Occurrence and Function of Hoogsteen Base Pairs in Nucleic Acids

by

Huiqing Zhou

Department of Biochemistry
Duke University

Date:_____

Approved:

_____
Hashim M. Al-Hashimi, Supervisor

_____
Jane S. Richardson

_____
Leonard D. Spicer

_____
Weitao Yang

Dissertation submitted in partial fulfillment of
the requirements for the degree of Doctor
of Philosophy in the Department of
Biochemistry in the Graduate School
of Duke University

2016

ABSTRACT

Occurrence and Function of Hoogsteen Base Pairs in Nucleic Acids

by

Huiqing Zhou

Department of Biochemistry
Duke University

Date:_____
Approved:

_____
Hashim M. Al-Hashimi, Supervisor

_____
Jane S. Richardson

_____
Leonard D. Spicer

_____
Weitao Yang

An abstract of a dissertation submitted in partial
fulfillment of the requirements for the degree
of Doctor of Philosophy in the Department of
Chemistry in the Graduate School of
Duke University

2016

Abstract

The double helix provides the structural basis for the storage and transfer of genetic information. The B-form double helix adopted by DNA can dynamically accommodate Watson-Crick and Hoogsteen base-pairing, through ≈180° flips of the purine base between *anti* and *syn* conformations. There is growing evidence that Hoogsteen base pairs play important roles in DNA replication, recognition, damage and mispair repair. This thesis examines the occurrence of Hoogsteen base pairs in DNA and RNA duplexes.

A structure-based survey of existing Hoogsteen base pairs in the Protein Data Bank revealed a diversity of Hoogsteen base pairing modes with stronger preferences for A–T versus G–C bps, TA versus GG steps, and also enrichment at terminal ends with a preference for 5'-purine. The survey also suggests that Hoogsteen base pairs induce a small but significant degree of DNA bending (~14°) directed toward the major groove.

As there were documented difficulties in modeling Hoogsteen versus Watson-Crick pairing by crystallography, we collaborated with the laboratories of Drs. Jane and David Richardson and identified potential Hoogsteen base pairs that were mis-modeled as Watson-Crick base pairs. These studies suggested that Hoogsteen base pairs are more prevalent than previously thought. We developed solution NMR method that relies on site-specific enrichment with $^{13}C/^{15}N$ labeled nucleotides to characterize stable or

transient Hoogsteen base pairs in large DNA-protein complexes under solution conditions. Application of this methodology to a complex formed between DNA and the Integration Host Factor protein reveals that the base pair at a sharply kinked site, which forms a Hoogsteen base pair based on X-ray crystallographic analysis, forms Watson-Crick base-pairing in solution with enhanced line-broadening; this result could indicate that there is potentially enhanced chemical exchange between Watson-Crick and Hoogsteen base pairs at the sharply kinked site with elevated Hoogsteen population (≈50%).

In stark contrast to B-form DNA, we found that Hoogsteen base pairs are strongly disfavored in A-form RNA duplexes. As a result, N^1-methyl adenosine and N^1-methyl guanosine, which occur in DNA as a form of alkylation damage, and in RNA as a posttranscriptional modification, have dramatically different consequences. They create G–C^+ and A–U Hoogsteen base pairs in duplex DNA that maintain the structural integrity of the double helix, but block base pairing all together and induce local duplex melting in RNA, providing a mechanism for potently disrupting RNA structure through posttranscriptional modifications. The markedly different propensities to form Hoogsteen base pairs in B-DNA and A-RNA may help meet the opposing requirements of maintaining genome stability on one hand, and dynamically modulating the structure of the epitranscriptome on the other.

Contents

List of Tables

List of Figures

List of Abbreviations

DNA – deoxyribonucleic acid

RNA – ribonucleic acid

WC –Watson-Crick

G – guanine

C – cytosine

A – adenine

T – thymine

U – uridine

bp(s) – base pair(s)

2′-OH – 2′-hydroxyl

HG – Hoogsteen

NMR – nuclear magnetic resonance

NOE – nuclear Overhauser effect

CSP – chemical shift perturbations

RD – relaxation dispersion

HSQC – heteronuclear single-quantum coherence spectroscopy

HMQC – heteronuclear multiple-quantum coherence spectroscopy

IHF – integration host factor

FP – fluorescence polarization

EMSA – electrophoretic mobility shift assay

m^1A – N^1-methylated adenosine

m^1G – N^1-methylated guanosine

UV – ultraviolet

1. Introduction

1.1 Historical account of Hoogsteen base pairs

1.1.1 Discovery of Hoogsteen base pairs

Back in 1953, Watson and Crick proposed the most well-known double helix structure for the deoxyribonucleic acid (DNA)[1]. The double helix structure model of DNA not only revealed the overall double-stranded, and helical appearance of the DNA molecule, but also provided the most fundamental structural feature in the central dogma: the specific Watson-Crick (WC) base-pairing[1] between the most plausible tautomeric forms of purine and pyrimidine nucleobases – guanine (G) with cytosine (C) and adenine (A) with thymine (T) – through formation of complementary hydrogen bonds (H-bonds) (Figure 1.1). This proposed structure and the base-pairing scheme fulfilled the overall shape and repeating pattern from the contemporary fiber diffraction data, explained the observation about the overall composition of DNA where A≈T and G≈C[2] and provided a structural basis for DNA self-replication and the potential role of DNA as the genetic information carrier[3]. Another important note from Watson and Crick is that: "It is probably impossible to build this structure with a ribose sugar in place of the deoxyribose, as the extra oxygen atom would make too close a van der Waals contact", which is probably the first note about the additional 2′-hydroxyl (2′-OH) group in the ribonucleic acid (RNA) that can lead to a different structure compared to DNA.

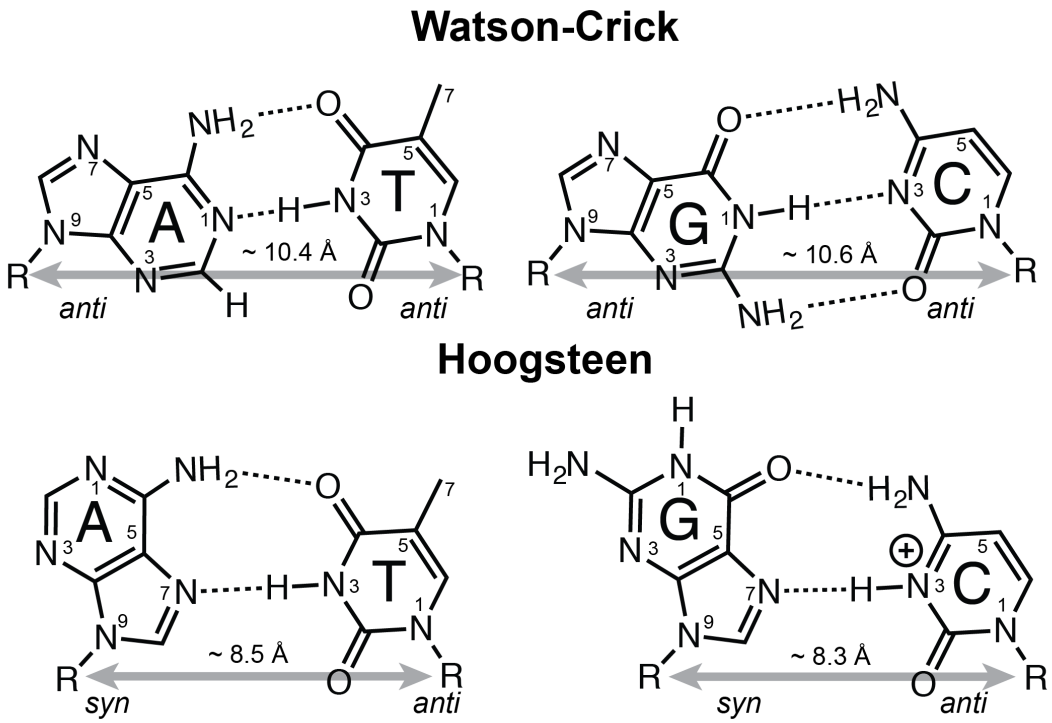

Figure 1.1: Chemical structures of WC and HG bps.

However, the data obtained by the fiber diffraction at that time were not nearly resolved enough to justify an atomic structure, as noted by Watson and Crick: "it (i.e. the structure) must be regarded as unproved until it has been checked against more exact results". Since the WC double helix model was proposed, there had been advances in the crystallography techniques that allowed solving single crystal structures of isolated nucleobase derivatives with atomic resolution. Strikingly, the first single crystal structure of the co-crystallized nucleobase derivatives (9-methyladenine and 1-methylthymine), reported by Dr. Karst Hoogsteen in 1959, presented a markedly different form (referred to now as the "Hoogsteen base pair") rather than WC base-pairing. In Hoogsteen (HG) base pairs (bps), both adenine and thymine bases remained their most probable tautomeric form as in WC bps, but had altered H-bonding pattern where the A-N7---HN3-T H-bond formed, replacing A-N1---HN3-T in A–T WC[4]. Hoogsteen also proposed G–C^+ HG base-pairing which required protonation of cytosine-N3 or guanine-N7 to form a second H-bond (Figure 1.1). He argued that it was more plausible for C-N3 to get protonated so as to preserve the base-pairing specificity; in other words, if G-N7 gets protonated, G^+–A can easily form a mispair that resembles A–T or G–C geometry in the helix[5]. The H-bonding geometry of HG bps brings the two paired bases into closer proximity compared to WC geometry, resulting in a constriction of C1′–C1′ distance by ≈2.5 Å[5]. To form HG bps in the antiparallel duplexes, the purine

base would be flipped around the glycosidic bond by ≈180°, resulting in the glycosidic bond angle (χ) adopting a *syn* rather than *anti* conformation in WC bps.

As the first experimental evidence of an isolated bp configuration with atomic resolution, the observation of HG base-pairing heated up the skepticism as to the correctness of the WC base-pairing scheme and the proposed double helix structure[6,7]. Several subsequent attempts were made to obtain co-crystals of adenine and thymine (or uridine) derivatives, all of which failed to yield WC bps but favored HG bps instead[8-10]. In sharp contrast to A–T HG bps, single crystal structures of 9-ethylguanine and 1-methylcytosine or 1-methyl-5-bromocytosine co-crystals and other derivatives revealed WC base-pairing[11]. As predicted earlier by Pauling and his colleagues[12] that G–C bps were stabilized by three and not two hydrogen bonds as proposed by Watson and Crick[1,3]. Arnott and Wilkins re-analyzed the diffraction data using HG bps as a model. They concluded that although it was possible for HG bps to be in a right-handed double helix, the HG structure did not maintain the structural regularity as well as the WC model due to loss of linearity in HG H-bonds[13]; and the structure with a mix of A–T HG and G–C WC bps did not satisfy the regularity in the diffraction pattern. During this time, other possible helical structures were also proposed that raised doubt on the WC double helix model, including the parallel duplex rather than *anti*-parallel with reverse A–T and G–C WC bps proposed by Donohue[14] and the single crystal structure of a left-

handed Z-form DNA with short CG repeats sequence reported by Rich and co-workers[15].

The controversy was partially resolved when Rich and colleagues reported the single crystal X-ray structure of the AU and GC dinucleoside phosphates in 1973[16,17] and both structures verified the antiparallel, right-handed double helix feature and WC base-pairing scheme proposed by Watson and Crick. The mystery of the DNA structure was not settled until 1980 when Drew, Dickerson and co-workers solved the single crystal structure of a DNA dodecamer containing both A–T and G–C WC bps using heavy atom X-ray crystallography[18,19]. During the 1970s, there were solution studies by nuclear magnetic resonance (NMR) spectroscopy[20-25] that indicated both G–C and A–T/U could form WC bps in double and triple helices in tRNA and polynucleotide complexes in the absence of the crystallographic conditions. Taken together the crystallographic and solution NMR studies, it has been widely accepted until now that WC bps are the most prevalent bps observed in nucleic acids, and the specific base-pairing scheme provides the structural basis for the template-based biological transactions: DNA replication, transcription and translation.

To be clear, the dominance of WC base-pairing does not preclude the co-existence of HG bps. Indeed, HG bps do represent an alternative pairing scheme that can expand the structural and functional versatility of duplex DNA beyond that which can be achieved based only with WC base-pairing. HG bps can form in not only triplexes

and quadruplexes but also in canonical duplex DNA, including in naked DNA duplexes with AT-rich sequences, DNA in complex with quinoxaline bis-intercalators and proteins[26]. Recently, our lab reported transient HG bps exist in canonical DNA duplexes in solution with low populations (≈ 1%) and short life time (≈1 ms). In stark contrast, little has been found in literature about HG formation in the A-form RNA duplex, which is known to be more compressed and rigid. Formation of HG can be an intrinsic difference between DNA and RNA duplexes. This chapter will provide an overview of the occurrence of HG bps in DNA and characterization techniques; with a focus on solution NMR.

1.1.2 Hoogsteen base pairs in naked duplexes

In the early 1980s, models were put forward for a Z-DNA structure that are exclusively comprised of HG bps[27], particularly for AT-rich sequences that frequently exhibited unusual diffraction patterns when dried (referred to as D- or E- type X-ray diffraction patterns)[28-30]. This form of DNA required helical structures with 7–7.5 bps per turn, which cannot be stereo chemically achieved by right-handed B-form DNA. Spectroscopic studies of poly(rA)-poly(rU) sequences that bear substituents at the adenine C2 position that sterically block WC base-pairing also suggested formation of duplexes with parallel or anti-parallel chain polarity, in which strands are held together by HG or reverse HG base-pairing, respectively[31-33].

The ability to prepare large quantities of highly pure DNA samples in a facile manner, in parallel with developments in $^{13}C/^{15}N$ isotopic enrichment and solution state NMR spectroscopy of nucleic acids, resulted in the high-resolution X-ray and NMR structure determination of diverse DNA sequences in the 1980s and 1990s showing WC B-form DNA duplexes. However, spectroscopic evidence for HG bps continued to mount in the 1990s and 2000s in the context of A-T rich sequences[34], in poly(dG-dC)-poly(dG-dC) sequences at low pH as possible intermediates along the B-to-Z DNA transition[35], as well as in non-canonical DNA regions such as closing bps of apical loops[36,37]. But it was not until 2002, when Subirana and colleagues reported the first crystal structure of an AT-repeat not capped by GC bps, that the first single crystal X-ray structure of a naked DNA duplex containing exclusively HG bps was resolved[38]. The structure of $d(AT)_3$ revealed an *anti*-parallel right-handed double helix made up exclusively of HG bps with an overall structure similar to that of B-form DNA (Figure 1.2).[38,39] Key differences included a change in the position of the helical axis relative to the bps, reduction in helical radius and C1′–C1′ by ≈2.5–3.0 Å, altered hydrogen bonding donor/acceptor pattern in the major and minor grooves, narrower and less electronegative minor groove; which favors hydrophobic interactions and distinct helix stacking and hydration patterns relative to B-DNA. Together, these features provide a distinct physicochemical presentation of the genetic code for potential sequence-specific recognition by the cellular machinery. Similar HG structures were subsequently

reported for related sequences d(ATATATCT)[40], d(CGATATATATAT)[41] and in a slightly different AT-rich sequence d(ATTAAT)[42].

It is important to note that there have not yet been any naked AT-repeats DNA oligonucleotides that crystallized as a WC helix; in contrast, solution state NMR studies of the above DNA sequences in aqueous solution[39] or under the same conditions used to grow crystals argued against formation of an HG helix (H.Z. *et al.*, unpublished data). Instead, the solution NMR studies suggest that the AT-repeats DNA duplexes are in favor of prototypical WC B-form double helices while the duplexes are highly unstable in solution that feature a melting temperature nearly below 5°C. This indicates that though crystal packing could play an important role in stabilizing the HG double helix, one cannot exclude the possibility that WC to HG transition can occur more readily and frequently in AT-repeat sequences compared to other sequences.

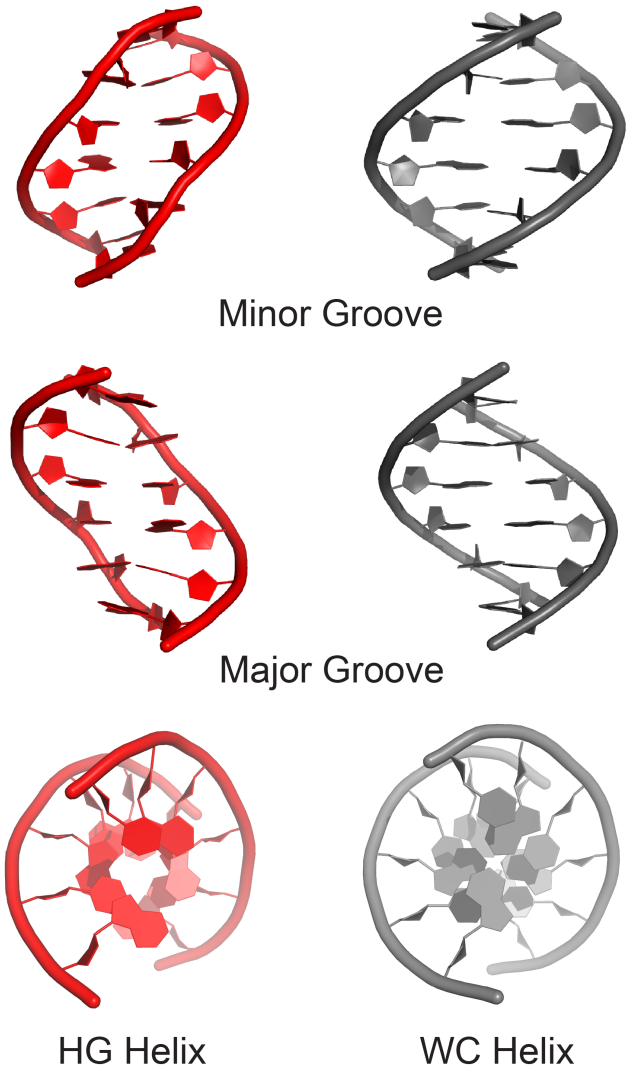

Figure 1.2: Comparison of B-form-like HG helix and WC helix.

1.1.3 Hoogsteen base pairs in DNA-antibiotic complexes

In 1984, Rich and colleagues reported the single crystal X-ray structure of a DNA double helix with sequence d(CGTACG) bound to triostin A[43]- a cyclic octadepsipeptide anti-tumor antibiotic containing two quinoxaline rings that binds DNA and inhibits replication and transcription *in vivo* (Figure 1.3)[44,45]. This was the first structure of a peptide antibiotic in complex with an oligonucleotide. The structure showed that the two quinoxaline rings bis-intercalate in the minor groove of the DNA double helix and surround the WC G–C bps, disrupting stacking interactions to the central A-T bps (Figure 1.3). Remarkably, although the two central A–T bps are not covered by the two triostin A molecules, they form HG rather than the WC bps. This marked the first crystallographic observation of the co-existence of WC and HG bps within the same duplex. No direct contacts are observed between the antibiotic and the exposed WC face of the A–T bases. Rather, the helical constriction at the HG bps appears to stabilize the complex by allowing close packing of the oligonucleotide around the end of the triostin A. Thus, several favorable van der Waals contact would be lost if the deoxyribose rings were further apart, as in WC bps. Similar structures were subsequently reported for DNA bound to the related echinomycin antibiotic[46], and for triostin A bound to d(GCGTACGC)[47], which featured two central A-T HG bps, and two terminal G–C^+ HG bps, marking the first crystallographic observation of protonated G–C^+ HG bps within a duplex (Figure 1.3).

Soon after, chemical footprinting studies preformed in solution showed that sites that form HG bps in X-ray structures of DNA-echinomycin complexes are hyperreactive to diethyl pyrocarbonate (DEPC)[48], which preferentially reacts with exposed N7 atoms of *syn* purines in non-canonical Z-DNA[49] and cruciform loops[50]. However, these results were challenged by footprinting studies employing DEPC and other reagents that target thymines, that showed little change in thymine chemical reactivity when replacing adenine with 7-deazaadenine, which has a diminished ability to form HG bps[51-54]. Moreover, oligonucleotides containing 7-deazaadenine and 7-deazaguanine bound echinomycin with affinity comparable to that of their unmodified counterparts, suggesting that HG bps are not essential for binding[54]. These studies argued that hyperreactivity does not arise from formation of HG bps but rather from unwinding and extension of the DNA helix upon drug binding.

Subsequent NMR studies by Feigon, Patel, and their co-workers confirmed formation of HG bps in DNA-antibiotic complexes[55,56], although their occurrence was shown to be highly dependent on sequence, temperature, and pH[57-59]. A-T bps generally form WC, and if they ever form HG, they do so transiently at physiological temperatures. Even the terminal HG A-T bps were only favored in DNA-antibiotic complexes having purine 5′ and pyrimidine 3′ to CG (i.e. ACGT, GCGC) and only at low pH for G–C^+ HG bps.[55,57,60,61]

Despite many studies, to date, it remains unclear whether quinoxaline antibiotics stabilize HG bps DNA bps *in vivo* and whether this is related in any way to their biological activity.

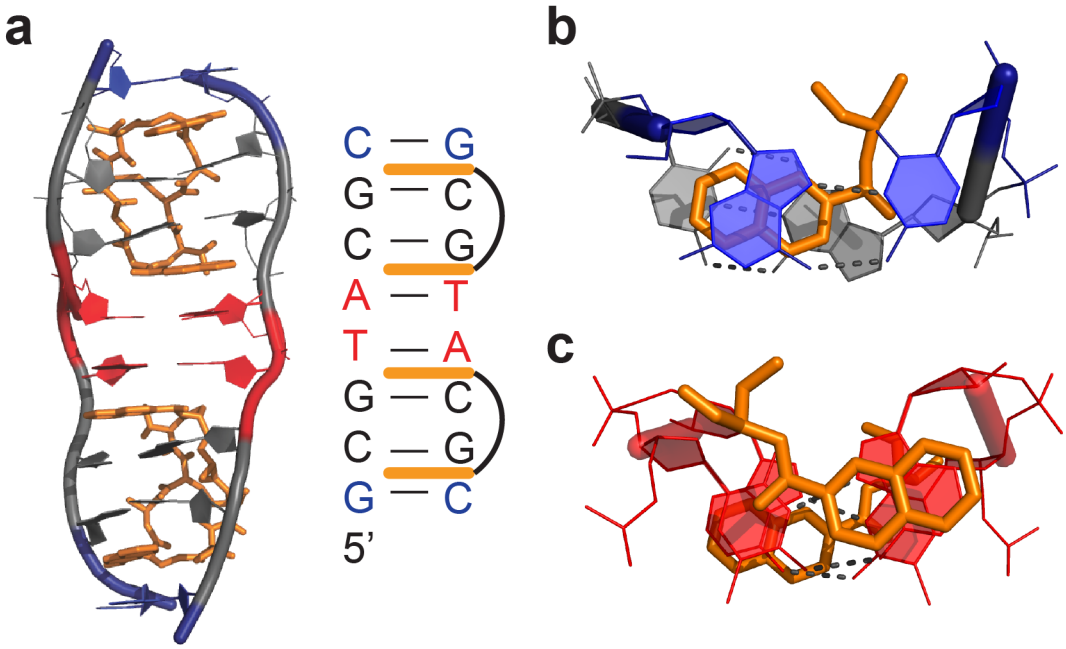

Figure 1.3: HG bps in DNA-quinoxaline bis-intercalator complex

(a) X-ray structure of DNA in complex with triostin A (PDBID: 1VS2). Zoom in view of (b) the terminal G–C+ HG bp and (c) helical A–T HG bps in the complex.

1.1.4 Hoogsteen base pairs in DNA-Protein complexes

In the late 1990s, X-ray structures emerged showing that certain proteins bind and in some cases specifically recognize HG bps embedded in B-form DNA (Figure 1.4). These studies raised the possibility that proteins exploit the unique structural and chemical features of HG bps in sequence-specific DNA recognition, and therefore, provided evidence for a functional role for HG bps *in vivo*.

The first crystallographic observation of HG bps in a protein-DNA complex was reported by Rice *et al.* in 1996[62] who visualized a single A-T HG bp immediately adjacent to a nicked site in the X-ray structure of a highly bent (>160 degrees) 35 bp DNA bound to the integration host factor (IHF) protein (Figure 1.4). Interestingly, a hydrogen bond was observed between the backbone amide group of an arginine residue and N3 of the *syn* A, suggesting specific recognition of the WC face in the HG bp. However, the nick is involved in crystal packing with a neighboring molecule in the complex and HG formation helps move the phosphate backbone away from a neighboring molecule. In addition, the protein makes specific contacts with N3 of an *anti* A in a symmetric site in the DNA lacking the nick, suggesting interactions that are specific for WC rather than HG base-pairing. Moreover, NMR studies of IHF binding to a shorter recognition sequence containing the first nicked site argue against the presence of an A–T HG bp in solution[63].

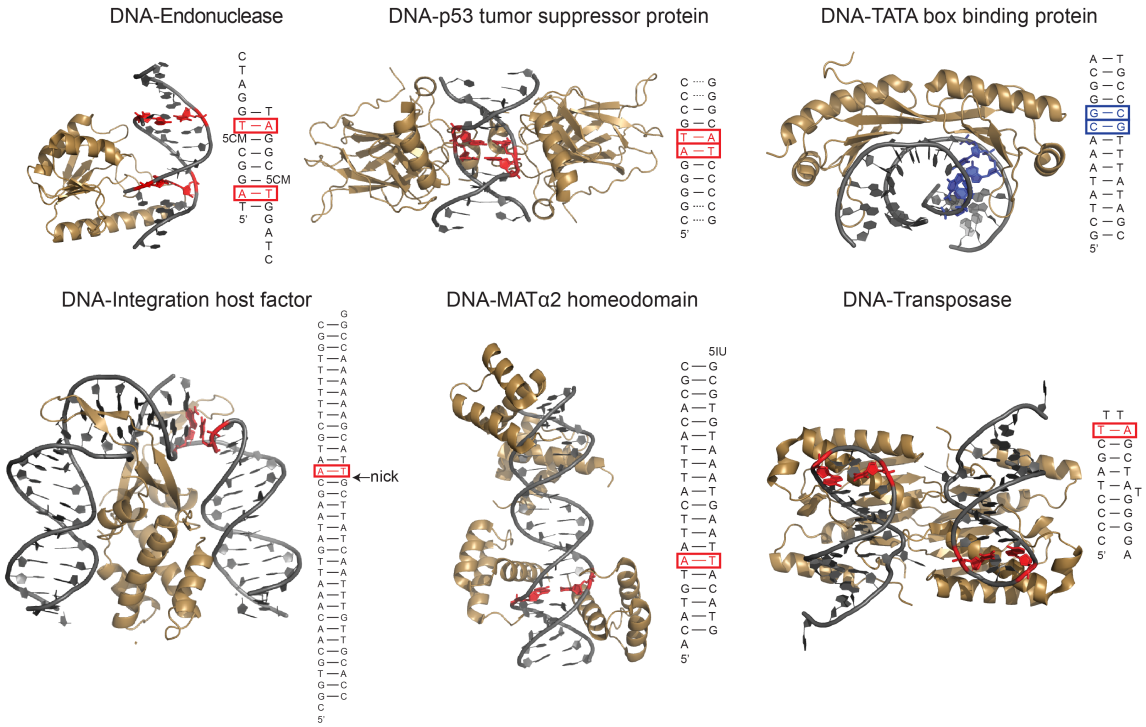

Figure 1.4: HG bps in DNA-protein complexes

upper left to right PDBIDs: 1ODG, 3IGL, 1QN3; lower left to right PDBIDs: 1IHF, 1K61, 2A6O

Subsequent X-ray structures of TATA elements bound to the TATA box-binding protein (TBP) revealed a G–C$^+$ HG bp in the mutant TATAAAC box in a region of DNA unwinding and intercalation.[64] No direct contacts were observed between the *syn* guanine base and the protein. However, the HG bp appears to contribute to binding by preventing steric clashes between the protein leucine 72 and the guanine exocyclic NH_2, while still preserving favorable van der Waals contacts with two neighboring phenylalanine residues. A second G–C$^+$ HG bp was observed but attributed to crystal packing forces. Interestingly, the ≈150-fold weaker binding affinity observed for TBP to this mutant TATA box,[65] which could be correlated to the selection of a transient HG over a WC bp at that site,[66] has been implicated in the transcriptional regulation of the human osteocalcin gene.[67] This observation suggests a biological role for the formation of a G–C$^+$ HG bp at the mutant promoter site.

Both IHF and TBP induce large distortions in the DNA, which could facilitate formation of HG bps. In contrast, Wolberger and colleagues observed a single A-T HG bp within an otherwise undistorted B-form WC duplex in the X-ray structure of MATα2 homeodomain non-specifically bound to DNA.[68] Van der Waals contacts were observed between an arginine side chain and the *syn* adenine base as well as the sugar-phosphate backbone of the adenine and the neighboring thymine. Once again, the HG bp appears to avoid unfavorable steric clashes that would otherwise arise with a WC bp. The HG bp is accommodated within the duplex DNA without inducing major distortions, even for

the directly neighboring bps. The ease with which HG bps could seamlessly fit within B-DNA raised the possibility that HG bps may have been incorrectly assigned to be WC bps due to misinterpretation of ambiguous electron density at medium to low resolution.[68]

Two neighboring A-T HG bps were subsequently observed in structures of a palindromic CATG/CATG sequence bound to the DNA binding domain of p53.[69] Although no direct contacts are observed with the *syn* adenines, the formation of the HG bps results in a narrowed minor groove in the region flanking the CATG site, leading to enhanced negative electrostatic potential that is further stabilized by insertion of the positively charged arginine side chains. Remarkably, these HG bps adopt WC conformation in X-ray structures with a longer spacer length[69] or a different intervening sequence[70,71] between DNA half-sites, which is accompanied by a different organization between p53 dimers, altered DNA helix conformation, and that also yield different DNA-tetramer binding affinities.[69] These studies suggest that WC and HG bps likely exist in equilibrium with each other and that their selection in DNA-p53 complexes is largely dictated by the nature of the DNA binding sequence.

1.1.5 Hoogsteen base pairs in damaged DNA

By the 1960s, it had become clear that DNA could be damaged by external and endogenous factors and that this in turn may be linked to disease states such as cancer.[72]

During the 1970s and 1980s, enzymes that recognize and repair damaged DNA began to be uncovered, resulting in great interest in characterizing the structure of damaged DNA.[73] These studies showed that HG base pairing provides an important mechanism for stacking and hydrogen bonding, in cases where the WC face of the purine bases are damaged, preventing favorable WC base-pairing.

Patel and co-workers showing that guanine adducts in the WC edge or the C8 position strongly favor a syn base orientation reported the first evidence for HG-type bps in damaged DNA in the late 1980s in solution NMR studies.[74-77] Subsequent NMR studies showed HG-type pairing in various purine lesions, including WC face alkylation adducts (e.g. 1,N2-propanoguanine[78,79] and 1,N2-ethenoguanine[80]), the bulky guanine C8 mutagenic adduct aminofluorine-C8-guanine[77,81], and the common mutagenic lesion N1-methyladenine (Figure 1.5).[82] The direct observation of HG base-pairing (rather than extra helical states) in a wide variety of lesions in naked DNA in the 1990s and 2000s established HG bps as an energetically closer alternative to WC bps.

There is great speculation and experimental evidence that HG-type pairs play important roles in DNA damage and mismatch repair. For example, it is likely that the enzyme AlkB, which repairs the mutagenic lesion N1-methyladenine, initially recognizes the HG bp between N1-methyladenine and thymine (Figure 1.5)[82,83] before flipping out the damaged purine for oxidative demethylation. The flipping of one purine base to a *syn* conformation is also often observed in purine-purine mismatches, where

the *syn-anti* bp configuration affords a shorter helical radius that can be more readily accommodated within B-DNA as compared to the *anti-anti* configuration. There is X-ray structural evidence that the DNA mismatch repair enzyme MutS specifically recognizes HG type purine-purine and purine-pyrimidine mismatches, even though they may not be the dominant conformation in unbound DNA, by making specific hydrophobic and hydrogen bonding minor groove contacts with the *syn* adenine/guanine base in A–C, A–A, and G–G mismatches.[84] The recognition of the increased population of *syn-anti* rather than *anti-anti* configuration in certain mismatched bps may help the enzyme discriminate against undamaged *anti-anti* WC bps. Thus, HG bps not only provide a mechanism for maintaining the overall structural integrity of damaged or incorrectly replicated DNA, they can play an important role in DNA repair mechanisms.

It is noteworthy that HG bps have also been observed in DNA containing non-natural modifications in the sugar-phosphate backbone, including the addition of an ethylene bridge between C3'and C5' in bicyclo-DNA, which fixes the gamma backbone torsion angle to a non-canonical orientation,[85] a single-residue substitution of sugar O4' with a methylene group,[86] or in dinucleotide d(TA) analogs containing a nonionic diisopropylsilyl-modified backbone at very low temperatures.[87]

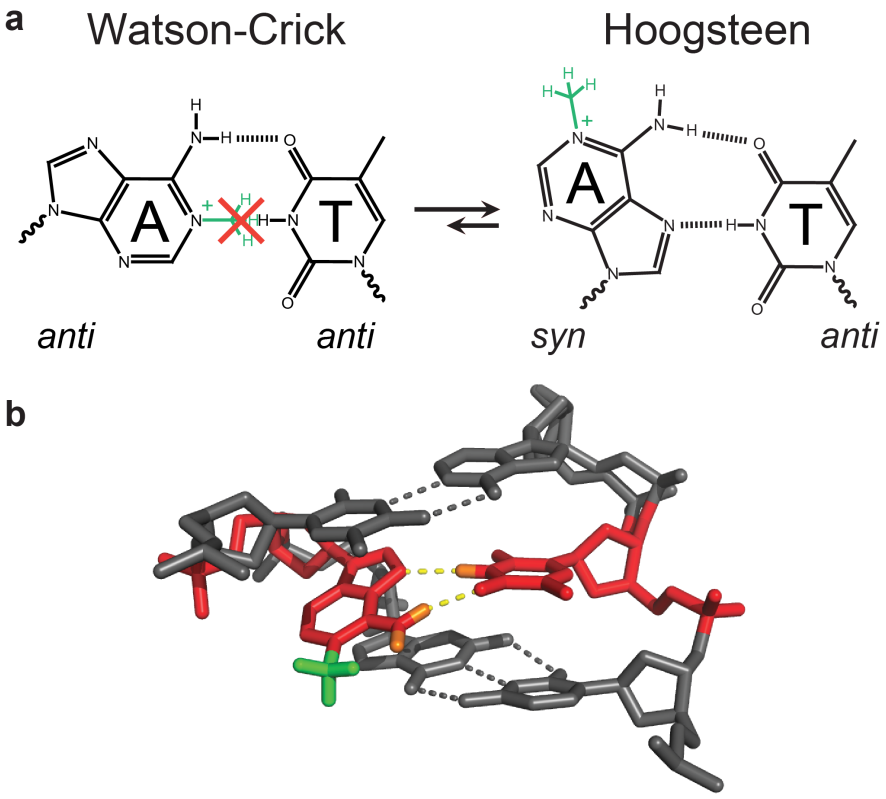

Figure 1.5: m^{1}A–T HG bp in DNA.

(a) Chemical structure of m^{1}A–T HG and WC bps. (b) m^{1}A–T HG bp in the structure of damaged DNA in complex with the AlkBH2 repair enzyme (PDBID: 3H8O).

1.1.6 Hoogsteen base pairs in DNA replication

WC bps were the most important aspect of the DNA double helix structure because, as succinctly stated in the very last sentence of Watson and Crick's 1953 Nature paper, "it immediately suggests a possible copying mechanism for the genetic material."[3] Four years later, Kornberg discovered the enzyme that catalyzed template DNA replication[88] and ensuing biochemical and structural studies established that DNA polymerases replicate DNA by WC pairing dNTP with the template strand.

During the 1990s, studies revealed that certain families of DNA polymerases (the X and Y families) contributed to damage-induced mutagenesis. It was later shown that some members of the Y family efficiently bypass DNA damage by replicating the template DNA via HG rather than WC base pairing. HG-based replication was first visualized in X-ray structures of an archaeal DNA Polη homolog, Dpo4, by Yang and colleagues nearly a decade ago.[89] The structure showed that Dpo4 replicates UV cross-linked thymine dimers by forming a HG bp between the 5′ thymine and an incoming dATP, thus avoiding backbone distortion and allowing discrimination against guanine and pyrimidines.[89]

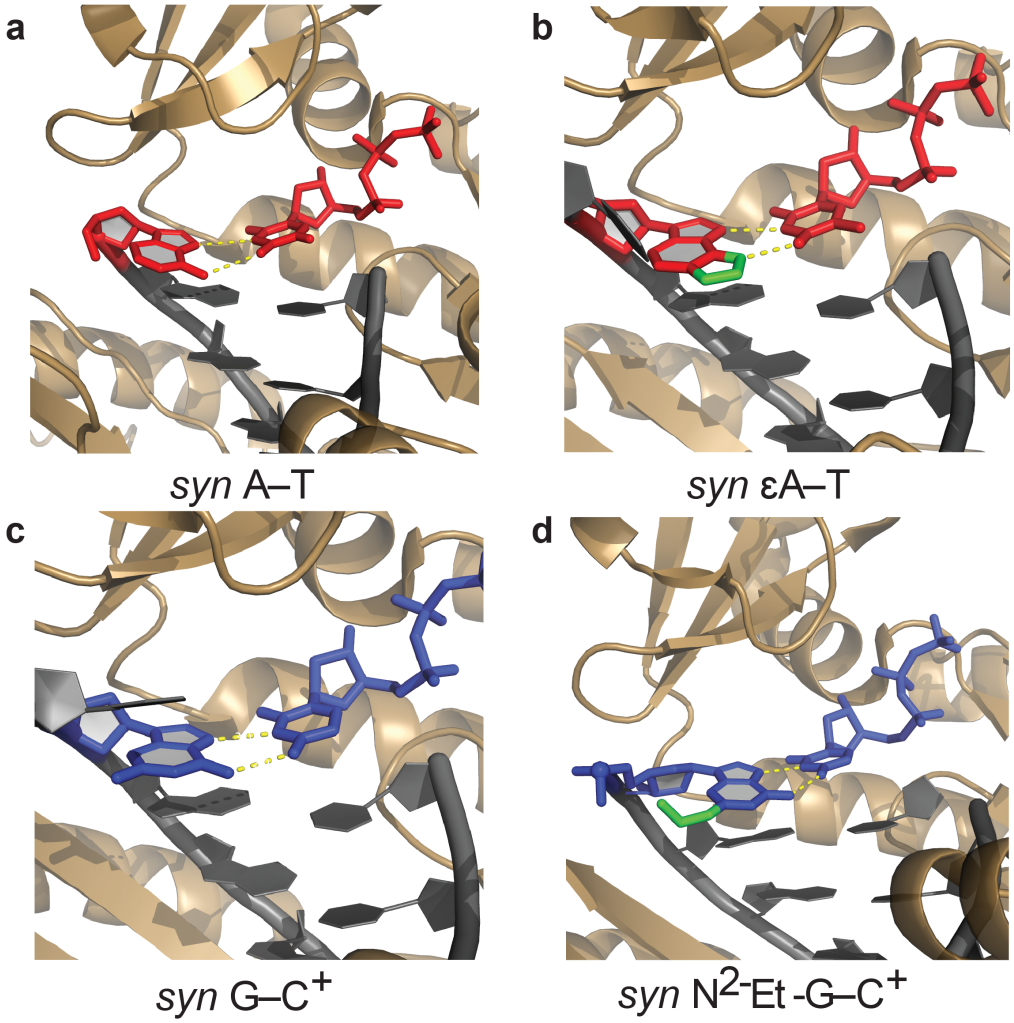

Figure 1.6: HG bps at replication active sites with Pol iota.

(a) a template A and an incoming dTTP (PDB ID: 1T3N), (b) a template 1,N6-ethenoadenine (eA) and an incoming dTTP (PDB ID: 2DPJ), (c) a template G and an incoming dCTP (PDB ID: 2ALZ), and (d) a template N2-ethylguanine (N2-Et-G) and an incoming dCTP (PDB ID: 3EPG). The damaged sites in eA and N2-Et-G are highlighted in green.

Aggarwal and co-workers subsequently showed using X-ray crystallography and biochemical experiments that another member of this family, DNA Polι, employs HG base-paring as a general mechanism to replicate both damaged and undamaged DNA.[90] A striking X-ray structure of Polι showed a template adenine in the active site of the enzyme adopting a *syn* conformation and forming an HG bp with an incoming dTTP (Figure 1.6).[90] The ability to insert the correct nucleotide across an adenine base also provided a rationale for prior biochemical studies showing a much higher efficiency of correct base incorporation across a templating adenine than across a templating thymine, which in fact favors G misincorporation because of its high propensity for forming of *anti*-G–T wobble bps.[91] This raised HG bps to a prominent position reserved previously only for WC bps; they provided a basis for copying DNA.

The proposal that Polι replicates DNA via HG base-pairing was quickly met with skepticism. In an accompanying News and View article, Wang[92] pointed out that based on the weak electron density for the active site A–T bp, it is difficult to resolve a WC from a HG conformation. He also questioned the ability of such a polymerase to form protonated G–C$^+$ bps at physiological pH, given the low intrinsic pK_a of cytosine N3 ≈ 4.2 – 4.4[93]. Aggarwal *et al.* later put the matter to rest by (i) solving X-ray structures of Polι unambiguously showing a protonated G–C$^+$ HG bp at pH 6.5, reinforcing their hypothesis that Polι has evolved to favor HG base-pairing by constraining the backbone C1′-C1′ distance between template and incoming nucleotide[94] and (ii) showing selective

inhibition of DNA synthesis by Polι but not by other polymerases when using 7-deazaadenine or 7-deazaguanine, which are incapable of forming HG base-pairing, as the templating residue.[95] Several other structures capturing DNA synthesis by Polι followed, ultimately demonstrating that major purine alkylation and oxidation lesions, including 1,N6-ethenoadenine,[96] N2-ethylguanine,[97] O6-methylguanine,[98] 8-oxoguanine,[99] adopted a *syn* conformation and, where possible, formed HG type bps with incoming complementary pyrimidine and purine nucleotides (Figure 1.7) (reviewed in Makarova *et al*.[100]). These observations in conjunction with biological studies showing that Polι was important for cell survival in the presence of alkylating agents[101,102] and oxidative stress,[103] provide the most compelling evidence to date for a biological function for HG bps in duplex DNA.

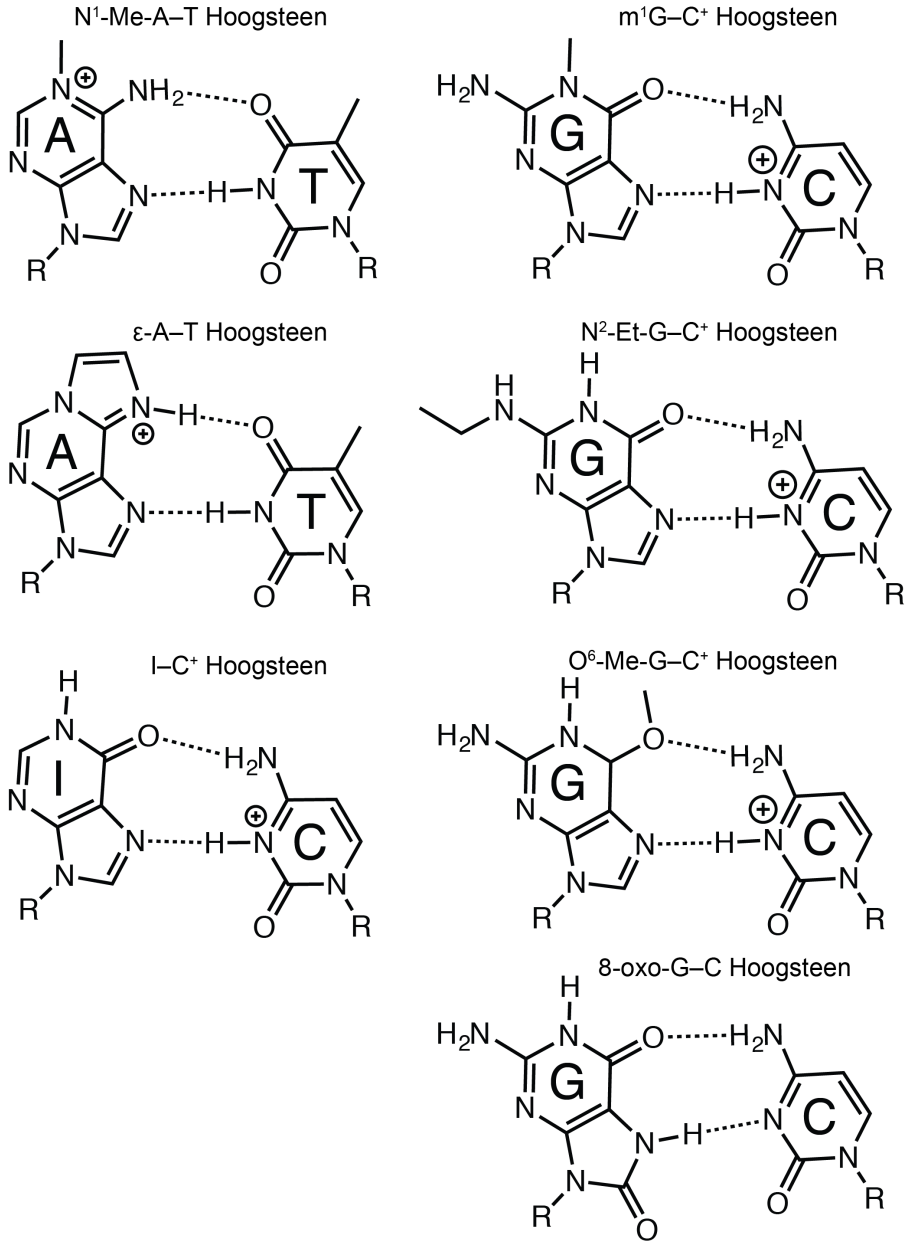

Figure 1.7: Chemical structures of HG bps in damaged DNA.

