JOURNAL OF VIROLOGY, Sept. 1990, p. 4264–4273
0022-538X/90/094264-10$02.00/0

Vol. 64, No. 9

Contrasts in Codon Usage of Latent versus Productive Genes of Epstein-Barr Virus: Data and Hypotheses

SAMUEL KARLIN,* B. EDWIN BLAISDELL, AND GABRIEL A. SCHACHTEL
Department of Mathematics, Stanford University, Stanford, California 94305-2125

Received 8 March 1990/Accepted 7 June 1990

Epstein-Barr virus (EBV) has two different modes of existence: latent and productive. There are eight known genes expressed during latency (and hardly at all during the productive phase) and about 70 other ("productive") genes. It is shown that the EBV genes known to be expressed during latency display codon usage strikingly different from that of genes that are expressed during lytic growth. In particular, the percentage of S3 (G or C in codon site 3) is persistently lower (about 20%) in all latent genes than in nonlatent genes. Moreover, S3 is lower in each multicodon amino acid form. Also, the percentage of S in silent codon sites 1 of leucine and arginine is lower in latent than in nonlatent genes. The largest absolute differences in amino acid usage between latent and nonlatent genes emphasize codon types SSN and WWN (W means nucleotide A or T and N is any nucleotide). Two principal explanations to account for the EBV latent versus productive gene codon disparity are proposed. Latent genes have codon usage substantially different from that of host cell genes to minimize the deleterious consequences to the host of viral gene expression during latency. (Productive genes are not so constrained.) It is also proposed that the latency genes of EBV were acquired recently by the viral genome. Evidence and arguments for these proposals are presented.

Codon bias among synonymous codons has been documented for many genes in many species (e.g., references 9, 25, 26, 36, 47, and 70). In brief, the following observations and hypotheses have been made: codon usage is organism specific (25); highly expressed genes prefer codons corresponding to the most abundant isoaccepting tRNA species (35); codon usages in bacterial genes and yeast genes differ from codon usage in multicellular organisms (36). Other studies have examined codon choices in relation to codon context reflecting influences between two (or more) contiguous codons (7, 28, 44, 62, 72). These influences include G+C content (38) and balances of strong versus weak base pair bonding (7, 27), translational efficiency and fidelity (36), and maintenance of DNA and RNA secondary structure (69).

Many recent publications focus on the S composition of DNA in the first, second, and third codon positions (S is the single-letter base code for strong hydrogen-bonding bases, i.e., G or C, and W denotes the complementary weakly hydrogen-bonding bases, A or T). These publications concerned comparisons of the S3 (codon site 3, S-value) content of related genes (36), S content reflected in chromosomal banding (dark and light regional dispositions [3, 4, 23, 34]), CpG occurrences (methylation and HTF islands [6]), S content in relation to DNA melting temperature (3), and S mutational pressures and effects of DNA polymerase and repair systems (23, 66). Inquiries have also been made concerning S frequencies with reference to protein classifications based on function, tissue or organelle specificity, time of expression in organism development or in the cell cycle, abundance and stability, and evolutionary age (24, 64).

The diversity of S frequencies across species raises challenging questions on mechanisms and evolution. For the human collection of 294 genes (from GenBank sources but culled for repetitions and required to be of minimal 130-codon length), the average S frequencies at sites 1, 2, and 3 were 56, 41, and 63%, respectively, and were similar for other vertebrates (see Table 1). The *Drosophila melanogaster* collection (71 genes) has similar values of 55, 42, and 65%, respectively. The high S percentages in site 3 are especially intriguing. However, for the human gene data, the histogram of S3 frequencies reveals a multimodal distribution with its major mode at about 72% and subsidiary modes near 42 and 52% (see Fig. 1). The *Escherichia coli* (410 genes) S frequencies at the three codon sites are 61, 41, and 56%, but yeast (for *Saccharomyces cerevisiae*, 193 genes) possesses a disparate set of frequencies: 46, 37, and 38%. Of course, the nature of the sampling bias inherent in the gene collections for these species is unknown.

The S3 frequencies were ascertained for established genes and open reading frames (ORFs) in the complete genome of the human herpesviruses (for herpes simplex virus type 1 [HSV-1], 70 genes are 68.3% S [48]; for varicella-zoster virus [VZV], 67 genes are 46.0% S [17]; and for the Epstein-Barr virus [EBV] B95-8 strain, 78 genes are 59.9% S [1]). We dealt with almost all genes of these viruses, and a sampling bias would be minimal. The overall genome S3 frequencies are 83% for HSV, 68% for EBV, and 41% for VZV (see Table 1). A histogram of the S3 frequencies for the individual EBV genes (see Fig. 1) presents a bimodal plot with peaks at 52 and 72%. By contrast, the corresponding S3 histograms for the HSV-1 and VZV genes are unimodal and of similar narrow distribution (see Fig. 1). However, the means of S3 frequencies for the last two viruses differ drastically (40%). This large difference also holds for highly conserved amino acid segments (G. Schachtel, P. Bucher, B. E. Blaisdell, E. M. Mocarski, and S. Karlin, submitted for publication).

The S3-frequency histogram for EBV reveals a striking difference for genes of the EBV latent state. These invariably exhibit S3 frequencies in the smaller modal range, 45 to 55% (see Fig. 1 and Table 1), while the large majority of the genes assigned to EBV productive growth exhibit S3 frequencies lying near the principal mode (72%). This prominent contrast of S3 levels effectively distinguishes latent from nonlatent genes of EBV. Even more telling, the S3

* Corresponding author.

TABLE 1. Base percentages in several sequence sets[a] pooled over genes and amino acids

Gene set	No. of ORFs, codons	Base[b] % in codon site: 1							2							3						
		T	C	A	G	S	Y	M	T	C	A	G	S	Y	M	T	C	A	G	S	Y	M
HSV-1	70 37,883	14.2	30.8	17.3	37.7	68.5	45.0	48.1	24.9	31.4	22.5	21.2	52.6	56.4	53.9	10.2	45.1	7.0	37.6	82.8	55.3	52.2
VZV	67 35,473	21.1	21.7	26.5	30.7	52.4	42.8	48.2	28.6	26.1	27.8	17.5	43.6	54.7	53.9	30.5	20.2	28.3	21.0	41.2	50.7	48.5
EBV	76 37,289	16.2	27.7	23.0	33.1	60.9	43.9	50.7	25.6	29.2	25.3	19.9	49.1	54.8	54.5	17.6	36.6	14.8	31.0	67.6	54.2	51.4
EBV-non[c]	57 33,486	16.8	26.9	23.5	32.8	59.7	43.8	50.3	27.0	28.1	25.8	19.1	47.1	55.1	53.8	17.0	38.0	13.0	31.9	70.0	55.0	51.2
EBV-lat[c]	7 4,090	14.8	32.2	21.5	31.6	63.7	46.9	53.6	22.7	30.9	24.7	21.8	52.6	53.5	55.6	25.5	25.1	24.2	25.2	50.4	50.6	49.5
Human	294 132,536	16.9	23.6	26.9	32.6	56.2	40.5	50.5	26.7	22.4	32.1	18.8	41.2	49.1	54.5	20.6	33.0	16.5	29.8	62.9	53.7	49.5
Mouse	181 65,902	17.3	23.3	28.0	31.3	54.6	40.6	51.3	27.7	22.4	31.4	18.5	40.9	50.1	53.8	21.7	32.6	17.2	28.6	61.1	54.3	49.8
Chicken	59 23,409	14.9	21.8	29.5	33.9	55.7	36.7	51.3	27.3	22.1	33.7	16.9	39.1	49.4	55.8	20.2	33.2	15.2	31.4	64.6	53.4	48.4
D. melanogaster	71 35,729	16.7	22.9	27.8	32.5	55.5	39.6	50.7	25.5	23.9	32.9	17.8	41.6	49.3	56.7	19.9	34.6	15.0	30.4	65.1	54.5	49.7
Yeast	193 96,942	21.6	15.3	32.1	31.0	46.3	36.9	47.3	28.2	22.6	34.8	14.4	37.0	50.8	57.4	33.9	19.9	27.9	18.2	38.2	53.8	47.9
E. coli	410 159,311	14.2	24.2	25.0	36.6	60.8	38.3	49.2	29.3	22.7	30.2	17.8	40.5	51.9	52.9	26.2	27.8	17.4	28.7	56.4	54.0	45.1

[a] The sequence sets are described in the text.
[b] Base two-letter codes: S, C or G; W, A or T; Y, C or T; R, A or G; M, C or A; K, G or T.
[c] The seven latent ORFs of lengths ≥200 are EBNA 1 to EBNA 4, EBNA 6, LMP, and LMP2A. The 57 nonlatent ORFs are all others with a length ≥200 codons.

frequency of the pool of all latent genes is substantially below that of the pool of all nonlatent genes for each degenerate amino acid form (see Table 4). A degenerate amino acid form is a set of synonymous codons with the first two nucleotides in common.

