Acta Cryst D
BIOLOGICAL
CRYSTALLOGRAPHY

ISSN 1399-0047

Sequence-dependent structural changes in a self-assembling DNA oligonucleotide

Maithili Saoji[a] and Paul J. Paukstelis[a,b]*

[a]Department of Chemistry and Biochemistry, University of Maryland, College Park, MD 20742, USA, and [b]Center for Biomolecular Structure and Organisation, Maryland NanoCenter, University of Maryland, College Park, MD 20742, USA. *Correspondence e-mail: paukstel@umd.edu

Received 9 September 2015
Accepted 15 October 2015

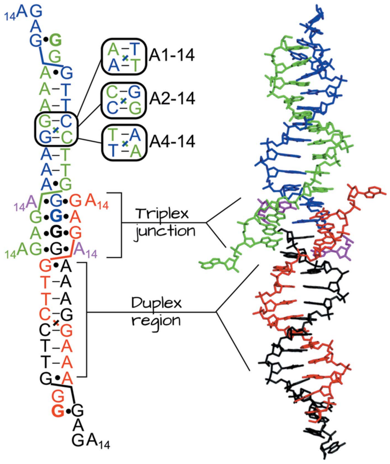
Edited by Z. Dauter, Argonne National Laboratory, USA

Keywords: DNA structure; self-assembly; sheared G–A base pair; base-triple interactions; sequence–structure relationship.

PDB references: continuously hydrogen-bonded 14-mer DNA, 5bxw; 5bzy; 5bz7; 5bz9

Supporting information: this article has supporting information at journals.iucr.org/d

DNA has proved to be a remarkable molecule for the construction of sophisticated two-dimensional and three-dimensional architectures because of its programmability and structural predictability provided by complementary Watson–Crick base pairing. DNA oligonucleotides can, however, exhibit a great deal of local structural diversity. DNA conformation is strongly linked to both environmental conditions and the nucleobase identities inherent in the oligonucleotide sequence, but the exact relationship between sequence and local structure is not completely understood. This study examines how a single-nucleotide addition to a class of self-assembling DNA 13-mers leads to a significantly different overall structure under identical crystallization conditions. The DNA 13-mers self-assemble in the presence of Mg^{2+} through a combination of Watson–Crick and noncanonical base-pairing interactions. The crystal structures described here show that all of the predicted Watson–Crick base pairs are present, with the major difference being a significant rearrangement of noncanonical base pairs. This includes the formation of a sheared A–G base pair, a junction of strands formed from base-triple interactions, and tertiary interactions that generate structural features similar to tandem sheared G–A base pairs. The adoption of this alternate noncanonical structure is dependent in part on the sequence in the Watson–Crick duplex region. These results provide important new insights into the sequence–structure relationship of short DNA oligonucleotides and demonstrate a unique interplay between Watson–Crick and noncanonical base pairs that is responsible for crystallization fate.



1. Introduction

The programmable interactions and structural predictability inherent in the complementary B-form duplex has made DNA one of the most widely used biomolecules for programmed self-assembly (Jones *et al.*, 2015). Oligonucleotides with complementary sequences can self-assemble in solution and have been successfully used in the rational design and construction of two-dimensional and three-dimensional DNA-based nanostructures (Winfree *et al.*, 1998; Seeman, 2003; Goodman *et al.*, 2005; Rothemund, 2006; Andersen *et al.*, 2009; Dietz *et al.*, 2009; Douglas *et al.*, 2009; Zheng *et al.*, 2009; Ke *et al.*, 2012; Tian *et al.*, 2014). Although complementary DNA oligonucleotides form predictable Watson–Crick duplexes, it has been known from some of the earliest structural studies that DNA can be both conformationally and structurally diverse (Drew *et al.*, 1988). Depending on the environmental conditions, B-form DNA can undergo conformational transitions to the A-form and the Z-form (Drew *et al.*, 1980, 1988; Dickerson & Drew, 1981; Rich & Zhang, 2003). Additionally, a variety of non-B-form DNA motifs have been characterized *in vivo*, including DNA cruciform, hairpin structures, triplexes

http://dx.doi.org/10.1107/S1399004715019598

and quadruplexes (Lee *et al.*, 1979; Lilley, 1981; Panayotatos & Wells, 1981; Sundquist & Klug, 1989; Bacolla & Wells, 2004; Burge *et al.*, 2006; Zhao *et al.*, 2010). One of the major areas of DNA structural biology over the course of several decades has been in understanding how non-Watson–Crick base pairs, or mismatches, could be accommodated in otherwise normal DNA helices or are responsible for forming alternate DNA structures (Peyret *et al.*, 1999; Tikhomirova *et al.*, 2006; Rossetti *et al.*, 2015).

G–A base pairs are one of the most well characterized non-Watson–Crick base pairings that can be readily integrated into the B-form duplex (Kan *et al.*, 1983; Patel *et al.*, 1984; Brown *et al.*, 1986, 1989; Hunter *et al.*, 1986; Privé *et al.*, 1987; Leonard *et al.*, 1990; Nikonowicz & Gorenstein, 1990; Li *et al.*, 1991*a*). Structural studies have revealed that these G–A base pairs can adopt up to four different base-pairing combinations depending on the local sequence and environment (Li *et al.*, 1991*b*). The two most prevalent types represented in the Nucleic Acid Database (Coimbatore Narayanan *et al.*, 2013) include the type I pair involving the Watson–Crick edges of the bases and the type IV sheared G–A pair involving the guanosine sugar edge and the adenosine Hoogsteen edge (Li *et al.*, 1991*a*; Greene *et al.*, 1994). However, the type and the stability of the G–A base pair formed is highly dependent on the local sequence (Cheng *et al.*, 1992). The type I pairing is favored for the d(AGAT)$_2$ sequence owing to an additional interstrand hydrogen bond between the N2 amino group of the paired G and O2 of the thymidine in the flanking A–T pair (Privé *et al.*, 1987). Sheared G–A base pairs are favored in d(YGAR)$_2$ sequences. In nearly all cases the sheared G–A pairs are found in tandem (GA/AG) and are thermodynamically quite stable within a canonical duplex due to the interstrand stacking between the sheared base pairs and the extensive intrastrand stacking between the sheared pairs and the flanking base pairs (Li *et al.*, 1991*b*).

