J. Mol. Biol. (1992) **225**, 397–431

Three-dimensional Structures of Bulge-containing DNA Fragments

Leemor Joshua-Tor[1]†, Felix Frolow[1], Ettore Appella[2], Håkon Hope[1]‡ Dov Rabinovich[1] and Joel L. Sussman[1]§

[1]*Department of Structural Biology*
The Weizmann Institute of Science
Rehovot 76100, Israel

[2]*Laboratory of Cell Biology*
National Cancer Institute
National Institutes of Health
Bethesda, MD 20205, U.S.A.

(Received 30 August 1991; accepted 14 January 1992)

The three-dimensional structure of a DNA tridecamer d(CGCAGAATTCGCG)$_2$ containing bulged adenine bases was determined by single crystal X-ray diffraction methods, at 120 K, to 2·6 Å resolution. The structure is a *B*-DNA type double helix with a single duplex in the asymmetric unit. One of the bulged adenine bases loops out from the double helix, while the other stacks in to it. This is in contrast to our preliminary finding, which indicated that both adenine bases were looped out. This revised model was confirmed by the use of a covalently bound heavy-atom derivative. The conformation of the looped-out bulge hardly disrupts base stacking interactions of the bases flanking it. This is achieved by the backbone making a "loop-the-loop" curve with the extra adenine flipping over with respect to the other nucleotides in the strand. The looped-out base intercalates into the stacked-in bulge site of a symmetrically related duplex. The looped-out and stacked-in bases form an A·A reversed Hoogsteen base-pair that stacks between the surrounding base-pairs, thus stabilizing both bulges. The double helix is frayed at one end with the two "melted" bases participating in intermolecular interactions. A related structure, of the same tridecamer, after soaking the crystals with proflavin, was determined to 3·2 Å resolution. The main features of this *B*-DNA duplex are basically similar to the native tridecamer but differ in detail especially in the conformation of the bulged-out base. Accommodation of a large perturbation such as that described here with minimal disruption of the double helix shows both the flexibility and resiliency of the DNA molecule.

Keywords: DNA structure; X-ray diffraction; bulged bases; frameshift mutation

1. Introduction

Bulged nucleotides can occur in double helical stretches of both DNA and RNA, and may have profound biological consequences. For example, insertion or deletion of bases in coding regions of DNA can cause a shift in the translational reading frame. This mutation, termed a frameshift, generally produces a protein with an incorrect sequence following the site of mutation, usually resulting in a non-functioning gene product. Frameshifts that result from unpaired nucleotides can arise from recombination processes or from displacement of bases during replication (Streisinger *et al.*, 1966), e.g. when the template strand contains an unpaired base, the progeny strand will contain a deletion; conversely, when the non-template strand has an extra base, subsequent daughter strands will contain an insertion. These types of mutations occur more frequently at runs of identical bases, because correct base-pairs can form distal to the mutation

†Present address: Division of Chemistry, California Institute of Technology, Pasadena, CA 91125, U.S.A.
‡Permanent address: Department of Chemistry University of California, Davis, CA 95616, U.S.A.
§Author to whom all correspondence should be addressed.

0022-2836/92/100397-35 $03.00/0

© 1992 Academic Press Limited

Table 1
Bulges in DNA; n.m.r. results

DNA sequence	Bulged nucleotide	Conformation of the bulge	Reference
(5′CGCA**A**GAATTCGCG3′)$_2$	A	Stacked-in	Patel *et al.*, 1982; Gorenstein *et al.*, 1987; Nikonowicz *et al.*, 1989, 1990
(5′CGCAGAGCTCGCG3′)$_2$	A	Stacked-in	Hare *et al.*, 1986
(5′CGCGAAATTT**A**CGCG3′)$_2$	A	Stacked-in	Roy *et al.*, 1987
(5′CCGGAATTC**A**CGG3′)$_2$	A	Stacked-in	Kalnik *et al.*, 1989*a*
(5′CCGA**G**AATTCCGG3′)$_2$	A	Stacked-in	Kalnik *et al.*, 1989*a*
5′CTG**A**CCCATC3′ 3′GAC–GGGTAG5′	A	Stacked-in	Woodson & Crothers, 1989
5′CTGC**G**CCATC3′ 3′GACG–GGTAG5′	G	Stacked-in	Woodson & Crothers, 1988*b*,*c*
5′CAAA**C**AAAG3′ 3′TTT–TTTC5′	C	Looped-out	Morden *et al.*, 1983
(5′CCG**C**GAATTCCGG3′)$_2$	C	Looped-out/stacked-in†	Kalnik *et al.*, 1989*b*
5′CTGGT**G**CGG3′ 3′GACC–CGCC5′	T	Stacked-in	van den Hoogen *et al.*, 1988*b*
(5′CCGGAATTC**T**CGG3′)$_2$	T	Looped-out	Kalnik *et al.*, 1990
(5′CCG**T**GAATTCCGG3′)$_2$	T	Looped-out/stacked-in†	Kalnik *et al.*, 1990

† Depending on temperature.

site stabilizing the bulged duplex region. All frameshifts of this type are fixed and propagate to future generations. Alternative mechanisms for frameshift mutations have been proposed and reviewed (Kunkel, 1990; Ripley, 1990). With the advent of better *in vitro* enzymatic assays, and more powerful physical methods to study molecules that play a role in frameshift mutations, there has been a renewed effort to understand this process in greater detail.

Bulges have been observed far more frequently in RNA than in DNA. They are likely to play a key role in RNA tertiary structure and have been shown to be important in protein–RNA interactions (Milligan & Uhlenbeck, 1989; Witherell & Uhlenbeck, 1989; Wu & Uhlenbeck, 1987). It has been suggested that in nuclear mRNA splicing, the branchpoint nucleotide, which carries out nucleophilic attack on the 5′ splice site in certain classes of introns (Guthrie, 1991), is a bulge in models containing both intermolecular (Parker *et al.*, 1987) and intramolecular (group II) (Michel *et al.*, 1989; Cech, 1990) helices.

In comparison to the structural characterization of single base mismatches in double-helical DNA (Kennard & Hunter, 1989; Shakked & Rabinovich, 1986), relatively little is known about the structure of a bulged, non-complementary base within the DNA duplex. The effect of a bulged nucleotide on the helix could result in several alternative secondary structures. An extra base could stack into the double helix or loop out into solution, keeping all the other base-pairs intact, or conceivably could cause a misalignment in the vicinity of the extra base. The existence of unpaired (or extrahelical) bases in nucleic acids was first postulated by the pioneering work of Fresco & Alberts (1960). They introduced the notion that non-complementary residues in helical polynucleotides may be looped out, i.e. unstacked and exposed to solvent, while the rest of the nucleotides remain in a normal Watson–Crick type double helix. Model building on their part showed that looping out of a uridine residue is readily achieved by rotation about its two adjacent phosphodiester bonds. The only significant alteration to the DNA structure would be a reduction in the distance between the phosphate groups around the loop. In addition, loops of more than one residue can be constructed. A reduction in thermal stability in these bulged helices, relative to that of the perfect analog, is observed (Lomant & Fresco, 1975) and is proportional to the mole fraction of non-complementary residues.

Other physical studies on bulge-containing RNAs have been reported, including hybridization and c.d.† (Fink & Crothers, 1972*b*), photohydration (Lomant & Fresco, 1973), drug binding studies (White & Draper, 1987) and n.m.r. (Hartel *et al.*, 1981; Lee & Tinoco, 1980; van den Hoogen *et al.*, 1988*a*), which all demonstrate the existence of looped-out bulges. Thermodynamic studies were done in an attempt to measure the destabilization of the RNA double helix by a bulge defect (Fink & Crothers, 1972*a*) and the effect of nearby sequences on the stability (Lomant & Fresco, 1975; Longfellow *et al.*, 1990). c.d. measurements (Gray *et al.*, 1980), mixing curve analyses and nuclease digestion (Evans & Morgan, 1982, 1986) have shown the existence of unpaired bases in DNA double helices. Photodimerization of pyrimidines flanking unpaired thymidine bases is a strong indication that the unpaired thymidine bases are in a looped-out

†Abbreviations used: c.d., circular dichroism; n.m.r., nuclear magnetic resonance; u.v., ultraviolet; F_{obs}, observed structure factor; F_{calc}, calculated structure factor; r.m.s., root-mean-square; MPD, 2-methyl-2,4-pentanediol.

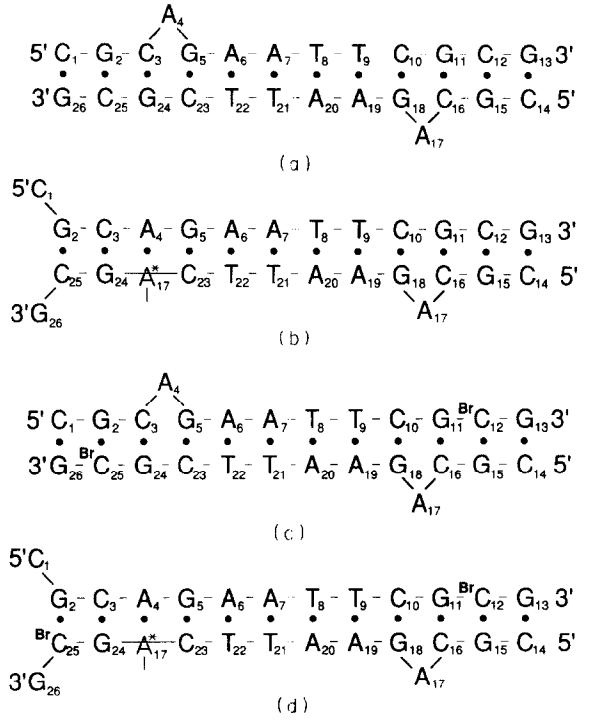


Figure 1. A representation of 2 possible models of the secondary structure of the DNA tridecamer d(CGCAGAATTCGCG)$_2$ and its 5-BrC derivative. (a) and (c) The fully looped-out model for the native and 5-BrC derivative, respectively; (b) and (d) the stacked-in/looped-out model, respectively. The distance between the bromine atoms is 1 base-pair further apart in the stacked-in/looped-out model relative to the looped-out model.

conformation. Thermodynamic data for bulges in an A·T tract indicate that the identity of the bulge and its sequence environment may influence the magnitude of the duplex destabilization (LeBlanc & Morden, 1991). n.m.r. structural analyses of bulge-containing oligodeoxyribonucleotides are summarized in Table 1. Based on the above physical studies, and gel-mobility studies (Bhattacharyya & Lilley, 1989; Hsieh & Griffith, 1989; Rice & Crothers, 1989), certain trends have emerged. Purines appear to stack into the double helix, causing it to kink. Pyrimidines, on the other hand, both bulge out from and stack into the double helix, depending on temperature and sequence, i.e. flanking bases as well as bases that are not adjacent to the bulge. Sequence context does not seem to have an effect on the conformation of extra purines (see Table 1). A detailed three-dimensional X-ray structure of bulged DNA fragments may provide an understanding of interactions involved in producing these aberrations in the DNA molecule, and how they may be detected and repaired *in vivo*, once they are formed.

We have reported on the structure of a DNA tridecamer d(CGCAGAATTCGCG)$_2$ (Fig. 1(a)) containing bulged adenine bases (Joshua-Tor *et al.*, 1988). Here, we describe the refinement of this sequence and the structure determination and refinement of this same fragment after crystals were soaked in a solution of proflavin. The refinement of the structure of the soaked crystals led to a correction of our original model, which we have confirmed by studies with a covalently bound heavy-atom derivative. We will discuss the conformation of the bulged region in detail, as seen by X-ray diffraction, its effect on the structure of the double helix, the interactions involved in stabilizing its conformation and a comparison of these features with those of the proflavin-soaked tridecamer structure.

2. Materials and Methods

(a) *Crystallization of the oligonucleotides*

The DNA oligomers were synthesized and purified as described (Saper *et al.*, 1986). The oligonucleotides were crystallized by the vapor diffusion sitting drop method (McPherson, 1985) on siliconized Corning depression glass plates. Each droplet contained 2·4 mg oligonucleotide/ml, 10 mM-sodium cacodylate (pH 7·0), 10% (v/v) 2-methyl-2,4-pentanediol (MPD), 15 mM-$MgCl_2$ and 1·5 mM-spermine·HCl for the native 13-mer d(CGCAGAATTCGCG), or 0·9 mM-spermine·HCl for the 12-Br-13-mer d(CGCGAATTCG^{Br}CG). The drops were equilibrated at 4 °C against 0·5 ml of a reservoir solution that contained 25% MPD and 0·2 M-NaCl in water for both the 13-mer and the 12-Br-13-mer. Native crystals grew typically to approximately 0·8 mm × 0·3 mm × 0·3 mm within 10 to 14 days as elongated triangular prisms; however, the shorter more spherical crystals proved to diffract better. Crystals of 12-Br-13-mer were of the same shape.

(b) *Soaking of crystals and determination of the ratio proflavin : duplex*

Crystals of the native 13-mer were soaked in a solution of 3,6 diaminoacridine (proflavin) hemisulfate salt (Sigma P-2508) at 4 °C. Crystals that acquired a yellow color due to proflavin were then washed several times in mother liquor in order to remove excess DNA and proflavin. They were then dissolved and the u.v. spectrum of this solution and of solutions of the oligonucleotide and of proflavin separately were taken using a Hewlett Packard diode array spectrophotometer model 8450A. Ratios were determined for 2 different soaks of proflavin, a longer soak and a shorter, less concentrated soak. The longer soak was done using a 1·2 mM-solution for about 18 h and the shorter soak in a 0·6 mM-solution for 2 h. The 2 h soak was used for X-ray data collection.

(c) *Data collection and crystal data*

Cryotemperature data collection techniques developed by Hope (1988) were used for data collection as all crystals appeared to be highly sensitive to X-ray irradiation even at 4 °C. In this manner, complete data sets were collected, for each experiment, using a single crystal, with virtually no decay, at approx. 120 K.

(i) *Native 13-mer*

A single crystal was coated with a viscous oil (Exxon Paratone-N with no additives), removing all the mother liquor solution covering the crystal. It was then picked up with a thin glass fiber, which was fitted into a copper pin and transferred to the goniometer head mounted on a Rigaku AFC5-R rotating anode diffractometer operated at 15 kW. The crystal was placed directly in a cold stream of freshly boiled liquid N_2. The cooling system was a basic

Table 2

Completeness of data

A. *As a function of resolution for the native tridecamer*

Shell (Å)	Possible	Observed†	% Observed	$>3\sigma$ (shell)	% $>3\sigma$ (shell)	$>3\sigma$ (sphere)	% $>3\sigma$ (sphere)
10·0-8·0	47	45	96	45	96	45	96
8·0-6·0	127	123	97	119	94	164	94
6·0-4·0	553	512	93	455	82	619	85
4·0-3·0	1034	964	93	695	67	1314	75
3·0-2·8	409	383	94	168	41	1482	68
2·8-2·6	573	520	91	158	28	1640	60
2·6-2·0	3328	2929	88	516	16	2156	36

B. *As a function of resolution for the proflavin-soaked tridecamer*

Shell (Å)	Possible	Observed†	% Observed	$>2\sigma$ (shell)	% $>2\sigma$ (shell)	$>2\sigma$ (sphere)	% $>2\sigma$ (sphere)
10·0-8·0	47	46	98	45	96	45	96
8·0-6·0	150	138	92	138	92	183	93
6·0-4·0	576	515	89	500	87	683	88
4·0-3·2	774	632	82	571	74	1254	81
3·2-3·0	355	242	68	193	54	1447	76
3·0-2·8	441	254	56	171	39	1618	69
2·8-2·5	571	266	47	133	23	1751	60

† Positive reflections.

Table 3

Completeness of data as a function of resolution for 12-Br-13-mer

Shell (Å)	Possible	Obs† (1)	% Obs (1)	Obs (2)	%Obs (2)	$>4\sigma$ merge (shell)	% $>4\sigma$ merge (shell)	$>4\sigma$ merge (sphere)	% $>4\sigma$ merge (sphere)
10.0-8.0	47	48	100	48	100	39	83	39	83
8.0-6.0	130	128	98	128	98	87	67	126	71
6.0-4.0	564	533	95	533	95	220	39	346	47
4.0-3.0	1049	457	44	983	94	301	29	647	36

The space group for this crystal is actually $P21$ although it has pseudo $C2$ symmetry, since $h+k = 2n+1$ reflections are very weak. (1) and (2) denote the first and second data sets, respectively.

†Positive reflections.

Enraf Nonius low-temperature apparatus, fitted with a home-made outlet tube of evacuated glass tubing and styrofoam, which was constructed to allow relatively free motion of the χ circle while keeping the cold gas stream at a fixed position with respect to the crystal. Data were collected with CuKα radiation at 120 K using an ω scan at a speed of 5°/minute to 2·0 Å resolution (1 Å = 0·1 nm). In all, 6792 reflections were collected, of these 5530 are unique, R_{sym}† for all positive data to 2·0 Å is 0·048. The space group is $C2$ with unit cell dimensions $a = 78{\cdot}48$ Å, $b = 42{\cdot}84$ Å, $c = 25.16$ Å, $\beta = 99{\cdot}36°$ with 1 DNA duplex/asymmetric unit. The number of reflections at different resolution ranges is shown in Table 2.

(ii) *12-Br-13-mer*

Data were collected using the same shock-cooling technique described for the native 13-mer, except that the crystal was mounted on a glass spatula. The spatula was made by attaching a very thin piece of glass of approximately the size and shape of the crystal to be measured, to a thin glass fiber as described above. Two data sets, from the same crystal, were collected on a Rigaku AFC5-R rotating anode diffractometer operating at 10 kW using CuKα radiation and an ω scan to 3·5 Å resolution at two different speeds (4°/min and 2°/min). In all, 1274 and 2352 reflections were collected, respectively, in this case. 1217 and 2234 are unique. The space group is $P2_1$ and the unit cell dimensions $a = 78{\cdot}67$ Å, $b = 42{\cdot}99$ Å, $c = 25{\cdot}09$ Å, $\beta = 100{\cdot}70°$ with 2 DNA duplexes/asymmetric unit with pseudo 2-fold symmetry between them (resulting in this space group being pseudo $C2$ and virtually isomorphous to the native crystals). Table 3 shows the number of reflections at different resolution ranges.

(iii) *Proflavin-soaked 13-mer*

The shock-cooling techniques were used for the proflavin-soaked crystals. They were coated with oil as described above and mounted on a glass spatula. Only the shorter soaked crystals were useful for data collection as the longer soaked crystals diffracted poorly. Data were collected on a multiwire Siemens/Xentronics area detector mounted on a Rigaku rotating anode operating at 10 kW to 2·5 Å resolution (360 frames, the framewidth being 0·5° and exposure was 300 s/frame). In all, 8318 reflections were collected, of these 2608 were unique reflections, R_{sym} for all data to 2·5 Å being 0·021. Significant data, based on counting statistics, were obtained only to 3·2 Å resolution (as assessed by the program XENGEN (Howard *et al.*, 1987)). The space group is $C2$ with unit cell dimensions $a = 80{\cdot}81$ Å, $b = 42{\cdot}97$ Å, $c = 25{\cdot}57$ Å, $\beta = 99{\cdot}6°$. Table 2

†R_{sym} is the unweighted squared R-factor on intensity.

shows the number of reflections at different resolution ranges.

(d) *Structure solution and model building*

The structure was solved *via* ULTIMA (Rabinovich & Shakked, 1984) as described by Joshua-Tor *et al.* (1988). An alternative search procedure was conducted in order to verify the "spin" of the molecule around the helix axis based on the orientation and position of the helix axis from ULTIMA. The molecule was rotated around its helix axis in 5° steps and the molecule subjected to rigid body refinement at each step using the program CORELS (Sussman *et al.*, 1977). The *R*-factor† and linear correlation coefficient‡ were then plotted against the helix spin angle, at various resolution ranges (20 to 10 Å; 15 to 8 Å; 15 to 6 Å). The results agreed with those of ULTIMA.

The initial model was a *B*-DNA dodecamer of the sequence d(CGCGAATTCGCG)$_2$ built from *B*-DNA helix parameters based on fiber diffraction studies (Arnott *et al.*, 1980). Fitting of the model to the electron density maps was done initially on a Vector General black and white computer graphics display and later on an Evans & Sutherland PS390 color graphics system using the interactive computer graphics program FRODO (Jones, 1978; Pflugrath *et al.*, 1984). A dictionary of optimal bond lengths, bond angles and fixed torsion angles for nucleic acids was used in order to be able to interactively add extra bases *via* the SAM and REFI options in FRODO. This dictionary was modified by us based on fiber diffraction data (Arnott *et al.*, 1980) for DNA structures and acridine dyes.

(e) *Refinement techniques*

Several different methods were used during the course of refinement of the native and the proflavin-soaked 13-mers. Initially, the constrained-restrained least-squares (CORELS) program (Sussman *et al.*, 1977) was used. This was followed by the Hendrickson–Konnert restrained least-squares refinement program (Hendrickson & Konnert, 1981) as modified by Westhof for nucleic acids, NUCLSQ (Westhof *et al.*, 1985). As the program X-PLOR (Brünger *et al.*, 1987) became available, we used it mainly for refinement of the looped-out/stacked-in model together with NUCLSQ (see Results and Discussion for details of the refinement). All computations were carried out on a VAX 11/780 or a MicroVAX 3600 computer, except for X-PLOR refinement, which was run on a CONVEX C220 computer.

(f) *Difference Fouriers for determining bromine atom positions*

Two 12-Br-13-mer data sets using $F>4{\cdot}0\sigma$ reflections between 10 and 4 Å resolution were merged using the program PROTEIN (Steigemann, 1974). PROTEIN was also used in order to calculate a difference Fourier map $(F(\mathrm{Br})-F(\mathrm{Nat}))\alpha_{\mathrm{mod}}$, where $F(\mathrm{Br})$ and $F(\mathrm{Nat})$ are the derivative and native structure factors, respectively, using phases (α_{mod}) once from a fully looped-out model (model A) and once from the stacked-in/looped-out model (model B: see Results and Discussion). This calculation was repeated with the CCP4 package of programs (from the Daresbury Laboratory, U.K.) yielding very similar results.

(g) *Calculation of helix parameters*

Most of the helix parameters were calculated with the NEWHEL90 library of programs (R. E. Dickerson, personal communication). Conventions and definitions of helix parameters used are those of the 1989 Cambridge conventions (Dickerson *et al.* 1989), described in detail by Dickerson *et al.* (1985). The frayed bases were omitted. Parameters for the A·A base-pair were calculated using the appropriate looped-out adenine (A17) from the proper symmetry-related molecule, which is marked A17*. The overall helix axis was calculated using all $C_{(1')}$ to $C_{(1')}$ vectors along each strand except from A17* and all purine-$N_{(9)}$ and pyrimidine-$N_{(1)}$ vectors except for A17*, for which $C_{(2)}$ was used instead, because of the special geometry of the base-pair. The long axis for each base-pair was taken as the $C_{(8)}$–$C_{(6)}$ vector and for A4·A17* as $C_{(8)}$ $N_{(3)}$. Comparisons of helix parameters were done with the Dickerson–Drew native room temperature dodecamer and not with the low-temperature (16 K) dodecamer structure. The 2 crystal structures are quite similar but the room temperature structure was refined at higher resolution and therefore parameters are more accurately determined.

