DNA Sequence of the Major Capsid Protein Gene of Herpes Simplex Virus Type 1

By A. J. DAVISON*† AND J. E. SCOTT
MRC Virology Unit, Institute of Virology, University of Glasgow, Church Street, Glasgow G11 5JR, U.K.

(Accepted 16 July 1986)

SUMMARY

The DNA sequence of the region of the herpes simplex virus type 1 genome encoding the major capsid protein was determined. The predicted protein contains 1374 amino acid residues and has a molecular weight of 149075. Comparisons of the amino acid sequence of this protein with those predicted from the published DNA sequences of two other herpesviruses, varicella-zoster virus and Epstein–Barr virus, resulted in the identification of the major capsid protein gene in each genome.

The virion of herpes simplex virus (HSV) comprises an enveloped capsid containing a large linear double-stranded DNA genome bearing about 80 genes. In pioneering work, Wildy et al. (1960) concluded that the capsid is an icosahedron with a surface lattice of 150 hexavalent capsomeres at the faces and 12 pentavalent capsomeres at the vertices. Vernon et al. (1974) implied that the hexavalent capsomeres may be trimers, but subsequent work supported the view that they are hexamers (Palmer et al., 1975; Almeida et al., 1978; Furlong, 1978; Steven et al., 1986). In general agreement with Wildy et al. (1960), the latter authors visualized the hexavalent capsomere as a cylindrical cone about 15 nm in height and 12 nm in diameter, with a tapering central indentation about 4 nm wide at the outer surface. They concluded tentatively from theoretical considerations of protein density that the hexamer subunit is a single copy of the major capsid protein, which has been identified as a polypeptide with an apparent molecular weight of 155000 (Spear & Roizman, 1972; Cohen et al., 1980).

A slight difference in mobility on polyacrylamide gels between the major capsid proteins of HSV types 1 and 2 (HSV-1 and HSV-2) enabled the approximate location of the major capsid protein gene to be determined by analysis of recombinants between the two closely related viruses (Morse et al., 1978; Marsden et al., 1978). Anderson et al. (1979) demonstrated that a 6 kb mRNA mapping in this region of the HSV-1 genome is the most abundant viral transcript associated with polyribosomes late in infection. Costa et al. (1981) showed that this mRNA is not spliced, and ascertained approximate map locations for the 5' and 3' ends. These workers were able to translate the 6 kb mRNA in vitro into a protein with an apparent molecular weight of 155000, which Costa et al. (1984) showed is the major capsid protein.

In this paper, we describe the gene arrangement in a 9371 bp DNA fragment (KpnI i) containing the HSV-1 major capsid protein gene and, by comparing amino acid sequences, identify the major capsid protein genes of two other herpesviruses: varicella-zoster virus (VZV) and Epstein–Barr virus (EBV).

† Present address: Laboratory of Viral Diseases, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland 20892, U.S.A.
A recombinant plasmid containing the \textit{KpnI} fragment from HSV-1 strain 17 DNA (Davison & Wilkie, 1983) was propagated in \textit{Escherichia coli} DH1. The DNA sequence of \textit{KpnI} was determined using the M13-dideoxynucleotide technology (Sanger \textit{et al.}, 1980). Ninety-three percent of the sequence was obtained on both strands. DNA sequences were compiled and analysed using the programs of Staden (1982) modified by P. Taylor for a DEC PDP11/44 computer operating under the RSX11M system. Amino acid sequences and compositions were obtained using the translation program of Taylor (1986). Homologous proteins were identified using a rapid search program which was derived from the EMBLSCAN program described by Bishop & Thompson (1984), and amino acid sequences were aligned using the direct alignment program of Taylor (1984).

The location of \textit{KpnI} in the HSV-1 genome is shown in Fig. 1(a). Fig. 1(b) summarizes the approximate locations of transcripts previously mapped in this region. Analysis of the DNA sequence of \textit{KpnI} indicated the presence of three complete open reading frames encoding proteins with molecular weights of 149075 (149K), 34269 (34K) and 24230 (24K), and parts of two others encoding proteins with molecular weights of at least 53442 (>53K) and 22468 (>22K). These open reading frames show a similar codon usage to that of the HSV-1 thymidine kinase gene (McKnight, 1980), indicating that they probably encode proteins. In addition to depicting the arrangement of protein-coding regions, Fig. 1(c) shows positions of the sequence AATAAA, which is part of the signal for polyadenylation of eukaryotic mRNAs (Fitzgerald & Shenk, 1981). The location of the 6 kb mRNA indicates that 149K is the major capsid protein.
The DNA sequence of the region of *KpnI i* containing the major capsid protein gene and the predicted amino acid sequence are shown in Fig. 2; the remainder of the *KpnI i* sequence will be published elsewhere.

