BIOLOGICAL CRYSTALLOGRAPHY
ISSN 1399-0047

The novel double-folded structure of d(GCATGCATGC): a possible model for triplet-repeat sequences

Arunachalam Thirugnanasambandam, Selvam Karthik, Pradeep Kumar Mandal‡ and Namasivayam Gautham*

CAS in Crystallography and Biophysics, University of Madras, Guindy Campus, Chennai, Tamil Nadu 600 025, India.
*Correspondence e-mail: n_gautham@hotmail.com

Received 4 May 2015
Accepted 22 July 2015

Edited by Z. Dauter, Argonne National Laboratory, USA

‡ Current address: Institut Européen de Chimie et Biologie (IECB), 2 Rue Robert Escarpit, 33607 Pessac CEDEX, France.

Keywords: double-folded structure; triplet repeat; minor-groove tetrad; bi-loop; DNA.

PDB reference: novel double-fold structure of d(GCATGCATGC), 4zkk

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

The structure of the decadeoxyribonucleotide d(GCATGCATGC) is presented at a resolution of 1.8 Å. The decamer adopts a novel double-folded structure in which the direction of progression of the backbone changes at the two thymine residues. Intra-strand stacking interactions (including an interaction between the endocylic O atom of a ribose moiety and the adjacent purine base), hydrogen bonds and cobalt-ion interactions stabilize the double-folded structure of the single strand. Two such double-folded strands come together in the crystal to form a dimer. Inter-strand Watson–Crick hydrogen bonds form four base pairs. This portion of the decamer structure is similar to that observed in other previously reported oligonucleotide structures and has been dubbed a 'bi-loop'. Both the double-folded single-strand structure, as well as the dimeric bi-loop structure, serve as starting points to construct models for triplet-repeat DNA sequences, which have been implicated in many human diseases.

1. Introduction

Single-crystal and solution-state studies of DNA oligonucleotides have revealed a variety of unusual three-dimensional structures that differ from the well known Watson–Crick double helices. These include triplexes (Felsenfeld *et al.*, 1957), cruciform structures (Panayotatos & Wells, 1981), quadruplexes (Sen & Gilbert, 1988) and slipped structures (Pearson *et al.*, 1998). Among these, the structures of the heptamer d(GCATGCT) (Leonard *et al.*, 1995; Thorpe *et al.*, 2003) are of particular relevance to the present report. This molecule assumes a single-stranded folded, or loop, conformation. Two symmetry-related strands dimerize to form a DNA quadruplex structure, with two quartets of Watson–Crick-paired G–C bases stacked on top of each other. Solution and crystal structures of a series of octamer sequences, both linear and circularized, show the recurrence of this structural motif in a variety of situations (Salisbury *et al.*, 1997; Escaja *et al.*, 2000, 2003, 2007; Viladoms *et al.*, 2009, 2010).

As part of our continuing studies on unusual DNA structures, we have crystallized and solved the structure of the decadeoxyribonucleotide d(GCATGCATGC). This is a self-complementary sequence with purine–pyrimidine repeats. It was designed to study the effect of inverting the sequence d(CGTACGTACG), which has been shown to crystallize as left-handed Z-DNA in the presence of cobalt hexammine (Brennan & Sundaralingam, 1985; Brennan *et al.*, 1986). Previous reports indicated that alternating decamers which start with a purine base, such as the present one, assume the A-form of DNA (Ban & Sundaralingam, 1996). The present decamer, however, crystallizes in a double-folded 'S' shape.

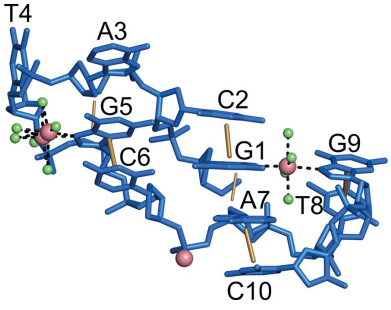

Table 1
Data-collection and processing statistics.

Values in parentheses are for the outer shell.

Diffraction source	BM14, ESRF
Wavelength (Å)	1.604
Temperature (K)	100
Detector	MAR Mosaic 225 mm CCD
Crystal-to-detector distance (mm)	103
Rotation range per image (°)	1
Total rotation range (°)	360
Space group	$P6_222$
a, b, c (Å)	34.64, 34.64, 89.60
α, β, γ (°)	90, 90, 120
Mosaicity (°)	0.31
Resolution range (Å)	24.92–1.80 (1.86–1.80)
Total No. of reflections	108161 (3466)
No. of unique reflections	3284 (240)
Completeness (%)	98.0 (83.7)
Multiplicity	32.9 (14.4)
Anomalous multiplicity	19.8 (8.0)
$\langle I/\sigma(I)\rangle$	36.6 (3.8)
$R_{\rm merge}$ (%)	5.4 (66.1)†/9.2 (71.3)‡
$R_{\rm meas}$ (%)	5.6 (70.7)†/9.4 (73.9)‡
$CC_{1/2}$ (%)	99.9 (93.9)
Overall B factor from Wilson plot (Å^2)	30.3

† Friedel pairs treated independently. ‡ Friedel pairs treated as equivalent reflections.

Table 2
Structure solution and refinement.

Values in parentheses are for the outer shell.

Resolution range (Å)	24.92–1.80 (1.98–1.80)
Completeness (%)	95.9
No. of reflections, working set	5358 (1079)
No. of reflections, test set	541 (124)
Final $R_{\rm work}$	0.229 (0.281)
Final $R_{\rm free}$	0.246 (0.294)
No. of cobalt ions	3
No. of solvent atoms	24
R.m.s. deviations	
Bonds (Å)	0.004
Angles (°)	0.557
Average B factors (Å^2)	
DNA atoms	42.2
Ions	38.2
Solvent atoms	40.3

The 'GCAT' tetrad appears to be a strong determinant of the structure, and a portion of the decamer is identical to the folded structure of the heptamers and octamers mentioned above. In addition, this portion of the decamer dimerizes with its symmetry-related neighbour through GC tetrads and forms a local quadruplex structure, as in the above structures. The presence of an extra three bases in the present sequence, however, introduces an extra fold into the structure. We have used this structure to build two models of 'triplet-repeat' DNA sequences, in which a triplet of bases, such as (CAG), is repeated many times. Such sequences have been implicated in several human genetic disorders (Ashley & Warren, 1995).

