J Mol Evol (1997) 44:214–225

JOURNAL OF MOLECULAR EVOLUTION

Retroviral Oligonucleotide Distributions Correlate with Biased Nucleotide Compositions of Retrovirus Sequences, Suggesting a Duplicative Stepwise Molecular Evolution

Ivan Laprevotte,[1] Sophie Brouillet,[2] Christophe Terzian,[2] Alain Hénaut[2]

[1] Laboratoire Rétrovirus et Rétrotransposons des Vertébrés, UPR 43 CNRS, Université Paris 7, Hôpital Saint Louis, 16 rue de la Grange aux Belles, 75475 Paris Cedex 10, France
[2] Centre de Génétique Moléculaire, CNRS, 91198 Gif sur Yvette Cedex, France

Received: 3 January 1996 / Accepted: 27 March 1996

Abstract. A computer-assisted analysis was made of 24 complete nucleotide sequences selected from the vertebrate retroviruses to represent the ten viral groups. The conclusions of this analysis extend and strengthen the previously made hypothesis on the Moloney murine leukemia virus: The evolution of the nucleotide sequence appears to have occurred mainly through at least three overlapping levels of duplication: (1) The distributions of overrepresented (3–6)-mers are consistent with the universal rule of a trend toward TG/CT excess and with the persistence of a certain degree of symmetry between the two strands of DNA. This suggests one or several original tandemly repeated sequences and some inverted duplications. (2) The existence of two general core consensuses at the level of these (3–6)-mers supports the hypothesis of a common evolutionary origin of vertebrate retroviruses. Consensuses more specific to certain sequences are compatible with phylogenetic trees established independently. The consensuses could correspond to intermediary evolutionary stages. (3) Most of the (3–6)-mers with a significantly higher than average frequency appear to be internally repeated (with monomeric or oligomeric internal iterations) and seem to be at least partly the cause of the bias observed by other researchers at the level of retroviral nucleotide composition. They suggest a third evolutionary stage by slippage-like stepwise local duplications.

Key words: Computer-assisted analysis — Retrovirus nucleotide sequence — Stepwise duplicative molecular evolution — Core consensus — TG/CT excess — Symmetry — Tandem repeat — Cryptic simplicity — Slippage — Nucleotide composition

Introduction

Repetitive nucleotide sequences are common in eukaryotic genomes. Repeat units vary in length from one to several thousand base pairs (Wooster et al. 1994). They show perfect repeats or a ‘‘cryptic simplicity,’’ both often suggesting an evolution by scrambled slippage-like events associated with point-by-point substitutions of one base by another in the DNA molecule (reviewed in Tautz et al. 1986). In addition, other mechanisms like unequal cross-over, transposition, or gene conversion (reviewed in Dover 1982; Golding and Glickman 1985) can account for the homogeneity of eukaryotic sequences and can also be at work in eukaryotic viruses. In fact, the existence of a bias of oligopurine sequences in eukaryotic viruses has already been demonstrated (Beasty and Behe 1988).

A computer-aided analysis of the *gag* region of the feline leukemia virus (FeLV B) and that of the Moloney strain of the murine leukemia virus (Mo-MuLV) has already been carried out (Laprevotte et al. 1984; Laprevotte 1989). This work was complemented with the analysis of the Mo-MuLV complete nucleotide sequence

Correspondence to: I. Laprevotte

Table 1.

Subfamilies	Groups	Viruses	EMBL accession No.
Oncovirinae	E-type	Bovine leukemia	K02120
		Human T-cell leukemia type I	D13784
		Human T-cell leukemia type II	M10060
	Mammalian C-type	Feline leukemia (subgroup A)	M18247
		AKV murine leukemia	J01998
		Moloney murine leukemia	J02255
	Avian-leukosis sarcoma	Rous sarcoma (Prague strain subgroup (C)	J02342
	B-type	Mouse mammary tumor	M15122
	D-type	Simian Mason-Pfizer	M12349
		Simian SRV-1	M11841
		Simian SRV-2	M16605
Lentivirinae	Human immunodeficiency type I	Aids-associated	K02007
		HIV I	K03455
		African	K03454
	Human immunodeficiency type II	Isolate SBLISY	J04498
		ROD isolate	M15390
		HIV2BEN	M30502
	Simian immunodeficiency	African green monkey	M29975
		Macaque	M33262
		Sooty mangabey	X14307
	Nonprimate lentiviruses	Equine infectious anemia	M16575
		Ovine lentivirus	M31646
		Visna lentivirus	M10608
Spumavirinae		Simian foamy type I	X54482

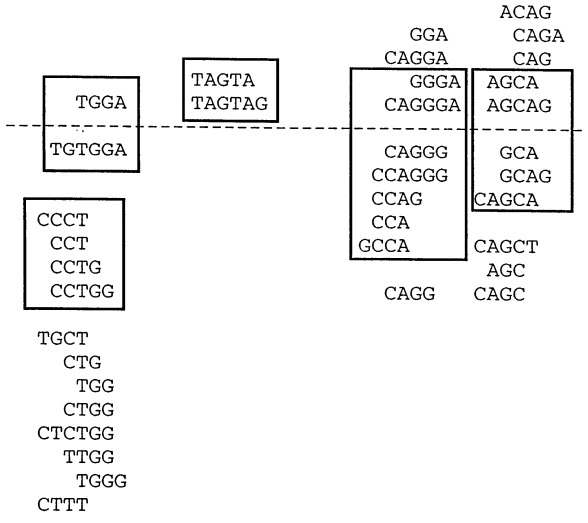

Fig. 1. Type I (3–6)-mers in K02007 (HIV I, AIDS-associated), the same sequence as in Fig. 2. The corresponding type I subset (see text) covers 47% of the sequence. The *n*-mers above the *dotted line* are also displayed at the upper part of Fig. 2. Boxed *n*-mers are displayed as a consensus in Fig. 3 (see text).

and the human spumaretrovirus (HSRV) (Laprevotte 1992). Quite a number of oligomers are overrepresented and internally repeated (with monomeric or oligomeric internal iterations), so that it can be assumed that viral genomes are subject to the same evolutionary mechanisms that are now known to be often operating in eukaryotic genomes. In addition, the Mo-MuLV nucleotide sequence follows the universal rule of a trend toward TG/CT excess (Ohno and Yomo 1990), and its overrepresented oligomers share a core consensus regularly scattered throughout the whole sequence. This sequence has been exhaustively analyzed and exhibits three overlapping levels of oligomeric repetitions. It could be that local repetitions have stemmed from (an) originally tandemly repeated oligonucleotide(s) (Ohno 1988). Moreover, the core consensus could correspond to an intermediary evolutionary stage (Southern 1972; Ohno 1988).

Recently (Bronson and Anderson 1994), all complete retroviral sequences in the EMBL database were examined with the goal of assessing possible relationships between the biased nucleotide composition of retroviral genomes, the amino acid composition of retroviral proteins, and evolutionary strategies used by these viruses; 24 selected sequences representing the ten groups in the three subfamilies of retroviruses were examined in detail. We have used these very 24 sequences to verify that the conclusions of the previous studies of FeLV B, Mo-MuLV, and HSRV hold for all the subfamilies of retroviruses. Moreover, we have found a correlation of the biased monomeric and dimeric frequencies of retroviruses with numerous repeated nucleotide motifs that appear to have evolved through local repetition. Finally, an overall similarity was found for the overrepeated (3–6)-base-long oligonucleotides in the 24 sequences; there are two common core consensuses, and other consensuses particular to some sequences or groups of sequences.

