Proc. Natl. Acad. Sci. USA
Vol. 95, pp. 15194–15199, December 1998
Biochemistry

Structural code for DNA recognition revealed in crystal structures of papillomavirus E2-DNA targets

HAIM ROZENBERG*, DOV RABINOVICH*, FELIX FROLOW†‡, RASHMI S. HEGDE§, AND ZIPPORA SHAKKED*¶

*Department of Structural Biology, Weizmann Institute of Science, Rehovot 76100, Israel; †Department of Chemical Services, Weizmann Institute, and ‡Department of Molecular Microbiology and Biotechnology, Tel-Aviv University, Ramat-Aviv, Tel-Aviv 69978 Israel; and §Program in Structural Biology and Department of Biochemistry, Skirball Institute, New York University Medical Center, New York, NY 10016

Communicated by Paul B. Sigler, Yale University, New Haven, CT, October 12, 1998 (received for review July 30, 1998)

ABSTRACT **Transcriptional regulation in papillomaviruses depends on sequence-specific binding of the regulatory protein E2 to several sites in the viral genome. Crystal structures of bovine papillomavirus E2 DNA targets reveal a conformational variant of B-DNA characterized by a roll-induced writhe and helical repeat of 10.5 bp per turn. A comparison between the free and the protein-bound DNA demonstrates that the intrinsic structure of the DNA regions contacted directly by the protein and the deformability of the DNA region that is not contacted by the protein are critical for sequence-specific protein/DNA recognition and hence for gene-regulatory signals in the viral system. We show that the selection of dinucleotide or longer segments with appropriate conformational characteristics, when positioned at correct intervals along the DNA helix, can constitute a structural code for DNA recognition by regulatory proteins. This structural code facilitates the formation of a complementary protein–DNA interface that can be further specified by hydrogen bonds and nonpolar interactions between the protein amino acids and the DNA bases.**

Papillomaviruses are DNA tumor viruses that infect a variety of mammalian species, including humans. Transcriptional regulation and replication in this family involves the sequence-specific binding of the papillomavirus E2 protein to several sites in the viral genome. In the well-studied bovine papillomavirus type 1 (BPV-1), the E2 protein binds to a consensus dodecameric sequence of the general form ACCG**NNNN**CGGT ($\mathbf{N}_4$ is variable), which is found at several locations on the viral genome. These DNA targets vary in their binding affinities over a 300-fold range depending on the identity of the central noncontacted region (1). Because the sequence involved in indirect effects can be varied independently, BPV-1 provides an attractive system for studying the relationship between DNA deformation/deformability and the specific affinity of a eukaryotic transcription factor for a gene regulatory element.

The crystal structure of the DNA-binding domain of the BPV-1 E2 protein complexed to a DNA target incorporating the conserved nucleotides (italicized), CCG*ACCG*ACGT*CGGT*CG, has been determined in two crystal forms at 1.7- and 2.5-Å resolution (2, 3). In the complex, the DNA is severely bent toward the protein by the interaction of a pair of symmetrically disposed α-helices with the two recognition sites in symmetrically related regions of the major groove. The specific interface is made up of a network of interactions in which more than one direct contact exists between each of the base pairs in the ACC/GGT trinucleotides of the conserved regions and the discriminating amino acids. These contacts could not all occur at once without significant deformation of the DNA target. Hence, sequence-dependent deformation and/or deformability of the DNA appear crucial for binding specificity in this system.

To study the role of DNA structure and deformability in this regulatory system, it is essential to compare the detailed three-dimensional structure of the E2-DNA target in its free state with that of the DNA in its complex with the E2 protein. We have investigated several DNA oligomers incorporating the consensus tetranucleotide motifs (ACCG/CGGT) by using x-ray crystallography. Here we present the x-ray analyses of three crystal structures; two of the dodecamer *ACCG*ACGT*CGGT*, which is identical in sequence to the central 12-bp region of the bound target (2), and one of a second dodecamer *ACCG*GTAC*CGGT*, which differs from the other two by the composition of the central 4 bp. The second sequence was chosen to study the effect of sequence variations in the central noncontacted region on DNA structure, which could affect its binding affinity to the E2 protein.

Our x-ray analyses show that all of the dodecameric structures adopt a new variant of the B form of DNA characterized by a roll-induced writhe and a helical repeat of 10.5 bp per turn. Comparison between the free and bound targets provides structural insight at atomic resolution into the possible mechanism of DNA deformation on protein binding and demonstrates that sequence-dependent DNA structure and inherent deformability are essential determinants in the recognition of the E2-DNA binding sites by their cognate protein.

MATERIALS AND METHODS

DNA Synthesis, Purification, and Crystallization. Oligonucleotides were prepared by the phosphoramidite method on an Applied Biosystems 380B DNA synthesizer and purified by reverse-phase HPLC using a C_4 matrix. The collected fractions were dialyzed against deionized water and then lyophilized.

Crystals of the DNA dodecamer ACCGACGTCGGT were obtained in triclinic and rhombohedral forms. Triclinic crystals of ACCGACGTCGGT (denoted as E2-P1) were grown at 4°C by the hanging-drop method from 5-μl drops containing 2 mg/ml DNA, 10 mM $MgCl_2$, 2 mM spermine tetrachloride, 20 mM sodium cacodylate buffer (pH 7), and 5% (wt/vol) 2-methyl-2,4-pentanediol (MPD) equilibrated against a reservoir of 17% (wt/vol) MPD in 100 mM sodium cacodylate buffer (pH 7).

