Key words: *herpesvirus/methylation/episomes*

Deviations from Expected Frequencies of CpG Dinucleotides in Herpesvirus DNAs May Be Diagnostic of Differences in the States of Their Latent Genomes

By R. W. HONESS,[1]* U. A. GOMPELS,[1] B. G. BARRELL,[2]
M. CRAXTON,[1] K. R. CAMERON,[1] R. STADEN,[2] Y.-N. CHANG[3]
AND G. S. HAYWARD[3]

[1] *Division of Virology, National Institute for Medical Research, Mill Hill, London NW7 1AA,*
[2] *MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, U.K. and*
[3] *Virology Laboratories, Department of Pharmacy and Molecular Sciences, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, U.S.A.*

(*Accepted 22 November 1988*)

SUMMARY

The DNA sequences of genomes from G + C-rich and A + T-rich lymphotropic herpesviruses [i.e. gammaherpesviruses; Epstein–Barr virus and herpesvirus saimiri (HVS)] are deficient in CpG dinucleotides and contain an excess of TpG and CpA dinucleotides relative to frequencies predicted from their mononucleotide compositions. In contrast, for sequences from genomes of G + C-rich and A + T-rich neurotropic herpesviruses (i.e. alphaherpesviruses; herpes simplex virus and varicella-zoster virus) and human cytomegalovirus (HCMV; a betaherpesvirus) the mean observed frequencies of these dinucleotides are close to those expected from their mononucleotide compositions. Comparisons between DNA sequences that encode proteins conserved in all these viruses also show that sequences of these lymphotropic viruses are CpG-deficient whereas the homologous genes from the neurotropic viruses and the HCMV are not. Analyses of local variations in dinucleotide frequencies reveal some occurrences of clustered CpG dinucleotides in generally deficient genomes (e.g. upstream of the thymidylate synthase gene of HVS) and locally CpG-deficient regions within some generally non-deficient genomes (e.g. the major immediate early genes of human, simian and murine CMVs). A relative deficiency in CpG and an excess of TpG and CpA dinucleotides is a diagnostic feature of higher eukaryotic DNA sequences that have been subjected to methylation of cytosine residues in CpG doublets with the resulting increase in mutations to give TpG (and thereby its complement, CpA). The available evidence implicates the latent genome as the site of methylation of these herpesviruses. We conclude that in the neurotropic herpesviruses the normal latent precursors to infectious progeny are not methylated whereas there is local methylation of the immediate early locus in the latent genomes of CMVs, and the latent genomes of these lymphotropic herpesviruses are extensively methylated.

INTRODUCTION

All herpesviruses have double-stranded DNA genomes of more than 100 kbp, which are transcribed, replicated and packaged into complex icosahedral nucleocapsids within the nuclei of their eukaryotic host cells during productive cycles of virus growth. Productive or lytic cycles of herpesvirus replication are cytocidal and virus gene expression typically proceeds in at least three main sequential phases: immediate early (IE or α), delayed early (E/DE or β) and late (γ) (Roizman & Batterson, 1984).

In addition to these productive cycles of virus growth which occur during the initial acute infections of their normal multicellular hosts, herpesviruses commonly persist thereafter in latent infections, during which virus gene expression is highly restricted. Virus which is maintained in this latent form may then give rise to new rounds of productive infection and thus also provides a reservoir for the infection of naive contacts (Wildy *et al.*, 1982; Roizman & Sears, 1987). However, herpesviruses differ remarkably in the composition and structure of their genomes and in their biological properties and the diseases with which they are associated. A subdivision into three major virus subgroups (alpha-, beta- and gammaherpesvirus) has been suggested, based upon differences in the growth of the viruses *in vitro* and in the nature of the tissues involved in the acute and latent infections *in vivo* (Roizman, 1982; Honess & Watson, 1977; Honess, 1984). The human herpesviruses provide representatives of each of these three major subgroups.

The agents of recurrent oral and genital infections (herpes simplex virus types 1 and 2; HSV-1, HSV-2) and the virus of chickenpox and shingles (varicella-zoster virus; VZV) are representatives of the alpha- or neurotropic herpesviruses. Virus from the acute lesions caused by these viruses is seeded to neurons of the sensory ganglia serving these primarily affected sites and establishes a latent infection within these cells. Periodic recurrences of productive virus replication at peripheral sites served by these neurons may then occur. In contrast, the Epstein–Barr virus (EBV; the agent of glandular fever or infectious mononucleosis and the prototype of the so-called lymphotropic or gammaherpesviruses) produces a primary infection of the oropharynx with subsequent non-productive infection of circulating B cells. Both *in vivo* and *in vitro* the virus can be shown to persist as a multi-copy episome with limited virus gene expression within dividing B cell populations (Kieff *et al.*, 1983; Dambaugh *et al.*, 1986). It is evident that productive virus replication in the oropharynx must occur to enable the progeny virus to sustain the natural cycle of transmission, and it has recently been suggested that undifferentiated epithelial cells in the oropharynx are also the relevant site of the latent infection and virus persistence (see Allday & Crawford, 1988 for a brief review of the evidence for this view). Whatever the site of the biologically relevant latent infection, the intact EBV genome [and the genomes of the lymphotrophic viruses of New World monkeys, e.g. herpesvirus saimiri (HVS); see Fleckenstein & Desrosiers, 1982] can certainly persist as an episome within a dividing cell population. Finally, the cytomegaloviruses (CMVs) of man (human CMV; HCMV) and of other animals currently constitute the third major subgroup, the betaherpesviruses. Once again, there is a requirement for virus replication in the oropharynx to explain the normal transmission properties of the virus. The sites and properties of the latent or persistent infection are, however, uncertain (see for example, Mercer *et al.*, 1988).

We are interested in the relationships between the diverse molecular and biological properties of herpesviruses. Following the development of relatively rapid and efficient methods of DNA sequencing, an increasingly valuable database of sequences from herpesvirus genomes is becoming available. Complete sequences of the genomes of the human gammaherpesvirus, EBV (human herpesvirus 4; Baer *et al.*, 1984), the A + T-rich human alphaherpesvirus, VZV (human herpesvirus 3; Davison & Scott, 1986), HSV-1 (a G + C-rich alphaherpesvirus; McGeoch *et al.*, 1988) and a betaherpesvirus (HCMV; human herpesvirus 5) are now available (see Table 1 for a summary). We have also sequenced some 40 kbp from the 111 kbp coding sequences of the A + T-rich gammaherpesvirus HVS (Table 1) and fragmentary information is available for some other members of the group.

One of the clear conclusions from earlier analyses of these sequences was that the large differences in the mean mononucleotide compositions of herpesvirus genomes were not produced by selective pressures acting via the protein-coding functions of these sequences (Honess, 1984) and that there need be no functionally significant linkage between the biological properties of a given herpesvirus and its mean base composition. However, in the course of further analyses of herpesvirus DNA sequence we have noted that the frequency and distribution of some dinucleotides does appear to correlate with some biologically relevant properties. We present and discuss the significance of these observations in this paper.

METHODS

DNA sequences. The sources and some relevant properties of DNA sequences used for the present analyses are summarized in Table 1. The initial data collection and analysis were completed in January 1987 [presented in April 1987 at the meeting of the Society for General Microbiology (St Andrews, U.K.) and July 1987 at the 12th International Herpesvirus Workshop (University of Pennsylvania, Philadelphia, Pa., U.S.A.)], with the addition of the simian CMV (SCMV) IE region in August 1987. The sequence of the HSV-1 (mP17) genome has since been completed (McGeoch *et al*., 1988; D. J. McGeoch, personal communication) as has the sequence of the HCMV strain AD169 genome (B. G. Barrell *et al*., unpublished results) and additional regions of the HVS genome have been sequenced.

Analyses of DNA and presentation of results. Two sets of programs [ANALYSEQ (Staden, 1984, 1986) and Molecular Genetics and Sequencing (MGS) (W. Greer, P. Gillett & R. Mott, National Institute for Medical Research, London)] were used to count the observed frequencies of mononucleotides and dinucleotides in samples of DNA sequences and to calculate the expected frequencies of dinucleotides as the simple products of these observed mononucleotide frequencies. Deviations of the observed from the expected frequencies of dinucleotides are displayed in a number of ways; first, as the number of times the observed frequency (O) exceeds the expected frequency (E) (i.e. when O > E, O/E = + fold excess), or is less than the expected frequency (i.e. when E > O, E/O = − fold deficit). Excesses and deficits of equivalent magnitude have equivalent scaling in this display and we prefer it to the more usual plot of O/E for all values of O. Since these displays both use ratios of observed and expected frequencies they are corrected for differences in the absolute values of expected frequencies for sequences of differing mononucleotide composition (Fig. 1, 3 and 5). In order to examine correlations between the deviations from expected frequencies of independent pairs of dinucleotides, the absolute magnitudes of these differences were computed (i.e. O − E; e.g. Fig. 2). A modified version of the ANALYSEQ program (R. Staden, unpublished) permitted the computation and graphic display of these deviations of observed from expected frequencies within a sliding window (e.g. Fig. 6).

All the analyses presented here were of a single strand of the listed DNA sequence; in most cases this was the message-sense strand for small samples of coding sequences. Reading frames in the complete DNA sequences of herpesvirus genomes are distributed between both DNA strands and the majority of these sequences have coding functions. In situations where repetitions of a sequence feature made a significant contribution to the sequence under analysis (e.g. the *Bam*HI W repeats of EBV) analyses were performed separately for the repetitive and non-repetitive portions of the sequence.

RESULTS

Mean deviations of observed from expected frequencies of dinucleotides in large samples of DNA sequences from genomes of alpha-, beta- and gammaherpesvirus

A summary of the mean deviations of observed from expected frequencies of the 16 possible dinucleotides in samples of sequences from G + C-rich and A + T-rich alphaherpesviruses, HSV-1 (68·3% G + C) and VZV (46% G + C), from a betaherpesvirus (HCMV; 57% G + C) and from G + C-rich and A + T-rich gammaherpesviruses, EBV (60% G + C) and HVS (35% G + C), is presented in Fig. 1. The most striking feature of these plots is the highly significant deficits in the observed occurrence of CpG dinucleotides in sequences of EBV and HVS, combined with an excess of TpG + CpA and ApG + CpT. In contrast, the sequences of HSV, VZV and HCMV have observed frequencies of CpG, TpG and CpA dinucleotides very close to those predicted from their corresponding mononucleotide composition, but have a deficit in ApG + CpT dinucleotides. All these sequences have the small deficit in ApT + TpA and the excess in ApA + TpT which have been noted as general properties of DNA sequences from very diverse sources (see, for example Nussinov, 1984). The dinucleotide frequency measurements derived here from DNA sequence data for HSV-1 closely resemble previous results based upon measurements of nearest neighbour base frequencies by Subak-Sharpe and colleagues (Subak-Sharpe, 1967; Russell & Subak-Sharpe, 1977). These previous studies also showed that other alphaherpesviruses, pseudorabies virus (74% G + C) and equine abortion/rhinopneumonitis virus (54·4% G + C), had frequencies of CpG and of TpG + CpA close to values expected from unbiased associations between mononucleotides. The fragmentary sequence data available for genes from these viruses confirm that they are not deficient in CpG dinucleotides (see Table 1 for sources; results not shown).

