Key words: HSV-1/DNA sequence/gene organization

The DNA Sequences of the Long Repeat Region and Adjoining Parts of the Long Unique Region in the Genome of Herpes Simplex Virus Type 1

By LISE J. PERRY† AND DUNCAN J. McGEOCH*

MRC Virology Unit, Institute of Virology, University of Glasgow, Church Street, Glasgow G11 5JR, U.K.

(Accepted 9 August 1988)

SUMMARY

We have determined the DNA sequence of the long repeat region (RL) in the genome of herpes simplex virus type 1 (HSV-1) strain 17, as 9215 bp of composition 71.6% G + C. In addition, the sequences of parts of the long unique region (UL) adjacent to the terminal (TRL) and internal (IRL) copies of RL were determined (2611 and 3836 bp, respectively). Gene organization in these regions of UL was deduced from the sequences and other available data. It was proposed that the region of UL sequenced, adjacent to TRL, contains three complete genes, none with significant previous characterization, and that the region of UL adjacent to IRL also contains three genes, one encoding the immediate early protein IE63. The RL sequence contains one well characterized gene, for the protein IE110, whose organization we have described previously. Between the downstream end of the IE110 gene and UL there is a 3500 bp segment of RL in which we did not find convincing protein-coding sequences, and which thus remains of obscure functionality. Upstream of the IE110 gene is a region previously proposed by others to contain a gene. However, our sequence data are not compatible with their interpretation. We do consider it possible that the region is protein-coding, but regard gene organization here as still unresolved.

INTRODUCTION

The genome of herpes simplex virus type 1 (HSV-1) consists of a linear, double-stranded DNA of some 152000 bp (see McGeoch et al., 1988a). The DNA is regarded as consisting of two covalently linked segments, termed the long (L) and short (S) regions. Each of these contains a unique sequence (UL and US) which is flanked by a pair of oppositely oriented repeat sequences (RL and RS). The terminal and internal copies of RL are termed TRL and IRL; similarly for RS. This structure is shown in Fig. 1. The genome also possesses a terminal redundancy of 400 bp, termed the a sequence, and at the internal 'joint' between the L and S segments there are one or more further copies of the a sequence, oppositely oriented to the terminal copies. In a productively infected cell, an inversion process operates at the internal a sequence so that the progeny virion DNA population consists of a mixture of four sequence-orientation isomers which differ in the relative orientations of their L and S segments. One isomer is designated as the prototype for mapping purposes (Roizman, 1979).

We have been engaged in general sequence determination of the genome of HSV-1 strain 17, and have published sequences for US (McGeoch et al., 1985), RS (McGeoch et al., 1986) and UL (McGeoch et al., 1988a). In this paper we report the complete sequence for RL, of 9200 bp, together with adjacent parts of UL. This work has allowed precise definition of the boundaries between UL and the flanking copies of RL, and of the organization of genes in UL adjacent to those junctions. RL contains one well characterized gene, encoding the immediate early (IE) transcriptional activator IE110, whose sequence we have previously described (Perry et al.,...
Functionality of the remainder of RL, however, remains obscure. Recently, new interest in this region of the genome has arisen with the finding in latently infected neurons of RNA species transcribed from RL (Stevens et al., 1987; Rock et al., 1987; Spivack & Fraser, 1987), and we have examined the possible coding potential of these RNAs in light of the sequence data.

**METHODS**

The following plasmid-cloned fragments of HSV-1 strain 17 DNA were used for DNA sequence determination: BamHI b, BamHI e and BamHI k, each cloned in the BamHI site of pAT153 (from F. J. Rixon & A. J. Davison), HpaI s plus v cloned in the EcoRI/BamHI sites of pAT153 (from C. M. Preston & J. Russell), and XhoI c cloned in the XhoI site of pMK16 (from N. D. Stow).

