Gene, 119 (1992) 207–213

23.

GENE 06661

Compositional bimodality and evolution of retroviral genomes

(Retrovirus; isochores; CpG islands; viral genes; base composition)

Serguei Zoubak*, Alla Rynditch* and Giorgio Bernardi

Laboratoire de Génétique Moléculaire, Institut Jacques Monod, Paris, France

Received by F. Galibert: 13 April 1992; Accepted: 26 April 1992; Received at publishers: 11 June 1992

SUMMARY

The compositional distributions of genomes, genes (and their third codon positions) and long terminal repeats from retroviruses of warm-blooded vertebrates are characterized by a striking bimodality which is accompanied by a remarkable compositional homogeneity within each retroviral genome. A first, major class of retroviral genomes is GC-rich, whereas a second, minor class is GC-poor. Representative expressed viral genomes from the two classes integrate in GC-rich and GC-poor isochores, respectively, of host genomes. The first class comprises all oncoviruses (except B-types and some D-types), the second, lentiviruses, spumaviruses, as well as B-type and some D-type oncoviruses (e.g., mouse mammary tumor virus and simian retroviruses type D, respectively). The compositional bimodal distribution of retroviral genomes and the accompanying compositional homogeneity within each retroviral genome appear to be the result of the compositional evolution of retroviral genomes in their integrated form.

INTRODUCTION

Previous investigations from our laboratory have shown that the integration of expressed retroviral genomes into mammalian genomes is isopycnic and compartmentalized, i.e., takes place in isochore families (or compositional compartments) of the host genomes that are close in composition (and therefore in buoyant density) to the integrated viral genomes. It should be recalled that mammalian genomes, as well as avian genomes, are mosaics of long DNA segments (> 300 kb on the average), the isochores, that are homogeneous in base composition and belong to a small number of families which cover a broad GC range (Bernardi et al., 1985; Bernardi, 1989).

An isopycnic integration was first shown for the genome of BLV (Kettmann et al., 1979; 1980), then for the genomes of MMTV (Salinas et al., 1987), RSV (Rynditch et al., 1991) and HTLV-I. Three of these retroviral genomes, BLV, RSV and HTLV-I, are GC-rich and integrate in the GC-rich isochores of the host genome. One, MMTV, is GC-poor and integrates in the GC-poor isochores of the host genome. Interestingly, eight out of nine HBV sequences integrated in the hepatocarcinoma Alexander cell line were also found in the GC-richest isochores which matched compositionally the HBV sequences (Zerial et al., 1986).

Several points should be stressed: (*i*) that the isopycnic integration was demonstrated by detecting the retroviral sequences in compositional fractions (as obtained by Cs_2SO_4 density gradient centrifugation in the presence of

Correspondence to: Dr. G. Bernardi, Laboratoire de Génétique Moléculaire, Institut Jacques Monod, 2 Place Jussieu, 75005 Paris, France. Tel. (33-1) 43 29 58 24; Fax (33-1) 44 27 79 77.

* Permanent address: (S.Z. and A.R.) Institute of Molecular Biology and Genetics, Ukrainian Academy of Sciences, 252627 Kiev, Ukraine. Tel. (7-044) 266 34 98.

Abbreviations: BLV, bovine leukemia virus; bp, base pair(s); FIV, feline immunodeficiency virus; GC, % of guanine+cytosine; HBV, human hepatitis B virus; HTLV-I, human T-cell leukemia virus type 1; kb, kilobase(s) or 1000 bp; LTR, long terminal repeat; MMTV, mouse mammary tumor virus; nt, nucleotide(s); ORF, open reading frame; *onc*, oncogene(s); *reg*, gene(s) involved in regulation of viral transcription; RSV, Rous sarcoma virus.

sequence-specific DNA ligands; see the references quoted above) of host DNA preparations 50–100 kb in size; (*ii*) that for each virus all the independent host cell isolates gave essentially the same results; and (*iii*) that the isopycnic integration of retroviral sequences is paralleled by the finding that mobile sequences, like SINES and LINES (Singer, 1982), are also found in genome regions which are close in composition to those of mobile sequences (Soriano et al., 1983); these are GC-rich for SINES and GC-poor for LINES. In other words, the situation found for integrated retroviral sequences and for mobile sequences is similar to that found for coding sequences in general, since the latter are compositionally correlated with the genome regions in which they are located (Bernardi et al., 1985; Bernardi, 1989).