We previously reported X-ray crystal structures of several DNA 13-mers that self-assemble into a continuously base-paired three-dimensional DNA lattice (Paukstelis *et al.*, 2004; Saoji *et al.*, 2015). Each 13-mer is hydrogen-bonded to one neighbor through a hexameric self-complementary duplex as well as being hydrogen-bonded to two other neighbors through parallel homopurine noncanonical base pairs. We analysed the sequence requirements in the duplex region of the DNA 13-mer by looking at all possible Watson–Crick base pairs in this region (Saoji *et al.*, 2015). We determined the X-ray crystal structures of 12 different sequence variants and also demonstrated that we could make heterogeneous crystals containing two different 13-mer sequences. To allow discrimination during gel electrophoresis of dissolved crystals, we added an additional adenosine to the 3′ end of one DNA in the mixture. These 14-mers crystallized under the same Mg^{2+} conditions and most displayed the same hexagonal crystal habit. However, four of the oligonucleotides crystallized with different crystal habits (Figs. 1*a*–1*e*).

Here, we describe the crystal structures of these four oligonucleotides and examine the role that the sequence plays in the adoption of the alternate crystal form. Our analysis suggests that the sequence in the 14-mer duplex region and the identity of the added nucleotide are necessary to promote this alternate structure. Remarkably, the added A14 residue from another strand makes tertiary contacts with the guanosine adjacent to a single sheared A–G pair, resulting in a conformation similar to tandem sheared G–A pairs. Together with a series of purine base triples, these interactions are responsible for the formation of the alternate crystal form. This study is a step forward in the understanding of the complex sequence–structure relationship of DNA oligonucleotides.

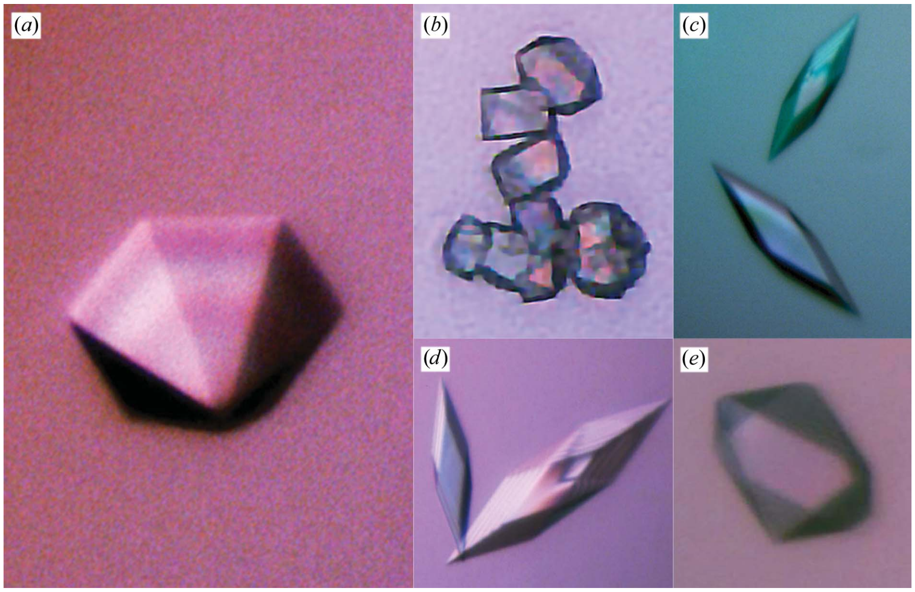


Figure 1
13-mer and 14-mer crystals. (*a*) The 13-mer DNAs crystallize with a hexagonal unipyrimidal crystal habit. Under identical crystallization conditions the four 14-mer DNAs A1-14, A2-14, A3-14 and A4-14 crystallize with the habits shown in (*b*), (*c*), (*d*) and (*e*), respectively. All of the 14-mers crystallized in the same space group with almost identical unit-cell parameters.

2. Materials and methods

2.1. DNA synthesis and purification

The four DNA 14-mers A1-14 [5′-d(GGAAAATTTGGAGA)], A2-14 [5′-d(GGAAACGTTGGAGA)], A3-14 [5′-d(GGAAAGCTTGGAGA)] and A4-14 [5′-d(GGAAATATTGGAGA)] were synthesized on a 1 µmol scale (Integrated DNA Technologies, Coralville, Iowa, USA) and were purified by 20% (19:1) polyacrylamide gel electrophoresis, electroeluted and ethanol-precipitated as described previously (Paukstelis *et al.*, 2004). The A3-14 (BrU9) oligonucleotide was synthesized using standard phosphoramidite chemistry on an Expedite 8909 DNA

Table 1
Sequence-dependent crystallization.

