J Mol Evol (1992) 35:420–428

Journal of
Molecular Evolution

Mo-MuLV Nucleotide Sequence Exhibits Three Levels of Oligomeric Repetitions, Suggesting a Stepwise Molecular Evolution

Ivan Laprevotte

UPR 41 CNRS "Recombinaisons Génétiques," Centre Hayem Hôpital Saint-Louis, 75475 Paris cedex 10, France

Summary. An exhaustive computer-assisted analysis of the Moloney murine leukemia virus nucleotide sequence shows numerous deviations in the oligomeric distribution, suggesting three overlapping levels of a stepwise duplicative evolution. (1) The sequence fits the universal rule of TG/CT excess which has been proposed as the construction principle of all sequences, and maintains some degree of symmetry between the two complementary strands. (2) Oligomeric repeating units share a core consensus regularly scattered throughout the sequence. This consensus is not merely predictable from the doublet frequencies and codon usage, but could correspond to an intermediary stage in a so-called periodic-to-chaotic transition. (3) Probable stepwise local duplications could be accounted for by slippagelike mechanisms. Comparison with the *h*uman *s*puma*r*etro*v*irus (HSRV) shows similar segments in the overrepresented oligomers of the two sequences. The intermediary stage of transition oligomeric repeating units is not so clearly suggested in HSRV, perhaps because of numerous stepwise local duplications. In any case, a common evolutionary origin for the two viruses is not ruled out.

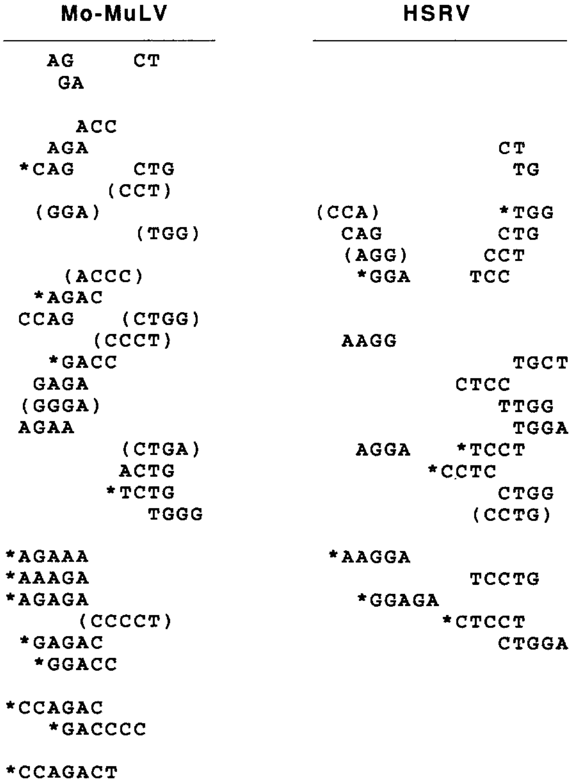
Key words: Mo-MulV — HSRV — DNA — Computer-assisted analysis — Core consensus — TG/CT excess — Symmetry — Markov analysis

Introduction

Internal nucleotide repetitions appear to be a widespread feature both in coding and noncoding DNA sequences (Doolittle and Sapienza 1980; Orgel and Crick 1980; Novak 1984; Ohno 1984; Greaves and Patient 1985; Tautz et al. 1986). Deletions/duplications/insertions and multiple-copy families can be accounted for by diverse putative mechanisms like unequal crossover (Smith 1976), transposition (reviewed in Shepherd et al. 1984), slippagelike mechanisms (reviewed in Tautz et al. 1986), duplicative transposition, or gene conversion (reviewed in Dover 1982; Golding and Glickman 1985). Previously (Laprevotte 1989), intragenic duplications have been screened by a detailed computer-assisted analysis of two related nucleotide sequences: the feline leukemia virus *gag* gene together with its flanking sequences, compared with the aligned sectors of the Moloney strain of the murine leukemia virus (Mo-MuLV). Overrepresented oligomers are scattered throughout the whole range of both sequences, sharing a core consensus which is found to be evenly distributed and mixed with scrambled short-scale repetitions as well as purine or pyrimidine stretches. This suggests an evolution by local repetitions that could stem from an original tandemly repeated oligonucleotide (Ohno 1984; Southern 1972; Ohno and Epplen 1983) unless the homogeneity of the sequence can be explained by (an)other mechanism(s) such as gene conversion.

The present paper extends this study to the whole of the plus strand of the 8332-base-long Mo-MuLV DNA sequence corresponding to the viral RNA (Shinnick et al. 1981). It brings out a molecular pattern similar to what was found in the *gag* portion of the sequence (Laprevotte 1989). Markov chain analysis shows that such a pattern is not merely predictable from the doublet frequencies.