Under these circumstances, an analysis of the composition of viral genomes that integrate into the host genome should be of interest (*i*) because it should allow classifying viral genomes according to their presumed integration in different compositional compartments of the host genome; and (*ii*) because it might shed light on the expression and evolution of both the viral genomes investigated and of the host genomes.

In the present paper we have studied these problems by investigating the compositional distribution of retroviral genomes from warm-blooded vertebrates, of their genes and of their LTRs. The retroviral sequences investigated here were obtained from Release 69.0 (September 1991) of GenBank. The ACNUC retrieval system (Gouy et al., 1984) was used. Computer analysis was performed by using ANALSEQ, a software developed by the Laboratoire de Biométrie of the University of Lyon for the statistical analysis of nt sequences.

RESULTS AND DISCUSSION

(a) The compositional distribution of retroviral genomes, genes, third-codon positions, and LTRs

Fig. 1A shows a histogram of the compositional distribution of all complete genomes from retroviruses that are currently available (see Table I for a list). Genomes from different isolates of the same viruses or from very closely related viruses were pooled together in order not to bias the distribution. The histogram is characterized by two peaks, centered at 53% GC, and at 43% GC, respectively. The bimodality of the histogram of Fig. 1A is found again in the histograms of Figs. 1B and 1C, which concern retroviral genes and third-codon positions, respectively, from the same genomes. In the latter case, the distribution is more spread out and the limits between the two classes are less well defined. This depends upon the fact that the third-codon position of coding sequences that are below 45% GC are lower in GC than the first and second positions, whereas the opposite is true for coding sequences that are above 45% GC (D'Onofrio and Bernardi, 1992).

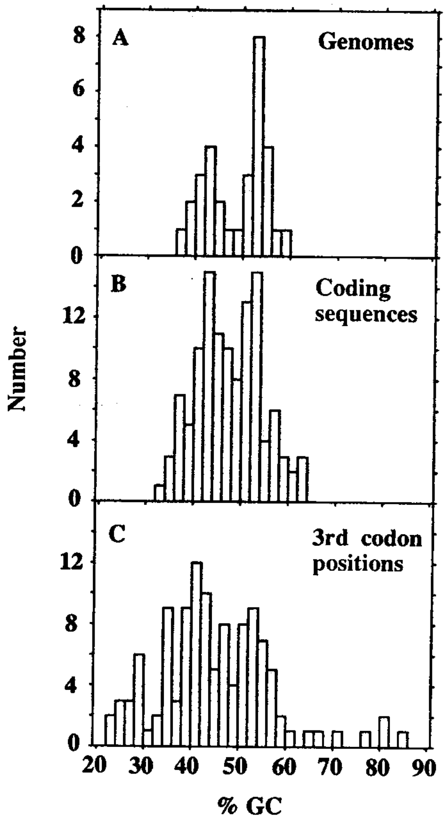

Fig. 1. Compositional distributions of (A) the 31 complete retroviral genomes of Table I; (B) their 115 coding sequences; and (C) the corresponding third-codon positions. Numbers of (A), (B) and (C) are plotted against %GC. Each bar in the histograms has a width corresponding to a 2% GC interval.

The histograms of Fig. 2, which concern *gag*, *pol*, and *env* genes from the genomes of Table I show that the genes derived from low-GC genomes are all GC-poor, whereas those derived from high-GC genomes are all GC-rich. In other words, different genes from the same retroviral genomes are very close to each other in compositional properties. Again, histograms of third-codon positions are more spread out than those of coding sequences, as expected, but the two classes of genes are still very evident.