2. Materials and methods

2.1. Crystallography

PAGE-purified synthetic DNA d(GCATGCATGC) and other chemicals were purchased from Sigma–Aldrich Pvt. Ltd, Bangalore, India and were used without further purification. Crystals were grown by the hanging-drop vapour-diffusion method at 293 K from a drop consisting of 1 m*M* DNA, 50 m*M* sodium cacodylate trihydrate buffer pH 7.0, 10 m*M* cobalt chloride hexahydrate. The drop was equilibrated against 40% 2-methyl-2,4-pentanediol (MPD) in the well. Regular hexagonal bipyramidal crystals of dimensions 0.08 × 0.09 × 0.08 mm were observed after about five weeks. These were mounted on a cryoloop and flash-cooled in liquid nitrogen, with the mother liquor as the cryoprotectant, before being shipped for data collection.

X-ray diffraction data were collected at 100 K on the BM14 beamline at ESRF, Grenoble, France. In order to facilitate SAD phasing, the data were collected at a wavelength of 1.604 Å near the absorption peak of cobalt. Diffraction images obtained using a MAR Research CCD detector were processed using *iMosflm* (Battye *et al.*, 2011). The data-collection and processing statistics are given in Table 1.

The structure was solved using the single anomalous diffraction technique (SAD) with cobalt as the anomalous scatterer. Anomalous differences were measured to full resolution (1.80 Å) and were analysed using *phenix.xtriage* (Zwart *et al.*, 2005). From the peak data set, phases were obtained using *phenix.autosol* (Adams *et al.*, 2010) with an FOM of 0.54. Density modification improved this to 0.73. This phase information was used to build an initial model in *phenix.autobuild* (Adams *et al.*, 2010). The model was extended and refined manually using *Coot* (Emsley & Cowtan, 2004). Further refinement was carried out in *phenix.refine* (Afonine *et al.*, 2012). Several cycles of refinement were carried out, with a manual check of the electron density each time to add water molecules or to modify portions of the DNA model. In the final cycles, TLS refinement was carried out (Winn *et al.*, 2001). Examination of the structure showed that it could be considered as three domains: (i) the nucleotides G1, C2 and A3, (ii) the nucleotides G5, C6 and A7 and (iii) the nucleotides T8, G9 and C10. The nucleotide T4, which makes no nonbonded interactions with any other residue, either intra-strand or inter-strand, was not part of any domain. Friedel pairs were treated as independent reflections. The final model contains one decameric DNA chain, three cobalt ions (one with a partial occupancy of 50%) and 24 water molecules in the asymmetric unit. The final values of $R_{\rm work}$ and $R_{\rm free}$ were 22.9 and 24.6%, respectively. Refinement statistics are given in Table 2. The final model, along with the electron density, is shown in Supplementary Fig. S1.

Structural parameters such as sugar-puckering, torsion-angle, backbone, groove and helical parameters were calculated using *X3DNA* (Lu & Olson, 2008) and *Curves+* (Lavery *et al.*, 2009). Figures were generated using *PyMOL* (Schrödinger) and *UCSF Chimera* (Pettersen *et al.*, 2004).

2.2. Circular dichroism

Circular-dichroism (CD) spectra for the decamer were obtained using a Jasco J-815 CD spectrometer at the

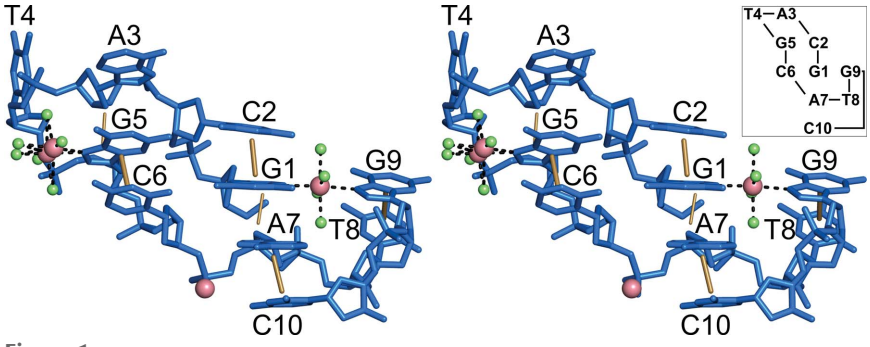

Figure 1
Wall-eyed stereoview of the double-folded structure of the decamer. Water molecules (except those coordinating the ion) are omitted for clarity. Stacking interactions are indicated as rods: thick rods indicate base–base stacking and thin rods indicate sugar–base interactions. Cobalt ions are shown as spheres. Metal-ion coordination is shown in dashed lines. Inset: schematic representation of the stacking and the connectivity.

Department of Biotechnology, IIT Madras, Chennai, India. The spectra were measured in the range 220–320 nm at 0.5 nm intervals at 293 K. The sample consisted of 50 μ*M* DNA in 50 m*M* sodium cacodylate buffer pH 7.0. This was titrated against $CoCl_2$ at concentrations ranging from 0 to 2000 μ*M*.

