Acta Cryst. (1998). D**54**, 1367–1376

Structure of the E2 DNA-Binding Domain from Human Papillomavirus Serotype 31 at 2.4 Å

Dirksen E. Bussiere,[a] Xiangpeng Kong,[b]† David A. Egan,[c] Karl Walter,[c] Thomas F. Holzman,[c] Frank Lindh,[c]‡ Terry Robins[d]§ and Vincent L. Giranda[b,e]*

[a]*Division of Scientific Information, Analysis, and Management, Pharmaceutical Products Division, Abbott Laboratories, 100 Abbott Park Road, Abbott Park, IL 60064, USA,* [b]*Laboratory of Protein Crystallography, Pharmaceutical Products Division, Abbott Laboratories, 100 Abbott Park Road, Abbott Park, IL 60064, USA,* [c]*Protein Biochemistry Group, Pharmaceutical Products Division, Abbott Laboratories, 100 Abbott Park Road, Abbott Park, Illinois 60064, USA,* [d]*Infectious Disease Research, Pharmaceutical Products Division, Abbott Laboratories, 100 Abbott Park Road, Abbott Park, IL 60064, USA, and* [e]*Cancer Research, Pharmaceutical Products Division, Abbott Laboratories, 100 Abbott Park Road, Abbott Park, IL 60064, USA. E-mail: girandav@abbott.com*

(*Received 24 December 1997; accepted 15 April 1998*)

Abstract

The papillomaviruses are a family of small double-stranded DNA viruses which exclusively infect epithelial cells and stimulate the proliferation of those cells. A key protein within the papillomavirus life-cycle is known as the E2 (Early 2) protein and is responsible for regulating viral transcription from all viral promoters as well as for replication of the papillomavirus genome in tandem with another protein known as E1. The E2 protein itself consists of three functional domains: an N-terminal trans-activation domain, a proline-rich linker, and a C-terminal DNA-binding domain. The first crystal structure of the human papillomavirus, serotype 31 (HPV-31), E2 DNA-binding domain has been determined at 2.4 Å resolution. The HPV DNA-binding domain monomer consists of two β–α–β repeats of approximately equal length and is arranged as to have an anti-parallel β-sheet flanked by the two α-helices. The monomers form the functional *in vivo* dimer by association of the β-sheets of each monomer so as to form an eight-stranded anti-parallel β-barrel at the center of the dimer, with the α-helices lining the outside of the barrel. The overall structure of HVP-31 E2 DNA-binding domain is similar to both the bovine papillomavirus E2-binding domain and the Epstein–Barr nuclear antigen-1 DNA-binding domain.

1. Introduction

The papillomaviruses are a family of small double-stranded DNA viruses which exclusively infect epithelial cells and stimulate the proliferation of normally quiescent cells (zur Hausen, 1991). Papillomaviruses are species-specific and are unable to replicate effectively in organisms other than their primary host (Galloway, 1994). They are also difficult to culture in undifferentiated cell strains, being intimately adapted to the highly differentiated state of epithelial cells (Galloway, 1994). Papillomaviral infections have been seen in species as varied as chaffinch, monkeys, mice, cows and humans (Murphy *et al.*, 1995). Infection by human papillomavirus (HPV) can lead to several possible outcomes, varying from cytologically inapparent infection, or benign warts, to a dysplasia which can progress after several years to an invasive cancer (Murphy *et al.*, 1995; Dillner *et al.*, 1995). The primary determinants of the outcome of infection are both the HPV serotype and the site of infection. Current treatment for refractory condylomata acuminata (genital warts) varies from surgical excision to injection of interferons (Stone, 1995). Injection of interferon is associated with complete clearance of the injected warts in 36% of patients due to the heightened immune surveillance triggered by interferon (Eron *et al.*, 1986). The lack of an effective, non-invasive treatment for HPV makes development of such therapeutics a priority.

The genome of HPV consists of 7.9 kb of DNA and codes for eight proteins. These eight proteins serve in varying roles: capsid structural proteins (designated L1 and L2), proteins responsible for maintaining a proliferative state in infected cells (E6 and E7 proteins), and control proteins in the viral life-cycle (E1, E2, and E8 proteins) (Galloway & McDougall, 1989). The E2 protein regulates viral transcription from all viral promoters as a *trans*-acting transcriptional activator (Barsoum *et al.*, 1992). It is also required in tandem with the E1 protein for the replication of the papillomavirus genome (Lusky *et al.*, 1993).

The E2 protein itself consists of three functional domains: an approximately 160 amino acid N-terminal *trans*-activation domain, a proline-rich linker of variable

† Present address: New York University, Department of Biochemistry, 550 First Avenue, New York, NY 10016, USA.
‡ Present address: Abbott Diagnostics Division, 100 Abbott Park Road, Abbott Park, IL 60064, USA.
§ Present address: Specialty Laboratories, Inc., 2211 Michigan Avenue, Santa Monica, CA 90404, USA.

composition, and a C-terminal domain of approximately 85 amino acids (Giri & Yaniv, 1988; Gauthier *et al.*, 1991). This C-terminal domain is responsible for mediating both DNA-binding and dimerization of the protein (Bedrosian & Bastia, 1990). This arrangement of domains remains constant throughout all known E2 proteins of papillomavirus (Hegde *et al.*, 1992); the consistent arrangement of domains within the E2 proteins, as well as their importance in the HPV life-cycle, makes the E2 protein an attractive target in the development of therapeutics against papillomavirus.

We describe here the high-resolution crystal structure of the DNA-binding domain, residues 291–371, of the E2 protein from human papillomavirus type-31 (hereafter HPV-31) at 2.4 Å resolution. The crystal structure of HPV-31 E2 DNA-binding domain is compared with the previously determined crystal structure of the bovine papillomavirus type-1 (BPV-1 or BPV) E2 protein complexed with DNA and to the crystal structure of the Epstein–Barr nuclear antigen-1 (EBNA-1 or EBNA) (Bochkarev *et al.*, 1995), which is a structural homologue of the E2 DNA-binding domain proteins. It is also compared with the NMR structure of HPV-31 E2 DNA-binding domain (Liang *et al.*, 1996). A model for the binding of the HPV-31 E2 DNA-binding domain to DNA will also be presented.

