ELSEVIER

Virus Research 60 (1999) 1–19

Virus Research

Strand-specific compositional asymmetries in double-stranded DNA viruses

Andrei Grigoriev *

Max-Planck-Institute for Molecular Genetics, Ihnestr. 73, Berlin 14195, Germany

Received 8 April 1998; received in revised form 21 November 1998; accepted 21 November 1998

Abstract

Analysis of 22 complete sequences of double-stranded DNA viruses reveals striking compositional asymmetries between leading and lagging, and between transcribed and non-transcribed strands. In all bi-directionally replicated genomes analyzed, the observed leading strand GC skew (measuring relative excess of guanines versus cytosines) is different from that in the lagging strand. In most of these genomes GC skew switches polarity close to replication origins. GC skew changes linearly across adenovirus linear genomes, which replicate from one end. In papillomavirus, GC skew is positive in one half of the genome where transcription and replication proceed in the same direction, and is close to zero in the other half with divergent transcription and replication. Possible contributions of these two processes (and associated repair mechanisms) as well as other potential sources of strand bias in the observed asymmetries are discussed. Use of cumulative skew plots for genome comparisons is demonstrated on the example of herpes simplex virus.

Keywords: Strand asymmetry; Nucleotide substitution; Genome comparison; Replication; Transcription; dsDNA viruses

1. Introduction

Viral genomes were among the first sequenced, starting with simian virus 40 (Fiers et al., 1978; Reddy et al., 1978). Recently, the focus of the whole-genome sequencing efforts has shifted to the larger genomes of archaea, bacteria and eucaryotes. Fascinating features of genome organization, such as strand compositional asymmetries, have been found in these organisms (Lobry, 1996; Blattner et al., 1997; Freeman et al., 1998; Grigoriev, 1998a). It is interesting to analyze whether there are similar global composition patterns in viral genomes and how such patterns may be related to genome organization and replication

* Present address: Genome Pharmaceuticals Corporation, Lochhamer Str. 29, Martinsried 82152, Germany. Tel.: +49-89-8996770; fax: +49-89-89967710.

E-mail address: andrei.grigoriev@gpc-ag.com (A. Grigoriev)

0168-1702/99/$ - see front matter
PII: S0168-1702(98)00139-7

mode of a virus. Composition changes reflect nucleotide substitution forces acting in vivo during the processes of DNA synthesis, transcription and repair. Knowledge of compositional bias may help us understand better the molecular mechanisms of these processes and their impact on the evolution of genomes.

While discontinuous replication has been reported for both leading and lagging strands of DNA in vivo (Wang and Chen, 1994), they replicate asymmetrically in vitro, so there is a potential for differential mutation. If substitution rates for both leading and lagging strands were equal, the complementary nucleotides would be present in one (Watson or Crick) strand of a chromosome with equal frequencies, i.e. one would detect $[A] = [T]$ and $[G] = [C]$ in one strand, where [N] is a number of occurrences of a nucleotide N within a sequence segment (Lobry, 1995; Sueoka, 1995). However, in the real genomic sequences, one often observes $[N_1] \neq [N_2]$ in one strand for two complementary nucleotides N_1 and N_2. Furthermore, if $[N_1] - [N_2] > 0$ in the leading, $[N_1] - [N_2] < 0$ in the lagging strand. Such biases have been reported for simian virus 40 (SV40) and polyomavirus strain A-2 (Smithies et al., 1981) and interpreted as evidence of asymmetry in mutation pressure in SV40 (Filipski, 1990) because of a polarity switch at the replication origin.

Later, it was shown (Lobry, 1996) that in long stretches of genomic DNA of *Escherichia coli*, *Bacillus subtilis* and *Haemophilus influenzae* there is a change in polarity of $[N_1] - [N_2]$ at the sites of replication origin and terminus. For *E. coli*, this has been confirmed for $[G] - [C]$ on a whole-genome scale (Blattner et al., 1997). These observations have been recently further extended and generalized (Grigoriev, 1998a) to demonstrate that the nucleotide composition of a microbial chromosome changes at two points separated by about a half of its length and these points coincide with sites of replication origin and terminus for all bacteria where such sites are known. The leading strand has been found to contain more guanines than cytosines in 12 microbial genomes.

Similar partitioning of a chromosome into segments rich in purine or keto-bases has been reported for a smaller set of these genomes (Freeman et al., 1998). Differences between estimates of bias with $[N_1] - [N_2]$ and aggregate measures like purine excess have been recently discussed (Grigoriev, 1998b). The basis of long- and short-range compositional asymmetries has been interpreted in terms of biases in replication, repair and transcription (Beletskii and Bhagwat, 1996; Lobry, 1996; Blattner et al., 1997; Francino and Ochman, 1997; Freeman et al., 1998; Grigoriev, 1998a).

In contrast, the excess of guanine over cytosine has been found to change gradually across differentially replicated genome segments of vertebrate mitochondria and correlate with the time DNA spends single-stranded during synthesis (Grigoriev, 1998a). This paper continues the study and concentrates exclusively on compositional asymmetries found in the sequences of several double-stranded DNA viruses and their relation to the viral genome organization.

Such asymmetries are shown here to correlate with the viral replication mechanisms, as is the case with bacteria and mitochondria, despite clear differences between these genomes. Thus compositional biases may reflect fundamental properties of DNA metabolism, observed on a whole-genome scale. Many bi-directionally replicated viruses also exhibit guanine versus cytosine excess in the leading strand (downstream of replication origin). Such excess changes linearly across adenovirus linear DNA replicating from one end. And in papillomavirus genomes, with genes transcribed from one strand, the behavior of this excess suggests that transcription and replication (and associated repair processes) may produce comparable effects in observed biases.

2. Materials and methods

For the set of 22 viral genomes, the complete sequences were obtained from GenBank (Table 1). The method of cumulative skew diagrams (Grigoriev, 1998a) was used to study nucleotide composition asymmetries. These diagrams present plots of cumulative skew calculated as a sum of $([N_1] - [N_2])/([N_1] + [N_2])$, where $[N_1]$ and $[N_2]$ are counts of two complementary nucleotides in adja-

Table 1
Viral genomes analyzed

Genome	References	Accession number
Adenovirus type 2	Roberts et al., 1986	J01917
Adenovirus type 5	Chroboczek et al., 1992	*M73260*
Adenovirus type 40	Davison et al., 1993	L19443
Adenovirus type 12	Sprengel et al., 1994	X73487
Human papillomavirus 1A (HPV-1A)	Danos et al., 1982	V01116
Human papillomavirus 11 (HPV-11)	Dartmann et al., 1986	M14119
Human papillomavirus 16 (HPV-16)	Seedorf et al., 1985	K02718
Human papillomavirus 33 (HPV-33)	Cole and Streeck, 1986	M12732
Bovine papillomavirus 1 (BPV-1)	Chen et al., 1982	X02346
Bovine papillomavirus 2 (BPV-2)	Groff et al., 1986(unpublished)	M20219
Simian virus 40 (SV40)	Reddy et al., 1978; Fiers et al., 1978	J02400
Polyomavirus strain A-2	Griffin et al., 1981	J02288
Polyomavirus strain A-3	Deininger et al., 1980	J02289
Polyomavirus bovis	Schuurman et al., 1991	M74843
JC virus (JCV)	Frisque et al., 1984	J02226
Human papovavirus BK (BKV)	Seif et al., 1979	J02038
Human cytomegalovirus (HCMV)	Bankier et al., 1991	X17403
Epstein-Barr virus (EBV)	Baer et al., 1984	V01555
Herpes simplex virus type 1 (HSV-1)	McGeoch et al., 1988	X14112
Herpes simplex virus type 2 (HSV-2)	Dolan et al., 1998	Z86099
Human herpesvirus type 6 (HHV-6)	Gompels et al., 1995	X83413
Human herpesvirus type 7 (HHV-7)	Nicholas, 1996	U43400

cent non-overlapping windows, from an arbitrary start to a given point in a sequence. Diagrams of cumulative purine excess (adopted from Freeman et al., 1998) represent plots of ([G] + [A]) − ([C] + [T]) in the same sequence windows. Coding sequences and direction of their transcription were taken from the CDS fields of respective GenBank files.

3. Results

3.1. Cumulative skew diagrams

A skew function measures relative excess of one nucleotide against its complementary counterpart in a sequence window on one DNA strand. Let us consider guanine-cytosine skew, or GC skew (AT skew defined analogously to what follows). GC skew is calculated as ([G] − [C])/([G] + [C]), where [G] and [C] stand for counts of guanines and cytosines in the window. This skew varies in the interval [− 1; 1]; its value close to 1 corresponds to a large excess of G versus C (so that both [G] − [C] and [G] + [C] approach [G]), while − 1 indicates the opposite bias. Zero skew reflects intrastrand parity of G and C within a window. The G + C composition in a window is taken into account in such a way that, e.g. a twofold excess of G versus C produces a skew value of 1/3, regardless of the G + C composition.

A plot of GC skew in 60-kb windows is given in Fig. 1A for the SV40 genome, whose compositional asymmetry is well-known: one half of the sequence is G-rich and the other is C-rich (Smithies et al., 1981; Filipski, 1990). One can see that GC skew changes sign somewhere near the 50% coordinate (global polarity switch) and there are multiple local switches throughout the plot. It is unclear which of these switches in the middle of the plot is actually the global one. Increasing the window size lowers the number of switches but hides the exact coordinate of the global switch. Smoothing the plot by averaging GC skew in sliding windows does not remove most of the local switches (not shown).