DNA sequences were obtained by cloning randomly sheared fragments of each DNA into M13mp8 (Messing & Vieira, 1982; Deininger, 1983) and sequencing clones by the dideoxynucleotide chain termination method (Sanger et al., 1977, 1980; Biggin et al., 1983; McGeoch et al., 1985, 1986, 1988a). Sequences were assembled using the database system of Staden (1982), as implemented by P. Taylor for a PDPI1/44 computer under RSX11M.

Programs used for the examination and interpretation of DNA and protein sequences were as previously described (McGeoch et al., 1988a) except for a method we used to examine DNA sequences for characteristics associated with protein-coding function. This consisted of the following. Starting at an arbitrary position, the sequence under examination was divided into successive non-overlapping triplets. Running plots of the percentage G + C contents were then drawn separately for the sets of first, second and third position residues in the triplets, using in each case a window usually of 50 or 100 residues and for successive windows adding increments of 20 residues. All three plots were made on the same panel. Because of the characteristics of codon sets used in HSV-1 genes, this 'three-phase' plot typically gave distinct G + C contents of the three phases for HSV-1 coding sequences (see Results).

**RESULTS**

*The DNA sequences of RL and adjacent parts of UL*

This paper reports the DNA sequences of two separate parts of the genome of HSV-1 strain 17: first, a region of 13051 bp which includes the rightmost part of UL and the whole of IRUL, and second, a 3164 bp region which includes the leftmost part of UL and an adjacent part of TRUL.
Fig. 2. DNA sequence of the rightmost part of U₁ plus IR₁. The sequence is listed as the rightward 5′ to 3′ strand only, from the BamHI site representing the left extremity of BamHI b to the right extremity of the a′ sequence (see Fig. 1), and is numbered according to our complete genome sequence (McGeoch et al., 1988a). Proposed encoded amino acid sequences are given in single letter code, with rightward reading sequence above the corresponding DNA sequences and leftward sequences below. Designations of genes are given at the left of the first line containing each gene’s encoded amino acid sequence, regardless of orientation. Mapped positions of transcripts are given for gene UL54 (Whitton et al., 1988a). Proposed encoded amino acid sequences are given in single letter code, with rightward sequence (see Fig. 1), and is numbered according to our complete genome sequence (McGeoch et al., 1988a). *

These locations are outlined in Fig. 1, and the sequences are listed in Fig. 2 and 3. Because we have now determined the complete sequence of HSV-1 DNA, the numbering system used throughout this paper is that for the complete rightward 5′ to 3′ strand of the prototype orientation of the genome, as described by McGeoch et al. (1988a).

The sequence of the rightmost part of U₁ and all of IR₁ was obtained from four plasmid-cloned HSV-1 fragments, as indicated in Fig. 1. The major fragments used were BamHI b (residues 113322 to 126372), XhoI c (residue 123028 to the extremity at residue 126372 of the sequence shown in Fig. 2) and BamHI k (residue 123459 to the extremity at residue 126372). In addition, because of a problem with interpretation of part of the BamHI b sequence, to be detailed below, additional data were obtained with an independently cloned HpaI s plus v DNA. The sequence given in Fig. 2 overlaps at its right end with previously published sequences for HSV-1 strain 17 (Davison & Wilkie, 1981; McGeoch et al., 1986). The sequence shown in Fig. 2 runs from the BamHI recognition site at residue 113322 to the rightward end of the a′ sequence at 126372. Two instances of length heterogeneity were found between the overlapping plasmid-cloned fragments used to construct the sequence. First, starting at residue 114376, the BamHI b clone showed a homopolymer tract of seven C residues, while at the same location the HpaI s plus v clone showed an additional residue, with a tract of eight Cs. This variability occurs within the protein-coding region of gene UL54 and, as
BamHI b clone exhibited nine Gs. As described previously, this locus lies within an intron in the IE110 gene (Perry residues 117158 to 117338 are shown directly repeated copies of a 17 residue sequence, plus number was set to 10 for the figure).

possessed different copy numbers of the repeat family, from four to more than 20. We then found a subfragment of UL3

A second type of heterogeneity was also encountered in determining the sequence in Fig. 2. At

The two possible start codons for translation of gene UL2 are marked.