What are processes and mechanisms that might explain the coexistence of distinct codon dialects associated with the two life states of EBV? Is there a corresponding divergence for proteins specific to the B-lymphocyte cell, the primary host to these viruses? What is the nature of the competitive demands between a virus and host for nucleotide, ribosome, and tRNA resources and of their influences on codon choices? Codon usage in association with latency can be considered in other viruses with multiple life modalities, e.g., distinctions in codon usage between the lysogenic versus lytic genes of λ bacteriophage.

It is convenient to have available some functional information on the latent genes of EBV. In the infected cell, EBV persists principally as a circular episome inside the nucleus. The latent presence of EBV with restricted copy numbers in infected B lymphocytes presumably induces immortalization (as it does in lymphoblastoid cell lines). Only in poorly characterized oropharyngeal tissue does the virus enter the productive cycle (50). There are about 85 substantial ORFs in the completely sequenced 172,282-base-pair (bp) genome of the B95-8 strain of EBV (1). In the latent state, about 10% of the viral genome is expressed as stable cytoplasmic poly(A) mRNA, including genes encoding the membrane proteins LMP1 and LMP2 and the nuclear antigens EBNA1 to EBNA6 (43, 56, 58). Current knowledge about EBV latent genes is limited; the production of EBNA1 is essential to the replication of EBV genomic plasmids (55). EBNA1 protein also transactivates the 21- by 30-bp *oriP* repeat region for enhancing transcriptional activation with various promoters (45, 68). Apparently, EBNA2 and LMP1 contribute to the immortalizing capacity of the virus (67, 71). The functions of other latent gene products are unknown (42, 67). Two genes, BZLF1 and BMLF1, are intimately associated with the initiation and further development of the productive cycle (16). The latent genes contain many long tandem repeats (amino acid and also DNA) and many distinctive charge configurations (8, 31). Neither of these characteristics is found in productive genes.

Detailed below (see Tables 1 to 6) are the statistics on codon preferences of EBV genes highlighting differences in silent-site S frequencies. The compilations comparing latent and nonlatent groupings of the genes are done for complete genes pooled over amino acids (see Fig. 1 and Table 3) and for each amino acid form pooled over genes (see Table 4). The listings also provide the corresponding pooled S frequencies for HSV-1 and VZV genes. Amino acid frequencies are compared below (see Table 6).

DATA AND RESULTS

Percentages of bases at various codon sites, pooling all genes and all amino acids, are shown in Table 1 for three herpesviruses and, for comparison, six nonviral species collections. Throughout this paper, we use the standard one-letter base code and one-letter amino acid code. Table 1 presents summary statistics in the four-letter alphabet (T, C, A, and G) and in the three two-letter alphabets. In all viral genomes and sequence collections of multicellular organisms, the departure of S3 from 50% dominates the departures of S1 and S2 from 50%.

The percentage of S3 is strikingly lower in all latent than in nonlatent genes, pooled over amino acids. Figure 1 shows histograms of the percentages of S3 for the individual genes of the complete genomes of HSV-1, VZV, and EBV (the complete sequence of human cytomegalovirus was not available to us) and, for comparison, for large collections of human, mouse, chicken, *Drosophila melanogaster*, yeast, and *E. coli* genes. The distributions of HSV-1 and VZV are very compact, with few outliers but widely divergent modes at 82 and 42%, respectively. The distributions of the collections of human, mouse, and chicken genes are much broader, with indications of multimodality. The distribution of EBV, unlike that of the other herpesviruses, appears bimodal. Summary statistics for the distributions are given in Table 2.

The EBV collection includes virtually all the ORFs identified by Baer et al. (1) and two recently identified multiply spliced latent genes, EBNA5 (20) and LMP2A (43, 58). Of the 19 genes in the lower modal range of the EBV distribution, 8 (EBNA1 through EBNA6, LMP, and LMP2A) are recognized to be latent and 1 (BZLF1) induces disruption of

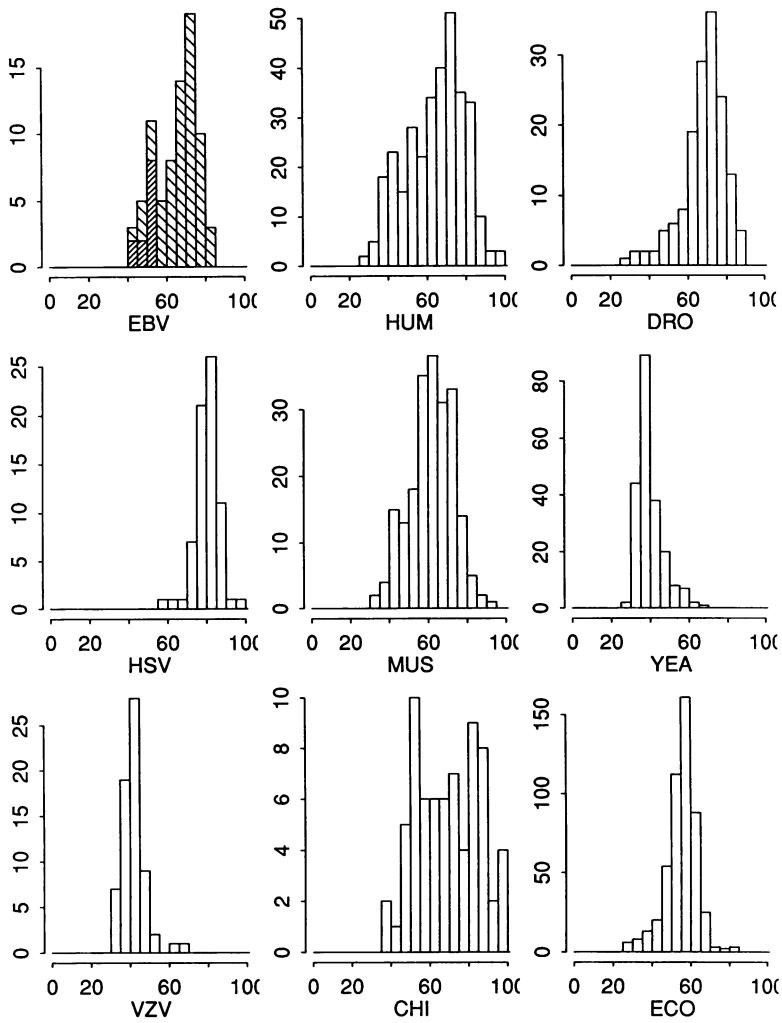

FIG. 1. Histograms of codon site 3 S3 percentages, pooled over amino acids, in several gene collections: complete genomes of EBV, HSV-1, and VZV and samples of human (HUM), mouse (MUS), chicken (CHI), *D. melanogaster* (DRO), yeast (YEA), and *E. coli* (ECO) genes. Counts of known latent genes in EBV are distinguished by denser shading.

the latent state. Three more (BLRF3, BERF2a, and BERF3) are probably leader exons spliced to BERF1 (EBNA3), BERF2b (EBNA4), and BERF4 (EBNA6), respectively. Fused transcripts BLRF3-BERF1 (10) and BERF3-BERF4 (12) have been found. In summary, in EBV, every gene known to be associated with latency lies in the lower modal range.

