Published online 26 September 2006 *Nucleic Acids Research, 2006, Vol. 34, No. 19* **5353–5360**
doi:10.1093/nar/gkl693

Crystal structures of oligonucleotides including the integrase processing site of the Moloney murine leukemia virus

Sherwin P. Montaño, Marie L. Coté[1], Monica J. Roth[1] and Millie M. Georgiadis*

Department of Biochemistry and Molecular Biology, Indiana University School of Medicine, 635 Barnhill Dr., Indianapolis, IN 46202, USA and [1]Department of Biochemistry, University of Medicine and Dentistry of New Jersey–Robert Wood Johnson Medical School, 675 Hoes Lane, Piscataway, NJ 08854, USA

Received June 19, 2006; Revised August 22, 2006; Accepted September 7, 2006

ABSTRACT

In the first step of retroviral integration, integrase cleaves the linear viral DNA within its long terminal repeat (LTR) immediately 3′ to the CA dinucleotide step, resulting in a reactive 3′ OH on one strand and a 5′ two base overhang on the complementary strand. In order to investigate the structural properties of the 3′ end processing site within the Moloney murine leukemia virus (MMLV) LTR d(TCTTTCATT), a host-guest crystallographic method was employed to determine the structures of four self-complementary 16 bp oligonucleotides including LTR sequences (underlined), d(TTTCATTGCAATGAAA), d(CTTTCATTAATGAAAG), d(TCTTTCATATGAAAGA) and d(CACAATGATCATTGTG), the guests, complexed with the N-terminal fragment of MMLV reverse transcriptase, the host. The structures of the LTR-containing oligonucleotides were compared to those of non-LTR oligonucleotides crystallized in the same lattice. Properties unique to the CA dinucleotide step within the LTR sequence, independent of its position from the end of the duplex, include a positive roll angle and negative slide value. This propensity for the CA dinucleotide step within the MMLV LTR sequence to adopt only positive roll angles is likely influenced by the more rigid, invariable 3′ and 5′ flanking TT dinucleotide steps and may be important for specific recognition and/or cleavage by the MMLV integrase.

INTRODUCTION

Integration of a DNA copy of the retroviral genome into the host chromosome is an essential step in the life cycle of the retrovirus [reviewed in Refs (1,2)]. The single-stranded RNA genome is replicated by the retrovirally encoded reverse transcriptase (RT) resulting in a double-stranded DNA copy flanked by direct repeat sequences referred to as long terminal repeats (LTRs) including U3, R and U5 sequences (3,4). Prior to integration of the retroviral DNA, the retroviral integrase specifically cuts 3′ of the CA dinucleotide step resulting in a free 3′ OH and a 2 bp 5′ overhang. Although cleavage of the target host DNA requires no sequence specificity (5), in recent work, the sites of insertion of retroviral DNA into the host genome have been shown not to be random, as originally proposed, but to include active genes for HIV-1 and promoters of active genes for Moloney murine leukemia virus (MMLV) (6–9).

Although retroviral DNA forms circular species including an LTR–LTR junction, the intermediate DNA molecule required for integration has been shown to be a double-stranded linear molecule with 3′ recessed ends (10,11). Repeat sequences found at the termini of retroviral DNA are required for recombination; however, sequence conservation among different retroviruses is limited to a CA dinucleotide most frequently located 2 bp from the termini within these repeats. Both the CA sequence and its position relative to the termini of the retroviral DNA are required for specific cleavage and integration to occur (11–14). Integration studies performed on MMLV showed that the spacing of 2 bp from the end of the LTR is not absolutely essential as mutants including 4 or 1 bp instead of 2 bp 3′ of CA on the U5 end were cleaved and integrated (12,14). Mutant LTRs including the 5′ CTTT sequence with deletions of 2 or 8 bp from the 3′ end were not integrated (15). Given adjacent CA and TA steps with TA positioned 2 bp from the end and CA 4 bp from the end, CA was preferred over the TA as the site of processing (12,14). No mutational analysis of the 5′ flanking sequence has been reported for the MMLV LTR.

Mutations at some positions within the 5′ and 3′ sequences flanking the critical CA dinucleotide step within the U3 and U5 HIV-1 LTR have significant effects on the 3′ processing activity including a mutation within the U5 LTR replacing GCAGT with TCAGT (16). This is of interest in considering the structural properties of the HIV-1 U5 LTR (17) compared

*To whom correspondence should be addressed. Tel: +1 317 278 8486; Fax: +1 317 274 4686; Email: mgeorgia@iupui.edu

to those of the MMLV U5 LTR as the sequence found in MMLV is TCATT, and HIV-1 integrase does not specifically cut the MMLV U5 LTR (18). When mutations of either C or A of the dinucleotide step were made, the processing activity was almost completely eliminated (16). Spacing requirements for the positioning of the CA dinucleotide within the HIV-1 LTR are similar to those reported for MMLV in that the CA dincucleotide could be positioned between 2 and 6 bp from the end and still result in integration of the retroviral DNA (18).

Using a host-guest crystallographic approach developed in our laboratory, we have determined 4 crystal structures of sequences derived from the MMLV LTR. As previously reported, this approach involves the use of an N-terminal fragment of MMLV RT 'host' to complex an oligonucleotide 'guest' (19–23). These structures are the first duplex DNA structures of oligonucleotide sequences derived from the 3′ end processing site of MMLV LTR to be reported.

MATERIALS AND METHODS

Crystallization and data collection

The N-terminal fragment of MMLV RT was expressed in *Escherichia coli* and purified as previously detailed (19). The oligonucleotides were synthesized by Trilink Biotechnologies, Inc. and purified by HPLC on a Hamilton PRP-1 column (7 μm, 7.0 × 305 mm) using the same conditions as described in earlier work (19). All the protein–DNA complexes were prepared and crystallized in the same manner: The protein was diluted to 0.65 mM using 0.1 M HEPES (pH 7.5), 0.3 M NaCl from a 1.4 mM stock that was in 50 mM MES (pH 6.0), 0.3 M NaCl, 1 mM DTT. The oligonucleotides were solubilized in 10 mM HEPES (pH 7.0), 10 mM $MgCl_2$ to make a 2.5 mM stock. The protein was incubated with the DNA on ice for an hour to form the complex. A 1:2 molar ratio of protein to DNA was used with final concentrations of 0.43 and 0.86 mM, respectively. Initial crystals of the complexes were obtained at 20°C from vapor diffusion hanging drops consisting of 1 μl of complex and 1 μl of crystallization solution containing 8% (w/w) PEG4000, 0.05 M *N*-[2-acetamido]-2-iminodiacetic acid (ADA) (pH 6.5) and 0.005 M magnesium acetate. Crystals used for X-ray diffraction data collection were obtained by microseeding techniques. Crystallographic parameters and data processing statistics are shown in Table 1.