Designation	Sequence†	Crystal habit
A1-14-T	GGAAAATTTGGAGT	Hexagonal
A1-14-G	GGAAAATTTGGAGG	Hexagonal
A1-14-C	GGAAAATTTGGAGC	Hexagonal
A2-14-T	GGAAACGTTGGAGT	None
A2-14-G	GGAAACGTTGGAGG	Hexagonal
A2-14-C	GGAAACGTTGGAGC	Hexagonal
A3-14-T	GGAAAGCTTGGAGT	None
A3-14-G	GGAAAGCTTGGAGG	Hexagonal
A3-14-C	GGAAAGCTTGGAGC	Hexagonal
A4-14-T	GGAAATATTGGAGT	None
A4-14-G	GGAAATATTGGAGG	None
A4-14-C	GGAAATATTGGAGC	None
B6-14-A	GGACACGTGGGAGA	Hexagonal
B7-14-A	GGACAGCTGGGAGA	Hexagonal
E1-14-A	GGATAATTAGGAGA	Hexagonal
E3-14-A	GGATAGCTAGGAGA	Hexagonal
C9-14-A	GGAGAATTCGGAGA	Microcrystals
C11-14-A	GGAGAGCTCGGAGA	Clusters
B1-14-A	GGAATATATGGAGA	None
B2-14-A	GGAATCGATGGAGA	None
B3-14-A	GGAATGCATGGAGA	None
B4-14-A	GGAATTAATGGAGA	None
A5-14-A	GGAACATGTGGAGA	None
A6-14-A	GGAACCGGTGGAGA	None
A7-14-A	GGAACGCGTGGAGA	None
A8-14-A	GGAACTAGTGGAGA	Clusters
A9-14-A	GGAAGATCTGGAGA	Hexagonal
A10-14-A	GGAAGCGCTGGAGA	None
A11-14-A	GGAAGGCCTGGAGA	None
A12-14-A	GGAAGTACTGGAGA	None

† The position of sequence variability is indicated in red.

synthesizer (PerSeptive BioLabs) with reagents from Glen Research (Sterling, Virginia, USA). The purified DNA samples were dialyzed against deionized water and the concentration was adjusted to 260 µ*M*. The oligonucleotides used to examine the effect of sequence on crystallization (Table 1) were synthesized on a 100 nmol scale, dissolved in deionized water and used without purification.

2.2. Crystallization

The DNA oligonucleotides were crystallized by sitting-drop vapor diffusion. Prior to crystallization, DNA samples (260 µ*M*) were heated at 95° for 2 min and cooled to room temperature. Samples were mixed (1:1) with crystallization buffer (120 m*M* magnesium formate, 50 m*M* lithium chloride, 10% 2-methyl-2,4-pentanediol) in a 4 µl drop. The reservoir contained 400 µl crystallization buffer. The crystal plates were incubated at 22°C. Crystals appeared in 16–20 h and grew to average dimensions of 250 × 75 × 100 µm.

2.3. Data collection and structure determination

Crystals were harvested in nylon loops, washed sequentially in crystallization buffer containing 30 and 40% 2-methyl-2,4-pentanediol and flash-cooled in liquid nitrogen. Native data sets were collected on beamline 24-ID-E at the Advanced Photon Source, Argonne National Laboratory. Data were indexed and integrated using *XDS* (Kabsch, 2010) and scaled using *AIMLESS* (Evans & Murshudov, 2013). Several data sets had a relatively low completeness owing to the crystal orientation. However, each crystal type was highly isomorphous with respect to the unit-cell dimensions (the r.m.s.d. of the unit-cell dimensions was ≤0.1 Å), allowing the merging of observations from multiple crystals to improve the completeness. Phases were initially determined using an A3-14 (BrU9) derivative with data collected on beamline 24-ID-C at the Advanced Photon Source. Phases were determined by single-wavelength anomalous dispersion with the substructure sites identified by *HySS* in the *PHENIX* crystallographic software package (Grosse-Kunstleve & Adams, 2003; Afonine *et al.*, 2012). Models were built in *Coot* (Emsley *et al.*, 2010). The other three 14-mer structures were solved by molecular replacement using the completed A3-14 structure as a search model. Refinement was performed with *PHENIX* (Afonine *et al.*, 2012). Water molecules and ions were added manually during the refinement process. Following converged refinement in *PHENIX*, all of the structures were run through the *PDB_REDO* pipeline (Joosten *et al.*, 2014) with tenfold cross-

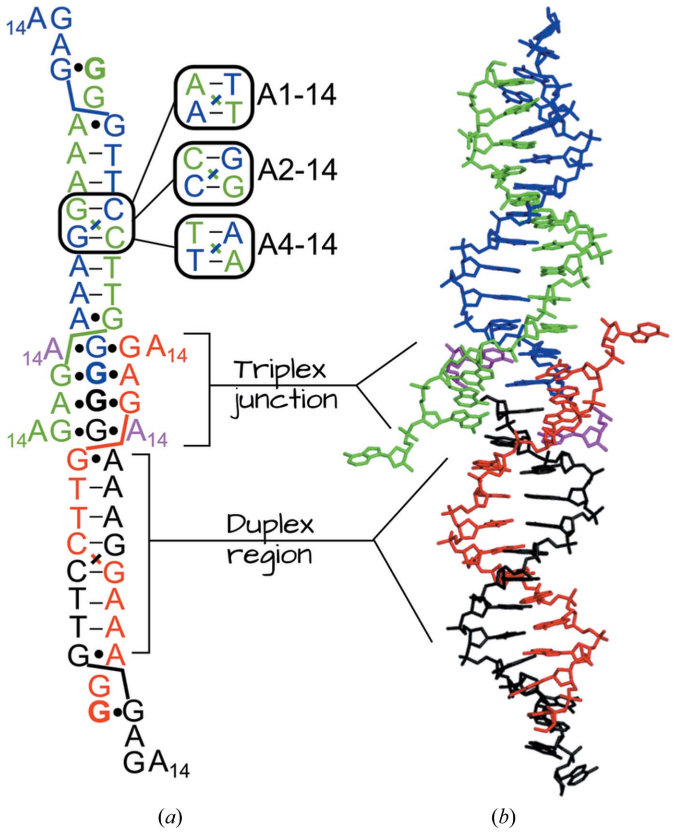


Figure 2
Overview of 14-mer crystal structures. (*a*) Secondary structure of 14-mer crystals. The A3-14 sequence is diagrammed and the sequence differences in the other oligonucleotides are shown. Each DNA 14-mer is hydrogen-bonded to five identical molecules related by crystallographic symmetry indicated in different colours. Interactions between DNA molecules lead to the formation of two distinct regions of base pairing. The duplex region is formed from residues A3–G10 of partner strands, and the triplex junction is formed by residues G1–G2 of one duplex (black/red), G11–G13 of the coaxially stacked duplex (green/blue) and the A14 residue of a neighboring duplex (magenta). The 5′ nucleotide of each strand is denoted in bold. (*b*) The overall three-dimensional arrangement of the 14-mers is shown in (*a*).

