Volume 14 Number 4 1986 Nucleic Acids Research

Complete DNA sequence of the short repeat region in the genome of herpes simplex virus type 1

Duncan J.McGeoch[1,2], Aidan Dolan[2], Sally Donald[2] and Dieter H.K.Brauer[3]

Institute of Virology, Church Street, Glasgow G11 5JR, UK

Received 10 January 1986; Accepted 4 February 1986

ABSTRACT

We report the complete DNA sequence of the short repeat region in the genome of herpes simplex virus type 1, as 6633 base pairs of composition 79.5% G+C. This contains immediate early gene 3, encoding the IE175 protein, an important transcriptional activator of later virus genes. The IE175 coding region was identified as a 3894 base sequence of 81.5% G+C DNA. The base composition of this gene is thus the most extreme yet determined, and the IE175 predicted amino acid composition is correspondingly biased, most notably with an alanine content of 20.9%. Functionally important regions of the IE175 polypeptide were tentatively identified by comparison with the sequence of the homologous protein from varicella-zoster virus and from locations of ts mutations, and were correlated with properties of the amino acid sequence. Aspects of the evolution of such an extreme composition DNA sequence were discussed.

INTRODUCTION

The DNA of herpes simplex virus type 1 (HSV-1) is a linear double stranded molecule of some 155,000 base pairs, regarded as comprising two covalently joined segments, termed the long and short regions (1,2). Each of these consists of an unique sequence flanked by inverted repeat sequences, as shown in Figure 1. This paper is concerned with the DNA sequence of the short repeat element (R_S) of HSV-1, which is 6.6 kb in size.

Only one gene lies completely within R_S. This is immediate early (IE) gene 3, which encodes the transcriptional activator protein IE175 (also called ICP 4) (3). IE175 is expressed immediately after infection by HSV-1, and is a large phosphoprotein, which accumulates in the nucleus of the infected cell (4-7). Experiments with mutants of HSV-1 carrying ts IE175 genes have shown that the presence of active IE175 is necessary

for the virus transcriptional programme to proceed beyond the IE phase, that IE175 acts at the transcriptional regulatory level and is required throughout the infectious cycle, and that the protein might also be involved in switching off transcription of the IE genes (8-10). More recently, it has been found that IE175, when expressed by a recombinant plasmid transfected into culture cells, can also activate non herpesvirus genes, including adenovirus genes and the rabbit β-globin gene (11,12). Thus, it is possible that IE175 may act through some rather general, cellular transcription regulatory system, and interest in the protein and in how it acts is accordingly broadened.

The mechanism of action of IE175 is at present unclear. It is not resolved whether it acts directly (that is, by interacting with a sequence on the target DNA near the site of transcription initiation) or indirectly (for instance, by interacting with a cellular protein which ultimately influences trancriptional activity) (13-15). Recently, our view of transcriptional activation by HSV-1 IE proteins has been complicated by two types of result. First, it has been shown that another IE protein, IE110 (or ICP 0), can also activate transcription of delayed early genes (16,17). Secondly, studies with cells which constitutively express IE175 have demonstrated that IE175 does not efficiently activate all delayed early genes when these are carried in the viral genome (18). The specificity of IE175 action implied by the latter work is apparently at odds with the rather general effects on activation of non herpesvirus genes.

HSV-1 DNA has an overall base composition of 67% G+C (1). For the R_S region, however, the G+C content is near 80%, both overall and in the IE175 coding sequence. This is by far the highest G+C content for any large sequence yet analysed, so far as we are aware. We think that the IE175 gene base composition lies near the attainable limit for protein coding DNA. As such, it constitutes a paradigm for the organization of genetic material subjected to this form of evolutionary development.

In this paper we present the complete DNA sequence of the R_S region of HSV-1, strain 17, and deduce the encoded amino acid sequence of IE175. The data are examined from two viewpoints. First, we attempt to evaluate functional importance of regions

within the predicted amino sequence of IE175 and, second, we propose possible evolutionary mechanisms which could have been involved in generating the extreme base composition. Apart from the IE175 gene transcribed region, R_S contains a number of previously identified functional entities, which are not further dealt with in this paper.

MATERIALS AND METHODS

(a) Recombinant plasmids. The following recombinant plasmids carrying HSV-1 strain 17 restriction fragments were used for sequence analysis: BamHI n and BamHI k cloned in the BamHI site of pAT153, obtained from F.J. Rixon and A.J. Davison respectively, and XhoI c cloned in the XhoI site of pMK16, obtained from N.D. Stow.

(b) DNA sequence analysis. Early experiments used chemical degradation sequencing (19). However, most of the analysis was performed by the M13/chain terminator method (20). Sets of M13 clones were generated by isolating an appropriate, large, plasmid-carried DNA restriction fragment of HSV-1 DNA, shearing by sonication and cloning into the SmaI site of M13mp8 (21,22). In addition, some M13 clones carrying restriction fragments were used for specific target regions. Sets of radioactive fragments for sequence analysis were prepared using [α^{32}P]-dATP as label. Products were fractionated by electrophoresis in 6% polyacrylamide, buffer gradient gels as described by Biggin et al. (23), except that gels contained additional urea, to 9 M. Gels were covalently bonded to one glass plate (24), and were fixed and dried before autoradiography.

Because of the extreme base composition of HSV-1 R_S, the incidence and the severity of sequencing artefacts were notably high. 6% polyacrylamide gels containing 7 M urea and 50% formamide were used with limited success to resolve artefacts. Sequencing reactions with dITP substituted for dGTP, and with polymerase reactions carried out at 37°C, were also used. The most successful method for resolving artefacts was use of a 6% polyacrylamide gel containing 9 M urea, 90 mM Tris-borate pH 8.3, 2 mM EDTA, with a jacket through which water at 80-85°C was pumped. This device was employed extensively.