Recently, I have proposed to use cumulative skew diagrams (Grigoriev, 1998a), which simplify

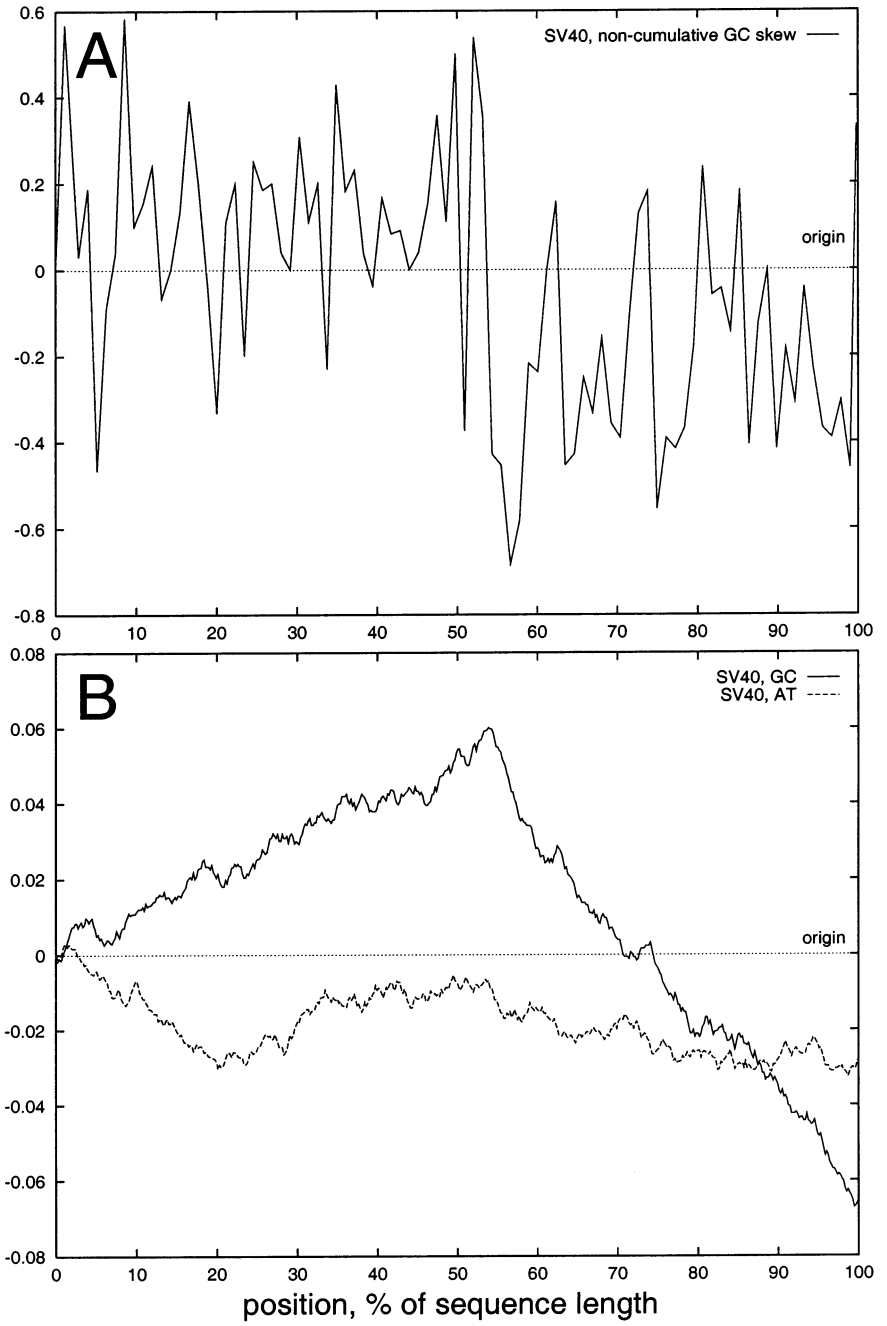

Fig. 1. (A) Non-cumulative GC skew (in 60-bp windows, smaller window sizes produce considerable fluctuations) and (B) cumulative GC and AT diagrams (in 10-bp windows) plotted along the genome of SV40. Replication origin location is at 0/100% (circular genome).

the analysis by replacing global and local switches by global and local extrema. Such diagrams plot cumulative skew, calculated as a sum of skew values in adjacent windows from an arbitrary start to a given point in a sequence. To minimize the influence of a window size W and genome length L on cumulative plots, the skew values are multiplied by W/L. For all genomes analyzed here, the window size was selected to be sufficiently small ($W/L < 0.005$; 10–300-bp windows were used, depending on a genome size). The plots with such small windows are practically the same as those with a window size of one nucleotide (not shown). Then cumulative skew is equivalent to a numerical integration of the skew function over the genome length.

The resulting integral skew plots are named cumulative GC (or AT) diagrams. An example of cumulative GC and AT skew diagrams of SV40 is given in Fig. 1B. It is clearly seen that multiple local polarity switches on the non-cumulative GC skew plot in Fig. 1A correspond to local minima and maxima on the GC diagram. The global maximum at 54% separates two genome segments with the opposite deviations from the parity rule [G] = [C]. Assuming normally distributed deviation within these segments, the slopes of the opposite linear trends on the GC diagram correspond to respective mean GC skews of the segments.

The observed asymmetry must be placed in context. The 5243-bp SV40 genome replicates bidirectionally from a single origin, spanning the region from 5191 to 83 bp (Li and Kelly, 1984). Similarly replicated microbial genomes can be identified by a characteristic inverted V-shape (if the sequence start is selected at the origin site) of their GC diagrams (Grigoriev, 1998a). The same shape can be seen in the GC diagram of SV40 (Fig. 1B), with the global minimum located exactly at the replication origin (0 or 100% in the diagram, since the genome is circular). The global maximum is at the opposite end of the circle (54% in the diagram). GC skew is positive for the leading (left half of the GC diagram) and negative for the lagging strand, as is the case with microbial genomes (Grigoriev, 1998a).

The two segments of the GC diagram also correspond to the divergently transcribed coding sequences of SV40. It is worth noting that the slopes of the two halves of the GC diagram V-shape are different (in contrast, most of the diagrams in bacterial genomes are more symmetrical, hence the term, V-shape). The excess of guanine over cytosine in the leading strand in the late mRNA region of SV40 is nearly half of the excess of cytosine over guanine in the lagging strand in the early mRNA region.

No comparable clear-cut partitioning of the genome with respect to the origin and terminus sites occurs in the AT diagram (Fig. 1B). Irregular behavior of AT (in contrast to GC) diagrams has also been found in microbial genomes (Grigoriev, 1998a). GC diagrams of related genomes of JC virus, BK virus and polyomavirus bovis show very similar behavior to that of SV40. Their AT diagrams feature more pronounced centrally-located global maxima, as do those of polyomavirus strain A-2 and A-3 (not shown). The GC diagrams of the latter two strains are similar to the AT diagram of SV40 (Fig. 1B).

3.2. Analysis of different genomes

In what follows, different diagram types, found in 22 viral genomes analyzed, are presented, together with brief interpretation to relate genome features with diagram behavior. For all sequences, the start was selected at "the basepair 1" of a respective GenBank file (Table 1). In fact, one feature of the diagrams is that when a starting point of a circular sequence is changed, the relative distance between global peaks of one genome on the x-axis does not alter and any global minimum remains a global minimum, although its position is shifted accordingly. Analogously, the total skew (the point where a plot intersects the right-hand vertical axis, corresponding to the sum of skew over the whole genome) would be the same for any starting point. To allow for plotting graphs of several genomes on the same diagram, the basepair coordinate is replaced by a percentage of the genome size on the x-axis.

3.2.1. Adenovirus

Two strands of linear DNA of human adenovirus (identical origins at each end of the

genome) replicate in a fashion that leaves one of them in a single-stranded state while another is being duplicated (Horwitz, 1976; Stillman, 1981). This means that the displaced strand may be subject to different mutation pressure than the template strand. Assuming a constant speed of replication, this pressure will change along the sequence, as the time the displaced DNA spends single-stranded decreases linearly from one end of the molecule to another. Integration of a linear function results in a second-order polynomial, a parabola.

Remarkably, the GC diagram of the 34-Kbp genome of adenovirus type 40 (Fig. 2A) has a shape very close to parabolic ($R^2 = 0.99$). It points upwards, reflecting a decrease in the skew value from positive to negative along one strand, consistent with the replication mode (Grigoriev, 1998a). The parabolic trendline reaches its global maximum (meaning that the GC skew equals zero) at about 44% of the sequence. The AT diagram is also indicative of a linear increase (left to right) of AT skew in about 20% of the sequence on each end of the genome and a positive AT skew for the most of the remaining 60%. Effects of transcription do not seem to be clearly pronounced, except for the late DBP and E4 mRNA regions (diagram positions 60–70% and 92–100%, respectively), oriented opposite to the majority of the genes. These regions correspond to the plot segments where the main deviations from the parabolic GC diagram shape can be seen.

Comparisons with the diagrams of adenovirus types 5 and 12 (36 and 34 Kbp, respectively) show that the main trends in the diagram behavior of adenoviral genomes are conserved (Fig. 2B). Also, the diagrams for adenovirus type 2 (not shown) are almost identical to those of the closely related adenovirus type 5.

