J. theor. Biol. (2001) **208,** 475–491
doi:10.1006/jtbi.2000.2233, available online at http://www.idealibrary.com on IDEAL®

Double-stranded RNA as a Not-self Alarm Signal: to Evade, most Viruses Purine-load their RNAs, but some (HTLV-1, Epstein-Barr) Pyrimidine-load

A. D. Cristillo*, J. R. Mortimer, I. H. Barrette, T. P. Lillicrap and D. R. Forsdyke†

Department of Biochemistry, Queen's University, Kingston, Ontario, Canada K7L3N6

(*Received on* 15 *August* 2000, *Accepted in revised form on* 2 *November* 2000)

For double-stranded RNA (dsRNA) to signal the presence of foreign (non-self) nucleic acid, self-RNA–self-RNA interactions should be minimized. Indeed, self-RNAs appear to have been fine-tuned over evolutionary time by the introduction of purines in clusters in the loop regions of stem-loop structures. This adaptation should militate against the "kissing" interactions which initiate formation of dsRNA. Our analyses of virus base compositions suggest that, to avoid triggering the host cell's dsRNA surveillance mechanism, most viruses purine-load their RNAs to resemble host RNAs ("stealth" strategy). However, some GC-rich latent viruses (HTLV-1, EBV) pyrimidine-load their RNAs. It is suggested that when virus production begins, these RNAs suddenly increase in concentration and impair host mRNA function by virtue of an excess of complementary "kissing" interactions ("surprise" strategy). Remarkably, the only mRNA expressed in the most fundamental form of EBV latency (the "EBNA-1 program") is purine-loaded. This apparent stealth strategy is reinforced by a simple sequence repeat which prefers purine-rich codons. During latent infection the EBNA-1 protein may evade recognition by cytotoxic T-cells, not by virtue of containing a simple sequence amino acid repeat as has been proposed, but by virtue of the encoding mRNA being purine-loaded to prevent interactions with host RNAs of either genic or non-genic origin.

1. Introduction

Double-stranded RNA (dsRNA) produces sequence-specific gene silencing in a wide variety of organisms. Although the mechanism is not understood, often silencing appears to operate at the post-transcriptional level resulting in inactivation or degradation of a specific mRNA species, leaving other mRNA species intact (Fire, 1999; Hamilton & Baulcombe, 1999). The response to dsRNA seems likely to have arisen as part of an intracellular mechanism for self/not-self-discrimination (Sharp, 1999). Consistent with this, dsRNA has long been known as a powerful inducer of the interferons, which relay alarm signals to other cells, thus inducing a general antiviral state (Kumar & Carmichael, 1998).

Protein synthesis is inhibited by very low concentrations of dsRNA (Ehrenfeld & Hunt, 1971; Hunter *et al.*, 1975). This involves activation of dsRNA-dependent protein kinase (PKR), which inhibits a protein involved in the initiation of protein synthesis. Potential evasive viral

* Present address: National Institutes of Health, 10 Center Drive, Bldg 10, Room 6C209, Bethesda, MD 20892, U.S.A.
† Author to whom correspondence should be addressed. E-mail: forsdyke@post.queensu.ca; full text versions of some of the cited references may be found at URL: http://post.queensu.ca/ ~ forsdyke/bioinfor.htm

0022–5193/01/040475 + 17 $35.00/0

strategies would include the acceptance of mutations to avoid formation of dsRNA, and inhibition of cell components required for the formation of, or the response to, dsRNA (Elia *et al*., 1996; Mittelstein Scheid, 1999).

One molecule of dsRNA can trigger interferon induction (Marcus, 1983). Yet dsRNA is formed transiently in large quantities during normal protein synthesis. This involves base pairing between anticodons at the tips of stem loops in tRNAs with cognate codons in mRNAs. However, the latter pairing involves at the most only five contiguous base pairs (Bossi & Roth, 1980), whereas more than 20 base pairs are required to activate PKR *in vitro* (Robertson & Mathews, 1996; Tian *et al*., 2000). While tRNA–mRNA interactions obviously do not trigger intracellular alarms, the fact that tRNA–mRNA interactions occur *so efficiently* in the cytosol suggests that mRNA–mRNA interactions might occur with *equal* efficiency (Izant & Weintraub, 1984; Melton, 1985; Bull *et al*., 1998). Among the RNA species of a potential virus host cell there might be two whose members, by chance, happened to have enough base complementarity for formation of a mutual duplex of a length sufficient to trigger alarms. Thus, there would have been an evolutionary selection pressure favouring mutations in host RNAs which decrease the possibility of their interaction with other "self"-RNAs in the same cell.

Indeed, this appears to have been assisted by "purine-loading" the loop regions of RNAs, thus avoiding the initial loop–loop "kissing" reactions which precede more complete formation of dsRNA (Eguchi *et al*., 1991). The excess of purines, observed both at RNA and at DNA levels (in mRNA-synonymous DNA strands), is sufficient to be detected as deviations from Chargaff's second parity rule (%A $\cong$ %T and %G $\cong$ %C for single strands; Forsdyke & Mortimer, 2001). These local deviations, or "skews", are found in a wide variety of organisms and their viruses, and can facilitate the identification of potential DNA open reading frames (ORFs), their transcription direction (Smithies *et al*., 1981; Dang *et al*., 1998; Bell *et al*., 1998; Bell & Forsdyke, 1999; Lao & Forsdyke, 2000), and origins of replication (Rocha *et al*., 1999).

Major questions remain. How does a virus trigger the dsRNA alarm resulting in a virus-specific hostile host response (Sharp, 1999)? While the majority of RNAs of many organisms and their viruses have purine-loading, why do certain viruses pyrimidine-load (Cristillo, 1998; Cristillo *et al*., 1998; Bell & Forsdyke, 1999)? We here present an analysis of the base composition of the genomes of various retroviruses and herpes viruses, which casts some light on these problems. Surprisingly, our results suggest adaptive roles for the simple sequence elements of viruses, and for the repetitive elements and non-genic DNA of their hosts.

2. Chargaff Difference Analysis

Transcribed duplex DNA has an mRNA-synonymous strand and an mRNA template strand. If transcription is to the right of the site of RNA polymerase initiation (promoter), the "top" strand (i.e. the sequence recorded in GenBank) is the mRNA-synonymous strand. Szybalski and co-workers (1966) showed that mRNA-synonymous strands (and hence the corresponding mRNAs) have purine-rich clusters. Combining this observation with Chargaff's second parity rule, it follows for mRNA-synonymous strands that purines in the clusters might be balanced by an equal number of dispersed pyrimidines, or that there might be small deviations from the second rule ("Chargaff differences") in favour of purines, as is indeed found.

Chargaff differences are simply the differences between the numbers of the classical Watson–Crick pairing bases in a nucleic acid segment ("AT-skew", "GC-skew"). The sign of the differences depends on the direction of subtraction, which in some previous work was determined alphabetically. For some purposes, Chargaff differences are best expressed as positive or negative base excesses, which may be combined to provide an index of the degree of purine-loading, with purine excesses scoring positive and pyrimidine excesses scoring negative (Lao & Forsdyke, 2000). If the open reading frames (ORFs) in a sequence are known, purine-loading indices may be calculated either directly from ORF base composition, or from codon-usage tables (Nakamura *et al*., 1999). The indices are then $1000[(A - T)/N]$ for the W bases (A and T) and $1000[(G - C)/N]$ for the S bases (G and C),

where N is the total number of bases (i.e. N = W + S). These two values are then summed to obtain a value for the overall purine-loading index (bases per kb).

Chargaff differences (absolute or %) may be calculated as A − T [or as (A − T)/W], and as G − C [or as (G − C)/S]. Here, A, T, G and C can be the frequency of each base in 1 kb sequence windows. This approach makes no assumption about the disposition of ORFs, and can be applied to uncharted DNA. When an ORF is located, values for windows whose centres overlap the ORF can be averaged to obtain an approximate value for that ORF. For the importance of 1 kb window sizes and other details see Dang *et al.* (1998) and Bell & Forsdyke (1999).

For AT-rich genomes, purine-loading is usually with respect to adenine, whereas for GC-rich genomes, purine-loading is usually with respect to guanine. In thermophilic bacteria, whatever the (G + C)%, purine-loading involves both purines (Lao & Forsdyke, 2000). An organism in which one or both Chargaff differences reflect the purine-loading of a significant excess of mRNAs is held to comply with "Szybalski's transcription direction rule" (Bell & Forsdyke, 1999).

3. Extremes of Positive and Negative Purine-loading

Indexes of purine-loading calculated from codon usage tables, although not taking into account 5′ and 3′ non-coding sequences, demonstrate the universality of the purine-loading phenomenon. Figure 1 shows how purine-loading is distributed across all species for which sequences of more than 3 ORFs and 2500 bases were available. The value for all human genes (excluding mitochondria) is 42 bases/kb, meaning that, on average, there are 42 more purines than pyrimidines for every kilobase of coding sequence. The shoulder with negative purine-loading values (pyrimidine-loading) corresponds mainly to mitochondrial genes, which are disproportionately abundant in GenBank.