Table 2
Data-collection and refinement statistics.

Values in parentheses are for the highest resolution shell.

	A1-14	A2-14	A3-14	A4-14	A3-14-Br
Data collection					
Wavelength (Å)	0.97919	0.97919	0.97919	0.97919	0.91940
Detector	ADSC Quantum 315	ADSC Quantum 315	ADSC Quantum 315	ADSC Quantum 315	PILATUS 6M
Space group	$P3_121$	$P3_121$	$P3_121$	$P3_121$	$P3_121$
No. of crystals	4	4	3	2	1
Average unit-cell parameters					
$a = b$ (Å)	26.01	25.99	25.83	25.96	26.27
c (Å)	122.02	121.53	123.08	121.53	123.30
$\alpha = \beta$ (°)	90	90	90	90	90
γ (°)	120	120	120	120	120
Resolution	22.52–2.03 (2.09–2.03)	22.51–2.10 (2.17–2.10)	41.02–2.15 (2.23–2.15)	22.48–2.40 (2.53–2.40)	122.34–1.99 (2.09–1.99)
$\langle I/\sigma(I)\rangle$	13.9 (1.0)	16.3 (3.0)	14.7 (2.5)	8.9 (2.6)	14.8 (1.0)
$CC_{1/2}$	0.997 (0.883)	0.987 (0.978)	0.978 (0.959)	0.993 (0.954)	0.999 (0.866)
$R_{\rm p.i.m.}$	0.036 (0.45)	0.047 (0.12)	0.061 (0.15)	0.055 (0.13)	0.023 (0.38)
No. of reflections	3436 (268)	3035 (291)	2965 (281)	2131 (299)	3818 (507)
Completeness (%)	99.4 (97.3)	97.8 (97.4)	99.9 (99.9)	99.1 (98.7)	99.1 (96.0)
Multiplicity	10.3 (6.9)	10.6 (4.6)	6.8 (4.5)	3.8 (3.5)	5.0 (3.8)
Anomalous completeness (%)	N/A	N/A	N/A	N/A	95.2 (77.9)
Anomalous multiplicity	N/A	N/A	N/A	N/A	2.7 (1.8)
Refinement					
Resolution (Å)	22.52–2.03 (2.07–2.03)	23.51–2.10 (2.15–2.10)	41.02–2.15 (2.20–2.15)	22.48–2.40 (2.46–2.40)	
No. of reflections	3092 (206)	2716 (202)	2626 (178)	1898 (117)	
Average $R_{\rm free}$	0.294	0.313	0.265	0.312	
R factor	0.233 (0.502)	0.270 (0.454)	0.244 (0.421)	0.245 (0.523)	
$R_{\rm free}$	0.293 (0.538)	0.311 (0.416)	0.262 (0.606)	0.312 (0.641)	
No. of atoms					
DNA	293	293	293	293	
Ion	2	2	2	2	
Water	13	13	13	6	
R.m.s.d., bond lengths (Å)	0.007	0.006	0.008	0.005	
R.m.s.d., bond angles (°)	1.567	1.406	1.887	1.220	
PDB code	5bz7	5bz9	5bxw	5bzy	

validation applied owing to the small number of reflections in these data sets. Average $R_{\rm free}$ values for these ten different test sets are reported in Table 2. Coordinates and structure factors have been deposited in the Protein Data Bank (Berman *et al.*, 2000).

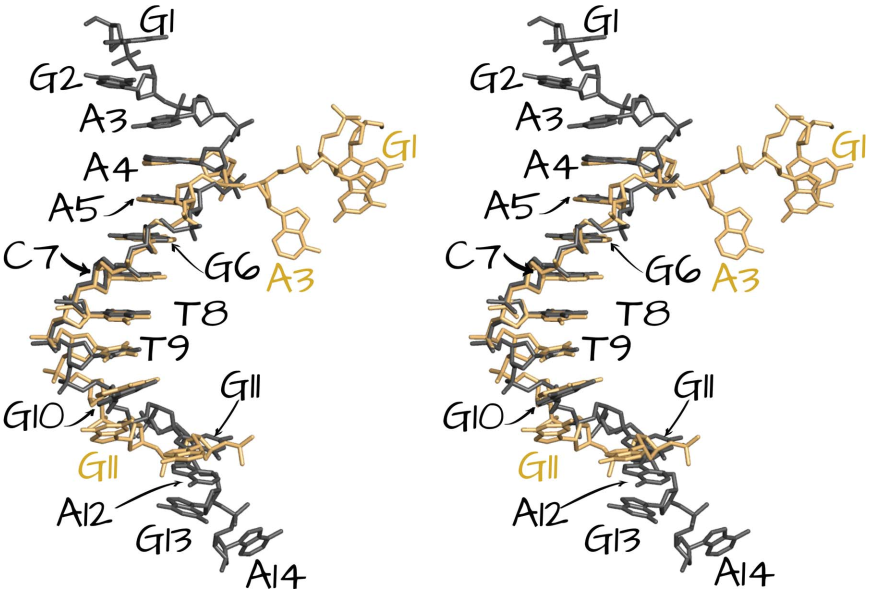


Figure 3
Structural comparison of A3-13 and A3-14 monomers. The X-ray crystal structures of A3-13 (gold) and A3-14 (black) superposed using residues A4–T9 of the duplex region. Label colors correspond to the model color. The largest deviations between the structures are in the 5′-most (G1–A3) and 3′-most (G10–A13) residues.