(c) Computing. Computing was performed with a DEC PDP 11/44 under RSX11M, as previously described (25,26).

RESULTS

(a) The DNA sequence of R_S. Figure 1 shows the HSV-1 prototype genome structure, with nomenclature of regions, and an expansion of the internal copy of R_S, designated IR_S. The sequence analysis reported here used three plasmid cloned fragments of HSV-1 strain 17 DNA, XhoI c, BamHI k and BamHI n, whose locations are indicated in Figure 1. We have previously described the sequence of BamHI y (wholly contained in R_S) and of BamHI x (the TR_S counterpart of BamHI n); the upstream end of the IE175 gene is contained in these sequences (27). The downstream end of the IE175 gene sequence has been determined by Davison and Wilkie (28). Figure 2 lists our DNA sequence of the whole IR_S region, starting with the nucleotide adjacent to U_S and ending after the a' sequence. The 5' terminus of IE175 mRNA is at residue 1176 and the 3' terminus is at residue 5435; the mRNA is unspliced

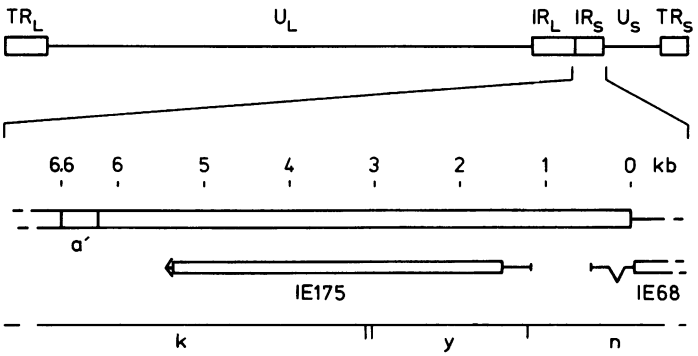


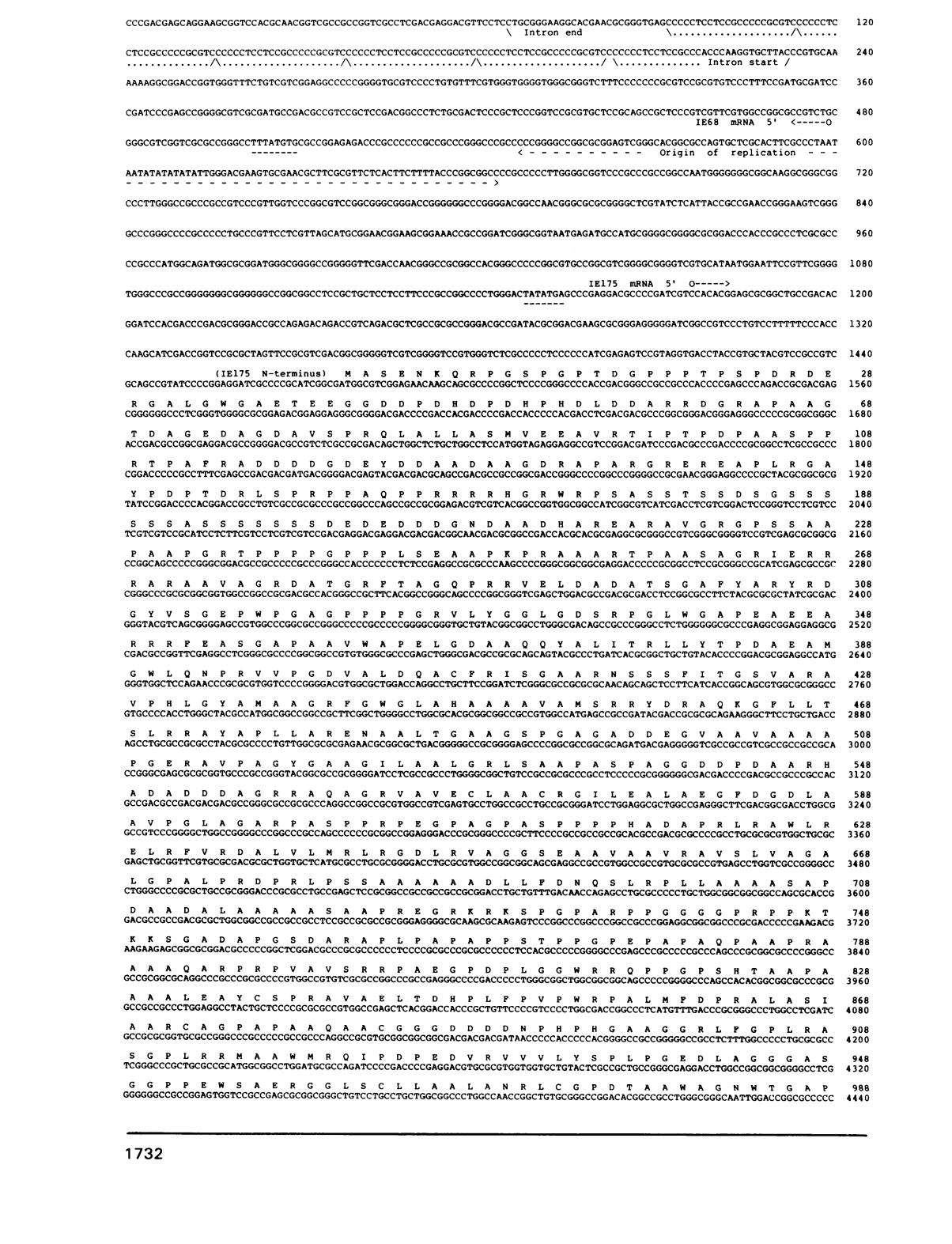
Figure 1. The genome of HSV-1. The upper part of the figure depicts the genome of HSV-1 (2). Unique sequences (U_L and U_S) are shown as solid lines and major repeat elements as boxes (TR_L and IR_L; TR_S and IR_S). The lower part of the figure expands the 6.6 kbp IR_S region, with numbering as in the sequence listing (Figure 2). The location of IE175 mRNA is indicated (5' to 3': right to left), with the predicted coding region as an open box. Part of the nearby IE68 mRNA is also shown. The location of the 400 bp a' sequence is shown. The a' sequence is an opposite orientation copy of the a sequences present as direct repeats at the genome termini. Locations of BamHI fragments used in sequence analysis are indicated on the bottom line. The XhoI c fragment, also used, includes the whole expanded region.