Moreover, comparison with the *h*uman *s*puma*r*et*rov*irus (HSRV) sequence shows an overall similarity of the overrepresented oligomers; such similarity suggests an analogous evolutionary model for both sequences.



```
    Mo-MuLV                 HSRV
   AG    CT
    GA

     ACC
   AGA                             CT
 *CAG    CTG                        TG
       (CCT)
  (GGA)               (CCA)       *TGG
         (TGG)         CAG         CTG
                        (AGG)     CCT
    (ACCC)               *GGA    TCC
  *AGAC
 CCAG   (CTGG)
      (CCCT)            AAGG
   *GACC                            TGCT
  GAGA                          CTCC
(GGGA)                             TTGG
 AGAA                               TGGA
        (CTGA)           AGGA  *TCCT
        ACTG                 *CCTC
       *TCTG                       CTGG
          TGGG                   (CCTG)

*AGAAA                 *AAGGA
*AAAGA                           TCCTG
*AGAGA                   *GGAGA
     (CCCCT)                   *CTCCT
 *GAGAC                            CTGGA
  *GGACC

*CCAGAC
   *GACCCC

*CCAGACT
```

Fig. 1. Overrepresented 2- through 7-mers in Mo-MuLV and HSRV. Each possible n-base-long motif is observed at x nonoverlapping positions in the sequence. The probability P ($\geqslant$x) for it to be found at least at these x positions in a shuffled sequence (of the same length and the same base composition) is given by the binomial law:

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

N is the total number of positions in the sequence (sequence length $-$ n + 1); P is the probability for the motif to be found at any position, i.e., the product of the actual compositions (in the sequence) of the bases included in the n-mer (0-order Markov chain). In a shuffled sequence, the average number m of n-base-long motifs having a given value of P ($\geqslant$x), is P ($\geqslant$x) $\times$ 4^n, 4^n being the total number of n-base-long oligomers. The n-mers are referred to as overrepresented when m < 0.01. This corresponds to a very high number of occurrences. Indeed the Poisson law with m = 0.01 shows that < 0.01 shuffled sequences ($1 - e^{-0.01}$) have $\geqslant$ 1 overrepresented n-mer(s). Complementary motifs are on the same row in the corresponding vertical half. In addition, a zero and first-order Markov chain analysis has been performed in order to calculate the residual values (r.v.) between observed and expected numbers of nonoverlapping positions. (See text and Table 1). *: r.v. 0 order $\geqslant$ +3 and r.v. 1st order $\geqslant$ +3. Motifs in parentheses: r.v. 0 order $\geqslant$ +3 and r.v. 1st order $\leqslant$ +1.

Selection of the Overrepresented Base Oligomers, Taking Into Account the Nucleotide Composition of Mo-MuLV Sequence

Overrepresented oligomers (dimers through heptamers) in the Mo-MuLV sequence are shown in Fig. 1 (left). A threshold value of the number of positions in the sequence is calculated for each observed oligonucleotide motif by using P the probability for the motif to occur at any position of a shuffled sequence; P is, as usual (Day and Blake 1982), the product of the actual relative frequencies of the bases included in the motif. The n-mers referred to as overrepresented correspond to an exceptionally high number of occurrences, so that <0.01 shuffled sequences could have at least one oligonucleotide of the same length with at least the threshold number of occurrences. (See Fig. 1 legend.)

These overrepresented n-mers show three overlapping subsets, as previously shown for the *gag* region (Laprevotte 1989): (1) purine and (2) pyrimidine (essentially C) stretches and (3) a set of alternating purine and pyrimidine short runs which are part of a CCAGACC consensus sequence. In addition, 12 n-mers out of 32 include TG and/or CT that is complementary to CA and AG in the consensus and is in agreement with the universal rule of TG/CT excess (Nussinov 1982; Ohno and Yomo 1990).

A certain degree of symmetry between the two complementary DNA strands is suggested by the following data that show double-strand characteristics in a single DNA strand: 3 × 2 overrepresented n-mers (AG/CT, CAG/CTG, and CCAG/CTGG) are complementary; also complementary are CAGG/CCTG in Fig. 2 (see below); moreover, purines and pyrimidines have strictly equal numbers of positions in Mo-MuLV (Table 5). Such symmetry has been shown in other sequences (Nussinov 1982; Yomo and Ohno 1989).

The same method as in Fig. 1 has been additionally used to calculate an averaged threshold value of the number of positions of any n-mer, taking into account the base composition of the whole sequence and the length of the motif, whatever the base composition of the motif may be. For any n-base-long oligonucleotide observed at any position, the probability P for it to be found is now an averaged value given by p^n; p is a weighted average of the relative frequencies of the four bases (Fig. 2 legend). This method aims 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 a zero-order Markov chain. The majority of the n-mers shown in Fig. 1 (24/32) are also

Mo-MuLV	HSRV
	AA
	AT
	TA
	AAA
	(AAG)
	(AGA)
	GAA
	CAA
	TAA TTA
	AAT ATT
CC	ATA TAT
(CCA)	AAAA
CCC	AAAG
*GAC	AAAT
GCC	AAGA
	AATA TATT
AACC	AATT
(AAGA)	ACAA
ACCA	AGAA
(CAGG) (CCTG)	ATAT
(CCAA)	ATTA
(CCCA)	CAAA
CCCC	GAAA
(CCTC)	TAAA TTTA
(GCCC)	TATA
GGAC	TTAA
	TTAT
(ACCCC)	
AGCCC	AAAAG
CCAAG	AAAAT
(CCACC)	AAATA
CCAGA	AACAA
(CCCTG)	AAGAA
GACCC	AATAT ATATT
	*AGAAA
*AAGCCC	*ATCAA
(CCCCCT)	GAAAT
CCCCTT	TAAAA
	TTAAA
*CCCAGAC	TTATA
	AAAGAA
	AAGAAA
	AGAAAA
	*GAAGAA
	*AGAAAT
	AAATAT
	*AACAAT
	*AATCAA
	*TATATT