Cylindrical projections of the DNA structure were done using a program written by T. Larsen (personal communication) with an assumed cylindrical radius of 10 Å.

3. Results and Discussion

(a) *Refinement of the DNA tridecamer structure*

The refinement of the DNA tridecamer structure was carried out initially as described by Joshua-Tor *et al.* (1988). Subsequently, a more careful refinement was initiated, which used additional and more restrictive restraints on the geometry in order to obtain a better fit to canonical stereochemistry. The guanine base, G24, opposite one of the bulges was built in a *syn* conformation in order to fit the electron density maps. The program NUCLSQ (Westhof *et al.*, 1985) was used for least-squares refinement alternating with molecular model building using FRODO (Jones, 1978; Pflugrath *et al.*, 1984) into electron density maps on an Evans & Sutherland PS390 computer graphics system.

At this stage of the refinement, electron density maps looked satisfactory with a reasonably good *R*-factor and stereochemistry (23·0% for 10 to 2·6 Å, $F_{\mathrm{obs}}>0\sigma$, r.m.s. deviation from ideality of 0·016 Å for bond lengths with 17 water molecules). A portion of the electron density map, in the vicinity of one of the bulges (A17) is shown in Figure 2(a) and (b). Two questionable features about the structure determination still remained based on this model. First, the electron density between the sugar moiety of one of the unpaired bases (A4) and the phosphate group of the residue 3′ to it (G5) was inadequate to fit all the atoms (Fig. 2(c)), although an omit map (Fig. 2(d)) of this bulged base indicated that the sugar was positioned correctly. The two phosphate groups around the unpaired base as well as its sugar moiety, were too close, resulting in very tight non-bonded contacts.

† $R=\dfrac{\Sigma\,||F_{\mathrm{obs}}|-|F_{\mathrm{calc}}||}{\Sigma|F_{\mathrm{obs}}|}$.

‡ $cc=\dfrac{n\Sigma F_{\mathrm{obs}}F_{\mathrm{calc}}-\Sigma F_{\mathrm{obs}}\Sigma F_{\mathrm{calc}}}{\sqrt{n\Sigma F_{\mathrm{obs}}^2-(\Sigma F_{\mathrm{obs}})^2}\sqrt{n\Sigma F_{\mathrm{calc}}^2-(\Sigma F_{\mathrm{calc}})^2}}$.

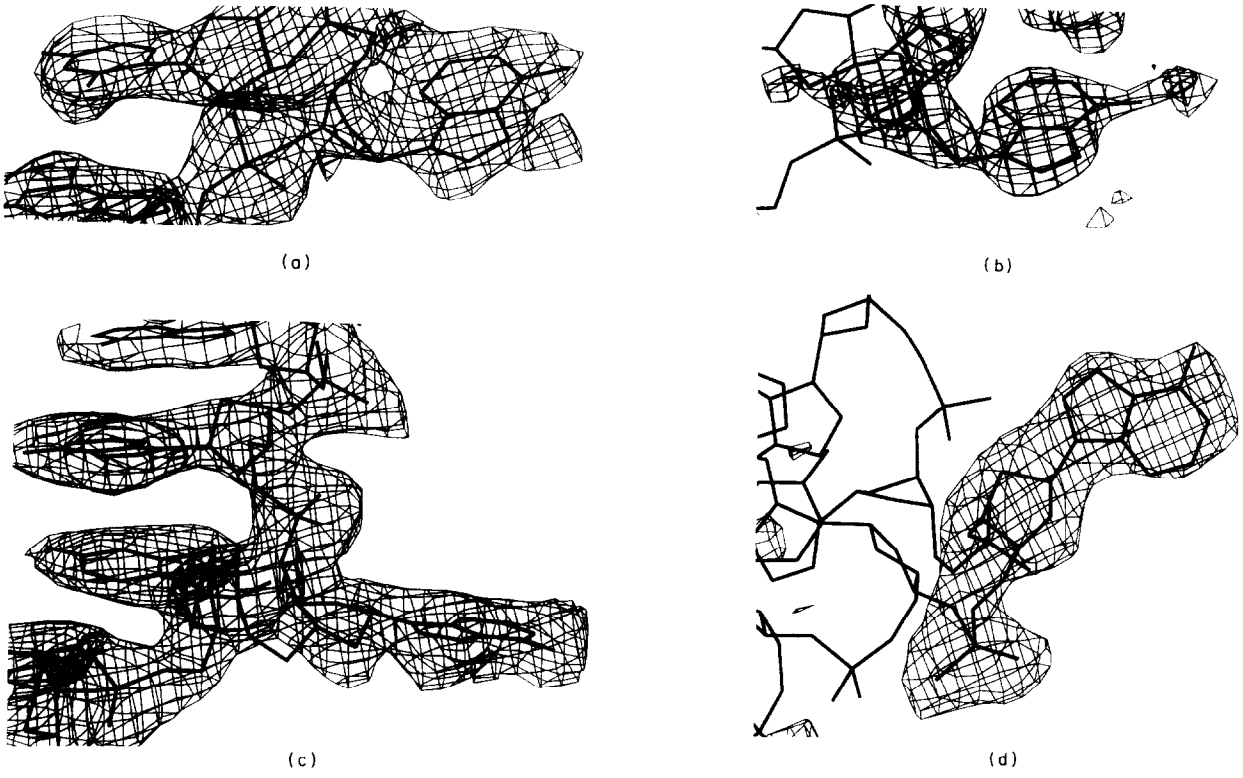


Figure 2. (a) Electron density map ($2F_{obs}-F_{calc}$) around one of the bulges (A17) in the looped-out model (lowest contour level 1·2σ); (b) difference Fourier ($F_{obs}-F_{calc}$) map in which A17 was omitted in the computation of the calculated structure factors of the looped-out model (lowest contour level 1·5σ); (c) electron density map ($2F_{obs}-F_{calc}$) around A4 in the looped-out model (lowest contour level 1·0σ). Note that not all of the backbone atoms fit into the map and that the phosphate groups flanking the bulge are very close; (d) omit map for A4 in the looped-out model (lowest contour level 1·5σ).

We were unable to refine this region so that it would have both acceptable geometry and, simultaneously, have all atoms fit the electron density maps. Second, there was quite a large unaccountable peak, in the minor groove near the first nucleotide, C1, in both the $2F_{obs}-F_{calc}$ map and in the difference map.

Proflavin-soaked crystals were subjected to similar experimental protocols, i.e. X-ray data were collected at 120 K. Starting from the looped-out model of the tridecamer, we expected that the structure determination of the proflavin-soaked crystals to be straightforward, which turned out not to be the case. The looped-out model of the tridecamer was initially converted into fractional co-ordinates because the unit cell of the soaked crystals was somewhat larger than those of the native cell (7·6% v/v). It was then subjected to rigid body refinement using the program CORELS (Sussman *et al.*, 1977) at 10 to 6 Å. The *R*-factor dropped from 36·3% to 30·4% within five cycles of refinement. Using simulated annealing crystallographic refinement with the program X-PLOR at 10 to 3 Å caused the *R*-factor to drop from 43·3% to 30·3% and resulted in an "expansion" of the structure towards the crystallographic 2-fold at ($\frac{1}{2}$,0,$\frac{1}{2}$). Inspection of different electron density maps revealed several severe problems with the model used, especially in the region of one of the extra bases, A17, and around G24. The structure was then refined by first omitting G24 and then omitting A4 (the other extra adenosine base). Maps based on these refinements showed no density for A4 or G24. The other unpaired base, A17, seemed to be interacting with C3, and C3 and G5 appeared to be connected through electron density. This led to substantial rebuilding of parts of the model such that A4 was inserted into the double helix instead of C3, C3 replaced G2, G2 replaced C1, and C1 was removed at this stage. On the opposite strand, A17 from a neighboring molecule (designated A17*) was modelled to occupy the density that previously was used by G24, thus bringing it into a position to pair with A4. The backbone between C23 and G24 was extended around this extra adenine base (A17*), G24 replaced C25, C25 replaced G26, and G26 was removed. Thus, there was no room for the terminal base-pair C1·G26, since if it were paired it would "bump" into the same base-pair of a 2-fold related neighboring molecule. As in the previous fully looped-out model, there are still 12 base-pairs, except that one of them arises from the pairing of the two extra bases. This suggested that the terminal base-pair must be frayed. A large and initially unexplained peak in the $2F_{obs}-F_{calc}$ and in the $F_{obs}-F_{calc}$ maps extending from the phosphate group of G2 revealed the position of the first nucleotide of one strand. Density for the G26 phosphate group was also very apparent and accounted for the unaccounted for peak previously described for the fully looped-out model. Another large peak was

located in the minor groove of a neighboring molecule and was subsequently identified as the sugar and base of nucleotide G26. At first, C1 and G26 were not included in the refinement in order to minimize bias in the electron density maps, since this fraying of the terminal base-pair C1·G26 seemed rather unusual. Later, after refinement was continued using X-PLOR, and density became clearer, these two nucleotides were included.

Thus, refinement of this proflavin-soaked tridecamer structure led to a different crystal structure than our looped-out model of the first DNA tridecamer structure, although no proflavin molecule could be located. In the soaked crystals, one extra adenine base is looped-out from the double helix, while the other stacks into the double helix. The looped-out adenine base forms an A·A base-pair with the stacked-in adenine base from a neighboring molecule. The result is a stacked-in/looped-out structure rather than a fully looped-out structure. Both are shown schematically in Figure 1(a) and (b) for comparison. Such drastic conformational changes are unlikely to occur in crystals without cracking them, so we decided to examine how well this stacked-in/looped-out structure fit the original native X-ray data. This was initially done by subjecting both the fully looped-out model A and the stacked-in/looped-out model B to the same refinement scheme, i.e. simulated annealing and restrained least-squares using the program X-PLOR (Brünger *et al.*, 1987) using reflections with $F_{obs} > 3\sigma$ and 10 to 2·6 Å resolution. The *R*-factor dropped from 28·9% to 25·6% for model A and from 41·5% to 23·3% for model B, with somewhat better stereochemistry for model B. Various electron density and omit maps, calculated to assess the results of these refinements, all showed the superiority of model B. Although the *R*-factor and the electron density maps indicated a solution that favored model B, both *R*-factors were sufficiently low so as not to reject the original model A at this point. Confirmation of the correct structure, using a bromine derivative is discussed in the next section. Refinement of a "grossly wrong model" to a reasonably low *R*-factor has been reported for a DNA decamer structure (Heinemann & Alings, 1991). Here, it should be added that the two models are very similar throughout a significant portion of the structure, the central part being virtually identical. Based on these results, model B was confirmed as the correct structure for the tridecamer, and refinement was continued.

A simulated annealing procedure was used for the refinement of model B against the native X-ray data. This included a preparation stage, which uses diffraction data together with energy minimization, followed by heating to 3000 K and running molecular dynamics for 3 picoseconds with a timestep of 0·5 femtoseconds. The system was then cooled to 300 K and running molecular dynamics for 1·5 picoseconds (the so-called fast cooling procedure). Charges were turned off during molecular dynamics. Following this, several cycles of restrained energy minimization were run once again. During the early stages of the X-PLOR refinement, backbone torsion angles were weakly restrained in order to drive them to acceptable values for *B*-DNA; however, following this, these restraints were excluded for the rest of the refinement. Hydrogen bonds were not restrained. Towards the end of the refinement and after water molecules were added, the molecular dynamics steps were excluded from the protocol, i.e. only the least-squares-energy minimization refinement protocol was used. Geometry was tightened at times by decreasing the weights for the structure factors (WA in X-PLOR) and a weighting scheme of the form $wt = \{90\text{-}900(\sin\theta/\lambda\text{-}0\text{·}16667)\}^{-2}$ was applied at later stages of the refinement to give approximately equal weights to ΔF^2 as a function of $\sin\theta/\lambda$ ranges. Since the dictionary of ideal bond lengths, bond angles and torsion angles used for X-PLOR and NUCLSQ proved to be considerably different, especially with respect to sugar geometry, both dictionaries were modified on the basis of average values obtained from well-refined crystal structures (Saenger, 1984). Other values of the parameter file used for X-PLOR were adjusted in order to obtain proper geometry. A few cycles of NUCLSQ were run at the last stage of refinement using the weighting scheme described above. The modified dictionary was used and a 0·35 parameter shift per cycle was employed. No torsion angle or base-pair hydrogen bond was restrained.

During all stages of the refinement, electron density maps were calculated and the refined models examined to follow the course of the refinement and to rebuild portions of the molecule where needed. The electron density maps were of the type $2F_{obs} - F_{calc}$, $F_{obs} - F_{calc}$, $F_{calc} - F_{obs}$ and omit maps, in which the region of the model to be examined, was omitted from the computation of the calculated structure factors. Omit-refine maps were used, in which several cycles of least-squares refinement were run after omitting the part of the structure being examined. Water molecules were added in a conservative manner, i.e. only peaks that were above a 3σ cutoff in the difference map, and that were between 2·4 and 3·5 Å (distances up to 4·0 Å were sometimes considered) from acceptor/donor atoms were identified as putative solvent molecules. Solvent molecules were continuously examined by omit maps and were deleted if they no longer fit the criteria after several cycles of refinement. Occupancies were not refined, i.e. they were kept at a value of 1·0 during refinement so that *B* values could indicate reliability, apart from one solvent sitting on a 2-fold axis in which the occupancy was set to 0·5. Individual isotropic temperature factors were refined during the later stages of refinement. These were allowed to drop to a minimum value of 2·0 Å^2. Average group *B* values for phosphate groups, sugar molecules and bases are given in Table 4. The final *R*-factor for the native tridecamer structure is 18·6% for 1640 reflections in the range 10 to 2·6 Å (see Table 5).

Table 4

Average group B-factors (Å^2) for phosphate groups, sugars and bases of the tridecamer and those of the proflavin-soaked tridecamer

Residue	Phosphate		Sugar		Base	
	Native	Soaked	Native	Soaked	Native	Soaked
1	33·2†	36·8†	28·7	35·5	25·0	37·5
2	34·5	34·3	14·6	12·0	5·5	2·0
3	13·4	16·7	5·0	8·3	2·3	2·0
4	21·3	15·3	18·5	13·0	5·5	2·0
5	12·4	16·6	14·0	12·0	4·3	2·0
6	11·7	19·2	14·8	12·8	13·4	2·3
7	36·6	19·8	33·7	17·6	23·0	2·7
8	39·0	29·0	14·6	15·1	16·8	2·3
9	26·1	27·6	19·0	19·3	5·0	2·5
10	18·9	26·4	20·2	19·4	8·7	4·0
11	10·7	24·3	25·7	18·8	21·0	2·9
12	17·9	29·6	13·9	25·4	17·7	11·9
13	21·0	25·9	22·2	19·5	26·7	13·1
14	26·3†	28·8†	20·6	22·6	15·1	24·6
15	13·7	26·5	26·2	31·4	24·2	14·5
16	33·0	35·3	38·1	23·3	32·1	2·2
17	28·9	25·7	20·2	16·3	3·9	2·1
18	24·5	21·8	14·5	13·4	4·8	2·2
19	15·1	23·6	17·6	16·6	16·5	2·4
20	26·6	22·5	23·2	16·3	9·1	2·0
21	7·8	17·7	21·6	15·4	15·9	2·0
22	22·0	29·4	12·4	16·2	3·5	2·0
23	18·4	25·5	10·2	13·6	5·7	2·0
24	16·3	17·0	14·3	8·0	2·0	2·0
25	31·5	19·6	7·2	12·9	8·7	2·0
26	9·5	20·2	8·2	21·4	12·2	16·6
Average	21·9	24·4	18·4	11·9	12·6	6·8

†These values are for $O_{(5')}$ since there are no 5′ phosphate groups.

The reason the *R*-factor for the previous model is lower than the corrected one appears to be due to several different reasons. Firstly, the resolution of the previous model was only to 2·8 Å resolution, while that of the corrected model is 2·6 Å resolution. Although a somewhat higher F_{obs}/σ cutoff was used in the current refinement, resulting in an improved quality of the electron density maps, there were still significantly more reflections used in the current refinement compared to the previous model (1640 versus 1113 reflections). Not enough care was taken with respect to the geometry of the previous model, especially in the backbone torsion angles, thus making it possible, in some sense, to more easily lower the r.m.s. deviation from ideality for the bond lengths and bond angles. This also may have resulted in lowering the *R*-factor. In addition, the inclusion of 50 water molecules may have both masked some of the errors in the model and further "artificially" lowered the *R*-factor. Both the geometry and the addition of water molecules was dealt with in a much more conservative and rigorous manner in the refinement of the current corrected structure. It should be noted, however, that there was no significant difference in the treatment of the temperature factors in both cases.

Table 5

Refinement results for the native tridecamer and for the proflavin-soaked tridecamer

	Native	Soaked
R-Factor (%)	18·6	22·2
Resolution (Å)	10–2·6	10–3·2
Cutoff	$F>3\sigma(F)$	$F>2\sigma(F)$
Number of reflections	1640	1254
Number of non-hydrogen DNA atoms	528	528
Number of water molecules	23	11
r.m.s. deviation from ideality:		
bond length distances (Å)	0·014	0·017
bond angle distances (Å)	0·019	0·025

(b) *Bromine atom positions distinguish between the looped-out* versus *the stacked-in/looped-out structures*

In order to verify experimentally that model B was correct, we used an essentially isomorphous derivative with 5-bromocytosine (5-BrC) replacing cytosine in the penultimate position (Joshua-Tor, 1991). As model B differs from model A by the presence of an extra A·A base-pair between the brominated cytosine bases, the 5-BrC bases on opposite strands in model B would be nine base-pairs apart, while in model A only eight (see Fig. 1(c) and (d)). By calculating difference electron density maps between the derivative and the native structure factors, using phases from model A and model B, the positions of the bromine atoms could be located (Fig. 3). Although the map calculated with model A phases has several additional noise peaks, both maps show the location of the bromine atoms clearly at the same positions in space, and

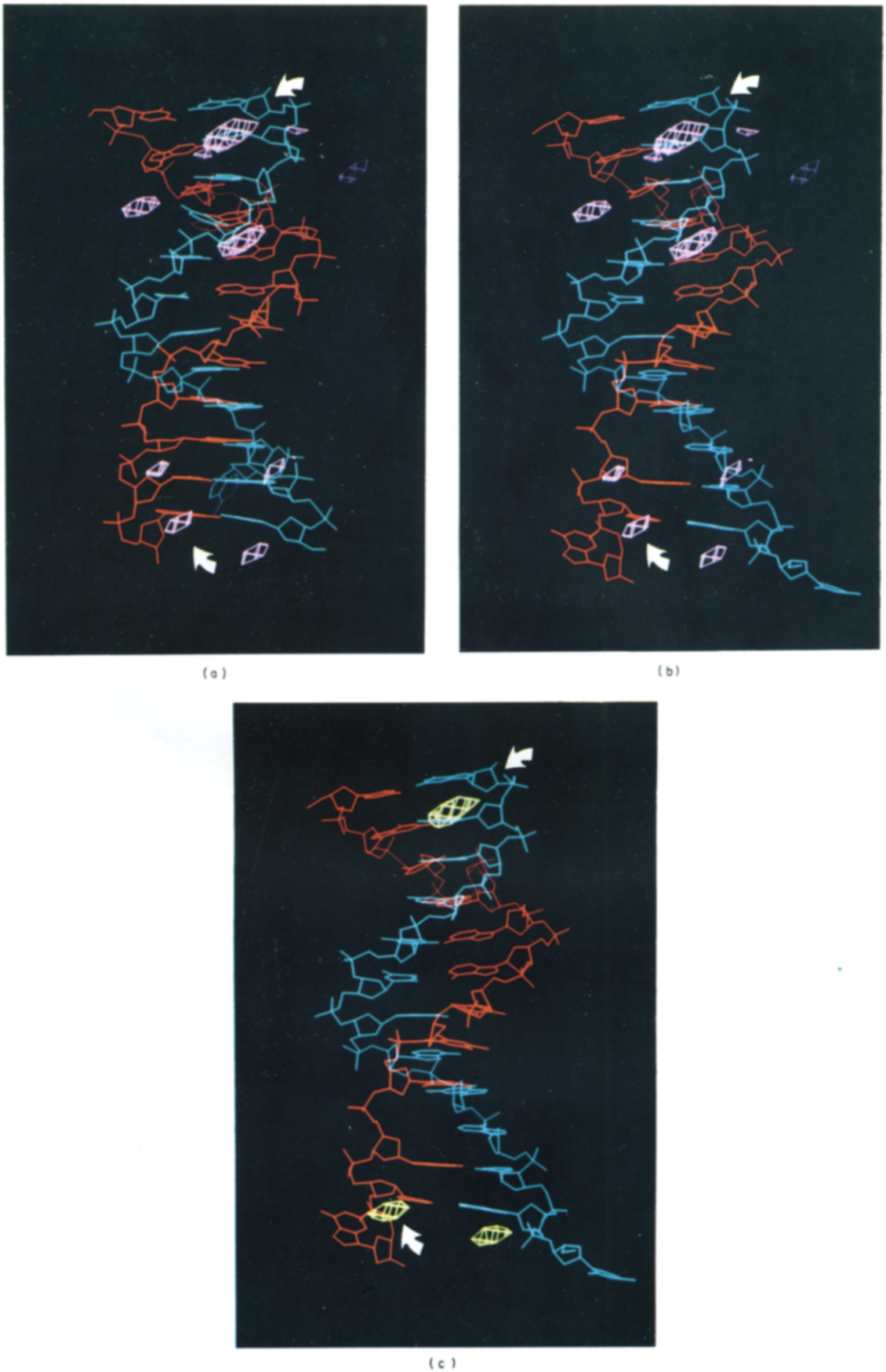
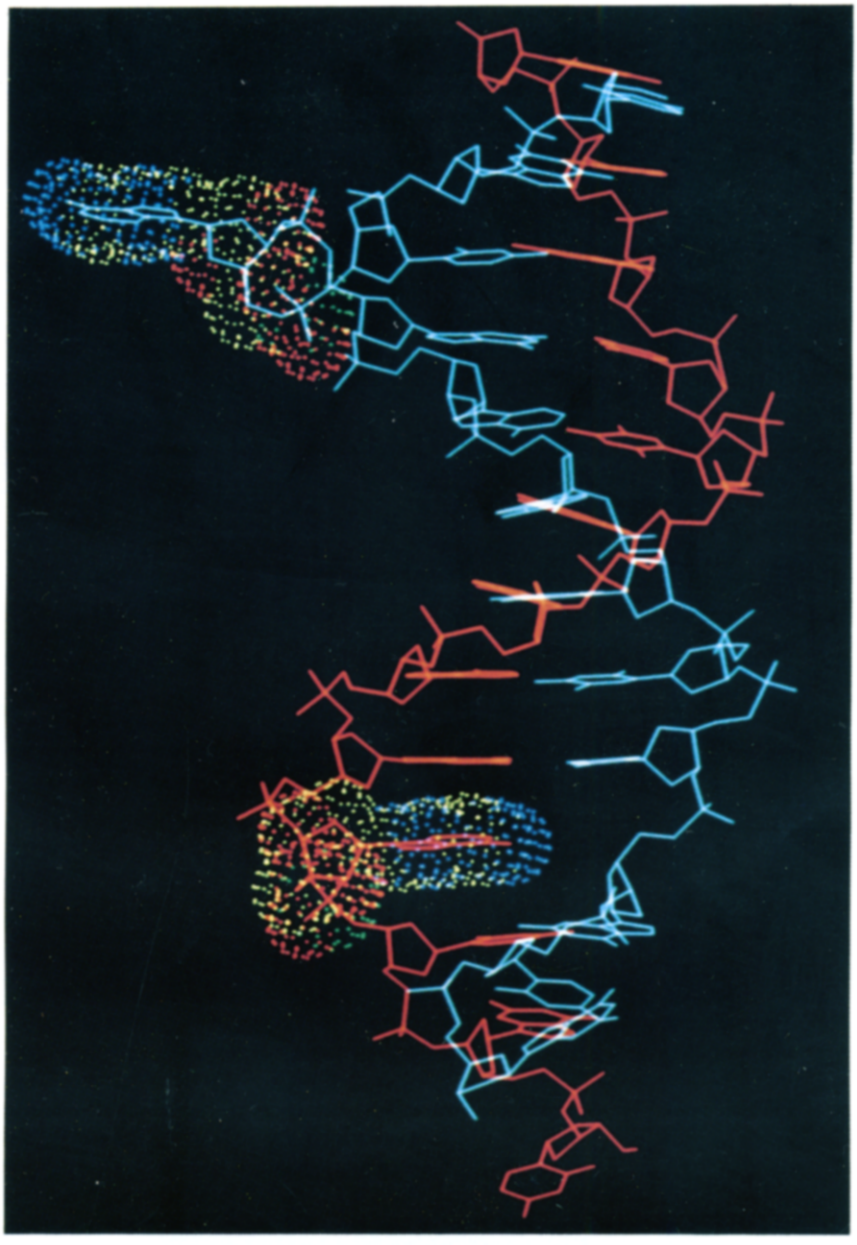


Figure 3. Difference maps with amplitudes (F(Br) − F(Nat)) showing the positions of the bromine atoms in the DNA tridecamer d(CGCAGAATTCG^{Br}CG)$_2$, which was synthesized and purified as described (Saper *et al.*, 1986). (a) Map calculated using phases from model A (the fully looped-out model, lowest contour level at 2·7σ) superimposed on model A; (b) same map superimposed on model B (the stacked-in/looped-out model); (c) map calculated using phases from model B superimposed on model B (lowest contour level at 3·3σ). The 3rd peak in this map arises from the bromine on a 2-fold related C25 base and appears in the map shown in (a) and (b) as well. Arrows indicate the 2 bromine peaks from 1 asymmetric unit.



