Predicted α -helix/ β -sheet secondary structures for the zinc-binding motifs of human papillomavirus E7 and E6 proteins by consensus prediction averaging and spectroscopic studies of E7

Christopher G. ULLMAN*, Parvez I. HARIS*, Denise A. GALLOWAY†, Vincent C. EMERY‡ and Stephen J. PERKINS*§

*Department of Biochemistry and Molecular Biology, Royal Free Hospital School of Medicine, Rowland Hill Street, London NW3 2PF, U.K.,

†Department of Tumor Biology, Fred Hutchinson Cancer Research Center, 1124 Columbia Street, Seattle, WA 98104-2092, U.S.A., and ‡Department of Virology, Royal Free Hospital School of Medicine, Rowland Hill Street, London NW3 2PF, U.K.

The E7 and E6 proteins are the main oncoproteins of human papillomavirus types 16 and 18 (HPV-16 and HPV-18), and possess unknown protein structures. E7 interacts with the cellular tumour-suppressor protein pRB and contains a zinc-binding site with two Cys-Xaa₂-Cys motifs spaced 29 or 30 residues apart. E6 interacts with another cellular tumour-suppressor protein p53 and contains two zinc-binding sites, each with two Cys-Xaa₂-Cys motifs at a similar spacing of 29 or 30 residues. By using the GOR I/III, Chou–Fasman, SAPIENS and PHD methods, the effectiveness of consensus secondary structure predictions on zinc-finger proteins was first tested with sequences for 160 transcription factors and 72 nuclear hormone receptors. These contain Cys₂His₂ and Cys₂Cys₂ zinc-binding regions respectively, and possess known atomic structures. Despite the zinc- and DNA-binding properties of these protein folds, the major α -helix structures in both zinc-binding regions were correctly identified. Thus validated, the use of these prediction methods with 47 E7 sequences indicated four well-defined α -helix (α) and β -sheet (β) secondary structure elements in the order $\beta\beta\alpha\beta$ in the zinc-

binding region of E7 at its C-terminus. The prediction was tested by Fourier transform infrared spectroscopy of recombinant HPV-16 E7 in H₂O and ²H₂O buffers. Quantitative integration showed that E7 contained similar amounts of α -helix and β -sheet structures, in good agreement with the averaged prediction of α -helix and β -sheet structures in E7 and also with previous circular dichroism studies. Protein fold recognition analyses predicted that the structure of the zinc-binding region in E7 was similar to a $\beta\beta\alpha\beta$ motif found in the structure of Protein G. This is consistent with the E7 structure predictions, despite the low sequence similarities with E7. This predicted motif is able to position four Cys residues in proximity to a zinc atom. A model for the zinc-binding motif of E7 was constructed by combining the Protein G coordinates with those for the zinc-binding site in transcription factor TFIIS. Similar analyses for the two zinc-binding motifs in E6 showed that they have different α/β secondary structures from that in E7. When compared with 12 other zinc-binding proteins, these results show that E7 and E6 are predicted to possess novel types of zinc-binding structure.

INTRODUCTION

The high-risk human papillomavirus types 16 and 18 (HPV-16 and HPV-18) are frequently found in cervical carcinoma tissues and have been associated with the generation and maintenance of cervical neoplasia. The early gene region E6–E7 of HPV-16 is maintained intact after integration into host DNA and is actively expressed in carcinoma cells, leading to the production of the HPV-transforming proteins E6 and E7, which together can immortalize human fibroblasts. Both E7 and E6 bind zinc in one and two atypical zinc-binding sites respectively, characterized by the sequence Cys-Xaa₂-Cys-Xaa_{29–30}-Cys-Xaa₂-Cys. The function and structure of this region are largely unknown, but there is evidence to suggest that it is involved in protein–protein interactions with key transcription factors [1–4].

E7 is a 98-residue nuclear phosphoprotein that has sequence and functional similarity to proteins in other DNA tumour viruses such as adenovirus E1A and SV40 large tumour antigen. These similarities occur outside the zinc-binding site in the N-terminal region of E7, and are involved in *trans*-activation and transformation. The N-terminal region of E7 is responsible for binding important cellular proteins to deregulate the cell cycle. In this region the conserved motif LXCXE is essential for binding to the underphosphorylated form of pRB, the retinoblastoma tumour suppressor protein, which regulates the action of the transcription factor E2F by complex formation [2–6]. High-risk

HPV E7 binds pRB more efficiently than low-risk HPV E7, and will release E2F from a pRB–E2F complex. The dissociation of the pRB–E2F complex by E7 requires the C-terminal region of E7, which binds with lower affinity and at a different site on pRB than the N-terminus of E7. In this region, the zinc-binding site is necessary for E7 dimerization, protein stability and the full transformation potency of E7 [2–4,7].

E6 is a 150-residue protein that is also located within the nucleus of infected cells and shares functional similarities with SV40 large tumour antigen and adenovirus E1B 55 kDa protein [2–4]. In E6 both the N-terminal and C-terminal zinc-binding regions are necessary for *trans*-activation and transformation, and both are involved in protein–protein interactions. E6 binds to p53, a cellular tumour suppressor protein that causes growth arrest and apoptosis in cells after DNA damage [2–4,8]. E6 from high-risk HPV binds p53 with higher affinity than that from low-risk HPV, and mediates the degradation of p53 through the ubiquitin-dependent system [2–4,9,10]. High-risk E6 also prevents p53 from binding to DNA, and abrogates the p53-mediated repression of transcription *in vivo* [11,12], which might contribute to the oncogenic potential of E6. However, E6 co-operates with *ras* independently of p53 [13,14]. Mutagenesis studies have revealed that the C-terminal region of HPV-16 E6 and both regions in HPV-18 E6 are necessary for binding to p53, and that the N-terminal region houses the p53 degradation function [13,14].

