Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Use of an N-terminal fragment from Moloney murine leukemia virus reverse transcriptase to facilitate crystallization and analysis of a pseudo-16-mer DNA molecule containing G–A mispairs

Marie L. Coté, Sarah J. Yohannan† and Millie M. Georgiadis*

Waksman Institute and Department of Chemistry, Rutgers University, 190 Frelinghuysen Road, Piscataway, NJ 08854, USA

† Present address: Department of Chemistry and Biochemistry, Molecular Biology Institute, University of California, Los Angeles, Los Angeles, CA, USA.

Correspondence e-mail: georgiadis@mbcl.rutgers.edu

Received 6 April 2000
Accepted 5 June 2000

PDB Reference: Maloney murine leukemia virus reverse transcriptase–nucleic acid complex, 1d1u.

Complexation with the N-terminal fragment of Moloney murine leukemia virus reverse transcriptase offers a novel method of obtaining crystal structures of nucleic acid duplexes, which can be phased by molecular replacement. This method is somewhat similar to the method of using a monoclonal antibody Fab fragment complexed to the molecule of interest in order to obtain crystals suitable for X-ray crystallographic analysis. Here a novel DNA structure including two G–A mispairs in a pseudo-hexadecamer determined at 2.3 Å resolution in a complex with the N-terminal fragment is reported. This structure has an asymmetric unit consisting of the protein molecule bound to the blunt end of a DNA 6/10-mer, which is composed of a six-base strand (5′-CTCGTG-3′) and a ten-base strand (3′-GAGCACGGCA-5′). The 6/10-mer is thus composed of a six-base-pair duplex with a four-base single-stranded overhang. In the crystal structure, the bases of the overhang are reciprocally paired (symmetry element $-x - 1, -y, z$), yielding a doubly nicked pseudo-hexadecamer primarily B-form DNA molecule, which has some interesting A-like structural features. The pairing between the single strands results in two standard (G–C) Watson–Crick pairs and two G–A mispairs. The structural DNA model can accommodate either a standard *syn* or a standard *anti* conformation for the 5′-terminal adenine of the ten-base strand of the DNA based on analysis of simulated-annealing omit maps. Although the DNA model here includes nicks in the phosphodiester backbone, modeling of an intact phosphodiester backbone results in a very similar DNA model and indicates that the structure is biologically relevant.

1. Introduction

A novel crystallographic approach to the analysis of nucleic acid structure has resulted from structural studies directed toward understanding processive DNA synthesis by Moloney murine leukemia virus reverse transcriptase (MMLV RT). The approach involves complexation of the nucleic acid sequence of interest to the N-terminal fragment of MMLV RT and subsequent crystallographic analysis of the resulting complex. As previously reported, the fragment includes the fingers and palm domains of MMLV RT and binds nucleic acid in a novel site in the fingers domain and not in the polymerase active site. Binding of the nucleic acid duplex to the fingers domain involves the highly conserved residues Asp114, Arg116, Asn119 and Gly191, and may play a mechanistic role in processive DNA synthesis by RT (Najmudin *et al.*, 2000). The

DNA atoms interacting with the protein include minor-groove base atoms and sugar atoms from the $n - 2$ and $n - 3$ template-strand positions as well as the 3′-OH of the primer strand, where n is the template base that would pair with an incoming nucleotide. Thus, interactions are limited to the three terminal base pairs of the duplex. Both ends of the oligonucleotide duplex contact the protein and the intervening nucleotides are free of the sort of contacts that are seen in crystal structures of DNA only. Also, the interactions observed between the protein and the DNA would be possible for any DNA or RNA sequence.

We have determined crystal structures for three different DNA duplexes complexed to the N-terminal fragment in three distinct crystal lattices and find that the protein–DNA interactions are quite similar in each case (Najmudin *et al.*, 2000). The crystal lattices include monoclinic forms I and II having unit-cell parameters a = 61.14, b = 38.43, c = 129.75 Å, β = 100.58° and a = 65.87, b = 63.59, c = 73.40 Å, β = 102.91°, respectively. The third lattice is orthorhombic (form IV), with unit-cell parameters a = 54.74, b = 145.49, c = 46.74 Å. The oligonucleotide sequences include 3′-CATGCATG-5′ crystallized in form I and II lattices, 3′-TTTCATGCATG-5′ in the form II lattice and the aforementioned 6/10-mer sequences in the form IV lattice. For historical reasons, the third lattice described above is referred to as form IV. We previously reported preliminary crystallographic studies on form III crystals (Sun *et al.*, 1998). The lattices that we have obtained to date include DNA duplexes of 8-mers with and without single-stranded overhangs in addition to a pseudo-16-mer. Fig. 1 shows DNA associated with its protein molecules in two of the crystal lattices that we have obtained. The form IV lattice shown in Fig. 1(*b*) contains the pseudo-hexadecamer, which has two G–A mispairs resulting from pairing of single-stranded overhangs and is an example of a novel DNA structure. The focus of the present study is the analysis of the pseudo-hexadecamer and of the G–A mispairs contained in the DNA molecule.

The G–A mispair is intriguing since it is known to be the most common cause of transversion mutations. Fersht *et al.* (1982) showed that the G–A mispair is incorporated into DNA with a frequency only eight times less often than the most frequent mispair (G–T); it is the most prevalent purine–purine mispair. The G–A mispair is also far less efficiently repaired than are other mispairs, as seen in experiments with simian kidney cells (Brown & Jiricny, 1988). In addition, the G–A mispair is interesting structurally since either purine is capable of adopting a *syn* conformation with respect to its sugar moiety. Theoretical and NMR studies have indicated that five possible conformational motifs of G–A mispairs may occur in rRNA (Chuprina & Poltev, 1983; Poltev & Shulyupina, 1986; Gautheret *et al.*, 1994). However, in X-ray crystallographic structures of otherwise unmodified B-DNA containing G–A mispairs, only the four following conformations have been reported: the standard (non-sheared) G(*anti*)–A(*anti*) (Privé *et al.*, 1987), the sheared G(*anti*)–A(*anti*) (Shepard *et al.*, 1998; Gao *et al.*, 1999), G(*anti*)–A(*syn*) (Brown *et al.*, 1986; Webster *et al.*, 1990) and G(*syn*)–A(*anti*) (Brown *et al.*, 1989). The conformational variabilities observed in the G–A mispair have been attributed to pH, structure, thermodynamic, hydrogen-bonding and base-stacking effects. The G–A mispaired dodecamer d(CGCR_1AATTR_2GCG)$_2$ results in a G(*syn*)–A(*anti*) conformation when R_1 = A and R_2 = G and the pH is 6.6 (Brown *et al.*, 1989; Leonard *et al.*, 1990); when R_1 = G and R_2 = A and the pH is 7.4, the G(*anti*)–A(*syn*) conformation results (Brown *et al.*, 1986). Theoretical studies indicate that there is roughly a 1 kcal mol^{-1} energy difference between the G(*anti*)–A(*anti*) and the G(*anti*)–A(*syn*) conformations (Chuprina & Poltev, 1983; Keepers *et al.*, 1984; Poltev & Shulyupina, 1986).

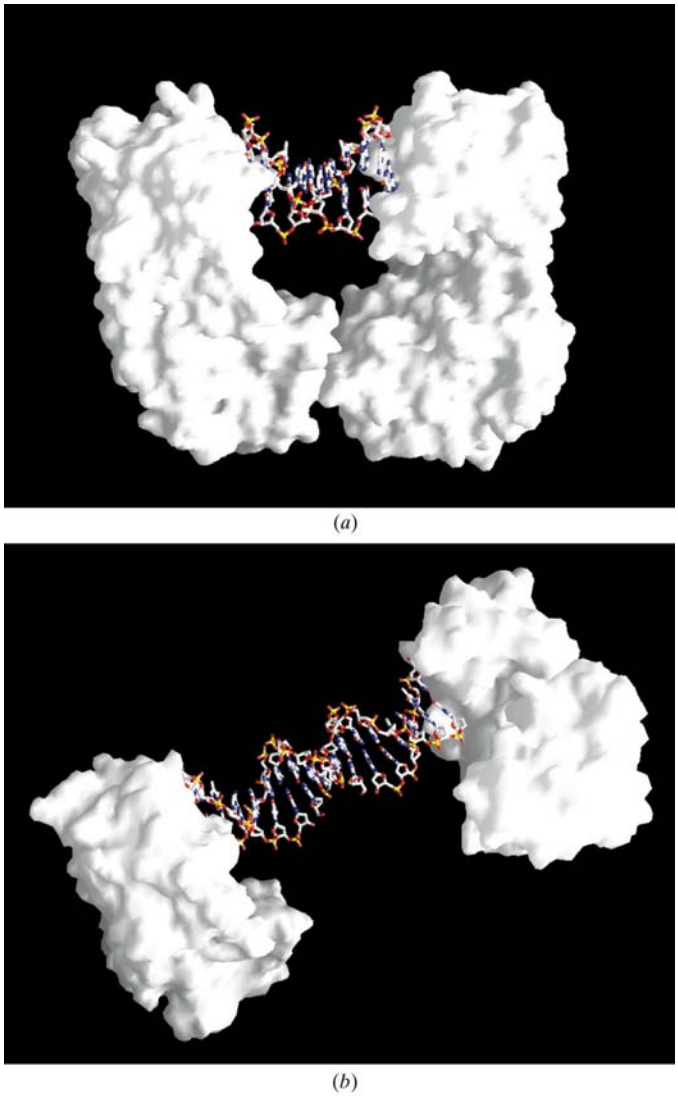

Figure 1
Molecular-surface renderings (Nicholls *et al.*, 1991) of the N-terminal fragment of MMLV RT with the intervening DNA shown as stick models. The protein is colored in white; the atoms of the DNA are colored according to type: phosphorus is yellow, oxygen is red and nitrogen is blue. (*a*) The form II crystal structure has distinct 'A' and 'B' protein molecules, and the associated 8/8-mer DNA lies between them. (*b*) The form IV structure has one protein molecule and one DNA 6/10-mer in its asymmetric unit. The reciprocally paired 6/10-mers form the pseudo-hexadecamer, as shown in the figure.