2.3. Model building of triplet-repeat DNA and molecular-dynamics simulations

The crystal structure was manipulated using *Coot* (Emsley & Cowtan, 2004) to build the models by cutting and joining appropriately, while preserving the core structural motif and all of its stabilizing interactions, both intra-strand and inter-strand. The folded structure of the single strand was first used to construct a model of a hexamer with the sequence AGCAGC. Multiple copies of the hexamer were arranged to build a model of an 18-mer with the sequence $(AGC)_6$. The structure of the crystallographic dimers, with the tetrads, was used to build a quadruplex model, starting with a four-stranded structure with each strand formed by the sequence AGC. Multiple copies of this structure were joined to form the final model to give a structure with four strands, each constituted of $(AGC)_2$ repeats. The models were optimized by energy minimization using *Amber*9 (Case *et al.*, 2006) with the ff99bsc0 force field (Wickstrom *et al.*, 2009). The nonbonded cutoff was set to 12 Å. Each model was placed in a box and filled with TIP3P water and sufficient Na^{2+} ions to neutralize the negative charges on the phosphate groups. 2500 cycles of steepest-descent energy minimization were followed by 2500 cycles of conjugate-gradient minimization. The initial and final energies of the models were 245 700.0 and −35 707.0 kcal mol^{-1}, respectively, for the 18-mer and 90 662.0 and −20 212.0 kcal mol^{-1}, respectively, for the quadruplex. Following the minimization, both models were subjected to molecular-dynamics simulations at a temperature of 300 K using the same program and force field, with all other conditions being the same. The total simulation time was 50 ns for each model. A similar set of molecular-dynamics calculations, *i.e.* minimization followed by 50 ns simulation, was also carried out on the single-stranded folded hexamer $d(AGC)_2$ and on the decamer as in the crystal but without the metal ions. The energies before and after minimization of these models are −14 411.0 and −24 718 kcal mol^{-1}, respectively, for the hexamer and −49 453 and −50 361 kcal mol^{-1}, respectively, for the decamer.

3. Results and discussion

3.1. The structure of d(GCATGCATGC)

The decamer d(GCATGCATGC) crystallizes as a folded chain (Fig. 1). The backbone reverses direction twice, making nearly 180° turns at T4 and T8. A portion of this structure has been observed in a number of sequences, and it constitutes a distinct DNA conformational motif termed the 'bi-loop' (Leonard *et al.*, 1995; Salisbury *et al.*, 1997; Escaja *et al.*, 2000, 2003, 2007; Thorpe *et al.*, 2003; Viladoms *et al.*, 2009, 2010). The backbone torsion angles of the present structure (Supplementary Table S1) are in approximate agreement with either of the well known A-type and B-type double-helical structures, except at the T4 and T8 residues and, to a lesser extent, at the preceding and succeeding residues. In T4 the deviation from a standard A-type or B-type double helix is apparently a result of the torsion angle α (O3′—P—O5′—C5′) adopting a *trans* conformation rather than a *gauche*$^-$ conformation and the torsion angle γ (O5′—C5′—C4′—C3′) adopting a *gauche*$^-$ conformation rather than a *gauche*$^+$ conformation. Likewise, in T8 α is *gauche*$^+$ and ε (C4′—C3′—O3′—P) of the preceding residue A7 is *gauche*$^-$ not *trans*. These changes, along with other less distinct alterations in the other angles, lead to drastic changes in the overall backbone conformation at the thymine residues. The residues G1, C2 and A3 are positioned in one direction of progression (forming one section of the double-folded structure), with a sharp turn at T4. The residues G5, C6 and A7 then proceed in the opposite, antiparallel direction, forming the second section. Another turn at T8 results in the terminal residues G9 and C10 proceeding in yet another direction. They form the third section.

3.2. Intramolecular interactions

There are three clearly identifiable sets of intramolecular interactions that stabilize the folded structure. These are stacking interactions, hydrogen bonds and ionic interactions. Intramolecular stacking interactions result in three sets of well stacked bases (Fig. 1 and Supplementary Table S2). The first set consists of the following four bases: G1 and C2 from the first section of the structure, A7 from the second section and C10 from the third section. G1 and C2 form a stacked pair, as do A7 and C10. The endocyclic sugar atom O4′ from A7 stacks on the imidazole ring of G1, similar to the sugar–base stacking interaction at the pyrimidine–purine base steps in left-handed Z-DNA (Wang *et al.*, 1979). Thus, the stacked sequence of bases is as follows: C2, G1, A7 and C10. The second set of

stacked bases consists of A3 from the first section, which makes a sugar–base stacking interaction with G5 from the second section. This base then stacks on C6. This structural motif, consisting of the above two stacks of bases (omitting C10, which is present only in the decamer), and specifically including the recruitment of a base from the antiparallel portion of the same strand by means of the sugar–base interaction, is a feature, perhaps a defining feature, of bi-loop structures (Supplementary Table S3 and Fig. 2). The third stacked set of bases consists of T8 and G9.

Hydrogen bonds constitute the second group of stabilizing intra-strand interactions. Three such interactions are observed. They connect the N2 atom of G1 to the O2 atom of C6 (3.13 Å), the N2 atom of G5 to the O2 atom of C2 (3.32 Å), and the terminal O5′ of G1 to the pendant phosphate atom O2 of T8 (2.61 Å).

Metal-ion coordination provides a third potent force for stabilizing the folded structure, particularly one part of it. There are three Co atoms in the asymmetric unit. All three of them are fully coordinated, with octahedral coordination geometry (Fig. 1). The first has the N7 atoms of G1 and G9 at two opposite vertices of the coordination octahedron, thus bringing together these two residues, which are on the opposite ends of the strand, and stabilizing the fold. The other vertices are occupied by well defined water molecules. The second Co atom partially occupies two sites, with occupancies of 51 and 49%, respectively (Supplementary Fig. S2). In both positions it coordinates to the N7 atom of G5. The other vertices of the octahedron are again occupied by water. The third ion occupies a special position along the twofold symmetry axis. Its coordination shell is completed by three water molecules and their symmetry-related copies.

Water-mediated interactions between atoms of the DNA strand constitute yet another stabilizing force. Supplementary Table S4 gives a list of these interactions. In addition, bases not paired with the symmetry-related neighbouring strand form hydrogen bonds to solvent water.