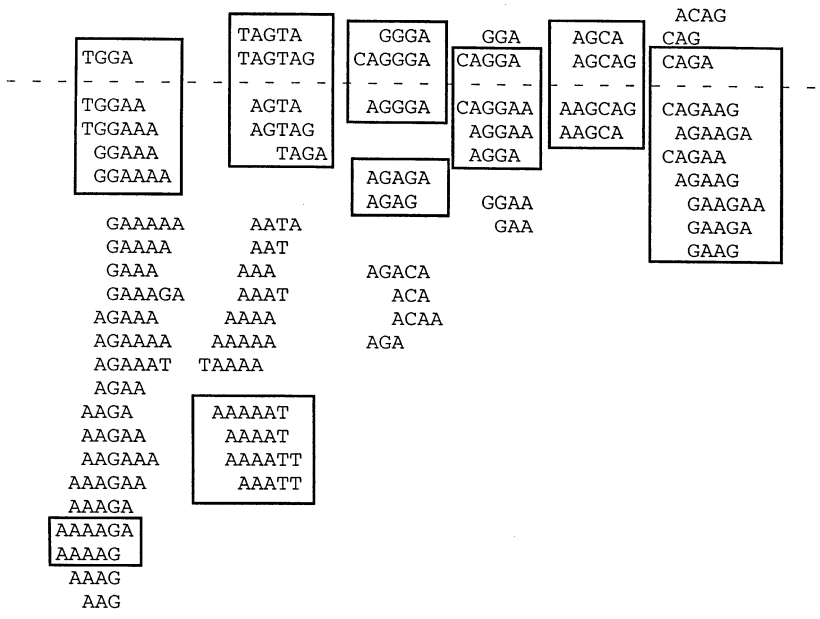

Fig. 2. The type II (3–6)-mers from K02007 (HIV I, AIDS-associated). The *n*-mers that are above the *dotted line* are also overrepresented as calculated for Figs. 1, 3, and 4; they are also displayed in Fig. 1. *Boxed n*-mers can be displayed as a consensus (see text). The sequence is 9,737 nucleotides in length with 3,445 A, 1,738 C, 2,377 G, and 2,177 T. The dimeric composition is AA: 1,143; AC: 549; AG: 1,029; AT: 724; CA: 789; CC: 388; CG: 88; CT: 473; GA: 809; GC: 462; GG: 671; GT: 434; TA: 704; TC: 338; TG: 589; TT: 546. The type II subset (see text) covers 57% of the sequence.

Materials and Methods

The names, the abbreviations, and EMBL accession numbers of the 24 retroviral sequences analyzed in this work, as well as the names of the subfamilies and groups of retroviruses, are listed in Table 1.

Overrepresented (3–6)-mers that we will call here ''nucleotide motifs of type I'' are selected as follows: Each *n*-nucleotide-long possible motif is observed at *x* (nonoverlapping) positions in the sequence. The probability $P(\geqslant x)$ for each motif to be found in at least these *x* positions in a shuffled sequence (same length and same base composition) is given by the binomial law:

$$P(\geqslant x) = 1 - \left[\sum_{i=0}^{x-1} \frac{N!}{i!(N-i)!} P^i (1-P)^{(N-i)} \right]$$

where *N* is the total number of positions in the sequence [(length of the sequence) − *n* + 1] and *P* is the probability that a motif is found in any given position, that is, the product of observed compositions (in the sequence) of nucleotides found in the *n*-mer. In a shuffled sequence, for a given value of $P(\geqslant x)$, the average number *m* of *n*-long motifs is $P(\geqslant x) * 4^n$, where 4^n is the total number of *n*-long motifs. When $m = 0.01$, *n*-mers are considered overrepresented; this corresponds to a very significant number of occurrences. Indeed, Poisson's law with $m = 0.01$ shows that less than 0.01 shuffled sequences ($1 - e^{-0.01}$) contain at least one such overrepresented *n*-mer.

Those (3–6)-mers not necessarily overrepresented but occurring at at least an averaged significant threshold number of positions are called here ''nucleotide motifs of type II''; they are selected as follows: as in (Laprevotte 1992), an average threshold value corresponding to a number of nonoverlapping positions of (3–6)-mers was calculated for each retroviral sequence and each length of oligomers. Calculations are identical to those of type I motifs, except that *P*, the probability for any *n*-mer to be observed in a given position, is $P = p^n$; the probability *p*, of finding any nucleotide at any position is a weighed average, $p_a^2 + p_c^2 + p_g^2 + p_t^2$, i.e., the sum of squared observed proportions of A, C, G, and T in the sequence (Day and Blake 1982). For any *n*-mer, the threshold value is chosen so that $m < 0.01$ ($m = P(\geqslant x) * 4^n$).

Results

For each sequence, the (3–6)-mers of type I can be grouped together in consensuses. In the set of 24 sequences, these motifs share general features and two core consensuses. Consensuses specific to certain retrovirus groups are compatible with the phylogenetic tree

For each of the 24 sequences, (3–6)-mers of type I (that are significantly overrepresented) were listed using the binomial law (see Fig. 1 and Laprevotte 1992). A threshold value for the number of (nonoverlapping) positions in the sequence is calculated for each observed motif, using the probability *P* that this motif occurs at any position in a random sequence with the same length and the same base composition: *P* (Day and Blake 1982) is the product of relative observed frequencies (in the whole sequence) of nucleotides contained in the motif. Reported *n*-mers of type I correspond to a very significant number of positions, such that less than 0.01 random sequences would have at least one oligonucleotide of the same length with at least the threshold number of positions (Materials and Methods).

Figure 1 shows the result obtained for a HIV I virus (K02007). Oligonucleotides above the dotted line are common in Figs. 1 and 2. In order to present the results clearly, motifs perfectly included in a consensus are boxed; any of these motifs is aligned with at least another one with a minimum of three nucleotides overlapping, priority being given to longest overlaps; a motif must not be found in another consensus unless the two consensuses can be merged.