Fig. 2. Nonoverrepresented 2- through 7-mers in Mo-MuLV or HSRV nucleotide sequence, which, however, occur at at least an averaged threshold number of positions. The calculation is the same as for Fig. 1, except that P is the probability for any observed n-mer to occur at one position: P = p^n; the probability p for any given base to be found at any position is a weighted average, $pa^2 + pc^2 + pg^2 + pt^2$, the sum of the squared fractions of A, C, G, and T, respectively (Day and Blake 1982). For any n-mer, the threshold value is also chosen so that $m < 0.01$ ($m = P(\geqslant x) \times 4^n$). The motifs are displayed in the same way as in Fig. 1; * and parentheses as in Fig. 1.

found to be over this averaged threshold value. They are shown in Table 1 (in addition to GAC, see below) and are identified as "overrepeated" di- through heptamers. None of the 8- through 10-mers are so "overrepeated." The "overrepeated" ≥11-mers were found to be included in one of two sets of direct repeats: a 75-base-long tandemly repeated sector (bases 7933–8007 repeated at positions 8008–8082), or the R region occurring at positions 1–68 at the 5′ end and at 8265–8332 at the 3′ end of the genome (Shinnick et al. 1981). The displayed 2- through 7-mers prove to be "overrepeated" even if only one of each of these 75- and 68-base-long direct repeats is taken into account in the sequence.

A set of 2- to 7-base-long oligomeric motifs have a number of positions greater than the averaged threshold value but are not found to be overrepresented when using the usual calculation method (Fig. 1 legend). They are shown in Fig. 2 (left). They show the same overlapping subsets as the motifs of Fig. 1, but with an overrepresentation of C-strings. Actually C is represented 70 times out of 118 in Fig. 2, instead of 39 out of 128 in Fig. 1, with a relative frequency of 0.29 in the whole sequence (Table 5). This suggests substantial iterations of certain letter types or of certain short-length oligomers, as discussed below for Mo-MuLV and HSRV.

The Markov Chain Rule Does not Completely Predict the Frequencies of the Overrepresented Oligonucleotides in the Mo-MuLV Sequence

To address the issue of whether the frequencies of the overrepresented 3- through 7-mers can be predicted from the frequencies of the included oligomers, a Markov chain analysis was additionally performed (Phillipps et al. 1987a,b; Table 1).

Table 1 shows a first-through (n − 2)-order Markov chain analysis of the so-called "overrepeated" 3- through 7-mers in addition to GAC. None of these n-mers any longer fulfills the criterion of overrepresentation defined in the legend of Fig. 1 and corresponding to an exceptionally high number of occurrences. This indicates some correlation between overrepetitions of n-mers and the frequencies of the included oligomers, particularly the dimer frequencies. However, in order to demonstrate residual deviations of the repeat occurrence distributions, Table 1 takes into account the successive values of P (≥x) (Fig. 1 legend), which is the probability for each n-mer to have at least its observed number of occurrences in the sequence; Table 1 also shows the corresponding residual values (r.v., see Table 1) that evaluate the deviations of the observed numbers of occurrences from the expected numbers (even for the lowest values of the expected numbers). In fact three subsets can be delineated:

1) CCT, CCCT, CCCCT, ACCC, and GGGA show P (≥x) > 0.05 (0.089–0.67) with r.v. < +2, even at the first-order level, so their repetitiveness appears to derive from the dimer distribution.
2) Twelve n-mers have P (≥x) <0.01 (3.9×10^{-5} to 6×10^{-3}) and r.v. > +2, in the first-order col-

Table 1. Markov chain analysis of the overrepeated oligomers of Mo-MuLV[a]