Structure determination and refinement

Initial phasing for all of the structures was obtained by molecular replacement with AMoRe (24) using the protein fragment from a final refined structure that was solved previously (PDB accession code of 1N4L) as the search model (22). Following molecular replacement, the model was subjected to rigid body refinement using data from 20 to 4 Å. All crystallographic refinement was performed using CNS (25) using the nucleic acid parameter files developed by Parkinson *et al*. (26). Five percent of the reflections were initially flagged for the calculation of R_{free} and used to monitor the progress of the refinement. Using the full resolution range of the data, two rounds of simulated annealing, B-factor refinement and energy minimization calculations were done to obtain the initial electron density map of the DNAs. Nucleic Acid Builder (27) was then used to generate a coordinate file of a B-form DNA with the appropriate sequence. This model was then manually positioned into the density and backbone torsion angles in the nucleotides from this model were adjusted. Typically, the first 3 bp were modeled initially and then refined. The next five were then fitted into the density after the second round of refinement. Water-picking and refinement of the waters were done during the

Table 1. Crystallographic and refinement data for the LTR structures

	A	B	C	D
Crystallographic data				
Space group	$P2_12_12$	$P2_12_12$	$P2_12_12$	$P2_12_12$
Cell dimensions (Å)				
a	54.72	55.12	54.79	54.24
b	146.13	145.87	145.90	145.68
c	46.86	46.79	46.93	46.94
Resolution (Å)	2.25	2.3	2.2	2.35
No. of reflections measured	68 062	88 476	66 981	79 074
No. of unique reflections	18 291	17 275	18 321	16 118
R_{sym}	0.068 (0.40)	0.069 (0.38)	0.043 (0.22)	0.045 (0.37)
Completeness (%)	98.4 (96.9)	98.7 (99.6)	92.0 (98.0)	99.5 (98.7)
I/σ	19.7 (2.9)	20.9 (4.2)	32.6 (5.5)	26.2 (3.9)
Refinement statistics				
R	22.37	23.62	23.14	24.36
R_{free}	28.76	26.28	25.99	26.54
$Distance_{RMSD}$ (Å)	0.0066	0.0060	0.0059	0.0064
$Angle_{RMSD}$ (°)	1.37	1.20	1.24	1.22
$<B>_{protein}$	40.60	40.57	38.28	46.54
$<B>_{DNA}$	69.82	62.62	57.59	59.44
Ramachandran plot				
Most favorable (%)	93.0	92.1	91.6	87.0
Add. allowed (%)	6.0	6.0	6.0	11.2
Gen. allowed (%)	0.5	1.4	0.9	0.9
Disallowed (%)	0.5	0.5	1.4	0.9

third and fourth rounds. Multiple rounds of refinement and manual model rebuilding in O (28) were then employed to obtain the final refined structure. The structural refinements were judged to be converged from the R_{work} and R_{free} values as well as from the difference Fourier maps (Fo-Fc, 2Fo-Fc) and simulated annealing omit maps.

For the LTR-B structure, positive density at 3.0 σ was observed near the base, sugar and phosphate groups of A13 and A14 (the 3rd and 4th base from the 3′ end of the molecule). Since the density was too close to the said moieties, this suggested an alternate conformation at these positions. Alternate conformations of these two bases and their respective sugar and phosphate groups were then built into the positive density. The two DNA molecules were set at half-occupancy and subjected to refinement. The difference Fourier maps (Fo-Fc) generated did not contain any positive or negative peaks around the conformations built, indicating that the model, indeed, has two conformations. Final refinement statistics are shown in Table 1.

The crystal structures have been deposited with PDB accession numbers as follows: LTR-A, 2FVP; LTR-B, 2FVQ; LTR-C, 2FVR; and LTR-D, 2FVS.

RESULTS AND DISCUSSION

Design of the LTR oligonucleotides

The host-guest crystallographic method takes advantage of minor groove binding interactions of the N-terminal fragment of the MMLV RT to the ends of duplex DNA. This method facilitates crystallization and analysis of DNA sequences of interest as previously described (20–23,29,30). Briefly, the interactions of the duplex DNA are limited to the minor groove and backbone of the terminal 3 bp of a 16 bp duplex, which is bound on either end by an N-terminal fragment molecule (see Supplementary Data and Table S1 for a full description of protein–DNA interactions). Thus, the middle 10 bp of the 16 bp duplex are free of interactions with the protein. The asymmetric unit, the unique repeating unit of the crystal, includes one protein molecule and one half of the 16 bp duplex. The 16 bp duplex is bisected by a crystallographic 2-fold rotation axis, as shown in Figure 1A. As a consequence of this symmetry, the electron density observed is an average of the two halves of the DNA molecule. Although we have successfully used this method to analyze asymmetric DNA sequences (22), the method is better suited to the analysis of symmetric sequences.