3. Results and discussion

3.1. Structural similarities and overview

We determined the X-ray crystal structures of four 14-mer DNA oligonucleotides differing by one base pair in the self-complementary duplex region (Fig. 2*a*). The structures were highly isomorphous to each other, with an average r.m.s.d. of 0.40 Å for all identical aligned atoms and 0.58 Å for the backbone atoms. All of the DNAs crystallized with one molecule in the asymmetric unit, with crystal symmetry generating interstrand hydrogen-bonding and base-stacking interactions. Each strand in the crystal forms hydrogen bonds to five other strands to form two distinct regions of nucleobase interactions (Figs. 2*a* and 2*b*). The B-form duplex region is formed from residues A3 through G10 of two strands, and the triplex junction is formed from G1 and G2 of one strand, G11–G13 of two different strands and A14 of another strand. End-to-end stacking of the triplex junction regions leads to columns of axially aligned stacks of helices that interact with adjacent stacks only through the A14 residues (Supplementary Fig. S1). For convenience, we will restrict our structural description to the A3-14 structure, noting that the only substantial differences in the other structures are the central base pairs of the duplex region.

3.2. B-form duplex region capped by sheared A–G pairs

The B-form duplex region is formed through base-pairing interactions between A3 and G10 of two DNA strands (Supplementary Table S1). The central six base pairs of the helix are composed of the self-complementary base pairs A4–T9, A5–T8 and G6–C5. These six nucleotides are structurally isomorphous to the duplex region in the parent A3-13 structure, with an r.m.s.d. of 0.69 Å (Fig. 3; Saoji *et al.*, 2015). The sugar-phosphate backbone of residue A4 shows the greatest variability between the two structures (r.m.s.d. of 3.40 Å for backbone atoms) and is the result of a significantly different conformation 5′ to the A4 nucleotide. In all of the 13-mer structures that we determined, residues G1–A3 are flipped out of the helical axis toward the major groove of the duplex, where they are positioned to make noncanonical interactions with G10–A12 of another strand. In the 14-mer structures, A3 remains stacked with A4 and is base-paired with G10 in a type IV sheared base pair. This results in a duplex region containing six self-complementary base pairs flanked on either end by A–G pairs.

3.3. Tertiary interactions fulfil a structural role to generate tandem G–A base pairs

The secondary-structural and tertiary-structural environment surrounding the sheared A3–G10 base pair establishes a local structure that is highly similar to the tandem sheared GA/AG steps that have previously been observed in B-form helices (Cheng *et al.*, 1992; Chou *et al.*, 1992, 2003; Chou, Cheng *et al.*, 1994; Chou, Zhu *et al.*, 1994), with the sheared A3–G10 base pair being structurally equivalent to the second base pair (Figs. 4*a* and 4*b*). This base pair is formed through the Hoogsteen edge of A3 (N6 and N7) and the sugar edge of G10 (N2 and N3) and displays the characteristic base-pair buckling (Fig. 4*a*). The nonplanarity of the base pair leads to the formation of a potential interstrand hydrogen bond between N6 of A3 and O2 of T9, although the geometry is not ideal (Supplementary Fig. S2). Like previous solution structures, interstrand and intrastrand stacking interactions play an important part in stabilizing the A3–G10 pairing. Despite the relatively large twist angle (60.3°) at the A3A4/T9G10 step, there are significant intrastrand stacking interactions between A3 and A4 (4.85 Å^2 overlap based on polygon projections using *X3DNA*; Lu & Olson, 2008). Intrastrand stacking interactions are even more pronounced for the partner strand, with T9–G10 stacking having an 8.10 Å^2 overlap. The overall 12.95 Å^2 overlap at this base-pair step is the single largest in the entire structure, suggesting that the capping A–G pairs provide significant stability to the duplex ends. Interstrand stacking interactions, which are one of the hallmarks of structures having tandem sheared G–A base pairs, are also present. G2 of one strand stacks with G10 of the partner strand. Overall, the structural environment around the A3–G10 pair is remarkably similar to previous solution structures (Fig. 4*a*), including the presence of several phosphate linkages in the B_{II} conformer (G2, A3 and G10), which is a hallmark of tandem sheared G–A structures (Chou *et al.*, 1992). The major difference is the lack of the first G–A pair. Interestingly, tertiary contacts between A14 from a different column of coaxially stacked helices and G2 maintain base-stacking interactions and lead to a similar overall structure.