	n-order Markov chain					
Oligonucleotide	0 order	1st order	2nd order	3rd order	4th order	5th order
AG	1.3×10^{-5} (+4)					
CT	4.0×10^{-10} (+6)					
ACC	8.3×10^{-6} (+4)	2.5×10^{-2} (+2)				
AGA	4.8×10^{-9} (+6)	4.1×10^{-2} (+2)				
CAG	1.2×10^{-5} (+4)	2.7×10^{-3} (+3)				
CCT	6.8×10^{-11} (+6)	6.7×10^{-1} (0)				
CTG	1.9×10^{-12} (+7)	2.4×10^{-2} (+2)				
GAC	4.4×10^{-4} (+3)	5.1×10^{-3} (+3)				
ACCC	1.3×10^{-5} (+4)	1.9×10^{-1} (+1)	4.1×10^{-1} (0)			
AGAC	1.8×10^{-5} (+4)	6.0×10^{-3} (+3)	8.4×10^{-1} (−1)			
CCAG	1.8×10^{-5} (+4)	6.9×10^{-2} (+2)	7.5×10^{-1} (−1)			
CCCT	1.7×10^{-9} (+6)	5.2×10^{-1} (0)	2.4×10^{-1} (+1)			
GACC	7.4×10^{-8} (+5)	1.5×10^{-3} (+3)	4.1×10^{-1} (0)			
GAGA	6.4×10^{-7} (+5)	3.5×10^{-2} (+2)	1.2×10^{-1} (+1)			
GGGA	8.7×10^{-7} (+5)	2.6×10^{-1} (+1)	2.3×10^{-1} (+1)			
AGAA	2.5×10^{-7} (+5)	1.0×10^{-2} (+2)	9.7×10^{-2} (+1)			
AGAAA	2.1×10^{-6} (+5)	2.6×10^{-3} (+3)	4.4×10^{-2} (+2)	2.1×10^{-1} (+1)		
AAAGA	2.1×10^{-6} (+5)	2.6×10^{-3} (+3)	1.5×10^{-1} (+1)	8.5×10^{-2} (+1)		
AGAGA	6.8×10^{-8} (+5)	2.3×10^{-3} (+3)	1.0×10^{-1} (+1)	4.2×10^{-1} (0)		
CCCCT	6.1×10^{-8} (+5)	2.8×10^{-1} (+1)	8.9×10^{-2} (+1)	2.2×10^{-1} (+1)		
GAGAC	5.5×10^{-6} (+4)	1.5×10^{-3} (+3)	3.8×10^{-2} (+2)	1.6×10^{-1} (+1)		
GGACC	2.0×10^{-7} (+5)	2.6×10^{-3} (+3)	7.3×10^{-2} (+2)	1.7×10^{-1} (+1)		
CCAGAC	2.6×10^{-7} (+5)	3.9×10^{-5} (+4)	3.5×10^{-2} (+2)	5.7×10^{-4} (+3)	2.9×10^{-2} (+2)	
GACCCC	1.1×10^{-6} (+5)	5.6×10^{-3} (+3)	3.3×10^{-2} (+2)	5.7×10^{-2} (+2)	1.4×10^{-1} (+1)	
CCAGACT	5.9×10^{-7} (+5)	6.1×10^{-5} (+4)	3.6×10^{-3} (+3)	3.5×10^{-4} (+4)	1.7×10^{-2} (+2)	4.4×10^{-1} (0)

[a] The displayed values are the probabilities P (≥x) calculated by the method described in the legend of Fig. 1, except that the probability P for one n-base-long motif to occur at one position is now given by a first-through (n − 2)-order Markov chain (Phillips et al. 1987a,b). For example, when a third-order Markov chain is applied to CCAGAC, P is obtained by multiplying the observed relative frequencies (overlapping positions included) of the overlapping set of tetramers and by dividing the result by the relative frequencies of the two overlapping transition trimers, that is

$$P(CCAGAC) = \frac{p(CCAG) \times p(CAGA) \times p(AGAC)}{p(CAG) \times p(AGA)}.$$

The "0 order" column refers to calculations with P equal to the product of the actual compositions of the included bases (legend of Fig. 1). In parentheses are indicated residual values (r.v.) given by (Phillips et al. 1987a) : r.v. = $\sqrt{2[(ob)\ln(ob/ex) - (ob - ex)]}$, where ob is the observed number of nonoverlapping positions of the indicated n-mer and ex is the expected number P × N, N being the total number of positions (8332 − n + 1). The magnitude of r.v., that is given the sign of (ob − ex), indicates the extent of deviation from the expected number and has been rounded to the nearest integer. GAC is added to the "overrepeated" n-mers displayed in this table, taking into account a χ^2 analysis of the trimeric distribution and the first-order Markov chain rule. (See the text.)

umn, so they clearly do not fit values predicted by a first-order Markov chain. Of these, five (underlined) have (P [≥x] × 4^n) < 1 (Fig. 1 legend), while eight are part of the above-mentioned CCAGACC consensus with one mismatch in three of them (CAG, GAC, AGAC, GACC, GAGAC, GGACC, CCAGAC, CCAGACT) and four have 3 or 4 bases matching the consensus, and may have been extended by monomeric or dimeric duplications (**AGAAA, AAAGA, AGAGA, GACCCC**).

3) ACC, AGA, CTG, CCAG, GAGA, and AGAA show intermediary values: P (≥x) = 0.01–0.069 (<0.05 for all but one) with r.v. = +2, in the first-order column.

Apart from CCCT, GGGA, and CCCCT, all of the "overrepeated" 4- through 7-mers show a r.v. gradually decreasing to ≤+2, from the zero-through-the (n−2)-order column.

In addition, a more classical method has been used to compare the nonoverlapping repeat occurrence distribution of all possible trimeric motifs in the sequence, with the distribution of the expected values calculated, as explained above, by a first-order Markov chain. Observed numbers (ob) are compared with the corresponding expected values (ex) by using a χ^2 test which shows significant deviations in the trimeric distribution (36 degrees of freedom; $p < 10^{-9}$). Only CAG and GAC have both ob > ex and $(ob - ex)^2/ex > 4.1$ (namely, 8.3 and

Table 2. Distribution of the heptamers that could stem from a (CAGAC)n tandem repetition in the Mo-MuLV nucleotide sequence[a]

Oligomer	No. of matching bases[b]	Observed no. of occurrences[c]	Expected no. of occurrences[d]	Residual value[e]
CCAGACC	7	3	0.91	+2
	6	43	11.40	+7
	5	145	101.00	+4
CAGACCA	7	3	0.82	+2
	6	19	11.40	+2
	5	150	101.00	+5
AGACCAG	7	2	0.69	+1
	6	21	11.40	+3
	5	145	101.00	+4
GACCAGA	7	2	0.69	+1
	6	31	11.40	+5
	5	148	101.00	+4
ACCAGAC	7	4	0.82	+3
	6	28	11.40	+4
	5	141	101.00	+4

[a] Distribution of the five heptamers that could stem from (CAGAC)n tandem repetitions, and homologous heptamers [with one or two mismatch(es)].