The three sets of intra-strand interactions mentioned above, along with hydrogen bonds to solvent water, could be sufficient to induce the double-folded structure in the single strand. The dimeric structure described below then may be a consequence of two such folded structures coming together. Support for this possibility comes from the report (Escaja *et al.*, 2000) of the solution structures of the cyclic octamers d<pCATTCATT> and d<pTGCTCGCT>. The NMR spectra of these sequences at low concentration indicate a monomeric structure with a 'dumbbell' shape. This report suggests that Watson–Crick base pairs form the 'handle' of the dumbbell. However, the NMR spectra are ambiguous on this point and the interactions could be as seen in the present decamer structure. Circular-dichroism spectra of the titration of the decamer with $CoCl_2$ (Fig. 3) indicate that at low DNA concentration there is a possible transition from a B-type helical structure to another folded conformation. However, it is possible that both structures are double-stranded. We note further that in the crystal structure the Co atom plays no role in the formation of the single-fold loop by the bases G1–A7. This structural motif may thus be stabilized by the stacking interactions and the hydrogen bonds alone, and it is therefore not possible to unequivocally interpret the changes in the CD spectra, which correlate with the presence of the cobalt ions, to indicate the formation of this structure. MD simulations of a single strand of the decamer, as seen in the crystal, but without Co^{2+} ions, showed that the double-fold structure persisted throughout the simulation time of 50 ns (Supplementary Fig. S3) and that most of the stacking interactions mentioned above were retained.

The structure of the sequence d(GCATGCT) crystallized in the presence of $CoCl_2$ (PDB entry 1qzl; C. J. Cardin, Y. Gan, J. H. Thorpe, C. S. M. Teixeira, B. C. Gale & M. I. A. Moraes,

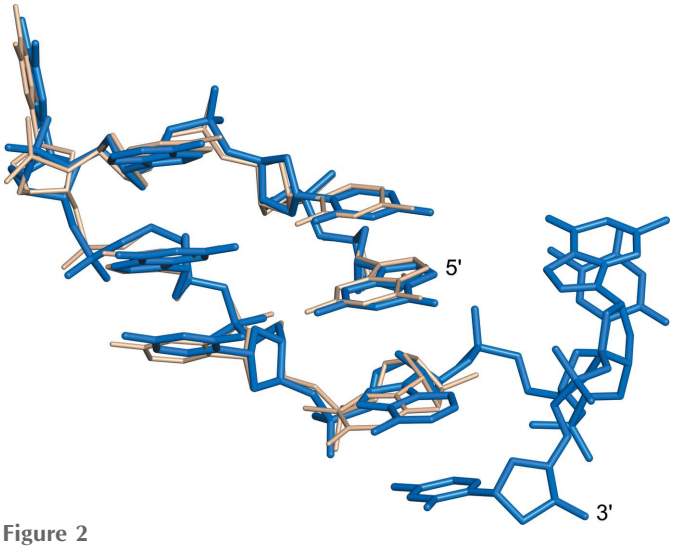

Figure 2
Least-squares superposition of a heptamer (PDB entry 1qyl; GCATGCT + V^{3+}) on the present structure. The backbone-atom r.m.s.d. is 0.80 Å. Details of the superposition of the present structure on other reported bi-loop structures are given in Supplementary Table S3.

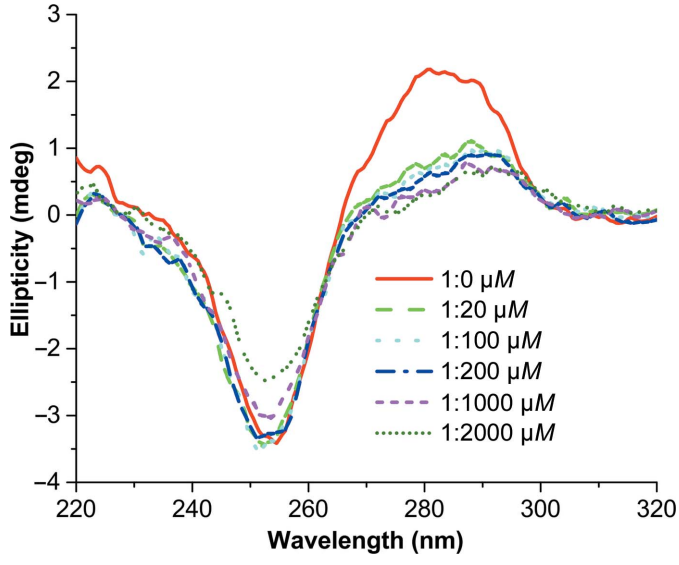

Figure 3
CD spectra of the decamer titrated with $CoCl_2$. The key indicates the ratio of DNA to cobalt chloride.

unpublished work) presents an interesting variation of the structural theme. The six-base single-fold motif is formed by the first three bases G, C and A from one molecule and the first three bases from another, symmetry-related, molecule. The rest of the heptamer superposes on the last three bases T8, G9 and C10 of the present decamer (Fig. 4). Thus, the heptamer assumes the same structure as sections 2 and 3 of the decamer, rather than sections 1 and 2. The cobalt ion, which coordinates G1 and G9 in the decamer, serves in the heptamer structure to bring together G1 and G5* of a symmetry-related pair of molecules.

3.3. Intermolecular interactions

The decamer forms a dimer with its symmetry-related neighbour (Fig. 5). Four Watson–Crick G–C base pairs are formed: G1–C2*, C2–G1*, G5–C6* and C6–G5* (where * indicates bases belonging to the molecule related by symmetry). The four bases G1, C2*, G5* and C6 form a hydrogen-bonded tetrad, called the 'minor-groove tetrad' (Escaja *et al.*, 2003), reminiscent of the motif seen in DNA quadruplex structures (Leonard *et al.*, 1995). The other four bases, C2, G1*, G5 and C6*, also form a similar tetrad. The two tetrads are stacked one on the other. The bases A7 and A3* (and A3 and A7*) continue the stack onto the next layer on either side of the central tetrad. The infinite stacked column then continues with C10 and C10** (from the next dimer) and C10* and C10*** (Fig. 6). The stacking is additionally stabilized by water-mediated interactions of the DNA atoms with the cobalt ions. In the crystal, adjacent columns are arranged approximately perpendicular to each other and form large voids. The water molecules that must be present in these spaces are invisible in the X-ray structure, presumably owing to their high mobility.

