S.; El-Sagheer, A. H.; Brown, T.; Schofield, C. J., Hydroxylation of Methylated CpG Dinucleotides Reverses Stabilisation of DNA Duplexes by Cytosine 5-Methylation. *Chem. Commun.* **2011,** *47*, 5325-5327.
34. Erfurth, S. C.; Peticolas, W. L., Melting and Premelting Phenomenon in DNA by Laser Raman Scattering. *Biopolymers* **1975,** *14*, 247-264.
35. Movileanu, L.; Benevides, J. M.; Thomas Jr, G. J., Determination of Base and Backbone Contributions to the Thermodynamics of Premelting and Melting Transitions in B DNA. *Nucleic Acids Res.* **2002,** *30*, 3767-3777.
36. Sanstead, P. J.; Stevenson, P.; Tokmakoff, A., Sequence-Dependent Mechanism of DNA Oligonucleotide Dehybridization Resolved through Infrared Spectroscopy. *J. Am. Chem. Soc.* **2016,** *138*, 11792-11801.
37. Sanstead, P. J.; Tokmakoff, A., Direct Observation of Activated Kinetics and Downhill Dynamics in DNA Dehybridization. *J. Phys. Chem. B* **2018,** *122*, 3088-3100.
38. Peng, C. S.; Jones, K. C.; Tokmakoff, A., Anharmonic Vibrational Modes of Nucleic Acid Bases Revealed by 2D IR Spectroscopy. *J. Am. Chem. Soc.* **2011,** *133*, 15650-15660.
39. Banyay, M.; Sarkar, M.; Gräslund, A., A Library of IR Bands of Nucleic Acids in Solution. *Biophys. Chem.* **2003,** *104*, 477-488.
40. Krummel, A. T.; Zanni, M. T., DNA Vibrational Coupling Revealed with Two-Dimensional Infrared Spectroscopy: Insight into Why Vibrational Spectroscopy Is Sensitive to DNA Structure. *J. Phys. Chem. B* **2006,** *110*, 13991-14000.

41. Hithell, G.; Ramakers, L. A.; Burley, G. A.; Hunt, N. T., Applications of 2D-IR Spectroscopy to Probe the Structural Dynamics of DNA. *Frontiers and Advances in Molecular Spectroscopy*, Elsevier: Amsterdam, **2018**; pp 77-100.
42. Chung, H. S.; Khalil, M.; Smith, A. W.; Tokmakoff, A., Transient Two-Dimensional IR Spectrometer for Probing Nanosecond Temperature-Jump Kinetics. *Rev. Sci. Instrum.* **2007,** *78*, 063101.
43. Jones, K. C.; Ganim, Z.; Tokmakoff, A., Heterodyne-Detected Dispersed Vibrational Echo Spectroscopy. *J. Phys. Chem. A* **2009,** *113*, 14060-14066.
44. Craig, M. E.; Crothers, D. M.; Doty, P., Relaxation Kinetics of Dimer Formation by Self Complementary Oligonucleotides. *J. Mol. Biol.* **1971,** *62*, 383-401.
45. Pörschke, D.; Uhlenbeck, O.; Martin, F., Thermodynamics and Kinetics of the Helix-Coil Transition of Oligomers Containing GC Base Pairs. *Biopolymers* **1973,** *12*, 1313-1335.
46. Patel, D. J.; Hilbers, C., Proton Nuclear Magnetic Resonance Investigations of Fraying in Double-Stranded d-ApTpGpCpApT in Aqueous Solution. *Biochemistry* **1975,** *14*, 2651-2656.
47. Wartell, R. M.; Benight, A. S., Thermal Denaturation of DNA Molecules: A Comparison of Theory with Experiment. *Phys. Rep.* **1985,** *126*, 67-107.
48. SantaLucia, J., A Unified View of Polymer, Dumbbell, and Oligonucleotide DNA Nearest-Neighbor Thermodynamics. *Proc. Natl. Acad. Sci. U.S.A.* **1998,** *95*, 1460-1465.
49. Sanstead, P. J.; Tokmakoff, A., A Lattice Model for the Interpretation of Oligonucleotide Hybridization Experiments. *J. Chem. Phys.* **2019,** *150*, 185104.
50. Bernasconi, C., *Relaxation Kinetics*. Academic Press, Inc.: New York. **1976**.
51. Kramers, H. A., Brownian Motion in a Field of Force and the Diffusion Model of Chemical Reactions. *Physica* **1940,** *7*, 284-304.
52. Sikorav, J.-L.; Orland, H.; Braslau, A., Mechanism of Thermal Renaturation and Hybridization of Nucleic Acids: Kramers' Process and Universality in Watson− Crick Base Pairing. *J. Phys. Chem. B* **2009,** *113*, 3715-3725.
53. Cho, C.; Urquidi, J.; Singh, S.; Robinson, G. W., Thermal Offset Viscosities of Liquid H_2O, D_2O, and T_2O. *J. Phys. Chem. B* **1999,** *103*, 1991-1994.
54. Jäger, M.; Nguyen, H.; Crane, J. C.; Kelly, J. W.; Gruebele, M., The Folding Mechanism of a β-Sheet: The WW Domain. *J. Mol. Biol.* **2001,** *311*, 373-393.
55. Kubelka, J.; Hofrichter, J.; Eaton, W. A., The Protein Folding 'Speed Limit'. *Curr. Opin. Struct. Biol.* **2004,** *14*, 76-88.
56. Marky, L. A.; Breslauer, K. J., Calculating Thermodynamic Data for Transitions of Any Molecularity from Equilibrium Melting Curves. *Biopolymers* **1987,** *26*, 1601-1620.
57. Kumar, A. T.; Zhu, L.; Christian, J.; Demidov, A. A.; Champion, P. M., On the Rate Distribution Analysis of Kinetic Data Using the Maximum Entropy Method: Applications to Myoglobin Relaxation on the Nanosecond and Femtosecond Timescales. *J. Phys. Chem. B* **2001,** *105*, 7847-7856.
58. Hashimoto, H.; Hong, S.; Bhagwat, A. S.; Zhang, X.; Cheng, X., Excision of 5-Hydroxymethyluracil and 5-Carboxylcytosine by the Thymine DNA Glycosylase Domain: Its Structural Basis and Implications for Active DNA Demethylation. *Nucleic Acids Res.* **2012,** *40*, 10203-10214.
59. Hashimoto, H.; Zhang, X.; Cheng, X., Selective Excision of 5-Carboxylcytosine by a Thymine DNA Glycosylase Mutant. *J. Mol. Biol.* **2013,** *425*, 971-976.

