Proc. Natl. Acad. Sci. USA
Vol. 87, pp. 6693–6697, September 1990
Biophysics

Crystal structure and sequence-dependent conformation of the A·G mispaired oligonucleotide d(CGCAAGCTGGCG)

(x-ray crystallography/DNA structure)

GORDON D. WEBSTER*, MARK R. SANDERSON*, JANE V. SKELLY*, STEPHEN NEIDLE*†, PETER F. SWANN‡, BEN F. LI‡, AND IAN J. TICKLE§

*Cancer Research Campaign Biomolecular Structure Unit, The Institute of Cancer Research, Sutton, Surrey SM2 5NG, United Kingdom; ‡Department of Biochemistry, University College London, Gower Street, London WC1E 6BT, United Kingdom; and §Department of Crystallography, Birkbeck College, Malet Street, London WC1E 7HX, United Kingdom

Communicated by Richard E. Dickerson, May 14, 1990 (received for review February 20, 1990)

ABSTRACT **The crystal structure of the dodecanucleotide d(CGCAAGCTGGCG) has been determined to a resolution of 2.5 Å and refined to an *R* factor of 19.3% for 1710 reflections. The sequence crystallizes as a B-type double helix, with two G(*anti*)·A(*syn*) base pairs. These are stabilized by three-center hydrogen bonds to pyrimidines that induce perturbations in base-pair geometry. The central AGCT region of the helix has a wide (>6 Å) minor groove.**

Mispairing in DNA can occur through genetic recombination events or as errors in replication (1). These mismatches have mutational consequences unless they are corrected by repair processes such as excision repair (2), which in *Escherichia coli* involves the *uvrD* gene product DNA helicase II and the *dam* methylase (3, 4). The A·G mismatch, which results in transversions, appears to have specific excision repair mechanisms associated with it, at least in *E. coli* (2), with the *MutT* protein appearing to prevent G·A pairing during replication (5). A·G repairs are less efficiently repaired in some mammalian cells than other heterogeneous mismatches, possibly because of structural differences between them (6).

A number of distinct base-pairing possibilities have been suggested for A·G pairing (7, 8), including the involvement of imino tautomers for either adenine or guanine. NMR spectroscopy has shown that G(*anti*)·A(*anti*) mismatches occur in some oligonucleotides (9, 10) with sequences around the mismatches of 5′-R-A(=G)-G(=A)-Y-3′ and 5′-R-A(=G)-R-3′, respectively (where = donates a mispair, R is a purine, and Y is a pyrimidine). Alternative G(*syn*)·A(*anti*) mispairing has been found at low pH by NMR methods in oligonucleotides with A(=G) flanked at both 5′ and 3′ sides by either guanine or cytosine (11). X-ray crystallographic studies have revealed both of these possibilities, as well as the less expected G(*anti*)·A(*syn*) arrangement (12, 13), in the sequence d(CGCGAATTAGCG). The G(*anti*)·A(*anti*) pairing has been found for the two contiguous A·G base pairs in the crystal structure of the sequence d(CCAAGCTTGG) (14, 15) and G(*syn*)·A(*anti*) in the dodecanucleotide d(CGCAAATTGGCG) (16). The latter structure has by implication protonation at N1 of the adenine to achieve the observed base pairing.

5′- CGCAAGCTGGCG
GCGGTCGAACGC
Scheme I

We present here the crystal structure of the dodecanucleotide duplex d(CGCAAGCTGGCG) (Scheme I) with A·G base pairs at positions 4 and 9 of the duplex.¶ These are both of the G(*anti*)·A(*syn*) type. This sequence differs substantially from the classic "Dickerson–Drew" type by not having a central A+T-rich region, with its characteristic minor groove.

MATERIALS AND METHODS

The dodecamer was synthesized by the phosphotriester method and purified by reverse-phase HPLC. Crystals were grown at 5°C at pH 7.2 from 10-μl droplets containing 0.75 mM dodecamer, 25 mM $MgCl_2$, and 2.5 mM spermine in 30 mM sodium cacodylate as buffer with a reservoir at 32% (vol/wt) 2-methyl-2,4-pentandiol. Crystals grew to a usable size in 10–14 days. Unit-cell dimensions are a = 25.29 Å, b = 41.78 Å, and c = 64.76 Å, with four duplex molecules in the unit cell of space group $P2_12_12_1$.