2. Methods

2.1. *Purification of HPV-31 E2 DNA-binding domain*

A gene fragment encoding the E2 DNA-binding domain of HPV-31 which consisted of the C-terminal 82 amino acids was cloned into the pET-3b vector for overexpression in *Escherichia coli*. Cells of *E. coli* strain BL21(DE3)pLysS transformed with the E2 expression plasmid were grown at 310 K on LB media. 16 g of *E. coli* cell paste was subjected to a freeze–thaw cycle followed by the addition of 100 ml of buffer *A* [50 m*M* Tris–HCl, pH 8.3, 100 m*M* NaCl, 1 m*M* EDTA (ethylenediaminetetraacetic acid), 5 m*M* DTT (dithiothreitol), 1 m*M* PMSF (phenylmethylsulfonyl fluoride)]. All subsequent steps were performed at 277 K unless otherwise stated. The cell paste was then sonicated for 6 min on ice to promote lysis of the cells. Lysis was confirmed by microscopy. This was followed by the addition of 10 µl of benzonase to depolymerize the DNA. After reacting for 30 min, the preparation was centrifuged at 35 000*g* for 60 min. The supernatant was applied to an S-Sepharose column and the column was washed with buffer *A* until the A_{280} dropped to below 0.2 a.u. (absorbance units). A linear gradient of buffer *A* to buffer *A* plus 500 m*M* NaCl was run with 5 column volumes at a flow rate of 2 ml min^{-1} and fractions were collected every 2 min. The absorbance profile was recorded and SDS–PAGE was run to confirm the identity of appropriate fractions. The fractions containing E2 DNA-binding domain were pooled and dialyzed against buffer *A*. The pooled fractions were then applied to a Mono-S column; a linear gradient of buffer *A* to buffer *A* plus 500 m*M* NaCl was run at a flow rate of 2 ml min^{-1} and 3 min fractions were collected. The E2 DNA-binding domain fractions were pooled and dialyzed against 50 m*M* phosphate buffer, pH 7.0, 100 m*M* NaCl, 10 m*M* DTT (buffer *B*) for the purpose of storage. The protein solution was then concentrated to a final concentration of approximately 20 mg ml^{-1} with Amicon filters.

2.1.1. *Crystallization*. The E2 DNA-binding domain was crystallized using the hanging-drop vapor-diffusion method (Ollis & White, 1990; Ducruix & Giegé, 1992). The hanging drops were formed by mixing the protein solution with the well solution in a 1:1 ratio to yield a final volume of 10–20 µl. The hanging drops were then sealed over the 1 ml well solution in a Linbro 24-well tissue-culture plate. The initial protein solution contained 15–25 mg ml^{-1} of E2 DNA-binding domain in buffer *B*. Protein crystals were obtained with a well solution of 100 m*M* citrate, pH 4.0–5.5, and 45–65% saturated ammonium sulfate. The crystals grew to 1.6 mm in the longest dimension and crystallized in space group $P6_122$ with unit-cell dimensions of $a = b = 45.89$ and $c = 195.64$ Å with a monomer in the asymmetric unit (Matthews, 1968). Crystals were transferred and stored in a 100 m*M* citrate buffer, pH 4.0, with 60% saturated ammonium sulfate. Crystals were cyroprotected immediately prior to data collection by serially transferring them to storage solutions containing 5, 10, 15 and 20% of glycerol for 10 min each.

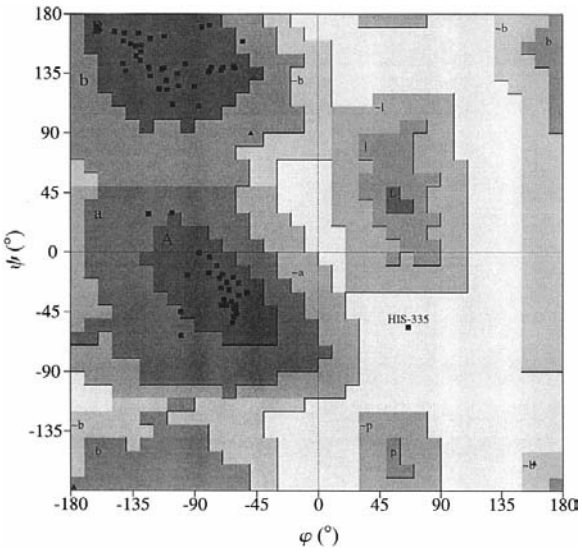

Fig. 1. Ramachandran plot of HPV-31 E2 DNA-binding domain. The plot was produced with the program *PROCHECK* (Laskowski *et al.*, 1993). Glycine residues are shown as triangles. Decreasing probability levels of conformations are indicated by the decreasing grey scale: core (dark gray), allowed (gray), generously allowed (pale gray) and disallowed (white).

2.1.2. *Data collection*. X-ray intensities on the cyro-protected crystal were collected at 90 K on an RAXIS II imaging-plate system, using monochromated Cu $K\alpha$ radiation from a Rigaku RU-200 rotating anode. Data were processed using the *DENZO* software package (Otwinowski, 1993). Data collection statistics are shown in Table 1. The crystals were stable in the X-ray beam for more than 24 h.

2.1.3. *Molecular replacement*. The E2 DNA-binding domain structure was solved by the molecular replacement method using the program *AMoRe* (Navaza, 1994). A search model was constructed using the coordinates of the bovine papillomavirus E2 DNA-binding domain. The sequences of the two proteins were aligned and non-identical residues between the two proteins were mutated to alanines. Regions which were suspected to be significantly different between the two proteins due to sequence or structural considerations were deleted from the search model. These excluded regions were residues 325–340 and 368–371. The cross-rotation function (Rossmann, 1972) was calculated using data from 15 to 4 Å resolution and a 3–19 Å Patterson search radius. The model cell was calculated as $P1$ with all cell dimensions set to 50 Å to reduce spurious peaks. The Shannon factor was set to 3.0, the box scale to 4.0, and Bessel limit was set to 6.0. The map was calculated with an angular grid of 2.5°. The correct peak was the seventh largest peak in the map, at a level of 3.2σ. Following the determination of the proper rotation matrix by *AMoRe*, the translation matrix was determined. The translation-function search, calculated with a resolution range of 15–4 Å, yielded a correct result with an R factor of 0.54 and a correlation coefficient of 0.35. The translation results were examined and the molecular twofold was found to correspond to a crystallographic twofold with the inter-subunit β-strand interactions held intact.