Figure 2 shows a subset of these data (494 species), which corresponds to all eukaryotic viruses, with the exclusion of plant viruses. Most viruses have positive purine-loading indices, which are often greater than the average for

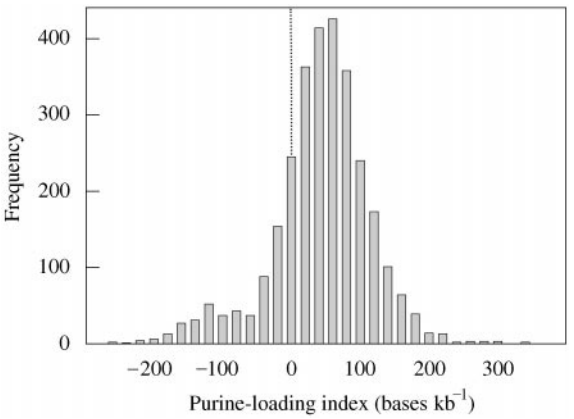

FIG. 1. Distribution of purine-loading among species. The purine-loading of coding regions was calculated from codon usage tables for all species (2958) represented in the August 1999 release of the GenBank database by more than three genes or more than 2500 bases. The purine-loading index (bases kb^{-1}) for a particular species was calculated as the sum of 1000[(G − C)/N] and 1000[(A − T)/N], where G, C, A, and T correspond to the numbers of individual bases, and *N* corresponds to the total number of bases, in the codon usage table. This measure of the purine-loading of RNAs disregards 5′ and 3′ non-coding sequences, including poly(A) tails.

human ORFs (vertical dashed line). There is a tendency for members of certain viral groups to be either extremely pyrimidine-loaded or extremely purine-loaded. Among retroviruses, at one extreme (highly pyrimidine-loaded) are human T-cell leukaemia (lymphotropic) virus-1 (HTLV-1) and some similar retroviral species, whereas at the other extreme are human immunodeficiency virus-1 (HIV-1) and some similar retroviral species (highly purine-loaded).

While the value for the agent of classical acute infectious hepatitis (Hepatitis A virus; 37) is close to the human average (42), that for the agent of classical serum hepatitis (Hepatitis B virus), which produces a chronic infection and requires reverse transcriptase for replication, is highly negative (−127). Hepatitis virus C, which usually produces chronic infections is negative (−27). Hepatitis virus D, which requires coinfection by Hepatitis B virus in order to be packaged, is very positive (+321). Hepatitis virus E, which causes acute infections in humans and emerges periodically from an unknown source (and hence may be chronic in that source), is very negative (−140).

Members of the Herpes virus group show less extreme Chargaff differences. The shoulder in

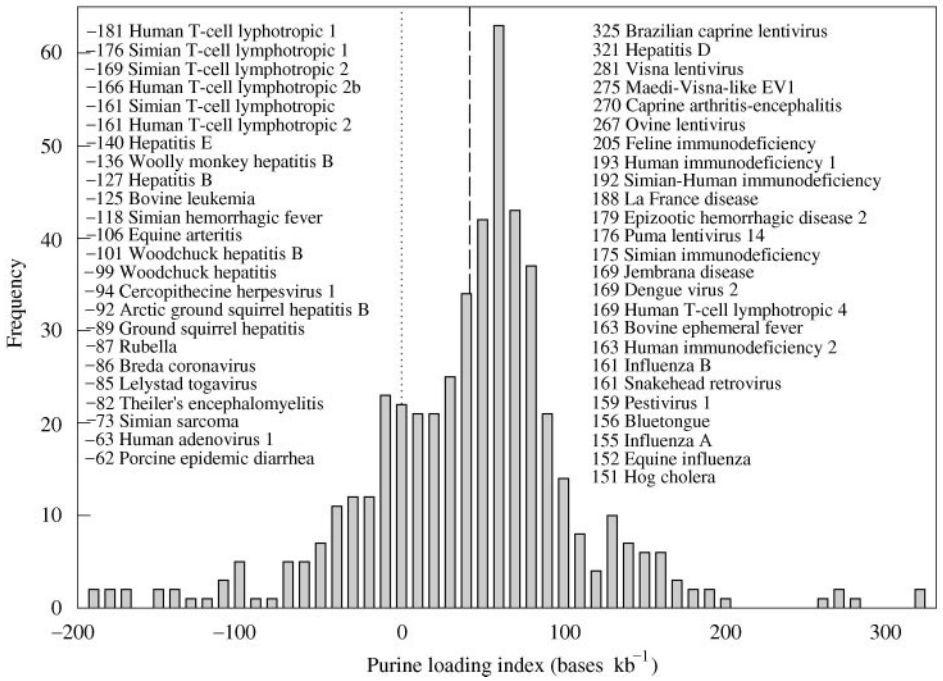

FIG. 2. Distribution of purine-loading indices among 494 viruses of eukaryotes, excluding plants. Viruses at the extremes of the distribution are listed in the figure. For example, Brazilian caprine lentivirus (325) and Hepatitis D virus (321) fall in the same interval, and provide a frequency value of 2 at the extreme right of the distribution. Positive indices were observed for murine polyoma virus (32), RSV (54), Vaccinia virus (64), SFV-1 (73), and SV40 (103).

Fig. 2 from −50 to −20 includes *Herpes simplex* virus-1 (HSV-1), Epstein-Barr virus (EBV) and human cytomegalovirus (HCMV)—all pyrimidine-loaded. Many other herpes-related viruses show overall purine-loading. The main ascending limb in the range +10 to +40 includes Varicella-Zoster virus (VZV), human herpesvirus 6 (HHV6), *Herpes saimiri* (HVS) and Ictalurid herpesvirus 1. Human herpesvirus 8 (HHV8; Kaposi sarcoma-associated virus) has a value of −1.

Several species with negative purine-loading indices produce chronic infections (see Section 8) and have high (G + C)% in ORFs. In general, there is an inverse relationship between species (G + C)% and purine-loading index. Thus, a linear regression plot of average (G + C) percentages calculated from codon usage tables for each viral species used in Fig. 2, relative to the corresponding purine-loading indices, has a downward slope which is significantly different from zero ($p < 0.0001$; data not shown). A more detailed analysis was made of the retroviral and herpes groups.

4. Extreme Purine-loading Indices in Retroviruses

In many genomes transcription directions vary so that, while total base composition of the "top" strand of DNA reflects Chargaff's second parity rule (e.g. G ≅ C), Chargaff differences for leftward-transcribing ORFs (e.g. C > G) tend to cancel out the differences for rightward-transcribing ORFs (e.g. G > C). Thus, with genome-sized sequence windows, compliance with Szybalski's transcription-direction rule, assessed as Chargaff differences, is not usually evident. However, retroviral genomes are transcribed entirely in one direction (considered by convention to be to the right), and mere examination of total base composition can illustrate some major points.

Table 1 compares the "top" strands of four retroviruses whose (G + C) percentages vary from 39.2 (Simian foamy virus-1; SFV-1; Kupiec *et al*., 1991), to 54.4% (Rous sarcoma virus; RSV; Schwartz *et al*., 1983). Three of the genomes obey Szybalski's transcription direction rule for rightward transcription (purines > pyrimidines). For AT-rich genomes (SFV-1, HIV-1; Ratner *et al*.,

TABLE 1
Chargaff differences of retroviral genomes

	Base composition					Chargaff differences†				
	W bases		S bases			W bases		S bases		Purine loading
Virus*	A	T	C	G	(C + G)%	A − T		G − C		index‡
SFV-1	4195	3696	2480	2601	39.2	A > T	38.5	G > C	9.3	48
HIV-1	3411	2163	1772	2373	42.6	A > T	128.4	G > C	61.8	190
HTLV-1	1983	1951	2932	1534	53.2	A > T	3.8	C > G	− 166.4	− 163
RSV	2216	2035	2362	2704	54.4	A > T	19.4	G > C	36.7	56

*SFV-1, simian foamy virus type 1 (Genbank accession number X54482); HIV-1, human immunodeficiency virus 1 (K03455); HTLV-1 human T cell leukaemia (lymphotropic) virus 1 (D13784); RSV, Rous sarcoma virus (D10652).

† Chargaff differences ("base skews" from equifrequency) are expressed as bases kb^{-1}. Values were calculated for each entire genome, which was not split into sub-windows. A, T, C, and G refer to numbers of bases.

‡ The sum of the purine excesses for the W bases and for the S bases, expressed as bases kb^{-1}. In the case of HTLV-1 there is a net pyrimidine excess, so the value is negative.

1985), A > T and G > C with the W bases providing the largest deviation from the parity rule. Similarly, for the GC-rich RSV genome, A > T and G > C, with the S bases providing the largest deviation. Calculations for different parts of the sequences (1 kb windows) show that these purine-rich patterns are sustained throughout, particularly in the case of the base pairs with the largest deviations from the parity rule (data not shown). A similar compliance with Szybalski's rule has been noted in the case of some other AT-rich viruses (SV40, polyoma, vaccinia; Smithies *et al.*, 1981; Bell & Forsdyke, 1999), again with the W bases providing the largest deviation. In contrast, HTLV-1 (Malik *et al.*, 1988) is GC-rich, yet C > G. There is only a weak tendency for compliance with Szybalski's rule with respect to the W bases. The C-rich pattern is sustained throughout the genome, affecting all ORFs.