Intensity data were collected on an Enraf–Nonius FAST area detector at Birkbeck College for a crystal of size 0.5 × 0.06 × 0.02 mm mounted in a sealed capillary tube. A total of 9842 reflections were collected to a resolution of 2.2 Å at 7°C. Only the 2495 reflections to 2.5 Å (with a merging R factor of 11.9%) were used in the analysis since the merging agreement factor for the (mostly weak) reflections in the range of 2.2–2.5 Å was unacceptably high. A total of 1710 unique reflections with $F_o > 3\sigma\ (F_o)$ were obtained.

The structure was solved by molecular replacement using the established structure of the dodecamer d(CGCGAATTCGCG) (17). The initial stages of refinement were undertaken using the CORELS constrained-restrained refinement package (18). The coordinates of the Dickerson dodecamer d(CGCGAATTCGCG) were refined as a rigid body in the present mismatch unit cell for six cycles yielding an R factor of 50.3% at 5-Å resolution. The sequence of this dodecamer was then altered to that of the mismatch structure using the GEMINI molecular graphics program (19), written by A. Beveridge (Institute of Cancer Research), and substituting the new bases, overlaid in the positions of the old. The A·G mismatched base pairs were included in the G(*anti*)·A(*anti*) conformation.

Further cycles of rigid-body refinement were done, increasing the resolution to 3 Å, before breaking up the structure into groups consisting of phosphates, sugars, and bases, a total of 48 rigid groups. Successive cycles of refinement using CORELS dropped the R factor to 33.9% at a resolution of 2.5 Å.

At this point, a $2F_0 - F_c$ map was generated and the structure was checked and fitted to the density as necessary. An "omit" map, with the phases from the structure with the A·G mismatches omitted, was also examined, which clearly showed that both the mismatches were in the G(*anti*)·A(*syn*)

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

†To whom reprint requests should be addressed.

¶The atomic coordinates and structure factors have been deposited in the Protein Data Bank, Chemistry Department, Brookhaven National Laboratory, Upton, NY 11973 (reference 1 DNM).

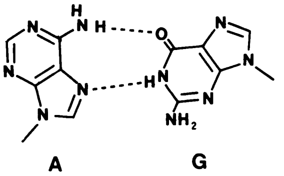

FIG. 1. A(*syn*)·G(*anti*) base-pairing arrangement, as found in this study.

conformation. The existing structure was altered accordingly, and further cycles of refinement were interspersed with periodic checks on the structure by viewing the model and its corresponding density map on the graphics and adjusting the model for best fit. The *R* factor dropped to 29.8% at a resolution of 2.5 Å before changing to the restrained-refinement program NUCLSQ (20, 21) for the remainder of the refinement. On further refinement the *R* factor dropped to 27.4%.

Difference Fourier maps revealed the positions of 47 water molecules with acceptable hydrogen-bonding geometry and the structure was refined to a final *R* factor of 19.3% at a resolution of 2.5 Å for the 3σ data.

As a check, parallel refinements using the procedure described above were made with the two A·G mismatched bases completely omitted from the structure. By using omit maps, the conformations of the A·G mismatched base pairs could be unambiguously assigned as G(*anti*)·A(*syn*) in both mismatches. As an extra precaution, the structure was further refined with both G(*anti*)·A(*anti*) and G(*syn*)·A(*anti*) mismatches, both of which failed to refine as well as the G(*anti*)·A(*syn*) model, with significantly higher (>2%) *R* factors. Maps were generated with the PROTEIN (22) package and displayed by using TOM, a version of FRODO (23), amended to run on a Silicon Graphics IRIS workstation by C. M. Cambillau (Marseilles).