2.1.4. *Refinement*. The search model was then subjected to rigid-body refinement using *X-PLOR* version 3.1 (Brünger, 1992), which led to reasonable improvement in the R factor (R = 0.51). Inspection of the $2F_{obs} - F_{calc}$ and $F_{obs} - F_{calc}$ maps at this point allowed for reconstruction of deleted portions of the search model. The structure then underwent several rounds of a standard refinement protocol using diffraction data from 10 to 2.4 Å with a 1σ cutoff. Each round consisted of Powell minimization, simulated annealing to 4000 K, and grouped temperature-factor refinement. A new Ramachandran force-field, which restrains the structure to adhere to known rotamers, was utilized in later rounds of the refinement.† Each round of refinement was followed by model rebuilding using *O* (Jones *et al.*, 1991). At the point at which the R factor fell below 30%, 74 well ordered waters and two structural sulfates were also located in the electron density and built into the model. This model underwent several more rounds of the refinement protocol, ending with individual temperature-factor refinement of the final model. The refinement converged at an R = 21.1% and R_{free} = 29.7%. The final structure was well defined, save for residues 329–336, where significant disorder was seen; this region is still included in the final model. There is one outlier in the Ramachandran plot (His335 in the disordered loop) as calculated by *PROCHECK* (Laskowski *et al.*,1993), with 96% of all residues in the protein lying in most favored regions (Fig. 1); the overall G factor of the refined model was 0.23 (Laskowski *et al.*, 1993). An Eisenberg profile analysis of the structure gives a combined score of 73.62, an acceptable score for a protein of this size (Luthy *et al.*, 1992). The final $2F_{obs} - F_{calc}$ electron-density map, shown in Fig. 2, was contiguous in all areas save for the disordered loop.

Table 1. *Diffraction data and final model statistics*

Data collection	
Total No. of observations	60427
No. of unique reflections	5649
% of data > 1σ (overall/last shell)	90.0/88.4
% of data > 13σ (overall/last shell)	83.2/79.1
R_{merge} (%) (overall/last shell)	6.5/18.5
Refinement	
No. of protein atoms	742
No. of reflections (10–2.4 Å)	4293
R_{free} value (%) (10–2.4 Å)	29.7
R value (%) (10–2.4 Å)	21.1
Average B factors (Å^2)	
Main chain	26.4
Side chains	25.5
Water	28.8
R.m.s. deviations	
Bonds (Å)	0.014
Angles (Å)	2.7
Improper angles (°)	2.2
Dihedral angles (°)	23.2
Luthy–Eisenberg profile score (dimer)	73.6
G factor (monomer)	0.23

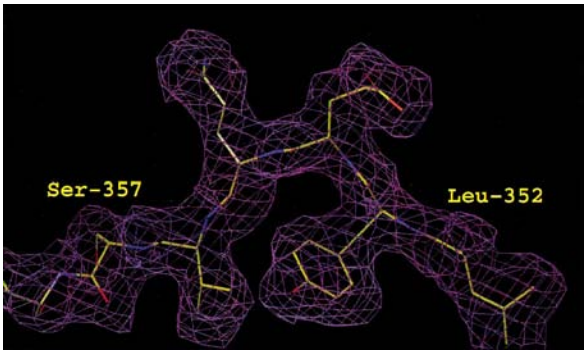

Fig. 2. $2F_{obs} - F_{calc}$ electron-density map of the HPV-31 E2 DNA-binding domain contoured at 1.5σ. The region connecting the first α-helix, $\alpha1$, to the second β-sheet, $\beta2$, is shown. The electron density has been displayed using the map-cover option of the program *O* (Jones *et al.*, 1991).

† Bussiere, Kuszewski, Gronenborn & Clore, unpublished results.

The final structure was confirmed *via* standard 'omit' maps, each representing 10% of the structure. As a final check against model bias, as well as to confirm the presence of the sulfate ions, the structural S-atom positions were determined by calculation of sulfur anomalous diffraction maps calculated using native data with phases provided by the final atomic model with all S atoms removed. Peaks (all greater than 3σ) were seen for both structural sulfates as well as all cysteine and methionine S atoms. This confirmed that the structural sulfates were indeed sulfates and not structural waters, as oxygen has a low anomalous signal at the wavelength used.

The structure of the HPV-31 E2 DNA-binding domain monomer with both structural waters and sulfates has been deposited with the Protein Data Bank (Bernstein *et al.*, 1977).†

3. Results and discussion

3.1. *Monomer structure*

The monomer consists of two β–α–β repeats of approximately equal length (Fig. 3). These repeats are arranged to form a four-stranded anti-parallel β-sheet at one side of the monomer and to decorate the other side with the two α-helices. The connectivity of secondary-structure elements within the monomer is shown in Fig. 3. The first helix (α1) corresponds to the recognition helix whose residues form the protein–DNA interactions which allow the protein to specifically recognize its cognate DNA sequence (Hegde *et al.*, 1992). The 11-residue loop between β2 and β3, which in the BPV structure interacts with the phosphodiester backbone of DNA, is significantly disordered within the crystal structure which has no DNA. This disorder is also seen in the protein structure as determined by NMR (Liang *et al.*, 1996). Also of interest are the approximately identical lengths of each of the secondary-structural elements. Both α-helices and all four β-strands are of the same approximate length. This arrangement of secondary-structural elements imparts a pseudo-twofold symmetry on each monomer.

Each monomer contains a hydrophobic core comprised of residues donated from the four-stranded β-sheet as well as the two amphipathic helices. Several of these core residues are well conserved across species (Fig. 4). An example of this is Leu305, located at the N-terminal portion of α1, which is invariant over all known papillomaviruses. A more detailed examination of observed sequence homologies will be given in a following section.

† Atomic coordinates and structure factors have been deposited with the Protein Data Bank at Brookhaven National Laboratory (Reference: 1A7G).

3.2. *Dimer structure*

The functional *in vivo* state of the E2 protein is that of a homodimer, although there is evidence that the full-length form of the protein also forms heterodimers with a naturally occurring truncated form of E2 which lacks the *trans*-activation domain (Chin *et al.*, 1988). These heterodimers are incapable of *trans*-activation and thereby repress the transcriptional activation activity of the full-length E2 protein. It has also been shown that the DNA-binding domain is the prime mediator of this dimerization (Bedrosian & Bastia, 1990). Within the homodimer, the four-stranded anti-parallel β-sheet from each monomer associates with its twofold symmetry-mate to form a dimer with an eight-stranded anti-parallel β-barrel at the center of the molecule and the α-helices lining the outside of the β-barrel (Fig. 3). The twofold-related recognition helices from each monomer are separated by approximately 24 Å, thereby allowing