Table 1. *Summary of sources and properties of DNA sequences from*

Subgroup	Virus (strain)	No.	Nature of sequence
α_1	HSV-1 (mP17)	1	U_S* component
	(mP17)	2	R_S† component
	(mP17)	3	U_L‡ component
	(mP17)	4	R_L‖ component
	(KOS)	5	U_L gene for alkaline exonuclease
	(12-7)	6	U_L gene for IE gene 5
	(mP17)	7	U_L gene for Vmw 65K (0·669 to 0·685 m.u.)¶
	(F)	8	U_L trans-inducing factor
	(HFEM)	9	U_L region encoding glycoprotein H (0·28 to 0·32 m.u.)
	(mP17)	10	U_L region encoding glycoprotein H (0·28 to 0·32 m.u.)
	(KOS)	11	U_L region encoding glycoprotein C (0·63 to 0·65 m.u.)
	(KOS)	12	U_L gene for glycoprotein B
	(KOS)	13	U_L gene for fusion function (0·732 to 0·745 m.u.)
	(KOS)	14	U_L early transcription unit (0·65 m.u.)
	(MP)	15	U_L thymidine kinase gene
	HSV-2 (333)	16	U_L region encoding 38K subunit of ribonucleotide reductase
	(333)	17	U_L region encoding 140K subunit of ribonucleotide reductase
α_1	Pseudorabies virus	18	U_L region encoding glycoprotein III homologue of glycoprotein C
		19	U_S gene encoding glycoprotein X
α_2	VZV (OKA)	20	Complete sequence
β	HCMV (AD169)	21	U_S component
		22	U_L *Hin*dIII F
		23	U_L gene encoding a homologue of glycoprotein B
		24	R_L component
		25	Complete sequence
	(Towne)	26	U_L gene for 67K tegument phosphoprotein (0·37 to 0·39 m.u.)
	(AD169)	27	U_L transforming region (0·123 to 0·14 m.u.)
	(Towne)	28	U_L major IE genes (region 1 and region 2)
	(Eisenhardt)	29	Early gene in R_L
β	SCMV (Colburn)	30	U_L major IE genes
β	MCMV (Smith)	31	U_L major IE genes
γ_1	EBV (B95-8)	32	Complete sequence
γ_2	HVS [11(Onc)]	33	*Sal*I + *Sma*I C fragment from right of L-DNA (contains 160K protein gene, and TS gene)
		34	Regions of L-DNA encoding glycoprotein H homologue
		35	Region of L-DNA encoding thymidine kinase homologue
		36	Region of L-DNA encoding 52K polypeptide (IE gene)
		37	Region of L-DNA encoding *Hin*dIII G protein (IE gene 1)
		38	Fragments of 350 bp distributed through L-DNA
		39	Terminal H-DNA repeat unit

genomes of representative alpha-, beta- and gammaherpesviruses

Length (bp)	Composition (percentage G + C)	Reference (entry in EMBO database)
12979	64·3	McGeoch *et al.* (1985) (HEHSV1SU)
6632	79·5	McGeoch *et al.* (1986) (M12345)
107943	66·9	McGeoch *et al.* (1988) (NCE)§
9215	71·6	Perry & McGeoch (1988) (NCE)
1000	67·2	Costa *et al.* (1983) (HE1EXO)
2560	69·1	Watson & Van de Woude (1982) (NCE)
2609	65·3	Dalrymple *et al.* (1985) (HEHSV165)
2522	64·9	Pellett *et al.* (1985) (HEHSV1AT)
6400	65	Gompels & Minson (1986) (NCE)
3740	65·7	McGeoch & Davison (1986) (HEHSV1GH)
2697	65·9	Frink *et al.* (1983) (HE1GC)
3755	64·2	Bzik *et al.* (1984) (HE1GB)
2041	62	Debroy *et al.* (1985) (HEHSV173)
1557	61·9	Draper *et al.* (1982) (HE1ETUK)
1799	63·3	McKnight (1980) (HERPES)
1890	62·6	Galloway & Swain (1984) (HE2P38K)
3998	66·4	Swain & Galloway (1986) (M12700)
1722	74·2	Robbins *et al.* (1986) (HEHSSGGI)
2023	69·4	Rea *et al.* (1985) (M10986)
124884	46·0	Davison & Scott (1986) (NCE)
43275	57·7	Weston & Barrell (1986) (HEHCMVU)
20349	58·8	Kouzarides *et al.* (1987) (NCE)
3125	53·1	Cranage *et al.* (1986) (HEHCMVGB)
6220	47·3	Greenaway & Wilkinson (1987) (NCE)
229355	57·2	B. G. Barrell *et al.* (unpublished results)
3072	50·4	Davis & Huang (1985) (M11911)
2848	40·1	Nelson *et al.* (1984) (HEADTRAN)
3847	53·4	Stenberg *et al.* (1984, 1985) (HEIE1) (HEHS5IER)
1170	47·1	Hutchinson & Tocci (1986) (NCE)
7916	45·7	Y.-N. Chang & G. S. Hayward (unpublished results) (NCE)
3360	50·1	Keil *et al.* (1987) (NCE)
172282	60	Baer *et al.* (1984) (EBV)
16500	35·6	Honess *et al.* (1986) (M13190); Bodemer *et al.* (1986) (NCE); Cameron *et al.* (1987) (NCE); R. W. Honess & K. R. Cameron (unpublished data) (NCE)
3927	32·2	U. A. Gompels, M. A. Craxton & R. W. Honess (unpublished data) (NCE)
2911	36·4	U. A. Gompels, M. A. Craxton & R. W. Honess (unpublished data) (NCE)
2236	36·5	U. A. Gompels, M. A. Craxton & R. W. Honess (unpublished data) (NCE)
1857	33·3	J. Nicholas, M. A. Craxton & R. W. Honess (unpublished data) (NCE)
3000	36·0	Gompels *et al.* (1988) (NCE)
1444	70·8	Bankier *et al.* (1985) (HSVSH)

* U_S, Short unique.

† R_S, Short repeat.

‡ U_L, Long unique.

§ NCE, No current entry.

|| R_L, Long repeat.

¶ m.u., Map unit.

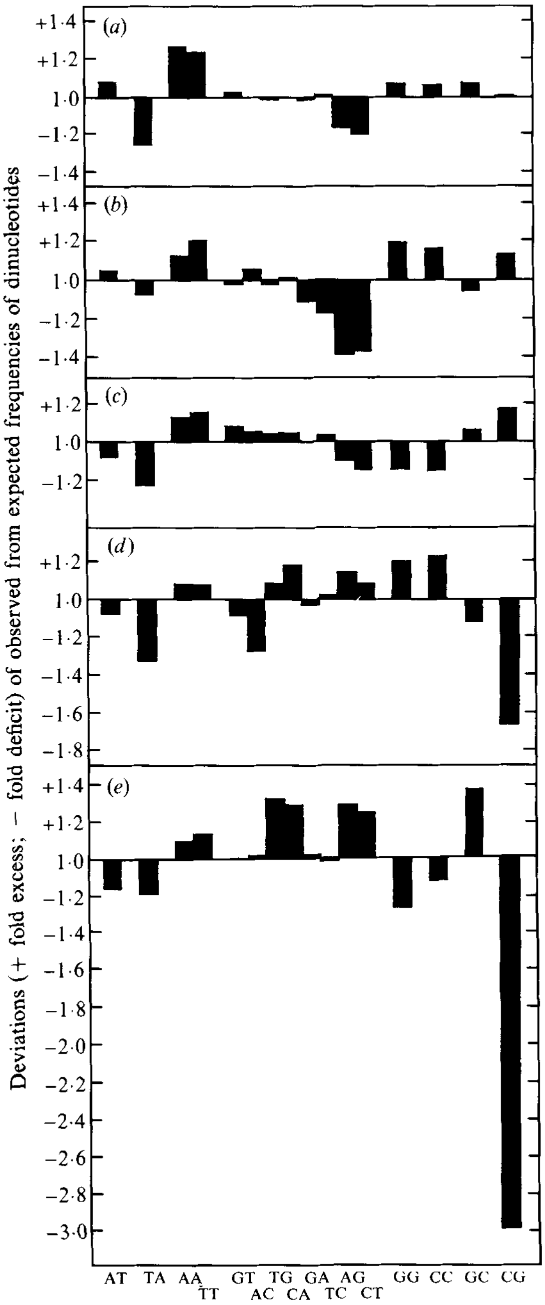


Fig. 1. Deviations of observed from expected frequencies of dinucleotides in DNA sequences from representatives of G + C-rich and A + T-rich alphaherpesviruses (HSV-1 and VZV), a betaherpesvirus (HCMV) and G + C-rich and A + T-rich gammaherpesviruses (EBV and HVS). The filled bars indicate the magnitude of deviations of observed occurrences (O) of each dinucleotide from occurrences expected (E) in random DNA sequences with the observed mononucleotide compositions (for O > E, O/E = + fold excess; for E > O, E/O = − fold deficit). The complete DNA sequences of (*a*) the HSV-1 strain mP17 genome (152260 nucleotides, 68·3% G + C), (*b*) the VZV genome (Davison & Scott, 1986; 124884 nucleotides, 46·1% G + C), (*c*) the HCMV strain AD169 genome (B. G. Barrell *et al.*, unpublished results; 229355 nucleotides, 57·2% G + C), (*d*) the B95-8 strain of EBV (172282 nucleotides, 60·1% G + C) and (*e*) 30 kbp of the 111 kbp L-DNA sequences of the HVS genome, with a mean composition of 35% G + C were included in the analysis shown. In addition, available samples of sequences from independent isolates of HSV and HCMV were analysed in a similar way, but are not illustrated (a summary of the sequences examined is given in Table 1).

The deficiency in CpG dinucleotides in gammaherpesvirus genomes is correlated with a compensating excess of TpG + CpA dinucleotides

Measurements of nearest neighbour base frequencies first established that within vertebrate DNAs the dinucleotide CpG occurs three- to fivefold less often than expected from their mean mononucleotide compositions. This relative deficiency has been convincingly related to the properties of the DNA methylation systems of higher eukaryotes and the instability of the major methylated product (Razin *et al.*, 1984). The major DNA methylation system modifies cytosine in CpG pairs to give 5-methyl-CpG and a substantial fraction of the available CpG dinucleotides are methylated. However, 5-methylcytosine is deaminated at high frequency to give thymine (Coulondre *et al.*, 1978), resulting in a net loss of 5-methyl-CpG to TpG and, after replication, CpA dinucleotides. In all the genomes so far analysed, the observed deficit in CpG dinucleotides is correlated with the extent of DNA methylation and a proportional excess of the products predicted from the resulting biased mutations (TpG + CpA) (Bird, 1980). Moreover, it is now apparent that vertebrate DNA is a mosaic of relatively G + C-rich sequences, in which CpG dinucleotides occur at close to the expected frequencies, and A + T-rich sequences in which CpG dinucleotides are deficient, i.e. that residual CpG dinucleotides are concentrated into 'islands' (Bernardi *et al.*, 1985). The majority of CpG dinucleotides in CpG islands are not methylated, whereas the residual CpG dinucleotides in the CpG-deficient fractions of the genomes are efficiently methylated (Bird, 1986).