[b] Number of aligned nucleotides matching the displayed heptamer (no gap allowed).

[c] Number of nonoverlapping observed positions (ob) in Mo-MuLV nucleotide sequence.

[d] Expected number of positions (ex) given by P × N, N being the total number of positions in Mo-MuLV sequence (8326), and P being the probability for the motif to occur at any position. In the case of 7 matching bases, P is the product of the actual compositions (in Mo-MuLV) of the bases included in the displayed heptamer. In the case of 6/7 or 5/7 matches,

$$P = \frac{7!}{n!(7 - n)!} p^n(1 - p)^{(7-n)}$$

n is equal to 6 or 5, respectively, and

$$p = pa^2 + pc^2 + pg^2 + pt^2$$

(legend of Fig. 1).

[e] Residual value (r.v.) as calculated in the legend of Table 1.

$$\text{r.v.} = \sqrt{2[(ob)\ln(ob/ex) - (ob - ex)]}.$$

7.03, respectively), so GAC has been added to CAG in the "overrepeated" n-mers displayed in Table 1. These results are in agreement with the fact that CAG and GAC are the only two trimers of Table 1 with a residual value $> +2$ (r.v. $= +3$) in the first-order column, which value corresponds to P ($\geq$x) equal to 2.7×10^{-3} and 5.1×10^{-3}, respectively ($P[\geq x] \times 4^3 < 1$).

Finally, the zero- and first-order Markov chain analysis has been extended to all of the tri- through heptamers displayed in Fig. 1 and in Fig. 2 in order to calculate the residual values (r.v.) between observed and expected numbers of nonoverlapping positions (Table 1). Fifteen of them (*) have a r.v. 0 order $\geq +3$ and a r.v. 1st order $\geq +3$ and clearly do not fit values predicted by the frequencies of the included dimers; 12 have already been discussed in Table 1; the other 3 (TCTG, AAGCCC, and CCCAGAC) clearly show similarities with the "overrepeated" oligomers. Twenty-one motifs (in parentheses) have a r.v. 0 order $\geq +3$ and a r.v. 1st order $\leq +1$ and so have a repetitiveness deriving mainly from the dimer distribution; 12 have already been discussed in Fig. 2; nine are shown in Fig. 1 (five already discussed in Table 1); the majority seem to result from monomeric and dimeric iterations.

From the data discussed above, it appears that for most of the "overrepeated" 3- through 7-mers, the observed frequency is not merely a consequence of dimer distribution. However, it has been shown that there still exists some correlation between overrepetitions of oligomers and the dimer frequencies, and as the majority of the "overrepeated" oligomers share a core consensus, one can assume that the dimer distribution could be partly the consequence of tandem repetitions including CAGAC.

Remnants of Putative Tandem Repetitions Including CAGAC Are Scattered Throughout the Mo-MuLV Sequence Without Any Stringent Correlation With the Coding Nucleotide Sectors and Codon Usage

Table 2 shows the distribution of the heptamers that could stem from (CAGAC)n tandem repetitions, and of homologous heptamers (with one or two mismatches). All of the 15 observed numbers of

Table 3. Distribution of (non)coding sectors according to their rate of similarity with (CAGAC)n in Mo-MuLV nucleotide sequence[a]

	No. of matching nucleotides in each sector								
	0	1	2	3	4	5	6	≥7	
No. of observed noncoding sectors[b]	43	5	5	7	2	8	7	8	85
No. of expected noncoding sectors[b]	(36.87)	(4.68)	(6.27)	(6.27)	(7.23)	(10.31)	(5.53)	(7.86)	
No. of observed coding sectors[b]	304	39	54	52	66	89	45	66	715
No. of expected coding sectors[b]	(310.13)	(39.33)	(52.73)	(52.73)	(60.78)	(86.69)	(46.48)	(66.14)	
	347	44	59	59	68	97	52	74	800

[a] Two-dimensional distribution of contiguous 10-base-long nucleotide sectors according to their location in the coding or noncoding part of Mo-MuLV, and to their number of bases matching the putative (CAGAC)n tandem repetition (See the text.) The numbers in parentheses have been calculated as usual in order to perform a χ^2 test. For example

$$36.87 = \frac{85 \times 347}{800}$$

Eight hundred is the sum of all the observed values, the partial sums in rows and columns being indicated in the right and bottom margins.

