Acta Cryst. (1996). D52, 766–775

Formation of (C·G)*G Triplets in a B-DNA Duplex with Overhanging Bases

DOMINIQUE VLIEGHE,[a] LUC VAN MEERVELT,[a]* ALAIN DAUTANT,[b] BERNARD GALLOIS,[b] GILLES PRÉCIGOUX[b] AND OLGA KENNARD[c]

[a]*Department of Chemistry, Katholieke Universiteit Leuven, Celestijnenlaan 200F, B-3001 Heverlee, Belgium,* [b]*Unité de Biophysique Structurale, EP CNRS, Université de Bordeaux II, 33405 Talence, France, and* [c]*Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, England. E-mail: luc.vanmeervelt@chem.kuleuven.ac.be*

(Received 18 January 1996; accepted 5 March 1996)

Abstract

Crystallization of a DNA double helix with overhanging bases at the 5′-ends of both strands, results in the formation of two crystallographically independent (C·G)*G triplets. In a previous report [Van Meervelt, Vlieghe, Dautant, Gallois, Précigoux & Kennard (1995). *Nature (London)*, **374**, 742–744] the unique molecular packing of the duplex and the Hoogsteen hydrogen-bond pattern and parallel backbone orientation of the guanine-containing strands in the triplets was described. The fine structural details and hydration of the d(GCGAATTCG) crystal structure refined to 2.05 Å ($R = 0.168$, 86 water molecules, two Mg^{2+} cations) are now presented. Helical parameters, stacking effects, the geometry at the duplex–triplex junction, and the hydration of the minor groove are discussed and compared with related theoretical and crystal structures.

1. Introduction

The existence of triple helices was established in 1957, when optical density studies showed that complexes of poly(U)·poly(A)·poly(U) form a three-stranded structure (Felsenfeld, Davies & Rich, 1957). Conformational data for triple helical DNA was obtained in the mid 1970's by X-ray fibre diffraction studies on poly[d(T)]·poly[d(A)]·poly[d(T)] (Arnott & Selsing, 1974). These were interpreted in terms of an A-DNA like triple helix with the two poly[d(T)] strands oriented antiparallel and with sugar puckering in the C3′-*endo* region.

Interest in triple helices has recently been strengthened by novel applications such as antigene technology, which take advantage of the capacity for sequence recognition of double helical nucleic acids when specific hydrogen bonds are formed between Watson–Crick base pairs and the bases of a third strand. Experimental studies using nuclear magnetic resonance, infrared and Raman spectroscopy, as well as theoretical methods such as molecular modelling and dynamics have indicated different conformations of triple helices depending on the composition of the three strands (Radhakrishnan, de los Santos & Patel, 1991; Cheng & Pettitt, 1992; Laughton & Neidle, 1992*a*; Raghunathan, Miles & Sasisekharan, 1993; Piriou, Ketterlé, Gabarro-Arpa, Cognet & Le Bret, 1994; Radhakrishnan & Patel, 1994; for review see, Sun & Hélène, 1993).

In a study of (C·G)*G homopolymer triplexes [where '·' denotes Watson–Crick base pairing and '*' (reverse) Hoogsteen pairing] using molecular modelling and dynamics both parallel (Van Vlijmen, Ramé & Pettitt, 1990) and antiparallel (Laughton & Neidle, 1992*b*) triple helices were considered. Molecular modelling and Raman spectroscopy indicated that in homopolymer (C·G)*G triple helices, the third strand is parallel with respect to the homopurine strand of the Watson–Crick double helix, with Hoogsteen hydrogen bonding between the two purine strands (Ouali *et al.*, 1993). Both the $d(C)_n$ and third strand have C2′-*endo*, while the Watson–Crick $d(G)_n$ strand has C3′-*endo* sugar puckering.

To date, no high-resolution structure of a polynucleotide triple helix is known. The crystal structure of a 2:1 peptide nucleic acid–DNA triplex has been reported (Betts, Josey, Veal & Jordan, 1995), but crystals formed by nucleic acid triplexes produce only fibre-diffraction patterns (Liu, Miles, Parris & Sasisekharan, 1994). While interactions between a Watson–Crick base pair and a single base of a neighbouring helix have been observed in other crystal structures, these interactions do not display triplet behaviour. Either a single base interacts in the minor groove of a Watson–Crick base pair (Wing *et al.*, 1980; Joshua-Tor *et al.*, 1992; Leonard & Hunter, 1993), the third base is not coplanar with the base pair, and no Hoogsteen-like hydrogen-bond pattern is observed (Ramakrishnan & Sundaralingam, 1993), or the triplet is part of a pseudo four-way helix–helix junction (Spink, Nunn, Vojtechovsky, Berman & Neidle, 1995). Triplets have been observed in protein–DNA structures. A (C·G)*G triplet was found in a crystal structure of the *E. coli* catabolite gene activator

Table 1. *Statistics of data collection and refinement*

No. reflections measured	22402
No. unique reflections	2824
R_{symm} (%)	3.6
Reflections with $I > 3\sigma(I)$ (%)	96
Mean $I/\sigma(I)$	11.5
Completeness (%)	
16.0–2.04 Å	87
2.11–2.04 Å	71
Multiplicity	
16.0–2.04 Å	7.9
2.11–2.04 Å	8.1
Resolution (Å)	8.0–2.05
No. water molecules	86
R factor $[F > 4\sigma(F)]$	0.168
R.m.s. deviations from ideal stereochemistry	
Bond distances (Å)	0.015
Bond angles (°)	3.21

protein (CAP)† complexed with a 30-base-pair DNA sequence with overhanging G residues (Schultz, Shields & Steitz, 1991).

We have recently presented a novel way of obtaining structural details of DNA triplets by the use of overhanging bases (Van Meervelt *et al.*, 1995). We describe here in detail the crystal and molecular structure of the d(GCGAATTCG). The nonamer crystallizes in the B-DNA conformation with unpaired guanine bases at its ends. Two crystallographically independent (C·G)*G base triplets are formed by interaction of the overhanging guanine bases with the terminal C·G base pairs of neighbouring double helices.

2. Experimental procedures

Details of the synthesis, crystallization and data collection for the nonamer d(GCGAATTCG) have been reported previously (Van Meervelt *et al.*, 1995). The unit-cell dimensions are $a = 22.238(4)$, $b = 37.010(2)$, $c = 54.100(2)$ Å with space group $P2_12_12_1$. Statistics of data collection and refinement are given in Table 1.