(a)

Fig. 4.

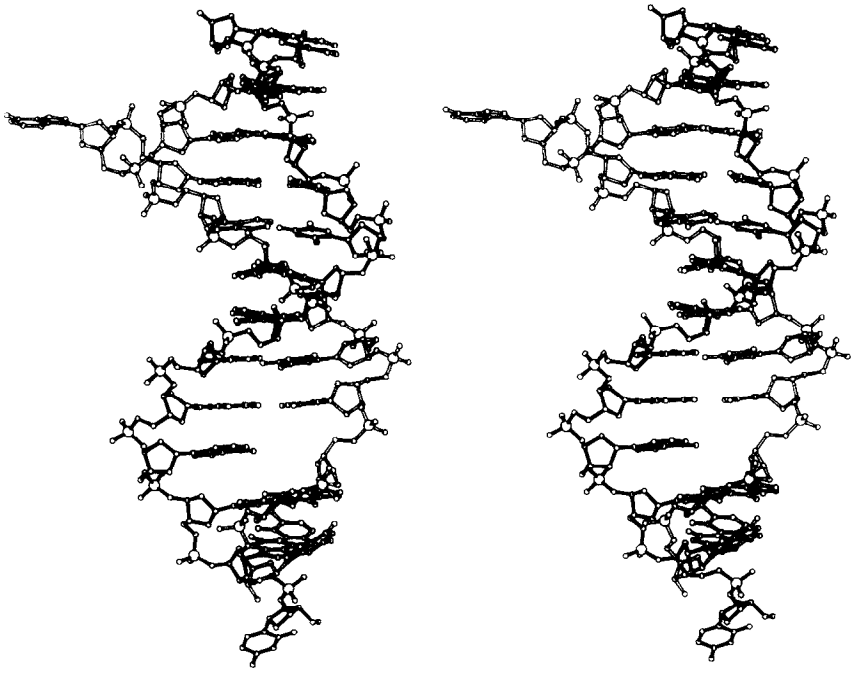
nine base-pairs apart, directly confirming the stacked-in/looped-out structure. The fact that the same peaks were obtained when phasing on either model verifies the correctness of model B. Furthermore, it eliminates the possibility that there is some gross disorder, with the two structures coexisting in the crystal.

(c) *The three-dimensional structure of the tridecamer*

The refined tridecamer structure, shown in Figure 4, is a *B*-type double helix with one duplex in the asymmetric unit. One extra adenine base loops out from the double helix, while the other stacks in. The double helix is frayed at one end, i.e. a potential terminal G·C base-pair is disrupted with each base participating in separate intermolecular interactions. This still resulted in a 12 base-pair double helix (Fig. 4), but different from what we previously reported (Joshua-Tor *et al.*, 1988). The looped-out adenine base interacts with a neighboring duplex by intercalating into the major groove, between bases opposite the stacked-in adenine base of that molecule, forming an A·A base-pair with it (Fig. 5). The average twist is 34° and the average rise per residue is 3·3 Å. This results in about 10·3 residues per turn of the helix.

(i) *The conformation of the looped-out bulge*

This structure gives us the first detailed look at a conformation of a looped-out bulge site in a DNA



(b)

Figure 4. (a) The structure of the tridecamer with strand 1–13 shown in red and strand 14–26 in blue. The 2 extra adenine bases are highlighted with van der Waals' dot surfaces: (b) a stereo diagram of the tridecamer structure.

duplex, although several theoretical models have been built (Fresco & Alberts, 1960; Olson *et al.*, 1985). At the bulge site, adenine base A17 loops out with the backbone making what we call a loop-the-loop curve (since it reminded us of a roller coaster loop-the-loop). The path of the chain in the 5′ to 3′ direction looking perpendicular to the helix axis and towards the major groove is as follows (Fig. 6): the helix continues downwards towards the phosphate group of A17, of which the oxygen atoms are pointing down, then goes up to the sugar, $O_{(4')}$ now pointing in the opposite direction to the rest of the sugars in that strand, up further to the phosphate group of G18 to approximately the same "height" as the sugar of C16 (when looking parallel to the helix axis), of which the oxygen atoms are pointing up this time, and then the backbone continues downwards to the sugar of G18. The looped-out nucleoside is flipped over, as can be seen from the opposite orientation its sugar moiety has relative to the other sugars in this strand. The distance between the phosphate groups surrounding the looped-out base (those of A17 and G18) is only 5·1 Å, which is closer than the normal distance of approximately 7.0 Å for *B*-type structures and approximately 5·9 Å for *A*-type DNA structures (Saenger, 1984). This feature was predicted by Fresco & Alberts (1960) in their model-building study and is supported by the increased salt dependence of t_m for bulge-containing DNA fragments relative to ordinary DNA fragments (Evans & Morgan, 1982). The bases 5′ and 3′ to the bulge are well stacked, as if they belong to consecutive nucleotides (see Fig. 7). The helical rise is 2·9 Å between these two flanking base-pairs, and their twist angle is 35°. They do, however, show some distortion, with a rather large buckle of 12° and 17° (Table 6).

The looped-out base, A17, is stabilized by intermolecular hydrogen bonds and by base stacking. It forms hydrogen bonds with A4, the stacked-in extra base from a symmetry-related molecule, and is also stabilized by stacking interactions with bases of this same neighboring molecule above and below, in particular with the guanine (G24: see Fig. 5(b), and the next section).

However, this loop-out is likely to correspond to one of the lower energy conformations of a bulge in a DNA molecule. There is probably an equilibrium between the stacked-in and looped-out states that depends on various environmental factors, e.g. temperature as seen in the n.m.r. studies for unpaired pyrimidines in DNA helices (Kalnik *et al.*, 1989*a*, 1990), or ionic strength, as in the increased salt dependence of t_m for DNA fragments containing bulges relative to ordinary DNA (Evans & Morgan, 1982), indicating that the phosphate groups are closer to one another in the bulged DNA. Our structure at the looped-out bulge site agrees with the latter observation, since the two phosphate groups 5′ and 3′ to the bulged nucleotide are in closer proximity than in the normal distance in DNA and is consistent with earlier work (Fresco & Alberts, 1960). This putative equilibrium of bulge to stacked conformation is also very likely to depend on other molecules that can interact with the bulged

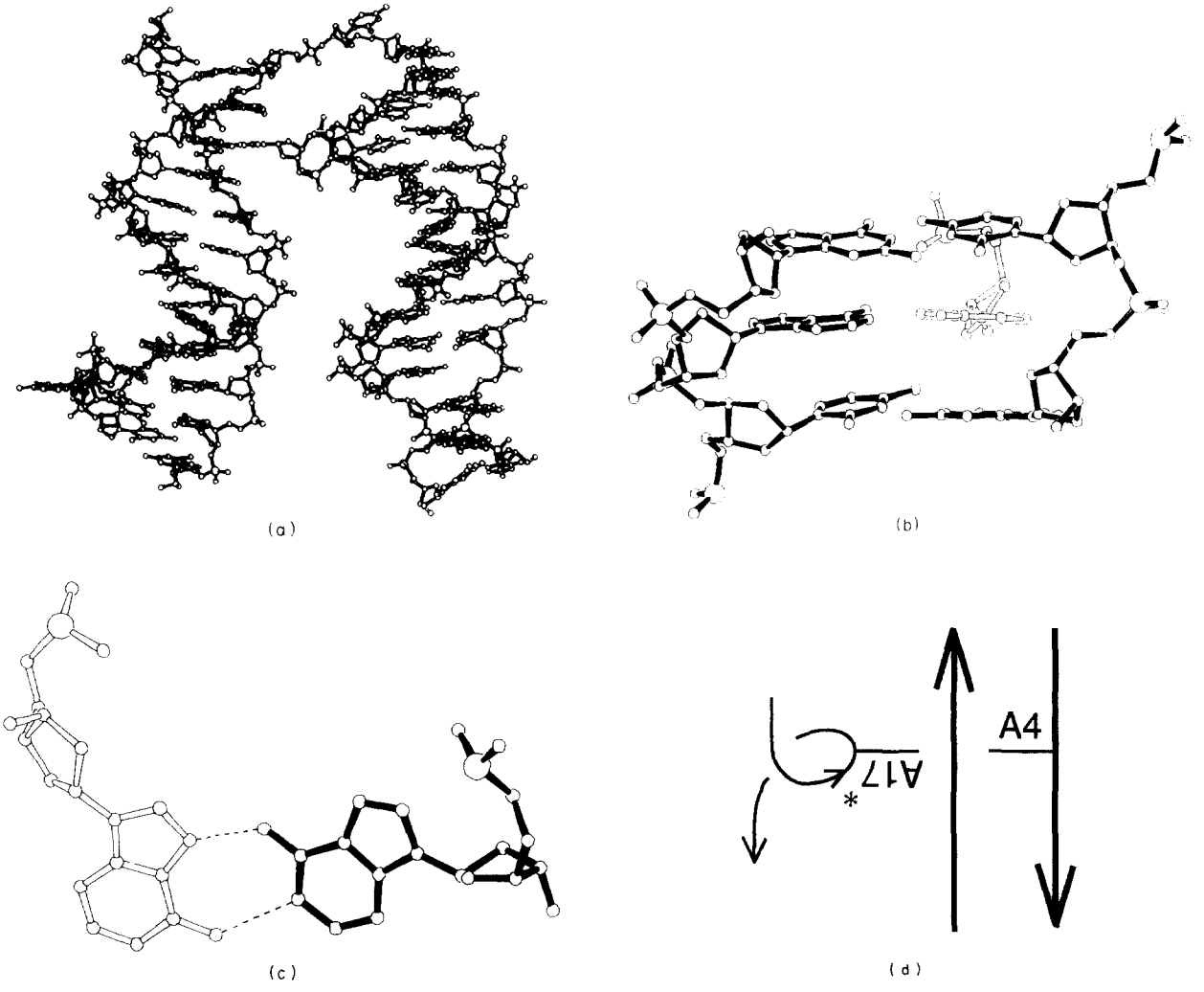


Figure 5. (a) One of the extra adenine bases from one molecule (drawn as open bonds) looping-out and forming an A·A base-pair with the other extra adenine base from a neighboring molecule (drawn as filled bonds); (b) a close-up on the intermolecular interaction; (c) the stacked-in adenine, A4, from one molecule (drawn as filled bonds) base-pairs with its Watson–Crick surface, while the looped-out base, A17, from a symmetry-related molecule (drawn as open bonds) uses the Hoogsteen surface; (d) a drawing depicting the directionality of the strands carrying the paired extra bases. Although these 2 strands are parallel, since A17 is flipped over and the backbone makes a loop, the 2 bases are in an antiparallel orientation.

DNA. Hydrogen bonding of a looped-out base to an amino acid residue of a nearby protein may stabilize a looped-out conformation in a similar manner as observed in this crystal structure where the interaction is with a nucleotide of a nearby DNA molecule. Stacking interactions with aromatic amino acid residues, other nucleotides or perhaps aromatic mutagens as well, may stabilize a looped-out base. This latter type of stabilization is observed in this crystal structure as well, since the looped-out base is sandwiched between two nucleotides in a neighboring double helix.

Fresco & Alberts (1960) suggested that the extra base can be looped out of the helix by rotation about its two adjacent phosphodiester bonds, opening a vacancy into which an adjacent residue

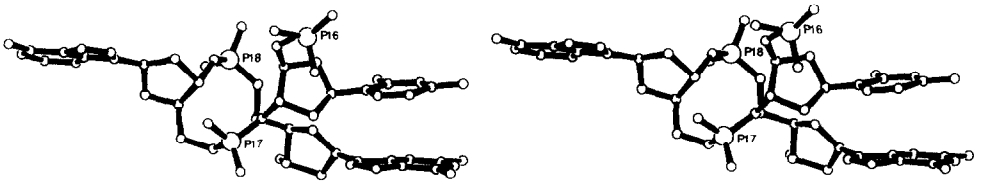


Figure 6. Stereo drawing of the bulge showing the loop-the-loop path the backbone takes around the looped-out base (see the text).

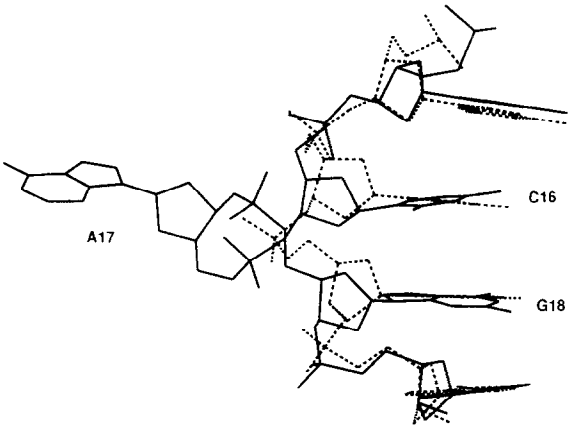


Figure 7. The bulge site around A17 (thick lines) superimposed upon a regular *B*-DNA double helix derived from fiber X-ray studies (broken lines). The nucleotides above and below the bulge are stacked upon each other in much the same way as in a regular *B*-DNA helix.

could move. Careful examination of the backbone torsion angles of the tridecamer (Table 7) revealed that these are not the only torsion angles involved in the bulge observed. In order to get a feeling for the way this strand can accommodate a looped-out base, one can envision the following schematic scenario starting from a *B*-DNA duplex: the extra base is turned out from a normal intrahelical position to point out from the helix axis. This leaves the two bases flanking the extra base in place, about 6·8 Å apart. By bringing these two bases closer together, there would be a gain in stacking energy. This is achieved by the extra nucleoside flipping over to an opposite orientation. It results in a movement of the $O_{(3')}$ of this sugar towards the 5′ direction, thus dragging along with it the rest of the chain one step upstream (in the 5′ direction of the strand). An alternative way of visualizing this is that, since the 3′ flanking base moves upstream, the looped-out nucleoside flips over to relieve the strain by increasing the distance between the two phosphate groups. This is drawn schematically in Figure 8(a). In order to examine which torsion angles are involved, physical models were built followed by computer model building. A stepwise construction of the bulge is shown in Figure 8(b).

Table 7

Backbone torsion angles (°) for the tridecamer

Residue	α	β	γ	δ	ε	ζ	χ
1			161	165	257	191	26†
2	21	138	101	158	205	296	244
3	243	91	129	98	240	222	217
4	317	116	63	146	214	220	249
5	333	127	23	136	236	174	267
6	279	110	79	151	206	230	224
7	333	161	24	148	185	254	272
8	281	152	70	99	161	293	231
9	244	165	112	152	152	280	242
10	289	238	37	162	121	293	271
11	260	222	64	154	174	259	277
12	318	167	27	126	223	173	254
13	343	126	24	156			255
14			85	163	242	253	242
15	211	87	118	173	183	238	236
16	191	231	107	137	230	166	297
17	243	90	308	143	322	301	267
18	98	150	180	96	165	291	233
19	295	195	40	151	179	242	260
20	324	160	43	149	157	264	267
21	261	179	89	147	226	213	255
22	297	99	93	110	215	290	236
23	352	232	280	153	194	275	245
24	311	188	45	144	223	169	282
25	63	253	199	117	218	343	204
26	283	160	82	172			304

†This angle is 211 for the other possible conformation (*anti*) for C1 (see the text).

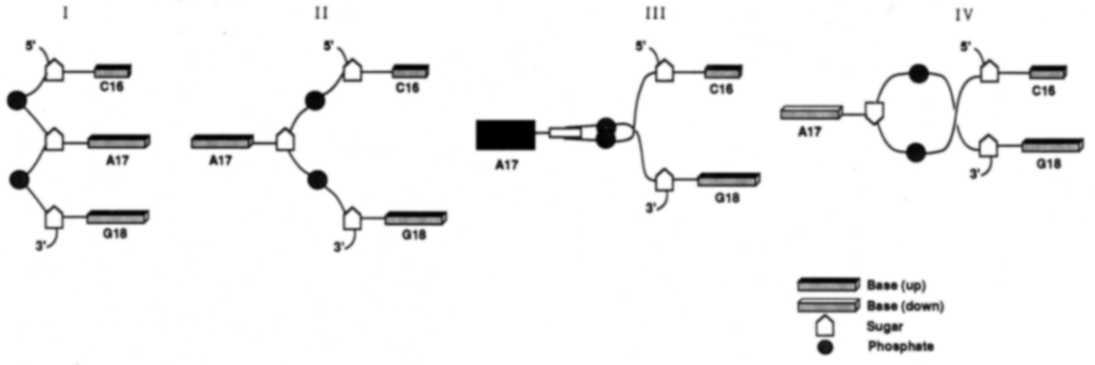
The angle around $O_{(3')}$–P of the nucleotide 5′ to the looped-out base, ζ of C16 is in an *ap* conformation. This is observed in other crystal structures, in UpA(2) (Sussman *et al.*, 1972), in d(pTpT) (Camerman *et al.*, 1976), in d(TpA) (Wilson & Al-Mukhtar, 1976) and at the TpA step in d(ATAT) (Viswamitra *et al.*, 1978). These all display an extended backbone structure rather than a helical arrangement. Another torsion angle that plays a role in the conformation of the bulge is P–$O_{(5')}$, α, of G18, which is $+sc$ rather than $-sc$. This is observed

Table 6

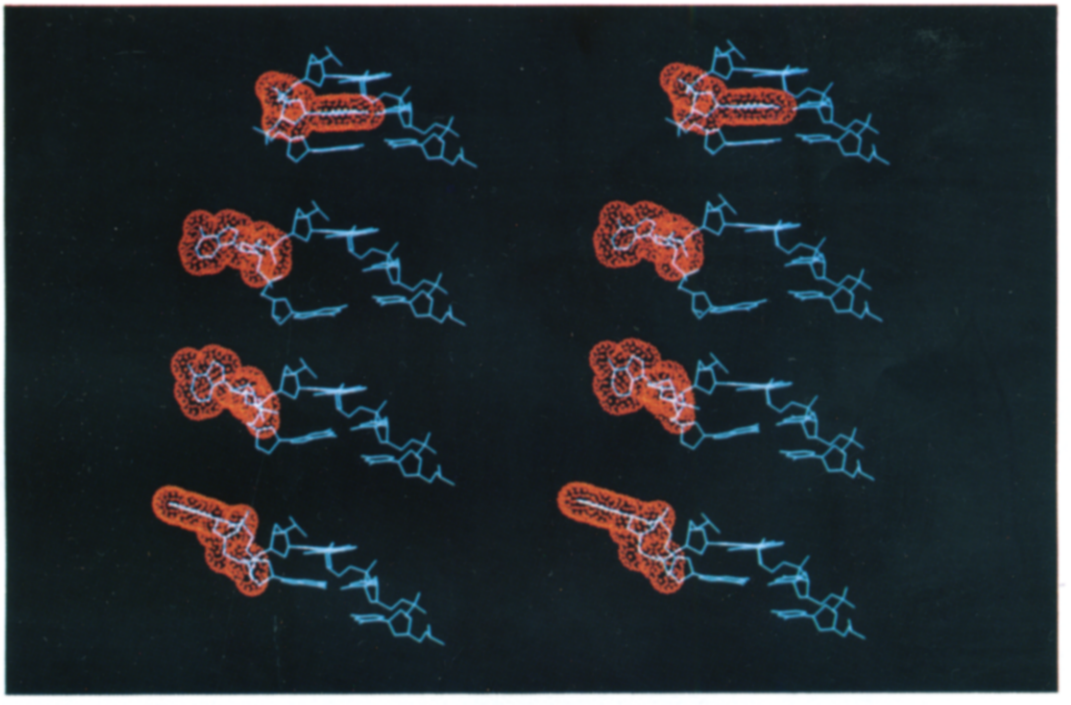
Helix parameters for the tridecamer and those for the proflavin-soaked tridecamer

Base step	Twist (Ω) Native	Twist (Ω) Soaked	Rise (D_z) Native	Rise (D_z) Soaked	Buckle (κ) Native	Buckle (κ) Soaked	Slide (D_y) Native	Slide (D_y) Soaked	x displ (d_x) Native	x displ (d_x) Soaked	y displ (d_y) Native	y displ (d_y) Soaked	Base-pair
G2-C3	38·2	36·0	2·8	3·6	−9·9	2·5	0·6	0·6	3·3	3·7	0·7	0·2	G2·C25
C3-A4	29·5	36·5	2·8	3·0	3·1	0·8	1·9	2·1	3·6	3·8	−0·7	−1·1	C3·G24
A4-G5	5·3	5·0	3·9	3·7	8·6	1·8	0·0	−0·2	2·7	2·4	−0·4	−0·4	A4·A17*
G5-A6	41·5	37·2	3·4	3·7	−4·8	−2·3	0·4	0·5	0·9	0·9	0·0	−0·3	G5·C23
A6-A7	39·9	40·5	3·4	3·5	4·8	−3·7	0·1	−0·4	0·4	−0·4	0·5	0·5	A6·T22
A7-T8	33·6	33·1	3·4	3·3	−6·1	−2·7	−0·6	−0·4	0·6	0·2	0·8	0·6	A7·T21
T8-T9	37·9	37·3	3·3	3·0	1·4	−6·5	0·1	0·0	0·6	1·0	0·1	0·2	T8·A20
T9-C10	41·1	38·9	2·9	3·1	−5·5	4·8	−0·5	−0·3	0·8	1·2	−0·2	−0·5	T9·A19
C10-G11	34·8	37·7	2·9	3·8	11·7	8·3	0·0	0·0	0·8	0·8	−1·5	−1·5	C10·G18
G11-C12	27·6	40·5	3·6	3·2	17·1	6·0	0·0	0·6	0·2	0·1	−2·2	−2·1	G11·C16
C12-G13	43·6	31·5	3·6	4·0	11·2	7·3	1·7	0·4	0·5	−0·3	−3·1	−1·9	C12·G15
					5·6	−6·8			−0·8	−1·2	−1·9	−1·8	G13·C14

* From a symmetry-related molecule.