Thus, we have designed four different symmetric sequences designated A, B, C and D that are derived from the MMLV LTR 3′ end processing site for analysis as shown in Figure 1B. In designing these sequences, we have included maximally 8 bp from the LTR in a self-complementary oligonucleotide as in sequences B and C, 7 bp in sequence A and 5 bp in sequence D. All designs include the CA dinucleotide, which is the site of 3′ end processing by the retroviral integrase, as well as the flanking LTR sequence. The CA dinucleotide is positioned in the middle portion of the oligonucleotide and is free of interactions with the protein allowing us to evaluate the structural properties of the 'naked' LTR sequence. Sequence B includes the LTR–LTR junction sequence formed in the circularized form of the LTR, which is not a substrate

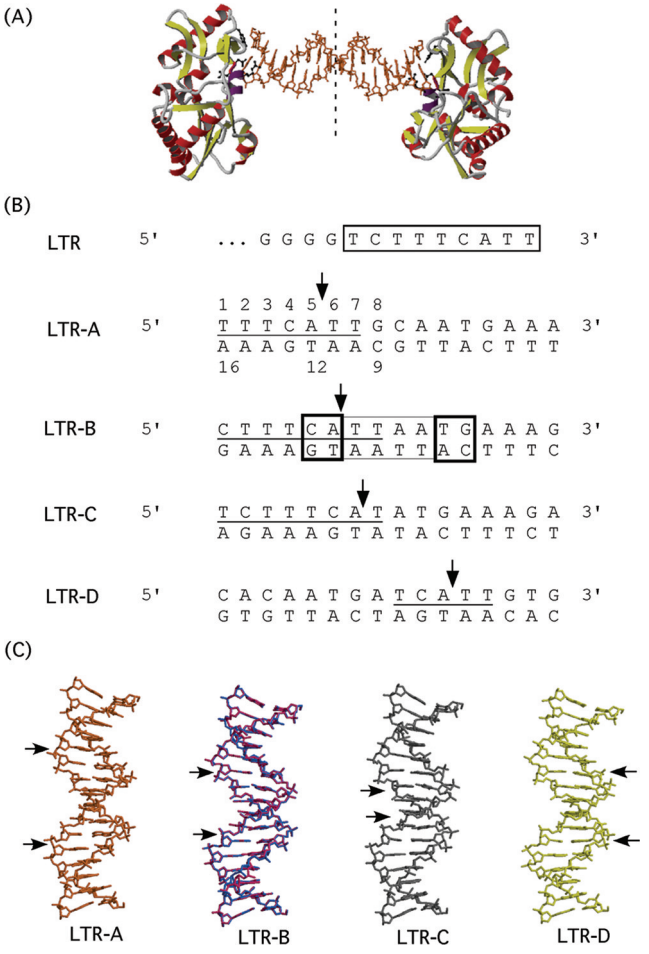

Figure 1. (**A**) Crystal structure of a 16mer DNA duplex in complex with the N-terminal fragment of Moloney murine leukemia virus reverse transcriptase. The ribbon-rendering was done with MOLSCRIPT (42). The DNA structure shown here is that of the LTR-A. The asymmetric unit in this lattice contains one protein molecule and one-half the DNA molecule. The dashed line marks the crystallographic 2-fold axis. (**B**) The LTR-A, -B, -C and -D sequences for which we have determined crystal structures are shown. These sequences are derived from the 3′-processing site of the actual LTR. The bases underlined are those in common with the actual LTR sequence. In LTR-B, the eight base-pairs that are boxed in is the LTR-LTR junction and outlined with a darker box is the CA cut site. The numbering scheme for the bases is shown in the LTR-A sequence. (**C**) The structures of LTR-A (orange), -B (blue and red), -C (black) and -D (green) are shown as stick renderings. Arrows indicate the site of cleavage immediately 3′ of the CA dinucleotide step.

for integration. In part, our interest in studying these sequences was to determine whether there are differences in the LTR dinucleotide steps within the LTR–LTR junction as compared to these steps in other sequences. In sequence D, the CA dinucleotide step is positioned 5 bp from the 3′ end making it the most similar to an integrase substrate. To our knowledge, this is the first study in which related sequences, namely those within d(TCTTCATT), have been studied in the same crystal lattice while in different positions with respect to the ends of the oligonucleotide allowing us to evaluate sequence specific aspects associated with the dinucleotide steps TT, TC, CA and AT comprising the 3′ end processing site of the MMLV LTR.

Analysis of the DNA molecules

The DNA structures reported here may be classified as B-form DNA as shown in Figure 1C. The electron density for the DNA in all of these structures is well defined in initial Fo-Fc difference maps. The best ordered electron density was obtained for the LTR-B structure for which two conformations are observed and are referred to as LTR-B1 and LTR-B2 (see Materials and Methods for more details). The structures of LTR-B1 and -B2 differ in the positions of A13 and A14 and their associated sugar and phosphate moieties as shown in Figure 2A but are otherwise the same. In particular, the base positions for A14 in B1 versus B2 are shifted by ~2 Å and appear to facilitate stacking interactions with either A13 or A15 in the structure. A14 is base paired to T3, which is hydrogen-bonded to R116 in the protein. Thus, the two distinct conformations observed for A14 may be a relevant property of this sequence or may be facilitated by the interactions of T3 with the protein. The LTR-C structure also includes the d(TCCCT), but in this case A13, equivalent to A14 in that it is found in the middle of the A-tract, is base-paired to T4, which is not hydrogen-bonded to the protein, and a single conformation was observed for this sequence.

Superpositioning of all C1' atoms in the four structures (26), independent of the sequence in each base pair, results in pairwise root mean square deviations (RMSDs) of 0.57–0.88 Å. Differences in the superimposed structures can be attributed largely to differences in sequence at each position within the 16mer. However, the similar sequence composition in the different structures potentially makes them more similar to each other than to the structures of unrelated sequences that we have determined in this same lattice for which pairwise superpositioning of C1' atoms results in RMSDs >1 Å.

In order to assess the degree of structural similarity, the same LTR sequences within the different crystal structures have been superimposed d(TTTCATT) from the LTR-A and -B1 structures (RMSD 0.63 Å), d(CTTTCAT) from LTR-B1 and -C structures (RMSD 0.49 Å), d(TCATT) from LTR-A, -B1 and -D structures (RMSDs 0.44–0.55 Å) and finally d(TCAT) from all of the structures (RMSDs from 0.42 to 0.49 Å) (Figure 2). The RMSD for d(TTTCATT) in LTR-B1 and LTR-B2 is 0.36 Å, for d(CTTTCAT) is 0.37 Å and for d(TCATT) is 0.23 Å. The structures for the same sequences within LTR-A, -B, -C or -D are quite similar to one another, as might be expected given that the sequences are the same and that the structures have all been solved in the same lattice. However, of particular interest are the structural differences that occur in the phosphodiester backbone atoms associated with A of the CA dinucleotide step and the T immediately 3′ of this A (Figure 2B and C).