Rhombohedral crystals of ACCGACGTCGGT (denoted as E2-R3) were grown at 19°C by using the hanging-drop method

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

Data deposition: The atomic coordinates and the structure factors reported in this paper have been deposited in the Nucleic Acid Database, Rutgers, The State University of New Jersey, Wright and Reiman Labs, Department of Chemistry, 610 Taylor Road, Piscataway, NJ 08854-8087 (NDB structure ID codes bd0001, bd0002, and bd0003).

A Commentary on this article begins on page 15163.

¶To whom reprint requests should be addressed. e-mail: cshaked2@weizmann.weizmann.ac.il.

combined with macroseeding. The best crystals were obtained by seeding tiny plate-like crystals (0.2×0.1 mm^2) into 5-μl drops containing 1.8 mg/ml DNA, 70–80 mM $MgCl_2$, 1–2 mM spermine·tetrachloride, 20 mM sodium cacodylate buffer (pH 7), and 6% (wt/vol) MPD. The drops were equilibrated against a reservoir of 23% MPD in 100 mM sodium cacodylate buffer (pH 7).

Crystals of the DNA dodecamer ACCGGTACCGGT (denoted as E2-P4) were grown at 4°C from 5-μl drops containing 2 mg/ml DNA, 32 mM $MgCl_2$, 3–10 mM $MnCl_2$, 7 mM spermine·tetrachloride, 20 mM sodium cacodylate buffer (pH 7), and 5% (wt/vol) MPD equilibrated against 1-ml solutions of 18% MPD in 100 mM sodium cacodylate buffer (pH 7).

Data Collection and Structure Determination. All crystals were covered with Exxon Paraton oil and flash-cooled to 100–120 K. Data from E2-R3 were collected on a Rigaku R-AXIS IIC detector mounted on a Rigaku flourescent resonance chromatography rotating-anode generator by using Ni-filtered CuKα radiation focused by Yale-type mirrors. Data were processed with DENZO-SCALEPACK (4). Data from a crystal of E2-P1 were measured on a Rigaku AFC5R diffractometer mounted on a Rigaku Ru300 rotating-anode generator by using graphite-monochromated CuKα radiation. Data from a crystal of E2-P4 were collected on a Xentronics area detector mounted on a Rigaku rotating-anode generator Ru300 by using graphite-monochromated CuKα radiation. The data were processed by the XDS package (5). Crystal data and intensity statistics for the three crystals are given in Table 1.

The crystal structures of E2-R3 (space group *R*3) and E2-P4 (space group $P4_1$ or $P4_3$), each with one duplex in the asymmetric unit, were solved by ULTIMA (6) using fiber-based B-DNA as a search model. In the case of E2-P4, the space group $P4_3$ was established by the structure solution. The structure of E2-P1 with three duplexes in the triclinic unit cell was determined by a procedure for solving large structures that combines the molecular Fourier-transform method (MFT) with a modified version of ULTIMA (7).

Structure Refinement. All three structures were subjected first to a rigid-body refinement by X-PLOR (8) using stereochemical DNA parameters from Parkinson *et al.* (9), dividing each duplex first into 12 rigid groups and then into 70 groups (phosphate, sugar, and base), gradually raising the resolution range from around 15–5 Å to 8–3.5 Å. In view of the limited resolution of the data in the cases of E2-P1 and E2-P4, strong restraints were applied on several geometrical parameters.

The next stage of the refinement by X-PLOR involved alternated cycles of energy minimization and refinement of individual isotropic temperature factors, an overall anisotropic temperature factor, and bulk solvent correction. In the case of E2-R3, two rounds of simulated annealing were done before and during the individual-atom refinement. After each round of several cycles of refinement, the electron-density maps were examined to manually correct the model and locate solvent molecules using the program O (10). Solvent peaks that satisfied hydrogen-bonding criteria and were adequately represented in the electron density were included in the model as oxygen atoms with restrained positions. Several magnesium ions were identified in the structure of E2-R3 on the basis of their octahedral hydration shell. The free R-factor was monitored throughout the process to verify the validity of each step of the refinement procedure with the exception of E2-P4 because of limited data. The refinement of E2-R3 was continued with SHELXL-97 package (11) using the model refined by X-PLOR with the exclusion of solvent molecules. The refinement was based on F^2 and included all measured data to 1.6 Å (506 reflections were used for free R test). Electron-density maps were calculated with Sigma-A-weighted Fourier coefficients (12). The solvent structure was rebuilt in several rounds by using both the automated water-location option of the package and manual fitting. The final 10 cycles of refinement included all data. The refinement results of the three structures are given in Table 1.