3.2.2. Human cytomegalovirus

The large (nearly 230 Kbp) human cytomegalovirus (HCMV) genome replicates bi-directionally (Hamzeh et al., 1990), starting at the origin site *oriLyt* located between 92 210 and 93 715 bp (Anders et al., 1992; Masse et al., 1992). Its GC diagram also shows a V-shape (Fig. 3A), although it is inverted, compared to the SV40 diagram (Fig. 1A). This reflects the fact that HCMV replication origin is placed at about 40% of the plot. The global minimum on the GC diagram is located at $\sim$96 Kbp, 1.5% of the genome length away from the middle of *oriLyt*. The AT diagram has a global maximum at *oriLyt* but its global minimum is at position 84%.

Gene orientation bias is detected to the right of *oriLyt* (Table 2). In the 20% interval, centered around the global minimum on the GC diagram, most of the genes are transcribed in one direction (Table 2). So (unlike the situation with SV40) there is no obvious relationship between the global switch of GC skew and gene orientation in HCMV.

3.2.3. Epstein-Barr virus

Two replication origins are known in the 172-Kbp long Epstein-Barr virus (EBV) genome. Circle-to-circle replication during the latent phase starts from the *oriP* origin (Yates et al., 1984) located between 7315 and 9312 bp. The *oriLyt* origin between the divergent BHLF 1 and BHRF 1 genes at 52 Kbp functions during lytic-phase DNA synthesis (Hammerschmidt and Sugden, 1988). The diagrams of EBV (Fig. 3B) are more complex than those presented above. The global minimum and maximum on the GC diagram are some 2 Kbp upstream of *oriLyt* and *oriP* locations, respectively.

Both diagrams can be divided into three different parts. The region between 7 and 52% on the GC diagram is a V-shape whose arms are quite different, especially remarkable is the 'jagged' left half where short stretches of positive GC skew are visible at regular intervals against the constant background of negative GC skew. One likely reason for these distortions is the 3072-bp repeat, occurring 12 times only in this stretch of the EBV genome. The same 'jagged' pattern is seen on the AT diagram. In the region between positions 29 and 54% the AT and GC diagrams display opposite trends. The third part, outside the GC V-shape displays a rather irregular skew behavior.

Practically all of the EBV genes between 0 and 48% are transcribed in one direction (left to right), traversing both origins. The genes between 86 and 100% are in the opposite orientation. For these

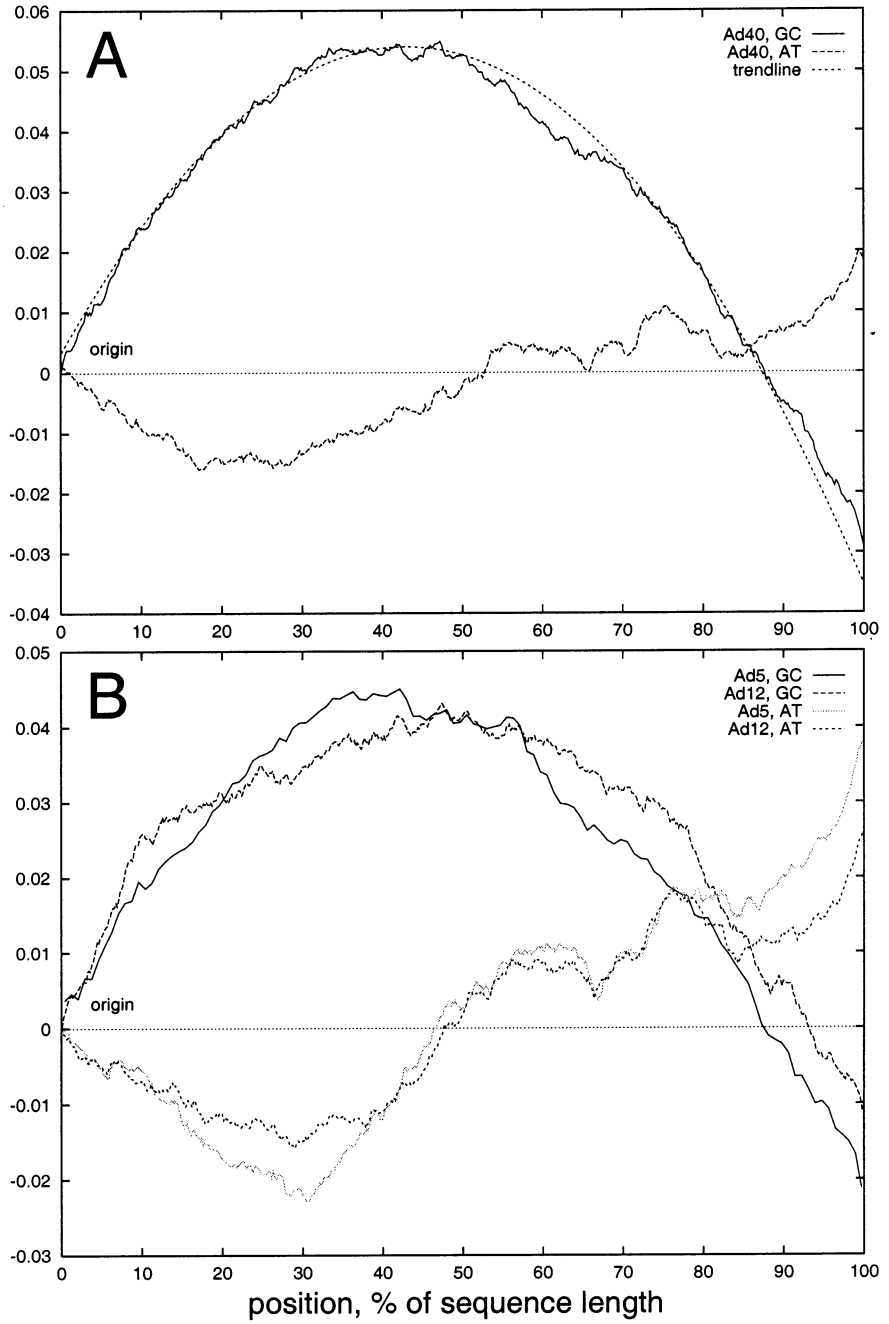

Fig. 2. GC and AT diagrams for adenovirus type 40 (A), 5 and 12 (B). Replication origin location for the plotted strand is at 0% (linear genomes).

two regions, the diagram behavior reflects the directionality of *oriLyt*-driven replication and characteristic repeat structure of EBV. In the remaining 38% of the genome, there are frequent changes in the direction of transcription and this part of diagrams is harder to interpret.

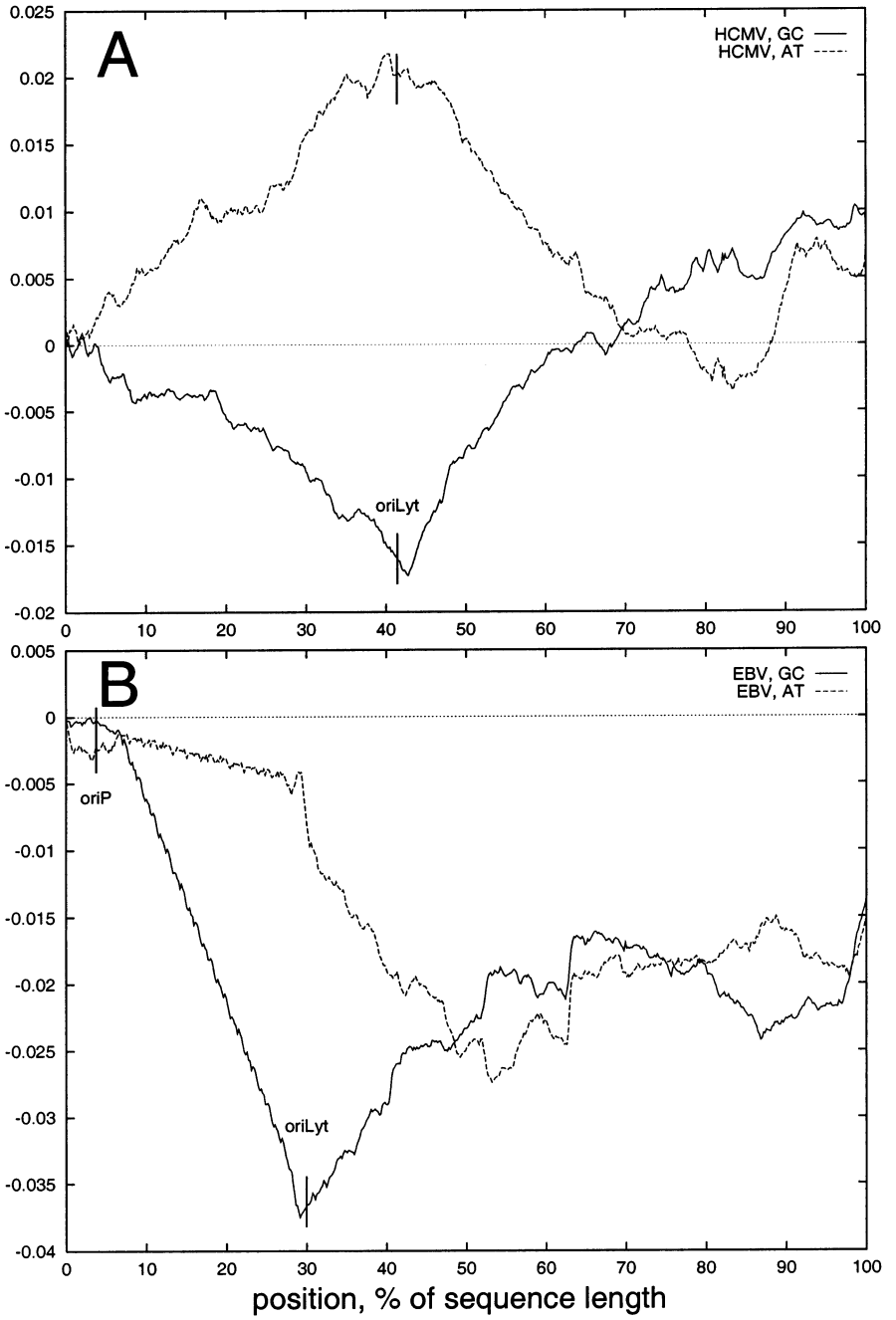

Fig. 3. GC and AT diagrams for cytomegalovirus (A) and Epstein-Barr virus (B). Locations of *oriP* and *oriLyt* origins are marked by vertical ticks, crossing the plots.