The structure was solved by molecular-replacement techniques (*AMoRe*, Navaza, 1994). The starting model was based on the Dickerson–Drew dodecamer (CGCGAATTCGCG) (Wing *et al.*, 1980; Dickerson & Drew, 1981) with all but those residues present in the nonamer deleted. No good solution was found until the cut-off for Patterson vectors used was lowered to 12 Å using data between 5 and 8 Å. The solution gave an initial R factor of 0.43 and no bad contacts in *X-PLOR* (Brünger, Kuriyan & Karplus, 1987).

Energy minimization and B-factor refinement led to an R factor of 0.26. Hydrogen-bond restraints were used for the Watson–Crick base pairs. Water molecules were included using the following criteria: distances to the nearest atoms should be less than 3.5 Å, B factors should not be higher than 70 Å^2 and water molecules should be detectable in the $2F_o - F_c$ electron-density maps. In a first round 40 water molecules were added; refinement of their positions and B factors gave an R of 0.21. Six waters were removed because they did not appear in the new $2F_o - F_c$ map. In a second round 30 water molecules were added, giving an R factor of 0.191. Re-inspection of the positions of the waters and examination of both $F_o - F_c$ and $2F_o - F_c$ maps showed the presence of two cations. These are presumed to be Mg^{2+} ions, because the concentration of the Mg^{2+} ions is four times higher than the concentration of Na^+ ions in the initial crystallization solution; however, this cannot be established from the electron-density maps. Gradually, further solvent molecules were added. The refinement converged at $R = 0.168$ for the 2379 reflections between 8 and 2.05 Å $[F_{obs} > 4\sigma(F_{obs})]$ with 86 solvent molecules treated as O atoms and two Mg^{2+} ions. Atomic coordinates have been deposited with the Protein Data Bank at Brookhaven (Abola, Bernstein, Bryant, Koetzle & Weng, 1987).*

3. Results and discussion

Bases were labelled G1–G9 in the 5′ to 3′ direction on strand 1, G10–G18 on strand 2. Overall helical parameters, torsion angles, sugar pseudorotation parameters and groove widths were calculated with the *Newhel*93 program (Dickerson, personal communication). Calculations across a base pair and local helical parameters are calculated with the program *RNA* (Babcock, Pednault & Olson, 1993, 1994). Calculations of parameters for the base triplets were performed with *OCL* and *Morcad* (Gabarro-Arpa, Cognet & Le Bret, 1992; Le Bret, Gabarro-Arpa, Gilbert & Lemarèchal, 1991). All parameters are calculated according to the Cambridge nomenclature conventions (Dickerson *et al.*, 1989) unless stated otherwise.

3.1. *Helix structure*

The nonamer d(GCGAATTCG) adopts an eight-base-pair B-DNA double helical structure with an unpaired-guanine base at each end (Fig. 1*a*). The global helical twist of the octamer duplex is 33.6° giving 10.7 residues per turn and a helical rise of 3.5 Å.

† Abbreviations used: CAP, catabolite gene activator protein; C_{WC}, cytosine involved in Watson–Crick pairing with guanine in a triplet; G_{WC}, guanine involved in Watson–Crick pairing in a triplet; G_H, guanine involved in Hoogsteen or Hoogsteen-like pairing in triplets; r.m.s., root mean square.

* Atomic coordinates and structure factors have been deposited with the Protein Data Bank, Brookhaven National Laboratory (Reference: 208D). Free copies may be obtained through The Managing Editor, International Union of Crystallography, 5 Abbey Square, Chester CH1 2HU, England (Reference: HE0157).

Table 2. *Sugar–phosphate backbone and glycosidic torsion angles, and pseudorotation parameters for the nonamer d(GCGAATTCG)*

Backbone torsion angles are defined according to the IUPAC–IUB (1983) recommendations. The torsional angles are defined as: P-α-O5′-β-C5′-γ-C4′-δ-C3′-ε-O3′-ζ-P. χ is the torsional angle O1′—C1′—N1—C2 for pyrimidines and O1′—C1′—N9—C4 for purines. The pseudorotation parameters τ_M and P are defined according to Altona & Sundaralingam (1972). All torsional angles are given in degrees. The values for γ indicated with a * are not used in the calculations of the mean and the standard deviation because of the end effect of the 5′-terminal guanosine nucleotide.

Base	α	β	γ	δ	ε	ζ	χ	τ_M	P
G1	—	—	321*	158	234	189	255	45	161
C2	280	166	52	133	206	278	249	30	151
G3	289	172	39	138	225	174	285	44	144
A4	312	140	42	139	169	271	262	29	166
A5	278	190	52	127	177	274	256	33	134
T6	290	171	62	111	183	260	237	36	115
T7	299	170	60	127	195	269	244	37	131
C8	306	171	37	139	267	169	273	40	138
G9	270	155	37	138	—	—	265	22	165
G10	—	—	16*	155	221	179	272	43	167
C11	308	153	42	128	195	281	238	31	136
G12	292	167	50	134	255	161	272	45	131
A13	302	143	34	138	172	273	259	26	166
A14	298	175	47	120	177	263	250	38	122
T15	299	169	56	115	177	262	238	45	118
T16	296	187	55	130	189	274	245	35	135
C17	295	177	41	145	227	181	276	47	147
G18	301	139	48	135	—	—	253	37	142
Mean	295	165	47	134	204	235	257	37	143
SD	11	15	9	12	31	48	14	7	17

Table 2 gives the sugar–phosphate backbone and glycosidic torsion angles together with sugar pseudorotation parameters. The torsion angles α, β, γ and δ are, respectively, in their usual (−)*gauche*, *trans*, (+)*gauche* and (+)*anticlinal* range. Depending on torsion angles ε and ζ the backbone can adopt two different conformational types named B_I [ε, ζ *trans*, (−)*gauche*] and B_{II} [ε, ζ (−)*gauche*, *trans*] (Fratini, Kopka, Drew & Dickerson, 1982). The residues G1, G3, C8, G10, G12 and C17 have backbone type B_{II} (mean ε, ζ are 238 and 175°, respectively), while the others have backbone type B_I (with a mean for ε and ζ of 184 and 270°, respectively). In the dodecamer only two residues, G10 and G22 (comparable with G9 and G18 in the present structure), adopt the B_{II} backbone type. The glycosidic conformations are all *anti*, with an average torsion angle $\chi = 257°$. The mean pseudorotation angle is 143°, corresponding to a C2′-*endo*-C1′-*exo* conformation.