The arrangement of genes in *KpnI i* predicted from the DNA sequence is largely in agreement with that deduced from published transcript mapping data. The latter indicate that the mRNAs encoding 149K and 34K form a 3'-coterminal family; that is, they have unique 5' termini but share the same 3' terminus. Thus, since the mRNA encoding 149K terminates 1.4 kbp downstream from the translational stop codon, it is somewhat surprising to find the sequence AATAAA close downstream from the stop codon (Fig. 2). This observation is rendered more intriguing by the presence of AATAAA close downstream from the stop codons of the VZV and EBV major capsid protein genes identified below. However, only certain AATAAA elements are involved in polyadenylation, and the transcript mapping data clearly indicate that the mRNAs encoding 149K and 34K are 3'-coterminal. Nevertheless, it has not been ruled out that polyadenylation of a minor proportion of 149K transcripts occurs close to the stop codon.

The gene upstream of that encoding 149K specifies a particularly hydrophobic protein (24K); the coding sequence partially overlaps the promoter region of the 149K gene (Fig. 2). The absence of a potential polyadenylation site close downstream from the 24K coding region implies that the transcript may be 3'-coterminal with those encoding 149K and 34K; that is, about 6.8 kb in size. Costa et al. (1981) detected a 7 kb mRNA of undetermined orientation, not shown in Fig. 1 (b), which at least partially overlaps the 6 kb mRNA and the region encoding 24K. Hence, it is possible that the 7 kb mRNA encodes 24K. Alternatively, the transcript specifying 24K may have been present in low abundance, and therefore not detected by Costa et al. (1981). Earlier versions of the DNA sequence encoding the carboxy terminus of 24K and the non-coding region at the 5' end of the 149K gene in HSV-1 strain KOS (Frink et al., 1981; Costa et al., 1984) contain several errors which were corrected without comment by Costa et al. (1985a). The remaining differences between the HSV-1 strain 17 and strain KOS sequences are located between the 24K and 149K coding regions, and thus may represent genuine sequence differences between the two strains.

Comparisons of the amino acid sequence of 149K with protein sequences derived from the complete VZV and EBV DNA sequences (Davison & Scott, 1986; Baer et al., 1984) resulted in the identification of a homologue of the 149K gene in each genome. These genes, VZV gene 40 and EBV BcLF1 according to the published nomenclature, are therefore excellent candidates for the major capsid protein genes of VZV and EBV. The predicted molecular weights of 154971 and 153909, respectively, for the encoded proteins correspond well with the apparent molecular weights of the major capsid proteins, which were estimated at 150000 for VZV (Grose, 1980) and 160000 for EBV (Dolyniuk et al., 1976). VZV gene 40 is located in the region of the VZV genome which hybridized to HSV-1 DNA fragments containing the major capsid protein gene (Davison & Wilkie, 1983).

An alignment of the predicted amino acid sequences of the HSV-1, VZV and EBV major capsid proteins is shown in Fig. 3. The degree of homology between each pair of proteins was calculated from Fig. 3 by expressing the number of matched residues as a percentage of 1453 (the number of residues from the first residue of the VZV protein to the last residue of the EBV protein, including blank characters inserted by the alignment program). On this basis, the degree of homology between the HSV-1 and VZV proteins is 50%, whereas that between the HSV-1 and EBV proteins is 27% and that between the VZV and EBV proteins is 24%. This result illustrates the closer relationship between HSV-1 and VZV, which belong to the same subfamily of the herpesviruses (the *Alphaherpesvirinae*), than between either virus and EBV, which is a member of a different subfamily (the *Gammaherpesvirinae*). The 19% of residues which are identical in the three proteins are distributed unevenly throughout the sequences; the longest completely conserved region is six residues in length. Eight percent of residues are conserved only between the HSV-1 and EBV proteins, whereas 5% are shared only by the VZV and EBV proteins. It is doubtful, however, that this indicates a closer evolutionary relationship between the HSV-1 and EBV proteins than between those of VZV and EBV, since the small difference is more readily explained as an effect of base composition on the coding potential of
Fig. 2. DNA sequence of the region of KpnIi encoding the major capsid protein (149K in Fig. 1c). The position of the 5' end of the mRNA defined by Dennis & Smiley (1984) and Costa et al. (1984) is shown. Asterisks indicate the locations of the TATA element upstream of the 5' end of the mRNA and an AATAAA sequence downstream from the protein-coding region. The amino acid sequence of the entire major capsid protein and the carboxy terminal region of 24K are shown in one-letter amino acid code (see Table 1). Nucleotide and amino acid residue numbers are given on the right.