3.2.4. *Human herpesvirus*

Human herpesvirus HHV-6 has a linear genome of 159 Kbp; bi-directional replication starts from the origin *oriLyt*, between 67 617 and 67 993 bp (Dewhurst et al., 1993). The global minimum on the GC diagram (Fig. 4A) coincides

Table 2
Distribution of coding sequences in different genome regions with respect to a diagram global extremum (GE)

Genome	GE, %% coordinate	0%→GE (left part)	GE→100% (right part)	20% interval centered on GE
HCMV	42.5 (Fig. 3A)	38/37[a]	45/83	11/24
HHV-6	43 (Fig. 4)	16/36	32/36[b]	7/14
HSV- 1	55 (Fig. 5)	18/22	22/15	7/4[c]

[a] Ratios n_R/n_L are shown, where n_R is a number of coding sequences (CDS in GenBank files) transcribed from left to right, and n_L is a number of right-to-left CDS (plotted strand is their transcribed strand).
[b] This difference (n_R-n_L) of four CDS is attributable to the interval 82→100%.
[c] This ratio may be misleading, since one of the right-to-left CDS is the 9.5 Kbp long UL36 (very large tegument protein).

with the position of *oriLyt*. There is a plateau of zero GC skew between 82 and 95%, which is also present on the diagram of another human herpesvirus, HHV-7, although the V-shape of the latter is more shallow. HHV-7 has a genome of 145 Kbp and is related to HHV-6 (Nicholas, 1996). Its origin *oriLyt* has been recently localized to a position about 62.5 Kbp (van Loon et al., 1997), close to the GC diagram minimum. The AT diagrams for both viruses (Fig. 4B) are in the opposite phase to the GC diagrams, displaying inverted V-shapes. The plateau is only seen on the AT diagram of HHV-6, it is much less pronounced in HHV-7, and the global peak on AT diagram of the latter is shifted to the left some 7% of the genome length.

Gene orientation bias is seen to the left of *oriLyt*, in contrast to HCMV (Table 2). As in HCMV, 2/3 of the human herpesvirus (HHV) genes are transcribed in one direction in the 20% interval, centered around the global minimum on the GC diagram (Table 2). Near the diagram plateau, transcription divergent with replication is locally favored (Table 2).

3.2.5. Herpes simplex virus

Herpes simplex virus type 1 (HSV-1) has a linear genome of 152 Kbp, and its origin sequences have been located in three places: *oriL* in the middle of L region (Spaete and Frenkel, 1985) and two copies of *oriS* in the inverted repeats flanking the S region (Stow, 1982). Non-conservative homologous recombination may be involved in circularization of the genome prior to replication (Yao et al., 1997). The GC diagrams (Fig. 5A) reflect this in that they show an inverted V-shape, corresponding to the origin location near the right-hand end of V and replication termination in the middle of the genome (global maximum at 55%). A small plateau between 30 and 46% on the diagram, corresponding to the GC skew close to zero, contains *oriL*. The same features are seen on the GC diagram (Fig. 5A) of HSV-2 (155 Kbp). *oriL* and *oriS* are very close to the main local minima on the AT diagram (Fig. 5B).

Gene orientation pattern is different from those in HCMV and HHV: no comparable bias in one arm of the V-shape is detected and direction of transcription frequently changes across HSV genomes (Table 2). This may be related to a relatively high number of distortions on the diagrams: their V-shapes are 'degraded', compared to the other herpesviruses above.

The AT diagrams of HSV-1 and HSV-2 (Fig. 5B) are almost identical. Such diagram similarity allows one to identify some of the differences between these genomes purely on the basis of plot shapes. The main compositional differences between HSV-1 and HSV-2 are found in left and right-hand terminal portions of the diagrams, corresponding to the TR_L and U_S regions. These contain the most diverged parts of the two genomes (McGeoch et al., 1991). Potential sites of genome rearrangements undergone by herpesviruses can be identified as local switches in the slope of a diagram (where the leading and lagging strand may have been recently swapped as a result of an inversion, or changed length via translocations and insertions/deletions). The most prominent shift on the GC diagram is that of the local

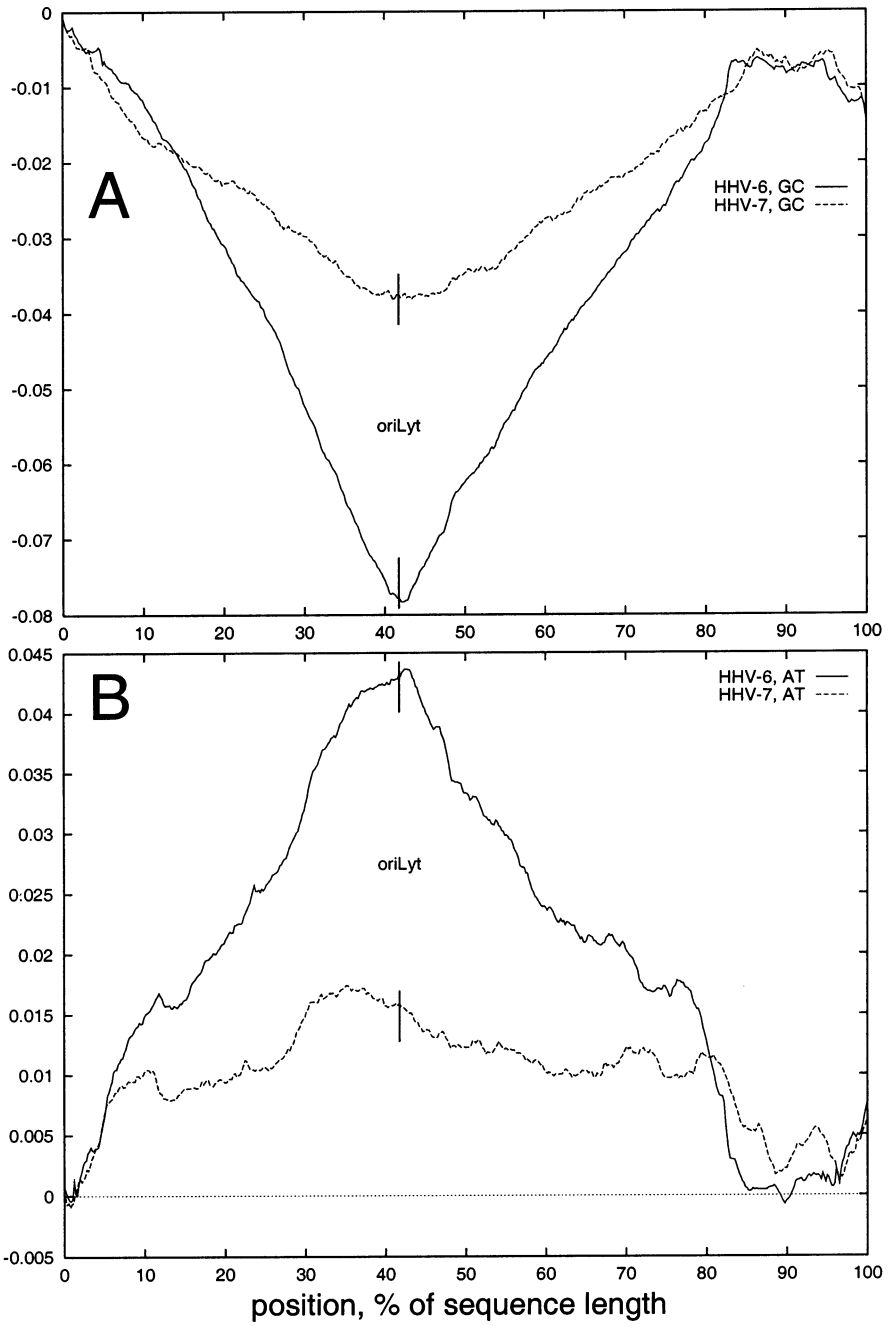

Fig. 4. GC (A) and AT (B) diagrams for human herpesvirus type 6 and 7. Locations of *oriLyt* origins are marked by vertical ticks, crossing the plots.

maxima near the 90% position. It indicates a deletion in HSV-1, since its local maximum is shifted to the right: indeed, there is a deletion in the US4 gene, to which nearly all the length difference between these genomes is attributable (Dolan et al., 1998).