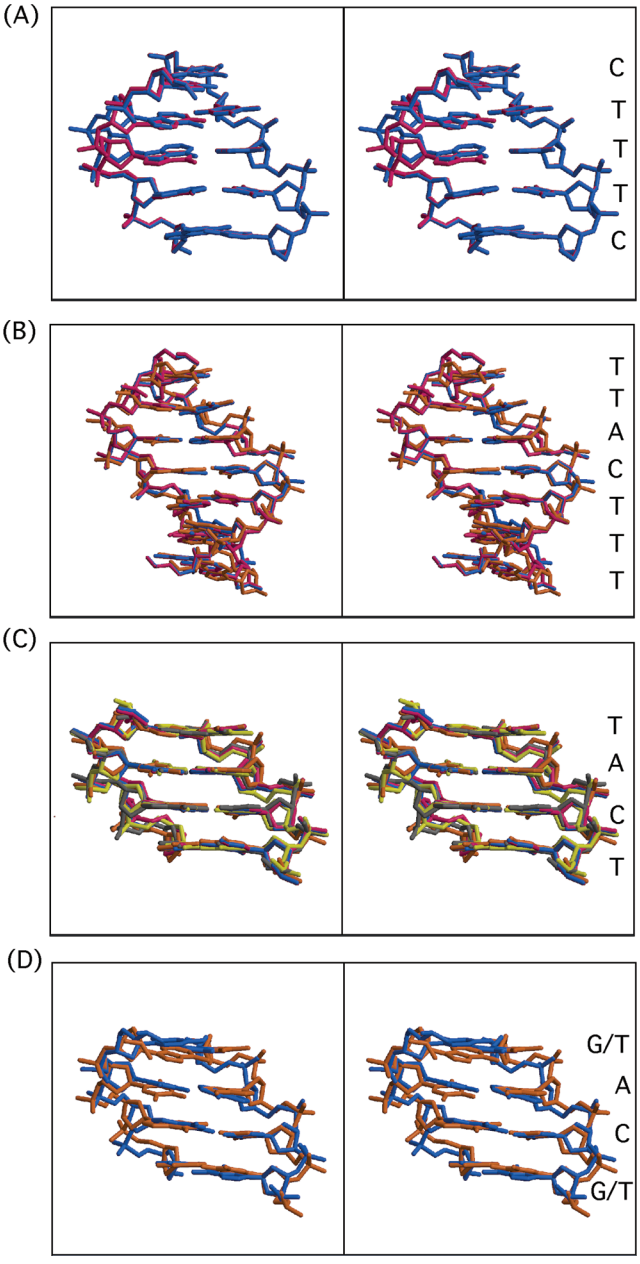

Figure 2. Stereorenderings of the LTR 3′ end processing sequences found in the different structures. The structures of (**A**) the first 5 bp of LTR-B1 (blue) and LTR-B2 (red) are superimposed highlighting the regions of conformational variation within this structure, (**B**) d(TTTCATT) from LTR-A (orange) and LTR-B (blue and red); (**C**) d(TCAT) from all four structures are superimposed. (**D**) The superimposed structures of d(TCAT) from the MMLV LTR-A structure (shown in orange) and d(GCAG) from one of the nine best calculated HIV-1 U5 LTR NMR structures (shown in light blue). The structures shown are the MMLV LTR and HIV-1 U5 LTR (17) with the most similar roll angles for the CA dinucleotide.

Analysis of the dinucleotide steps within the LTR

While it would clearly be of interest to compare our LTR structures to those of 'naked' DNA structures including d(TTCATT) sequences, there are currently no 'naked' DNA structures in the NDB or PDB including this sequence or even d(TCAT) sequences. In protein–DNA complexes including d(TCAT), these steps are involved in interactions with the protein (see Supplementary Data for more detail). Thus, we have performed a comparative analysis of our LTR structures and other non-LTR structures that we have determined in the same lattice.

Base pair step and helical parameters of the dinucleotide steps in the four LTR structures have been analyzed, specifically focusing on the dinucleotide steps within d(TTCATT).

In order to distinguish sequence-specific from position-specific parameters within our oligonucleotide structures, we have also compared the parameters of dinucleotide steps within the LTR structures to those found in three structures of unrelated sequences crystallized in the same lattice (PDB accession codes 1N4L, 2FJW, 1ZTW, which were determined at 2.0, 1.95 and 1.8 Å, respectively, with *R*-values ranging from 0.23 to 0.24 and R_{free} from 0.25 to 0.27). Accordingly, base-pair step roll, helical incline, slide and twist values were calculated using the program 3DNA (31) and compiled for each dinucleotide position within seven different oligonucleotides crystallized in the same lattice as shown in Tables 2 and 3 and Supplementary Tables S2 and S3.

Roll angles (Table 2) for dinucleotide steps 1, 2 and 3, which are involved in interactions with the protein in our structures, are all very similar in the different structures. In addition, dinucleotide steps 4 and 7 in the different structures appear to have a preference for a positive and negative roll angle, respectively. The average roll angle for position 4 in the LTR sequences is 4.5°, while for the non-LTR sequences it is 3.6°. For position 7, the average roll angle is −2.6° for the LTR sequences and −4.1° for the non-LTR sequences. Dinucleotide steps 5 and 6 show no particular preference for positive or negative roll angles; for these steps, the average roll angles are −0.1 and 0.9° for LTR sequences and −2.2 and 0.1° for non-LTR sequences, respectively. Helical incline values (Supplementary Table S2) follow very similar trends to those observed for the roll angles.

With the exception of the first three dinucleotide steps in which there are clear patterns consistent in both LTR and non-LTR sequences, slide values (Table 3) do not appear to exhibit position specific effects such as those seen for the roll angle at dinucleotide positions 4 and 7. Overall slide values within the LTR sequences ranging from −0.9 to 0.7 Å are somewhat smaller than those observed in the non-LTR sequences, −1.2 to 0.7 Å. Twist values (Supplementary Table S3) range from 27 to 42° with average values in both LTR and non-LTR sequences in each step position ranging from 30° to 36°. No patterns in twist values were observed for specific positions within the oligonucleotides or for a specific dinucleotide step.