RESULTS

Base-Pair Geometry. The overall structure of the dodecamer is an anti-parallel double helix with A·G base pairs between position 4 on strand 1 and position 21 on strand 2, and between position 9 on strand 1 and position 16 on strand 2. Both A·G base pairs in this structure have G(*anti*)·A(*syn*) conformations (Figs. 1 and 2). The fit of bases to the electron density is unequivocal; alternative anti or syn conformations would involve significant (>2 Å) base movements. There is no suggestion of alternative conformers in the crystal structure or of a disordered situation in their vicinity. The base pairings have hydrogen bonding between N7(adenine) and N1(guanine) and between N6(adenine) and O6(guanine) with distances A(4)N7 · · · N1G(21) of 3.03 Å, A(4)N6 · · · O6G(21)

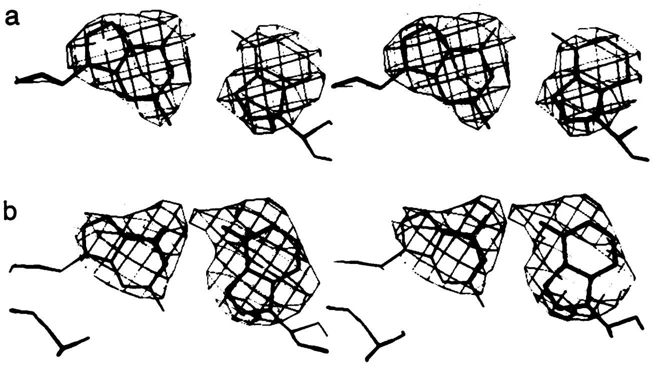

FIG. 2. "Omit" electron density calculated in the plane of the A·G base pairs with the base pair omitted from the structure-factor calculation. The contours are drawn at equal-arbitrary intervals and stereo views are shown. (*a*) A(4)·G(21) base pair. (*b*) G(9)·A(16) base pair.

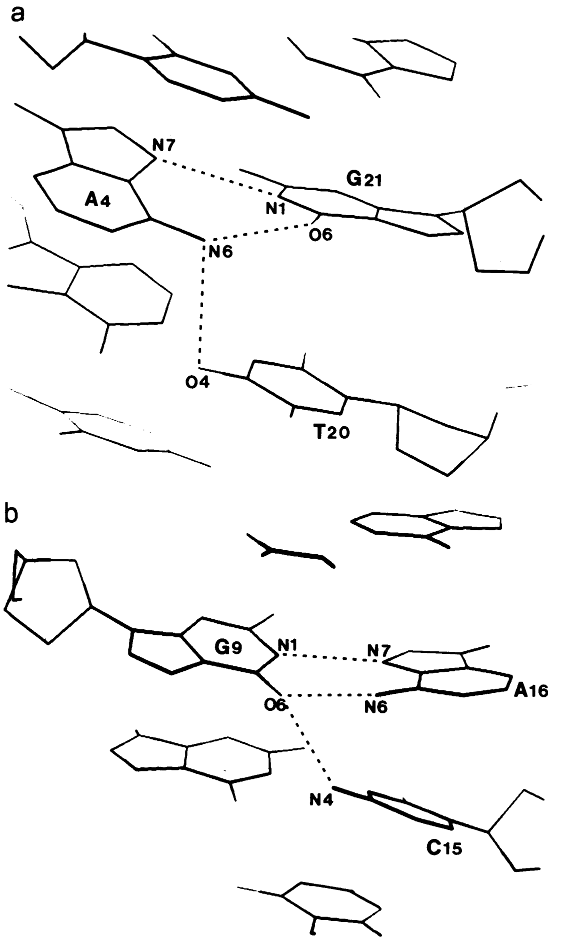

FIG. 3. Views of the two A·G base pairs and their adjacent bases. Dashed lines represent hydrogen bonds. (*a*) A(4)·G(21) base pair. (*b*) G(9)·A(16) base pair.

Table 1. Helical parameters

	Twist, deg.	Inclination, deg.	Roll, deg.	Tilt, deg.	Propeller twist, deg.	Slide, Å
C·G					−12	
	42	−1	2	0		0
G·C					−23	
	35	4	−3	3		1
C·G					0	
	35	8	2	2		1
A·G					−19	
	40	6	1	2		1
A·T					−14	
	32	4	16	−2		0
G·C					−16	
	30	0	−4	−2		0
C·G					−12	
	38	−1	8	3		1
T·A					9	
	37	4	−3	−1		2
G·A					−15	
	32	6	5	−1		−1
G·C					−20	
	38	6	−10	−2		1
C·G					−20	
	40	1	1	1		0
G·C					−11	

Parameters were calculated with the NEWHELIX program using the Cambridge conventions (24).