A general correlation between the relative deficiency in CpG and the excess in TpG and CpA in herpesvirus DNAs was evident from Fig. 1. However, one of the advantages of an analysis of herpesvirus genomes is that a range of clearly homologous genes are encoded by highly divergent nucleotide sequences. It is therefore possible to examine the relationship between deviations of observed from expected frequencies of dinucleotides for these sets of homologous genes from CpG-deficient and non-deficient genomes of different mononucleotide composition. Data summarizing the relationships between deviations of observed from expected frequencies of CpG versus TpG + CpA dinucleotides for large samples of DNA sequences from HSV, VZV, EBV, HVS and CMV, together with data derived from homologous genes from these five viruses, are given in Fig. 2. There is a good correlation between CpG deficit and TpG + CpA excess in coding sequences for proteins of all recognized regulatory and functional classes in these G + C-rich and A + T-rich gammaherpesviruses. Sequences encoding homologous genes from G + C-rich and A + T-rich alphaherpesviruses are not deficient in CpG dinucleotides. The majority of the sequences of the HCMV contain CpG and TpG + CpA at frequencies that do not deviate markedly from random expectation (Fig. 1). More surprisingly, there is a significant local deficit of CpG, correlated with an excess of TpG and CpA, in the region of the dominant immediate early genes of HCMV (Fig. 2, IE; see also Fig. 5 and 6).

Local variations in the frequencies of CpG dinucleotides in CpG-deficient genomes

A relative deficiency in the occurrence of CpG correlated with an excess of TpG + CpA is reasonable *a priori* evidence for methylation as the mechanism responsible for CpG depletion. As noted above, exposure to methylation is not uniform in vertebrate DNA, and residual CpG dinucleotides are not randomly distributed. The selective retention of CpG dinucleotides in some regions of DNA and their loss from others may provide an indirect measure of differences in exposure to methylation systems or of a requirement for retention of a fraction of these residues. There are clear examples of this phenomenon in the papovaviruses, which are profoundly CpG-depleted and where residual CpG dinucleotides are concentrated in the regulatory sequences which separate early and late transcription units (unpublished observations). We have therefore examined available herpesvirus DNA sequences for intragenomic heterogeneity in the distribution of CpG dinucleotides.

A representation of the major coding and non-coding features of the linear form of the EBV genome is shown in Fig. 3, relative to displays of the variations in mononucleotide content and in the deviations of observed from expected occurrences of GpC (filled histograms) and CpG (hatched histograms) dinucleotides. All occurrences of CpG were below expectation, but there are significant differences in the magnitude of the deviation in different regions of the EBV genome. There is no simple relationship between the variations in mononucleotide composition of different regions of the genome and the deviations of observed from expected dinucleotide frequencies. Most major repetitive elements are CpG-deficient, as are the best characterized *cis* recognition frequencies (oriP and the major latent cycle promoter sequences in *Bam*HI C; see Discussion). The only sequences that can currently be presumed to contain a *cis* recognition element and which are not markedly CpG-deficient are the repeated sequences at the genome termini (two copies are indicated at the right-hand end of the genome in the map shown here).

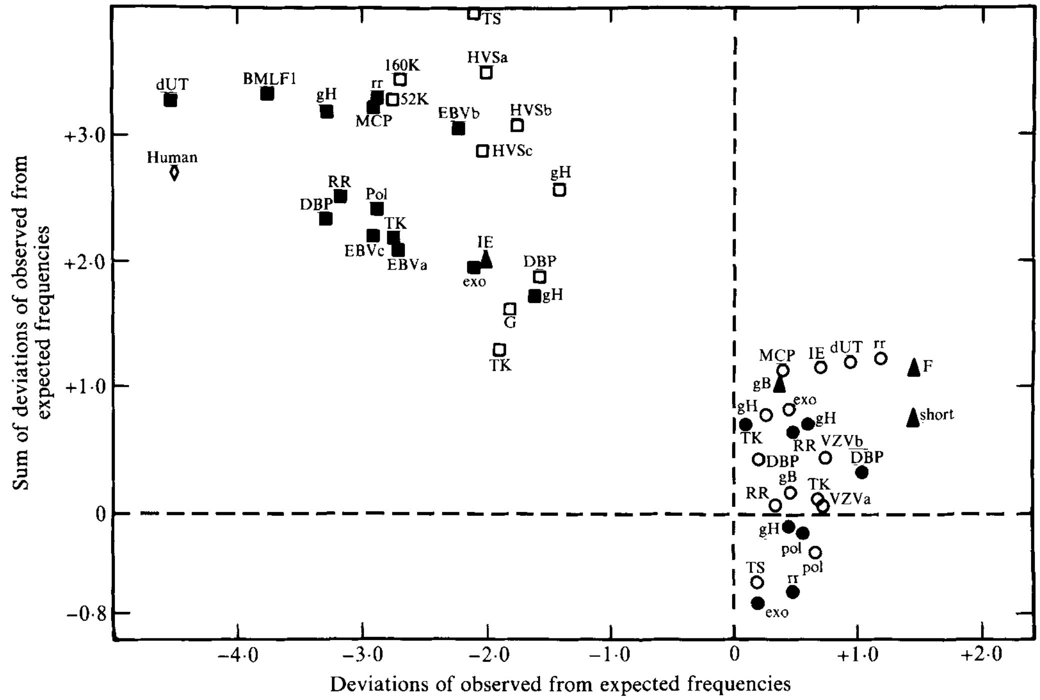


Fig. 2. Relationships between deviations of observed (subscript O) from expected (subscript E) frequencies of CpG [CpG(%)$_O$ − CpG(%)$_E$] (abscissa) and the sum of deviations of observed from expected frequencies of TpG and CpA {[TpG(%)$_O$ − TpG(%)$_E$] + [CpA(%)$_O$ − CpA(%)$_E$]} (ordinate) in samples of DNA sequences from representative herpesviruses, including regions that encode homologous protein products. The sequences of EBV analysed (■) are from the complete DNA sequence of Baer *et al.* (1984) with the reading frames identified as described in Fig. 3, except that the BMLF1 reading frame is annotated as such. In addition, data are given for regions defined by nucleotides 1 to 11500 (EBV a), 54 to 170000 (EBV c) and 55500 to 62000 (EBV b; BFLF2 and LF1 through BFRF1 to RF3). The sources of data for HVS (□) are given in Table 1. TS signifies the coding sequences for TS, G the region of the IE gene in *Hin*dIII G, 52K the coding sequences of the 52K IE gene; data points also shown for the whole 30000 nucleotide sample of HVS (HVS a), the rightmost 16500 nucleotides of HVS L-DNA (HVS c) and the region of this latter sequence that encodes homologues of the genes in EBV b (HVS b). For VZV (○) data points from the sequence of the whole genome (VZV a) and for the region of reading frames 23 to 27 inclusive (VZV b; homologous to EBV b and HVS b) are shown in addition to individual reading frames. For HCMV (▲), analyses from the sequence of the U$_S$ and R$_S$ regions (short), the *Hin*dIII F fragment (F), IE regions 1 and 2 (IE; see also Fig. 5 and 6) and glycoprotein B (gB) are given. A mean value for the total DNA from human spleen is also indicated (◇; human).

Other sites at which observed frequencies of CpG approach the expected values are distributed throughout coding sequences across the virus genome.

One possible basis for the selective retention of CpG dinucleotides within coding sequences would be via a constraint on the nature of permissible amino acid substitutions consistent with maintaining an essential function. The EBV genome contains many genes that are conserved throughout the herpesviruses (Fig. 2 and 3) as well as genes that have no recognized counterpart in other members of the group. We therefore examined the distribution of CpG dinucleotides relative to the distribution of conserved amino acid sequences in alignments of EBV genes encoding the major capsid protein (BcLF1), glycoprotein H (BXLF2), DNA polymerase (BALF5) and the major DNA-binding protein (BALF2) with the homologous genes of VZV and HSV. There was no simple relationship between the distribution of residual CpG dinucleotides in the EBV sequence and the distribution of amino acid residues that are conserved in these alignments (not shown). Whilst some of the most profoundly CpG-deficient regions of the EBV