1.1.7 Transient Hoogsteen base pairs in DNA duplexes

The earliest fiber X-ray diffraction studies of DNA highlighted its polymorphic nature and the ability of the double helix to adopt different forms depending on environmental conditions and sequence contexts. Subsequent studies showed that DNA does indeed come in many different forms and that even B-DNA is not rigid, but rather, can undergo large deformations and thermal fluctuations in a sequence-dependent biologically important manner.[104] This flexibility was not confined to the weakly constrained sugar and phosphodiester backbone, but also includes the WC bps themselves.[105] Chemical probing and hydrogen exchange studies spanning the 1970s-1990s established that WC bps break apart and open at millisecond timescales and that the open state exists in at most ≈0.002% abundance for AT or ≈0.00008% for GC bps.[106-110] There are now several X-ray structures that capture these open states of the bps when bound to proteins that establish their functional significance.

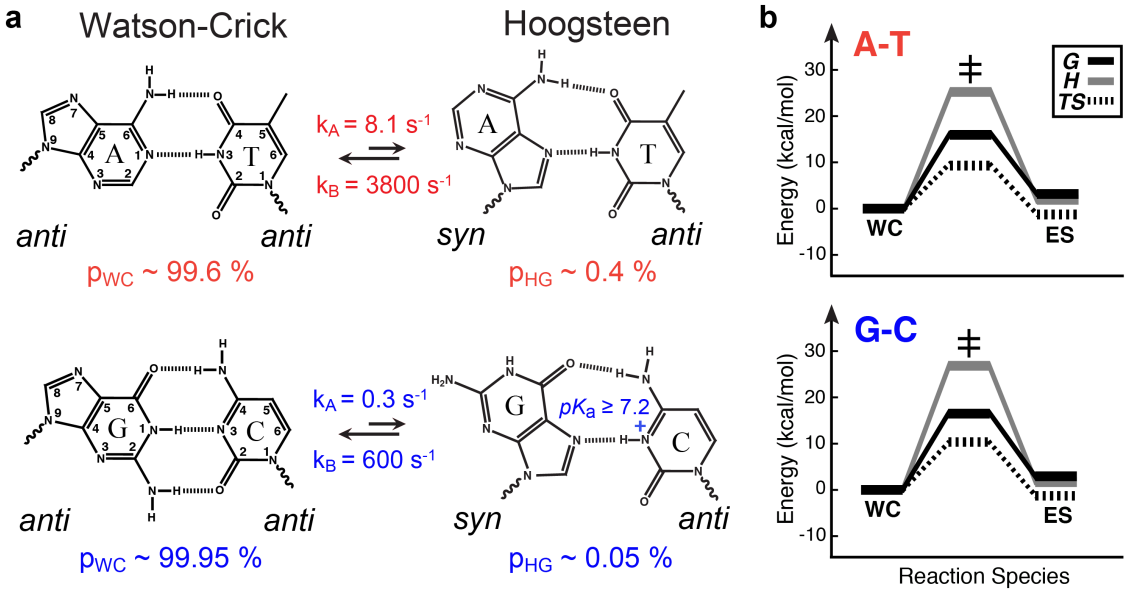

Figure 1.8: Transient HG in duplex DNA.

(a) A two-state exchange between WC and HG A–T and G–C1 HG base-pairs showing the relative populations and exchange rate constants obtained from R1ρ relaxation dispersion experiments at pH 6.8 and the estimated pKa for a HG G–C^+ base-pair[66,111]. (b) Corresponding thermodynamic profiles obtained from a temperature dependence of R1q relaxation dispersion at pH 5.4 (G, free energy; H, enthalpy; TS, entropy)[66].

NMR studies from our laboratory showed that both A–T and G–C WC bps can transiently undergo excursions toward HG bps in duplex DNA.[66,112] The transient HG bps were characterized with the use of recently developed NMR R1ϱ spectroscopic methods that make it possible to observe and structurally characterize fleeting states of macromolecules.[113-115] The transient HG bps (Figure 1.8) had populations of ≈ 0.1 - 1%, making them nearly three orders of magnitude more abundant than the open state, with the G–C$^+$ HG bps being less abundant than their A–T counterparts at physiological pH by a factor ranging between 20 and 100. The transient HG bps have lifetimes on the order of hundreds of microseconds to milliseconds (≈ 0.3 to 1.5 ms), which are significantly longer than the lifetimes of bp open states found to be in the nanosecond range.[110] It is remarkable that the HG bps are energetically less favorable than WC counterparts by a mere ≈3 kcal/mol in the case of A–T bps, roughly the equivalent of one strong hydrogen-bond. These energetic differences are small compared to forces that exist in cells due to proteins, binding torsional stress, and supercoiling or those applied due to crystal packing forces; or that arise from changing pH and ionic conditions. Studies suggest that the transient HG bps occurs universally across all DNA sequence contexts, in a non-cooperative manner, and with small albeit significant sequence-specific differences in population and lifetimes[116].

The picture that emerges is one, in which every bp in DNA exists as rapid superposition of WC and HG bps, with external parameters operating on the DNA

resolving one or the other base-pair. This helps explain the long and controversial observation of WC versus HG – small changes in conditions can favor one form over the other. It is striking that the difference in the abundance of transient G–C^+ and A–T bps mirrors the differences in efficiency observed in Polι replication of A/T versus G/C. The HG bps transiently expose the WC faces of purines, and may potentially help explain the much greater abundance of N1 methylation in adenine versus guanine. Most importantly, the observation of transient HG bps in duplex DNA, with comparable energetics to WC, raises the possibility that HG bps exist in much greater abundance *in vivo*, particularly in A–T rich regions of the genome. When combined with the current difficulties in resolving WC from HG based on X-ray diffraction data, it may well be the case that there are more HG bps in X-ray structures currently deposited in the PDB that have gone undetected, particularly for A–T bps.

1.1.8 Hoogsteen base pairs in RNA

The Watson-Crick (WC) double helix is not only the dominant structure of genomic DNA but also the most common structural element in RNA. Li *et al* find more than 50 percent of RNA in metazoan cells were double-stranded and the double-stranded RNA occur widespread in functional RNAs including mRNAs, long non-coding RNAs, rRNAs, tRNAs and transposable RNAs[117]. Double-stranded RNA forms the basic structural building block for secondary, tertiary and higher-order RNA

structures and is common in RNA architectures, intermolecular RNA-RNA interactions such as the mini-helix formed between codon and anti-codon as well as kissing dimers. Double-stranded RNAs also play important roles in gene expression regulation, translation and RNA interference pathways[118]. It has long been recognized that the canonical double helices formed by RNA (A-form) and DNA (B-form) differ in terms of their structure. The A-form helix differs from the B-form helix, and results in a stiffer helix characterized by a lower rise and twisting per bp step, larger rolling and displacement of bps away from the helical axis, wider helical diameter as well as both narrower and deeper major grooves. These differences arise because the 2′-hydroxyl (2′-OH) group in the ribose sugar ring in RNA sterically disfavors the C2′-endo sugar pucker that is preferred in B-DNA due to steric contacts between the 2′-OH and 3′-O[119] and the electronic effect of the 2′-OH group such as the electronegativity favors the C3′-endo sugar pucker[120]. This in turn brings the O5′ and O3' linking adjacent nucleotides into closer proximity in A-form as compared to B-form helix effectively compressing and rigidifying the A-form helix, widening its helical diameter, and displacing bps away from the helical axis. The observation of A–U and G–C WC bps in RNA by Rich et al[17] created a false sense of comfort suggesting that as recent studies have shown that, in contrast to B-form DNA, HG bps are not likely to form in A-form RNA, as suggested by the lack of HG base-pairs in A-form RNA duplexes in over 1000 high resolution crystal structures surveyed in the PDB.

Although HG bps were rarely observed in A-form RNA double helixes, they do occur in other structural contexts in RNA[121,122], at tertiary contacts in ribosomal RNA and transfer RNA and in RNA triplexes; reverse rA–rU HG bps can form within duplexes or triplexes[121,123]. Structures of HG bps can differ when they occur in different structural contexts, for example, HG or reverse HG bps in triplexes have the purine base adopting *anti* rather than *syn* conformation; reverse A–U HG bps form one different H-bond (A-N6H---O2-U) compared to A–U HG bps (A-N6H---O4-U) and the C1′–C1′ distance (9.5 Å) is less constricted than that in HG bps (~ 8.5 Å) that arises due to the *trans* orientation of the nucleobases. This thesis focuses on structures and biological implications of HG bps in the A-form RNA double helices, in comparison with HG in B-form DNA.

1.2 Characterization of Hoogsteen base pairs by NMR

As the work in this thesis focuses on characterizing HG bps in solution by NMR methods, the sections below are to provide a general introduction to cover the basis for the NMR approaches that will be applied in the following chapters.

1.2.1 Chemical shift

In NMR spectroscopy, chemical shift (δ) is the resonating frequency (Ω) of a nuclear spin in reference to the resonance signal (Ω_{ref}) from a standard molecule in the presence of a magnetic field with the strength of B_0 (in unit of Tesla or MHz); chemical shift is usually measured in the unit of parts per million (p.p.m.) as in Equation 1.1[124]:

$$\delta = \frac{\Omega - \Omega_{ref}}{\omega_0} \times 10^6 \qquad (1.1)$$

where ω_0 is the precessional frequency, (i.e. Larmor frequency) of a spin at a given static field with the strength of B_0 related by the gyromagnetic ratio (γ) of the spin ($\omega_0 = -\gamma B_0$). Practically ω_0 refers to the carrier frequency, for example, in a 600 MHz spectrometer, $\omega_0(^1H)$≈600 MHz or ω_0≈14.1 T. The presence of the static field B_0 can induce motions in electrons surrounding a given spin in the sample; such changes in the electron density thus generate a secondary local field, which will be experienced by the nucleus in addition to the static field B_0. The phenomenon is called "chemical shielding"[124]. Given the chemical shielding effect, the Larmor frequency of a spin can be described by Equation 1.2 in isotropic solution[124]:

$$\omega = -\gamma(1-\sigma)B_0 \quad (1.2)$$

where σ is the isotropic shielding constant. Note that chemical shielding is anisotropic when the overall tumbling of the molecule is slowed down which can be described by the chemical shift anisotropy (CSA) tensor. In solution NMR, CSA effect has minimal effect on chemical shifts due to the averaging by the fast molecular tumbling; however, it still contributes to the relaxation properties of a spin, for example, as used in the Transverse relaxation-optimized spectroscopy[125] (TROSY) experiment.

The chemical shielding property endows chemical shift the power to fingerprint spins that have different local electronic environments, ranging from small organic molecules to macromolecules and therefore to provide structural information. It has been used widely as a probe to study conformations in proteins and nucleic acids. For example, $^{13}C\alpha$ and $^{13}C\beta$ chemical shifts correlate with secondary structure formation (α-helix, β-helix) relative to the random coil conformation in protein[126]. Meanwhile, the imino ^{1}H resonances in nucleic acids have unique chemical shifts (≈12–13 p.p.m. for G-H1 and ≈13–14.5 p.p.m. for T/U-H3) that are far from the other proton resonances, and can directly detect base-pairing in DNA or RNA secondary or tertiary structures[127].

Unique chemical shift signatures for HG compared to WC bps arise from the unique HG structure in duplexes: i) *syn* purine conformation, ii) unique HG H-bonding and iii) cytosine protonation in G–C$^+$ HG bps at pH below the C-N3 pKa (≈7)[111]. Accompanying the purine flipping from *anti* to *syn* are downfield shifts by ≈2–4 p.p.m.

in the chemical shifts of purine-C8 and purine-C1′. Such shifts have been observed in chemically modified HG bps in duplex DNA[66], G^{syn}–G^{anti} mispairs in DNA quadruplexes[128] and G^{syn}–C^{anti} WC bp in Z-DNA[129]. For the aromatic carbon C8, the origin of such downfield shift accompanying the change in the χ angle are less likely due to the ring-current effect[130,131], but rather, can be attributed to the paramagnetic contributions from the chemical bonds. DFT studies have indicated that when the purine base flips from *anti* to *syn*, there are increased interactions between the *p* orbitals on C8 and those on O4′, resulting in decreased utilization of C8 *p* orbitals in bonding with N7 and N9. This results in a deshielding effect on C8 due to less paramagnetic contribution from the molecular orbitals localized on C8–N7 and C8–N9 chemical bonds[130]. The downfield shift in C1′ was suggested to originate likely from the torsion angle effect (i.e. change in the relative orientation of the base dipole relative to the sugar C1′), as well as the lengthening of the glycosidic bond C1′ – N9 (by ≈0.01Å) for *syn* compared to *anti* conformation, which can result in downfield shift in C1′[128,131].

By forming A–T HG H-bonding, it was reported that T-N3 and T-H3 in A–T HG H-bonds undergo an upfield shift relative to WC bp by ≈1–2 p.p.m. due to weaker H-bonding observed in chemically modified HG bps or HG bps in DNA-echinomycin complex[66,112]; the upfield shift in T-H3 is smaller in magnitude (< 1 p.p.m.) for HG bps in T–A–T in DNA triplexes[132]. In G–C^+ H-bonding, G-N1 presents ≈2 p.p.m. upfield shift due to loss of H-bonding; both imino and amino protons of the protonated cytosine (C^+-

H3 and C^+-H4) feature ≈ 2 p.p.m. downfield shift arising due to the protonation event[132]. In addition, the protonated cytosine also induced downfield shifted C^+-C6 (≈3 p.p.m.) and upfield shifted C^+-C5[111].

1.2.2 Longitudinal spin relaxation and NOE

The Nuclear Overhauser Effect (NOE) is the spin polarization transfer through a cross-relaxation mechanism between spins that have dipole-dipole interactions strongly depending on their distances in space. NOE provides a powerful tool in solution NMR applications for obtaining distance information in biomolecules. To understand the NOE effect, basic concepts of spin relaxation and dipole-dipole interactions will be introduced first.

Longitudinal relaxation of isolated spins: In the presence of an external magnetic field, spins align themselves along the magnetic field depending on their energy state (the "Zeeman effect"). For spin-1/2 nuclei, such as ^1H, ^{13}C and ^{15}N, they align either parallel or anti-parallel to the external field, which results in two populated states with quantized energies. At thermal equilibrium, the energy difference between the two energy states is $\Delta E = -\hbar\gamma B_0$, which dictates the populations of spins (n_α and n_β) at the two states (|α> and |β>, Figure 1.9) by the Boltzmann distribution ($\frac{n_\alpha}{n_\beta} = e^{\frac{\Delta E}{kT}}$). The net consequence of the Zeeman effect is a net bulk magnetization (M_z) along the Z-

direction, which is proportional to the population between |α> and |β> states[133], while $M_{xy} = 0$.

$$M_z = \frac{1}{2}\hbar\gamma(n_\alpha - n_\beta) \tag{1.3}$$

The magnitude of M_z dictates the amount of NMR signal can be observed. When the thermal equilibrium gets perturbed, for example, excited by a hard radiofrequency pulse, the spin system will relax back to the equilibrium (i.e. Boltzmann population distribution) through relaxation, including: (i) populations of different spin states back to the Boltzmann distribution (i.e. thermal equilibrium) (ii) complete loss of coherence in the spin system[133]. The former refers to the longitudinal (or spin-lattice) relaxation whereas the latter refers to the transverse (or spin-spin) relaxation.

In the longitudinal relaxation, the rate of population change with the rate constant *W* for an isolated spin at time *t* can be written as equation (1.4) assuming the first order process[133].

$$\frac{dn_\alpha}{dt} = -Wn_\alpha + Wn_\beta; \frac{dn_\beta}{dt} = -Wn_\beta + Wn_\alpha \tag{1.4}$$

Combining Equations (1.3) and (1.4), the longitudinal relaxation for the bulk net magnetization M_z follows:

$$\frac{dM_z(t)}{dt} = -R_1(M_z(t) - M_z^0) \tag{1.5}$$

where R_1 is the first-order longitudinal relaxation rate constant. The longitudinal relaxation requires a time-dependent magnetic field fluctuating at the Larmor frequency (ω_0) and the energy transfers from the spin system to the surrounding environment.

Spectral density: In a bulk sample in solution, each individual spin experiences stochastic variations in the local magnetic fields (B_{loc}) which arises due to the thermal molecular motions such as inter-molecular collisions. Both the magnitude and direction of B_{loc} vary randomly with time in an isotropic liquid sample. As only the frequencies that are close to the Larmor frequency (ω_0) that can cause relaxation, one needs to find out the amount of frequencies that exist around ω_0 in the random variation of B_{loc}. This can be assessed by the correlation function and spectral density function.

The random variation of the local field in the bulk sample can be described by the correlation function $G(t, \tau)$:

$$G(t,\tau) = \frac{1}{N}\sum_{i=1}^{N} B_{loc,i}(t)B_{loc,i}(t+\tau)$$
$$G(\tau) = \frac{1}{N}\sum_{i=1}^{N} B_{loc,i}(0)B_{loc,i}(\tau) \qquad (1.6)$$

where N is the total number of spins in the sample, t is time and τ is the time interval.

Generally the correlation function does not vary with the time point t due to the random behavior of the motion; rather, it depends on the time interval τ so the function can also be simplified as $G(\tau)$ at a time $t = 0$ that is arbitrarily chosen. In one simple scenario where molecules are spherical and freely tumbling and the solvent only

provides a medium with certain viscosity, such molecular motion (i.e. the rotational diffusion) can be expressed as a simple exponential correlation function G(τ); or its reduced form g(τ).

$$\begin{aligned} &G(\tau) = \overline{B_{loc}^2} \exp(-\frac{|\tau|}{\tau_c}), \\ &\overline{B_{loc}^2} = G(0), \\ &g(\tau) = \exp(-\frac{|\tau|}{\tau_c}) \end{aligned} \tag{1.7}$$

where τ_c is the rotational correlation time.

Fourier transforming the correlation function results in the spectral density function J(ω). In the example above, the resulting spectral density for the free tumbling, spherical molecule is a Lorentzian-shape function in Equation (1.8) with a center at ω=0.

$$\begin{aligned} &g(\tau) = \exp(-\frac{|\tau|}{\tau_c}) \xrightarrow{FT} j(\omega) = \frac{2\tau_c}{1+\omega^2\tau_c^2} \\ &j(\omega_0) = \frac{2\tau_c}{1+\omega_0^2\tau_c^2} \end{aligned} \tag{1.8}$$

The spectral density describes the probability that molecular motions with an angular frequency ω are present in the overall molecular tumbling; $J(\omega_0)$ represents the amount of molecular motions that can cause relaxation. Based on the Equation (1.8), the most rapid relaxation occurs when $\omega_0\tau_c = 1$. For a spherical molecule, τ_c mainly depends on molecular weight and the solvent viscosity by the Stoke's law,

$$\tau_c = \frac{4\pi\eta r^3}{3kT}; r \approx \sqrt[3]{\frac{3M}{4\pi\rho N_a}} + r_w \tag{1.9}$$

where η is the viscosity of the solvent, r is the hydrodynamic radius that can be estimated from the molecular weight (M), average density (ρ) and the hydration radius (r_w).

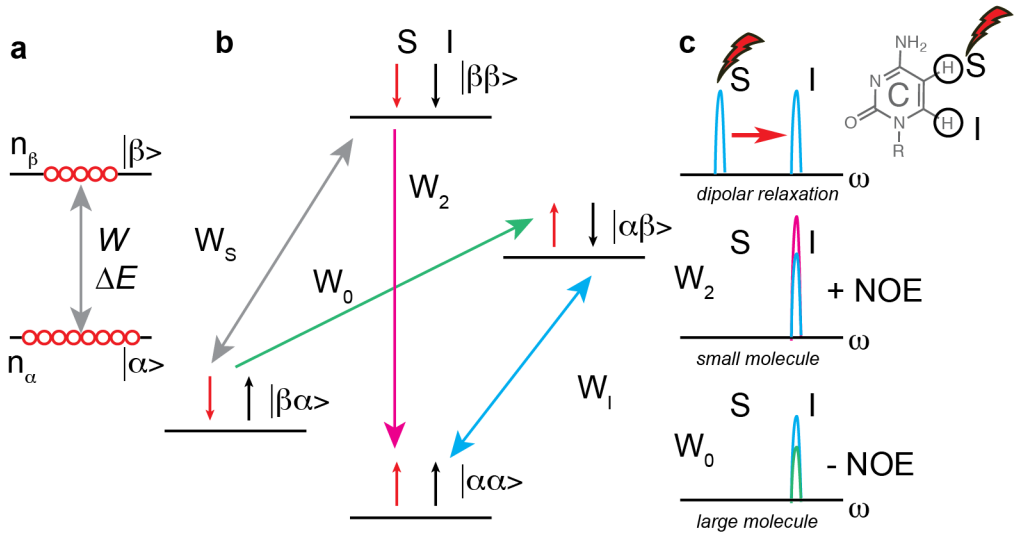

Figure 1.9: Cross-relaxation mechanisms for steady-state NOE.

(a) Population of spin states for an isolated spin. (b) Energy diagram and spin states for two dipole-dipole coupled spins I and S are shown with relaxation mechanisms in steady-state NOE experiments. (c) Example for illustration of 1D NOE difference experiment.

Dipolar coupling: The dipolar coupling originates from each spin generating a magnetic field that is parallel to the nuclear spin vector; two spins that are close in distance can experience each other's magnetic field, resulting in a slightly different local magnetic field on each spin depending on the orientation of both magnetic dipoles. The dipole-dipole coupling constant (b_{IS}) of two spins (I and S) are given by the Equation (1.10).

$$b_{IS} = -\frac{\mu_0}{4\pi}\frac{\gamma_I \gamma_s \hbar}{r_{IS}^3} \tag{1.10}$$

where μ_0 is the permeability of vacuum ($\mu_0 = 4\pi \times 10^{-7}$ H m^{-1}); r_{IS} is the distance between the two spins and γ_I and γ_S are their gyromagnetic ratios.

Relaxation of two dipole-dipole coupled spins: For two dipole-dipole coupled spins I and S (e.g. cytosine-H6 and H5 in Figure 1.9), the energy diagram consists of four energy states due to the dipolar coupling of spin states between I and S and only spin S is irradiated with a continuous radiofrequency pulse. At the end of the pulse, the populations at $|\alpha\rangle$ and $|\beta\rangle$ states of spin S are equalized (Figure 1.9) which is out of equilibrium, whereas spin I is not perturbed. From the diagram, it is clear that spin S has two channels to return back to equilibrium where $|\alpha\rangle$ state is more populated than $|\beta\rangle$ state: zero (W_0) and double (W_2) quantum transition pathways. In either of the relaxation pathway, not only the population of spin S undergoes transition but also spin I. In the double quantum pathway, the spin system transition is $|\beta\beta\rangle \rightarrow |\alpha\alpha\rangle$; for the spin I, starting from the Boltzmann distribution, has more $|\beta\rangle$ state spins transitioning to $|\alpha\rangle$

state; resulting in an increased net magnetization based on Equation (1.3) and in turn more intense NMR signal of spin I (i.e. positive NOE). In the zero quantum pathway, the spin system transition is $|\beta\alpha\rangle \rightarrow |\alpha\beta\rangle$; for the spin I, more $|\alpha\rangle$ state becoming $|\beta\rangle$ state resulting in a decreased net magnetization based on Equation (1.3) and in turn a less intense NMR signal (i.e. negative NOE). Although spin I was not excited by the initial pulsing, it still experiences an increased or decreased z-magnetization induced by the cross-relaxation of spin S under the condition that spin I and S are coupled through a dipole-dipole interaction. The Solomon equation[124] best describes the relaxation rate for the dipole-coupled two spins as shown in the Equation (1.11).

$$\frac{d\Delta I_z(t)}{dt} = -R_{1I}\Delta I_z(t) - \sigma_{IS}\Delta S_Z(t),$$
$$\frac{d\Delta S_z(t)}{dt} = -R_{1S}\Delta S_z(t) - \sigma_{IS}\Delta I_Z(t) \qquad (1.11)$$

where,

$$\Delta I_z(t) = I_z(t) - I_z^0,$$
$$\Delta S_z(t) = S_z(t) - S_z^0,$$
$$R_{1I} = W_0 + 2W_I + W_2,$$
$$R_{1S} = W_0 + 2W_S + W_2,$$
$$\sigma_{IS} = W_2 - W_0$$

R_{1I} and R_{1S} are the longitudinal relaxation rate constants for spin I and S, respectively and σ_{IS} is the cross-relaxation rate constant of the two dipolar-coupled spins. Based on the Equation (1.11), σ_{IS} is the difference between the two relaxation rate constants; the cross-relaxation only occurs when σ_{IS} is not equal to zero. The above

relaxation rate constants can be further described by the spectral density function[124,133] as given by Equation (1.12).

$$\begin{aligned} R_{1I} &= (1/24)\{j(\omega_I - \omega_S) + 3j(\omega_I) + 6j(\omega_I + \omega_S)\}, \\ R_{1S} &= (1/24)\{j(\omega_I - \omega_S) + 3j(\omega_S) + 6j(\omega_I + \omega_S)\}, \\ \sigma_{IS} &= (1/24)\{-j(\omega_I - \omega_S) + 6j(\omega_I + \omega_S)\} \end{aligned} \tag{1.12}$$

The cross-relaxation rate constant in a homonuclear spin system ($\gamma_I = \gamma_S = \gamma$), σ_{IS} from NOE is given by Equation (1.13),

$$\sigma_{IS}^{NOE} = \frac{\hbar^2 \mu_0^2 \gamma^4 \tau_c}{160\pi^2 r_{IS}^6}\left(-1 + \frac{6}{1 + 4\omega_0^2 \tau_c^2}\right) \tag{1.13}$$

NOE is proportional to the inverse sixth power of distance and to the fourth power of the gyromagnetic ratio. In 2D nuclear Overhauser spectroscopy (NOESY), the strength of the NOE signal between two nearby protons can be detected if they are within ≈5Å of each other.

NOESY cross-peaks were used to characterize WC versus HG bps. The NOESY cross-peaks unique to HG bps include strong intra-nucleotide H1′–H8 NOE for *syn* purine (≈2.5 Å) which is comparable to the distance between H5 and H6 in cytosine (≈2.5 Å), (i)A-H2–(i-1)H1′/H2′ and (i)A-H2–(i-1)H6/H8 for *syn* adenosine, and H8–H3, A-H6/C$^+$-H4–H3, (i)H3–(i+1)/(i-1)H1/H3, (i)A-H6/C$^+$-H4–(i+1)/(i-1)H1/H3 NOEs for connectivity involving imino or amino protons in both G–C and A–U/T HG bps[66,132,134]. Absence of the canonical sequential (i-1)H1′–(i)H8 NOE is also expected for *syn* purines due to the base flip.

1.2.3 Spin relaxation in the rotating frame ($R_{1\rho}$)

A different set of NMR techniques[135,136], referred to as spin relaxation in the rotating frame ($R_{1\rho}$) or simply relaxation dispersion (RD) can be used to probe short-lived and low-populated HG bps[66]. The $R_{1\rho}$ RD experiment[115,137] measures the line broadening contribution (R_{ex}) to the transverse relaxation rate (R_2) during a relaxation period in which a continuous radiofrequency (RF) field is applied with variable power (ω_{SL}) and frequency (ω_{RF}) for a spin of interest. ω_{RF} is very close to the Larmor frequency (Ω) of the observed state for the spin of interest and the difference between ω_{RF} and Ω gives the offset ($\Delta\Omega$) along the z-direction; ω_{SL} represents the strength of the B_1 field on the transverse plane (Figure 1.10). The vector sum of the offset and the B_1 field defines the strength and direction of the effective field (Figure 1.10). The $R_{1\rho}$ is the rotating frame relaxation rate constant that characterizes the exponential decay of the projection of the magnetization about the effective magnetic field as a function of time (Figure 1. 10).

$$
\begin{aligned}
&R_{1\rho} = R_1 \cos^2\theta + (R_2 + R_{ex})\sin^2\theta, \\
&\theta = \tan^{-1}(\frac{\omega_{SL}}{\Delta\Omega}), \\
&\Delta\Omega = \Omega - \omega_{RF}; \Omega = p_{GS}\Omega_{GS} + p_{ES}\Omega_{ES}
\end{aligned}
\qquad (1.14)
$$

p_{GS} and p_{ES} are populations of GS and ES, respectively; Ω_{GS} and Ω_{ES} are the precessing Larmor frequencies of GS and ES states.

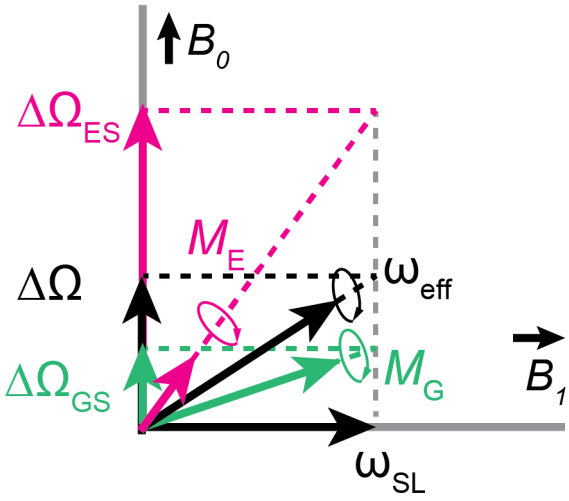

Figure 1.10: Offsets, effective fields and magnetizations in $R_{1\rho}$ RD experiment.

Figure courtesy of Dawn Merriman.

In presence of the applied RF field, the evolution of magnetization for a spin in a 2-state exchange can be described by the Bloch-McConnell (BM) equations[138,139]. McConnell modified the first-order kinetic Bloch equations by introducing the terms describing chemical exchange and provided a set of differential equations that describe the time evolution of magnetizations of the GS (M_{Gx}, M_{Gy}, M_{Gz}) and the ES (M_{Ex}, M_{Ey}, M_{Ez}).

$$\frac{d}{dt}\begin{pmatrix} M_{Gx} \\ M_{Ex} \\ M_{Gy} \\ M_{Ey} \\ M_{Gz} \\ M_{Ez} \end{pmatrix} = \begin{pmatrix} -k_f - R_2 & k_b & -\Delta\Omega_{GS} & 0 & 0 & 0 \\ k_f & -k_b - R_2 & 0 & -\Delta\Omega_{ES} & 0 & 0 \\ \Delta\Omega_{GS} & 0 & -k_f - R_2 & k_b & -\omega_{SL} & 0 \\ 0 & \Delta\Omega_{ES} & k_f & -k_b - R_2 & 0 & -\omega_{SL} \\ 0 & 0 & \omega_{SL} & 0 & -k_f - R_1 & k_b \\ 0 & 0 & 0 & \omega_{SL} & k_f & -k_b - R_1 \end{pmatrix} \begin{pmatrix} M_{Gx} \\ M_{Ex} \\ M_{Gy} \\ M_{Ey} \\ M_{Gz} \\ M_{Ez} \end{pmatrix} + R_1 \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ M_{G0} \\ M_{E0} \end{pmatrix} \qquad (1.15)$$

$$\Delta\Omega_{GS} = \Omega_{GS} - \omega_{RF}; \Delta\Omega_{ES} = \Omega_{ES} - \omega_{RF}$$

k_f is the forward rate constant from GS to ES and k_b is the backward rate constant from ES to GS. The BM equations can be extended to describe *n*-site exchange[140].

In the $R_{1\rho}$ RD experiment, the $R_{1\rho}$ relaxation rate constants were calculated by fitting peak intensities versus relaxation delay durations to a single exponential decay[115]. Uncertainty in the fitted $R_{1\rho}$ values (one s.d.) were derived using a Monte-Carlo method[141]. $R_{1\rho}$ data were fit to simulated $R_{1\rho}$ values given by the solution to the BM equations[140] at each given $\Delta\Omega$ and ω_{SL} combination to extract exchange parameters of interest, including the population of the ES (p_{ES}), the rate constant for conformational exchange ($k_{ex} = k_f + k_b$), and the difference between the chemical shifts of the ES and GS ($\Delta\omega = \omega_{ES} - \omega_{GS}$). Residual sum of squares were minimized using a bounded least-squares

algorithm[142] to give best-fit exchange parameters. The uncertainty in the chemical exchange parameters was calculated as the standard error of the fit[141].

2. Occurrence and structural features of Hoogsteen base pairs in DNA duplexes

2.1 Introduction

The unique HG base-pairing geometry can give rise to characteristic structural consequences that differ from WC geometry in duplex DNA. HG-unique H-bonding exposes unique set of H-bond donors and acceptors to the major groove which would be otherwise inaccessible as in WC bps; the *syn* purine and constriction of C1′–C1′distance can lead to altered backbone conformations which can be utilized in recognition by proteins or ligands.

HG bps have been observed in several DNA-protein[62,64,68,69] and DNA-ligand (e.g. quinoxaline bis-intercalators[47,143-145]) complexes where they can contribute to recognition (Figure 2.1). For example, in X-ray structures of bent TATA elements in complex with the TATA box-binding protein, two consecutive G– C HG bps help avoid a steric clash between the guanine exocyclic NH_2 group and a nearby leucine side chain[64]. Two consecutive A–T HG bps in X-ray structures of DNA in complex with the tumor suppressor protein p53 are thought to contribute to a narrowed minor groove and a more negative electrostatic potential surface that may favor insertion of positively charged Arg248[69]. HG bps have also been observed in chemically modified DNA, including N2-propanoguanine[78], 1,N2-ethylguanine[80], and N1-methyl adenine[82], and 8-amino-purine[146-148] where they may contribute to damage accommodation, recognition, and repair. There is also strong evidence that some members of the Y-family "low

fidelity polymerases" replicate DNA using HG pairing as the dominant mechanism, providing a means for bypassing lesions on the WC face during replication[90,95,100]. In the complex structures of DNA-quinoxaline bis-intercalators (i.e. triostin A and echinomycin), HG bps are observed flanking the intercalation sites at both the interior and termini of the duplex. Solution NMR studies on this complex with various DNA sequences provided evidence for dynamic HG bps in solution[55,56]. Theoretical studies suggested that the stabilization of HG bps flanking intercalation site and its dependence on DNA sequence could result from favorable van der Waals[149] and stacking[150] interactions.

Despite that there are above examples of HG base pairs suggested potentially significant role of HG in DNA structure-and-function, there has not been any comprehensive survey of occurrence of HG base pairs and their structural consequences. Moreover, recent studies employing NMR spin relaxation in the rotating frame ($R_{1\rho}$)[135,136,151] have shown that G–C$^+$ and A–T HG bps exist transiently in duplex DNA[66,111] across a variety of sequence and positional contexts[116]. These transient HG bps form with strong sequence-specific energetic preferences that are comparable to the sequence-specific variations in WC stability[152] potentially providing a new basis for sequence-specific DNA transactions[152]. These transient HG bps have populations of ≈0.5% and lifetimes of ≈1ms[66,116] but can increase considerably in modified bases such as inosine[153]. Observation of transient HG bps as well as HG bps in DNA-ligand complex in solution

can help rule out HG bps that occur in X-ray structures are totally due to crystallographic artifacts; though HG bps exist low in population, 0.5% reflects ≈3 million HG bps in human genome. These findings suggest HG bps can exist in much higher abundance than ever thought which can largely contribute to DNA structure polymorphism and to biological function, and emphasize the need to rigorously study the occurrence and preferences of HG bps in duplex DNA.

In order to provide a framework for guiding future explorations of the occurrence and functional importance of HG bps in duplex DNA, we comprehensively examined the Protein Data Bank (PDB)[154] to survey the occurrence and structural features of HG bps in duplex DNA[155].

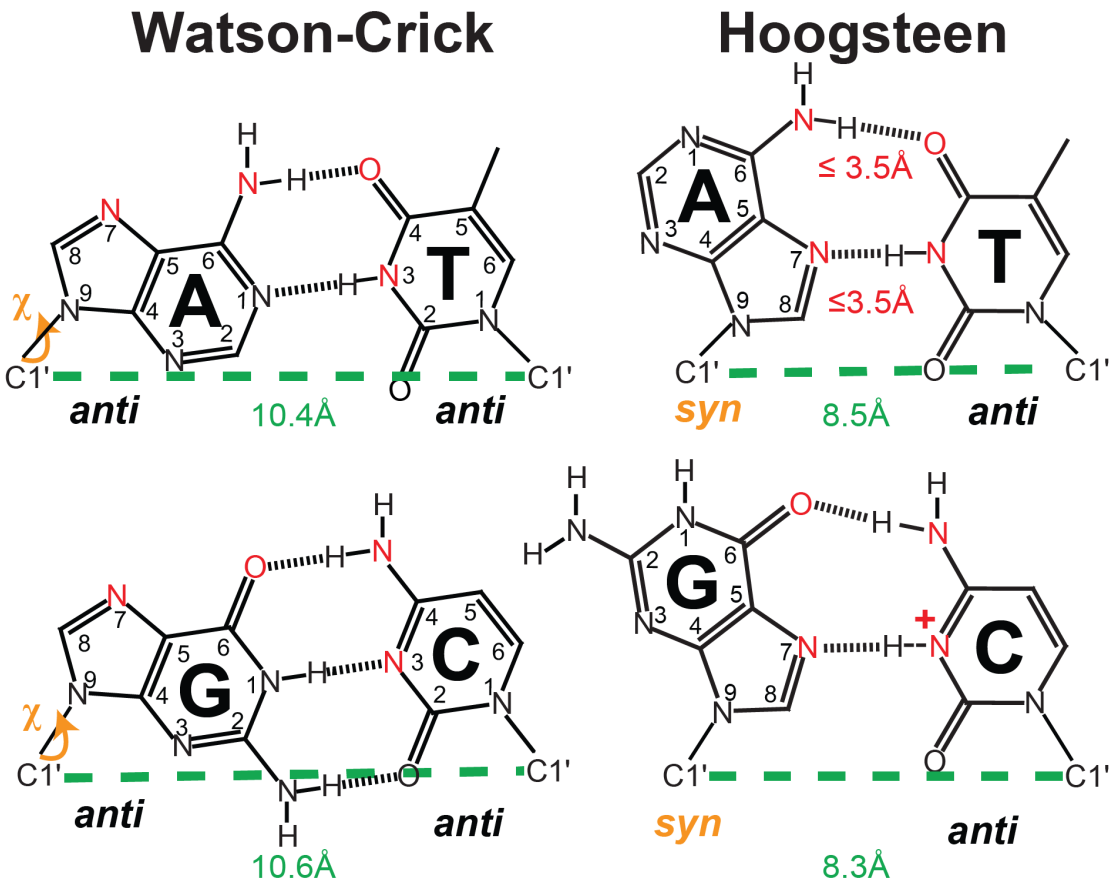

Figure 2.1: HG criteria for the PDB survey.

2.2 Methods

2.2.1 Survey protocol

The survey was carried out based on "HG criteria" that include unique structural features of HG bps in antiparallel DNA duplexes (Figure 2.1):

(i) HG hydrogen bonding: Both AN7---TN3 and AN6---TO4 distances ≤ 3.5Å for A–T; both GN7---CN3 and GO6---CN4 distances ≤ 3.5Å for G–C.

(ii) Constricted C1′–C1′ distance: The distance between the C1′ atoms of the purine and pyrimidine pair is restricted to ≤ 9.5Å, which is midway between the average distances observed for WC bps (≈10.5Å)[17] and HG bps (≈8.5Å)[5,43]. (Note that the constricted C1′–C1′ distance doesn't necessarily entail a shortened P-P distance across the helix).

(iii) *syn* purine: The *syn* glycosidic torsion angle (χ angle) of the purine base is in the range 0° ≤ χ ≤ 90°.

Details of the survey protocol has been described in the published work[155]. Briefly, X-ray crystal structures of DNA (by 9/4/2013, resolution ≤ 3.0Å) were downloaded from the Protein Data Bank (PDB)[154] and processed through an in-house program to find base pairs and generate structural parameters by the 3DNA program[156,157] including local base-pair parameters (shear, stretch, stagger, buckle, propeller, opening), the C1′–C1′ distance across the bp, heavy atom distances in H-bonds and sugar-phosphodiester backbone torsion angles (ν_0, ν_1, ν_2, ν_3, ν_4, α, β, γ, δ, ε, ζ)

(Figure 2.2). "HG criteria" were subsequently applied to all bps to identify HG bps. These bps feature not only HG bps in DNA duplexes but also those involved in base triples, tertiary interactions, WC bps in Z-DNA, as well as distorted WC-like bps. Despite that other structural contexts of HG bps could also be interesting and worth studying, they were not subjected to analysis in this survey to reduce complexity and keep the focus on antiparallel DNA duplexes.

It is well documented that WC and HG bps can be difficult to distinguish due to ambiguous electron densities[68,69,92,158] and yet they are often modeled by default as WC bps. This leaves open the possibility that there are actual HG bps mismodeled as WC. It is also possible that bps are often modeled as single states (HG or WC) when in fact partial occupancies of WC and HG might fit the density even better. Although the R-factor has been widely used as a metric for assessing the overall quality of structure model fitting, it is not applicable to evaluate the local density and fitting of an individual bp. To assess potential modeling errors in the bps, we manually examined the electron density maps and average B-factors (averaged over all sugar-phosphate and base heavy atoms i.e. C, N, O and P if not specified) of a given bp. Real-space correlation coefficient (RSCC) and σ-weighted $2F_o$-F_c map values for the base were also used to evaluate bps that are more susceptible to dynamics and thus modeling errors, where the RSCC values range between 0 to 1 for zero and perfect correlation between the F_c and $2F_o$-F_c maps, respectively.

To statistically analyze the sequence- and position-specific preferences of HG bps, and their structural features, redundancies in the identified HG bps were accounted by excluding redundant bps that are surrounded by identical adjacent WC bps in both 5' and 3' direction, involved in DNA duplexes with identical lengths, and when applicable, are bound to the same protein or ligand. This yielded a set of "non-redundant HG bps". As controls, two sets "adjacent WC bps" and "control WC bps" that consists of WC bps immediately adjacent to or more than one bp away from helical HG bps, respectively, were also generated. For each of the three sets of bps ("non-redundant HG", "adjacent WC", and "control WC"), population-weighted distributions were constructed for local base-pair parameters (i.e. shear, stretch, stagger, buckle, propeller, opening), C1′–C1′ distances, heavy atom distances in H-bonds, as well as sugar (ν_0–ν_4) and phosphodiester backbone torsion angles (α, β, γ, δ, ε, ζ).

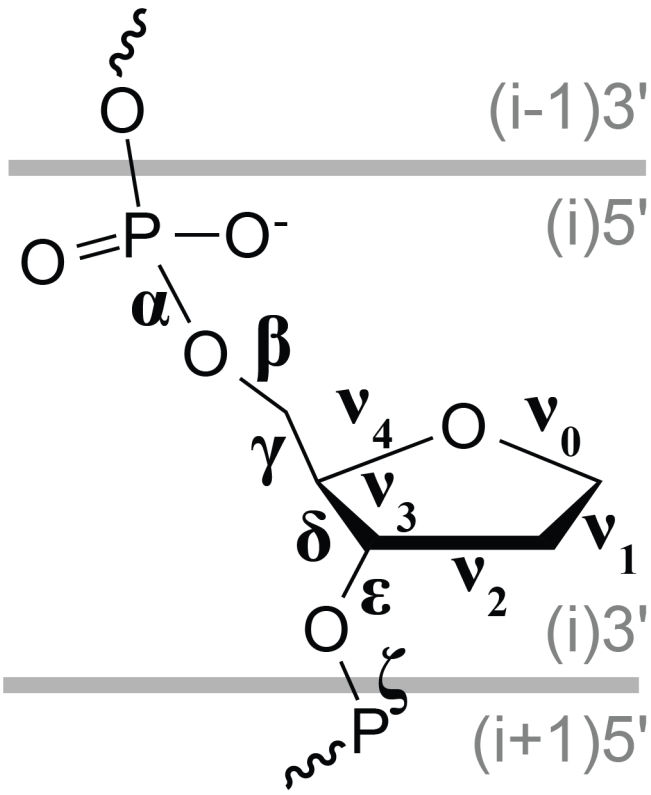

Figure 2.2: Sugar-phosphodiester backbone torsion angles.

2.2.2 Analysis of local structure

The impact of HG bps on local structure can be reflected by altered base-pairing geometries and/or sugar-phosphodiester backbone conformations at the HG bp itself or the surrounding WC bps. We examine this impact by analyzing the local structures of "non-redundant HG bps" and "adjacent WC bps" in comparison with those of "control WC bps". Firstly, we constructed 1D histogram distributions for local structural parameters (including sugar-phosphodiester torsion angles, local base-pair parameters, C1′–C1′ distances and heavy atom distances in H-bonds) for the "non-redundant HG", "adjacent WC" and "control WC" sets. Two-dimensional distribution plots for any given two parameters were generated to visualize different trends in clustering, if any, among the three sets of bps. More quantitatively, a recently introduced REsemble approach[159] was then used to measure the similarity between "control WC" distribution and those of "non-redundant HG" and "adjacent WC". Here, the overlap between two distributions (T and P) is computed using the square root of the Jensen-Shannon divergence (Ω^2)[160] given by Equation 2.1[159]:

$$\Omega^2\left(w_i^T(m), w_i^P(m)\right) = S\left(\frac{w_i^T(m)+w_i^P(m)}{2}\right) - \frac{1}{2}\left[S\left(w_i^T(m)\right) + S\left(w_i^P(m)\right)\right] \quad (2.1)$$

where $w_i^T(m)$ and $w_i^P(m)$ are the corresponding population weights for the ith bin for a given bin size m of distributions T and P. The term $S(w_i) = -\sum w_i(m) log_2 w_i(m)$ is the

information entropy[159]. The value of Ω is then computed as a function of bin size (m) that is used to build the histogram distribution. The resulting values are summed over K different bin sizes and normalized against a zero-overlap condition (Ω = 1 for all bin sizes) according to Equation (2.2),

$$\sum_K \Omega(w^T, w^P) = \frac{\sum_m \Omega(w_i^T(m), w_i^P(m))}{K} \quad (2.2)$$

The value of $\sum_K \Omega(w^T, w^P)$ provides a measure of similarity between distributions T and P and ranges between 0 and 1 for perfect and zero similarity, respectively[159]. REsemble was used to compare 1D distributions of sugar-phosphodiester backbone torsion angles (α, β, γ, δ, ε, ζ, ν_0, ν_1, ν_2, ν_3, ν_4) and local base-pair parameters (shear, stretch, stagger, buckle, propeller twist, opening), C1′–C1′ distances and heavy atom distances in H-bonds).