The molecule can be divided into three main parts: the central AATT region of the double helix and two novel duplex–triplex junctions. The central AATT part closely resembles that of the parent dodecamer in its principal helical parameters (Fig. 2). Only one significant difference is observed in this region between the two structures, namely opening of the base pair

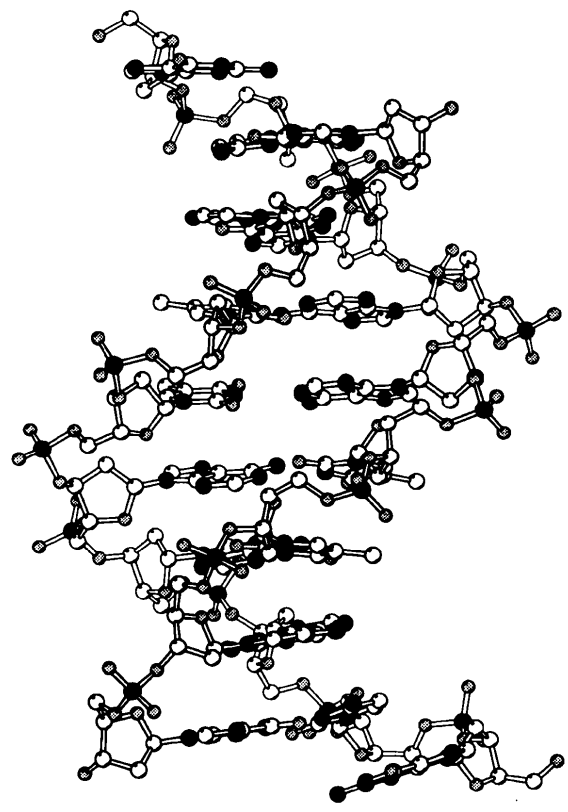


Fig. 1. (*a*) Global helical structure of d(GCGAATTCG), viewed into the minor groove. (*b*) Schematic representation of (C·G)*G triplet formation by overlapping duplexes. (*c*) Hoogsteen hydrogen-bond pattern in (C·G)*G triplets as observed in d(GCGAATTCG).

T6–A14. The width of the minor groove is similar, with a mean value of 4.7 Å (P–P distance minus twice the radius of P) compared to 4.1 Å for the dodecamer. A least-squares fit between the central AATT parts of the nonamer and the parent dodecamer gives a root-mean-square (r.m.s.) deviation of 0.4 Å.

In this fit, larger deviations are found at the ends. R.m.s. deviations of 1.3 and 0.9 Å are observed, respectively, for base pairs C2·G18 and G9·C11 which are involved in triplet formation, and of 1.6 and 1.7 Å for the neighbouring pairs (G3·C17 and C8·G12). At this point, it is difficult to decide whether these observations are the consequence of triplet formation or whether they are caused by the difference in sequence length (the dodecamer has two extra base pairs at each end).

The r.m.s. differences for the adjacent base pairs are reflected in the deviations in helical parameters compared to the same positions in the dodecamer: they show a smaller propeller twist, higher values of roll, shift and slide for the steps with the neighbouring A·T base pairs. Inclination and X and Z displacement also have higher values (data not shown). In the case of the triplet-forming base pairs, most of these parameters follow the opposite trend.

A computational and experimental study of double–triple helical junctions of a parallel triple helix (Chomilier *et al.*, 1992) shows a higher twist (34°) and very high rise (4.5 Å) for the double helical step that precedes the double helix–triple 5′-junction, while the junction itself adopts a lower twist (31°) and a much lower rise (3.3 Å). The puckering of the triple helical part and the adjacent Watson–Crick base pair is in the C3′-endo region. These conclusions differ from the present structure which contains two double helix–triple helix 5′-junctions (following the notations of Chomilier *et al.*, 1992). The twist follows the same trend, but is much larger (40°) and close to the value of the dodecamer. For one junction a lower rise (3.3 Å) is followed by a higher rise (3.4 Å), while for the other the