LTR specific dinucleotide steps, TT, TC, CA and AT, that are not involved in interactions with the protein and thus represent 'naked' conformations within the LTR structures are found in several different dinucleotide step positions. The CA dinucleotide step is found at positions 4, 5 and 6; TT step at 4, 6 and 7; AT step at 5, 6 and 7; and TC step at 4, 5 and 7. (The dinucleotide step at position 4 from the duplex end includes the 4th and 5th nucleotides from the end of the duplex and similarly for the other steps). The CA dinucleotide step within the LTR sequence has a positive roll angle in all of the different structures. Note that the CA step in LTR-A at position 7, which has a negative roll angle, is not within the LTR sequence in that oligonucleotide, i.e. it does not have flanking 5′ and 3′ TT dinucleotide steps. AT dinucleotide steps within the LTR sequence have negative or zero roll angles, while TT and TC have both positive and negative roll angles.

For the MMLV LTR CA dinucleotide steps, we note a correlation of positive roll angle and negative slide that is independent of its position within the oligonucleotide for this particular dinucleotide within the LTR sequences. In each case, the positive roll value for the CA dinucleotide step exceeds that of the average roll angle for that step. The largest roll angle is observed in the LTR-A structure with a value of 9.9°. The magnitude of this value may in part result from the presence of ordered water molecules in the minor groove in this structure as shown in Figure 3 and discussed

Table 2. Roll angles for dinucleotide steps in LTR versus non-LTR sequences crystallized in the same lattice

	LTR-A		LTR-B1		LTR-B2		LTR-C		LTR-D		1N4L (23)		2FJW (29)		1ZTW (30)	
1	TT/AA	7.4	CT/AG	5.4	CT/AG	5.9	TC/GA	5.4	CA/TG	**3.4**	CT/AG	4.8	CT/AG	4.8	CT/AG	5
2	TT/AA	6.6	TT/AA	3.2	TT/AA	1.7	CT/AG	6	AC/GT	5.2	TT/AA	3.4	TT/AA	3.6	TT/AA	5.2
3	TC/GA	3.4	TT/AA	6.4	TT/AA	3.5	TT/AA	3.2	CA/TG	**2.9**	TT/AA	3.6	TG/CA	**8**	TA/TA	4.8
4	*CA/TG*	**9.9**	TC/GA	−0.4	TC/GA	4.6	TT/AA	3.4	AA/TT	4.8	TT/AA	4.6	GA/TC	2.9	AA/TT	3.2
5	AT/AT	−2.3	*CA/TG*	**0.9**	*CA/TG*	**0.4**	TC/GA	0.4	AT/AT	0	TT/AA	−3.7	AA/TT	−2.3	AT/AT	−0.6
6	TT/AA	−0.1	AT/AT	−1.7	AT/AT	−1.6	*CA/TG*	**4.9**	*TG/CA*	**2.8**	TA/TA	1.4	AT/AT	−0.8	TT/AA	−0.2
7	TG/CA	−**2.2**	TT/AA	−3.2	TT/AA	−3.6	AT/AT	−3.6	GA/TC	−0.6	AA/TT	−5.3	TG/CA	**0.2**	TC/GA	−7.3

LTR CA/TG steps are italicized and underlined. Roll angles (°) for both LTR and non-LTR CA/TG dinucleotide steps are in bold text. Dinucleotide steps below the horizontal line, steps 4–7, are free of interactions with the protein. The vertical line separates LTR from non-LTR structures. Non-LTR structures include 1N4L, 2FJW and 1ZTW (denoted by PDB accession code) with references to the original manuscripts.

Table 3. Slide values for dinucleotide steps in LTR versus non-LTR sequences crystallized in the same lattice

	LTR-A		LTR-B1		LTR-B2		LTR-C		LTR-D		1N4L (23)		2FJW (29)		1ZTW (30)	
1	TT/AA	0.4	CT/AG	0.2	CT/AG	0.2	TC/GA	0.1	CA/TG	**0.4**	CT/AG	0.1	CT/AG	0	CT/AG	0
2	TT/AA	0.1	TT/AA	−0.7	TT/AA	−0.6	CT/AG	0	AC/GT	−0.8	TT/AA	−0.7	TT/AA	−0.6	TT/AA	−0.4
3	TC/GA	0.2	TT/AA	−0.3	TT/AA	0.3	TT/AA	−0.2	CA/TG	**0.7**	TT/AA	0.2	TG/CA	**0.6**	TA/TA	0.6
4	*CA/TG*	−**0.3**	TC/GA	0.1	TC/GA	−0.2	TT/AA	−0.3	AA/TT	−0.4	TT/AA	−0.2	GA/TC	−0.1	AA/TT	−0.9
5	AT/AT	−0.3	*CA/TG*	−**0.8**	*CA/TG*	−**0.9**	TC/GA	0.3	AT/AT	−0.5	TT/AA	−0.2	AA/TT	−0.2	AT/AT	−0.4
6	TT/AA	−0.7	AT/AT	−0.3	AT/AT	−0.3	*CA/TG*	−**0.2**	*TG/CA*	−**0.3**	TA/TA	−0.5	AT/AT	−1.2	TT/AA	−1.1
7	TG/CA	**0.2**	TT/AA	−0.8	TT/AA	−0.8	AT/AT	−0.8	GA/TC	−0.6	AA/TT	−0.5	TG/CA	**0.2**	TC/GA	0.7

Designations for values are the same as those given in Table 2.

below. The combination of positive roll and negative slide has specific consequences for the positioning of hydrogen bonding atoms present within the CA step. In the minor groove of a CA dinucleotide step with negative roll and positive slide, O2 of C points towards the 3′ A. Whereas, for the CA dinucleotide with positive roll and negative slide, the O2 points away from the 3′A in the minor groove (Figure 2C) resulting in an O2 of C to N3 of A distance that is ~0.7 Å longer (in LTR-A) than that for the same atoms in the CA dinucleotide with negative roll. Within the major groove, the opposite trends are seen for the positioning of hydrogen-bonding atoms, namely that for the CA step with positive roll, the hydrogen bonding atoms in C and A are closer together than are those in the CA step with negative roll. The structural properties imparted by the positive roll angle and negative slide value of the CA dinucleotide step may provide a basis for recognition of the MMLV LTR by integrase.