X-ray crystal structures of B-DNA containing a modified G or A within the mispair seem to indicate that base sequence in concert with hydrogen bonding may direct the G–A mispair conformation. In these modified structures, one of the bases of the G–A mispair will generally adopt the *syn* conformation. In the d(CGCR_1AATTR_2GCG)$_2$ dodecamer, if R_1 = A and R_2 = (O8)G, the modified mispair conformation is (O8)G(*syn*)–A(*anti*), the same conformation as that observed in the unmodified dodecamer, albeit with an additional C—H···O contact to the O8 (McAuley-Hecht *et al.*, 1994). A similar result is seen in the X-ray crystal structure of the dodecamer when it is changed to R_1 = G and R_2 = (O8)A. In that structure, the modified mispair conformation is G(*anti*)–(O8)A(*syn*), the same as that observed in the unmodified dodecamer, albeit with greater hydrogen bonding owing to the creation of two pseudo-symmetric reverse three-center hydrogen bonds (Leonard *et al.*, 1992). Again in the same dodecamer sequence, when R_1 = G and R_2 = (εd)A, where (εd)A = 1,N^6-ethenoadenosine, the crystal structure DNA has its G–A mispairs in the G(*anti*)–(εd)A(*syn*) conformation. In that structure, the creation of an additional nitrogen acceptor on the (εd)A and the ability to form an extra C—H···O contact to O6 of guanine ensures the already predisposed mispair conformation (Leonard *et al.*, 1994). The crystal structure of the modified d[CGAGAATTC(O^6Me)GCG]$_2$ has its mispairs in the (O^6Me)G(*anti*)–A(*syn*) conformation (Ginell *et al.*, 1994).

Here, we report the helical properties of the pseudo-hexadecameric DNA duplex structure, as well as specific details regarding the G–A mispairs, the effects of contacts with the protein and the deoxyribose-ring conformations. In addition, the advantages and limitations of the approach for obtaining novel nucleic acid structures complexed with the N-terminal fragment of MMLV RT are discussed.

2. Materials and methods

2.1. Crystallization and data collection

The bacterially expressed N-terminal fragment from Moloney murine leukemia virus reverse transcriptase and oligonucleotides were purified as previously described (Sun *et al.*, 1998). The crystals were obtained from a 1:2:8 molar ratio of protein:DNA:ddCTP. The complex of protein, DNA and ddCTP was formed at 277 K for 1.5 h, with final concentrations of 0.45 m*M* protein, 0.9 m*M* oligonucleotide and 3.6 m*M* ddCTP. The nucleotide was originally included with the goal of obtaining a ternary complex. However, we later determined that the single-stranded overhangs are base-paired in this crystal lattice. Thus, the nucleotide was not required for crystallization and is not bound to the enzyme in the crystal structure of the complex. We have since obtained the same crystals without added nucleotide in the crystallizations (Coté & Georgiadis, unpublished results). Crystals of the fragment–DNA complex were obtained at 293 K using 1 µl each of complex solution and a crystallization solution consisting of 10% PEG 4000, 0.10 *M* NaCl and 0.05 *M* ADA pH 6.5 in vapour-diffusion hanging drops. The plate-like 250 × 200 × 60 µm form IV crystals grew in 1–3 d. Microseeding was required in order to grow single diffraction-quality crystals.

Table 1
Data-collection statistics for form IV crystals.

Average $I/\sigma(I)$, percentage completeness and R_{sym} are given for all data in the resolution range specified. $R_{sym} = \sum|I - \langle I\rangle|/\sum I$.

Resolution (Å)	Average $I/\sigma(I)$	Completeness (%)	R_{sym}
50.0–4.95	33.5	97.5	0.034
4.95–3.93	35.1	99.5	0.051
3.93–3.44	30.0	99.3	0.063
3.44–3.12	23.4	99.2	0.079
3.12–2.90	16.5	99.2	0.114
290–2.73	11.9	98.7	0.143
2.73–2.59	8.6	99.0	0.184
2.59–2.48	7.2	99.2	0.208
2.48–2.38	5.9	98.9	0.241
2.38–2.30	4.6	98.5	0.278
Total (50.0–2.30)	22.8	98.9	0.069

Refinement statistics for *syn versus anti* complex models

	Syn	*Anti*
R_{work} (%)	22.9	23.0
R_{free} (%)	28.5	28.6
R.m.s.d. values		
Bond lengths (Å)	0.09	0.09
Bond angles (°)	1.3	1.3
Dihedrals (°)	23.8	23.8
Improper torsions (°)	1.1	1.1

X-ray crystallographic analysis was performed on an R-AXIS IV image-plate detector with Cu $K\alpha$ radiation at 108 K using an Oxford Cryocool System. Data were collected to Bragg spacings of 2.3 Å, processed with *DENZO* and scaled with *SCALEPACK* (Otwinowski, 1993) including all data. Form IV crystallizes in the space group $P2_12_12$, with unit-cell parameters a = 54.74, b = 145.49, c = 46.74 Å. The crystals were stabilized in 20% ethylene glycol, 16% PEG 4000, 0.1 *M* NaCl, 0.05 *M* ADA pH 6.5 for cryocooling. Statistics for data collected from a cryocooled crystal are given in Table 1.

2.2. Structure determination and refinement

AMoRe (Navaza, 1994) was used to obtain a molecular-replacement solution for 8–4 Å data collected at 108 K using the 'A' protein molecule from the refined structural model of form IIa crystals (Najmudin *et al.*, 2000). The Euler rotation angles (α, β, γ) and translations (x, y, z) for the final molecular-replacement solution for form IV were 337.02, 147.36, 24.13° and 0.10, 0.14, 0.07 Å, respectively. The correlation coefficient (r) and R value for this solution were 71.6 and 32.9%, respectively.

Initial rigid-body refinement using *REFMAC* (Murshudov *et al.*, 1997) with 8–4 Å data yielded R_{work} and R_{free} values of 47.3 and 52.3%, respectively. Subsequent refinement of the protein in *REFMAC* using all data in the resolution range 8–2.3 Å and overall isotropic B factors reduced R_{work} and R_{free}

to 36.0 and 40.1%, respectively. A standard B-form DNA 6/10-mer was created using *INSIGHT II* (Biosym Technologies, 1993) and then positioned and rebuilt into the well defined difference electron density. Subsequent refinement invoking positional and individual *B*-factor refinement yielded R_{work} and R_{free} values of 25.1 and 32.7%, respectively.

The addition of 175 water O atoms along with modest rebuilding yielded R_{work} and R_{free} values of 22.5 and 31.4%, respectively, for the form IV structure containing either DNA model using all data from 50–2.3 Å resolution and a bulk-solvent correction in *REFMAC*. The near 9% disparity between R_{work} and R_{free} indicated possible model bias; the model was therefore subjected to simulated annealing in *CNS* using all data in the resolution range 50–2.3 Å (Brunger *et al.*, 1998). This caused the DNA to move approximately 0.2–0.5 Å in the model, especially in the region of the single-stranded overhang of the 10-mer. Modest rebuilding and further individual *B*-factor refinement including a bulk-solvent correction in *CNS* gave final R_{work} and R_{free} values of 23.0 and 28.6%, respectively, for either the model containing the *syn* adenine or the *anti* adenine. Standard *CNS* parameter files were used in the refinement. The overall σ_A coordinate errors based on R_{work} and R_{free} are 0.28 and 0.18 Å , respectively (Read, 1986).

Because the symmetry-related pairing between the single strands resulted in two G–A mispairs, consideration was given to the possibility that either the guanine (G10) or the adenine (A7) base in the single-stranded overhang adopted a *syn* conformation with respect to its sugar moiety (see Fig. 2). For most of the refinement process, our structure retained the DNA modeled with both purines of the mispair in standard *anti* conformations since the electron-density maps easily accommodated them. In an attempt to confirm the conformations of the bases of the mispair, models incorporating the standard G(*anti*)–A(*anti*), the G(*anti*)–A(*syn*) and the G(*syn*)–A(*anti*) conformational possibilities were all separately created and subjected to simulated annealing. Any model with the mispaired guanine (G10) in the *syn* conformation could be readily ruled out, since difference electron-density maps clearly showed that a *syn* guanine in the mispair would have the majority of its base lying in a region of negative electron density (see Fig. 3*a*). Thus, no attempt was made to further analyze the structure with G10 in a *syn* conformation, and it seemed that the mispairs were either in the standard G(*anti*)–A(*anti*) or G(*anti*)–A(*syn*) conformation. Simulated-annealing omit maps were calculated for both the standard G(*anti*)–A(*anti*) and the G(*anti*)–A(*syn*) models. A firm conclusion as to the absolute conformation of the adenine could not be made from either refinement statistics (see Table 1) or electron-density maps (see Fig. 3*b*). In addition, a sheared G(*anti*)–A(*anti*) model was created and superpositioned onto the standard G(*anti*)–A(*anti*) model such that their adenine bases coincided. There is no evidence that a slip dislocation has occurred in our structure, which would yield the very rare sheared G(*anti*)–A(*anti*) conformation (see Fig. 3*c*).