(a)



(b)

Figure 8. (a) A stepwise construction of the looped-out bulge. I, Initial regular *B*-type structure. II, A17 swings out, leaving C16 and G18 in place. Flipping over of A17 and movement upstream from G18 is shown here in 2 steps: III, A17 flipped halfway, such that the plane of A17 base lies roughly on the plane of the helix axis and perpendicular to the other bases, the plane of its sugar moiety is also roughly perpendicular to the other sugars. G18 begins to move toward C16. IV, A17 sugar and base are flipped over relative to the other bases in the strand, flanking phosphate groups are pushed apart and the flanking bases C16 and G18 are stacked on top of each other; (b) same steps as in (a) using computer model building. I, Starting with a double helix of the sequence d(C16-A17-G18) d(C9-C10-G11) in a *B* conformation (Arnott *et al.*, 1980); numbering was chosen so that it would correspond to numbering at the looped-out bulge site in the tridecamer. II, A17 swings out mainly by changing the α and γ of both A17 and G18. Specifically, A17α changes from a $-sc$ to a $+sc$ conformation, A17γ from $+sc$ to $-ac$, G18α changes from $-sc$ to $+sc$ as well, and G18γ from $+sc$ to $-ac$. III, The main backbone torsion angles that were changed here are A17α, back to a $-sc$ conformation, A17ε from $-ac$ to $+ac$ and G18α to an *ap* conformation. This moves the base of G18 towards C16 halfway up to a normal stacking distance. IV, Finally, C16ζ was changed to a $+ap$ conformation, A17α to $-ac$, A17ζ to a $-sc$ conformation and G18γ to an *ap* conformation, which completed flipping of the extra base, and allowed proper stacking between the 2 flanking bases. It should be noted that there may be other combinations of torsion angles that might have the same effect.

in the crystal structure of UpA(1) (Sussman *et al.*, 1972), in the second phosphate group of ApA^+pA^+ (Suck *et al.*, 1973), and in ApA (the Ca^{2+} salt) (Einspahr *et al.*, 1981). These all display a conformation corresponding to a non-helical π turn in the backbone, which is found in the anticodon and T loops of yeast $tRNA^{Phe}$, where α is in an *ap* conformation (Kim & Sussman, 1976). The glycosidic angle χ of A17 also changes somewhat in the tridecamer structure (from $\sim -120°$ in *B*-DNA to

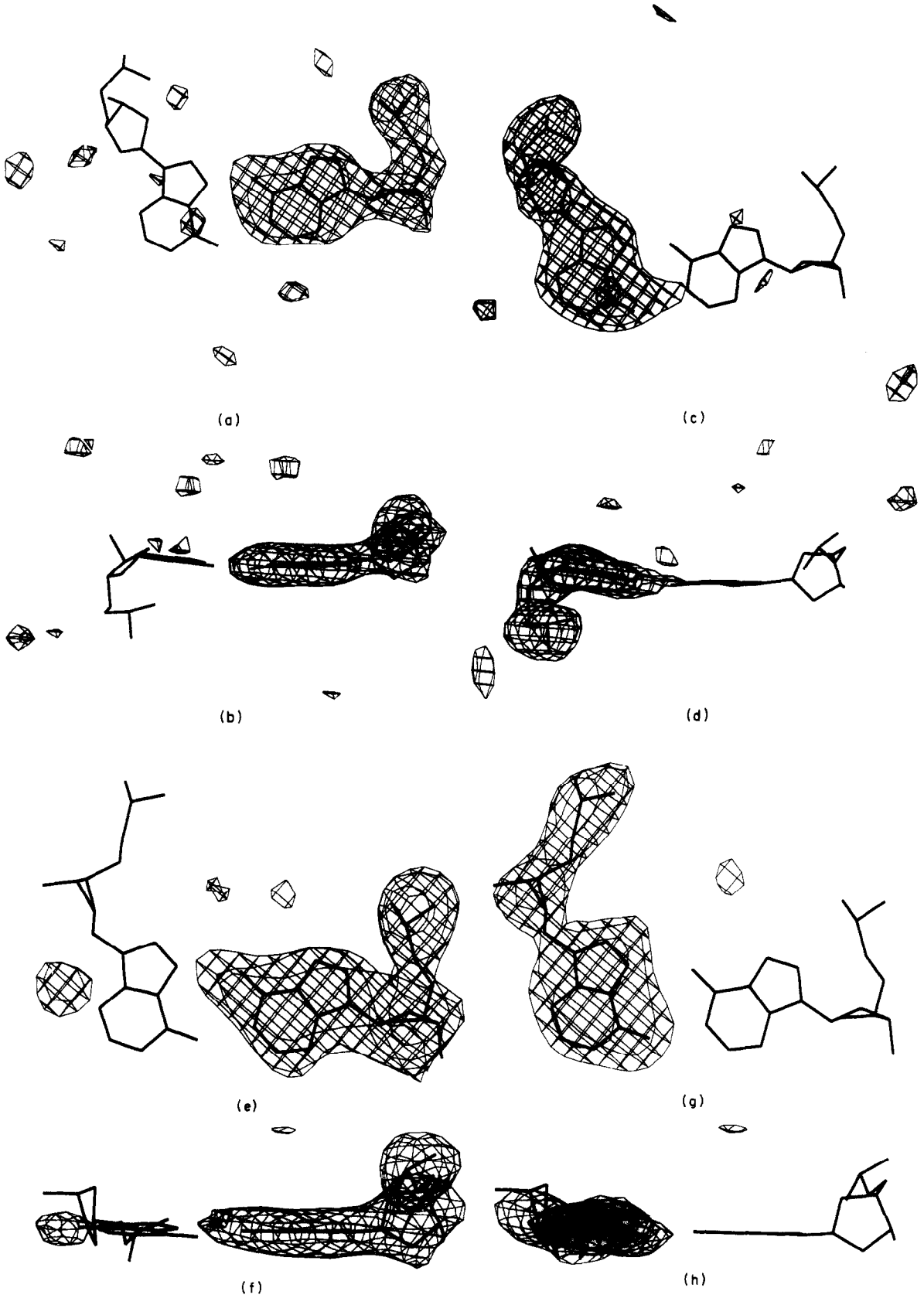


Figure 9. Omit maps for each of the extra adenine bases (lowest contour level 3σ) for the native tridecamer from a view looking down the base-pair (top) and from an edge-on view (bottom). (a) and (b) A4 omitted; (c) and (d) A17 omitted; and for the the proflavin-soaked tridecamer (lowest contour level 3σ) from a view looking down the base-pair (top) and from an edge-on view (bottom). (e) and (f) A4 omitted. (g) and (h) A17 omitted.

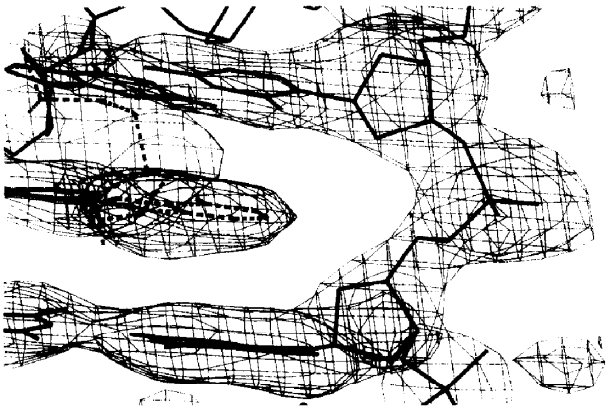


Figure 10. Electron density map ($2F_{obs} - F_{calc}$) at the site of intercalation of the looped-out base, A17, between C23 and G24 of a neighboring helix (lowest contour level 1σ).

$-63°$), probably in order to improve the intermolecular interactions with the neighboring molecule, i.e. the A·A base-pair and stacking. Still, one should be cautious about overinterpreting the significance of the precise values of these torsion angles since, given the resolution of this structure, some of the torsion angles may be changed while still yielding a good fit to the electron density.

(ii) *The A·A base-pair*

The bulged base is stabilized in two ways, one of which is intermolecular hydrogen bonding. The looped-out base (A17) from one duplex extends through the major groove of another duplex and forms a base-pair with the stacked-in extra adenine base of that duplex. This is a novel type of base base interaction in DNA crystal structures. The inserted adenine base (A4) uses the Watson–Crick side for pairing, while the looped-out base (A17*) uses the Hoogsteen surface. Specifically, the amino group ($N_{(6)}$) of adenine base A4 forms a hydrogen bond with $N_{(7)}$ of A17*, and the amino group of A17* forms a hydrogen bond to $N_{(1)}$ of A4 (Fig. 5(c)). The strand carrying A4 and that carrying looped-out A17* with which A4 base-pairs, run in parallel directions but, since A17 is flipped over as discussed above, the two nucleotides A4 and A17* are oriented antiparallel to one another as illustrated in Figure 5(d). Omit maps for these two extra bases are shown in Figure 9.

(iii) *The looped-out base inserts into a neighboring duplex*

The looped out base, A17, from one duplex is stacked between the bases of a neighboring double helix, and it intercalates into this symmetry-related molecule through the major groove (Fig. 5). Therefore, the two bases (C23 and G24) on the opposite strand from the stacked-in extra adenine base (A4) to which the looped-out base (A17*) hydrogen bonds, increase their separation to 6·5 Å to allow A17* to intercalate. However, the twist angle between these base-pairs remains close to that of consecutive base-pairs and is 38°. A17 is therefore stabilized by intermolecular stacking as well as hydrogen bonding. An electron density map ($2F_{obs} - F_{calc}$) map for this region is shown in Figure 10. Stacking is very good both with the guanine base (G24) on one side and with the cytosine base (C23) on the other, as discussed in subsection (viii), below. Its base-pair partner, A4, on the other hand, stacks well with the guanine base (G5) below but appears unstacked with the cytosine base (C3) above it, which is the normal case for a pyrimidine–purine step in *B*-DNA. The base-pair helical rise (between $C_{(1')}$–$C_{(1')}$ vectors) between C3·G24 and A4·A17* is 2·8 Å and between A4·A17* and G5·C23 is 3·9 Å†. Without the looped-out base intercalating between C23 and G24, these two bases would probably stack on top of each other. This might result in kinking or bending of the helix, a feature that was shown by n.m.r. and gel mobility studies (Goren-

†These values do not add to 6·5 Å because the helical rise is defined as the distance between base-pairs along the helix axis.

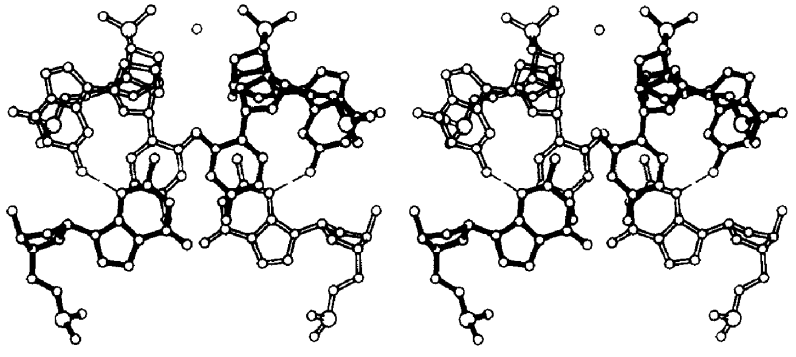


Figure 11. Stereo drawing of the intermolecular interaction between G26 and G2 from molecules related by a 2-fold axis. One molecule is drawn with filled bonds and the other with open bonds. The G2·C25 base-pair and G26 of each molecule is shown. The amino group of each G26 makes a hydrogen bond, drawn as a broken lines, with $N_{(3)}$ of G2 in the minor groove of the 2-fold related molecule. A water molecule (or ion) lies on the 2-fold axis, bridging the phosphate groups of the symmetry-related G26 nucleotides.

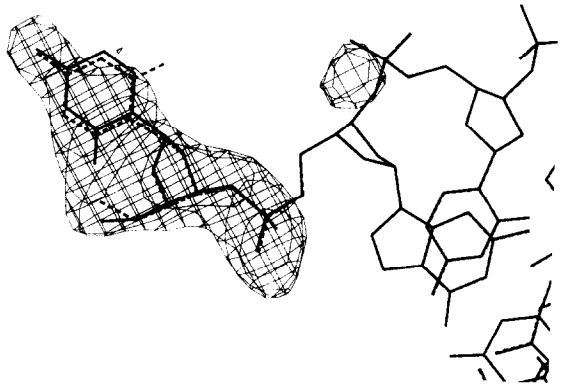


Figure 12. Omit-refine map of C1 (lowest contour level 2·2σ). The *syn* conformation for C1 is drawn as continuous lines and the *anti* conformation as broken lines.

stein *et al.*, 1987; Nikonowicz *et al.*, 1989, 1990; Patel *et al.*, 1982; Wang & Griffith, 1991). This is consistent with stacked-in extra bases discussed in the Introduction, and from models, based on the n.m.r. results, built by molecular dynamics and energy minimization (Hirshberg *et al.*, 1988). Hence, it is tempting to imagine A17* as having the effect of "straightening" out a bent helix by this intercalation. A similar type of effect could, of course, result from intercalation of other aromatic molecules, e.g. acridines.

(iv) *Frayed ends*

Another feature of the structure is the fraying of the molecule at one end. As the tridecamers in the crystal structure are stacked end-to-end, and as there is space due to crystal packing for only 12 base-pairs, a potential base-pair at one end is frayed apart. The terminal guanine base (G26) lies in the minor groove of a neighboring molecule (see Fig. 11). Its amino group ($N_{(2)}$) forms a hydrogen bond with $N_{(3)}$ of G2 of the symmetry-related molecule ($N_{(2)}$–$N_{(3)}$, 2·9 Å). The phosphate groups of two of these guanine bases related by a 2-fold axis, are bridged by an assigned water molecule. However, when omitting all of the water molecules and calculating a difference map, the peak between these two phosphate groups is the highest peak in the difference map. It is significantly higher than all the rest of the peaks that were interpreted as water molecules, most of which have similar values. It is therefore reasonable to interpret this peak as a hydrated magnesium ion, in which two of the water molecules were replaced by two phosphate oxygen atoms. Its melted partner C1 extends out to a different direction within the minor groove of yet another neighboring molecule and away from its own double helix. The backbone around the 3′ phosphate group (of G2) adopts torsion angles characteristic of extended structures. Namely, ζ of C1 is in an *ap* conformation as was described for ζ of C16, and α of G2, the torsion around the P–$O_{(5')}$ bond, is in a $+sc$ conformation, as is the case for G18α flanking the bulge. At this resolution, it is not clear from the electron density maps whether this cytosine base is in an *anti* or a *syn* conformation. However, the appearance of the maps do favor a *syn* conformation over the *anti* conformation. The *anti* conformation, although possible, does not fit the electron density maps as well as the cytosine base in the *syn* conformation even when using phases of the *anti* model. In addition, an omit-refine difference map calculated after omitting C1 and the adjacent phosphate group (of G2) and running several cycles of refinement shows that the *syn* conformation fits the electron density better than the *anti* conformation (Fig. 12). Two intermolecular and one intramolecular hydrogen bonds may form if this cytosine base is in a *syn* conformation. One intermolecular hydrogen bond is between its $N_{(4)}$ amino hydrogen and $N_{(3)}$ of G15 in the minor groove of a symmetry-related duplex ($N_{(4)}$–$N_{(3)}$ distance at 2·8 Å). The other requires the protonation of the $N_{(3)}$ of the cytosine base, which can then form a hydrogen bond to the carbonyl oxygen atom ($O_{(2)}$) of C14 of that same symmetry-related molecule ($N_{(3)}$–$O_{(2)}$ distance 2·9 Å: see Fig. 13). Although the pK_a value for $N_{(3)}$ of cytosine as a free nucleotide is 4·54 (Saenger, 1984), pK_a values are likely to change significantly when there is a stabilizing association with other moieties, as for example in a base-pair. A self base-paired parallel structure is observed for d(CpG) in which one of the two cytosine bases is protonated, where the pH of crystallization was 5·3 (Coll *et al.*, 1987*b*). In addition, protonation of cytosine in solution at pH 7 was proposed from c.d. studies (Gray, 1974; Gray *et al.*, 1984). An additional intramolecular hydrogen bond may form between the carbonyl $O_{(2)}$ of C1 and its 5′ hydroxyl group. In the *anti* conformation, the same hydrogen bond may form between the amino group of the cytosine base and $N_{(3)}$ of G15 ($N_{(4)}$–$N_{(3)}$ distance being 2·6 Å), but this is the only interaction. The gain in energy in forming an additional hydrogen bond ($\sim$3 to 6 kcal mol^{-1} per bond) may well stabilize the *syn* conformation over the *anti* conformation in this case. Another form of stabilization of this frayed cytosine residue is by stacking interactions between two such cytosine bases related by a 2-fold axis of symmetry. The electron density indicates the existence of an additional minor conformation for this cytosine base. In this conformation, C1 lies at the end of the major groove of the molecule, but no apparent interaction was found that can stabilize this particular conformation. It should be noted that this minor conformation was not included in the refinement.

(v) *Packing and intermolecular stacking interactions*

The molecules stack lengthwise, although in alternating directions (head-to-head and tail-to-tail) along the *ac* diagonal (Fig. 14(a)) in endless columns of double helices. Stacking is between molecules related by 2-fold symmetry. For one end of the molecule, the twist angle between the two molecules is 28° and the rise is 3·4 Å. At the other end, the stacking is between the second base-pair of each molecule, i.e. base-pair G2·C25 stacks on top of

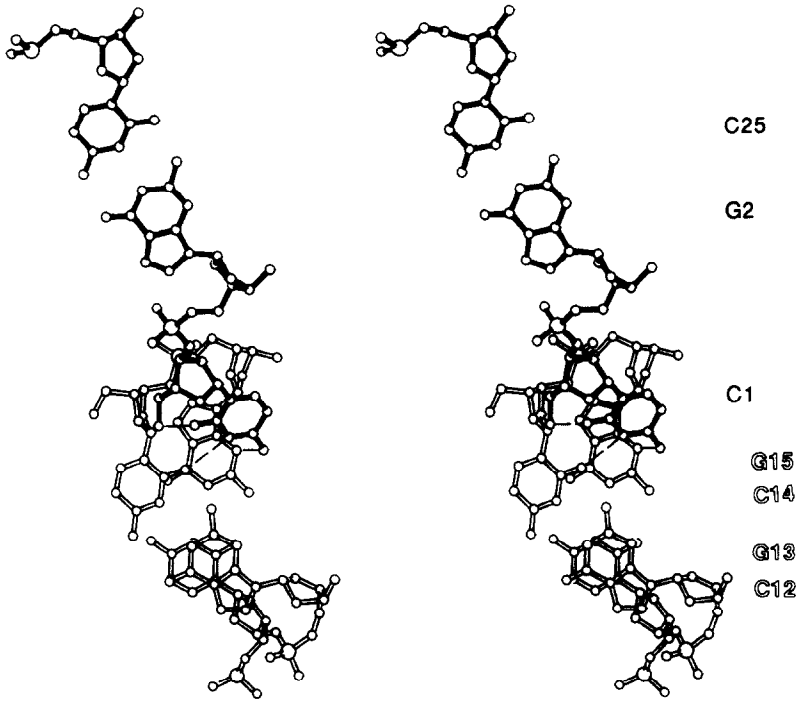


Figure 13. Stereo drawing of the hydrogen bonding interactions (drawn as broken lines) of the frayed cytosine, C1, in the *syn* conformation. C1, G2 and C25 from one duplex is drawn with filled bonds and C12, G13, C14 and G15 from a neighboring duplex are drawn with open bonds. The amino group ($N_{(4)}$) of C1 forms a hydrogen bond with $N_{(3)}$ of G15 in the minor groove. Another hydrogen bond forms between $O_{(2)}$ of C1 and its own $O_{(5')}$. The 3rd hydrogen bond requires protonation of $N_{(1)}$ of the frayed cytosine base (see the text) and forms with the carbonyl ($O_{(2)}$) of C14.

another G2·C25 related by 2-fold symmetry. On this end, the twist angle is −66° with a rise of 3·3 Å. Edge-on views of these two intermolecular stacking interactions are shown in Figure 14(b) and (c). The other type of intermolecular interactions was discussed in detail above and involves the hydrogen bonding between the two extra adenine bases, i.e. the A·A base-pairs, and the intermolecular hydrogen bonding in which the frayed bases are involved. The A·A base-pairs form between molecules related by a 2-fold screw axis. Several solvent molecules are involved in hydration bridges between helices as well. These will be discussed in detail below.