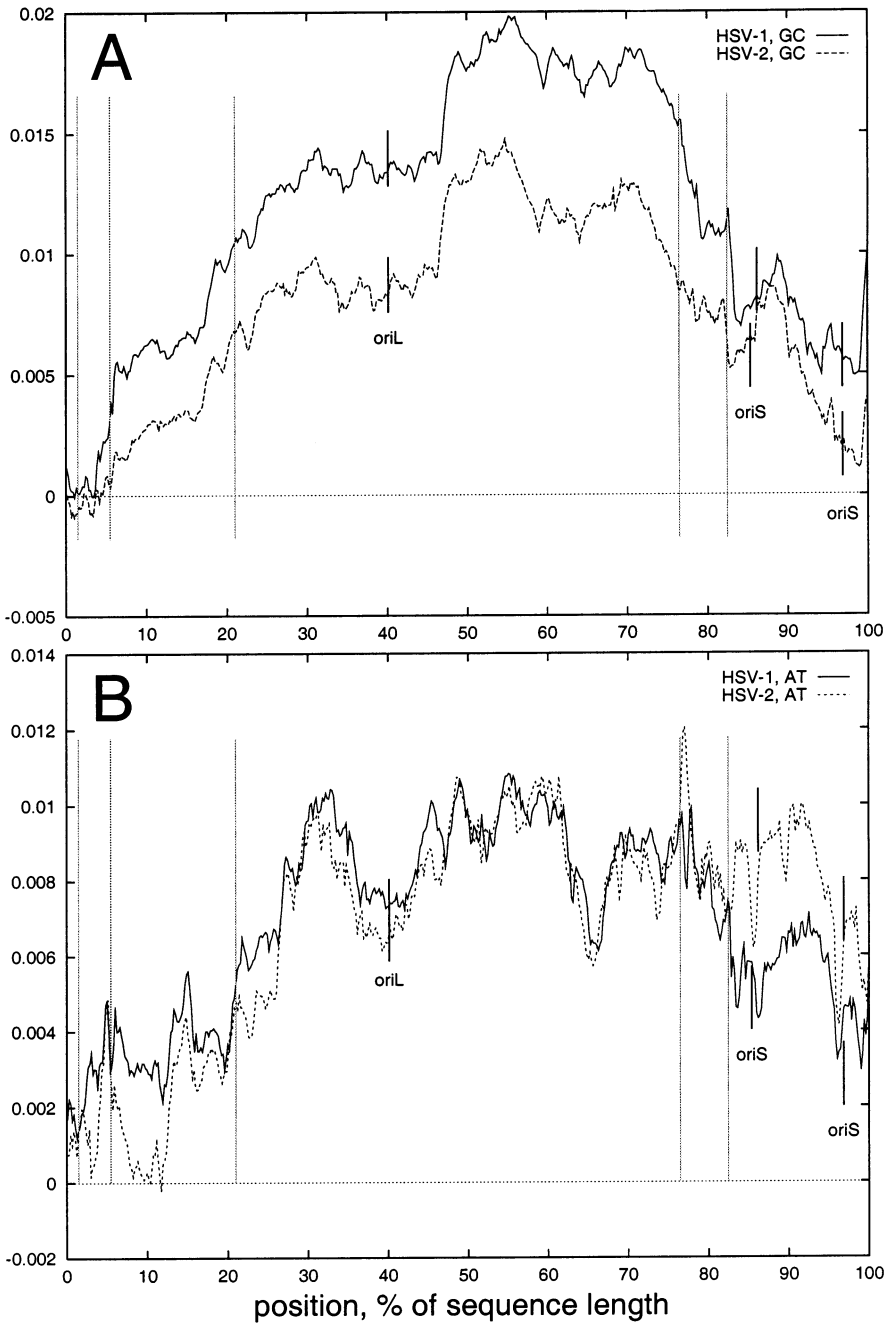

Fig. 5. GC (A) and AT (B) diagrams for herpes simplex virus type 1 and 2. Locations of *oriL* and *oriS* origins are marked by vertical ticks, crossing the plots. Longer vertical lines indicate points of potential genome rearrangements (see text).

Positions corresponding to the left-hand coordinates of a few (by far not all) other potential rearrangements, are shown in Fig. 5 as long vertical lines. While the AT diagram local differences at these positions are clear, they are often almost negligible in the respective portions of the GC

diagram. In the actual sequences, the marked positions correspond to the regions near (from left to right): the end of the first exon in RL2 gene; between the LAT intron and repeat reiteration set 3; the second exon of UL15 gene; IR_L upstream of UL56 gene; and the start of IR_S region. Some of these differences are discussed in more detail in Dolan et al. (1998).

3.2.6. Papillomavirus

Human papillomavirus types 1A, 11, 16, 33 and bovine papillomavirus types 1 and 2 have circular genomes of about 7.9 Kbp. Replication starts from the origin close to the upstream regulatory region (Auborn et al., 1994; Flores and Lambert, 1997), corresponding to 0 or 100% on the diagrams (Fig. 6). While the replication is bidirectional, the orientation of papillomavirus genes is very different: they all are transcribed from one strand. So if there are separate replication and transcription-induced biases, they should act in the same direction in one half of a papillomavirus genome, and in the opposite directions in the other half. Under this model, an explanation of the observed diagram behavior in Fig. 6 may be that the steeper slopes on the left reflect a sum of the net effects of replication and transcription, and the right-hand parts of the diagrams correspond to their subtraction.

If this model is correct, the nearly zero slope on the right of the HPV- 1A GC diagram (Fig. 6A, and similar diagrams of bovine papillomavirus 1 (BPV-1) and bovine papillomavirus 2 (BPV-2), not shown) suggests that a contribution of transcription is comparable to that of replication in these genomes. They almost cancel each other out in the region between 50 and 100% on the HPV-1A diagram ([G] = 758, [C] = 773), and their concerted effects result in a significant guanine excess ([G] = 900, [C] = 690) in the other half of the genome.

The GC diagram of HPV-16 (Fig. 6A, and similar diagrams for HPV-11 and HPV-33, not shown) also behaves differently to the left and to the right of the position 40%. The right-hand part contains several local minima and maxima between 75 and 92% positions, occupied by the major capsid protein gene. The AT diagrams (Fig. 6B) of these genomes also display much steeper slopes in the left half of the sequence, in agreement with the model above.

3.3. Comparison with purine excess diagrams

In addition to the skew diagrams above, one can analyze genome composition in terms of cumulative excesses of purine or keto-bases (Freeman et al., 1998). The differences between these measures and cumulative skew have been discussed recently (Grigoriev, 1998b). Cumulative purine excess can be presented as cumulative purine diagrams, plotting [G] + [A] − [C] − [T] instead of skew in the same sequence windows.

For polyomavirus genomes, purine diagrams are very similar (Fig. 7A). Except for the first 4% on the diagram and small local distortions, the plots are almost parallel. The differences seen on GC and AT diagrams between SV40 (clear central peak on the GC diagram) and polyomavirus strains A-2 and A-3 (clear central peak on the AT diagram) are lost on the purine diagram. This is due to the fact that purine excess is equivalent to a sum of GC and AT skew in numerical integration with small windows (Grigoriev, 1998b).

Effects of such summation are visible on the purine diagrams of other genomes (Fig. 7B). In adenovirus, linear fit ($R^2 = 0.98$ in the left and $R^2 = 0.95$ in the right half of the diagram) is better than parabolic ($R^2 = 0.93$, compare with $R^2 = 0.99$ in Fig. 2A). In HCMV, the prominent global extreme of GC and AT diagrams around 40% in Fig. 3A are replaced by a smaller minimum in Fig. 7B, and additional extremes appear near 18 and 92% coordinates, where local maxima of the AT diagram reside. On the purine diagram of HHV-7, the global minimum is shifted away from *oriLyt* to the 49% coordinate, and the global maximum is at 80%.

In EBV, behavior of the purine diagram is similar to that of GC (Fig. 2B), although in the region 30–100% the former plot is more shallow (not shown). For papillomavirus genomes, purine excess diagrams somewhat resemble the AT diagrams (Fig. 6B), with the steeper slopes corresponding to the first 50% of the sequence (not shown).