5. Extreme Purine-loading Indices in Herpes Viruses

Table 2 shows that, like HTLV-1, GC-rich members of the herpes virus family (EBV, HSV-1) do not follow Szybalski's transcription direction rule (i.e. the majority of mRNAs are pyrimidine-loaded). This applies strongly to the S bases (e.g. C > G when transcription is to the right, and G > C when transcription is to the left), but less so to the W bases.

For most ORFs of AT-rich members of the Herpes virus family (HVS, VZV), A > T when transcription is to the right (i.e. they follow the transcription-direction rule), but the S bases do not follow the rule (C > G when transcription is to the right). When transcription is to the left, both the W and the S bases follow Szybalski's rule (pyrimidines > purines), but this is most evident in the case of the S bases. On balance, like AT-rich retroviruses (Table 1), the AT-rich herpes viruses seem to follow the rule.

From Table 2, it is evident that an individual ORF may be enriched for one of the W bases and one of the S bases, with four possible combinations (GA, GT, CA, CT). Table 3 shows that the ORFs of herpes viruses are distributed over all four groups, but with significant biases (bold number quartets with subscripts designating group). Rightward-transcribed ORFs which do not follow Szybalski's rule with respect to both the W and the S bases would be in the CT group. This group dominates in the cases of the two viruses of highest (G + C)% (EBV and HSV-1). Rightward-transcribed ORFs which follow Szybalski's rule with respect to both the W and the S bases would be in the GA group. This group is poorly represented in EBV and HSV-1. For EBV there are 37 rightward ORFs in the CT group, and only ten in the GA group. Since assignment of functions to ORFs is not yet complete, whether the group biases relate to function remains for future study. That there can be a

TABLE 2
*Distribution of pyrimidine- and purine-loading among leftward- and rightward-transcribing ORFs of various herpes-related viruses**

Direction of transcription		To left		To right	
Virus†	C + G %	W bases	S bases	W bases	S bases
HVS	35.0	T > A 22:15 ($p = 0.020$)	C > G 28:9 ($p = 0.00016$)	A > T 35:5 ($p < 0.00003$)	C > G 28:12 ($p = 0.0035$)
VZV	46.1	T > A 19:15 ($p = 0.15$)	C > G 27:7 ($p = 0.001$)	A > T 24:13 ($p = 0.02$)	C > G 25:12 ($p = 0.04$)
EBV	60.1	T > A 23:21 ($p = 0.24$)	G > C 30:14 ($p = 0.00007$)	T > A 49:37 ($p = 0.08$)	C > G 64:22 ($p < 0.00001$)
HSV	68.3	A > T 20:17 ($p = 0.15$)	G > C 29:8 ($p = 0.0007$)	T > A 24:17 ($p = 0.037$)	C > G 32:9 ($p = 0.00019$)

* Each sequence was examined using 1 kb windows moving in steps of 0.1 kb. Values for windows whose centres overlapped an ORF were averaged to obtain a value for that ORF. Each data set shows the relative proportion of ORFs with positive or negative Chargaff differences (i.e. skew such that there is either purine-, or pyrimidine-loading of the corresponding mRNAs), and the probability (p) that the asymmetry in numbers of positive and negative ORFs is not significant. The Wilcoxan signed ranks test performed with Minitab software (Meyer & Krueger, 1994) takes into account the magnitudes of Chargaff differences.

† HVS, *Herpes saimiri* virus (X64346); VZV, Varicella-Zoster virus (X04370); EBV, Epstein-Barr virus (V01555); HSV-1, *Herpes simplex* virus-1 (X14112).

relationship is suggested by an ORF in the GA group, which encodes the EBNA-1 latency protein.

6. Gene Encoding the Major Latency Transcript Obeys the Rule

Figure 3 shows Chargaff difference analysis of the section of the GC-rich Epstein-Barr virus (EBV) genome from which a major latency-associated transcript (encoding EBNA-1 protein) is derived. Whereas, like the majority of EBV genes, most neighbouring genes are pyrimidine-loaded (C > G when transcription is to the right; G > C when transcription is to the left), the rightward-transcribed gene encoding EBNA-1 protein follows Szybalski's rule (G > C; A > T), and very dramatically so.

The EBNA-1-encoding gene is exceptional. It is the *only* viral gene expressed in the most fundamental type of EBV latency (the "EBNA-1 only program"; Thorley-Lawson *et al*., 1996). Among the other EBV latency-associated genes, those encoding EBNAs 2–6 purine-load their mRNAs only with respect to the W bases, and much less dramatically than the gene encoding EBNA-1; those encoding LMP1 and LMP2 pyrimidine-load their mRNAs with respect to both the W bases and the S bases (data not shown).

7. Simple Sequence Repeats Reinforce Compliance with Szybalski's Rule

It appears that, unlike most other EBV genes, the rightward-transcribed gene encoding EBNA-1 has been under pressure to accepted mutations which increase the purine content of the top (mRNA-synonymous) strand. If this were not possible without disrupting protein functional domains, the gene might have locally increased its content of purine-rich codons in inter-domain regions. Indeed, the EBNA-1 gene has a long "simple sequence" region (Karlin *et al*., 1988) containing exclusively either glycine codons (GGN), or alanine codons (GCN). Table 4 shows that choices of third bases (N) in these codons are almost exclusively purines (Karlin *et al*., 1990). Although the EBNA-1 gene (*BKRF1*) without

the simple sequence is already slightly purine-loaded, the additional purines in the simple sequence greatly increase Chargaff difference values in favour of Szybalski's rule (Table 5).

In several other members of the Herpes virus family there are similar purine biases in long (>100 amino acids) simple sequence-encoding regions within genes which may be latency associated (Tables 4 and 5). These include *ORF 48* of HVS (T cell tropic), which is located in the HVS genome in a similar position to the EBNA-1-encoding gene in the genome of EBV (B cell tropic).

The ORF encoding the latency-associated nuclear antigen (LANA) of HHV8 also contains a long simple sequence repeat (Rainbow *et al*., 1997). Figure 4 shows Chargaff differences in the region of the ORF. Being leftward-transcribed, it obeys Szybalski's rule in having an excess of pyrimidines in the top strand (C > G; T > A), whereas most neighbouring genes, whatever their transcription direction, disobey the rule. Thus, there is purine-loading of LANA mRNA which is reflected in codon choice (data not shown).

Human herpes virus 6 (GenBank accession number X83413) has a 117 codon simple sequence in ORF *LJ1*, which displays purine-loading with respect to the S bases, and has a CpG island; such association with CpG islands (Table 5) is an expected feature of the promoter regions of latency-associated genes with hypomethylated CpGs (Honess *et al*., 1989; Tao *et al*., 1998).

Intriguingly, there is a long region of repetitive DNA in an ORF of unknown function (*ORF 50*) in Ictalurid Herpesvirus 1 (channel catfish virus; GenBank M75136); here again, codon-choice (with respect to glycine, valine and alanine) suggests purine-loading (data not shown).

8. Disruption of Host Traffic

Viruses can have acute or chronic (persistent) patterns of host infection (Villarreal *et al*., 2000). Certain acutely lytic viruses (e.g. Hepatitis A, Vaccinia) purine-load their RNAs in compliance with Szybalski's rule, whereas viruses causing chronic (sometimes sub-clinical) infections tend to pyrimidine-load their RNAs (e.g. Hepatitis B; Bell & Forsdyke, 1999). Prolonged and profound clinical latency is a characteristic of some viruses that pyrimidine-load (Tables 1–3; Fig. 3). In contrast to individuals latently infected with HIV-1, most individuals infected with HTLV-1 remain asymptomatic and live normal lives. Cytotoxic T cells appear able to target only peptides from the Tax protein (Gould & Bangham, 1998). Furthermore, HTLV-1 is likely to transfer between individuals when integrated in host DNA within intact cells. Virions alone show low infectivity. Similarly, Herpes simplex-related viruses permanently infect many individuals in their host species, who are often asymptomatic (Baer *et al*., 1984; Davison & Scott, 1986; McGeoch *et al*., 1988; Albrecht *et al*., 1992).

Thus, certain persistent GC-rich viruses appear to risk interactions with host RNAs, which would be initiated through complementary base pairing between loops. *In vivo*, C-rich loops of virus RNAs would interact with G-rich loops of the host RNAs [just an *in vitro* poly(rC) interacts *rapidly* at *low temperatures* with mRNA-synonymous DNA strands; Szybalski *et al*., 1966]. When in a functionally latent state most virus mRNAs would not be transcribed; thus the risk would be minimized. When triggered to move from the latent state to one of rapid productive cytolysis, the viruses would transcribe RNAs which, when released from the nucleus, might suddenly flood the cytosol with RNAs "driving on the wrong side of the road". The multiplicity of distracting loop–loop RNA interactions might slow host cell "traffic" and impair defence responses, including those triggered by any dsRNA which was formed. This "surprise" strategy might be of adaptive value to the virus.