Final verification of the protein model using *PROCHECK* (Laskowski *et al.*, 1993) shows 88.4% of non-Gly/Pro residues residing in most favored regions, 10.2% in additional allowed regions and 0.9% in generously allowed regions. Val223 lies in a disallowed region in form IV, as seen in the previously reported uncomplexed model of the N-terminal fragment (Georgiadis *et al.*, 1995) as well as our other complexed forms (Najmudin *et al.*, 2000). In the final model, the electron density in the $\beta4$–$\beta5$ loop region (residues 100–109) in the protein is weak and completely lacks side-chain density for Asp107. Another highly mobile loop region, $\beta8$–$\beta9$ (residues 173–187), also has weak electron density, with little side-chain density for Arg173 or Met177. Difference electron-density maps indicated that the side chain of Tyr64 should be modeled in two different conformations, each at half-occupancy. No such indication was apparent from the electron-density maps regarding possible partial occupancy for the adenine (A7) at the single-stranded terminus. The deposited coordinate file (1d1u) retains A7 in the *anti* conformation; the coordinates for the *syn* conformation are available upon request.

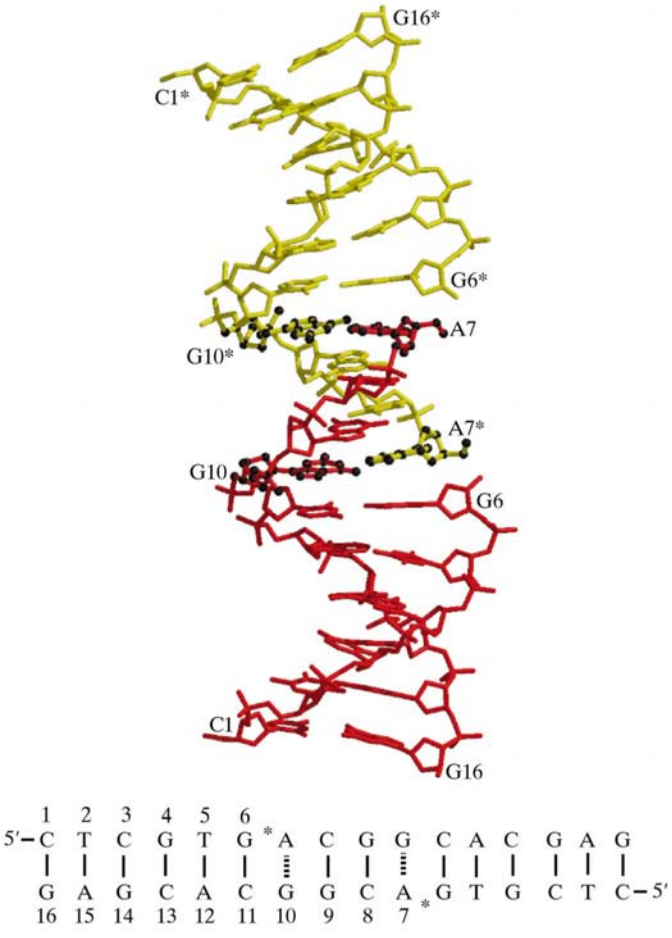

Figure 2
View (Kraulis, 1991) of the DNA pseudo-hexadecamer in the form IV structure with its numbering scheme. The two reciprocally paired 6/10-mers are shown with different coloring for emphasis. Strands (5′-CTCGTG-3′) are numbered 1–6 starting from the 5′ end and strands (3′-GAGCACGGCA-5′) are numbered 7–16, also starting from the 5′ end. Symmetry-related nucleotides are designated with asterisks. The atoms of the G–A mispairs are shown with black spheres and the bases of the mispairs are so labeled.

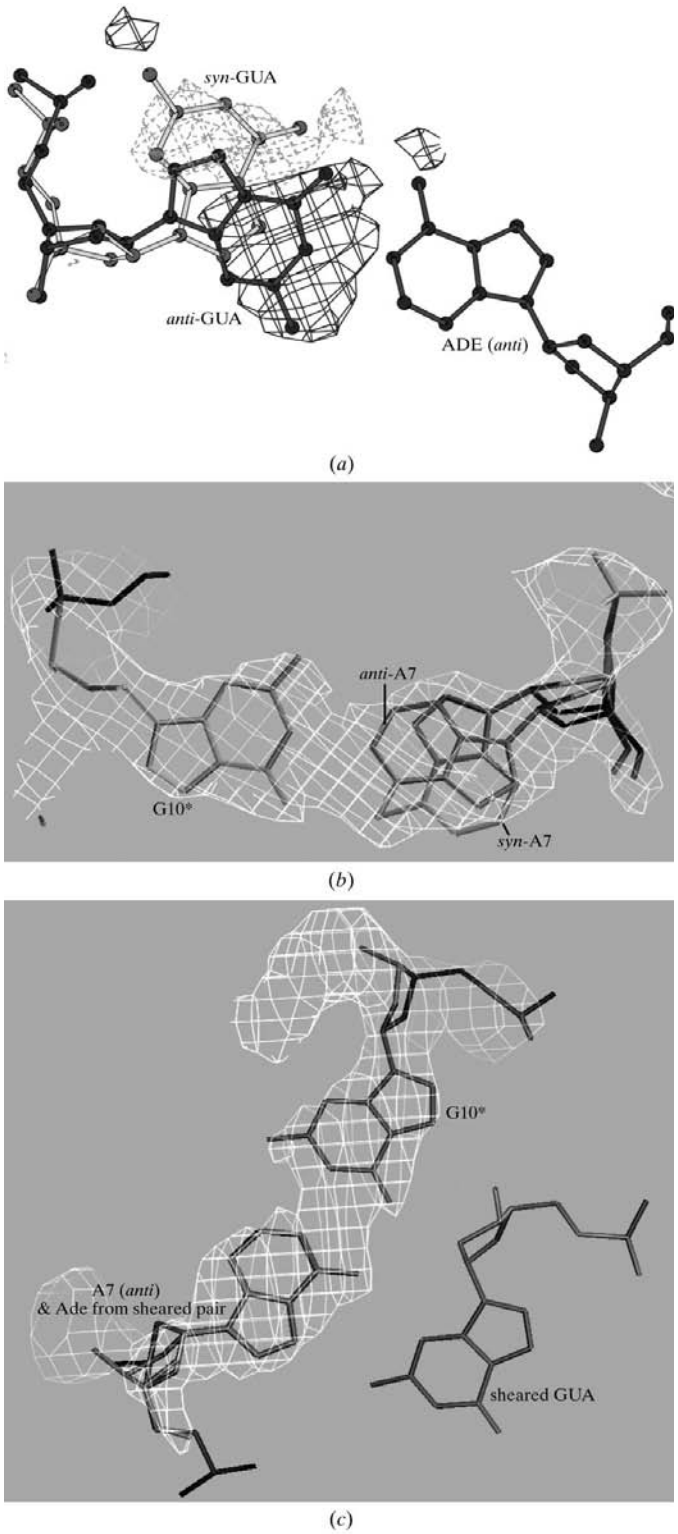

Figure 3
(*a*) $F_o - F_c$ and $2F_o - F_c$ electron-density maps of G10 in the G–A mispair. The G(*syn*) conformation is shown in light grey and its covering density is the negative contour (2.5σ) $F_o - F_c$ map, shown in dashed light-grey lines (Esnouf, 1997). The G(*anti*) conformation of G10 is shown in black and its covering density at 3.0σ is the $2F_o - F_c$ map, shown in black contiguous lines. (*b*) $2F_o - F_c$ simulated-annealing omit map (Jones *et al.*, 1991) with both standard G(*anti*)–A(*anti*) and G(*anti*)–A(*syn*) models for the mispairs in the pseudo-hexadecamer. The map is contoured at 1.5σ. (*c*) The same $2F_o - F_c$ map (Jones *et al.*, 1991) as in (*b*) showing that a sheared G–A mispair conformation does not fit the density.

3. Results and discussion

3.1. Description of the structure

The protein fragment's overall form resembles that of a semi-clenched right-hand fist, with a fingers and palm domain. We have used the same secondary-structural assignments for the protein in form IV as for the uncomplexed fragment (Georgiadis *et al.*, 1995). The DNA-binding site for the fragment is located in the fingers domain, which principally involves residues comprising the αD helix (see Fig. 4*a*). The oligonucleotide is bound to conserved residues in much the same manner as that seen in our other crystal forms (Najmudin *et al.*, 2000).

In contrast to our previously reported crystal forms (Sun *et al.*, 1998; Najmudin *et al.*, 2000), form IV has one protein molecule and one DNA molecule (the 6/10-mer) in the asymmetric unit (see Fig. 4*a*). In the crystal structure, the 6/10-mer forms a pseudo-hexadecamer through crystallographic symmetry (symmetry element $-x - 1, -y, z$), where the bases of the single-stranded overhang reciprocally pair to yield two typical Watson–Crick (G–C) base pairs and two G–A mispairs (see Figs. 2 and 4*b*). As mentioned, the terminal adenine in the single-stranded overhang may adopt either the *syn* or the *anti* conformation. The DNA model shown in Figs. 2 and 4 retains the adenine in the *anti* conformation.

3.2. Crystal packing, protein–protein and protein–DNA interactions

Based on results of surface-area calculations using *NACCESS* (Hubbard & Thornton, 1993), it is seen that a significant contribution to the packing is made by the formation of the pseudo-hexadecamer, burying a surface of 736 Å^2. Another significant packing interaction involves protein–protein contacts *via* twofold rotationally related molecules, which buries a surface of 813 Å^2. The twofold rotation places the molecules such that the palm regions are stacked palm-to-palm, with the fingers domains facing away from each other.