Packing in this crystal is not as tight as seen in the *B*-DNA decamers, which also pack in a *C*2 crystal lattice with end-to-end stacking (Privé *et al.*, 1991), or in the phosphorothionate hexamer, crystallized in a $P2_12_12_1$ crystal form (Cruse *et al.*, 1986), both of which, however, stack head-to-tail, since they are related by unit cell translation and not by a 2-fold axis seen in the tridecamer. The columns of helices do not come as close to one another as in the decamers. A view of the packing down the helix axis of the tridecamer is shown in Figure 15. In order to get a measure for how tight the packing is, one can calculate the unit cell volume per DNA molecular mass in the cell, i.e. the DNA "packing volume" for the different crystals. Two of the decamers (GA and CG decamers in Privé *et al.* (1991)) have a packing volume of 2·02 and

Table 8

Geometry of the possible three-centered hydrogen bonds in the major groove

	Heavy-atom distance			H atom distance		Angles				
Base-pair	R12	R31	R23	R24	R14	A12	A31	A23	SAN	Propeller
A6·T22	3·7	3·2	2·7	1·8	2·9	101	101	155	357	−8·8
T9·A19	3·4	2·9	3·0	2·0	2·6	95	100	164	359	−27·5

The Table is in the same format as an equivalent Table in Yanagi *et al.* (1990). Atom $3(N_{(6)})$ and $2(O_{(4)})$ are on the same base-pair and atom $1(O_{(4)})$ is on the same strand as $2(O_{(4)})$ on an adjacent base-pair. Hydrogen atom 4(H) is bonded to $3(N_{(6)})$. R12, R31 and R23 are the distances (in Å) between atoms 1 and 2, 3 and 1, and 2 and 3, respectively. A12, A31 and A23 are the angles defined by atoms 1, 4 and 2, by atoms 3, 4 and 1 and by atoms 2, 4 and 3, respectively. SAN is the sum of 3 angles about central atom 4, as a measure of planarity of the 3-center cluster. Propeller is the propeller twist angle of the first base-pair.

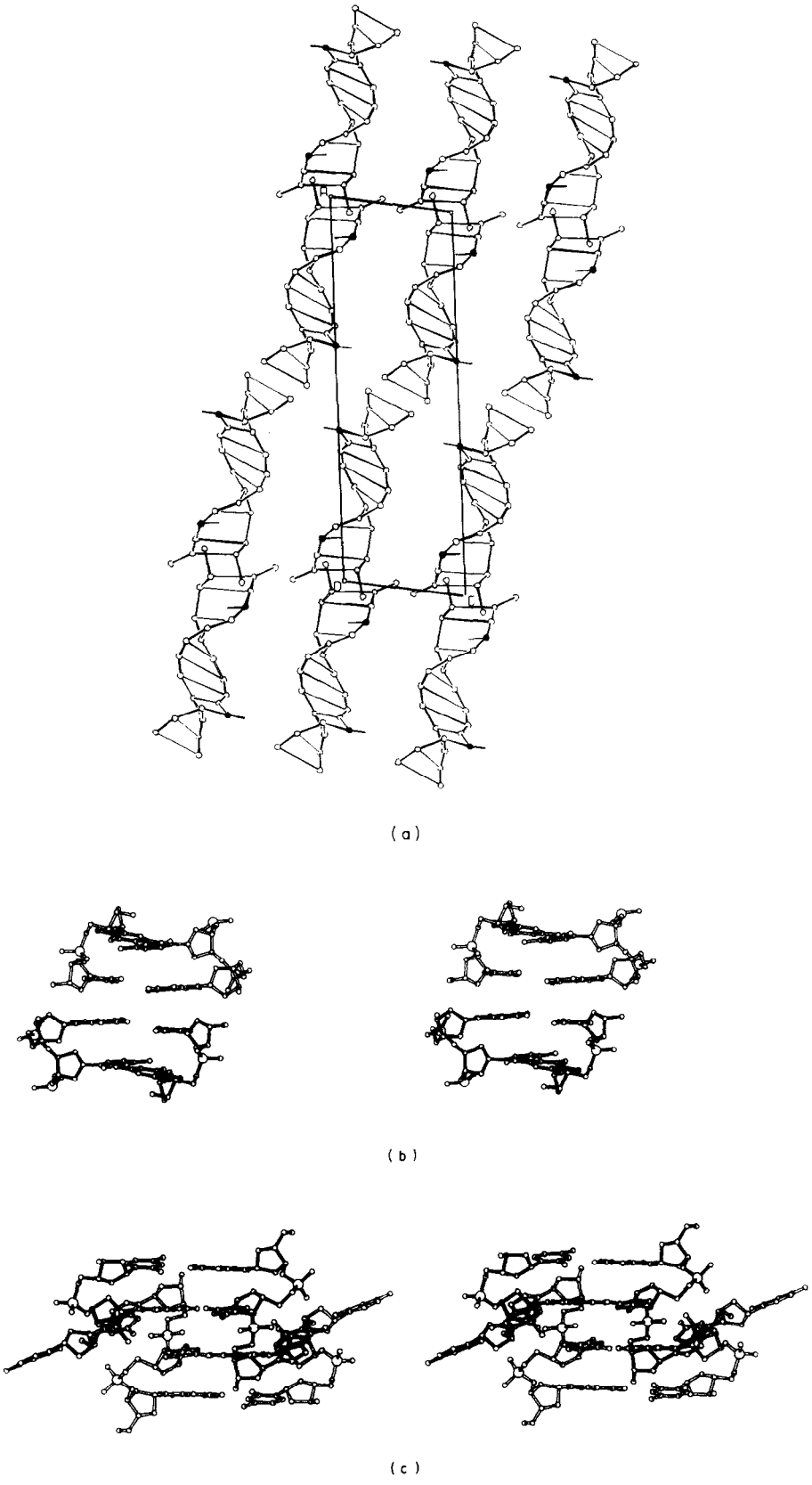


Figure 14. (a) A packing diagram of the tridecamer structure showing one layer of molecules as viewed down the crystallographic *b* axis. Each strand is represented by lines connecting the $C_{(1')}$ atoms, and the base-pairs are drawn as straight lines connecting the $C_{(1')}$ atoms on opposite strands. For the 2 extra adenine bases, a line connects the $C_{(1')}$ (shown as filled circles) and $N_{(1)}$ atoms. (b) Stereo drawing depicting the stacking interactions between 2 molecules (one drawn with open bonds and the other with filled bonds) around the 2-fold axis at $(\frac{1}{2},0,\frac{1}{2})$; (c) same as (b) around the 2-fold axis at the origin. The terminal base-pair is frayed in each molecule and the stacking is between the 2nd base-pair, C2–G25, in each molecule.

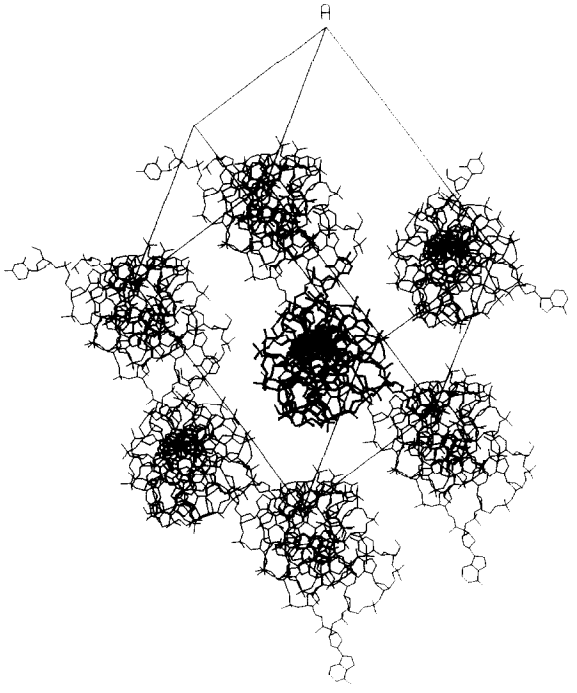


Figure 15. A packing diagram of the tridecamer structure in a view down the helix axis. The looped-out bases are seen protruding into neighboring helices: notice the looped-out base from the top helix inserting into the center helix and the looped-out base from the center helix inserting into the bottom right helix.

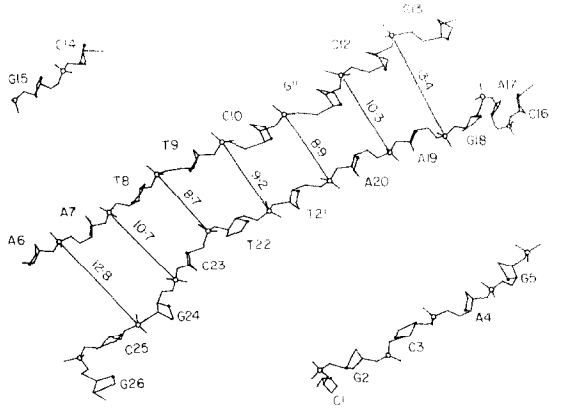


Figure 16. Cylindrical projection of the native DNA tridecamer structure showing phosphorus–phosphorus distances across the minor groove to be narrower in the center of the structure, in the vicinity of the AT-rich region, and more open at the 2 ends of the duplex (GC-rich region).

2·07 Å^3 mol g^{-1}, respectively, while for the tridecamer this value is 2·64 Å^3 mol g^{-1}, implying that it is not as tightly packed. The value for the Dickerson–Drew dodecamer is 2·29 Å^3 (g/mol DNA)$^{-1}$. Indeed, the decamers are the best diffracters among the four, while the tridecamer is the worst.

(vi) *Narrowing of the minor groove*

The minor groove near the middle of the structure at the A+T-rich region is narrowed to a width of 8·7 Å between phosphorus atoms across the groove. Narrowing of the minor groove is characteristic of A+T-rich sequences. This most likely arises from the high propeller twist in that region, which affects the width of the minor groove (Fratini *et al.*, 1982). Propeller twist in this structure is discussed in the following section. The minor groove opens up towards the G+C-rich ends of the duplex and is more pronounced at the frayed end of the duplex than at the other side. A plot of the P–P distances across the minor groove is shown in Figure 16.

(vii) *Propeller twist*

Figure 17 displays the observed propeller twist angles for the tridecamer and those of the Dickerson–Drew native dodecamer. Both structures have an increased propeller twist at the center AATT region. A large propeller twist is generally observed for runs of A residues (A-rich) (Coll *et al.*, 1987*a*; DiGabriele *et al.*, 1989; Fratini *et al.*, 1982; Nelson *et al.*, 1987). Curiously though, the propeller twist for two of the A·T base-pairs (A20·T8 and A19·T9) is closer to that of the A-tract sequences, i.e. around −28°, while for the other two base-pairs it is lower. The propeller twist of A7·T21 is similar to the value for the A·T base-pairs in the Dickerson–Drew dodecamer at around −17° (some of the dodecamer crystal structures, e.g. at 16 K, do exhibit larger propeller twist angles of around −23°) and for A6·T22 it is closer to the average for other sequences (−9°). The increased propeller twisting positions the adenine base amino group towards its 3′ side and within hydrogen bonding

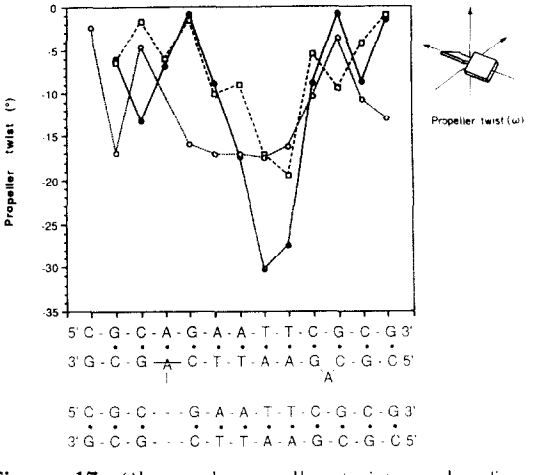


Figure 17. Observed propeller twist angles for the native tridecamer (filled circles), the proflavin-soaked tridecamer (open squares) and the Dickerson–Drew native dodecamer (open circles). All structures exhibit a large propeller twist in the AATT region. Sequences are given for the tridecamers (top) and for the dodecamer (bottom). Note that the 2 strands of the dodecamer were swapped for comparison with the tridecamer structures for reasons explained in the discussion about helix bending (see the text).

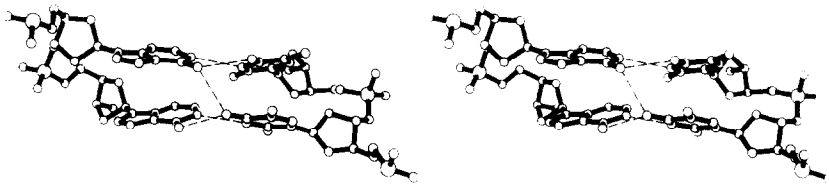


Figure 18. Stereo drawing of 2 A·T base-pairs exhibiting a strong propeller twist. A bifurcated hydrogen bond may form in the major groove between the amino group, $N_{(6)}$, of A19 and the $O_{(4)}$ carbonyl group of T9 and the $O_{(4)}$ carbonyl group of T8 of the flanking base-pair. Hydrogen bonds are shown as broken lines.

distance of the thymine carbonyl group of the following base-pair. Thus, this amino group may form a bifurcated (3-centered) hydrogen bond in the major groove with two successive thymine bases of the opposite strand, a feature that is found in the A-tract sequences mentioned above. In the tridecamer structure, two such bifurcated hydrogen bonds may form. One between the amino $N_{(6)}$ hydrogen atom of A19 with the carbonyl $O_{(4)}$ atom of T9 ($N_{(6)}$–$O_{(4)}$ distance 3·0 Å) and the carbonyl $O_{(4)}$ atom of T8 ($N_{(6)}$–$O_{(4)}$ distance 2·9 Å). This involves the two base-pairs that exhibit the strongest propeller twist angles and is shown in Figure 18. The other involves the amino hydrogen atom of A6 and $O_{(4)}$ of T22 (of its Watson–Crick base-pair: $N_{(6)}$–$O_{(4)}$ distance 2·7 Å) and $O_{(4)}$ of T21 ($N_{(6)}$–$O_{(4)}$ distance 3·2 Å). By generating hydrogen atoms in idealized positions, the geometry of the three-centered hydrogen bonds was examined. This shows the geometry to fit the criteria for a three-centered hydrogen bond described by Yanagi *et al.* (1991) and based on data listed in Table 8 (Jeffrey & Mitra, 1984). In order for a potential three-centered hydrogen bond to be accepted as such, each of the three angles about the central hydrogen atom should be greater than 90° and their sum should be approximately 360° (indicating the coplanarity of the 4 atoms involved). Distances should be within reasonable hydrogen bonding distances. Of course, these bifurcated hydrogen bonds do not form the zig-zag system of hydrogen bonds down the major groove that is characteristic of the A-tract sequences, since a minimum of three successive adenine bases (or thymine) are needed for this purpose (Coll *et al.*, 1987*a*; Nelson *et al.*, 1987).

(viii) *Base-pair stacking*

In general, base stacking in the tridecamer is typical of DNA in the *B* form, i.e. good intrastrand stacking for purine–pyrimidine and purine–purine steps and poor stacking for pyrimidine–pyrimidine and pyrimidine–purine steps. Figure 19 shows stereo plots for all base-pair steps, including steps involving the interduplex A·A base-pair and the two non-bonded steps from one helix to the next. They are all displayed in a view perpendicular to the mean plane of the upper base-pair shown with filled bonds. The direction is 5′ to 3′ on going from top to bottom on the right strand. These were compared with base-pair step stereos of the Dickerson–Drew native dodecamer (Fig. 7 in Dickerson & Drew (1981)). They prove to be very similar in most cases. It is remarkable that adding such perturbations as two extra adenosine bases in the structure has so little effect on the base-pair stacking scheme. It should be stressed here that the Dickerson–Drew dodecamer was never used for solving or refining the structure and that only an Arnott-type regular *B*-DNA double helix was used for these purposes. Therefore, it is especially interesting to note how "normal" this C-G base-pair step looks from the view across the bulge, i.e. flanking A17 (Fig. 19(j)). The two terminal base-pair steps do differ, however, from the dodecamer. Specifically, the stacking between base-pairs G2·C25 and C3·G24 (Fig. 19(b)) is rather poor in contrast to what would be expected for a G-C step. The slide between the two base-pairs is also quite apparent from this Figure. Displacement is even more noticeable for the last C-G step between C12·G15 and G13·C14. Note the path of the backbone of the second strand pulling to the left between C14 and G15. Changes in stacking at the ends of the molecule may arise from differences in packing requirements in the tridecamer relative to the Dickerson–Drew dodecamer. Stacking in a central A-T step, namely between A7·T21 and T8·A20, is better than the stacking of the equivalent step in the native dodecamer. A combination of low helical twist angle (of 28°) and large displacement in the $-x$ direction (towards the minor groove) for base-pair G11·C16 results in good stacking between G11 and C12 in the first strand but relatively poor stacking of the other two bases, C16 and G15 in the second strand. Two special base-pair steps take place in the ladder of base-pairs of the tridecamer and involve the A·A base-pair and flanking base-pairs. As was mentioned earlier, there is good stacking between G24 and the looped-out A17* (A17 from another molecule), while C3 and A4 are not stacked (Fig. 19(c)). A17* also stacks well with C23, as does A4 with G5 (Fig. 19(d)). Stacking interactions between molecules is shown in Figure 19(a) and (m) and was discussed earlier.

(ix) *Helix bending and base-pair displacement*

The double helix of the tridecamer is bent by about 22°. The direction of the bend is roughly similar to the bend in the Dickerson–Drew dodecamer. Another way of illustrating the axial bending of the tridecamer double helix is by plotting the x and y direction cosines of the normal to the best mean plane through each base-pair. This is shown in

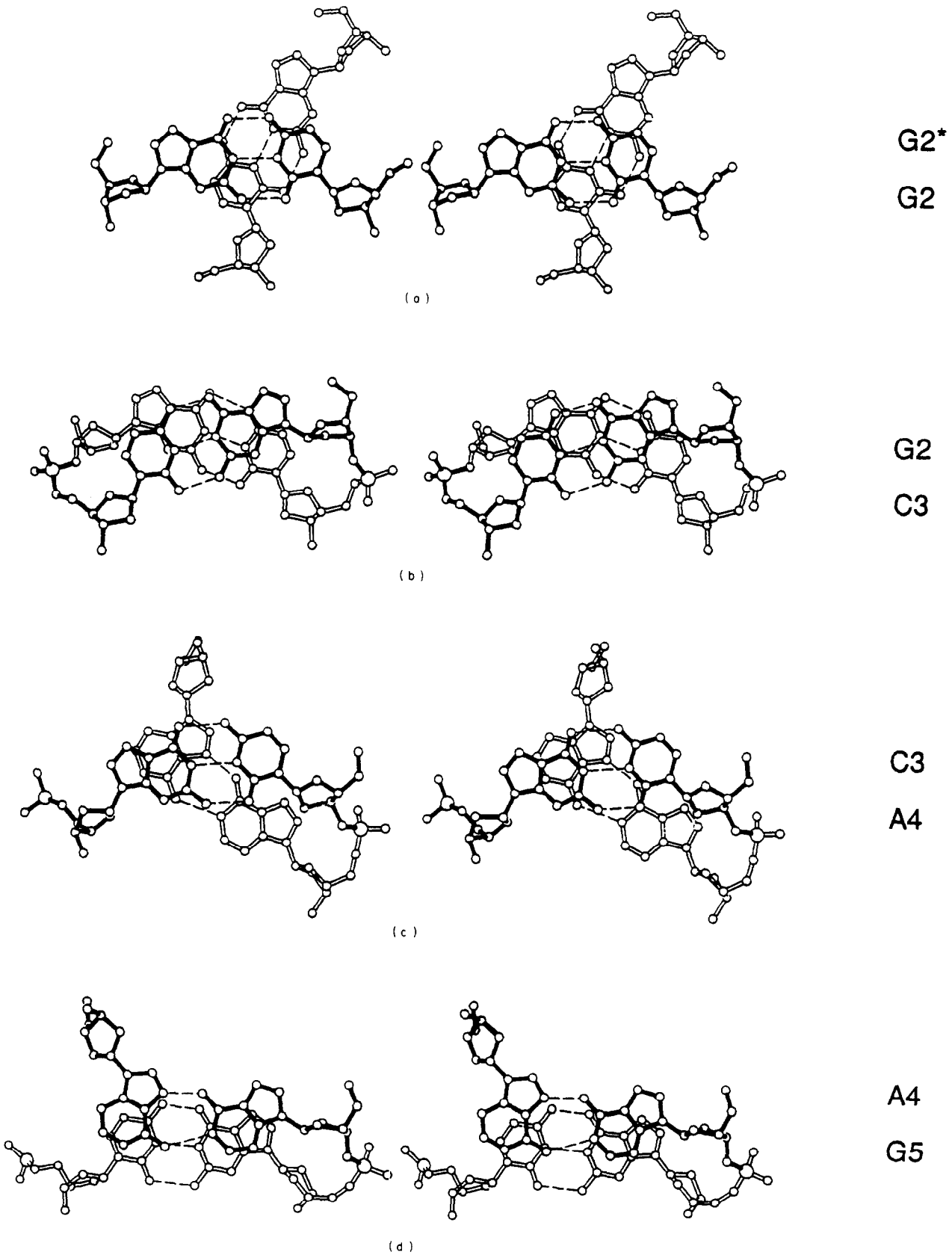


Fig. 19.