9. The Role of Simple-Sequence Repeats in Latency-Associated Gene Products

In the functionally latent state most viral mRNAs would not be expressed, and so would not be available to interact with host mRNAs. However, although HTLV-1 has no latency-specific transcripts, most herpesviruses do. Remarkably, often herpesvirus transcripts include purine-biased simple sequence elements (see Section 7).

Like those of other members of the herpes virus family, the EBV genome is very compact with little intergenic DNA; this suggests an evolutionary selection pressure to eliminate

TABLE 3
Distribution of Chargaff differences among ORFs of Herpes-related viruses*

Virus†		Transcription to the left					Transcription to the right				
			A>T	T>A	*p*‡	Rule§		A>T	T>A	*p*	Rule
HVS 35%		37 →	15	22	0.02	+	40 →	35	5	<0.0003	+
	G>C	9	$\mathbf{1}_{GA}$	$\mathbf{8}_{GT}$	0.006	+	12	$\mathbf{9}_{GA}$	$\mathbf{3}_{GT}$	0.054	+
	C>G	28	$\mathbf{14}_{CA}$	$\mathbf{14}_{CT}$	0.21	–	28	$\mathbf{26}_{CA}$	$\mathbf{2}_{CT}$	<0.001	+
	p	0.0002	0.001	0.037			0.003	0.001	0.39		
	Rule	+	+	+			–	–	+		
HHV6 42%		70 →	38	32	0.23	–	49 →	26	23	0.45	+
	G>C	28	$\mathbf{3}_{GA}$	$\mathbf{25}_{GT}$	<0.001	+	28	$\mathbf{8}_{GA}$	$\mathbf{20}_{GT}$	0.045	–
	C>G	42	$\mathbf{35}_{CA}$	$\mathbf{7}_{CT}$	<0.001	–	21	$\mathbf{18}_{CA}$	$\mathbf{3}_{CT}$	<0.001	+
	p	0.015	<0.001	<0.001			0.47	0.008	0.008		
	Rule	+	+	–			+	–	+		
VZV6 46%		34 →	15	19	0.151	+	37 →	24	13	0.022	+
	G>C	7	$\mathbf{2}_{GA}$	$\mathbf{5}_{GT}$	0.045	+	12	$\mathbf{9}_{GA}$	$\mathbf{3}_{GT}$	0.063	+
	C>G	27	$\mathbf{13}_{CA}$	$\mathbf{14}_{CT}$	0.447	+	25	$\mathbf{15}_{CA}$	$\mathbf{10}_{CA}$	0.089	+
	p	0.001	0.002	0.071			0.038	0.124	0.054		
	Rule	+	+	+			-	-	-		
HCMV 57%		126 →	62	64	0.104	+	82 →	44	38	0.405	+
	G>C	80	$\mathbf{35}_{GA}$	$\mathbf{45}_{GT}$	0.381	+	40	$\mathbf{15}_{GA}$	$\mathbf{25}_{GT}$	0.015	–
	C>G	46	$\mathbf{27}_{CA}$	$\mathbf{19}_{CT}$	0.037	–	42	$\mathbf{29}_{CA}$	$\mathbf{13}_{CT}$	0.002	+
	p	0.0008	0.22	<0.001			0.111	0.001	0.015		
	Rule	–	–	–			–	–	+		

non-functional sequences. The long simple sequence (encoding Gly–Ala repeats) in the EBNA-1 gene has been explained by Karlin *et al.* (1988, 1990; Karlin, 1995) as an adaptation operative at the protein level. However, the Gly–Ala region can be removed experimentally without affecting the known functions of EBNA-1 (Yates & Camiolo, 1988; Summers *et al.*, 1997).

The paradox appeared to be resolved by evidence that the Gly–Ala region functions in *cis* at the protein level to inhibit antigen processing for MHC presentation (Levitskaya *et al.*, 1995, 1997; Mukherjee *et al.*, 1998). However, in order to express the protein these authors used a vector which first had to express the corresponding mRNA; this was then translated into the protein. Their evidence is consistent with the Gly–Ala region being simply a device for purine-loading a foreign mRNA ("non-self") to make it appear like host mRNA ("self"); this might subvert intracellular self/not-self-discrimination (Forsdyke, 1994, 1995a, b, 1999).

EBV 60%		44 →	21	23	0.24	+	86 →	37	49	0.081	–
	G>C	30	$\mathbf{10}_{GA}$	$\mathbf{20}_{GT}$	0.015	+	22	$\mathbf{10}_{GA}$	$\mathbf{12}_{GT}$	0.425	–
	C>G	14	$\mathbf{11}_{CA}$	$\mathbf{3}_{CT}$	0.008	–	64	$\mathbf{27}_{CA}$	$\mathbf{37}_{CT}$	0.066	–
	p	0.00007	0.105	<0.001			<0.0001	<0.001	<0.001		
	Rule	–	+	–			–	–	–		
HSV-1 68%		37 →	20	17	0.15	–	41 →	17	24	0.037	–
	G>C	29	$\mathbf{15}_{GA}$	$\mathbf{14}_{GT}$	0.80	–	9	$\mathbf{5}_{GA}$	$\mathbf{4}_{GT}$	0.594	+
	C>G	8	$\mathbf{5}_{CA}$	$\mathbf{3}_{CT}$	0.69	–	32	$\mathbf{12}_{CA}$	$\mathbf{20}_{CT}$	0.014	–
	p	0.0007	0.002	0.008			0.0002	0.017	0.002		
	Rule	–	–	–			–	–	–		

* Chargaff differences were calculated as in Table 2. Bold numbers with subscripts are ORFs in a given category. For example, of 37 HVS ORFs transcribed to the left, a minority (15) have A > T. Of this 15, one has G > C and 14 have C > G. Thus, of the 37 ORFs, one is in the GA group, eight are in the GT group, 14 are in the CA group, and 14 are in the CT group.

† Abbreviated names of viruses with their (G + C)%. HHV6, human herpesvirus 6 (X83413); HCMV, human cytomegalovirus (X17403).

‡ Probability (*p*) values at the bottom of columns refer to the proportions of ORFs with G-excess over C, relative to ORFs with C-excess over G. Probability (*p*) values at the right of the rows refer to the proportions of ORFs with A-excess over T, relative to ORFs with T-excess over A. The significance of departures from equifrequency was calculated using the Wilcoxan signed-ranks test.

§ " + " refers to compliance with Szybalski's rule (e.g. excess purines when transcription is to the right). " – " refers to a deviation from Szybalski's rule (e.g. excess pyrimidines when transcription is to the right). Designation of " + " or " – " is based on the sum of Chargaff difference values, which closely corresponds to the relative numbers of " + " and " – " ORFs.

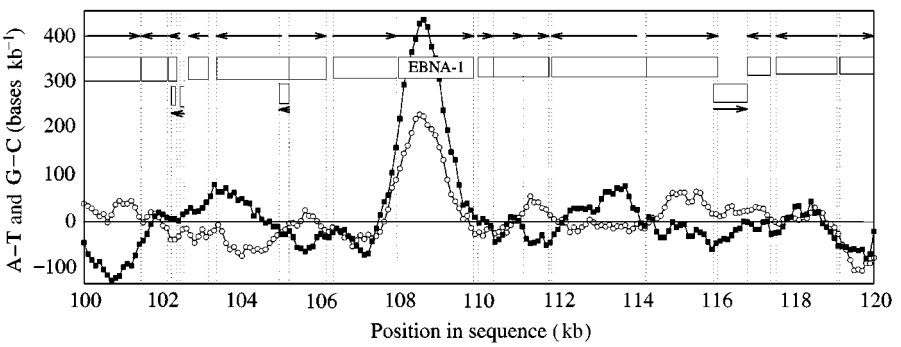

FIG. 3. Chargaff difference analysis of a section of the Epstein-Barr virus genome. ORFs are shown as open boxes with arrows indicating transcription direction. A, T, G, and C are the number of bases counted in 1 kb windows which were moved along the sequence in 0.1 kb steps. Each data point corresponds to the middle of a window. Chargaff differences (%) are expressed either as G − C (filled squares), or as A − T (open circles). The major ORF encoding Epstein-Barr nuclear antigen-1 (EBNA-1) is labelled.