Results are shown in Fig. 3 for the 24 sequences according to such a principle, where groups of motifs were replaced with consensuses between square brackets (the number next to a consensus represents the number of

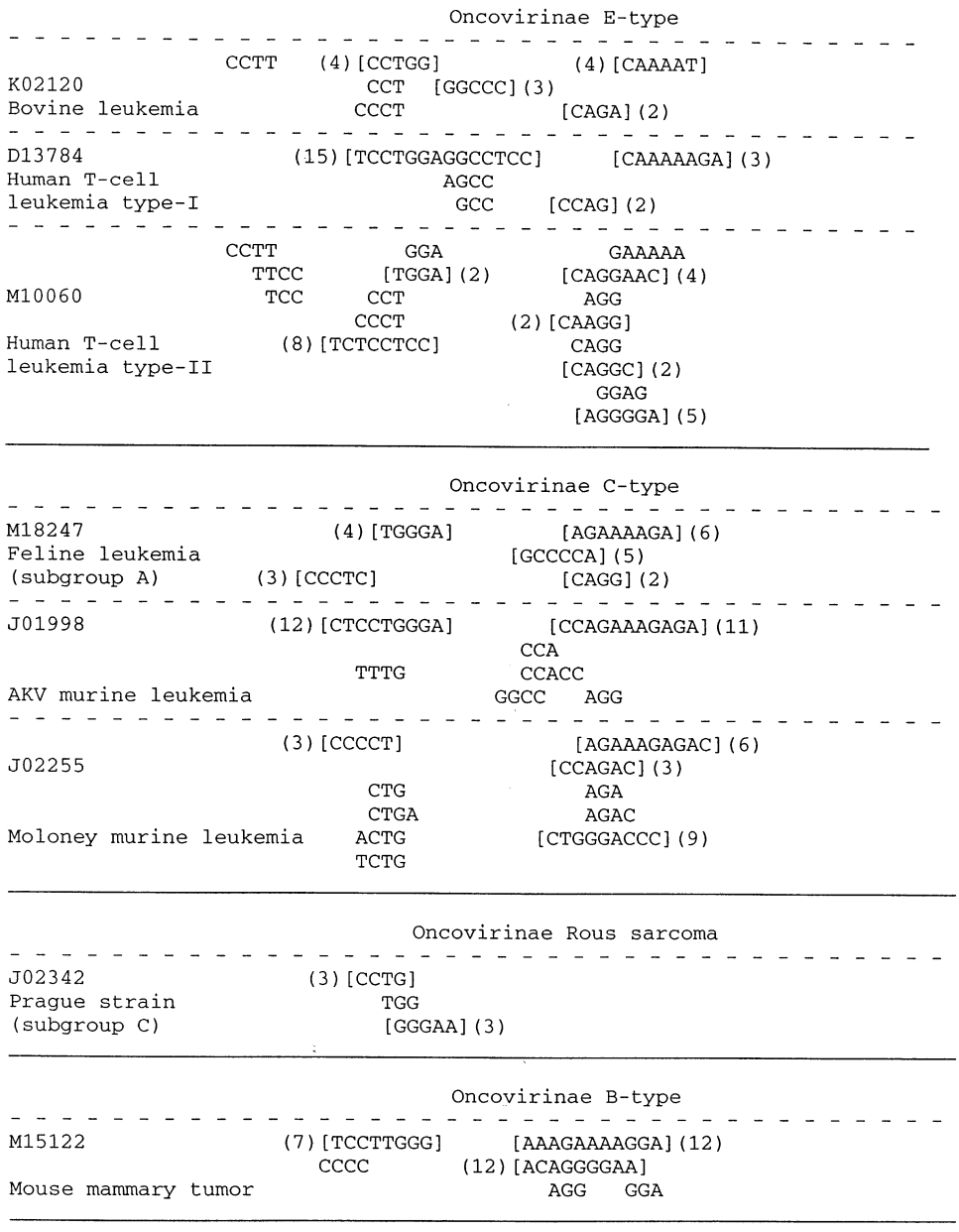

Fig. 3. Type I (3–6)-mers in the 24 sequences referred to as in Table 1. The sequences belonging to the same group are only separated by a *dotted line.* The computing method for selecting overrepresented (type I) *n*-mers is detailed in Materials and Methods. The *bracketed se quences* are *n*-mers included in a consensus in order to simplify the presentation of the data (see text); the *number near a bracketed sequence* indicates the number of *n*-mers included in the consensus.

n-mers involved in this consensus). General characteristics can be drawn from these results. In the set of type I *n*-mers, alternating purine and pyrimidine stretches can be emphasized that can be displayed clearly in two subsets, on both sides of CTG (or TTG) and of CAG, respectively. Two general core consensuses show up, CCTGG and CAGR (R: purine); both of these core consensuses are found in most groups of retroviruses and at least one of them is found in each group. Moreover, within each group, retroviral sequences can be grouped in pairs or triplets according to common *n*-mers or to consensuses, further emphasizing the existence of particular relationships. For instance, for oncovirinae of type E, CCTT and CCCT link K02120 and M10060 together; GGCCY (Y: pyrimidine) and CAAAA link K02120 and D13784 together; TCCT and TGGA link D13784 and M10060 together; AAAA is found in those three sequences. AGAAA, AAAGA, and TGGGA link all three oncovirinae of type C; GAGAC, CCAGAC, and GGACCC found in J02255, i.e., Mo-MuLV, are not found in M18347, i.e., FeLV of type A; however, it was established earlier that GACC and ACCC are some of the frequent tetramers in *gag* regions of FeLV B and Mo-MuLV (Laprevotte 1989). The same type of correlation can be drawn from examining motifs of type I in the rest of Fig. 3 (isolated or grouped by consensus)