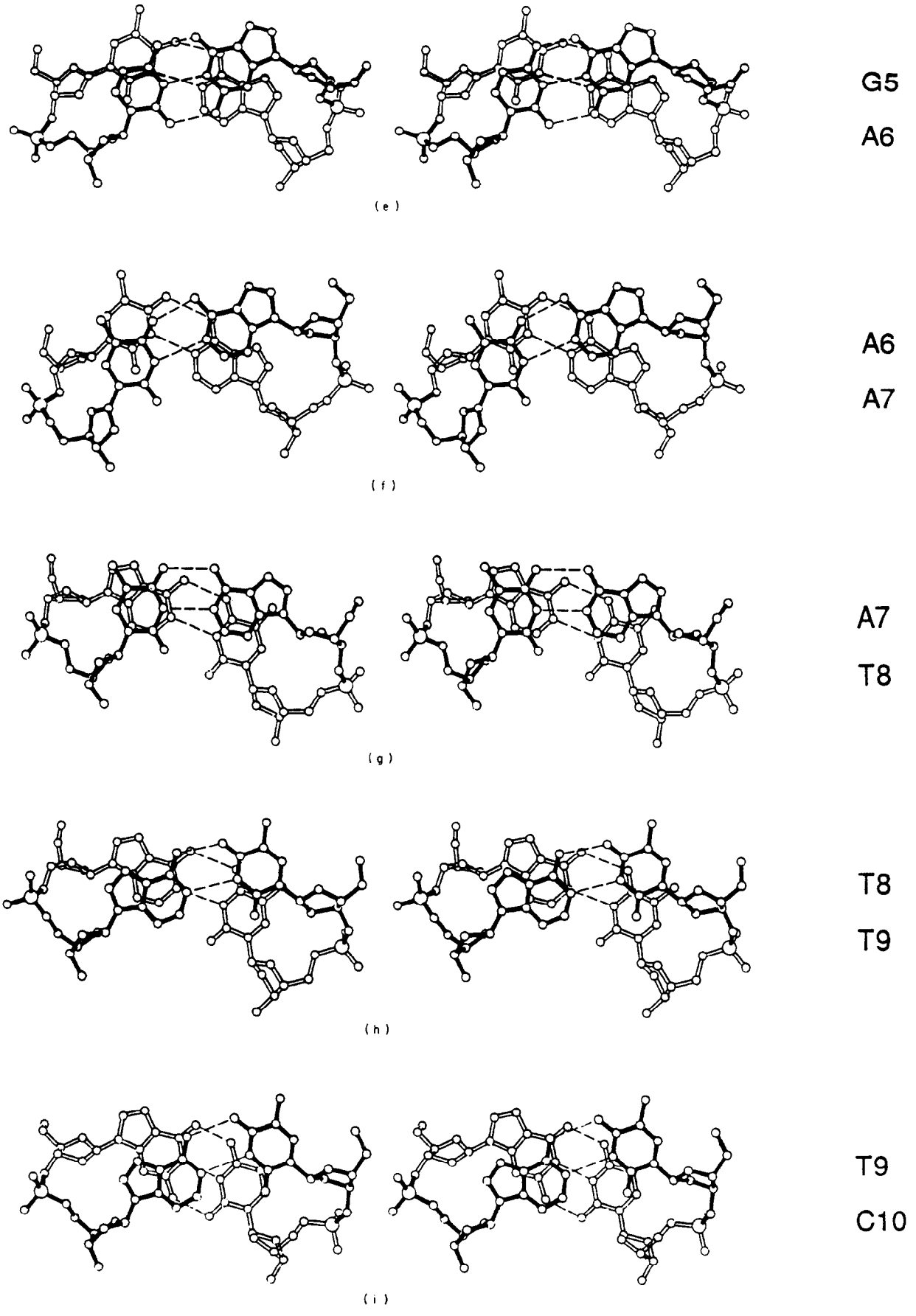


Fig. 19.

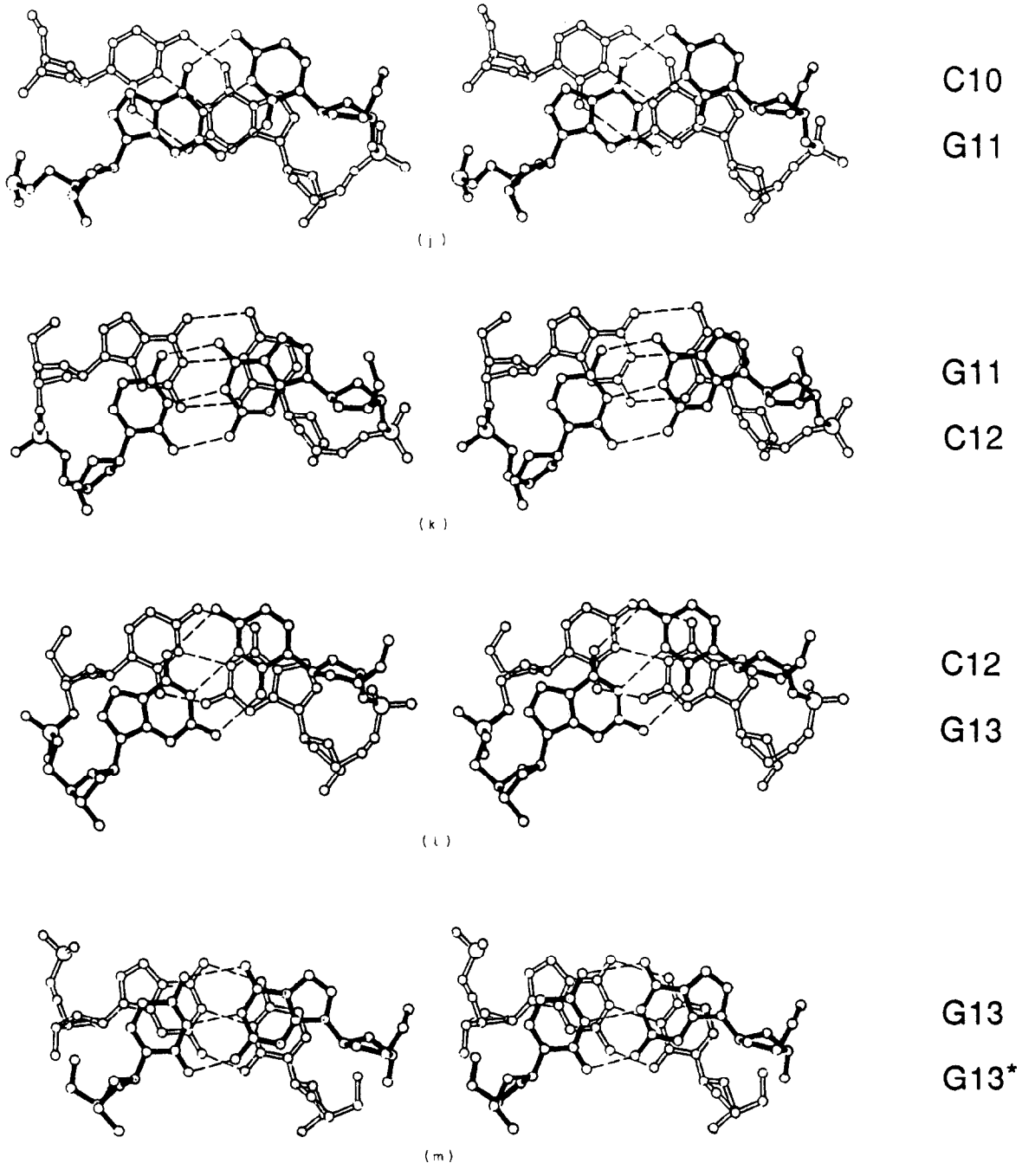


Figure 19. Stereo drawings of all the base-pair steps in the DNA tridecamer structure, including the 2 non-bonded steps between helices. Each view is perpendicular to the mean plane of the upper base-pair shown with filled bonds. The 5′ to 3′ direction is downwards for the right strand and upwards for the left strand. The base on the right hand side of each pair is indicated. (a) The non-bonded step between helices G2*·C25*/G2·C25; (b) G2·C25/C3·G24; (c) C3·G24/A4·A17*; (d) A4·A17*/G5·C23; (e) G5·C23/A6·T22; (f) A6·T22/A7·T21; (g) A7·T21/T8·A20; (h) T8·A20/T9·A19; (i) T9·A19/C10·G18; (j) the base step flanking the bulge C10·G18/G11·C16; (k) G11·C16/C12·G15; (l) C12·G15/C13·G14; (m) the non-bonded step between helices C13·G14/C13*·G14* (nucleotides from symmetry-related duplexes are marked by an asterisk).

Figure 20. The points represent the tip of a base-pair normal vector drawn from a common origin and looking down the helix axis. The distance between any two vector tips is the arc sine of the angle between the mean planes through each base-pair and is therefore a combination of the roll and tilt angles. In the tridecamer structure, one immediately notices a progression from right to left, which indicates bending of the helix axis. Another feature that is apparent from this plot is that little changes occur in the middle of the structure while most of the changes occur at the ends of the molecule; in particular, between vector 11 (normal to G11·C16 at the 5′ end of the bulge) and vector 12 (normal to

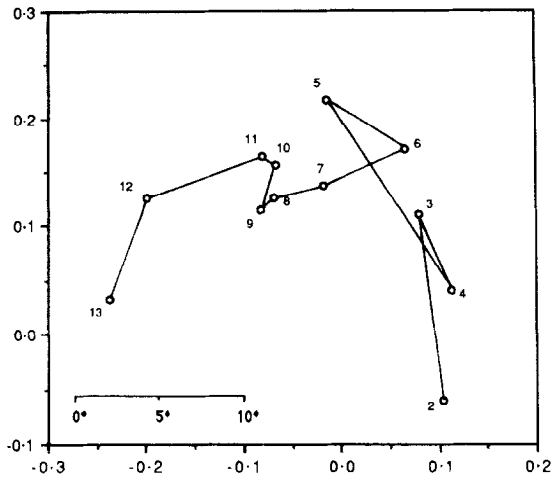


Figure 20. Polar diagram of normals to the mean plane of the base-pairs, projected onto a plane perpendicular to the best global helix axis. The points (numbered 2 to 13 according to the residue number of 1 member of the base-pair) represent the tip of normal vectors, with all vectors brought to a common origin and viewed down the best helix axis. The distance from the origin to each point corresponds to the sine of the angle between the mean plane of the base-pair and the perpendicular to the best helix axis. Distances between points in this projection can be converted approximately into angles by using the scale at the bottom of the plot. The progression of the points from right to left illustrates the bending of the helix.

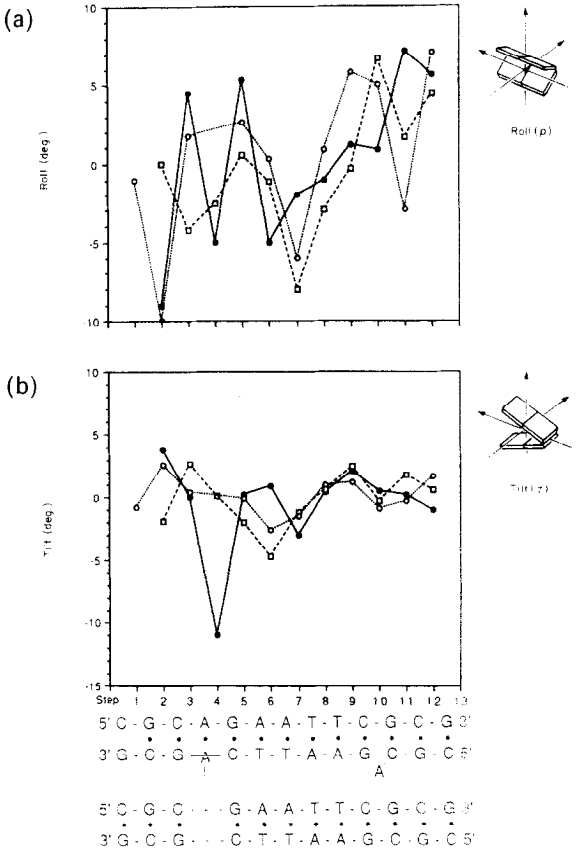


Figure 21. (a) Variation in roll angles between base-pairs for the native tridecamer (filled circles), the proflavin-soaked tridecamer (open squares) and the Dickerson–Drew dodecamer (open circles). Sequences are drawn for the tridecamers (top) and for the dodecamer (bottom). (b) Variation in tilt angles between base-pairs for the native tridecamer (filled circles), the proflavin-soaked tridecamer (open squares) and the Dickerson–Drew dodecamer (open circles). Sequences are given for the tridecamers (top) and the dodecamer (bottom).

C12·G15), a step that also has the largest roll angle. Surprisingly, there is very little change across the bulge, as can be seen by examining the change between vectors 10 (normal to C10·G18) and 11. However, there is a large change also between vectors 4 and 5, probably due to tilt, which will be discussed below. For the Dickerson–Drew dodecamer, when using the numbering system used by the authors, one finds a similar progression but in the opposite direction (Fig. 41 in Dickerson *et al.* (1985)). Since the direction is only due to the numbering scheme, the sequence being symmetric, and since the packing is different in any case, it is more natural to compare the two structures when they are aligned such that their curvatures lie in the same direction. For this reason, the two strands of the dodecamer were swapped in the comparison plots (or, in other words, the 1st base-pair in the tridecamer would be compared to the last base-pair of the Dickerson–Drew dodecamer; the 2nd to the penultimate, and so forth). One such plot displays the behavior of roll angles from one base-pair to the next in the tridecamer and in the dodecamer (Fig. 21(a)). The roll angle becomes generally more positive going along the sequence and up the helix. This indicates a compression of the major groove when progressing along the sequence. The bending in the tridecamer, as in the dodecamer, is due mainly to the roll component. Figure 21(b) depicts the behavior of tilt angles between base-pairs in the tridecamer and in the dodecamer for comparison. Most of the tilt angles are relatively small, apart from the angle between the special A·A base-pair and the following G·C base-pair. This tilt is mainly due to the neighboring looped-out adenine, which comes into the double helix from the major groove at a slight angle.

Packing forces are most likely responsible for producing this particular bend in the helix. The bend is similar to the bend in the dodecamer structure, but the packing is very different. Although it is not completely clear what bends the helix, one explanation could be that in order for A17, the looped-out base, to be in the right position to hydrogen bond to the stacked-in extra adenine base from another molecule, it pulls the backbone of C16, thus somewhat bending the helix. Hence, while it straightens out one side of another double helix, it actually bends its own double helix. The other forces would be those of end-to-end stacking with other duplexes. The bending of the upper part of the molecule (towards C12·G15) is also characterized by a relatively large y displacement in the direction of strand 2 (C14–G26) as can be seen in Table 6. The

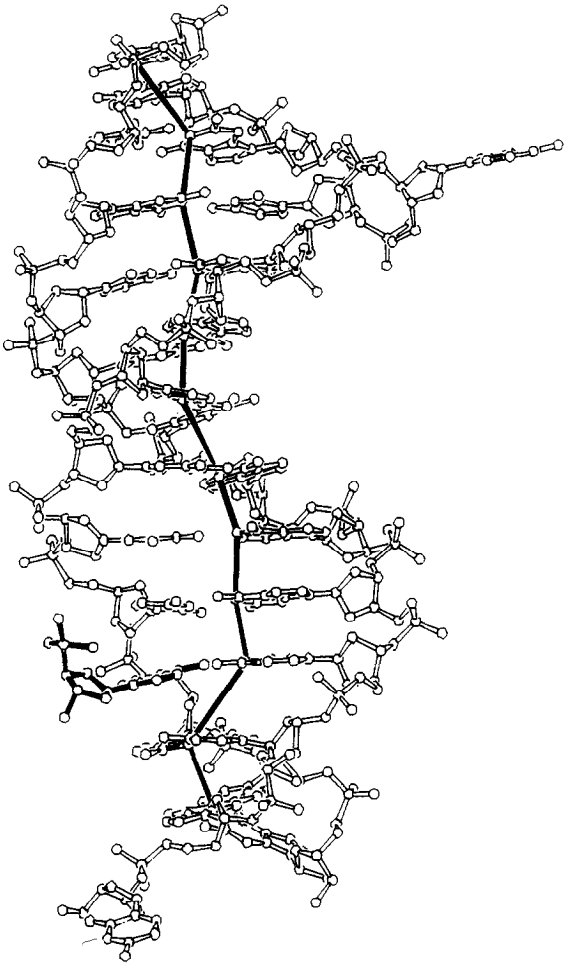


Figure 22. View of the structure of the tridecamer showing the large displacement of the 1st 2 base-pairs in the direction of the major groove towards the intercalating base (drawn with filled bonds). A spine made up of lines connecting the $N_{(1)}$ atoms of the purines from one base-pair to the next is drawn to help illustrate this.

magnitude of this displacement is of course dependent on the position of the helix axis and so in a way, and in this instance, it illustrates the bending of the helix. The slide, however, is irrespective of the choice of the helix axis. The relatively large positive slide between the last two base-pairs towards the other direction (strand 1†) may indicate a straightening out at this end.

Another type of helix distortion seems to arise from intermolecular interactions in which the looped-out bulge is the major culprit. Figure 22 shows the molecule in a view that accentuates the rather large displacement of the first two base-pairs in the direction of the major groove (in the positive x direction). From this view, it is apparent that this displacement results in improved stacking interactions between the guanine base (G24) sandwiching the bulge and between the bulged base (A17*). Since A17* is restrained by its own helix, it seems that G24 compensates by stretching out towards A17*, resulting in a displacement of about 3 Å from the helix axis.

†Strand 1 is defined as the strand from C1 to G13 (see Fig. 1).

(x) *Structured solvent molecules*

Due to the limited resolution of this crystal structure, only 23 solvent molecules could be located unambiguously. These are given in Table 9 together with their contacts with acceptor and/or donor atoms and are shown in Figure 23. Water molecules were added conservatively as described in Materials and Methods. Four solvent molecules seem to be involved in hydration bridges between helices. One of these (W15) makes a bridge between phosphate groups of G26 belonging to two symmetry-related molecules and sits on a 2-fold axis (see Fig. 11). As was discussed previously, this assigned water peak was interpreted as a hydrated magnesium ion, since it is at a significantly higher level in the map relative to the other assigned water molecules even for a peak on a special position, i.e. on a 2-fold axis. The somewhat closer distance to the phosphate oxygen atoms of the two G26 nucleotides may also indicate that these oxygen atoms replace two of the water molecules in the hydration shell of the ion. A water molecule was located between the phosphate groups of A4 and that of C25 of a symmetry-related molecule. Two solvent molecules form a hydration bridge between the phosphate groups of G18 and that of G15. All the water molecules are first hydration shell molecules, with at least one interaction with DNA atoms. Of these, 12 water molecules interact with phosphate oxygen atoms, which are the principal site of hydration in nucleic acids

Table 9
Water–DNA and water–water contacts in the tridecamer

Water	Interaction	Distance (Å)	B (Å^2)
1	G11 $O_{(6)}$; C16 $N_{(4)}$; G15 $O_{(6)}$	2·5; 2·9; 3·1	5·7
2	C23 $N_{(4)}$	3·0	9·4
3	W14; G24 $O_{(2P)}$	3·8; 4·0	2·0
4	T8 $O_{(2P)}$	2·7	4·8
5	G18 $O_{(1P)}$; W20	2·8; 2·8	3·0
6	A20 $N_{(6)}$; A20 $N_{(7)}$	2·8; 2·8	2·0
7	G11 $O_{(2P)}$; C10 $O_{(3')}$	2·2; 3·2	19·0
8	C12 $O_{(2P)}$	3·3	8·6
9	A19 $O_{(2P)}$; A19 $O_{(1P)}$; A19 $O_{(5')}$	2·5; 3·4; 3·4	6·9
10	C25 $O_{(2P)}$; A4 $O_{(2P)}$	3·5; 3·8	19·6
11	A4 $N_{(7)}$; A4 $N_{(6)}$	2·5; 2·9	2·0
12	A4 $N_{(3)}$; G26 $O_{(3')}$; W13	3·3; 3·5; 3·5	2·0
13	G5 $N_{(2)}$; G5 $N_{(3)}$; W12	3·0; 3·3; 3·5	18·2
14	C25 $O_{(2P)}$; W3	3·3; 3·8	2·0
15	G26 $O_{(2P)}$; G26 $O_{(2P)}$†	2·4; 2·4	2·0
16	C10 $O_{(3')}$	3·4	39·9
17	A7 $N_{(6)}$	3·6	3·2
18	C10 $O_{(2P)}$	3·4	2·0
19	G18 $O_{(2P)}$	3·8	11·0
20	W5; G15 $O_{(1P)}$	2·8; 3·7	42·1
21	T21 $O_{(1P)}$	2·1	2·0
22	G5 $N_{(7)}$; A4 $N_{(7)}$	2·9; 3·9	8·3
23	T22 $O_{(4)}$; A6 $N_{(6)}$; G5 $O_{(6)}$	2·4; 3·2; 3·2	13·9

†Between 2 symmetry-related molecules.

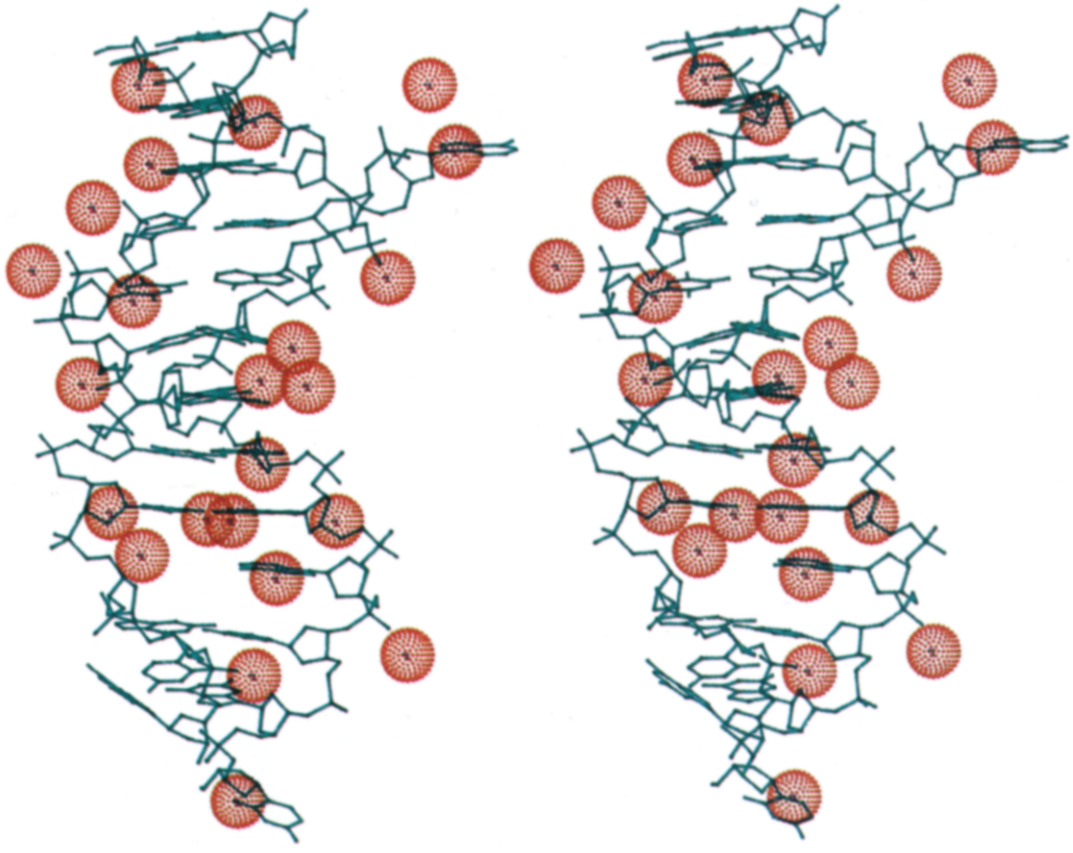


Figure 23. Hydration of the tridecamer. The 23 solvent molecules located in the crystal structure are shown as van der Waals' dotted spheres.