If selection had been acting at the protein level to conserve the Gly–Ala region there should *not* have been such extreme codon bias (Table 4). On the other hand, the bias might have arisen by the amplification of an initially small Gly–Ala-coding segment, which just happened to have the

TABLE 4

Codon usages of genes containing long simple sequences in Herpes simplex-related viruses*

Virus	Gene	Codons†		Complete protein	Less the simple sequence	Simple sequence alone	Human average‡ (%)
EBV	*BKRF1* (EBNA-1)	Gly	GGG	63	11	52	24
			GGA	144	43	101	25
			GGT	25	24	1	16
			GGC	19	19	0	34
		Ala	GCG	4	3	1	10
			GCA	85	2	83	23
			GCT	6	6	0	26
			GCC	8	8	0	40
HVS	*ORF 48*	Gly	GGG	34	2	32	24
			GGA	55	7	48	25
			GGT	1	1	0	16
			GGC	6	6	0	34
		Glu	GAG	92	10	82	58
			GAA	75	29	46	42
	ORF 73	Gly	GGG	2	2	0	24
			GGA	15	7	8	25
			GGT	1	1	0	16
			GGC	1	1	0	34
		Glu	GAG	2	1	1	58
			GAA	128	3	125	42
		Arg	AGG	2	2	0	20
			AGA	17	11	6	20
			CGG	1	1	0	21
			CGA	1	1	0	11
			CGT	4	3	1	8
			CGC	1	1	0	19
		Ala	GCG	2	2	0	10
			GCA	8	6	2	23
			GCT	43	6	37§	26
			GCC	0	0	0	40
VZV	*ORF 11*	Gly	GGG	6	6	0	24
			GGA	28	16	12	25
			GGT	11	11	0	16
			GGC	5	5	0	34
		Glu	GAG	58	14	44	58
			GAA	30	30	0	42
		Asp	GAT	20	19	1	47
			GAC	35	16	19	53
		Ala	GCG	31	11	20	10
			GCA	19	19	0	23
			GCT	13	13	0	26
			GCC	13	13	0	40

* Values are the absolute number of codons in each protein segment.

† The main amino acids contributing to each simple sequence are listed (e.g. the EBNA-1 protein simple sequence has alternating glycines and alanines).

‡ Data are from 17 625 human genes (Nakamura *et al.*, 1999). Percentage distributions within each codon family are shown.

§ Pyrimidine-loading by virtue of this alanine codon is more than offset by the large excess of glutamate codons in *ORF 73*.

TABLE 5

Contribution of long simple sequences to Chargaff differences of ORFs of Herpes simplex-related viruses

Chargaff difference formula*					A − T			G − C		
Virus	Gene	CpG island (local)†	ORF length in codons: Complete ORF	ORF length in codons: Simple sequence	Complete ORF	Less the simple sequence	Simple sequence alone	Complete ORF	Less the simple sequence	Simple sequence alone
EBV	*BKRF1*	−	642	238	A > T (141.2)	A > T (46.2)	A > T (95.0)	G > C (270.0)	G > C (86.2)	G > C (187.4)
HVS	*ORF 48*	−	797	301	A > T (135.9)	A > T (31.8)	A > T (104.1)	G > C (225.0)	G > C (25.9)	G > C (199.1)
	ORF 73	−	407	183	A > T (270.3)	A > T (73.7)	A > T (196.6)	G > C (110.5)	C > G (−11.5)	G > C (122.0)
VZV	*ORF 11*	+	819	102	A > T (11.0)	T > A (−18.7)	A > T (29.7)	G > C (57.8)	G > C (5.3)	G > C (52.5)
HHV6	*LJ1*	+	321	117	T > A (−140.2)	T > A (−77.9)	T > A (−62.3)	G > C (145.4)	C > G (−36.3)	G > C (181.7)

* Chargaff differences (bases kb^{-1}) were calculated directly from the ORFs (and not by summing windows within the ORF).

† CpG islands are arbitrarily defined as >80 CpG dinucleotides/1 kb sequence window. The presence of a CpG island in regulatory regions suggests activity during viral latency. Absence of a local CpG island may mean that the promotor operating during latency is distant from the ORF (Tao *et al*., 1998). HSV-1 is not generally CpG depleted and has no ORF with a simple sequence >70 codons.

purine bias. However, we note that in several other members of the Herpes virus family there are similar purine biases in long (>100 amino acids) simple sequence-encoding regions within genes likely to be latency-associated (see Section 7). These biases might also be a consequence of a selection pressure for purine-loading of mRNA to assist maintenance of latency. For example, the LANA protein of HHV8 stabilizes latency by preventing p53-mediated apoptosis (Friborg *et al*., 1999).

10. Messenger RNAs as "Antibodies"

So why is EBNA-1 mRNA purine-loaded? Distracted by the *messenger* role of mRNA molecules, we may fail to note that the diverse spectrum of cell mRNA species, like the diverse spectrum of antibodies in serum, constitutes a repertoire of specificities with the potential to react with complementary sections of non-self RNA "antigens". Just as interactions between antibody and foreign antigen provoke *extracellular* inflammatory responses to the antigen, so interactions between host RNA and foreign RNA might provoke *intracellular* responses to foreign RNA, which could include gene silencing. If EBNA-1 mRNA ("sense") in latent EBV-infected cells were not purine-loaded to avoid "kissing" interactions), then it is possible that a self-RNA species would have a sufficient degree of complementarity ("antisense") to progress beyond kissing interactions. Molecules of dsRNA of a length sufficient to alert host defence systems might then be formed (Suzuki *et al*., 1999). The alarm might serve to increase MHC protein expression since only newly synthesized MHC proteins bind peptides efficiently for presentation to T cells (Townsend *et al*., 1990). Thus, through its purine-loading of EBNA-1 mRNA ("stealth" strategy; Cristillo *et al*., 1996), EBV would fail to provoke gene silencing or increase MHC expression. This would impede the MHC-dependent cytotoxic T cell response (Callan *et al*., 1998), and so assist maintenance of the latent state.

There are reports of natural antisense RNAs derived from overlapping genes with different transcriptional orientations (Vanhée-Brossollet & Vaquero, 1988). In the light of the present thesis, such transcripts should not normally

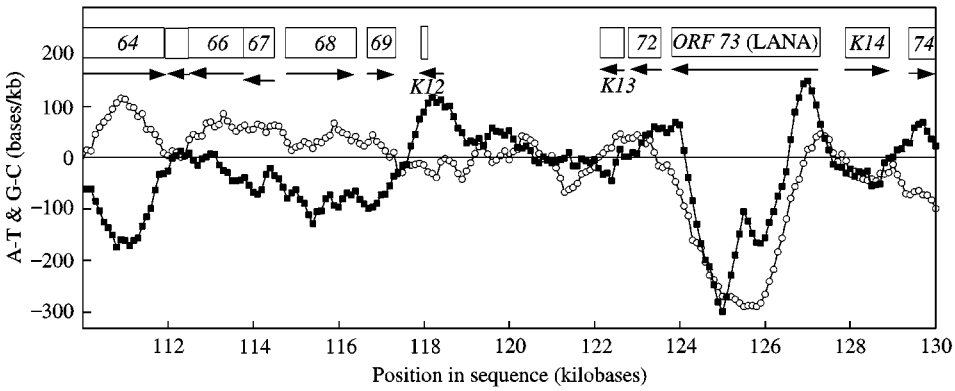

FIG. 4. Chargaff difference analysis of a section of the genome of Kaposi's sarcoma associated herpesvirus (HHV8). ORFs are numbered. Those with the prefix "K" may be unique to this virus. The major latency-associated nuclear antigen (LANA) is encoded by *ORF 73*. Note that, since transcription is to the left, the upper (template) strand of *ORF 73* is pyrimidine-loaded, so that the lower (mRNA-synonymous) strand, and hence the corresponding mRNA, would be purine-loaded. Other details are as in Fig. 3.

coexist in the same cell or intracellular compartment, or should be special cases for which there are adaptations to prevent the inadvertent firing, or response to, dsRNA alarms.

11. Charge Cluster Domains Decrease Immunogenicity of Other Domains?

The use of simple sequence to purine-load mRNAs means that, at the protein level, the simple sequences often contain runs of charged amino acids (e.g. Glu, Asp). Karlin (1995) refers to such regions as "hyper-charge runs", and notes that "for most of the hypercharge runs [in proteins] there is considerable *variation* in codon usage, which suggests an important function for these charge runs" (our parentheses and italics). However, our studies show that the variation is restricted to purine-rich codons (Table 4), which is more consistent with selection acting at the nucleic acid level.

Many of the codons characteristic of the triplet expansion diseases, some of which generate charge runs, are also purine-rich (Green & Wang, 1994; Hancock & Santibanez-Koref, 1998). We suggest that charge runs themselves may *not* have an important function with respect to the function of the end-product (protein) of the ORF in which they locate (although they may affect protein solubility). When attempting to relate a protein's sequence to its biological function, the possibility that the major selective pressure has been at the nucleic acid level must be considered (Ball, 1973; Rocha *et al.*, 1999; Lao & Forsdyke, 2000). As the result, a protein of less than optimum function may be synthesized, or the protein sequence may have to counter-adapt to improve function in the face of a primary selection pressure operating at the level of the corresponding nucleic acid (Forsdyke, 1996; Forsdyke & Mortimer, 2001).

For example, to counter a tendency of its protein product to provoke autoimmune attack by cytotoxic T cells, there would be a selection pressure for a gene to purine-load its mRNAs, thus generating long charge-rich alpha-helices which might be irrelevant to the function of the protein itself (Dohlman *et al.*, 1993). In this respect, we note the prevalence of charge clusters in antigens implicated in various autoimmune diseases. The clusters do *not* coincide with major autoantigenic epitopes (Brendel *et al.*, 1991). This suggests that charge cluster domains may not be the *primary* cause of the diseases, as has been supposed, but may have evolved *in response* to the disease-provoking characteristics of other domains (i.e. the domains corresponding to autoantigenic epitopes).