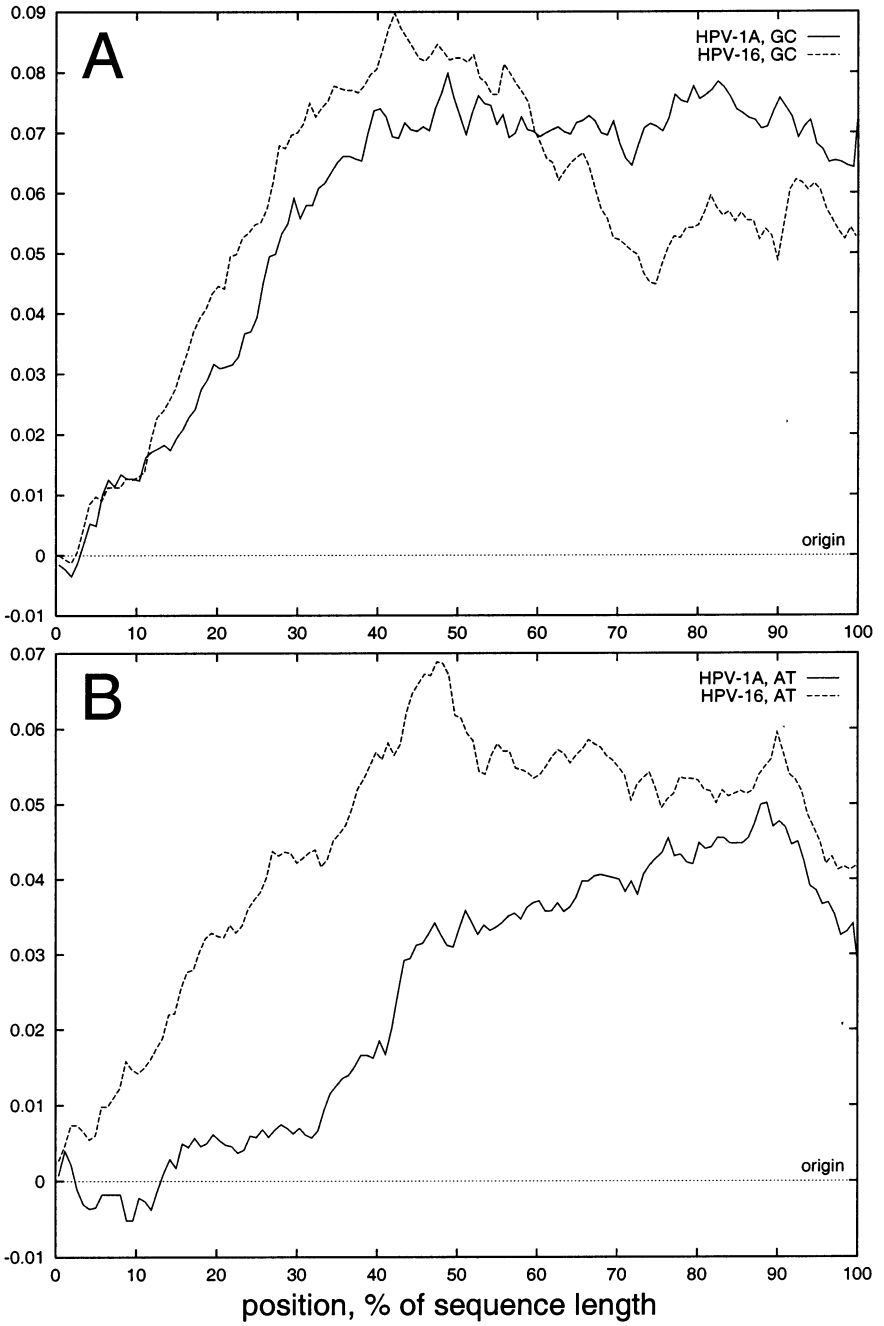

Fig. 6. GC (A) and AT (B) diagrams for human papillomavirus type 1A and 16. Replication origin location is at 0/100% (circular genomes).

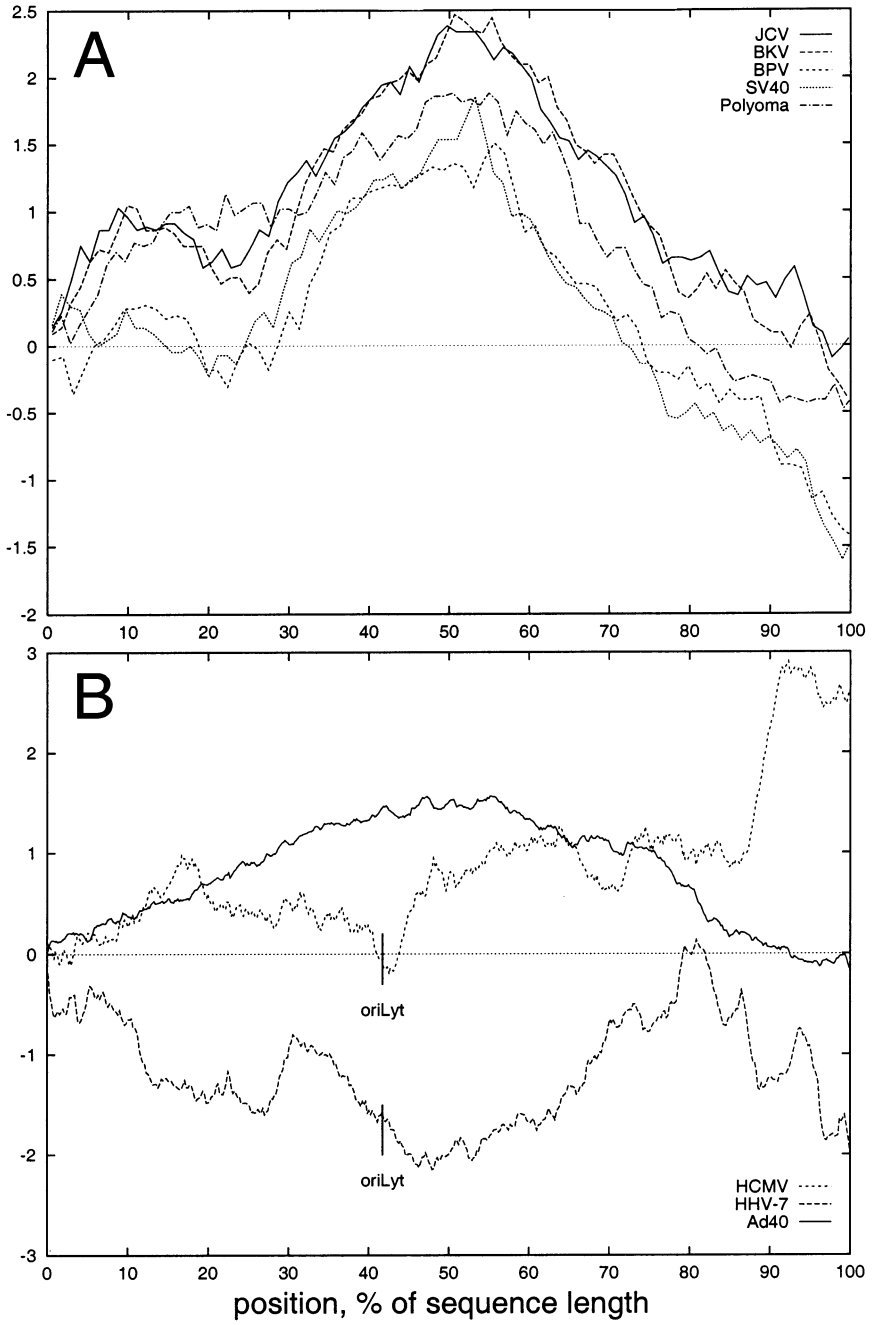

Fig. 7. Purine diagrams for polyomavirus genomes in 60-bp windows (A), adenovirus type 40, HCMV and HHV-7 in 300-bp windows (B). Locations of *oriLyt* are shown for the latter two genomes.

4. Discussion

Strand compositional bias and changes of bias across large genomic regions of 22 dsDNA viruses were analyzed using the method of cumulative skew diagrams. Asymmetries were observed between leading and lagging, and between transcribed and non-transcribed strands. The diagram behavior of these genomes suggests a link between the direction of replication/transcription and DNA composition.

4.1. Replication

The viral genomes analyzed vary in size (from 5 to 230 Kbp), genome organization and replication mechanisms. However, for all bi-directionally replicated genomes, the observed leading strand bias is different from that in the lagging strand. For most of these genomes, GC skew changes sign near the origin and terminus of replication, resulting in V-shaped diagrams (Fig. 1, Figs. 3–5). The same effect has been detected in microbial genomes, and has been hypothesized to result from asymmetries of DNA synthesis and repair (Lobry, 1996; Grigoriev, 1998a). Assuming a constant speed of a replication fork, equal segments of the leading strand should spend equal amounts of time single-stranded (being more prone to damage) during replication.

The very different GC diagrams of adenovirus demonstrate a linear increase of the GC skew towards the genome end, which remains single-stranded until the end of a replication round (left 5′ end of the sequence plotted in Fig. 2A). This also suggests that the bias may depend on the time DNA spends single-stranded as a replication template. Skew reaches zero near the middle of the sequence, so both origins located at opposite ends contribute to this linear change (each in its respective strand). This behavior is conserved in other adenoviruses (Fig. 2B). A similar linear skew change has been also reported for diagrams of vertebrate mitochondria where the heavy strand replication origin spends considerably longer time single-stranded during replication (Grigoriev, 1998a).

4.2. Replication and transcription

In papillomavirus genomes, replication proceeds in the direction of transcription in one half of a sequence, and opposite to transcription in another half. This is reflected in the diagrams of HPV-1A (Fig. 5), BPV-1 and BPV-2, where the latter orientation seems to decrease AT skew and almost annihilate GC skew. Analogous effects, albeit more local, are seen in HPV-16 (Fig. 5), HPV-11 and HPV-33. Since transcription also leaves a portion of non-transcribed strand single-stranded, the resulting mutational damage may either add to (when transcription is collinear with replication) or compensate (opposite orientation) the replication-induced effects.

Such interplay between transcription and replication may be involved in the observed differences between two arms of V-shaped diagrams of SV40, JC virus, BK virus and polyomavirus bovis, where GC skew in the late mRNA region is about a half of that in the late mRNA region (Fig. 1). Transcription-induced effects may be also responsible for the slight deviations from a perfect parabolic shape of the adenovirus GC diagram (Fig. 2A).

4.3. Transcription versus replication

The direction of transcription often coincides with that of replication in bacteria (Brewer, 1990) and viruses. This complicates the analysis, making it harder to distinguish between relative contributions of these two (and related repair) processes to the observed mutational changes. Skew changes sign near origins and termini of replication in many genomes, supporting a hypothesis of replication-induced bias (Lobry, 1996; Blattner et al., 1997; Grigoriev, 1998a). Studies at the level of individual bacterial genes (Beletskii and Bhagwat, 1996; Francino et al., 1996) and detected correlations (strongest in archaea) between gene orientation and purine excess (Freeman et al., 1998) favor transcription-based models (Francino and Ochman, 1997). While biases in the polyomavirus genomes can be explained purely on the basis of gene orientation, this does not seem to apply to the larger viral genomes (e.g. HCMV, HHV; Table 2), where the majority of the genes on both

sides of *oriLyt* are encoded in one orientation. Moreover, the diagrams of adenovirus (Fig. 2), HCMV (Fig. 3A), HHV (Fig. 4) and papillomavirus (Fig. 6) strongly support the view that transcription alone cannot account for the observed bias. The example of HPV-1A (Fig. 6A) actually suggests that the net contribution of transcription-related effects in this genome is probably close to that of replication, in terms of GC skew.