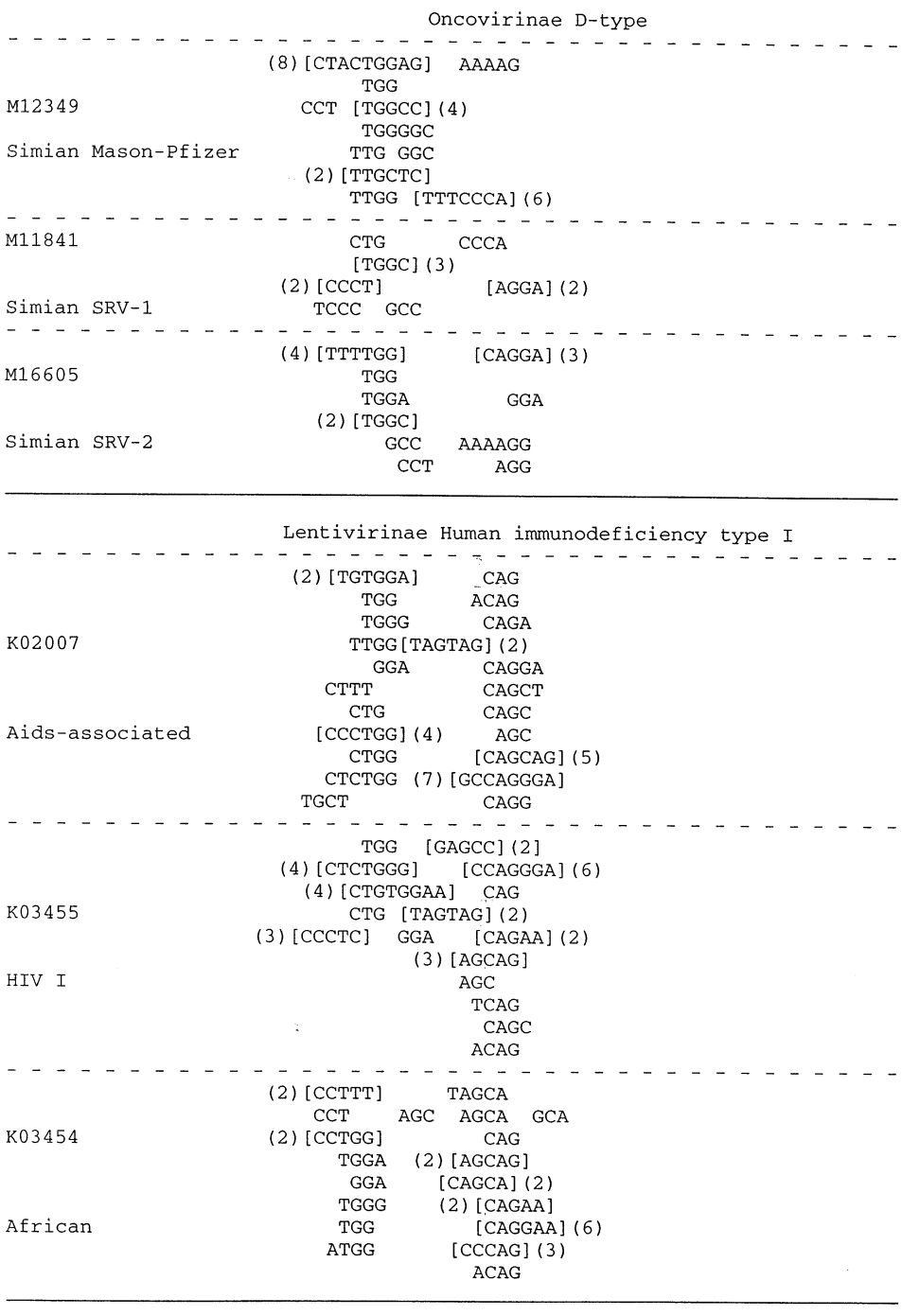

Fig. 3. Continued.

within each of the groups characterized here by several retroviral sequences.

A question remains: On the one hand, retrovirus LTRs (long terminal repeats) have multiplied motifs that may correspond to experimentally determined regulatory elements (Seto et al. 1989); these motifs could show up in the analysis especially when the two LTRs are included in the nucleotide file of the EMBL database (see Table 1). On the other hand, some overrepeated oligomers could include trimers representing codons for favorite amino acids within the open reading frames. Although the type I motifs cover a great part of the whole sequence (around 40%, see below), they could have distinct patterns of distribution in the LTRs that are essentially noncoding and in the rest of the sequences that are essentially coding. To address this issue, the type I (3–6)-mers have been listed in the parts of the sequences (referred to as deleted sequences) that are bounded by the LTR(s) (LTRs excluded) for the six HIV (1 and 2) and for three additional retroviruses (J01998, M33262 and X54482) representing oncovirinae, lentivirinae, and spumavirinae, respectively (data not shown). In the same way, the type

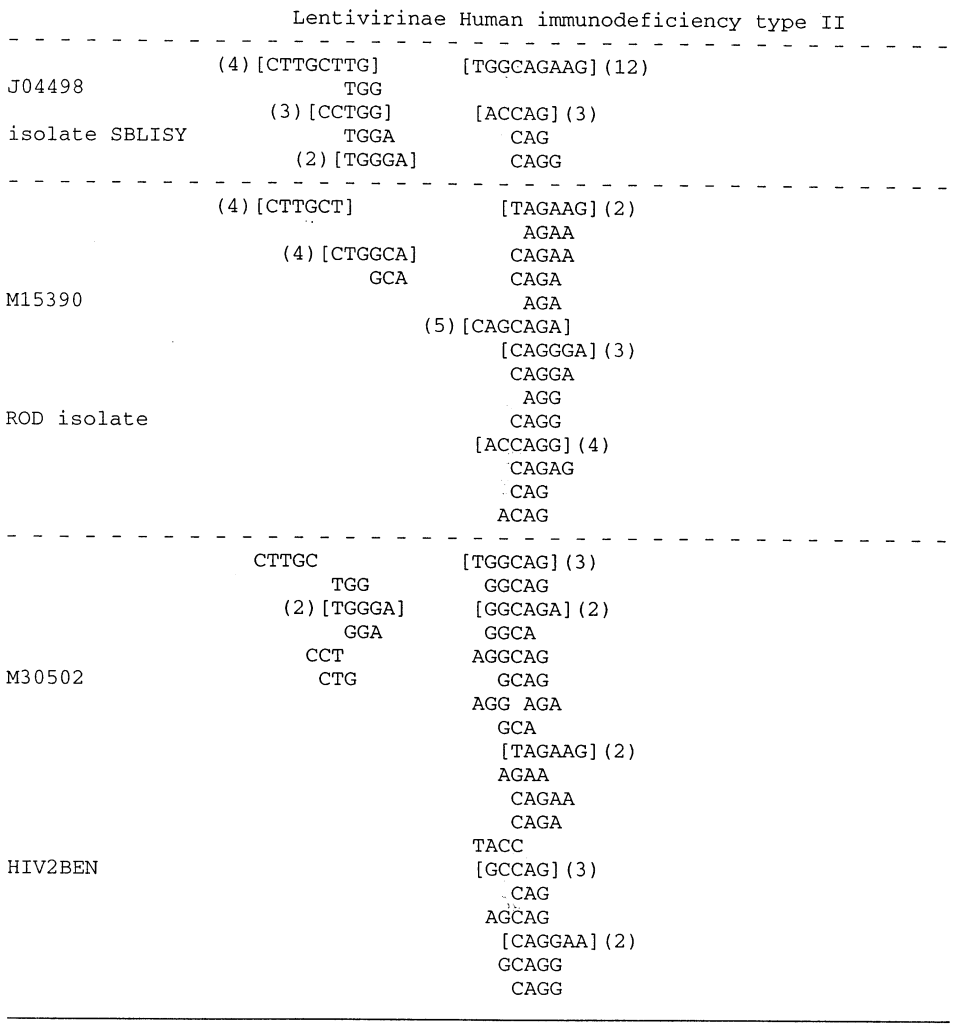

Fig. 3. Continued.

I trimers have been listed in the 15 additional deleted sequences. Some motifs disappear and others appear, but basically, all of the general features described above hold: the two core consensuses and alternating purine and pyrimidine stretches clearly displayed on both sides of CTG (or TTG) and of CAG, respectively. In the whole of the 24 sequences (complete or deleted), 176 type I trimers are overrepresented: Eight have a nearly equal relative number of occurrences in the LTR(s) and in the corresponding deleted sequence; 86 are not found or are relatively less represented in the corresponding deleted sequence; inversely, 82 are not found in the complete sequence or are relatively less represented in the corresponding LTR(s). Of the 25 motifs corresponding to these 176 type I trimers, the eight most represented in the 24 complete sequences are TGG (24 sequences), CCT (21), GGA (18), CAG (17), CTG (12), CCA (11), AGG (11), and GCA (eight). In the 24 deleted sequences, they are TGG (24 sequences), CCT (19), GGA (19), CAG (16), CCA (11), AGG (nine), GCA (nine), and CTG (seven). The sequences of HIV (1 and 2) are the most biased in favor of the type I trimers in the LTRs. This corresponds to quite a lot of type I motifs of the ''CTG'' group; actually, CTG is often found in the overlapping multiplicated motifs described in HIV 1 LTRs (Seto et al. 1989). However, as seen above, the general characteristics of the type I motifs also hold for the HIV (1 and 2) deleted sequences. Inversely, the two sequences that are the most biased in favor of the type I trimers in the deleted sequences are that of the Mo-MuLV and of the FeLV A. This raises the question (see above) of codons for favorite amino acids. However, a previous analysis of the *gag* regions of FeLV B and Mo-MulV (Laprevotte et al. 1984; Laprevotte 1989), complemented by the analysis of the Mo-MuLV complete nucleotide sequence (Laprevotte 1992), has shown three overlapping levels of duplications with short-range duplications sharing a core consensus (or its inverse) scattered throughout the sequences without any stringent correlation with codon usage and coding reading frames (see discussion below). Moreover, in Mo-MuLV complete sequence, it has been verified that the ''overrepeated'' 2–7-mers held even if only one of the 75- and 68-base-long direct repeats was taken into account in the sequence. Inversely, the ''over-