As an ad hoc resolution, the copy number was set to 10 for the figure.

Turning to TRi and the left extremity of U1, the sequence is shown in Fig. 3 of a SmaI/BamHI subfragment of BamHI e, running from the SmaI site at residue 8662 to the BamHI site at residue 11825.

Fig. 3. DNA sequence of the leftmost part of U1. The sequence is listed from the rightmost SmaI site in BamHI e to the right extremity of BamHI e, with conventions as described for Fig. 2. The two possible start codons for translation of gene UL2 are marked.
Table 1. Genes adjacent to ends of UL

<table>
<thead>
<tr>
<th>Gene</th>
<th>Translation position</th>
<th>No. of amino acids</th>
<th>Mr</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>UL1</td>
<td>9339-10011</td>
<td>224</td>
<td>24932</td>
<td>Possible N-terminal signal and/or transmembrane sequence (McGeoch et al., 1988a). Possibly a syn locus (Little &amp; Schaffer, 1981).</td>
</tr>
<tr>
<td>UL2</td>
<td>9886-10888</td>
<td>334</td>
<td>36326</td>
<td>Possible uracil-DNA glycosylase (Caradonna et al., 1987; also unpublished data of J. Mullaney, H. W. M. Moss &amp; D. J. McGeoch).</td>
</tr>
<tr>
<td></td>
<td>or 10156-10888</td>
<td>244</td>
<td>27327</td>
<td></td>
</tr>
<tr>
<td>UL3</td>
<td>10959-11664</td>
<td>235</td>
<td>25607</td>
<td>Possible near N-terminal signal and/or transmembrane sequence (McGeoch et al., 1988a).</td>
</tr>
<tr>
<td>UL45</td>
<td>113734-115270</td>
<td>512</td>
<td>55249</td>
<td>IE transcriptional modulating protein IE63 (Everett, 1986).</td>
</tr>
<tr>
<td>UL55</td>
<td>115496-116054</td>
<td>186</td>
<td>20491</td>
<td>Dispensable (MacLean &amp; Brown, 1987).</td>
</tr>
<tr>
<td>UL56</td>
<td>116923-116332</td>
<td>197</td>
<td>21182</td>
<td>Dispensable (MacLean &amp; Brown, 1987). Possibly a determinant of pathogenicity (Ben-Hur et al., 1988).</td>
</tr>
</tbody>
</table>

Our objective, having determined these sequences, was to interpret them in terms of locations of genes and their protein-coding regions. Only two previously well characterized genes were known to lie in the regions of the genome described here. These were the IE gene encoding the protein IE63 and located near the right end of UL, and the IE gene encoding the protein IE110 and located within RL (Whitton et al., 1983; Perry et al., 1986). We have previously described the organization and sequence of the latter (Perry et al., 1986) and, beyond listing the inferred amino acid sequence in Fig. 2, it is not further treated in the present paper.

We examined the sequences by looking for open reading frames (ORFs), and evaluating these by comparing codon usage with that of known HSV-1 genes with the program of Staden & McLachlan (1982), by examining triplet periodic trends in base composition and by looking for possible transcriptional control signals. In addition, comparisons were made with amino acid sequences encoded by proposed genes in the completely sequenced genomes of the related alphaherpesvirus varicella-zoster virus (VZV; Davison & Scott, 1986) and the gammaherpesvirus Epstein–Barr virus (EBV; Baer et al., 1984). The following sections describe in turn our evaluations of the leftmost part of UL, the rightmost part of UL, and RL.