(3). RNA polymerase II transcription initiation and termination signals are present at the terminal regions. The locations of other entities in the sequence are also noted in Figure 2.

The 6633 base pair sequence of R_S has an overall base composition of 79.5% G+C. Sequence determination for such DNA is particularly demanding, since the G+C rich sequences are prone to various sequencing gel artefacts. Another problem presents itself, in interpretation: since nonsense codons are A+T rich, the incidence in potential polypeptide coding sequences of out-of-frame stop codons is very low, and so the correct reading frame is not revealed by multiple blockage of other frames. In fact, the whole 3894 residue protein coding region of the IE175 gene contains just 11 out-of-frame nonsense codons in the appropriate orientation. We employed the codon usage evaluation program of Staden and McLachlan (36) to help authenticate the correct reading frame, and concluded that translation of IE175 commences with the first ATG in the transcribed region, at 1477, and terminates with TAA at 5371. This reading frame is strongly supported by codon usage comparison, using the US3 gene of HSV-1 as reference (25), except where it encodes amino acid residues 180 to 210 (data not shown). The latter region contains an inverted repeat and encodes multiple Ser residues, characteristics which evidently defeat the algorithm. The IE175 protein so defined contains 1298 amino acids and has a molecular weight of 132,835. An antiserum against an oligopeptide comprising residues 65 to 76 of this predicted amino acid sequence can precipitate native IE175 (37).

(b) Base composition of R_S and amino acid composition of IE175. The IE175 gene coding region presents the extraordinary base composition of 81.5% G+C. This impinges on our thinking about the gene in two ways. The first concerns the derivation of such a genetic entity: what events in evolution might have produced this result? This will be addressed briefly in the Discussion. The other aspect of the phenomenon, more immediately relevant to our evaluation of the nature of the IE175 protein, is the relationship between the extreme DNA base composition and the amino acid composition of encoded polypeptide.

Study of the DNA sequence shows clearly that requirements


Intron end
Intron start
IE68 mRNA 5'
Origin of replication
IE175 mRNA 5'
(IE175 N-terminus)

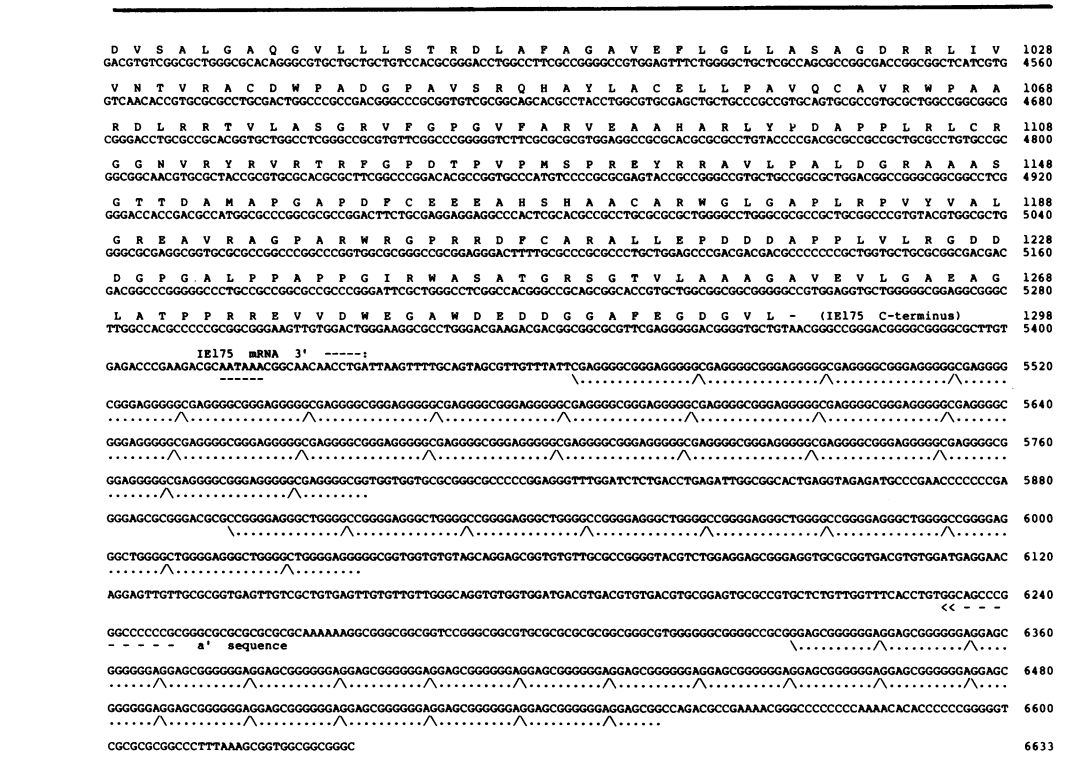


Figure 2. DNA sequence of HSV-1 short region. The DNA sequence of IR_S is listed as the 5' to 3' strand only, starting with the residue adjacent to U_S (27). Positions of the termini of IE175 mRNA are indicated at residues 1176 and 5435 (3), and the predicted IE175 amino acid sequence is shown. Other functional entities in IR_S include, upstream of the IE175 coding region: the 5' portion of the IE68 gene, with mRNA 5' terminus at residue 479 on the opposite strand, and intron between 232 and 65 (27,29); an origin of DNA replication (30); promoters for IE68 and IE175 genes (proposed TATA boxes underlined) and upstream activator regions for these genes (not marked) (31-33). Downstream of the IE175 gene lies the a' sequence, which contains sequences involved in site specific inversion and in packaging of nascent DNA (28,34,35). Four sets of tandemly repeated sequences are underlined as: \...../.