Abbreviations used: α , α -helix; β , β -sheet; FT-IR, Fourier transform infrared; HPV, human papillomavirus.

§ To whom correspondence should be addressed.

Despite much understanding of the functional properties of E7 and E6, their protein structures are unknown. Knowledge of these structures might facilitate the design of antiviral inhibitors to interfere with tumour growth, and provide insight into the role of low-risk and high-risk E7 and E6 in cancer development. Atomic structures are known for ten distinct small zinc-binding motifs, many of which interact with nucleic acids [15,16] although E7 and E6 are involved in protein-protein interactions. These provide the basis of a predictive approach for examining the protein structures of E7 and E6 [17–19]. Multiple sequence alignments for the zinc-binding transcription factors and nuclear hormone receptors were used to predict their averaged secondary structure and solvent accessibility, which were then compared against their atomic structures. Thus validated, multiple sequence alignments for 47 E7 and 57 E6 sequences were analysed to predict their averaged α -helix (α) and β -sheet (β) secondary structures and solvent accessibilities. The predictions for E7 were compared with Fourier transform infrared (FT-IR) [20] and CD spectroscopy data on E7 that provide experimental evidence of secondary structures. Protein fold recognition methods [19,21] can thread a sequence of interest through a protein fold and score it for topological compatibility, using a library of 254 known folds. In favourable cases a candidate fold corresponding to the sequence can be identified. In such a case its topology is further tested by comparison with the secondary structure predictions for consistency, and molecular graphics modelling is used to confirm that key residues are correctly located within the predicted structure. Using this approach, we show that (1) the zinc-binding motifs in E7 and E6 are predicted to have $\beta\beta\alpha\beta$ and $\alpha\beta\alpha\beta$ secondary structures respectively, (2) the $\beta\beta\alpha\beta$ structure of E7 is predicted to resemble the protein fold of Protein G and a model can be built for its zinc-binding motif, and (3) both motifs are predicted as novel types of zinc-binding structures.

MATERIALS AND METHODS

Sequence alignments of zinc-binding proteins

Sequences corresponding to two groups of zinc-binding motifs of known structure were extracted from databases. A total of 160 zinc-finger motifs from 25 transcription factor sequences were used ([22], together with XFIN_XENLA). A total of 72 nuclear-hormone-receptor domains with double zinc-finger motifs were used as follows (SwissProt and PIR accession names): ESTR_HUMAN, ESTR_MOUSE, ESTR_RAT, ESTR_CHICK, RRA1_HUMAN, RRA1_MOUSE, RRA_NOTVI, RRB2_HUMAN, RRB1_CHICK, RRB1_MOUSE, RRB_NOTVI, RRD_NOTVI, RRG1_HUMAN, RRG1_MOUSE, RRG2_XENLA, RRXA_HUMAN, RRXA_MOUSE, RRXB_MOUSE, RRXC_HUMAN, RRXG_MOUSE, RRXA_CHICK, E75C_DROME, E75A_DROME, E75B_DROME, A30226, EAR1_HUMAN, A32608, TLL_DROME, TR2_HUMAN, VDR_RAT, VDR_HUMAN, COTF_HUMAN, 7UP1_DROME, A35872, NUR7_MOUSE, KNIR_DROME, KNRL_DROME, EGON_DROME, MUSH2RIIBP, THA2_HUMAN, THB2_HUMAN, THAA_XENLA, THAX_XENLA, THB1_XENLA, THB6_XENLA, THA1_RAT, THB1_RAT, THA_CHICK, THB_CHICK, THA1_MOUSE, RATERBAA, ERBA_AVIER, ERR1_HUMAN, ERR2_HUMAN, EAR2_HUMAN, PRGR_HUMAN, PRGR_RABIT, PRGR_CHICK, GCRA_HUMAN, GCR_RAT, GCR_MOUSE, MCR_HUMAN, MCR_RAT, ANDR_HUMAN, ANDR_MOUSE, ANDR_RAT, PPAR_HUMAN, PPAR_MOUSE, PPAR_RAT, PPAR_XENLA, PPAS_XENLA, PPAT_XENLA. Sequences for 51 papillomavirus E7 proteins and 57 papillomavirus E6 proteins were obtained from the Entrez CD-ROM database, release 17.0 (June

1995; National Center for Biotechnology Information). Alignments were performed initially with the multiple alignment program MULTAL with a range of fixed and variable gap penalties [23], and refined by eye to maximize the occurrence of conserved or chemically similar residues and minimize the number of gaps. Similar residues are defined as follows: tiny, G, A, S; aliphatic, I, L, V, M; aromatic, H, F, Y, W; positive, R, H, K; negative, D, E; hydroxyl, S, T; amide, N, Q [24].

Predictions of secondary structure and accessibility, and fold recognition analyses

Protein crystal and NMR structures were identified by their Brookhaven database codes. Their observed secondary structures were classified with DSSP [25] into three secondary structure types, namely β -strands (DSSP codes E and B), α -helices (DSSP codes I, H or G), and loop regions (DSSP codes s, t and blank). The observed side-chain accessibilities to solvent were calculated by the Lee-Richards method in COMPARE [26], and classified as inaccessible if they were less than 30%. Protein structures were visualized using INSIGHT II (Biosym/MSI, San Diego, CA, U.S.A.) or SETOR [27] on Silicon Graphics R3000 and R4000 series workstations. A molecular model of E7 was constructed with HOMOLGY (Biosym/MSI) to merge the coordinates of Protein G and transcription factor TFIIS. Energy minimization of the structure with DISCOVER (Biosym/MSI) used the consistent-valence force field while fixing the position of the Cys side chains involved in zinc co-ordination. Iterations were made with the steepest-descent algorithm.