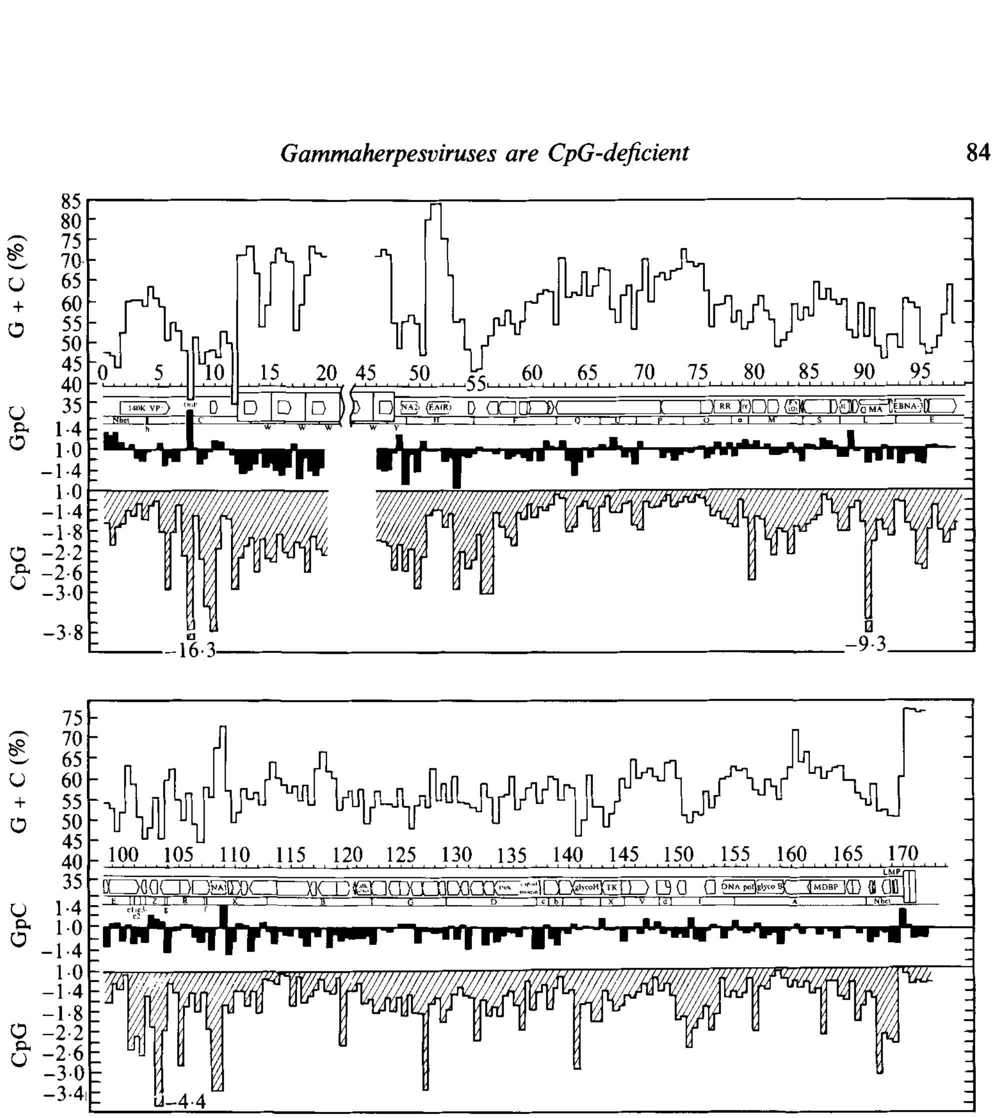


Fig. 3. Variations in nucleotide composition (percentage G + C; open histograms) and in deviations of observed from expected occurrences (+ fold excess; − fold deficit) of GpC (filled histograms) and CpG dinucleotides (hatched histograms) relative to the physical (*Bam*HI endonuclease cleavage sites are shown and the resulting fragments are annotated with their identifying letters) and genetic maps of the genomes from the B95-8 strain of EBV. The mononucleotide compositions and expected and observed occurrences of the dinucleotides were computed for non-overlapping 501-nucleotide intervals across the complete sequence of the EBV genome determined by Baer *et al.* (1984) (i.e. 172282 nucleotides; abscissa is marked at intervals of 1 kbp). The interpretation of the sequence to produce the genetic map, together with results of studies associating functions with products of the major open reading frames, has recently been summarized by Farrell (1989). Major open reading frames are represented by open arrows (amino to carboxy terminus) and are named in sequence relative to the *Bam*HI fragment containing their amino termini with their orientation on the prototype map indicated by R (rightward) or L (leftward). For example, the leftmost open reading frame begins in *Bam*HI N and is directed to the right, i.e. BNRF1. A number of reading frames with which functions have been associated, either directly or by virtue of homology with genes of known functions in other herpesviruses, are annotated. For example, the EBV-determined nuclear antigens (EBNAs), the latent membrane protein (LMP) and some early antigens are indicated [i.e. EBNA-1 exon in BKRF1, annotated NA-1; EBNA-2 includes the BYRF1 exon, annotated NA-2; EBNA-3 includes the BERF1 exon; LMP includes the BNLF1 exons; restricted early antigens EA(R) include the BHLF1 exon, and diffuse early antigens EA(D) include the BMLF1 exon]. A convenient summary of data identifying reading frames by their homology to genes of other herpesviruses has been presented by Davison & Taylor (1987) and by Gompels *et al.* (1988). These genes include BaRF1 and BORF2 with homology to the small (rr) and large (RR) subunits of ribonucleotide reductase, BLLF2 homologous to dUTPase (dU; dUT in Fig. 2), BGLF5 homologous to alkaline exonuclease (alk exo; exo in Fig. 2), BcLF1 homologous to the 150K capsid protein (MCP in Fig. 2), and BALF2, BALF4, BALF5, BXLF1 and BXLF2 encoding homologues of the major DNA-binding protein (MDBP; DBP in Fig. 2), glycoprotein B (glyco B; gB in Fig. 2), DNA polymerase (DNA-pol; pol in Fig. 2), thymidine kinase (TK), and glycoprotein H (glyco H, gH in Fig. 2), respectively.

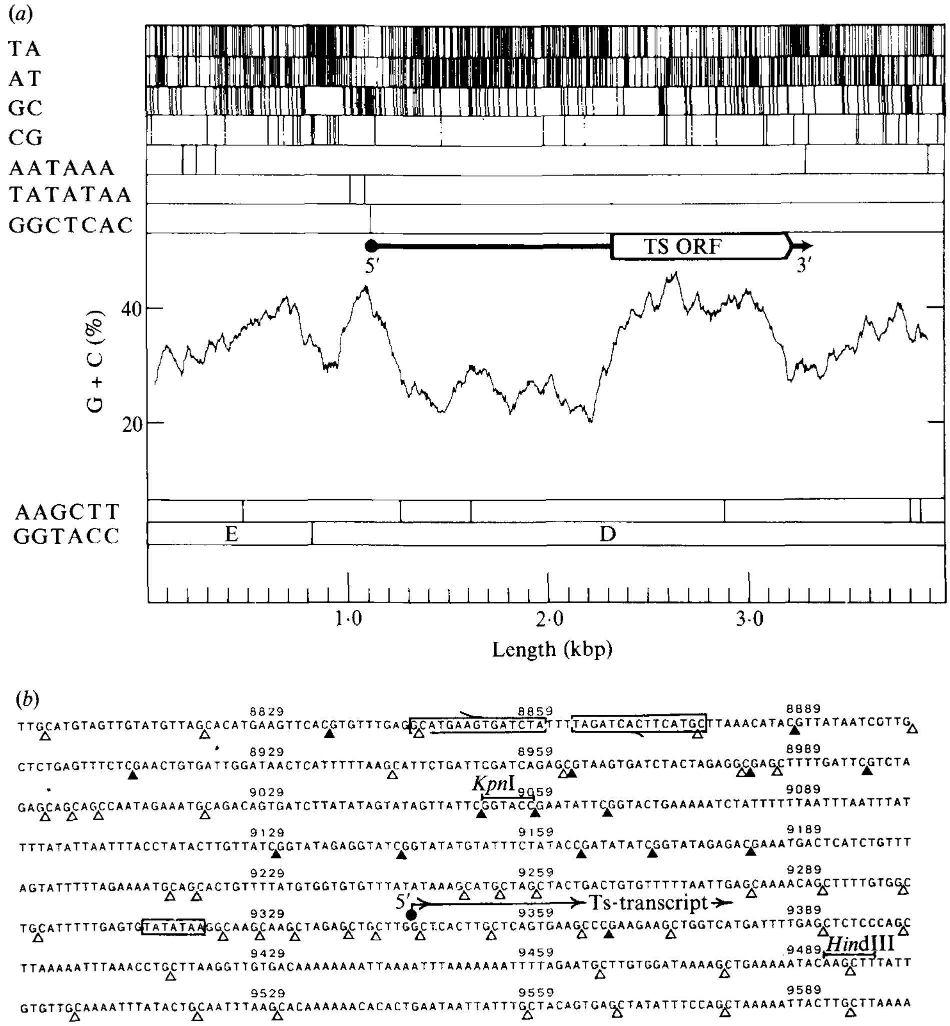

(a)
TA
AT
GC
CG
AATAAA
TATATAA
GGCTCAC
TS ORF
5′
3′
G + C (%)
40
20
AAGCTT
GGTACC
E
D
1·0
2·0
3·0
Length (kbp)
(b)
8829
8859
8889
TTGCATGTAGTTGTATGTTAGCACATGAAGTTCACGTGTTTGAGGCATGAAGTGATCTATTTTAGATCACTTCATGCTTAAACATACGTTATAATCGTTG
8929
8959
8989
CTCTGAGTTTCTCGAACTGTGATTGGATAACTCATTTTTAAGCATTCTGATTCGATCAGAGCGTAAGTGATCTACTAGAGGCGAGCTTTTGATTCGTCTA
KpnI
9029
9059
9089
GAGCAGCAGCCAATAGAAATGCAGACAGTGATCTTATATAGTATAGTTATTCGGTACCGAATATTCGGTACTGAAAAATCTATTTTTTAATTTAATTTAT
9129
9159
9189
TTTATATTAATTTACCTATACTTGTTATCGGTATAGAGGTATCGGTATATGTATTTCTATACCGATATATCGGTATAGAGACGAAATGACTCATCTGTTT
9229
9259
9289
AGTATTTTTAGAAAATGCAGCACTGTTTTATGTGGTGTGTTTATATAAAGCATGCTAGCTACTGACTGTGTTTTTAATTGAGCAAAACAGCTTTTGTGGC
5′
Ts-transcript
9329
9359
9389
TGCATTTTTGAGTGTATATAAGGCAAGCAAGCTAGAGCTGCTTGGCT.CACTTGCTCAGTGAAGCCCGAAGAAGCTGGTCATGATTTTGAGCTCTCCCAGC
HindIII
9429
9459
9489
TTAAAAATTTAAACCTGCTTAAGGTTGTGACAAAAAAAATTAAAATTTAAAAAAATTTTAGAATGCTTGTGGATAAAAGCTGAAAAATACAAGCTTTATT
9529
9559
9589
GTGTTGCAAAATTTATACTGCAATTTAAGCACAAAAAACACACTGAATAATTATTTGCTACAGTGAGCTATATTTCCAGCTAAAAATTACTTGCTTAAAA

Fig. 4. Local regions that are not CpG-deficient occur in globally CpG-deficient genomes, sequences upstream of the TS gene of HVS. (*a*) Summary of relevant features within a 4000 nucleotide sequence from the *Kpn*I D and *Kpn*I + *Sma*I E (position of *Kpn*I sites, GGTACC, and *Hin*dIII sites, AAGCTT, are indicated at the bottom of the panel) fragments of the L-DNA component of HVS DNA which includes the coding sequences for the TS gene. Occurrences of TpA, ApT, GpC and CpG dinucleotides, polyadenylation signals (AATAAA) and TATA box homologies (TATATAA), are indicated in the upper part of the panel, relative to the structure of the 2·2 kbp major late transcript which serves the TS gene (bold line with the position of the open reading frame, indicated as TS ORF, and with 5′ end at the first G of the GGCTCAC sequence). The continuous tracing below the schematic diagram of the transcript shows variations in the mean mononucleotide composition across the sequence. (*b*) Detailed annotation of an 800-nucleotide sequence in the region of the start site for the TS transcript. The sequence is numbered with respect to the *Sma*I site marking the conventional righthand end of HVS L-DNA (see Stamminger *et al.*, 1987; Cameron *et al.*, 1987) with the *Kpn*I site separating *Kpn*I + *Sma*I E and *Kpn*I D (in *a*) starting at nucleotide 9052. Occurrences of GpC (△) and CpG (▲) are marked, as in the perfect 15 bp palindrome (beginning at nucleotide 8844) and the TATA homology (boxed, beginning at nucleotide 9314) proximal to the start site for the major transcript (nucleotide 9343). The sequence is from our analyses of the HVS genome, but this region is included in the sequence published by Bodemer *et al.* (1986) and the differences between our sequence and the published sequence of Bodemer *et al.* (1986) do not affect the present results.