RESULTS AND DISCUSSION

Crystal Structures of the E2-DNA Targets: A Conformational Variant of B-DNA. The dodecamer *ACCG*ACGT*CGGT* has been analyzed in two crystal forms (*R*3 and *P*1) obtained by using crystallization conditions that differed in temperature and Mg^{2+} concentration (see *Materials and Methods*). The structures were determined at resolutions of 1.6 and 2.7 Å, respectively. The rhombohedral structure contains one duplex in the asymmetric unit and the triclinic structure contains three duplexes in the unit cell, thus providing us with four unique copies of the same molecule. The dodecamer *ACCG*GTAC*CGGT* has been crystallized in space group $P4_3$ with one unique duplex and has been determined at 2.8-Å resolution. The sequence differs from the complexed target and the other free dodecamer in the central noncontacted sequence (GGTACC versus GACGTC). Because the E2-dodecamer crystallizing in the *R*3 form has been determined at the highest

Table 1. Crystallographic data and refinement statistics

	Sequence		
	ACCGACGTCGGT		ACCGGTACCGGT
	Space group *R*3	Space group *P*1	Space group $P4_3$
Unit cell dimensions (Å and °)	a = b = 64.07, c = 44.68	a = 40.5, b = 40.1, c = 40.5 α = 82.6, β = 116.2, γ = 80.7	a = b = 40.2, c = 57.6
No. of independent DNA duplexes	1	3	1
Volume per base-pair, Å^3	1,471	1,576	1,939
Resolution limits, Å	23.6–1.6	19.0–2.7	25.5–2.8
No. of measured reflections with $I > 0$	68,426	5,668	13,447
No. of unique reflections ($\|F\| > 2\sigma(F)$)	8,888 (8,888)	5,415 (4,608)	2,230 (2,164)
Completeness of data, %	98.7 (95.0)	89.4 (75.5)	97.5 (86.8)
R_{sym} (I), %	5.2 (23.5)	6.5 (18.0)	5.6 (21.0)
Upper resolution shell, Å	1.60–1.62	2.8–2.7	2.9–2.8
R-factor/R_{work}/R_{free}, %	17.4/17.3/20.6	21.6/21.1/28.3	21.9
No. of DNA atoms/solvent molecules	486/162	1,458/18	486/8
rms deviations:			
Bond length, Å	0.01	0.01	0.01
Bond angle, °	2.3	1.1	1.3
Bonded B-factor, Å^2	4.1	2.9	4.1

The values in parentheses refer to the data of the corresponding upper resolution shells.

resolution, its structure will be referred to as the "free" DNA structure.

The free E2 dodecamer is extensively hydrated. The second hydration shell incorporates five Mg^{2+} ions (three cations located at the major groove and two at the minor groove) bound in octahedral geometry to six water molecules. Three of the hydrated ions link two or three neighboring molecules in the crystal, thus contributing to crystalline order and hence to the unusually high-resolution diffraction of the crystals. The hydrated duplex is shown in Fig. 1. A close-up view of the experimental electron density corresponding to a hydrated magnesium ion that links three duplex molecules is shown in Fig. 2. Magnesium ions have been used routinely in crystallization of DNA oligomers and their complexes with other molecules, but only high-resolution structures allow them to be unambiguously identified in electron-density maps. The modes of Mg-mediated interactions observed here may be relevant to cellular processes, where such cations are required for physiological binding of DNA to other molecules such as in DNA packaging and processing.

The four independent duplexes of *ACCG*ACGT*CGGT* and the single copy of *ACCG*GTAC*CGGT* are illustrated side-by-side in Fig. 3. The rms pairwise differences between the various molecules range from 0.7 to 1 Å. The global and mean local parameters of the five E2-helices are given in Table 2. All five helices are straight and display a gentle roll-induced writhe so that the base pairs are positively inclined to the helix axis by 4–7°, and the mean local roll angle between adjacent base pairs is close to 4°. We use the term "writhe" here to describe a structure where the local axes of the base pairs are nonparallel and precess around the global helix axis (see Fig. 4). It is therefore different from the writhing number (W_r) used to characterize the global folding of the helix axis of a supercoiled DNA. As a result of the writhe and the positive inclination of the base pairs, each of the E2-helices is continuously "bent" toward the major groove in a manner qualitatively similar to A-DNA and TA-DNA, where the corresponding inclination angles are 10–20° for the former and close to 50° for the latter (18). All helices are also similarly wound, with average helical repeats ranging from 10.3 to 10.5 bp per turn (Table 2).

The average local roll angle in previously reported B-DNA structures (19) and in general sequence fiber B-DNA (20) has been shown to be close to zero and as a result, the average inclination is close to zero. In addition, the average helix twist has been found to be close to 36° in free B-DNA crystal

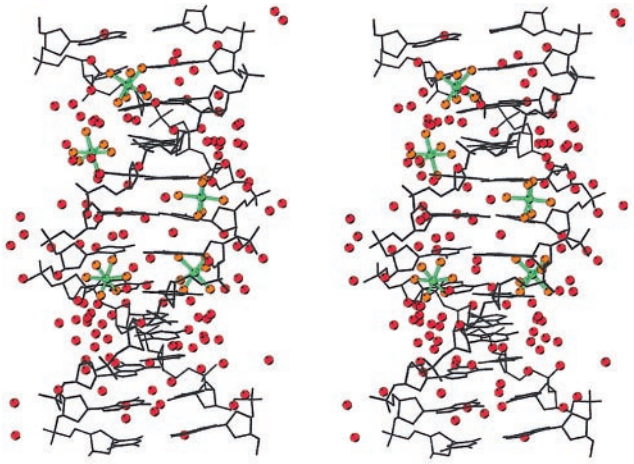

FIG. 1. A stereoview of the hydrated E2-R3 dodecamer. Water molecules are shown in red, the octahedrally coordinated magnesium ions in green, and the attached water molecules in orange. Only the 157 water molecules of the asymmetric unit are shown. Three hydrated Mg^{2+} ions are in the major groove and two are in the minor groove. The picture was produced with MOLSCRIPT 2.0 (13).