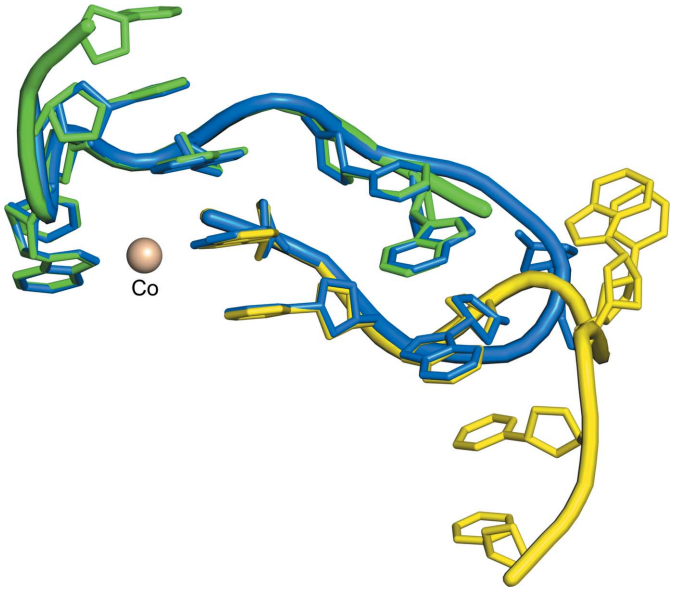

Figure 4
The heptamer d(GCATGCT) (PDB entry 1qzl; yellow) and its symmetry mate (green) superposed on the decamer (blue). The cobalt-ion position is the same in both structures.

The formation of a structural dimer by means of four Watson–Crick base pairs is a recurring feature in all of the octamer and heptamer sequences mentioned above. Structure-based alignment of the sequences (Fig. 7) shows the constant presence of a pyrimidine base, either T or C, in the centre. If we name this as position 0, then the bases at positions −3 and −2 are always complementary to each other, as are the bases at positions +2 and +1. This allows the formation of the Watson–Crick base pairs necessary to construct the dimer.

These packing interactions allow us to build a model of a quadruplex structure of triplet-repeat sequence DNA. This is described below, along with that built using the folded single-strand structure.

3.4. Models of triplet-repeat sequence DNA

DNA sequences in which a trinucleotide is repeated tens or hundreds of times have been implicated in several neurological disorders (Sutherland & Richards, 1995). Of the triplet-repeat sequences that have been shown to cause disease, CAG/CTG repeats have been linked to the largest number of

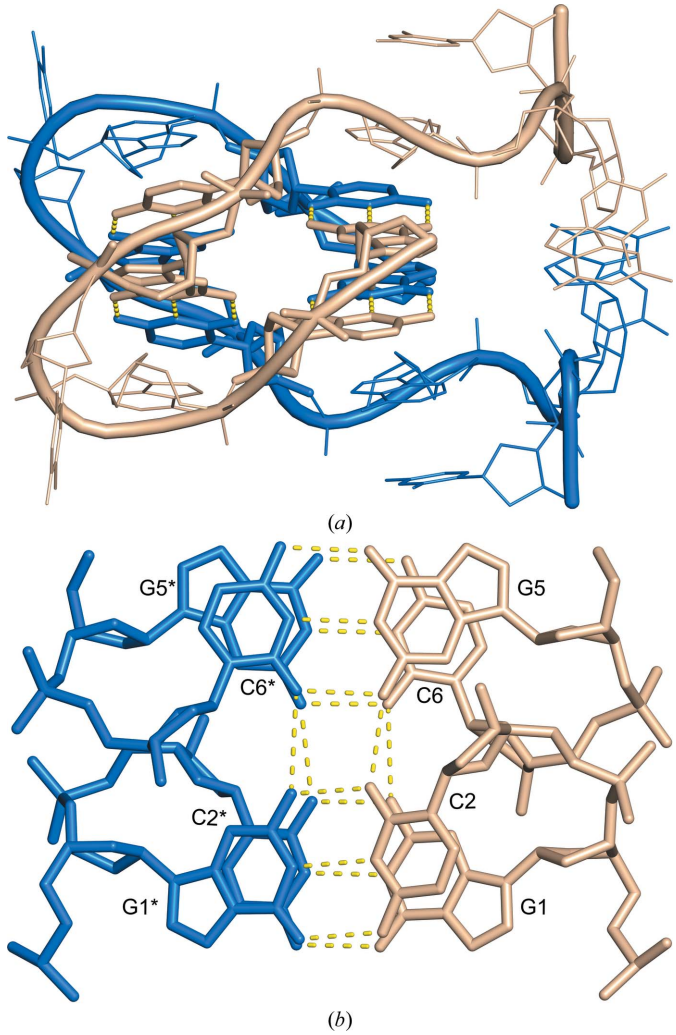

Figure 5
(*a*) The crystallographic dimer. (*b*) The Watson–Crick base-paired tetrads.

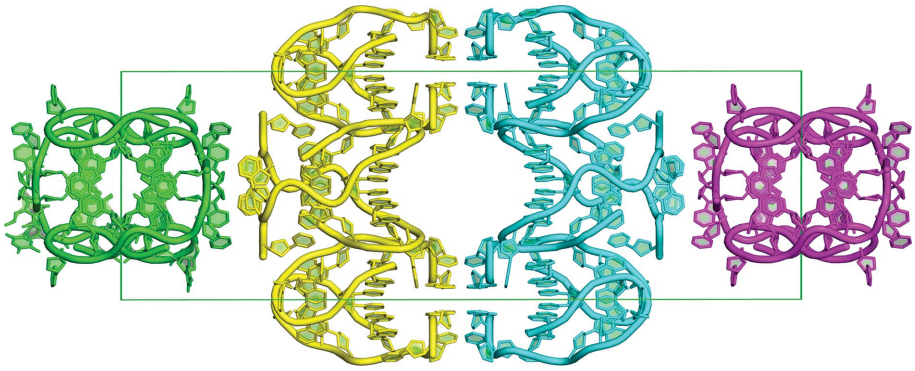

Figure 6
The unit-cell contents, showing the continuous base stacking extending in two approximately perpendicular directions through the crystal.