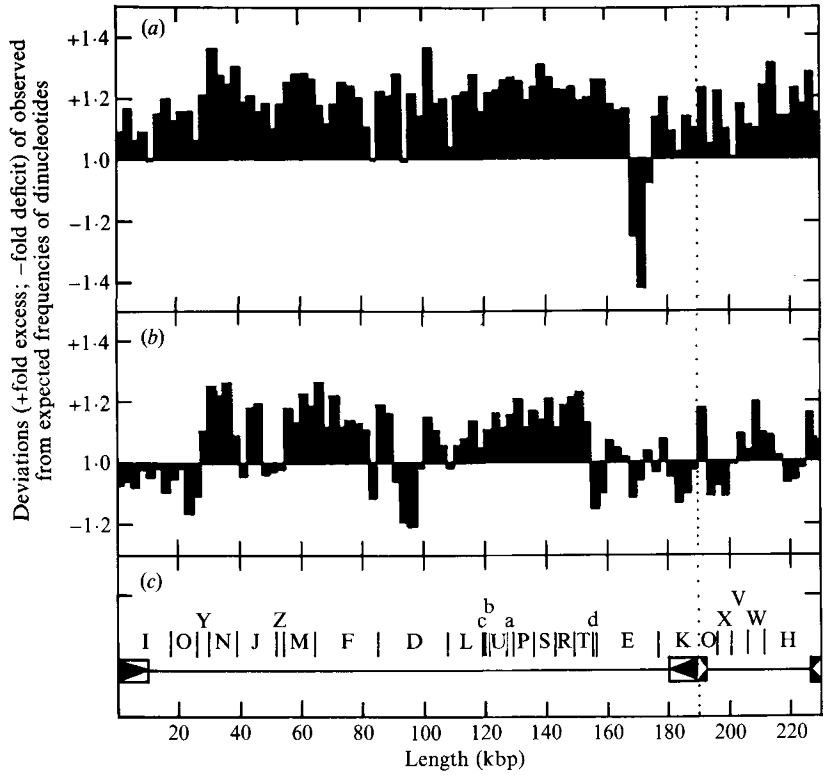
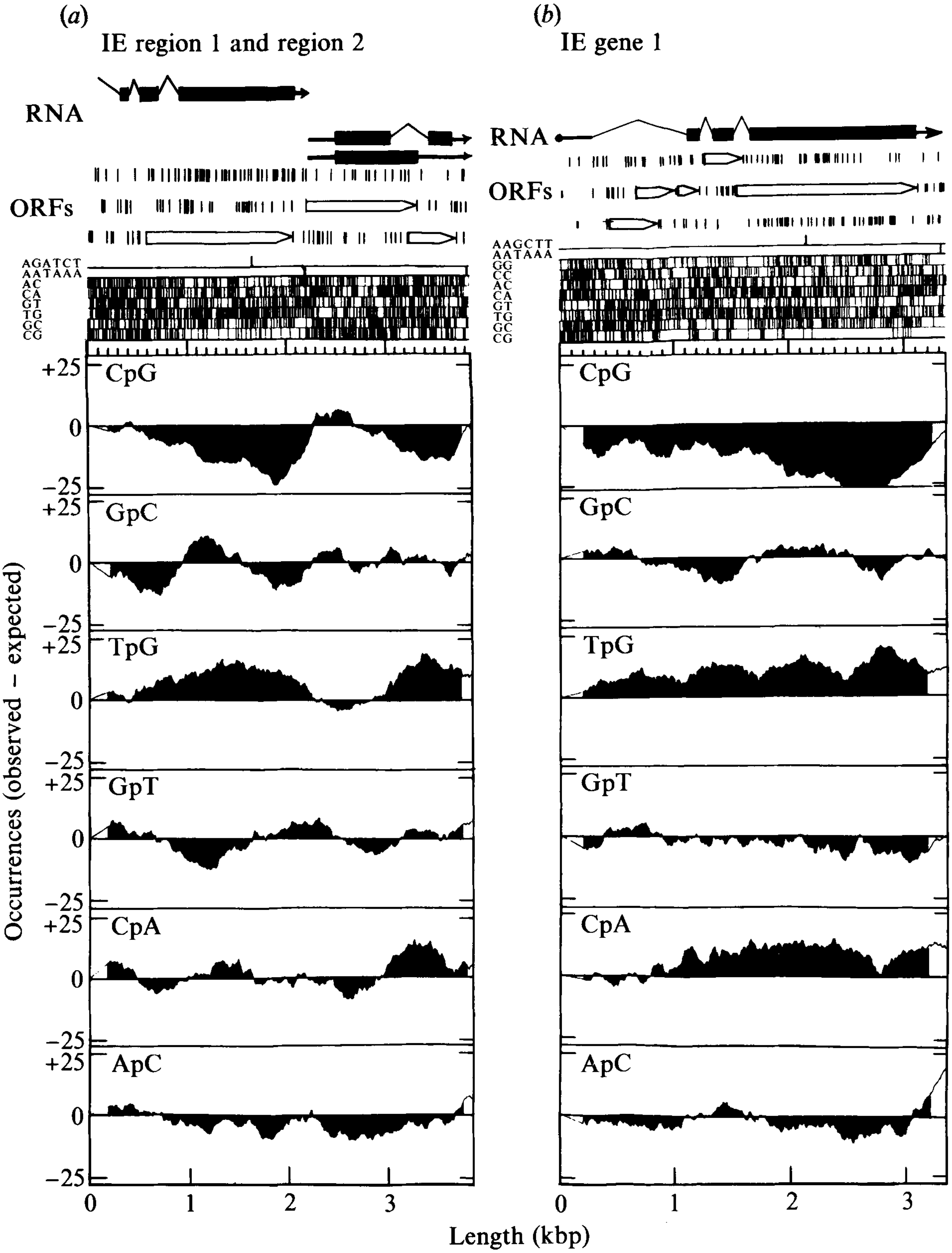


Fig. 5. Deviations of observed from expected frequencies of CpG (*a*) and GpC (*b*) dinucleotides in non-overlapping intervals of 2501 nucleotides across the complete sequence of the genome of HCMV, strain AD169 (deviations calculated as for Fig. 1 and 3; sequence from B. G. Barrell *et al.*, unpublished results). A map of the cleavage sites for *Hin*dIII and the resulting DNA fragments is shown (*c*) relative to the 'long' and 'short' subcomponents of the AD169 genome ('long' and 'short' inverted duplications are indicated by boxed arrows with the L-S junction indicated by the dotted vertical line). Some regional variations in the deviations of observed from expected dinucleotide frequencies are apparent, the most significant of which is a local CpG deficiency over the region of the major IE gene (within *Hin*dIII E; see Fig. 6 for analyses of the regions of the major IE genes in another isolate of HCMV and in SCMV and MCMV).

genome occur in coding and non-coding regions which have no homologues in the genomes of HSV and VZV, most of these regions contain repetitive nucleotide sequence elements. Within non-repetitive sequences, some of the least CpG-deficient regions of the EBV genome specify the least well conserved protein products (e.g. BOLF1, BPLF1; see for example Davison & Taylor, 1987). We conclude that regional variations in exposure to methylation are a major source of heterogeneity in the distribution of CpG dinucleotides in EBV DNA.

An examination of the available sequences of the HVS genome (Table 1, and unpublished results) revealed two examples of sequences which depart from the general pattern of CpG deficiency with TpG/CpA excess. The first example consists of a cluster of CpG dinucleotides localized in the unique sequences upstream of the start site for the major late transcript which serves the thymidylate synthase (TS) gene of HVS (Honess *et al.*, 1986; Bodemer *et al.*, 1986; E. P. Smith & R. W. Honess, unpublished results). The major features of this sequence are summarized in Fig. 4. The general level of CpG deficiency across this region is similar to the remainder of the genome (compare occurrences of CpG and GpC summarized at the top in *a*). However, a small cluster of CpG dinucleotides occurs in a relatively A + T-rich region between the TATA box of the HVS promoter and a prominent 15 bp dyad symmetry element (Fig. 4).


(a)
IE region 1 and region 2
RNA
ORFs
AGATCT
AATAAA
AC
CA
GT
TG
GC
CG
(b)
IE gene 1
RNA
ORFs
AAGCTT
AATAAA
GG
CC
AC
CA
GT
TG
GC
CG
CpG
GpC
TpG
GpT
CpA
ApC
+25
0
−25
Occurrences (observed − expected)
0
1
2
3
Length (kbp)