Predictions of the averaged three- or four-state secondary structure from the multiple sequence alignment were performed by the statistically based GOR I and GOR III methods (with no decision constant bias) and the Chou-Fasman method [28]. The SAPIENS prediction is an environment-dependent method based on the probabilities of amino acid substitution [29]. The PHD prediction is a neural-networking method, where the network is trained on a database of known protein structures to recognize relationships between secondary structures and sequences [30]. The predictive accuracy is defined as $100 \times (\text{number of correctly identified residue positions}) / (\text{number of residues in the NMR or crystal structure within the consensus sequence})$.

Predictions of the mean hydropathy from the multiple sequence alignments were performed with the Eisenberg consensus hydrophobicity scale [28], where residues were scored as hydrophobic (b) or hydrophilic (e) by majority vote. Predictions of solvent accessibilities by the amino acid environment-dependent substitution tables using SAPIENS [31] gave residues that were either 20% accessible or less (b), or more than 20% accessible (e). Accessibility predictions with the PHD neural network method [32] resulted in three-state assignments where less than 9% accessibility was denoted as inaccessible (b), 9–36% accessibility was intermediate (.), and more than 36% accessibility was exposed (e).

In fold recognition analyses, the 51 E7 sequences were subjected to optimal fitting to a library of 254 known protein folds by using THREADER [21]. Threading energies were computed in terms of pairwise interaction energies to evaluate the fit of each sequence to a particular fold. The comparison was represented as a Z-score [$= (\text{energy} - \text{mean}) / \text{S.D.}$]. A high structural match requires a Z-score of less than -2.7 and becomes highly significant if it is less than -3.0 . Averaged Z-scores for the 51 E7 sequence threadings were calculated with SUM-THREADER, which also gave the averaged position of each of the 254 folds in the 51 threadings [19].

In this study, results from experimental crystal or NMR structures are described as observed, whereas predicted results are stated as such. Comparisons are made between observed and predicted features to substantiate the use of these methods for E7 and E6 below. NMR and crystal structures for 13 Cys₂His₂ zinc-finger proteins gave secondary structures summarized in Figure

NMR and crystal structures for four nuclear hormone receptors have been reported, and five coordinate sets for two proteins are available for secondary structure analysis (Figure 2b). Two short β -strands B1 and B2 were observed in four structures, followed by two α -helices A1 and A2 in all the structures. A1 is involved in DNA binding. The five consensus predictions of Figure 2(c) all identified a predicted α -helix A1, whereas B1 and B2 were merged into a single prediction B1/2. Variable predictions were observed for A2, which was displaced or extended towards the C-terminus in four of the five predictions. This might be related to an extra C-terminal α -helix observed in the retinoid X receptor α [41], for which the sequences are included in Figures 2(a) and 2(c) but no coordinates were available for Figure 2(b). The prediction of an N-terminal β -strand B was not experimentally observed, possibly because of structural disorder. The predicted minor features B3 and B4 did not correspond to the observed structures. The accuracies of the five predictions were

Figure 3 Alignment of consensus structure predictions for 47 E7 protein sequences

The consensus sequence is defined by residue occupancies in over half the E7 sequences. The most frequently occurring 19 residues that occur at positions with more than 90% conservation are indicated above the sequences. The single zinc-binding motif at Cys⁵⁸-Cys⁶¹-Cys⁹¹-Cys⁹⁴ is indicated below the sequences. Consensus averaged predictions of the E7 structure by the GOR III, Chou-Fasman, SAPIENS and PHD methods are shown; the symbols are as in Figure 1(c). The predicted secondary structure is A1, B1, A2, B2, B3, A3 and B4 in that order. Consensus accessibility predictions by the hydropathy, SAPIENS and PHD methods are shown, using the symbols given in Figure 1(d).

reduced to 46–73%, and the consensus prediction was 58% accurate. Despite these lower accuracies, the analyses of Figure 2 showed that the three main secondary structures B1/B2, A1 and A2 were successfully detected.

The consensus observed solvent accessibility from the six structures (Figure 2d) showed that 34 out of 88 side chains were less than 30% exposed (Figure 2d), and many of these are well-conserved in the sequence alignment. The accuracies of the three hydropathy or accessibility predictions were 63–79% and the consensus prediction was 68% accurate, showing again that these methods performed well.

Averaged secondary structure prediction for papillomavirus E7

The multiple sequence alignment for 47 Cys₄ E7 sequences (Figure 3) included the following sequences: 43 human, 2 primate, 1 rat and 1 rabbit. Four deer, elk and bovine papillomavirus sequences were excluded (VE7-PAPVD, VE7-PAPVE, W7WLB2, W7WLEB) as they showed little similarity between residues 1 and 40 at the N-terminal region and do not contain the

LXCXE motif at residues 23–27. The consensus length of 98 residues is occupied in over half the E7 sequences. More than 90% conservation was observed for 22 out of 98 residues. The alignment can be divided into a moderately conserved N-terminus region (40 residues) and a well conserved C-terminus zinc-binding region (58 residues) separated by a gap region of variable size.