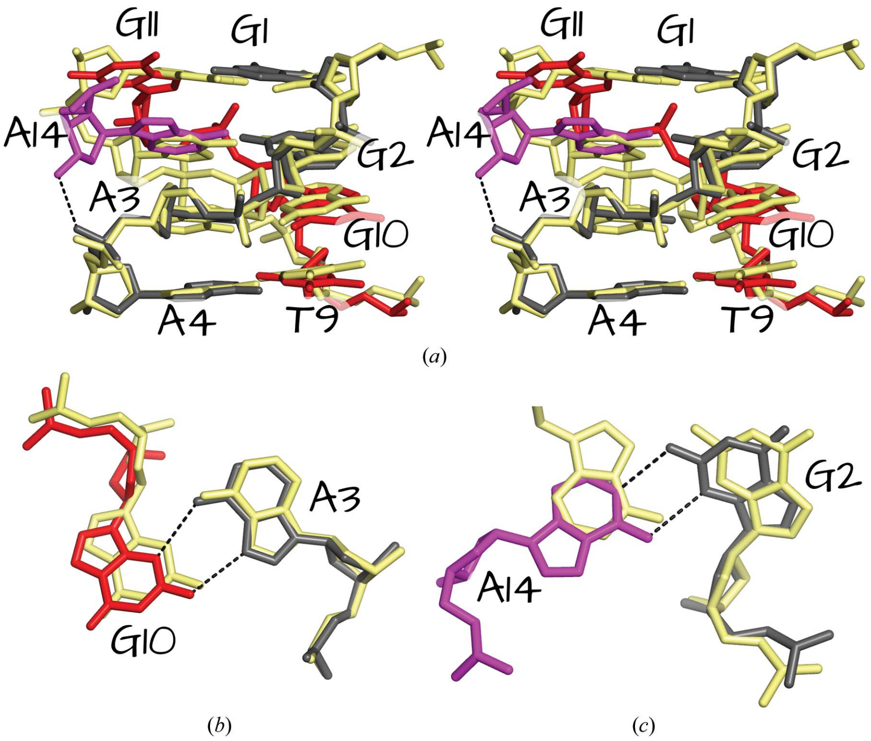


Figure 4
Comparison of GA/AG motifs. (*a*) The tandem sheared GA/AG base pairs flanked by Watson–Crick base pairs from a solution structure (yellow; PDB entry 175d; Chou, Cheng *et al.*, 1994) are shown superposed on G1–A4 from one strand (grey) and T9–G11 of the partner strand (red). A14 from a neighboring duplex is shown in magenta. A hydrogen bond between the A14 3′-OH and the A4 phosphate group is shown as a dashed line. (*b*) The A3–G10 base pair from the 14-mer structure is similar to the second base pair in the tandem sheared G–A solution structure. Hydrogen bonds are shown as dashed lines. (*c*) The tertiary contact between A14–G2 mimics the first sheared G–A base pair. The hydrogen bonds between the Watson–Crick face of A14 and the Hoogsteen face of G2 are shown as dotted lines.

The spacing between the G10 and G11 nucleobases allows A14 to stack between the A3 and G11 nucleotides from the partner strands, while forming a base pair with G2 from the minor-groove side (Figs. 4*a* and 4*c*). Unlike the tandem sheared G–A structures, this tertiary contact occurs between the Watson–Crick face of A14 and the sugar edge of G2, with N6 of A14 in an almost identical position to the first G–A in the solution structures (Fig. 4*c*). Along with these base-pairing and base-stacking interactions, A14 makes additional 3′-OH contacts with the A4 phosphate and is also involved in the base-capping interactions with the A3 sugar. This is an example of a tertiary-structure interaction providing a structural equivalence to a secondary-structure motif.

3.4. Triplex junction

The triplex junction connects two duplex segments into pseudo-infinite axially aligned helices. Two distinct triple interactions are present within the junction. Firstly, G2 is involved in a sugar-edge contact with A14 as described, but it also makes a single hydrogen bond through O6 to G13 N2 from the next coaxially stacked duplex on the major-groove side (Fig. 5*a*). This interaction is mediated in part by the hydrogen bonding of a solvent molecule that is within hydrogen-bonding distance of G2 N1, G13 N1, G1 O6 and the G11 phosphate of the partner duplex. This solvent molecule is present in all four crystal structures. Next, G1 is base-paired to G11 of the partner duplex through their Watson–Crick and Hoogsteen faces, respectively. A12 from a coaxially aligned duplex makes a single hydrogen bond to G1 also from the major-groove side (Fig. 5*b*). To our knowledge, this kind of all-purine triple interaction has not previously been observed in DNA. Additionally, these interactions provide an atypical example of end-to-end interactions in DNA. In this case, only the 5′-most residues (G1) are directly stacked, while the 3′-most residues (A14) are involved in the tertiary contacts that allow the parallel arrangement of an adjacent duplex. The purine base triples effectively 'stitch' the four strands together at the major grooves, without significant stacking interactions between the duplexes.

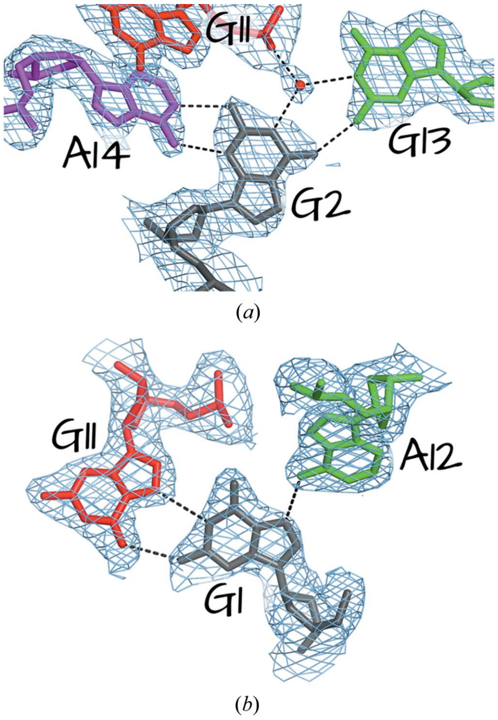


Figure 5
Purine base triples. The two purine base-triple interactions at the end of the duplex help to connect the stacked duplexes to form pseudo-infinite helices. (*a*) The first base triple is mediated by hydrogen bonding between the sugar edge of G2 and the Watson–Crick edge of A14, with an additional hydrogen bond between O6 of G2 and N2 of G13 from the adjacent stacked duplex. Additional hydrogen bonding is mediated by the interaction of the water molecule with G2 N1, G13 N1, G1 O6 and the G11 phosphate of the partner duplex. (*b*) The second base triple is formed between the Watson–Crick edge of G1 and the Hoogsteen edge of G11 from its duplex partner, along with a single hydrogen bond between G1 and A12 from a coaxially stacked duplex. In both (*a*) and (*b*) the hydrogen bonds are denoted by dashed lines, and a σ_A-weighted electron-density map ($2F_o - F_c$) contoured at 1.0σ is shown in blue.