TABLE 2. Summary statistics of the codon S3 frequencies in genes, pooled over amino acids

Gene source	S3 frequency (%)	
	Avg	SD
EBV	65.1	10.6
HSV	80.3	6.2
VZV	41.7	6.1
Human	64.6	14.9
Mouse	62.0	11.3
Chicken	69.1	16.0
D. melanogaster	68.2	11.4
Yeast	39.9	6.9
E. coli	55.1	8.1

The remainder of the data presented here examines in detail the difference in S percentage between latent and productive genes of EBV. Since we wish to make comparisons of S3 usage for individual genes, we limit the data to EBV genes encoding 200 amino acids or more in order to avoid problems of gross statistical fluctuations. These include 7 latent genes (EBNA1 through EBNA4, EBNA6, LMP, and LMP2A) and 57 other, putatively nonlatent, genes. Only 4 of the 57 nonlatent genes are lower than 55%: BLLF1a, BZLF2, BZLF1, and BRRF2. (In these analyses, all of the 240-residue [Gly from GGA or GGG and Ala from GCA] repeat region containing only 15.7% S3 has been removed from EBNA1 and the run of 42 Pro from EBNA2.) Pooling over genes and amino acids, the frequency of S3 in the latent genes is significantly lower than that in the 57-gene nonlatent set, 50.4% compared with 70.0%, but 4 to 5% higher in site 1 and in site 2 (Table 1).

The latent genes of EBV occur in three clusters (Table 3): (i) LMP1 (BNLF1) and LMP2 located proximal and strad-

TABLE 3. Locations, gene names, lengths, S3 usage (pooled over amino acids), quartiles of S3 percentages, and designations of known latency of ORFs of EBV

Coordinate	Gene	Length (amino acids)	S3 usage (%)	Quartile[a]
57	LMP2A[b]	497	52.7	L1
1,736	BNRF1	1,318	72.2	3
9,675	BCRF1	170	55.3	2
14,385	EBNA5	176	52.8	L1
48,429	BYRF1	470	43.4	L1
54,376	BHRF1	191	42.9	1
55,982	BFLF2	318	65.1	2
56,910	BFRF2	657	68.9	3
56,951	BFLF1	525	70.3	3
58,891	BFRF1	336	71.4	3
61,456	BFRF3	193	63.7	2
62,081	BPLF1	3,149	71.7	3
71,523	BOLF1	1,239	79.7	4
75,238	BORF1	364	74.2	4
76,407	BORF2	826	74.2	4
78,900	BaRF1	302	73.2	4
79,899	BMRF1	404	68.6	3
81,118	BMRF2	357	70.3	3
82,746	BMLF1	459	65.4	2
84,260	BSLF1	874	73.7	4
86,924	BSRF1	218	76.6	4
87,461	BLLF2	148	50.0	1
88,295	BLRF2	162	49.4	1
88,547	BLRF1	102	70.6	3
89,433	BLLF1a	907	50.9	1
89,569	BLLF3	278	77.3	4
92,243	BLRF3	119	53.8	L1
92,646	BERF1	839	53.4	L1
95,353	BERF2a	123	46.3	L1
95,725	BERF2b	840	52.5	L1
98,323	BERF3	148	52.0	L1
98,805	BERF4	872	50.3	L1
102,116	BZLF2	223	46.2	1
103,155	BZLF1	200	45.5	1
103,369	BRLF1	605	66.9	2
105,182	BRRF1	310	69.4	3
106,302	BRRF2	537	54.6	1
107,950	BKRF1	402	40.0	L1
109,958	BKRF2	137	63.5	2
110,275	BKRF3	281	66.5	2
111,107	BKRF4	226	58.4	2
111,833	BBLF4	809	78.5	4
114,204	BBRF1	613	74.9	4
115,843	BBRF2	313	78.0	4
116,784	BBLF3	201	70.6	3
117,416	BBLF2	555	70.8	3
119,137	BBRF3	405	80.7	4
120,750	BBLF1	75	54.7	1
120,932	BGLF5	470	66.8	2
122,328	BGLF4	455	72.7	4
123,944	BGLF3	332	63.9	2
124,938	BGRF1	325	63.4	2
125,866	BGLF2	336	61.9	2
126,854	BGLF1	507	62.5	2
128,347	BDLF4	225	71.1	3
129,188	BDRF1	387	77.3	4
130,365	BDLF3	234	59.4	2
131,130	BDLF2	420	60.7	2
132,403	BDLF1	301	68.1	3
133,324	BcLF1	1,381	71.4	3
137,862	BcRF1	618	69.4	3
139,642	BTRF1	425	70.6	3
140,919	BXLF2	706	67.3	2
143,041	BXLF1	607	59.0	2
144,860	BXRF1	248	60.9	2
145,416	BVRF1	570	77.7	4
147,927	BVRF2	605	70.6	3
149,782	BILF2	248	58.9	2
152,164	BILF1	312	73.4	4
153,702	BALF5	1,015	82.0	4
156,752	BALF4	857	76.0	4
159,312	BALF3	789	78.6	4
161,387	BALF2	1,128	83.5	4
164,858	BALF1	220	78.6	4
165,504	BARF1	221	68.3	3
166,498	LMP2A[b]	497	52.7	L1
167,001	BNLF2b	101	67.3	2
167,307	BNLF2a	60	68.3	3
168,163	BNLF1	386	54.1	L1

[a] Quartile of S3 percentage; L designates a known latent gene.
[b] LMP2A is entered twice: once for exon 1 at coordinate 166,498 and once for exons 2 to 8 at coordinate 57.

dling, respectively, the long terminal repeat; (ii) the coordinate range 12,000 to 50,000 containing the genes encoding EBNA5 derived from the 3-kbp tandem repeat region (coordinates 12,000 to 45,000) and EBNA2 (BYRF1) immediately following the 3-kbp repeats; and (iii) the 18-kbp stretch from coordinates 92,000 to 110,000 which includes the gene domains for EBNA3, -4, -6, and -1 (BLRF3, BERF1), (BERF2a, BERF2b), (BERF3, BERF4) BKRF1, and nonlatent low-S3 BZLF1, BZLF2, and BRRF2. At the other extreme, there are 13 genes with S3 frequencies exceeding 75%, and all but three are located in the coordinate region from 112,000 to 165,000, highlighting the succession of all five BALF ORFs. These 13 genes include the DNA polymerase (BALF5) and ribonucleotide reductase (BaRF1, BORF2). These are among the most highly conserved genes of EBV, HSV-1, and VZV. Here conservation refers to the amino acid level; codon usage in these genes is strikingly disparate. In fact, most genes of the U_L section of HSV-1 and VZV are similar but differ in S3 frequencies by 40%, and this applies even for highly conserved amino acid segments (Schachtel et al., submitted).

The percentage of S3 is strikingly lower in latent than in nonlatent genes for all 21 degenerate amino acid forms, pooled over genes. The percentages of S3 pooled over genes in the latent and nonlatent subsets are shown in Table 4. For all degenerate amino acid forms, the S3 percentage difference of nonlatent versus latent genes is positive and remarkably constant, 19.3 ± 5.6 (average ± standard deviation), omitting the outlying I3 (isoleucine) and F2 (phenylalanine). Degeneracy-2 and -4 forms for L, R, and S are listed separately. Phenylalanine seems to be special in these viruses, showing very little codon bias.