pathologies (Ashley & Warren, 1995). It may be noted that since the triplets are present as threefold to 20-fold repeats in normal conditions, and as 20-fold to 1000-fold repeats in pathological conditions (Fu *et al.*, 1991), the reading frame of the repeat sequences is ambiguous. Thus, $(CAG)_n$ may be also read as $(AGC)_{n-1}$ or $(GCA)_{n-1}$. If we include the complementary strand as well, then 18 of the 64 possible trinucleotides are represented by the five pathological repeat triplets identified in Ashley & Warren (1995). In the structure of the decamer d(GCATGCATGC), T4 is external to the loop made by G1, C2, A3, G5, C6 and A7. In building a model for triplet-repeat sequences, based on the structure of the decamer, the thymine base was excised and the α, β and γ backbone torsion angles of the G5 residue were manipulated to 'ligate' the A3 and G5 residues. These changes, together with excision of T8, G9 and C10, and manipulations of the backbone atoms at the A7 residue, results in the hexameric sequence 5′-d(AGCAGC)-3′ (Supplementary Fig. S4). Apart from the absence of T4, and minor changes in the backbone angles at two locations, the structure remains identical to that in the decamer crystal, preserving all of the intra-strand interactions, in particular the stacking and hydrogen-bond interactions. Multiple copies of this model hexamer were joined together to form a model of $(AGC)_6$, as seen in Fig. 8(*a*). Note that the hexamer loops are best placed in alternately opposite orientations. A similar model for the complementary sequence $(GCT)_n$ may also be constructed. As explained above, these are also models for $(CAG)_n$ repeats.

Nucleotides	:		$N_1^{\#}$	N_1	N	Y	N_2	$N_2^{\#}$	N			
Position	:		−3	−2	−1	0	1	2	3			
Decamer	:		G	C	A	T	G	C	A	T	G	C
1qzl	:		G*	C*	A*		G	C	A	T	G	C T
1qyl	:		G	C	A	T	G	C	T			
284d	:		A	T	T	C	A	T	T	C		
2k8z	:	T	C	G	T	T	G	C	T			
2k90	:	T	G	C	T	T	C	G	T			
1eu2	:	T	G	C	T	C	G	C	T			
1eu6	:	C	A	T	T	C	A	T	T			
1n96	:	C	G	C	T	C	A	T	T			
2hk4	:	C	C	G	T	C	C	G	T			
2k97	:	C	G	C	T	C	C	G	T			

Figure 7
Structure-based sequence alignment. The invariant central pyrimidine is at position 0. The bases at positions −2 and −3 (N_1 and $N_1^{\#}$) are complementary to each other, as are the bases (N_2 and $N_2^{\#}$) at positions 1 and 2. The bases at positions −1 and 3 form sugar–base stacking interactions with those at positions 1 and −3, respectively. * indicates a crystal symmetry-related molecule.

Molecular-dynamics simulations show that in both the single-stranded hexamer $d(AGC)_2$ model and the single-stranded 18-mer $d(AGC)_6$ model the conformation begins to fray within a simulation time of 50 ns, and the intra-strand interactions are not all retained (Supplementary Fig. S5). However, the fold between one CAG triplet and the next remains more or less intact. A possible deduction from these results is that the removal of thymine does not disrupt the essential folded nature of the structure and that the model that we have built for the triplet-repeat DNA is plausible. To summarize, the following facts may be tentatively adduced in support of the model. (i) The folded motif recurs in several crystal structures with different sequences. (ii) These sequences are similar to the triplet-repeat sequences, except for the presence of an extra 'central' pyrimidine nucleotide. (iii) In the crystal structures this nucleotide plays no role in the folded structure, being placed external to it, and makes no nonbonded contacts, either intra-strand nor inter-strand. (iv) The CD spectra indicate a change to, possibly, a folded structure in the presence of cobalt. (v) The molecular-dynamics simulations indicate that the folded motif, with the extra pyrimidine excised, is reasonably stable, although the intra-strand interactions are not all preserved.

We used similar cut-and-join manipulations of the Watson–Crick base-paired section of the crystallographic dimer, without altering any of the interactions in this region, to create a four-stranded structural model of triplet-repeat sequences (Fig. 8*b*). Molecular-dynamics simulations of this four-stranded model show that it is stable even after 50 ns and that the tetraplex retains most of the Watson–Crick base-pairing and base-stacking interactions (Supplementary Fig. S6). The support for this model is thus somewhat stronger than for the previous model. All five of the points mentioned above are also applicable here, with the fifth one being more supportive.

A number of crystal and solution NMR structures of triplet-repeat DNA oligomers have been reported in the literature. The solution structures of $d(GGA)_4$ (Matsugami *et al.*, 2001) and $d(GGA)_8$ (Matsugami *et al.*, 2003) show that these form compact structures that resemble the quadruplex 'G-plate' motifs (Arnott *et al.*, 1974). The solution structure of $d[(GGA)_2T]$ (Kettani *et al.*, 1999) is a dimer formed by two folded strands, reminiscent of the crystal structure of the

decamer reported in this paper. The hexamer $d(CCG)_2$ forms a Watson–Crick base-paired dimer in solution (Zheng *et al.*, 1996). However, four cytosine residues are placed outside the dimer and the structure is a distorted double helix. The structure of $d(GAC)_3$ in solution (Zheng *et al.*, 1996) is an irregular, parallel double helix with A–A, C–C and G–G base pairs. The crystal structure of $d[T(CCG)_3A]$ has recently been reported (Chen *et al.*, 2014). This is again a dimer formed by two folded strands, in effect constituting a four-stranded structure.