60. Ashwood, B.; Sanstead, P. J.; Dai, Q.; He, C.; Tokmakoff, A. 5-Carboxylcytosine and Cytosine Protonation Distinctly Alter the Stability and Dehybridization Dynamics of the DNA Duplex. *J. Phys.Chem. B* **2020**,124, 627-640.
61. Krężel, A.; Bal, W., A Formula for Correlating pKa Values Determined in D_2O and H_2O. *J. Inorg. Biochem.* **2004,** *98*, 161-166.
62. Zhang, L.; Lu, X.; Lu, J.; Liang, H.; Dai, Q.; Xu, G.-L.; Luo, C.; Jiang, H.; He, C., Thymine DNA Glycosylase Specifically Recognizes 5-Carboxylcytosine-Modified DNA. *Nat. Chem. Biol.* **2012,** *8*, 328-330.
63. Da, L.-T.; Shi, Y.; Ning, G.; Yu, J., Dynamics of the Excised Base Release in Thymine DNA Glycosylase During DNA Repair Process. *Nucleic Acids Res.* **2017,** *46*, 568-581.
64. Naydenova, E.; Dietschreit, J. C.; Ochsenfeld, C., Reaction Mechanism for the N-Glycosidic Bond Cleavage of 5-Formylcytosine by Thymine DNA Glycosylase. *J. Phys. Chem. B* **2019,** *123*, 4173-4179.
65. Hu, L.; Lu, J.; Cheng, J.; Rao, Q.; Li, Z.; Hou, H.; Lou, Z.; Zhang, L.; Li, W.; Gong, W., Structural Insight into Substrate Preference for Tet-Mediated Oxidation. *Nature* **2015,** *527*, 118-122.
66. Wang, R.; Luo, Z.; He, K.; Delaney, M. O.; Chen, D.; Sheng, J., Base Pairing and Structural Insights into the 5-Formylcytosine in RNA Duplex. *Nucleic Acids Res.* **2016,** *44*, 4968-4977.
67. Lercher, L.; McDonough, M. A.; El-Sagheer, A. H.; Thalhammer, A.; Kriaucionis, S.; Brown, T.; Schofield, C. J., Structural Insights into How 5-Hydroxymethylation Influences Transcription Factor Binding. *Chem. Commun.* **2014,** *50*, 1794-1796.
68. La Francois, C. J.; Jang, Y. H.; Cagin, T.; Goddard, W. A.; Sowers, L. C., Conformation and Proton Configuration of Pyrimidine Deoxynucleoside Oxidation Damage Products in Water. *Chem. Res. Toxicol.* **2000,** *13*, 462-470.
69. Franklin, R. E.; Gosling, R. G., The Structure of Sodium Thymonucleate Fibres. I. The Influence of Water Content. *Acta Crystallogr.* **1953,** *6*, 673-677.
70. Falk, M.; Hartman, K. A.; Lord, R., Hydration of Deoxyribonucleic Acid. III. A Spectroscopic Study of the Effect of Hydration on the Structure of Deoxyribonucleic Acid. *J. Am. Chem. Soc.* **1963,** *85*, 391-394.
71. Corongiu, G.; Clementi, E., Simulations of the Solvent Structure for Macromolecules. I. Solvation of B-DNA Double Helix at T= 300 K. *Biopolymers* **1981,** *20*, 551-571.
72. Duboué-Dijon, E.; Fogarty, A. C.; Hynes, J. T.; Laage, D., Dynamical Disorder in the DNA Hydration Shell. *J. Am. Chem. Soc.* **2016,** *138*, 7610-7620.
73. Irrera, S.; Portalone, G., First X-Ray Diffraction and Quantum Chemical Study of Proton-Acceptor and Proton-Donor Forms of 5-Carboxylcytosine, the Last-Discovered Nucleobase. *J. Mol. Struct.* **2013,** *1050*, 140-150.
74. Pidugu, L. S.; Dai, Q.; Malik, S. S.; Pozharski, E.; Drohat, A. C., Excision of 5-Carboxylcytosine by Thymine DNA Glycosylase. *J. Am. Chem. Soc.* **2019**, *141*, 18851-18861.
75. Khalil, M.; Demirdöven, N.; Tokmakoff, A., Coherent 2D IR Spectroscopy: Molecular Structure and Dynamics in Solution. *J. Phys. Chem. A* **2003,** *107*, 5258-5279.
76. Williams, S.; Causgrove, T. P.; Gilmanshin, R.; Fang, K. S.; Callender, R. H.; Woodruff, W. H.; Dyer, R. B., Fast Events in Protein Folding: Helix Melting and Formation in a Small Peptide. *Biochemistry* **1996,** *35*, 691-697.

TOC Graphic



Supporting Information

Oxidized Derivatives of 5-Methylcytosine Alter the Stability and Dehybridization Dynamics of Duplex DNA

Paul J. Sanstead[†‡§], Brennan Ashwood[†‡§], Qing Dai[†‡], Chuan He[†‡∥⊥], Andrei Tokmakoff*[†‡§]

[†]Department of Chemistry, [‡]Institute for Biophysical Dynamics, [§]James Franck Institute, [∥]Department of Biochemistry and Molecular Biology, and [⊥]Howard Hughes Medical Institute, The University of Chicago, Chicago, Illinois 60637, United States

*E-mail: tokmakoff@uchicago.edu

I. Application of the Standard All-or-None Model

It is standard practice to assume a model that describes the DNA duplex to single-strand transition and fit melting curves to determine the melting temperature (T_m) as a proxy for DNA duplex stability. Indeed we and many others have taken this exact approach in an attempt to evaluate the influence of these modifications on hybridization thermodynamics.[1-7] The standard model of DNA oligonucleotide melting assumes a two-state all-or-none description of base pairing where all possible base pairs for a given strand are either fully intact (duplex) or fully broken (single-strand). This assumption is typically well justified for short canonical sequences, which generally display melting curves that are sharply transitioning sigmoids symmetric about the inflection point.[8-9] It is clear from the appearance of the melting curves in Fig. 1f that the standard all-or-none description may fail for many of the oligonucleotides containing cytosine modifications, most notably for the X = fC sequence. The low temperature baseline is severely sloped, resulting in highly asymmetric melting curves.

Ignoring these complications and fitting sloping baselines to the melting curves in Fig. 1f results in the baseline corrected curves plotted in Fig. S1a. Sloping baselines on melting curves are a common observation and are attributed to such factors as thermal changes in solvent transmission, path length, sample evaporation, and drifts in spectrometer lamp intensity.[10] It is therefore routine practice to subtract linear fits to the baselines to correct for these measurement artifacts since they do not reflect melting.[8-9, 11] However, the degree to which the low temperature baselines are sloped for these oligonucleotides is so great that it is likely indicative of some change to the DNA dehybridization process itself. The severe slopes on many of the curves further complicate the already subjective task of baseline fitting, since it is difficult to resolve where the low temperature baseline ends. Such a consideration can influence the melting temperature, since T_m is commonly defined as the temperature at which the melting curve is equal to 0.5 and a large baseline correction can shift this point by several degrees depending on how the baselines are selected. Only the low temperature baseline is fit in this case since only 2-3 points can be sampled for the high temperature baseline before the D_2O solvent reaches the boiling point.

In the standard all-or-none model, these corrected curves are assumed to reflect the duplex fraction (θ_D) and T_m is defined as the temperature at which $\theta_D = 0.5$. Comparing the trends suggested by T_m determined in this way, one would conclude that the X = hmC and caC oligonucleotides do not differ meaningfully in stability relative to X = C, but that the fC and mC

nucleotides are slightly stabilizing, resulting in a modest increase in T_m. However, drawing conclusions from this T_m trend fails to consider potential deviations from the all-or-none model invoked to determine the melting temperature and disregards potentially valuable information contained within the temperature-dependent spectra and the shape of the melting curves. Currently there is no consensus on the thermodynamic influence of modified cytosine bases on the stability of duplex DNA.[12] The breakdown of the standard melting curve analysis routinely applied for canonical oligonucleotides is likely a major contributing factor to this lack of agreement.