The DNA is bound to the fingers domain of the N-terminal fragment as shown generally in Fig. 4 and in close-up in Fig. 5. In the form IV MMLV RT fragment–DNA complex, an ion-pair forms between Asp114 and Arg116, involving atoms $O^{\delta1}\cdots N^{\varepsilon}$ and $O^{\delta2}\cdots N^{\eta2}$, similar to that found in our other crystal forms (Najmudin *et al.*, 2000). This ion-pair potentially serves to delocalize charge and to position the Arg116 into the minor groove of the oligonucleotide. The ion-pair hydrogen-bonding distances for $O^{\delta1}\cdots N^{\varepsilon}$ and $O^{\delta2}\cdots N^{\eta2}$ are both 2.8 Å. The ion-pair formed in form IV is rather planar as well, having $O^{\delta1}—H\cdots N^{\varepsilon}$ and $O^{\delta2}\cdots N^{\eta2}$ angles of 146.2 and 172.3° (idealized H atom from *CONTACT*; Collaborative Computational Project, Number 4, 1994), indicative of a strong interaction (see Fig. 5).

Protein–DNA interactions that occur in the binding site of this structure involve a significant number of interactions with sugar and base atoms of G16. First, there are strong protein–DNA binding interactions with the 3′-OH of G16, encompassing two hydrogen bonds and two other

contacts whose distances range from 3.3–3.9 Å. These interactions involve Leu115 N, Gly191 O, Arg116 N and Gln113 O (in order of increasing distances). In addition, the side-chain atoms Asp114 $O^{\delta1}$ and Asn119 $N^{\delta2}$ make longer contacts of 4.0 and 3.9 Å, respectively. The N2 atom of G16 forms strong hydrogen bonds to Asp114 $O^{\delta2}$ and Arg116 $N^{\eta2}$ and a weaker contact of 3.8 Å to Tyr64. The N3 atom of G16 forms contacts with Asp114 $O^{\delta1}$ and $O^{\delta2}$, having distances of 3.5 and 3.6 Å, respectively. There is a weak contact of 4.0 Å from Arg116 N^{ε} to the O4′ atom of G16. The other bases involved in the protein–DNA binding are C1, T2 and C3. The side-chain atoms of Arg116 form strong hydrogen bonds to the O2 atoms of T2 and C3, with $N^{\eta2}\cdots$O2 (T2) having a distance of 2.8 Å and $N^{\eta1}\cdots$O2 (C3) having a distance of 3.2 Å. Interactions with Arg116 also occur with O4′ of C3. Table 2 lists the hydrogen-bonding interactions between the protein and the DNA (also see Fig. 5).

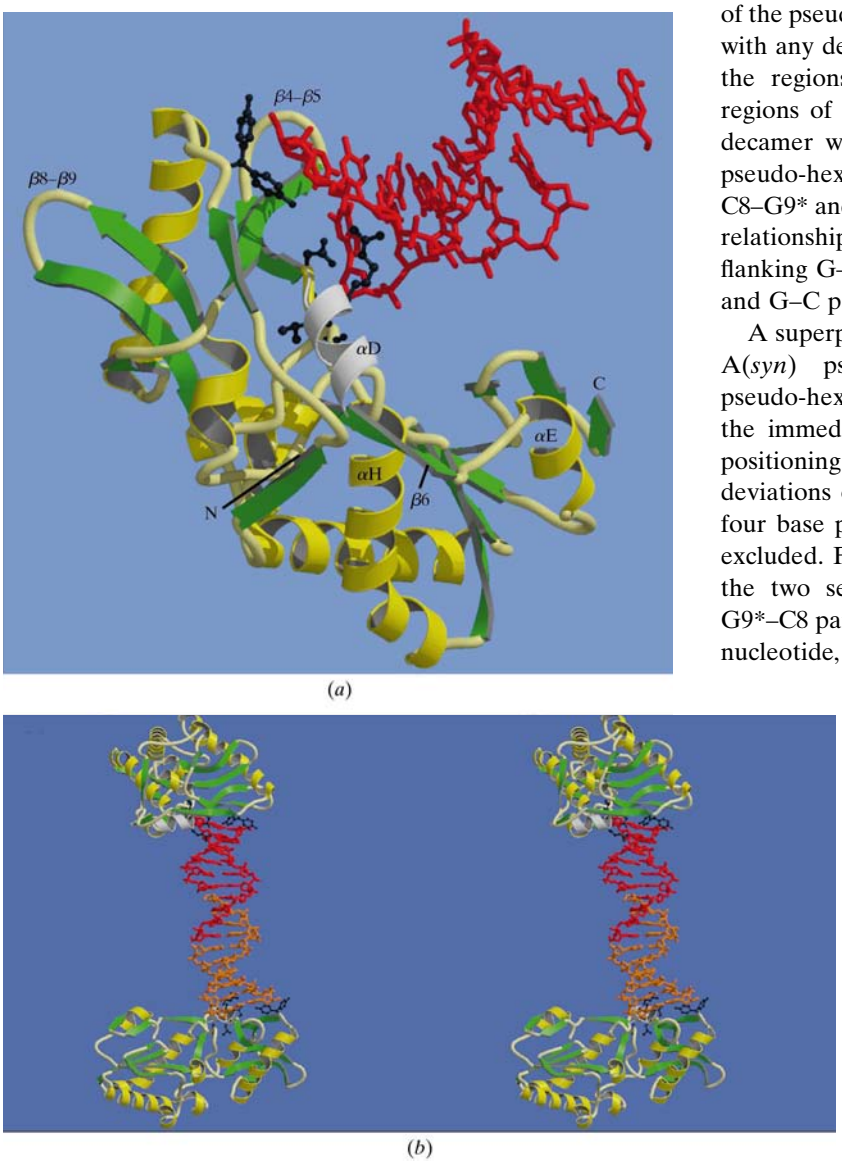

Figure 4
Views of (*a*) the asymmetric unit and (*b*) the symmetry-related molecules forming the pseudo-hexadecamer in the crystal structure of form IV. (*a*) A ribbon diagram (Kraulis, 1991; Merritt & Bacon, 1997) showing the form IV structure of the MMLV RT N-terminal fragment with its bound DNA (depicted with red stick models). The β-strands are shown in green, the coils in light yellow and the α-helices in yellow, with the exception of the αD helix, which is shown in white. Tyr64, Asp114, Leu115, Arg116 and Gly191, the principal residues comprising the oligonucleotide-binding site, are emphasized with dark ball-and-stick representations. (*b*) The stereoview (Merritt & Bacon, 1997) shown retains the color scheme of (*a*), with the exception that the symmetry-related 6/10-mer is colored in orange, thus emphasizing the pseudo-hexadecamer.

3.3. DNA analysis

3.3.1. The pseudo-hexadecamer helix. The overall structure of the pseudo-hexadecamer is essentially that of B-form DNA, with any deviations from the canonical structure occurring in the regions of the protein–DNA interactions and in the regions of the G–A mispairs. Fig. 2 shows the pseudo-hexadecamer with its numbering scheme. The symmetry in the pseudo-hexadecamer is such that it is bisected between the C8–G9* and G9–C8* pairs (where * denotes the $-x-1, -y, z$ relationship). Each mispair in the pseudo-hexadecamer has flanking G–C and C–G steps and the two mispairs have C–G and G–C pairs sandwiched between them.

A superpositioning in *O* (Jones *et al.*, 1991) of the G(*anti*)–A(*syn*) pseudo-hexadecamer onto the G(*anti*)–A(*anti*) pseudo-hexadecamer reveals nearly exact mapping except in the immediate region of the mispair. An O3′ atom superpositioning for the two pseudo-hexadecamers gave r.m.s. deviations of 0.06 Å for all nucleotides and 0.01 Å when the four base pairs involving the single-stranded overhang were excluded. Fig. 6 shows a close-up of the superpositioning of the two separate G–A mispair models with the flanking G9*–C8 pair. Note the nearly exact mapping of the guanosine nucleotide, which is indicative of the almost coincident mapping of the two pseudo-hexadecamer models.

The pseudo-hexadecamer was analyzed based on local parameters (Lu & Olson, 1999) both as an intact DNA duplex and as a doubly nicked molecule (the observed condition) using the program *3DNA* (Lu & Olson, unpublished results). It should be noted that in constructing an intact pseudo-hexadecamer for groove-analysis purposes the DNA structure was not allowed to change any of its base or base-pair conformations. Thus, for the major- and minor-groove analyses, idealized bridging phosphate groups were placed between nucleotides G6 and A7* and A7 and G6* (symmetry-related * molecules *via* $-x-1, -y, z$) using *O* (Jones *et al.*, 1991). Fig. 7 shows the O3′ atom superpositioning of the observed doubly nicked pseudo-hexadecamer onto a model hexadecamer retaining a completely intact phosphodiester backbone. The r.m.s. deviation for all atoms in common for the intact *versus* the doubly nicked models is 0.07 Å. Thus, it is clear that the overall DNA structure is unchanged by the addition of

Table 2
Protein–DNA hydrogen-bonding interactions.

The separate partial occupancy conformations of Tyr64 are arbitrarily designated with (i) and (ii).

Residue atom	Nucleotide atom	Distance (Å)
Tyr64(i) OH	C1 O5′	2.7
Tyr64(ii) OH	C1 O	3.0
Asp114 $O^{\delta2}$	G16 N2	3.2
Leu115 N	G16 O3′	2.8
Arg116 $N^{\eta1}$	C3 O2	3.2
Arg116 $N^{\eta2}$	G16 N2	3.1
Arg116 $N^{\eta2}$	T2 O2	2.8
Gly191 O	G16 O3′	3.3

bridging phosphate molecules and that the actual structure of the DNA is similar to an intact molecule. In addition, the DNA backbone torsion angles of the observed pseudo-hexadecamer were analyzed as described later (Schneider *et al.*, 1997).