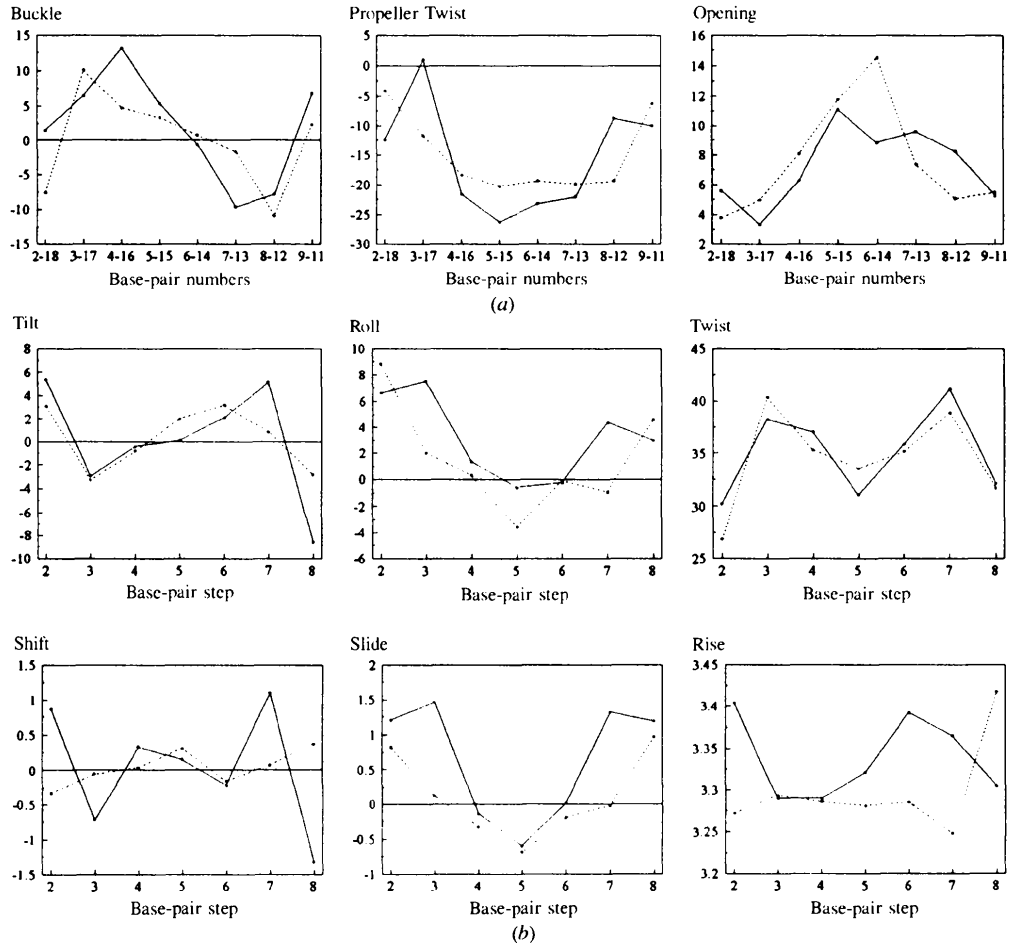


Fig. 2. Comparison of helical parameters calculated with the program *RNA* (Babcock *et al.*, 1993, 1994). (*a*) Interbase parameters and (*b*) Cartesian neighbouring base-pair parameters. Base-pair steps are the same as for the stacking (step 2 is the base-pair step C2pG3 ..., see legend of Fig. 4). Full lines indicate the nonamer parameters, dotted lines are for the dodecamer, d(CGCGAATTCGCG).

opposite is observed. Sugar puckers are in the C2′-*endo* region typical for B-DNA.

3.2. *Triplet formation*

Two crystallographically independent triplets are formed as a result of intermolecular contacts in the crystal. Nucleotide G1 interacts with the Watson–Crick base pair G9·C11 of a neighbouring duplex, while G10 similarly forms a triplet with base pair G18·C2 (Fig. 1*b*). Both overhanging bases interact in a Hoogsteen-like hydrogen-bond pattern (Fig. 1*c*) in the major groove of the Watson–Crick base pairs, with the sugar–phosphate backbones oriented parallel to those containing the Watson–Crick G nucleotides. The deoxyribose of each Hoogsteen nucleotide adopts a C2′-*endo* conformation. The hydrogen-bond distances between the third base and the Watson–Crick base pairs range from 2.7 to 2.9 Å, while between the Watson–Crick partners in the triplets the interatomic distances range from 2.7 to 2.9 Å.

Analysis of the parameters for adjacent single bases (Fig. 3) shows large variations in twist and shift. These changes are necessary for the correct alignment of the third base in the major groove of the Watson–Crick base pair. The adjustment of the shift parameters is important for the central positioning of the third base. Smaller differences are observed for the tilt and roll parameters, which bring the third base in the plane of the Watson–Crick base pair. This detailed analysis suggests that

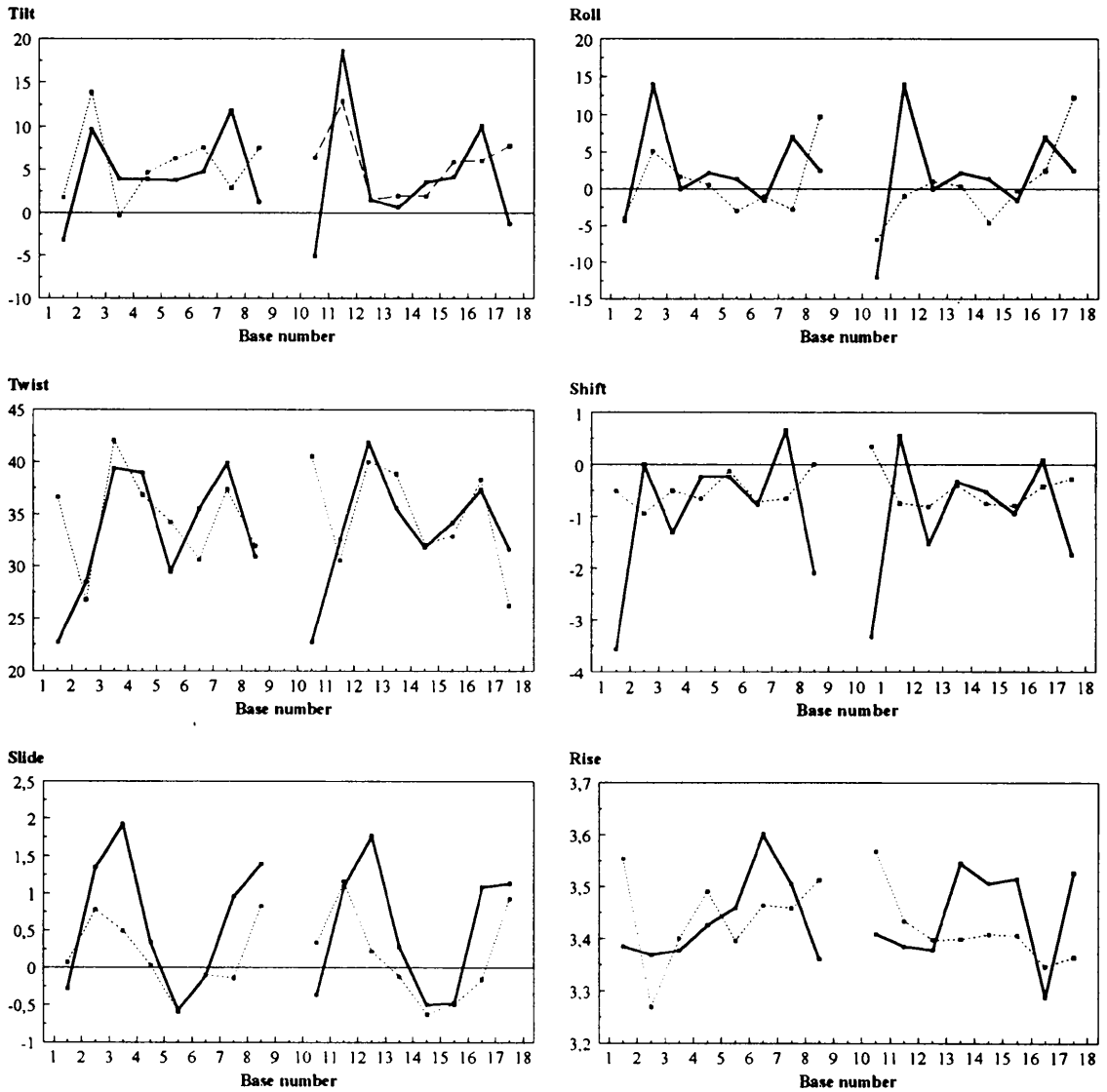


Fig. 3. Parameters for adjacent single bases using Cartesian coordinate frames calculated with the program *RNA* (Babcock *et al.*, 1993, 1994). Values for the nonamer are represented by a full line, dotted lines give data for the dodecamer, d(CGCGAATTCGCG).