A second observation with regard to parameters describing the dinucleotide steps is that the CA dinucleotide steps within the LTR sequences do not adopt as wide a range of roll angles as this step in non-LTR sequences, while the flanking TT dinucleotide steps adopt roll angle and slide values very similar to those found in non-LTR sequences. The LTR CA dinucleotide step roll angles vary from 0.93° to 9.9° and helical incline from 0.7° to 17.5°, while for the flanking TT dinucleotide steps, roll angles varied from −3.6° to 4.8° and helical incline values from −5.8° to 8.7°. The CA dinucleotide step in non-LTR DNA structures has previously been described as continuously flexible with highly correlated roll, slide and twist values (32). In this previous analysis of 60 'naked' DNA structures having both A and B backbone conformations, CA dinucleotide steps adopt both positive and negative roll values (~−10° to 10°) and exhibit primarily positive slide values (32). In an analysis of dinucleotide steps in oligonucleotides complexed with protein molecules, the CA dinucleotide step also adopts both positive and negative roll angles and was found to stand out as one of the most variable steps (33). In contrast, the TT/AA dinucleotide step has previously been reported to be the most rigid step adopting a smaller range of roll angles (~−5° to 5°) than other dinucleotide steps with predominantly negative slide values (32), while these same dinucleotide steps within oligonucleotides complexed to proteins adopt a much wider range of roll angles (33). Thus, although the CA dinucleotide step has been reported to be highly flexible (32–35), the CA dinucleotide steps in our structures appear to adopt a smaller subset of the range of roll values observed in non-LTR sequences, namely only positive roll angles. We attribute this property of the CA dinucleotide steps within the MMLV LTR sequences to the effects of the rigid 3′ and 5′ flanking TT dinucleotide steps.

The values of the roll angles and slide values for the TC and AT dinucleotide steps that each include 1 nt from the CA dinucleotide step are similar to those reported for the 'naked' DNA structures (32). The TC and AT dinucleotide steps adopt roll angles of −0.6° to 4.6° and 0° to −3.6° and slide values of −0.6 to 0.3 Å and −0.3 to −0.8 Å, respectively. Interestingly, the roll angles of the TC step within the LTR-B1 and B2 structures are −0.4° and 4.6°, while slide values are 0.1 and −0.2 Å, respectively, indicating that a change in backbone conformation results in an ~5° change in the roll angle and 0.3 Å shift in the slide value. Consistent with the findings from a previous analysis (32), the AT step also adopts a lower than average twist value.

The structural properties of the CA dinucleotide step in the LTR–LTR junction structure LTR-B do not differ significantly from those found in the other sequences as shown in Figure 2. The roll angles adopted for the CA dinucleotide step in both LTR-B1 and LTR-B2 structures are positive albeit smaller in magnitude than those of CA dinucleotides in the other LTR structures. Correspondingly, slide values are negative but larger in magnitude for the LTR-B structures than in the other LTR structures. However, retention of the positive roll angles and negative slide values suggests that structurally the CA dinucleotide steps within the LTR-B structures are similar to those in the other LTR structures. Thus, it does not appear to be differences in the structures or properties of the LTR sequences within an LTR–LTR junction mimic that account for the preference of the linear LTR as the substrate for integration.

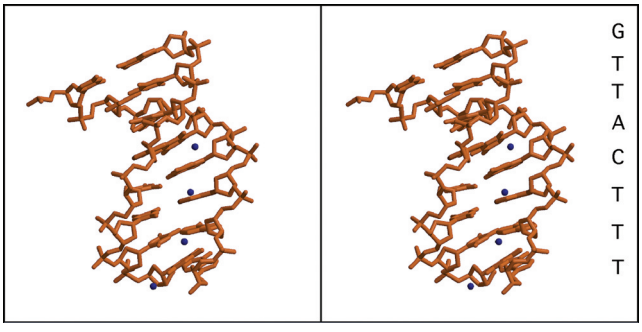

Figure 3. Hydration of LTR-A DNA. The water molecules that interact with the minor groove of the LTR-A DNA are represented as blue spheres. Only one half of the LTR-A DNA, the unique 8mer, is shown for clarity.

Solvent within the LTR-A structure

Another aspect of the DNA structures that may contribute to recognition by the retroviral integrase is the associated solvent structure. The arrangement of water molecules surrounding the DNA has long been reported to affect the DNA's structure and stability through interaction with the sugar-phosphate backbone and the nitrogen and oxygen atoms of the major and minor grooves (36). A narrow minor groove has the propensity to have an extensive spine of hydration formed by water molecules diagonally bridging two atoms in adjacent bases, which are subsequently bridged by water molecules that form a second layer (37,38). With a wide minor groove, a distinct spine of hydration was observed in which there is one hydrogen-bonded water molecule per base, resulting in a side-by-side ribbon of water molecules (38,39).

The LTR-A structure is the only structure that we have determined to date in which there is a partial spine of hydration including four water molecules as shown in Figure 3. The first water molecule is hydrogen bonded to the 5′ strand, to the O2 atom of thymine (1B, refers to residue 1 of chain B in the coordinate file) and O4′ atom of its sugar with distances

of 2.8 and 3.0 Å, respectively. Further interactions of this water molecule extend to the OH group of Tyr64 and Oδ2 of Asp114. The second and third water molecules interact only with the bases in the 3′ strand of the DNA. The second water molecule bridges the N3 atom of adenine (15D) and the sugar atom O4′ with bonding distances of 2.6 and 3.4 Å, respectively. In the third step, the third water molecule hydrogen bonds only to the N3 atom of adenine (14D). The hydrogen bonds formed by the fourth water molecule are interesting in that these are similar to the hydrogen bonding pattern observed in the original spine of hydration, wherein a water molecule bridges two atoms in adjacent bases. This fourth water molecule bridges the N2 atom of guanine (13D) and, in a longer range interaction, the N3 atom of the adenine (5B) from the opposite strand, with bonding distances of 2.6 and 3.5 Å, respectively. This fourth water molecule in fact bridges atoms within the CA dinucleotide step in this structure and likely contributes to its relatively large positive roll angle. Thus, solvent structure may play a role in the structural features associated with the CA dinucleotide step as seen in the LTR-A structure.

Comparison to the U5 HIV-1 LTR

In comparing the properties of the CA dinucleotide steps found in our MMLV LTR sequences with the CA dinucleotide step from the U5 HIV-1 LTR (17), we find that some base pair specific properties associated with the CA dinucleotide step are conserved. The roll values of the CA dinucleotide step in the nine best NMR structures of the U5 HIV-1 LTR (17) are all positive ranging from 0.5° to 11.9° and slide values are negative, −0.9° to −1.2°. However, the CA dinucleotide steps in the MMLV LTR structures are all B-form steps, whereas in the HIV-1 U5 LTR, this step is not B-form as analyzed in 3DNA (31).