Table 1. Predicted amino acid compositions (% by number) of the major capsid proteins of HSV-1, VZV and EBV

<table>
<thead>
<tr>
<th>Residue</th>
<th>HSV-1</th>
<th>VZV</th>
<th>EBV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ala (A)</td>
<td>13.8</td>
<td>9.0</td>
<td>9.6</td>
</tr>
<tr>
<td>Val (V)</td>
<td>8.7</td>
<td>6.4</td>
<td>8.6</td>
</tr>
<tr>
<td>Leu (L)</td>
<td>10.8</td>
<td>9.7</td>
<td>8.8</td>
</tr>
<tr>
<td>Ile (I)</td>
<td>1.9</td>
<td>5.7</td>
<td>4.5</td>
</tr>
<tr>
<td>Gly (G)</td>
<td>7.2</td>
<td>6.4</td>
<td>5.6</td>
</tr>
<tr>
<td>Pro (P)</td>
<td>6.9</td>
<td>6.4</td>
<td>5.2</td>
</tr>
<tr>
<td>Cys (C)</td>
<td>1.8</td>
<td>2.0</td>
<td>1.3</td>
</tr>
<tr>
<td>Met (M)</td>
<td>2.2</td>
<td>3.0</td>
<td>3.2</td>
</tr>
<tr>
<td>His (H)</td>
<td>3.3</td>
<td>3.4</td>
<td>3.4</td>
</tr>
<tr>
<td>Phe (F)</td>
<td>4.3</td>
<td>4.5</td>
<td>5.1</td>
</tr>
<tr>
<td>Tyr (Y)</td>
<td>3.1</td>
<td>3.4</td>
<td>4.1</td>
</tr>
<tr>
<td>Trp (W)</td>
<td>0.7</td>
<td>0.4</td>
<td>0.3</td>
</tr>
<tr>
<td>Ser (S)</td>
<td>3.6</td>
<td>4.9</td>
<td>6.3</td>
</tr>
<tr>
<td>Thr (T)</td>
<td>5.2</td>
<td>7.9</td>
<td>6.0</td>
</tr>
<tr>
<td>Lys (K)</td>
<td>1.4</td>
<td>1.5</td>
<td>3.0</td>
</tr>
<tr>
<td>Arg (R)</td>
<td>4.2</td>
<td>4.2</td>
<td>4.7</td>
</tr>
<tr>
<td>Glu (E)</td>
<td>4.4</td>
<td>6.0</td>
<td>4.9</td>
</tr>
<tr>
<td>Gln (G)</td>
<td>4.4</td>
<td>3.9</td>
<td>5.7</td>
</tr>
</tbody>
</table>

those codons which may specify one of several residues without detrimentally influencing protein function.

The amino acid compositions of the major capsid proteins of HSV-1, VZV and EBV are shown in Table 1. The G + C contents of the coding regions of the HSV-1 and VZV genes are 69% and 47% respectively. Although much of this substantial difference is accounted for by the third codon position, and thus has little effect on coding potential, it is apparent that amino acid residues specified by G + C-rich codons, particularly alanine, are more abundant in the HSV-1 protein, and those specified by A + T-rich codons, particularly isoleucine, are less abundant. The G + C content of the coding region of the EBV gene is 57%, and the EBV protein is similar in
In this paper, we have described the DNA sequence of the HSV-1 major capsid protein gene and identified the corresponding genes in the VZV and EBV genomes. The amino acid sequence of the HSV-1 major capsid protein predicted from the DNA sequence, and the identification of conserved, and therefore structurally important, regions will be essential in eventually obtaining a detailed description of herpesvirus capsid morphology and morphogenesis.

REFERENCES


(Received 10 June 1986)