Table 3. *Comparison of interbase parameters between G_{WC} and G_H in (C·G)*G triplets*

Interbase parameters are calculated according to the conventions adopted by Piriou *et al.* (1994): orthonormal reference frames are calculated for both bases with the x axis in the direction of the line passing through O6 and N7 for G_{WC}, and for G_H through O6 and N1, with the origin in the middle of the line joining both atoms. The z axis is perpendicular to the mean plane through the six-membered ring atoms, the y axis is defined so that the final orthonormal reference frame is right-handed.

Triplet	Buckle	Propeller twist	Opening	Shear	Stretch	Stagger
This structure						
(C2·G18)*G10	−4.8	5.7	0.0	0.0	2.9	0.1
(C11·G9)*G1	2.2	−0.4	1.6	0.0	2.7	−0.1
Triplet database†						
GCoGβpt	−8.4	−3.6	−0.6	0.0	2.8	−0.5
CAP–DNA complex‡						
(C17c·G15c)-*G16c	−2.6	9.2	−2.8	0.5	2.6	0.3
(C17d·G15d)-G16d	−13.0	21.0	−8.0	0.1	2.9	0.0

† *MORCAD* triplet database (Piriou *et al.*, 1994). In the namecode GCoGβpt the o refers to an unoccupied position (see Fig. 3 of Piriou *et al.*, 1994), β-stands for the β-anomer of the sugar of the third strand, p for the parallel orientation of the G_H in respect to G_{WC} and t for the *anti* orientation of the third strand base. ‡ Schultz *et al.* (1991).

helical parameters adjustments optimize the triplet formation.

The triplets are, as already stated, crystallographically independent. However, they are geometrically very similar as indicated by an r.m.s. deviation between the two sets of base atoms of 0.2 Å. Table 3 lists the interbase parameters of the third guanine bases (G_H) and the guanine of the Watson–Crick base pairs (G_{WC}). Small differences are observed for the rotational parameters (buckle, propeller twist, opening) but not for the translational parameters (shear, stretch, stagger). The triplets observed in the crystal are remarkably similar to the analogous triplet from the database of *Morcad*, which is based on fibre diffraction data (Piriou *et al.*, 1994) and to both triplets in the CAP–DNA complex (Schultz *et al.*, 1991). Differences are largest for rotational parameters suggesting more rotational than translational freedom in the Hoogsteen pairing.

In our earlier report (Van Meervelt *et al.*, 1995) we emphasized the existence of three hydrogen bonds in the triplets, with G_H spanning the Watson–Crick base pair. This situation contrasts, however, with the Hoogsteen pairing between guanines in a duplex containing a G·G mismatch (Skelly, Edwards, Jenkins & Neidle, 1993). There the guanines can be considered analogous to G_H and G_{WC} in the present structure, but in the absence of C_{WC} only two hydrogen bonds are possible and the relative positioning of the bases is accordingly different. A similar central positioning of the G_H is observed in the less planar (C·G)*G triplet in the CAP–DNA complex (Schultz *et al.*, 1991).

3.3. *Base stacking*

The central part of the nonamer displays similar stacking patterns to those of the dodecamer. Differences occur in the extreme steps (steps 1, 2, 3, 7 and 8, where step 1 is between the two triplets) as shown in Fig. 4. Steps 3 (Fig. 4*c*) and 7 (Fig. 4*d*) shown a similar stacking of C onto T and of G onto A. This is not observed in the analogous steps of the dodecamer, where the TpC step shows a destacking of T on C. The largest differences are observed for steps 2 (Fig. 4*b*) and 8 (Fig. 4*e*), a consequence of the doublet to triplet stacking. While in the dodecamer C3 (analogous to C2) destacks from G4 and G22 stacks heavily onto C21, the opposite is observed in the present structure: a greater extent of stacking of C2 onto G3, and a destacking of G18 on C17. This is explained by the fact that base C17 is located between the G bases of the triplet above. A similar effect is observed at the other end of the nonamer, with base C8 stacked between G9 and G1. The differences in stacking correlate with differences in helical parameters for these steps.

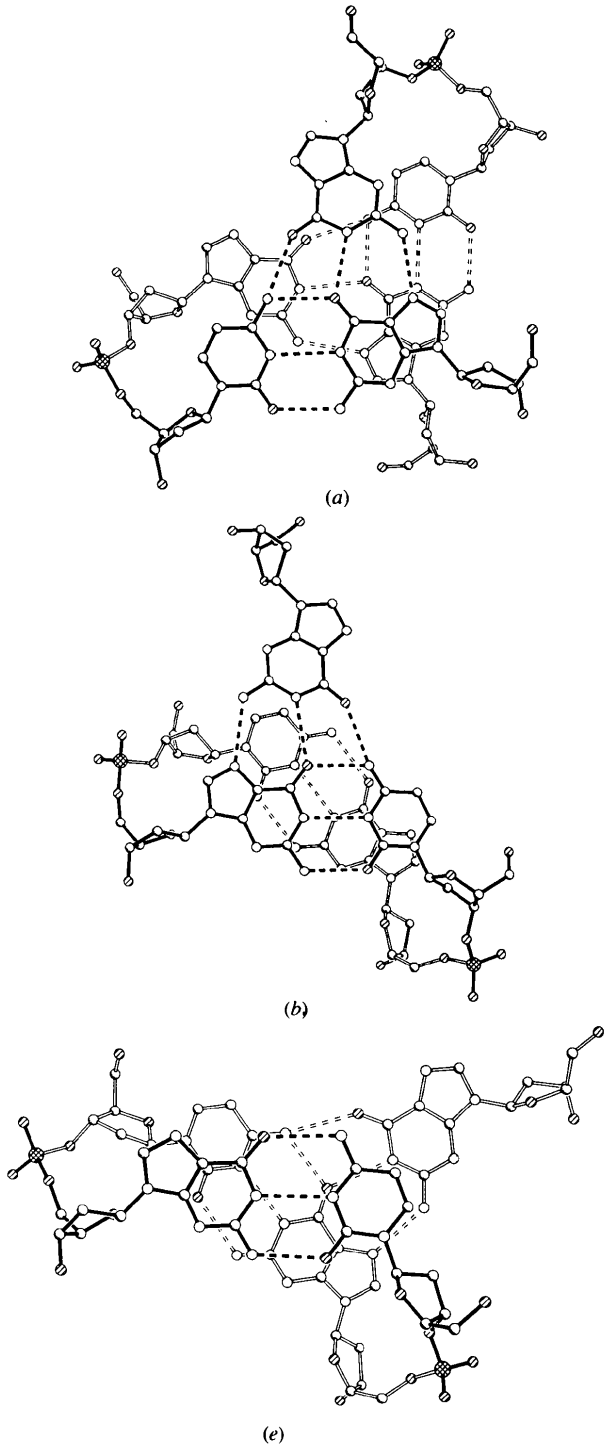
The stacking of G1 onto C2 and G10 onto C11 differs from that observed in the dodecamer (steps G2/C3, G13/C14) since it is clearly dominated by triplet formation by the terminal bases.