The experimentally determined structures of several RNA oligomers have also been suggested as models for triplet-repeat sequences (Supplementary Table S5). Almost all of them form A-type double helices with mismatched base pairs: U–U, A–A, U–G, C–A, G–G (Hoogsteen) and U–A (Hoogsteen). For example, the structure of the decamer r(GGCAGCAGCC)$_2$ (Kiliszek *et al.*, 2010), with two CAG repeats, has noncanonical A–A base pairs that fit in well and do not disrupt the helix to any significant extent. Likewise, the structure $r[UUGGGC(CAG)_3GUCC]_2$ (Yildirim *et al.*, 2013) is an A-type duplex with overhanging bases at either end and with A–A base pairs (internal loops).

Triplet repeats (or rather their expansion) may lead to disease in two ways. They may result in the production of abnormal proteins (Cha, 2000). Alternatively, they may disrupt the transcription and (after transcription as RNA) translation mechanisms (Cha, 2000;Mirkin, 2006; Bidichandani *et al.*, 1998; Ohshima *et al.*, 1998). It is in the latter pathway that the above variant structures for DNA, including our models, are relevant.

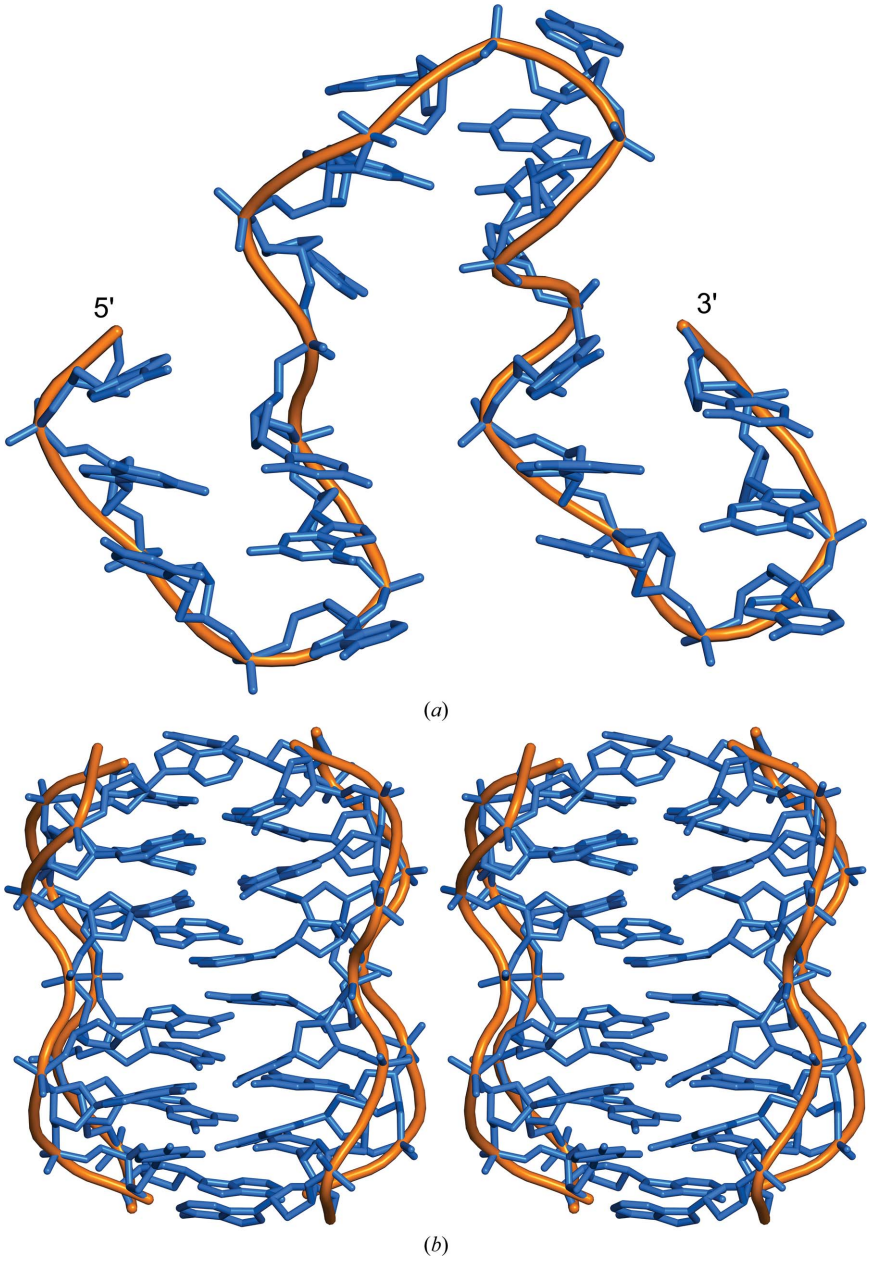

Figure 8
Models of AGC (or CAG) repeats. (*a*) Energy-minimized model of the 18-mer $(AGC)_6$. (*b*) Energy-minimized model of the tetraplex $d(AGCAGC)_4$.

4. Conclusion

This present structure increases the variety in the conformations known to be adopted by DNA sequences. The sequence is self-complementary, and could have been expected to form a fully Watson–Crick base-paired duplex, such as an A-type, B-type or Z-type helix. However, it adopts this unusual double-fold structure. Taken together with other similar structures adopted by similar sequences, this may point to a biological role for this conformation, for example as the consequence of genetic anomalies such as triplet-repeat expansion, which then lead to disease.

Acknowledgements

We gratefully acknowledge the following agencies: BM14 at ESRF, Grenoble, France and DBT, Government of India for access to the synchrotron facility and travel assistance; Department of Biotechnology, Indian Institute of Technology, Madras for access to the CD spectrometer; DST, Government of India for funds under the PURSE scheme; DST, Government of India under the FIST scheme; and UGC, Government of India under the CAS scheme for infrastructural funding. AT thanks UGC for a fellowship under the BSR scheme and KS thanks the DST for a fellowship under the INSPIRE scheme.

References

Adams, P. D. *et al.* (2010). *Acta Cryst.* D**66**, 213–221.

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.

Arnott, S., Chandrasekaran, R. & Marttila, C. M. (1974). *Biochem. J.* **141**, 537–543.