**Gene organization in the leftmost part of UL**

As shown in Fig. 1 and 3, we propose that the region sequenced at the left end of UL contains three complete genes, which are all rightward oriented. We designated these UL1, UL2 and UL3, and data on them are summarized in Table 1. These genes have had essentially no previous characterization, and no published information on transcript mapping is available for this region. We propose that transcription of the UL3 gene probably terminates downstream of the polyadenylation-associated consensus AATAAA at 11717. Transcription of UL1 and UL2 could terminate next to the AATAAA sequence at 10945; this, however, is situated atypically
Fig. 4. Comparisons of predicted amino acid sequences encoded by HSV-1, VZV and EBV. Optimal alignments produced by the program HOMOL (Taylor, 1984) are shown for amino acid sequences of UL1, UL2, UL3, UL54 and UL55 with the corresponding VZV sequences (from Davison & Scott, 1986). UL2 is also compared with its EBV counterpart (from Baer et al., 1984). Pairs of identical residues are marked with asterisks. In (a) to (d), only the most similar sections of the sequences are shown, whereas in (e) the whole of both the sequences are given.

close to the UL3 ORF. The alternative would be that UL1 and UL2 transcripts were 3'-coterminal with the UL3 transcript.

A weak point in identifying functional ORFs in the absence of other data is that it may not be possible to define satisfactorily the real translational start. For UL2, translation is shown as starting at the first candidate ATG initiation codon (marked 1 in Fig. 3) to give an overlap with the UL1 coding sequence. An alternative possibility is that translation might start at the next ATG (marked 2), and this would abolish the overlap.

The amino acid sequences of UL1, UL2 and UL3 are homologous to those of VZV genes 60, 59 and 58 respectively, which occupy corresponding positions in the genome of that virus (Davison & Scott, 1986). These relationships are illustrated in Fig. 4. Only UL2 has a counterpart with detectable amino acid similarity in the genome of the widely diverged EBV: this is the ORF BKRF3 (Baer et al., 1984), which is quite notably well conserved (Fig. 4). From their positions in the EBV genome, BKRF2 and BKRF4 are possible counterparts of UL1 and UL2 respectively, but do not exhibit any quantifiable sequence similarity.

Both the UL1 and UL3 amino acid sequences have pronounced hydrophobic regions near their N termini, which could represent signal sequences for translation on membrane-bound ribosomes or could act as transmembrane anchor sequences (see McGeoch et al., 1988a). The VZV counterparts also exhibit these features. The sequences do not possess other features of obvious note.
Not much information is available regarding functions of genes in this region of the HSV-1 genome. Little & Schaffer (1981) mapped a syn locus (that is, a mutation causing a syncytial plaque morphology) to a region bounded by residues 6071 and 9620 in our numbering system. UL1 is the only presently recognized gene which lies even partly within these limits. The suggestion above that the UL1 protein might be membrane-associated is of interest in this regard since other syn loci lie in genes either known or predicted to encode membrane-inserted proteins (Bzik et al., 1984; Debroy et al., 1985; Pogue-Geile & Spear, 1987). Caradonna et al. (1987) have shown that an elevated uracil-DNA glycosylase activity found in HSV-infected cells appears to be encoded by a region near the left end of UL, which we now see most likely corresponds to gene UL2. Recently, we have obtained genetic data which strongly support this assignment (unpublished work of J. Mullaney, H. W. M. Moss & D. J. McGeoch).

Gene organization in the rightmost part of UL

We propose that the region sequenced at the right extremity of UL contains three complete genes (see Fig. 1 and 2 and Table 1). Following our determination of the whole sequence of UL and interpretation of its gene content, we have designated these rightmost genes as UL54, UL55 and UL56 (McGeoch et al., 1988a). UL54 is the previously recognized IE gene which encodes the protein IE63 (Watson et al., 1979; Whitton et al., 1983). IE63 is a protein essential for virus growth, and acts as a modulator of the transcription of certain later HSV-1 genes (Sacks et al., 1985; Everett, 1986). It has an estimated $M_r$ of 63000, is produced immediately after infection, is localized to the nucleus, and is phosphorylated (Pereira et al., 1977; Marsden et al., 1978; Hay & Hay, 1980).