In the triplet stacking (step 1, Fig. 4*a*), both G_{WC} stack onto each other with the five-membered rings on top of the six-membered rings. The N4 atoms of the cytosine base in one triplet are stacked onto G_H of the other triplet.

3.4. *Hydration and thermal parameters*

X-ray crystallography can reveal the positions of tightly bound water molecules allowing us to examine the principles that underlie the hydration of DNA (Schneider, Cohen & Berman, 1992; Schneider *et al.*, 1993). The nonamer is heavily hydrated: 86 water molecules surround the oligonucleotide, with B factors ranging from 9.8 to 68.0 Å^2 (mean value 36.4 Å^2). Of these, 59 are in the first co-ordination shell and overall there are about 11 per base pair. As in the parent dodecamer, a spine of hydration is observed in the minor groove of the central base-pair steps, while water molecules in the major groove tend not to occupy bridging positions (Drew & Dickerson, 1981). Fig. 5 shows the hydration pattern in the minor groove of both structures. Minor differences are that no equivalent position in the nonamer structure is found for W78 in the dodecamer and that W58 has better bonding geometry compared to W88 in the dodecamer. In all other cases, there is a remarkable resemblance in distances and positions.

Most phosphate O atoms are hydrogen bonded to water molecules in a monodentate or bidentate fashion. Where two double helices are close together, water molecules bridge the sugar–phosphate backbones *e.g.* W23 connects A4(O2P) with A4(O2P) of a symmetry-equivalent molecule (distances are 2.8 and 3.2 Å, respectively). Interestingly, heavy hydration is observed in the vicinity of the major groove of the C2pG3 step, but not at the C11pG12 step which is equivalent by duplex symmetry.



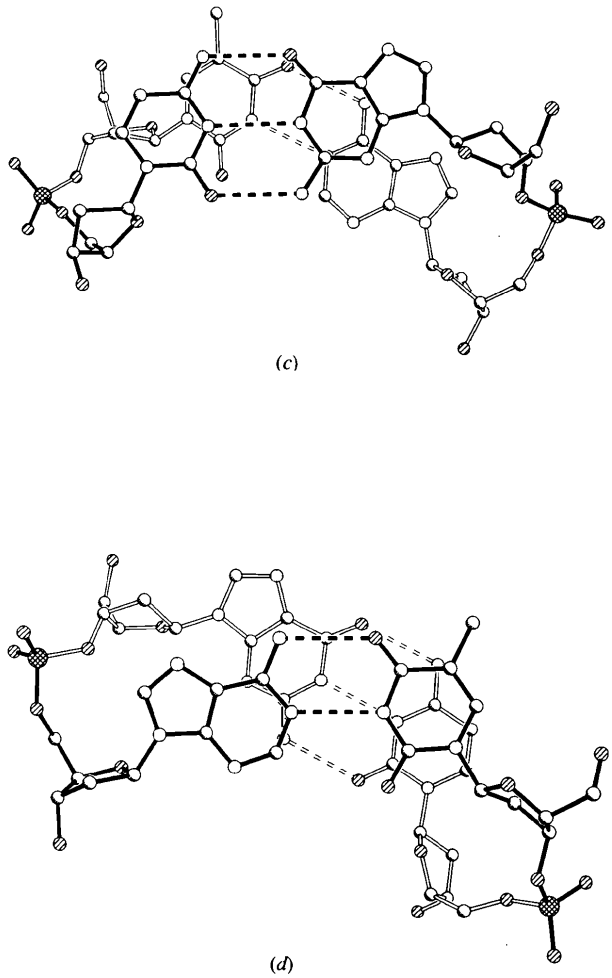


Fig. 4. Stacking of the extreme steps in d(GCGAATTCG). (*a*) Step 1 (C11·G9)*G1 on (C2·G18)*G10, (*b*) step 2 (C2·G18)*G10 on G3·C17, (*c*) step 3 G3·C17 on A4·T16, (*d*) step 7 T7·A13 on C8·G12 and (*e*) step 8 C8·G12 on (C11·G9)*G1. Covalent bonds of the base-pair/triplet closer to the viewer are filled.

Two metal ions, presumed to be Mg^{2+}, were be identified both near the triplets and occupying hexa-coordinate environments. The first ion is located in the vicinity of base pair C2·G18 and surrounded by six water molecules with distances ranging from 2.0 to 2.5 Å, accounting for the high hydration of this site. The second ion is in contact with four water molecules and also N7 of G1 and O2P of G12 of a symmetry-equivalent molecule. Interatomic distances range from 2.1 to 2.8 Å. In theoretical studies using *ab initio* quantum chemical methods (Jiang, Jernigan, Ting, Syi & Raghunathan, 1994) it is proposed that the preferred site of Mg^{2+} in a triplet is in the region between N7 and O6 of the third base. In the present study, the Mg^{2+} is more in the vicinity of N7 alone [Mg^{2+}–N7, 2.7 Å; Mg^{2+}–O6(G_H), 4.5 Å] and may be explained by the additional interaction with O2P(G12). The theoretically predicted position is, however, occupied by one of the water molecules of the Mg^{2+} cluster. The stacking of both triplets onto each other is further stabilized by water molecules [W29 bridging G18(O1P) with G9(N2) and W77 bridging G1(N2) with C2(O2) in the minor groove] and the waters of the second Mg^{2+} cluster, which connect both G_Hs by a network of hydrogen bonds. It is noteworthy that intramolecular (C·G)*G triplexes, often referred to as H′-DNA, are formed *in situ* in *E. coli* cells at neutral pH in the presence of Mg^{2+} (Kohwi, Malkhosyan & Kohwi-Shigematsu, 1992).