[b] Noncoding sectors span positions 1 to 290 and 7771 to 8330; these intervals correspond to the noncoding part of Mo-MuLV nucleotide sequence except for a few bases. Coding sectors span positions 621 to 7770, an interval which corresponds to the coding part of the sequences, starting from the beginning of p15 (Shinnick et al. 1981), apart from a few bases.

nonoverlapping positions are greater than expected and ten of the residual values (r.v.) that evaluate their deviations from the expected values are greater than +2. In addition, an analogous study (data not shown) has been performed for CCAGACT (r.v. = +5), and CAGACTG (r.v. = +3), included in CCAGACTG (r.v. = +4). Three r.v. out of six are > +2, so the hypothesis of diverse intermediate repeating units (that could have stemmed from an original tandemly repeated sequence, see below and Southern 1972) cannot be ruled out. In any case, about 25% of Mo-MuLV nucleotides match with (CAGAC)n and are included in at least one heptamer sharing ≥5 bases with a motif of Table 2 (overlapping positions included).

To see whether intragenic duplications could be the result of an evolution primarily constrained by the structures of the coded proteins, these matching nucleotides have been subsequently numbered, regardless of the position(s) of the corresponding heptamer(s), in two subsets of contiguous 10-base-long sectors covering the noncoding portion of Mo-MuLV on one hand, and the coding portion, starting with p15, on the other (Table 3). The χ^2 calculated from the data shown in Table 3 do not show any significant differences (7 degrees of freedom; p > 0.30) between coding and noncoding sectors.

In order to further investigate whether intragenic duplications could be the result of an evolution primarily constrained by the functional aspect of the coded proteins, relatively overused codons are listed in Table 4. Only three of these codons (CUG, ACC, AGA) are "overrepeated" trimers listed in Table 1, and only two (ACC, AGA) are part of the consensus CCAGACC. Only one (ACC) is regularly overused in all of the successive sectors of the coding part of Mo-MuLV, while the others suggest that the codon usage could be partly accounted for by local repetitions. Other trimers that could be part of (CAGAC)n, particularly CAG that could be in the same phase as ACC, are not overused codons. Thus some correlation exists in fact, as expected, between overrepetition of trimers and codon usage, inasmuch as the coding part makes up about 90% of the sequence; however, as previously shown (Laprevotte 1989), it is unlikely for the n-meric repetitions to be merely the consequence of the codon usage.

Moreover, an analysis of the framing positions of occurrences of CAGAC (they do not overlap and occur exclusively in coding regions) do not show significant differences (phase 1: 7 positions, 11 in phase 2 and 4 in phase 3; χ^2 with 2 degrees of freedom = 3.364; p > 0.10). This further suggests that the intragenic duplications are not the result of an evolution primarily constrained by the structures of the coded proteins.

Remnants of Putative (CAGAC)n Tandem Repetitions Could Have Been Shifted by Short-Scale Local Duplications/Deletions

Additional data fit the conclusion of the previous exhaustive analysis of the *gag* region (Laprevotte 1989)—that scrambled short-scale local duplications/deletions could have played an important role in the evolution of the Mo-MuLV nucleotide sequence by shifting the remnants of these putative tandem repetitions:

1) χ^2 analyses of the dimers close to (+1) CAG and (+1) GAC in the sequence (5′ and 3′ positions [−2 to −6] and [+4 to +8], respectively) have been done. For each 5′ or 3′ position a χ^2 anal-

Table 4. Overused codons in the coding parts of Mo-MuLV sequence[a]

All together	*gag*	*pol*	*env*
CUG (leucine)		CUG (leucine)	
UCU (serine)	UCU (serine)		
UCC (serine)			UCC (serine)
CCC (proline)		CCC (proline)	
ACC (threonine)	ACC (threonine)	ACC (threonine)	ACC (threonine)
GCC (alanine)		GCC (alanine)	
AGA (arginine)		AGA (arginine)	AGA (arginine)
GGA (glycine)	GGA (glycine)		

[a] Codon usage is studied in all the genes together (left column) and in each of them separately (*gag, pol,* or *env,* Shinnick et al. 1981). For each of the a codons specifying each amino acid, the corresponding expected number (ex) is given by b/a, where b is the total number of positions occupied by these codons. When ex ≥ 5, a χ^2 analysis (a $-$ 1 *df*) is done in the corresponding subset. When $p < 0.05$, the corresponding codons with ob (observed number of positions) > ex are screened. Only the codons with $(ob - ex)^2/ex > 6.64$ (χ^2; $p < 0.01$; 1 *df*) are displayed.

ysis has been performed by comparing the observed number of occurrences (ob) of each dimer and the expected number (ex) calculated from the relative frequency (overlapping dimeric motifs included) in the whole sequence. All but two χ^2 are <23.6. A significant deviation is only demonstrated for two sets of values ($p < 0.001$) corresponding to GACXX ($\chi^2 = 42.7$) and XXGAC ($\chi^2 = 55.7$). In both sets, some individual values ($[ob\text{-}ex]^2/ex > 3.8$) are clearly overrepresented and corrrespond to three subsets: (i) GACCA and CAGAC that are part of the consensus discussed above, (ii) GACCC, GAGAC, and GGGAC that could have been extended by monomeric or dimeric duplications and fit the internally iterative oligomers discussed above (Figs. 1, 2), and (iii) GACTG that fits the universal rule of TG/CT excess already mentioned.