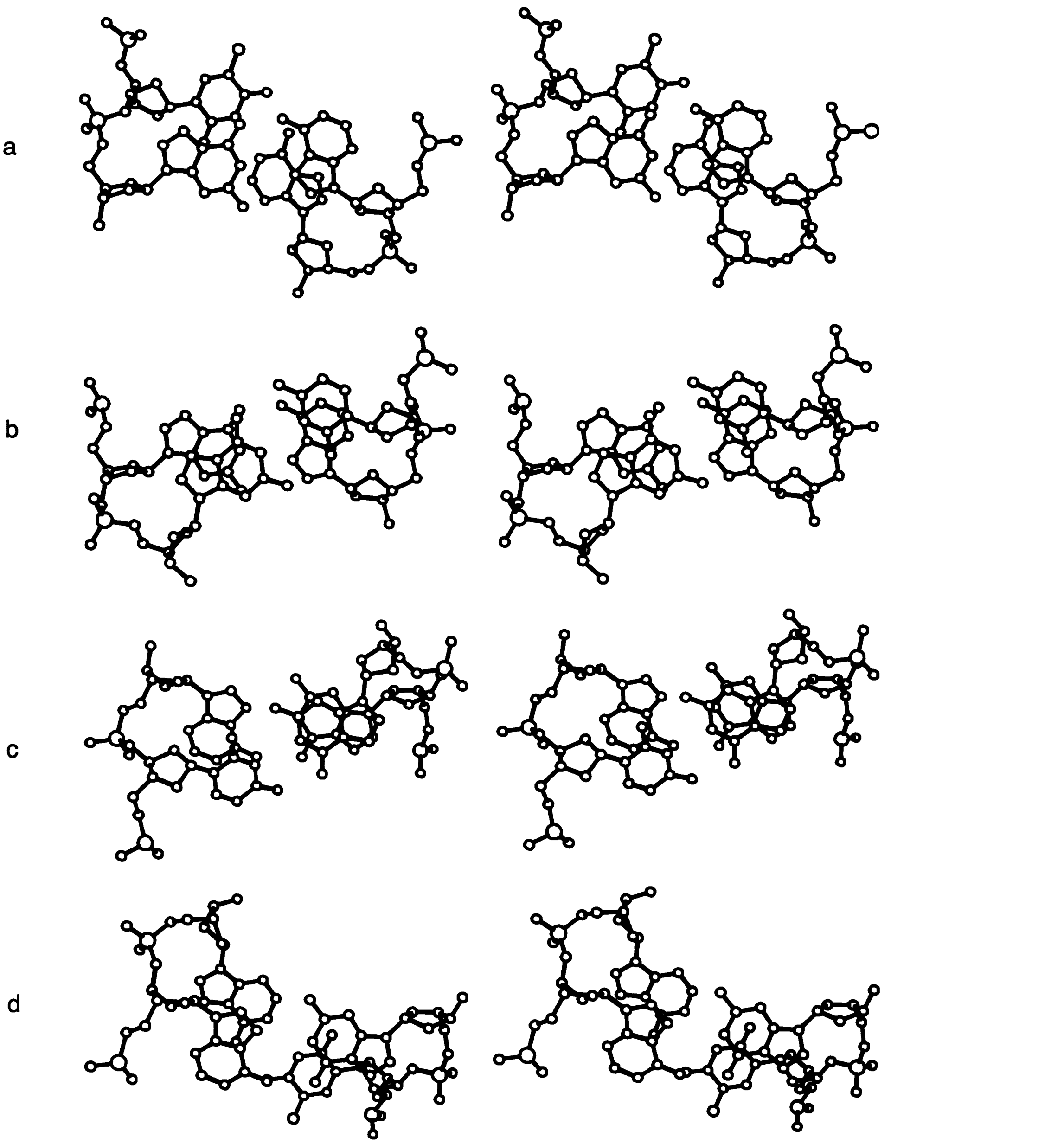

FIG. 4. Stereo views projected onto the base-pair planes showing overlaps of various bases on the A·G mismatch base pairs. (*a*) A(4)·G(21) and A(5)·T(20) base pairs. (*b*) C(3)·G(22) and A(4)·G(21) base pairs. (*c*) G(9)·A(16) and G(10)·C(15) base pairs. (*d*) T(8)·A(17) and G(9)·A(16) base pairs.

of 2.68 Å, A(16)N7 ··· N1G(9) of 3.12 Å, and A(16)N6 ··· O6G(9) of 2.61 Å (where numbers in parentheses are positions of bases). This consistent difference in length between the two hydrogen bonds in the A·G base pair may be significant in view of the lack of restraints between bases during the refinement.