The alignment of the N-terminal region of E7 was not as straightforward as that of the C-terminal region. The former contains the L²³XCXE²⁷ motif that binds to pRB and is conserved in sequences from other double-stranded DNA viral proteins and human cellular proteins [42,43]. Leu¹³ was almost completely conserved and seems to be essential for papillomavirus E7 function. Indeed, when Leu¹³ was mutated to Ser, the trans-forming function of the HPV-16 E7 protein in co-operation with HPV-16 E6 in human cells was abolished [43]. There are many negatively charged residues at the end of this region between residues 33 and 40.

The C-terminus of E7 contains the zinc-binding region, which is involved in E7 dimerization, transformation and *trans*-ac-

Figure 4 Graphical representation of the consensus secondary-structure predictions for the E7 protein of papillomavirus

Residue numbering is taken from Figure 3. The pRB-binding and zinc-binding motifs are highlighted. Tentative assignments of α -helix and β -strand are shown in parentheses. Symbols: ●, α -helix; ○, β -strand; broken line, turn/loop; dotted line, coil. The GOR III prediction in (a) indicates five predicted secondary-structure elements, whereas the Chou–Fasman prediction in (b) indicates seven and the PHD prediction in (c) indicates six.

tivation [4]. The zinc ligands occur at Cys⁵⁸, Cys⁶¹, Cys⁹¹ and Cys⁹⁴. This motif is 29 residues long with three exceptions (Figure 3). Seven hydrophobic aliphatic residues exhibit more than 90 % conservation, and five more exhibit more than 70 % conservation. This implies that the structure of this region is maintained by a hydrophobic core. Arg⁶⁶ and Arg⁷⁷ exhibit over 70 % conservation within this 29-residue segment. In transcription factors and nuclear hormone receptors, Arg residues are conserved where these form contacts with phosphate groups on DNA; however, the function of these Arg residues in E7 is not known.

Four secondary structure predictions for E7 are summarized in Figures 3 and 4, as follows.

(1) In the N-terminus region, two α -helices A1 and A2 and one β -strand B1 were predicted only by majority votes (Figure 3). Varied structure predictions in this region have also been reported by others [36,44]. Figure 4 shows that these features are not reproducibly observed, and the assignments are therefore tentative. The L²³XCXE²⁷ pRB-binding motif was associated with the possible predicted β -strand B1. The varied outcome of these predictions suggests that the N-terminal pRB-binding region might have a structure that is only fully formed in the complex between E7 and pRB.

(2) In the C-terminus region, four features B2, B3, A3 and B4

(denoted as $\beta\beta\alpha\beta$ below) with significant sizes were now clearly predicted by all four methods (Figure 4). The reproducibility of the four predictions is reminiscent of the clear α -helix prediction in Figure 1(c). Previous predictions had identified only B2, B3 and A3 by the GOR I method and two others in an analysis of 12 E7 sequences [36]. The predicted β -strand B2 (nine residues) precedes the zinc-binding motif at Cys⁵⁸–Cys⁶¹. The predicted elements B3 (seven residues), A3 (twelve residues) and B4 (four residues) occur between the two Cys–Xaa₂–Cys motifs. Both Cys–Xaa₂–Cys motifs are well predicted to occur at turn or loop structures.

FT-IR spectroscopy of recombinant HPV-16 E7 protein

The prediction of the secondary structure of E7 was tested by FT-IR spectroscopy. Recombinant HPV-16 E7 was purified from a yeast expression system [33]. Purification under mild reducing conditions to prevent intra- and inter-molecular disulphide bonds resulted in protein that migrated at an apparent molecular mass of approx. 50 kDa through a G75 Sephadex superfine column, as previously observed and assigned to multimers of E7 [34,45]. Denaturing SDS/PAGE showed that HPV-16 E7 was intact up to the refolding stage, after which 80 % possessed a band at an apparent molecular mass of 19 kDa and

Figure 5 FT-IR spectra of recombinant HPV-16 E7 protein

The absorbance spectra are shown in (a) and (c), and the second derivative spectra are shown in (b) and (d). Data were obtained at a concentration of 14 mg/ml E7 in H₂O buffer (a, b) and in ²H₂O buffer (c, d).

20% had bands at 14 or 16 kDa, which are presumed to be cleaved products even in the presence of protease inhibitors or freshly repurified urea. Multiple bands seen after treatment with urea have also been observed previously [45]. The omission of the urea denaturation stage resulted in much reduced quantities (less than 15%) of cleaved E7 at a single apparent molecular mass of 16 kDa. The abnormal migration of HPV-16 E7 at an apparent molecular mass of 19 kDa has been reported previously [46,47]. This is known to result from the presence of Asp⁴ in the N-terminal region of HPV-16 E7, because the removal of Asp⁴ causes E7 to migrate normally during electrophoresis [48,49].