The 5' and 3' termini of IE63 mRNA have been mapped, and it has been shown that the transcript is not spliced (Whitton et al., 1983). Transcription starts at residue 113596 and terminates at 115295 downstream of the AATAAA sequence at 115277. We consider that the IE63 coding sequence consists of the ORF from ATG at 113734 to TAG at 115270, giving a polypeptide of 512 amino acids and $M_r$ 55249. As mentioned above, this ORF was disrupted in the first clone of the region which we sequenced. We consider that the results available on this gene and its product constitute reasonable, circumstantial evidence for our choice of the eight residue version of the C tract starting at 114376 being correct. In addition, we have recently determined the sequence of the corresponding gene in herpes simplex virus type 2 (HSV-2) and this is in agreement with our assignment (unpublished data of D. J. McGeoch, C. Cunningham & A. Dolan).

As previously outlined by Davison & Taylor (1987), using our IE63 sequence, the amino acid sequence of IE63 is homologous to that encoded by VZV gene 4, and both of these exhibit a distant relationship with the EBV BMLF1 sequence. The BMLF1 protein also acts as a transcriptional trans-activator (Lieberman et al., 1986; Wong & Levine, 1986). The HSV-1/VZV homology is illustrated in Fig. 4. This similarity occurs only from about residue 300 of IE63 to its C terminus. Some intrinsic aspects of the IE63 amino acid sequence show a similar division into two regions with distinct properties. First, the sequence is particularly hydrophilic in its first part, to about residue 210. Secondly, the content of alanine, glycine, proline and arginine residues is very high in a region centred on residues 125 to 242. These are amino acids which can be specified by codons containing only G and C residues, and their high frequency reflects the high G + C content of the corresponding part of the gene. This lack of homogeneity in properties of the IE63 amino acid sequence and in its relationship to the VZV counterpart is similar to the situation previously described for another transcriptional regulatory protein of HSV-1, IE175 (McGeoch et al., 1986).

Two specific sequence elements found in IE63 appear worth noting. First, at residues 37 to 49 there is a serine-rich tract followed by several acidic residues. This resembles a region in IE175 which has been shown to be a target for phosphorylation (DeLuca & Schaffer, 1988). Secondly, there is a set of four histidine and four cysteine residues distributed through the final 52 residues of the sequence and we suggest that these might form a 'zinc-finger' DNA-binding structure (Berg, 1986).
As shown in Fig. 1 and 2 and in Table 1, we propose that the two rightmost genes in UL are UL55, oriented rightward, and UL56, oriented leftward. Nothing is known definitely about the functions of genes UL55 and UL56. Neither is essential for virus replication in tissue culture, since an HSV-1 variant with this region deleted is viable (MacLean & Brown, 1987). No published transcript mapping data are available. Transcription of UL55 is envisaged as terminating downstream of the AATAAA sequence at 116098, and of UL56 downstream of one or both of the closely spaced AATAAA sequences on the leftward 5' to 3' strand at 116201 and 116205. The amino acid sequence of UL55 is homologous to that encoded by VZV gene 3 (Fig. 4). UL56 has no detectably similar VZV counterpart, but two apparently unrelated genes occupy the corresponding position in the VZV genome. There are no detectable counterparts of UL55 and UL56 in EBV. Recently, we have detected counterparts to both of these ORFs in HSV-2 DNA sequences (D. J. McGeoch, C. Cunningham & A. Dolan, unpublished data).

**Organization of the long repeat region**

The complete R_L sequence, including one copy of the a sequence, contains 9215 residues as listed in Fig. 2. The base composition is 71.6% G + C, substantially higher than for the adjacent parts of U_L discussed in this paper (65.7% and 61.8% at the left and right extremities of U_L, respectively). Excluding the a sequence, there are within R_L six sets of contiguous direct repeats (tandem reiterations) as marked in Fig. 2 and 5. In addition to these, other imperfectly repeated
sequences and also homopolymeric runs occur. This tendency towards 'simple' sequence is particularly marked in the region comprising around 700 bp of $R_L$ adjacent to $U_L$.