There are two distinct three-center major-groove hydrogen bonds in this structure (Fig. 3). Both are from a base in an A·G base pair, and both involve a strand-2 pyrimidine base on the 5′ side of an A·G pair in an interstrand network. The first has a A(4)N6 ··· O4T(20) hydrogen bond of 2.89 Å as well as the hydrogen bond of the A(4)N6 ··· O6G(21) A·G base pair. Each hydrogen bond involves a distinct hydrogen atom on N6; thus the three-center hydrogen-bonding network is not a bifurcated one. The angle O6–N6–O4 is 103° and the individual N6–H6 ··· oxygen angles are likely to be around 140° (as judged by calculation of likely hydrogen atom positions). The second three-center network involves G(9)O6 ··· N4C(15) (2.82 Å) and the G(9)O6 ··· N6A(16) A·G hydrogen bond. The angle subtended by G(9)O6 and the 3′ side of guanine at G(10)O6 with the hydrogen atom on N4C(15) is ≈100°; this hydrogen atom is in a position implying that it is shared between the two hydrogen bonds G(9)O6 ··· N4C(15) and G(10)O6 ··· N4C(15).

Both mismatched base pairs have large propeller twists (Table 1), as do the majority of base pairs in the structure as a whole. It is notable that the C·G base pairs tend to be more twisted than the two A·T base pairs. Base-pair steps on the 3′ sides of the mismatches have large roll angles, albeit in different directions. Base pairs 5 and 6 are rolled open by almost 16° to the major groove whereas base pairs 10 and 11 are rolled by 10° to the minor groove. Base-pair steps immediately adjacent to the mismatches have small rolls. Base pairs 7 and 8, almost in the center of the helix and between the two mismatches, have an 8° major groove roll.

The stacking patterns of bases around the mismatches are shown in Fig. 4. It is remarkable that a number of the intrastrand base steps are very poorly stacked, with the two Y-3′·5′-R steps

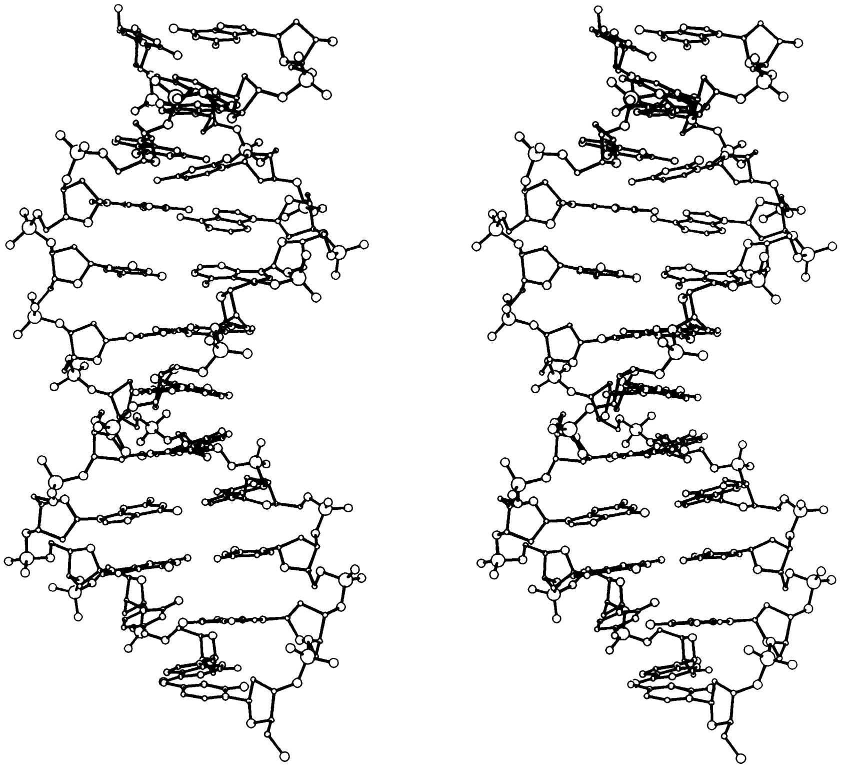

FIG. 5. Stereo view of the dodecanucleotide double helix.