The intact duplex model yields an average minor-groove width of 7.0 Å, with values in the range 6.3–7.6 Å. This is somewhat wider than the 5.8 Å average minor-groove width seen in standard B-form DNA and is likely to be a consequence of a number of factors. The short six-base strand has several interactions with the protein, some of which are quite strong, along with the absence of bridging phosphates in the center of the structure, which is likely to relax the helical tension. Heinemann *et al.* (1992) describe the crystal structure

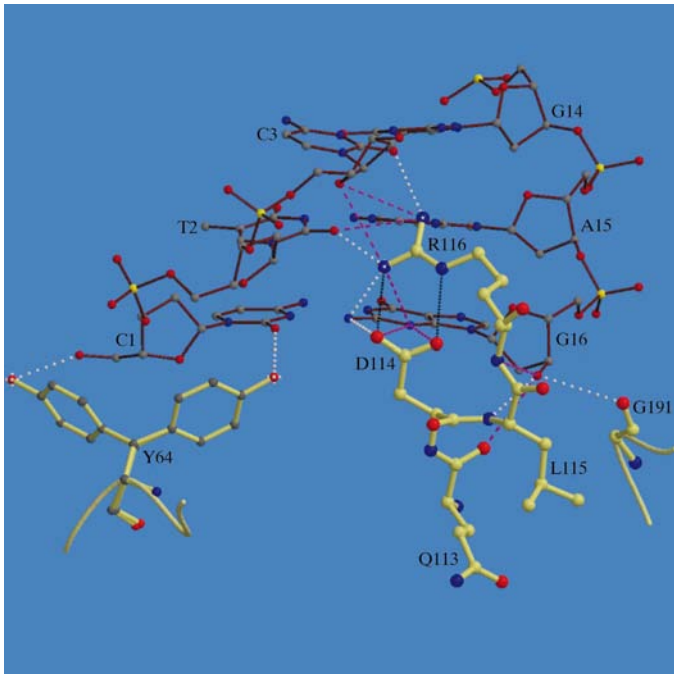

Figure 5
Schematic representation (Kraulis, 1991; Merritt & Bacon, 1997) of the interactions in the protein–DNA binding site of form IV. The hydrogen-bonding distances between 2.4 and 3.3 Å are indicated with white dotted lines. The non-bonded contacts ranging from 3.3 to 3.7 Å are indicated with longer-dashed magenta lines. Also shown with black dotted lines is the ion-pair formed by D114 $O^{\delta1}$ with R116 N^{ε} and D114 $O^{\delta2}$ and R116 $N^{\eta2}$ as observed in form IV.

Table 3
C1′–C1′ distances and glycosidic parameters.

Y, pyrimidine; R, purine. Values for the pseudo-hexadecamer model retaining the G(*anti*)–A(*syn*) conformation are given in parentheses.

	λ(1)† (°)	λ(2)† (°)	C1′–C1′‡ (Å)	RN9–YN1§ (Å)	RC8–YC6¶ (Å)
C–G	54.7	53.5	10.7	9.0	9.9
T–A	59.1	61.8	10.3	8.9	10.0
C–G	62.2	53.6	10.3	8.7	9.8
G–C	62.2	54.0	10.6	9.0	10.1
T–A	60.1	49.5	11.0	9.3	10.2
G–C	62.2	54.0	10.6	9.0	10.1
A–G	49.8 (29.7)	45.7 (48.5)	12.4 (12.2)	10.4 (9.9)	10.9 (9.2)
C–G	52.4	56.1	10.8	9.1	10.1
G–C	56.1	52.4	10.8	9.1	10.1
G–A	45.7 (48.5)	49.8 (29.7)	12.4 (12.2)	10.4 (9.9)	10.9 (9.2)
C–G	52.9	63.8	10.5	9.0	10.1
A–T	49.5	60.1	11.0	9.3	10.2
C–G	54.0	62.2	10.6	9.0	10.1
G–C	53.6	62.2	10.3	8.7	9.8
A–T	61.8	59.1	10.3	8.9	10.0
G–C	53.5	54.7	10.7	9.0	9.9

† λ: the angle between the C1′—YN1 or C1′—RN9 glycosidic bonds and the bp C1′–C1′ line. ‡ Distance between C1′ atoms for each base pair. § Distance between RN9/YN1 atoms for each base pair. ¶ Distance between RC8/YC6 atoms for each base pair.

of a self-complementary B-form DNA decamer containing solely G and C nucleotides and report its unusually wide minor groove. They attribute the wider minor groove to the sliding of the base pairs along their long axes. Our pseudo-hexadecamer is 68% G/C and as a result may also have some predisposition to a wider minor groove. The average minor-groove depth is 5.1 Å, with values in the range 4.1–6.0 Å, which compares favorably with standard B-form DNA. The average major-groove width for the intact duplex is 11.4 Å, with values in the range 9.6–12.4 Å. The average major-groove depth is 4.9 Å and the average diameter is 19.3 Å.

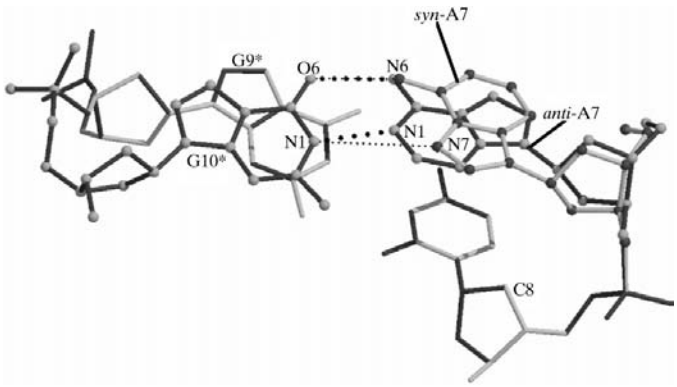

Figure 6
A close-up of the superpositioning (Jones *et al.*, 1991) of the G(*anti*)–A(*syn*) pseudo-hexadecamer model onto the G(*anti*)–A(*anti*) model, showing the nearly exact match between the models with the exception of the conformation of the mispairs. The *anti* model is depicted (Kraulis, 1991) with black bonds and grey atoms denoting the mispaired adenine. The hydrogen bonds of the *anti* mispair are drawn with heavy dotted black lines. The *syn* model is depicted with grey bonds and black atoms denoting its mispaired adenine. The hydrogen bonds of the *syn* mispair are drawn with smaller dotted lines. The C8–G9* pair is shown without atomic spheres. The atoms involved in the hydrogen bonding are labeled.

In B-form DNA, when the bases of a G–A mispair both retain the standard *anti* conformation, longer intra-base-pair C1′–C1′ distances occur and usually fall far outside the canonical range 10.5 ± 0.2 Å. Longer C1′–C1′ intra-base-pair distances may lead to increased buckling and twisting in the structure (Chuprina & Poltev, 1983; Keepers *et al.*, 1984). Privé *et al.* (1987), however, report no excessive bulging or twisting in their G(*anti*)–A(*anti*) mispaired DNA decamer despite the observed 12.5 Å C1′–C1′ distance in the mispairs. Any bulging or excessive twisting that could occur in their structure is alleviated by the significant propeller twist (24.8°) in the mispairs and the gradual increase/decrease in the P–P distances of the neighboring approaching/departing Watson–Crick pairs. In our pseudo-hexadecamer the lack of bridging phosphates between G6 and A7* and A7 and G6* is likely to alleviate any prospective bulging, as the C1′–C1′ intra-base-pair distance is 12.4 Å for the G(*anti*)–A(*anti*) model and 12.2 Å for the G(*anti*)–A(*syn*) model, both of which values are far closer to the Privé value than to the canonical value. The average intra-base-pair C1′–C1′ distance for the remaining base pairs in the pseudo-hexadecamer is 10.6 Å, most of which lie in a tight range around the canonical value of 10.5 Å. An exception to this is the 11.0 Å C1′–C1′ distance in the T5–A12 pair. The pseudosymmetry about the glycosyl bonds in the base pairs generally holds and the angular values fall within the accepted range 52–62° (Rosenberg *et al.*, 1976; Seeman *et al.*, 1976), except for A5 and the mispaired bases. The G(*anti*)–A(*anti*) angles are closer to the Watson–Crick range. Table 3 shows the C1′–C1′ bond distance data and the glycosidic bond parameters for the two pseudo-hexadecamer models.

A least-squares fitting of the bases of the pseudo-hexadecamer to standard bases gives r.m.s. deviations in the range 0.009–0.027 Å. The least deviation was found in C1 and G16, whereas the largest deviation was observed in G4. The small r.m.s. deviations in C1 and G16 are likely to be a consequence of their being extremely well defined since they form such strong interactions with the protein. Based on a global analysis with *CURVES* (Lavery & Sklenar, 1997) there is some minor bending in the pseudo-hexadecamer, as evidenced by the global axis-curvature value of 22.6°. The path length for the pseudo-hexadecamer is 50.2 Å and its end-to-end length is 49.0 Å; there is an overall 2.3% shortening of the molecule (Lavery & Sklenar, 1997).

Regarding the local base-pair helical parameters, the mean rise in either of our models is 3.3 Å; however, the differences (between the models) in rise that do occur reside in the region of the mispair. In the G(*anti*)–A(*syn*) model, the steps are more evenly spaced in this region than those of the G(*anti*)–A(*anti*) model. The mean helical twist is 35°, close to the standard B-form DNA value; however, there are wide variations among values in the mispair region. Table 4 shows the helical parameters for the pseudo-hexadecamer, noting the differences between the mispair models with parentheses.