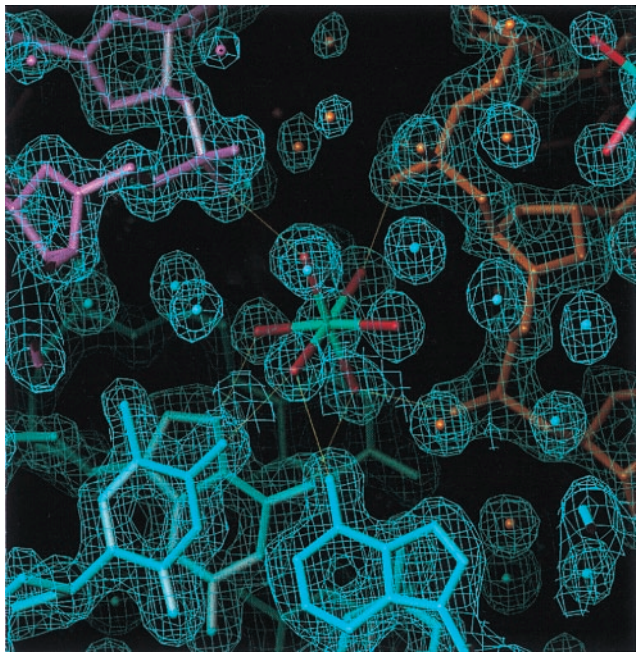

FIG. 2. A view of a hexahydrated magnesium ion (in green) bridging three DNA duplexes with the corresponding electron density (2*Fo*-*Fc*) map at 1.6 Å resolution contoured at 1.5σ. Hydrogen bonds between the Mg-coordinated water molecules (shown in red) and base or backbone donor/acceptor atoms from three helices (shown in blue, orange, and magenta) are drawn as yellow lines. The figure was prepared with O 5.10 (10).

structures, except in a few cases (ref. 21 and references therein). Thus, the structure observed here, characterized by a roll-induced writhe and a helical repeat of 10.5 bp per turn represents a conformational variant of B-DNA that we term Writhed B-DNA or WB-DNA. It should be emphasized that continuous writhe need not change the overall direction of a DNA double helix, unless it is joined to helices with significantly different writhe properties (15, 18). This kind of B-DNA structure was proposed to explain the phenomenon of DNA bending induced by phased adenine tracts (22, 23). The groove dimensions of the five helices also deviate significantly from canonical B-DNA (20). The major groove is deeper by 1.5 Å and the minor groove is wider by 1 Å with respect to regular B-DNA, whereas the depth of the minor groove and the width of the major groove are similar to that of B-DNA (Table 2).

The common conformational characteristics of the free DNA targets do not appear to be induced by crystal-packing forces because all five helices are embedded in different crystalline environments and are relatively "loosely" packed (see volume per base pair in Table 1) in comparison to other DNA structures (24). The observation that all of these molecules adopt similar global shapes may be attributed to the identity of the common tetranucleotides. It is possible, however, that this kind of structure represents more closely the conformation of general sequence B-DNA in solution.

Comparison Between the Free and Bound E2 Helices. The 16-bp DNA target in its complex with the E2-DNA binding domain is significantly bent toward the protein, forming a tight concave interface (2). Two views of the complexed target and the free dodecamer are displayed in Fig. 4. Each view shows the best-fitted global-helix axis (continuous green line) and the local axes of the base pairs (shown in red). The overall helix axis in the free molecule is essentially straight, as indicated by the best fitted global axis. The writhe of the local axes around the global axis is highlighted by the inclined red arrows spiraling around the continuous green line. Also shown for comparison are the analogous views of general sequence

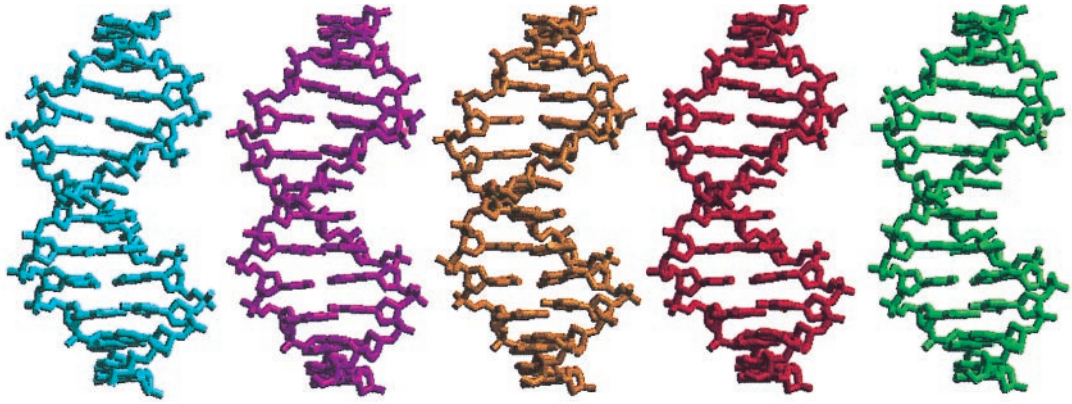

FIG. 3. The five unique E2 helices of the free DNA crystal structures. From left to right: E2-R3 (cyan), the three E2-P1 molecules (magenta, orange, and red), and E2-P4 (green). Despite of differences in crystal packing and base sequence, all five duplexes exhibit similar global structures with rms deviations ranging from 0.7 to 1.0 Å using all common atoms. The figure was drawn with SETOR (14).