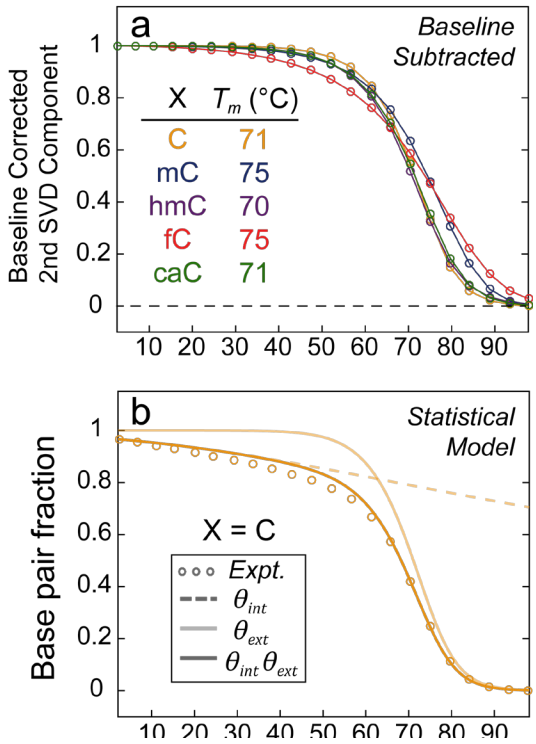


Figure S1: (a) The baseline corrected melting curves following linear fits to the upper baselines of the melting curves in Fig. 1f. The T_m reported in the figure corresponds to the temperature where the corrected melting curve is equal to 0.5. (b) Comparison of the uncorrected 2nd SVD component for the canonical sequence against the statistical model of base pairing from Ref. 13. The internal fraction (θ_{int}) models the average fraction of intact base pairs among duplexes while the external fraction (θ_{ext}) models the fraction of dimers among all DNA strands. The product $\theta_{int}\theta_{ext}$ models the melting curve.

Statistical thermodynamic descriptions in which many possible base pair conformations are considered explicitly offer a more detailed model of DNA hybridization beyond the all-or-none assumption. For example, Fig. S1b shows the melting curve simulated using such an approach that employs nearest-neighbor (NN) parameters and simple polymer physics to model the oligonucleotide dimer ensemble.[13] In this picture, the melting curve is considered to track the total fraction of intact base pairs in the DNA ensemble and this fraction is expressed as the product of an internal (θ_{int}) and external (θ_{ext}) base pair fraction.[14] The internal fraction is calculated as the ensemble average fraction of intact base pairs among duplexes while the external fraction is defined as the fraction of duplexed DNA strands among all DNA strands. As seen in Fig. S1b, the statistical model predicts considerable asymmetry towards the low temperature side of the melting curve for the X = C sequence and θ_{int} is the origin of this effect. The experimental melting curve in Fig. S1b is scaled to the lowest temperature point of the modeled melting curve but it is not baseline corrected in any way. Therefore the apparent residual mismatch in the low temperature baseline slope between the model and experiment is accounted for, at least in large part, by the sloping baseline artifacts intrinsic to spectroscopic melting curves discussed above. The deviation due to this effect is clearly small relative to the overall asymmetry of the melting curve imparted by θ_{int}.

Unfortunately the NN parameters have only been determined for the canonical DNA nucleobases and this model cannot be applied to the modified oligonucleotide sequences directly. However, the insight that the asymmetry in the X = C melting curve originates from an accumulating loss of base pairing contacts within the DNA duplex ensemble with increasing temperature motivates an interpretation of the melting curves in Fig. 1f that considers such effects and provides direct evidence for the manner in which the standard all-or-none model fails to describe the dehybridization of oligonucleotides containing modified cytosine bases.

II. Assignment of Spectroscopic Changes along the Low Temperature Baseline

If the deviations in the shape of the melting curves for these sequences are due to a reduction in base pairing contacts within the duplex ensemble as suggested by the statistical model for the canonical sequence above, then one would expect to observe evidence for a loss of base pairing in the infrared spectra sampled along the low temperature baseline and these changes should be proportional to the degree of asymmetry in the melting curves. Returning to the FTIR

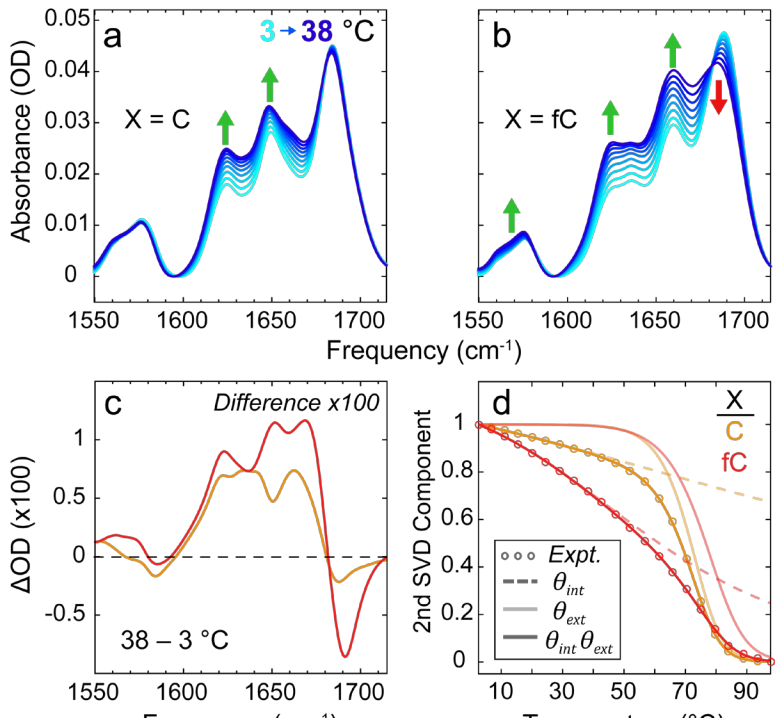


Figure S2: (a) The temperature-dependent FTIR spectrum for the X = C sequence between 3-38 °C. (b) The corresponding spectra for the X = fC sequence. Green arrows highlight intensity gain while red arrows indicate intensity loss. (c) The difference spectrum between the 38 and 3 °C spectra for the X = C sequence plotted in orange and the X = fC sequence plotted in red. (d) Fitting the product of θ_{int} and θ_{ext} to the X = C and fC melting curves. The internal fraction is plotted as a dashed line while the external fraction is plotted as a faded line.

temperature series, Fig. S2a plots the spectra measured for the X = C sequence from 3-38 °C while Fig. S2b plots the corresponding spectra for the X = fC sequence. This temperature range lies well within the low temperature baseline region for both sequences. For the X = C sequence, the G ring mode absorptions below 1600 cm^{-1} as well as the G carbonyl peak centered at 1685 cm^{-1} show minimal changes, but there is an increase in intensity at 1620 cm^{-1} corresponding primarily to A ring mode absorption and at 1650 cm^{-1} corresponding to overlapping absorptions from C, G, and T. The X = fC sequence shows a similar pattern of increasing intensity between 1600 to 1670 cm^{-1}, but a modest increase in G ring mode absorption below 1600 cm^{-1} as well as a loss in intensity at 1685 cm^{-1} are observed as well. Green arrows in Fig. S2a,b highlight these intensity gains while red arrows indicate intensity loss. To further emphasize the changes to the FTIR spectrum outlined

across this temperature range, the difference between the spectrum measured at 38 °C and 3 °C for each sequence is shown in Fig. S2c.

As discussed in the main text, a growth in intensity of G ring mode absorptions accompanied by a loss of intensity at 1685 cm^{-1} are signatures of decreasing C:G base pairing. The minimal change in these features for the canonical sequence suggests that base pairing contacts within the central CpG domain are not disrupted significantly prior to duplex dissociation. In contrast, the fCpG domain appears to show a reduction in base pairing between 3 and 38 °C. The intensity growth observed between 1600 and 1670 cm^{-1} for both sequences likely reflects a loss of contacts at the TA termini of the duplex as well as temperature-dependent changes in solvation of the central XpG domain. The greater increase observed near 1660 cm^{-1} for the X = fC sequence also reflects a reduction in X:G pairing, since the 1685 cm^{-1} G carbonyl peak in a Watson-Crick pair shifts to lower frequency in the unpaired nucleotide.