3.3.2. Analysis of the G–A mispairs. Several otherwise unmodified DNA-only X-ray crystal structures containing G–A mispairs are known (Brown *et al.*, 1986, 1989; Privé *et al.*, 1987; Webster *et al.*, 1990; Shepard *et al.*, 1998; Gao *et al.*, 1999). Our structure is unique in that it is a protein–DNA complex whose DNA forms a pseudo-hexadecamer containing two G–A mispairs. In addition, our G–A mispairs reside in a G–C rich region and have the aforementioned nicks at the mispair steps. In the published otherwise unmodified B-DNA crystal structures cited earlier, all of the phosphodiester backbones are intact and each contains at least one flanking

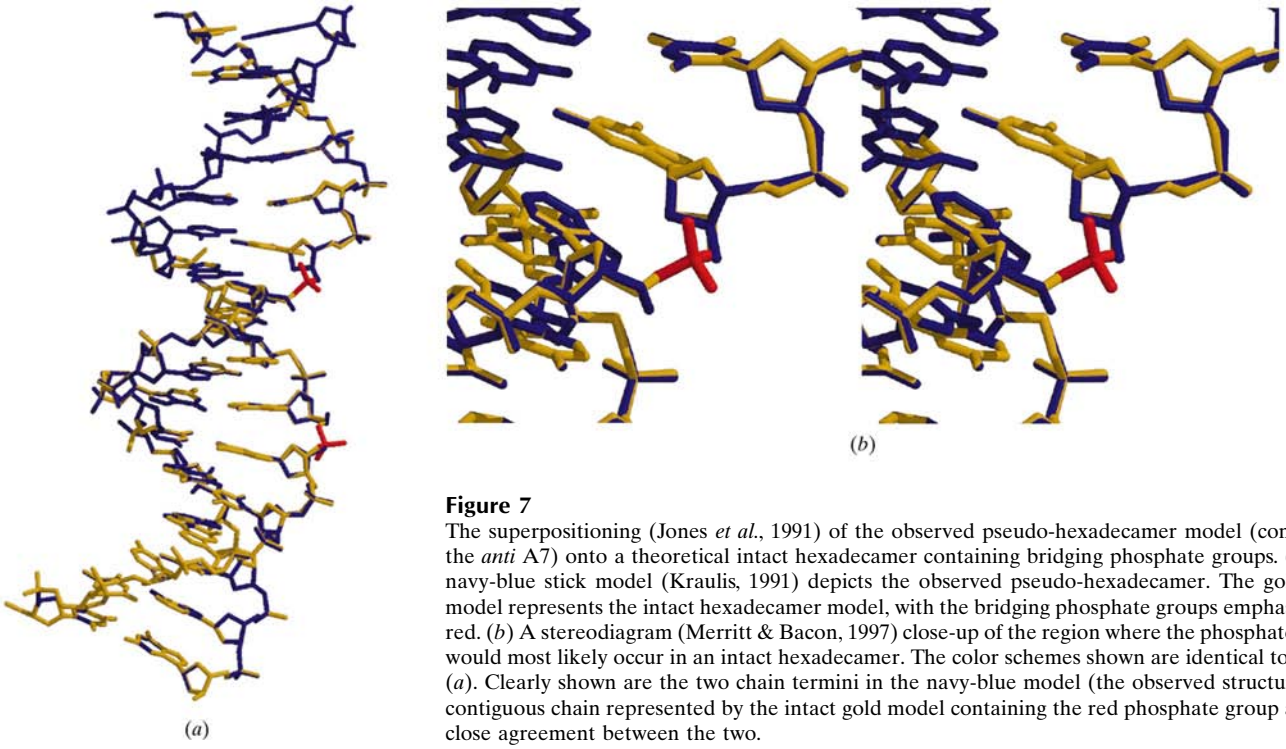

Figure 7
The superpositioning (Jones *et al.*, 1991) of the observed pseudo-hexadecamer model (containing the *anti* A7) onto a theoretical intact hexadecamer containing bridging phosphate groups. (*a*) The navy-blue stick model (Kraulis, 1991) depicts the observed pseudo-hexadecamer. The gold stick model represents the intact hexadecamer model, with the bridging phosphate groups emphasized in red. (*b*) A stereodiagram (Merritt & Bacon, 1997) close-up of the region where the phosphate group would most likely occur in an intact hexadecamer. The color schemes shown are identical to that in (*a*). Clearly shown are the two chain termini in the navy-blue model (the observed structure), the contiguous chain represented by the intact gold model containing the red phosphate group and the close agreement between the two.

Table 4
Local base-pair helical parameters.

The values in parentheses denote those values associated with the G(*anti*)–A(*syn*) pseudo-hexadecamer model. The twist angle cannot be reliably calculated for the *syn* model using available parameters at this time.

Step	*X*-energy-dispersive (Å)	*Y*-energy-dispersive (Å)	Rise (Å)	Inclin. (°)	Tip (°)	Twist (°)
CT/AG	−0.5	−2.5	3.4	4	9	36
TC/GA	−0.2	1.0	2.9	−4	−10	36
CG/CG	−0.2	0.5	3.3	9	−1	38
GT/AC	−1.6	0.2	3.5	2	6	33
TG/CA	0.7	−0.2	3.3	−1	−17	38
GA/GC	−0.6 (1.1)	2.8 (−0.2)	3.0 (3.3)	4 (0)	−8 (−5)	28
AC/GG	1.1 (7.4)	−1.7 (1.0)	3.6 (3.2)	−5 (18)	−6 (7)	34
CGCG	0.3	0.0	3.3	17	0	40
GG/AC	1.1 (7.4)	1.7 (−1.0)	3.6 (3.2)	−5 (18)	6 (−7)	34
GC/GA	−0.6 (1.1)	−2.8 (0.2)	3.0 (3.3)	4 (0)	8 (5)	28
CA/TG	0.7	0.2	3.3	−1	17	38
AC/GT	−1.6	−0.2	3.5	2	−6	33
CG/CG	−0.2	−0.5	3.3	9	1	38
GA/TC	−0.2	−1.0	2.9	−4	10	36
AG/CT	−0.5	2.5	3.4	4	−9	36
Average	−0.1	0.0	3.3	2 (5)	0	35
S.d.	0.8	1.6 (1.1)	0.2	6 (8)	9	3

A–T pair. In those structures the conformations of the bases of the mispairs are unequivocally characterized, even though studies indicate little energy difference between the G(*anti*)–A(*anti*) and the G(*anti*)–A(*syn*) conformations (Chuprina & Poltev, 1983; Keepers *et al.*, 1984). Interestingly, a modified dodecamer d[CGAGAATTC(O^6Me)GCG]$_2$, containing (O^6Me)G–A mispairs in a G–C rich region, has a crystal structure wherein the mispairs adopt the (O^6Me)G(*anti*)–A(*syn*) conformation (Ginell *et al.*, 1994), but its NMR structure retains the mispairs in the (O^6Me)G(*anti*)–A(*anti*) conformation (Patel *et al.*, 1986). This disparity is likely to be a consequence of crystal packing and suggests that the environment near the G–A mispair makes either conformation possible in the structure. In the dodecamer structure, the mispaired adenine is flanked on each side by a guanine. In our pseudo-hexadecamer the mispaired adenine is flanked on its 3′ side by a cytosine. So although the two models are not identically comparable, the notion of either a *syn* or an *anti* conformation for our A7 is consistent given the (O^6Me)G dodecamer and the energy data along with the fact that it resides at a strand terminus.

Gautheret *et al.* (1994) showed that a sheared G–A mispair conformation can occur in rRNA when another sheared non-Watson–Crick pair is flanking 5′ to the adenosine. Thus far in RNA crystal structures, the sheared G–A conformation has predominantly been seen when tandem purine–purine mispairs occur (Cheng *et al.*, 1992; Baeyens *et al.*, 1996). Recently, however, B-DNA crystal structures containing sheared G–A mispairs have been reported (Shepard *et al.*, 1998; Gao *et al.*, 1999). The unusual zipper-like DNA duplex reported by Shepard *et al.* (1998) is unique since it contains a centromeric sequence (GAAA) which upon dimerization results in two isolated sheared G–A mispairs flanking a central region of intercalated adenosines. The structures reported by Gao *et al.* (1999) also contain centromeric sequences with tandem sheared G–A mispairs. Within a regular B-DNA helix, an adenine in an isolated sheared G–A mispair cannot exist without severe helix distortion or disruption, since a Watson–Crick pair cannot follow in a step on its 5′ side. In our pseudo-hexadecamer, since the terminal adenosine is not connected on its 5′ side, we considered the possibility that a sheared G(*anti*)–A(*anti*) conformation could exist in the mispair given the presence of the flanking guanosine with which it could form favorable stacking interactions. A G–A mispair with a sheared G(*anti*)–A(*anti*) conformation was created and its adenine base was superimposed onto the *anti* adenine of our standard G(*anti*)–A(*anti*) structural model using *O* (Jones *et al.*, 1991). Fig. 3(*c*) shows the electron-density map for our structure, with our G(*anti*)–A(*anti*) model and the superimposed sheared G–A model, clearly illustrating that our mispairs have not undergone a slip dislocation and indeed adopt one of the more standard G–A mispair conformations.