of amino acid coding are not a primary cause of the extreme base composition of the DNA. Thus, both coding and non coding DNA regions have base compositions near 80% G+C. In protein coding DNA, the codon set available at this base composition is sufficiently biased that restrictions on the possible amino acid composition of encoded protein must exist. Within the protein

Table 1. Codon catalogue of the IE175 gene

TTT Phe	6	TCT Ser	1	TAT Tyr	2	TGT Cys	0
TTC Phe	16	TCC Ser	24	TAC Tyr	17	TGC Cys	15
TTA Leu	0	TCA Ser	1	TAA ---	1	TGA ---	0
TTG Leu	2	TCG Ser	29	TAG ---	0	TGG Trp	21
CTT Leu	0	CCT Pro	0	CAT His	0	CGT Arg	2
CTC Leu	13	CCC Pro	84	CAC His	17	CGC Arg	78
CTA Leu	1	CCA Pro	3	CAA Gln	0	CGA Arg	6
CTG Leu	84	CCG Pro	74	CAG Gln	19	CGG Arg	49
ATT Ile	1	ACT Thr	0	AAT Asn	1	AGT Ser	1
ATC Ile	10	ACC Thr	15	AAC Asn	10	AGC Ser	20
ATA Ile	0	ACA Thr	0	AAA Lys	0	AGA Arg	1
ATG Met	11	ACG Thr	20	AAG Lys	8	AGG Arg	3
GTT Val	1	GCT Ala	2	GAT Asp	3	GGT Gly	1
GTC Val	14	GCC Ala	161	GAC Asp	97	GGC Gly	83
GTA Val	1	GCA Ala	8	GAA Glu	4	GGA Gly	2
GTG Val	47	GCG Ala	100	GAG Glu	53	GGG Gly	56

coding region, the separate base compositions for each codon position are: first position, 81.9% G+C; second, 66.2%; and third, 96.3%. This pronounced triplet periodicity does not exist in non coding DNA regions. This distribution is consistent with retention of a near maximum variety of encoded amino acids at the given, extreme base composition. The extreme biases of the third codon positions are strikingly demonstrated by the codon usage catalogue of Table 1. These characteristics of the sequence are thoroughly in keeping with the hypothesis that the observed DNA base composition is the result of an evolutionary force acting directly on the DNA, rather than indirectly through amino acid sequence requirements, and that perpetuation of changes produced by this force is secondarily influenced by other, functional constraints, such as the nature of codons specified. For our present purpose of considering the IE175 amino acid sequence, we use this model as a framework.

The deduced amino acid composition of IE175 is listed in Table 2. The amino acid sequence shows a near equality of acidic and basic residues. The most common amino acid is Ala, at 20.9% of the total, and the four most common species are Ala, Pro, Gly and Arg, together comprising 54.9% of all amino acids. The least common amino acids are Lys, Asn, Ile and Met. The most common four species are those which possess codons containing only G and

Table 2. Amino acid composition of IE175

Residue	Number	%	Residue	Number	%
Ala A	271	20.9	Leu L	100	7.7
Arg R	139	10.7	Lys K	8	0.6
Asn N	11	0.8	Met M	11	0.8
Asp D	100	7.7	Phe F	22	1.7
Cys C	15	1.2	Pro P	161	12.4
Gln Q	19	1.5	Ser S	76	5.9
Glu E	57	4.4	Thr T	35	2.7
Gly G	142	10.9	Trp W	21	1.6
His H	17	1.3	Tyr Y	19	1.5
Ile I	11	0.8	Val V	63	4.9

C residues; the codon usage list (Table 1) shows that such codons are employed almost to the exclusion of other codons for these amino acids. The high levels of these four amino acids, and of Ala in particular, are thus the major biases of amino acid composition associated with the extreme DNA base composition. Thus, we think that functioning of the protein tolerates 20.9% Ala rather than requires this level. The particularly high level of Ala, compared with Pro, Gly and Arg, may reflect the relatively innocuous nature of this residue as a substitute for other species. We correlate the low levels of Lys, Asn, Ile and Met with the fact that these are four of the six species for which all codons contain at least two A or T residues. We suppose that the alternative strongly basic amino acid, Arg, is generally used in preference to Lys. These arguments gain force when IE175 is compared with its varicella-zoster virus (VZV) homologue, in the next section.

(c) Homology between IE175 and VZV 140 proteins. Known protein sequences were searched for any with homology to IE175, but the only homologue found was the VZV 140 protein, which is the VZV equivalent of IE175 (38,41). HSV-1 and VZV are both members of the Alphaherpesvirinae sub-family (42). Like IE175, VZV 140 can transactivate foreign plasmid-borne genes (12). The two amino sequences are shown aligned in Figure 3. Extent of homology varies widely along the sequences. Based on this, we have divided each sequence into five regions.