The percentage of S in silent codon sites 1 of L (leucine) and R (arginine) is lower in latent than in nonlatent genes, pooled over genes. It is well known that L and R have silent choice between S and W in codon site 1 but that S (serine) does not have such a choice in either site 1 or site 2. Table 5 shows that usage of the degeneracy-2 forms with W in site 1 in preference to degeneracy-4 forms with S in site 1 is greater in latent than in nonlatent genes for L and R and differs to about the same extent as does the usage of S3 for both amino acids. A similar preference for the degeneracy-2 form is shown for VZV compared with that of HSV-1. The usages of the degeneracy-4 and degeneracy-2 forms of S (serine) are

TABLE 4. S3 usage in 21 degenerate amino acid forms, pooled over genes, for nonlatent versus latent genes of EBV and for HSV-1 versus VZV

Amino acid[a]	% S3 usage in genes					
	Nonlatent	(dif1)[b]	Latent	HSV-1	(dif2)[c]	VZV
V4	78.8	22.8	56.0	84.2	48.2	36.0
P4	62.0	23.5	38.5	85.0	34.2	50.8
T4	68.6	22.1	46.5	89.7	39.7	50.0
A4	72.5	22.3	50.2	89.9	38.8	51.1
G4	67.4	15.7	51.7	82.3	43.5	38.8
L4	79.4	23.0	56.4	85.4	46.4	39.0
L2	72.1	10.6	61.5	83.6	48.4	35.2
R4	71.6	25.5	46.1	84.4	39.9	44.5
R2	65.9	16.6	49.3	74.7	37.9	36.8
S4	59.4	8.7	50.7	83.9	36.3	47.6
S2	74.2	33.0	41.2	85.2	38.9	46.3
Y2	68.2	18.7	49.5	78.4	40.9	37.5
H2	69.1	18.7	50.4	81.8	41.5	40.3
Q2	75.7	19.5	56.2	82.0	44.5	37.5
N2	65.4	14.4	51.0	88.2	41.9	46.3
K2	74.7	15.8	58.9	72.3	43.0	29.3
D2	66.0	16.5	49.5	81.5	42.2	39.3
E2	73.5	23.1	50.4	79.2	42.3	36.9
C2	69.3	15.5	53.8	64.4	35.9	28.5
F2	46.0	4.0	42.0	47.7	31.8	15.9
I3	54.3	24.5	29.8	67.2	49.2	18.0
Avg	68.3	18.8	49.5	79.6	41.2	38.4
SD	8.0	6.4	7.2	9.9	4.6	9.6
Mdn	69.1	18.7	50.4	82.3	41.5	38.8
1/7	62.0	14.4	42.0	72.3	36.3	29.3
6/7	74.7	23.5	56.2	85.4	46.4	47.6

[a] One-letter code for degenerate amino acid forms with appended degeneracy. The average is unweighted. Mdn, Median; 1/7, first septile (value of rank 4); 6/7, sixth septile (value of rank 18).
[b] dif1, Nonlatent value minus latent value.
[c] dif2, HSV-1 value minus VZV value.

close in all collections to the ratio of 4:2 expected for random codon usage.

The percentage of S3 is generally lower in latent than in nonlatent genes for each degenerate amino acid form for each gene. Since the sample size for some amino acid forms and genes can be quite small, the data concerning the percentages of S3 are restricted to those genes of more than 300 residues. They number 50 and include 6 latent genes. The rank of the S3 percentage for each of these 6 genes among the total of 50 is generally low for each degenerate amino acid form (except isoleucine).

The largest absolute differences in amino acid usage between latent and nonlatent genes emphasize SSN and WWN, pooled over genes. The emphasis on SSN and WWN in the largest absolute differences in amino acid usage between latent and nonlatent genes (Table 6 and 7) is caused mainly by an unusually high occurrence of P (proline) in latent genes, especially in the genes EBNA2, -3, -4, and -6. Amino acid Q (glutamine) is similarly high in EBNA3, EBNA4, and EBNA6. (P- and Q-rich regions have been associated with activating functions in several eucaryotic transcription factors, e.g., Sp1 and Oct1 [for a review, see reference 51].) Unlike EBNA2 through EBNA6, EBNA1 carries significantly high numbers of G (glycine) and R (arginine) residues. These residues may contribute to its being the only EBNA known to bind to DNA at the *oriP* region (39). (Recall that the 240-residue Gly and Ala repeat has been removed from EBNA1 and the 42-residue P run from EBNA2.) None of these four amino acids is significantly high in a nonlatent gene.

TABLE 5. Percentages of degeneracy-4 and degeneracy-2 forms in degeneracy-6 amino acids, pooled over genes

Amino acid (codon type)	EBV		HSV	VZV
	Nonlatent	Latent		
L4 (CTN)	86.6	73.7	89.0	49.2
L2 (TTR)	13.4	26.3	11.0	50.8
R4 (CGN)	65.1	54.3	92.5	76.1
R2 (AGR)	34.9	45.7	7.5	23.9
S4 (TCN)	67.3	64.7	67.9	74.2
S2 (AGY)	32.7	35.4	32.1	25.8

DISCUSSION AND HYPOTHESES

We report in this paper a striking contrast in the S3 frequencies of the genes of the latent state in EBV versus the genes of the productive state (see Data and Results for details). The histogram of S3 frequencies for the EBV genes shows a bimodal distribution (Fig. 1) with all of the latent genes occurring in the lower modal range, whereas the bulk of the genes expressed in the EBV productive state occur near the principal mode. Contrary to the EBV case, the corresponding histograms for the HSV-1 and VZV genes are unimodal (Fig. 1), with a similar compact shape but with means differing by a very large value, 40%. The pooled S3 frequencies of the EBV nonlatent genes average 70%, whereas the corresponding latent gene frequencies average 50% (Table 1 and Fig. 1). All known latent genes, without exception (EBNA1 to EBNA6, LMP1, and LMP2), and four other genes (BZLF1, BZLF2, BRRF2, and BLLF1a) have S3 frequencies in the range from 40 to 55%. The difference in codon usage is even more compelling because the S3 frequencies for latent genes versus nonlatent genes are reduced for all degenerate amino acid forms by about the same amount, an average of 19% (Table 4). The S content in site 3 of latent genes (Table 1) compared with that of lytic genes is lower by 20% but in sites 1 and 2 is higher by an average of about 5%. The increase in sites 1 and 2 largely reflects a higher percentage of proline (CCN) and glycine (GGN) residues in the latent genes compared with that in the lytic genes.

Codon preferences in evolutionary terms have been discussed from two main perspectives: in relation to translational accuracy and efficiency and in relation to selective and nonselective substitution biases operational during the DNA replication and repair process. Attributions of codon bias to the translational machinery pertain to codon-anticodon interactions (7, 25, 27) and to tRNA abundances (2, 35). Variation in tRNA availabilities is proffered by several authors (e.g., references 26, 36, 60, and 61) as a key factor of codon bias for the highly expressed genes of yeast and *E. coli*. The data for multicellular eucaryotic genomes in this context are sparse. Compartmental heterogeneity in eucaryotes (isochores) emphasizing S or W regional preeminence has been promulgated for many vertebrate genomes (5, 63). In this context, the data on human protein sequences featuring a preponderance of high S3 frequencies suggest that most proteins are encoded from S-rich isochores where DNA dissociation propensities are more restricted.

Our proposal to account for the marked disparity in codon

TABLE 6. Ordered differences in amino acid usage between nonlatent and latent genes of EBV, pooled over genes

Amino acid[a]	% Amino acid usage in genes		Base code[c]
	Nonlatent	Latent (dif)[b]	
P4	7.1	13.4 (−6.3)	SS
G4	6.8	9.1 (−2.3)	SS
Q2	3.7	5.9 (−2.2)	SW
L2	1.4	2.3 (−0.9)	WW
R2	2.3	3.1 (−0.8)	WS
W1	1.1	1.8 (−0.7)	WS
D2	4.6	4.9 (−0.3)	SW
M1	2.0	2.2 (−0.2)	WW
H2	2.5	2.6 (−0.1)	SW
S2	2.5	2.5 (0.0)	WS
I3	3.8	3.7 (0.1)	WW
E2	5.5	5.1 (0.4)	SW
R4	4.2	3.8 (0.4)	SS
S4	5.3	4.7 (0.6)	WS
C2	2.1	1.5 (0.6)	WS
T4	6.4	5.7 (0.7)	WS
Y2	3.0	2.1 (0.9)	WW
N2	3.2	2.1 (1.1)	WW
K2	3.3	2.0 (1.3)	WW
V4	6.7	5.4 (1.3)	SW
F2	3.9	2.5 (1.4)	WW
A4	9.3	7.1 (2.2)	SS
L4	9.3	6.5 (2.8)	SW

[a] One-letter code for degenerate amino acid form with appended degeneracy.
[b] dif, Latent value minus nonlatent value.
[c] One-letter base code in S-W alphabet of bases in codon sites 1 and 2 of the degenerate amino acid form.