Superpositioning of the C1' atoms of the d(GCAG) from the U5 HIV-1 LTR and d(TCAT) from the LTR-A structure, which has very similar roll and helical incline angles (roll angles of 10.3° versus 9.7° and incline angles of 16.1° and 17.5° for HIV-1 U5 LTR and MMLV LTR-A, respectively) to one of the NMR structures, results in an RMSD of 0.6 Å as shown in Figure 2D. The superimposed structures show distinct differences that likely result from the differences in the flanking sequences. The only base that superimposes well is the C of the CA dinucleotide step. In particular, a structural feature that is not conserved in the MMLV and HIV-1 LTR sequences is the interstrand stacking of the A15 and G21 within the CA dincleotide step. This may be a unique structural feature associated with the specific sequence d(AGCAGT) found at the end of the U5 HIV-1 LTR. If the G immediately 5′ to the CA dinucleotide is replaced with a T, HIV-1 IN fails to efficiently cleave and integrate the substrate (16). As the MMLV LTR includes a T immediately 5′ to the CA dinculeotide step and there is no interstrand base stacking of the G and A in our structures, it may be that the 5′ T confers a different structure. Thus, the flanking sequences may be a distinguishing feature in the recognition of HIV-1 IN and MMLV IN of LTR substrates.

Implications for MMLV integrase recognition and mechanism

Based on the mutational studies performed on the LTRs of retroviruses including MMLV and HIV-1, the CA step has been identified as a critical factor in the processing of this end since this sequence is preferentially cleaved. Previous studies showed strong evidence that the ends of the DNA are distorted, consistent with unpairing and unstacking, resulting from interactions of integrase and that this distortion is a critical step in the processing reaction (40,41). The structural differences observed in the backbone atoms of A of the CA step and the 3′ T within the d(TCAT) sequence are consistent with inherent flexibility in the region of the processing site that may play a role in the ability of integrase to distort the LTR ends.

Integration of retroviral DNA is a very complex process. Although biochemical studies have provided insights regarding the cleavage and strand transfer mechanisms, the mechanism of the initial recognition of the integrase protein to the LTR has yet to be detailed. In this study, a conserved positive roll angle and negative slide value are associated with the CA dinucleotide step within the MMLV LTR 3′ end processing site, which impart specific structural features that may be necessary for the MMLV integrase to recognize and subsequently cut 3′ of the CA step. Specifically, the positions of hydrogen bonding atoms within the major and minor groove for the CA dinucleotide step that result from the positive roll angle as compared to that for the same dinucleotide step with a negative roll angle suggests a feature that might allow integrase to recognize this dinucleotide step.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank the National Institutes of Health for financial support of this work (GM55026 to M.M.G. and GM070837 to M.J.R.). Funding to pay the Open Access publication charges for this article was provided by institutional funds available to M.M.G.

Conflict of interest statement. None declared.

REFERENCES

1. Grandgenett,D.P. and Mumm,S.R. (1990) Unraveling retrovirus integration. *Cell*, **60**, 3–4.
2. Van Maele,B. and Debyser,Z. (2005) HIV-1 integration: an interplay between HIV-1 integrase, cellular and viral proteins. *AIDS Rev*., **7**, 26–43.
3. Gilboa,E., Mitra,S., Goff,S.P. and Baltimore,D. (1979) A detailed model of reverse transcription and tests of crucial aspects. *Cell*, **18**, 93–100.
4. Telesnitsky,A. and Goff,S.P. (1997) Reverse transcriptase and the generation of retroviral DNA. In Coffin,J.M., Hughes,S.H. and Varmus,H.E. (eds), *Retroviruses*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, pp. 121–150.
5. Hughes,S.H., Shank,P.R., Spector,D.H., Kung,H.J., Bishop,J.M., Varmus,H.E., Vogt,P.K. and Breitman,M.L. (1978) Proviruses of avian