Region 1 comprises residues 1 to 314 of IE175. The two proteins are not homologous in this region, except for a possible, limited N terminal homology and a region (186-204 of

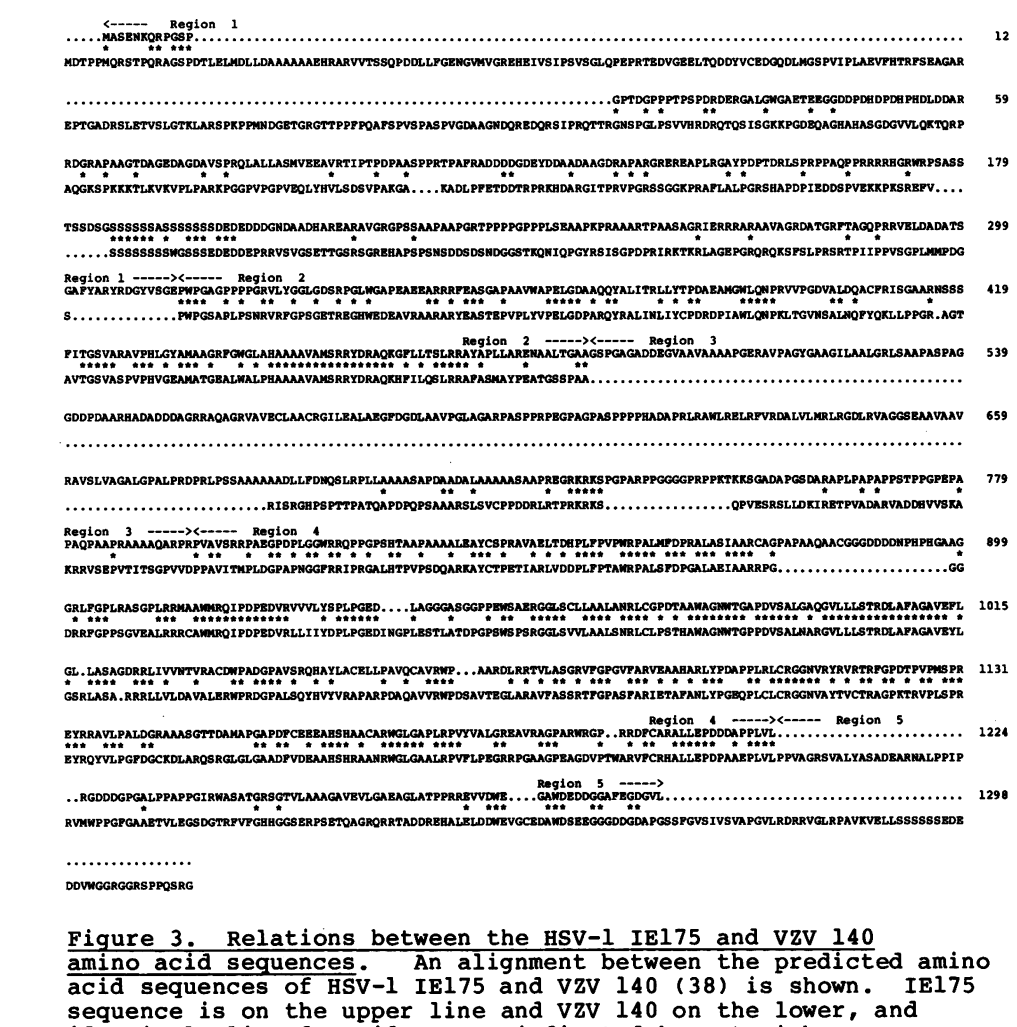


Figure 3. Relations between the HSV-1 IE175 and VZV 140 amino acid sequences. An alignment between the predicted amino acid sequences of HSV-1 IE175 and VZV 140 (38) is shown. IE175 sequence is on the upper line and VZV 140 on the lower, and identical aligned residues are indicated by asterisks. Introduced padding characters are shown as dots. Residue numbering for IE175 (Figure 2) is at the right, and boundaries of proposed regions are shown. This alignment was produced by first identifying major homologies with a matrix comparison program (39), then aligning corresponding subsections with an optimal alignment program (40). Alignment in largely non homologous regions is to some extent arbitrary.

IE175) consisting of a Ser rich tract followed by acidic residues. In addition, VZV 140 region 1 is 153 residues longer than IE175 region 1. Another indication of the non conserved nature of region 1 is given by the limited sequence data

available for the HSV-2 counterpart of HSV-1 IE175: the predicted 47 residue N terminal sequence is not homologous to IE175 sequence (43). Region 2 (residues 315-484 of IE175) is clearly homologous in the two sequences. The proteins have near identical lengths in this region (IE175, 170 residues; VZV 140, 169 residues). Region 3 of IE175 contains amino acids 485-796, and shows little detectable homology to VZV 140. In addition, region 3 of IE175 is 214 residues longer than the VZV counterpart. Region 4 is the largest of our divisions and comprises residues 797-1224 of IE175. It contains a 23 residue insertion in IE175 relative to VZV 140 (residues 875-897 of IE175) and several smaller additions/deletions. Otherwise it is clearly homologous between the two proteins. Region 5 comprises residues 1225 to 1298 (that is, the C terminus) of IE175. Apart from some limited homology at the C terminus of IE175, this region is non homologous. In addition, region 5 of VZV 140 is 88 residues longer than the IE175 region.

We have thus defined two quite distinct large scale classes of sequence within each of these proteins. One class (regions 1, 3 and 5) shows little or only very sporadic homology, and in addition exhibits substantial length differences between the two proteins. The other class (regions 2 and 4) shows very extensive homology, with only limited addition/deletion changes. Following section (b), above, we note that the VZV 140 gene has a lower G+C content (64.1%), and that VZV 140 contains less Ala (10.9% as against 20.9%) and more Lys (2.7% against 0.6%) than IE175. Other amino acids show lesser changes.

(d) Evaluation of functional importance of regions in IE175.