4.4. Potential sources of bias

For single-stranded DNA, rates of spontaneous deamination of C or 5-methylcytosine are more than 100-fold compared to a double-stranded molecule (Frederico et al., 1990). If a deaminated base in a single-stranded replication template is paired subsequently with A, that would lead to a relative abundance of G and T on the template strand. Also, it has been shown that transcription causes some fourfold increase in spontaneous deamination in the non-transcribed strand (Beletskii and Bhagwat, 1996).

Mismatch repair may also be involved in the observed bias. For instance, C-C mispair is a very poor substrate for the methyl-directed repair pathway (Su et al., 1988). If this mispair occurs during synthesis of the lagging strand and is not repaired, the next round of replication will result in a $C \rightarrow G$ transversion in the leading strand, contributing to its positive GC skew. Transcription-coupled repair may be involved in the observed strand asymmetries via unequal exposure of the two strands to damage and differential opportunity for repair during mRNA synthesis (for a review, see Francino and Ochman, 1997). Such repair is more efficient on the non-transcribed strand in the removal of pyrimidine dimers (Oller et al., 1992).

Fluctuations in dNTP precursor pools occur throughout the cell cycle (see Meuth, 1989, for review) and have been implicated in formation of isochores (Wolfe, 1991). It is unclear whether such fluctuations may affect strand composition differentially (especially in short viral sequences) but asymmetric replication in the presence of variable levels of competing substrates may favor point mutations on one strand. Regularly spaced repeats can also affect strand composition, introducing biases. Such repeats, together with the combined effects of co-existing origins may give rise to the more complex diagrams of HSV and EBV.

4.5. GC skew, AT skew and purine excess

Compared to GC diagrams, plots of cumulative AT skew have a variable behavior. For some organisms, they are in the opposite phase to the respective GC diagrams (e.g. HCMV, HHV-6 and HHV-7, Fig. 3A and Fig. 4), for others, in the same phase (e.g. HSV-1 and HSV-2, Fig. 5), while a mixture of these can be seen for adenoviruses and EBV (Fig. 2, Fig. 3B). Similar variability was reported for microbial AT diagrams, in contrast to a consistent behavior of GC diagrams (Grigoriev, 1998a). It remains unclear what causes this irregularity. Also, for closely related genomes, local distortions in AT diagrams are more pronounced than those in GC diagrams (Fig. 4).

Another method of compositional analysis, plotting cumulative excesses of purine ($[G]+[A]$ versus $[C]+[T]$) or keto-bases ($[G]+[T]$ versus $[C]+[A]$), has been described recently (Freeman et al., 1998). However, their approach does not distinguish between the important different contributions of the GC and AT skews, presented in this paper. Purine excess is equivalent to a sum of GC and AT skews, when integrating over small windows (or is exactly the same over windows of size 1); analogously, keto excess is equivalent to GC skew minus AT skew (Grigoriev, 1998b). As a consequence, plots of such aggregate measures as purine or keto excess may be affected by phase and amplitude differences of cumulative GC and AT skews. For example, the purine diagram of adenovirus neither favors a replication-related parabolic fit (Fig. 7B), in contrast to the GC diagram (Fig. 2A), nor can it be explained in terms of gene orientation. The purine diagram minimum is shifted away from *oriLyt* in HHV-7 but not in HCMV (Fig. 7B), although patterns of gene orientation around *oriLyt* are similar in these genomes. The GC diagram resemblance between HCMV and HHV (Fig. 2A, Fig. 4A) is practically lost on their purine diagrams.

One puzzling observation is that the global maximum on the AT diagrams of polyomavirus strains A-2 and A-3 coincides with the replication terminus, but the global maximum on the GC diagrams does not. In fact, polyomavirus strains A-2 and A-3 are the only exceptions out of nearly a hundred eucaryotic, microbial, viral, mitochondrial and chloroplast sequences analyzed (Grigoriev, 1998a and unpublished data), where a V-shaped AT diagram is not accompanied by a similar trend in a GC diagram. This is surprising, since other polyomaviruses (SV40, JC virus, BK virus and bovine polyomavirus) have clear V-shapes on their GC diagrams. Purine diagrams for polyomavirus genomes are very close (Fig. 7A) and it is tempting to explain this similarity by accumulation of pyrimidines on the transcribed strand via transcription-coupled repair (Francino and Ochman, 1997; Freeman et al., 1998). However, given the variety of presented diagrams and implied connections with replication models, such single-mechanism explanation does not seem to be adequate.

4.6. Genome comparisons

Large-scale genome comparisons can be facilitated using cumulative diagrams, even in cases of complex skew patterns, as shown for HSV-1 and HSV-2 genomes (Fig. 5). Relative distortions in the diagrams of closely related genomes reflect local variations in nucleotide substitutions and rearrangements that took place since the organisms or strains had diverged. It is, in fact, much simpler to create a cumulative skew diagram with a clear and compact representation of the genome-wide differences than to align sequences of two or more related viral genomes of >150 Kbp in length (alignment of diagrams can be easily done manually). See Grigoriev (1998a) for an example of graphical comparison of two *E. coli* strains (4.6 Mbp long).

While this paper was undergoing a review, I learned about work by Mrazek and Karlin (1998), who analyzed several bacterial and viral genomes, plotting non-cumulative GC skew in 10- and 50-Kbp sliding windows. They detected strong strand bias only in β-herpesviruses, not in α- or γ-herpesviruses. This is likely to be a result of the observed differences in the behavior of the herpesvirus diagrams (Figs. 3–5), given the lower sensitivity of the non-cumulative skew plots (Fig. 1). In agreement with the results reported here, they observed positive GC skews in the leading strand and inconsistent behavior of AT skew. They also listed a number of additional possible sources of compositional asymmetry (e.g. helicase primase bias, differences in signal or binding sites in the two strands, biases in gene density and selective amino acid/codon constrains).

Even more recently, Fijalkowska et al. (1998) have presented results on differential fidelity of leading and lagging strand replication in *E. coli*. If similar effects take place during viral replication, they may also contribute to the observed strand biases. Overall, it is likely that multiple independent and organism-specific factors related to gene expression and relative asymmetries and fidelities of DNA synthesis and repair are combined with natural selection to produce mutational sequence biases. The described approach may be a useful tool for finding relationships between sequence composition and global features of genome organization and for evolutionary studies of driving forces of genome-wide sequence change.

Acknowledgements

I would like to thank H. Lehrach for support of this project at MPIMG, I. Ivanov for critical reading of the manuscript and helpful discussions, and S. Meier-Ewert for useful advice. The interpretation of the results was greatly improved by stimulating comments from one of the anonymous referees.

References

Anders, D.G., Kacica, M.A., Pari, G., Punturieri, S.M., 1992. Boundaries and structure of human cytomegalovirus ori-Lyt, a complex origin for lytic-phase DNA replication. J. Virol. 66, 3373–3384.

Auborn, K.J., Little, R.D., Platt, T.H., Vaccariello, M.A., Schildkraut, C.L., 1994. Replicative intermediates of human papillomavirus type 11 in laryngeal papillomas: site of

replication initiation and direction of replication. Proc. Natl. Acad. Sci. USA 91, 7340–7344.

Baer, R.J., Bankier, A.T., Biggin, M.D., Deininger, P.L., Farrell, P.J., Gibson, T.J., Hatfull, G.F., Hudson, G.S., Satchwell, S.C., Seguin, C., Tuffnell, P.S., Barrell, B.G., 1984. DNA sequence and expression of the B95-8 Epstein-Barr virus genome. Nature 310, 207–211.

Bankier, A.T., Beck, S., Bohni, R., Brown, C.M., Cerny, R., Chee, M.S., Hutchinson III, C.A., Kouzarides, T., Martignetti, J.A., Preddie, E., Satchwell, S.C., Tomlinson, P., Weston, K.M., Barrell, B.G., 1991. The DNA sequence of the human cytomegalovirus genome. DNA Seq. 2, 1–12.

Beletskii, A., Bhagwat, A.S., 1996. Transcription-induced mutations: increase in C to T mutations in the non-transcribed strand during transcription in *Escherichia coli*. Proc. Natl. Acad. Sci. USA 93, 13919–13924.

Blattner, F.R., Plunkett, G., Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., Gregor, J., Davis, N.W., Kirkpatrick, H.A., Goeden, M.A., Rose, D.J., Mau, B., Shao, Y., 1997. The complete genome sequence of *Escherichia coli* K-12. Science 277, 1453–1474.

Brewer, B.J., 1990. Replication and the transcriptional organization of the *Escherichia coli* chromosome. In: Drilica, K., Riley, M. (Eds.), The Bacterial Chromosome. ASM, Washington, DC, pp. 61–83.

Chen, E.Y., Howley, P.M., Levinson, A.D., Seeburg, P.H., 1982. The primary structure and genetic organization of the bovine papillomavirus type 1 genome. Nature 299, 529–534.

Chroboczek, J., Bieber, F., Jacrot, B., 1992. The sequence of the genome of adenovirus type 5 and its comparison with the genome of adenovirus type 2. Virology 186, 280–285.

Cole, S.T., Streeck, R.E., 1986. Genome organization and nucleotide sequence of human papillomavirus type 33, which is associated with cervical cancer. J. Virol. 58, 991–995.