Ashley, C. T. Jr & Warren, S. T. (1995). *Annu. Rev. Genet.* **29**, 703–728.

Ban, C. & Sundaralingam, M. (1996). *Biophys. J.* **71**, 1222–1227.

Battye, T. G. G., Kontogiannis, L., Johnson, O., Powell, H. R. & Leslie, A. G. W. (2011). *Acta Cryst.* D**67**, 271–281.

Bidichandani, S. I., Ashizawa, T. & Patel, P. I. (1998). *Am. J. Hum. Genet.* **62**, 111–121.

Brennan, R. G. & Sundaralingam, M. (1985). *J. Mol. Biol.* **181**, 561–563.

Brennan, R. G., Westhof, E. & Sundaralingam, M. (1986). *J. Biomol. Struct. Dyn.* **3**, 649–665.

Case, D. A. *et al.* (2006). *Amber*9. University of California, San Francisco, USA.

Cha, J.-H. J. (2000). *Trends Neurosci.* **23**, 387–392.

Chen, Y.-W., Jhan, C.-R., Neidle, S. & Hou, M.-H. (2014). *Angew. Chem. Int. Ed.* **53**, 10682–10686.

Emsley, P. & Cowtan, K. (2004). *Acta Cryst.* D**60**, 2126–2132.

Escaja, N., Gelpí, J. L., Orozco, M., Rico, M., Pedroso, E. & González, C. (2003). *J. Am. Chem. Soc.* **125**, 5654–5662.

Escaja, N., Gómez-Pinto, I., Pedroso, E. & González, C. (2007). *J. Am. Chem. Soc.* **129**, 2004–2014.

Escaja, N., Pedroso, E., Rico, M. & González, C. (2000). *J. Am. Chem. Soc.* **122**, 12732–12742.

Felsenfeld, G., Davies, D. R. & Rich, A. (1957). *J. Am. Chem. Soc.* **79**, 2023–2024.

Fu, Y.-H., Kuhl, D. P. A., Pizzuti, A., Pieretti, M., Sutcliffe, J. S., Richards, S., Verkert, A. J. M. H., Holden, J. J. A., Fenwick, R. G. Jr, Warren, S. T., Oostra, B. A., Nelson, D. L. & Caskey, C. T. (1991). *Cell*, **67**, 1047–1058.

Kettani, A., Bouaziz, S., Skripkin, E., Majumdar, A., Wang, W., Jones, R. A. & Patel, D. J. (1999). *Structure*, **7**, 803–815.

Kiliszek, A., Kierzek, R., Krzyzosiak, W. J. & Rypniewski, W. (2010). *Nucleic Acids Res.* **38**, 8370–8376.

Lavery, R., Moakher, M., Maddocks, J. H., Petkeviciute, D. & Zakrzewska, K. (2009). *Nucleic Acids Res.* **37**, 5917–5929.

Leonard, G. A., Zhang, S., Peterson, M. R., Harrop, S. J., Helliwell, J. R., Cruse, W. B. T., Langlois d'Estaintot, B., Kennard, O., Brown, T. & Hunter, W. N. (1995). *Structure*, **3**, 335–340.

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

Matsugami, A., Okuizumi, T., Uesugi, S. & Katahira, M. (2003). *J. Biol. Chem.* **278**, 28147–28153.

Matsugami, A., Ouhashi, K., Kanagawa, M., Liu, H., Kanagawa, S., Uesugi, S. & Katahira, M. (2001). *J. Mol. Biol.* **313**, 255–269.

Mirkin, S. M. (2006). *Curr. Opin. Struct. Biol.* **16**, 351–358.

Ohshima, K., Montermini, L., Wells, R. D. & Pandolfo, M. (1998). *J. Biol. Chem.* **273**, 14588–14595.

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

Pearson, C. E., Wang, Y.-H., Griffith, J. D. & Sinden, R. R. (1998). *Nucleic Acids Res.* **26**, 816–823.

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C. & Ferrin, T. E. (2004). *J. Comput. Chem.* **25**, 1605–1612.

Salisbury, S. A., Wilson, S. E., Powell, H. R., Kennard, O., Lubini, P., Sheldrick, G. M., Escaja, N., Alazzouzi, E., Grandas, A. & Pedroso, E. (1997). *Proc. Natl Acad. Sci. USA*, **94**, 5515–5518.

Sen, D. & Gilbert, W. (1988). *Nature (London)*, **334**, 364–366.

Sutherland, G. R. & Richards, R. I. (1995). *Proc. Natl Acad. Sci. USA*, **92**, 3636–3641.

Thorpe, J. H., Teixeira, S. C. M., Gale, B. C. & Cardin, C. J. (2003). *Nucleic Acids Res.* **31**, 844–849.

Viladoms, J., Escaja, N., Frieden, M., Gómez-Pinto, I., Pedroso, E. & González, C. (2009). *Nucleic Acids Res.* **37**, 3264–3275.

Viladoms, J., Escaja, N., Pedroso, E. & González, C. (2010). *Bioorg. Med. Chem.* **18**, 4067–4073.

Wang, A. H.-J., Quigley, G. J., Kolpak, F. J., Crawford, J. L., van Boom, J. H., van der Marel, G. & Rich, A. (1979). *Nature (London)*, **282**, 680–686.

Wickstrom, L., Okur, A. & Simmerling, C. (2009). *Biophys. J.* **97**, 853–856.

Winn, M. D., Isupov, M. N. & Murshudov, G. N. (2001). *Acta Cryst.* D**57**, 122–133.

Yildirim, I., Park, H., Disney, M. D. & Schatz, G. C. (2013). *J. Am. Chem. Soc.* **135**, 3528–3538.

Zheng, M., Huang, X., Smith, G. K., Yang, X. & Gao, X. (1996). *J. Mol. Biol.* **264**, 323–336.

Zwart, P., Grosse-Kunstleve, R. & Adams, P. (2005). *CCP4 Newsl. Protein Crystallogr.* **43**, 27–35.