3.5. Sequence requirements for the alternate crystal form

To understand why the addition of a single 3′ adenosine could result in a significantly different structure under identical crystallization conditions, we set out to understand the sequence requirements for the alternate crystal form in the context of the determined crystal structures. We screened 30 variants of the 14-mer oligonucleotides by altering the nucleobase identities at positions mediating key interactions in the structures. We probed these interactions in three different groups. In the first group, we screened all four of the oligonucleotides described here, but with different nucleobase identities at the added 14th residue. Out of the 12 DNA oligonucleotides screened, seven crystallized with a hexagonal unipyramidal habit and the other five failed to show any crystals (Table 1). Notably, three of these sequences were the variants of A4-14, which did not crystallize as the 13-mer in our original study (Saoji *et al.*, 2015). These results suggest that an adenosine at the 14th position is a requirement for the alternate crystal form. This is consistent with our structural observations of the G2–A14 tertiary base pair. Simple modeling with different nucleotide identities at position 14 indicate that pyrimidines would be unable to pair with G2 without significant backbone clashes, while a guanosine at this position would present incompatible hydrogen-bonding partners.

In the second group we examined six sequences with different self-complementary base pairs formed by positions 4 and 9 adjacent to the A3–G10 pairing (Table 1). We observed crystals in all cases, with sequences having Y4–R9 base pair exhibiting the hexagonal crystal habit, while the other two sequences that had a G4–C9 base pair forming microcrystals or irregular crystals. Although we have not yet been able to ascertain whether the G4–C9-containing crystals belong to one of the two crystal forms, these results indicate that the sequence rules observed for the formation of tandem sheared G–A base pairs (Cheng *et al.*, 1992) also apply to these crystals. Only sequences with a thymidine 5′ to G10 adopt the alternate crystal form described here, although it is possible that a cytosine at this location could promote the alternate crystal form. The strong stacking interactions between T9 and G10, along with geometric constraints, were previously suggested as reasons for the presence of the pyrimidine 5′ to guanosines in the tandem sheared pairs (Cheng *et al.*, 1992; Chou *et al.*, 1992, 2003). Our structural and crystal screening results support this analysis, but also indicate that a potential A3–T9 interstrand hydrogen bond (Supplementary Fig. S2) may help to stabilize these structures. Notably, the hydrogen-bond acceptor at O2 would be present with either pyrimidine at position 9.

Finally, we screened several sequence variants at the A5–T8 base pair. Based on the local sequence rules for the formation of the sheared G–A pair, we anticipated that this position should have little impact on the interactions necessary to form the alternate crystal form. Interestingly, ten of the 12 sequences screened in this group failed to crystallize, while the remaining two sequences formed only poor crystals (Table 1). This somewhat surprising result may be explained in several ways. Firstly, this may indicate that the significant stacking interactions between T8 and T9 (7.56 $\text{\AA}^2$) are required to adopt the conformation necessary to form the alternate crystal form, although this does not appear to be the case for solution structures containing tandem sheared G–A pairs. Secondly, our previous work established that the A5–T8 base pair is an important determinant for crystallization and crystallization speed in the context of 13-mers (Saoji *et al.*, 2015). It is possible that this base pair may have a more fundamental role in the formation of the short self-complementary duplex that is a common feature of the 13-mer and 14-mer structures.

Altogether, the sequence study and the structural observations strongly suggest that the presence of A14–G2, A4–T8 and A5–T9 are all critical for the formation and stabilization of the alternate crystal form.

4. Concluding remarks

Here, we have described how minor sequence variations in a DNA oligonucleotide can lead to drastically different crystal structures. Our work has demonstrated that the identity of an added 3′ nucleotide, as well as specific sequence rules in the Watson–Crick duplex region, are responsible for promoting the formation of distinct noncanonical base pairs. This study provides a new look at understanding the complex sequence–structure relationship of DNA oligonucleotides, particularly in an environment that is not exclusively Watson–Crick. Outside common DNA motifs such as G-quadruplexes and the i-motif, non-Watson–Crick interactions in functional nucleic acid structures have generally been reserved for RNA. The structures presented here along with other noncanonical DNA structures hint at the potential for DNA to adopt many 'alternate' structures that in some cases may have biological roles.

Acknowledgements

This work is based upon research conducted on the Northeastern Collaborative Access Team beamlines at the Advanced Photon Source, which are supported by award GM103403 from the National Center for Research Resources at the National Institutes of Health. Use of the Advanced Photon Source is supported by the US Department of Energy, Office of Basic Energy Sciences under Contract No. DE-AC02-06CH11357. This work was supported by an NSF CAREER award to PJP (DMR1149665).

References

Afonine, P. V., Grosse-Kunstleve, R. W., Echols, N., Headd, J. J., Moriarty, N. W., Mustyakimov, M., Terwilliger, T. C., Urzhumtsev, A., Zwart, P. H. & Adams, P. D. (2012). *Acta Cryst.* D**68**, 352–367.

Andersen, E. S., Dong, M., Nielsen, M. M., Jahn, K., Subramani, R., Mamdouh, W., Golas, M. M., Sander, B., Stark, H., Oliveira, C. L. P., Pedersen, J. S., Birkedal, V., Besenbacher, F., Gothelf, K. V. & Kjems, J. (2009). *Nature (London)*, **459**, 73–76.

Bacolla, A. & Wells, R. D. (2004). *J. Biol. Chem.* **279**, 47411–47414.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.

Brown, T., Hunter, W. N., Kneale, G. & Kennard, O. (1986). *Proc. Natl Acad. Sci. USA*, **83**, 2402–2406.

Brown, T., Leonard, G. A., Booth, E. D. & Chambers, J. (1989). *J. Mol. Biol.* **207**, 455–457.

Burge, S., Parkinson, G. N., Hazel, P., Todd, A. K. & Neidle, S. (2006). *Nucleic Acids Res.* **34**, 5402–5415.

Cheng, J.-W., Chou, S.-H. & Reid, B. R. (1992). *J. Mol. Biol.* **228**, 1037–1041.

Chou, S.-H., Cheng, J.-W., Fedoroff, O. & Reid, B. R. (1994). *J. Mol. Biol.* **241**, 467–479.

Chou, S.-H., Cheng, J.-W. & Reid, B. R. (1992). *J. Mol. Biol.* **228**, 138–155.