T(8)-G(9) and C(15)-A(16) having essentially no overlap at all. Adenine-16 lacks any significant stacking on both 5′ and 3′ sides; there is a large slide associated with this step (Table 1).

Helix and Its Conformational Features. The helix is of overall B type (Fig. 5) with an average helical twist of 39.6°, which thus defines it as a 9-fold helix. Individual helical twist angles vary over almost 13°, with no significant tendency for lower values at R-3′-5′-Y steps. The minor groove of the helix is of approximately constant width, with an average of >6 Å (Fig. 6), and thus there is no narrowing of groove width in the central region of the structure.

DISCUSSION

A·G Base Pairing. The finding of A·G base pairs in the G(*anti*)·A(*syn*) conformation is in accord with the assignments from the crystallographic analysis of d(CGCGAATT-AGCG) (12, 13), although as yet this alone of the three possible anti–syn possibilities has not been reported by NMR

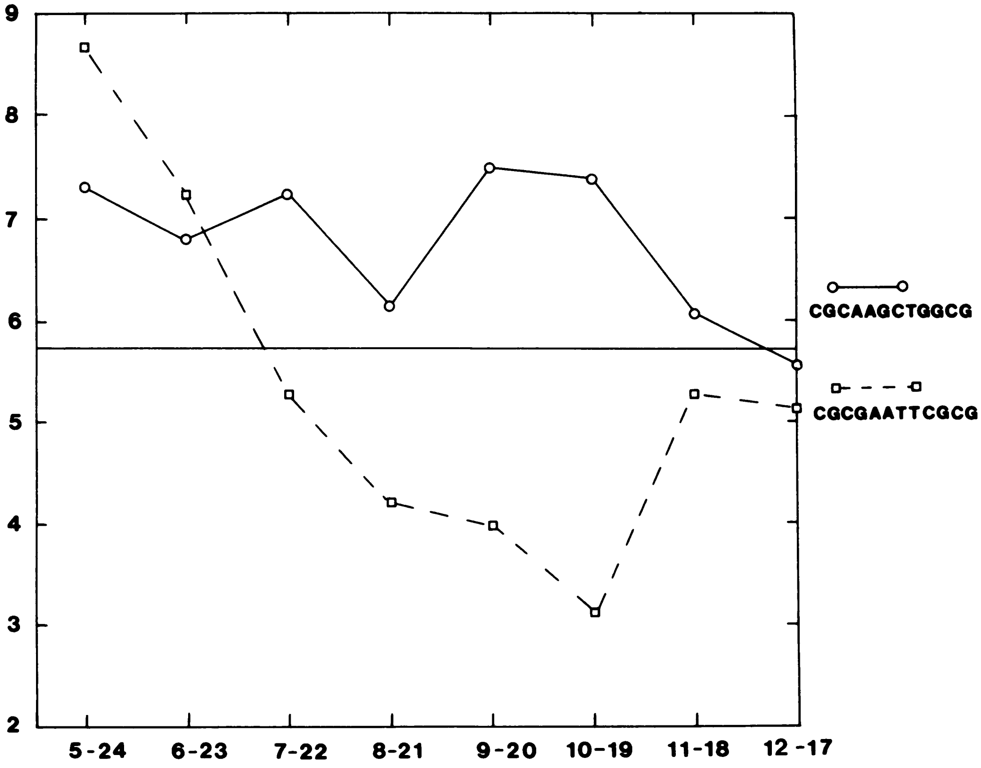

FIG. 6. Plot of the minor groove width in the structure (○—○), compared to that for the native dodecanucleotide (17) (□—□). The horizontal line represents the minor groove width for canonical B-DNA. Groove width is defined as the P–P distance minus the sum of the phosphate group radii (5.8 Å). Horizontal and vertical axes denote interstrand phosphate pairs and groove width, in Å respectively.