Unlike EBV, HSV-1 does not show CpG suppression (indicating no general methylation of CpGs), and there are no long simple sequence regions. However, HSV-1 would be expected to have pyrimidine-rich loops in most mRNAs (Tables 2 and 3). Intriguingly, the main HSV-1

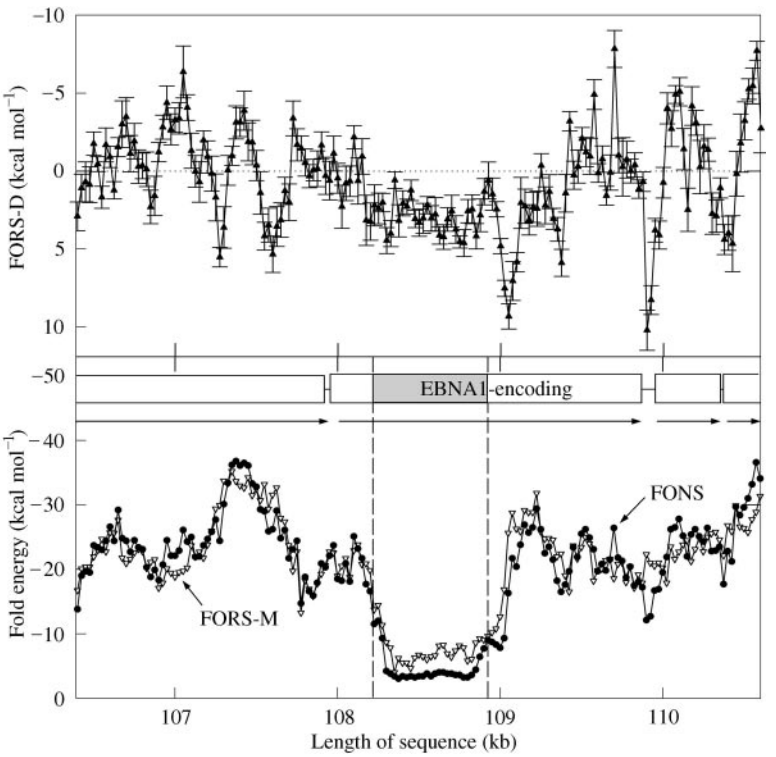

FIG. 5. Fold analysis of a segment of the Epstein-Barr virus genome containing the EBNA-1-encoding gene (labelled box with horizontal arrow indicating transcription to the right). The region of the Gly–Ala repeat is marked by shading and two vertical dashed lines. FONS, values for the folding of the natural sequence. High negative FONS values (e.g. -40 kcal mol^{-1}) correspond to high folding potential (stem-loop potential; Forsdyke, 1998). FORS-M, values for the base composition-dependent component of the FONS values. FORS-D (upper plot with standard errors of the mean), values for the base order-dependent component of the FONS values (such that FONS = FORS-M + FORS -D).

RNAs transcribed during latency correspond to the "*antisense*" strand (Croen *et al*., 1987; Stevens *et al*., 1987). We predict that any parts of these latency-associated transcripts which persist in the cell would be relatively purine rich (Goldenberg *et al*., 1997; Arthur *et al*., 1998). A similar prediction applies to certain antisense transcripts found in EBV latency (Karran *et al*., 1992; Brooks *et al*., 1993).

12. Low Stem-Loop Potential of the Gly–Ala Repeat-Encoding Region

Simple sequence elements in proteins, as found in the trinucleotide expansion diseases [e.g. polyglutamine tracts encoded by poly(CAG)], sometimes cause intracellular protein aggregation. It is possible that such aggregates are responsible for the underlying pathology (Hancock & Santibanez-Koref, 1998; Forsdyke, 2000). However, trinucleotide repeats are often capable of adopting stem-loop conformations, which at the RNA level can activate PKR. This may contribute to the disease mechanism (Tian *et al*., 2000). In the case of the gene encoding EBNA-1, the region of the repeat has very low stem-loop potential, as revealed by a sustained low negative value for the folding of the natural sequence (FONS value for the mRNA-synonymous strand; Fig. 5). This is contributed to both by the base composition of the repeat (FORS-M value), and by its base order (FORS-D value; Forsdyke, 1998). In the corresponding mRNA, most parts encoding known functions of the protein may adopt compact secondary structures, whereas the part encoding the simple sequence repeat and the region on its immediate 3′ side may have a structure more available for intermolecular interactions.

13. A Role for Non-Genic DNA?

The threshold for binding to PKR is approximately 15 trinucleotide repeats (45 bases), which

corresponds to a dsRNA segment of approximately 22 bases (Tian *et al*., 2000). There are 4^{22} possible combinations in the universe of 22 base sequences, of which half ($4^{22}/2$) complements the other half. A virus encoding 10 mRNAs of average length 1021 bases, would have 10 000 (approx. 4^7) "windows" of 22 bases, any one of which could potentially act as an RNA "antigenic determinant" in the host cell. Assuming 10 000 different host mRNA species in a cell, there would be $10\,000 \times 1000$ (approx. 4^{12}) potential complementary RNA "windows" in host mRNAs. With a much higher mutation rate, a virus species might appear capable of adapting to ensure that its 4^7 specificities did not complement the host's 4^{12} specificities. Various factors militate against this.

First, a high degree of polymorphism among host transcripts (Sunyaev *et al*., 2000) would make it likely that what a member of a virus species "learned" (through mutation) on one member of its host species, would not be applicable to the next member of the host species which it encountered (Forsdyke, 1991; Forsdyke, 2000). Second, due to a low level of read-through transcription (failure of transcription termination) of host mRNAs, a low level of transcription of extragenic DNA occurs (Heximer *et al*., 1998). Thus, the maximum potential repertoire of "RNA antibodies" would be limited only by genome size (approx. 4^{16} potential specificities in humans). Indeed, one function of the promoters of repetitive DNA elements (e.g. human *Alu* elements) might be to provide such read-through transcription, as has been observed (Manley & Colozzo, 1982; Feuchter *et al*., 1992). It would be of adaptive advantage to the host to activate such promoters under conditions of cell stress (heat shock or viral infection); again, this is observed (transcription by RNA polymerase III; Jang & Latchman, 1989; Liu *et al*., 1995).

Thus, non-genic "junk" DNA (Dang *et al*., 1998) can be viewed in much the same way as we view the diverse genes encoding the variable regions of immunoglobulin antibodies. Just as B-cells capable of synthesizing a unique anti-self antibody would be eliminated during somatic time to prevent self-reactivity, so junk DNA would be screened over evolutionary time (by positive selection of individuals in which favourable mutations had been collected together by recombination) to decrease the probability of two complementary "self" transcripts interacting to form dsRNA segments of more than 21 bases. High polymorphism of non-genic DNA (Beck *et al*., 1996; Nickerson *et al*., 1998) would make it difficult for viruses to anticipate the RNA "antibody" repertoire of future hosts (Forsdyke, 1999, 2000a, b).

We thank J. Gerlach for assistance with computer configuration, J. T. Smith for statistical advice, and G. McPherson for assistance with the Silicon Graphics Computer maintained by Base4 BioInformatics Inc., Mississauga. The National Research Council of Canada, Academic Press, Cold Spring Harbor Laboratory Press, and Elsevier Publishing Corporation gave permission for the display of full-text versions of some of the cited references at our internet site (http://post.queensu.ca/ ~ forsdyke/bioinfor.htm).

REFERENCES

ALBRECHT, J. C., NICHOLAS, J., BILLER, D., CAMERON, K. R., BIESINGER, B., NEWMAN, C., WITTMANN, S., CRAXTON, M. A., COLEMAN, H., FLECKENSTEIN, B. & HONESS, R. W. (1992). Primary structure of the *Herpes saimiri* genome. *J. Virol.* **66,** 5047–5058.

ARTHUR, J. L., EVERETT, R., BRIERLEY, I. & EFSTATHIOU, S. (1998). Disruption of the 5′ and 3′ splice sites flanking the major latency-associated transcripts of herpes simplex virus type 1: evidence for alternative splicing in lytic and latent infections. *J. Gen. Virol.* **79,** 107–116.

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

BALL, L. A. (1973). Secondary structure and coding potential of the coat protein gene of bacteriophage MS2. *Nat. New Biol.* **242,** 44–45.

BECK, S., ABDULLA, S., ALDERTON, R. P., GLYNNE, R. J., GUT, I. G., HOSKING, L. K., JACKSON, A., KELLY, A., NEWELL, W. R., SANSEAU, P., RADLEY, E., THORPE, K. L. & TROWSDALE, J. (1996). Evolutionary dynamics of non-coding sequences within the class II region of the human MHC. *J. Mol. Biol.* **255,** 1–13.

BELL, S. J., CHOW, Y. C., HO, J. Y. K. & FORSDYKE, D. R. (1998). Correlation of CHI orientation with transcription indicates a fundamental relationship between recombination and transcription. *Gene* **216,** 285–292.

BELL, S. J. & FORSDYKE, D. R. (1999). Deviations from Chargaff's second parity rule correlate with direction of transcription. *J. theor. Biol.* **197,** 63–76.