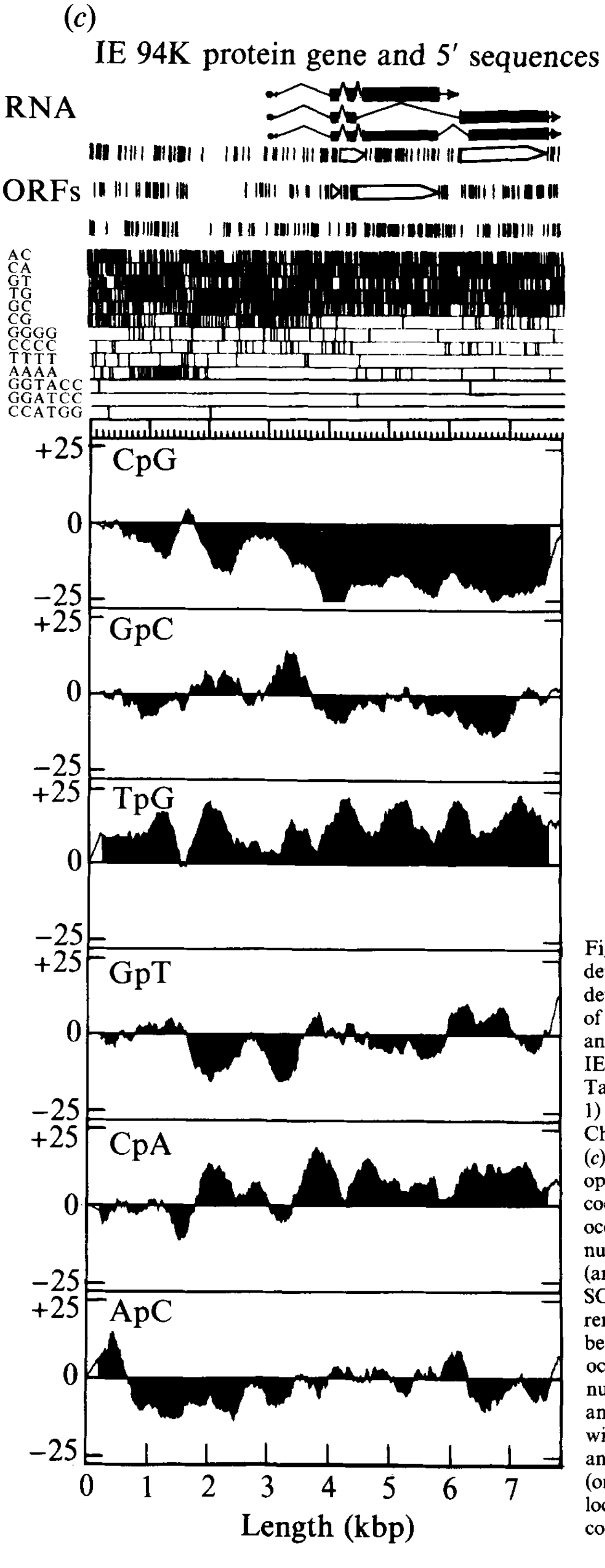


Fig. 6. Local CpG deficiencies occur in non-deficient genomes. Regions of relative CpG deficiency and TpG/CpA excess in sequences of IE regulatory genes in CMV genomes. The analyses presented are of sequences from the IE regions of HCMV (strain Towne; see Table 1) (*a*), MCMV (strain Smith; see Table 1) (*b*), and SCMV (strain Colburn; Y.-N. Chang & G. S. Hayward, unpublished results) (*c*). In each panel, the major transcripts and open reading frames (bars indicate stop codons) are indicated at the top, together with occurrences of reference restriction endonuclease cleavage sites and of dinucleotides (and homotetranucleotides in the case of the SCMV sequence). The six plots forming the remainder of each panel show the difference between the observed and expected number of occurrences (observed − expected) of the dinucleotides CpG, GpC, TpG, GpT, CpA and ApC in a 401-nucleotide sliding window with data points plotted every five nucleotides and with the plots scaled from +25 to −25 (ordinates). In each case a highly significant local deficiency in occurrences of CpG is correlated with an excess of TpG and CpA.

The second example of a set of sequences that has properties which differ from the majority of the HVS genome are those of the terminal H-DNA repeat units. The most common form of this repetitive unit from HVS strain 11 DNA consists of 1444 bp with a mean composition of 70·9% G + C (Bankier *et al.*, 1985). Mature progeny genomes each contain more than 30 copies of this sequence distributed between the left- and right-hand molecular ends. The only function attributed to these highly reiterated sequences is that of providing appropriately spaced recognition sites for the cleavage/packaging of progeny DNA molecules (see Stamminger *et al.*, 1987). The dinucleotide frequencies of the repetitive unit show a number of significant departures from random expectations, including a deficit in CpG dinucleotides (O/E = 7·6%/12·6%). This relative deficit is less than that observed for the majority of the HVS genome and there is not a compensating excess of TpG and CpA (O/E = 6·2%/4·9% and 5·1%/5·3%, respectively). Moreover, the CpG deficiency is not the only significant departure from random expectation (there are significant excesses of ApG and CpT, O/E = 9·7%/5·8% and 7·1%/4·6% respectively; and deficits of ApC and GpT, O/E = 2·6%/5·3% and 3·3%/4·9% respectively). We conclude that DNA methylation does not have a major influence on the base sequence of the H-DNA repeat.

Local variations in dinucleotide frequencies in alpha- and betaherpesvirus genomes

Within the available samples of sequences from genomes of the alpha- and betaherpesviruses, the only clearly significant local CpG deficiency correlated with a TpG + CpA excess occurs over the major IE genes of CMV (Fig. 5). Data for the major IE genes of HCMV were also noted in Fig. 2; however, the sequences of IE genes of a murine CMV (MCMV) and a simian CMV (SCMV) have also been completed (Table 1). A comparison of the occurrences of selected di- and oligonucleotides relative to the major open reading frames and mRNA species proposed for each of these sequences is presented in Fig. 6 (upper portions of *a*, *b* and *c*). Also shown in Fig. 6 are measures of the differences between observed and expected frequencies of selected dinucleotides across these sequences (lower parts of *a*, *b* and *c*). A significant CpG deficit with a correlated TpG + CpA excess occurs over the coding sequences of each of these IE genes. The HCMV and SCMV sequences encode proteins which are clearly related, and although these proteins encoded by the 'MCMV sequence are not obviously related to proteins of the primate viruses (Keil *et al.*, 1987), they are presumed to have analogous roles as regulatory cofactors.

DISCUSSION

In this paper we have identified some major differences in the frequencies of dinucleotides that occur in the genomes of biologically distinct herpesviruses. Specifically, DNA sequences of A + T-rich and G + C-rich gammaherpesviruses are generally deficient in CpG dinucleotides with a compensating excess in occurrences of TpG and CpA. In contrast, these dinucleotides occur at frequencies close to those expected from random associations between mononucleotides in the genomes of A + T-rich and G + C-rich alphaherpesviruses. A localized deficiency of CpG correlated with an excess of TpG and CpA occurs in the region of the major IE genes of three different CMVs (i.e. betaherpesviruses). Our interest in these observations is that they suggest a simple criterion which may discriminate in a significant way between properties required of the latent genomes of alpha-, beta- and gammaherpesviruses and the tissues that harbour them. We summarize our interpretation of these observations in Table 2 and state the main steps involved in reaching these interpretations below. We also consider experimental data which support these interpretations and outline some predictions based upon them.

The major logical elements in our interpretation are as follows. (i) The only known mechanism for producing a selective depletion of CpG dinucleotides with a compensating increase in TpG and CpA is via biased mutations resulting from deamination of 5-methylcytosine in methylated CpG dinucleotides. (ii) Herpesviruses are not known to encode DNA methylation systems, and virion DNA from productively infected cells is not detectably methylated. Differences in the methylation of latent genomes by host cell methylation systems are therefore the most likely sources of differential loss of CpG and gain of TpG and CpA dinucleotides. (iii) Methylated latent genomes must be, or have been, the regular precursors to infectious progeny in viruses

Table 2. *Summary of subgroup-specific differences in the occurrence of CpG dinucleotides in herpesvirus genomes and their interpretation*

Subgroup	Examples	Dinucleotide frequency	Conclusion	Inference
Alphaherpesvirus	HSV, VZV, Equine abortion virus.	CpG at expected frequency. ApG/CpT deficiency.	Latent precursors to infectious progeny are not methylated.	Latent genomes are maintained in non-dividing, terminally differentiated tissues. No *de novo* methylation.
Betaherpesvirus	HCMV, MCMV, SCMV	Global CpG at expected frequency. Local deficiency in region of IE genes.	Local methylation of IE regions of latent genomes.	Latently infected tissue capable of limited *de novo* methylation; immediate early regions selectively exposed.
Gammaherpesvirus	EBV, HVS	Global deficiency in CpG, some local retention at expected frequency.	Generalized methylation of latent genomes with some privileged sites protected.	Latent site in dividing cells capable of efficient *de novo* methylation. Majority of genome exposed to methylating system.

with CpG-deficient genomes, and cannot be the normal precursors to infectious progeny in viruses which are not CpG-deficient. (iv) *De novo* methylation of DNA is a relatively inefficient process and there are significant tissue-specific differences in the rates of *de novo* methylation. Virus genomes that routinely become extensively methylated must therefore remain stably associated with a dividing cell population through multiple cell divisions.

Experimental investigations with EBV and with HVS have shown that genomes of each of these viruses are maintained as multi-copy episomes within proliferating lymphoblastoid cells (see Introduction). In the case of EBV this property has been linked to the possession of a specific *cis* recognition sequence (the plasmid origin, oriP) and the expression of a single origin-binding function (EBNA-1) which permits stable maintenance of an intact virus genome as a plasmid which is replicated by host cell enzymes (Reisman *et al.*, 1985; Yates *et al.*, 1985; Rawlins *et al.*, 1985). These EBV and HVS episomes have each been shown to become extensively methylated in a process requiring many cell divisions. However, even in highly methylated molecules, a number of sites have been shown to remain undermethylated (hypomethylated), including sequences comprising the terminal H-DNA repeats of HVS DNA. The DNA progeny produced by induction or reactivation from cultures containing methylated EBV or HVS episomes are not methylated (Desrosiers *et al.*, 1979; Desrosiers, 1982; Kintner & Sugden, 1981; Larocca & Clough, 1982; Szyf *et al.*, 1985; Dyson & Farrell, 1985).

In contrast, the latent genomes of HSV in neuronal cells are not detectably methylated (Dressler *et al.*, 1987). Moreover, there is no evidence for the functional equivalent of a plasmid origin in the genome of HSV or VZV. The only known 'origin' sequences in these alphaherpesviruses are sequences utilized as origins of lytic cycle DNA replication. The activation of these lytic cycle origins requires the presence of multiple DE replicative functions

(Wu *et al.*, 1988), none of which have been shown to be expressed constitutively in latently infected cells (e.g. Deatly *et al.*, 1987). The absence of detectable methylation of HSV in neurons is not due to the inability of host cell methylation systems to act on this substrate. Multiple CpG residues within G + C-rich sequences appear to be optimal targets for these methylation systems (Bolden *et al.*, 1985) and integrated copies of HSV DNA in cell lines have been shown to become methylated (e.g. Clough *et al.*, 1982). Moreover, there is at least one report of a lymphoblastoid cell line in which apparently intact HSV genomes have been shown to become methylated (Youssoufian *et al.*, 1982). Thus, there is reasonable evidence showing that the genomes of EBV and HVS routinely become methylated, that they can be reactivated from methylated precursors and that HSV is not methylated in its normal latent site, the neuron.

A number of specific and general predictions arise from our interpretation of these observations. First, significant CpG clusters in non-repetitive sequences of CpG-deficient genomes should indicate regions of latent virus genomes that are relatively protected from methylation. The site 5′ to the HVS TS gene should provide an appropriate system to test this proposition. Conversely, the CpG deficiency localized over the region of the major IE gene of the CMVs suggests that this region alone is routinely exposed to methylation in the latent state. Such localized exposure is difficult to imagine if the latent template is replicated as an episome, but can be imagined if constitutive expression of the IE gene from an unreplicated template increases its relative accessibility. Another obvious general prediction, which should also be testable, is that the transposition of sequences from a non-CpG-deficient virus into a CpG-deficient virus should lead to high rates of methylation-induced mutation of the transposed sequences associated with passage through the latent state.