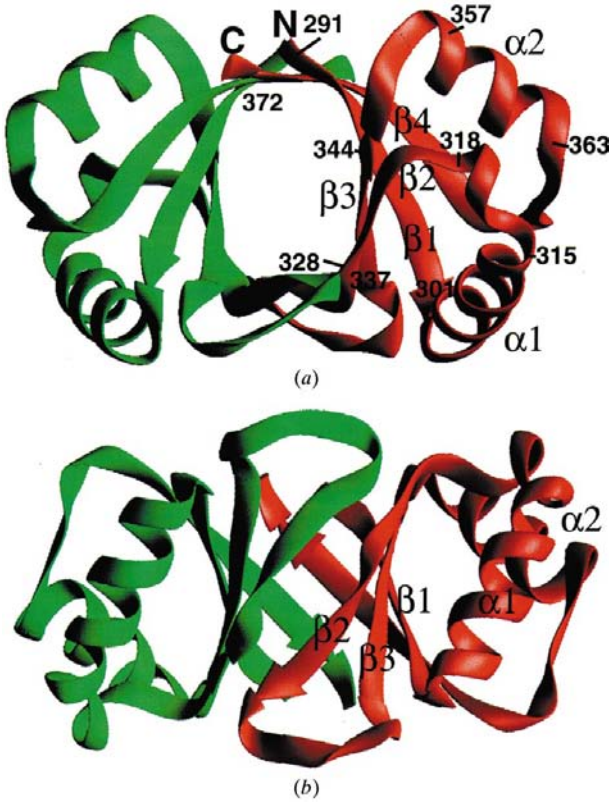

Fig. 3. Overall fold of HPV-31 E2 DNA-binding domain monomer and dimer. (*a*) Ribbon diagram of the tertiary structure of the HPV-31 E2 DNA-binding domain dimer looking down on the core β-barrel, with the α1 recognition helix positioned at the bottom of the diagram. The N- and C-termini are labeled, as are amino-acid residues at the beginning and end of secondary-structure elements. (*b*) View orthogonal to (*a*), showing a side view of the core β-barrel; this orientation clearly shows the DNA-binding surface. One monomer is shown in green, the other in red. Secondary-structural elements are labeled. The drawing was generated using the program *RIBBONS* (Carson, 1987).

each recognition helix to interact with the major groove of the palindromic cognate DNA site (Steitz, 1990).

The formation of the eight-stranded β-barrel on dimerization results in the creation of an inter-monomeric hydrophobic core distinct from the intra-molecular hydrophobic core of each monomer. The extended intermonomeric hydrophobic core mediates dimerization: formation of the homodimer occludes 1304 $Å^2$ of surface area, corresponding to approximately 138 kJ mol^{-1} (33 kcal mol^{-1}) of dimerization energy due to the hydrophobic effect (Richards, 1977). Other forces also contribute to the dimerization energy, however. An aromatic network, formed by two highly conserved tryptophans (Trp326 and Trp328) and their symmetry mates in the adjacent monomer, also serve to significantly stabilize the dimer. Aromatic networks are quite common in small DNA-binding proteins and are postulated to be stabilizing forces in protein folding and dimerization (Burley & Petsko, 1985). Such networks are characterized by aromatic groups stacking roughly perpendicular to one another, forming a herring-bone pattern (Burley & Petsko, 1985). This is precisely what is seen in HPV-31 E2 DNA-binding domain, where the Trp326 indole rings, one from each monomer, stack parallel to one another and the indole rings of the second pair of tryptophans (Trp328) stack perpendicular to the first pair. Unlike other DNA-binding proteins, where inter-monomer salt links are seen stabilizing monomer–monomer contacts, HPV-31 E2 DNA-binding domain has few such electrostatic interactions. Each monomer itself is stabilized by numerous inter-strand hydrogen bonds, but only a few such bonds of this type are seen between monomers. These latter hydrogen bonds, only eight in number, seal together the edges of each half-barrel or monomer. A few additional inter-monomer hydrogen bonds are also formed from side chains on β2 and β4.

The intermonomeric hydrophobic core at the interior of the β-barrel is not completely contiguous. A small 4 Å cavity is present at the approximate center of the dimer, proximal to the aromatic network (Fig. 5). This cavity can be accessed from both sides of the homodimer, and is occupied by a sulfate ion in the crystal structure. This sulfate ion has its S atom placed at a crystallographic special position and coordinates to the His297 and Thr341 side chains from each monomer that border the cavity. The size and polar character of the cavity is suitable for occupancy by a sulfate ion, a phosphate ion or a structural water. Examination of the sequences of other E2 DNA-binding domains from known serotypes of HPV suggests that this cavity does not occur in all serotypes of E2 DNA-binding domains. In most of the E2 DNA-binding domains, residue 297 (proximal to the N-terminus) is either a histidine, glutamine or a hydrophobic group such as an isoleucine or a valine. The exception to this is cottontail rabbit papillomavirus (CRP) E2 DNA-binding domain, which has a cysteine residue at position 297. In E2 DNA-binding domains possessing a histidine or glutamine at position 297, position 341 will always contain a threonine residue thereby forming the necessary coordination groups within the cavity. The other set of E2 DNA-binding

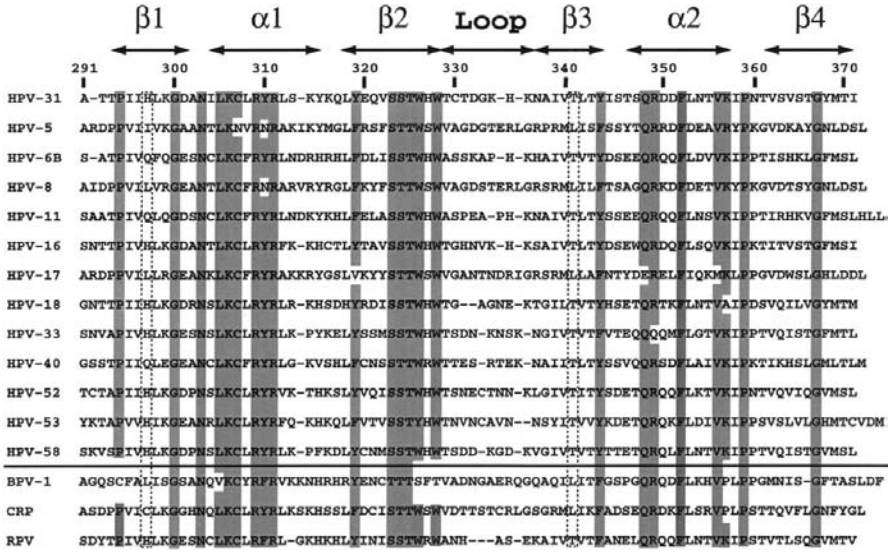

Fig. 4. Sequence alignment of the DNA-binding domain of the E2 protein from various human papillomavirus serotypes, along with sequences from serotypes from other species such as bovine papillomavirus (BPV-1), cottontail rabbit papillomavirus (CRP), and rhesus papillomavirus (RPV). Homologous residues are shaded, while residues which would be indicative of the counter-ion pore are shown in a dashed outline; serotypes which contain the fingerprint of the pore will have a histidine at position 297 and threonine at position 341. Other serotypes, which most likely have a contiguous hydrophobic core, will have a hydrophobic residue at the positions equivalent to 297 and 341.

domains have hydrophobic residues at position 341. This hydrophobic arrangement is present in the BPV-1 structure and results in a contiguous hydrophobic core (Fig. 5). This coupled variance of residues at positions 297 and 341 appears to be the hallmark of the whether the protein contains the cavity. An examination of HPV serotypes predicts that serotypes 16, 18, 30, 33, 34, 35, 39, 42, 45, 51, 52, 56, 57 and 58 possess this cavity and that serotypes 5, 8, 17, 40 and 41 do not; rhesus papillomavirus would also be predicted to have the cavity, while both bovine papillomavirus – as confirmed by the crystal structure (Hegde *et al.*, 1992) – and cottontail rabbit papillomavirus do not.