BOSSI, L. & ROTH, J. R. (1980). The influence of codon context on genetic code translation. *Nature* **286,** 123–127.

BRENDEL, V., DOHLMAN, J., BLAISDELL, B. E. & KARLIN, S. (1991). Very long charge runs in systemic lupus erythematosus-associated autoantigens. *Proc. Natl Acad. Sci. U.S.A.* **88,** 1536–1540.

BROOKS, L. A., LEAR, A. L., YOUNG, L. S. & RICKINSON, A. B. (1993). Transcripts from the Epstein-Barr virus *Bam*H1 A fragment are detectable in all three forms of virus latency. *J. Virol.* **67,** 3182–3190.

BULL, J. J., JACOBSON, A., BADGETT, M. R. & MOLINEUX, I. J. (1998). Viral escape from antisense RNA. *Mol. Microbiol.* **28,** 835–846.

CALLAN, M. F. C., TAN, L., ANNELS, N., OGG, G. S., WILSON, J. D. K., O'CALLAGHAN, C. A., STEVEN, N., MCMICHAEL, A. J. & RICKINSON, A. B. (1998). Direct visualization of antigen-specific $CD8^+$T cells during the primary immune response to Epstein-Barr virus in vivo. *J. Exp. Med.* **187,** 1395–1402.

CREON, K. D., OSTROVE, J. M., DRAGOVIC, L. J., SMIALEK, J. E. & STRAUS, S. E. (1987). Latent *Herpes simplex* virus in human trigeminal ganglia. Detection of an immediate early gene "anti-sense" transcript by in situ hybridization. *N. Engl. J. Med.* **317,** 1427–1432.

CRISTILLO, A. D. (1998). Characterization of G_0/G_1 switch genes in cultured T lymphocytes. Ph.D. Thesis, Queen's University, Kingston, Ontario.

CRISTILLO, A. D., HEXIMER, S. P. & FORSDYKE, D. R. (1996). A "stealth" approach to inhibition of lymphocyte activation by oligonucleotide complementary to the putative G_0/G_1 switch regulatory gene *GOS30/EGEI/NGFI-A*. *DNA Cell Biol.* **15,** 561–570.

CRISTILLO, A. D., LILLICRAP, T. P. & FORSDYKE, D. R. (1998). Purine-loading of EBNA-1 mRNA avoids senseantisense "collisions". *FASEB J.* **12,** A1453. Abstract # 828.

DANG, K. D., DUTT, P. B. & FORSDYKE, D. R. (1998). Chargaff differences correlate with transcription direction in the bithorax complex of *Drosophila*. *Biochem. Cell Biol.* **76,** 129–137.

DAVISON, A. J. & SCOTT, J. E. (1986). The complete DNA sequence of Varicella-Zoster virus. *J. Gen Virol.* **67,** 1759–1815.

DOHLMAN, J. G., LUPAS, A. & CARSON, M. (1993). Long charge-rich alpha-helices in systemic autoantigens. *Biochem. Biophys. Res. Commun.* **195,** 686–696.

EGUCHI, Y., ITOH, T. & TOMIZAWA, J. (1991). Antisense RNA. *Annu. Rev. Biochem.* **60,** 631–652.

EHRENFELD, E. & HUNT, T. (1971). Double-stranded poliovirus RNA inhibits initiation of protein synthesis by reticulocyte lysates. *Proc. Natl Acad. Sci. U.S.A.* **68,** 1075–1078.

ELIA, A., LAING, K. G., SCHOFIELD, A., TILLERAY, V. J. & CLEMENS, M. J. (1996). Regulation of the double-stranded RNA-dependent protein kinase PKR by RNAs encoded by a repeated sequence of the Epstein-Barr virus genome. *Nucl. Acids Res.* **24,** 4471–4478.

FEUCHTER, A. E., FREEMAN, J. D. & MAGER, D. L. (1992). Strategy for detecting cellular transcripts promoted by human endogenous long terminal repeats: identification of a novel gene (CDC4L) with homology to yeast CDC4. *Genomics* **13,** 1237–1246.

FIRE, A. (1999). RNA-triggered gene silencing. *Trends Genet.* **15,** 358–363.

FORSDYKE, D. R. (1991). Early evolution of MHC polymorphism. *J. theor. Biol.* **150,** 451–456.

FORSDYKE, D. R. (1994). Relationship of X chromosome dosage compensation to intracellular self/not-self discrimination: a resolution of Muller's paradox? *J. theor. Biol.* **167,** 7–12.

FORSDYKE, D. R. (1995a). Entropy-driven protein self-aggregation as the basis for self/not-self discrimination in the crowded cytosol. *J. Biol. Sys.* **3,** 273–287.

FORSDYKE, D. R. (1995b). Fine tuning of intracellular protein concentrations, a collective protein function involved in aneuploid lethality, sex determination and speciation? *J. theor. Biol.* **172,** 335–345.

FORSDYKE, D. R. (1996). Different biological species "broadcast" their DNAs at different (G + C)% "wavelengths". *J. theor. Biol.* **178,** 405–417.

FORSDYKE, D. R. (1998). An alternative way of thinking about stem-loops in DNA. A case study of the human *G0S2* gene. *J. theor. Biol.* **192,** 489–504.

FORSDYKE, D. R. (1999). Heat shock proteins as mediators of "danger" signals: implications of the slow evolutionary fine-tuning of sequences for the antigenicity of cancer cells. *Cell Stress Chaperones* **4,** 205–210.

FORSDYKE, D. R. (2000). Double-stranded RNA and/or heat-shock as initiators of chaperone mode switches in diseases associated with protein aggregation. *Cell Stress Chaperones* **5,** 375–376.

"FORSDYKE, D. R. (2001). *Search for a Victorian. The Origin of Species, Revisited*. Montreal: McGill-Queen's University Press."

FORSDYKE, D. R. & MORTIMER, J. R. (2001). Chargaff's legacy. *Gene* **261,** 127–137.

FRIBORG, J., KONG, W.-P., HOTTINGER, M. O. & NABEL, G. J. (1999). P53 inhibition by the LANA protein by KSHV protects against cell death. *Nature* **402,** 889–894.

GOLDENBERG, D., MADOR, N., BALL, M. J., PANET, A. & STEINER, I. (1997). The abundant latency-associated transcripts of Herpes simplex virus type 1 are bound to polyribosomes in cultured neuronal cells and during latent infection in mouse trigeminal ganglia. *J. Virol.* **71,** 2897–2904.

GOULD, K. G. & BANGHAM, C. R. M. (1998). Virus variation, escape from cytotoxic T lymphocytes and human retroviral persistence. *Sem. Cell Dev. Biol.* **9,** 321–328.

GREEN, H. & WANG, N. (1994). Codon reiteration and the evolution of proteins. *Proc. Natl Acad. Sci. U.S.A.* **91,** 4298–4302.

HAMILTON, A. J. & BAULECOMBE, D. C. (1999). Role of a species of small antisense RNA in post-transcriptional gene silencing in plants. *Science* **286,** 950–951.

HANCOCK, J. M. & SANTIBANEZ-KOREF, M. F. (1998). Trinucleotide expansion diseases in the context of micro- and mini-satellite evolution. *EMBO J.* **17,** 5521–5524.

HEXIMER, S. P., CRISTILLO, A. D., RUSSELL, L. & FORSDYKE, D. R. (1998). Expression and processing of *G_0/G_1 Switch Gene 24* (*GOS24/TIS11/TTP/NUP475*) RNA in cultured human blood mononuclear cells. *DNA Cell Biol.* **17,** 249–263.

HONESS, R. W., GOMPELS, U. A., BARRELL, B. G., CRAXTON, M., CAMERON, K. R., STADEN, R., CHANG, Y.-N. & HAYWARD, G. S. (1989). Deviations of expected frequencies of CpG dinucleotides in Herpesvirus DNAs may be diagnostic of differences in the states of their latent genomes. *J. Gen. Virol.* **70,** 837–855.

HUNTER, T., HUNT, T., JACKSON, R. J. & ROBERTSON, H. D. (1975). The characteristics of inhibition of protein synthesis by double-stranded ribonucleic acid in reticulocyte lysate. *J. Biol. Chem.* **250,** 409–417.

IZANT, J. G. & WEINTRAUB, H. (1984). Inhibition of thymidine kinase gene expression by anti-sense RNA: a molecular approach to genetic analysis. *Cell* **36,** 1007–1015.

JANG, K. L. & LATCHMAN, D. S. (1989). HSV infection induces increased transcription of Alu repeated sequences by RNA polymerase III. *FEBS Lett.* **258,** 255–258.

KARLIN, S. (1995). Statistical significance of sequence patterns in proteins. *Curr. Opin. Struct. Biol.* **5,** 360–371.

KARLIN, S., BLAISDELL, B. E., MOCARSKI, E. S. & BRENDEL, V. (1988). A method to identify distinctive charge configurations in protein sequences, with application to human herpesvirus polypeptides. *J. Mol. Biol.* **205,** 165–177.