(Saenger, 1984), two of these also have an interaction with backbone ($O_{(5')}$ or $O_{(3')}$) oxygen atoms, one bridging to the adjacent $O_{(5')}$ oxygen and one to the adjacent $O_{(3')}$ oxygen atom. There are two more water molecules in this structure involved in hydration of backbone oxygen atoms. Of the seven water molecules hydrating the major groove, there are two (W6 and W11) that interact both with the $N_{(6)}$ amino group and the $N_{(7)}$ atom of the same adenine base and one water molecule that interacts only with an $N_{(6)}$ amino group of adenine (W17). One water molecule (W23) in the major groove bridges between $N_{(6)}$ of A6 and $O_{(4)}$ of T22 as well as $O_{(6)}$ of G5. Another molecule (W1) bridges $O_{(6)}$ of G11 and $N_{(4)}$ of C16 of the same base-pair. There is a water molecule (W22) interacting with G5 $N_{(7)}$ but has a weak interaction also with $N_{(7)}$ of A4. $N_{(4)}$ of C23 is hydrated as well. Only two solvent molecules have been located in the minor groove of this structure, one between $N_{(3)}$ and amino $N_{(2)}$ both of G5 and one bridging between a backbone oxygen $O_{(3')}$ atom and $N_{(3)}$ of A4 belonging to a symmetry-related duplex. Hence, the minor groove spine of hydration, first described by Drew & Dickerson (1981), and often present in *B*-DNA crystal structures, was not identified in this structure nor in other structures at a resolution of 2·5 Å or less (Privé *et al.*, 1991). This may be due to the limited resolution.

(d) *The structure of the proflavin-soaked DNA tridecamer*

Proflavin belongs to the family of acridine dyes that are flat polycyclic molecules that have been observed to cause frameshift mutations (Brenner *et al.*, 1961). They are known to bind preferentially to DNA sequences that contain bulges (Woodson & Crothers, 1988*a*). These aromatic molecules can intercalate into the double helix, leading to the hypothesis that this may be the mechanism inducing frameshift mutations (Lerman, 1963). Alternatively, they may bind to bases that loop out from the double helix by stacking interactions, thus stabilizing the looped-out structure and increasing the probability for a mutation. They may adopt several other "outside" binding modes, as seen in some crystal structures of small oligonucleotides and dinucleoside phosphate groups complexed with proflavin (Berman *et al.*, 1979; Westhof *et al.*, 1988). Diverse frameshift specificities have been observed *in vivo* and *in vitro* for acridine-induced frameshifts, as reviewed by Ripley (1990), which may reflect the differences in acridine–DNA interactions or the influence of different proteins involved in DNA metabolism.

It was our intention to examine, in the solid state, how proflavin may bind to a DNA molecule containing extra bases. We considered two alterna-

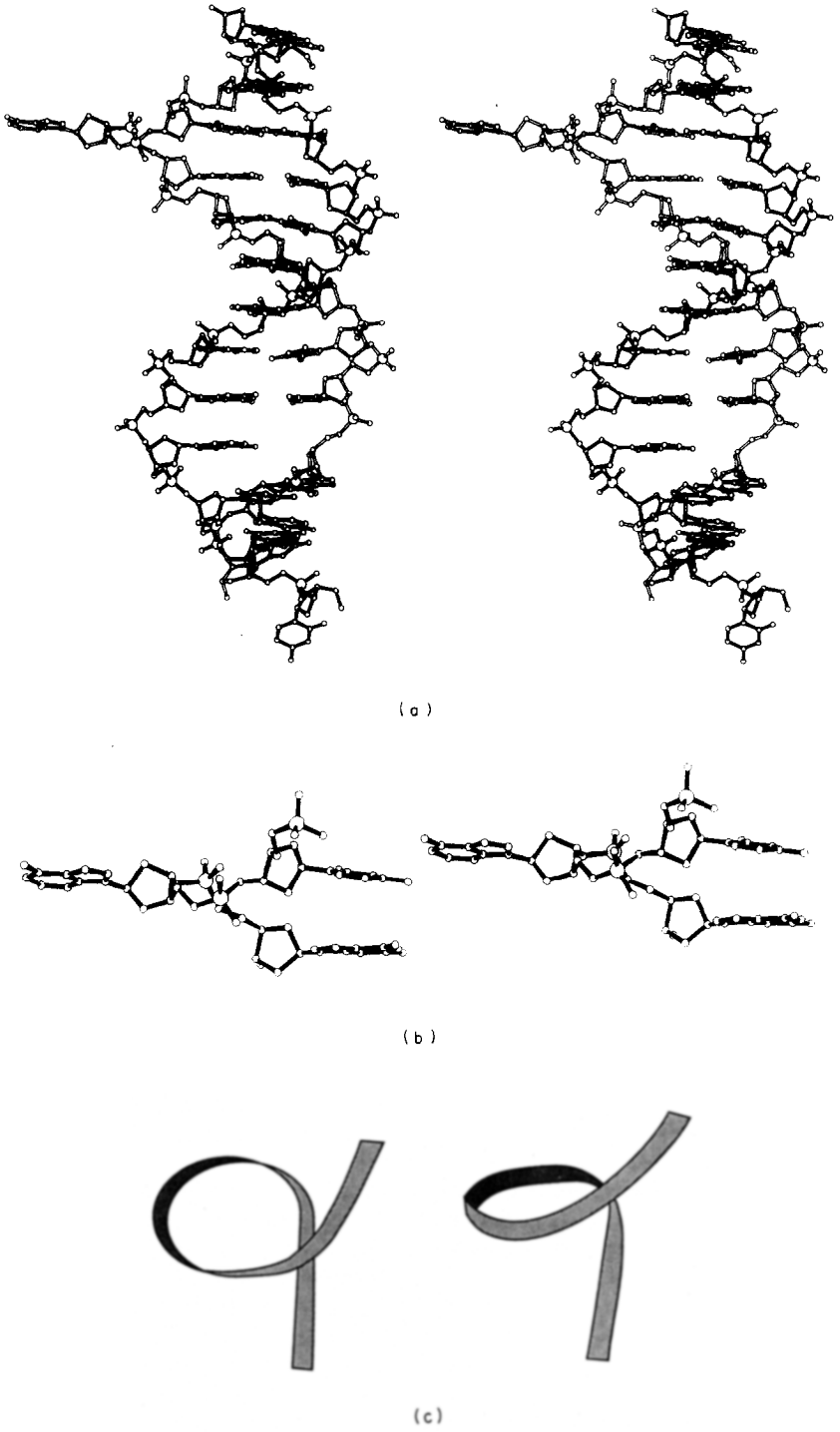


Figure 24. (a) Stereo drawing of the proflavin-soaked tridecamer structure; (b) stereo drawing of the looped-out bulge (A17) in the proflavin-soaked tridecamer; (c) a representation of the bulge conformations in the native (left) and in the proflavin-soaked (right) tridecamer crystal structures. The backbone is depicted as a ribbon whose loop is oriented approximately perpendicular to the plane of the base-pairs in the native structure, while it lies flat, parallel to the bases, in the proflavin-soaked tridecamer structure.

tive experimental strategies. One was to co-crystallize a DNA fragment containing an unpaired base, like the tridecamer described above, with proflavin. The other was to soak the mutagen into the existing crystals of the tridecamer. The latter is by far easier, although perhaps somewhat biased, the reason being that the proflavin molecule may have less effect on the stabilization of the extra bases as they may be stabilized by the lattice, as turned out to be the case here. Nevertheless, the second pathway was the first we tried.

Solution and model rebuilding of the proflavin-soaked tridecamer was described above as it was an integral part of the solution of the native tridecamer structure. The same general strategy of refinement was used as in the native tridecamer, i.e. X-PLOR

refinement with initial restraints of backbone torsional angles that were later released; intervention with manual refitting to electron density maps; tightening of weights on the geometry; conservative addition of water molecules in subsequent stages (11 molecules were added); and refinement of isotropic temperature factors. Average group B values for phosphate groups, sugars and bases for the proflavin-soaked tridecamer are given in Table 4. Many of these values, especially for the bases, tend to reach the minimum value allowed (2·0 Å^2). This may be due to the cutoff in resolution (10 to 3·2 Å). Several cycles of NUCLSQ were run at the last stage of the refinement, which resulted in further improvement of geometry. Although X-PLOR moved sugar puckers to various and different values there was no apparent change in R-factor or quality of the maps when restraining them to $C_{(2')}$ endo puckers. Therefore, in the course of the NUCLSQ refinement, all sugar puckers were restrained to $C_{(2')}$ endo. Proflavin could not be located in the difference maps. This is not surprising, since it would have a rather low occupancy due to the low incorporation of the drug in the crystals used for data collection (see below). Of course, it is possible that proflavin does not bind to any specific site and therefore cannot be located in the crystal structure. The final R-factor is 22·2% for data between 10 and 3·2 Å resolution (see Table 5).

(e) *Incorporation of proflavin*

The ratio of proflavin to duplex in the crystals was determined by means of u.v. absorbance measurements for a longer soak and a shorter, less concentrated soak (see Materials and Methods). For both soaks, the crystals acquired the yellow color of the drug, which remained even after multiple washings of the crystals and after there was no yellow "halo" around the crystals. A ratio of 1·5 proflavin molecules per duplex was established for the longer soak, while the ratio for the shorter soak was about 0·15 proflavin molecule per duplex. Reflections from crystals from the longer soak were too broad and diffraction was poor. As the crystals from the shorter soak were used for X-ray data collection, it is not surprising that the drug molecule could not be seen. Nevertheless, it was this structure that led to the correct model for the native tridecamer. Furthermore, even though this was done at lower resolution and is not as well refined as the native tridecamer, a fact that precludes discussion of the structure in any great detail, the unit cell dimensions are somewhat larger and the packing constraints are different enough to impose significant and observable changes in conformation.

(f) *The three-dimensional structure of the proflavin-soaked tridecamer*

The structure of the proflavin-soaked tridecamer is, in general, very similar to the structure of the tridecamer described above and is shown in Figure 24(a). It is a B-type double helix as well, with an average twist angle of 34° and average rise per residue of 3·4 Å. One extra adenine base is looped out from the double helix and forms an A·A base-pair with the other extra stacked-in adenine base of a neighboring helix in much the same way as the native tridecamer, and one end of the molecule is frayed as well. The main difference is that the helix is somewhat "expanded" relative to the native tridecamer (see Fig. 25). As a result, there are some changes in local conformation, i.e. the structure is more relaxed and regular, helix and base-pair parameters being in many cases closer to the Dickerson–Drew dodecamer (although this dodecamer was never used for refinement) than the native structure. This is expressed by the difference in cell dimensions between the two, which may be due to the flash cooling procedure, which often causes a different shrinkage of the cell upon cooling (L. Joshua-Tor & F. Frolow, unpublished results).

(i) *The conformation of the looped-out bulge*

As the packing is not as tight as in the native crystals, the bulged-out nucleotide (A17) has more space. It is more easily positioned relative to a neighboring molecule to form an A·A base-pair with

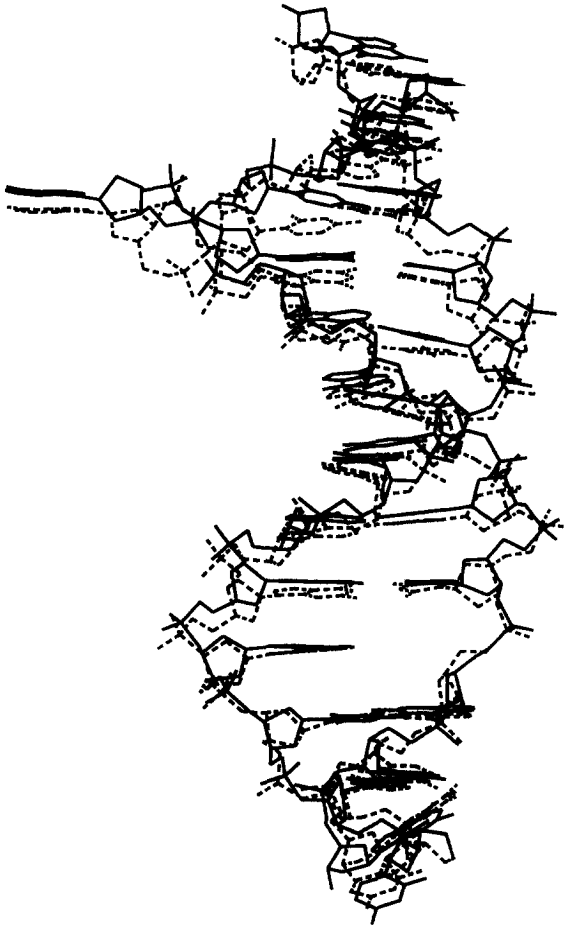


Figure 25. The proflavin-soaked tridecamer, drawn in thick lines, is "expanded" relative to the native double helix, drawn with thin lines.

Table 10

Distances (Å) between atoms around the bulge site in the native and in the proflavin-soaked tridecamers

Atom 1	Atom 2	Native	Proflavin-soaked
16C1′	17C1′	7·9	9·1
17C1′	18C1′	9·1	9·7
16C1′	18C1′	4·6	5·8
17P	18P	5·1	5·5

the stacked-in extra adenine base of that molecule, and the glycosyl angle, χ of A17, is closer to a normal value, $-130°$, than that observed in the native structure. Therefore, the backbone at the bulge adopts a somewhat different conformation. The backbone loop that the bulge makes, lies flat and parallel to the base-pairs (Fig. 24(b)). On going from 5′ to 3′, the path the backbone takes in the structure of the proflavin-soaked crystals is: from the sugar of C16 the backbone goes down and out to the phosphate group of A17, out further to the sugar, then back inward to the phosphate group of G18 and then down and in to its sugar. A representation of the bulge conformations for the native and proflavin-soaked structures is shown in Figure 24(c). Most everything is further apart in this structure, as can be seen from various distances for both structures in Table 10. However, the phosphate groups flanking the bulge are still relatively close together, at 5·5 Å, when compared to normal *A* or *B*-type helices. This is again consistent with the increased salt dependence for bulge-containing DNA fragments, as predicted by Fresco & Alberts (1960), and as discussed for the native tridecamer.

(ii) *Intermolecular interactions*

The inter-helical A·A base-pair is very similar to the equivalent base-pair in the native structure. Omit maps for both bases of this base-pair are shown in Figure 9.

As in the native structure, the looped-out base is like an intercalator into the strand opposite A4. As such, it is stabilized by stacking interactions with C23 and G24. Stacking with cytosine base C23 is even better in the proflavin-soaked structure.

The same fraying of one of the terminal base-pairs is seen in this structure. In general, these bases interact with the same neighboring molecules as in the native structure but in detail these interactions are slightly different. G26 makes two hydrogen bonds in the minor groove of a 2-fold related molecule. One with its amino group, $N_{(2)}$, to $N_{(3)}$ of symmetry-related G2 ($N_{(2)}$–$N_{(3)}$ 2·5 Å) and the other hydrogen bond is between $N_{(3)}$ of G26 and the amino group, $N_{(2)}$, of the same base, i.e. G2 ($N_{(3)}$–$N_{(2)}$ 2·9 Å). This second hydrogen bond may be very weak in the native structure, since the distance between G26 $N_{(3)}$ atom and the amino $N_{(2)}$ of symmetry-related G2 for that structure is 3·6 Å and therefore was not discussed there. Two phosphate groups of 2-fold related G26 nucleotides are held together by what was assigned as a water molecule, which can also be an ion, such as magnesium, for the same reasons as in the native structure.

As for the cytosine, C1, it does not extend as far into the minor groove of the neighboring molecule as in the native tridecamer and so the hydrogen bond between its amino $N_{(4)}$ to $N_{(3)}$ of G15 is very weak ($N_{(4)}$–$N_{(3)}$, 3·5 Å). It is also in a *syn* conformation with the carbonyl $O_{(2)}$ atom hydrogen bonding to the free 5′ hydroxyl group.

Stacking between helices bears a close resemblance to the native tridecamer in this aspect of the structure as well. The twist angle of one end (G13–C14) is virtually the same (28°) with a somewhat larger rise of 3·6 Å, and for the other end (G2–C25) the twist is very close ($-59°$ *versus* $-66°$), again with a slightly larger rise of 3·5 Å.

(iii) *Helix parameters*

Many of the structural parameters of the proflavin-soaked tridecamer are more regular than the native structure and it is even more similar to the Dickerson–Drew dodecamer, see for example the behavior of roll and tilt angles of the three structures in Figure 21. The minor groove is not as narrow as in the native tridecamer, with a width of 9·4 Å between phosphorus atoms across the groove of the A·T-rich region, which is even wider than in the dodecamer. Two of the A·T base-pairs exhibit an increased propeller twist but not as high as in the native structure and, in fact, more similar to the dodecamer ($\sim 18°$), the other two A·T base-pairs have less of a propeller twist ($\sim 10°$) as does the native structure (see Fig. 17).

Base stacking, in most cases, is very similar to the native structure. Differences do exist, however. Stacking of A17* (from another duplex) with C23 is improved. The AT step between A7·T21 and T8·A20 is different. There are small differences in the GA step between G5·C23 and A6·T22, the TC step between T9·A19 and C10·G18 and the GC step between G11·C16 and C12·G15. The large *y* displacement that was apparent for the last CG step between C12·G15 and G13·C14 in the native structure is considerably smaller in this structure. The base-pair stack flanking the bulge looks, again, like a normal CG step.

The double helix of the proflavin-soaked tridecamer is bent by about 19°, which is similar to the native tridecamer and to the dodecamer. But the bend, which as described above, is an effect of the roll angles between base-pairs and is more gradual than in the native tridecamer. Roll angles along the sequence are plotted in Figure 21(a) together with those of the native tridecamer and the dodecamer for comparison. Tilt angles are very similar to the dodecamer in this case and are relatively small (Fig. 21(b)). The large tilt resulting from the "intercalation" of the looped-out base into the helix in the native structure disappears here, since the bulge is in a better orientation to interact with the stacked-in extra base A4. The *y* displacement at the top part of the helix, namely at base-pair C12·G15, is not as

large as in the tridecamer, but the x displacement of base-pairs C3·G24 and G2·C25 towards the major groove is similar (see Table 6).

(iv) *Solvent molecules*

For this structure, only 11 solvent molecules were located. We find the same bridging solvent (W10) between phosphate groups of symmetry-related G26 nucleotides as found in the native tridecamer. Here too, peak height and the close distance to the oxygen atoms suggest that this is an ion rather than a water molecule. Another solvent peak was located just above these phosphate groups (W4). There are three other water molecules, W3, W5 and W7, which are involved in hydration of the backbone, one of which (W7) bridges between a phosphate oxygen atom and $N_{(7)}$ of A4. Three water molecules are located in the major groove, W2, W9 and W11, and one, W6, hydrates the sugar $O_{(4')}$ and $N_{(3)}$ of G13. Two water molecules, W1 and W8, which are located in the minor groove, are typical for *B*-DNA structures and belong to the so-called "spine of hydration". They bridge between thymine $O_{(2)}$ and adenine $N_{(3)}$ of the opposite strand and the adjacent base-pair at the AA and TT steps; however, a water molecule that would normally bridge the two (at the AT step) was not located here.

4. Conclusions

The most surprising observation that can be drawn from the crystal structures of the tridecamer and the proflavin-soaked tridecamer is both the flexibility and the overwhelming resiliency of the *B*-DNA double helix. Mismatches, in general, can be accommodated with only small deviations from geometry of normal Watson–Crick helices (Kennard & Hunter, 1989). It has been shown that placing even a bulky purine-purine G·A mismatch into the DNA double helix, with bases in the *anti* conformation, does not create a bulge in the structure but rather can be accommodated by the base-pair being pushed further into the major groove (Privé *et al.*, 1987, 1991). In the tridecamer, although the perturbation is potentially even larger, because of the extra nucleotide, the double helix is flexible enough to accommodate the bulge without a major disruption of the structure. Deviations from regularity are reminiscent of those observed in other *B*-DNA helices. Even locally, bases flanking the looped-out base stack as if they belong to consecutive nucleotides in a normal *B*-DNA geometry. This, as well as other deformations caused by packing effects, i.e. interactions with other duplexes in the crystal lattice by stacking interactions or hydrogen bonding interactions, emphasize the inherent flexibility of DNA in the *B*-form as well as its overall conformational resilience.

Novel features are evident in this nucleic acid structure that enhance our understanding of the DNA molecule and the processes in which it is involved. Among these are the two different conformations of looped-out bulges seen in detail and the types of interactions that may be involved in their stabilization. The intercalation of a nucleotide from one DNA molecule into another is an example of an unexpected way to straighten out and stabilize a double helix containing a stacked-in bulge. Although in this case it is another DNA molecule that binds to and interacts with the bulge site and alters its conformation, it is conceivable that other molecules, i.e. amino acid residues, drug molecules, etc., could mimic this effect. These could, in principle, interact with the bulges in a similar fashion, i.e. hydrogen bonding and stacking, shedding new light on a mechanistic view of frameshift mutations. Indeed, ethidium and acridines, both of which bind preferentially to bulge-containing sequences (Nelson & Tinoco, 1985; Woodson & Crothers, 1988*a*) are aromatic molecules capable of both hydrogen bonding and stacking.

Extra purines are stacked into the double helix, based on n.m.r. studies in solution (see Table 1). The two crystal structures described here show that they may also loop out, particularly if there are some stabilizing factors. Based on our results, it is not yet possible to determine *a priori* whether an extra base will be looped out or stacked in to the double helix, regardless of the nature of the extra base. The conformation that bulges adopt seems to be very much dependent on other molecular interactions. It may depend also on other environmental factors, such as temperature, as seen in n.m.r. structures. An example may be the interaction between a bulge site and a polymerase. Although we do not know the detailed conformation of the bulge site in the context of the polymerase, we do know that the same DNA sequence may have different frameshift error rates for different polymerases (Kunkel & Soni, 1988). This implies that there is a different conformation, or different interactions of the proposed extra base with the different polymerases, which may arise by protein-induced conformational changes in the DNA. This feature, in which DNA undergoes conformational changes upon interaction with proteins is observed in structures of DNA–protein complexes (Kim *et al.*, 1990; Suck *et al.*, 1988) and accentuates the structurally flexible and dynamic nature of this molecule. Based on the conformations that bulge sites could adopt, a mechanism for frameshift mutagenesis has been proposed (Berman, Sussman, Joshua-Tor, Revich & Ripley, personal communication).

In conclusion, this work has dealt with the conformation of extra bases in DNA, the effect they have on the overall structure of the DNA molecule and implications for frameshift mutagenesis. Both the structure of the tridecamer and of the proflavin-soaked tridecamer have one of their bulged adenine bases looping out from the double helix. They intercalate into the other bulge site of neighboring helices and are stabilized both by stacking interactions and by hydrogen bonding *via* formation of an A·A base-pair with the extra stacked adenine base. The conformation of the bulge is somewhat different for the two structures, but the inter-

molecular interactions and their affect on the structure of the double helix are basically the same. In one case, the backbone makes a loop-the-loop curve at the bulge site, flipping the sugar to the opposite direction, while in the other case, the loop at the backbone lies parallel to the base-pairs. Both allow the bases, above and below the looped-out base, to be stacked upon each other as if they belonged to consecutive nucleotides. The stacked-in extra base is stabilized by an "intercalator", which happens to be the looped-out base from another molecule. If the helix were kinked, as found in n.m.r. and gel mobility studies, the intercalation causes it to straighten out. These observations manifest the conformational flexibility and resilience of the DNA molecule. In addition, the body of structural information collected on bulge-containing DNA fragments from the n.m.r. solution structures described in the Introduction and the crystallographic studies presented here show that the conformation of the bulge itself and of the double helix carrying this perturbation is dependent on several factors. These include the nature of the bulged nucleotide, the sequence context in which the bulge lies, on environmental factors and on its possibility to undergo various interactions with other factors in its surroundings.