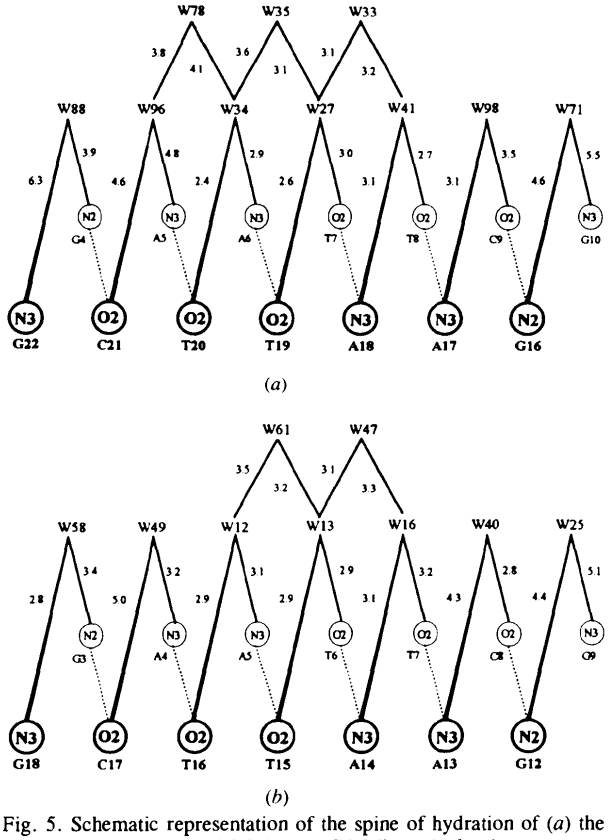


Fig. 5. Schematic representation of the spine of hydration of (*a*) the parent dodecamer d(CGCGAATTCGCG) and (*b*) the nonamer. Base atoms are encircled, water molecules are indicated with W and their follow-up number. Hydrogen-bonded distances are given in Å.

Fig. 6 shows an analysis of the mean *B* factors for the phosphates, sugars and bases. The classical trend $B_{phosph} > B_{sugar} > B_{base}$ is observed. Smaller *B* factors occur for the central AATT part, while they are higher for the outer residues.

3.5. *Crystal packing*

The packing in the nonamer crystals is quite different from that observed in other B-DNA crystal structures. The volume per base pair of 1391 $Å^3$ (calculated for eight base pairs; for nine base pairs a volume per base pair of 1237 $Å^3$) indicates a tightly packed crystal form as observed in orthorhombic dodecamers (Drew *et al.*, 1981; Nelson, Finch, Luisi & Klug, 1987; Yoon, Privé, Goodsell & Dickerson, 1988). Instead of the minor-groove interactions between terminal base pairs in the packing of the parent dodecamer and its analogues, major groove contacts and the end-to-end stacking typical of other B-DNA oligomers build here zigzag helical columns in which successive helices are related by a twofold screw axis.

This structure does not follow the correlation between volume per base pair and mean twist angle or mean number of base pairs per turn as described by Baikalov, Grzeskowiak, Yanagi, Quintana & Dickerson (1993): the mean number of base pairs per turn (10.7) is closer to the value for less dense crystal forms and polymeric B-DNA. The corresponding part of the dodecamer has 10.9 base pairs per turn and the similarity between the two conformations suggests that it is determined by the sequence rather than its environment, since in both packing arrangements it is held suspended in the crystal by interactions at the ends of the helices.

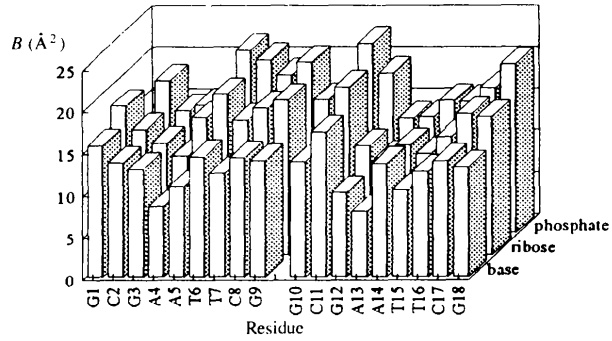


Fig. 6. Average thermal *B* factors ($Å^2$) for all nucleotides in d(GCGAATTCG). Atoms of bases, sugars and phosphates are grouped separately.

4. Conclusions

The detailed molecular structure of d(GCGAATTGC) in the solid state contains two main features of wider application to nucleic acids. These arise as a result of interactions between adjacent molecules in the crystal and from the conformation of the double helix itself, respectively.

(G·C)*G triplets are created in the crystal when the 5′ guanine bases of the nonamer, which overhang the double helical core, each make contact in the major groove with a terminal base pair of the neighbouring duplex. Both of these crystallographically independent triplets are coplanar and closely similar, and these properties together with the arrangement of hydrogen bonds connecting the component bases suggest a preferred geometry possibly common to longer fragments of triple helix.

The adjoining parts of the duplex are also indicative of likely backbone distortions and base stacking at duplex–triplex junctions in DNA. This particular result differs from theoretical predictions (Chomilier *et al.*, 1992), but whether it is a consequence of the crystal environment or the sequence is not at present evident. Further experimental work to adapt the sequence of the present nonamer template to produce other types of triplet, for example (T·A)*T, is in progress and may clarify this point.

The regions next to the two triplet-forming guanines in the crystal are occupied only by solvent molecules and allow space to extend the overlap with a second nucleotide without seriously disrupting the packing of the DNA. This would produce a dinucleotide fragment of triplex which could be used as a basis for modelling infinite triple helices. X-ray analysis of crystals of d(GGCCAATTGG), which can form two (G·C)*G triplets, is under way.

The AATT sequence is common to both the nonamer and the dodecamer d(CGCGAATTCGCG). In both structures crystal packing in these regions is free from interactions with neighbouring DNA molecules which are confined to contacts at the ends and effectively suspend the central portion of the duplex. The two structures appear similar in these regions and this is evident also from the correspondence between their helical parameters, base stacking, minor-groove dimensions and hydration. The comparison provides strong evidence that the conformation observed in each case is a property of the sequence itself rather than its environment.

Financial support by the Research Council of the K. U. Leuven (Belgium) (LVM) is gratefully acknowledged. This work has been accomplished with a fellowship from the IWT (DV) and the ERASMUS program. We thank S. A. Salisbury for his help with the preparation of the manuscript and H. Reynaers for his continued support. LVM is a Senior Research Associate of the National Fund for Scientific Research (Belgium).

References

Abola, E. E., Bernstein, F. C., Bryant, S. H., Koetzle, T. F. & Weng, J. (1987). *Crystallographic Databases - Information Content, Software Systems, Scientific Applications*, edited by F. H. Allan, G. Bergerhoff & R. Seivers, pp. 107–132. Cambridge: Data Commission of the IUCr.