B-DNA structure (20), where inclination and roll angles are close to zero and the local axes are nearly parallel to the straight global helix axis.

In contrast to the free DNA helix, the protein-bound DNA is significantly bent in a direction that compresses the minor groove at the central region and the major groove at the two flanking regions (Fig. 4). The overall axis follows a gentle left-handed superhelical trajectory rather than bending within a plane. The local axes of the base pairs display significant writhe with respect to the curved global axis, particularly at the recognition tetranucleotide segments. The similarity and dissimilarity between the various regions also are reflected by the corresponding rms differences. The rms difference between the conserved ACCG/CGGT regions of the bound and free helices (using all atoms) is 0.8 Å, which is comparable to the value obtained between the two equivalent domains of the free DNA molecule (0.6 Å), whereas the rms difference between the central ACGT regions of the free and bound helices is much larger (1.8 Å).

The changes in the relative motions of the base pairs at each step in the free and the bound helices are described in terms of their local parameters displayed in Fig. 5. These parameters, which include two rotations (roll and helix twist) and one translation (slide) have been shown to be most variable and sequence-dependent in DNA crystal structures and hence pivotal to DNA deformation and bending (27, 28).

The variations in the local parameters demonstrate that the geometrical features at the conserved and symmetrically disposed contact zones (ACCG/CGGT, shaded areas in Fig. 5.) are similar in the free and complexed DNA targets. The roll angles at these regions are mostly positive, i.e., bending toward the major groove. The values for the protein-bound helix are slightly larger than those for the free target as a result of the close interaction with the protein. The helix twist and slide patterns at the conserved 4-bp sequences also are very similar for the free and bound structures. Hence, the intrinsic sequence-induced features of the free ACCG/CGGT regions constitute structural identity elements that are recognized by the protein.

By contrast, there are major differences between the two structures at the central ACGT region, which is not contacted by the protein. The roll angles at the three core steps are negative, i.e., bending into the minor groove, in contradistinction to the positive values of the free target. However, the relative fluctuations in the parameters, which display alternation between low and high values, are almost identical in the free and protein-bound helices. The same kind of pattern is seen by the helix-twist and slide variations. Thus, the transition from the free to the bound conformation is achieved by a global motion at the central three steps that involves closing of the base pairs toward the minor groove by 10° coupled with helix winding of 5° and a positive sliding of 1 Å.

What is the energy cost associated with the conformational change imposed by protein binding? Energy calculations performed on the crystal-structure coordinates by a procedure described previously (29) show that the average difference in base-stacking energy is about 0.2 kcal/mole (1 kcal = 4.18 kJ) per helix step in favor of the free target (H.R. and Z.S., unpublished data). The small energy difference supports our hypothesis that it is the deformable nature of the noncontacted base sequence that is critical for recognition in this system, as discussed below.

Table 2. Average helix parameters

DNA sequence	Crystal form	Inclination, °	Roll, °	Helix twist, °	Rise, Å	Slide, Å	D_x, Å	Groove width, Å		Groove depth, Å	
								major	minor	major	minor
ACCGACGTCGGT	*R*3	6.6	3.7	34.4	3.3	0.05	−0.54	11.1	7.2	9.9	7.8
ACCGACGTCGGT-1	*P*1	5.0	3.4	34.5	3.3	0.14	−0.41	11.3	6.7	9.7	7.9
ACCGACGTCGGT-2		4.4	2.8	34.8	3.3	0.11	−0.41	11.4	6.3	9.8	7.9
ACCGACGTCGGT-3		4.9	3.4	34.7	3.3	0.01	−0.49	11.1	6.8	9.8	7.8
ACCGGTACCGGT	$P4_3$	4.4	2.5	34.5	3.3	−0.02	−0.60	11.5	6.6	9.9	7.8
B-DNA (20)		2.9	1.5	36.0	3.4	0.53	0.57	11.4	5.9	8.3	8.5

All parameters were calculated with FREEHELIX (15). The detailed definition and the nomenclature of the various helical parameters are given by Dickerson *et al*. (16). Inclination measures the rotation of the base pair long axis with respect to a plane perpendicular to the global helix axis. The local parameters were calculated with respect to the local axes of the base pair steps as follows. Roll measures the closing of adjacent base pairs toward the major groove (positive roll) or the minor groove (negative roll). Helix twist is the relative rotation of two successive base pairs about the helix axis. Rise is the relative translation of two successive base pairs parallel to the helix axis. Slide is the relative displacement of the base pairs along their long axes. D_x is the perpendicular distance from the C6-C8 vector of a base pair to the global helix axis. Groove dimensions were calculated as in ref. 17.