We previously used REsemble to compare 1D torsion angle distributions in RNA[159]. Here, we extended the analysis to compare multi-dimensional probability distributions consisting of six phosphodiester backbone torsion angles, five sugar torsion angles, and six local base-pair parameters. This was necessary because two pairs of 1D distributions (A and B versus a and b) could exhibit perfect overlap in 1D (A = a and B = b) yet exhibit zero overlap in 2D (AB ≠ ab). To maintain computational efficiency, 6D, 5D, and 6D REsemble was used to compare similarities between adjacent

WC bps (or HG bps if applicable) and the control WC bps of six phosphodiester backbone torsion angles (α, β, γ, δ, ε, ζ), five sugar torsion angles (ν_0, ν_1, ν_2, ν_3, ν_4), and six local base-pair parameters (shear, stretch, stagger, buckle, propeller twist, opening) distributions, respectively. Because HG bps have a distinct reference frame relative to WC bps and by definition have a constricted C1′–C1′ distance, the local base-pair parameters and C1′–C1′ distances for HG bps were not computed and compared with control WC bps. As a control, we measured the similarity between the control WC bp distributions and distributions obtained by randomly picking entries from the control WC bps such that the total number of entries equals that in the HG distribution.

To carry out multi-dimensional REsemble analyses, the translational local base-pair parameters (i.e. shear, stretch and stagger) between -2.5Å and 2.5Å were linearly converted to the range of 0° to 360° to be consistent with the torsional angle distribution range. The multi-dimensional distribution was constructed by using the same bin size to bin each parameter. The bin size (*m)* used in the REsemble analysis was varied between 15° and 360° with an increment of 15° (K = 24). Note that data points near the edges of the angle distribution (i.e. -180° and 180°) can lead to overestimation of $\sum_K \Omega$ (e.g., -179° and 179° differ by 358° in binning but only differ by 2° in reality). All of the local structural parameters including the converted translational base-pair parameters, are distributed in the middle of the range (-180° to 180°) except the backbone torsion angles β and ε, which have major distributions near the edges[155]. To minimize adverse effects

from data near the edges, all backbone torsion angles[155] (α, β, γ, δ, ε, ζ) between -180° and 0° were reflected onto the 180° to 360° region by addition of 360°, while those in the 0° to 180° region remain unchanged[155]. This yielded a final distribution between 0° and 360° with all angles distributed away from the edges.

2.2.3 Analysis of global structure

To assess the impact of HG bps on the DNA structure, we adopted the inter-helical Euler angle protocol developed for describing relative orientations of RNA A-form helices across junctions[87,161]. Other program such as "Curves" was also commonly used to analyze global bending of a helix; however, it is not suitable for measuring bending or kinking around single bp. Here, three inter-helical Euler angles (α_h, β_h, γ_h) are computed which describe the relative orientation of two helices across a given junction, in this case, a single or tandem HG/HG-like bps. For a given target DNA structure containing HG bps, we define a corresponding lower helix H1 and upper helix H2 to be the helices at the 5' and 3' sides, respectively, of the *syn* purine base in the HG bp (Figure 2.3). The inter-helical Euler angles describe the orientation of H2 relative to H1 across the junction of HG/HG-like bps and are determined by computing the rotation matrix that is required in order to rotate H2 so that it is in perfect coaxial alignment with H1. The approach has been described elsewhere in A-form RNA[87,161,162]. Here we provide a brief description emphasizing those differences that relate to bending in B-form DNA. β_h

is the inter-helical bend angle between H2 and H1, and ranges between 0° and 180°. α_h and γ_h are defined as 'twist' and 'arc' angles of H2 around the H2 and H1 helical axes, respectively and range between -180° and 180° (Figure 2.3). The inter-helical Euler angles (α_h, β_h, γ_h) are computed relative to a reference idealized B-form linear helix with 10 bps per turn consisting of two consecutive and perfectly coaxial helices (iH1 and iH2). This reference B-form helix was constructed using the 3DNA program[163] and the helix axis was oriented along the z-direction (Figure 2.3). The C1′–C1′ vector across the WC bp in iH1 immediately neighboring the junction was oriented along the y-axis with the major groove facing the +x direction (Figure 2.3). H1 in the target DNA structure was superimposed onto iH1 using heavy atoms (i.e. P, N, O, C) in the sugar-phosphodiester backbone. Next, reference helix iH2 was superimposed onto the resulting target helix H2 to yield iH2'. A rotation matrix R(α_h, β_h, γ_h) was then computed to transform iH2' back to iH2 using the EULER-RNA program (https://sites.google.com/site/hashimigroup/resources)[87,164]. The direct output (α_h^0, β_h^0, γ_h^0) from EULER-RNA was then converted based on the current definition of inter-helical Euler angles by: $if\ \beta_h^0 \geq 0, (\alpha_h, \beta_h, \gamma_h) = (\alpha_h^0, \beta_h^0, \gamma_h^0)$; $if\ \beta_h^0 < 0, (\alpha_h, \beta_h, \gamma_h) = (\alpha_h^0 \pm 180°, -\beta_h^0, \gamma_h^0 \pm 180°)$

In this reference frame, γ_h corresponds to the angle between the x-axis of the reference frame and the projection of the H2 helix axis onto the x-y plane (Figure 2.3) and represents the bending direction of H2 relative to the aligned WC bp (Figure 2.3); -

$90° \leq \gamma_h \leq 90°$ indicates bending towards the major groove whereas $-180° < \gamma_h < -90°$ and $90° < \gamma_h < 180°$ reflect bending towards the minor groove (Figure 2.3). Note that the bending direction (major or minor groove) may vary depending on the choice of the reference bp. For example, the direction may be different relative to a reference bp in H2, where α_h and not γ_h specifies the direction of bending of H1 relative to H2. A complete description of the bending direction requires all three Euler angles. The inter-helical twist angle $\zeta_h = \alpha_h + \gamma_h$ describes the relative twist between H1 and H2, and is equal to zero for a perfectly coaxial helix in B-form DNA. $\zeta_h > 0°$ and $\zeta_h < 0°$ represents under- and over-twisting, respectively[161]. The three angles β_h, γ_h, and ζ_h provide a complete angular description of the two helices.

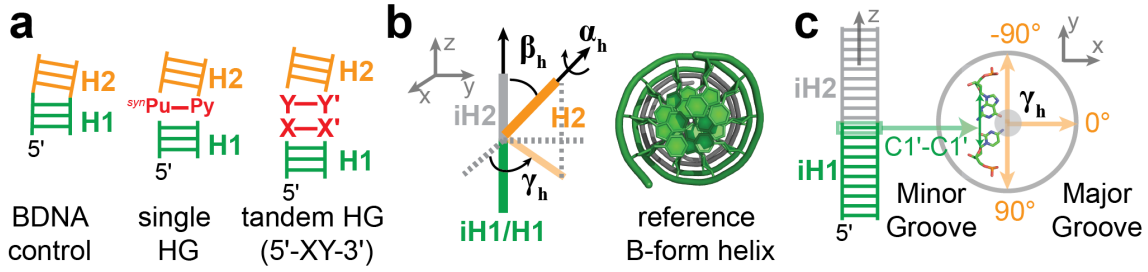

Figure 2.3: Definition of reference frame and Euler angles in bending analysis.

(a) Definition of upper, lower helices and junctions in control DNA and DNA containing HG bps. (b) Definition of Euler angles for the description of helix bending. (c) Reference frame for bending direction (i.e. towards major or minor groove).

The above approach for computing bend and twist angles assumes an idealized B-form geometry for the two helices. In RNA, the A-form geometry has been shown to be highly robust across different sequence contexts, and to a very good approximation, WC bps surrounding WC bps can be modeled assuming an idealized A-form geometry[87,161,162]. There can be greater variability in local structural parameters in B-form DNA based on analyses of X-ray structures[165] and MD trajectories[166]. We previously showed that the computed inter-helical Euler angles will not be reliable if the target helices superimpose with idealized helices with RMSD > 2Å[161]. In the current study, six out of fifteen structures yielded superposition RMSD > 2Å and were excluded from analysis. Among the remaining nine structures, five contained helices with terminal or non-canonical bps, which were used in the superposition. To evaluate the robustness of this approach, we compared the inter-helical angles computed when varying the number of bps (2 versus 3 bps) and types of heavy atoms (with or without sugar atoms C1′/C2'/O4') used in the superposition and found very small variations (≤2° for β_h and ≤6° for α_h and γ_h). In addition, as a negative control, we computed inter-helical angles for 11 X-ray structures of B-form naked DNA duplexes that don't contain HG bps and show no significant localized bending. These structures were used in a prior survey of duplex DNA structure[167]. Each duplex was sub-divided into two coaxially stacked 3 bp helices denoted H1 and H2 (Figure 2.3). Inter-helical Euler angles (α_h, β_h, γh) were then computed for H1 and H2 in all 11 structures. As a positive control, we computed inter-

helical angles for two bent DNA structures that do not contain HG bps, including an A6-T6 A-tract sequence (PDBID: 1FZX)[168] and the nucleosome particle (PDBID: 3UT9)[169]. The inter-helical Euler angles were then computed at the bent site and compared to the bending angle reported in literature as determined by other methods (e.g. Curves[170]).

2.3 Results and Discussion

2.3.1 Structural polymorphism in Hoogsteen base pairs

Our survey identified a total of 106 A–T and 34 G–C HG bps that satisfy all three HG criteria in DNA duplexes. In addition, the survey identified 22 A–T and 16 G–C HG-like bps that satisfy one or two of the HG criteria (Figure 2.4). Note that 91% of HG and 74% of HG-like bps have averaged B-factors over all heavy atoms (i.e. C, N, O and P) on the base and sugar-phosphodiester backbone ≤ 60 (Figure 2.4), indicating that they are reasonably well defined by the crystallography data.

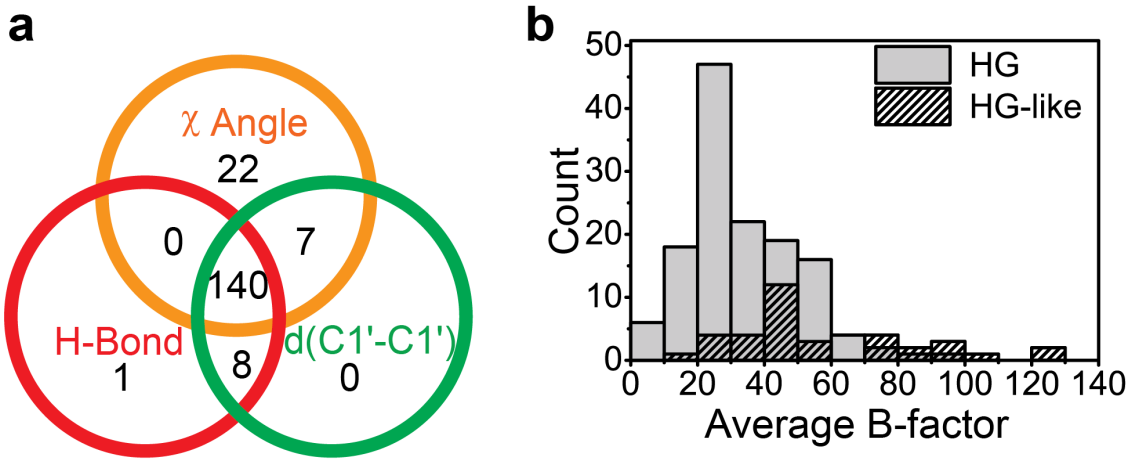

Figure 2.4: Summary of HG bps and their B-factors from the PDB survey.

(a) Statistics of number of HG and HG-like bps. (b) Histogram distribution of the average local B-factors of HG and HG-like bps.

Table 1: Sequence and biological contexts of HG and HG-like bps.

	PDBID	Sequence Context	Resolution (Å)	Biological Context	Context Function	Crystal Contact
HG base pairs	1RSB	5'-ATATAT-3'	2.17	naked DNA	N.A.	blunt-end stacking (HH)
	1XVK	5'-GCGTACGC-3'	1.26	echinomycin	antitumor antibiotic	blunt-end stacking (HH)
	1XVN	5'-ACGTACGT-3'	1.50	echinomycin	antitumor antibiotic	blunt-end stacking (HH)
	3EY0	5'-ATA	2.52	pentamidine	antiprotozoal drug	blunt-end stacking (HH)
	1VS2	5'-GCGTACGC-3'	2.00	triostin A	antitumor antibiotic	blunt-end stacking (HH)
	3H8O	T(MA7)G	2.00	alpha-ketoglutarate-dependent dioxygenase alkB homolog 2 (ABH2)	oxidoreductase; methylation lesion repair	
	3IGM	GCA-3'	2.20	apicomplexan apetala2 (AP2) domain	transcription regulator protein	
	2XCS	5'-AGC	2.10	S. aureus gyrase and antibacterial agent	type IIA topoisomerase; introduction of negative supercoils in DNA	protein contact
	4BUL	5'-AGC	2.60	S. aureus gyrase and antibacterial agent	type IIA topoisomerase; introduction of negative supercoils in DNA	protein contact
	1F2I	5'-ATG	2.35	fusion Cys2His2 zinc-finger protein	DNA binding	blunt end stacking (HH)
	3VOK	5'-ATG	2.00	heme-regulated transporter regulator (HrtR)	transcription regulator	
	1T3N	5'-AGG	2.30	human polymerase ι	error-prone Y family polymerase; replicate DNA	
	2ALZ	5'-TGG	2.50	human polymerase ι	error-prone Y family polymerase; replicate DNA	
	3GV5	TTC	2.00	human polymerase ι	error-prone Y family polymerase; replicate DNA	
	4EBD	5'-CTG	2.57	human polymerase ι	error-prone Y family polymerase; replicate DNA	
	4EYI	5'-(DG8)GG	2.90	human polymerase ι	error-prone Y family polymerase; replicate DNA	
	1K61	TAA	2.10	Mating-type protein α-2 (MATα2) homeodomain	DNA binding	
	4ATI	AAC-3'	2.60	Microphthalmia-associated transcription factor (MITF)	DNA binding	
	2ATA	5'-AAG	2.20	human p53 core domain	tumor suppressor protein	blunt-end stacking (HH)
	3IGK	CATG	1.70	human p53 core domain	tumor suppressor protein	
	3KZ8	CATG	1.91	human p53 core domain	tumor suppressor protein	
	2ODI	5'-AAC	1.45	BcnI	typeII restriction endonuclease; recognize and excise DNA	sticky-end stacking (WH)
	3N7B	5'-AGT	2.65	SgrAI		blunt-end stacking (WH)
	2IBK	5'-TCATGA	2.25	sulfolobus solfataricus P2 DNA polymerase IV (Dpo4)	error-prone Y family polymerase; replicate DNA	
	3V6J	(EFG)GA	2.30	sulfolobus solfataricus P2 DNA polymerase IV (Dpo4)	error-prone Y family polymerase; replicate DNA	
	1QN3	ACGG	1.95	TATA-box binding protein	transcription initiation factor TFIID-1; specifically binds TATA-box DNA	
	1QN6	AGG	2.10	TATA-box binding protein	transcription initiation factor TFIID-1; specifically binds TATA-box DNA	
	1QNB	TGG	2.23	TATA-box binding protein	transcription initiation factor TFIID-1; specifically binds TATA-box DNA	
	2VIH	CTTTTAG	2.10	IS608 transposase	DNA binding and excision	
	2VJU	CTTTAG	2.40	IS608 transposase	DNA binding and excision	
	2XM3	CTTCAG	2.30	ISDra2 transposase	DNA binding and excision	
	1ODG	5'-TAGGC(5CM)TG	2.80	very-short-patch repair (Vsr) enzyme	nucleotide excision repair of G•T mismatches	
HG-like base pairs	1QP5	CGG	2.60	naked DNA	N.A.	
	239D	GGG-3'	2.05	naked DNA	N.A.	
	2PIS	GAA(FFD)TT	2.80	naked DNA	N.A.	
	329D	CGG	2.70	naked DNA	N.A.	
	4E8X	CGG-3'	2.18	ruthenium complex	N.A.	
	1LWW	TAC	2.10	human 8-oxoguanine DNA glycosylase	base excision DNA repair	
	3GYH	TAG	2.80	alkyltransferase-like (ATL) protein	DNA alkylation damage repair	
	2WT7	5'-AAT	2.30	heterodimeric MafB: cFos	leucine zipper transcription factor	blunt-end stacking (WH)
	1K7A	5'-ACA	2.80	Ets domain of Ets-1	Ets family transcription activator	
	4I2O	TAT	1.77	FixK2 protein	transcription regulator	
	1IHF	CAA	2.50	integration host factor (IHF)	DNA binding; architectural factor	protein contact
	4AUW	5'-TAA	2.90	bZIP homodimeric MafB	transcription factor	sticky-end stacking (HT)
	1DE9	5'-(3DR)GA	3.00	human major apurinic/apyrimidinic endonuclease (APE1)	base excision DNA repair of apurinic/apyrimidinic DNA	
	3G2D	5'-CAG	2.30	Mth212 exodeoxyribonuclease	uridine/abasic endonuclease, 3'->5' exonuclease	
	3ODH	ATA-3'	2.30	OkrAI endonuclease	restriction endonuclease	stacking (WH)
	1S0M	5'-AT(BPA)A	2.70	sulfolobus solfataricus P2 DNA polymerase IV (Dpo4)	error-prone Y family polymerase; replicate DNA	
	3V6H	(EFG)GA	2.30			
	1OZJ	5'-GTA	2.40	Smad3 MH1 DNA binding domain	transcription factor; DNA binding	

The vast majority of the HG and HG-like bps (88%) are found in structures of duplex DNA in complex with proteins and/or ligands. Among these bps, most cases (96%) are not in direct contact with the protein or ligand and 66% are located at the duplex terminal ends. In contrast, all 16 HG bps observed in naked DNA duplexes correspond to pure HG helices of AT-repeats. No HG bps neighboured by WC bps on either end in naked DNA duplexes are observed. Many of these HG and HG-like bps, especially those located at duplex termini, do not appear to be documented in the primary literature.

The survey identifies a total of 178 HG and HG-like bps (128 A–T and 50 G–C bps) that correspond to ≈0.3% of all 51485 A–T and G–C DNA bps in the PDB (as of 9/4/2013). Interestingly, this overall abundance of HG bps compares favorably to the population ≈0.5% (at pH ≈6.8) measured by NMR relaxation dispersion for transient HG bps in duplex DNA in solution[66]. This suggests that the HG bps captured by NMR and X-ray crystallography are subject to similar energetic forces and that the differences in the experimental conditions do not lead to substantial changes in the overall abundance of HG versus WC bps. However, we cannot rule out that the environmental factors influence the distribution of HG bps and their specific location.

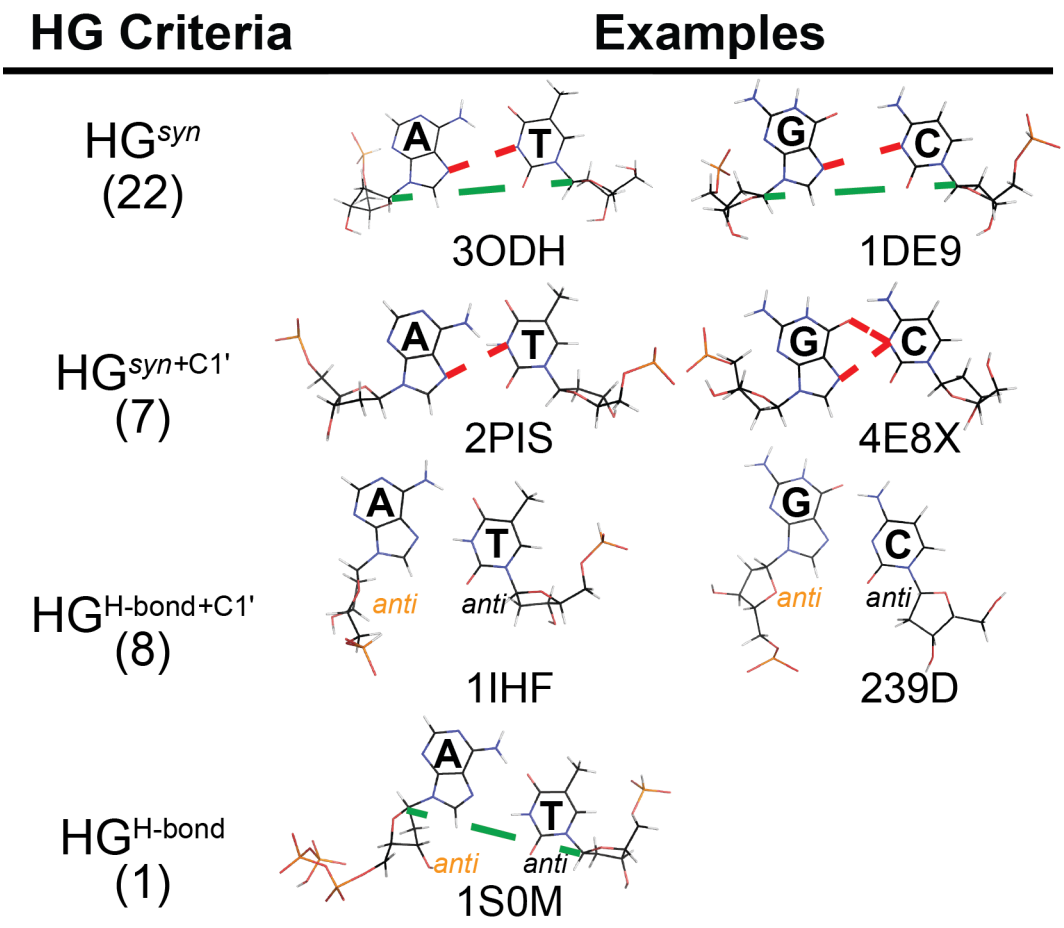

Figure 2.5: Statistics and examples for HG-like bps from the PDB survey.

The 38 HG-like bps exclude the entries that only satisfy the constricted C1′–C1′ distance criteria, which we consider 'distorted WC' bps. Most of the HG-like bps satisfy only the *syn* purine criterion (22 HG^{syn}) or both the *syn* purine and constricted C1′–C1′ distance (7 $HG^{syn+C1'}$). The HG^{syn} represents a partially open HG bp in which the purine and pyrimidine bases are not brought into proximity following the purine flip, explaining the absence of HG-type H-bonding (Figure 2.5). Interestingly, the HG^{syn} conformation falls along a WC-to-HG transition pathway previously proposed based on peak conjugate refinement simulations[66] and by Φ-value analysis[116]. The $HG^{syn+C1'}$ bps feature deviations in shear and do not support HG-type H-bonds (Figure 2.5). Approximately 62% of the HG^{syn} and $HG^{syn+C1'}$ bps are located at or near the duplex terminal ends. Another 8 bps satisfy the HG H-bonds and constricted C1′–C1′ distance but feature *anti* rather than *syn* purine base ($HG^{Hbond+C1'}$). Seven of these $HG^{Hbond+C1'}$ bps are located at a nicked site in the IHF-DNA complex and another one occurs at the terminal end of a left-handed DNA duplex (Figure 2.5). We also find one bp that only satisfies the HG H-bonding criterion (Figure 2.5). This corresponds to an *anti* purine base with a C3'-endo sugar pucker that forms HG H-bonds through rearrangement of the sugar-phosphodiester backbone in a manner analogous to Z-DNA. Note that we cannot rule out that some of these bps arise due to modeling errors and ambiguous density. Indeed, it was previously shown that an open G–U pair without hydrogen bonding in the ribozyme structure (PDBID: 1CX0) could be more confidently assigned to be a

reverse wobble G–U bp after single-residue remodeling with ERRASER[171]. The unpublished work from Hintze *et al*. suggested that densities of nine structures fit WC equally well and two had potential modeling errors where WC fit the density better.

2.3.2 Structure and sequence preferences of Hoogsteen base pairs

We examined a set of 54 non-redundant HG bps (Methods, Table 1) to assess the statistical significance of their position and sequence-preferences. In this non-redundant set, HG bps are ≈4-fold more enriched in A–T versus G–C bps at both helical (by ≈5-fold) and terminal (by ≈3-fold) sites (Figure 2.6). Similar preferences are observed when considering all 140 HG bps. The increased preference for A–T versus G–C^+ HG bps is consistent with the ≈8-fold greater abundance of transient A–T versus G–C^+ HG bps measured in duplex DNA in solution by NMR relaxation dispersion[66,111]. The lower abundance of G–C^+ HG bps can be attributed to the loss of one H-bond as well as by the requirement to protonate cytosine N3 (pKa ≈7.2)[111]. The average pH and standard deviations in the crystallization conditions for structures containing A–T or G–C HG both have pH≈6±1 as compared to the overall average pH≈7±1 of all DNA structures in the PDB with resolution ≤ 3.5Å (Figure 2.6). Interestingly, three structures of DNA-quinoxaline bis-intercalator complexes with G–C HG bps are collected under rather acidic conditions (pH=4.5)[43,143]. In these cases, it is possible that the lower pH helps

contribute toward the stabilization of the HG bps. Other structures with only A–T HG bps from similar complexes are observed at higher pH≈6.

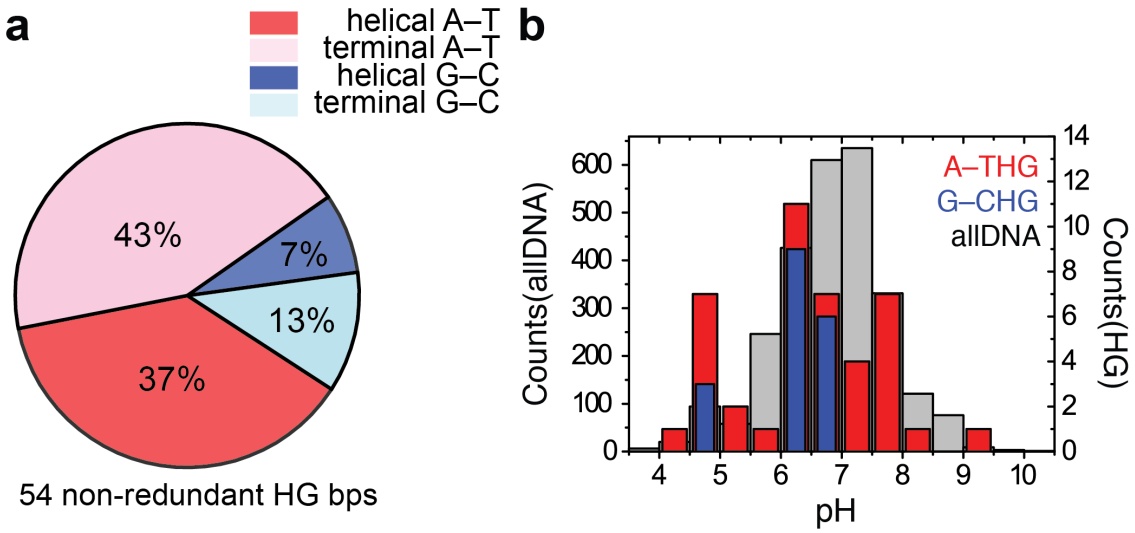

Figure 2.6: Bp types, sequence contexts and pH conditions for HG bps.

The overall ratio between bps in the interior of DNA helices and bps at helix termini is approximately 6:1. The same ratio for non-redundant HG bps identified in the survey is approximately 1:1 (Figure 2.6), implying that HG bps are enriched at terminal ends. The terminal bps should be treated with caution given that they have increased susceptibility to dynamics and intrinsic structural noise. However, close examination of the electron density maps (see Methods) at these terminal HG bps reveals that most have good electron density and low to moderate average B-factors. The examination of non-redundant HG and HG-like bps reveals that 66% of all terminal HG and HG-like bps are involved in crystal contacts with nearby DNA or proteins in the crystal lattice while the remaining 34% of terminal HG bps are observed within active sites of polymerases, endonucleases or as the apical loop closing bps showing no contact with nearby molecules. It is possible that these terminal HG bps are stabilized, at least in part, by crystal packing forces. All of the HG bps in crystal contact with nearby DNA molecules show end-to-end stacking, 86% of which involves very similar blunt-ended HG-to-HG stacking with 2-fold symmetry while 14% are attributed to HG-to-WC stacking with translational symmetry in the crystal contact (Table 1). Enrichment of one specific type of stacking in crystal contacts indicates that favorable stacking could be a reason for the enrichment of HG bps at terminal ends. Nevertheless, one should not solely consider these HG bps as crystallographic artifacts. It may well be that packing in the crystal lattice could have parallels in vivo including in nucleosomes, chromatin, and

possibly other stressed cellular DNA. This suggests that HG bps may play unique roles at termini-involved DNA biochemical transactions such as active sites of DNA polymerases, nucleases, transposases and ligases involved in DNA replication, recombination and various damage repair pathways. This also suggests that transient HG bps may also be more abundant at terminal ends, possibly as intermediates that have been observed accompanying end-fraying events[152,172]. This preference for A–T HG bps at terminal ends is consistent with early studies showing that isolated 9-methyladenine and 1-methylthymine bases prefer to associate as HG rather than WC bps[4,26].

Interestingly the *syn* purine bases in terminal HG bps show a 12-fold preference for being at the 5' versus the 3'-end. A recent study showed that sequence-specific variations in the stabilities of HG bps in duplexes can be attributed to variations in WC stability[116,152]. These A–T HG preferences may also mirror variations in WC stabilities and indeed prior NMR studies suggest weaker stability for 3' versus 5' terminal guanine WC bps[173]. End-to-end intermolecular stacking could also favor 5'-purine A–T HG bps as this results in a TA dinucleotide step formed in the crystal contact which favors HG[116,174].

Although the limited number of HG bps does not allow a statistically rigorous analysis of sequence-specific preferences, we note a few observations. First, we only observe either single HG bps surrounded by WC bps or two tandem (TA, AT or CG) HG bps that are palindromic. There are also three entire helices that are formed exclusively

of HG bps in sequences consisting of three[38-40] or five AT repeats[41]. We do not observe three or more consecutive HG bps that are surrounded by WC bps. In addition, the TA step (HG bp is underlined) is the most frequently observed HG step representing 47% of all single HG bp-involved steps while the GG step is completely absent. These sequence-specific preferences are in good agreement with studies showing that HG bps favor AT-rich sequences[26,66,116] and NMR relaxation dispersion studies showing the greatest abundance of transient HG bp is in TA steps and lowest abundance is in GG steps[116].

2.3.3 Impact of Hoogsteen base pairs on local B-form DNA structure

Prior structural[39,40,68] and computational studies[174,175] have shown that HG bps can be accommodated within B-form helices of WC bps without significantly distorting the base-pairing geometry and sugar-phosphodiester backbone of neighboring WC bps. Nevertheless, studies have reported small perturbations induced by HG pairing including α/γ torsion angles with gauche$^+$/gauche$^-$ rather than the common gauche$^-$/gauche$^+$ at the HG bp[39,68]. Our survey provides an opportunity to examine any local perturbations that may be induced by formation of HG bps.

To this end, we compared 1D histogram distributions of sugar (ν_0-ν_4) and phosphodiester backbone (α, β, γ, δ, ε, ζ) torsion angles for three data sets (i) non-redundant HG bps ("HG") (ii) WC bps adjacent to HG ("adjacent WC") and (iii) WC bps surrounded by WC bps ("control WC"). Visual inspection reveals relatively broad

distributions with no discernable differences between the three sets of distributions (Figure 2.7). To more quantitatively compare the similarities between the distributions, we used the REsemble approach[159], which was recently developed to measure the extent of similarity between histogram distributions (see Methods). In this approach, the similarity between two distributions is measured by computing $\Sigma\Omega$ that ranges between 0 and 1 for maximum and minimum similarity, respectively (see Methods)[159]. In general, we observe high similarity between the three datasets with $\Sigma\Omega \leq 0.2$ for HG versus control WC and adjacent WC versus control WC with the values being generally lower for sugar torsion angles ($\Sigma\Omega < 0.1$). However, among these small differences, the relatively larger deviations ($\Sigma\Omega > 0.1$) are in γ and ζ for HG versus control WC and in α for adjacent WC versus control WC, suggesting that HG bps more likely induce changes in these torsion angles.

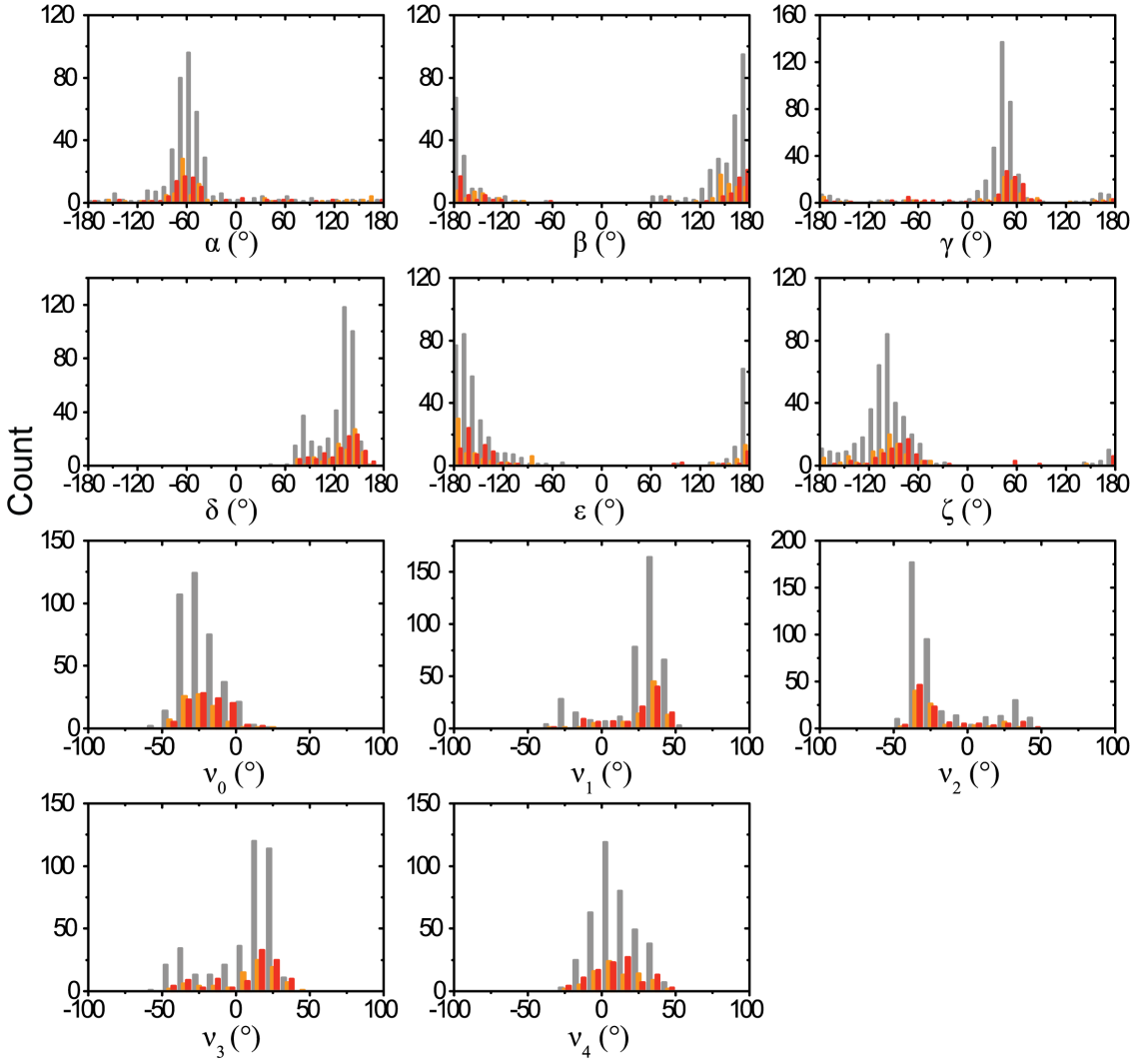

Figure 2.7: 1D distribution of HG backbone torsion angles compared to WC.

Overlay of backbone angle distributions of HG (in red), adjacent WC bps (in orange) and control WC bps (in grey).

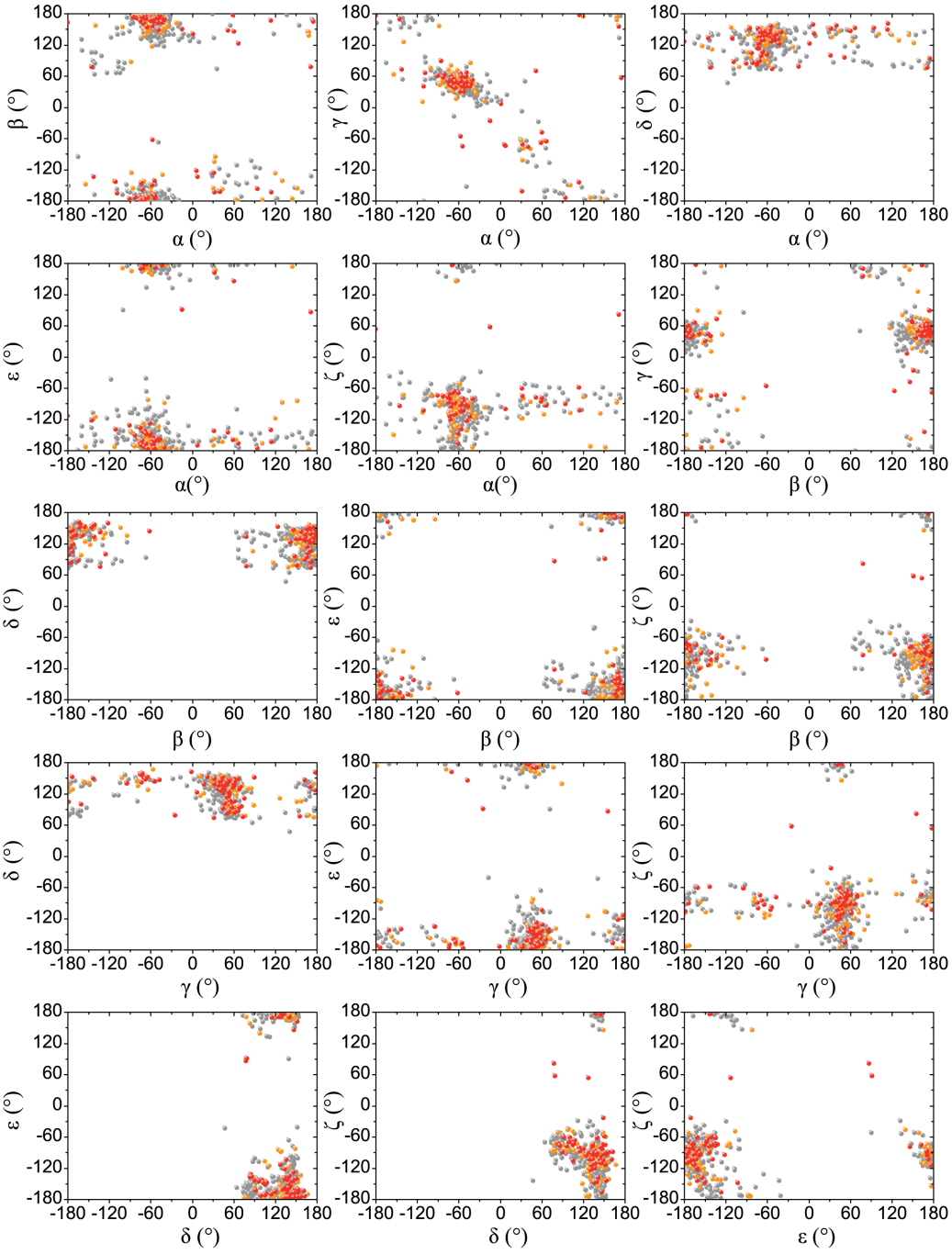

Figure 2.8: 2D scatter plot of HG backbone torsion angles compared to WC.

Although only small deviations are apparent when comparing 1D (Figure 2.7) and 2D (Figure 2.8) distributions of backbone torsion angles, there could be more significant deviations induced by HG bps that are not captured because they involve small correlated variations in different torsion angles. To examine this, we used REsemble to compare 5D sugar (ν_0, ν_1, ν_2, ν_3, ν_4) and 6D phosphodiester (α, β, γ, δ, ε, ζ) distributions across the various sets of bps. Note that carrying out higher dimensionality comparisons (e.g. 11D encompassing both sugar and phosphodiester torsion angles) is very computationally costly (see Methods). Although the probability for overlap decreases rapidly for multi-dimensional distributions, this inherent decrease in overlap is taken into account by evaluating the overlap between the control WC distribution and sub-distributions of its own (see Methods). Based on this analysis, we find that 6D distributions (see Methods) of six phosphodiester backbone torsion angles for both HG bps ($\Sigma\Omega_{HGbp} = 0.42$) and adjacent WC bps ($\Sigma\Omega_{adWC} = 0.37$) deviate significantly from control WC bps as compared to the subset WC bps taken from the control WC bps ($\Sigma\Omega_{subWC} = 0.25$) (Figure 2.9). This indicates that torsion angles in HG bps and adjacent WC bps deviate from the control WC bps even when taking into account intrinsic statistical deviations in the multidimensional distribution of torsion angles for the control WC bps (Figure 2.9). Similar but smaller deviations are observed for the 5D distributions of sugar torsion angles (ν_0, ν_1, ν_2, ν_3, ν_4) (Figure 2.9).

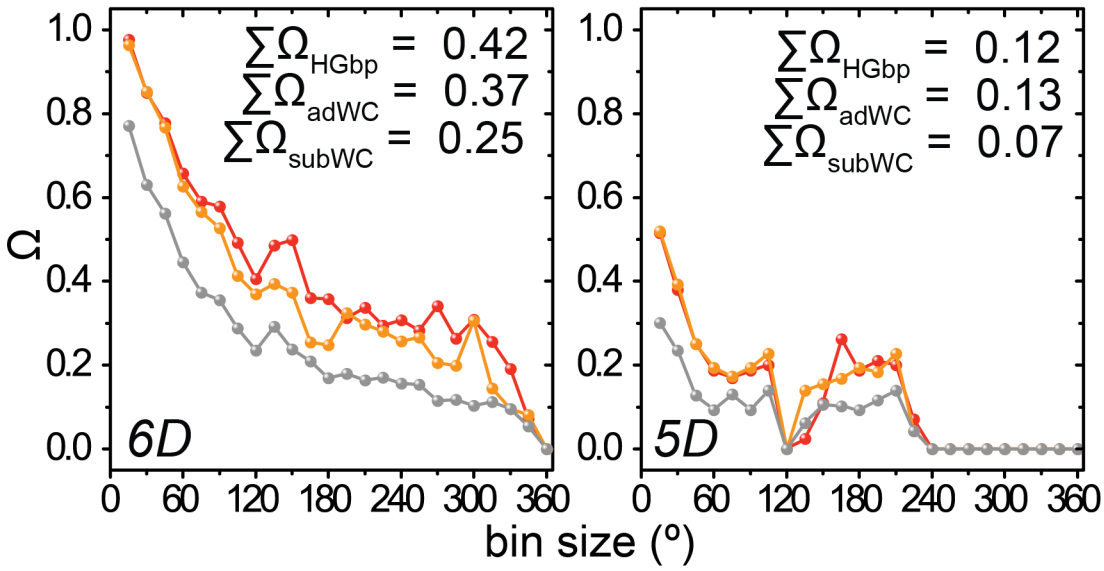

Figure 2.9: Multi-dimensional REsemble analysis on HG and WC bps.

REsemble analyses of HG (in red), adjacent WC bps (in orange) and control WC bps (in grey) are shown for six backbone torsion angles (left) and sugar dihedral angles (right).

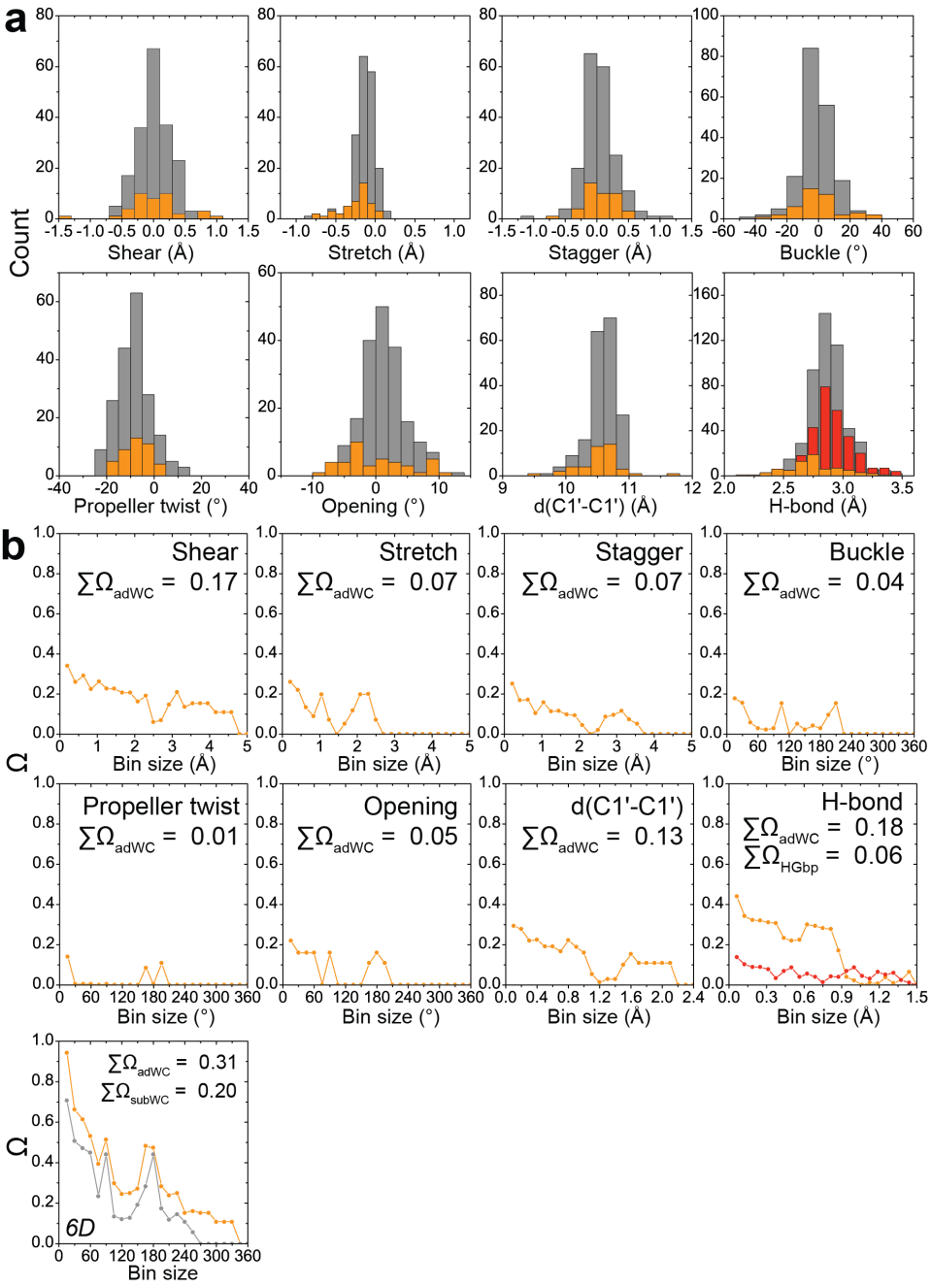

Figure 2.10: Histogram and 1D REsemble analyses of local-bp parameters.