These homology relations between IE175 and VZV 140 are summarised, with respect to the IE175 sequence, in the top part of Figure 4. In considering the implications of this homology distribution, it is important to note that the protein coding regions of the two genes differ by 17.4 percentage points in their G+C contents (such variation is a widespread phenomenon of herpesvirus genomes - see Discussion). Thus, in diverging from a common ancestor, the two genomes have been subjected to a very extensive process of point mutation, as well as to large scale addition/deletion changes. Conserved amino acid sequences have

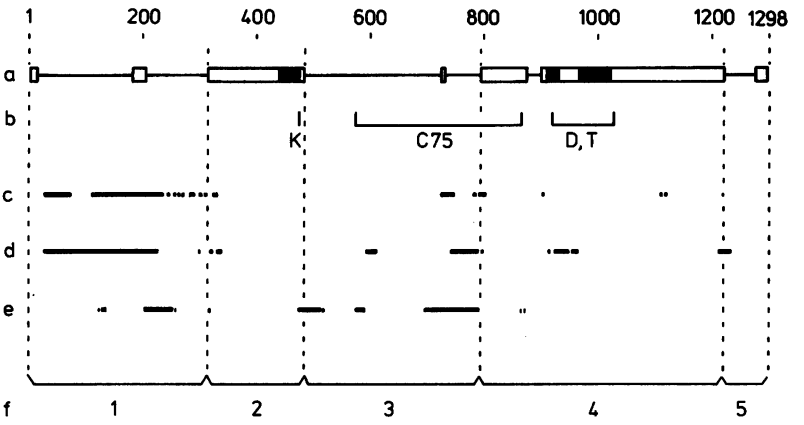


Figure 4. Characteristics of the IE175 amino acid sequence. Line (a) depicts the IE175 polypeptide, numbered as in Figure 2. Solid lines indicate regions without clear homologues in VZV 140. Clearly homologous sections are shown as boxes, with the most homologous sections filled (at least 16 residues identical out of 20). Line (b) shows the mapped positions of four ts mutants (44,45). Line (c) summarises "hydropathy" analyses (46): hydropathy sums were made for successive 50-residue windows, and regions with hydrophilic character greater than 4% of possible range from the mean shown as solid lines. Line (d) summarises an aspect of secondary structure prediction (47): sums of predicted coil content were made with 50-residue windows, and regions more than 10% of whole range above the mean shown as solid lines. In line (e), sums of (Ala + Pro + Gly + Arg) contents were made with 50-residue windows, and regions with a higher content of these amino acids than 10% of whole range above the mean were marked. Line (f) and the vertical dotted lines indicate regions proposed on basis of homology with VZV 140.

therefore survived comprehensive mutational "probing", and so very probably represent regions of particular functional importance for the IE175 protein. Conversely, the poorly conserved, variable length regions may be of lesser functional significance.

In other comparisons between HSV-1 and VZV protein sequences, we have observed several instances in which members of an homologous pair differ substantially in length at a terminal region, and these can be rationalised as representing in one member an addition of an adjunct protein structure, possibly of no great functional significance (41). The length differences of regions 1 and 5 appear to fall in this category. However, the

large size difference of 214 amino acids in region 3, in the middle of the sequence, is much more compelling, and makes attractive the speculative idea that the conserved regions, 2 and 4, represent separate physical (and possibly functional) domains, with region 3 having, at least in part, a spacer function in which precise chain length is not important.

Another type of external evidence relating to functionality comes from the mapping of ts mutations. As shown in Figure 4, the best mapped ts mutations (tsK, tsD and tsT; refs. 44,45) lie in or close to the gene regions encoding the most conserved amino acid sequences. Thus, these protein regions certainly contain essential functional or structural elements. The wider bracket assigned to tsC75 is of less diagnostic value.

We have examined three local attributes of amino acid sequences within IE175, namely, degree of hydrophilicity or hydrophobicity, predicted secondary structure, and content of frequently occurring amino acids, and have correlated these data with the homology results. The results of these analyses are shown in Figure 4 as summary forms indicating localities of particularly hydrophilic sequence, of particularly high predicted random coil content, and particularly rich in Ala, Pro, Gly and Arg. From the first two of these measures it is clear that region 1 of IE175 is distinct from the rest of the protein, containing very hydrophilic sequences, and with a high random coil content as predicted by the program of Garnier et al. (47).

We presented above the rationale that the gross amino acid composition of IE175 in part reflects evolution of the gene to an extremity of DNA base composition, most noticeably in that the four most frequent amino acids are those specified by codons containing only G and C. We examined the distribution of these residues, and found that localities of highest (Ala + Pro + Gly + Arg) content lie predominantly in the large non conserved regions 1 and 3 (Figure 4). An alternative presentation of the same phenomenon can be made by examination of local G+C content of the DNA, and in particular G+C content summed for the second codon position only: the highest local areas are within DNA specifying regions 1 and 3 (data not shown). We view this as indicating that these localities of the gene show particular plasticity in

mutation to a high G+C content - that is, they may encode polypeptide whose composition is not critical to IE175 function. This is supported by the work of Schröder et al. (48), who described an HSV-1 mutant with an internal, in-frame deletion in the IE175 gene. This removed codons 209 to 236, but the resulting virus was still viable. The sequence removed is in a part of region 1 with a high (Ala + Pro + Gly + Arg) content.

DISCUSSION

Our DNA sequence analysis gave an IE175 polypeptide molecular weight of 132,835, much lower than the previous gel electrophoretic estimate of 175,000 (49), which we consider resulted from the limitations of denaturing gel electrophoresis as a sizing technique and from the atypical amino acid composition of IE175. Nonetheless, IE175 is evidently a large and complex protein, and may well contain more than one physical domain. Our division into regions could indicate the basis of such a domain structure. We used three aspects of the amino acid sequence of IE175 to evaluate functional importance of regions of the protein, and to correlate with the external data of homology and mapping of mutations. We recognize that these tests represent imperfect tools. In particular, the high predicted coil content of region 1 may represent as much a failure of the predictive algorithm as a real prediction ("coil" is defined by Garnier et al. (47) as not α-helix, not extended, and not turn structures). Additionally, the three measures utilised are presumably not completely independent. We emphasize that the tests were used at low resolution over large stretches of sequence, to derive general characteristics of regions, and we consider that the conclusions reached are strongly supported by the correlations between the various measures, internal and external. None of these analyses indicate mechanisms for IE175's action. However, they do represent a major increase in resolution of the anatomy of this protein, and will provide a basis for molecular genetic analysis of function.