FT-IR spectroscopy was performed on HPV-16 E7 in both H₂O and ²H₂O buffers to monitor solvent effects [20,37] (Figure 5). The broad amide I band (Figures 5a and 5c) corresponds primarily to C=O stretching vibrations. This occurs at 1650 cm⁻¹ in H₂O buffer and undergoes a large shift to 1642 cm⁻¹ in ²H₂O buffer. In ²H₂O buffer the disappearance of the amide II band at 1550 cm⁻¹ shows that H-²H exchange in E7 is rapid, indicating a solvent-accessible structure. A band at 1577 cm⁻¹ now becomes visible, which is associated with the absorption of amino acid side chains. The second-derivative spectrum (Figures 5b and 5d) reveals the substructure of the broad amide I band. In H₂O buffer, a strong band occurs at 1654 cm⁻¹ with a shoulder at 1645 cm⁻¹. The 1654 cm⁻¹ band can be assigned to an α -helix, as frequencies at 1650–1658 cm⁻¹ are usually considered to be characteristic of α -helices [20]. However, overlap from random coil structure can also occur at this frequency in H₂O buffer. The band at 1634 cm⁻¹ is characteristic of a β -sheet, and the weak band at 1687 cm⁻¹ can be attributed to antiparallel β -strands. The weak components at 1669–1679 cm⁻¹ can be assigned to turns. In ²H₂O buffer the main band is shifted by 12 cm⁻¹ to 1642 cm⁻¹. This is a large shift for a protein that contains a small

Table 1 Secondary structure of HPV-16 E7 determined by consensus predictions and quantitative FT-IR and CD spectroscopy data

Method	Conformation			
	α -helix (%)	β -sheet (%)	Turn (%)	Coil (%)
GOR III prediction	33	20	27	20
Chou-Fasman prediction	52	34	14	—
SAPIENS prediction	37	19	—	44
PHD prediction	17	24	—	59
Means $\pm$ S.D. for four predictions	35 $\pm$ 14	24 $\pm$ 7	21 $\pm$ 9	41 $\pm$ 20
FT-IR spectroscopy* (Figure 5)	28 $\pm$ 3	38 $\pm$ 8	16	18
CD spectroscopy [36] †	29 $\pm$ 5	11 $\pm$ 5	28 $\pm$ 5	33 $\pm$ 5

* Errors in the quantification of FT-IR spectra were taken from [38].

† Errors in the CD quantitative analyses were taken from [36]. The data relate to E7 in the presence of zinc.

proportion of random coil structure. However, proteins with large contents of secondary structure can exhibit large shifts in their amide I bands in the presence of ²H₂O buffers as the consequence of highly solvent-exposed α -helix structures [37]. The band shift observed for E7 is thus attributed to the presence of solvent-exposed α -helices. This is consistent with the hydrophathy and accessibility predictions of Figure 3 for E7, in particular for A2.

The observation of resolved bands for both α -helix and β -sheet shows that both secondary structures are present in E7. Quantification with the FT-IR data in H₂O buffer [38] gave 28% α -helix, 38% β -sheet, 16% turn and 18% random coil. This is within the error of the averaged structure predictions in Table 1, and supports the analyses of Figures 3 and 4. Similar results had been obtained from CD spectroscopy with recombinant HPV-16 E7 expressed in *Escherichia coli* [36], although a lower content of β -sheet was reported. The differences in β -sheet contents are attributed to the greater difficulty in quantitative analyses of β -sheet than α -helix [38].

Given the multiple protein species observed after urea unfolding, it is possible that small amounts of degraded E7 affect the FT-IR spectra. FT-IR spectra of degraded proteins exhibit narrow absorbance bands between 1620 and 1690 cm⁻¹ that correspond to aggregated intermolecular β -sheet structures formed by unfolded polypeptide chains [37]. The E7 FT-IR spectra do not exhibit such narrow bands, with the exception of a minor band observed in H₂O buffer at 1615 cm⁻¹ (Figure 5b) but not in ²H₂O buffer, which might arise from amino acid side chains. In addition, FT-IR spectra of HPV-16 E7 prepared in ²H₂O buffer without a urea unfolding stage were identical with those in Figure 5 (results not shown).

Protein fold recognition analyses of the E7 zinc-binding region

Having verified the secondary-structure prediction for E7 by FT-IR spectroscopy, protein fold recognition methods were now applied to the well-conserved C-terminal zinc-binding region. As this had given a clear secondary structure prediction $\beta\beta\alpha\beta$ in Figure 4, it was reasoned that this region had a good chance of being matched to a known fold. Residues 51–98 in each of the 51 E7 sequences (E7^{51–98}) were tested for compatibility with a library of 254 known protein folds. As observed in other THREADER analyses [19], the pairwise energy Z-scores varied

7(a). This hybrid structure was energy-minimized and then used as a template for the side chains of HPV-16 E7. This model of E7 (Figure 8) was again energy-minimized until analysis of its structure for connectivity and bad contacts using PROCHECK [50] showed that its stereochemistry was satisfactory. Structure analyses of the E7 model showed that, whereas A3 can be identified in the position specified in Figure 6, B2, B3 and in particular B4 were all shorter than expected. A suitable zinc-binding site had been incorporated in this E7 model. Although the model offers a satisfactory account of its known properties, an atomic structure determination for E7 is required to confirm its authenticity.

Averaged secondary-structure prediction for papillomavirus E6

Structure predictions were performed for 57 E6 sequences (Figure 9) to assess whether the two zinc-binding regions were similar in structure to that of E7. The E6 sequences were highly conserved and 32 residues in a consensus length of 150 residues showed more than 90% conservation. Comparison of the two zinc-binding regions in E6 shows that (1) their sequences are closely matched in length, (2) a single gap occurs in both regions at the same position, (3) they exhibit the greatest residue conservation, and (4) eight conserved zinc-binding Cys ligands and other pairs of conserved or similar residues occur at the same relative positions in both regions. The conserved pairs of residues include Ile/Leu/Val²⁹ and Ile/Leu/Val¹⁰², Leu³⁸ and Leu¹¹¹, Ile/Leu/Val⁵³ and Phe¹²⁶, Glu/Asp⁴² and Glu¹¹⁵, and Gly/Ala⁶² and Gly/Ala¹³⁵. Further highly conserved residues in the two zinc-binding regions are, however, not related to each other. Only Leu³⁸ and Leu¹¹¹ are highly conserved in a position analogous with that at position 65 in E7 (Figure 3). This suggests that the zinc-binding domains of E6 and E7 are similar but might not be structurally identical.