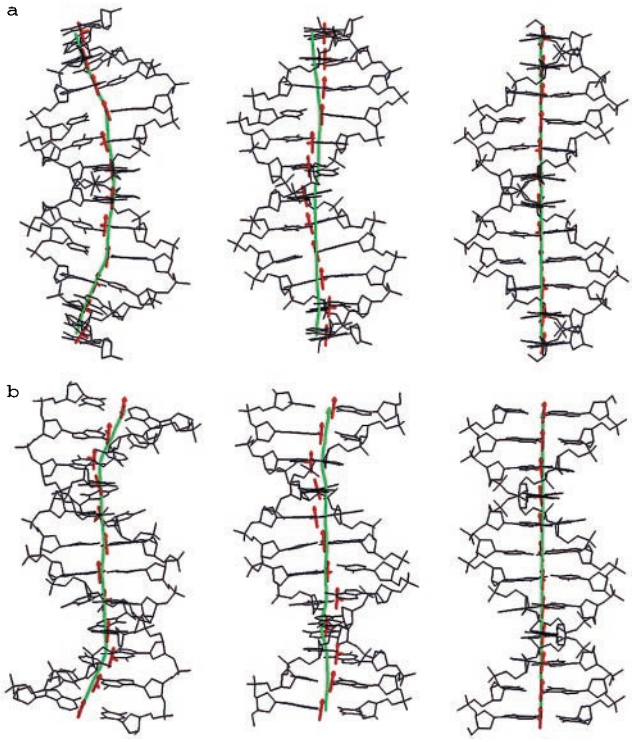

FIG. 4. Two views of (from left to right) the protein-bound E2-DNA (only the central 12 bp are shown), the E2-R3 dodecamer, and fiber B-DNA of the same sequence (20). (*a*) View perpendicular to the 2-fold axis of the double helix. (*b*) View along the 2-fold axis. The green line represents the best-fitted global axis and the red, short vectors represent the local axes of the base pairs. The global and local axes were calculated by CURVES 5.2 (25). In the free and bound E2 targets, the writhe of the local axes around the global axis is illustrated by the inclined red arrows precessing around the green line. The images were prepared with CURVES 5.2 (25) and RASMOL 2.6 (26).

Comparisons with Other Regulatory Systems. The central 6-bp sequence, GACGTC, analyzed in the context of the E2-DNA target is also an essential region of two other DNA regulatory elements studied in their complexed form by using x-ray crystallography. These are the *met* operator (..TA-*GACGTC*TA*GACGTC*TA..), which is the target of the *met* repressor (30), and the ATF/CREB site (AT*GACGTC*AT), which is recognized by the yeast transcription factor GCN4 and by other proteins related to the CREB-ATF family (ref. 31 and references therein). In contrast to the protein-bound E2 target, where the hexameric sequence is not contacted by the protein and is bent toward the minor groove, both the *met* and ATF/CREB hexameric sites make direct contacts to their cognate proteins accompanied by considerable bending toward the major groove. Significant bending toward the major groove also was observed in solution for the ATF/CREB site in both its free state and when complexed to the bZIP segment of GCN4, whereas no bend was observed when the same DNA site was complexed to a member of the CREB-ATF family (32).

Binding experiments performed on the DNA binding domains of the E2 proteins from bovine and human papillomaviruses (BPV-1 and HPV-16) with a series of DNA targets, including sequences that incorporate the hexameric spacer GACGTC, suggest that the flexibility of this particular sequence is important for its binding affinity to the BPV-1 protein (33).

All of the above data, together with the present findings on the E2-DNA structures, suggest that the GACGTC sequence is highly deformable and capable of being straight or bent in either direction. The significant variations in binding affinities

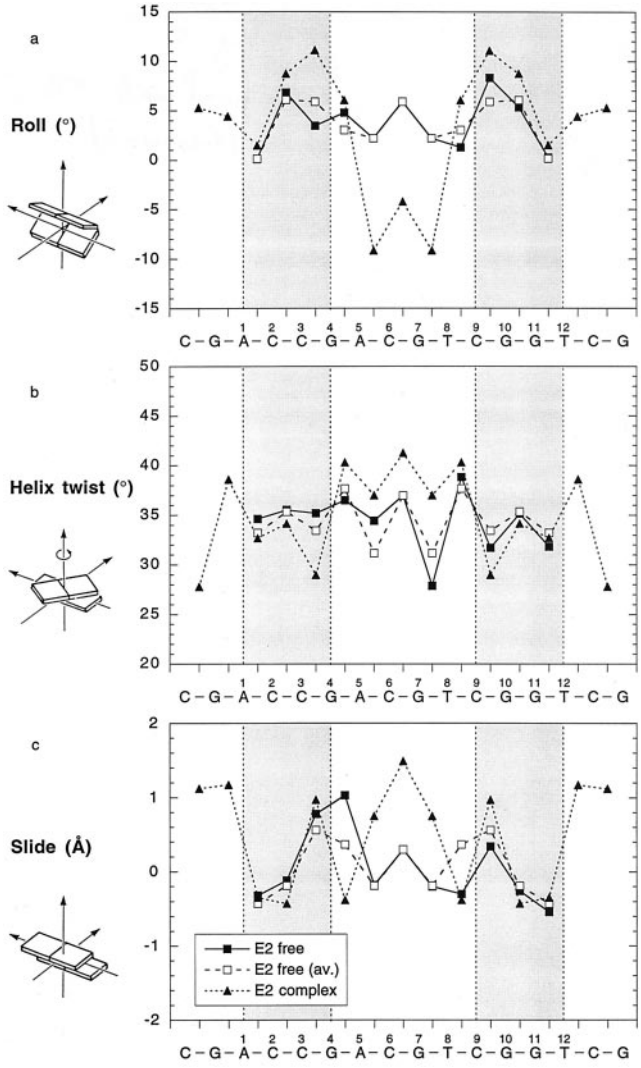

FIG. 5. Comparison of local parameters between the free E2 dodecamer (■) and the bound E2 16-mer (▲). End-to-end average values of the free dodecamer are also shown (□). Parameters shown are roll (*a*), helix twist (*b*), and slide (*c*). The parameters were derived by FREEHELIX (15) The various definitions are given in Table 2 and indicated by the corresponding drawings at the left. The consensus ACCG/CGGT regions are shaded.