DNA has come a long way from being thought of as a beautiful but rather repetitive and boring molecule, to what is now seen as a dynamic and fine-structured entity, not yet disclosing all of its secrets to us.

We thank Drs Dinshaw Patel, Craig Ogata, William Weis, Helen Berman, Christian Oefner and Zvi Livneh for valuable discussions in various stages of this research, Dr Wayne Hendrickson for use of computing facilities, and Drs Philip Gottlieb and Douglas Rees for critical reviews of the manuscript. Coordinates (code name 1D31) were deposited in the Protein Data Bank, Brookhaven National Laboratory (Bernstein *et al.*, 1977), and in the Nucleic Acid Database, Rutgers University, NJ. This work was supported by grants (to J.L.S.) from the United States-Israel Binational Science Foundation (BSF), Jerusalem, Israel and the Kimmelman Center for Biomolecular Structure and Assembly, Rehovot, Israel.

References

Arnott, S., Chandrasekaran, R., Birdsall, D. L., Leslie, A. G. W. & Ratliff, R. L. (1980). Left-handed DNA helices. *Nature (London)*, **283**, 743–745.

Berman, H. M., Stallings, W., Carrell, H. L., Glusker, J. P., Neidle, S., Taylor, G. & Achari, A. (1979). Molecular and crystal structure of an intercalation complex: proflavin-cytidylyl-(3′,5′)-guanosine. *Biopolymers*, **18**, 2405–2429.

Bernstein, F. C., Koetzel, T. F., Williams, G. J. B., Meyer, E. F., Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Schimanouchi, T. & Tasunmi, M. (1977). The protein database: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535–542.

Bhattacharyya, A. & Lilley, D. M. (1989). The contrasting structure of mismatched DNA sequences containing looped-out bases (bulges) and multiple mismatches (bubbles). *Nucl. Acids Res.* **17**, 6821–6840.

Brenner, S., Barnett, L., Crick, F. H. C. & Orgel, L. (1961). The theory of mutagenesis. *J. Mol. Biol.* **3**, 121–124.

Brünger, A. T., Kuriyan, J. & Karplus, M. (1987). Crystallographic R factor refinement by molecular dynamics. *Science*, **235**, 458–460.

Camerman, N., Fawcett, J. K. & Camerman, A. (1976). Molecular structure of a deoxyribose-dinucleotide, sodium thymidylyl-(5′-3′)-thymidylate-(5′) hydrate (pTpT), and a possible structural model for polythymidylate. *J. Mol. Biol.* **107**, 601–621.

Cech, T. R. (1990). Self-splicing of group I introns. *Annu. Rev. Biochem.* **59**, 543–568.

Coll, M., Frederick, C. A., Wang, A. H.-J. & Rich, A. (1987*a*). A bifurcated hydrogen-bonded conformation in the d(A·T) base-pairs of the DNA dodecamer d(CGCGAAATTTGCG) and its complex with distamycin. *Proc. Nat. Acad. Sci., U.S.A.* **84**, 8385–8389.

Coll, M., Solans, X., Font-Altaba, M. & Subirana, J. A. (1987*b*). Crystal and molecular structure of the sodium salt of the dinucleotide duplex d(CpG). *J. Biomol. Struct. Dynam.* **4**, 797–811.

Cruse, W. B. T., Salisbury, S. A., Brown, T., Cosstick, R., Eckstein, F. & Kennard, O. (1986). Chiral phophorothioate analogues of *B*-DNA . The crystal structure of Rp-d[Gp(S)CpGp(S)CpGp(S)C]. *J. Mol. Biol.* **192**, 891–905.

Dickerson, R. E. & Drew, H. R. (1981). Structure of a *B*-DNA dodecamer. II. Influence of base sequence on helix structure. *J. Mol. Biol.* **149**, 761–786.

Dickerson, R. E., Kopka, M. L. & Pjura, P. (1985). In *Biological Macromolecules and Assemblies* (Jurnak, F. & McPherson, A., eds), vol. 2, pp. 471–494. Wylie, New York.

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., Saenger, W., Shakked, Z., Sklenar, H., Soumpasis, D. M., Tung, C.-S., Wang, A. H.-J. & Zhurkin, V. B. (1989). Definitions and nomenclature of nucleic acid structure components. *J. Mol. Biol.* **206**, 787–791.

DiGabriele, A. D., Sanderson, M. R. & Steitz, T. A. (1989). Crystal lattice packing is important in determining the bend of a DNA dodecamer containing an adenine tract. *Proc. Nat. Acad. Sci., U.S.A.* **86**, 1816–1820.

Drew, H. R. & Dickerson, R. E. (1981). Structure of a *B*-DNA dodecamer. III. Geometry of hydration. *J. Mol. Biol.* **151**, 535–556.

Einspahr, H., Cook, W. J. & Bugg, C. E. (1981). Conformational flexibility in single-stranded oligonucleotides: crystal structure of a hydrated calcium salt of adenylyl-(3′-5′)-adenosine. *Biochemistry*, **20**, 5788–5794.

Evans, D. H. & Morgan, A. R. (1982). Extrahelical bases in duplex DNA. *J. Mol. Biol.* **160**, 117–122.

Evans, D. H. & Morgan, A. R. (1986). Characterization of imperfect DNA duplexes containing unpaired bases and non-Watson–Crick base-pairs. *Nucl. Acids Res.* **14**, 4267–4280.

Fink, T. R. & Crothers, D. M. (1972*a*). Free energy of imperfect nucleic acid helices. I. The bulge defect. *J. Mol. Biol.* **66**, 1–12.

Fink, T. R. & Crothers, D. M. (1972*b*). Local effects of partial adenine-N1-oxidation on poly A–poly U and poly A–2 poly U helix conformations. *Biopolymers*, **11**, 127–136.

Fratini, A. V., Kopka, M. L., Drew, H. R. & Dickerson, R. E. (1982). Reversible bending and helix geometry

in a *B*-DNA dodecamer: CGCGAATT^{Br}CGCG. *J. Biol. Chem.* **257**, 14686–14707.

Fresco, J. R. & Alberts, B. M. (1960). The accommodation of non-complementary bases in helical polyribonucleotides and deoxyribonucleic acids. *Proc. Nat. Acad. Sci., U.S.A.* **46**, 311–321.

Gorenstein, D. G., Schroeder, S. A., Miyasaki, M., Fu, J. M., Roongta, V., Abuaf, P., Metz, J. T. & Jones, C. R. (1987). ^{32}P NMR and two-dimensional NMR spectra of nucleic acids and 2D NOESY-constrained molecular mechanics calculations for structural solution of duplex oligonucleotides. *Bull. Magn. Reson.* **8**, 137–146.

Gray, D. M. (1974). A circular dichroism study of poly dG, poly dC, and poly dG:dC. *Biopolymers.* **13**, 2087–2102.

Gray, D. M., Vaughan, M., Ratliff, R. L. & Hayes, F. N. (1980). Circular dichroism spectra show that repeating dinucleotide DNAs may form helices in which every other base is looped-out. *Nucl. Acids Res.* **8**, 3695–3707.

Gray, D. M., Cai, T. & Ratliff, R. L. (1984). Circular dichroism measurements show that C·C$^+$ base-pairs can coexist with A·T base-pairs between antiparallel strands of an oligodeoxynucleotide double-helix. *Nucl. Acids Res.* **12**, 7565–7579.

Guthrie, C. (1991). Messenger RNA splicing in yeast: clues to why the spliceosome is a ribonucleoprotein. *Science*, **253**, 109–240.

Hare, D., Shapiro, L. & Patel, D. J. (1986). Extrahelical adenosine stacks into right-handed DNA: solution conformation of the d(C-G-C-A-G-A-G-C-T-C-G-C-G) duplex deduced from distance geometry analysis of nuclear Overhauser effect spectra. *Biochemistry*, **25**, 7456–7464.

Hartel, A. J., Willie-Hazeleger, G., van Boom, J. H. & Altona, C. (1981). Conformational analysis of a modified ribotetranucleoside triphosphate: m_2^6A-U-m_2^6-A-U studied in aqueous solution by nuclear magnetic resonance at 500 MHz. *Nucl. Acids Res.* **9**, 1405–1423.

Heinemann, U. & Alings, C. (1991). The conformation of a *B*-DNA decamer is mainly determined by its sequence and not by crystal environment. *EMBO J.* **10**, 35–43.

Hendrickson, W. A. & Konnert, J. H. (1981). In *Biomolecular Structure, Function, Conformation and Evolution* (Srinivasan, R., Subramanian, E. & Yathindra, N., eds), pp. 43–57. Pergamon Press, Oxford.

Hirshberg, M., Sharon, R. & Sussman, J. L. (1988). A kinked model for the solution structure of DNA tridecamers with inserted adenosines: Energy minimization and molecular dynamics. *J. Biomol. Struct. Dynam.* **5**, 965–979.

Hope, H. (1988). Cryocrystallography of biological macromolecules: a generally applicable method. *Acta Crystallogr. sect. B*, **44**, 22–26.

Howard, A. J., Gilliland, G. L., Finzel, B. C., Poulos, T. L., Ohlendorf, D. H. & Salemme, F. R. (1987). The use of an imaging proportional counter in macromolecular crystallography. *J. Appl. Crystallogr.* **20**, 383–387.

Hsieh, C.-H. & Griffith, J. D. (1989). Deletions of bases in one strand of duplex DNA, in contrast to single-base mismatches, produce highly kinked molecules: possible relevance to the folding of single-stranded nucleic acids. *Proc. Nat. Acad. Sci., U.S.A.* **86**, 4833–4837.

Jeffrey, G. A. & Mitra, J. (1984). Three-centered (bifurcated) hydrogen bonding in structures of amino acids. *J. Amer. Chem. Soc.* **106**, 5546–5553.

Jones, T. A. (1978). A graphics model building and refinement system for macromolecules. *J. Appl. Crystallogr.* **11**, 268–272.

Joshua-Tor, L. (1991). Three-dimensional structures of bulge-containing DNA fragments. Ph.D. thesis, Weizmann Institute of Science.

Joshua-Tor, L., Rabinovich, D., Hope, H., Frolow, F., Appella, E. & Sussman, J. L. (1988). The three-dimensional structure of a DNA duplex containing looped out bases. *Nature (London)*, **334**, 82–84.

Kalnik, M. W., Norman, D. G., Swann, P. F. & Patel, D. J. (1989*a*). Conformational of adenosine bulge-containing deoxytridecanucleotide duplexes in solution: extra adenosine stacks into duplex independent of flanking sequence and temperature. *J. Biol. Chem.* **264**, 3702–3712.

Kalnik, M. W., Norman, D. G., Zagorski, M. G., Swann, P. F. & Patel, D. J. (1989*b*). Conformation transitions in cytidine bulge-containing deoxytrideca-nucleotide duplexes: extra cytidine equilibrates between looped out (low temperature) and stacked (elevated temperature) conformations in solution. *Biochemistry*, **28**, 294–303.

Kalnik, M. W., Norman, D. G., Li, B. F., Swann, P. F. & Patel, D. J. (1990). Conformational transitions in thymidine bulge-containing deoxytridecanucleotide duplexes. *J. Biol. Chem.* **265**, 636–647.

Kennard, O. & Hunter, W. N. (1989). Oligonucleotide structure: a decade of results from single crystal X-ray diffraction studies. *Quart. Rev. Biophys.* **22**, 327–379.

Kim, S. H. & Sussman, J. L. (1976). π turn is a conformational pattern in RNA loops and bends. *Nature (London)*, **260**, 645–646.

Kim, Y., Grable, J. C., Love, R., Greene, P. J. & Rosenberg, J. M. (1990). Refinement of EcoRI endonuclease crystal structure: a revised protein chain tracing. *Science*, **249**, 1307–1309.

Kunkel, T. A. (1990). Misalignment-mediated DNA synthesis errors. *Biochemistry*, **29**, 8003–8011.

Kunkel, T. A. & Soni, A. (1988). Mutagenesis by transient misalignment. *J. Biol. Chem.* **263**, 14784–14789.

LeBlanc, D. A. & Morden, K. M. (1991). Thermodynamic characterization of deoxyribooligonucleotide duplexes containing bulges. *Biochemistry*, **30**, 4042–4047.

Lee, C. H. & Tinoco, I. (1980). Conformation studies of 13 trinucleoside diphosphates by 360 MHz PMR spectroscopy. A bulged base conformation. I base protons and H_1 protons. *Biophys. Chem.* **11**, 283–294.

Lerman, L. S. (1963). The structure of the DNA-acridine complex. *Proc. Nat. Acad. Sci., U.S.A.* **49**, 94–102.

Lomant, A. J. & Fresco, J. R. (1973). Polynucleotides XIV: Photochemical evidence for an extrahelical solvent-accessible environment of non-complementary residues in polynucleotide helices. *J. Mol. Biol.* **77**, 345–354.

Lomant, A. J. & Fresco, J. R. (1975). Structural and energetic consequences of noncomplementary base oppositions in nucleic acid helices. *Progr. Nucl. Acid Res. Mol. Biol.* **15**, 185–218.

Longfellow, C. E., Kierzek, R. & Turner, D. H. (1990). Thermodynamic and spectroscopic study of bulge loops in oligoribonucleotides. *Biochemistry*, **29**, 278–285.

McPherson, A. (1985). In *Diffraction Methods for*

Biological Macromolecules (Wyckoff, H. W., Hirs, C. H. W. & Timasheff, S. N., eds), pp. 112–120. Academic Press, New York.

Michel, F., Umesono, K. & Ozeki, H. (1989). Comparative and functional anatomy of group II catalytic introns — a review. *Gene*, **82**, 5–30.

Milligan, J. F. & Uhlenbeck, O. C. (1989). Determination of RNA-protein contacts using thiophosphate substitutions. *Biochemistry*, **28**, 2849–2855.

Morden, K. M., Chu, Y. G., Martin, F. H. & Tinoco, I., Jr (1983). Unpaired cytosine in the deoxyoligonucleotide duplex $dC(A)_3C(A)_3G$-$dC(T)_6G$ is outside of the helix. *Biochemistry*, **22**, 5557–5563.

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

Nelson, J. W. & Tinoco, I. (1985). Ethidium ion binds more strongly to a DNA double helix with a bulged cytosine base than to a regular double helix. *Biochemistry*, **24**, 6416–6421.

Nikonowicz, E. P., Roongta, V., Jones, C. R. & Gorenstein, D. G. (1989). Two-dimensional ^1H and ^{31}P n.m.r. spectra and restrained molecular dynamics structure of an extrahelical adenosine base tridecamer oligodeoxyribonucleotide duplex. *Biochemistry*, **28**, 8714–8725.

Nikonowicz, E. P., Meadows, R. P. & Gorenstein, D. G. (1990). n.m.r. structural refinement of an extrahelical adenosine tridecamer $d(CGCAGAATTCGCG)_2$ via a hybrid relaxation matrix procedure. *Biochemistry*, **29**, 4193–4204.

Olson, W. K., Marky, N. L., Srinivasan, A. R., Do, K. D. & Cicariello, J. (1985). In *Molecular Basis of Cancer, Part A: Macromolecular Structure, Carcinogens, and Oncogenes*, pp. 109–121. Alan R. Liss Inc., New York.

Parker, R., Siliciano, P. G. & Guthrie, C. (1987). Recognition of the TACTAAC box during mRNA splicing in yeast involves base-pairing to the U2-like snRNA. *Cell*, **49**, 229–239.

Patel, D. J., Kozlowski, S. A., Marky, L. A., Rice, J. A., Broka, C., Itakura, K. & Breslauer, K. J. (1982). Extra adenosine stacks into the self-complementary d(CGCAGAATTCGCG) duplex in solution. *Biochemistry*, **21**, 445–451.

Pflugrath, J. W., Saper, M. A. & Quiocho, F. A. (1984). In *Methods and Applications in Crystallographic Computing* (Hall, S. & Ashiaka, T., eds), pp. 404–407. Clarendon Press, Oxford.

Privé, G. G., Heinemann, U., Chandrasegaran, S., Kan, L.-S., Kopka, M. L. & Dickerson, R. E. (1987). Helix geometry, hydration, and G·A mismatch in a *B*-DNA decamer. *Science*, **238**, 498–504.

Privé, G. G., Yanagi, K. & Dickerson, R. E. (1991). Structure of the *B*-DNA decamer C-C-A-A-C-G-T-T-G-G and comparison with isomorphous decamers C-C-A-A-G-A-T-T-G-G and C-C-A-G-G-C-C-T-G-G. *J. Mol. Biol.* **217**, 177–199.

Rabinovich, D. & Shakked, Z. (1984). A new approach to structure determination of large molecules by multidimensional search methods. *Acta Crystallogr. sect. A*, **40**, 195–200.

Rice, J. A. & Crothers, D. E. (1989). DNA bending by the bulge defect. *Biochemistry*, **28**, 4512–4516.

Ripley, L. S. (1990). Frameshift mutation: determinants of specificity. *Annu. Rev. Genet.* **24**, 189–213.

Roy, S., Sklenar, V., Appella, E. & Cohen, J. S. (1987). Conformational perturbation due to an extra adenosine base in a self-complementary oligodeoxynucleotide duplex. *Biopolymers*, **26**, 2041–2052.

Saenger, W. (1984). *Principles of Nucleic Acid Structure*. Springer-Verlag, New York.

Saper, M. A., Eldar, H., Mizuuchi, K., Nickol, J., Appella, E. & Sussman, J. L. (1986). Crystallization of a DNA tridecamer d(C-G-C-A-G-A-A-T-T-C-G-C-G). *J. Mol. Biol.* **188**, 111–113.

Shakked, Z. & Rabinovich, D. (1986). The effect of the base sequence on the fine structure of DNA double helix. *Progr. Biophys. Mol. Biol.* **47**, 159–195.

Steigemann, W. (1974). Ph.D. thesis. Technische Universitaet, Munchen.

Streisinger, G., Okada, Y., Emrich, J., Newton, J., Tsugeta, A., Terzaghi, E. & Inouye, M. (1966). Frameshift mutations and the genetic code. *Cold Spring Harbor Symp. Quant. Biol.* **31**, 77–84.

Suck, D., Manor, P. C., Germain, G., Schwalbe, C. H., Weimann, G. & Saenger, W. (1973). X-ray study of helix, loop and base pair stacking in trinucleoside diphosphate ApApA. *Nature New Biol.* **246**, 161–165.

Suck, D., Lahm, A. & Oefner, C. (1988). Structure refined to 2·0 Å of a nicked DNA octanucleotide complex with DNase I. *Nature (London)*, **332**, 465–468.

Sussman, J. L., Seeman, N. C., Kim, S. H. & Berman, H. M. (1972). The Crystal structure of a naturally occurring dinucleotide phosphate uridylyl 3′,5′-adenosine phosphate. Models for RNA chain folding. *J. Mol. Biol.* **66**, 403–421.

Sussman, J. L., Holbrook, S. R., Church, G. M. & Kim, S.-H. (1977). A structure-factor least squares refinement procedure for macromolecular structures using constrained and restrained parameters. *Acta Crystallogr. sect. A*, **33**, 800–804.

van den Hoogen, Y. T., van Beuzekom, A. A., de Vroom, E., van der Marel, G. A., van Boom, J. H. & Altona, C. (1988*a*). Bulge-out structures in the single-stranded trimer AUA and in the duplex (CUGGUGCGG)·(CCGCCCAG). A model-building and NMR study. *Nucl. Acids Res.* **16**, 5013–5030.

van den Hoogen, Y. T., van Beuzekom, A. A., van den Elst, H., van der Marel, G. A., van Boom, J. H. & Altona, C. (1988*b*). Extra thymidine stacks into the d(CTGGTGCGG)·d(CCGCCCAG) duplex. An NMR and model-building study. *Nucl. Acids Res.* **16**, 2971–2986.

Viswamitra, M. A., Kennard, O., Jones, P. G., Sheldrick, G. M., Salisbury, S., Falvello, L. & Shakked, Z. (1978). DNA double helical fragment at atomic resolution. *Nature (London)*, **273**, 687–688.

Wang, Y.-H. & Griffith, J. D. (1991). Effects of bulge composition and flanking sequence on the kinking of DNA by bulged bases. *Biochemistry*, **30**, 1358–1363.

Westhof, E., Dumas, P. & Moras, D. (1985). Crystallographic refinement of yeast aspartic acid transfer RNA. *J. Mol. Biol.* **184**, 119–145.

Westhof, E., Hosur, M. V. & Sundaralingam, M. (1988). Nonintercalative binding of proflavin to Z-DNA: structure of a complex between d(5BrC-G-5BrC-G) and proflavin. *Biochemistry*, **27**, 5742–5747.

White, S. A. & Draper, D. E. (1987). Single bases bulges in small RNA hairpins enhance ethidium binding and promote an allosteric transition. *Nucl. Acids Res.* **15**, 4049–4064.

Wilson, H. R. & Al-Mukhtar, J. (1976). Structure of thymidylyl-3′,5′-deoxyadenosine. *Nature (London)*, **263**, 171–172.

Witherell, G. W. & Uhlenbeck, O. C. (1989). Specific RNA binding by Qβ coat protein. *Biochemistry*, **28**, 71–76.

Woodson, S. A. & Crothers, D. M. (1988*a*). Binding of 9-aminoacridine to bulged-base DNA oligomers from a frameshift hot spot. *Biochemistry*, **27**, 8904–8914.

Woodson, S. A. & Crothers, D. M. (1988*b*). Preferential location of bulged-guanosine internal to a G-C tract by ^{1}H-NMR. *Biochemistry*, **27**, 436–445.

Woodson, S. A. & Crothers, D. M. (1988*c*). Structural model for an oligonucleotide containing a bulged guanosine by n.m.r. and energy minimization. *Biochemistry*, **27**, 3130–3141.

Woodson, S. A. & Crothers, D. M. (1989). Conformation of a bulge-containing oligomer from a hot spot sequence by NMR and energy minimization. *Biopolymers*, **28**, 1149–1177.

Wu, H. N. & Uhlenbeck, O. C. (1987). Role of a bulged A residue in a specific RNA-protein interaction. *Biochemistry*, **26**, 8221– 8227.

Yanagi, K., Privé, G. G. & Dickerson, R. E. (1991). Analysis of local helix geometry in three *B*-DNA decamers and eight dodecamers. *J. Mol. Biol.* **217**, 201–214.

Edited by P. E. Wright