Table 2. Sequence environments for A·G mismatched oligonucleotides whose structures have been determined

Sequence	Conformation	pH	Ref(s).
	From NMR studies		
AGAT	A(*anti*)·G(*anti*)	7	9
GAG	A(*anti*)·G(*anti*)	7	10
CAC	A(*anti*)·G(*anti*)	7	11
	A(*anti*)·G(*syn*)	<7	
	From x-ray studies		
TAG	A(*syn*)·G(*anti*)	7.4	12, 13
AGAT	A(*anti*)·G(*anti*)	7(?)	14, 15
CAA	A(*anti*)·G(*syn*)	6.6	16
CAC	A(*syn*)·G(*anti*)	7.2	This work

A·G base pairs are indicated in bold type.

studies to be present in solution (9–11). Table 2 lists those oligomers for which x-ray or NMR studies have provided an assignment. The observation of A(*syn*)·G(*anti*) in two different crystal structures (refs. 12 and 13 and this study), both in the sequence 5′-YAR, suggests that it is a major form at and greater than pH 7.0 (i.e., at physiological pH). The crystals of the sequence d(CGCAAATTGGCG) were grown at pH 6.6, and the finding of A(*anti*)·G(*syn*) protonated base pairs in the resulting structure (16), which has identical flanking bases as the present one, suggests that pH (and, therefore, the tautomeric equilibria of the bases) plays a major role in addition to sequence. The stabilization of the base-pairing arrangement in the present structure by three-center hydrogen bonds to pyrimidines is probably also possible with appropriate purines in their place. The x-ray and NMR evidence consistently indicates that 5′-RAR sequences tend to prefer A(*anti*)·G(*anti*) base pairs. Theoretical studies have suggested (25–27) that the energy difference between the different conformers is likely to be ≈1 kcal/mol; clearly factors such as hydrogen bonding in addition to normal base–base interactions can force one type to be stabilized. It has been pointed out (6, 12) that the particular repair capabilities of the A·G mispair may well be due to the syn conformation adopted by adenosine not being adequately recognized by repair enzymes. If the features of the G(*anti*)·A(*syn*) as found here are indeed adopted by biological DNA, then the fact that the major-groove region close to the mismatch does not have its full complement of hydrogen-bond donors and acceptors (due to their involvement in the three-center hydrogen bonds) might diminish active enzyme recognition of the lesion.

Three-Center Hydrogen Bonds. The existence of three-center hydrogen bonds was first observed in oligonucleotide crystal structures containing an oligo(dA)·oligo(dT) sequence (28, 29) with bifurcated interstrand hydrogen bonds from N6 of adenines to O4 of thymines on the 3′ side. The arrangement at A(4) · · · T(20) in the present structure is thus very similar, except that here the hydrogen bond is clearly not bifurcated. The G(9) · · · C(15)hydrogen bond here has precedence in that found recently in a G+C-rich decamer structure (30). The angular criteria defined in this study for three-center hydrogen-bond networks are amply fulfilled by the present structure. It seems that this type of interaction is readily formed by a variety of DNA sequence types, with the observation of minor-groove three-center hydrogen bonds in the mismatched decamer (14), suggesting that runs of contiguous non-Watson–Crick base pairs may induce quite distinct interactions. It is not clear why three-center interactions are not present in the other G(*anti*)·A(*syn*) structures (12, 13); it is tempting to speculate that the particular structural features of the central AATT sequence in this structure do not enable these additional interactions to take place.

The three-center hydrogen bonding in d(CGCAAGCTGGCG) has resulted in a number of the structural features noted in this paper. Thus, both A·G base pairs are highly propeller twisted to attain the distance to the 3′ base for the additional hydrogen bonding. This also produces marked base tilt (Table 1), again to achieve the additional contacts. On the other hand, the wide minor groove in the central AGCT region is likely to be an intrinsic consequence of this sequence rather than of the mispairing or its effects. No significant changes to the narrow AATT region in the Dickerson–Drew type of sequence has been observed when A·G mismatches are included (12, 13). We note that the AGCT mixed-sequence wide groove is representative of canonical B-DNA and is to our knowledge the first dodecamer to display such features.

We are grateful to Joel Sussman and Eric Westhof for advice on the use of their programs, to Tom Brown for discussions, and to Tom Blundell for access to facilities at Birkbeck College. This work was supported by the Cancer Research Campaign (grants to S.N. and P.F.S.).