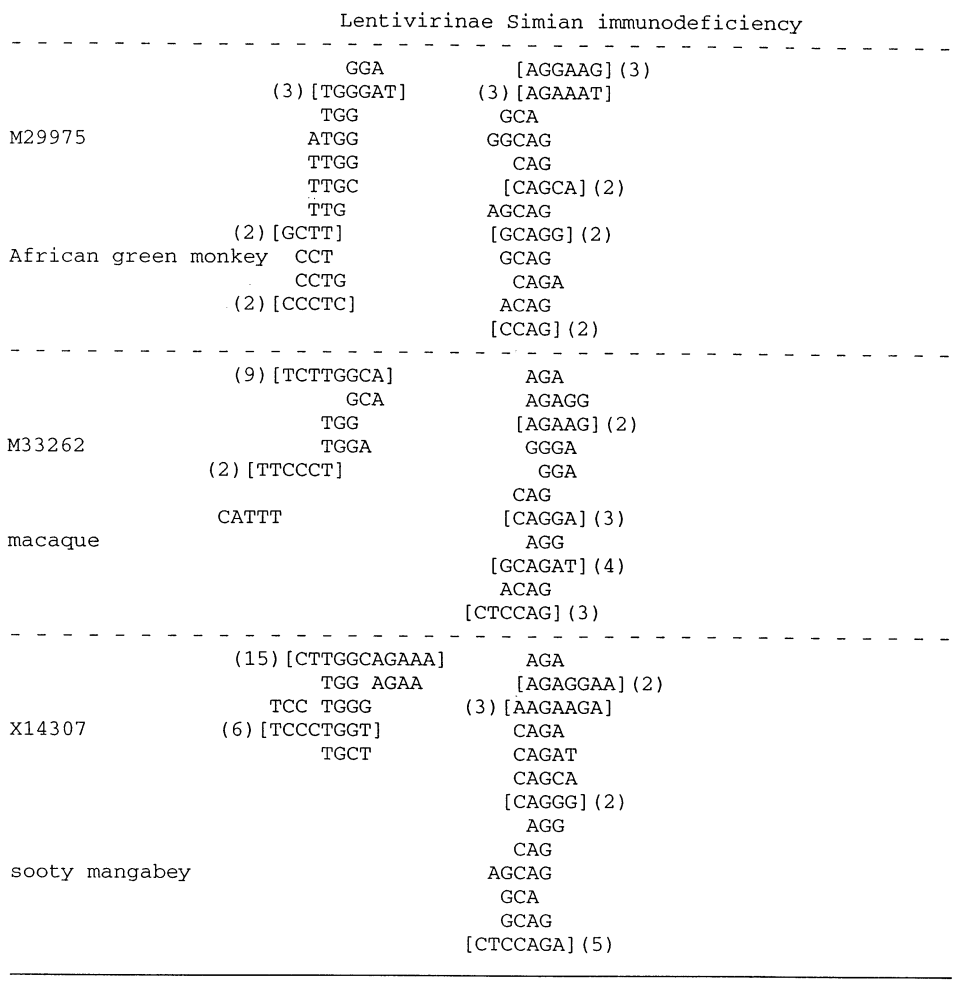

Fig. 3. Continued.

repeated'' 11-mers (or longer) were found to be included in these repeats.

A question arises: One of the subsets of type I *n*-mers can be displayed on both sides of CTG; CT and TG are known to tend to be generally overrepresented in nucleotide sequences (Ohno 1988). The other subset can be displayed on both sides of CAG; CA and AG are complementary to CT and TG. It is possible that the two general core consensuses CCTGG and CAGR that show up are not specific to vertebrate retroviruses but found more widely in other species. A study (Terzian et al. data to be published) of retrotransposons in drosophila, yeast, etc., shows consensuses that are often different than those described here and shows that the retroviruses segregate in a distinct subset as concerns their oligomeric distributions. For example, Fig. 4 shows several consensuses distinct from those described for vertebrate retroviruses in the case of the retrotransposon Del of *Lilium henryi,* even though the shown *n*-mers contain a lot of CT and AG as well as three TG and CA.

In each of the 24 retroviral sequences, both the nucleotide and dinucleotide composition of the whole sequence is compared to that in the set of type I (3–6)-mers as follows: a computer program locates nucleotides in the sequence occurring in the type I motifs (overlapping positions included), thus defining a type I nucleotide subset. The linear correlation coefficient calculated from the pairwise comparisons between the 4 × 24 nucleotide numbers in the 24 type I subsets taken together, on the one hand, and in the 24 complete sequences, on the other hand, is 0.53 (in contrast with 0.93 for the type II subsets defined in the same way, see below). Table 2 shows also that type I (3–6)-mers in a sequence are not merely predictable from the dinucleotide composition (such a composition is considered as a ''genomic signature,'' Karlin and Burge 1995). Indeed, for each retrovirus, the correlation between the dinucleotide distribution in the subset of type I (3–6)-mers and that of the whole sequence is null or looser (from −0.0365 to 0.89) than the correlation evaluated for type II *n*-mers (see below). The correlation is null or weak (−0.125 to 0.502) for type I only oligomers.