Finally, the most interesting prediction concerns the properties of Marek's disease virus (MDV) of chickens. This virus has long been regarded as a lymphotropic or gammaherpesvirus because of its association with lymphomas in chickens and as a result of extensive studies of cell lines derived from these lymphomas which have suggested parallels to EBV (Payne, 1982; Nonoyama, 1982). Thus, multiple copies of the virus genome are detectable in cell lines in which a minority of cells appear to be undergoing productive infection, virus genomes in some of these cell lines have apparently been shown to be episomal and to become methylated (Hirai *et al.*, 1981; Kaschka-Dierich *et al.*, 1979; Kanamori *et al.*, 1987). However, recent studies of the genetic organization of MDV [and of its near relative the herpesvirus of turkeys (HVT)] have shown that the genomes of these viruses are grossly collinear with those of the alphaherpesviruses HSV and VZV (Buckmaster *et al.*, 1988). There are a number of genetic permutation events necessary to relate the gene orders of these alphaherpesviruses to the order of conserved genes observed in the gammaherpesviruses EBV and HVS (Gompels *et al.*, 1988). Furthermore, pairwise comparisons between the predicted protein sequences encoded by the MDV genome with those of HSV/VZV and EBV also reveal a much greater degree of similarity to products of the alphaherpesviruses and a number of the proteins encoded by MDV/HVT have only been recognized in other alphaherpesviruses [Buckmaster *et al.*, 1988; we have also found that the MDV 'A' antigen sequence presented by Coussens & Velicer (1988) is clearly homologous to HSV-1 glycoprotein C). Most relevant to our interest here, available samples of sequences of MDV and HVT are clearly not deficient in CpG, have no excess of TpG and CpA but do have a deficit in ApG/CpT (our analyses of sequences provided by Dr A. Buckmaster & Dr L. J. N. Ross; and in Coussens & Velicer, 1988). There is, therefore, a discrepancy between the classification of MDV/HVT as a gammaherpesvirus based upon a view of some of its biological properties and objective measures of relatedness which place MDV with members of the alphaherpesvirus lineage in a phylogenetically based system. If our interpretation of the source of CpG deficiency is correct, we must also predict that the methylated MDV DNA sequences in lymphoid cells cannot be the latent intermediate in the natural transmission cycle of the virus. The epidemiologically significant form of the virus must be transmitted from the progeny of a persistent, lytic infection or by reactivation from a latent genome which is not methylated. The resolution of this issue will require a much clearer view of the criteria that should be applied in attempting to understand the relationships between molecular and biological properties of herpesviruses. We believe that the analysis of dinucleotide frequencies provides a simple molecular characteristic which may well correlate with some significant biological properties and should be considered in attempts to understand the evolution of the herpesviruses.

A number of colleagues provided us with copies of DNA sequences in advance of publication or with tapes of published sequences prior to their general availability in public databases. In particular we thank Drs D. J. McGeoch, A. J. Davison, K. Weston and T. Kouzarides. Drs J. L. N. Ross and A. Buckmaster provided mononucleotide and dinucleotide frequency data on some of their unpublished sequences from Marek's disease virus DNA.

REFERENCES

ALLDAY, M. J. & CRAWFORD, D. H. (1988). Role of epithelium in EBV persistence and pathogenesis of B-cell tumours. *Lancet* i, 855–857.

BAER, R., BANKIER, A. T., BIGGIN, M. D., DEININGER, P. L., FARRELL, P. J., GIBSON, T. J., HATFULL, G., HUDSON, G. S., SATCHWELL, S. C., SÉGUIN, C., TUFFNELL, P. S. & BARRELL, B. G. (1984). DNA sequence and expression of the B95-8 Epstein–Barr virus genome. *Nature, London* **310**, 207–211.

BANKIER, A. T., DIETRICH, W., BAER, R., BARRELL, B., COLBERE-GARAPIN, F., FLECKENSTEIN, B. & BODEMER, W. (1985). Terminal repetitive sequences in herpesvirus saimiri virion DNA. *Journal of Virology* **55**, 133–139.

BERNARDI, G., OLOFSSON, B., FILIPSKI, J., ZERIAL, M., SALINAS, J., CUNY, G., MEUNIER-ROTIVAL, M. & RODIER, F. (1985). The mosaic genome of warm-blooded vertebrates. *Science* **228**, 953–958.

BIRD, A. P. (1980). DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Research* **8**, 1499–1504.

BIRD, A. P. (1986). CpG-rich islands and the function of DNA methylation. *Nature, London* **321**, 209–213.

BODEMER, W., NILLER, H. H., NITSCHE, N., SCHOLZ, B. & FLECKENSTEIN, B. (1986). Organization of the thymidylate synthase gene of herpesvirus saimiri. *Journal of Virology* **60**, 114–123.

BOLDEN, A. H., NALIN, C. M., WARD, C. A., POONIAN, M. S., McCOMAS, W. W. & WEISSBACH, A. (1985). DNA methylation: sequences flanking C–G pairs modulate the specificity of the human DNA methylase. *Nucleic Acids Research* **13**, 3479–3494.

BUCKMASTER, A. E., SCOTT, S. D., SANDERSON, M. J., BOURSNELL, M. E. G., ROSS, N. L. J. & BINNS, M. M. (1988). Gene sequence and mapping data from Marek's disease virus and herpesvirus of turkeys: implications for herpesvirus classification. *Journal of General Virology* **69**, 2033–2042.

BZIK, D. J., FOX, B. A., DELUCA, N. A. & PERSON, S. (1984). Nucleotide sequence specifying the glycoprotein gene, gB, of herpes simplex virus type 1. *Virology* **133**, 301–314.

CAMERON, K. R., STAMMINGER, T., CRAXTON, M., BODEMER, W., HONESS, R. W. & FLECKENSTEIN, B. (1987). The 160,000-Mr virion protein encoded at the right end of the herpesvirus saimiri genome is homologous to the 140,000-Mr membrane antigen encoded at the left end of the Epstein-Barr virus genome. *Journal of Virology* **61**, 2063–2070.

CLOUGH, D. W., KUNKEL, L. M. & DAVIDSON, R. L. (1982). 5-Azacytidine-induced reactivation of a herpes simplex thymidine kinase gene. *Science* **216**, 70–73.

COSTA, R. H., DRAPER, K. G., BANKS, L., POWELL, K. L., COHEN, G., EISENBERG, R. & WAGNER, E. K. (1983). High-resolution characterization of herpes simplex virus type 1 transcripts encoding alkaline exonuclease and a 50,000-dalton protein tentatively identified as a capsid protein. *Journal of Virology* **48**, 591–603.

COULONDRE, C., MILLER, J. H., FARABAUGH, P. J. & GILBERT, W. (1978). Molecular basis of base substitution hotspots in *Escherichia coli*. *Nature, London* **274**, 775–780.

COUSSENS, P. M. & VELICER, L. F. (1988). Structure and complete nucleotide sequence of the Marek's disease herpesvirus gp57-65 gene. *Journal of Virology* **62**, 2373–2379.

CRANAGE, M. P., KOUZARIDES, T., BANKIER, A. T., SATCHWELL, S., WESTON, K., TOMLINSON, P., BARRELL, B., HART, H., BELL, S. E., MINSON, A. C. & SMITH, G. L. (1986). Identification of the human cytomegalovirus glycoprotein B gene and induction of neutralizing antibodies via its expression in recombinant vaccinia virus. *EMBO Journal* **5**, 3057–3063.

DALRYMPLE, M. A., McGEOCH, D. J., DAVISON, A. J. & PRESTON, C. M. (1985). DNA sequence of the herpes simplex virus type 1 gene whose product is responsible for transcriptional activation of immediate-early promoters. *Nucleic Acids Research* **13**, 7865–7879.

DAMBAUGH, T., HENNESSY, K., FENNEWALD, S. & KIEFF, E. (1986). The virus genome and its expression in latent infections. In *The Epstein-Barr Virus: Recent Advances*, pp. 13–45. Edited by M. Epstein & B. Achong. London: Heinemann.

DAVIS, M. G. & HUANG, E-S. (1985). Nucleotide sequence of a human cytomegalovirus DNA fragment encoding a 67-kilodalton phosphorylated viral protein. *Journal of Virology* **56**, 7–11.

DAVISON, A. J. & SCOTT, J. E. (1986). The complete DNA sequence of varicella-zoster virus. *Journal of General Virology* **67**, 1759–1816.

DAVISON, A. J. & TAYLOR, P. (1987). Genetic relations between varicella-zoster virus and Epstein–Barr virus. *Journal of General Virology* **68**, 1067–1079.

DEATLY, A. M., SPIVACK, J. G., LAVI, E. & FRASER, N. W. (1987). RNA from an immediate early region of the type 1 herpes simplex virus genome is present in the trigeminal ganglia of latently infected mice. *Proceedings of the National Academy of Sciences, U.S.A.* **84**, 3204–3208.

DEBROY, C., PEDERSON, N. & PERSON, S. (1985). Nucleotide sequence of a herpes simplex virus type 1 gene that causes cell fusion. *Virology* **145**, 36–48.

DESROSIERS, R. C. (1982). Specifically unmethylated cytidylic-guanylate sites in herpesvirus saimiri DNA in tumor cells. *Journal of Virology* **43**, 427–435.

DESROSIERS, R. C., MULDER, C. & FLECKENSTEIN, B. (1979). Methylation of herpesvirus saimiri DNA in lymphoid tumor cells. *Proceedings of the National Academy of Sciences, U.S.A.* **76**, 3839–3843.

DRAPER, K. G., FRINK, R. J. & WAGNER, E. K. (1982). Detailed characterization of an apparently unspliced beta herpes simplex virus type 1 gene mapping in the interior of another. *Journal of Virology* **43**, 1123–1128.

DRESSLER, G. R., ROCK, D. L. & FRASER, N. W. (1987). Latent herpes simplex virus type 1 DNA is not extensively methylated *in vivo*. *Journal of General Virology* **68**, 1761–1765.

DYSON, P. J. & FARRELL, P. J. (1985). Chromatin structure of Epstein–Barr virus. *Journal of General Virology* **66**, 1931–1940.

FARRELL, P. J. (1989). Epstein-Barr virus genome. In *Advances in Viral Oncology*, vol. 8, pp. 103–132. Edited by G. Kleine. New York: Raven Press.

FLECKENSTEIN, B. & DESROSIERS, R. C. (1982). Herpesvirus saimiri and herpesvirus ateles. In *The Herpesviruses*, vol. 1, pp. 253–332. Edited by B. Roizman. New York & London: Plenum Press.

FRINK, R. J., EISENBERG, R., COHEN, G. & WAGNER, E. K. (1983). Detailed analysis of the portion of the herpes simplex virus type 1 genome encoding glycoprotein C. *Journal of Virology* **45**, 634–647.

GALLOWAY, D. A. & SWAIN, M. A. (1984). Organization of the left-hand end of the herpes simplex virus type 2 BglII-N fragment. *Journal of Virology* **49**, 724–730.

GOMPELS, U. A. & MINSON, A. (1986). The properties and sequence of glycoprotein H of herpes simplex virus type 1. *Virology* **153**, 230–247.

GOMPELS, U. A., CRAXTON, M. A. & HONESS, R. W. (1988). Conservation of gene organization in the lymphotropic herpesviruses, herpesvirus saimiri and Epstein-Barr virus. *Journal of Virology* **62**, 757–767.