The only well characterized gene in $R_L$ is that for $IE_{110}$, whose sequence and organization we described previously (Perry et al., 1986). Chou & Roizman (1986) and Ackermann et al. (1986) have presented evidence that there exists another gene upstream of the $IE_{110}$ gene and in the same orientation, and their analysis is compared below with our data. Recently, there have been a number of reports concerning transcripts of HSV-1 found in latently infected neurons (Stevens et al., 1987; Rock et al., 1987; Spivack & Fraser, 1987). These are copied from $R_L$, from the region downstream of the $IE_{110}$ gene and also overlapping it, as shown in Fig. 2 and 5. The major purpose of this section is to report on the examination, in light of our sequence data, of the possible functional organization of the regions of $R_L$ flanking the $IE_{110}$ gene. To state the outcome first: we have not been able to construct to our satisfaction a model of the type obtained for $U_S$ and $U_L$, where the sequence was largely filled with proposed functional ORFs.

Other ORFs of significant size do exist in $R_L$, apart from the coding sequences of the $IE_{110}$ gene (see Fig. 5). However, they do not, by our tests, exhibit characteristics expected for genuine polypeptide-coding sequences. There are also five possible polyadenylation-associated sequences AATAAA and ATTAAA in $R_L$ which are not at present assigned to known mRNAs (see Fig. 5). Two pairs of these are each close: at residues 118004 and 118020 rightward oriented, and at 120688 and 120728 leftward oriented.

We have found the three-phase plot of DNA base composition (described in Methods) to be a particularly useful method for displaying the coding characteristics of HSV-1 DNA sequences, and we use this technique here to illustrate our evaluations. Fig. 6 shows a standard three-phase plot for the HSV-1 DNA sequences reported in this paper, together with further sequences adjacent to the sections of $U_L$ reported, and the whole of $R_S$. The plot shows several features. Within the coding region of a given gene, the three phases generally exhibit widely separated G + C contents, corresponding to the previously well known observation that in HSV-1 genes the set of third positions in codons has the highest G + C content and the set of second positions the lowest G + C content. These features can be readily rationalized in terms of the nature of the genetic code. The existence of this pattern in an uncharacterized sequence can then allow assignment of the direction and frame of amino acid coding. Within this marked overall trend, a number of local divergences can be seen, for instance in gene $UL54$ and in the $IE_{110}$ gene. Gene $UL56$ presents a clear exception to the pattern. We consider that these 'anomalies' represent either particular requirements of amino acid composition, or perhaps some undefined processes in the molecular evolution of HSV-1 DNA. Turning to sequences generally regarded as non-coding, we see that the pronounced separation of phases is not present. This applies to the control regions upstream of gene $UL54$ and the $IE_{110}$ and $IE_{175}$ genes. Again, local 'anomalies' do exist. Most marked is a region in the $a'$ sequence, where the pronounced separation of one phase from the other two results from the occurrence of a 12 bp reiterated sequence.

The features just described for coding and non-coding sequences are found throughout $U_L$ and in $U_S$, with few departures (data not shown). Thus, typical coding and non-coding sequences in the HSV-1 genome can be readily distinguished by this method. When the region of $R_L$ between $U_L$ and the $IE_{110}$ gene was examined, we found that there were no large separations of phase base compositions. By this test, we concluded that this region does not contain large, functional ORFs with codon contents characteristic of other parts of the genome. We could not, however, exclude the existence of coding ORFs which have an atypical codon content or are small.

Recently, Wagner et al. (1988) published the sequence of part of the $R_L$ component of HSV-1 strain KOS, covering the region giving rise in latently infected neurons to the latency-associated transcript (LAT), and corresponding to residues 119281 to 121566 of our sequence (Fig. 2). Since the strain 17 $R_L$ sequence and the KOS strain sequence were exchanged before publication, differences between them most probably do not reflect trivial errors. There are 10 addition/deletion differences between the two sequences, all downstream of the $IE_{110}$ gene sequences, and six of these would be frameshifting (see Fig. 5).