When unmodified G–A mispairs adopt the standard G(*anti*)–A(*anti*) conformation, the intra-base-pair hydrogen bonds occur between G N1···N1 A and G O6···N6 A. In our G(*anti*)–A(*anti*) model, these distances are 2.5 and 2.7 Å, respectively. In a standard G(*anti*)–A(*syn*) model the hydrogen bonding involves G N1···N7 A and G O6···N6 A. In our G(*anti*)–A(*syn*) model these distances are 3.8 and 2.8 Å, respectively. The rather long (3.8 Å) N1···N7 distance in our G(*anti*)–A(*syn*) model could be a consequence of the opening between the bases created by the 2.9 Å intramolecular contact between O5′ and N3 of the adenine or of the greater stacking interaction with G6. In either of our G(*anti*)–A(*anti*) or G(*anti*)–A(*syn*) mispair models, the base-pair hydrogen bonding of the flanking G–C and C–G pairs warrants comment. In the flanking G9–C8* pair the N2···O2, N1···N3 and O6···N4 hydrogen-bond distances are 2.9, 3.0 and 3.1 Å, respectively. These distances are slightly longer than canonical values, but are comparable to those observed in other B-DNA structures with G–A mispairs (Brown *et al.*, 1986, 1989; Webster *et al.*, 1990). There is also a 3.4 Å N2···N3 distance in the G9–C8* pair. In the flanking C11–G6 pair, however, the 10° opening and the 1.0 Å shearing are reflected in the intra-base-pair hydrogen-bonding distances O2···N2, N3···N1 and N4···O6, which are 2.6, 3.0 and 3.4 Å, respectively. There is a very close N3···N2 distance of 2.7 Å, suggesting a bifurcated N2 donor, which is likely to be influenced by the favorable stacking between the terminal G6 and A7* bases.

3.4. Effects of protein–DNA interactions

The greatest deviations from the canonical B-form structure seen in our pseudo-hexadecamer occur most often in the region of the G–A mispair or where there are protein–DNA interactions. Considering the intra-base-pair parameters, the greatest shear occurs at G6–C11, the last pair of the intact 6/10-mer and the base pair directly below the mispair (see Fig. 2 for the numbering scheme of the pseudo-hexadecamer).

Table 5
Intra-base-pair parameters.

The values in parentheses represent the data for the G(*anti*)-A(*syn*) pseudo-hexadecamer model. The opening angle cannot be reliably calculated for the *syn* model using available parameters at this time.

	Shear (Å)	Stretch (Å)	Stagger (Å)	Buckle (°)	Propeller (°)	Opening (°)
C–G	0.0	0.1	−0.3	−11	−2	3
T–A	−0.6	0.3	0.9	−22	−10	10
C–G	0.6	0.0	0.3	−8	−20	11
G–C	0.6	0.3	0.3	−13	−10	9
T–A	0.8	0.4	0.9	−14	−10	2
G–C	1.0	0.2	0.3	−2	−12	10
A–G	0.4	0.6	0.2	5 (−6)	−16 (−17)	−9
C–G	−0.4	0.2	0.3	−7	−10	3
G–C	0.4	0.2	0.3	7	−10	3
G–A	−0.4	0.6	0.2	−5 (6)	−16 (−17)	−9
C–G	−1.0	0.2	0.3	2	−12	10
A–T	−0.8	0.4	0.9	14	−10	2
C–G	−0.6	0.3	0.3	13	−10	9
G–C	−0.6	0.0	0.3	8	−20	11
A–T	0.6	0.3	0.9	22	−10	10
G–C	0.0	0.1	0.3	11	−2	3
Average	0.0	0.2	0.4	0	−11	5
S.d.	0.6	0.2	0.4	12	5	7

The greatest stretching (1.3 Å) occurs in the mispair itself. The largest staggering occurs at T2–A15 and T5–A12. In the instance of T2–A15, the 0.9 Å stagger is likely to be a consequence of the strong hydrogen bond between Arg116 $N^{\eta 2}$ and the O2 of T2. In the case of T5–A12, the larger stagger is likely caused by a combination of factors. The T5 base is in the penultimate step of the short six-base strand of the intact duplex, is the first base of the 6/10-mer with no strong protein interaction and is able to form stable interactions with the detached G6 above it. The greatest buckling (–22°) and propeller twist (–20°) occur at T2–A15 and C3–G14, respectively, again likely to be a consequence of the strong hydrogen bonding of Arg116 to the O2 atoms of T2 and C3. The largest opening occurs between the same bases as well. Table 5 summarizes the intra-base-pair data, noting in parentheses the differences between the mispair models. The deviations observed owing to the interactions with the protein are modest and deviations of this magnitude have been observed in nucleic acid structures in the absence of complexation. The interactions with the protein are limited to the ends of the duplex involving the first three base pairs as described above. There are no direct interactions with the intervening ten base pairs of the pseudo-hexadecamer.

3.5. Deoxyribose-ring conformations

The pseudo-hexadecamer has several sugar-conformation parameters which exhibit A-like character, especially in the puckering of the deoxyribose rings. This is noteworthy in an otherwise B-form DNA molecule. Attempts were made to fit more commonly observed puckering conformations (*e.g.* $P \simeq 160°$; C2′-*endo*) to those rings whose parameters lay outside common B-form ranges. Each attempt caused significant increases in the $R_{\rm work}$ and $R_{\rm free}$ values. All the bases with A-like puckerings or phosphodiester torsion angles are on the short strand of the 6/10-mer, which is the strand with the greatest interaction with the protein. The C1 base has a standard C2′-*endo* pucker and P = 160°; however, the remaining five bases in the strand (T2, C3, G4, T5 and G6) have acute P angles with C4′-*exo* (T2) and C3′-*endo* (C3, G4, T5 and G6) sugar-pucker conformations. Although C3′-*endo* sugar puckers are observed in A-form DNA almost exclusively, they are not the predominant motif in B-form DNA. C4′-*exo* puckers are more commonly seen in RNA than DNA. In our structure, Arg116 reaches far into the minor groove of the 6/10-mer and not only forms the previously discussed hydrogen bonds with the O2 atoms of T2 and C3 but also has two interactions with O4′ of C3. The fixed angle formed by $N^{\eta 1}$—C^{ζ}—$N^{\eta 2}$ in Arg116 serves to lock not only the O2 atoms of T2 and C3 but also the O4′ atom of C3. The interaction of Arg116 with O4′ is shared virtually equally, as evidenced by the respective O4′$\cdots N^{\eta 1}$ and O4′$\cdots N^{\eta 2}$ bond distances of 3.65 and 3.71 Å. Analysis of the backbone torsion angles, specifically in a χ/δ scattergram, of T2, C3 and G4 shows these bases to have A-DNA character, with the T5 and G6 bases lying in the region bordering A and BI character (Schneider *et al.*, 1997). This is reminiscent of the HIV-1 RT–DNA complex structures where the DNA has A-like character in the polymerase active site, then gradually becomes more B-like approximately six steps behind (Jacobo-Molina *et al.*, 1993; Huang *et al.*, 1998). Table 6 presents the sugar-conformation parameters for the pseudo-hexadecamer. Since the pseudo-hexadecamer is symmetric, only the data for strand I appears.

3.6. Complexation as a means of obtaining novel nucleic acid structures

Complexation with a molecule that will limit the conformational flexibility of the molecule of interest and provide additional packing interactions has been used in the pursuit of structure determinations of a number of complexes including the HIV-1 RT–DNA (Clark *et al.*, 1995) and the HIV-1 gp120–CD4 complexes (Kwong *et al.*, 1999). In each of these cases, a monoclonal Fab fragment was included in the complex of interest in order to obtain crystals suitable for a structure determination. However, in the case of the gp120 complex structure, the monoclonal Fab fragment was one of many variables screened that contributed to the ability to obtain crystals. A novel approach for the determination of nucleic acid structures involving complexation with the N-terminal fragment of MMLV RT provides a conformational limitation on the DNA duplexes that we have examined. In addition, complexation with the fragment molecules provides protein–protein and protein–DNA interactions that allow the DNA molecules to pack in several different crystal lattices.

The N-terminal fragment can be expressed in *Escherichia coli* and purified using standard methods with high yields (approximately 5–10 mg of purified protein per litre of cell culture) as previously reported (Sun *et al.*, 1998). The fingers-domain binding site will accommodate blunt-ended duplexes in addition to those with single-stranded overhangs. There is

Table 6
Sugar-conformation parameters.

All sugar pucker designations are identical for either pseudo-hexadecamer model. The values in parentheses are those for the G(*anti*)–A(*syn*) model.

Base	v0† (°)	v1‡ (°)	v2§ (°)	v3¶ (°)	v4†† (°)	Tm‡‡ (°)	*P*§§ (°)	Pucker
C	−22	34	−32	20	1	34	160	C2′-*endo*
T	−19	−5	25	−37	35	37	48	C4′-*exo*
C	−6	−15	29	−33	25	33	29	C3′-*endo*
G	−3	−19	32	−34	23	34	23	C3′-*endo*
T	−9	−14	31	−37	29	37	33	C3′-*endo*
G	−2	−23	38	−40	26	41	21	C3′-*endo*
A	−19 (−26)	34 (39)	−35 (−37)	24 (23)	−3 (1)	36 (40)	166 (159)	C2′-*endo*
C	−14	27	−29	21	−4	29	170	C2′-*endo*
G	−27	40	−37	22	3	40	157	C2′-*endo*
G	−24	36	−34	20	2	37	158	C2′-*endo*
C	−25	37	−35	21	2	37	158	C2′-*endo*
A	−22	36	−36	23	−1	37	163	C2′-*endo*
C	−15	29	−31	23	−5	32	171	C2′-*endo*
G	−20	33	−32	21	−1	34	163	C2′-*endo*
A	−29	39	−33	17	7	38	151	C2′-*endo*
G	−19	33	−33	22	−2	34	164	C2′-*endo*

† v0, C4′—O4′—C1′—C2′. ‡ v1, O4′—C1′—C2′—C3′. § v2, C1′—C2′—C3′—C4′. ¶ v3, C2′—C3′—C4′—O4′. †† v4, C3′—C4′—O4′—C1′. ‡‡ Tm, amplitude of pseudorotation of the sugar ring. §§ *P*, phase angle of pseudorotation of the sugar ring.

no preference for sequence or type of nucleic acid. Reverse transcriptase is a DNA polymerase that can synthesize DNA using either RNA or DNA as a template and will therefore bind RNA–DNA and DNA–DNA duplexes. Although we have not attempted to obtain complexes of RNA-containing duplexes, modeling studies suggest that the fingers-domain binding site will equally well accommodate an RNA–DNA or RNA–RNA duplex. In each case that we have examined the DNA duplex is essentially B-form with some interesting deviations noted in this study for the pseudo-hexadecamer with G–A mispairs. We also note that attempts to obtain crystals suitable for a structure determination of the pseudo-hexadecamer alone have to date proven unsuccessful.