3.3. *Comparison with the NMR structure*

Comparison of the crystal structure of HPV-31 E2 DNA-binding domain with its counterpart NMR structure (PDB accession 1DHM) proved interesting. The two dimer structures superimpose with an r.m.s. (root-mean-square) deviation of 1.9 Å. However, the individual monomers from each structure can be superimposed with a lower r.m.s. deviation of 1.5 Å, indicating that there is a difference in the intermonomer transform between the two dimers. Within a monomer, the greatest region of deviation occurs in the region of the disordered loop between β2 and β3, which is also poorly defined in NMR structure (Liang *et al.*, 1996). The core regions, such as the internal portions of the β-barrel, overlap quite nicely in the two structures. For example, the Trp326 and Trp328 indole rings from both structures superimpose almost perfectly. Other side chains in the NMR structure show slight deviations from the crystal structure, and these deviations in side-chain orientation increase as one moves out from the center of the molecule towards the solvent. A typical comparison of surface residues is as follows: in Tyr30, the Cα atoms of the two structures are offset by approximately 1.3 Å,

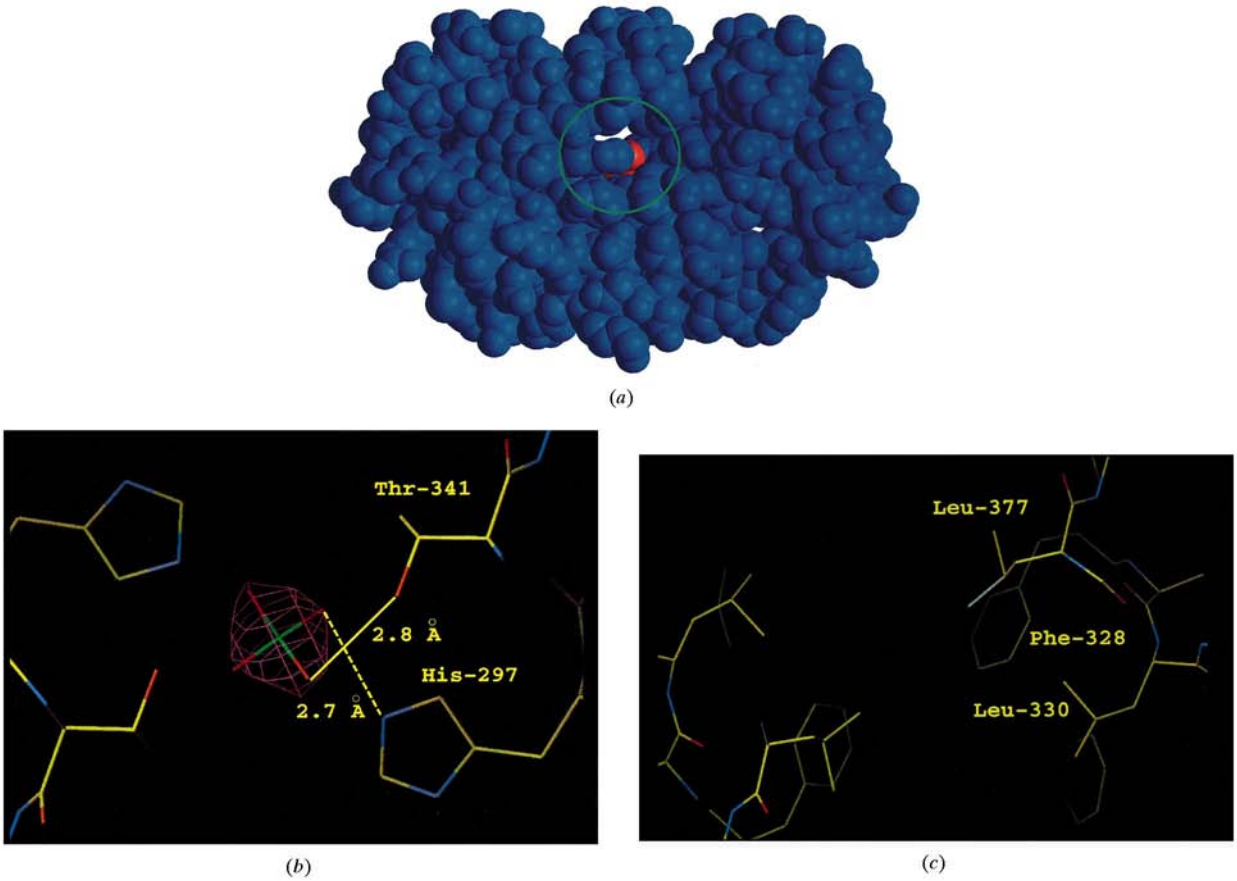

Fig. 5. Views of the hydrophilic cavity. (*a*) Space-filling model of the HPV-31 E2 DNA-binding domain showing the cavity. van der Waals radii are set to standard values for each atom type; protein atoms are shown in blue and the sulfate molecule is shown in red. The model was generated using the program *RIBBONS* (Carson, 1987). (*b*) Internal chemical environment of the cavity. The key histidine and threonine residues and their symmetry mates as well as the sulfate ion are shown. Electron density from the sulfur anomalous map corresponding to the position of the sulfate ion is shown in magenta; the sulfur peak is contoured at 6σ. Putative hydrogen-bonding distances between donor and acceptor atoms are drawn in solid and dashed yellow lines, with the solid yellow line lying above the sulfur, and the dashed yellow line lying below the sulfur. (*c*) Internal chemical environment of BPV-1 E2 DNA-binding domain, which lacks the cavity, shown in the same orientation as (*b*). In this case, the histidine and threonine residues are replaced by leucines. The model was generated using the program *O* (Jones *et al.*, 1991).

and while the side chains superimpose along the axis of the ring, the phenyl ring in the NMR structure is rotated approximately 80° in relation to the ring in the crystal structure. According to a recent report by Brünger (1997), a crystal structure of this resolution will have an average coordinate error of 0.3 Å, while a comparable NMR structure will have, at best, an average coordinate error of 0.8–0.9 $Å^3$. Therefore, these differences may be due to the higher coordinate error of an NMR structure or to true motion within the protein in solution. Also, slight perturbations in protein structure may be enforced by the crystal matrix, especially in side chains involved in crystal contacts. The chemical composition of the mother liquor, in this case high salt, will also affect the protein within the crystal matrix. In both cases, it is expected that residues close to or on the protein surface would be readily influenced by both of these factors. This is probably what is seen in the case of the HPV-31 E2 DNA-binding domain, where the most variation is seen in the side chains of surface residues. And, as the tertiary folds in the two structure determinations are essentially identical, it is expected that the crystal structure represents the more precise view of the molecule.