The compositional distribution of two other sets of retroviral genes, regulatory genes (which will be called here *reg* for short) and v-*onc* genes seemed not to be bimodal (Fig. 3). This is, however, due to the fact that the *reg* genes present in the genomes studied here were predominantly associated with GC-poor genomes, whereas v-*onc* genes were predominantly associated with GC-rich genomes. A closer analysis showed that the only *reg* genes present in a GC-rich retroviral genome, those of HTLV-I are GC-rich. In the case of v-*onc* genes, most of them are GC-rich and are present in GC-rich genomes. Two of them (*ros* and *yes*)

TABLE I

List of the retroviral genomes investigated in the present work[a]

Mnemonic	Name	% GC	bp	Subfamily or type[b]
*FIV	Feline immunodeficiency viruses	36.68	9471	L
EIAVCG	Equine infectious anemia virus	38.51	8407	L
SMFVGENOME	Simian foamy virus type 1	39.20	12972	S
OLVCG	Ovine lentivirus (SA-OMVV)	40.60	9256	L
CEAVCG	Caprine arthritis encephalitis virus	41.03	9180	L
VLVCG	Visna lentivirus Icelandic strains LV1-1 and 1514	41.49	9202	L
*HIV1	Human immunodeficiency viruses type 1	42.00	9387	L
*SRV	Simian type D retroviruses	42.84	8163	D
MMTPROCG	Mouse mammary tumor virus	43.32	9900	B
*SIV	Simian immunodeficiency viruses	43.79	9828	L
BIM127	Bovine immunodeficiency-like virus	44.98	8482	L
*HIV2	Human immunodeficiency viruses type 2	45.36	9695	L
ACSUR2CG	Avian sarcoma virus UR2	47.90	3165	O
PCSLTRA	Human retrovirus type D (SMRV-HLB; SMRV-H)	49.56	8785	D
ACSY73CG	Avian sarcoma virus Y73	50.24	3718	O
FCVF6A	Feline leukemia virus subgroup A (FeLV-FAIDS)	50.73	8440	O
PCBCG	Baboon endogenous virus	51.18	8018	C+D
PCGGPE	Gibbon ape leukemia virus	52.34	8088	O
ACSASVXX	Avian sarcoma virus CT10	53.20	2428	O
ALVCG	Avian leukemia virus ALV-RSA	53.28	7286	O
*MuLV	Murine leukemia viruses	53.27	8333	O
*MuSV	Murine osteosarcoma viruses	53.40	3909	O
*HTLV-1 2	Human T-cell leukemia viruses (type 1, 2)	53.62	8807	O
MLFRO	Friend spleen focus-forming virus	53.75	6296	O
ALRCG	Rous sarcoma virus (Prague strain subgroup C)	53.84	9625	O
*MoMSV	Moloney murine sarcoma viruses	54.21	5831	O
BLVCG	Bovine leukemia virus	54.43	8714	O
PCS	Simian sarcoma virus	55.11	5159	O
MLAPRO	Abelson murine leukemia virus	55.21	5894	O
AC2E21CG	Avian retrovirus MH2E21	56.39	2630	O
ACF	Fujinami sarcoma virus	59.73	4788	O

[a] Mnemonics %GC and bp are from GenBank (release 69.0 September 1991). Asterisked mnemonics concern pooled genomes from different isolates of the same virus or from closely related viruses (a complete list can be provided upon request).

[b] L, Lentivirinae; S, Spumavirinae; O, Oncovirinae; B and D are retroviruses of types B and D, respectively (Coffin, 1990), C+D, a retrovirus of mixed type (Kato et al., 1987).

are relatively GC-poor and belong to genomes (from avian sarcoma viruses UR2 and Y73, respectively) which are at the borderline between GC-poor and GC-rich genomes. An analysis of all complete v-*onc* genes from GenBank, including genes from retroviral genomes only partially sequenced, showed that this larger set comprised both GC-poor and GC-rich genes, the latter being predominant (Fig. 3).

As far as LTR sequences from the genomes of Table I are concerned, they also show a bimodality (Fig. 4A). If, however, LTRs are divided into two groups, those belonging to GC-poor (Fig. 4B) and those belonging to GC-rich retroviral genomes (Fig. 4C), it is clear that a minority of GC-poor genomes comprise relatively GC-rich LTRs (this is the case for simian retrovirus type D, and immunodeficiency viruses, except for the feline immunodeficiency virus, FIV), and a minority of GC-rich genomes comprise GC-poor LTRs (this class is exclusively associated with avian retroviral genomes, including RSV).