The complexation approach is potentially quite general and will allow for crystallographic analysis of a number of nucleic acid molecules that have proven refractory to crystallization efforts. In addition, this approach provides the opportunity to compare the structural features of DNA sequences that have been analyzed in different helical forms with the same sequence in a more B-like conformation. To this end, we have obtained crystals of a Z-form DNA duplex, 5′-CGCGCGCG-3′, complexed with the N-terminal fragment of MMLV RT in a previously characterized lattice (Coté & Georgiadis, unpublished results).

Another significant advantage of our approach includes phasing of the resultant complexes, which can be preformed by molecular replacement using one of the fragment molecules in complex with DNA that we have already reported (PDB accession codes 1d1u, 1qai or 1qaj) as a search model. Following rebuilding and positional refinement of the protein model, the electron density for the DNA duplex was easily interpreted, allowing rapid rebuilding and adjustment of the DNA model in the structures that we have determined. The initial phasing does not include any model bias for the nucleic acid. Of course, the refinement problem is potentially more complicated and time-consuming for the nucleic acid complexed to the fragment than for the nucleic acid alone. However, the ability to obtain unbiased phases undoubtedly outweighs the more difficult refinement in an assessment of the feasibility of the approach.

In conclusion, the DNA structure of the pseudo-hexadecamer reported here contains two G–A mispairs in a unique G–C-rich environment. The DNA structure has been obtained in a complex with the N-terminal fragment of MMLV RT and has some very interesting features that deviate from B-form DNA. Although we can clearly determine that the guanine of the mispair is in the *anti* conformation, the mispaired adenine can be equally well modeled as *syn* or *anti* conformation. The short six-base strand that is conformationally constrained by strong hydrogen bonding to a protein intervening into its minor groove results in nucleotides with A-like sugar puckers and backbone torsion angles in an otherwise B-like DNA molecule. Based on the analysis of an intact model for the phosphodiester backbone, our model, which includes nicks in the phosphodiester backbone, presents a reasonable model of a biologically relevant DNA structure containing G–A mispairs.

We thank Xiang-Jun Lu and Bohdan Schneider for assistance with analysis of the DNA, Helen Berman and Wilma Olson for helpful discussions and critical reading of the manuscript, Shabir Najmudin for helpful discussions, Dick Leidich and Greg Listner for technical support of the X-ray equipment and Attilio Defalco for general computing assistance. This work was supported by a grant GM55026 (to MMG) from the National Institutes of Health.

References

Baeyens, K., De Bondt, H. L., Pardi, A. & Holbrook, S. (1996). *Proc. Natl Acad. Sci. USA*, **93**, 12851–12855.
Biosym Technologies (1993). *INSIGHT II User Guide.* San Diego, CA: Biosym Technologies.
Brown, T., Hunter, W. N., Kneale, G. & Kennard, O. (1986). *Proc. Natl Acad. Sci. USA*, **83**, 2402–2406.
Brown, T. C. & Jiricny, J. L. (1988). *Cell*, **54**, 705–711.
Brown, T., Leonard, G. A., Booth, E. D. & Chambers, J. (1989). *J. Mol. Biol.* **207**, 455–457.

Brunger, A. T., Adams, P. A., Clore, G. M., Gros, P., Grosse-Kunstleve, R. W., Jiang, J.-S., Kuszewski, J., Nilges, N., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). *Acta Cryst.* D**54**, 905–921.
Cheng, J.-W., Chou, S.-H. & Reid, B. (1992). *J. Mol. Biol.* **228**, 1037–1041.
Chuprina, V. P. & Poltev, V. I. (1983). *Nucleic Acids Res.* **11**, 5205–5222.
Clark, A. D. J., Jacobo-Molina, A., Clark, P., Hughes, S. H. & Arnold, E. (1995). *Methods Enzymol.* **262**, 171–185.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Esnouf, R. (1997). *J. Mol. Graph.* **15**, 132–143.
Fersht, A. R., Knill-Jones, J. W. & Tsui, W.-C. (1982). *J. Mol. Biol.* **156**, 37–51.
Gao, Y.-G., Robinson, H., Sanishvili, R., Joachimiak, A. & Wang, A. H.-J. (1999). *Biochemistry*, **38**, 16452–16460.
Gautheret, D., Konings, D. & Guttel, R. R. (1994). *J. Mol. Biol.* **242**, 1–8.
Georgiadis, M. M., Jessen, S. M., Ogata, C. M., Telesnitsky, A., Goff, S. P. & Hendrickson, W. A. (1995). *Structure*, **3**, 879–892.
Ginell, S. L., Vojtechovsky, J., Gaffney, B., Jones, R. & Berman, H. M. (1994). *Biochemistry*, **33**, 3487–3493.
Heinemann, U., Alings, C. & Bansel, M. (1992). *EMBO J.* **11**, 1931–1939.
Huang, H., Chopra, R., Verdine, G. L. & Harrison, S. C. (1998). *Science*, **282**, 1669–1675.
Hubbard, S. J. & Thornton, J. M. (1993). *NACCESS* program. Department of Biochemistry and Molecular Biology, University College, London.
Jacobo-Molina, A., Ding, J., Nanni, R. G., Clark, A. D. Jr, Lu, X., Tantillo, C., Williams, R. L., Kamer, G., Ferris, A. L., Clark, P., Hizi, A., Hughes, S. H. & Arnold, E. (1993). *Proc. Natl Acad. Sci. USA*, **90**, 6320–6324.
Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. (1991). *Acta Cryst.* A**47**, 110–119.
Keepers, J. W., Schmidt, P., James, T. L. & Kolman, P. A. (1984). *Biopolymers*, **23**, 2901–2929.
Kraulis, P. J. (1991). *J. Appl. Cryst.* **24**, 946–950.
Kwong, P. D., Wyatt, R., Desjardins, E., Robinson, J., Culp, J. S., Hellonig, B. D., Sweet, R. W., Sodroski, J. & Hendrickson, W. A. (1999). *J. Biol. Chem.* **274**, 4115–4123.
Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993). *J. Appl. Cryst.* **26**, 283–290.
Lavery, R. & Sklenar, H. (1997). *CURVES* 5.2: *Helical Analysis of Irregular Nucleic Acids*.
Leonard, G. A., Booth, E. D. & Brown, T. (1990). *Nucleic Acids Res.* **18**, 5617–5623.
Leonard, G. A., Guy, A., Brown, T., Teoule, R. & Hunter, W. N. (1992). *Biochemistry*, **31**, 8415–8420.
Leonard, G. A., McAuley-Hecht, K. E., Gibson, N. J., Brown, T., Watson, W. P. & Hunter, W. N. (1994). *Biochemistry*, **33**, 4755–4761.
Lu, X.-J. & Olson, W. K. (1999). *J. Mol. Biol.* **285**, 1563–1575.
McAuley-Hecht, K. E., Leonard, G. A., Gibson, N. J., Thompson, J. B., Watson, W. P., Hunter, W. N. & Brown, T. (1994). *Biochemistry*, **33**, 10266–10270.
Merritt, E. A. & Bacon, D. J. (1997). *Methods Enzymol.* **277**, 505–524.
Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.
Najmudin, S., Coté, M. L., Sun, D., Yohannan, S., Montano, S. P., Gu, J. & Georgiadis, M. M. (2000). *J. Mol. Biol.* **296**, 613–632.
Navaza, J. (1994). *Acta Cryst.* A**50**, 157–163.
Nicholls, A., Sharp, K. & Honig, B. (1991). *Proteins Struct. Funct. Genet.* **11**, 281–296.
Otwinowski, Z. (1993). *Proceedings of the CCP4 Study Weekend. Data Collection and Processing*, edited by L. Sawyer, N. Isaacs & S. Bailey, pp. 56–62. Warrington: Daresbury Laboratory.
Patel, D. J., Shapiro, L., Kozlowski, S., Gaffney, B. L. & Jones, R. A. (1986). *J. Mol. Biol.* **188**, 677–692.
Poltev, V. I. & Shulyupina, N. V. (1986). *J. Biomol. Struct. Dyn.* **3**, 739–765.
Privé, G. G., Heinemann, U., Chandrasegaran, S., Kan, L.-S., Kopka, M. L. & Dickerson, R. E. (1987). *Science*, **238**, 498–504.
Read, R. J. (1986). *Acta Cryst.* A**42**, 140–149.
Rosenberg, J. M., Seeman, N. C., Day, R. O. & Rich, A. (1976). *J. Mol. Biol.* **104**, 145–167.
Schneider, B., Neidle, S. & Berman, H. M. (1997). *Biopolymers*, **42**, 113–124.
Seeman, N. C., Rosenberg, J. M., Suddath, F. L., Park Kim, J.-J. & Rich, A. (1976). *J. Mol. Biol.* **104**, 109–144.
Shepard, W., Cruse, W. B., Fourme, R., de la Fortelle, E. & Prange, T. (1998). *Structure*, **6**, 849–861.
Sun, D., Jessen, S., Liu, C., Liu, X., Najmudin, S. & Georgiadis, M. M. (1998). *Protein Sci.* **7**, 1575–1582.
Webster, G. D., Sanderson, M. R., Skelly, J. V., Neidle, S., Swann, P. F., Li, B. F. & Tickle, I. J. (1990). *Proc. Natl Acad. Sci. USA*, **87**, 6693–6697.