The (3–6)-mers of type II correlate with the base and dimer compositions of the sequences, but are not merely predictable from the nucleotide composition

As previously explained (Laprevotte 1992), the type II motifs defined above (occurring at at least an averaged

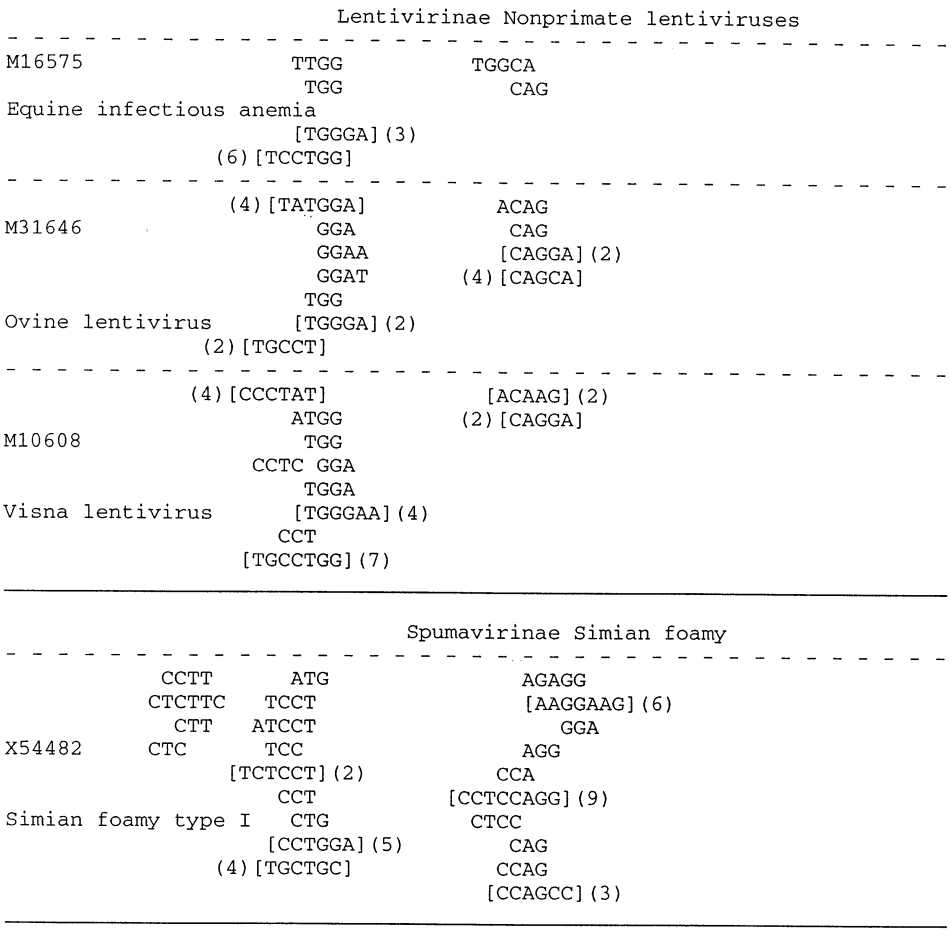

Fig. 3. Continued.

significant threshold number of positions) are computed in order to take into account putative overwhelming duplications of nucleotides or of oligonucleotides which could have biased the base composition of the sequence and would not be detectable by the zero-order Markov chain used to select the type I motifs. As exemplified in Fig. 2, the motifs of type II appear to be internally repeated (with monomeric or oligomeric internal iterations) and to include a great proportion of the most frequent nucleotide in the whole sequence. This suggests that the biased nucleotide frequencies of retroviruses could be the consequence of short-range repeats of nucleotide motifs of one to a few bases in length; e.g., GAAAAA and AAAAAT look like A duplications, A being the most frequent letter in the HIV 1 sequence (K02007); e.g., TAGTAG would be a trimer duplication, e.g., AGAGA a dimer duplication, etc. Analogous patterns are found in the other 23 sequences, A or C being the most frequently observed letter (Bronson and Anderson 1994; Van Hemert and Berkhout 1995); for instance, D13784 (HTLV 1) which is C-rich (0.35) has C-rich motifs of type II.

For each sequence, the WINDOW program from the GCG sequence analysis software package (Genetic Computer Group 1992) has been run in order to study the

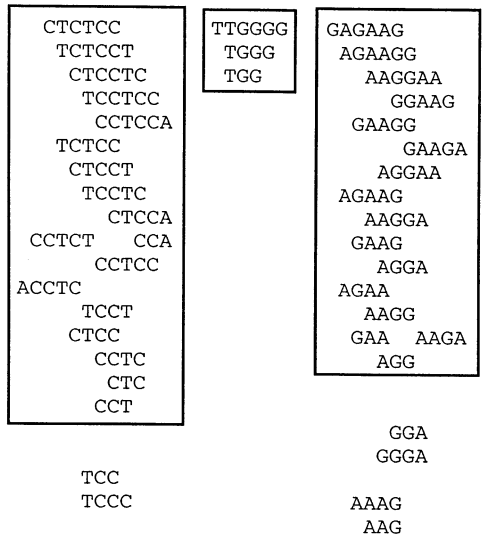

Fig. 4. Type I (3–6)-mers in Del sequence: gypsy-like retrotransposon of the plant *Lilium henryi*. The sequence is 9,345 bases long with 2,927 A, 1,765 C, 1581 G, and 3,072 T. The *n*-mers as selected as for Fig. 1 and Fig. 3. *Boxed n*-mers can be displayed as a consensus (see text and Fig. 1).

Table 2. Linear correlation coefficient, for each sequence, between the occurrences of the dimers (overlapping positions included) in the whole of, or parts of each type I and/or II subset, and the corresponding occurrences in the whole of the sequence

Sequence	Words type I not II	Words type I	Words type I and II	Words type II	Words type II not I
D13784	0.001	0.789	0.865	0.959	0.960
J01998	0.312	0.883	0.936	0.919	0.783
J02255	0.214	0.890	0.830	0.905	0.870
J02342	0.144	0.806	0.802	0.891	0.814
J04498	0.193	0.734	0.739	0.917	0.916
K02007	0.223	0.518	0.683	0.920	0.900
K02120	0.045	0.780	0.902	0.946	0.941
K03454	0.352	0.603	0.695	0.928	0.910
K03455	0.220	0.562	0.744	0.923	0.904
M10060	−0.125	0.802	0.836	0.971	0.976
M10608	0.372	0.601	0.643	0.944	0.941
M11841	−0.088	−0.037	0.280	0.836	0.810
M12349	−0.019	0.180	0.470	0.843	0.760
M15122	0.313	0.638	0.600	0.827	0.793
M15390	0.241	0.774	0.772	0.916	0.913
M16575	0.167	0.167	0.517	0.903	0.903
M16605	−0.043	0.010	0.473	0.791	0.792
M18247	0.124	0.844	0.846	0.877	0.857
M29975	0.225	0.781	0.797	0.911	0.908
M30502	0.170	0.798	0.781	0.919	0.912
M31646	0.502	0.633	0.536	0.942	0.944
M33262	0.234	0.741	0.696	0.887	0.880
X14307	0.172	0.703	0.706	0.909	0.902
X54482	0.062	0.391	0.537	0.867	0.871