TABLE 7. Ordered differences in amino acid usage between HSV-1 and VZV, pooled over genes

Amino acid[a]	% Amino acid usage in:		Base code[c]
	HSV-1	VZV (dif)[b]	
L2	1.1	4.9 (−3.8)	WW
I3	2.9	5.7 (−2.8)	WW
N2	2.3	4.2 (−1.9)	WW
K2	1.8	3.6 (−1.8)	WW
T4	6.0	7.4 (−1.4)	WS
S4	4.2	5.4 (−1.2)	WS
R2	0.7	1.6 (−0.9)	WS
Y2	2.6	3.4 (−0.8)	WW
F2	3.5	4.1 (−0.6)	WW
M1	1.7	2.1 (−0.4)	WW
Q2	3.1	3.5 (−0.4)	SW
E2	4.8	5.1 (−0.3)	SW
C2	1.8	2.1 (−0.3)	WS
H2	2.5	2.6 (−0.1)	SW
W1	1.1	1.1 (0.0)	WS
D2	5.4	5.3 (0.1)	SW
S2	2.0	1.9 (0.1)	WS
V4	7.1	6.9 (0.2)	SW
G4	7.9	6.0 (1.9)	SS
P4	8.7	5.8 (2.9)	SS
R4	7.8	4.8 (3.0)	SS
L4	8.8	4.9 (3.9)	SW
A4	12.5	7.4 (5.1)	SS

[a] One-letter code for degenerate amino acid form with appended degeneracy.
[b] dif, HSV-1 value minus VZV value.
[c] One-letter base code in S-W alphabet of bases in codon sites 1 and 2 of the degenerate amino acid form.

usage between latent and nonlatent genes of EBV centers on two principal concepts and related observations, coexistence and history. In a state of coexistence with hosts and other life forms, there are advantages to divergent codon choices during both transcriptional and translational processes akin to a partition of the cellular resources into separate ecological niches. For history, the latent genes in EBV are a more recent importation into the EBV lytic genome.

Coexistence of latent EBV in human B cells. An assessment of nucleotide contents for proteins specific to B-lymphocyte cells, including CD20, CR2 (B-cell receptor protein of EBV), Fc epsilon receptor CD23, and the complement component C3 totalling 3,388 codons, revealed levels of S3 frequencies for each of these genes concordant with those of the overall human protein collection (data not shown). We might surmise that as one of several means of limiting competition for various host resources, the latent genes of EBV, in contradistinction to the human genes, opted for divergent codon usage with latent genes concentrating on 50% S3 frequencies, compared with the human gene average of about 60%. Moreover, since the human genome is assayed at about 40% S content (59), a surplus of W nucleotides can be expected in the nucleotide pool. From this point of view, maintaining a virus-host coexistence utilizing more W base pairs at codon site 3 affords the EBV latent virus a means to reduce competition for the human tRNA supply. Along these lines, there are manifest contrasts in amino acid percentages. Thus, proline (encoded SSN) residue use is high, in excess of 10%, for latent genes (LMP2 excepted) but seemingly less than 6% for proteins specific to B cells. In agreement, among amino acids of the latent genes, proline codons are highest in W3 nucleotides (Table 4).

One can speculate that the latent genes in their ancestral regime already favored site 3 weakly bonding nucleotides or that they evolved these codon choices to facilitate DNA disassociation and transcription opportunities and thus to provide a more viable latent state. For productive ends, the EBV virus, seeking to proliferate and disseminate mature viral progeny, would expropriate human resources of all kinds. In particular, the codon choices among EBV lytic genes quite similar to those of human genes can foster severe virus-host competition.

The historical scenario. It is tempting to hypothesize that the EBV progenitor originally lived as a pure lytic virus, whereas the latent genes are of a more recent vintage sequestered into EBV at multiple stages from the human genome (or from other mammalian or viral sources). What is the evidence and rationale for this hypothesis? The following observations of the EBV latent state are germane. (i) No extant homolog to the latent genes is found in any other herpesvirus (48). (ii) Long-range splicing and processing or genomic rearrangement events seems to be unique to several latent EBV genes among herpesviruses (65). (iii) There are arguments for a human transposon ancestry of the *oriP* region (39). (iv) Significantly long tandem DNA repeats are prominent in all latent genes (8). (v) Latent genes are not or are very little expressed during productive growth. (vi) There are a common orientation and clustering of the EBNA genes in the EBV genome (Table 3). (vii) Multiple charge clusters of opposite signs uniquely characterize the EBNA proteins among all proteins encoded from EBV genes (8). (viii) Latent stages in other viruses exhibit low values of S3.

(i) Active latent genes are unique to EBV. Sequence comparisons among the human herpesviruses EBV, VZV, HSV-1, and human cytomegalovirus have revealed significant

similarities for about 5 to 15 structural genes, all of the lytic phase (18, 48; M. S. Chee, A. T. Bankier, S. Beck, R. Bohni, C. M. Brown, R. Corny, T. Horsnell, C. A. Hutchinson, T. Kouzarides, I. A. Martignetti, E. Preddie, S. C. Satchwell, P. Tomlinson, K. Weston, and B. G. Barrell, manuscript submitted). No substantial similarity was discerned relating to the EBV latent genes. Herpesvirus saimiri, squirrel monkey, classified in the gammaherpesvirus family, as is EBV, has also not revealed any counterpart to the EBNA genes. It is telling that of the genes in the lower modal range of S3 frequencies—the genes of the latent state and the other four ORFs, BLLF1 (glycoprotein [gp 350]), BZLF1 (the principal gene whose expression disrupts the latent state), BZLF2, and BRRF2—none possess homologs in any of the herpesviruses.

(ii) Long-range splicing seems to be unique to the EBNA genes. In herpesviruses, many proteins of multiple exons appear to be dispensable for productive growth (19). Moreover, no extensive primary transcripts have been observed in HSV-1, VZV, or cytomegalovirus. However, it is confirmed that all EBNA genes are encoded in the same orientation and are composed of a number of common and unique exons (10–13, 57, 65) putatively extracted from a primary transcript extending up to 90 kbp. EBV resides latently in B cells equipped with machinery for implementing long-range genomic rearrangements without long transcripts being involved, as occurs with V-D-J recombination in the production of antibody molecules. A more extreme alternative relates to the possibility of "jumping polymerase" transcription or of "transplicing" mechanisms, for which precedents are known (14, 15).

(iii) The *oriP* region and latent genes may be of transposon source. In the work of Karlin and Blaisdell (39), following a detailed sequence analysis of the patterns, length, phases, and composition of the 21- by 30-bp *oriP* tandem palindromic repeat units, we proposed that the *oriP* region is a relic of an ancestral transposon acquired from the human host or from some other autonomous replicating source. Since *oriP* requires EBNA1 for its function (54, 55), EBNA1 was subsequently incorporated by some means into the EBV genome. Possible modes of binding EBNA1 to the *oriP* elements mediated by its singular charge distribution are described in the work of Karlin and Blaisdell (39). Perhaps proteins like EBNA1 are used in host replication events. However, hybridization experiments (30) have, as yet, failed to detect a remnant of the EBNA1 gene in the human or mouse genome, presumably because the hybridization capabilities were compromised by the 700-bp region of exclusive glycine and alanine nucleotides that allowed primarily glycine- and alanine-rich regions to be selected. We are not aware of hybridization tests screening for similarity to the *oriP* region of the human genome. Contrasted with the entire EBV genome, which is overall only 41% W, the *oriP* region is 63% W, the highest concentration of W in EBV among segments of at least 200 bp in length.

The reduced S3 frequencies in the latent genes resemble those of transposable genes in the following sense. Studies of codon preference in the *Drosophila* gene collection (137 genes; average S3, 60.2%) revealed a strikingly low value (average, 39.5%) for transposon genes (*gypsy*, *hobo*, *copia*, and *P*), comprising the most W-biased functional group of *Drosophila* genes. These comparisons were made among groups of genes for developmental control, various classes of cytoplasmic enzymes, nuclear proteins, membrane receptors, extracellular (secreted) proteins, and other proteins (data not shown).