Examination of histogram distributions and REsemble analyses for local base-pair parameters (shear, stretch, stagger, buckle, propeller, opening), C1′–C1′ distances, heavy atom distances in H-bonds reveals slight differences in base-pair geometries between adjacent WC and control WC datasets. In particular, although the differences are small based on 1D REsemble analysis ($\Sigma\Omega < 0.2$), 6D REsemble analysis shows local base-pair parameters in the adjacent WC deviate from control WC bps ($\Sigma\Omega_{adWC} = 0.31$), compared to that of a subset taken from control WC bps ($\Sigma\Omega_{subWC} = 0.20$) (Figure 2.10). The largest deviations are observed for shear, opening, C1′–C1′ distances as well as heavy atom distances in H-bonds for adjacent WC bps compared to control WC bps (Figure 2.10). Based on 1D histogram distributions, the adjacent WC bps tend to have larger shear and opening, together with shorter heavy atom distances in H-bonds as compared to control WC bps (Figure 2.10). These perturbations on the WC bps adjacent to HG are consistent with the observation of exchange broadening of aromatic resonances in WC bps adjacent to HG bps that are trapped by N1-methylation which suggests enhanced dynamics[66,111].

2.3.4 Impact of Hoogsteen base pairs on global B-form DNA structure

In crystal structures of DNA-IHF (e.g. PDBID: 1IHF) and DNA-TBP (e.g. PDBID: 1QN3) complexes, HG bps are observed near sharp kinks in the DNA[62,64]. Interestingly, kinking across HG helices with sticky ends has also been observed in the absence of

proteins or ligands in a coiled-coil DNA structure with sequence d(CGATATATATAT) (PDBID: 2AF1) where the $(AT)_5$ HG bps form a linear HG helix but there is a kink between two HG helices at the junction of two intermolecular G–C bps formed by the sticky ends[41]. To examine whether HG bps are more generally associated with DNA bending, we manually examined all DNA structures (total of 15) containing HG bp(s) flanked by at least two WC bps.

Among these 15 DNA structures, 10 contained a single HG bp while 5 contained tandem HG bps. Interestingly, we find evidence for bending across all single and tandem HG bps. In some cases, the ability to characterize global bending is obscured by local structural distortions arising due to presence of nicks in DNA strands and protein/ligand interactions. However, clear signs of bending are observed for the complexes of DNA-MATα2 (PDBID: 1K61) (Figure 2.11), DNA-p53 (PDBID: 3KZ8) (Figure 2.11), and DNA-AlkBH2 (PDBID: 3H8O) that show very little local deviation from the reference B-form structure with superposition RMSD of H1 and H2 < 2Å (see below and Figure 2.12).

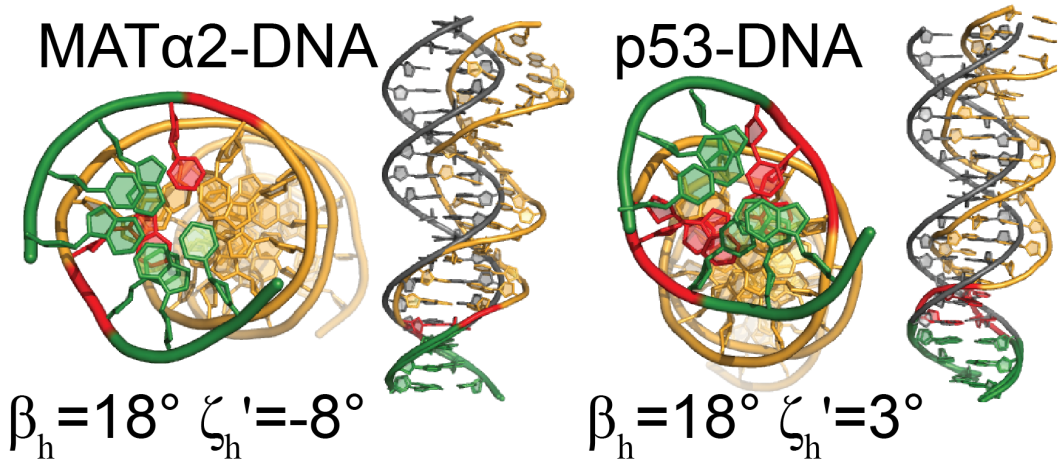

Figure 2.11: Examples showing kinking at HG bps.

PDBIDs: 1K61 and 3KZ8.

To put these observations on a quantitative footing, we adapted our approach for computing inter-helical Euler angles in RNA junctions to compute inter-helical Euler angles in DNA[87,161] (see Methods). We first benchmarked this approach on 13 DNA structures that do not contain HG bps and that show either the absence or presence of bending. The computed inter-helical angles were consistent with previous analyses of these DNA duplexes. The linear DNA duplexes yielded an average computed inter-helical bend angle of $\beta_h = 6^o$ with a narrow standard deviation ($\sigma = 2^o$) and average inter-helical twist angle of $\zeta_h = -3^o$ with a narrow $\sigma = 2^o$. The computed averages and standard deviations of β_h and ζ_h are insensitive to having one or two WC bps as the junction between H1 and H2. As expected these angles are near zero when considering the uncertainty in the computed inter-helical bend angle of $\approx 5^o$ arising due to superposition inaccuracy for short helical segments[161]. For the bent duplexes, the computed inter-helical Euler angles capture the degree and direction of bending previously reported in the solution structure of A6-T6 A-tract DNA (PDBID: 1FZX)[168] and the nucleosome DNA (PDBID: 3UT9)[169] (Supplementary Information).

Table 2: Bend angles for helical HG bps.

PDB	Residue	Sequence	H1 RMSD (Å)	H2 RMSD (Å)	α_h (°)	β_h (°)	γ_h (°)	ζ_h (°)
1K61	E. 4:10	**TGT**AATT	0.8	0.9	-18	18	46	-8
	F. 40:34	**ACA**TTAA						
3KZ8	C. 2:9	**GGC**ATGCC	0.8	0.8	20	18	56	3
	D.19:12	**CCG**TACGG						
3IGL	A. 3:10	**GGC**ATGCC	0.8	0.8	20	18	56	4
	B. 10:3	**CCG**TACGG						
1QNB	E. 207:213	**AAT**GGGC	2.4	1.3	17	60	-1	-21
	F. 222:216	**TTA**CCCG						
1QN3	E. 206:213	**AAA**CGGGC	2.9	1.2	33	77	10	-29
	F. 223:216	**TTT**GCCCG						
1QN6	E. 207:213	**ATA**GGGC	2.4	1.1	17	61	-3	-22
	F. 222:216	**TAT**CCCG						
1IHF	C. -32:-26	**AGC**AATG	2.8	0.7	20	57	-9	-25
	D. 32:26	**TCG**(n)TTAC						
3H8O	B. 266:270	**AT**(MA7)GC	1.2	0.8	-7	19	41	-3
	C. 278:274	**TA**TCG						
1VS2	A. 2:7	**CG**TACG	2.9	2.9	-16	23	20	-69
	B. 7:2	**GC**ATGC						
1XVK	A. 2:7	**CG**TACG	3.1	3.1	-14	24	22	-65
	B. 7:2	**GC**ATGC						
2PIS	C. 3:7	**CG**AA(FFD)	0.8	0.9	-29	13	61	-4
	D. 24:20	**GC**TT(FFD)						
2PIS	D. 17:21	**GA**A(FFD)T	0.6	0.9	10	6	29	2
	C. 10:6	**CT**T(FFD)A						
1LWW	E. 26:30	**CT**ACC	1.0	1.0	-10	7	48	2
	D. 5:1	**GA**TGG						
4I2O	X. 3:7	**CT**ATC	0.7	0.8	-162	11	-160	1
	W. 26:22	**GA**TAG						
329D	B. 20:24	**GC**GGT	0.7	0.8	29	12	7	0
	A. 5:1	**CG**CCA						

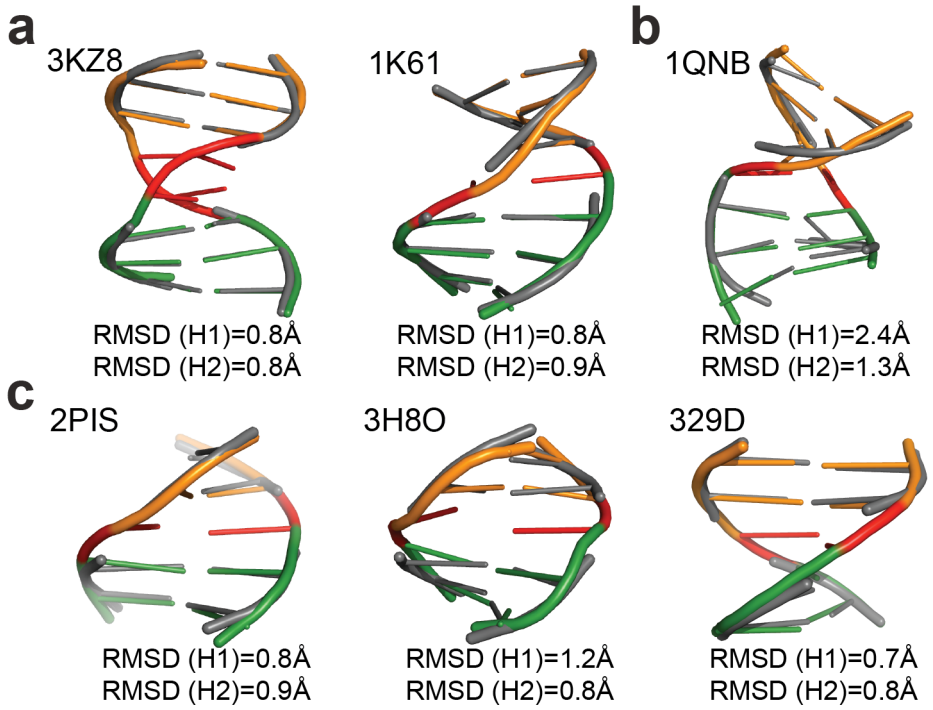

Figure 2.12: Deviation of helices next to HG bp from idealized B-form reference.

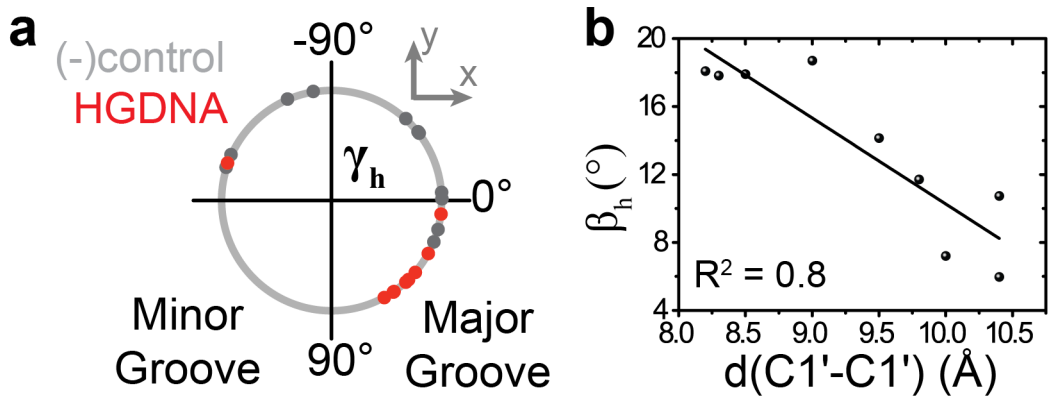

Figure 2.13: Major-groove directed bending and correlation between HG bend angle and the C1′–C1′ distance.

By comparison with the negative control results ($\beta_h = 6^o \pm 2^o$), we observed consistently higher degrees of bending in the HG-containing duplexes overall (average $\beta_h = 37^o$ with a range of $\sigma = 23^o$) (Table 2). Substantial bending is observed for structures that show little local distortion (superposition RMSD < 2Å) from the idealized B-form helix (average $\beta_h = 14^o$ with $\sigma = 5^o$) (Table 2). A smaller degree of bending is observed for two A–T bps in two similar DNA-p53 complexes when they adopt WC ($\beta_h \approx 8^o$ in structure with PDBID: 3KMD) rather than HG ($\beta_h \approx 18^o$ in structure with PDBID: 3KZ8) geometry. Significant bending ($\beta_h \approx 14^o$) is also observed at a HG-like bp, which doesn't satisfy HG H-bonding found in a naked DNA duplex (PDBID: 2PIS) containing the modification of 3-fluorobenzene[176].

These results imply that HG bps are associated with a modest degree of DNA bending though contributions from environmental factors such as protein/ligand interaction and crystal packing cannot be ruled out. This is consistent with spin relaxation dispersion showing that the population of transient HG bps in CA steps increases in longer A-tracts, which are known to induce DNA bending[66]. This increase in global bending may arise in part from the correlated local variations in sugar and phosphodiester torsion angles identified by REsemble (Figure 2.9). It should be noted that the computed inter-helical bend angles are in principle subject to uncertainties arising due to small number of bps flanking the HG bp available for superposition as well as local distortions in the helical structure (see Methods). However, control

calculations examining the effect of number of bps and the types of atoms used in the superposition suggest that the observed bending can be robustly defined (see Methods).

Interestingly, the HG bending is consistently directed towards the major groove ($-90° \leq \gamma_h \leq 90°$) as compared to the more random bending directions observed for the negative control linear B-form DNA structures (Figure 2.13). This directional bending is observed consistently across naked DNA and diverse DNA-protein complexes containing HG and HG-like bps (Table 2) under different crystallization conditions. The only exception is a HG-like bp (PDBID: 4I2O) showing bending towards the minor groove that also features the least constricted C1′–C1′ distance (see below). In addition, we observe an inverse correlation ($R^2 \approx 0.8$) between the degree of bending (β_h) and the C1′–C1′ distance of HG and HG-like bps in intact DNA helices (Figure 2.13). This trend holds even for duplexes that have only minor local distortions (superposition RMSD to idealized B-form < 2Å). This suggests a mechanism for correlating HG and bending. In particular, it becomes increasingly difficult to accommodate a WC bp geometry under more constricted C1′–C1′ distances due to steric clashes that arise between the bases. This steric clash can be released by forming HG bps, which in turn makes possible a range of conformations with variable constricted C1′–C1′ distances, sugar-backbone distortions, and DNA bend angles. It is interesting to note that constricted C1′–C1′ distances naturally lead to a narrowed minor groove of the bp which can result in favourable electrostatic interactions through minor groove recognition as observed for

example with Arg248 in the structure of the p53-DNA complex[69,177]. It is also important to note that DNA bending does not necessarily require constriction of the C1′–C1′ distance; bending may arise also due to local translation or rotation of WC base-pair steps (e.g. in roll, tilt and/or propeller twist) as proposed for bending in the nucleosome DNA[169] and A-tract DNA[168,178]. For example, the average C1′–C1′ distance in the curved nucleosome structure (PDBID: 1KX5) is ≈10.6Å ± 0.3Å.

In contrast to the bend angle, HG bps do not apparently lead to significant changes in the inter-helical twist angle ζ_h (see Methods). For example, a single HG bp leads to over-twisting by ≈8º in the complex structure of DNA-MATα2 (PDBID: 1K61) and ≈3º in the complex structure of DNA-AlkBH2 (PDBID: 3H8O) (Table 2). Additional data is needed to examine the effects of single HG bps, though based on these two observations, one would predict that single HG bps are favored by positive DNA supercoiling as arises for example, in front of RNA polymerases during transcription and in front of the replication fork during DNA replication. On the other hand, the observed inter-helical twist angle (ζ_h = -3º) across tandem HG bps in the p53-DNA complex (Figure 2.11) is within error of control B-form DNA structures (ζ_h = -3º ± 2º); thus no evident tendency to over- or under- twisting of tandem HG bps can be concluded from the current survey. Similarly in five other structures containing single HG-like bps (Table 2), the computed inter-helical twist angles range from -4º to 2º, which shows no preference for over- or under- twisting.

HG-mediated DNA bending may provide a new mechanism for indirect DNA sequence-specific recognition. Many DNA binding proteins bind DNA as oligomers and interact with multiple sites along the DNA. HG-induced bending could play topological roles defining the geometrical presentation of distant binding sites along the DNA duplex.

2.4 Supplementary Information

2.4.1 Multi-dimensional REsemble analysis

The multi-dimensional REsemble approach was used to compare distributions of local structural parameters between HG bps or adjacent WC bps and control WC bps. The 6D, 5D and 6D REsemble analyses (see below for the procedure) were carried out for six backbone torsion angles (α β, γ, δ, ε, ζ), five sugar dihedral angles (ν_0, ν_1, ν_2, ν_3, ν_4) and six local base-pair parameters (shear, stretch, stagger, buckle, propeller, opening) respectively. Instead of binning 1D data, multi-dimensional REsemble requires binning in multiple dimensions simultaneously. For example, six-dimensional REsemble analysis of backbone dihedral angles (α, β, γ, δ, ε, ζ) for any two datasets (i.e. HG bps against control WC bps) includes the following steps[159]:

(i) bin the six-dimensional space by the given bin size *m*: if *m*=15°, the total number of bins is $K = (\frac{360}{15})^6$;

(ii) count the number of 6D data points (i.e. six backbone dihedral angles) that fall into each bin for two datasets T and P i.e. HG and control WC bps, respectively;

(iii) calculate the overlap between the distributions of T and P using the square root of the Jensen-Shannon divergence (Ω^2) (See Methods, Eq (2));

(iv) vary the bin size m from 15° to 360° in increments of 15° and for each m, repeat (i) to (iii) to obtain a value of Ω for each bin size;

(v) plot Ω as a function of bin size m (Fig 3B) and calculate the normalized Ω over bin size ($\sum_K \Omega$) that measures the similarity between T and P, ranging between 0 and 1 for perfect and zero similarity, respectively;

(i)-(v) were performed to compare HG versus control WC bps ($\Sigma\Omega_{HGbp}$); WC bps adjacent to HG bps versus control WC ($\Sigma\Omega_{adWC}$); subset of control WC versus control WC bps ($\Sigma\Omega_{subWC}$), respectively.

The same procedure was used for 5D-REsemble employing five sugar dihedral angles (ν_0, ν_1, ν_2, ν_3, ν_4) and 6D-REsemble for the six local base-pair parameters (shear, stretch, stagger, buckle, propeller, opening).

2.4.2 Inter-helical analysis of bent DNA controls

The inter-helical Euler angles were computed for DNA structures that do not contain HG but show significant helical bending (see Methods), including an A6-T6 A-tract dodecamer DNA (PDBID: 1FZX)[168] and a minor-groove-inward bending sequence

in the nucleosome DNA (PDBID: 3UT9)[169]. The inter-helical Euler angles were computed across a junction of one or more bps as described in Methods for the sites of bending as reported in previous studies[168,169]. The bend angle and direction of bending (major versus minor groove) of the two structures were then directly compared with those reported previously by using other methods. We find that both the amplitude and direction of bending computed by the inter-helical Euler angle approach are in good agreement with those reported in literature.

3. Chemical shift fingerprints of Hoogsteen base pairs in DNA and DNA-protein complexes

3.1 Introduction

Prior studies have documented that there can be significant ambiguity in modeling HG versus WC bps when solving X-ray structures of DNA[68,92,158]. For example, Aggarwal and coworkers showed using X-ray crystallography and biochemical experiments that human DNA Polι, a member of the Y-family polymerases, employs HG base-paring to replicate damaged and undamaged DNA. The X-ray structure of Polι showed a template *syn* adenine in the active site forming a HG bp with an incoming dTTP. This structure was met with some skepticism. In particular Wang pointed out[92] that based on the weak electron density for the active site A–T base-pair, it is difficult to resolve a WC from a HG geometry. There have been other studies noting the difficulty in resolving HG versus WC bps in X-ray structures protein-DNA complexes including the DNA-homeodomain[68] and DNA-p73 complexes[158].

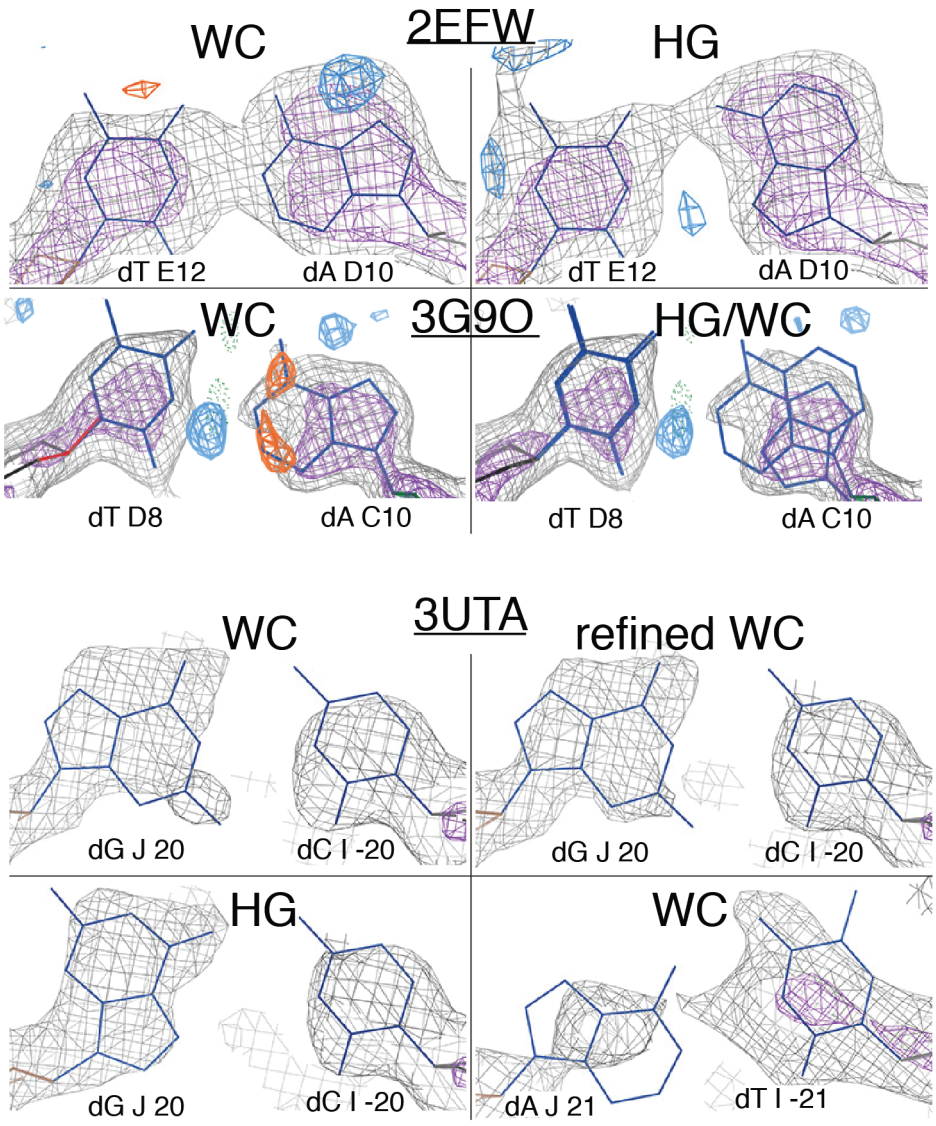

Figure 3.1: Potential HG bps mis-modeled as WC bps.

Figure courtesy of Dr. Bradley J. Hintze.

The difficulty in resolving WC versus HG bps in X-ray structures of DNA raises the possibility that there are examples in which HG bps were mistakenly modeled as WC. Dr. Bradley J. Hintze in the laboratories of Drs. Jane and David Richardson examined this possibility through the development of automated tools that can survey the protein databank in search for such modeling errors. Running these methods on DNA structures in the PDB reveals many ambiguous cases, and some examples where HG or partial HG/WC bps may have potentially been mis-modeled as WC bps.

Two examples are complexes of DNA with the replication terminator protein (RTP)(PDB: 2EFW) [179] and glucocorticoid receptor (GR) (PDB: 3G9O)[180] (Figure 3.1). Flipping the purine base followed by refinement resulted in HG bps that fit the $2mF_o$-DF_c maps better as judged by elimination or strong reduction of the difference peaks (Figure 3.1). In these models of the RTP-DNA complex, three dA–dT bps in an A-tract fully occupy HG bp whereas GR favored partial 25%: 75% HG: WC occupancies (Figure 3.1). In both cases, the R and R-free differences between the WC and HG or WC/HG refinements were negligible, which is not surprising considering the resolution and that only a few purines were modified while leaving the rest of the structure, DNA and protein, as modeled. In addition, also examined were bps in X-ray structures of the nucleosome particle [169,181,182]. Even for the highest resolution X-ray structure for which the electron density is publicly available (2.07 Å PDBID: 3UTA[169]; note that the density for the 1.94 Å structure[183] PDBID:1KX5 is not publicly available), the density around the

DNA for many bases is highly ambiguous e.g. I-21–J21 dA–dT bp (Figure 3.1). This may reflect extensive DNA dynamics. Interestingly, the I-20–J20 dG–dC bp (Figure 3.1) at the major groove-bending site was marginally modeled better as HG in comparison to WC. Again, R and R-free differences between the two refinements with the I-20–J20 dG–dC bp being WC or HG are negligible, leaving open the possibility of HG bps. Moreover, the refinement with J20-dG in the *syn* conformation showed an improved fit to the density (Figure 3.1).

Even if HG and WC bps can be resolved by X-ray maps, one cannot rule out that crystal packing forces or unique aspects of the crystallization conditions (temperature, co-solutes, metals) bias the WC-HG equilibrium. Therefore, there is a need for independent approaches for characterizing WC versus HG bps and WC-HG dynamics under solution conditions.

Unlike X-ray crystallography, WC and HG bps can in principle readily be resolved with the use of NMR spectroscopy[66]. In particular, unique NMR chemical shift signatures as well as NOE-based distance connectivity can be employed to distinguish WC versus HG bps. In addition, relaxation dispersion based approaches can in principle be used to characterize WC-HG dynamics. However, application of these NMR-based approaches can prove difficult for large DNA-protein complexes owing to severe spectral overlap combined with unfavorable relaxation properties owing to the large molecular weight of these complexes; which results in significant line-broadening and

losses in sensitivity. Here, we propose an approach involving the use of nucleotide specific $^{13}C/^{15}N$ labeling DNA samples in concert with 7-deazapurine-substitution binding measurements for characterizing HG bps in large DNA-protein complexes. The approach is demonstrated on the integration host factor (IHF)-DNA complex for which prior X-ray studies provided evidence for a HG bp.

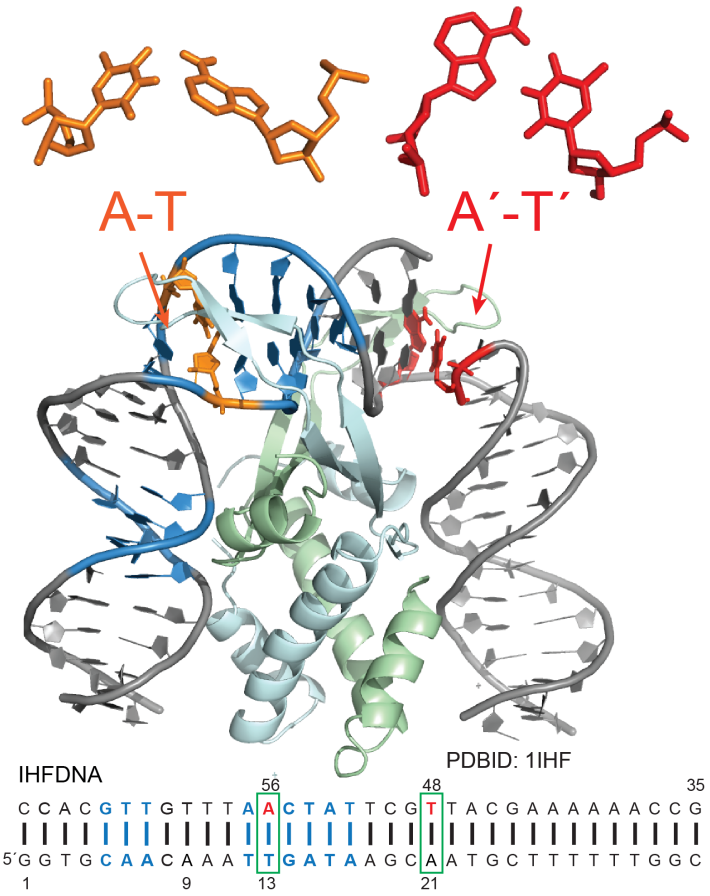

Figure 3.2: X-ray structure of the IHF-DNA complex.

The consensus sequence for specific protein recognition is shown in blue with the two sharp kinks highlighted by the green rectangle.

IHF is an architectural protein[184] that aids in the compaction of prokaryotic genomes by inducing significant DNA bending. IHF was recently found to play essential roles in the Cas1-Cas2-medilated spacer integration by binding and bending the CRISPR leader sequence[185]. It binds to ≈35 bp DNA duplexes containing the consensus sequence WATC**A**RNNNNTTR, where W = A or T; N = A, T, C or G; R = A or G; **A** = site of suspected HG. IHF specifically binds the H′ site of bacteriophage λ (K_d ≈10^{-9} M) with 10^3-10^4 greater affinity than a random sequence [186,187]. A high resolution X-ray structure of the IHF-DNA complex reveals little to no direct sequence-specific contacts with the DNA indicating that sequence-specificity is mainly achieved by indirect readout mechanisms[62,184]. The IHF-DNA crystal structure (PDB: 1IHF) features a HG bp adjacent to a nicked site that was introduced to aid crystallization[62] (Figure 3.2). However, a distorted WC A–T bp is observed at the pseudo-symmetry related site, which does not have an adjacent nick (Figure 3.2). The HG bp is associated with an unusual backbone of the terminal-like dT, resulting the base orientations similar to those in Z-form DNA with the purine being *anti* rather than *syn* conformation; interestingly, the sugar pucker of dT is C3′-endo. H-bonding is observed between the A-N3 in the HG bp with a nearby arginine residue (Arg62); a similar interaction is also found between the A-N3 in the distorted A–T WC bp at the pseudo-symmetric site and Arg63 which could help stabilize the *anti* A conformation.

In this chapter, we developed an NMR approach to characterize HG bps in large protein-DNA complexes, which uses chemically synthesized DNA samples with site-specifically $^{13}C/^{15}N$ labeled residues to overcome the spectral overlap problem. This allows robust measurements of NMR chemical shifts and other parameters, which can be used to resolve WC versus HG bps as well as help define their dynamic equilibrium [66,112]. We used this approach to examine two bps in a ≈40 kDa complex between a highly bent 34 bp duplex DNA and the integration host factor (IHF) protein (Figure 3.2) [62]. We combined this NMR-based approach with deazapurine-substitution binding affinity measurements to further explore the importance of HG bps in determining the DNA-protein binding affinity.

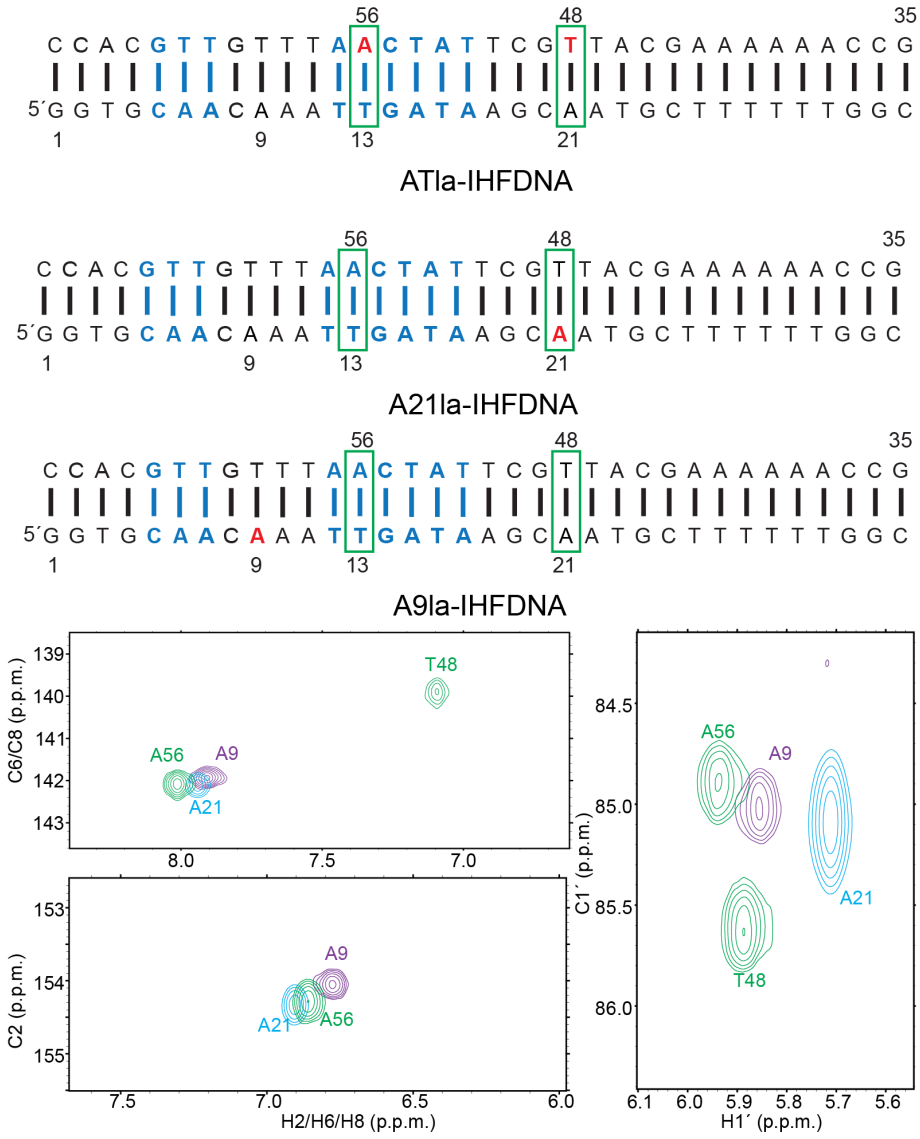

Figure 3.3: Site-specifically $^{13}C/^{15}N$ labeled IHF-DNA.

Residues with $^{13}C/^{15}N$ labeling are colored red in the DNA sequence.

3.2 Materials and Methods

3.2.1 Sample preparation

NMR buffer: All DNA or DNA-protein complex samples were buffer exchanged into the HEPEs buffer consisting of 25mM HEPES, 100 mM NaCl, 0.1 mM EDTA with pH 7.0 and 10% D_2O three times using a centrifugal concentrator (EMD Millipore with 3kDa cut-off) until containing >99.9% of the desired buffer.

DNA and IHF protein samples: IHF-DNA strands with single (A21 or A9) or double (A56+T48) sites uniformly $^{13}C/^{15}N$-labeled, and those with 7-deazaA21 or 7-deazaA56 were purchased from the Keck Oligo Synthesis Resource (W.M. Keck Foundation) with Glen-Pak™ DNA cartridge purification. Unlabeled DNA and 5'-Fluorescein-dT labeled strands, were purchased from IDT (Integrated DNA Technologies, Inc.). The IHF protein sample (≈1 mM) was provided by Dr. Ying Zhang (the laboratory of Dr. Phoebe Rice, University of Chicago) in the initial storage solution consisting the HEPEs buffer and 20% glycerol to maintain protein stability.

DNA-IHF complex for NMR: DNA and protein samples were diluted to below 0.05 mM with ≈ 1.8 mL volume with 1:1 molar ratio; diluted DNA was titrated into the protein sample (≈0.6 mL each time, 3 times with 5 min time interval) with slow pipette mixing. During this process, the solution appeared clear and no visible precipitation formed. The complex solution was then subjected to buffer exchange and yielded the NMR sample containing 0.3 mM DNA-IHF complex in the HEPEs buffer.

3.2.2 Fluorescence polarization assay for measurement of binding affinity

The fluorescence polarization (FP) assay was set up on a Clariostar monochromator microplate reader (BMG Labtech) with a black polypropylene 384 well plate (Corning Inc.). The DNA strand that does not contain any chemical modifications is labeled with fluorescein-dT at the 5'-end of the sequence and then annealed with the complementary strand to make the duplex DNA. The concentration of fluorescein-labeled DNA was kept constant (at 0.05, 0.5 or 1.5 nM) with the protein concentration varying from 0.1 – 30 nM. The FP assay was carried out at room temperature with the excitation wavelength set at ≈ 482nm and emission wavelength of ≈ 540nm for detection. FP values were calculated by the detected parallel and perpendicular light intensities by Equation (3.1).

$$FP = \frac{I_{//} - I_{\perp}}{I_{//} + I_{\perp}} \tag{3.1}$$

The FP values (*FP*) and protein concentration (x, nM) were then fit by the quadratic Equation (3.2), where D is the constant DNA concentration (in nM) and A,B are constant parameters, to obtain the dissociation constant (K_d) for the binding.

$$FP = A + B \times (D + K_d + x - \sqrt{(D + K_d + x)^2 - 4Dx}) \tag{3.2}$$

3.2.3 NMR experiments

NMR data were collected on a 700 Bruker Avance III spectrometer equipped with a triple-resonance HCN cryogenic probe, and 600 MHz Bruker NMR spectrometer equipped with an HCN cryogenic probe. Data were processed and analyzed using NMRpipe[188] and SPARKY (T. D. Goddard and D. G. Kneller, SPARKY 3, University of California, San Francisco), respectively. Chemical shifts data were obtained using 2D HSQC, TROSY; resonance assignments were analyzed using ^{15}N-edited NOESY and HCN experiments.

3.3 Results

3.3.1 NMR Characterization of site-specifically labeled DNA duplex

We obtained DNA duplex samples lacking any nicks in which either A21 (**A'**) or A56 (**A**)+T48 (**T'**) were uniformly ^{13}C/^{15}N labeled (Figure 3.3) and probed the conformation of these dA–dT bps. All of the latter residues are located in a region of sharp DNA kinking. We also obtained DNA samples with labeled residue A9 that is located outside an area of sharp kinking (Figure 3.3). The 1D ^{1}H spectrum of these three DNA duplexes were as expected, identical with well resolved imino proton resonances that are consistent with a WC B-form DNA duplex. Excellent quality 2D NH and CH HSQC TROSY-based NMR spectra could be obtained for all three samples clearly showing only a subset of resonances involving the labeled nucleotides (Figure 3.3). The

^{1}H, ^{13}C, and ^{15}N chemical shifts were all consistent with the values expected for a WC B-DNA double helix. Interestingly, the **A** pseudo-symmetry related **A'** exhibit different chemical shifts consistent with their distinct environments (Figure 3.3). These results demonstrate the feasibility of studying much larger DNA duplexes by NMR than is typically possible with the use of nucleotide-specifically labeled samples.

To examine whether WC⇌HG exchange also occurs in such large DNA duplexes in the absence of bound protein, we carried out RD measurements on residue A56. Indeed, we observed significant RD at A56-C1′ (Figure 3.4) that are consistent with transient HG bps at 37 °C (p = 0.33±0.06%, k_{ex} = 1482±479 s^{-1} and Δω = 4.17±0.28 p.p.m.). These results demonstrate that WC⇌HG equilibrium exists robustly in DNA duplexes and are not restricted to short duplexes that have been typically studied using NMR.

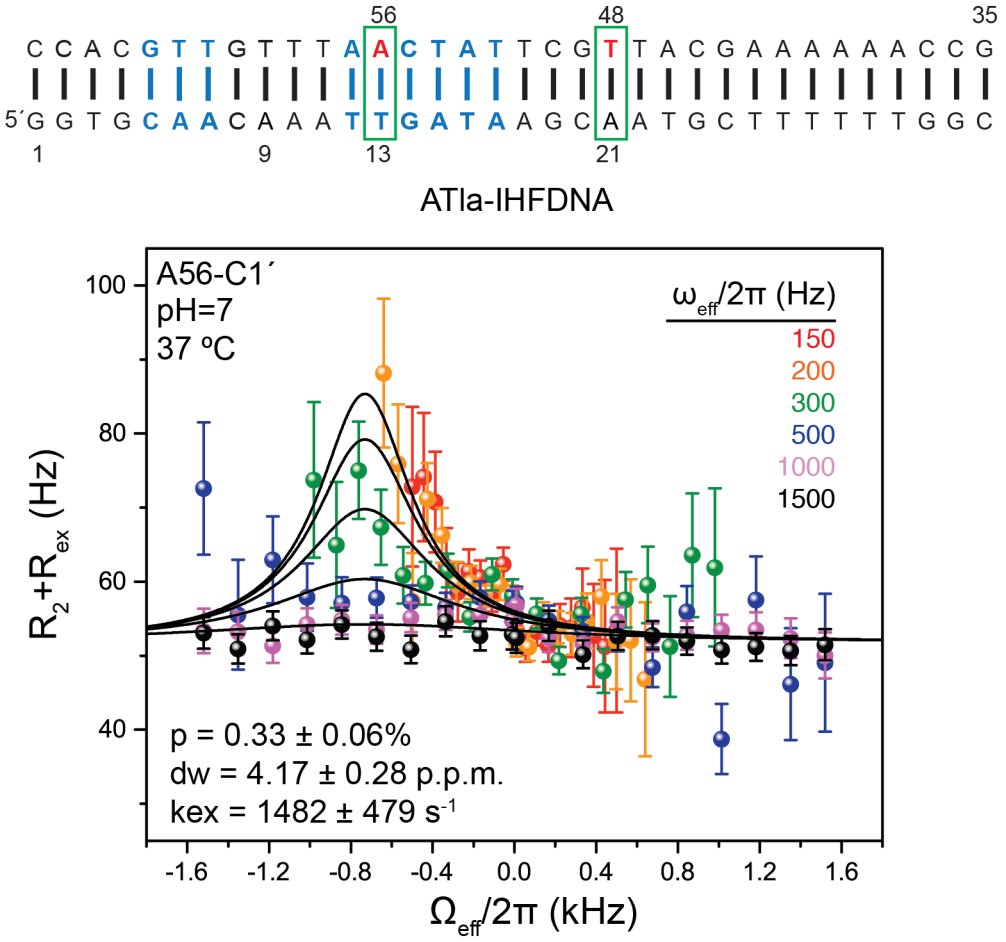

Figure 3.4: Transient HG on A56-C1′ in ATla-IHFDNA.

Off-resonance RD profiles showing R_2+R_{ex} as a function of spin lock offset (Ω_{eff} $2\pi^{-1}$Hz) and power (ω_{eff} $2\pi^{-1}$Hz, in insets). Error bars represent experimental uncertainty (one s.d.) estimated from mono-exponential fitting of n = 2 independently measured peak intensities using a Monte-Carlo based method.

3.3.2 Characterization of IHF-DNA complex formation

We used three approaches to characterize complex formation between the DNA duplexes and IHF protein (provided by Dr. Ying Zhang from the Phoebe lab, University of Chicago). First, we used an electrophoretic mobility shift assay (EMSA) and observed the expected migration retardation when mixing the DNA duplexes with increasing concentration of IHF with saturation observed at 1:1 ratios consistent with the high affinity K_d of ≈10^{-9} M (Figure 3.5). Second, using a fluorescence polarization (FP) assay in which the DNA duplex is tagged with 5′-fluorescein-dT, we measured a K_d = 1.76±0.92 nM that is in excellent agreement with published values (≈1.9 nM)[189] (Figure 3.5).

Finally, we characterized IHF-DNA complex formation using NMR. 1D ^{1}H NMR spectra revealed clear changes in the DNA spectrum upon addition of IHF consistent with complex formation and large changes in the DNA structure due to severe DNA bending (Figure 3.5). Complex formation resulted in large chemical shift perturbations (CSPs) in DNA resonances that are very similar for **A** and **A′** indicating that the two adenines adopt similar conformations in the complex (Figure 3.5). We observed sharp T′-H3 imino proton resonance at 37 °C, consistent with stable WC or HG **A′–T′** H-bonding and inconsistent with significant base opening or bp distortions which would lead to line broadening due to solvent exchange.