3.4. *Model for DNA binding*

As with several other DNA-binding proteins, E2 DNA-binding domains bind specifically to two turns of double-stranded B-form DNA and recognize a palindromic sequence of $ACCGN_4CGGT$ *via* the major groove, which is found in multiple copies within papillomavirus genome (Li *et al.*, 1989; Alexander & Phelps, 1996). A model of the HPV-31 E2 DNA-binding domain bound to cognate DNA was constructed by superimposing the Cα backbones of HPV-31 E2 DNA-binding domain and the BPV-1 E2 DNA-binding domain:DNA cocrystal structure (Hegde *et al.*, 1992). The model thus generated (Fig. 6) suggests that HPV-31 E2 DNA-binding domain binds DNA in a manner equivalent to its bovine cousin. The model shows no steric clashes between the protein and the DNA, and both α1 recognition helices and the loop between β2 and β3 are suitably positioned to recognize the specific bases in the cognate DNA sequence. The recognition helices are positioned within the major groove of the DNA, while the loops ride the phosphate backbone on either side of the dyad axis. The loop side chains are able to interact with either the major or minor grooves or the phosphate backbone. The overall curvature of the contact face of HPV-31 E2 DNA-binding domain matches that of the bovine DNA site, with both approximating a circle with a 45 Å radius (Hegde *et al.*, 1992). Examination of the model suggests that several residues within the protein are responsible for DNA recognition. These interactions can be subdivided into sequence-specific and non-specific types. Asp301, a member of β1, and Asn303, Lys306, Cys307, Tyr310 and Arg311, all members of the α1 recognition helix, are suitably positioned to interact specifically with the bases of the DNA. Residues such as Arg309, Ser324 and Thr325 are more suitably positioned to interact with the phosphate backbone. Lys306 and Arg309 are found to interact with a structural sulfate ion, which in this case is serving as a chemical surrogate for the phosphodiester backbone of DNA. Examination of the electrostatic profile of HPV-31 E2 DNA-binding domain (Fig. 7) confirms this arrangement, showing the major portion of positive electrostatic potential to be clustered around the α1 recognition helix and the flexible loop.

Fig. 6. Model of the HPV-31 E2 DNA-binding domain bound to cognate DNA. The recognition helix, α1, fits precisely into the major groove, while the loop interacts with phosphate backbone. Very few interactions with the minor groove are seen in the model, consistent with published data (Hegde *et al.*, 1992). The drawing was generated using the program *RIBBONS* (Carson, 1987).

3.5. *Conservation of E2 DNA-binding domain primary sequence and secondary structure*

Perhaps the most striking feature of the papillomavirus E2 DNA-binding domain proteins is their primary and secondary structural invariance. As shown in Fig. 4, there is considerable sequence similarity between E2 DNA-binding domains from various papillomavirus serotypes. For example, E2 DNA-binding domain from HPV-31 and from BPV-1 share an amino-acid sequence identity of approximately 31%. This sequence homology/identity extends over the other HPV serotypes as well as in rhesus and rabbit papillomaviruses. This identity is clustered around areas of key secondary structures. These areas are involved in either dimerization or DNA-binding. Each area of sequence homology is identified in Fig. 4. Examples of homologies important in maintaining the structure include: (i) Trp326 and Trp328, which are involved in the formation of the stabilizing aromatic network; (ii) Gly300, which is part of the tight turn between the first β-strand β1 and the α1

recognition helix; (iii) Pro294, which caps the first β-sheet, β1. These invariant residues allow structurally critical interactions. The sequence homology in the areas responsible for DNA recognition show some variation, although many conserved residues are seen. Foremost of these is Cys307, which is found in the BPV-1 E2 DNA-binding domain and the E2 DNA-binding domains from most human papilloma serotypes. This cysteine is involved in key interactions with the DNA binding site in the BPV-1 E2 DNA-binding domain:DNA complex crystal structure, donating a hydrogen bond to a guanine on one strand of the DNA and accepting a hydrogen bond from an adenine on the opposite strand (Hegde *et al.*, 1992). This cysteine is involved in protein:DNA interactions in human papillomavirus as well. Oxidation of this residue results in ablation of DNA binding in E2 protein from both BPV-1 and HPV serotypes (Prakash *et al.*, 1992). This is unusual as cysteines are rarely observed in coordinating to DNA (Hegde *et al.*, 1992). Another example of conserved interactions is seen in the triad of residues 309-RYR-311. Both the arginines and the tyrosine are positioned to interact with the bases and the phosphate backbone of the recognition sequence in the BPV-1 cocrystal structure. These residues are suitably positioned in HPV-31 to form the same sort of contact. The 323-SST-325 sequence is also invariant amongst all papillomavirus E2 DNA-binding domain sequences. As previously mentioned, these residues are all positioned to interact with the phosphate backbone of the DNA (Hegde *et al.*, 1992).