The presence of two distinct origins of replication operational in two different life states seems to be unique to EBV among herpesviruses. The three independent *oriP* elements of HSV-1 are all associated with productive growth. No distinct latent genes have unambiguously emerged with HSV-1, VZV, human cytomegalovirus, and even herpesvirus saimiri. For example, a presumed latent gene of HSV-1, LAT (S3 percentage, 73.6%), does not, when mutated, appear to corrupt a viable HSV-1 life performance (33). Moreover, no protein product of LAT has been reported as yet.

(iv) There are many significant tandem repeats in the latent genes. Significant tandem multiple DNA repeats with few errors, extant in only 11 out of ~80 genes, are found in all 8 latent genes of EBV (e.g., see references 1 and 41). After removal of the significantly long repeat elements, the set of latent genes shows a large and consistent difference in codon usage from the set of lytic genes (see Results). The variability across strains in the number of copies with few errors suggests that many of these repeats are recent amplifications, perhaps the products of responses to challenges by the human immune system or to transposition events. Of the three nonlatent genes (BLLF1, BZLF1, and BPLF1) that carry significant DNA repeats, the first two possess relatively low S3 frequencies of 51 and 46%, respectively, consonant with the nature of the S frequencies among the latent genes (Table 3).

(v) Latent genes are hardly expressed in the productive cycle. Adhering to the historical scenario in the development of the EBV genome, the apparatus of the productive cycle existed prior to the introduction of the latent genes. In this perspective, latent genes find meager expression during productive growth, since they contribute no essential function to it.

(vi) Latent genes are clustered in the EBV genome. The organization (in the coordinates of the B95-8 strain) of the EBNA latent genes in the EBV genome can be regarded as localized to two clusters (Table 3). The coordinate range of 7,000 to 50,000 includes *oriP* (7,300 to 9,300), the 3-kbp repeats from coordinates 12,000 to 45,000 (which afford possible multiple copies of the EBNA5 gene), and EBNA2 (coordinates 48,000 to 50,000). Two promoter sequences of the putative long transcripts of the EBNA genes have been mapped to the 3-kbp repeats and to coordinates near 11,650 preceding them (10). The 18-kbp segment from coordinates 92,243 to 110,000 contains 12 genes in the order BLRF3, BERF1 (EBNA3), BERF2a, BERF2b (EBNA4), BERF3, BERF4 (EBNA6), BZLF2, BZLF1, BRLF1, BRRF1, BRRF2, and BKRF1 (EBNA1). All but two of these (BRLF1 and BRRF1) have S contents between 40 and 55% (Table 3). EBNA3, EBNA4, and EBNA6 all lack a start ATG codon, which can be remedied by invoking the two-exon hypothesis (12, 52, 53, 56) joining ORFs BLRF3 and BERF1 (these indeed are found joined together in a processed mRNA among the collection secured by Bodescot et al. [10]), joining exon BERF2a with BERF2b (found in a protein [56]), and joining exons BERF3 and BERF4 (the latter also found in a processed mRNA [12]). The two-exon unions are generally concordant in lengths, in amino acid and codon usage, in significant charge and repeat structures, and in overall S3 frequencies and may have a common ancestor.

In the third group of latent genes, LMP1 (three exons) and LMP2 (eight exons) are proximal and straddle, respectively, the LTR; the region of the long terminal repeat is spliced out in achieving LMP2. It is conceivable that the total region of coordinates about from 1 to 50,000 (encompassing part of

LMP2, the *oriP* region, the 3-kbp repeats, and EBNA2) occurred as a single addition to the EBV lytic genome. From this perspective, the origin of lytic replication of EBV recently mapped within the coordinate range of 53,000 to 54,500 (29) was ab initio proximal downstream to the long terminal repeat, a genomic organization not uncommon to many virion structures. Following this scenario, coordinates 1 to 50,000 and 92,000 to 112,000 were not part of an ancestral EBV lytic genome. Without exception (cf. section ii), the latent gene transcripts are all composed of at least two exons while few other genes of the productive state or of any of the herpesviruses invoke elaborate mechanisms of gene splicing. Herpesviruses intrinsically engage many fewer splicing mechanisms compared with the corresponding processes with eucaryotic proteins. The phenomenon of substantial splicing for the latent genes accords with the thesis of their recent derivation from the human host or some other eucaryotic source. Only two genes of HSV-1 have been established to engage to some extent in splicing, and both of these unite relatively close exons (48). Thus, it appears that productive herpesviruses employ little of the RNA processing machinery of the host.

(vii) There are many multiple charge clusters in latent genes. Karlin et al. (40) introduced methods to identify statistically significant clusters, runs, and periodic patterns of charged residues in protein sequences. Studies of the charge distribution of a large number of proteins revealed that most significant charge configurations occur in eucaryotic DNA viruses. A remarkable example is EBNA1 of EBV. This protein carries four statistically significant separate charge clusters, two anionic and two cationic. More specifically, the EBNA1 protein is distinguished in having, bracketing a long exclusively glycine-alanine 240-residue repetitive stretch, peptide domains of 40 to 80 residues that alternate in being highly positively charged and highly negatively charged (8). This structure probably allows these proteins broad flexibility for simultaneous binding to *oriP* and also in maintaining other strong ionic protein-protein interactions. Furthermore, the dimeric nature of EBNA1 (46) would permit, facilitated by the charge clusters, its simultaneous cooperative binding to both repeat regions of *oriP* and consequent induction of the EBV latent replication cycle (39).

Only EBNA1 to EBNA4 and EBNA6 among EBV proteins carry separate charge clusters of each sign. LMP1 carries a negative-charge cluster and a mixed-charge cluster. The only other protein of EBV, apart from those of the latent state with multiple charge clusters (a negative-charge cluster and one of mixed sign) is BMLF1, an important transactivator of the lytic cycle. It seems that the conjunction in the latent and immediate-early lytic genes of multiple charge configurations with nonoverlapping repeat regions may have a coordinated function (8). Parenthetically, the proteins encoded by LMP2, BZLF1, and BLLF1a each carry a single-charge cluster. Only 14 of all putative proteins of EBV have at least one single-charge cluster (8, 40).

(viii) There are similar S3 differences in latent and lytic genes in other viruses. The lysogenic active genes of the λ phage (*cI*, *rex a*, *rex b*, and *int*) have, respectively, S3 frequencies of 43, 40.5, 43.8, and 41% compared with those of lytic genes of lengths ≥100 codons, which average 55.7%. Thus, in terms of S3 frequencies, the difference of lysogenic and lytic genes of λ phage is parallel to the difference of latent and lytic genes of EBV. Only three contiguous genes of λ phage, ea 47, ea 31, and ea 59 (functions unknown), have S3 contents (average, 28%) smaller than those of the active genes of the lysogenic state. The λ phage structural genes (tail, capsid, and head components) have average S3 silent-site frequencies in excess of 68%. Lysogenic genes of λ phage show diminished S content at all codon sites (1, 2, and 3) compared with the S content of the lytic genes, a phenomenon that is compatible with a regional W versus S nucleotide preference.

It is familiar (37) that the λ phage genome divides into an S-rich half and a W-rich half, but this assessment is subject to some heterogeneity of nucleotide disposition. For example, gene N, mapped to the W half, whose product helps to turn off *c*II, is thereby involved in control of lysogeny, and is expressed in the lytic phase, carries 49.6% S3 content. The exonuclease, beta, and gamma recombination genes in the same genomic half average 55% S3.

Parenthetically, on the basis of organizational details and the split disposition of W versus S content in the λ phage genome, T. Cox (unpublished data) has advocated an historical chronology for the acquisition of the lysogenic λ phage life state subsequent to a progenitor lytic state. However, the proposal is moot since the λ-like phages p434 and p21 share many genes with authentic λ and embody a similar genomic partitioning into predominant W against S halves, suggesting a more remote ancestry for the whole λ genome (32).