Partial denaturation mapping has shown that the highest G+C region in the HSV-1 genome is the R_S element (50). Here we have reported the sequence of R_S: 6633 base pairs of 79.5% G+C DNA.

Table 3. Comparison of codon catalogue base compositions

Gene	% (G+C) for each codon position		
	1st	2nd	3rd
HSV-1 IE175	81.9	66.2	96.3
VZV 140	68.1	57.2	67.0
T. thermophilus isopropylmalate dehydrogenase	72.8	48.6	89.4
Pseudorabies virus glycoprotein	69.8	50.1	93.8

Within this lies the protein coding region of one gene only, comprising 3894 base pairs of 81.5% G+C. Remarkably, this sequence is transcribed by host cell RNA polymerase II and is translated apparently by the unmodified host cell machinery in the newly infected cell. Sequences of other genes of a relatively high G+C content have been determined, although other HSV genes so far analysed are generally around 65% G+C (see, for instance, ref. 25). The two previously published, extreme sequences are of the isopropylmalate dehydrogenase gene of the bacterium Thermus thermophilus (51) and of a pseudorabies virus glycoprotein gene (52). However, the coding regions of these are 70.3 and 71.2% G+C respectively, more than 10 percentage points below the IE175 gene value. The codon set of the IE175 gene shows a third position G+C content of 96.3% (Table 3); that is, a near saturation of the redundant third positions by G and C. On the other hand, the second position value is 66.2%, and this is consistent with maximal retention of a wide range of encoded amino acid types. However, comparison with the codon position compositions for the two genes mentioned above shows that most of the difference between them and the IE175 codon set is in the second position and to a lesser extent the first position (Table 3). We view this as indicating that evolution to the present IE175 gene must have involved, to an unequalled extent, changes in proportions of encoded amino acids.

One imagines that evolution of a gene in this manner must create problems of functionality in the encoded protein and that these problems, although evidently soluble, must affect the relative fitnesses of progressively higher G+C versions of the genotype. This emphasizes that powerful evolutionary forces must

underlie such genome compositional effects. To put the biased base composition of HSV-1 R_S into context: this is only one example of a characteristic herpesvirus phenomenon of large scale base composition variation (53). This is seen at the whole genome level (for instance, HSV-1 (67% G+C) versus VZV (46% G+C; ref. 54)), and within one genome (for instance, R_S versus the adjacent U_S region of HSV-1, which is 64.3% G+C; ref. 25).

We have no real idea of the nature of the evolutionary forces which produce these effects. However, we outline here proposed elements of their mechanisms of implementation. We class them functionally as a mutation producer, a biasing mechanism and a disseminator. The mutation producer could be herpesvirus DNA polymerase (55,56). A biasing mechanism is necessary to give directionality of base change to the effect. In order to generate the full range of herpesvirus whole genome base compositions (32 - 75% G+C; ref. 56), the biasing mechanism must be capable of changing its directionality. Herpesvirus DNA polymerase could supply the biasing mechanism, by favouring introduction of appropriate mismatched residues (56). However, we now propose a model in which bias is introduced by recombination in DNA molecules within an infected cell, by a form of biased gene conversion which allows (in the case of evolution to high G+C) preferential survival of alleles containing G and C residues instead of A and T at any given position, or, equivalently, favours survival of some class of G+C rich sequence. This mechanism is of the same class as proposed mechanisms of multigene family evolution in eukaryotes ("molecular drive"; 57), which is of some aesthetic satisfaction. Involvement of recombination (biased or not) is seen as a mechanism for disseminating changes through an intracellular herpesvirus genome population. This is not an essential element of the scheme, for whole genome compositional variation, given a sufficient mutation rate. However, we have to invoke recombination (inter- or intra-genomic) between repeated regions to explain differences in base composition between such repeats and adjacent non repeated sequences, as in the case of HSV R_S. Since the population size of repeat sequences, for the HSV genome structure, is twice that of unique sequences, mutations in the

repeat population should arise at twice the rate per sequence site, and recombinational fixation can then accelerate the rate of population base compositional change. Alternatively, if rate of recombination is lower relative to mutation rate, so that only a proportion of directionally favoured mutations becomes fixed in the population, then intragenomic recombination provides for repeats a class of mutation spread not available to unique sequences.

ACKNOWLEDGEMENTS

We owe thanks to J.H. Subak-Sharpe for support and critical evaluation of the manuscript, to A.J. Davison for discussion and to C.M. Preston for criticism of the manuscript; also to P. Taylor for computing support, and to A.J. Davison, F.J. Rixon and N.D. Stow for plasmids.

[1]To whom correspondence should be addressed
[2]Members of the MRC Virology Unit
[3]Present address: Department of Molecular Biology, Massachusetts General Hospital, Boston, MA 02114, USA

REFERENCES

1. Kieff, E.D., Bachenheimer, S.L. & Roizman, B. (1971). J. Virol. 8, 125-132.
2. Roizman, B. (1979). Cell 16, 481-494.
3. Rixon, F.J., Campbell, M.E. & Clements, J.B. (1982). EMBO J. 1, 1273-1277.
4. Watson, R.J., Preston, C.M. & Clements, J.B. (1979). J. Virol. 31, 42-52.
5. Pereira, L., Wolff, M.H., Fenwick, M. & Roizman B. (1977). Virology 77, 733-749.
6. Marsden, H.S., Stow, N.D., Preston, V.G., Timbury, M.C. & Wilkie, N.M. (1978). J. Gen. Virol. 26, 389-410.
7. Preston, C.M. (1979). J. Virol. 32, 357-369.
8. Preston, C.M. (1979). J. Virol. 29, 275-285.
9. Watson, R.J. & Clements, J.B. (1980). Nature (London) 285, 329-330.
10. Dixon, R.A.F. & Schaffer, P.A. (1980). J. Virol. 36, 189-203.
11. Everett, R.D. (1983). Nucl. Acids Res. 11, 6647-6666.
12. Everett, R.D. & Dunlop, M. (1984). Nucl. Acids Res. 12, 5969-5978.
13. Freeman, J.M. & Powell, K.L. (1982). J. Virol. 44, 1084-1087.
14. Everett, R.D. (1984). Nucl. Acids Res. 12, 3037-3056.