The four predictions of Figure 9 indicate that both zinc-binding regions of E6 have different structures from that in E7. The predictions suggest that the first region in E6 has a well-defined $\alpha\beta\alpha\beta$ motif (A1-B1-A2-B2 in Figure 9) and that the second region has a similar $\alpha\beta\alpha\beta$ motif (A3-B4-A4-B5 in Figure 9), although this second prediction was weaker and majority votes were required for identification. The two regions were joined by a possible β -strand B3, which was predicted only by a majority vote. The two regions were initiated by the predicted α -helices A1 and A3, and the predicted β -strands B1 and B4 were located at the first Cys-Xaa₂-Cys motif of each zinc-binding region. The second $\alpha\beta$ pair in each region was contained within each zinc-binding loop in E6. Figure 9 shows that the α -helices in each predicted $\alpha\beta\alpha\beta$ motif possessed amphipathic properties, whereas the β -strands were more hydrophobic. Even though E6 and E7 have similar lengths of zinc-binding region, E6 is predicted to have a double $\alpha\beta\alpha\beta$ secondary structure, which is distinct from the $\beta\beta\alpha\beta$ structure predicted for E7. THREADER analyses of the 57 full E6 sequences and for each of the separate zinc-binding regions showed no correlation with known α/β protein folds.

Conclusions

Predicted structures for the zinc-binding regions of E7 and E6

E5, E7 and E6 in papillomaviruses are among the smallest known transforming proteins [2,4]. We have previously analysed the structure of the transmembrane protein E5 by a joint secondary-structure prediction and FT-IR approach to show that it has a motif related to connexin of gap junctions [51], which is now supported by biochemical evidence [52]. In the

present study, we have proposed a rationally based structural model for the E7 C-terminal zinc-binding motif, although an atomic-structure determination will be required to verify it. The model assumes that each monomer is structurally independent and binds a single zinc atom. It might be useful for the planning of an atomic-structure determination for the zinc-binding region of E7, and mutagenesis studies to determine the molecular basis of low-affinity pRB binding, E2F release from a pRB-E2F complex, and E7 dimerization.

The protein structure prediction strategy in this study was initially validated by reference to other small zinc-binding proteins of known structure. Predicted structures for the transcription factor and nuclear hormone receptors were successfully correlated with their observed NMR and crystal structures. Only the major features in these structures were identified in the predictions. In application to E7, a $\beta\beta\alpha\beta$ secondary structure was well predicted in the C-terminal zinc-binding region. The predictions were variable for the N-terminal region of E7, and it is possible that this is less defined structurally. The existence of both α -helix and β -sheet structures in E7 was experimentally supported by FT-IR and CD spectroscopy. Protein fold recognition analyses showed that the zinc-binding region of E7 was well correlated with the $\beta\beta\alpha\beta$ fold of Protein G. A plausible model could then be constructed by the combination of a $\beta\beta\alpha$ motif from Protein G with the zinc-binding site and a β -strand from transcription factor TFIIS. This predicted model for E7 (Figure 8) explains why mutations or deletions of residues within this zinc-binding region lead to structural instability [43,53–55], as these substitutions disrupt its well-ordered structure. It is interesting that an E7 mutant containing only residues 67–98 possesses *trans*-activation activity and binds zinc as effectively as native E7 [56]. The predicted E7 model shows that residues 67–98 correspond neatly to A3 and B4, both of which are presumably able to retain their structures as a complete motif in this mutant. E7 forms dimers whose formation might be mediated by the chelation of zinc. It is not known whether single Cys-Xaa₂-Cys motifs from two separate E7 molecules form a zinc-binding site [34,45], or whether the binding of zinc creates a dimerization interface by altering the protein structure so that two monomers can fit together.

The similar lengths of 29 residues for the two Cys-Xaa₂-Cys motifs in E6 and that in E7 did not correspond to a similar predicted structure for the zinc-binding region in E6. Sequence similarity is not high between these motifs in E6 and E7. Analyses resulted in the prediction of a double $\alpha\beta\alpha\beta$ structure for E6, which is distinct from the $\beta\beta\alpha\beta$ structure predicted for E7.

Structures of small zinc-binding proteins

The use of two Cys-Xaa₂-Cys motifs separated by 29 or 30 residues to co-ordinate zinc make E7 and E6 distinct from other related zinc-binding motifs. The use of zinc in E7 and E6 can be explained on the basis that Cys–zinc–Cys bridges will stabilize the E7 and E6 protein structures in the reducing environment of the cell nucleus where disulphide bridges do not form [47,57]. The interest in these structures resides in the protein–protein interactions that E7 and E6 form with the tumour-suppressor proteins pRB and p53 respectively, in contrast with the interactions that classical zinc-finger structures make with DNA and RNA [45]. The zinc-binding region of E7 interacts with low affinity with the C-terminus of pRB and prevents interactions between pRB and E2F [7].