pattern of distribution of the most frequent nucleotide in the whole sequence (and of the corresponding dimer) in successive windows of 100 nucleotides (with a shift of one or three bases) covering the whole range of the 24 sequences (data not shown). As compared with shuffled sequences, clear clusters in specific regions were not found. Such clusters could have been expected in case of biased amino acid compositions, in particular in the *gag* region (Laprevotte et al. 1984). Together with the fact that type II motifs cover a great part of the whole sequence (around 50%, see below), these results suggest that the internal repetitions of type II motifs are short-range iterations scattered throughout the sequence; they fit the previous conclusion (Laprevotte 1992) that the ''overrepeated'' motifs (of both type I and II) in the Mo-MuLV nucleotide sequence are not merely the consequence of the codon usage. The linear correlation coefficient calculated from the pairwise comparisons between the 4 × 24 nucleotide numbers in the 24 type II subsets taken together (defined in the same way as type I subsets, see above), on the one hand, and in the 24 complete sequences, on the other hand, is 0.93 (in contrast with 0.53 for the type I subsets). Moreover, for each sequence, as shown by Table 2, a good correlation (0.791 to 0.971) holds between the dimer distribution in the type II subset and in the whole sequence. This good correlation holds for type II–only oligomers (0.760 to 0.976) but is weaker (0.280 to 0.936) when oligomers are common to type I and II. However, it could be that the (3–6)-mers of type II are merely predictable from the nucleotide compositions of the sequences. Table 3 shows the number of occurrences, for each of the 24 subsets, of the most frequent nucleotide in the whole sequence. This number is always above the mean value obtained from 100 shuffled sequences. For 19 sequences out of 24, it is above the maximum value obtained from these 100 shuffled sequences. For five sequences, the number is close below the maximum value. For these five sequences, the same calculation was performed with the total number of bases in the subset. It shows that the number is greater in four of the five sequences than the maximum number obtained from the 100 shuffled sequences; for the fifth sequence (accession number M10060) the value is close to the maximum (4,533 vs 4,555 with a mean of 3,834.9). This demonstrates that the (3–6)-mers of the 24 type II subsets are not merely predictable from the nucleotide compositions of the sequences, and could be, as suggested above, at least partly the consequence of the evolution of short nucleotide motifs by local repetition.

Discussion

Conclusions drawn from the Mo-MuLV analysis (Laprevotte 1992) can be extended in light of the data analysis presented here.

1. The type I (3–6)-mers can be displayed in two subsets on both sides of CTG (or TTG) and of CAG, respectively. Two general complementary consensuses,

Table 3. The most represented base in each of the 24 retroviral sequences: number of occurrences in the type II subset, compared to 100 corresponding shuffled sequences[a]

	Composition						
Sequence	A	C	G	T	Value	Max	Mean
D13784	1,983	*2,932*	1,534	1,951	2,321	2,405	2,194.7
J01998	2,135	*2,416*	2,056	1,767	1,884	1,235	578.8
J02255	2,143	*2,395*	2,025	1,769	1,804	1,213	482.5
J02342	2,300	2,421	*2,761*	2,143	1,625	1,228	505.2
J04498	*3,291*	1,940	2,396	2,009	2,800	2,773	2,420.1
K02007	*3,445*	1,738	2,377	2,177	2,996	2,817	2,619.0
K02120	1,898	*2,879*	1,864	2,073	2,242	2,345	2,023.2
K03454	*3,333*	1,632	2,179	2,032	2,955	2,745	2,586.2
K03455	*3,411*	1,773	2,370	2,164	2,998	2,873	2,590.2
M10060	2,171	*3,187*	1,629	1,965	2,567	2,576	2,418.8
M10608	*3,416*	1,420	2,399	1,968	2,982	2,899	2,686.2
M11841	*2,503*	1,944	1,528	2,198	1,870	1,710	1,276.5
M12349	*2,544*	2,059	1,620	2,334	1,856	1,593	1,184.9
M15122	*3,020*	2,139	2,298	2,668	2,242	1,924	1,369.4
M15390	*3,314*	1,972	2,401	1,984	2,930	2,794	2,442.6
M16575	*2,984*	1,378	1,860	2,185	2,488	2,500	2,242.2
M16605	*2,377*	1,818	1,521	2,043	1,571	1,553	1,122.7
M18247	*2,334*	2,332	1,950	1,824	1,499	982	487.5
	2,334	*2,332*	1,950	1,824	1,589	1,129	470.5
M29975	*3,321*	1,905	2,450	2,118	2,805	2,866	2,416.5
M30502	*3,506*	2,132	2,598	2,123	3,049	2,959	2,576.9
M31646	*3,520*	1,415	2,343	1,978	3,161	2,963	2,812.1
M33262	*3,540*	1,988	2,607	2,400	2,991	2,880	2,590.3
X14307	*3,481*	1,891	2,568	2,301	2,934	2,863	2,555.5
X54482	*4,195*	2,480	2,601	3,696	3,618	3,227	2,855.5

[a] *Sequence:* EMBL accession number as shown in Table 1; *Composition:* number of each base in the whole sequence (maximum value is italicized; for M18247, two neighboring values are italicized); *Value:* number of positions in the subset of type II (3–6)-mers for the base most represented in the whole sequence; *Max:* maximum value for the corresponding number observed in 100 corresponding shuffled sequences; *Mean:* mean values for the corresponding number observed in the 100 shuffled sequences

CCTGG and CAGR, show up that contain CT and TG or the complementary dimers CA and AG. This result is consistent with the universal rule of a trend towards TG/CT excess which was proposed as a generative principle of sequences (Ohno and Yomo 1990) and with the persistence of a certain degree of symmetry between the two strands of proviral DNA. Actually, as previously discussed (Laprevotte 1992), complementary overrepresented oligomers should be expected to be located at the complementary strands of the proviral DNA. At some evolutionary stage(s), the existence of one (several) original tandemly repeated sequence(s) and some reversed duplications can be assumed (Nussinov 1982) in an ancestral retrovirus and/or a possible original eukaryotic sequence (Temin 1980). One can also imagine an evolutionary trend toward a duplicative expansion of only one of these two complementary (''CTG'' or ''CAG'') oligomeric subsets, independently at the two strands of (a) double-stranded DNA intermediate(s); this also could account for the just-mentioned degree of symmetry.

2. Features common to type I (3–6)-mers, in particular the two general core consensuses CCTGG and CAGR, support the hypothesis of a common evolutionary origin of vertebrate retroviruses (distinct consensus cores were found in retrotransposons of other plant or animal species). Secondary consensuses more specific to certain sequences are compatible with phylogenetic trees established independently (Xiong and Eickbush 1990; Bronson and Anderson 1994). Both types of consensuses could correspond to intermediary evolutionary stages (Southern 1972; Ohno 1988). Short tandem repeats can give rise to longer oligomeric repeats as deducted by Southern in 1972. Indeed, point mutations and/or short fragment duplications can affect shorter repeat units, leading to longer and more imperfect repeats.

   On average, the subset of type I motifs covers 43% of the retroviral sequence (from 25% to 55%), suggesting that intermediary stages of duplication could have covered most of the sequences. Correlation studies show that type I motifs are not merely predictable from the sequence dinucleotide composition (considered as a ''genomic signature,'' Karlin and Burge 1995).

3. Type II (3–6)-mers appear to be internally repeated whether one or more nucleotide(s) is repeated. They seem to be at least partly the cause of the bias observed at the level of nucleotide compositions of retroviruses (Bronson and Anderson 1994). On average,

the subset of type II motifs (overlapping that of type I motifs) covers 52% of the retroviral sequence (from 32% to 61%), suggesting that sequences evolved mainly through local duplications such as a slippage-like mechanisms (Tautz et al. 1986). Such an evolutionary process would be superimposed to other duplication stages proposed earlier on.