Historical order. In view of the foregoing observations and discussions, a speculative but consistent historical sequence on the course of development of the latent state in EBV goes as follows. The *oriP* region was acquired initially via a transposon vehicle, as discussed earlier. EBNA1 known to activate replication from *oriP* was subsequently acquired into the EBV genome. It is well established that the conjunction of *oriP* and EBNA1 are sufficient to maintain an episomal existence in vitro. In fact, autonomous replication is robust in various vector constructs incorporating only *oriP* and EBNA1 (21) and putatively in Burkitt lymphoma and nasopharyngeal cell lines (42). We suggest that EBNA2 through EBNA6 and the LMPs were acquired at a later stage (or over several stages) and that their proteins serve to promote associations with host replication factors and DNA polymerase complexes in facilitating the EBV latent episomal maintenance. LMP1, a nuclear membrane-associated protein abundant in the latent state (1), is thought to anchor the EBV episome, thereby also contributing to a more stable latent life cycle process. Moreover, it has been proposed that EBNA2 through EBNA6 stimulate B-cell growth factor production and appropriate receptor protein expression (42) and thus help in perpetuating and distributing EBV episomes to other B cells.

ACKNOWLEDGMENTS

We benefited much from discussions and comments on the manuscript from M. Botchan, M. Calos, A. Campbell, and especially E. Mocarski.

S.K. was supported in part by Public Health Service grants GM39907-02 and GM10452-26 and by National Science Foundation grant DMS86-06244. G.A.S. was supported in part by DAAD Ref. 312.

LITERATURE CITED

1. **Baer, R., A. T. Bankier, M. D. Biggin, P. L. Deininger, P. J. Farrell, T. J. Gibson, G. Hatfull, G. S. Hudson, S. C. Satchwell, C. Séguin, P. S. Tuffnell, and B. G. Barrell.** 1984. DNA sequence and expression of the B95-8 Epstein-Barr virus genome. Nature (London) **310**:207–211.
2. **Bennetzen, J. L., and B. D. Hall.** 1982. Codon selection in yeast. J. Biol. Chem. **257**:3026–3031.

3. **Bernardi, G., and G. Bernardi.** 1986. Compositional constraints and genome evolution. J. Mol. Evol. **24**:1–11.
4. **Bernardi, G., D. Mouchirond, C. Gautier, and G. Bernardi.** 1988. Compositional patterns in vertebrate genomes: conservation and change in evolution. J. Mol. Evol. **28**:7–18.
5. **Bernardi, G., B. Olofsson, J. Filipski, M. Zerial, G. Cuny, M. Meunier-Rotival, and F. Rodier.** 1985. The mosaic genome of warm-blooded vertebrates. Science **228**:953–957.
6. **Bird, A. P.** 1986. CpG-rich islands and the function of DNA methylation. Nature (London) **321**:209–213.
7. **Blaisdell, B. E.** 1983. Choice of base at silent codon site 3 is not selectively neutral in eucaryotic structural genes: it maintains excess short runs of weak and strong hydrogen bonding bases. J. Mol. Evol. **19**:226–236.
8. **Blaisdell, B. E., and S. Karlin.** 1988. Distinctive charge configurations in proteins of the Epstein-Barr virus and possible functions. Proc. Natl. Acad. Sci. USA **85**:6637–6641.
9. **Blake, R. D., and P. W. Hinds.** 1984. Analysis of the codon bias in E. coli sequences. J. Biomol. Struct. Dynam. **2**:593–606.
10. **Bodescot, M., O. Brison, and M. Perricaudet.** 1986. An Epstein-Barr virus transcription unit is at least 84 kilobases long. Nucleic Acids Res. **14**:2611–2620.
11. **Bodescot, M., B. Chambraud, P. Farrell, and M. Perricaudet.** 1984. Spliced RNA from the IRI-U2 region of Epstein-Barr virus: presence of an opening reading frame for a repetitive polypeptide. EMBO J. **3**:1913–1917.
12. **Bodescot, M., and M. Perricaudet.** 1986. Epstein-Barr virus mRNAs produced by alterative splicing. Nucleic Acids Res. **17**:7103–7114.
13. **Bodescot, M., M. Perricaudet, and P. J. Farrell.** 1987. A promoter for the highly spliced EBNA family of RNAs of Epstein-Barr virus. J. Virol. **61**:3424–3430.
14. **Boothroyd, J. C.** 1989. Transplicing of RNA. Nucleic Acids Mol. Biol. **3**:216–230.
15. **Borst, P., R. Beane, and H. F. Tabek.** 1989. A fused chimeric protein made in human cells. Cell **58**:421–422.
16. **Chevallier-Greco, A., E. Manet, P. Chavrier, C. Mosnier, J. Daillie, and A. Sergeant.** 1986. Both Epstein-Barr virus (EBV) encoded transactivating factors, EB1 and EB2, are required to activate transcription from an EBV early promoter. EMBO J. **5**:3243–3249.
17. **Davison, A. J., and J. E. Scott.** 1986. The complete DNA sequence of varicella-zoster virus. J. Gen. Virol. **67**:1759–1816.
18. **Davison, A. J., and P. Taylor.** 1987. Genetic relations between varicella-zoster virus and Epstein-Barr virus. J. Gen. Virol. **68**:1067–1081.
19. **DeLuca, N. A., and P. A. Schaffer.** 1988. Physical and functional domains of the herpes simplex virus transcriptional regulatory protein ICP4. J. Virol. **62**:732–743.
20. **Dillner, J., B. Kallin, H. Alexander, I. Ernberg, M. Uno, Y. Ono, G. Klein, and R. A. Lerner.** 1986. An Epstein-Barr virus (EBV)-determined nuclear antigen (EBNA5) partly encoded by the transformation-associated Bam WYH region of EBV DNA: preferential expression in lymphoblastoid cell lines. Proc. Natl. Acad. Sci. USA **83**:6641–6645.
21. **Dubridge, R. B., P. Tang, H. C. Hsia, P. Leong, J. H. Miller, and M. P. Calos.** 1987. Analysis of mutation in human cells by using an Epstein-Barr virus shuttle system. Mol. Cell. Biol. **7**:379–387.
22. **Farrell, P. J.** 1989. Epstein-Barr virus genome. Adv. Viral Oncol. **8**:103–132.
23. **Filipski, J.** 1987. Correlation between molecular clock ticking, codon usage, fidelity of DNA repair, chromosome banding and chromatin compactness in germline cells. FEBS Lett. **217**: 184–186.
24. **Filipski, J., J. Salinas, and F. Rodier.** 1987. Two distinct compositional classes of vertebrate gene-bearing DNA stretches, their structures and possible evolutionary origin. DNA **6**:109–118.
25. **Grantham, R., C. Gautier, M. Gouy, M. Jacobzone, and R. Mercier.** 1981. Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res. **9**:r43–r74.
26. **Grantham, R., C. Gautier, M. Gouy, R. Mercier, and A. Pave.** 1980. Codon catalog usage and the genome hypothesis. Nucleic Acids Res. **8**:r49–r62.
27. **Grosjean, H., and W. Fiers.** 1982. Preferential codon usage in prokaryotic genes: the optimal codon-anticodon interaction energy and the selective codon usage in efficiently expressed genes. Gene **18**:199–209.
28. **Gutman, G. A., and G. W. Hatfield.** 1989. Nonrandom utilization of codon pairs in Escherichia coli. Proc. Natl. Acad. Sci. USA **86**:3699–3703.
29. **Hammerschmidt, W., and B. Sugden.** 1988. Identification and characterization of oriLyt, a lytic origin of DNA replication of Epstein-Barr virus. Cell **55**:427–433.
30. **Heller, M., E. Flemington, E. Kieff, and P. Deininger.** 1985. Repeat arrays in cellular DNA related to the Epstein-Barr virus IR3 repeat. Mol. Cell. Biol. **5**:457–465.
31. **Hennessy, K., and E. Kieff.** 1983. One of two Epstein-Barr virus nuclear antigens contains a glycene-alanine copolymer domain. Proc. Natl. Acad. Sci. USA **80**:5665–5669.
32. **Highton, P. J., and M. Whitfield.** 1975. Similarities between the DNA molecules of bacteriophages 414, λ, and 21, determined by denaturation and electron microscopy. Virology **63**:438–446.
33. **Ho, D. Y., and E. S. Mocarski.** 1989. Herpes simplex virus latent RNA (LAT) is not required for latent infection in the mouse. Proc. Natl. Acad. Sci. USA **86**:7596–7600.
34. **Holmquist, G. P.** 1989. Evolution of chromosome bands: molecular ecology of noncoding DNA. J. Mol. Evol. **28**:469–486.
35. **Ikemura, T.** 1981. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. J. Mol. Biol. **151**:389–409.
36. **Ikemura, T.** 1985. Codon usage and tRNA content in unicellular and multicellular organisms. Mol. Biol. Evol. **2**:13–34.
37. **Inmar, R. B., and R. L. Baldwin.** 1964. Helix random coil transitions in DNA homopolymer pairs. J. Mol. Biol. **8**:452–469.
38. **Jukes, T. H., and V. Bhushan.** 1986. Silent nucleotide substitutions and G+C content of some mitochondrial and bacterial genes. J. Mol. Evol. **24**:39–44.
39. **Karlin, S., and B. E. Blaisdell.** 1987. A model for the development of the tandem repeat units in the EBV ori-P region and a discussion of their possible function. J. Mol. Evol. **25**:215–229.
40. **Karlin, S., B. E. Blaisdell, E. S. Mocarski, and V. Brendel.** 1989. A method to identify distinctive charge configurations in protein sequences, with application to human herpesvirus polypeptides. J. Mol. Biol. **205**:165–177.
41. **Karlin, S., and G. Ghandour.** 1985. Comparative statistics for DNA and protein sequences: single sequence analysis. Proc. Natl. Acad. Sci. USA **82**:5800–5804.
42. **Klein, G.** 1989. Viral latency and transformation: the strategy of Epstein-Barr virus. Cell **58**:5–8.
43. **Laux, G., M. Perricaudet, and P. J. Farrell.** 1988. A spliced Epstein-Barr virus gene expressed in latently transformed lymphocytes is created by circularization of the linear viral genome. EMBO J. **7**:769–774.
44. **Lipman, D. J., and W. J. Wilbur.** 1983. Contextual constraints on synonymous codon choice. J. Mol. Biol. **163**:363–376.
45. **Luka, J., T. Kreofsky, G. R. Pearson, K. Hennessy, and E. Kieff.** 1984. Identification and characterization of a cellular protein that cross-reacts with the Epstein-Barr virus nuclear antigen. J. Virol. **52**:833–838.
46. **Luka, J., T. Lindahl, and G. Klein.** 1978. Purification of the Epstein-Barr virus-determined nuclear antigen from Epstein-Barr virus-transformed human lymphoid cell lines. J. Virol. **27**:604–611.
47. **Maruyama, T., T. Gojobori, S. Aota, and T. Ikemura.** 1986. Codon usage tabulated from the GenBank genetic sequence data. Nucleic Acids Res. **14**:r151–r197.
48. **McGeoch, D. J., M. A. Dalrymple, A. J. Davison, A. Dolan, M. C. Frame, D. McNab, L. J. Perry, J. E. Scott, and P. Taylor.** 1988. The complete DNA sequence of the long unique region of the genome of herpes simplex virus type I. J. Gen. Virol. **69**:1531–1574.
49. **McGeoch, D. J., A. Donald, S. Donald, and F. J. Rixon.** 1985.