III. Modeling Melting Curves with respect to θ_{int} and θ_{ext}

These comparisons between the canonical sequence and the X = fC sequence that displays the most atypical melting behavior are consistent with the interpretation of the melting curves as the product of an internal and external base pair fraction, $\theta(T) = \theta_{int}\,\theta_{ext}$. We therefore propose a model that considers both of these effects to account for the shape of the observed melting curves in Fig. 1f. The duplex [*D*] to single-strand [*S*] transition is modeled as a two-state process as in the all-or-none description, but within the duplex ensemble the opening and closing of base pairs is allowed. For simplicity we treat all base pair sites equally and assume that a single rate constant for breaking a base pair k_b and a single rate constant for forming a base pair k_f adequately describes the opening/closing of base pair contacts within the duplex and that the base pairing equilibrium at a given site is described by the equilibrium constant (or statistical weight)

$$s = k_f / k_b \tag{S1}$$

in analogy to classic models of the helix-to-coil transition in biomolecules.[15-17] Fig. 4 in the main text is an overview of this reaction scheme. The dissociation constant for the overall duplex to single-strand transition is then,

$$K_d = \frac{[S]^2}{[D]} = \frac{k_d}{k_a} \tag{S2}$$

The external fraction θ_{ext} is defined as the fraction of duplexed DNA strands out of the total concentration of all DNA

$$\theta_{ext} = \theta_D = \frac{2[D]}{[C_{Tot}]} \tag{S3}$$

where $[C_{Tot}]$ is the total concentration of oligonucleotide strands and θ_D is the fraction of strands that have at least one intact base pair. Given that the fraction of single-strands is $1 - \theta_D$, K_d can be expressed with respect to the duplex fraction as

$$K_d = \frac{2\left[C_{Tot}\right](1-\theta_D)^2}{\theta_D} \tag{S4}$$

Rearranging this expression to solve for θ_D results in

$$\theta_D = \frac{4[C_{Tot}] + K_d - \sqrt{K_d^2 + 8K_d[C_{Tot}]}}{4[C_{Tot}]} \tag{S5}$$

In deriving an expression for the internal fraction, the many possible base pairing configurations that can contribute to the duplex state can be grouped into a series of substates D_i which differ by their total number of intact base pairs, i. If the fundamental rates k_f and k_b are considered as average values that are independent of the initial state D_i or base pair position in the sequence, we can then treat the distribution of population between these states as a series of equilibria governing the formation or disruption of single base pairs within the duplex.

$$\left[D_1\right] \underset{d_1 k_b}{\overset{d_1 k_f}{\rightleftharpoons}} \left[D_2\right] \underset{d_2 k_b}{\overset{d_2 k_f}{\rightleftharpoons}} \cdots \underset{d_{L-2} k_b}{\overset{d_{L-2} k_f}{\rightleftharpoons}} \left[D_{L-1}\right] \underset{d_{L-1} k_b}{\overset{d_{L-1} k_f}{\rightleftharpoons}} \left[D_L\right] \tag{S6}$$

where $[D_i]$ is the concentration of duplex species with i total intact base pairs and L is the total number of nucleotides in the DNA strand. As a result the total duplex concentration is

$$\left[D\right] = \sum_{i=1}^{L} \left[D_i\right] \tag{S7}$$

There are a number of possible ways in which to add or break a base pair starting from any one initial configuration. The degeneracy factor, d_i is a count of the number of possible ways to form an additional base pair from state i (or break a base pair from state $i + 1$)

$$d_i = \frac{L!}{i!(L-i-1)!} \tag{S8}$$

such that d_i is minimized at d_1 and d_L, and is maximized at $i = L/2$.

Infrared experiments are sensitive to changes in the overall fraction of opened/closed base pairing contacts within the duplex ensemble. Across all duplex species, the total concentration of opened, [o] and closed, [c] base pair sites is

$$\begin{aligned} [o] &= \sum_{i=1}^{L} (L-i)[D_i] \\ [c] &= \sum_{i=1}^{L} i[D_i] \end{aligned} \tag{S9}$$

Since the internal fraction is defined as the fraction of intact base pairs among all possible base pairs within duplexed DNA,

$$\theta_{int} = \theta_c = \frac{[c]}{L[D]} \tag{S10}$$

where $L[D] = [o] + [c]$ gives the total concentration of base pair sites within duplexes and θ_c is the fraction of closed base pair sites out of all possible sites. The fraction of closed base pair sites, θ_c can be expressed in terms of the ratio of opened to closed base pair sites, $Q_{oc} = [o]/[c]$.

$$\theta_c = \frac{1}{Q_{oc}+1} \tag{S11}$$

The melting curve is assumed to track the overall fraction of intact base pairs among all DNA oligonucleotides as a function of temperature, $\theta(T)$.[14] This quantity is determined by the product $\theta_{int}\theta_{ext}$ which we have shown is given by $\theta_c\theta_D$ within the context of the proposed scheme in Fig. 4 in the main text.

The ratio Q_{oc} can be expressed in terms of the equilibrium constant for the elementary reaction of base pair formation,

$$s = [D_{i+1}] / [D_i] \tag{S12}$$

by recognizing that any step, $[D_i]$ along the reaction in eq. S6 can be expressed in terms of the concentration of $[D_L]$ through,

$$\left[D_i\right] = s^{i-L}\left[D_L\right] \tag{S13}$$

Substituting eq. S13 into the expressions for $[o]$ and $[c]$ in eq. S9 above, the ratio of opened to closed base pairs within the duplex ensemble, Q_{oc} can be written in terms of the elementary base pairing equilibrium constant, s

$$Q_{oc} = \frac{\sum_{i=1}^{L}\left(L-i\right)s^{i-L}}{\sum_{i=1}^{L} i s^{i-L}} \tag{S14}$$

Recognizing that s equates to an internal equilibrium constant, K_{int} and that K_d equates to an external equilibrium constant, K_{ext}, the temperature dependence of each is determined by an internal/external standard free energy

$$K_i(T) = \exp\left[-\Delta G_i^{\circ} / RT\right] = \exp\left[-\Delta H_i^{\circ} / RT\right]\exp\left[\Delta S_i^{\circ} / R\right] \qquad (i = int, ext) \tag{S15}$$

where R is the ideal gas constant and the standard enthalpy ΔH_i° and entropy ΔS_i° are assumed to be temperature independent. In addition to these four thermodynamic parameters a scaling factor, A is introduced that accounts for the fact that the 2nd SVD component that contains the melting curve for each sequence has been arbitrarily normalized to the lowest temperature point. Fig. S2d shows fits of this model to the X = C and fC sequences. The internal and external fractions are plotted by the dashed and faded lines, respectively, with the fractions associated with X = C plotted in gold and the fractions associated with X = fC plotted in red. As can be seen, the product $A\theta_{int}\theta_{ext}$ represented by the solid lines fits the experimental melting curves well and suggests that the shape of the curve in both cases can be reasonably described by this model.