The structure predictions for the E6 and E7 zinc-binding proteins of the papillomaviruses have resulted in motifs that have not previously been documented for ten other zinc-binding

proteins [15]. These motifs in E7 and E6 have not been described as 'zinc-fingers' as this term is best reserved for the classical DNA-binding transcription factor and nuclear hormone receptor proteins. The zinc-binding loops of E7 and E6 are among the longest so far characterized, with 29 or 30 residues between the two Cys-Xaa₂-Cys motifs. Two further structures are available. A much larger loop is found in the Cys-Xaa₂-Cys-Xaa₈₀-Cys-Xaa₂-Cys zinc-binding motif of the guanine-nucleotide-exchange factor human Mss4, which is mostly β -sheet [58]. Another zinc-binding motif in the DNA-binding protein gp32 is based on a His-Xaa₁₂-Cys-Xaa₉-Cys-Xaa₂-Cys sequence [59].

The ten zinc-binding structures have been classified into six groups [15], as follows.

(1) The classical zinc-finger structure has a $\beta\beta\alpha$ motif with a Cys₂His₂ zinc co-ordination in which both His residues occur on the same α -helix (Figure 1), with a finger 12 residues long between the pairs of zinc-binding residues.

(2) The nuclear hormone receptors, together with GATA-1 and LIM, form a group of three structures based mostly on a single or double $\alpha\alpha$ motif, with a Cys₄ or Cys₃His zinc co-ordination, and fingers 13–18 residues long (Figure 2).

(3) The equine herpes virus RING finger and the cystine-rich subdomain of protein kinase C contain a double $\beta\beta\beta\alpha$ motif with either Cys₄ or Cys₃His zinc co-ordination in which each loop 16 and 17 residues long contributes to the structure of the other motif.

(4) The zinc-binding structure of transcription factor TFIIS has a $\beta\beta\beta$ motif with a Cys₄ zinc co-ordination and 24 residues in the loop region.

(5) GAL4 has six Cys residues held within an $\alpha\alpha$ motif, and these co-ordinate two zinc atoms in a cluster with a six-residue loop between pairs of zinc-binding residues.

(6) The retroviral 'zinc knuckle' protein with a Cys₃His zinc co-ordination has four to ten residues in the loop between the two pairs of zinc binding residues and no readily discernible secondary structures.

Of those ten structures, five possess mixed α/β structures, three contain only α or β structure, and two have neither. It is clear that the predicted $\beta\beta\alpha\beta$ and $\alpha\beta\alpha\beta$ and the observed β_{12} and $\beta\beta\alpha\beta\alpha\alpha$ zinc-binding motifs of E7, E6, Mss4 and gp32 respectively constitute further unique folds that are distinct from the six groups described above [15], and indicate further diversity in the structures of small zinc-binding protein folds.

We thank Dr. R. B. Sim and Professor D. Chapman for useful discussions, and Dr. Y. J. K. Edwards, Dr. K. F. Smith, Mr. N. C. Brissett and Mr. A. S. Nealis for computational support. C. G. U. thanks the Science and Engineering Research Council for a Ph.D. studentship, and the Royal Free Hospital School of Medicine and the Peter Samuel Royal Free Fund for further support.