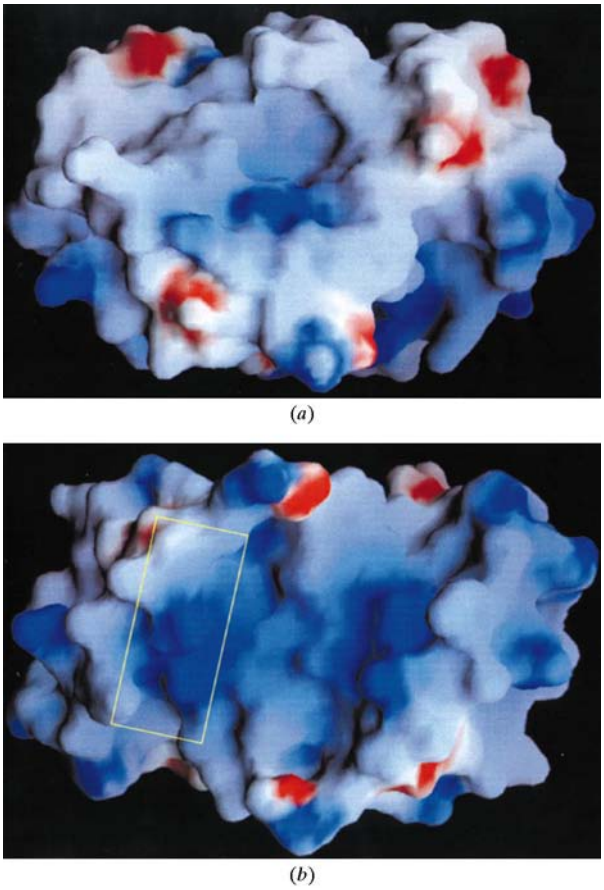

Fig. 7. Electrostatic potential of the HPV-31 E2 DNA-binding domain dimer. This diagram illustrates the solvent-accessible surface of the protein, color-coded for the surface electrostatic potential, with red representing negative electrostatic potential and blue representing positive electrostatic potential, each contoured at 7 kbT. (*a*) View equivalent to Fig. 2(*a*). The DNA-binding surface is positioned at the bottom of the protein. (*b*) View equivalent to Fig. 2(*b*). The strong positive electrostatic potential of the DNA-binding surface can clearly be seen, with the strongest points of positive potential positioned approximately in the center of the α1 recognition helix. Illustrations were produced using the program *GRASP* (Nicholls *et al.*, 1991).

The flexible loop between β2 and β3 shows extensive sequence variability. The residues within this area show a strong propensity to be positively charged, with histidines, arginines and lysines all strongly represented. Given that this area has been shown to be involved in non-specific binding of the DNA phosphate backbone (Hegde *et al.*, 1992), one would expect that there would be little selective pressure to maintain a particular sequence so long as the positive electrostatic complementarity for binding to DNA was maintained. Analysis of the available sequences also predicts that the lengths of these loops will vary considerably between species, in some species being quite long (approximately 11 residues in HPV-31) and in others being shorter, as in the case of rhesus papillomavirus where it is predicted to be five residues in length.

Extending from the sequence homology, the two known experimental structures of E2 DNA-binding domains are extremely similar. As shown in Fig. 8, the Cα skeleton of BPV-1 (Hegde *et al.*, 1992) and HPV-31 E2 DNA-binding domain can be superimposed with an r.m.s. deviation of 1.12 Å over all Cα atoms. Given the high-level of structural homology between these two E2 DNA-binding domains and the level of sequence homology between all E2 proteins, it is reasonable to assume that other E2 DNA-binding domains, regardless of serotype, will maintain a homologous structure. This structural homology extends to the DNA-binding domain of EBNA-1 (Bochkarev *et al.*, 1995). EBNA-1 is a protein encoded by Epstein–Barr virus, which, like the E2 DNA-binding domain, is crucial for viral DNA replication and governs DNA replication from the viral origin of replication (Yates *et al.*, 1985). Epstein–Barr virus itself is a ubiquitous human γ-herpesvirus with no known evolutionary relationship to human papillomavirus (Grossman & Laimins, 1996). The Cα atoms of the HPV-31 E2 DNA-binding domain and EBNA-1 'core' region (Bochkarev *et al.*, 1995) superimpose with an r.m.s.d. of 2.0 Å, over all homologous Cα atoms. It has been purported that this similarity in structure is due to their similar functions within their respective viruses, particularly their roles in activating DNA replication

and in DNA bending (Bochkarev *et al.*, 1996). Interestingly, the hydrophobic cores of these two proteins are also similar, with both sharing an aromatic network. The side chain of Trp326 of HPV-31 DNA-binding domain is nearly superimposable upon Tyr561 of the EBNA-1 DNA-binding domain, despite the fact that the Cα atoms of these residues are in distinctly different locations. Also, in both the HPV-31 E2 and EBNA-1 DNA-binding domains, there is a distinctive kink in the second β-strand which causes the β-sheet to twist and change its direction. This kink results from an unusual hydrogen-bonding arrangement between the second and third β-strands. This kink in the β-sheet orients the loop between β-strands 2 and 3 into a position which is suitable for interaction with the DNA phosphate backbone. However, unlike the E2 DNA-binding domain where the α1 recognition helix is the sole mediator of sequence-specific protein–DNA interactions, in the EBNA-1 DNA-binding domain sequence-specific interactions are spread over several structural elements (Ambinder *et al.*, 1990; Chen *et al.*, 1994). In essence, the E2 DNA-binding domain represents the core protein with localized DNA recognition and functional characteristics, while the EBNA-1 DNA-binding domain has added functionality as well as additional structural elements. These elements separate EBNA-1 into a core domain, which binds the inner portion of the EBNA-1 binding site, and a flanking domain, which is unique to EBNA-1 and binds the outer portion of the longer EBNA-1 cognate DNA site (Bochkarev *et al.*, 1995).

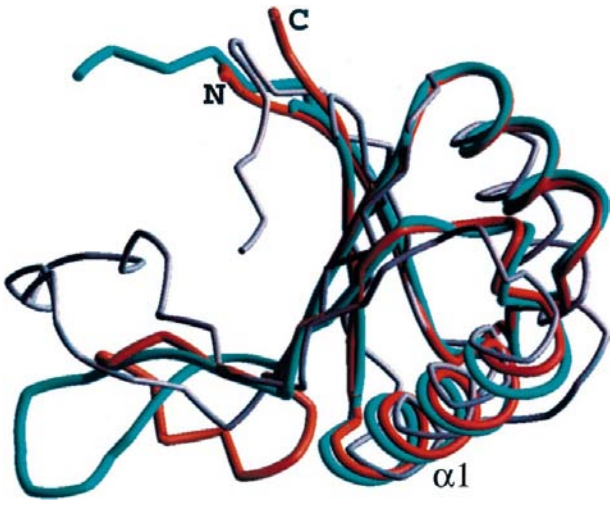

Fig. 8. Superposition of the Cα traces of HPV-31 (red) and BPV-1 (blue) E2 DNA-binding domain monomers with the EBNA-1 (magenta) DNA-binding domain monomer 'core'. HPV-31 E2 DNA-binding domain superimposes with BPV-1 E2 DNA-binding domain with an r.m.s. deviation of 1.12 Å, and superimposes with EBNA-1 E2 DNA-binding domain with an r.m.s. deviation of 2.0 Å. Superpositions were performed using the program *O* (Jones *et al.*, 1991), and the traces were generated using the program *RIBBONS* (Carson, 1987).

3.6. *Implications of the protein structure*

Two main features stand out among the E2 DNA-binding domains: their pivotal importance within the viral lifecycle and their relative structural invariance between serotypes. Grossman and Laimins have recently proposed the possibility that this 'barrel–helix' motif seen in the E2 DNA-binding domain will be seen in other origin-binding proteins (Grossman & Laimins, 1996). They go on to suggest that the similarity between EBNA-1 and the E2 DNA-binding domains might be an example of convergent evolution and that the 'barrel–helix' structure and the distortions to DNA which it inflicts are vitally necessary to each protein's role in directing origin-dependent replication (Grossman & Laimins, 1996). Regardless of the reason for the structural homology, given the similarity between both the BPV-1 and HPV-31 E2 DNA-binding domains as well as the EBNA-1 DNA-binding domain, it is likely that other E2 and EBNA DNA-binding domains from within each viral family will adhere to the same approximate core structure.