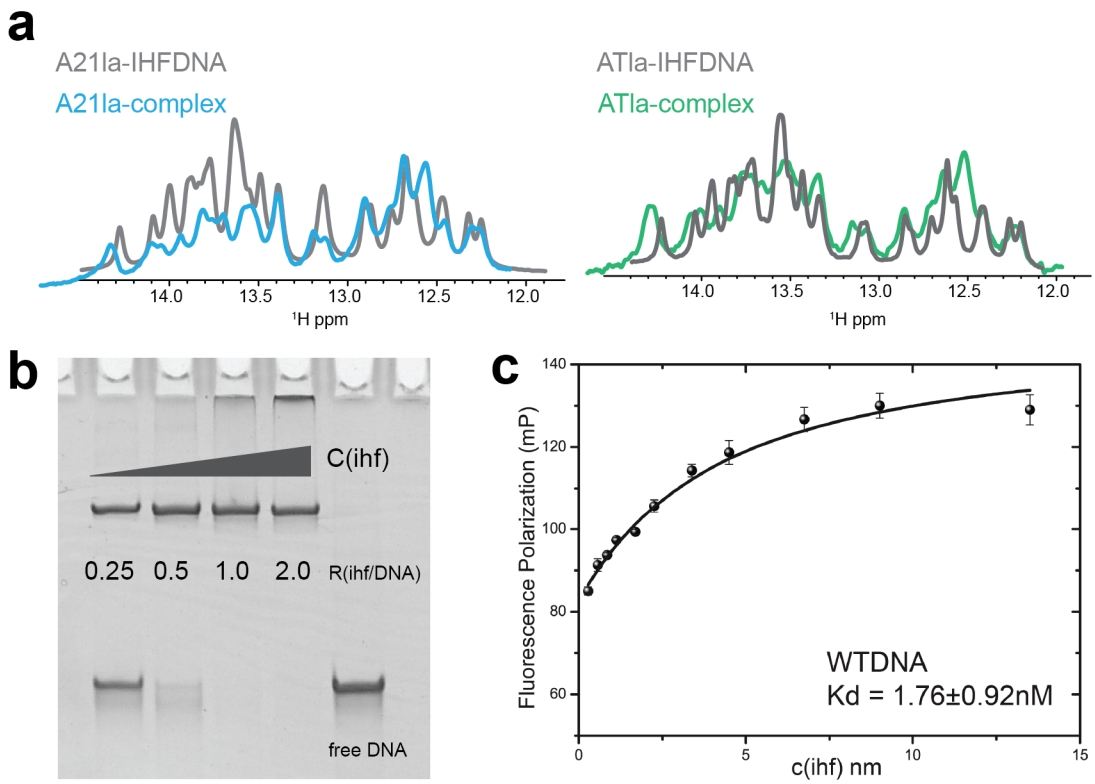

Figure 3.5: Formation of IHF-DNA complex by ^{1}H NMR, EMSA and FP assays.

(a) Imino ^{1}H NMR spectra overlay between free DNA and DNA-IHF complexes. (b) EMSA result for 1:1 binding stoichiometry (c) FP binding curve for DNA-IHF complex.

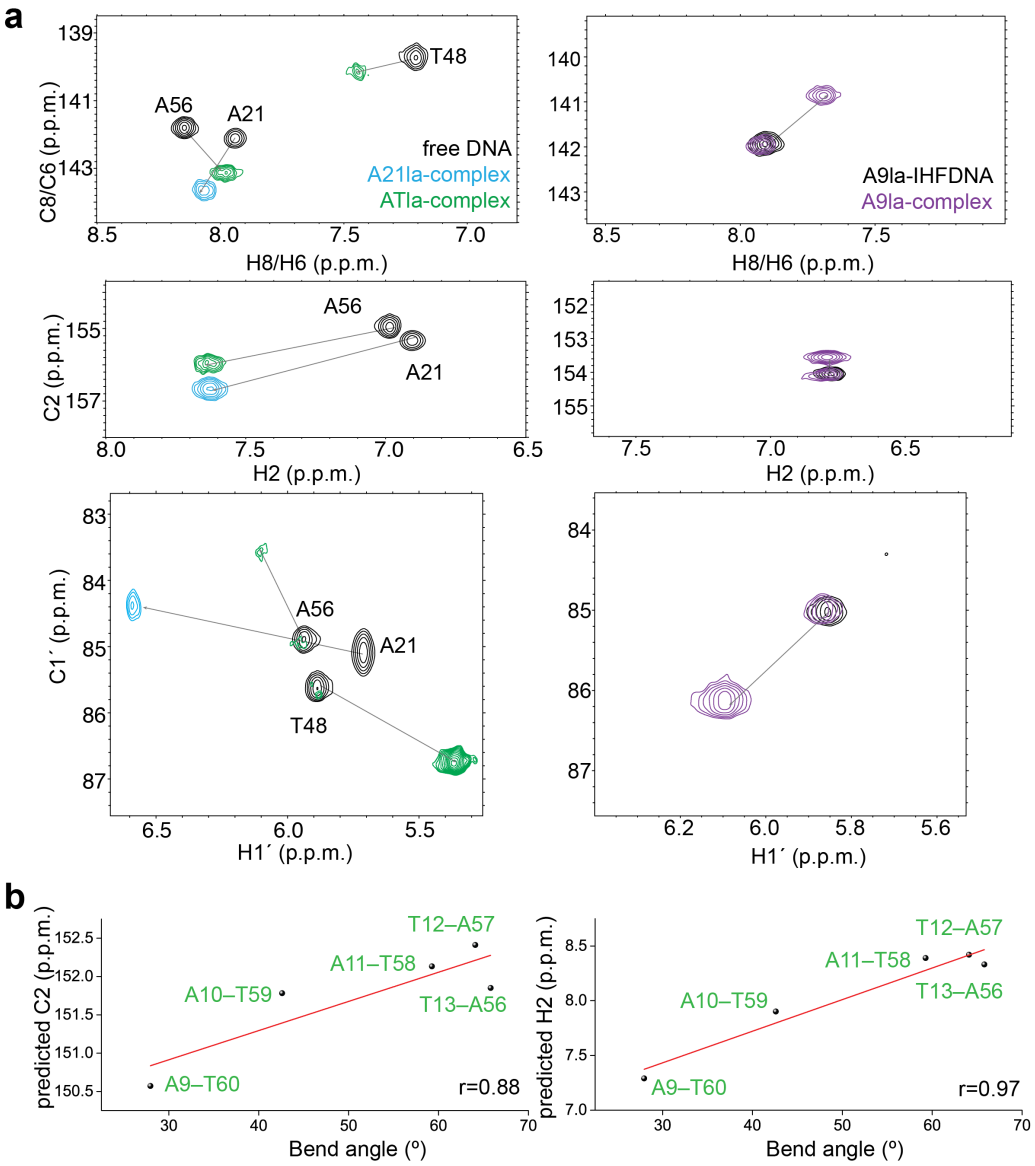

Figure 3.6: Chemical shift perturbations on DNA upon IHF protein binding.

(a) NMR spectra of chemical shift perturbations upon complex formation. (b) Correlation of adenine-C2H2 chemical shifts predicted by QMMM based on the X-ray structure (PDBID: 1IHF) with the bend angle computed at the corresponding bp. Pearson's r is shown for each linear correlation.

Importantly, while we do observe a downfield shift in the A'-C8 and A-C8 which are consistent with a HG bp, the magnitude of the shift (≈ 1.5 ppm) is smaller than would be expected (≈ 3 ppm). Furthermore, an upfield shift (≈ 0.7 ppm) is observed for the **A'**-C1' and **A**-C1' which is opposite what would be expected for a HG bp (Figure 3.6). This analysis should however be interpreted cautiously. In particular, our expectations for HG bps are based on HG bps within B-form duplexes. There can be other contributions toward the chemical shift arising due to protein-DNA contacts and other changes in the DNA structure, particularly the severe bending induced by IHF binding. Indeed, we observe significant perturbations for the control A9-C8 and A9-C1' which adopts a WC bps at site with minimal kinking (Figure 3.6) showing that complexation alone can induce significant changes in the chemical shift. Finally, we note that we observe significant changes in the base **A'**-C2 and **A**-C2 but not for the control site A9-C2 (Figure 3.6). These perturbations may reflect DNA bending as inferred from QM/MM calculations (in collaboration with the laboratory of Dr. David Case, Rutgers University) preformed on the IHF-DNA complex (Figure 3.6). These QM/MM calculations accurately predict A–T and G–C HG chemical shift perturbations in the A_6-DNA duplex (data not shown).

3.3.3 NOESY data reveal a WC bp for A'-T' in the DNA-IHF complex.

We carried out ^{15}N-edited NOESY experiments (Figure 3.7) to more directly examine obtain structural information regarding the nature of base pairing at **A′–T′** in the DNA-IHF complex. Interestingly, we observed all of the NOE distance-based connectivity expected for WC pairing (A21H2–T48H3). We do not observe NOE cross peaks expected for HG bps (A21H8–T48H3). These results, together with CSPs and imino resonance of T48 suggest that **A′–T′** adopts a WC bp in the IHF-DNA complex.

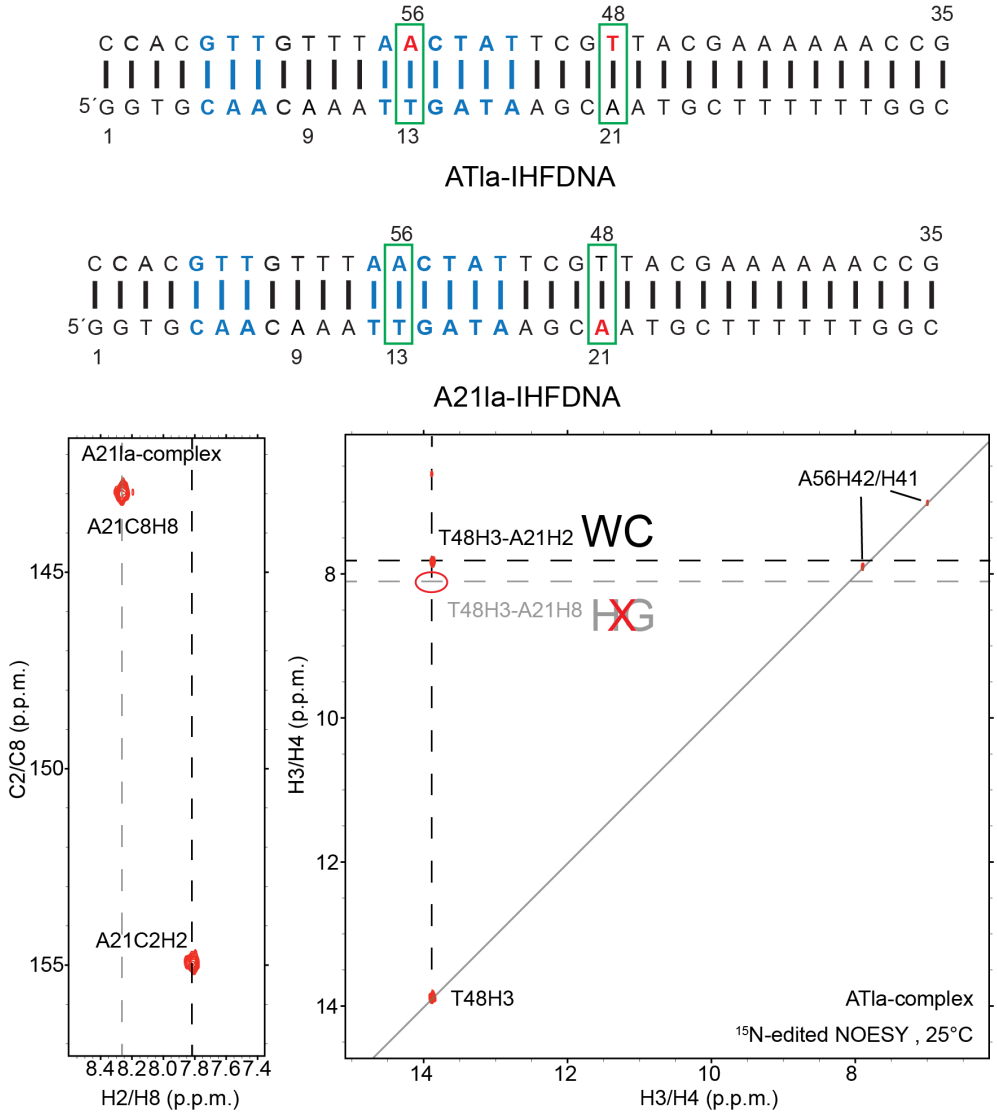

Figure 3.7: NOE evidence for WC pairing at the A′–T′ site.

3.3.4 Chemical exchange in IHF-DNA complex.

We applied $R_{1\rho}$ RD NMR experiment[136,137,141] to transient HG bp formation at the kink site in the IHF-DNA complex. Due to the low sensitivity arising due to the low sample concentration (≈0.3 mM) and large molecular size (≈ 40 kDa), a limited amount of RD data were collected (Figure 3.8). Although the data were not sufficient to quality to allow for accurate fitting with the 2-state BM equation, we clearly observe enhanced chemical exchange at the **A′**-C1′ (Figure 3.8) with elevated p_{ES} (≈ 50%) with a downfield shifted Δω (≈ 1.1 p.p.m.) which is smaller in magnitude that would be expected for a HG bp in duplex DNA[66,116]. This unique chemical shift could arise from the unique environment in the DNA-IHF complex or could reflect fast exchange that would reduce Δω by a factor of 2. In contrast, we did not observe detectable RD on the asymmetric kink site **A**-C8 (Figure 3.8). **A** site locates in the consensus sequence while the **A′** site does not. This contrasting observation of chemical exchange on **A** versus **A′** sites could reveal different dynamics in protein-DNA interactions. One hypothesis is that half DNA that has the consensus sequence is more tightly bound to the protein and shows less dynamics. We cannot exclude the possibility that the difference arises from different types of spins that were measured. The dynamics can be localized on the backbone due to the kinking but minimally gets experienced by the nucleobase.

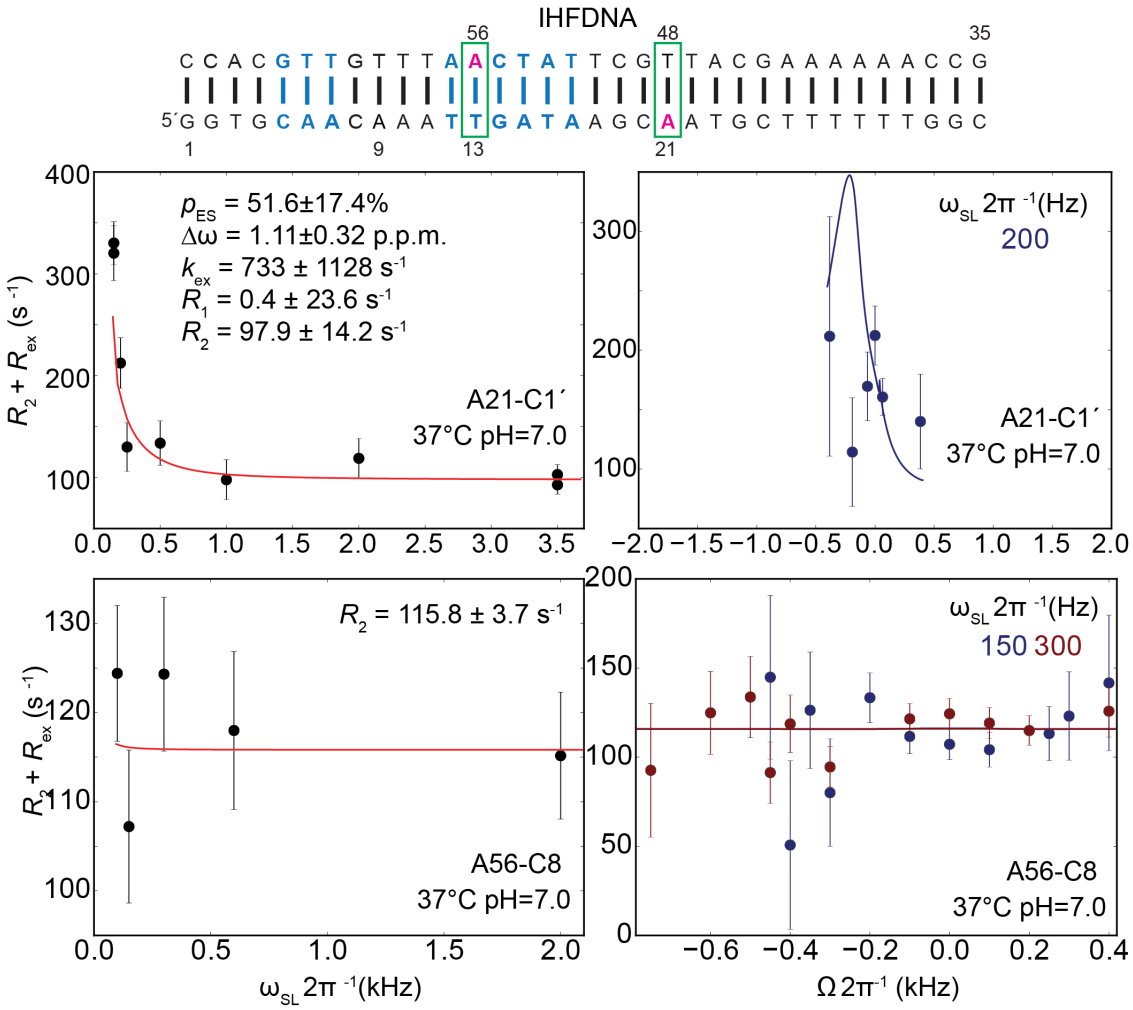

Figure 3.8: On- and Off-resonance RD profiles for A-C1′ and A-C8 in IHF-DNA complex.

DNA sequence is also shown with the residues measured in RD colored in magenta.

3.3.5 7-deaza-adenine substitutions minimally affect the IHF-DNA binding affinity.

If formation of a HG bp is an important determinant of binding affinity, substitution of a given site with 7-deazapurine should significantly destabilize the HG bp and thereby reduce the DNA-protein binding affinity. Prior studies suggest that due to the knocking out of a hydrogen bond, 7-deazapurine should destabilize the HG bp relative to the WC by at least ≈1 kcal/mol. This ten-fold reduction in the transient HG bp population should in principle translate into a ten-fold reduction in apparent DNA-protein binding affinity to a conformation that features a 100% stabilized HG bp.

To examine the importance of HG bps at either the at **A′** or **A** sites, we used the same FP assay (Figure 3.9) to measure the impact on K_d following 7-deazaadenine substitutions. Substituting either **A′** or **A** with 7-deazaadenine does not show significant impact on the measured K_d (1.8 ± 0.6 nM) relative to that of the control DNA (K_d ≈1.9 ± 0.2 nM).

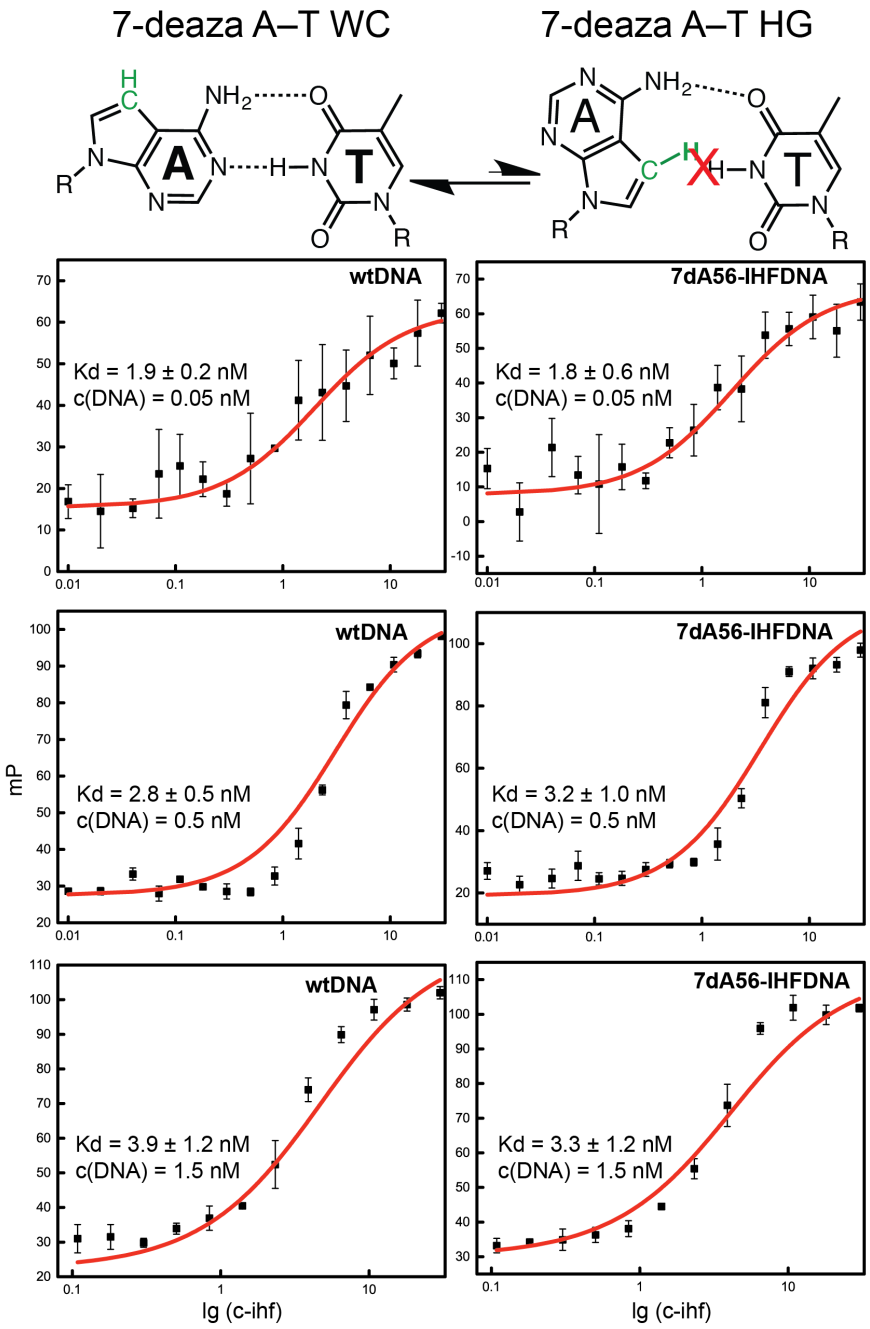

Figure 3.9: Minimal impact of 7-deazapurines on IHF-DNA binding affinity.

3.4 Discussion

Based on NMR and binding affinity measurements, our results strongly suggest that the **A′–T′** forms a WC bp in the IHF-DNA complex and not a HG bp as observed in the X-ray structure. Rather, the HG bp in DNA-IHF complex is most likely stabilized by the adjacent knick that was used to aid crystallization. Our survey of HG bps shows them to be prevalent in such environments, including for example near terminal ends. These findings highlight the importance of not only examining HG bps that could have been mis-modeled as WC, but also, inversely, to confirm the identity of HG bps modeled in X-ray structures using solution based methods.

Our results highlight the power of using site-labeled DNA samples to characterize the structural and dynamic properties of large DNA complexes that are otherwise inaccessible to conventional NMR characterization. Of particular interest is the ability to resolve WC versus HG bps or opened states, using robust NMR experiments, including measurements of imino resonances and distance-based NOESY cross peak signatures. In addition to aiding the characterization of HG bps suspected in such protein-DNA complexes, these approaches can provide unique insights into the structural and dynamic behavior of protein-DNA complexes under solution conditions. They also provide an important avenue for cataloguing valuable chemical shift structure relationships that can be harnessed in the future to aid chemical shift based characterization of DNA structure and dynamics.

While our results rule out stabilization of HG bps as a major species at **A'** site, we cannot rule out that complex formation promotes transient HG bps. Indeed, we did observe significant evidence for chemical exchange in the complex that could be attributed to WC-HG exchange. However, the poor quality of the data could not allow robust fit to the 2-state exchange model but clearly hints an enhanced population ≈50% with slow-to-intermediate exchange rate. We have evidence for such dynamics in the DNA in complex with glucocorticoid receptor based on the crystallographic analysis (Figure 3.1).

Finally, we note that binding affinity measurements with 7-deazapurine substitutions provide an even simpler approach for surveying protein-DNA complexes suspected of having HG bps. Together, the NMR and 7-deazapurine substitution binding affinity measurements provide a new basis for robustly characterizing HG bps in DNA-protein complexes under solution conditions.

4. Hoogsteen base pairs are strongly disfavored in A-form RNA duplexes.

4.1 Introduction

The Watson-Crick (WC) double helix is the most common structural element in RNA and the dominant structure of genomic DNA. It provides the basis for templated replication, transcription, and translation, and also serves as a scaffold that defines the 3D structure of DNA, RNA, and their protein complexes. The canonical double helices formed by RNA (A-form) and DNA (B-form) differ in several important respects (Figure 4.1). In B-form DNA (B-DNA), the five-membered deoxyribose ring is flexible and favors the C2′-endo sugar pucker (Figure 4.1). In contrast, due to the sugar 2′-hydroxyl group (2′-OH), the sugar in A-RNA is more rigid and adopts an alternative C3′-endo conformation[119,120] (Figure 4.1). This in turn brings the oxygen atoms (O5′ and O3′) adjoining sequential nucleotides into closer proximity effectively compressing and rigidifying the A-form helix, widening its helical diameter, and displacing base pairs (bps) away from the helical axis[119,190](Figure 4.1). In addition, B-DNA and A-RNA differ considerably with respect to their deformability, with B-DNA being generally more flexible[191]. For example, B-DNA is more bendable than A-RNA, and this property is of fundamental importance for many biochemical processes including the tight compaction of genome within the nucleus in higher order organisms.

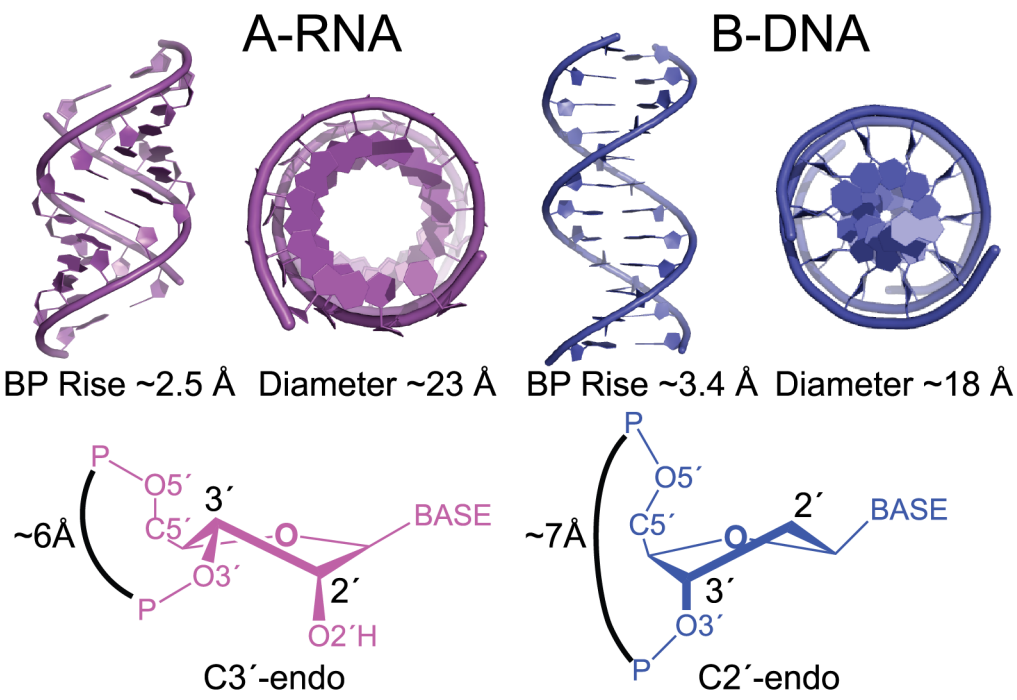

Figure 4.1: Comparison of A-form RNA and B-form DNA structures.

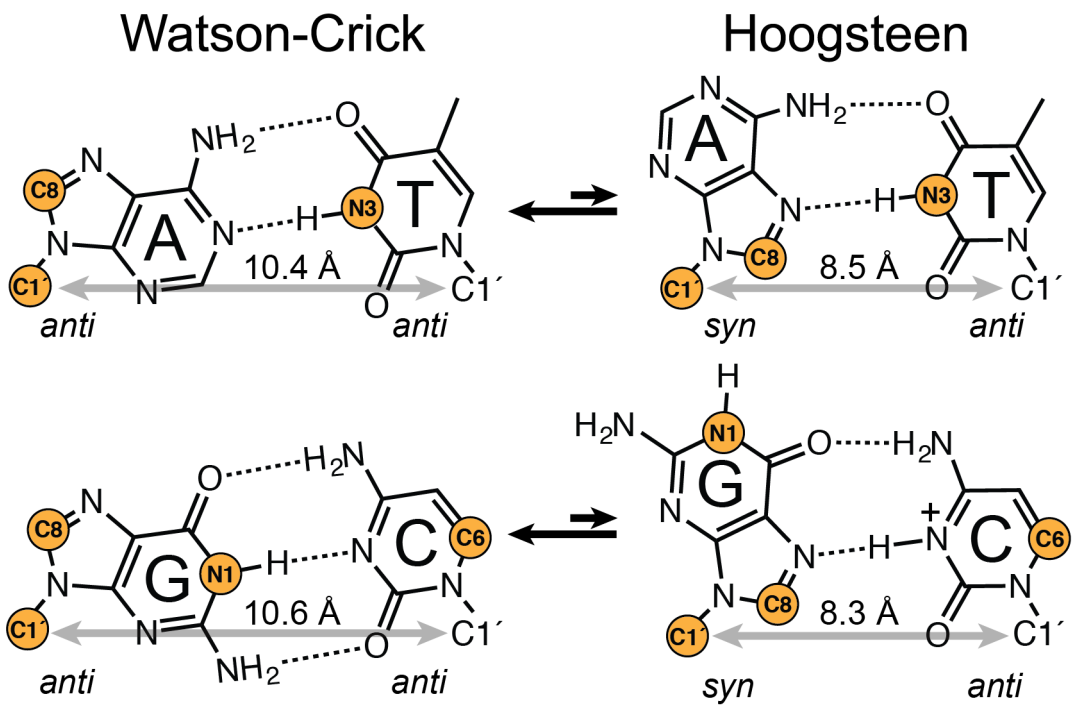

Figure 4.2: Probes for WC–HG exchange in RD measurements.

Recently, NMR studies have uncovered a new dynamic property in canonical B-DNA; Watson-Crick (WC) dG–dC and dA–dT bps exist in a dynamic equilibrium with alternative Hoogsteen (HG) bps[4,66]. A HG bp is created by rotating a WC purine base ≈180° around the glycosidic bond to adopt a *syn* rather than *anti* conformation[4] (Figure 4.2). The two bases are also brought into closer proximity by ≈2.0–2.5 Å to form a unique set of hydrogen bonds (H-bonds) (Figure 4.2). HG bps exist transiently (lifetimes typically 0.1–1 ms) and in low abundance (populations typically <3%) in naked canonical B-DNA[66,116]. However, dA–dT and dG–dC$^+$ HG bps can become the dominant configuration (for review see ref. [26]) in DNA-protein[69] and DNA-small molecule[47] complexes where they contribute to DNA recognition, in damaged nucleotides where they contribute to damage accommodation and repair[83,192,193], and in the active sites of translesion synthesis polymerases that use HG pairing to bypass damage during DNA replication[100]. Purine-purine HG bps have also been shown to play important roles in DNA replication errors and in DNA damage accommodation and repair[194,195].

Here, we set out to study WC-HG dynamics in canonical A-RNA duplexes. We show that unlike the canonical B-DNA double helix, rA–rU and rG–rC$^+$ HG bps are strongly disfavored in A-RNA duplexes. As a result, while the DNA double helix can absorb damaged nucleotides incapable of forming WC bps such as N^1-methyl deoxyadenosine (m^{1}dA) and N^1-methyl deoxyguanosine (m^{1}dG) by forming HG bps; the same methyl marks, N^1-methyl adenosine (m^{1}rA) and N^1-methyl guanosine (m^{1}rG),

acting as a posttranscriptional modification in RNA, block base pairing altogether. This provides a direct mechanism for potently modulating the structure of the epitranscriptome. Our results indicate that HG-dependent DNA biochemical transactions may not be as readily supported in RNA duplexes and identify a unique dynamic property in B-DNA that may help enhance its ability to function as the repository of genetic information.

4.2 Methods

4.2.1 Sample preparation

NMR buffer: All RNA and DNA samples were buffer exchanged at least three times using a centrifugal concentrator (EMD Millipore) until containing >99.9% of the desired buffer, which unless stated otherwise, consisted of 15mM sodium phosphate, 25 mM NaCl, 0.1 mM EDTA with pH 5.4 or 6.8 and 10% D_2O.

Uniformly $^{13}C/^{15}N$ labeled RNA and DNA samples: hp-A_6-RNA and single-strands of the E-gc and TAR-UUCG^{GU} were prepared using *in vitro* transcription with uniformly $^{13}C/^{15}N$-labeled ribonucleotide triphosphates (Cambridge Isotope Laboratories), T7 polymerase (Takara Mirus Bio Inc.) and synthetic DNA templates (Integrated DNA Technologies, Inc.), purified by 20% (w/v) denaturing polyacrylamide gel electrophoresis (PAGE) and electro-eluted into 20 mM Tris buffer (pH = 8) followed by ethanol precipitation as described previously[196]. The uniformly labeled T6-strand in

A_6-DNA^{rG} and uniformly labeled A_6-DNA were prepared by the primer-extension approach[197] using uniformly $^{13}C/^{15}N$-labeled deoxyribonucleotide triphosphates (Silantes) as previously described[66].

m^1A and m^1G containing oligonucleotides: Oligonucleotides were purchased from Keck Oligo Synthesis Resource (W.M. Keck Foundation) with Glen-Pak™ DNA/RNA cartridge purification (A_6-RNA^{m1A}, A_2-RNA^{m1A}, gc-RNA^{m1A}, A_6-DNA^{m1G}, A_6-DNA^{m1rA}, gc-DNA^{m1A}, hp-A_6-RNA^{m1A}, hp-A_6-DNA^{m1A}, hp-A_6-DNA^{m1G}, hp-gc-RNA^{m1A}, and hp-gc-DNA^{m1A}) Midland Certified Reagents with reverse-phase (RP) HPLC purification (A_6-DNA^{m1A} and A_2-DNA^{m1A}), and GE Healthcare Dharmacon Inc. with RP-HPLC purification (A_6-RNA^{m1G}, A_6-DNA^{m1rG} and hp-A_6-RNA^{m1G}). To minimize Dimroth rearrangement of m^1A into m^6A[198,199], all DNA and RNA oligonucleotides containing m^1A were synthesized and deprotected using the UltraMild protocol (http://www.glenresearch.com/Technical/TB_UltraMild_Deprotection.pdf; Glen Research Corporation).

Assessing purity of m^1A and m^1G containing oligonucleotides: Samples were assessed using 20% denaturing PAGE, MALDI Mass Spectrometry, Liquid Chromatography-Mass Spectrometry (LC-MS) and NMR spectroscopy. For hairpin constructs hp-gc-RNA^{m1A} and hp-A_6-RNA^{m1A}, we obtained evidence for incomplete base deprotection during synthesis based on observation of additional imino proton and acetyl group (the N4 protecting group on the cytosine) resonances and NOE talk

between the two. Evidence for the acetyl groups was also obtained by quantitative mass spectrometry (LC-MS). We suspect that incomplete deprotection arises due to formation of stable secondary structure in these hairpin constructs during the UltraMild deprotection step. These impurities could be effectively eliminated by synthesizing individual single strands of duplex versions of the hairpin sequence (gc-RNA^{m1A} and A$_6$-RNA^{m1A}).

In all cases, the NMR chemical shifts of the *N*^{1}-methyl group and base moieties (A-C2, N1C, N1H) were consistent with m^{1}A with no evidence for Dimroth rearrangements[198], which lead to formation of m^{6}A. In particular, we observed ≈4 p.p.m. upfield shift in m^{1}A-C2, consistent with base protonation, which is expected for m^{1}A but not m^{6}A. This, as well as the observation of two amino protons (H61 and H62) with distinct chemical shifts involved in HG H-bonding in DNA (see Figure 4.11) indicates a major positively charged amine tautomer rather than a neutral imine tautomeric form[200]. Nevertheless, we cannot rule out the existence of the neutral imine tautomeric form transiently and/or in low-abundance.

Unmodified oligonucleotides: Unmodified RNA oligonucleotides were synthesized using an in-house MerMade 6 Oligo Synthesizer employing 2′-TBDMS RNA phosphoramidites (ChemGenes) on 1 μmol standard synthesis columns (1000 Å) from BioAutomation using the option to leave the final 4,4′-dimethoxytrityl (DMT), the 5′-protection group on for the cartridge purification. The oligonucleotide was cleaved from

each 1 μmol column using ≈1 mL ammonia methylamine (1:1 ratio of 30% ammonium hydroxide and 30% methylamine) followed by 2-hour incubation at room temperature to allow base deprotection. The solution was then subjected to airflow until complete evaporation, leaving the desired product oligonucleotide as dried crystals. The crystals were then dissolved in 115 μL DMSO, mixed with 60 μL TEA and 75uL TEA⊛3HF, and incubated at 65°C for 2.5 hours for 2′-deprotection. The reaction was quenched using Glen-Pak™ RNA quenching buffer and loaded onto Glen-Pak™ RNA cartridges (Glen Research Corporation) for purification following the online protocol (http://www.glenresearch.com/Technical/GlenPak_UserGuide.pdf). Samples were ethanol precipitated and exchanged into NMR buffer. Unmodified DNA oligonucleotides were purchased from Integrated DNA Technologies with standard desalting.

Site-specifically $^{13}C/^{15}N$-labeled samples: The A_6 strand of A_6-DNA^{rA16} containing C8-$^{13}C/^{15}N$-labeled adenosine was synthesized using an in-house Solid-phase Oligonucleotide Synthesizer (BioAutomation MerMade 6), C8-$^{13}C/^{15}N$-labeled adenosine phosphoramidite (see below), and unlabeled DNA phosphoramidites (ChemGenes) using 1 μmol scale 1000 Å CPG DNA columns (BioAutomation). The synthesized oligonucleotides were cleaved and deprotected as described above for unlabeled RNA oligonucleotides and purified with the Glen-Pak™ RNA cartridge (Glen Research

Corporation) followed by ethanol precipitation and exchange into the desired NMR buffer.

Synthesis of 8-^{13}C-adenosine phosphoramidite: The 8-^{13}C-adenine nucleobase was synthesized according to a published procedure[201]. The protection of the exocyclic amino group with a benzoyl moiety and the conversion to the 5′-O-DMT-2′-O-TOM-protected 8-^{13}C-adenosine 3′-O-phosphoramidite was accomplished according to published procedures[202,203]. A detailed description on the chemical synthesis of 8-^{13}C-purine RNA phosphoramidite building blocks will soon be published elsewhere.

4.2.2 NMR experiments

Resonance assignment: NMR data were collected on an 800 MHz Varian DirectDrive2 spectrometer equipped with a triple resonance HCN cryogenic probe; 700 Bruker Avance III spectrometer equipped with a triple-resonance HCN cryogenic probe; and 600 MHz Varian Inova NMR spectrometer equipped with a Bruker HCPN cryogenic probe. Data were processed and analyzed using NMRpipe[188] and SPARKY (T. D. Goddard and D. G. Kneller, SPARKY 3, University of California, San Francisco), respectively. Resonances were assigned using conventional 2D HSQC, HMQC, NOESY and HCN experiments.

Chemical shift perturbations (CSPs) induced by m^1A or m^1G for each residue ($\Delta\omega_{residue}$) were calculated using Equation (4.1)[204] from the average Euclidean distance of all measured CSP ($\Delta\omega_C$, $\Delta\omega_N$, and $\Delta\omega_H$):

$$\Delta\omega_{residue} = \sqrt{\frac{1}{N}\sum_{i}^{N}\left(\frac{\gamma_i}{\gamma_{^1H}}\Delta\omega_i\right)^2} \tag{4.1}$$

where γ_i is the gyromagnetic ratio of the i^{th} nucleus (C, H or N), N is the total number of CSPs measured for each residue, and $\Delta\omega_i$ is the difference in chemical shifts (in p.p.m.) for the i^{th} nucleus between the m^1A or m^1G modified and unmodified duplex. Residues with $\Delta\omega_{residue} \geq 0.1$ p.p.m. are highlighted on the duplexes in Figure 4.8. An average CSP ($\Delta\omega_{avg}$) was calculated for each duplex by averaging $\Delta\omega_{residue}$ for two base pairs above and below the modified bp.

^{13}C and ^{15}N $R_{1\rho}$ relaxation dispersion: ^{13}C and ^{15}N $R_{1\rho}$ RD experiments were performed at 600 MHz (14.1 T) and 700 MHz (16.4 T) Bruker spectrometers as previously described[66,112,115] using spinlock powers (ω_{SL} $2\pi^{-1}$ Hz) and offset frequencies (Ω $2\pi^{-1}$ Hz). Magnetization of the spins of interest was allowed to relax under an applied spinlock for the following durations: [0–120 ms] for N1/N3 in hp-A_6-RNA and E-gc; [0–60 ms] for C8/C1′ in hp-A_6-RNA, E-gc, TAR-$UUCG^{GU}$, A_6-DNA, A_6-DNA^{rA} and A_6-DNA^{rG}.

4.2.3 Analysis of $R_{1\rho}$ data

Fitting of ^{13}C and ^{15}N $R_{1\rho}$ data: Experimental $R_{1\rho}$ relaxation rate constants were calculated by fitting peak intensities versus relaxation delay durations to a single exponential decay[115]. Uncertainty in the fitted $R_{1\rho}$ values (one s.d.) were derived using a Monte-Carlo method[141]. $R_{1\rho}$ data were fitted to simulated $R_{1\rho}$ values given by the solution to the Bloch–McConnell (BM) equations[140] at each given Ω and ω_{SL} combination. Residual sum of squares were minimized using a bounded least-squares algorithm[142] to give best-fit exchange parameters. The uncertainty in the chemical exchange parameters was calculated as the standard error of the fit[141]. A 2-state exchange model was used to fit the $R_{1\rho}$ RD profiles of A_6-DNA, A_6-DNA^{rA} and A_6-DNA^{rG} with the initial magnetization aligned either along the effective field of the ground (for slow exchange with $k_{ex}\,\Delta\omega^{-1} < 1$) or average (for fast exchange with $k_{ex}\,\Delta\omega^{-1} \geq 1$) state[139]. For the dA16-C8 RD data measured in A_6-DNA at low temperatures, both protocols yielded acceptable fits but resulted in different exchange parameters given the slower exchange rate. The exchange parameters obtained from average alignment protocol were selected based on a van't Hoff analysis[66]. For the dC15-C6 RD data measured in A6-DNA^{rG}, a 3-state chemical exchange model without minor exchange with average alignment was statistically favored over 2-state models. In all cases, Akaike information criterion (AIC)[205] and Bayesian information criterion (BIC)[205] were used to select the models.

4.2.4 Analysis of chemical shift and NOESY data

Chemical shifts and NOESY cross-peaks were used to characterize WC versus HG bps. The NOESY cross-peaks unique to HG bps include strong intra-nucleotide H1′–H8 NOE for *syn* purine, (i)A-H2–(i-1)H1′/H2′ and (i)A-H2–(i-1)H6/H8 for *syn* adenosine, and H8–H3, A-H6/C$^+$-H4–H3, (i)H3–(i+1)/(i-1)H1/H3, (i)A-H6/C$^+$-H4–(i+1)/(i-1)H1/H3 NOEs for connectivity involving imino or amino protons in both G–C and A–U/T HG bps[66,132,134]. Absence of the canonical sequential (i-1)H1′–(i)H8 NOE is also expected for *syn* purines due to the base flip. As described previously[66], the HG bps are also characterized by a unique set of chemical shifts relative to WC bps including downfield shifted purine-C8 and purine-C1′, protonated cytosine-C6 (≈3 p.p.m.) and upfield shifted protonated cytosine-C5, guanine-N1 and thymine-N3 (≈1–2 p.p.m.). The C1′ chemical shifts is also sensitive to sugar pucker. In A-RNA, deviations from the A-form C3′-endo toward C2′-endo sugar pucker leads to an upfield shift (≈4 p.p.m.)[206,207]. In addition, deviations from the A-form conformation due to loss of stacking and bulging out of nucleotides results in a downfield shift in the base C6/C8 and cytosine-C5 and upfield shift on sugar-C1′[141].

4.2.5 Density functional theory geometry optimizations and CS calculations

Density functional theory (DFT) calculations[131] using Gaussian 09c (Gaussian, Inc.) were performed as previously described[66] to compute chemical shifts for WC and

HG bps in A-RNA and B-DNA. In all cases, protons were added using PyMOL (https://www.pymol.org/) and the phosphate backbone truncated leaving only the nucleoside motifs for each bp[66]. Calculations were performed on rA–rU HG bp obtained from snapshots of rA16–rU9 bp in biased MD simulation of hp-A_6-RNA, rA–rU HG bp from the X-ray structure of P4-P6 domain of the Group I intron RNA (PDBID: 1L8V[208]), tertiary rG–rC^+ HG bp in the structure of 23S ribosomal RNA-protein complex (PDBID: 3U56[209]) and rG^{syn}–rG^{anti} mispair in duplex RNA structure (PDBID: 3CZW[210]). Reference rA–rU or rG–rC WC bps were taken from MD snapshot or from the same X-ray structures used to obtain HG bps. Two runs of geometry optimizations were carried out using the B3LYP functional with 3-21G and 6-311+G(2d,p) basis sets, respectively, with all heavy atoms (C, N and O) frozen. Carbon chemical shifts were computed using the GIAO method within the B3LYP/6-311+G(2d,p) basis set[66] on the converged configuration after the second run of optimization. The isotropic carbon chemical shift (ω_{13C}) was referenced to that of TMS (ω_{TMS} = 182.4656 p.p.m.), which was optimized and computed at the same level of theory. The carbon chemical shifts computed for reference WC bps were subtracted from those computed for HG bps to yield chemical shift changes upon HG formation ($\Delta\omega = \omega_{HG}-\omega_{WC}$).