Danos, O., Katinka, M., Yaniv, M., 1982. Human papillomavirus 1a complete DNA sequence: a novel type of genome organization among papovaviridae. EMBO J. 1, 231–236.

Dartmann, K., Schwarz, E., Gissmann, L., zur Hausen, H., 1986. The nucleotide sequence and genome organization of human papilloma virus type 11. Virology 151, 124–130.

Davison, A.J., Telford, E.A., Watson, M.S., McBride, K., Mautner, V., 1993. The DNA sequence of adenovirus type 40. J. Mol. Biol. 234, 1308–1316.

Deininger, P.L., Esty, A., Laporte, P., Hsu, H., Friedmann, T., 1980. The nucleotide sequence and restriction enzyme sites of the polyoma genome. Nucleic Acids Res. 8, 855–860.

Dewhurst, S., Dollard, S.C., Pellett, P.E., Dambaugh, T.R., 1993. Identification of a lytic-phase origin of DNA replication in human herpesvirus 6B strain Z29. J. Virol. 67, 7680–7683.

Dolan, A., Jamieson, F.E., Cunningham, C., Barnett, B.C., McGeoch, D.J., 1998. The genome sequence of herpes simplex virus type 2. J. Virol. 72, 2010–2021.

Fiers, W., Contreras, R., Haegemann, G., Rogiers, R., Van de Voorde, A., Van Heuverswyn, H., Van Herreweghe, J., Volckaert, G., Ysebaert, M., 1978. Complete nucleotide sequence of SV40 DNA. Nature 273, 113–120.

Fijalkowska, I.J., Jonczyk, P., Tkaczyk, M.M., Bialoskorska, M., Schaaper, R.M., 1998. Unequal fidelity of leading strand and lagging strand DNA replication on the *Escherichia coli* chromosome. Proc. Natl. Acad. Sci. USA 95, 10020–10025.

Filipski, J., 1990. Evolution of DNA sequence, contributions of mutational bias and selection to the origin of chromosomal compartments. In: Ole, G. (Ed.), Advances in Mutagenesis Research 2. Springer, Berlin.

Flores, E.R., Lambert, P.F., 1997. Evidence for a switch in the mode of human papillomavirus type 16 DNA replication during the viral life cycle. J. Virol. 71, 7167–7179.

Francino, M.P., Chao, L., Riley, M.A., Ochman, H., 1996. Asymmetries generated by transcription-coupled repair in enterobacterial genes. Science 272, 107–109.

Francino, M.P., Ochman, H., 1997. Strand asymmetries in DNA evolution. Trends Genet. 13, 240–245.

Frederico, L.A., Kunkel, T. A., Shaw, B.R., 1990. A sensitive genetic assay for the detection of cytosine deamination: determination of rate constants and the activation energy. Biochemistry 29, 2532–2537.

Freeman, J.M, Plasterer, T.N., Smith, T.F., Mohr, S.C, 1998. Patterns of genome organization in bacteria. Science 279, 1827.

Frisque, R.J., Bream, G.L., Cannella, M.T., 1984. Human polyomavirus JC virus genome. J. Virol. 51, 458–469.

Gompels, U.A., Nicholas, J., Lawrence, G., Jones, M., Thomson, B.J., Martin, M.E., Efstathiou, S., Craxton, M., Macaulay, H.A., 1995. The DNA sequence of human herpesvirus-6: structure, coding content, and genome evolution. Virology 209, 29–51.

Griffin, B.E., Soeda, E., Barrell, B.G., Staden, R., 1981. In: Tooze, J. (Ed.), DNA Tumor Viruses, 2nd ed. Sequence and analysis of polyoma virus DNA. Cold Spring Harbor Laboratory, Cold Spring Harbor, pp. 843–910.

Grigoriev, A., 1998. Analyzing genomes with cumulative skew diagrams. Nucleic Acids Res. 26, 2286–2290.

Grigoriev, A., 1998. Genome arithmetic. Science 281, 1923 (summary, full text available at http://www.sciencemag.org/cgi/content/full/281/5385/1923a).

Hammerschmidt, W, Sugden, B., 1988. Identification and characterization of oriLyt, a lytic origin of DNA replication of Epstein-Barr virus. Cell 55, 427–433.

Hamzeh, F.M., Lietman, P.S., Gibson, W., Hayward, G.S., 1990. Identification of the lytic origin of DNA replication in human cytomegalovirus by a novel approach utilizing ganciclovir-induced chain termination. J. Virol. 64, 6184–6195.

Horwitz, M.S., 1976. Bidirectional replication of adenovirus type 2 DNA. J. Virol. 18, 307–315.

Li, J.J., Kelly, T.J., 1984. Simian virus 40 DNA replication in vitro. Proc. Natl. Acad. Sci. USA 81, 6973–6977.

Lobry, J.R., 1995. Properties of a general model of DNA evolution under no-strand-bias conditions. Mol. Biol. Evol. 40, 326–330.

Lobry, J.R., 1996. Asymmetric substitution patterns in the two DNA strands of bacteria. Mol. Biol. Evol. 13, 660–665.

Masse, M.J., Karlin, S., Schachtel, G.A., Mocarski, E.S., 1992. Human cytomegalovirus origin of DNA replication (oriLyt) resides within a highly complex repetitive region. Proc. Natl. Acad. Sci. USA 89, 5246–5250.

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.

McGeoch, D.J., Cunningham, C., McIntyre, G., Dolan, A., 1991. Comparative sequence analysis of the long repeat regions and adjoining parts of the long unique regions in the genomes of herpes simplex viruses types 1 and 2. J. Gen. Virol. 72, 3057–3075.

Meuth, M., 1989. The molecular basis of mutations induced by deoxyribonucleoside triphosphate pool imbalances in mammalian cells. Exp. Cell Res. 181, 305–316.

Mrazek, J., Karlin, S., 1998. Strand compositional asymmetry in bacterial and large viral genomes. Proc. Natl. Acad. Sci. USA 95, 3720–3725.

Nicholas, J., 1996. Determination and analysis of the complete nucleotide sequence of human herpesvirus-7. J. Virol. 70, 5975–5989.

Oller, A.R., Fijalkowska, I.J., Dunn, R.L., Schaaper, R.M., 1992. Transcription-repair coupling determines the strandedness of ultraviolet mutagenesis in *Escherichia coli*. Proc. Natl. Acad. Sci. USA 89, 11036–11040.

Reddy, V.B., Thimmappaya, B., Dhar, R., Subramanian, K.N., Zain, B.S., Pan, J., Ghosh, P.K., Celma, M.L., Weissman, S.M., 1978. The genome of simian virus 40. Science 200, 494–502.

Roberts, R.J., Akusjaervi, G., Alestroem, P., Gelinas, R.E., Gingeras, T.R., Sciaky, D., Pettersson, U., 1986. A consensus sequence for the adenovirus-2 genome. In: Doerfler, W. (Ed.), Adenovirus DNA. Martinus Nijhoff, Boston, pp. 1–51.

Schuurman, R., Sol, C., van der Noordaa, J., 1991. The complete nucleotide sequence of bovine polyomavirus. J. Gen. Virol. 71, 1723–1735.

Seedorf, K., Kraemmer, G., Duerst, M., Suhai, S., Roewekamp, W.G., 1985. Human papillomavirus type 16 DNA. Virology 145, 181–185.

Seif, I., Khoury, G., Dhar, R., 1979. The genome of human papovavirus BKV. Cell 18, 963–977.

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 G gamma and A gamma fetal globin gene region. Cell 26, 345–353.

Spaete, R.R., Frenkel, N., 1985. The herpes simplex virus amplicon: analyses of cis-acting replication functions. Proc. Natl. Acad. Sci. USA 82, 694–698.

Sprengel, J., Schmitz, B., Heuss-Neitzel, D., Zock, C., Doerfler, W., 1994. Nucleotide sequence of human adenovirus type 12 DNA: comparative functional analysis. J. Virol. 68, 379–389.

Stow, N.D., 1982. Localization of an origin of DNA replication within the TRS/IRS repeated region of the herpes simplex virus type 1 genome. EMBO J. 1, 863–867.

Stillman, B.W, 1981. Adenovirus DNA replication in vitro: a protein linked to the 5′ end of the nascent DNA strands. J. Virol. 37, 139–147.

Su, S.S., Lahue, R.S., Au, K.G., Modrich, P.., 1988. Mispair specificity of methyl-directed DNA mismatch correction in vitro. J Biol. Chem. 263, 6829–6835.

Sueoka, N., 1995. Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. J. Mol. Evol. 40, 318–325.

van Loon, N., Dykes, C., Deng, H., Dominguez, G., Nicholas, J., Dewhurst, S., 1997. Identification and analysis of a lytic-phase origin of DNA replication in human herpesvirus 7. J. Virol. 71, 3279–3284.

Wang, T.C., Chen, S.H., 1994. Okazaki DNA fragments contain equal amounts of lagging-strand and leading-strand sequences. Biochem. Biophys. Res. Commun. 198, 844–849.

Wolfe, K., 1991. Mammalian DNA replication: mutation biases and the mutation rate. J. Theor. Biol. 149, 441451.

Yao, X.D., Matecic, M., Elias, P., 1997. Direct repeats of the herpes simplex virus a sequence promote non-conservative homologous recombination that is not dependent on XPF/ERCC4. J. Virol. 71, 6842–6849.

Yates, J., Warren, N., Reisman, D., Sugden, B., 1984. A cis-acting element from the Epstein-Barr viral genome that permits stable replication of recombinant plasmids in latently infected cells. Proc. Natl. Acad. Sci. USA 81, 3806–3810.