GREENAWAY, P. J. & WILKINSON, G. W. G. (1987). Nucleotide sequence of the most abundantly transcribed early gene of human cytomegalovirus strain AD169. *Virus Research* **7**, 17–31.

HIRAI, K., IKUTA, K., KITAMOTO, N. & KATO, S. (1981). Latency of herpesvirus of turkey and Marek's disease virus genomes in a chicken T-lymphoblastoid cell line. *Journal of General Virology* **53**, 133–143.

HONESS, R. W. (1984). Herpes simplex and 'the herpes complex': diverse observations and a unifying hypothesis. *Journal of General Virology* **65**, 2077–2107.

HONESS, R. W. & WATSON, D. H. (1977). Unity and diversity in the herpesviruses. *Journal of General Virology* **37**, 15–37.

HONESS, R. W., BODEMER, W., CAMERON, K. R., NILLER, H.-H., FLECKENSTEIN, B. & RANDALL, R. E. (1986). The A + T-rich genome of herpesvirus saimiri contains a highly conserved gene for thymidylate synthase. *Proceedings of the National Academy of Sciences, U.S.A.* **83**, 3604–3608.

HUTCHINSON, N. I. & TOCCI, M. J. (1986). Characterization of a major early gene from the human cytomegalovirus long inverted repeat: predicted amino acid sequence of a 30kDa protein encoded by the 1.2kb mRNA. *Virology* **155**, 172–182.

KANAMORI, A., IKUTA, K., UEDA, S., KATO, S. & HIRAI, K. (1987). Methylation of Marek's disease virus DNA in chicken T-lymphoblastoid cell lines. *Journal of General Virology* **68**, 1485–1490.

KASCHKA-DIERICH, C., NAZERIAN, K. & THOMSSEN, R. (1979). Intracellular state of Marek's disease virus DNA in two tumour-derived chicken cell lines. *Journal of General Virology* **44**, 271–280.

KEIL, G. M., EBELING-KEIL, A. & KOSZINOWSKI, U. H. (1987). Sequence and structural organization of murine cytomegalovirus immediate-early gene 1. *Journal of Virology* **61**, 1901–1908.

KIEFF, E., DAMBAUGH, T., HUMMEL, M. & HELLER, M. (1983). Epstein-Barr virus transformation and replication. In *Advances in Viral Oncology*, vol. 3, pp. 133–182. Edited by G. Klein. New York: Raven Press.

KINTNER, C. & SUGDEN, B. (1981). Conservation and progressive methylation of Epstein-Barr virus DNA sequences in transformed cells. *Journal of Virology* **38**, 305–316.

KOUZARIDES, T., BANKIER, A. T., SATCHWELL, S. C., WESTON, K., TOMLINSON, P. & BARRELL, B. G. (1987). Large-scale rearrangement of homologous regions in the genomes of HCMV and EBV. *Virology* **157**, 397–413.

LAROCCA, D. & CLOUGH, W. (1982). Hypomethylation of Epstein-Barr virus DNA in the nonproducer B-cell line EBR. *Journal of Virology* **43**, 1129–1131.

McGEOCH, D. J. & DAVISON, A. J. (1986). DNA sequence of the herpes simplex virus type 1 gene encoding glycoprotein gH, and identification of homologues in the genomes of varicella-zoster virus and Epstein–Barr virus. *Nucleic Acids Research* **14**, 4281–4292.

McGEOCH, D. J., DOLAN, A., DONALD, S. & RIXON, F. J. (1985). Sequence determination and genetic content of the short unique region in the genome of herpes simplex virus type 1. *Journal of Molecular Biology* **181**, 1–13.

McGEOCH, D. J., DOLAN, A., DONALD, S. & BRAUER, H. K. (1986). Complete DNA sequence of the short repeat region in the genome of herpes simplex virus type 1. *Nucleic Acids Research* **14**, 1727–1745.

McGEOCH, D. J., DALRYMPLE, M. A., DAVISON, A. J., DOLAN, A., FRAME, M. C., McNAB, D., PERRY, L. J., SCOTT, J. E. & TAYLOR, P. (1988). The complete DNA sequence of the long unique region in the genome of herpes simplex virus type 1. *Journal of General Virology* **69**, 1531–1574.

McKNIGHT, S. L. (1980). The nucleotide sequence and transcript map of the herpes simplex virus thymidine kinase gene. *Nucleic Acids Research* **8**, 5949–5964.

MERCER, J. A., WILEY, C. A. & SPECTOR, D. H. (1988). Pathogenesis of murine cytomegalovirus infection: identification of infected cells in the spleen during acute and latent infections. *Journal of Virology* **62**, 987–997.

NELSON, J. A., FLECKENSTEIN, B., JAHN, G., GALLOWAY, D. A. & McDOUGALL, J. K. (1984). Structure of the transforming region of human cytomegalovirus AD169. *Journal of Virology* **49**, 109–115.

NONOYAMA, M. (1982). The molecular biology of Marek's disease herpesvirus. In *The Herpesviruses*, vol. 1, pp. 333–346. Edited by B. Roizman. New York & London: Plenum Press.

NUSSINOV, R. (1984). Doublet frequencies in evolutionary distinct groups. *Nucleic Acids Research* **12**, 1749–1763.

PAYNE, L. N. (1982). Biology of Marek's disease virus and the herpesvirus of turkeys. In *The Herpesviruses*, vol. 1, pp. 343–431. Edited by B. Roizman. New York & London: Plenum Press.

PELLETT, P. E., McKNIGHT, J. L. C., JENKINS, F. J. & ROIZMAN, B. (1985). Nucleotide sequence and predicted amino acid sequence of a protein encoded in a small herpes simplex virus DNA fragment capable of trans-inducing alpha genes. *Proceedings of the National Academy of Sciences, U.S.A.* **82**, 5870–5874.

PERRY, L. J. & McGEOCH, D. J. (1988). 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. *Journal of General Virology* **69**, 2831–2846.

RAWLINS, D., MILMAN, G., HAYWARD, S. & HAYWARD, G. (1985). Sequence specific DNA-binding of the Epstein-Barr virus nuclear antigen (EBNA-1) to clustered sites in the plasmid maintenance region. *Cell* **42**, 859–868.

RAZIN, A., CEDAR, H. & RIGGS, A. D. (1984). *DNA Methylation: Biochemistry and Biological Significance*. New York & Wien: Springer-Verlag.

REA, T. J., TIMMINS, J. G., LONG, G. W. & POST, L. E. (1985). Mapping and sequence of the gene for the pseudorabies virus glycoprotein which accumulates in the medium of infected cells. *Journal of Virology* **54**, 21–29.

REISMAN, D., YATES, J. & SUGDEN, B. (1985). A putative origin of replication of plasmids derived from Epstein-Barr virus is composed of two *cis*-acting components. *Molecular and Cell Biology* **5**, 1822–1832.

ROBBINS, A. K., WATSON, R. J., WHEALY, M. E., HAYS, W. W. & ENQUIST, L. W. (1986). Characterization of a pseudorabies virus glycoprotein gene with homology to herpes simplex virus type 1 and type 2 glycoprotein C. *Journal of Virology* **58**, 339–347.

ROIZMAN, B. (1982). The family Herpesviridae: general description, taxonomy, and classification. In *The Herpesviruses*, vol. 1, pp. 1–23. Edited by B. Roizman. New York & London: Plenum Press.

ROIZMAN, B. & BATTERSON, B. (1984). The replication of herpesviruses. In *General Virology*, pp. 497–526. Edited by B. Fields. New York: Raven Press.

ROIZMAN, B. & SEARS, A. E. (1987). An inquiry into the mechanisms of herpes simplex virus latency. *Annual Review of Microbiology* **41**, 543–571.

RUSSELL, G. J. & SUBAK-SHARPE, J. H. (1977). Similarity of the general designs of protochordates and invertebrates. *Nature, London* **266**, 533–536.

STADEN, R. (1984). Graphic methods to determine the function of nucleic acid sequences. *Nucleic Acids Research* **12**, 521–538.

STADEN, R. (1986). The current status and portability of our sequence handling software. *Nucleic Acids Research* **14**, 217–231.

STAMMINGER, T., HONESS, R. W., YOUNG, D. F., BODEMER, W., BLAIR, E. D. & FLECKENSTEIN, B. (1987). Organization of terminal reiterations in the virion DNA of herpesvirus saimiri. *Journal of General Virology* **68**, 1049–1066.

STENBERG, R. M., THOMSEN, D. R. & STINSKI, M. F. (1984). Structural analysis of the major immediate-early gene of human cytomegalovirus. *Journal of Virology* **49**, 190–199.

STENBERG, R. M., WITTE, P. R. & STINSKI, M. F. (1985). Multiple spliced and unspliced transcripts from human cytomegalovirus immediate-early region 2 and evidence for a common initiation site within immediate-early region 1. *Journal of Virology* **56**, 665–675.

SUBAK-SHARPE, J. H. (1967). Base doublet frequency patterns in the nucleic acid and evolution of viruses. *British Medical Bulletin* **23**, 161–168.

SWAIN, M. A. & GALLOWAY, D. A. (1986). Herpes simplex virus specifies two subunits of ribonucleotide reductase encoded by 3′-coterminal transcripts. *Journal of Virology* **57**, 802–808.

SZYF, M., ELIASSON, L., MANN, V., KLEIN, G. & RAZIN, A. (1985). Cellular and viral DNA hypomethylation associated with induction of Epstein-Barr virus lytic cycle. *Proceedings of the National Academy of Sciences, U.S.A.* **82**, 8090–8094.

WATSON, R. J. & VAN DE WOUDE, G. (1982). DNA sequence of an immediate-early gene (IE mRNA-5) of herpes simplex virus type 1. *Nucleic Acids Research* **10**, 979–991.

WESTON, K. & BARRELL, B. G. (1986). Sequence of the short unique region, short repeats and part of the long repeats of human cytomegalovirus. *Journal of Molecular Biology* **192**, 177–208.

WILDY, P., FIELD, H. J. & NASH, A. A. (1982). Classical herpes latency revisited. In *Virus Persistence, Symposium of the Society for General Microbiology* no. 33, pp. 133–167. Edited by B. W. J. Mahy, A. C. Minson & G. K. Darby. Cambridge: Cambridge University Press.

WU, C. A., NELSON, N. J., McGEOCH, D. J. & CHALLBERG, M. D. (1988). Identification of herpes simplex virus type 1 genes required for origin-dependent DNA synthesis. *Journal of Virology* **62**, 435–443.

YATES, J., WARREN, N. & SUGDEN, B. (1985). Stable replication of plasmids derived from Epstein–Barr virus in various mammalian cells. *Nature, London* **313**, 812–815.

YOUSSOUFIAN, H., HAMMER, S. M., HIRSCH, M. S. & MULDER, C. (1982). Methylation of the viral genome in an *in vitro* model of herpes simplex virus latency. *Proceedings of the National Academy of Sciences, U.S.A.* **79**, 2207–2210.

(*Received 19 September 1988*)