4.2.6 Analysis of UV melting data

The UV absorbance at 260nm (A_{260}) as a function of temperature was measured on a Shimadzu UV-3600 UV-Vis-NIR spectrophotometer using the 8-cell sample holder with a Fisher Isotemp Refrigerated Circulator to regulate sample temperature. All DNA and RNA oligonucleotides were diluted directly from NMR samples using the same NMR buffer (15mM phosphate, 25mM NaCl, 0.1mM EDTA at pH = 5.4 or 6.8) unless stated otherwise and triplicate measurements were carried out for each oligonucleotide simultaneously using a sample volume of 125 μL in each cell and an additional reference cell containing the same amount of buffer. The temperature was varied between 5°C and 90°C at a ramping rate of 1 °C min^{-1}.

The melting temperature (T_m) and enthalpy (ΔH) for duplex association and hairpin folding was obtained by fitting the melting curves to Equations (4.2) and (4.3)[211], respectively,

$$f = \frac{1 + 4e^{(1/T_m - 1/T)\Delta H/R} - \sqrt{1 + 8e^{(1/T_m - 1/T)\Delta H/R}}}{4e^{(1/T_m - 1/T)\Delta H/R}} \tag{4.2}$$

$$f = \frac{e^{(1/T_m - 1/T)\Delta H/R}}{1 + e^{(1/T_m - 1/T)\Delta H/R}} \tag{4.3}$$

T is the measured temperature and f is the fraction of the remaining duplex or folded hairpin over the total concentration of duplex or hairpin (C_T), which is proportional to the measured A_{260}.

The thermodynamic parameters ΔG and ΔS were calculated using Equation (4.4) for duplex association and Equation (4.5) for hairpin folding, respectively.

$$\Delta S = \Delta H / T_m - R\ln(C_T / 2); \Delta G = \Delta H - T\Delta S \quad (4.4)$$

$$\Delta S = \Delta H / T_m; \Delta G = \Delta H - T\Delta S \quad (4.5)$$

The fitting was carried out using nonlinear model fitting with Mathematica 10.0 (Wolfram Research). Errors in T_m and ΔH represent the standard deviation (one s.d.) from the triplicate measurements. The destabilization effects due to m^1A and m^1G in DNA or RNA were calculated by taking the difference in free energy for folding i.e. $\Delta\Delta G = \Delta G_{\mathrm{mod}} - \Delta G_{un\,\mathrm{mod}}$, $\Delta\Delta H = \Delta H_{\mathrm{mod}} - \Delta H_{un\,\mathrm{mod}}$ and $\Delta\Delta S = \Delta S_{\mathrm{mod}} - \Delta S_{un\,\mathrm{mod}}$.

Since the Dimroth rearrangement can occur for m^1A in both DNA and RNA under basic conditions[198] and high temperatures[199], melting experiments were repeated for all m^1A containing duplexes when restricting the temperature to <65°C and when using both neutral (pH = 6.8) and acidic (pH = 5.4) conditions. These control experiments yielded reproducible melting curves and fitted thermodynamic parameters at neutral or acidic pH conditions that are within experimental error. ^1H 1D NMR spectra recorded for the A6-DNA^{m1A} sample following melting showed insignificant changes and no evidence for Dimroth rearrangements.

4.2.7 Steric analysis and survey of HG bps in RNA

A–T/U and G–C WC bps were obtained from idealized B-DNA and A-RNA helices built using 3DNA[163]. 146 and 159 WC bps surrounded by at least one WC bp above and below were obtained from high resolution (< 2 Å) X-ray structures of A-RNA and B-DNA duplexes, respectively in the PDB. Purine bases were flipped around the glycosidic bond and inter-atomic distances measured using an in-house python script. Note that proximity of the exocyclic amino group on guanine to the phosphate group during the base flip was not considered a steric clash given the potential for H-bonding. The survey of HG bps in RNA was carried out following the same protocol reported for B-DNA[155]. Briefly, all RNA X-ray structures with resolution ≤ 3 Å were downloaded from the PDB on August 31 2014. The same in-house program was used to identify rA–rU and rG–rC bps in RNA structures using three HG criteria (H-bonding, constricted C1′–C1′ distance and *syn* purine)[155]. Redundancies defined as bps that are surrounded with the same sequence contexts and from same RNA or RNA-protein/ligand complexes were removed by manual inspection as described previously[155]. The survey identified a single rA–rU HG bp that reoccurs in four distinct X-ray structures of the P4-P6 domain of the *Tetrahymena thermophila* group 1 intron RNA (PDBID: 1GID[212], 1l8V[208], 1HR2[213], and 2R8S[214]). The RNA HG survey also identified several examples of long-range HG bps forming tertiary contacts, HG bps in triplexes, and reverse rA–rU HG bps within duplexes typically near rG–rA mismatches.

4.2.8 Biased and unbiased molecular dynamics simulations

Structure generation for MD simulation: hp-A_6-RNA, hp-A_6-RNA 3′→5′, and A_6-DNA helices were built using make-na[215] with all bases in WC conformation. In the case of hp-A_6-RNA, a duplex structure was generated using make-na and the UUCG loop attached and annealed using the CHARMM simulation package[216]. Rotating along the glycosidic bond angle χ by 180° created structures with HG conformation at A16.

Unbiased MD equilibrium simulations: All structures were simulated using constant temperature MD with CHARMM36 forcefield[217] and a generalized Born molecular volume (GBMV) implicit solvent[218]; parameters for m^1A were taken from Xu *et al* [219]. Integration used a velocity-Verlet algorithm with a timestep of 1 fs. The cutoff for non-bonded list generation was 21 Å, the cutoff for non-bonded interactions was 18 Å, and the onset of switching for non-bonded interactions occurred at 16 Å. The SHAKE algorithm was used to constrain the covalent bonds to hydrogen atoms involved. Each structure was heated to 300.0 K with harmonic constraints on all non-hydrogen atoms, heating occurred in1 ps increments of 1.0 K for a total of 300 ps steps, followed by 200 ps equilibration at 300.0 K. Harmonic constraints were then gradually removed during a sequence of 4 reductions for 50 ps each. Unbiased production-run simulations were then run for 3 ns without constraints for each system. Ten independent simulations with hp-A_6-RNA and A_6-DNA^{rA} with A16 in HG conformation were produced from independent

conformations obtained during the heating and equilibration method described above. A_6-DNA in HG was repeated twice.

Global RMSD was calculated from the single 3 ns trajectories of m^1A starting in HG for both hp-A_6-RNA and A_6-DNA,

$$RMSD = \sqrt{\frac{\sum_{i=1}^{N}(r_i(t) - r^R)^2}{N}} \tag{4.6}$$

in which $r_i(t)$ is the instantaneous coordinate of an atom and r^R is the position of the reference structure. H-bond presence was evaluated using CHARMM's COOR HBOND module for each trajectory with cutoff distance and angle of 3.6 Å, and 120° following Goldsmith *et al*[220].

Biased MD simulations: The protocols for minimization, heating, and solvation were identical to those used for the unbiased simulations. The biased molecular dynamics method[221] implemented in the CHARMM package was used to force conformational transitions between WC and HG states using a biasing potential W(ρ(t)) applied according to Equation (4.7),

$$W(\rho(t)) = \begin{cases} \frac{\alpha}{2}\left(\rho(t) - \rho_a(t)\right)^2, & \text{if } \rho(t) < \rho_a(t) \\ 0 & \text{if } \rho(t) \geq \rho_a(t) \end{cases}$$

where

$$\rho(t) = \left(\frac{1}{N(N-1)}\right)\sum_{i=1}^{N}\sum_{j\neq 1}^{N}\left(r_{ij}(t) - r_{ij}^R\right)^2 \tag{4.7}$$

$\rho(t)$ is a collective distance between the instantaneous (r_{ij}) and the reference structure (r_{ij}^{R}), and α the strength of the half-harmonic bias. In all cases, biases were placed between pairs of atoms that share a hydrogen-bond in the target structure, ensuring that the adenine base would not only perform the roughly 180° flip, but also form the definitive hydrogen-bonding structure of the desired WC or HG configuration. After the biased trajectories were generated, they were post-processed in CHARMM, outputting the χ-angle dependence of the relative interaction energy value in the absence of the bias. The relative interaction energy was calculated for the base pair that includes the flipping base as well as the base pairs above and below the flipping base. Angle-energy pairs were binned into 50 bins and the mean of the energy was evaluated within each bin. Plots of relative interaction energy as a function of the χ-angle were thus generated.

4.3 Results

4.3.1 Absence of conformational exchange in A-RNA

We used NMR spin relaxation in the rotating frame ($R_{1\rho}$)[136,137,141] to examine whether WC bps in A-RNA duplexes transiently adopt HG bps like in B-DNA. A dynamic equilibrium between a dominant ground state (GS) and short-lived low-abundance 'excited state' (ES) conformation can lead to line-broadening of NMR resonances if the conformational exchange occurs on the μs–ms timescale. The $R_{1\rho}$ experiment[137] measures this line broadening contribution (R_{ex}) to the transverse

relaxation rate (R_2) during a relaxation period in which a continuous radiofrequency (RF) field is applied with variable power (ω_{SL}) and frequency (ω_{RF}). The resulting dependence of R_2+R_{ex} on ω_{SL} and ω_{RF}, referred to as relaxation dispersion (RD), can be fitted to the Bloch-McConnell equations describing n-site exchange[140] to extract exchange parameters of interest, including the population of the ES (p_{ES}), the rate constant for conformational exchange ($k_{ex} = k_{forward} + k_{backward}$), and the difference between the chemical shifts of the ES and GS ($\Delta\omega = \omega_{ES}-\omega_{GS}$).

So far, RD studies have provided evidence for μs–ms conformational exchange in non-coding RNAs involving localized changes in secondary structure in and around non-canonical motifs (reviewed in ref. [141]). The RD contributions from such chemical exchange processes can mask the ability to detect WC⇌HG exchange. To hone in on WC⇌HG exchange in A-RNA, we carried out ^{13}C and ^{15}N $R_{1\rho}$ RD experiments[114,115] on an RNA duplex (hp-A_6-RNA) capped by a stabilizing apical loop lacking non-canonical motifs and containing the same sequence (A_6-DNA) for which we first reported transient HG bps in B-DNA[66] (Figure 4.3). We targeted purine-C8, C-C6, G-N1, T-N3 and sugar purine-C1′ sites (highlighted in orange in Figure 4.2), all of which have previously been shown to exhibit significant RD due to WC⇌HG chemical exchange in B-DNA[66,112,116] (Supplementary Information). In stark contrast to B-DNA, all RD profiles measured in in hp-A_6-RNA were flat with no signs of detectable conformational exchange on the μs–ms timescale (Figure 4.4). No RD was observed across a variety of rG–rC and rA–rU WC

bps, under low pH conditions (pH = 5.4) that allow optimal RD detection of WC⇌HG exchange in B-DNA[66,111], upon increasing the temperature (T = 35°C), and in the presence of 4 mM Mg^{2+} (at pH = 6.8 and T = 5 or 25°C) (Figures 4.4, 4.5 and 4.6).

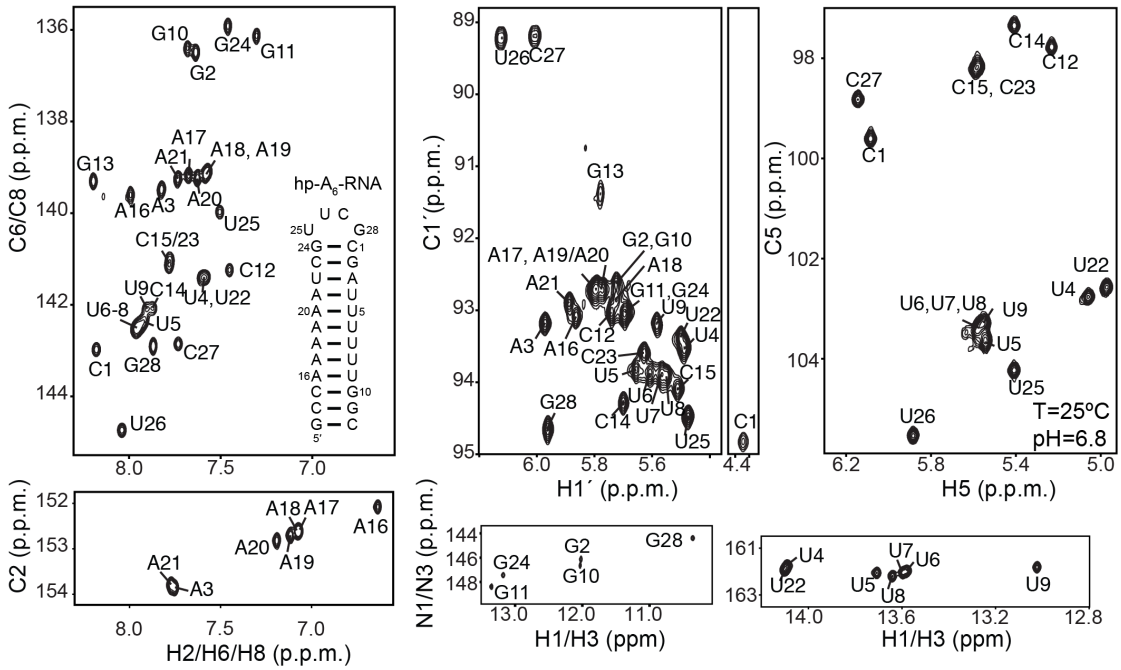

Figure 4.3: Resonance assignment of hp-A_6-RNA.

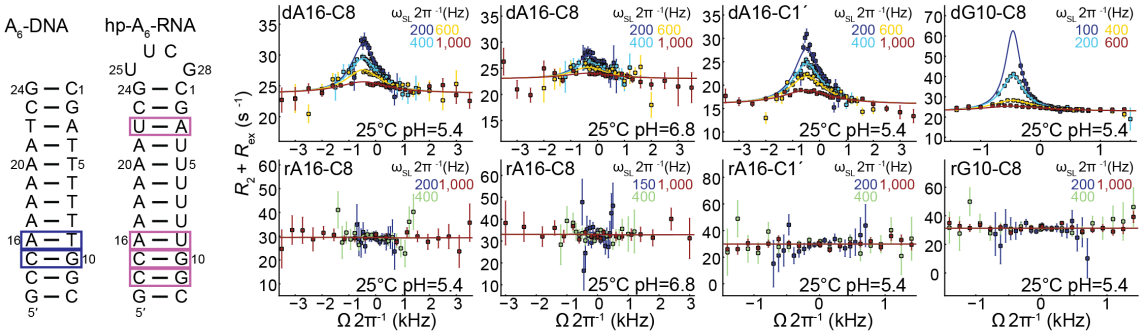

Figure 4.4: Comparison of chemical exchanges in A-RNA and B-DNA.

A_6-DNA and hp-A_6-RNA duplexes are shown with bps targeted in RD measurements highlighted. Off-resonance RD profiles showing R_2+R_{ex} as a function of spin lock offset (Ω $2\pi^{-1}$Hz, where $\Omega = \Omega_{obs} - \omega_{RF}$) and power ($\omega_{SL}$ $2\pi^{-1}$Hz, in insets). Error bars represent experimental uncertainty (one s.d.) estimated from mono-exponential fitting of n = 10 (A_6-DNA) and n = 6 (hp-A_6-RNA) independently measured peak intensities using a Monte-Carlo based method (Methods). Solid line represents a fit to two-state exchange[66].

To broaden the search for WC⇌HG exchange in A-RNA duplexes, we carried out additional RD measurements over a wide range of conditions (pH = 5.4–8.4 and T = 5–35°C) for another ten rG–rC and seven rA–rU bps embedded in distinct sequence and structural contexts in three additional RNA molecules, including a GC-rich hairpin (hp-gc^{GU}), elongated duplex (E-gc), the transactivation response element (TAR) and a mutant form of TAR (TAR-UUCG^{GU}) that is impaired from undergoing secondary structure chemical exchange[222] (Figure 4.5). In all cases we did not detect any signs of RD (Figure 4.6). These results, together with our studies[196,222,223] reporting flat RD profiles for RNA WC bps near non-canonical motifs and mismatches (wtTAR and P5abc, Figure 4.5) and for the reverse wobble rG^{syn}–rU mispairs in apical loops[224] stand in striking contrast to canonical duplex DNA in which we have robustly observed WC⇌HG exchange in all 35 dA–dT and dG–dC bps examined to date in a wide variety of positional and sequence contexts in eight different duplexes that have varying lengths and stabilities[66,116].

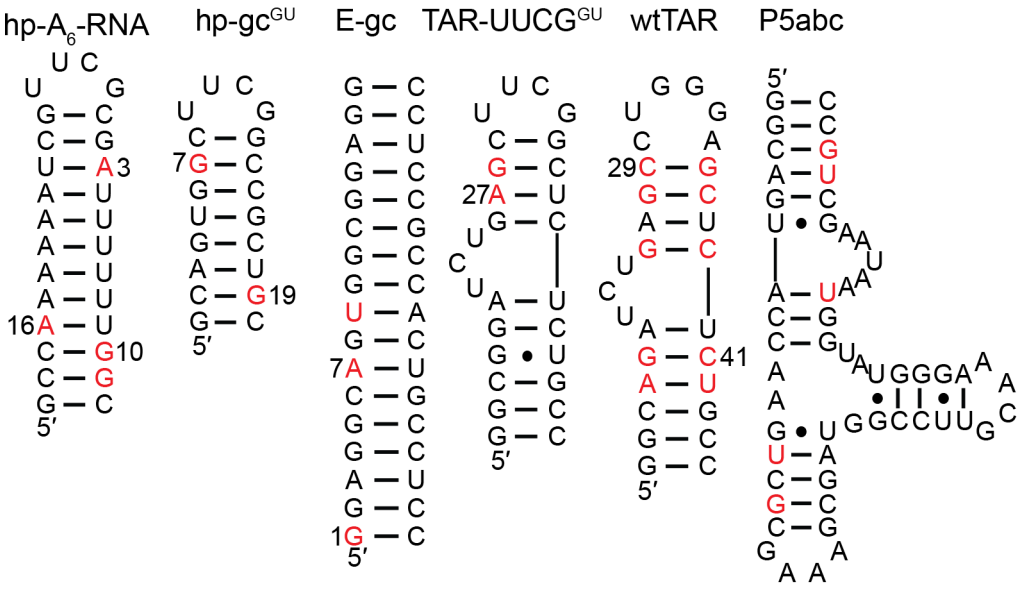

Figure 4.5: Secondary structures of A-RNA that lack conformational exchange with measured bps shown in red.

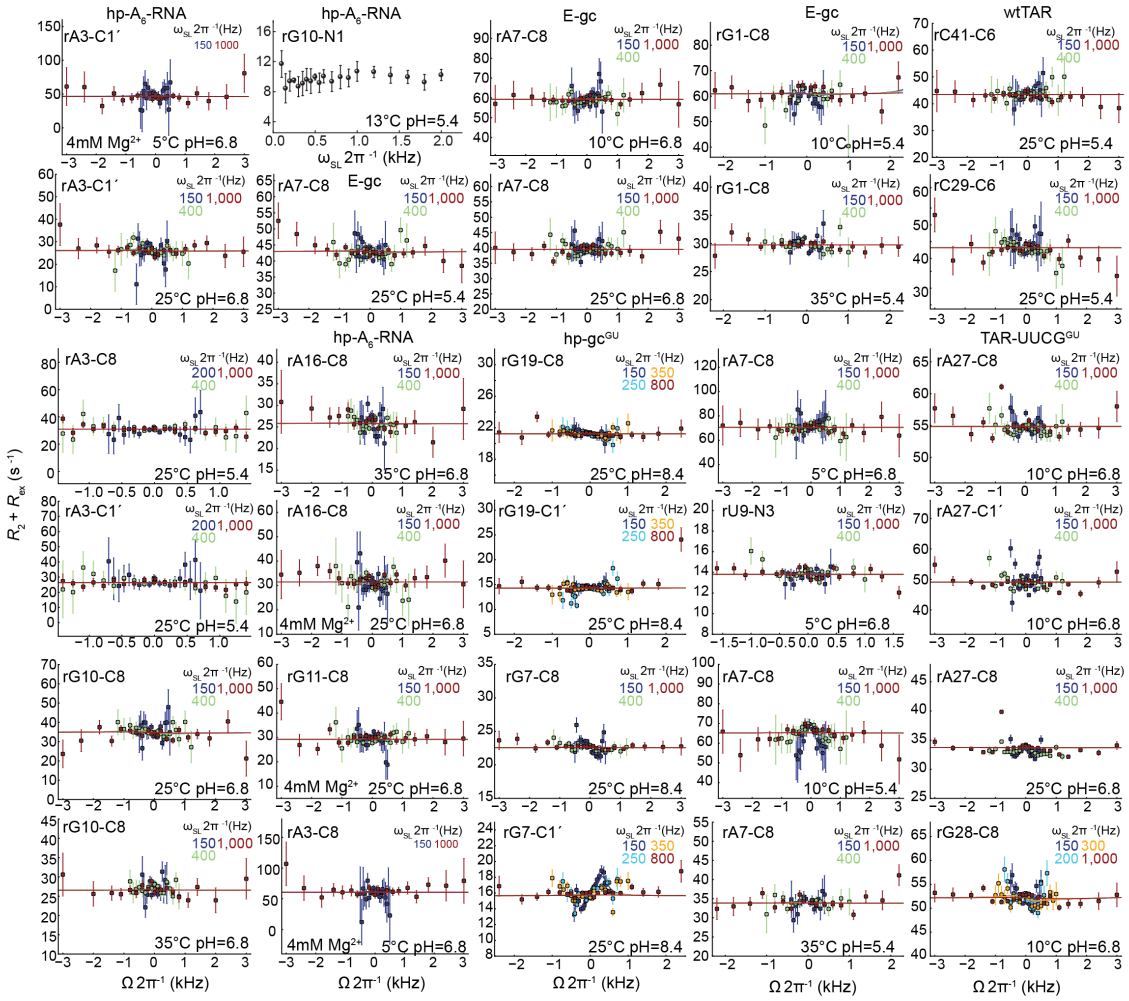

Figure 4.6: Lack of chemical exchange under various temperature, pH and Mg^{2+} conditions.

The lack of detectable WC⇌HG exchange in A-RNA could in principle result from small differences between the WC and HG NMR chemical shifts ($\Delta\omega < 0.5$ p.p.m. for carbon chemical shifts). However, based on density functional theory calculations (DFT)[66,130] and a survey of *syn* purine base chemical shifts in the Biological Magnetic Resonance Data Bank[127], it is highly unlikely that such a large transformation in base pairing would result in such small changes in chemical shifts for the different sugar (C1′) and base (C8, C6 and N1/N3) sites targeted for RD measurements (Supplementary Information). The absence of RD is unlikely to be due to the exchange rate falling outside the detection limits of the RD experiment given that flat profiles are observed over a wide range of temperatures and pH conditions (Figure 4.6) known to significantly alter the WC⇌HG exchange rate in B-DNA[66,116].

A more likely explanation is that HG bps are energetically disfavored in A-RNA duplexes, and have an abundance that falls below the detection threshold of the RD experiment (population < 0.01%). Indeed, a survey of X-ray structures of RNA duplexes in the Protein Data Bank (PDB)[154] failed to identify a single rG–rC$^+$ or rA–rU HG bp within continuous A-RNA duplexes out of a total of 123,935 rG–rC and rA–rU bps (Methods); while in sharp contrast, a similar survey conducted recently on B-DNA duplexes[155] identified 54 dG–dC$^+$ or dA–dU HG bps out of a much smaller set of 51,485 bps. The survey identified a single rA–rU HG bp (for example, PDBID: 1GID[212]) within an RNA duplex that fell well outside the A-form structural context being surrounded by

a bulge and internal loop. The survey did identify several examples of long-range rG–rC$^+$ and rA–rU HG bps forming tertiary contacts; rG–rC$^+$ and rA–rU HG bps in triplexes and reverse rA–rU HG bps within duplexes typically near rG–rA mismatches where purines adopt *anti* rather than *syn* conformation, as well as several examples of HG mispairs in A-RNA duplexes (e.g. rG^{syn}–rG^{anti}, and rG^{syn}–rA^{anti}) (Supplementary Information).

4.3.2 m^{1}A and m^{1}G modified A-RNA

If rG–rC$^+$ and rA–rU HG bps are indeed thermodynamically disfavored in A-RNA, they should prove more difficult to trap using chemical modifications known to stabilize dG–dC$^+$ and dA–dT HG bps in B-DNA[66]. We therefore examined whether HG bps could be stably trapped in A-RNA duplexes using m^{1}rA and m^{1}rG. These modified bases block WC pairing because of steric collisions with the methyl group and because the methylation knocks out one of the WC H-bonds (Figure 4.7). Both m^{1}dA and m^{1}dG occur in DNA due to alkylation damage[192,193]. In B-DNA, m^{1}dA and m^{1}dG are accommodated as m^{1}dA–dT and m^{1}dG–dC$^+$ HG bps[66,82,83] (Figure 4.7), which can in turn be recognized and repaired by damage repair enzymes[192,193]. m^{1}rA and m^{1}rG can also occur as a form of alkylation damage in RNA but they are also highly conserved post-transcriptional modifications in transfer and ribosomal RNAs that play critical structural and functional roles often by blocking WC base pairing[225-229]. m^{1}rG and m^{1}rA have been

shown to induce duplex-to-hairpin transitions in palindromic RNA sequences where the modified base favors an unpaired conformation within apical loops[230,231]. Recent genome-wide studies have shown m^{1}rA to be a dynamic reversible eukaryotic messenger RNA (mRNA) modification that can potentially play roles in epitranscriptomic regulation[232,233].

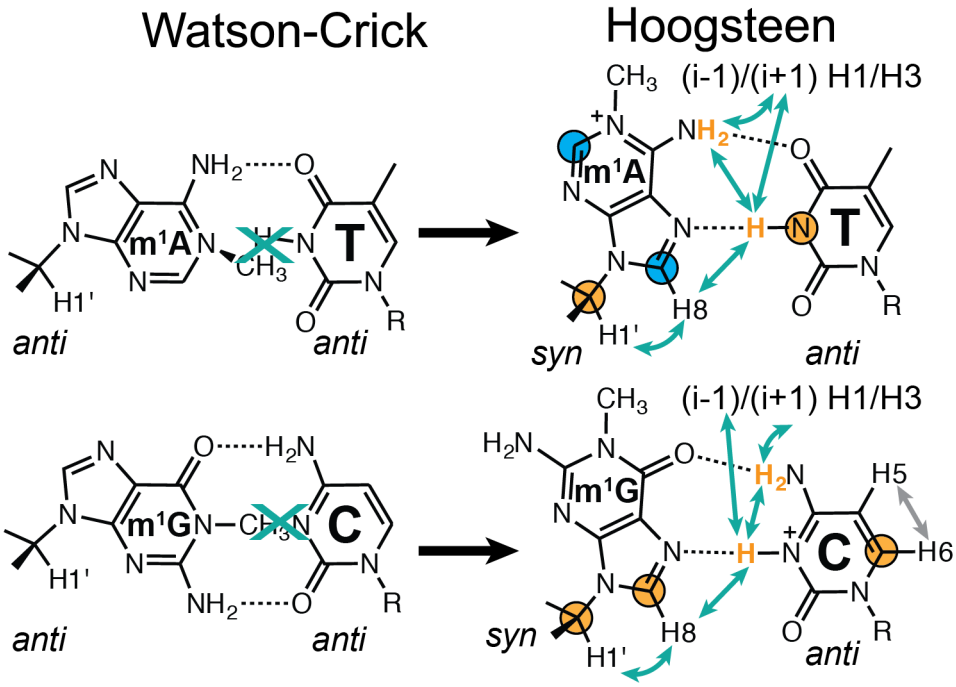

Figure 4.7: *N*^{1}-methylated HG bps and NOE signatures.

N^1-methylated purines trap HG bps in B-DNA. NMR chemical shift probes of HG bps are in orange and of purine methylation state in cyan. Arrows indicate characteristic HG NOE cross-peaks.

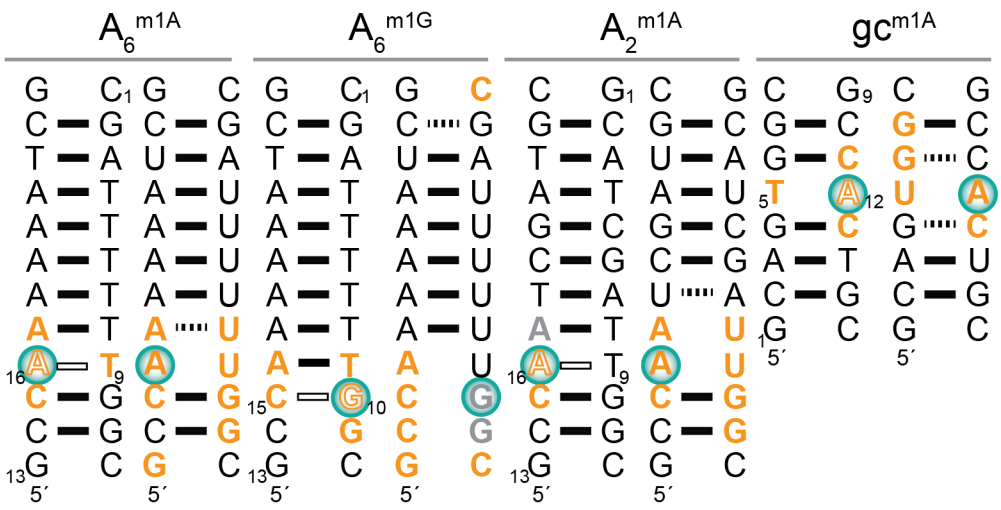

Figure 4.8: DNA and RNA duplexes for *N*[1]-methylation study.

Duplexes containing m[1]A or m[1]G (turquoise circles). *syn* or *anti* purines deduced by NMR are shown as open and filled letters, respectively. HG and partially melted bps as deduced by NMR are indicated using open and dashed lines, respectively. Residues showing significant chemical shift perturbations or line-broadening due to m[1]A or m[1]G are colored orange and grey, respectively.

In prior studies[66], we showed that m^{1}dA16 and m^{1}dG10 form stable m^{1}dA16–dT9 and m^{1}dG10–dC15$^+$ HG bps stabilized by unique H-bonds in A$_6$-DNA while minimally impacting neighboring WC bps as judged based on observation of HG-specific chemical shifts, Nuclear Overhauser effect spectroscopy (NOESY) cross-peaks, and imino resonances (highlighted in Figure 4.7). In contrast, we did not observe any NMR evidence for HG bps or *syn* purine bases in the corresponding A$_6$-RNA duplex containing m^{1}rA16 or m^{1}rG10 (Figures 4.8–4.11). This was the case despite the fact that A$_6$-DNA and A$_6$-RNA duplexes have very similar thermodynamic stabilities. Rather, the rA-C1′ chemical shifts falls in a region that is consistent with A-form helical residues (Figure 4.9 and Supplementary Information). We also observe continuous NOE distance-based connectivity between the m^{1}rA and its preceding residue that are consistent with an *anti* conformation for the purine base. These data, together with the absence of strong H1′–H8 NOEs expected for *syn* base (Figure 4.10) and imino and amino resonances indicative of H-bonding (Figure 4.11) suggest that in A$_6$-RNA, m^{1}rA adopts a predominantly unpaired *anti* conformation although we cannot rule out transient formation of *syn* base conformations. The resonances belonging to m^{1}rG in A$_6$-RNA were broadened out of detection suggesting extensive conformational exchange at the μs–ms timescale with no NMR evidence for HG pairing given the absence of downfield-shifted rC-H4 (Figure 4.11). However, we cannot exclude micro-to-millisecond exchange

between *syn* and *anti* conformations for the m^{1}rG base since the resonances are broadened out of detection[111].

Compared to A$_6$-DNA, m^{1}A and m^{1}G also induced more significant structural perturbations in A$_6$-RNA (Figure 4.8). We do not observe some of the imino resonances belonging to WC bps neighboring the modified site (Figure 4.11), which suggests loss of H-bonds and the melting of these base pairs. The modifications also induced more extensive chemical shift perturbations (Figure 4.12 and highlighted in orange in Figure 4.8) and line broadening (highlighted in grey in Figure 4.8) in the sugar and base resonances that extend to the partner strand. The direction of the perturbations is consistent with deviations from a helical conformation. The perturbations were particularly pronounced for m^{1}rG, which broadened all imino resonances out of detection at 35°C, consistent with significant melting of the entire duplex (Figure 4.13). Thus, HG bps are so sufficiently disfavored in A-RNA that m^{1}rA and m^{1}rG prefer to adopt predominantly non-helical conformations that disrupt the duplex structure.

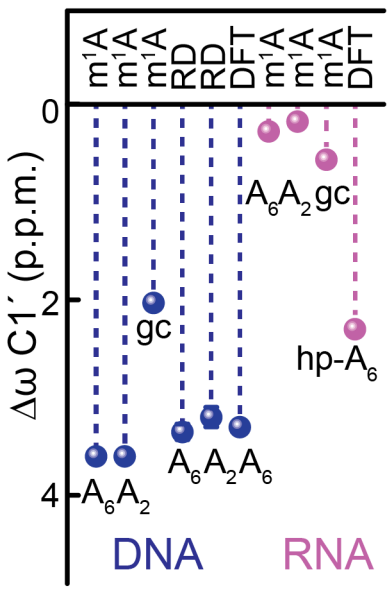

Figure 4.9: Comparison of chemical shift perturbations on C1′ upon N^1-methylation in DNA and RNA.

m^{1}A or m^{1}G induced purine-C1′ chemical shift perturbations ($\Delta\omega = \omega_{modified} - \omega_{unmodified}$) in A-RNA (violet) and B-DNA (blue). Shown for comparison are $\Delta\omega = \omega_{HG} - \omega_{WC}$ measured for transient dA–dT HG bps by RD ("RD") in unmodified DNA duplexes (error bars showing one s.d.) and computed for adenine residues using DFT (Methods).

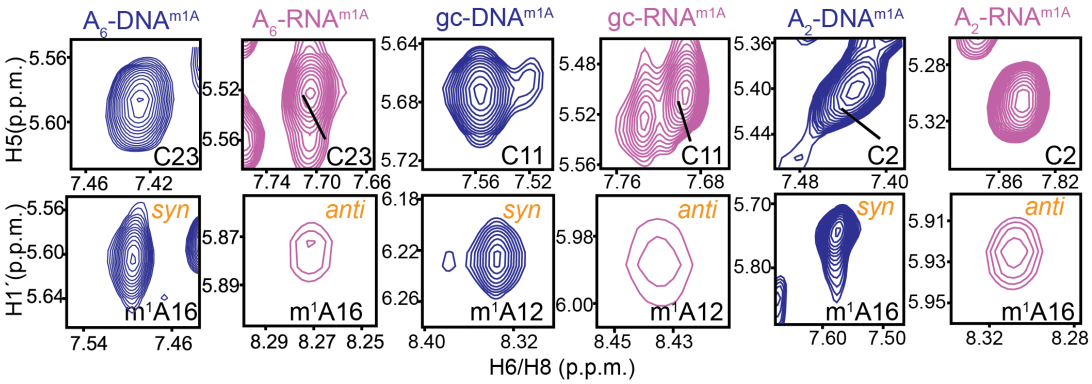

Figure 4.10: NOE signatures for *syn* purine in DNA and *anti* in RNA.

NOESY H1′–H8 cross-peaks showing *syn* purine bases in B-DNA but not A-RNA. Shown for reference is the cytosine base H5–H6 NOE with inter-atomic distance ≈2.5 Å.

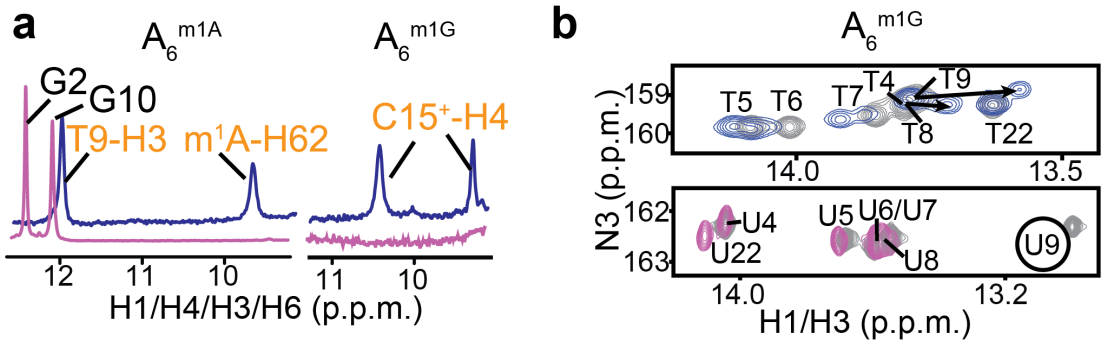

Figure 4.11: imino ^{1}H for HG H-bonding and impact on neighbouring bps.

(a) 1D ^{1}H spectra showing the imino/amino resonances expected for HG type H-bonds in A_6-DNA^{m1A} and A_6-DNA^{m1G} but not in methylated RNA at 5°C and 15°C, respectively. (b) Example showing m^{1}G induced loss of a WC imino resonance (highlighted in a circle) in A_6-RNA but not A_6-DNA in 2D NMR spectra.

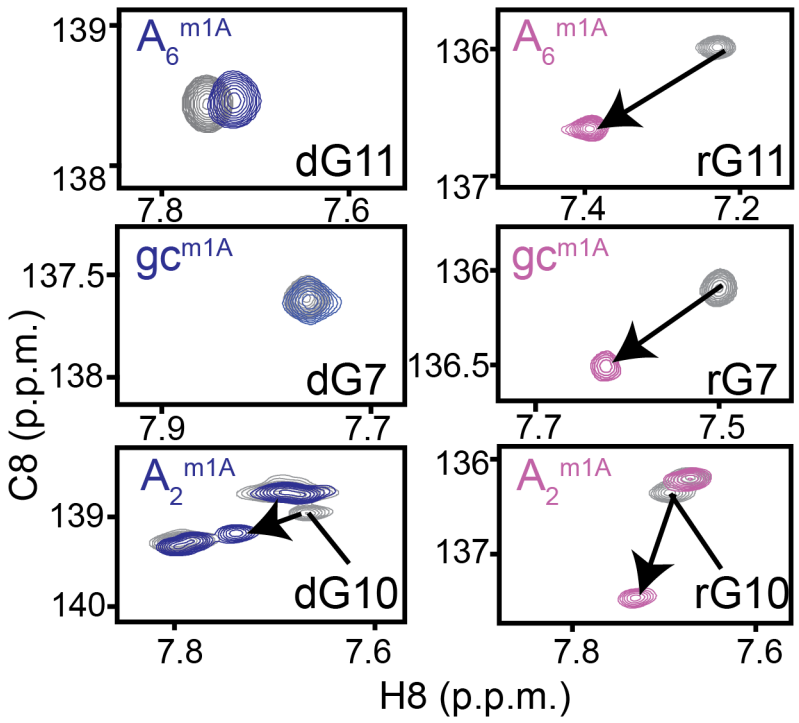

Figure 4.12: m^{1}A in RNA introduces larger structural perturbations than in DNA.

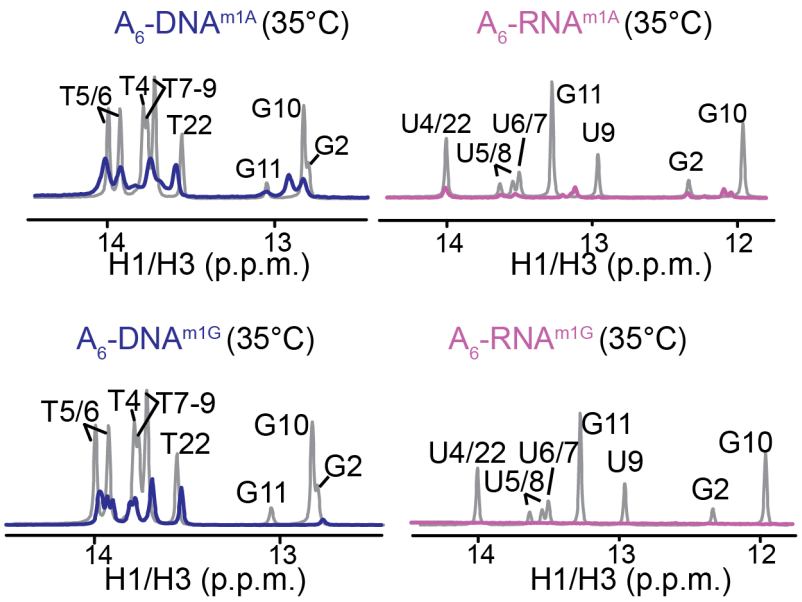

Figure 4.13: imino ^{1}H NMR shows more extensive helix melting of N^{1}-methylated RNA compared to DNA.

Similar results were obtained in GC-rich (gc^{m1A}) and scrambled (A_2^{m1A})[66] (B.S., H.Z., Y.X., H.M.A., unpublished) duplexes with m^{1}A consistently forming HG bps or adopting a *syn* conformation in B-DNA but not in A-RNA (Figure 4.10) and with the modification more significantly perturbing the structure of A-RNA as compared to B-DNA (Figure 4.12). The structural perturbations induced by m^{1}rA varied with sequence, and were either distributed across many WC bps (A_6-RNA and A_2-RNA) or more severe but localized to the modified and partner base (gc-RNA) (Figure 4.8). In all cases we did not observe any evidence for m^{1}rA or m^{1}rG induced duplex-to-hairpin transition based on spectral overlays with the unmodified counterparts.

We corroborated the more portent destabilization of A-RNA as compared to B-DNA duplexes by m^{1}A and m^{1}G using UV melting experiments. m^{1}dA destabilized A_6-DNA, A_2-DNA, and gc-DNA duplexes by $\Delta\Delta G$ = 1.8–3.4 kcal mol^{-1} (Figure 4.14) in good agreement with the relative stability of transient HG bps measured by NMR RD (2.1–4.3 kcal mol^{-1}) and prior UV-melting studies of m^{1}dA containing DNA duplexes (≈2 kcal mol^{-1})[234]. By comparison, m^{1}rA and m^{1}rG destabilized the corresponding A-RNA duplexes by a larger amount $\Delta\Delta G$ = 4.3–6.5 kcal mol^{-1}. Interestingly, this greater destabilization is comparable to the relative stability of the base-opened state[235]. This suggests that in A-RNA, the modification results in a conformation similar to the base-opened state, consistent with the NMR evidence for local melting. Greater destabilization (by ≈1.1–4.7 kcal mol^{-1}) of A-RNA as compared to B-DNA was robustly

observed across different duplex and hairpin contexts in the presence or absence of Mg^{2+} and with the destabilization being principally enthalpic (Figure 4.14). The potent m^1rA induced destabilization of duplex RNA is highly significant considering recent studies showing it to be a dynamic mRNA modification with roles in post-transcriptional gene regulation[232,233]. For comparison, the other well studied mRNA modification[236] N^6-methyladenosine (m^6A) which affects mRNA localization, stability, translation, and splicing destabilizes A-RNA by only 0.5–1.7 kcal mol^{-1} [237].

The more potent m^1A and m^1G destabilization of A-RNA as compared to B-DNA is unlikely to be due to differences in steric contacts involving the methyl group in a HG bp configuration (Supplementary Information). While the positive charge on m^1rA may affect stacking and H-bonding interactions, significant destabilization is also observed with the neutral m^1rG, and the m^1A destabilization is greater for A-RNA as compared to B-DNA (Figure 4.14). Rather, the greater destabilization observed in A-RNA is likely due to the higher energetic cost of forming HG bps in A-RNA as compared to B-DNA.

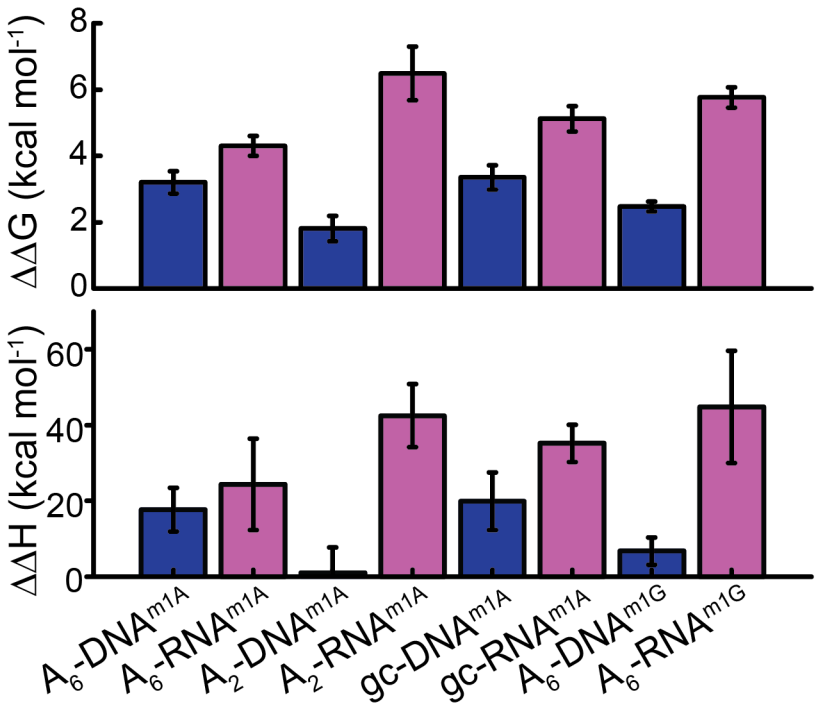

Figure 4.14: Destabilization by *N*[1]-methylation in DNA and RNA.

Free energy (ΔΔG) and enthalpy (ΔΔH) destabilization due to m[1]A and m[1]G in DNA (blue) and RNA (violet) duplexes measured by UV melting experiments with error bars, one s.d. (n = 3 independent measurements).