Chou & Roizman (1986) have published sequence data for part of $R_L$ of HSV-1 strain F, between the $IE_{110}$ gene and the $a$ sequence, and have presented evidence that this region
Fig. 6. Three-phase plots of HSV-1 sequences. Each panel represents 11400 residues from the HSV-1 DNA sequence. The G+C content of each phase is then plotted with a window of 100 (that is, corresponding to 300 residues in the HSV-1 sequence). Below each box is shown the part of the genome concerned, with arrows representing genes' ORFs (not transcripts). (a) Sequence shown in Fig. 3, with genes UL1 to UL3, extended to part way through UL8 (McGeoch et al., 1988a, b). (b, c) Continuous sequence, starting at genome residue 110001, within gene UL52, through the sequence shown in Fig. 2, and including the whole of IRs.
DNA sequence of HSV-1 RL

contains a gene oriented in the same sense as the IE110 gene. The proposed coding region of this gene was a 358 codon ORF. Ackermann et al. (1986) raised antisera against a repeating tripeptide sequence in the ORF, and showed that one such serum reacted with a protein (termed ICP 34.5) in extracts of infected cells. We have compared the strain F sequence with our sequence for the region (residues downstream of 124258 in Fig. 2), and have found a number of differences within the bounds of the ORF. These include a difference in copy number of the 9 bp tandem repeat family (10 complete copies in the strain F sequence, five in our sequence). There are also 20 points in our sequence at which an addition or deletion change breaks the proposed coding ORF (see Fig. 6). We re-examined our sequence data in the light of these results and did not find errors. Accordingly, we think that the differences most probably represent either divergence between the two virus strains, or sequencing errors in strain F. In either case, we are led to the conclusion that it is unlikely that the whole of the ORF in question is genuinely protein-coding.

If translation were to start at the ATG codon proposed for the strain F sequence (corresponding to position 125904 on the complementary strand of our sequence), then the reading frame in strain 17 changes from the strain F version after 20 codons, and terminates after 93 codons before the region used by Ackermann et al. (1986) to synthesize an antigenic peptide. If translation started at the next downstream ATG, at position 125858, then no stop codon is reached until 124793, which is downstream of the termination site for transcription proposed by Chou & Roizman, adjacent to ATTAAA at 124835. Translation in this strain 17 ORF is not coincident with the strain F ORF through the tandem repeat set.

Examination of the region by the three-phase plot (Fig. 6) showed that between residues 125170 and 125620 there is pronounced separation of the phases, which might indicate the presence of a functional ORF oriented in the same direction as the IE110 gene. This includes the 9 bp set of tandem reiterations mentioned above, and the ORF indicated by the three-phase analysis is coincident, across the reiterations, with the proposed ICP 34.5 ORF. We thus consider it possible that this region does contain protein-coding sequences, but the detail of functional organization remains obscure.

DISCUSSION

Comparison of the sequences from the extremities of the UL revealed that the junctions between UL and the two copies of RL are defined by a family of reiterated sequences. We think that within UL the two genes adjacent to the UL/RL junctions are transcribed away from the proximal copy of RL. This layout is similar to that seen in the short region, where at each end of US there is a gene transcribed from RS into US (Murchie & McGeoch, 1982). There exists in RS also a set of tandem reiterations close to the US/RS junction (but in this case not in immediate apposition to the junction). Transcription of those US genes starts within the repeat element, with the reiterated sequence family lying within an intron in each case (Rixon & Clements, 1982).