The two methods for computing the most repeated (3–6)-mers led to select two distinct, albeit overlapping, sets of motifs. The type I motifs are truly overrepresented and show general features for all of the retroviruses; they appear to be the consequence of ancestral evolutionary events. The type II motifs are correlated with the bias observed at the level of nucleotide compositions of the sequences; such a bias makes it possible to distinguish A-rich, C-rich, or G-rich retrovirus sequences and could account for more recent short-range duplicative (and possibly mutational, see below) events.

These results should spur further comparison with retrotransposons and with eukaryotic sequences in which these retrotransposons are integrated and from which they could originally come from. Such models of molecular evolution complement the conventional determination of phylogenetic trees.

Even though the model presented here gives a good account of the repetitive aspect of retroviral nucleotide sequences, another evolutionary process(es) may be considered, such as gene conversion (leading to homogeneity throughout DNA sequences, see discussion in Laprevotte 1989) and a converging evolution toward repeated motifs serving useful functions such as RNA packaging or RNA/protein interaction. Moreover, duplication mechanisms could favor particular nucleotide motifs. In the human genome, sequences like (dC-dA)*n*.(dG-dT)*n* vary in length (Weber 1990) and an increase of repeated CAG is correlated to some diseases (Han et al. 1994). Some types of human cancer involve unstable short tandem repeats (microsatellites): the repeat units are GT, CAG, TCTA, and TATC (Wooster et al. 1994). CT, TG, CA, and CAG are included in the general consensuses described above. Besides, AGCTG frequently observed on the immunoglobulin gene Cμ 5′ noncoding sequence (Ohno 1981) contains CTG. Assuming the cellular origin of retroviruses holds, these intriguing coincidences could be understood through a better knowledge of eukaryotic genomes.

The high frequency of A in genomes of lentiviruses (Table 3, Fig. 2) has been extensively studied though the underlying molecular mechanism is still unknown (Berkhout and van Hemert 1994). A recent study (Martinez et al. 1995) shows experimentally that with a high concentration of dNTP, a hypermutation from G to A takes place in the course of retrotranscription. Such a hypermutation could contribute to raising the A frequency. Moreover, in the same study, successions of A found in two clones suggest series of duplications by multiple strand transfer. In any case, several mechanisms acting together could increase the frequency of a nucleotide. Indeed, it seems difficult to explain the presence of repeats in oligomers of type II by a simple hypermutation from G to A. This would involve the existence of preferential regions for mutations, which should lead to postulating preexisting polypurines stretches (Beasty and Behe 1988). Actually, in lentiviruses, the high frequency of A goes along with a low frequency of C and a weaker effect on G and T frequencies (Table 3, Martinez et al. 1995; van Hemert and Berkhout 1995). Since the percentage of G is weakly or not decreased, there should have been a high percentage of G to begin with. Finally, comparisons with random sequences show that A-rich short-length repeats cannot only be due to the high frequency of A in lentiviral sequences.

Searching for possible secondary structures could help to partially assess the degree of symmetry between two complementary strands. More light can be shed on these problems as eukaryotic genome sequencing projects are progressing and retrotransposon genomes are being studied.

References

Beasty AM, Behe MJ (1988) An oligopurine sequence bias occurs in eukaryotic viruses. Nucleic Acids Res 16:1517–1528

Berkhout B, VanHemert FJ (1994) The unusual nucleotide content of the HIV RNA genome results in a biased amino acid composition of HIV proteins. Nucleic Acids Res 22:1705–1711

Bronson EC, Anderson JN (1994) Nucleotide composition as a driving force in the evolution of retroviruses. J Mol Evol 38:506–532

Day GR, Blake RD (1982) Statistical significance of Symmetrical and repetitive segments in DNA. Nucleic Acids Res 10:8323–8339

Dover G (1982) Molecular drive: a cohesive mode of species evolution. Nature 299:111–117

Genetic Computer Group (1992) Program manual for the GCG package, version 7, 575 Science Drive, Madison, WI 53711, USA

Golding GB, Glickman BW (1985) Sequence-directed mutagenesis: evidence from a phylogenetic history of human α-interferon genes. Proc Natl Acad Sci USA 82:8577–8581

Han J, Hsu C, Zhu Z, Longshore JW, Finley WH (1994) Overrepresentation of the disease associated (CAG) and (CGG) repeats in the human genome. Nucleic Acids Res 22:1735–1740

Karlin S, Burge C (1995) Dinucleotide relative abundance extremes: a genomic signature. Trends Genet 11:283–290

Laprevotte I (1989) Scrambled duplications in the feline leukemia virus *gag* gene: a putative pattern for molecular evolution. J Mol Evol 29:135–148

Laprevotte I (1992) Mo-MuLV nucleotide sequence exhibits three levels of oligomeric repetitions, suggesting a stepwise molecular evolution. J Mol Evol 35:420–428

Laprevotte I, Hampe A, Sherr CJ, Galibert F (1984) Nucleotide sequence of the *gag* gene and *gag-pol* junction of feline leukemia virus. J Virol 50:884–894

Martinez MA, Sala M, Vartanian JP, Wain-Hobson S (1995) Reverse transcriptase and substrate dependence of the RNA hypermutagenesis reaction. Nucleic Acids Res 23:2573–2578

Nussinov R (1982) Some indications for inverse DNA duplication. J Theor Biol 95:783–791

Ohno S (1981) (AGCTG) (AGCTG) (AGCTG) (GGGTG) as the primordial sequence of intergenic spacers: the role in immunoglobulin class switch. Differentiation 18:65–74

Ohn S (1988) Codon preference is but an illusion created by the construction principle of coding sequences. Proc Natl Acad Sci USA 85:4378–4382

Ohno S, Yomo T (1990) Various regulatory sequences are deprived of their uniqueness by the universal rule of TA/CG deficiency and TG/CT excess. Proc Natl Acad Sci USA 87:1218–1222

Seto MH, Brunck TK, Bernstein RL (1989) Overlapping redundant sextuplets identical with regulatory elements of HIV-1 and SV40. Nucleic Acids Res 17:2783–2800

Southern E (1972) Repetitive DNA in mammals. Symposia Medica Hoechst No 6. Schaltauer Verlag, Stuttgart, pp 19–27

Tautz D, Trick M, Dover GA (1986) Crypic simplicity in DNA is a major source of genetic variation. Nature 322:652–656

Temin HM (1980) Origin of retroviruses from cellular moveable genetic elements. Cell 21:599–600

Van Hemert FJ, Berkhout B (1995) The tendency of lentiviral open reading frames to become A-rich: constraints imposed by viral genome organization and cellular tRNA availability. J Mol Evol 41:132–140

Weber JL (1990) Informativeness of human (dC-dA)*n*.(dG-dT)*n* polymorphisms. Genomics 7:524–530

Wooster R, Cleton-Jansen AM, Collins N, Mangion J, Cornelis RS, Cooper CS, Gusterson BA, Ponder BAJ, von Deimling A, Wiestler OD, Cornelisse CJ, Devilee P, Stratton MR (1994) Instability of short tandem repeats (microsatellites) in human cancers. Nat Genet 6:152–156

Xiong Y, Eickbush TH (1990) Origin and evolution of retroelements based upon their reverse transcriptase sequences. EMBO J 9:3353–3362