2) The calculations displayed in Table 5 demonstrate that an oligopurine and an oligopyrimidine bias occurs in the sequence. This is in agreement with a previous analysis of 24 eukaryotic viruses including retroviruses (Beasty and Behe 1988). The oligomers discussed in Fig. 2 include essentially the base C (70/118) and, to a lesser extent, A (26/118), which are the two most represented in Mo-MuLV (Table 5). Taken all together, purine and pyrimidine runs, which are overrepresented in the sequence, contain even greater fractions of A and C (Table 5). Together with the data discussed above concerning the internally iterative n-mers and the dimers close to CAG and GAC, this suggests processes of slippagelike local duplications involving one or a few bases.

The Three Levels of Oligomeric Repetitions in Mo-MuLV as Compared With the Oligomeric Repetitions in HSRV

The data discussed above suggest a model of three imbricated levels of oligomeric repetitions in the Mo-MuLV sequence (or, at least partially, in an original viral and/or eukaryotic sequence, see below):

1) The sequence fits the universal rule of TG/CT excess which has been proposed as the construction principle of all sequences (Ohno and Yomo 1990). This fact together with the possible maintenance of some degree of symmetry between the two complementary strands makes it possible to hypothesize a universal internally repeated original sequence and one or more inverted duplications [and/or gene conversions] at some step(s) of the evolution (Nussinov 1982). It is interesting to point out that repetitive (TG/CA)n and (CT/AG)n sequences are found to be abundant in rodent and human genomes, but almost completely absent in bacterial genomes (Tripathi and Brahmachari 1991).
2) At an intermediary stage, possibly a so-called periodic-to-chaotic transition (Southern 1972; Ohno 1988), oligomeric repeating units may have been formed by the multiplication of segments taken out of the preexisting repeated array(s). Such units could account for the consensus CCAGACC and possibly other slightly different ones such as CCAGACTG.
3) Stepwise local duplications are suggested by numerous short-range tandem repetitions. They could be accounted for by slippagelike mechanisms (Tautz et al. 1986).

Although the above model gives a good account of the highly repetitive Mo-MuLV sequence, it nonetheless does not rule out other mechanisms, such as gene conversion (leading to homogeneity throughout DNA sequences, Laprevotte 1989) and a converging evolution toward repeated motifs serving useful functions. For example, it should be kept in mind that, intriguingly, GACC or its inverse GGTC is present in the vicinity of all nucleotide

Table 5. Base compositions of Mo-MuLV and HSRV and distribution of ≥10-base-long iterations of ≤2 bases[a]

Mo-MuLV			HSRV		
Base composition					
	In the whole sequence	In purine and pyrimidine runs		In the whole sequence	In purine and pyrimidine runs
C	0.29	0.31	A	0.33	0.45
A	0.26	0.29	T	0.28	0.13
G	0.24	0.19	G	0.20	0.32
T	0.21	0.21	C	0.18	0.10

≥10-base-long runs composed of ≤2 bases

Bases	Residual values	Bases	Residual values
A/G purines	+4	A/G purines	+4
C/T pyrimidines	+4	C/T pyrimidines	+2
A/C	+1	A/C	+3
A/T	+2	A/T	0
C/G	0	C/G	+1
G/T	−1	G/T	0

[a] Each purine or pyrimidine run consists of ≥10 contiguous purines or pyrimidines. The base composition of these runs taken together, in Mo-MuLV and HSRV, respectively, is compared with that of the whole corresponding sequence (upper part). In addition (bottom), for these purine or pyrimidine tracts, together with the four other possible ≥10-base-long runs composed of ≤2 bases, the observed and expected occurrences of each type of tracts are compared by calculating a residual value, as explained in Table 1. Let p_r be the fraction in the whole sequence of the two bases included in the run, p_y that of the two bases not included, and L the length of the sequence; the expected number is given by

$$\sum_{n=10}^{\infty} [p_r^n \times p_y^2 \times (L - n - 1)] + (2 \times p_r^n \times p_y) .$$

Practically, the program stops when the added up successive values no longer show any significant change.

substitutions and deletions of a mouse sarcoma virus (Van Beveren et al. 1981). Furthermore, a conserved sequence TGACTCT, found in both mouse and human genes, acts as a *ras*-responsive enhancer element, with a mutation to AGACTCT making it even more effective (Owen and Ostrowski 1990). The consensus oligomers could have some properties, e.g., unusual stability or affinity for some viral or host-encoded protein, and have therefore become overrepresented through positive selection of such point mutations as lead to greater similarity to the consensus sequence in given regions. However, duplicative processes appear to be a major source of genetic variation in all regions of the genome (Tautz et al. 1986).