Chou, S.-H., Chin, K.-H. & Wang, A. H.-J. (2003). *Nucleic Acids Res.* **31**, 2461–2474.

Chou, S.-H., Zhu, L. & Reid, B. R. (1994). *J. Mol. Biol.* **244**, 259–268.

Coimbatore Narayanan, B., Westbrook, J., Ghosh, S., Petrov, A. I., Sweeney, B., Zirbel, C. L., Leontis, N. B. & Berman, H. M. (2013). *Nucleic Acids Res.* **42**, D114–D122.

Dickerson, R. E. & Drew, H. R. (1981). *J. Mol. Biol.* **149**, 761–786.

Dietz, H., Douglas, S. M. & Shih, W. M. (2009). *Science*, **325**, 725–730.

Douglas, S. M., Dietz, H., Liedl, T., Högberg, B., Graf, F. & Shih, W. M. (2009). *Nature (London)*, **459**, 414–418.

Drew, H. R., McCall, M. J. & Calladine, C. R. (1988). *Annu. Rev. Cell Biol.* **4**, 1–20.

Drew, H., Takano, T., Tanaka, S., Itakura, K. & Dickerson, R. E. (1980). *Nature (London)*, **286**, 567–573.

Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. (2010). *Acta Cryst.* D**66**, 486–501.

Evans, P. R. & Murshudov, G. N. (2013). *Acta Cryst.* D**69**, 1204–1214.

Goodman, R. P., Schaap, I. A. T., Tardin, C. F., Erben, C. M., Berry, R. M., Schmidt, C. F. & Turberfield, A. J. (2005). *Science*, **310**, 1661–1665.

Greene, K. L., Jones, R. L., Li, Y., Robinson, H., Wang, A. H.-J., Zon, G. & Wilson, W. D. (1994). *Biochemistry*, **33**, 1053–1062.

Grosse-Kunstleve, R. W. & Adams, P. D. (2003). *Acta Cryst.* D**59**, 1966–1973.

Hunter, W. N., Brown, T. & Kennard, O. (1986). *J. Biomol. Struct. Dyn.* **4**, 173–191.

Jones, M. R., Seeman, N. C. & Mirkin, C. A. (2015). *Science*, **347**, 1260901.

Joosten, R. P., Long, F., Murshudov, G. N. & Perrakis, A. (2014). *IUCrJ*, **1**, 213–220.

Kabsch, W. (2010). *Acta Cryst.* D**66**, 125–132.

Kan, L. S., Chandrasegaran, S., Pulford, S. M. & Miller, P. S. (1983). *Proc. Natl Acad. Sci. USA*, **80**, 4263–4265.

Ke, Y., Ong, L. L., Shih, W. M. & Yin, P. (2012). *Science*, **338**, 1177–1183.

Lee, J. S., Johnson, D. A. & Morgan, A. R. (1979). *Nucleic Acids Res.* **6**, 3073–3091.

Leonard, G. A., Booth, E. D. & Brown, T. (1990). *Nucleic Acids Res.* **18**, 5617–5623.

Li, Y., Zon, G. & Wilson, W. D. (1991*a*). *Proc. Natl Acad. Sci. USA*, **88**, 26–30.

Li, Y., Zon, G. & Wilson, W. D. (1991*b*). *Biochemistry*, **30**, 7566–7572.

Lilley, D. M. (1981). *Nucleic Acids Res.* **9**, 1271–1290.

Lu, X.-J. & Olson, W. K. (2008). *Nature Protoc.* **3**, 1213–1227.

Nikonowicz, E. P. & Gorenstein, D. G. (1990). *Biochemistry*, **29**, 8845–8858.

Panayotatos, N. & Wells, R. D. (1981). *Nature (London)*, **289**, 466–470.

Patel, D. J., Kozlowski, S. A., Ikuta, S. & Itakura, K. (1984). *Biochemistry*, **23**, 3207–3217.

Paukstelis, P. J., Nowakowski, J., Birktoft, J. J. & Seeman, N. C. (2004). *Chem. Biol.* **11**, 1119–1126.

Peyret, N., Seneviratne, P. A., Allawi, H. T. & SantaLucia, J. (1999). *Biochemistry*, **38**, 3468–3477.

Privé, G. G., Heinemann, U., Chandrasegaran, S., Kan, L. S., Kopka, M. L. & Dickerson, R. E. (1987). *Science*, **238**, 498–504.

Rich, A. & Zhang, S. (2003). *Nature Rev. Genet.* **4**, 566–572.

Rossetti, G., Dans, P. D., Gomez-Pinto, I., Ivani, I., Gonzalez, C. & Orozco, M. (2015). *Nucleic Acids Res.* **43**, 4309–4321.

Rothemund, P. W. K. (2006). *Nature (London)*, **440**, 297–302.

Saoji, M., Zhang, D. & Paukstelis, P. J. (2015). *Biopolymers*, **103**, 618–626.

Seeman, N. C. (2003). *Nature (London)*, **421**, 427–431.

Sundquist, W. I. & Klug, A. (1989). *Nature (London)*, **342**, 825–829.

Tian, C., Li, X., Liu, Z., Jiang, W., Wang, G. & Mao, C. (2014). *Angew. Chem. Int. Ed.* **53**, 8041–8044.

Tikhomirova, A., Beletskaya, I. V. & Chalikian, T. V. (2006). *Biochemistry*, **45**, 10563–10571.

Winfree, E., Liu, F., Wenzler, L. A. & Seeman, N. C. (1998). *Nature (London)*, **394**, 539–544.

Zhao, J., Bacolla, A., Wang, G. & Vasquez, K. M. (2010). *Cell. Mol. Life Sci.* **67**, 43–62.

Zheng, J., Birktoft, J. J., Chen, Y., Wang, T., Sha, R., Constantinou, P. E., Ginell, S. L., Mao, C. & Seeman, N. C. (2009). *Nature (London)*, **461**, 74–77.