1. Modrich, P. (1987) *Annu. Rev. Biochem.* **56,** 435–466.
2. Lu, A.-L. & Chang, D.-Y. (1988) *Cell* **54,** 805–812.
3. Lu, A.-L., Clark, S. & Modrich, P. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 4639–4643.
4. Kramer, B., Kramer, W. & Fritz, H.-J. (1984) *Cell* **38,** 879–887.
5. Akiyama, M., Maki, H., Sekiguchi, M. & Horiuchi, T. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 3949–3952.
6. Brown, T. C. & Jiricny, J. (1988) *Cell* **54,** 705–711.
7. Topal, M. D. & Fresco, J. R. (1976) *Nature (London)* **263,** 285–293.
8. Traub, W. & Sussman, J. L. (1982) *Nucleic Acids Res.* **10,** 2701–2708.
9. Kan, L. S., Chandrasegaran, S., Pulford, S. M. & Miller, P. S. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 4263–4265.
10. Patel, D. J., Kozlowski, S. A., Ikuta, S. & Itakura, K. (1984) *Biochemistry* **23,** 3207–3217.
11. Gao, X. & Patel, D. J. (1988) *J. Am. Chem. Soc.* **110,** 5178–5182.
12. Brown, T., Hunter, W. N., Kneale, G. & Kennard, O. (1986) *Proc. Natl. Acad. Sci. USA* **83,** 2402–2406.
13. Hunter, W. N., Brown, T. & Kennard, O. (1986) *J. Biomol. Struct. Dyn.* **4,** 173–191.
14. Privé, G. G., Heinemann, U., Chandrasegaran, S., Kan, L.-S., Kopka, M. L. & Dickerson, R. E. (1987) *Science* **238,** 498–504.
15. Privé, G. G., Heinemann, U., Chandrasegaran, S., Kan, L.-S., Kopka, M. L. & Dickerson, R. E. (1988) in *Structure and Expression*, eds. Olson, W. K., Sarma, M. H., Sarma, R. H. & Sundaralingam, M. (Adenine, Schenectady, NY), Vol. 2, pp. 27–47.
16. Brown, T., Leonard, G. A., Booth, E. D. & Chambers, J. (1989) *J. Mol. Biol.* **207,** 455–457.
17. Dickerson, R. E. & Drew, H. R. (1981) *J. Mol. Biol.* **149,** 761–786.
18. Sussman, J. L., Holbrook, S. R., Church, G. M. & Kim, S.-H. (1977) *Acta Crystallogr.* **33,** 800–804.
19. Beveridge, A. (1989) GEMINI, *a Molecular Modeling Program for Silicon Graphics IRIS Workstations* (Hampden Data Systems, Foxcombe Court, Wyndyke Furlong, Abingdon Business Park, Abingdon, Oxon OX14 1DZ, U.K.).
20. Konnert, J. H. & Hendrickson, W. A. (1980) *Acta Crystallogr. Sect. A* **36,** 344–350.
21. Westhof, E., Dumas, P. & Moras, D. (1985) *J. Mol. Biol.* **184,** 119–145.
22. Steigemann, W. (1974) Dissertation (Technische Universität, Munich).
23. Jones, T. A. (1982) in *Computational Crystallography*, ed. Sayre, D. (Clarendon, Oxford), pp. 303–310.
24. Dickerson, R. E., *et al.* (1989) *EMBO J.* **8,** 1–4.
25. Chaprina, V. P. & Poltev, V. I. (1983) *Nucleic Acids Res.* **11,** 5205–5223.
26. Keepers, J. W., Schmidt, P., James, T. L. & Kollman, P. A. (1984) *Biopolymers* **23,** 2901–2929.
27. Rao, S. N., Singh, U. C. & Kollman, P. A. (1986) *Isr. J. Chem.* **27,** 189–197.
28. Nelson, H. C. M., Finch, J. T., Luisi, B. F. & Klug, A. (1987) *Nature (London)* **330,** 221–223.
29. Coll, M., Frederick, C. A., Wang, A. H.-J. & Rich, A. (1987) *Proc. Natl. Acad. Sci. USA* **84,** 8385–8389.
30. Heinemann, U. & Alings, C. (1989) *J. Mol. Biol.* **210,** 369–381.