of the wild-type BPV-1 E2 targets that differ in the central noncontacted base pairs (1) probably are related to the degree of deformability of this region.

The other E2-dodecamer of the present study, ACC*GGTACC*GGT, which contains a different hexameric spacer (italicized), has a global conformation similar to that of ACC*GACGTC*GGT (Table 2). However, the deformability of the two spacers may be different in a manner that could affect their bending and hence their binding properties.

Unlike the bovine papillomavirus E2 binding sites, the central noncontacted regions of the human papillomavirus E2 targets (e.g., oncogenic HPV-16) are A/T-rich, e.g., GAATTC, GTTTTC, or GAAATC (34). Such sequences have been shown to be intrinsically bent in solution in a direction that positions the minor groove of the A/T region at the concave side of the helix (35) just as required for the E2 complex. In such a case, the intrinsic structure of the DNA region, which is not contacted by the protein, appears crucial for the recognition process. This is supported by the binding studies of the same E2 targets to the bovine and the human papillomavirus E2 proteins (33).

Structural Code for DNA Recognition. The structural analysis of the present high-resolution structures provides a basis for some concepts about structural codes for DNA recognition by proteins. It is noteworthy that all of the A-C/G-T steps in both free and bound targets resist bending into the major groove. In the free dodecamer, the base pairs composing these steps are nearly parallel to each other, unlike most other sites where the base pairs close toward the major groove, as reflected by the positive roll angles (Fig. 5*a*). Similar characteristics of A-C/G-T sites were observed in other protein-bound DNA that display gentle roll-induced writhe, including the GCN4 target, which incorporates the same 6-bp sequence (31) and the DNA bound to MATα2 homeodomain (36). In the E2-complexed target, these sites display zero roll at the conserved tetranucleotides as they do for the free target. However, the two symmetric A-C/G-T sites of the central linker bend largely toward the minor groove (negative roll angles), as they tend to do in other protein–DNA complexes (28). Thus, such sites appear highly suitable for regions that resist major-groove bending or favor minor-groove bending.

Unlike the A-C/G-T step, C-G sites appear to be ideal for bending toward the major groove. All C-G sites display positive roll angles except for the single site of the central linker, which is negative in the complex and, together with the two adjoining A-C/G-T doublets, afford the required local bending into the minor groove. C-G sites, like the other two pyrimidine-purine sites (T-A and C-A/T-G), play here the important role of flexible "hinges" capable of adopting a large range of correlated conformations (28). In general, the large roll angles of such sites are associated with low helix twist and vice versa, as shown here and in previous studies (19, 28). In several of the natural bovine papillomavirus targets (1), C-G dinucleotides at the conserved regions are replaced by C-A/T-G, which are of similar conformational characteristics.

In the present systems, the A-C/G-T and C-G dinucleotides largely govern the structure of the free target and its deformation on protein binding. The relative motions of the bridging purine–purine sites (G-A/T-C and C-C/G-G) are dictated by their "dominant" partners, displaying intermediate roll angles and "absorbing" changes in helix twist necessitated by winding or unwinding of the flanking steps.

In addition to the specific geometrical features of the base pair doublets, their positioning or "phasing" along the double helix is critical for creating the global DNA conformation that is recognized by the protein. The large positive bends in the protein-bound DNA are localized at the two conserved CCG sites that interact directly with the protein. Because the two sites are only 7 bp apart—less than a complete helical repeat—they lead, together with the middle negative bend, to the left-handed superhelical curvature of the DNA (discussed above), thereby enabling the recognition elements of both the DNA and the protein to form an intimate sequence-specific interface. The presence of the "straight" A-C/G-T steps at each end of the target ensures the localization of the positive bends at the CCG sites. Any migration or diffusion of such motion along the DNA would impair the correct spacing between the centers of the local bends and hence the DNA trajectory required to form the complex with the protein.

Thus, the combination of base pair doublets or longer sequences with specific geometric propensities positioned at appropriate spacings along the double helix could constitute a structural code for DNA recognition by regulatory proteins. This structural code facilitates the formation of a complementary protein–DNA interface that is further specified by hydrogen bonds and nonpolar interactions between the protein amino acids and the DNA bases.

We thank our colleagues Linda Shimon and Joseph Gilboa for comments on the manuscript and R.E. Dickerson for the provision of FREEHELIX. This work was supported by the Helen and Milton Kimmelman Center for Macromolecular Assembly, the Israel Science Foundation administered by the Israel Academy of Sciences and Humanities, and the United States-Israel Binational Science Foundation (BSF). R.S.H. was partially supported by National Institutes of Health Grant CA66964. Z.S. holds the Helena Rubinstein Professorial Chair of Structural Biology.