KARLIN, S., BLAISDELL, B. E. & SCHACHTEL, G. A. (1990). Contrasts in codon usage of latent versus productive genes of Epstein-Barr virus: data and hypothesis. *J. Virol.* **64,** 4264–4273.

KARRAN, L., GAO, Y., SMITH, P. R. & GRIFFIN, B. E. (1992). Expression of a family of complementary-strand transcripts in Epstein-Barr virus-infected cells. *Proc. Natl Acad. Sci. U.S.A.* **89,** 8058–8062.

KUMAR, M. & CARMICHAEL, G. G. (1998). Antisense RNA: function and fate of duplex RNA in cells of higher eukaryotes. *Microbiol. Mol. Biol. Rev.* **62,** 1415–1434.

KUPIEC, J. J., KAY, A., HAYAT, M., RAVIER, R., PERIES, J. & GALIBERT, F. (1991). Sequence analysis of the simian foamy virus type 1 genome. *Gene* **101,** 185–194.

LAO, P. J. & FORSDYKE, D. R. (2000). Thermophilic bacteria strictly obey Szybalski's transcription direction rule and politely purine-load RNAs with both adenine and guanine. *Genome Res.* **10,** 228–236.

LEVITSKAYA, J., CORAM, M., LEVITSKY, V., IMREH, S., STEIGERWALD-MULLEN, P. M., KLEIN, G., KURILLA, M. G. & MASUCCI, M. G. (1995). Inhibition of antigen processing by the internal repeat region of the Epstein-Barr virus nuclear antigen-1. *Nature* **375,** 685–688.

LEVITSKAYA, J., SHARIPO, A., LEONCHIKS, A., CIECHANOVER, A. & MASUCCI, M. G. (1997). Inhibition of ubiquitin/proteasome-dependent protein degradation by the Gly-Ala repeat domain of the Epstein-Barr virus nuclear antigen 1. *Proc. Natl Acad. Sci. U.S.A.* **94,** 12 616–12 621.

LIU, W.-M., CHU, W.-M., CHOUDARY, P. V. & SCHMID, C. W. (1995). Cell stress and translational inhibitors transiently increase the abundance of mammalian SINE transcripts. *Nucl. Acids Res.* **23,** 1758–1765.

MALIK, K. T., EVEN, J. & KARPAS, A. (1988). Molecular cloning and complete nucleotide sequence of an adult T cell leukaemia virus/Human T cell leukaemia virus type I isolate of Caribbean origin. *J. Gen. Virol.* **69,** 1695–1710.

MANLEY, J. L. & COLOZZO, M. T. (1982). Synthesis in vitro of an exceptionally long RNA transcript promoted by an *Alu*I sequence. *Nature* **300,** 376–379.

MARCUS, P. (1983). Interferon induction by viruses: one molecule of dsRNA as the threshold for induction. *Interferon* **5,** 115–180.

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. *J. Gen. Virol.* **69,** 1531–1574.

MELTON, D. A. (1985). Injected antisense RNAs specifically block messenger RNA translation *in vivo*. *Proc. Natl Acad. Sci. U.S.A.* **82,** 144–148.

MEYER, R. K. & KRUEGER, D. D. (1994). *Minitab Computer Supplement*. New York: Macmillan College Publishing.

MITTELSTEN SCHEID, O. (1999). New tool for Swiss army knife. *Nature* **397,** 25.

MUKHERJEE, S., TRIVEDI, P., DORFMAN, D. M., KLEIN, G. & TOWNSEND, A. (1998). Murine cytotoxic T lymphocytes recognize an epitope in an EBNA-1 fragment, but fail to lyse EBNA-1-expressing mouse cells. *J. Exp. Med.* **187,** 445–450.

NAKAMURA, Y., GOJOBORI, T. & IKEMURA, T. (1999). Codon usage tabulated from the international DNA sequence databases: its status 1999. *Nucl. Acids Res.* **27,** 292.

NICKERSON, D. A., TAYLOR, S. L., WEISS, K. M., CLARK, A. G., HUTCHINSON, R. G., STENGARD, J., SALOMAA, V., VARTIAINEN, E., BOERWINKLE, E. & SING, C. (1998). DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. *Nat. Genet.* **19,** 233–240.

RAINBOW, L., PLATT, G. M., SIMPSON, G. R., SARID, R., GAO, S.-J., STOIBER, H., HERRINGTON, C. S., MOORE, P. S. & SCHULZ, T. F. (1997). The 222- to 234-kilodalton latent nuclear protein (LNA) of Kaposi's sarcoma-associated herpesvirus (human herpes virus 8) is encoded by *orf73* and is a component of the latency-associated nuclear antigen. *J. Virol.* **71,** 5915–5921.

RATNER, L., HASELTINE, W., PATARCA, R., LIVAK, K. J., STARCICH, B., JOSEPHS, S. F., DORAN, E. R., RAFALSKI, J. A., WHITEHORN, E. A., BAUMEISTER, K., IVANOFF, L., PETTEWAY, S. R., PEARSON, M. L., LAUTENBERGER, J. A., PAPIS, T. S., GHRAYEB, J., CHANG, N. T., GALLO, R. C., & WONG-STAAL, F. (1985). Complete nucleotide sequence of the AIDS virus. *Nature* **313,** 277–284.

ROBERTSON, H. D. & MATHEWS, M. B. (1996). The regulation of the protein kinase PKR by RNA. *Biochimie* **78,** 909–914.

ROCHA, E. P. C., DANCHIN, A. & VIARI, A. (1999). Universal replication biases in bacteria. *Mol. Microbiol.* **32,** 11–16.

SCHWARTZ, D. E., TIZARD, R. & GILBERT, W. (1983). Nucleotide sequence of Rous sarcoma virus. *Cell* **32,** 853–869.

SHARP, P. (1999). RNAi and double-strand RNA. *Genes Dev.* **13,** 139–141.

SMITHIES, O., ENGELS, W. R., DEVEREUX, J. R., SLIGHTOM, J. L. & SHEN, S. (1981). Base substitutions, length differences and DNA strand asymmetries in the human Gλ and Aλ fetal globin gene region. *Cell* **26,** 345–353.

STEVENS, J. G., WAGNER, E. K., DEVI-RAO, G. B., COOK, M. L. & FELDMAN, L. T. (1987). RNA complementary to a herpesvirus α gene mRNA is prominent in latently infected neurons. *Science* **235,** 1056–1059.

SUMMERS, H., FLEMING, A. & FRAPPIER, L. (1997). Requirements for Epstein-Barr nuclear antigen 1 (EBNA-1)-induced permanganate sensitivity of the Epstein-Barr latent origin of DNA replication. *J. Biol. Chem.* **272,** 26 434–26 440.

SUNYAEV, S. R., LATHE, W. C., RAMENSKY, V. E. & BORK, P. (2000). SNP frequencies in human genes: an excess of rare alleles and differing modes of selection. *Trends Genet.* **16,** 335–337.

SUZUKI, K., MORI, A., ISHII, K. J., SAITO, J., SINGER, D. S., KLINMAN, D. M., KRAUSE, P. R. & KOHN, L. D. (1999). Activation of target-tissue immune-recognition molecules by double-strand polynucleotides. *Proc. Natl Acad. Sci. U.S.A.* **96,** 2285–2290.

SZYBALSKI, W., KUBINSKI, H., & SHELDRICK, P. (1966). Pyrimidine clusters on the transcribing strands of DNA and their possible role in the initiation of RNA synthesis. *Cold Spring Harbor Symp. Quant. Biol.* **31,** 123–127.

TAO, Q., ROBERTSON, K. D., MANNS, A., HILDESHEIM, A. & AMBINDER, R. F. (1988). The Epstein-Barr virus major latent promoter Qp is constitutively active, hypomethylated, and methylation sensitive. *J. Virol.* **72,** 7075–7083.

THORLEY-LAWSON, D. A., MIYASHITA, E. M. & KAHN, G. (1996). Epstein-Barr virus and the B cell: that's all it takes. *Trends Microbiol.* **4,** 204–207.

TIAN, B., WHITE, R. J., XIA, T., WELLE, S., TURNER, D. H., MATHEWS, M. B. & THORNTON, C. A. (2000). Expanded CUG repeat RNAs form hairpins that activate the double-stranded RNA-dependent protein kinase PKR. *RNA* **6,** 79–87.

TOWNSEND, A., ELLIOTT, T., CERUNDULO, V., FOSTER, L., BARBER, B. & TSE, A. (1990). Assembly of MHC class 1 molecules analysed in vitro. *Cell* **62,** 285–295.

VANHÉE-BROSSOLLET, C. & VAQUERO, C. (1998). Do natural antisense transcripts make sense in eukaryotes? *Gene* **211,** 1–9.

VILLARREAL, L. P., DEFILIPPIS, V. R. & GOTTLIEB, K. A. (2000). Acute and persistent viral life strategies and their relationship to emerging diseases. *Virology* **272,** 1–6.

YATES, J. L. & CAMIOLO, S. M. (1988). Dissection of DNA replication and enhancer activation functions of Epstein-Barr virus nuclear antigen 1. *Cancer Cells* **6,** 197–205.