Altona, C. & Sundaralingam, M. (1972). *J. Am. Chem. Soc.* **94**, 8205–8212.

Arnott, S. & Selsing, E. (1974). *J. Mol. Biol.* **88**, 509–521.

Babcock, M. S., Pednault, E. P. D. & Olson, W. K. (1993). *J. Biomol. Struct. Dynam.* **11**, 597–628.

Babcock, M. S., Pednault, E. P. D. & Olson, W. K. (1994). *J. Mol. Biol.* **237**, 125–156.

Baikalov, I., Grzeskowiak, K., Yanagi, K., Quintana, J. & Dickerson, R. E. (1993). *J. Mol. Biol.* **231**, 768–784.

Betts, L., Josey, J. A., Veal, J. M. & Jordan, S. R. (1995). *Science*, **270**, 1838–1841.

Brünger, A. T., Kuriyan, J. & Karplus, M. (1987). *Science*, **235**, 458–460.

Cheng, Y.-K. & Pettitt, B. M. (1992). *J. Am. Chem. Soc.* **114**, 4465–4474.

Chomilier, J., Sun, J.-S., Collier, D. A., Garestier, T., Hélène, C. & Lavery, R. (1992). *Biophys. Chem.* **45**, 143–152.

Dickerson, R. E., Bansal, M., Calladine, C. R., Diekmann, S., Hunter, W. N., Kennard, O., von Kitzing, E., Lavery, R., Nelson, H. C. M., Olson, W. K., Saenger, W., Shakked, Z., Sklenar, H., Soumpasis, D. M., Tung, C.-S., Wang, A. H.-J. & Zhurkin, V. B. (1989). *EMBO J.* **8**, 1–4.

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

Drew, H. R. & Dickerson, R. E. (1981). *J. Mol. Biol.* **151**, 535–556.

Drew, H. R., Wing, R. M., Takano, T., Broka, C., Tanaka, S., Itakura, K. & Dickerson, R. E. (1981). *Proc. Natl Acad. Sci. USA*, **78**, 2179–2183.

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

Fratini, A. V., Kopka, M. L., Drew, H. R. & Dickerson, R. E. (1982). *J. Biol. Chem.* **257**, 14686–14707.

Gabarro-Arpa, J., Cognet, J. A. H. & Le Bret, M. (1992). *J. Mol. Graphics*, **10**, 166–173.

IUPAC–IUB Joint Commission on Biochemical Nomenclature (1983). *Eur. J. Biochem.* **131**, 9–15.

Jiang, S.-P., Jernigan, R. L., Ting, K.-L., Syi, J.-L. & Raghunathan, G. (1994). *J. Biomol. Struct. Dynam.* **12**, 383–399.

Joshua-Tor, L., Frolow, F., Appella, E., Hope, H., Rabinovich, D. & Sussman, J. L. (1992). *J. Mol. Biol.* **225**, 397–431.

Kohwi, Y., Malkhosyan, S. R. & Kohwi-Shigematsu, T. (1992). *J. Mol. Biol.* **223**, 817–822.

Laughton, C. A. & Neidle, S. (1992*a*). *J. Mol. Biol.* **223**, 519–529.

Laughton, C. A. & Neidle, S. (1992*b*). *Nucleic Acids Res.* **20**, 6535–6541.

Le Bret, M., Gabarro-Arpa, J., Gilbert, J. C. & Lemarèchal, C. (1991). *J. Chim. Phys.* **88**, 2489–2496.

Leonard, G. A. & Hunter, W. N. (1993). *J. Mol. Biol.* **234**, 198–208.

Liu, K., Miles, H. T., Parris, K. D. & Sasisekharan, V. (1994). *Nature Struct. Biol.* **1**, 11–12.

Navaza, J. (1994). *Acta Cryst.* A**50**, 157–163.

Nelson, H. C. M., Finch, J. T., Luisi, B. F. & Klug, A. (1987). *Nature (London)*, **330**, 221–226.

Ouali, M., Letellier, R., Sun, J.-S., Akhebat, A., Adnet, F., Liquier, J. & Taillandier, E. (1993). *J. Am. Chem. Soc.* **115**, 4264–4270.

Piriou, J. M., Ketterlé, Ch., Gabarro-Arpa, J., Cognet, J. A. H. & Le Bret, M. (1994). *Biophys. Chem.* **50**, 323–343.

Radhakrishnan, I. & Patel, D. J. (1994). *Biochemistry*, **33**, 11405–11416.

Radhakrishnan, I., de los Santos, C. & Patel, D. J. (1991). *J. Mol. Biol.* **221**, 1403–1418.

Raghunathan, G., Miles, H. T. & Sasisekharan, V. (1993). *Biochemistry*, **32**, 455–462.

Ramakrishnan, B. & Sundaralingam, M. (1993). *J. Mol. Biol.* **231**, 431–444.

Schneider, B., Cohen, D. & Berman, H. M. (1992). *Biopolymers*, **32**, 725–750.

Schneider, B., Cohen, D. M., Schleifer, L., Srinivasan, A. R., Olson, W. K. & Berman, H. M. (1993). *Biophys. J.* **65**, 2291–2303.

Schultz, S. C., Shields, G. C. & Steitz, T. A. (1991). *Science*, **253**, 1001–1007.

Skelly, J. V., Edwards, K. J., Jenkins, T. C. & Neidle, S. (1993). *Proc. Natl Acad. Sci. USA*, **90**, 804–808.

Spink, N., Nunn, C. M., Vojtechovsky, J., Berman, H. M. & Neidle, S. (1995). *Proc. Natl Acad. Sci. USA*, **92**, 10767–10771.

Sun, J.-S. & Hélène, C. (1993). *Curr. Opin. Struct. Biol.* **3**, 345–356.

Van Meervelt, L., Vlieghe, D., Dautant, A., Gallois, B., Précigoux, G. & Kennard, O. (1995). *Nature (London)*, **374**, 742–744.

Van Vlijmen, H. W. Th., Ramé, G. L. & Pettitt, B. M. (1990). *Biopolymers*, **30**, 517–532.

Wing, R., Drew, H., Takano, T., Broka, C., Tanaka, S., Itakura, K. & Dickerson, R. E. (1980). *Nature (London)*, **287**, 755–758.

Yoon, C., Privé, G. G., Goodsell, D. S. & Dickerson, R. E. (1988). *Proc. Natl Acad. Sci. USA*, **85**, 6332–6336.