4.3.3 Why are HG bps disfavored in A-RNA?

Why are HG bps so strongly disfavored in RNA as compared to DNA duplexes? The HG bp could in principle be disfavored in RNA due to the sugar 2′-OH at the purine residue. The 2′-OH helps bias the sugar pucker toward the C3′-endo conformation (Figure 4.1) due to unfavorable steric contacts between O2′and O3′ and electronic effects involving the 2′-OH group[120,238]. This in turn disfavors the *syn* purine base conformation even in nucleosides[239] and single-stranded polynucleotides[240] due to unfavorable base-sugar steric contacts (N3–H3′ and N3–O4′). The *syn* purine conformation may also destabilize water-bridged interactions involving the 2′-OH and N3 of the *anti* purine base[241]. To examine whether the mere presence of a 2′-OH group on the ribose moiety of the flipping purine base is sufficient to suppress WC⇌HG exchange, we carried out $R_{1\rho}$ RD experiments on site or strand specifically labeled A_6-DNA duplexes containing a single ribonucleotide, rA16 or rG10 (Figure 4.15 and Methods). These RD measurements were also of interest given that single ribonucleotides are frequently incorporated in DNA during replication and can have important biological consequences through mechanisms that are not fully understood[242].

Both rA16 and rG10 formed the expected rA16–dT9 and rG10–dC15 WC bps[243] and exhibited RD consistent with WC⇌HG exchange (Figure 4.15). The lower R_{ex} contribution observed for the rA16 and rG10 substituted samples relative to the unmodified DNA duplex can be attributed to ≈4-fold faster exchange rate (k_{ex} = 2325 s^{-1}

versus 595 s^{-1}) in the case of rA16 and a combination of slightly smaller $\Delta\omega$ (1.8 versus 2.1 p.p.m.) and transient HG population (0.8% versus 1.3%) in the case of rG10 (Figure 4.15). Neither rA16 nor rG10 significantly impacted the abundance of the transient HG bps relative to the unmodified A_6-DNA duplex (Figure 4.15) indicating that the purine sugar 2′-OH group alone cannot account for the lack of observable WC⇌HG exchange in A-RNA duplexes. We confirmed these findings by analyzing A_6-DNA duplexes containing N^1-methylated single ribonucleotide, m^1rA16 or m^1rG10. In both cases, we observe stably formed m^1rA16–dT9 and m^1rG10–$dC15^+$ HG bps (Figures 4.16 and 4.17). These data suggest that the destabilization of HG bps requires the broader A-form RNA helical context.

Next, we examined whether there were unique steric clashes that could disfavor *syn* purine bases within the compact A-RNA helix context that are absent in the more capacious B-form DNA helix. Indeed, flipping the purine base around the glycosidic χ-angle through a range of angles (160°–200°) that span *syn* base conformations found in RNA helices (Supplementary Information) resulted in greater steric clashes in A-RNA as compared to B-DNA. The additional base-sugar (N3–H3′ and N3–O4′) and base-backbone (N3–O5′) clashes observed in A-RNA arise due to both the C3′-endo sugar pucker and unique phosphodiester backbone conformation at the *syn* purine residue (Figure 4.18).

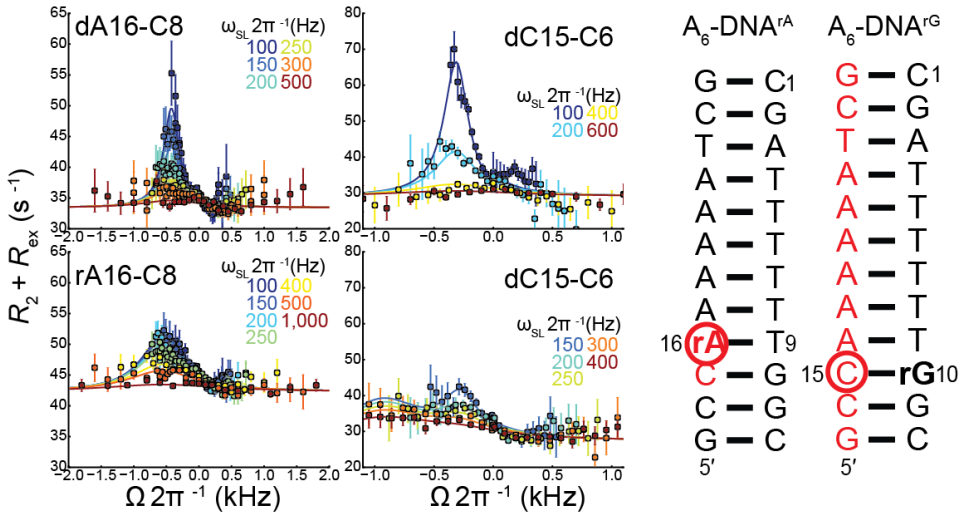

Figure 4.15: Transient HG bps in A6-DNA^{rA} and A6-DNA^{rG}.

Comparison of RD profiles measured in A$_6$-DNA^{rA} (lower left), A$_6$-DNA^{rG} (lower right), and A$_6$-DNA (upper). Error bars correspond to one s.d. estimated from mono-exponential fitting of n = 10 independently measured peak intensities using a Monte-Carlo based method (Methods). Note that the larger R_2 value in A$_6$-DNA^{rA} A16-C8 as compared to A$_6$-DNA likely reflects decreased flexibility in rA16. A$_6$-DNA^{rA} and A$_6$-DNA^{rG} duplexes with ^{13}C/^{15}N labeled residues (Methods) colored in red. rA16 and rG10 are shown in bold. Sites used as NMR RD probes are highlighted with a circle.

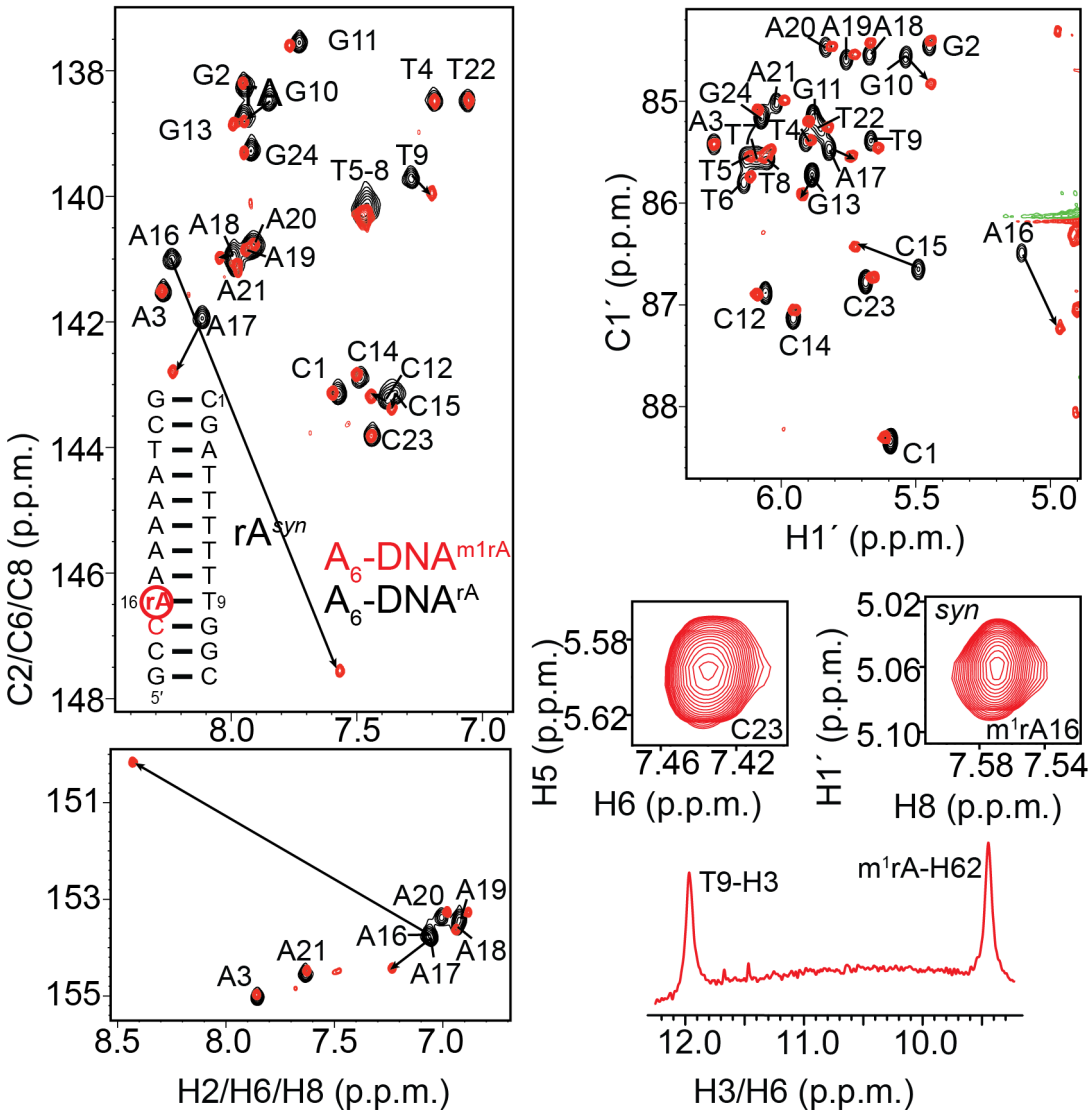

Figure 4.16: NMR evidence for m1rA–dT HG bp in A$_6$-DNA^{m1rA}.

Shown are 2D HSQC spectra of A$_6$-DNA^{m1rA} (in red) overlaid with unmodified (in black) A$_6$-DNA^{rA}, with arrows indicating significant chemical shift perturbations induced by the m^{1}rA16. Strong H1′-H8 NOE for *syn* m^{1}rA and HG imino/amino resonances.

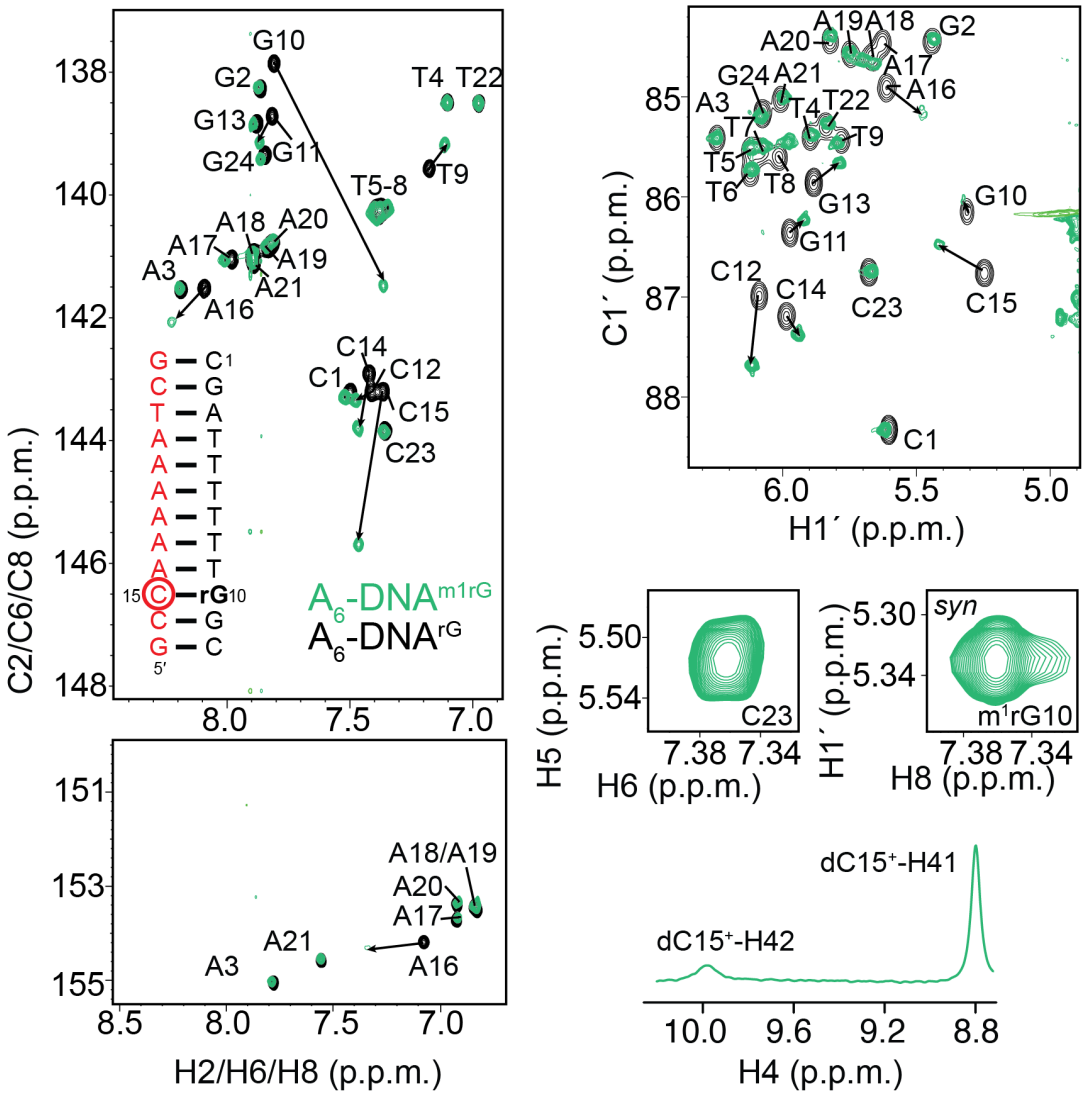

Figure 4.17: NMR evidence for m1rG–dC HG bp in A$_6$-DNA^{m1rG}.

Shown are 2D HSQC spectra of A$_6$-DNA^{m1rG} (in green) overlaid with unmodified (in black) A$_6$-DNA^{rG}, with arrows indicating significant chemical shift perturbations induced by the m^{1}rG10. Strong H1′-H8 NOE for *syn* m^{1}rG and HG amino resonances.

To further examine the energetics of the WC⇌HG transition, we carried out biased MD simulations on the A_6-DNA duplex and hp-A_6-RNA hairpin, as well as a 3′→5′ inverted sequence of the hp-A_6-RNA hairpin. A bias was applied on dA16 or rA16 starting in a WC bp configuration to force purine base flipping and a transition to a target HG configuration (Methods). The computed mean interaction energy (averaged over an ensemble of biased trajectories) as a function of the χ-angle along the WC⇌HG transition (Methods) reveals a clear two state transition in the case of B-DNA, consistent with previous results[66], whereas in the cases of A-RNA the resultant HG bp is significantly destabilized relative to its WC bp counterpart; for A-RNA, the energy profile in the *syn* region has much higher relative energies than in the case of DNA (Figure 4.19). In accord with this energetic destabilization, the simulations reveal that flipping the purine base in A-RNA is accompanied by major structural disruption of the surrounding base pairs, consistent with m^1rA induced NMR chemical shift perturbations, which were more pronounced for residues 3′ to the modified nucleotide (Figure 4.8). The extent of the disruption for the neighboring bp was far less significant in B-DNA.

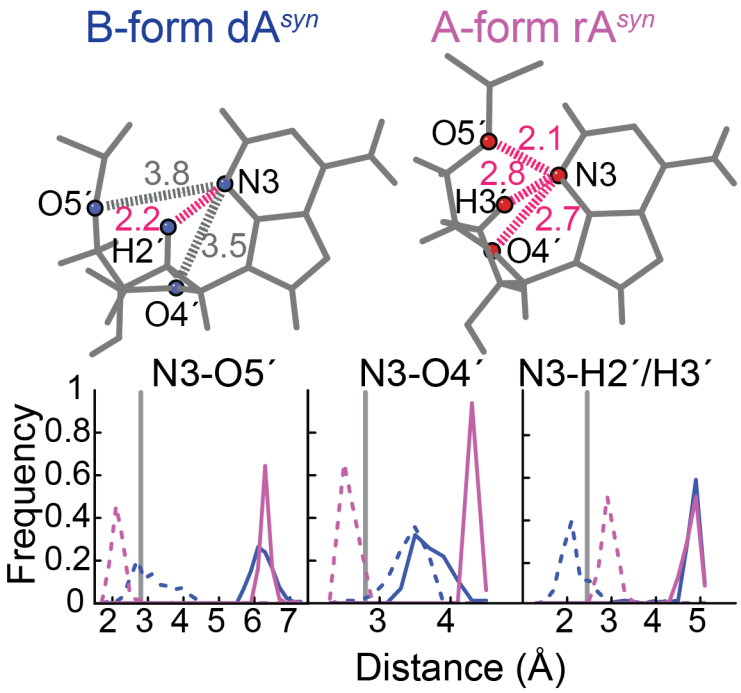

Figure 4.18: Steric analysis showing A-form helix disfavors *syn* purine conformation.

Inter-atomic distances (in Å) with unfavorable steric contacts in pink when rotating the purine base 180° around the glycoside bond in WC bps derived from idealized B-DNA and A-RNA duplexes (Methods) to adopt a *syn* conformation. Shown below are corresponding distance distributions in WC bps derived from X-ray structures of A-RNA (total 146) and B-DNA (total 159) duplexes before (solid line) and following (dashed line) 180° rotation of the purine base. The inter-atomic cut-off distance (grey line) was defined based on the van der Waals radii.

We corroborated these findings using unbiased MD simulations, which began with a HG bp embedded in various duplex and hairpin contexts (Methods). The HG H-bonding remained stable during the course of the simulation in the case of B-DNA, B-DNA containing a single rA, and B-DNA containing m^1dA (Table 3). In contrast, for A-RNA strong disruption of N7---H3-N3 HG H-bond between A16 and U9 was observed in cases of the hp-A_6-RNA hairpin (Table 3). In ≈35% of the trials in the case of 3′→5′ sequence hp-A_6-RNA, the HG bp even transitioned rapidly after equilibration back to a WC bp. Strikingly, in the case of m^1rA embedded in A-RNA, the HG bp caused melting of the A-form helix (Figure 4.19).

Taken together, these results indicate that HG bps are disfavored in the more compact A-RNA helix due to steric contacts that are difficult to alleviate without substantially perturbing the A-form helix structure.

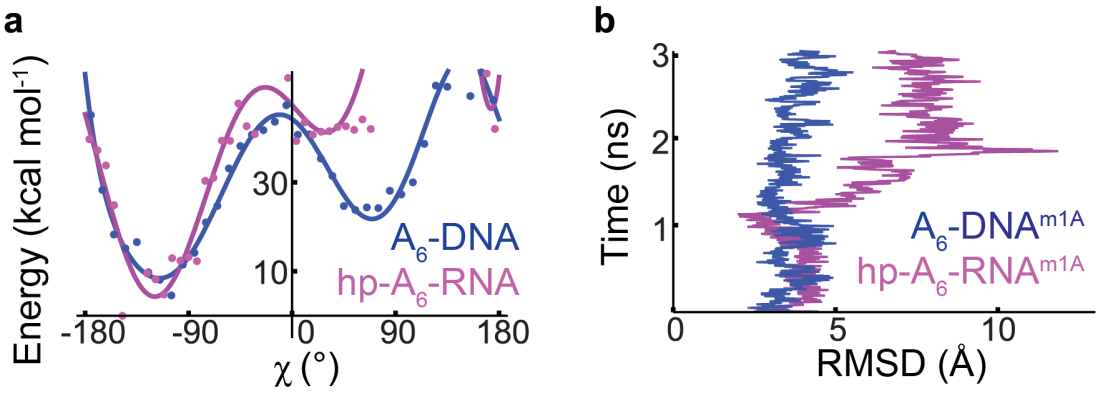

Figure 4.19: MD simulations showing A-form helix disfavors HG bp.

(a) Relative interaction energy versus χ-angle from biased MD trajectories of A_6-DNA (blue) and hp-A_6-RNA (violet). (b) Simulation time (ns) versus the global RMSD (Methods) for single A_6-DNA^{m1A} and hp-A_6-RNA^{m1A} trajectories depicting the destabilization of the RNA strand within the time of the simulation.

Table 3: HG H-bonding in unbiased MD simulations.

	N7---H-N3 (%)	O4---H-N6 (%)
A_6-DNA-HG	89.0±7.0	97.5±1.9
hp-A_6-RNA-HG (run1)	2.4	99.3
hp-A_6-RNA-HG (run2)	82.7	37.5
A_6-DNA^{m1A}-HG	61.6	99.7
hp-A_6-RNA^{m1A}-HG	6.6	37.5
A_6-DNA^{rA}-HG	97.5	98.5

4.4 Discussion

Duplex B-DNA can stably accommodate dA–dT and dG–dC$^+$ HG bps which can in turn play roles in sequence-specific DNA recognition, damage induction and repair, and in DNA replication. In contrast, our results indicate that rA–rU and rG–rC$^+$ HG bps are so unstable in the more compressed A-RNA that melting is preferred over the HG bp conformation. It remains to be seen whether the greater instability of HG bps in A-RNA as compared to B-DNA extends to purine-purine HG mispairs (Supplementary Information), which play important roles in replication[100,244] and translation errors[245], mismatch repair[26], as well in translational reprogramming[246,247].

The markedly different stability of the A–T/U and G–C$^+$ HG bp in RNA and DNA duplexes provides a basis for achieving opposing functions at the genome and transcriptome levels (Figure 4.20). If DNA did not have a capacity to form HG bps, and instead behaved like RNA, lesions such as m^{1}dA and m^{1}dG that block canonical WC base pairing could greatly destabilize the double helix and potentially cause genomic instability (Figure 4.20). The ability to form HG bps therefore endows DNA with an additional layer of chemical stability over its RNA counterpart that goes beyond resistance to hydrolysis due to the absence of the sugar 2′-OH group. On the other hand, the greater instability of HG bps in A-RNA gives rise to a chemical switch in the form of m^{1}rA and m^{1}rG that can potently modulate RNA structure (Figure 4.20). While it has long been recognized that m^{1}A and m^{1}G can modulate the structure and function of

tRNA, rRNA, and other non-coding RNAs[227-229,248], this functionality hinges on the unique instability of HG bps in A-RNA uncovered in this work.

For example, m^1rA9 has been shown to stabilize the native structures of human mitochondrial tRNAs by blocking helical rA–rU WC bps that would otherwise stabilize alternative secondary structures[248] (Figure 4.20). Likewise, m^1rG37 next to the anti-codon loop, which is highly conserved in most tRNAs that read the CNN codon, has been shown to prevent +1 frameshifting by blocking base-pairing between G37 and the first rC in the codon sequence[227,229] (Figure 4.20). If RNA behaved like DNA, such posttranscriptional modifications would simply create HG bps and fail to block base pairing and have their intended functional consequence (Figure 4.20).

m^1rA was recently shown to be a reversible mRNA modification in eukaryotic cells, from yeast to mammals, that can respond to changes in physiological conditions[232,233]. It is enriched in the 5′ UTR near start codons and was shown to promote translation through mechanisms that are not yet understood[232,233]. The formation of stable mRNA secondary structure around start codons has been shown to reduce translational efficiency[249,250]. Although it is unclear whether these m^1rA modifications target adenine nucleotides involved in WC base pairing, it is possible that m^1rA enhances translation in part by destabilizing secondary structure at the 5′ UTR near the start codons. Indeed, based on our results, m^1rA should also be capable of stabilizing alternative RNA secondary structures that feature bulged adenosines even if they are

disfavored by as much as ≈5 kcal mol^{-1} in the absence of the modification. At the same time, placement of m^1rA in an unpaired bulged conformation can make it accessible to demethylases for achieving efficient reversible control at the epitranscriptomic level (Figure 4.20). Further studies are needed to test this proposed mechanism for m^1A-enhanced translation.

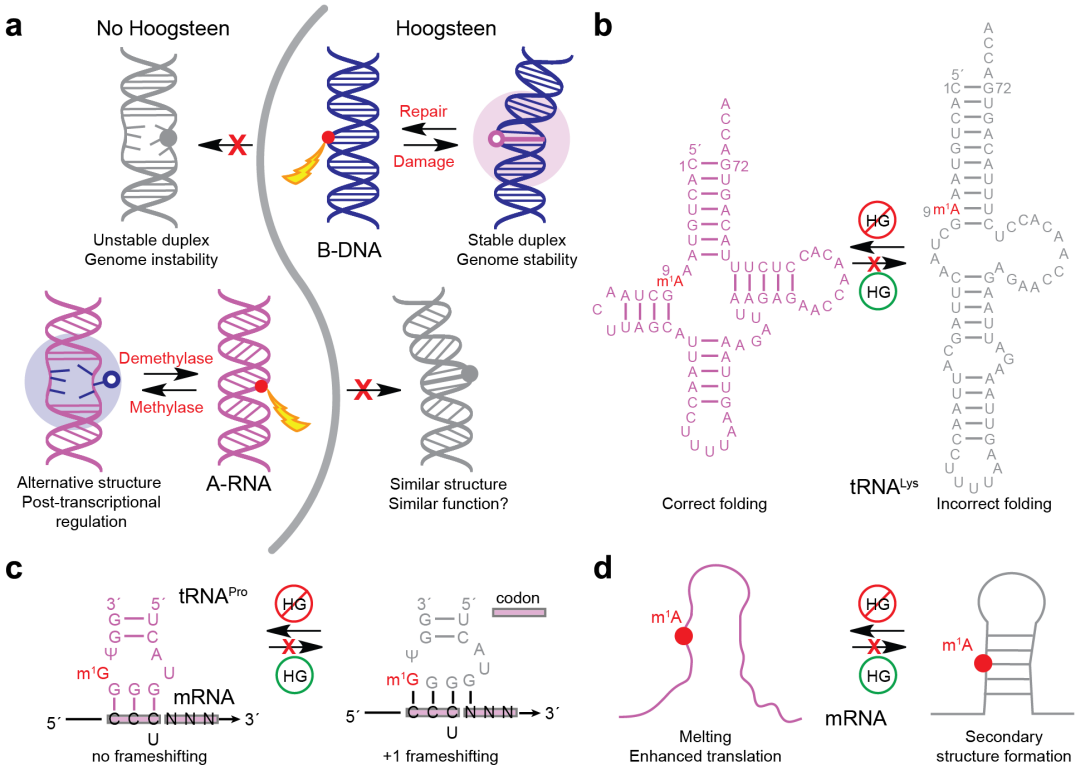

Figure 4.20: Different propensities to form HG bps in B-DNA and A-RNA enable contrasting roles at the genome and transcriptome level.

(a) In DNA, m^{1}dA or m^{1}dG damage is absorbed as HG bps that can be recognized by repair enzymes (in red). Had B-DNA lacked the ability to form HG bps, damage could result in duplex melting and genomic instability. In RNA, post-transcriptional modifications resulting in m^{1}rA and m^{1}rG block both WC and HG pairing, melting or modulating RNA secondary structure to favor functional states or effect epigenetic regulation. Had A-RNA had the ability to form HG, the m^{1}rA and m^{1}rG would form HG bps and potentially fail to more significantly alter RNA structure and function. (b) Highly conserved m^{1}rA9 in human mitochondrial tRNA^{Lys} blocks rA–rU

WC base pairing and stabilizes native tRNA structure in which m^1rA9 is in a single strand[248]. The m^1rA9 modification would not stabilize native tRNA structure if it were simply absorbed as a HG bp. (c) Highly conserved m^1rG37 next to the anti-codon loop[227] blocks base pairing between m^1rG37 and the first rC in the codon and prevents +1 frameshifting in $tRNA^{Pro}$, which could occur if m^1rG37 formed stable HG bp with rC. (d) Proposed mechanism for m^1rA enhanced translation through destabilization of secondary structure in the 5′ UTR of mRNA.

4.5 Supplementary Information

4.5.1 *syn* purines in A-form helices

The PDB survey identified one helical rA–rU HG bp in RNA which is surrounded by non-canonical motifs in the X-ray structure of the P4-P6 domain of the Group I intron RNA (PDBID: 1L8V[208]). Interestingly, the χ-angle (≈70º) for the *syn* rA is similar to χ-angles found in DNA HG bps (≈70º), which in turn differs by ≈170º from typical *anti* χ-angles in B-DNA but by ≈130º from typical *anti* χ-angles in A-RNA. Accommodation of this HG bp is accompanied by significant changes in the torsion angles α and γ which resemble those in Z-form DNA duplexes. In addition, the residue 3′ to the *syn* rA adopts a DNA-like C2′-endo sugar pucker. The survey also identified nine HG-like bps with *syn* purines in A-RNA duplexes that do not form HG-type H-bonds. Similar HG-like conformations were previously reported in B-DNA[155]. In these HG-like bps, the C1′–C1′ distance (≈10.5 Å) is not constricted as in HG bps (≈8.5 Å) but remains WC-like (≈10.7 Å). The *syn* purine is accommodated through changes in backbone angles α, γ, δ and ε. In addition, rG^{syn}–rA^{anti} and rG^{syn}–rG^{anti} HG mismatches were identified in A-RNA with even larger C1′–C1′ distance (11.3±0.1 Å). Here, the *syn* purine is accommodated through changes in the torsion angles α and γ. Interestingly, the χ-angle for these *syn* purine bases (≈15º) differs by ≈180º relative to the *anti* χ-angle in A-RNA (≈-160º). These results indicate that while *syn* purines can be accommodated within A-RNA, constriction of the C1′–C1′ distance may require substantial changes in

the A-RNA structure. We did not identify any additional steric clashes when adding a methyl group at the N^1-position of the *syn* purine base in A-RNA or B-DNA, indicating that the lower stability of HG bps in m^1rA and m^1rG containing A-RNA is not due to unique steric contacts involving the methyl group as verified in molecular dynamics (MD) simulations.

4.5.2 HG chemical shifts in RNA

In B-DNA, HG bps result in downfield shifts (≈3 p.p.m.) in the sugar-C1′, purine-C8, and dC-C6 carbon chemical shifts and smaller (≈1.8 p.p.m.) upfield shifts in the dG-N1 and dT-N3 imino nitrogen chemical shifts[66,112]. The downfield shift in sugar-C1′ [131,207] and purine-C8 are both attributed to the change in the χ-angle from *anti* to *syn* conformation whereas the upfield shift in dT-N3 and dG-N1 is attributed to weaker H-bonds[112]. The downfield shift in dC-C6 is due to protonation of dC-N3[111]. The absence of detectable RD in WC bps within A-RNA is unlikely to be due to smaller differences in chemical shifts between WC and HG for the various spins targeted for RD measurements. First, significant (1.4–5.8 p.p.m.) downfield shifted purine-C8 and/or purine-C1′ chemical shifts have been reported for *syn* rG in the UUCG tetraloop (e.g. BMRID: 16431[251]) which forms a reverse wobble bp ("UUCG"), rG^{anti}–rG^{syn} mispairs near a bulge in the HIV-1 Rev responsive element ("RRE")[252], transient HG bps in rA16 and rG10 substituted A_6-DNA ("RD"), and trapped HG bp in m^1rG10 substituted A_6-DNA.

Second, DFT calculations on a variety of HG bp configurations suggest significant differences in chemical shifts for WC and HG bps. These HG configurations include the rA–rU HG bp in P4-P6 (PDBID: 1L8V[208]); tertiary rG–rC$^+$ HG bp in the structure of 23S ribosomal RNA-protein complex (PDBID: 3U56[209]); snapshots of rA–rU HG bp from the biased MD simulations (Methods); and rG^{syn}–rG^{anti} HG mismatches (PDBID: 3CZW[210]). In all cases, we observe sizeable downfield shifts in purine-C8 (4.3±2.8 p.p.m.) and C1′ (2.8±1.9 p.p.m.) consistent with the values observed in BMRB. Finally, sizeable chemical shift changes are expected for base N1/N3 and protonated dC$^+$-C6 even for these specific HG configurations.

We also carried out DFT calculations on a number of different HG configurations in RNA that feature C3′-endo or C2′-endo sugar with and without N^1-methyl. These data show that for most configurations, large changes in chemical shifts are expected in either or both the base and sugar chemical shifts due to WC-HG transitions. In general, the calculations indicate that the N^1-methyl induces a downfield shift for rA-C8 in both *anti* and *syn* purine conformations, analogous to what is observed in DNA, but that it minimally affects the chemical shifts of *anti* or *syn* rG-C8 and the sugar-C1′.

5. Conclusions and Future Perspectives

DNA has been pictured, for decades, as being predominantly a right-handed B-form double helix composed of WC bps. The WC bps not only provide the basic mechanisms for templated replication, transcription, and translation, but also provide the basic architectural building block that defines the structure of DNA and its interactions with proteins. HG bps provide an alternative building block that can significantly alter the structural and chemical properties of the double helix and thereby expand its functional complexity. Important roles for HG bps have been identified in by-passing replication damage, in DNA damage accommodation and repair, and in DNA-protein recognition. However, little is known about the broader occurrence of HG bps in nucleic acid duplexes and their roles in DNA and RNA.

Our survey of HG bps in B-DNA shows that they can robustly exist in various structural and biological contexts; that they are enriched in AT-rich sequences, with stronger preferences for A–T versus G–C bps, TA versus GG steps, and also at terminal ends. The survey also suggests that Hoogsteen base pairs induce a small but significant degree of DNA bending (~14°) directed toward the major groove. Together, these studies suggest that HG bps may indeed be enriched in regions of severe DNA structural distortions, such as kinking, that may destabilize the WC geometry.

In collaboration with the laboratories of Drs. Jane and David Richardson, we have identified several examples of potential HG bps that were mis-modeled as WC bps.

Many of these suspected HG bps exist in regions that favor HG bps, such as AT-rich sequences or nucleosomes that feature sharp DNA bending. This motivated our development of new NMR methods for the sensitive and robust detection of HG bps in DNA-protein complexes under solution conditions. This method takes advantage of site-specific labeling to overcome the challenges arising due to spectral overlap in large systems. Application of this methodology to a complex formed between DNA and the Integration Host Factor protein suggests enhanced chemical exchange between Watson-Crick and Hoogsteen base pairs at a sharply kinked site which forms a HG base pair based on X-ray crystallographic analysis.

Neither condition used in X-ray crystallography nor NMR spectroscopy perfectly reproduce the conditions *in vivo* where DNA forms higher order chromatin structures. An important goal for the future will be to examine HG bps in chromatin. Methods based on solid-state NMR and chemical probing offer promising approaches for exploring the occurrence of HG bps in the more relevant *in vivo* environment. It is possible that tight compaction and torsional stress in chromatin creates an environment where HG bps are even more prevalent. Further studies are also required to understand the forces that stabilize HG bps in various contexts. To what extent are these effects due to destabilization of the WC bp versus stabilization of the HG bp? Are differences in stacking interactions major contributors to these forces or are there other electrostatic effects?

So far, studies have focused on HG bps in duplex B-DNA. We examined to what extent do HG bps also form in canonical A-form RNA duplexes. Surprisingly, we found no evidence for transient HG bps in A-RNA duplexes using NMR RD over a wide range of sequence and structural contexts and across a wide range of environmental conditions. Because of this inability to form HG bps, the posttranscriptional modifications such as m^1A and m^1G that block WC paring, significantly destabilize A-RNA duplexes. Thus while the ability to form HG bps endows DNA with an ability to absorb damage including m^1A and m^1G modifications, the more compressed A-form RNA suppresses HG pairing so strongly that melting is preferred over the HG conformation. This gives rise to a chemical switch in the form of m^1rA and m^1rG that can potently modulate the structure, and thus the function, of the epitranscriptome. Thus, the markedly different stabilities of the HG bp in RNA and DNA duplexes provides a basis for achieving opposing functions at the genome and transcriptome levels. If DNA did not have the capacity to form HG bps, and instead behaved like RNA, lesions such as m^1dA and m^1dG that block canonical WC base pairing would greatly destabilize the double helix and potentially cause genomic instability. Our studies indicate that the HG bps are disfavored in A-RNA due to steric effects arising due to the unique sugar pucker and phosphodiester backbone in A-RNA. Further studies are needed to examine whether purine-purine HG bps are also more destabilized in A-RNA as compared to B-

DNA and to explore other factors such as water interactions that could influence the relative stabilities of HG bps in DNA and RNA.

Based on our studies of HG bps in A-RNA, we proposed a model m^1rA enhanced translation that involves destabilizing secondary structure at the 5′ UTR near the start codons. This proposed mechanism for m^1A-enhanced translation could be tested in the future using RNA structure mapping *in vivo* approaches, analogous to those used to map out the effects of m^6A. These studies can also illuminate the structure of the mRNA at sites bearing these modifications. Our studies also suggest that m^1rA should also be capable of stabilizing alternative RNA secondary structures that feature bulged adenosines even if they are disfavored by as much as ≈5 kcal mol^{-1} in the absence of the modification. Additional studies should be carried out in the future to examine such potential effects for known m^1A sites in rRNA, tRNA, and mRNA. Additionally, it will be of interest to examine whether Dimroth rearrangement could occur *in vivo*. If this were true, m^1A and m^6A could inter-convert and provide another efficient mechanism for the epitranscriptomic regulation by switching the modified site into less or more potent helix destabilizers.

References

1 Watson, J. D. & Crick, F. H. C. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* **171**, 737-738, (1953).

2 Wyatt, G. R. THE NUCLEIC ACIDS OF SOME INSECT VIRUSES. *The Journal of General Physiology* **36**, 201-205, (1952).

3 Watson, J. D. & Crick, F. H. Genetical implications of the structure of deoxyribonucleic acid. *Nature* **171**, 964-967, (1953).

4 Hoogsteen, K. The structure of crystals containing a hydrogen-bonded complex of 1-methylthymine and 9-methyladenine. *Acta Crystallographica* **12**, 822-823, (1959).

5 Hoogsteen, K. The crystal and molecular structure of a hydrogen-bonded complex between 1-methylthymine and 9-methyladenine. *Acta Crystallographica* **16**, 907-916, (1963).

6 Donohue, J. Fourier Analysis and the Structure of DNA. *Science* **165**, 1091-1096, (1969).

7 Wilkins, M. H. F. *et al.* Fourier Analysis and the Structure of DNA. *Science* **167**, 1693-1702, (1970).

8 Haschemeyer, A. E. & Sobell, H. M. THE CRYSTAL STRUCTURE OF AN INTERMOLECULAR NUCLEOSIDE COMPLEX: ADENOSINE AND 5-BROMOURIDINE. *Proc. Natl. Acad. Sci. U. S. A.* **50**, 872-877, (1963).

9 Mathews, F. S. & Rich, A. The molecular structure of a hydrogen bonded complex of N-ethyl adenine and N-methyl uracil. *J. Mol. Biol.* **8**, 89-95, (1964).

10 Haschemeyer, A. E. & Sobell, H. M. The crystal structure of a hydrogen bonded complex of deoxyguanosine and 5-bromodeoxycytidine. *Acta Crystallographica* **19**, 125-130, (1965).

11 Sobell, H. M., Tomita, K. I. & Rich, A. The crystal structure of an intermolecular complex containing a guanine and a cytosine derivative. *Proc. Natl. Acad. Sci. U. S. A.* **49**, 885-892, (1963).

12 Pauling, L. & Corey, R. B. Specific hydrogen-bond formation between pyrimidines and purines in deoxyribonucleic acids. *Arch. Biochem. Biophys.* **65**, 164-181, (1956).

13 Arnott, S., Wilkins, M. H. F., Hamilton, L. D. & Langridge, R. Fourier synthesis studies of lithium DNA. *J. Mol. Biol.* **11**, 391-402, (1965).

14 Donohue, J. HYDROGEN-BONDED HELICAL CONFIGURATIONS OF POLYNUCLEOTIDES. *Proc. Natl. Acad. Sci. U. S. A.* **42**, 60-65, (1956).

15 Wang, A. H. *et al.* Molecular structure of a left-handed double helical DNA fragment at atomic resolution. *Nature* **282**, 680-686, (1979).

16 Day, R. O., Seeman, N. C., Rosenberg, J. M. & Rich, A. A crystalline fragment of the double helix: the structure of the dinucleoside phosphate guanylyl-3',5'-cytidine. *Proc. Natl. Acad. Sci. U. S. A.* **70**, 849-853, (1973).

17 Rosenberg, J. M. *et al.* Double helix at atomic resolution. *Nature* **243**, 150-154, (1973).

18 Wing, R. *et al.* Crystal structure analysis of a complete turn of B-DNA. *Nature* **287**, 755-758, (1980).

19 Drew, H. R. *et al.* Structure of a B-DNA dodecamer: conformation and dynamics. *Proc. Natl. Acad. Sci. U. S. A.* **78**, 2179-2183, (1981).

20 Kearns, D. R., Patel, D. J. & Shulman, R. G. High Resolution Nuclear Magnetic Resonance Studies of Hydrogen Bonded Protons of tRNA in Water. *Nature* **229**, 338-339, (1971).

21 Cross, A. D. & Crothers, D. M. Proton magnetic resonance study of single-stranded and double-helical deoxyribooligonucleotides. *Biochemistry* **10**, 4015-4023, (1971).

22 Wong, Y. P., Kearns, D. R., Reid, B. R. & Shulman, R. G. Investigation of exchangeable protons and the extent of base pairings in yeast phenylalanine transfer RNA by high resolution nuclear magnetic resonance. *J. Mol. Biol.* **72**, 725-740, (1972).

23 Crothers, D. M., Hilbers, C. W. & Shulman, R. G. Nuclear Magnetic Resonance Study of Hydrogen-Bonded Ring Protons in Watson-Crick Base Pairs. *Proc. Natl. Acad. Sci. U. S. A.* **70**, 2899-2901, (1973).

24 Patel, D. J. & Tonelli, A. E. Assignment of the proton nmr chemical shifts of the T–N3H and G–N1H proton resonances in isolated AT and GC Watson-Crick base pairs in double-stranded deoxy oligonucleotides in aqueous solution. *Biopolymers* **13**, 1943-1964, (1974).

25 Kallenbach, N. R., Daniel, W. E. & Kaminker, M. A. Nuclear magnetic resonance study of hydrogen-bonded ring protons in oligonucleotide helices involving classical and nonclassical base pairs. *Biochemistry* **15**, 1218-1224, (1976).

26 Nikolova, E. N. *et al.* A historical account of hoogsteen base-pairs in duplex DNA. *Biopolymers* **99**, 955-968, (2013).

27 Dickerson, R. E. *et al.* The anatomy of A-, B-, and Z-DNA. *Science* **216**, 475-485, (1982).

28 Davies, D. R. & Baldwin, R. L. X-ray studies of two synthetic DNA copolymers. *J. Mol. Biol.* **6**, 251-IN256, (1963).

29 Selsing, E., Arnott, S. & Ratliff, R. L. Conformations of poly[d(A-T-T)]·poly[d(A-A-T)]. *J. Mol. Biol.* **98**, 243-248, (1975).

30 Drew, H. R. & Dickerson, R. E. A new model for DNA containing A.T and I.C base pairs. *The EMBO Journal* **1**, 663-667, (1982).

31 Ikehara, M., Hattori, M. & Fukui, T. Synthesis and Properties of Poly(2-Methyladenylic Acid). *Eur. J. Biochem.* **31**, 329-334, (1972).

32 Ishikawa, F., Frazier, J., Howard, F. B. & Miles, H. T. Polyadenylate polyuridylate helices with non-Watson-Crick hydrogen bonding. *J. Mol. Biol.* **70**, 475-490, (1972).

33 Hattori, M., Ikehara, M. & Miles, H. T. Poly(2-methyl-N6-methyladenylic acid). Synthesis, properties, and interaction with poly(uridylic acid). *Biochemistry* **13**, 2754-2761, (1974).

34 Liu, K., Miles, H. T., Frazier, J. & Sasisekharan, V. A novel DNA duplex. A parallel-stranded DNA helix with Hoogsteen base pairing. *Biochemistry* **32**, 11802-11809, (1993).

35 Segers-Nolten, G. M. J., Sijtsema, N. M. & Otto, C. Evidence for Hoogsteen GC Base Pairs in the Proton-Induced Transition from Right-Handed to Left-Handed Poly(dG-dC)·Poly(dG-dC). *Biochemistry* **36**, 13241-13247, (1997).

36 Blommers, M. J. J., Van De Ven, F. J. M., Van Der Marel, G. A., Van Boom, J. H. & Hilbers, C. W. The three-dimensional structure of a DNA hairpin in solution. *Eur. J. Biochem.* **201**, 33-51, (1991).

37 Ronning, D. R. *et al.* Active Site Sharing and Subterminal Hairpin Recognition in a New Class of DNA Transposases. *Mol. Cell* **20**, 143-154, (2005).

38 Abrescia, N. G., Thompson, A., Huynh-Dinh, T. & Subirana, J. A. Crystal structure of an antiparallel DNA fragment with Hoogsteen base pairing. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 2806-2811, (2002).

39 Abrescia, N. G., González, C., Gouyette, C. & Subirana, J. A. X-ray and NMR studies of the DNA oligomer d(ATATAT): Hoogsteen base pairing in duplex DNA. *Biochemistry* **43**, 4092-4100, (2004).

40 Pous, J. *et al.* Stabilization by extra-helical thymines of a DNA duplex with Hoogsteen base pairs. *J. Am. Chem. Soc.* **130**, 6755-6760, (2008).

41 De Luchi, D., Tereshko, V., Gouyette, C. & Subirana, J. A. Structure of the DNA coiled coil formed by d(CGATATATATAT). *ChemBioChem* **7**, 585-587, (2006).

42 Acosta-Reyes, F. J., Alechaga, E., Subirana, J. A. & Campos, J. L. Structure of the DNA Duplex d(ATTAAT)₂ with Hoogsteen Hydrogen Bonds. *PLoS One* **10**, e0120241, (2015).

43 Wang, A. H. *et al.* The molecular structure of a DNA-triostin A complex. *Science* **225**, 1115-1121, (1984).

44 Ward, D. C., Reich, E. & Goldberg, I. H. Base Specificity in the Interaction of Polynucleotides with Antibiotic Drugs. *Science* **149**, 1259-1263, (1965).

45 Sato, K., Shiratori, O. & Katagiri, K. The mode of action of quinoxaline antibiotics. Interaction of quinomycin A with deoxyribonucleic acid. *The Journal of antibiotics* **20**, 270-276, (1967).

46 Ughetto, G. *et al.* A comparison of the structure of echinomycin and triostin A complexed to a DNA fragment. *Nucleic Acids Res.* **13**, 2305-2323, (1985).

47 Quigley, G. J. *et al.* Non-Watson-Crick G.C and A.T base pairs in a DNA-antibiotic complex. *Science* **232**, 1255-1258, (1986).