A plot of GC levels of genes and LTRs from the retroviral genomes against the GC levels of the corresponding genomes (Fig. 5), reveals that *gag*, *pol* and *env* genes have GC levels which are very close to those of the other genes from the genomes to which they belong; this is much less true for *reg* and *onc* genes which show a number of deviating points. In the case of *reg* genes, this deviation might be due to the fact that average GC values under consideration concern small-size genes; in the case of oncogenes, the deviation might be due to the fact that the transduced c-*onc* genes were slightly different in composition from those of the viral genomes and that transduction events were relatively recent, leaving insufficient time to the slightly different oncogenes to adjust to the GC level of the other retroviral genes. Finally, GC levels of LTRs from GC-rich

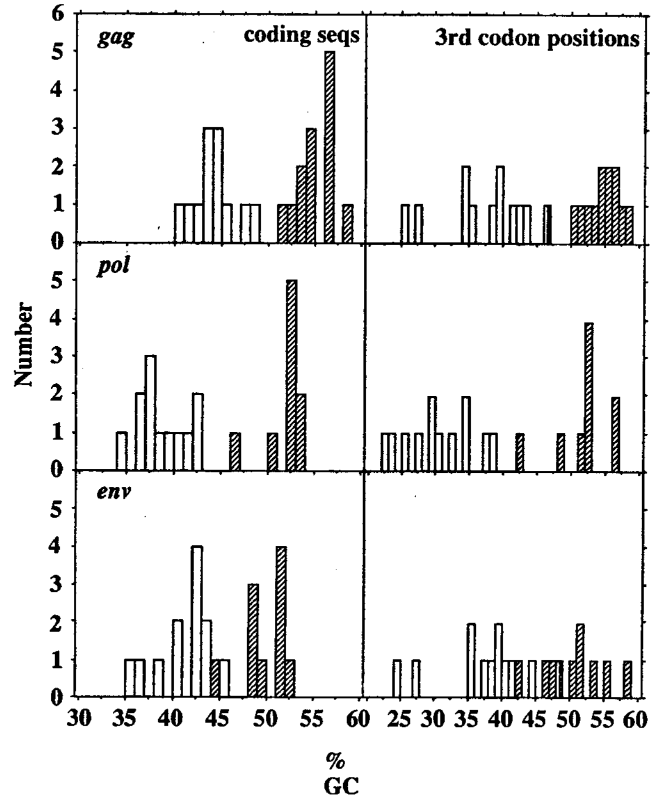

Fig. 2. Compositional distributions of retroviral genes *gag*, *env* and *pol* from the genomes of Table I (left panels) and of their third-codon positions (right panels). Open and hatched bars concern genes from the complete retroviral genomes belonging to the two classes shown in Fig. 1A (see text). Note that the abscissa scale is expanded compared with that of Fig. 1.

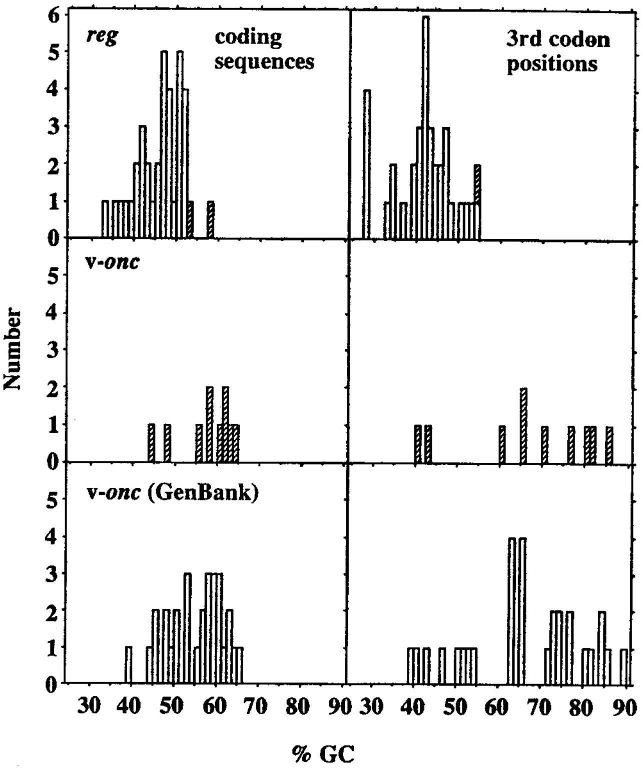

Fig. 3. Compositional distribution of retroviral genes *reg* and v-*onc* from the genomes of Table I and from GenBank (left panels) and their third-codon positions (right panels). Open and hatched bars in the four top panels concern genes from the two classes shown in Fig. 1A (see text). Note that the abscissa scale differs from that of Fig. 2.

retroviral genomes are comparable in composition to those of the corresponding genomes particularly in the case of GC-rich genomes, with the exception of the GC-poor LTRs associated with the GC-rich genomes of avian retroviruses (Fig. 5).