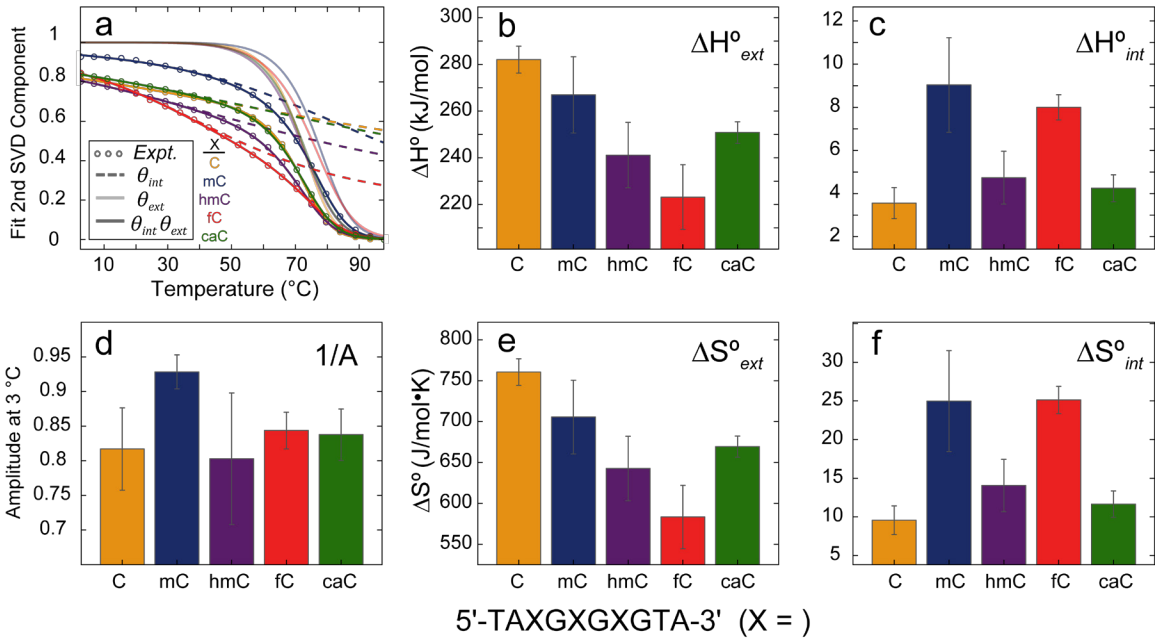


Figure S3: (a) Melting curves for each of the modified oligonucleotides fit to the product of $A\theta_{int}\theta_{ext}$. The experimental melting curves derived from the normalized 2nd SVD component are scaled by $1/A$ in the plot. The (b) external standard enthalpy, (c) internal standard enthalpy, (d) low temperature amplitude offset, (e) external standard entropy, and (f) internal standard entropy as a function of modification, X.

Applying this model to the melting curves measured across the set of modified sequences results in the external and internal thermodynamic parameters plus the amplitude offsets depicted in Fig. S3b-f. As can be seen in Fig S3a, the model describes the melting curves well regardless of the identity of X. The inverse of the amplitude offset, $1/A$, provides the proper normalization for the lowest temperature point (3 °C). Note that the commonly invoked all-or-none assumption in which the melting curve is assumed to range between 1 and 0 no longer applies. This standard melting curve normalization is rooted in the reasonable assumption that the duplex fraction for short oligonucleotides is 1 on the low temperature baseline and 0 on the high temperature baseline. However, when the melting curve is not strictly equivalent to θ_D but is instead better described by the product of an internal and external fraction of base pairing, it is inappropriate to apply the duplex fraction normalization to the entire melting curve. In other words, the assumption that θ_{int} corresponds to a value of 1 at the lowest temperature point sampled is arbitrary and unsupported by the experimental evidence. It is therefore necessary to include the amplitude offset A. Inspection

of the inverse amplitudes plotted in Fig. S3d reveals two distinct groupings within the confidence of the fit. The first includes X = C, hmC, fC, and caC and clusters around an average value of 0.83. In contrast, the X = mC sequence stands out from the remaining sequences with a low temperature offset of 0.93. In practice, we apply the mean low temperature scaling of 0.83 to the melting curves for all sequences except the X = mC sequence, which is scaled to a value of 0.93 at T = 3 °C. The amplitude scaled melting curves are then fit with respect to only the four thermodynamic parameters ΔH°_{ext}, ΔS°_{ext}, ΔH°_{int}, and ΔS°_{int}, as shown in Fig. 4 in the main text.

The melting curves in Fig. S3 are not baseline corrected in any way, and therefore any additional baseline artifacts such as those discussed above will be reflected in the fit to θ_{int}. However, as seen in Fig. S1b, these artifacts are small relative to the asymmetry imparted by θ_{int}. Furthermore one would anticipate baseline artifacts to be similar regardless of oligonucleotide sequence, suggesting any error introduced to θ_{int} across the oligonucleotide series represents a systematic offset that does not interfere with the ability to compare relative trends across the set of modifications.

It is important to note that θ_{int} should not be interpreted literally in the sense of a quantitative fraction of intact base pair sites among duplexes. The model describes internal base pairing in a discretized way such that any given site in the duplex can adopt only an opened or closed configuration. This binary picture is a simplifying assumption that captures the essential features of θ_{int}, but in reality base pairing is better described by a continuous coordinate, or set of coordinates, that tracks the loosening of hydrogen bonds, the disruption of base stacking, and changes in inter-base separation, not to mention the loss of base pairing contact that coincides with a fully extra-helical rotation of unpaired nucleotides. It is temperature dependent shifts in these continuous coordinates that are responsible for the changes to the FTIR spectrum along the low temperature baseline as well as the overall shape of the melting curve. A drop in the internal fraction with increasing temperature is therefore best interpreted as a general loosening of base pairing contacts, and θ_{int} can be thought of as reporting on the reduction in average base pairing character relative to the duplex ensemble measured at 3 °C.

IV. Discussion and Analysis of t-2D IR Spectra of the X = C and fC Sequences

Considering the 180 ns t-2D IR surface in Fig. 3d measured for the X = C sequence in more detail, reduced growth at 1575 and 1665 cm^{-1} relative to the intensity gain at 1620 cm^{-1} indicates

that the loss of T:A base pairing is a more significant contribution to the spectrum at 180 ns relative to 320 μs. This conclusion is also supported by comparatively large changes in the off-diagonal cross peak region between 1620-1690 cm^{-1}, which is most evident in the apparent reweighting of intensity on the 180 ns surface between the features highlighted by the purple and green arrows indicated on the spectrum in Fig. 3d. These features reflect a growth in intensity of thymine intramolecular cross peaks and further suggest loss of T:A base pairing.[18] Taken together, these observations indicate that the λ_{ns} process corresponds to a pre-dissociation reduction in base pairing character within the duplex with contributions from both T:A and C:G base pairs. However, the dominant contribution for the X = C sequence originates from fraying of the T:A termini, consistent with previous T-jump studies of canonical DNA oligonucleotides.[18-19] This assignment is also consistent with the characterization of the drop in internal base pairing fraction for this sequence as primarily originating from a loss of T:A base pairing, as discussed above.

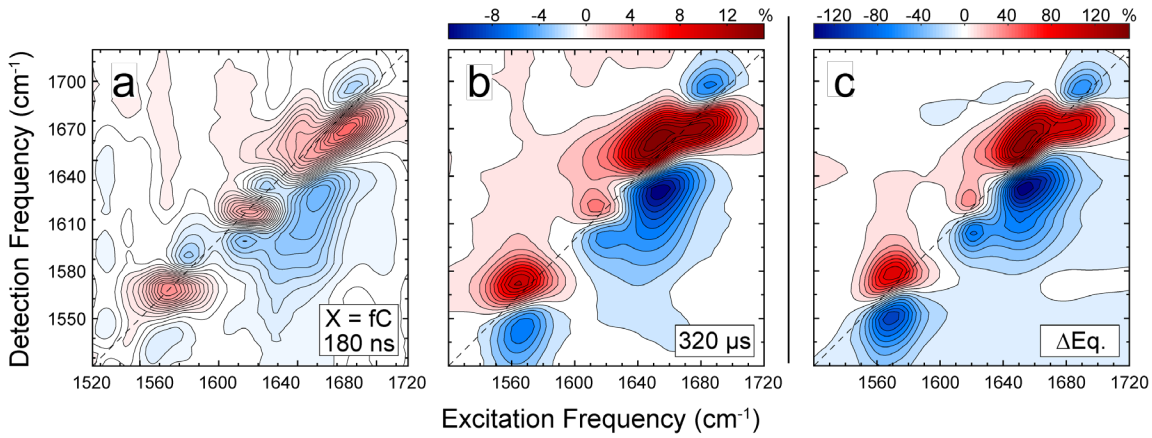


Figure S4: The (a) τ = 180 ns and (b) τ = 320 μs t-2D IR surfaces collected for the X = fC sequence. The waiting time was fixed at τ_2 = 150 fs and the polarization was set to ZZZZ for all measurements. For the t-2D IR measurements, T_i = 55 °C and ΔT = 15 °C. (c) Equilibrium difference spectrum between 10 and 90 °C.