REFERENCES

- Galloway, D. A. and McDougall, J. K. (1989) *Adv. Virus Res.* **37**, 125–171
- Mansur, C. P. and Androphy, E. J. (1993) *Biochim. Biophys. Acta* **1155**, 323–345
- Vousden, K. (1993) *FASEB J.* **7**, 872–879
- Ullman, C. G. and Emery, V. C. (1996) *Rev. Med. Virol.* **6**, 39–55
- Dyson, N., Guida, P., Münger, K. and Harlow, E. (1992) *J. Virol.* **66**, 6893–6901
- Weintraub, S. J., Chow, K. N. B., Luo, R. X., Zhang, S. H., He, S. and Dean, D. C. (1995) *Nature (London)* **375**, 812–815
- Patrick, D. R., Oliff, A. and Heimbrook, D. C. (1994) *J. Biol. Chem.* **269**, 6842–6850
- Lane, D. P. (1993) *Nature (London)* **362**, 786–787
- Band, V., Dalal, S., Delmolino, L. and Androphy, E. J. (1993) *EMBO J.* **12**, 1847–1852
- Scheffner, M., Werness, B. A., Huibregtse, J. M., Levine, A. J. and Howley, P. M. (1990) *Cell* **63**, 1129–1136
- Lechner, M. S., Mack, D. H., Finicle, A. B., Crook, T., Vousden, K. H. and Laimins, L. A. (1992) *EMBO J.* **11**, 3045–3052 (erratum: *EMBO J.* **11**, 4248)
- Molinari, M. and Milner, J. (1995) *Oncogene* **10**, 1849–1854
- Crook, T., Tidy, J. A. and Vousden, K. H. (1991) *Cell* **67**, 547–556
- Pim, D., Storey, A., Thomas, M., Massimi, P. and Banks, L. (1994) *Oncogene* **9**, 1869–1876
- Klug, A. and Schwabe, J. W. R. (1995) *FASEB J.* **9**, 597–604
- Berg, J. M. (1990) *Annu. Rev. Biophys. Biophys. Chem.* **19**, 405–421
- Russell, R. B. and Sternberg, M. J. E. (1995) *Curr. Biol.* **5**, 488–490
- Perkins, S. J., Smith, K. F., Williams, S. C., Haris, P. I., Chapman, D. and Sim, R. B. (1994) *J. Mol. Biol.* **238**, 104–119
- Edwards, Y. J. K. and Perkins, S. J. (1996) *J. Mol. Biol.* **260**, 277–285
- Surewicz, W. K., Mantsch, H. H. and Chapman, D. (1993) *Biochemistry* **32**, 389–394
- Jones, D. T., Taylor, W. R. and Thornton, J. M. (1992) *Nature (London)* **358**, 86–89
- Gibson, T. J., Postma, J. P. M., Brown, R. S. and Argos, P. (1988) *Prot. Eng.* **2**, 209–218
- Taylor, W. R. (1988) *J. Mol. Evol.* **28**, 161–169
- Taylor, W. R. (1986) *J. Mol. Biol.* **188**, 233–258
- Kabsch, W. and Sander, C. (1983) *Biopolymers* **22**, 2577–2637
- Šali, A. and Blundell, T. L. (1990) *J. Mol. Biol.* **212**, 403–428
- Evans, S. V. (1993) *J. Mol. Graphics* **11**, 134–138
- Perkins, S. J., Haris, P. I., Sim, R. B. and Chapman, D. (1988) *Biochemistry* **27**, 4004–4012
- Wako, H. and Blundell, T. L. (1994) *J. Mol. Biol.* **238**, 693–708
- Rost, B. and Sander, C. (1993) *J. Mol. Biol.* **232**, 584–599
- Wako, H. and Blundell, T. L. (1994) *J. Mol. Biol.* **238**, 682–692
- Rost, B. and Sander, C. (1994) *Proteins* **20**, 216–226
- Carter, J. J., Yaegashi, N., Jenison, S. A. and Galloway, D. A. (1991) *Virology* **182**, 513–521
- Patrick, D. R., Zhang, K., Defeo-Jones, D., Vuocolo, G. R., Maigetter, R. Z., Sardana, M. K., Oliff, A. and Heimbrook, D. C. (1992) *J. Biol. Chem.* **267**, 6910–6915
- Perkins, S. J. (1986) *Eur. J. Biochem.* **157**, 169–180
- Pahel, G., Aulabaugh, A., Short, S. A., Barnes, J. A., Painter, G. R., Ray, P. and Phelps, W. C. (1993) *J. Biol. Chem.* **268**, 26018–26025
- Jackson, M., Haris, P. I. and Chapman, D. (1991) *Biochemistry* **30**, 9681–9686
- Lee, D. C., Haris, P. I., Chapman, D. and Mitchell, R. C. (1990) *Biochemistry* **29**, 9185–9193
- Russell, R. B. and Barton, G. J. (1993) *J. Mol. Biol.* **234**, 951–957
- Luisi, B. F., Xu, W. X., Otwinowski, Z., Freedman, L. P., Yamamoto, K. R. and Sigler, P. B. (1991) *Nature (London)* **352**, 497–505
- Lee, M. S., Kliever, S. A., Provencal, J., Wright, P. E. and Evans, R. M. (1993) *Science* **260**, 1117–1121
- Chan, S.-Y., Bernard, H.-U., Ong, C.-K., Chan, S.-P., Hofmann, B. and Delius, H. (1992) *J. Virol.* **66**, 5714–5725
- Watanabe, S., Kanda, T., Sato, H., Furuno, A. and Yoshiike, K. (1990) *J. Virol.* **64**, 207–214
- Jones, R. E., Heimbrook, D. C., Huber, H. E., Wegrzyn, R. J., Rotberg, N. S., Stauffer, K. J., Lumma, P. K., Garsky, V. M. and Oliff, A. (1992) *J. Biol. Chem.* **267**, 908–912
- McIntyre, M. C., Frattini, M. G., Grossman, S. R. and Laimins, L. A. (1993) *J. Virol.* **67**, 3142–3150
- Imai, Y., Matsushima, Y., Sugimura, T. and Terada, M. (1991) *J. Virol.* **65**, 4966–4972
- Sato, H., Watanabe, S., Furuno, A. and Yoshiike, K. (1989) *Virology* **170**, 311–315
- Münger, K., Yee, C. L., Phelps, W. C., Pieterpol, J. A., Moses, H. L. and Howley, P. M. (1991) *J. Virol.* **65**, 3943–3948
- Armstrong, D. J. and Roman, A. (1993) *Biochem. Biophys. Res. Commun.* **192**, 1380–1387
- Morris, A. L., MacArthur, M. W., Hutchinson, E. G. and Thornton, J. M. (1992) *Proteins* **12**, 345–364
- Ullman, C. G., Haris, P. I., Kell, B., Cason, J., Jewers, R. J., Best, J. M., Emery, V. C. and Perkins, S. J. (1994) *Biochem. Soc. Trans.* **22**, 439S
- Oelze, I., Kartenbeck, J., Crusius, K. and Alonso, A. (1995) *J. Virol.* **65**, 7078–7083
- Phelps, W. C., Münger, K., Yee, C. L., Barnes, J. A. and Howley, P. M. (1992) *J. Virol.* **66**, 2418–2427
- Edmonds, C. and Vousden, K. H. (1989) *J. Virol.* **63**, 2650–2656
- Watanabe, S., Sato, H., Furuno, A. and Yoshiike, K. (1992) *Virology* **190**, 872–875
- Rawls, J. A., Pusztal, R. and Green, M. (1990) *J. Virol.* **64**, 6121–6129
- Greenfield, I., Nickerson, J., Sheldon, P. and Stanley, M. (1991) *Proc. Natl. Acad. Sci. U.S.A.* **88**, 11217–11221
- Yu, H. and Schreiber, S. L. (1995) *Nature (London)* **376**, 788–791
- Shamoo, Y., Friedman, A. M., Parsons, M. R., Konigsberg, W. H. and Steiz, T. A. (1995) *Nature (London)* **376**, 362–366