(b) The isopycnic integration of retroviral sequences

As an introduction to the discussion of the results just presented, it may be appropriate to recall a few points concerning the integration of retroviral sequences, as seen in a compositional perspective.

The isopycnic integration of expressed viral sequences found in our laboratory (see INTRODUCTION) contradicted the generally held opinion (for a review, see Weinberg, 1980) that viral integration in the host genome is random. This view was based on the primary structure of host-virus DNA junctions and on restriction analysis of flanking sequences (reviewed in Rynditch et al., 1991) which, in fact, did not rule out a regional specificity of integration, such as the isopycnic integration.

Other independent observations also contradicted the concept of random integration, in that they demonstrated a preference for the integration of expressed proviral sequences into transcriptionally active regions of the genome and in recombination-prone sites (see Rynditch et al., 1991).

We have recently shown that the isopycnic integration in the GC-rich regions of the mammalian genome and the integration in highly transcribed sequences are two sides of the same coin (Rynditch et al., 1991). Indeed, investigations from our laboratory (Bernardi et al., 1985; Bernardi, 1989; Mouchiroud et al., 1991; Aïssani and Bernardi, 1991a,b) have shown that the GC-richest regions of the mammalian (and avian) genome have, by far, the highest concentrations of genes and of CpG islands. The latter are regularly associated with constitutively expressed housekeeping genes (Gardiner-Garden and Frommer, 1989). These findings point to a high level of transcription in those regions which largely correspond to a rather open chromatin structure characterized by nucleosome-free stretches, absence of histone H1, acetylation of histones H2 and H4 (Tazi and Bird, 1990; see also Aïssani and Bernardi, 1991a,b). Interestingly, the GC-richest host genome regions are rich in repetitive sequences and microsatellites and are highly prone to recombination (see Bernardi, 1989; 1992).

The case of MMTV is different from those just discussed, but reinforces the isopycnic integration model, in

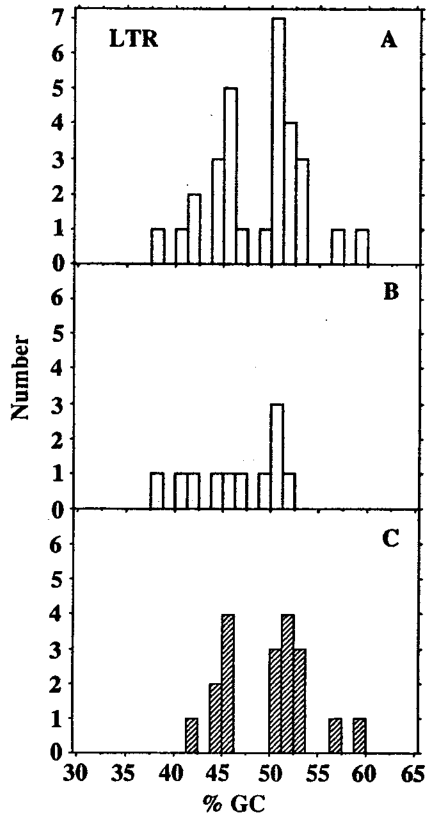

Fig. 4. Compositional distribution of LTR sequences from the retroviral genomes of Table I (A). The distributions of LTRs from GC-poor (open bars) and from GC-rich (hatched bars) genomes of Fig. 1A are shown separately in (B) and (C), respectively.

that these viral sequences, which are GC-poor, integrate into GC-poor sequences of the host genome (Salinas et al., 1987).