The 180 ns t-2D IR surface measured for the X = fC sequence in Fig. S4a likewise shows evidence of a pre-dissociation reduction in T:A and X:G base pairing, but the contribution from each is somewhat more balanced, as evidenced by the comparatively flat response across the transient surface. On the whole, the X = fC 180 ns t-2D IR surface suggests a more complex ns response containing additional features not observed for the X = C sequence. For example, the G

ring mode absorptions near 1575 cm^{-1} show a loss towards higher frequency and a gain towards lower frequency, consistent with a shift of these absorptions to lower frequency along the diagonal. The 1685 cm^{-1} loss feature is more diagonally elongated when X = fC. This inhomogeneous broadening of the line shape suggests a more varied distribution of base pairing environments. Finally, a gain feature above the diagonal at excitation frequency 1630 cm^{-1} and detection frequency 1655 cm^{-1} corresponds to an intramolecular cross peak to an additional carbonyl mode that is unique to the fC nucleobase.[1] These additional features are only observed in the 180 ns t-2D IR spectrum. The ns timescales characteristic of these changes suggest rapid interconversion between the available base pairing conformations, consistent with increased base pair mobility and enhanced structural fluctuations in the duplex. Differences in the 320 μs surfaces as a function of X are far more subtle. As observed for the X = C sequence, the high and low temperature equilibrium difference spectrum and the 320 μs t-2D IR spectrum are in good agreement for the X = fC sequence as well (Fig. S4b,c). In conjunction, these observations indicate that base pairing contacts within the X = fC sequence are more varied and dynamic than in the canonical sequence. The conclusions drawn from t-2D IR are consistent with the interpretation of the FTIR melting experiments discussed above in which the X = fC sequence displays the sharpest reduction in internal base pairing fraction with increasing temperature and provide insight that this drop corresponds to increased base pair mobility within the duplex.

Although less informative overall, the absolute value 2D IR surface is useful for quantifying the intensity changes associated with the four prominent difference features discussed in the main text (Fig. 3c) since this representation of the data is universally positively signed, resulting in difference spectra where negative intensity unambiguously corresponds to loss and positive intensity to gain. Fig. S5a-c shows the absolute value representations of the equilibrium and difference spectra measured for the X = C sequence in Fig. 3a-c. The color coded arrows in Fig. S5c highlight the four prominent difference features discussed in the main text. The corresponding surfaces measured for the X = fC sequence are shown in Fig. S5d-f. To quantify the changes in signal observed at early versus later times discussed in the main text and plotted in Fig. 3f, we determine the ratio of integrated peak volumes on the absolute value t-2D IR surfaces measured at 180 ns and 320 μs ($\phi_{ns} = \Delta S_{ns} / \Delta S_{\mu s}$) for each of the four indicated difference features as well as the integration across the entire t-2D IR surface.

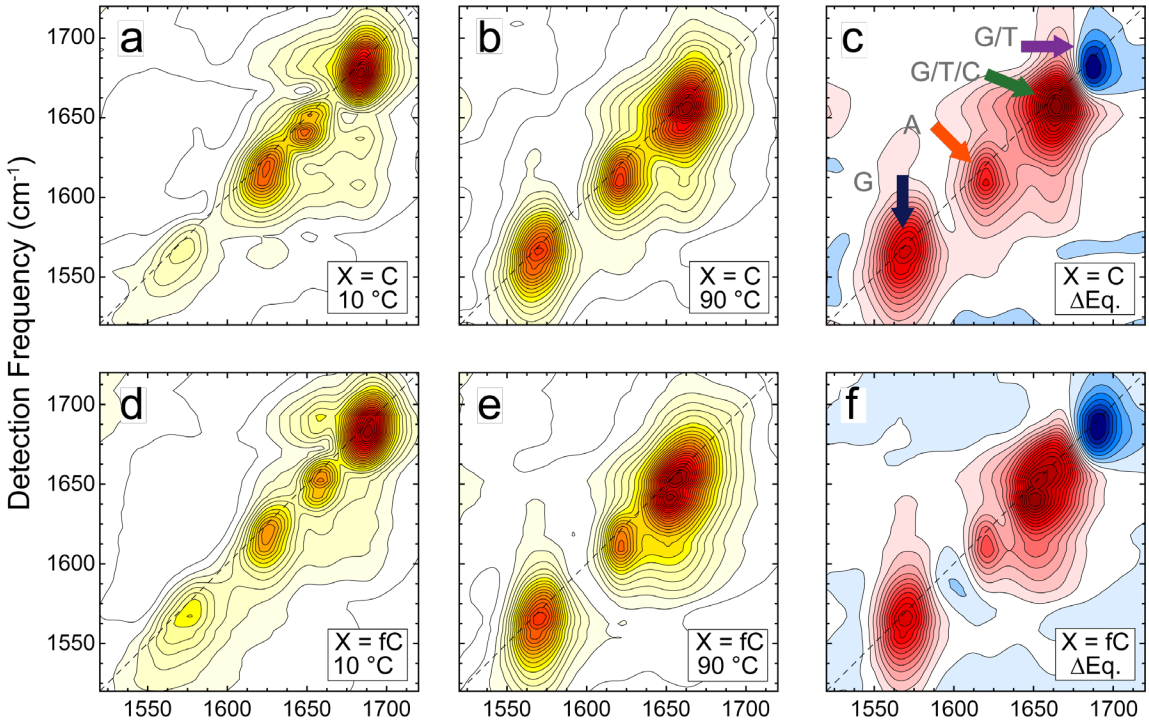


Figure S5: Absolute value equilibrium 2D IR spectra for the X = C sequence at (a) 10 °C and (b) 90 °C. (c) The absolute value difference spectrum between the 90 °C and 10 °C surfaces in panels (a) and (b). Color coded arrows highlight the same features discussed in the main text. (d-f) The corresponding surfaces for the X = fC sequence.

Fig. S6 shows the absolute value t-2D IR surfaces for both sequences at each delay. The integration boundaries for each feature are outlined by the color coded boxes in Fig. S6b. To compare the total change in transient signal to the pre-dissociation drop in base pairing suggested by the melting curves and quantified by the thermodynamic model, we calculate the percent change of the internal fraction relative to the overall difference in the melting curve ($\phi_{eq} = \Delta\theta_{int}/\Delta\theta$) across the temperature range from the T-jump. The similar behavior of these quantities supports the assignment of the reduction in θ_{int} to increasing base pair fluctuations within duplexed DNA and further suggests that the global model is a reasonable description.