A comparative study has been done on the *h*uman *s*puma*r*etro*v*irus (HSRV), using the plus-strand of the 11,158-base-long concatenated DNA sequence (Flügel et al. 1987; Maurer et al. 1988) corresponding to the viral RNA. HSRV-overrepresented oligomers are displayed in Fig. 1 (right). They have been selected on the same basis as those of Mo-MuLV. Only one motif (AAGGA) is "over-repeated" as defined earlier. HSRV-overrepresented n-mers are part of one out of two consensuses. The first CC(A)AGGAGA has CCA, CAG, and AGA in common with Mo-MuLV CCAGACC. The second, CCTCCTGGA, shows some complementarity with the first, shares CCT, CTGG, and GGA with the Mo-MuLV-repeated n-mers, and fits the universal rule of TG/CT excess. Though a major unique consensus is not so clearly apparent as in Mo-MuLV, a common evolutionary origin of Mo-MuLV and HSRV is not ruled out. It could be that stepwise local duplications have become prominent in HSRV. All but one (AAGGA) of the oligomers that have a number of positions greater than an averaged threshold value are not overrepresented (Fig. 2), which is different from Mo-MuLV, where the lists of the oligomers screened by the two methods of calculation are overlapping. In fact these oligomers in Fig. 2 are internally repetitive; they include A/G, A/C, and A/T (with 0.28 T in the whole sequence). Moreover, purine and A/C ≥10-base-long runs are overrepresented in HSRV (Table 5); the relative frequency of (A + G) is 0.53 in the whole sequence, 0.77 in the purine and pyrimidine runs taken together (Table 5). A more exhaustive study of HSRV is beyond the scope of this paper and deserves further investigation, in order to bring out putative cryptic tandem repetitions and possible correlations with secondary structures as suggested by the 5 × 2 inverse overrepresented motifs in Fig. 1 and the 6 × 2 inverse n-mers in Fig. 2.

This detailed analysis of the Mo-MuLV nucleotide sequence compared with HSRV, together with previous work focusing on the *gag* regions of Mo-MuLV and FeLV (Laprevotte 1989), could serve as a basis for further investigation. The analysis can be repeated on additional retrotransposons in order to verify that retroviruses have emerged via the same process during evolution. Also, since current opinion holds that retroviruses evolved from the genome of their hosts (Temin 1984), an analysis of the host DNA would be of interest as well.

Acknowledgments. I am grateful to J. Adam for typing the manuscript, to A. Kay for his critical reading of the manuscript, and to J.J. Kupiec, who suggested that I study the HSRV sequence. I am also grateful to F. Galibert for help during this work in his laboratory and to the referees for their comments. This work was supported by the French CNRS ("Centre National de la Recherche Scientifique").

References

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

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

Doolittle WF, Sapienza C (1980) Selfish genes, the phenotype paradigm and genome evolution. Nature (London) 284:601–603

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

Flügel RM, Rethwilm A, Maurer B, Darai G (1987) Nucleotide sequence analysis of the *env* gene and its flanking regions of the human spumaretrovirus reveals two novel genes. EMBO J 6:2077–2084

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

Greaves DR, Patient RK (1985) (AT)n is an interspersed repeat in the *Xenopus* genome. EMBO J 4:2617–2626

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

Maurer B, Bannert H, Darai G, Flügel RM (1988) Analysis of the primary structure of the long terminal repeat and the *gag* and *pol* genes of the human spumaretrovirus. J Virol 62:1590–1597

Novak U (1984) Structure and properties of a highly repetitive DNA sequence in sheep. Nucleic Acids Res 12:2343–2350

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

Ohno S (1984) Birth of a unique enzyme from an alternating reading frame of the preexisted, internally repetitious coding sequence. Proc Natl Acad Sci USA 81:2421–2425

Ohno 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, Epplen JT (1983) The primitive code and repeats of base oligomers as the primordial protein-encoding sequence. Proc Natl Acad Sci USA 80:3391–3395

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

Orgel LE, Crick FHC (1980) Selfish DNA: the ultimate parasite. Nature (London) 284:604–607

Owen RD, Ostrowski MC (1990) A nuclear factor that binds to *ras*-responsive enhancer elements is present in human tumor cells. Cell Growth Different 1:601–606

Phillips GJ, Arnold J, Ivarie R (1987a) Mono- through hexanucleotide composition of the *Escherichia coli* genome: a Markov chain analysis. Nucleic Acids Res 15:2611–2626

Phillips GJ, Arnold J, Ivarie R (1987b) The effect of codon usage on the oligonucleotide composition of the *E. coli* genome and identification of over- and underrepresented sequences by Markov chain analysis. Nucleic Acids Res 16:2627–2638

Shepherd NS, Schwarz-Sommer Z, Blumberg vel Spalve J, Gupta M, Wienand U, Saedler H (1984) Similarity of the *Cin*1 repetitive family of *Zea mays* to eukaryotic transposable elements. Nature (London) 307:185–187

Shinnick TM, Lerner RA, Sutcliffe JG (1981) Nucleotide sequence of Moloney murine leukaemia virus. Nature (London) 293:543–548

Smith GP (1976) Evolution of repeated DNA sequences by unequal crossover. Science 191:528–535

Southern E (1972) Repetitive DNA in mammals. Symposia Medica Hoechst No6, Schattauer Verlag, Stuttgart, New York, pp 19–27

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

Temin HM (1984) L'origine des rétrovirus. La Recherche 15: 192–203

Tripathi J, Brahmachari SK (1991) Distribution of simple repetitive (TG/CA)n and (CT/AG)n sequences in human and rodent genomes. J Biomol Struct Dynam 9:387–397

Van Beveren C, Van Straaten F, Galleshaw JA, Verma IM (1981) Nucleotide sequence of the genome of a murine sarcoma virus. Cell 27:97–108

Yomo T, Ohno S (1989) Concordant evolution of coding and noncoding regions of DNA made possible by the universal rule of TA/CG deficiency-TG/CT excess. Proc Natl Acad Sci USA 86:8452–8456

Received January 18, 1992/Revised May 6, 1992