15. ElKareh, A., Murphy, J.M., Fichter, T., Efstratiadis, A. & Silverstein, S. (1985). Proc. Nat. Acad. Sci., U.S.A. 82, 1002-1006.
16. Everett, R.D. (1984). EMBO J. 3, 3135-3141.
17. O'Hare, P. & Hayward, G.S. (1985). J. Virol. 53, 751-760.
18. Persson, R.H., Bacchetti, S. & Smiley, J.R. (1985). J. Virol. 54, 414-421.
19. Maxam, A.M. & Gilbert, W. (1980). Methods in Enzymology 65, 499-560.
20. Sanger, F., Coulson, A.R., Barrell, B.G., Smith, A.J.H. & Roe, B.A. (1980). J. Mol. Biol. 143, 161-178.
21. Deininger, P.L. (1983). Anal. Biochem. 129, 216-223.
22. Messing, J. & Vieira, J. (1982). Gene 19, 269-276.
23. Biggin, M.D., Gibson, T.J. & Hong, G.F. (1983). Proc. Nat. Acad. Sci., U.S.A. 80, 3963-3965.
24. Garoff, H. & Ansorge, W. (1981). Anal. Biochem. 115, 450-457.
25. McGeoch, D.J., Dolan, A., Donald, S. & Rixon, F.J. (1985). J. Mol. Biol. 181, 1-13.
26. Staden, R. (1982). Nucl. Acids Res. 10, 4731-4751.
27. Murchie, M.-J. & McGeoch, D.J. (1982). J. Gen. Virol. 62, 1-15.
28. Davison, A.J. & Wilkie, N.M. (1981). J. Gen. Virol. 55, 315-331.
29. Rixon, F.J. & Clements, J.B. (1982). Nucl. Acids Res. 10, 2241-2256.
30. Stow, N.D. & McMonagle, E.C. (1983). Virology 130, 427-438.
31. Mackem, S. & Roizman, B. (1982). Proc. Nat. Acad. Sci., U.S.A. 79, 4917-4921.
32. Cordingley, M.G., Campbell, M.E.M. & Preston, C.M. (1983). Nucl. Acids Res. 11, 2347-2365.
33. Preston, C.M., Cordingley, M.G. & Stow, N.D. (1984). J. Virol. 50, 708-716.
34. Mocarski, E.S. & Roizman, B. (1982). Cell 31, 89-97.
35. Stow, N.D., McMonagle, E.C. & Davison, A.J. (1983). Nucl. Acids Res. 11, 8205-8220.
36. Staden, R. & McLachlan, A.D. (1982). Nucl. Acids Res. 10, 141-156.
37. Palfreyman, J.W., Maclean, J.G. Messeder, E. & Sheppard, R.C. (1984). J. Gen. Virol. 65, 865-874.
38. Davison, A.J. & Scott, J.E. (1985). J. Gen. Virol. 66, 207-220.
39. Pustell, J. & Kafatos, F.C. (1982). Nucl. Acids Res. 10, 4765-4782.
40. Taylor, P. (1984). Nucl. Acids Res. 12, 447-456.
41. Davison, A.J. & McGeoch, D.J. (1986). J. Gen. Virol., in press.
42. Matthews, R.E.F. (1982). Intervirology 17, 1-199.
43. Whitton, J.L. & Clements, J.B. (1984). Nucl. Acids Res. 12, 2061-2079.
44. Preston, V.G. (1981). J. Virol. 39, 150-161.
45. Davison, M.-J., Preston, V.G. & McGeoch, D.J. (1984). J. Gen. Virol. 65, 859-863.
46. Kyte, J. & Doolittle, R.F. (1982). J. Mol. Biol. 157, 105-132.
47. Garnier, J., Osguthorpe, D.J. & Robson, B. (1978). J. Mol. Biol. 107, 327-356.

48. Schröder, C.H., DeZazzo, J., Knopf, K.W., Kaerner, H.C., Levine, M. & Glorioso, J. (1985). J. Gen. Virol. 66, 1589-1593.
49. Marsden, H.S., Crombie, I.K. & Subak-Sharpe, J.H. (1976). J. Gen. Virol. 31, 347-373.
50. Delius, H. & Clements, J.B. (1976). J. Gen. Virol. 33, 125-133.
51. Kagawa, Y., Nojima, H., Nukiwa, N., Ishizuka, M., Nakajima, T., Yasuhara, T., Tanaka, T. & Oshima, T. (1984). J. Biol. Chem. 259, 2956-2960.
52. Rea, T.J., Timmins, J.G., Long, G.W. & Post, L.E. (1985). J. Virol. 54, 21-29.
53. Honess, R.W. & Watson, D.H. (1977). J. Gen. Virol. 37, 15-37.
54. Ludwig, H., Haines, H.G., Biswal, N. & Benyesh-Melnick, M. (1972). J. Gen. Virol. 14, 111-114.
55. Hall, J.D., Coen, D.M., Fisher, B.L., Weisslitz, M., Randall, S., Almy, R.E., Gelep, P.T. & Schaffer, P.A. (1984). Virology, 132, 26-37.
56. Honess, R.W. (1984). J. Gen. Virol. 65, 2077-2107.
57. Dover, G. (1982). Nature (London) 299, 111-117.