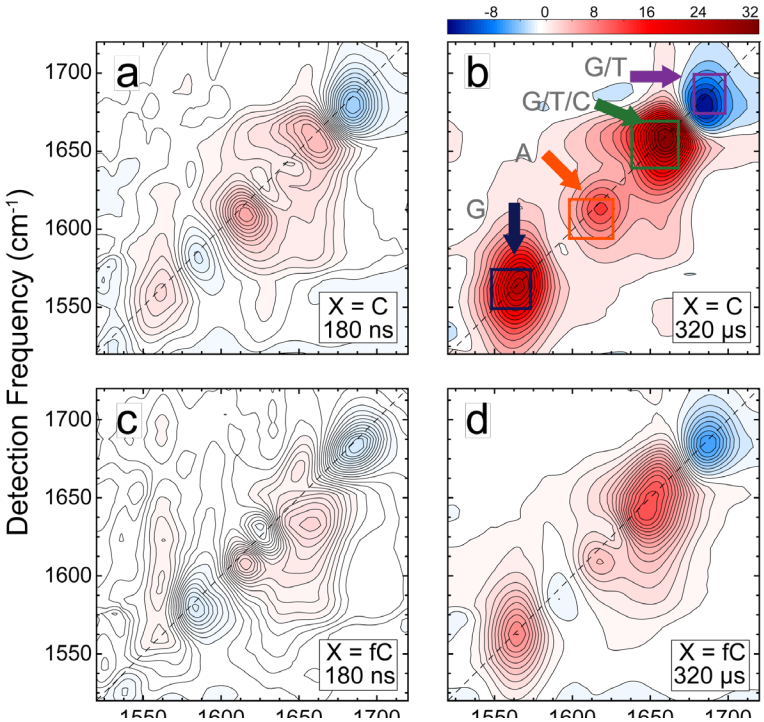


Figure S6: The (a) 180 ns and (b) 320 µs absolute value t-2D IR surfaces for the X = C sequence. Color coded boxes in panel (b) indicate the integration boundaries used to compute the ratio of integrated signal change at 180 ns versus 320 µs plotted in the bar graph in Fig. 3f in the main text. (c-d) The corresponding absolute value t-2D IR surfaces for the X = fC sequence.

V. Rate Domain Representation of t-HDVE Spectra and Determination of Observed Rates

Fig. S7 shows an illustrative set of rate spectra from the series of T-jump measurements on the canonical sequence outlined in Fig. 2a in the main text. A representative λ_{ns} and $\lambda_{\mu s}$ at each temperature is determined by taking the amplitude weighted mean across the series of rate peaks in the corresponding range. The ns range outlined in green in Fig. S7 has been scaled up by a factor of seven to highlight the comparatively low amplitude features associated with this response. While $\lambda_{\mu s}$ is observed to increase exponentially with increasing temperature, λ_{ns} does not appear to change meaningfully between T_f = 64-80 °C. Above T_f = 80 °C, the amplitude of the ns response diminishes to the point that the dominant feature in the ns range is the tail of the µs response. As a result, the λ_{ns} included in the prefactor of eq. 3 is taken to be the average across the four lowest temperature points.

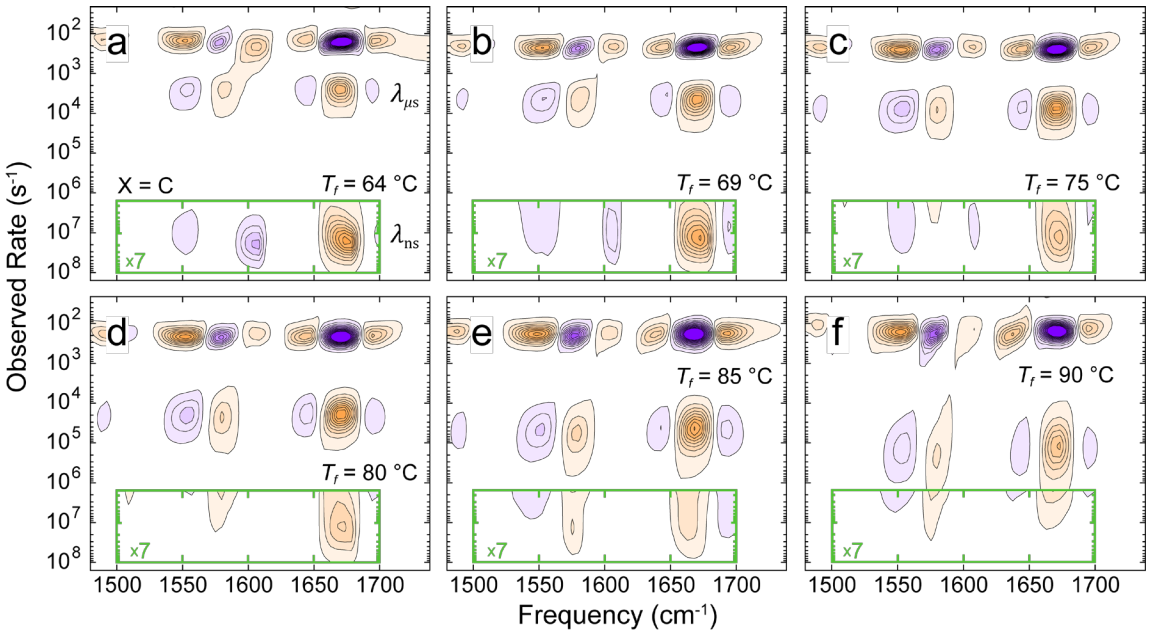


Figure S7: (a-f) The rate domain representation of the t-HDVE data collected across the duplex to single-strand transition for the X = C sequence. Orange represents positive amplitude while purple represents negative amplitude. The intensity inside the green box is multiplied by a factor of seven to highlight the response in the ns range.

VII. Modification and pH Dependent Stretched Exponential Kinetics

The steady-state and transient results presented in the main text suggest modification dependent and, for the X = caC sequence, pH dependent heterogeneity in local base pairing, the overall duplex state, and the hybridization transition state. As a check on the validity of the thermodynamic and kinetic framework proposed to account for our experimental observations, we return to the kinetic traces measured through t-HDVE to evaluate if the expected degree of disorder is reflected in the data independent of any explanatory model. Heterogeneity within the duplex and transition states is likely to translate to an increase in the variety of observed dissociation pathways. One might expect that this collection of pathways would skew asymmetrically towards faster rates of dehybridization assuming that increasing heterogeneity corresponds to a decrease in the average base pairing character within duplexes. Such a scenario would result in the observation of stretched exponential behavior in the T-jump data.[20-21]

Fig. S8a shows kinetic traces tracked at the maximum response of the t-HDVE spectrum acquired across the center of the melting transition for each of the modified oligonucleotides while

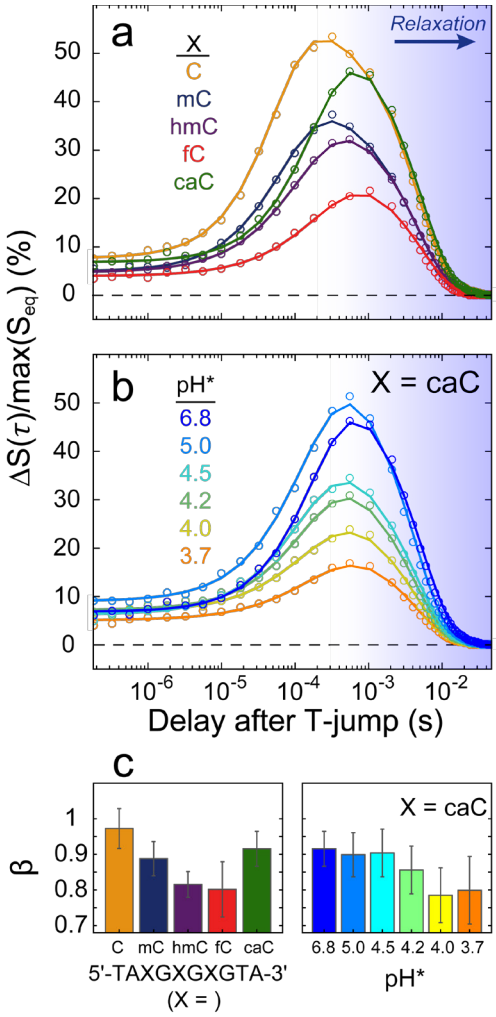


Figure S8: t-HDVE kinetic traces tracked at the maximum response of the T-jump across the center of the melting transition for (a) each of the modified oligonucleotides and (b) the series of pH* points measured for the X = caC sequence. (c) Trends in the stretching parameter from stretched exponential fits, $\Delta S / \max(S_{eq}) = B\exp[-(\lambda_{\mu s} t)^{\beta}]$ to the rise in each kinetic trace from panels (a) and (b). The observed rate, $\lambda_{\mu s}$ is read off from the rate domain representation of the data such that experimental traces are fit only with respect to the stretching parameter, β and amplitude, *B*. Error bars correspond to 95% confidence intervals from the fit.