We wish to thank Dr John Kuzewski (National Institutes of Health) for assistance in implementing the Ramachandran force-field in *X-PLOR*. We are grateful to Drs Alexey Bochkarev and Aled M. Edwards (McMaster University) for providing us with the Cα coordinates of the EBNA-1 protein. We also thank Dr Ho Sup Yoon for providing the clone which was used to produce the E2 DNA-binding domain protein fragment. We would also like to thank our colleagues, Drs Vicki Nienaber, Steven Muchmore, Chang Park, Cele Abad-Zapatero, and Jonathan Greer for their assistance in critiquing this manuscript. Finally, we thank Dr Axel Brünger for communicating his results regarding the precision of NMR and X-ray crystallography structures.

References

Alexander, K. A. & Phelps, W. C. (1996). *Biochemistry*, **35**, 9864–9872.

Ambinder, R. F., Shah, W. A., Rawlins, D. R., Hayward, G. S. & Hayward, S. D. (1990). *J. Virol.* **64**, 2369–2379.

Barsoum, J., Prakash, S. S., Han, P. & Androphy, E. J. (1992). *J. Virol.* **66**, 3941–3945.

Bedrosian, C. L. & Bastia, D. (1990). *Virology* **2**, 557–575.

Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. E., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.

Bochkarev, A., Barwell, J. A., Pfuetzner, R. A., Bochkareva, E., Frappier, L. & Edwards, A. M. (1996). *Cell*, **84**, 791–8.

Bochkarev, A., Barwell, J. A., Pfuetzner, R. A., Furey, W. J., Edwards, A. M. & Frappier, L. (1995). *Cell*, **83**, 39–46.

Brünger, A. T. (1992). *X-PLOR Version 3.1. A system for X-ray Crystallography and NMR*, Yale University Press, New Haven, Connecticut, USA.

Brünger, A. T. (1997). *Nature Struct. Biol.* **4**(Suppl.), 862–865.

Burley, S. K. & Petsko, G. A. (1985). *Science*, **229**, 23–28.
Carson, M. (1987). *J. Mol. Graphics*, **5**, 103–106.
Chen, M. R., Zong, J. & Hayward, S. D. (1994). *Virology*, **205**, 486–495.
Chin, M. T., Hirochika, R., Hirochika, H., Broker, T. R. & Chow, L. T. (1988). *J. Virol.* **62**, 2994–3002.
Dillner, L., Zellbi, A., Avall-Ludqvist, E., Heino, P., Eklund, C., Pettersson, C. A., Forslund, O., Hansson, B. G., Grandien, M. & Bistoletti, P. (1995). *Cancer Detect. Prev.* **19**, 381–393.
Ducruix, A. & Giegé, R. (1992). Editors. *Crystallization of Nucleic Acids and Proteins: a Practical Approach*. Oxford: IRL Press.
Eron, L. J., Judson, F., Tucker, S., Prawer, S., Mills, J., Murphy, K., Hickey, M., Rogers, M., Flannigan, S. & Hein, N. (1986). *N. Engl. J. Med.* **315**, 1059–1064.
Galloway, D. A. (1994). *Infect. Agents Dis.* **3**, 187–193.
Galloway, D. A. & McDougall, J. K. (1989). *Adu. Virus Res.* **37**, 125–171.
Gauthier, J. M., Dillner, J. & Yaniv, M. (1991). *Nucleic Acids Res.* **19**, 7073–7079.
Giri, I. & Yaniv, M. (1988). *EMBO J.* **7**, 2823–2829.
Grossman, S. R. & Laimins, L. A. (1996). *Trends Microbiol.* **4**, 87–89.
zur Hausen, H. (1991). *Virology*, **184**, 9–13.
Hegde, R. S., Grossman, S. R., Lainmins, L. A. & Sigler, P. B. (1992). *Nature (London)*, **359**, 505–512.
Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. (1991). *Acta Cryst.* A**47**, 110–119.
Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993). *J. Appl. Cryst.* **26**, 283–291.
Li, R., Knight, J., Bream, G., Stenlund, A. & Botchan, M. (1989). *Genes Dev.* **3**, 510–526.
Liang, H., Petros, A. M., Meadows, R. P., Yoon, H. S., Egan, D. A., Walter, K., Holzman, T. F., Robins, T. & Fesik, S. W. (1996). *Biochemistry*, **35**, 2095–2103.
Lusky, M., Hurwitz, J. & Seo, Y. (1993). *J. Biol. Chem.* **268**, 15795–15803.
Luthy, R., Bowie, J. U. & Eisenberg, D. (1992). *Nature (London)*, **356**, 83–85.
Murphy, F. A., Fauguet, M. A., Mayo, A. W., Jarvis, S. A., Ghabrial, M. D., Summers, G. P. & Marelli, D. H. L. (1995). Editors. *Sixth Report of the International Committee on Taxonomy of Viruses.* New York: Springer Verlag.
Matthews, B. W. (1968). *J. Mol. Biol.* **33**, 491–497.
Navaza, J. (1994). *Acta Cryst.* A**50**, 157–163.
Nicholls, A., Sharp, K. & Honig, B. (1991). *Proteins*, **11**, 281–296.
Ollis, D. L. & White, S. W. (1990). *Methods Enzymol.* **162**, 646–659.
Otwinowski, Z. (1993). *Oscillation Data Reduction Program*. In *Proceedings of the CCP4 Study Weekend: Data Collection and Processing, 29–30 January*, edited by L. Sawyer, N. Isaacs & S. Bailey, pp. 56–62. Warrington: Daresbury Laboratory.
Prakash, S. S., Grossman, S. R., Pepinsky, R. B., Laimins, L. A. & Androphy, E. J. (1992). *Genes Dev.* **6**: 105–116.
Richards, F. M. (1977). *Annu. Rev. Biophys. Bioeng.* **6**, 151–176.
Rossmann, M. G. (1972). Editor. *The Molecular Replacement Method: A Collection of Papers on the Use of Molecular Replacement*. New York: Gordon & Breach.
Steitz, T. A. (1990). *Q. Rev. Biophys.* **23**, 205–280.
Stone, K. M. (1995). *Clin. Infect. Dis.* **20**, S91-S97.
Yates, J. L., Warren, N. & Sugden, B. (1985). *Nature (London),* **313**, 812–815.