sarcoma virus are terminally redundant, co-extensive with unintegrated linear DNA and integrated at many sites. *Cell*, **15**, 1397–1410.
6. Bushman,F.D. (2003) Targeting survival: integration site selection by retroviruses and LTR-retrotransposons. *Cell*, **115**, 135–138.
7. Schroder,A.R., Shinn,P., Chen,H., Berry,C., Ecker,J.R. and Bushman,F. (2002) HIV-1 integration in the human genome favors active genes and local hotspots. *Cell*, **110**, 521–529.
8. Wu,X., Li,Y., Crise,B. and Burgess,S.M. (2003) Transcription start regions in the human genome are favored targets for MLV integration. *Science*, **300**, 1749–1751.
9. Mitchell,R.S., Beitzel,B.F., Schroder,A.R., Shinn,P., Chen,H., Berry,C.C., Ecker,J.R. and Bushman,F.D. (2004) Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. *PLoS Biol*., **2**, E234.
10. Brown,P.O., Bowerman,B., Varmus,H.E. and Bishop,J.M. (1989) Retroviral integration: Structure of the initial covalent product and its precursor, and a role for the viral IN protein. *Proc. Natl Acad. Sci. USA*, **86**, 2525–2529.
11. Lobel,L.I., Murphy,J.E. and Goff,S.P. (1989) The palindromic LTR–LTR junction of Moloney murine leukemia virus is not an efficient substrate for proviral integration. *J. Virol*., **63**, 2629–2637.
12. Colicelli,J. and Goff,S.P. (1988) Sequence and spacing requirements of a retrovirus integration site. *J. Mol. Biol*., **199**, 47–59.
13. Bushman,F.D. and Craigie,R. (1990) Sequence requirements for integration of Moloney murine leukemia virus DNA *in vitro*. *J. Virol*., **64**, 5645–5648.
14. Roth,M.J., Schwartzberg,P.L. and Goff,S.P. (1989) Structure of the termini of DNA intermediates in the integration of retroviral DNA: dependence on IN function and terminal DNA sequence. *Cell*, **58**, 47–54.
15. Colicelli,J. and Goff,S.P. (1985) Mutants and pseudorevertants of Moloney murine leukemia virus with alteration at the integration site. *Cell*, **42**, 573–580.
16. Esposito,D. and Craigie,R. (1998) Sequence specificity of viral end DNA binding by HIV-1 integrase reveals critical regions for protein-DNA interaction. *EMBO J*., **17**, 5832–5843.
17. Renisio,J.G., Cosquer,S., Cherrak,I., El Antri,S., Mauffret,O. and Fermandjian,S. (2005) Pre-organized structure of viral DNA at the binding-processing site of HIV-1 integrase. *Nucleic Acids Res*., **33**, 1970–1981.
18. Vink,C., van Gent,D.C., Elgersma,Y. and Plasterk,R.H. (1991) Human immunodeficiency virus integrase protein requires a subterminal position of its viral DNA recognition sequence for efficient cleavage. *J. Virol*., **65**, 4636–4644.
19. Sun,D., Jessen,S., Liu,C., Liu,X., Najmudin,S. and Georgiadis,M.M. (1998) Cloning, expression, and purification of a catalytic fragment of Moloney murine leukemia virus reverse transcriptase: crystallization of nucleic acid complexes. *Protein Sci*., **7**, 1575–1582.
20. Cote,M.L., Yohannan,S.J. and Georgiadis,M.M. (2000) 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. *Acta Crystallogr. D Biol. Crystallogr*., **56**, 1120–1131.
21. Cote,M.L. and Georgiadis,M.M. (2001) Structure of a pseudo-16-mer DNA with stacked guanines and two G-A mispairs complexed with the N-terminal fragment of Moloney murine leukemia virus reverse transcriptase. *Acta. Crystallogr. D Biol. Crystallogr*., **57**, 1238–1250.
22. Cote,M.L., Pflomm,M. and Georgiadis,M.M. (2003) Staying straight with A-tracts: a DNA analog of the HIV-1 polypurine tract. *J. Mol. Biol*., **330**, 57–74.
23. Najmudin,S., Cote,M.L., Sun,D., Yohannan,S., Montano,S.P., Gu,J. and Georgiadis,M.M. (2000) Crystal structures of an N-terminal fragment from Moloney murine leukemia virus reverse transcriptase complexes with nucleic acid: functional implications for template-primer binding to the fingers domain. *J. Mol. Biol*., **296**, 613–632.
24. Navaza,G. (1994) AMoRe: an automated package for molecular replacement. *Acta Crystallogr*., **A50**, 157–163.
25. Brunger,A.T., Adams,P.A., Clore,G.M., Gros,P., Grosse-Kunstleve,R.W. and Jiang,J.-S. (1998) Crystallography and NMR system (CNS): a new software system for macromolecular structure determination. *Acta Crystallogr. D Biol. Crystallogr*., **54**, 905–921.
26. Parkinson,G., Vojtechovsky,J., Clowney,L., Brunger,A.T. and Berman,H.M. (1996) New parameters for the refinement of nucleic acid-containing structures. *Acta Crystallogr. D Biol. Crystallogr*., **52**, 57–64.
27. Macke,T. and Case,D.A. (1998) In Leontes,N.B. and J. SantaLucia,J. (eds), *In Molecular Modeling of Nucleic Acids*. American Chemical Society, Washington, DC, pp. 379–393.
28. Jones,T.A., Zou,J.Y., Cowan,S.W. and Kjeldgaard,M. (1991) Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystalllogr. A*, **47**, 110–119.
29. Goodwin,K.D., Lewis,M.A., Tanious,F.A., Tidwell,R.R., Wilson,W.D., Georgiadis,M.M. and Long,E.C. (2006) A High-throughput, high-resolution strategy for the study of site-selective DNA binding agents: analysis of a 'highly twisted' benzimidazole-diamidine. *J. Am. Chem. Soc*., **128**, 7846–7854.
30. Goodwin,K.D., Long,E.C. and Georgiadis,M.M. (2005) A host-guest approach for determining drug-DNA interactions: an example using netropsin. *Nucleic Acids Res*., **33**, 4106–4116.
31. Lu,X.-J. and Olson,W.K. (2003) 3DNA: a software package for the analysis, rebuilding, and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res*., **31**, 5108–5121.
32. Hassan,M.A.E. and Calladine,C.R. (1997) Conformational characteristics of DNA: empirical classifications and a hypothesis for the conformational behaviour of dinucleotide steps. *Phil. Trans. R. Soc. Lond. A*, 43–100.
33. Olson,W.K., Gorin,A.A., Lu,X.J., Hock,L.M. and Zhurkin,V.B. (1998) DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc. Natl Acad. Sci. USA*, **95**, 11163–11168.
34. Packer,M.J., Dauncey,M.P. and Hunter,C.A. (2000) Sequence-dependent DNA structure: tetranucleotide conformational maps. *J. Mol. Biol*., **295**, 85–103.
35. Packer,M.J., Dauncey,M.P. and Hunter,C.A. (2000) Sequence-dependent DNA structure: dinucleotide conformational maps. *J. Mol. Biol*., **295**, 71–83.
36. Westhof,E. (1998) Water: an integral part of nucleic acid structure. *Annu. Rev. Biophys. Biophys. Chem*., **17**, 125–144.
37. Drew,H.R. and Dickerson,R.E. (1981) Structure of a B-DNA dodecamer: geometry of hydration. *J. Mol. Biol*., **151**, 535–556.
38. Prive,G.G., Yanagi,K. and 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.
39. Grzeskowiak,K., Yanagi,K., Prive,G.G. and Dickerson,R.E. (1991) The structure of B-helical C-G-A-T-C-G-A-T-C-G and comparison with C-C-A-A-C-G-T-T-G-G. The effect of base pair reversals. *J. Biol. Chem*., **266**, 8861–8883.
40. Katz,R.A., DiCandeloros,P., Kukolj,G. and Skalka,A.M. (2001) Role of DNA end distortion in catalysis by avian sarcoma virus integrase. *J. Biol. Chem*., **276**, 34213–34220.
41. Scottoline,B.P., Chow,S., Ellison,V. and Brown,P.O. (1997) Disruption of the terminal base pairs of retroviral DNA during integration. *Genes. Dev*., **11**, 371–382.
42. Kraulis,P.J. (1991) MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Cryst*., **24**, 946–950.