Sequence determination and genetic content of the short unique region of the genome of herpes simplex virus type 1. J. Mol. Biol. **181**:1–13.

50. **Miller, G.** 1985. Epstein-Barr virus, p. 563–589. *In* B. N. Fields, D. M. Knipe, J. L. Melnick, R. M. Chanock, B. Roizman, and R. E. Shope (ed.), Virology. Raven Press, New York.
51. **Mitchell, P. J., and R. Tjian.** 1989. Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. Science **245**:371–378.
52. **Petti, L., and E. Kieff.** 1988. A sixth Epstein-Barr virus nuclear protein (EBNA3B) is expressed in latently infected growth-transformed lymphocytes. J. Virol. **62**:2173–2178.
53. **Petti, L., J. Sample, F. Wang, and E. Kieff.** 1988. A fifth Epstein-Barr virus nuclear protein (EBNA3C) is expressed in latently infected growth-transformed lymphocytes. J. Virol. **62**:1330–1338.
54. **Rawlins, D. R., G. Milman, S. D. Hayward, and G. S. Hayward.** 1985. Sequence-specific DNA binding of the Epstein-Barr virus nuclear antigen (EBNA1) to clustered sites in the plasmid maintenance region. Cell **42**:859–868.
55. **Reisman, D., J. Yates, and B. Sugden.** 1985. A putative origin of replication of plasmids from Epstein-Barr virus is composed of two *cis*-acting components. Mol. Cell. Biol. **5**:1822–1832.
56. **Ricksten, A., B. Kallin, H. Alexander, J. Dillner, R. Fahraeus, G. Klein, R. Lerner, and L. Rymo.** 1988. *Bam*HI E region of the Epstein-Barr virus genome encodes three transformation-associated nuclear proteins. Proc. Natl. Acad. Sci. USA **85**:995–999.
57. **Sample, J., M. Hummel, D. Braun, M. Birkenback, and E. Kieff.** 1986. Nucleotide sequences of mRNAs encoding Epstein-Barr virus nuclear proteins: a probable transcriptional initiation site. Proc. Natl. Acad. Sci. USA **83**:5096–5100.
58. **Sample, J., D. Liebowitz, and E. Kieff.** 1989. Two related Epstein-Barr virus membrane proteins are encoded by separate genes. J. Virol. **63**:933–937.
59. **Shapiro, H. S.** 1976. Distribution of purine and pyrimidine in deoxyribonucleic acids, p. 241–281. *In* G. D. Fasman (ed.), CRC handbook of biochemistry and molecular biology, vol. 3. Nucleic acids. CRC Press, Inc., Cleveland.
60. **Sharp, P. M., and W.-H. Li.** 1986. An evolutionary perspective on synonymous codon usage in unicellular organisms. J. Mol. Evol. **24**:28–38.
61. **Shields, D. C., and P. M. Sharp.** 1987. Synonymous codon usage in Bacillus subtilis reflects both translational selection and mutational biases. Nucleic Acids Res. **15**:8023–8040.
62. **Shpaer, E. G.** 1986. Constraints on codon context in Escherichia coli genes: their possible role in modulating efficiency of translation. J. Mol. Biol. **188**:555–564.
63. **Smith, C. L., and C. R. Cantor.** 1986. Approaches to physical mapping of the human genome. Cold Spring Harbor Symp. Quant. Biol. **51**:115–122.
64. **Smith, T. F., W. W. Ralph, M. Goodman, and J. Czelusniak.** 1985. Codon usage in the vertebrate haemoglobins and its implications. Mol. Biol. Evol. **2**:390–398.
65. **Speck, S., and J. Strominger.** 1985. Analysis of the transcript encoding the latent Epstein-Barr virus nuclear antigen I: a potentially polycistronic message generated by long-range splicing of several exons. Proc. Natl. Acad. Sci. USA **82**:8305–8309.
66. **Sueoka, N.** 1988. Directional mutation pressure and neutral molecular evolution. Proc. Natl. Acad. Sci. USA **85**:2653–2657.
67. **Sugden, B.** 1989. An intricate route to immortality. Cell **57**:5–7.
68. **Sugden, B., and N. Warren.** 1989. A promoter of Epstein-Barr virus that can function during latent infection can be transactivated by EBNA1, a viral protein required for viral DNA replication during latent infection. J. Virol. **63**:2644–2649.
69. **Wada, A., and A. Sujama.** 1986. Local stability of DNA and RNA secondary structure and its relation to biological functions. Prog. Biophys. Mol. Biol. **47**:113–157.
70. **Wain-Hobson, S., R. Nussinov, R. J. Brown, and J. L. Sussman.** 1981. Preferential codon usage in genes. Genes **13**:355–364.
71. **Wang, F., C. D. Gregory, M. Rowe, A. B. Rickinson, D. Wang, M. Birkenbach, H. Kikutani, T. Kishimoto, and E. Kieff.** 1987. Epstein-Barr virus nuclear antigen 2 specifically induces expression of the B-cell activation antigen CD23. Proc. Natl. Acad. Sci. USA **84**:3452–3456.
72. **Yarus, M., and L. S. Folley.** 1985. Sense codons are found in specific contexts. J. Mol. Biol. **182**:529–540.