Fig. S8b shows the corresponding traces acquired across the set of pH* conditions measured for the X = caC sequence. The rise of each trace is fit to a stretched exponential function, $\Delta S / \max(S_{eq}) = B\exp[-(\lambda_{\mu s} t)^{\beta}]$. Time points below 180 ns are truncated to exclude the λ_{ns}

response. To reduce the number of fit parameters, the average observed rate constant $\lambda_{\mu s}$ is read off from the rate domain representation of the data such that only an amplitude *B* and stretching parameter β are fit. The stretching parameter can range between 1 (pure exponential) and 0, with a decreasing value of β indicating increasingly stretched exponential behavior. In practice, the trace beyond 180 ns is simultaneously fit to the sum of an exponential rise and decay to account for thermal re-equilibration. Fig. S8c shows the modification and pH dependent trends in β for the signal rise. The X = C sequence exhibits the most purely exponential kinetics, with β = 0.97, while the X = hmC and fC kinetic traces are comparatively stretched, with β = 0.82 and 0.80. The X = mC and caC sequences are once again intermediate cases, with β = 0.89 and 0.92. Overall, the trend in the stretching parameter as a function of X mirrors the trends observed in both the external and dissociation activation parameters, consistent with the thermodynamic and kinetic framework invoked to describe the modification dependent results. Likewise, the trend in β observed for the X = caC sequence as a function of pH* is in agreement with the thermodynamic and kinetic interpretation of the pH dependent results, with the stretching parameter following a titration profile centered at the pK_a of the caC carboxyl group and dropping to a value similar to that of the X = fC sequence below pH* ~4.0.

VIII. Citations

1. Dai, Q.; Sanstead, P. J.; Peng, C. S.; Han, D.; He, C.; Tokmakoff, A., Weakened N3 Hydrogen Bonding by 5-Formylcytosine and 5-Carboxylcytosine Reduces their Base-Pairing Stability. *ACS Chem. Biol.* **2015,** *11*, 470-477.

2. Raiber, E.-A.; Murat, P.; Chirgadze, D. Y.; Beraldi, D.; Luisi, B. F.; Balasubramanian, S., 5-Formylcytosine Alters the Structure of the DNA Double Helix. *Nat. Struct. Mol. Biol.* **2015,** *22*, 44-49.

3. Sumino, M.; Ohkubo, A.; Taguchi, H.; Seio, K.; Sekine, M., Synthesis and Properties of Oligodeoxynucleotides Containing 5-Carboxy-2′-Deoxycytidines. *Bioorg. Med. Chem. Lett.* **2008,** *18*, 274-277.

4. Münzel, M.; Lischke, U.; Stathis, D.; Pfaffeneder, T.; Gnerlich, F. A.; Deiml, C. A.; Koch, S. C.; Karaghiosoff, K.; Carell, T., Improved Synthesis and Mutagenicity of Oligonucleotides Containing 5-Hydroxymethylcytosine, 5-Formylcytosine and 5-Carboxylcytosine. *Chem.: Euro. J.* **2011,** *17*, 13782-13788.

5. Thalhammer, A.; Hansen, A. S.; El-Sagheer, A. H.; Brown, T.; Schofield, C. J., Hydroxylation of Methylated CpG Dinucleotides Reverses Stabilisation of DNA Duplexes by Cytosine 5-Methylation. *Chem. Comm.* **2011,** *47*, 5325-5327.

6. Wanunu, M.; Cohen-Karni, D.; Johnson, R. R.; Fields, L.; Benner, J.; Peterman, N.; Zheng, Y.; Klein, M. L.; Drndic, M., Discrimination of Methylcytosine from Hydroxymethylcytosine in DNA Molecules. *J. Am. Chem. Soc.* **2010,** *133*, 486-492.

7. Szulik, M. W.; Pallan, P. S.; Nocek, B.; Voehler, M.; Banerjee, S.; Brooks, S.; Joachimiak, A.; Egli, M.; Eichman, B. F.; Stone, M. P., Differential Stabilities and Sequence-Dependent Base Pair Opening Dynamics of Watson–Crick Base Pairs with 5-Hydroxymethylcytosine, 5-Formylcytosine, or 5-Carboxylcytosine. *Biochemistry* **2015,** *54*, 1294-1305.

8. Owczarzy, R., Melting Temperatures of Nucleic Acids: Discrepancies in Analysis. *Biophys. Chem.* **2005,** *117*, 207-215.

9. Mergny, J.-L.; Lacroix, L., Analysis of Thermal Melting Curves. *Oligonucleotides* **2003,** *13*, 515-537.

10. Amunson, K. E.; Anderson, B. A.; Kubelka, J., Temperature Effects on the Optical Path Length of Infrared Liquid Transmission Cells. *Appl. Spectrosc.* **2011,** *65*, 1307-1313.

11. Marky, L. A.; Breslauer, K. J., Calculating Thermodynamic Data for Transitions of any Molecularity from Equilibrium Melting Curves. *Biopolymers* **1987,** *26*, 1601-1620.

12. Hardwick, J. S.; Lane, A. N.; Brown, T., Epigenetic Modifications of Cytosine: Biophysical Properties, Regulation, and Function in Mammalian DNA. *BioEssays* **2018,** *40,* 1700199.

13. Sanstead, P. J.; Tokmakoff, A., A Lattice Model for the Interpretation of Oligonucleotide Hybridization Experiments. *J. Chem. Phys.* **2019,** *150*, 185104.

14. Wartell, R. M.; Benight, A. S., Thermal Denaturation of DNA Molecules: A Comparison of Theory with Experiment. *Phys. Rep.* **1985,** *126*, 67-107.

15. Zimm, B. H.; Bragg, J., Theory of the Phase Transition between Helix and Random Coil in Polypeptide Chains. *J. Chem. Phys.* **1959,** *31*, 526-535.

16. Craig, M. E.; Crothers, D. M.; Doty, P., Relaxation Kinetics of Dimer Formation by Self Complementary Oligonucleotides. *J. Mol. Biol.* **1971,** *62*, 383-401.

17. Poland, D.; Scheraga, H. A., *Theory of Helix-Coil Transitions in Biopolymers*. Academic Press: New York, 1970.

18. Sanstead, P. J.; Stevenson, P.; Tokmakoff, A., Sequence-Dependent Mechanism of DNA Oligonucleotide Dehybridization Resolved through Infrared Spectroscopy. *J. Am. Chem. Soc.* **2016,** *138*, 11792-11801.

19. Sanstead, P. J.; Tokmakoff, A., Direct Observation of Activated Kinetics and Downhill Dynamics in DNA Dehybridization. *J. Phys. Chem. B* **2018,** *122*, 3088-3100.

20. Edholm, O.; Blomberg, C., Stretched Exponentials and Barrier Distributions. *Chem. Phys.* **2000,** *252*, 221-225.

21. Vlad, M. O.; Ross, J.; Huber, D. L., Linear Free Energy Relations and Reversible Stretched Exponential Kinetics in Systems with Static or Dynamical Disorder. *J. Phys. Chem. B* **1999,** *103*, 1563-1580.