1. Li, R., Knight, J., Bream, G., Stenlund, A. & Botchan, M. (1989) *Genes Dev.* **3,** 510–526.
2. Hegde, R. S., Grossman, S. R., Laimins, L. A. & Sigler, P. B. (1992) *Nature (London)* **359,** 505–512.
3. Hegde, R. S., Wang, A.-F., Kim, S.-S. & Schapira, M. (1998) *J. Mol. Biol.* **276,** 797–808.
4. Otwinowski, Z. & Minor, W. (1997) in *Macromolecular Crystallography, Part A*, eds. Carter, C. W., Jr., & Sweet, R. M. (Academic, New York), Vol. 276, pp. 307–326.
5. Kabsch, W. (1988) *J. Appl. Crystallogr.* **21,** 916–924.
6. Rabinovich, D. & Shakked, Z. (1984) *Acta Crystallogr. A* **40,** 195–200.
7. Rabinovich, D., Rozenberg, H. & Shakked, Z. (1998) *Acta Crystallogr. D* **54,** 1336–1342.
8. Brünger, A. T., ed. (1992) *x-plor Version 3.1. A System for X-Ray Crystallography and NMR.* (Yale Univ. Press, New Haven, CT).
9. Parkinson, G., Vojtechovsky, J., Clowney, L., Brünger, A. T. & Berman, H. M. (1996) *Acta Crystallogr. D* **52,** 57–64.
10. Jones, T. A., Zou, J.-Y., Cowan, S. W. & Kjeldgaard, M. (1991) *Acta Crystallogr. A* **47,** 110–119.
11. Sheldrick, G. M. & Schneider, T. R. (1997) in *Macromolecular Crystallography, Part B*, eds. Carter, C. W., Jr., & Sweet, R. M. (Academic, New York) Vol. 277, pp. 319–343.
12. Read, R. J. (1986) *Acta Crystallogr. A* **42,** 140–149.
13. Kraulis, P. J. (1991) *J. Appl. Crystallogr.* **24,** 946–950.
14. Evans, S. V. (1993) *J. Mol. Graphics* **11,** 134–138.
15. Dickerson, R. E. (1998) *Nucleic Acids Res.* **26,** 1906–1926.
16. Dickerson, R. E., Bansal, M., Calladine, C. R., Diekmann, S., Hunter, W. N., Kennard, O., von Kitzing, E., Lavery, R., Nelson, H. C. M., Olson, W. K., *et al*. (1989) *EMBO J.* **8,** 1–4.
17. Heinemann, U., Alings, C. & Bansal, M. (1992) *EMBO J.* **11,** 1931–1939.
18. Guzikevich-Guerstein, G. & Shakked, Z. (1996) *Nat. Struct. Biol.* **3,** 32–37.
19. Gorin, A. A., Zhurkin, V. & Olson, W. (1995) *J. Mol. Biol.* **247,** 34–48.
20. Chandrasekaran, R. & Arnott, S. (1996) *J. Biomol. Struct. Dyn.* **13,** 1015–1027.
21. Shakked, Z., Guzikevich-Guerstein, G., Frolow, F., Rabinovich, D., Joachimiak, A. & Sigler, P. B. (1994) *Nature (London)* **368,** 469–473.
22. Calladine, C., Drew, H. & McCall, M. (1988) *J. Mol. Biol.* **201,** 127–137.
23. Maroun, R. C. & Olson, W. K. (1988) *Biopolymers* **27,** 585–603.
24. Dickerson, R. E. (1992) in *DNA Structures, Part A*, eds. Lilley, D. M. J. & Dalhberg, J. E. (Academic, New York), Vol. 211, pp. 67–111.
25. Lavery, R. & Sklenar, H. (1988) *J. Biomol. Struct. Dyn.* **6,** 63–91.
26. Sayle, R. & Milner-White, E. J. (1995) *Trends Biochem. Sci.* **20,** 374–376.
27. Shakked, Z. & Rabinovich, D. (1986) *Prog. Biophys. Mol. Biol.* **47,** 159–195.
28. Olson, W. K., Gorin, A. A., Lu, X.-J., Hock, L. M. & Zhurkin, V. B. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 11163–11168.
29. Eisenstein, M., Frolow, F., Shakked, Z. & Rabinovich, D. (1990) *Nucleic Acids Res.* **18,** 3185–3194.
30. Somers, W. S. & Phillips, S. E. (1992) *Nature (London)* **359,** 387–393.
31. Keller, W., Konig, P. & Richmond, T. (1995) *J. Mol. Biol.* **254,** 657–667.
32. Paolella, D. N., Palmer, C. R. & Schepartz, A. (1994) *Science* **264,** 1130–1133.
33. Hines, C. S., Meghoo, C., Shetty, S., Biburger, M., Brenowitz, M. & Hegde, R. S. (1998) *J. Mol. Biol.* **276,** 809–818.
34. Bedrosian, C. & Bastia, D. (1990) *Virology* **174,** 557–575.
35. Crothers, D. M., Haran, T. E. & Nadeau, J. G. (1990) *J. Biol. Chem.* **265,** 7093–7096.
36. Wolberger, C., Vershon, A. K., Johnson, L. B. & Pabo, C. O. (1991) *Cell* **67,** 517–528.