We have not yet mapped the 5' termini of transcripts for genes UL1 and UL56, and in general we have not attempted to predict transcription initiation sites from DNA sequence data. However, it is interesting to note that the UL1 and UL56 ORFs both possess an excellent candidate for a TATA box element, in UL close to the upstream UL/RL junction. For UL1, this is located at residues 9216 to 9223, and for UL56 at 117107 to 117113. We therefore speculate that transcription of these genes may initiate within UL, downstream of these sites. In strain HFEM of HSV-1 there is a deletion whose endpoints correspond to residues 117087 and 120640 in our sequence, and this been shown to confer an apathogenic phenotype on virus inoculated into mice intraperitoneally (Koch et al., 1987; Ben-Hur et al., 1988). The only functional elements which we can see the deletion removing are the presumed control sequences for gene UL56.

Our present understanding of the functional organization of the 3500 bp, UL-proximal section of RL remains tenuous. Parts of the sequence are relatively 'simple', and may well not have specific functions. From the tests we applied, it is evident that the sequence does not exhibit the characteristics of protein-coding sequences seen elsewhere in the genome. However, it is much
less feasible to establish firmly the negative conclusion, that the sequence is definitely not protein-coding, than it has been to establish the opposite for other regions of the genome. Instead of definitive evidence, we have an accumulation of indications, each of which by itself is relatively insecure. Thus, the absence of typical codon usage characteristics and of a protein-coding type of three-phase plot could represent instead an atypical history of molecular evolution, or the presence of multiple introns in a coding sequence. The numbers of addition/deletion differences registered between virus strains and clones support the view that the sequence is not protein-coding in the localities of these changes. However, these indicators are somewhat vitiated by the counter-example provided by our observation of a frameshifting mutation in a plasmid copy of the coding sequences of gene UL54.

In conclusion, we have not detected convincingly functional ORFs in this section of RL, but we cannot definitively exclude the existence of coding sequences. Our present opinion is that it is unlikely that LAT encodes a protein, and we do not see indications of any other genes. This leaves unresolved the significance of transcription of RL in latently infected neurons. Interestingly, there exists in the genome of VZV at the position corresponding to this part of the HSV-1 genome (between VZV genes 60 and 61; Davison & Scott, 1986) a 1400 bp sequence also without any assigned function. Turning to the section of RL upstream of the IE110 gene, we again have not reached any clear conclusion. From the number of frameshifting differences between our sequence and that of Chou & Roizman (1986), we consider that the organization proposed by those workers for a gene encoding ICP 34.5 is not likely to be correct. Our data could be consistent with there being protein-coding sequences in this locality, but we do not consider the analysis satisfactory enough to propose any definite gene layout.

Finally, we think that further experimental approaches will be needed to resolve the functional potential of RL. These could include genetic methods of analysis, and also cloning and characterization of cDNA clones. Our laboratory has undertaken determination of the sequence of HSV-2 RL, in the hope that comparisons of the sequences from the two serotypes will be informative.

We acknowledge the expert assistance of A. Dolan and D. McNab. We owe thanks to A. J. Davison, C. M. Preston, F. J. Rixon, J. Russell and N. D. Stow for supply of plasmids. We wish to thank L. J. E. Kattenhorn for help in preparing the paper, and A. J. Davison for a critical review. During the course of the work described, L. J. P. was in receipt of a grant from the Shell Trust for Higher Education.

Note added in proof. Since this paper was submitted for publication, the sequence of the uracil-DNA glycosylase gene of E. coli has been published (Varshney, U., Hutcheon, T. & van de Sande, J. H., 1988; Journal of Biological Chemistry 263, 7776-7784). The predicted amino acid sequence of the E. coli enzyme shows considerable similarity to that for HSV-1 UL2, confirming that UL2 does encode uracil-DNA glycosylase.

REFERENCES


KOC¥, H.-G., ROSEN, A., ERNST, F., BECKER, Y. & DARAI, G. (1987). Determination of the nucleotide sequence flanking the deletion (0.762 to 0.789 map units) in the genome of an intraperitoneally avirulent HSV-1 strain HFEM. *Virus Research* 7, 105–115.


(Received 20 May 1988)