48 Mendel, D. & Dervan, P. B. Hoogsteen base pairs proximal and distal to echinomycin binding sites on DNA. *Proc. Natl. Acad. Sci. U. S. A.* **84**, 910-914, (1987).

49 Herr, W. Diethyl pyrocarbonate: a chemical probe for secondary structure in negatively supercoiled DNA. *Proceedings of the National Academy of Sciences* **82**, 8009-8013, (1985).

50 Scholten, P. M. & Nordheim, A. Diethyl pyrocarbonate: a chemical probe for DNA cruciforms. *Nucleic Acids Res.* **14**, 3981-3993, (1986).

51 McLean, M. J. & Waring, M. J. Chemical probes reveal no evidence of Hoogsteen base pairing in complexes formed between echinomycin and DNA in solution. *J. Mol. Recognit.* **1**, 138-151, (1988).

52 Portugal, J., Fox, K. R., McLean, M. J., Richenberg, J. L. & Waring, M. J. Diethyl pyrocarbonate can detect a modified DNA structure induced by the binding of quinoxaline antibioties. *Nucleic Acids Res.* **16**, 3655-3670, (1988).

53 McLean, M. J., Seela, F. & Waring, M. J. Echinomycin-induced hypersensitivity to osmium tetroxide of DNA fragments incapable of forming Hoogsteen base pairs. *Proc. Natl. Acad. Sci. U. S. A.* **86**, 9687-9691, (1989).

54 Sayers, E. W. & Waring, M. J. Footprinting titration studies on the binding of echinomycin to DNA incapable of forming Hoogsteen base pairs. *Biochemistry* **32**, 9094-9107, (1993).

55 Gao, X. & Patel, D. J. NMR studies of echinomycin bisintercalation complexes with d(A1-C2-G3-T4) and d(T1-C2-G3-A4) duplexes in aqueous solution: sequence-dependent formation of Hoogsteen A1.cntdot.T4 and Watson-Crick T1.cntdot.A4 base pairs flanking the bisintercalation site. *Biochemistry* **27**, 1744-1751, (1988).

56 Gilbert, D. E., van der Marel, G. A., van Boom, J. H. & Feigon, J. Unstable Hoogsteen base pairs adjacent to echinomycin binding sites within a DNA duplex. *Proceedings of the National Academy of Sciences* **86**, 3006-3010, (1989).

57 Gilbert, D. E. & Feigon, J. The DNA sequence at echinomycin binding sites determines the structural changes induced by drug binding: NMR studies of

echinomycin binding to [d(ACGTACGT)]2 and [d(TCGATCGA)]2. *Biochemistry* **30**, 2483-2494, (1991).

58 Gilbert, D. E. & Feigon, J. Proton NMR study of the [d(ACGTATACGT)]2-2echinomycin complex: conformational changes between echinomycin binding sites. *Nucleic Acids Res.* **20**, 2411-2420, (1992).

59 Park, J.-Y. & Choi, B.-S. NMR Investigation of Echinomycin Binding to d(ACGTTAACGT)2: Hoogsteen versus Watson-Crick A·T• Base Pairing between Echinomycin Binding Sites. *J. Biochem. (Tokyo, Jpn.)* **118**, 989-995, (1995).

60 Gallego, J. *et al.* DNA Sequence-Specific Reading by Echinomycin: Role of Hydrogen Bonding and Stacking Interactions. *J. Med. Chem.* **37**, 1602-1609, (1994).

61 Gallego, J., Luque, F. J., Orozco, M. & Gago, F. Binding of Echinomycin to d(GCGC)2 and d(CCGG)2: Distinct Stacking Interactions Dictate the Sequence-Dependent Formation of Hoogsteen Base Pairs. *Journal of Biomolecular Structure and Dynamics* **12**, 111-129, (1994).

62 Rice, P. A., Yang, S., Mizuuchi, K. & Nash, H. A. Crystal structure of an IHF-DNA complex: a protein-induced DNA U-turn. *Cell* **87**, 1295-1306, (1996).

63 Dhavan, G. M., Lapham, J., Yang, S. & Crothers, D. M. Decreased imino proton exchange and base-pair opening in the IHF-DNA complex measured by NMR1. *J. Mol. Biol.* **288**, 659-671, (1999).

64 Patikoglou, G. A. *et al.* TATA element recognition by the TATA box-binding protein has been conserved throughout evolution. *Genes Dev.* **13**, 3217-3230, (1999).

65 Hoopes, B. C., LeBlanc, J. F. & Hawley, D. K. Contributions of the TATA box sequence to rate-limiting steps in transcription initiation by RNA polymerase II1. *J. Mol. Biol.* **277**, 1015-1031, (1998).

66 Nikolova, E. N. *et al.* Transient Hoogsteen base pairs in canonical duplex DNA. *Nature* **470**, 498-502, (2011).

67 Meyer, T., Gustafsson, J. A. N. Å. & Carlstedt-Duke, J. A. N. Glucocorticoid-Dependent Transcriptional Repression of the Osteocalcin Gene by Competitive Binding at the TATA Box. *DNA Cell Biol.* **16**, 919-927, (1997).

68 Aishima, J. *et al.* A Hoogsteen base pair embedded in undistorted B-DNA. *Nucleic Acids Res.* **30**, 5244-5252, (2002).

69 Kitayner, M. *et al.* Diversity in DNA recognition by p53 revealed by crystal structures with Hoogsteen base pairs. *Nat. Struct. Mol. Biol.* **17**, 423-429, (2010).

70 Malecka, K. A., Ho, W. C. & Marmorstein, R. Crystal structure of a p53 core tetramer bound to DNA. *Oncogene* **28**, 325-333, (2008).

71 Chen, Y., Dey, R. & Chen, L. Crystal Structure of the p53 Core Domain Bound to a Full Consensus Site as a Self-Assembled Tetramer. *Structure* **18**, 246-256, (2010).

72 Alexander, P. in *Advances in Cancer Research* Vol. Volume 2 (eds P. Greenstein Jesse & Haddow Alexander) 1-72 (Academic Press, 1954).

73 Friedberg, E. C. A brief history of the DNA repair field. *Cell Res* **18**, 3-7, (2008).

74 Patel, D. J., Shapiro, L., Kozlowski, S. A., Gaffney, B. L. & Jones, R. A. Covalent carcinogenic O6-methylguanosine lesions in DNA structural studies of the O6meG·A and O6meG·G interactions in dodecanucleotide duplexes. *J. Mol. Biol.* **188**, 677-692, (1986).

75 Kalnik, M. W., Li, B. F. L., Swann, P. F. & Patel, D. J. O6-Ethylguanine carcinogenic lesions in DNA: an NMR study of O6etG.cntdot.T pairing in dodecanucleotide duplexes. *Biochemistry* **28**, 6170-6181, (1989).

76 Kouchakdjian, M. *et al.* NMR studies of exocyclic 1,N2-propanodeoxyguanosine adducts (X) opposite purines in DNA duplexes: protonated X(syn).cntdot.A(anti) pairing (acidic pH) and X(syn).cntdot.G(anti) pairing (neutral pH) at the lesion site. *Biochemistry* **28**, 5647-5657, (1989).

77 Norman, D. *et al.* NMR and computational characterization of the N-(deoxyguanosin-8-yl)aminofluorene adduct [(AF)G] opposite adenosine in DNA: (AF)G[syn].cntdot.A[anti] pair formation and its pH dependence. *Biochemistry* **28**, 7462-7476, (1989).

78 Singh, U. S. *et al.* 1H NMR of an oligodeoxynucleotide containing a propanodeoxyguanosine adduct positioned in a (CG)3 frameshift hotspot of Salmonella typhimurium hisD3052: Hoogsteen base-pairing at pH 5.8. *Chem. Res. Toxicol.* **6**, 825-836, (1993).

79 Weisenseel, J. P., Reddy, G. R., Marnett, L. J. & Stone, M. P. Structure of an Oligodeoxynucleotide Containing a 1,N2-Propanodeoxyguanosine Adduct Positioned in a Palindrome Derived from the Salmonella typhimurium hisD3052 Gene: Hoogsteen Pairing at pH 5.2. *Chem. Res. Toxicol.* **15**, 127-139, (2002).

80 Shanmugam, G., Kozekov, I. D., Guengerich, F. P., Rizzo, C. J. & Stone, M. P. Structure of the 1,N2-ethenodeoxyguanosine adduct opposite cytosine in duplex DNA: Hoogsteen base pairing at pH 5.2. *Chem. Res. Toxicol.* **21**, 1795-1805, (2008).

81 Mao, B., Hingerty, B. E., Broyde, S. & Patel, D. J. Solution structure of the aminofluorene [AF]-intercalated conformer of the syn-[AF]-C8-dG adduct opposite dC in a DNA duplex. *Biochemistry* **37**, 81-94, (1998).

82 Yang, H., Zhan, Y., Fenn, D., Chi, L. M. & Lam, S. L. Effect of 1-methyladenine on double-helical DNA structures. *FEBS Lett.* **582**, 1629-1633, (2008).

83 Lu, L., Yi, C., Jian, X., Zheng, G. & He, C. Structure determination of DNA methylation lesions N1-meA and N3-meC in duplex DNA using a cross-linked protein-DNA system. *Nucleic Acids Res.* **38**, 4415-4425, (2010).

84 Natrajan, G. *et al.* Structures of Escherichia coli DNA mismatch repair enzyme MutS in complex with different mismatches: a common recognition mode for diverse substrates. *Nucleic Acids Res.* **31**, 4814-4821, (2003).

85 Bolli, M., Christopher Litten, J., Schu¨tz, R. & Leumann, C. J. Bicyclo-DNA: a Hoogsteen-selective pairing system. *Chem. Biol.* **3**, 197-206, (1996).

86 Isaksson, J. *et al.* The First Example of a Hoogsteen Basepaired DNA Duplex in Dynamic Equilibrium with a Watson-Crick Basepaired Duplex—A Structural (NMR), Kinetic and Thermodynamic Study. *Journal of Biomolecular Structure and Dynamics* **18**, 783-806, (2001).

87 Bailor, M. H., Sun, X. & Al-Hashimi, H. M. Topology links RNA secondary structure with global conformation, dynamics, and adaptation. *Science* **327**, 202-206, (2010).

88 Lehman, I. R., Bessman, M. J., Simms, E. S. & Kornberg, A. Enzymatic Synthesis of Deoxyribonucleic Acid: I. PREPARATION OF SUBSTRATES AND PARTIAL PURIFICATION OF AN ENZYME FROM ESCHERICHIA COLI. *J. Biol. Chem.* **233**, 163-170, (1958).

89 Ling, H., Boudsocq, F., Plosky, B. S., Woodgate, R. & Yang, W. Replication of a cis-syn thymine dimer at atomic resolution. *Nature* **424**, 1083-1087, (2003).

90 Nair, D. T., Johnson, R. E., Prakash, S., Prakash, L. & Aggarwal, A. K. Replication by human DNA polymerase-iota occurs by Hoogsteen base-pairing. *Nature* **430**, 377-380, (2004).

91 Tissier, A., McDonald, J. P., Frank, E. G. & Woodgate, R. polι, a remarkably error-prone human DNA polymerase. *Genes Dev.* **14**, 1642-1650, (2000).

92 Wang, J. DNA polymerases: Hoogsteen base-pairing in DNA replication? *Nature* **437**, E6-E7, (2005).

93 Izatt, R. M., Christensen, J. J. & Rytting, J. H. Sites and thermodynamic quantities associated with proton and metal ion interaction with ribonucleic acid, deoxyribonucleic acid, and their constituent bases, nucleosides, and and nucleotides. *Chem. Rev. (Washington, DC, U. S.)* **71**, 439-481, (1971).

94 Nair, D. T., Johnson, R. E., Prakash, L., Prakash, S. & Aggarwal, A. K. Human DNA Polymerase ι Incorporates dCTP Opposite Template G via a G.C+ Hoogsteen Base Pair. *Structure* **13**, 1569-1577, (2005).

95 Johnson, R. E., Prakash, L. & Prakash, S. Biochemical evidence for the requirement of Hoogsteen base pairing for replication by human DNA polymerase iota. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 10466-10471, (2005).

96 Nair, D. T., Johnson, R. E., Prakash, L., Prakash, S. & Aggarwal, A. K. Hoogsteen base pair formation promotes synthesis opposite the 1,N6-ethenodeoxyadenosine lesion by human DNA polymerase [igr]. *Nat. Struct. Mol. Biol.* **13**, 619-625, (2006).

97 Pence, M. G. *et al.* Lesion Bypass of N2-Ethylguanine by Human DNA Polymerase ι. *J. Biol. Chem.* **284**, 1732-1740, (2009).

98 Pence, M. G., Choi, J.-Y., Egli, M. & Guengerich, F. P. Structural Basis for Proficient Incorporation of dTTP Opposite O6-Methylguanine by Human DNA Polymerase ι. *J. Biol. Chem.* **285**, 40666-40672, (2010).

99 Kirouac, K. N. & Ling, H. Unique active site promotes error-free replication opposite an 8-oxo-guanine lesion by human DNA polymerase iota. *Proceedings of the National Academy of Sciences* **108**, 3210-3215, (2011).

100 Makarova, A. V. & Kulbachinskiy, A. V. Structure of human DNA polymerase iota and the mechanism of DNA synthesis. *Biochemistry (Mosc)* **77**, 547-561, (2012).

101 Johnson, R. E., Yu, S.-L., Prakash, S. & Prakash, L. A Role for Yeast and Human Translesion Synthesis DNA Polymerases in Promoting Replication through 3-Methyl Adenine. *Mol. Cell. Biol.* **27**, 7198-7205, (2007).

102 Plosky, B. S. *et al.* Eukaryotic Y-family polymerases bypass a 3-methyl-2′-deoxyadenosine analog in vitro and methyl methanesulfonate-induced DNA damage in vivo. *Nucleic Acids Res.* **36**, 2152-2162, (2008).

103 Petta, T. B. *et al.* Human DNA polymerase iota protects cells against oxidative stress. *The EMBO Journal* **27**, 2883-2895, (2008).

104 Rich, A. DNA comes in many forms. *Gene* **135**, 99-109, (1993).

105 Travers, A. A. The structural basis of DNA flexibility. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* **362**, 1423-1438, (2004).

106 Englander, S. W., Downer, N. W. & Teitelbaum, H. Hydrogen Exchange. *Annu. Rev. Biochem.* **41**, 903-924, (1972).

107 Gueron, M., Kochoyan, M. & Leroy, J.-L. A single mode of DNA base-pair opening drives imino proton exchange. *Nature* **328**, 89-92, (1987).

108 Guéron, M. & Leroy, J.-L. in *Methods Enzymol.* Vol. Volume 261 383-413 (Academic Press, 1995).

109 Russu, I. M. in *Methods Enzymol.* Vol. Volume 379 152-175 (Academic Press, 2004).

110 Coman, D. & Russu, I. M. Base Pair Opening in Three DNA-unwinding Elements. *J. Biol. Chem.* **280**, 20216-20221, (2005).

111 Nikolova, E. N., Goh, G. B., Brooks, C. L., III & Al-Hashimi, H. M. Characterizing the protonation state of cytosine in transient G•C Hoogsteen base pairs in duplex DNA. *J. Am. Chem. Soc.* **135**, 6766-6769, (2013).

112 Nikolova, E. N., Gottardo, F. L. & Al-Hashimi, H. M. Probing transient Hoogsteen hydrogen bonds in canonical duplex DNA using NMR relaxation dispersion and single-atom substitution. *J. Am. Chem. Soc.* **134**, 3667-3670, (2012).

113 Massi, F., Johnson, E., Wang, C., Rance, M. & Palmer, A. G. NMR R1ρ Rotating-Frame Relaxation with Weak Radio Frequency Fields. *J. Am. Chem. Soc.* **126**, 2247-2256, (2004).

114 Korzhnev, D. M., Orekhov, V. Y. & Kay, L. E. Off-Resonance R1ρ NMR studies of exchange dynamics in proteins with low spin-lock fields: an application to a Fyn SH3 domain. *J. Am. Chem. Soc.* **127**, 713-721, (2005).

115 Hansen, A. L., Nikolova, E. N., Casiano-Negroni, A. & Al-Hashimi, H. M. Extending the range of microsecond-to-millisecond chemical exchange detected in labeled and unlabeled nucleic acids by selective carbon R1ρ NMR Spectroscopy. *J. Am. Chem. Soc.* **131**, 3818-3819, (2009).

116 Alvey, H. S., Gottardo, F. L., Nikolova, E. N. & Al-Hashimi, H. M. Widespread transient Hoogsteen base pairs in canonical duplex DNA with variable energetics. *Nat Commun* **5**, 4786, (2014).

117 Li, F. *et al.* Global Analysis of RNA Secondary Structure in Two Metazoans. *Cell Reports* **1**, 69-82, (2012).

118 Cruz, J. A. & Westhof, E. The Dynamic Landscapes of RNA Architecture. *Cell* **136**, 604-609, (2009).

119 Neidle, S. *Principles of nucleic acid structure*. (Academic Press, 2010).

120 Brameld, K. A. & Goddard, W. A. Ab initio quantum mechanical study of the structures and energies for the pseudorotation of 5′-dehydroxy analogues of 2′-deoxyribose and ribose sugars. *J. Am. Chem. Soc.* **121**, 985-993, (1999).

121 Nagaswamy, U. *et al.* NCIR: a database of non-canonical interactions in known RNA structures. *Nucleic Acids Res.* **30**, 395-397, (2002).

122 Chawla, M., Oliva, R., Bujnicki, J. M. & Cavallo, L. An atlas of RNA base pairs involving modified nucleobases with optimal geometries and accurate energies. *Nucleic Acids Res.*, (2015).

123 Zhou, H. *et al.* m1A and m1G disrupt A-RNA structure through the intrinsic instability of Hoogsteen base pairs. *Nat. Struct. Mol. Biol.* **advance online publication**, (2016).

124 Cavanagh, J., Fairbrother, W. J., Palmer III, A. G. & Skelton, N. J. *Protein NMR spectroscopy: principles and practice*. (Academic Press, 1995).

125 Pervushin, K., Riek, R., Wider, G. & Wüthrich, K. Attenuated T2 relaxation by mutual cancellation of dipole–dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biological macromolecules in solution. *Proceedings of the National Academy of Sciences* **94**, 12366-12371, (1997).

126 Spera, S. & Bax, A. Empirical correlation between protein backbone conformation and C.alpha. and C.beta. 13C nuclear magnetic resonance chemical shifts. *J. Am. Chem. Soc.* **113**, 5490-5492, (1991).

127 Ulrich, E. L. *et al.* BioMagResBank. *Nucleic Acids Res.* **36**, D402-D408, (2008).

128 Greene, K. L., Wang, Y. & Live, D. Influence of the glycosidic torsion angle on 13C and 15N shifts in guanosine nucleotides: Investigations of G-tetrad models with alternating syn and anti bases. *J. Biomol. NMR* **5**, 333-338, (1995).

129 Sklenar, V., Bax, A. & Zon, G. Assignment of Z DNA NMR spectra of poly d(Gm5C) by two-dimensional multinuclear spectroscopy. *J. Am. Chem. Soc.* **109**, 2221-2222, (1987).

130 Ghose, R., Marino, J. P., Wiberg, K. B. & Prestegard, J. H. Dependence of 13C Chemical Shifts on Glycosidic Torsional Angles in Ribonucleic Acids. *J. Am. Chem. Soc.* **116**, 8827-8828, (1994).

131 Xu, X.-P. & Au−Yeung, S. C. F. Investigation of chemical shift and structure relationships in nucleic acids using NMR and density functional theory methods. *The Journal of Physical Chemistry B* **104**, 5641-5650, (2000).

132 de los Santos, C., Rosen, M. & Patel, D. NMR studies of DNA (R+)n•(Y-)n•(Y+)n triple helices in solution: imino and amino proton markers of T•A•T and C•G•C+ base-triple formation. *Biochemistry* **28**, 7282-7289, (1989).

133 Keeler, J. *Understanding NMR spectroscopy*. (John Wiley & Sons, 2011).

134 Sklenar, V. & Felgon, J. Formation of a stable triplex from a single DNA strand. *Nature* **345**, 836-838, (1990).

135 Palmer, A. G., III. Chemical exchange in biomacromolecules: past, present, and future. *J. Magn. Reson.* **241**, 3-17, (2014).

136 Sekhar, A. & Kay, L. E. NMR paves the way for atomic level descriptions of sparsely populated, transiently formed biomolecular conformers. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 12867-12874, (2013).

137 Palmer III, A. G. A dynamic look backward and forward. *J. Magn. Reson.*, (2016).

138 McConnell, H. M. Reaction Rates by Nuclear Magnetic Resonance. *The Journal of Chemical Physics* **28**, 430-431, (1958).

139 Bothe, J. R., Stein, Z. W. & Al-Hashimi, H. M. Evaluating the uncertainty in exchange parameters determined from off-resonance R1ϱ relaxation dispersion for systems in fast exchange. *J. Magn. Reson.* **244**, 18-29, (2014).

140 Trott, O. & Palmer III, A. G. Theoretical study of *R*1ϱ rotating-frame and R2 free-precession relaxation in the presence of *n*-site chemical exchange. *J. Magn. Reson.* **170**, 104-112, (2004).

141 Xue, Y. *et al.* in *Methods Enzymol.* Vol. Volume 558 (eds A. Woodson Sarah & H. T. Allain Frédéric) 39-73 (Academic Press, 2015).

142 Branch, M., Coleman, T. & Li, Y. A subspace, interior, and conjugate gradient method for large-scale bound-constrained minimization problems. *SIAM Journal on Scientific Computing* **21**, 1-23, (1999).

143 Cuesta-Seijo, J. A. & Sheldrick, G. M. Structures of complexes between echinomycin and duplex DNA. *Acta Crystallogr D Biol Crystallogr* **61**, 442-448, (2005).

144 Cuesta-Seijo, J. A., Weiss, M. S. & Sheldrick, G. M. Serendipitous SAD phasing of an echinomycin-(ACGTACGT)2 bisintercalation complex. *Acta Crystallogr D Biol Crystallogr* **62**, 417-424, (2006).

145 Pfoh, R., Cuesta-Seijo, J. A. & Sheldrick, G. M. Interaction of an echinomycin-DNA complex with manganese ions. *Acta Crystallographica Section F* **65**, 660-664, (2009).

146 García, R. G., Ferrer, E., Macías, M. J., Eritja, R. & Orozco, M. Theoretical calculations, synthesis and base pairing properties of oligonucleotides containing 8-amino-2′-deoxyadenosine. *Nucleic Acids Res.* **27**, 1991-1998, (1999).

147 Soliva, R. *et al.* DNA-triplex stabilizing properties of 8-aminoguanine. *Nucleic Acids Res.* **28**, 4531-4539, (2000).

148 Cubero, E. *et al.* Hoogsteen-Based Parallel-Stranded Duplexes of DNA. Effect of 8-Amino-purine Derivatives. *J. Am. Chem. Soc.* **124**, 3133-3142, (2002).

149 Singh, U. C., Pattabiraman, N., Langridge, R. & Kollman, P. A. Molecular mechanical studies of d(CGTACG)2: complex of triostin A with the middle A - T base pairs in either Hoogsteen or Watson-Crick pairing. *Proc. Natl. Acad. Sci. U. S. A.* **83**, 6402-6406, (1986).

150 Gallego, J., Ortiz, A. R. & Gago, F. A molecular dynamics study of the bis-intercalation complexes of echinomycin with d(ACGT)2 and d(TCGA)2: rationale for sequence-specific Hoogsteen base pairing. *J. Med. Chem.* **36**, 1548-1561, (1993).

151 Mittermaier, A. K. & Kay, L. E. Observing biological dynamics at atomic resolution using NMR. *Trends Biochem. Sci.* **34**, 601-611, (2009).

152 SantaLucia, J., Jr., Allawi, H. T. & Seneviratne, P. A. Improved nearest-neighbor parameters for predicting DNA duplex stability. *Biochemistry* **35**, 3555-3562, (1996).

153 Nikolova, E. N., Stull, F. & Al-Hashimi, H. M. Guanine to Inosine Substitution Leads to Large Increases in the Population of a Transient G•C Hoogsteen Base Pair. *Biochemistry* **53**, 7145-7147, (2014).

154 Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28**, 235-242, (2000).

155 Zhou, H. *et al.* New insights into Hoogsteen base pairs in DNA duplexes from a structure-based survey. *Nucleic Acids Res.* **43**, 3420-3433, (2015).

156 Olson, W. K. *et al.* A standard reference frame for the description of nucleic acid base-pair geometry. *J. Mol. Biol.* **313**, 229-237, (2001).

157 Lu, X. J. & Olson, W. K. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res.* **31**, 5108-5121, (2003).

158 Ethayathulla, A. S. *et al.* Structure of p73 DNA-binding domain tetramer modulates p73 transactivation. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 6066-6071, (2012).

159 Yang, S., Salmon, L. & Al-Hashimi, H. M. Measuring similarity between dynamic ensembles of biomolecules. *Nat. Methods* **11**, 552-554, (2014).

160 Fisher, C. K., Huang, A. & Stultz, C. M. Modeling Intrinsically Disordered Proteins with Bayesian Statistics. *J. Am. Chem. Soc.* **132**, 14919-14927, (2010).

161 Bailor, M. H., Mustoe, A. M., Brooks, C. L., III & Al-Hashimi, H. M. 3D maps of RNA interhelical junctions. *Nat. Protoc.* **6**, 1536-1545, (2011).

162 Musselman, C. *et al.* Impact of static and dynamic A-form heterogeneity on the determination of RNA global structural dynamics using NMR residual dipolar couplings. *J. Biomol. NMR* **36**, 235-249, (2006).

163 Lu, X. J. & Olson, W. K. 3DNA: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. *Nat. Protoc.* **3**, 1213-1227, (2008).

164 Bailor, M. H. *et al.* Characterizing the relative orientation and dynamics of RNA A-form helices using NMR residual dipolar couplings. *Nat. Protoc.* **2**, 1536-1546, (2007).

165 Olson, W. K., Gorin, A. A., Lu, X.-J., Hock, L. M. & Zhurkin, V. B. DNA sequence-dependent deformability deduced from protein–DNA crystal complexes. *Proceedings of the National Academy of Sciences* **95**, 11163-11168, (1998).

166 Dans, P. D., Pérez, A., Faustino, I., Lavery, R. & Orozco, M. Exploring polymorphisms in B-DNA helical conformations. *Nucleic Acids Res.* **40**, 10668-10678, (2012).

167 Svozil, D., Kalina, J., Omelka, M. & Schneider, B. DNA conformations and their sequence preferences. *Nucleic Acids Res.* **36**, 3690-3706, (2008).

168 MacDonald, D., Herbert, K., Zhang, X., Pologruto, T. & Lu, P. Solution structure of an A-tract DNA bend. *J. Mol. Biol.* **306**, 1081-1098, (2001).

169 Chua, E. Y., Vasudevan, D., Davey, G. E., Wu, B. & Davey, C. A. The mechanics behind DNA sequence-dependent properties of the nucleosome. *Nucleic Acids Res.* **40**, 6338-6352, (2012).

170 Lavery, R. & Sklenar, H. The definition of generalized helicoidal parameters and of axis curvature for irregular nucleic acids. *J. Biomol. Struct. Dyn.* **6**, 63-91, (1988).

171 Kapral, G. J. *et al.* New tools provide a second look at HDV ribozyme structure, dynamics and cleavage. *Nucleic Acids Res.* **42**, 12833-12846, (2014).

172 SantaLucia, J., Jr. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 1460-1465, (1998).

173 Kojima, C., Ono, A., Kainosho, M. & James, T. L. DNA duplex dynamics: NMR relaxation studies of a decamer with uniformly 13C-labeled purine nucleotides. *J. Magn. Reson.* **135**, 310-333, (1998).

174 Cubero, E., Luque, F. J. & Orozco, M. Theoretical study of the Hoogsteen-Watson-Crick junctions in DNA. *Biophys. J.* **90**, 1000-1008, (2006).

175 Cubero, E., Abrescia, N. G. A., Subirana, J. A., Luque, F. J. & Orozco, M. Theoretical Study of a New DNA Structure: The Antiparallel Hoogsteen Duplex. *J. Am. Chem. Soc.* **125**, 14603-14612, (2003).

176 Matsuda, S. *et al.* Efforts toward expansion of the genetic alphabet: structure and replication of unnatural base pairs. *J. Am. Chem. Soc.* **129**, 10466-10473, (2007).

177 Cho, Y., Gorina, S., Jeffrey, P. D. & Pavletich, N. P. Crystal structure of a p53 tumor suppressor-DNA complex: understanding tumorigenic mutations. *Science* **265**, 346-355, (1994).

178 Nelson, H. C., Finch, J. T., Luisi, B. F. & Klug, A. The structure of an oligo(dA)•oligo(dT) tract and its biological implications. *Nature* **330**, 221-226, (1987).

179 Vivian, J. P., Porter, C. J., Wilce, J. A. & Wilce, M. C. An asymmetric structure of the Bacillus subtilis replication terminator protein in complex with DNA. *J. Mol. Biol.* **370**, 481-491, (2007).

180 Meijsing, S. H. *et al.* DNA binding site sequence directs glucocorticoid receptor structure and activity. *Science* **324**, 407-410, (2009).

181 Luger, K., Mader, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8[thinsp]A resolution. *Nature* **389**, 251-260, (1997).

182 Richmond, T. J. & Davey, C. A. The structure of DNA in the nucleosome core. *Nature* **423**, 145-150, (2003).

183 Davey, C. A., Sargent, D. F., Luger, K., Maeder, A. W. & Richmond, T. J. Solvent Mediated Interactions in the Structure of the Nucleosome Core Particle at 1.9 Å Resolution†. *J. Mol. Biol.* **319**, 1097-1113, (2002).

184 Swinger, K. K. & Rice, P. A. IHF and HU: flexible architects of bent DNA. *Curr. Opin. Struct. Biol.* **14**, 28-35, (2004).

185 Nuñez, James K., Bai, L., Harrington, Lucas B., Hinder, Tracey L. & Doudna, Jennifer A. CRISPR Immunological Memory Requires a Host Factor for Specificity. *Mol. Cell* **62**, 824-833.

186 Wang, S., Cosstick, R., Gardner, J. F. & Gumport, R. I. The specific binding of Escherichia coli integration host factor involves both major and minor grooves of DNA. *Biochemistry* **34**, 13082-13090, (1995).

187 Hales, L. M., Gumport, R. I. & Gardner, J. F. Determining the DNA sequence elements required for binding integration host factor to two different target sites. *J. Bacteriol.* **176**, 2999-3006, (1994).

188 Delaglio, F. *et al.* NMRPipe: A multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR* **6**, 277-293.

189 Yang, S. W. & Nash, H. A. Comparison of protein binding to DNA in vivo and in vitro: defining an effective intracellular target. *The EMBO Journal* **14**, 6292-6300, (1995).

190 Seeman, N. C., Rosenberg, J. M., Suddath, F. L., Kim, J. J. P. & Rich, A. RNA double-helical fragments at atomic resolution: I. The crystal and molecular structure of sodium adenylyl-3′,5′-uridine hexahydrate. *J. Mol. Biol.* **104**, 109-144, (1976).

191 Lipfert, J. *et al.* Double-stranded RNA under force and torque: Similarities to and striking differences from double-stranded DNA. *Proceedings of the National Academy of Sciences* **111**, 15408-15413, (2014).

192 Aas, P. A. *et al.* Human and bacterial oxidative demethylases repair alkylation damage in both RNA and DNA. *Nature* **421**, 859-863, (2003).

193 Chen, F. *et al.* Adaptive response enzyme AlkB preferentially repairs 1-methylguanine and 3-methylthymine adducts in double-stranded DNA. *Chem. Res. Toxicol.*, (2016).

194 Kouchakdjian, M. *et al.* NMR structural studies of the ionizing radiation adduct 7-hydro-8-oxodeoxyguanosine (8-oxo-7H-dG) opposite deoxyadenosine in a DNA duplex. 8-Oxo-7H-dG(syn)•dA(anti) alignment at lesion site. *Biochemistry* **30**, 1403-1412, (1991).

195 Kuchino, Y. *et al.* Misreading of DNA templates containing 8-hydroxydeoxyguanosine at the modified base and at adjacent residues. *Nature* **327**, 77-79, (1987).

196 Dethoff, E. A., Petzold, K., Chugh, J., Casiano-Negroni, A. & Al-Hashimi, H. M. Visualizing transient low-populated structures of RNA. *Nature* **491**, 724-728, (2012).

197 Zimmer, D. P. & Crothers, D. M. NMR of enzymatically synthesized uniformly 13C15N-labeled DNA oligonucleotides. *Proc. Natl. Acad. Sci. U. S. A.* **92**, 3091-3095, (1995).

198 Macon, J. B. & Wolfenden, R. 1-Methyladenosine. Dimroth rearrangement and reversible reduction. *Biochemistry* **7**, 3453-3458, (1968).

199 Timofeev, E. N. *et al.* Oligodeoxynucleotides containing 2′-deoxy-1-methyladenosine and Dimroth rearrangement. *Helv. Chim. Acta* **90**, 928-937, (2007).

200 Pivovarov, V. B., Stepanian, S. G., Reva, I. D., Sheina, G. G. & Blagoi, Y. P. Infrared spectra and the structure of 1-methyladenine in an argon matrix and solutions. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* **51**, 843-853, (1995).

201 Longhini, A. P. *et al.* Chemo-enzymatic synthesis of site-specific isotopically labeled nucleotides for use in NMR resonance assignment, dynamics and structural characterizations. *Nucleic Acids Res.*, (2015).

202 Wenter, P., Reymond, L., Auweter, S. D., Allain, F. H. T. & Pitsch, S. Short, synthetic and selectively 13C-labeled RNA sequences for the NMR structure determination of protein–RNA complexes. *Nucleic Acids Res.* **34**, e79-e79, (2006).

203 Wunderlich, C. H. *et al.* Synthesis of (6-13C)pyrimidine nucleotides as spin-labels for RNA dynamics. *J. Am. Chem. Soc.* **134**, 7558-7569, (2012).

204 Williamson, M. P. Using chemical shift perturbation to characterise ligand binding. *Prog. Nucl. Magn. Reson. Spectrosc.* **73**, 1-16, (2013).

205 Wagenmakers, E.-J. & Farrell, S. AIC model selection using Akaike weights. *Psychonomic Bulletin & Review* **11**, 192-196.

206 Fürtig, B., Richter, C., Bermel, W. & Schwalbe, H. New NMR experiments for RNA nucleobase resonance assignment and chemical shift analysis of an RNA UUCG tetraloop. *J. Biomol. NMR* **28**, 69-79, (2004).

207 Fonville, J. M. *et al.* Chemical shifts in nucleic acids studied by density functional theory calculations and comparison with experiment. *Chemistry – A European Journal* **18**, 12372-12387, (2012).

208 Battle, D. J. & Doudna, J. A. Specificity of RNA–RNA helix recognition. *Proceedings of the National Academy of Sciences* **99**, 11676-11681, (2002).

209 Rich, A. The Era of RNA Awakening: Structural biology of RNA in the early years. *Q. Rev. Biophys.* **42**, 117-137, (2009).

210 Rypniewski, W., Adamiak, D. A., Milecki, J. & Adamiak, R. W. Noncanonical G(syn)–G(anti) base pairs stabilized by sulphate anions in two X-ray structures of the (GUGGUCUGAUGAGGCC) RNA duplex. *RNA* **14**, 1845-1851, (2008).

211 Crothers, D. M., Bloomfield, V. A. & Tinoco, I. *Nucleic acids: structures, properties, and functions*. (University science books, 2000).

212 Cate, J. H. *et al.* Crystal Structure of a Group I Ribozyme Domain: Principles of RNA Packing. *Science* **273**, 1678-1685, (1996).

213 Juneau, K., Podell, E., Harrington, D. J. & Cech, T. R. Structural basis of the enhanced stability of a mutant ribozyme domain and a detailed view of RNA–solvent Interactions. *Structure* **9**, 221-231, (2001).

214 Ye, J.-D. *et al.* Synthetic antibodies for specific recognition and crystallization of structured RNA. *Proceedings of the National Academy of Sciences* **105**, 82-87, (2008).

215 Macke, T. J. & Case, D. A. in *Molecular Modeling of Nucleic Acids* Vol. 682 *ACS Symposium Series* Ch. 24, 379-393 (American Chemical Society, 1997).

216 Brooks, B. R. *et al.* CHARMM: The biomolecular simulation program. *J. Comput. Chem.* **30**, 1545-1614, (2009).

217 Denning, E. J., Priyakumar, U. D., Nilsson, L. & MacKerell, A. D. Impact of 2′-hydroxyl sampling on the conformational properties of RNA: Update of the CHARMM all-atom additive force field for RNA. *J. Comput. Chem.* **32**, 1929-1943, (2011).

218 Lee, M. S., Feig, M., Salsbury, F. R. & Brooks, C. L. New analytic approximation to the standard molecular volume definition and its application to generalized Born calculations. *J. Comput. Chem.* **24**, 1348-1356, (2003).

219 Xu, Y., Vanommeslaeghe, K., Aleksandrov, A., MacKerell, A. D. & Nilsson, L. Additive CHARMM force field for naturally occurring modified ribonucleotides. *J. Comput. Chem.* **37**, 896-912, (2016).

220 Goldsmith, G., Rathinavelan, T. & Yathindra, N. Selective preference of parallel DNA triplexes Is due to the disruption of Hoogsteen hydrogen bonds caused by the severe nonisostericity between the G*GC and T*AT Triplets. *PLoS One* **11**, e0152102, (2016).

221 Paci, E. & Karplus, M. Forced unfolding of fibronectin type 3 modules: an analysis by biased molecular dynamics simulations. *J. Mol. Biol.* **288**, 441-459, (1999).

222 Lee, J., Dethoff, E. A. & Al-Hashimi, H. M. Invisible RNA state dynamically couples distant motifs. *Proceedings of the National Academy of Sciences* **111**, 9485-9490, (2014).

223 Yi Xue, B. G., Daniel Herschlag, Rick Russell, and Hashim M. Al-Hashimi. Visualizing Formation of an RNA Folding Intermediate through a Fast Highly Modular Secondary Structure Switch. *Nature Communication*, (2016).

224 Salmon, L. *et al.* Modulating RNA alignment using directional dynamic kinks: application in determining an atomic-resolution ensemble for a hairpin using NMR residual dipolar couplings. *J. Am. Chem. Soc.* **137**, 12954-12965, (2015).

225 Dunn, D. The occurence of 1-methyladenine in ribonucleic acid. *Biochim. Biophys. Acta* **46**, 198-200, (1961).

226 Saikia, M., Fu, Y., Pavon-Eternod, M., He, C. & Pan, T. Genome-wide analysis of N1-methyl-adenosine modification in human tRNAs. *RNA* **16**, 1317-1327, (2010).

227 Hagervall, T. G., Tuohy, T. M. F., Atkins, J. F. & Björk, G. R. Deficiency of 1-methylguanosine in tRNA from Salmonella typhimurium induces frameshifting by quadruplet translocation. *J. Mol. Biol.* **232**, 756-765, (1993).

228 Agris, P. F. in *Prog. Nucleic Acid Res. Mol. Biol.* Vol. Volume 53 (eds E. Cohn Waldo & Moldave Klvle) 79-129 (Academic Press, 1996).

229 Bjork, G. R., Wikstrom, P. M. & Bystrom, A. S. Prevention of translational frameshifting by the modified nucleoside 1-methylguanosine. *Science* **244**, 986-989, (1989).

230 Micura, R. *et al.* Methylation of the nucleobases in RNA oligonucleotides mediates duplex–hairpin conversion. *Nucleic Acids Res.* **29**, 3997-4005, (2001).

231 Tianming Yang, W. Q. A. C., Xiangrui Mai , Shui Zou and Esther C. Y. Woon. A methylation-switchable conformational probe for sensitive and selective detection of RNA demethylase activity. *Chem. Commun.*, (2016).

232 Dominissini, D. *et al.* The dynamic N1-methyladenosine methylome in eukaryotic messenger RNA. *Nature* **530**, 441-446, (2016).

233 Li, X. *et al.* Transcriptome-wide mapping reveals reversible and dynamic N1-methyladenosine methylome. *Nat. Chem. Biol.* **advance online publication**, (2016).

234 Yang, H. & Lam, S. L. Effect of 1-methyladenine on thermodynamic stabilities of double-helical DNA structures. *FEBS Lett.* **583**, 1548-1553, (2009).

235 Huang, Y., Weng, X. & Russu, I. M. Enhanced base-pair opening in the adenine tract of a RNA double helix. *Biochemistry* **50**, 1857-1863, (2011).

236 Meyer, K. D. & Jaffrey, S. R. The dynamic epitranscriptome: N6-methyladenosine and gene expression control. *Nat. Rev. Mol. Cell Biol.* **15**, 313-326, (2014).

237 Roost, C. *et al.* Structure and thermodynamics of N6-methyladenosine in RNA: a spring-loaded base modification. *J. Am. Chem. Soc.* **137**, 2107-2115, (2015).

238 Uesugi, S., Miki, H., Ikehara, M., Iwahashi, H. & Kyogoku, Y. A linear relationship between electronegativity of 2′-substituents and conformation of adenine nucleosides. *Tetrahedron Lett.* **20**, 4073-4076, (1979).

239 Haschemeyer, A. E. V. & Rich, A. Nucleoside conformations: an analysis of steric barriers to rotation about the glycosidic bond. *J. Mol. Biol.* **27**, 369-384, (1967).

240 Olson, W. K. Syn–Anti effects on the spatial configuration of polynucleotide chains. *Biopolymers* **12**, 1787-1814, (1973).

241 Sundaralingam, M. & Pan, B. Hydrogen and hydration of DNA and RNA oligonucleotides. *Biophys. Chem.* **95**, 273-282, (2002).

242 Williams, J. S. & Kunkel, T. A. Ribonucleotides in DNA: Origins, repair and consequences. *DNA Repair* **19**, 27-37, (2014).

243 DeRose, E. F., Perera, L., Murray, M. S., Kunkel, T. A. & London, R. E. Solution structure of the Dickerson DNA dodecamer containing a single ribonucleotide. *Biochemistry* **51**, 2407-2416, (2012).

244 Wu, W.-J. *et al.* How a low-fidelity DNA polymerase chooses non-Watson–Crick from Watson–Crick incorporation. *J. Am. Chem. Soc.* **136**, 4927-4937, (2014).

245 Topal, M. D. & Fresco, J. R. Base pairing and fidelity in codon-anticodon interaction. *Nature* **263**, 289-293, (1976).

246 Fernandez, I. S. *et al.* Unusual base pairing during the decoding of a stop codon by the ribosome. *Nature* **500**, 107-110, (2013).

247 Kimsey, I. & Al-Hashimi, H. M. Increasing occurrences and functional roles for high energy purine-pyrimidine base-pairs in nucleic acids. *Curr. Opin. Struct. Biol.* **24**, 72-80, (2014).

248 Helm, M., Giegé, R. & Florentz, C. A Watson−Crick Base-Pair-Disrupting Methyl Group (m1A9) Is Sufficient for Cloverleaf Folding of Human Mitochondrial tRNALys *Biochemistry* **38**, 13338-13346, (1999).

249 Kertesz, M. *et al.* Genome-wide measurement of RNA secondary structure in yeast. *Nature* **467**, 103-107, (2010).

250 Mortimer, S. A., Kidwell, M. A. & Doudna, J. A. Insights into RNA structure and function from genome-wide studies. *Nat. Rev. Genet.* **15**, 469-479, (2014).

251 Ferner, J. *et al.* Structures of HIV TAR RNA–Ligand Complexes Reveal Higher Binding Stoichiometries. *ChemBioChem* **10**, 1490-1494, (2009).

252 Peterson, R. D. & Feigon, J. Structural Change in Rev Responsive Element RNA of HIV-1 on Binding Rev Peptide. *J. Mol. Biol.* **264**, 863-877, (1996).

Biography

Huiqing Zhou was born on January 23, 1989 in Zoucheng, Shandong Province, China and raised in the Tai'an City in Shandong Province. She attended Nankai University for college where she worked with Prof. Pingchuan Sun as an undergraduate research assistant and earned a Bachelor of Science in Chemistry in June of 2011. Huiqing started her graduate study under the guidance of Prof. Hashim Al-Hashimi in the University of Michigan in July of 2011 and obtained her Master's degree of Science in Chemistry in August, 2013. She transferred to Duke University with her advisor in January of 2014 and continued her doctoral research at Duke major in Biochemistry with the China Scholarship Council Fellowship.

Publications:

1. Zhou, H.; Kimsey, I. J.; Nikolova, E. N.; Sathyamoorthy, B.; Grazioli, G.; McSally, J.; Bai, T.; Wunderlich, C. H.; Kreutz, C.; Andricioaei, I.; Al-Hashimi, H. M. (2016) m^1A and m^1G disrupt A-RNA structure through the intrinsic instability of Hoogsteen base pairs. ***Nat. Struct. Mol. Biol.*, advance online publication**, (2016).
2. Sathyamoorthy, B.; Zhou, H.; Xue, Y.; Al-Hashimi, H. M. (2016) Solution NMR Structure and Dynamics of DNA Duplexes Containing N1-Methyladenosine Provides Insights into Hoogsteen Base Pair Directed Dynamics. *Manuscript in Preparation*.
3. Zhou, H.; Hintze B.J.; Kimsey, I. J.; Sathyamoorthy, B.; Yang S.; Richardson, J.S.; Al-Hashimi, H. M. (2015) New insights into Hoogsteen base pairs in DNA duplexes from a structure-based survey. ***Nucleic Acids Research,*** 43, 3420-3433.
4. Goh, G.B.; Hulbert B.S.; Zhou, H.; Brooks. C.L. 3rd. (2014) Constant pH molecular dynamics of proteins in explicit solvent with proton tautomerism. ***Proteins***, 82, 1319-1331.
5. Nikolova, E.N.; Zhou, H.; Gottardo F.L.; Alvey, H. S.; Kimsey, I. J.; Al-Hashimi, H. M. (2013) A Historical Account of Hoogsteen Base-Pairs in Duplex DNA. ***Biopolymers***, 99, 955-968.