(c) The compositional distribution of retroviral sequences

The main result obtained in the present work is the demonstration of a bimodal distribution of GC levels of retroviral sequences. One class is GC-rich, covers a 48–60% GC range and comprises genomes like those of BLV, RSV and HTLV-I, which are known to be integrated in the GC-richest components of the bovine, hamster and human genomes, respectively (see INTRODUCTION). Another class is GC-poor, covers a 38–46% GC range and comprises genomes like that of MMTV, which is known to be integrated in the GC-poor fractions of the mouse genome.

The bimodality just discussed is also found for retroviral genes *gag*, *pol*, and *env*. This is the consequence of the compositional homogeneity of the retroviral sequences derived from the same genome. Two other sets of genes showed a distribution which seemed to be unimodal in our sample of retroviral genomes. In the case of *reg* genes, this was due to their predominant presence in GC-poor genomes; in the case of v-*onc* genes, to the fact that our sample almost exclusively comprised v-*onc* genes from GC-rich genomes. In fact, these two classes of genes are also characterized by a bimodal distribution, as shown by the facts that *reg* genes associated with GC-rich genomes (like HTLV-I) are GC-rich and that GC-poor v-*onc* genes exist (Fig. 3, bottom frame). A bimodality was also found for LTRs, but as already pointed out, the correspondence of GC-rich and GC-poor LTRs with GC-rich and GC-poor genomes, respectively, was not always observed, in particular in the case of the GC-poor LTRs which are present in the GC-rich retroviral genomes of birds. A more detailed analysis of LTRs is currently under way in order to better understand such situations.

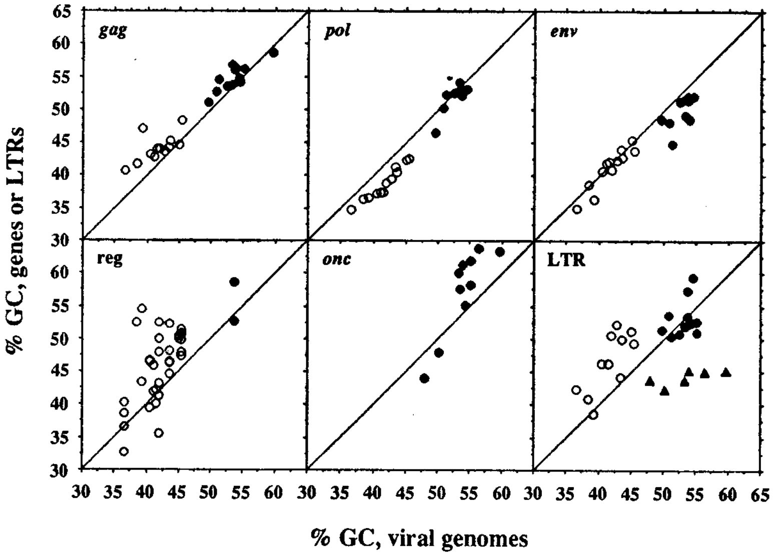

Fig. 5. Correlations between GC levels of retroviral genes and LTR sequences and GC levels of the corresponding retroviral genomes are shown for the GC-poor (open symbols) and GC-rich genomes (blackened symbols). GC-poor LTRs associated with GC-rich viral genomes are shown as blackened triangles. Diagonal (45°) lines correspond to identical GC levels of LTRs and viral genomes.

The question should now be considered which sub-families (oncovirinae, lentivirinae and spumavirinae; Coffin, 1990) of retroviruses belong to each class. The GC-poor class comprises lentiviruses and spumaviruses, which have no oncogenes, but contain genes for regulatory proteins. This class also comprises oncoviruses of the B-type and some D-types, which also do not contain oncogenes. The GC-rich class comprises all oncoviruses, except for those of B-type and some D-types, namely all retroviruses containing oncogenes, like RSV, and some retroviruses which do not contain oncogenes, like BLV and HTLV-I. It should be noted that the human retroviruses type D correspond to genomes which are at the borderline between the two classes.

The above considerations indicate that compositional classes correspond to well-defined groups of retroviruses. In turn, these groups justify considering the existence of a major and a minor class of retroviruses. This conclusion cannot be simply drawn on the basis of the histograms of Fig. 1, because they depend upon the sample of genomes which are available in data banks. If, however, one considers which sub-families of retroviruses are represented in each class, the minor and the major classes indicated by the histograms appear to exist, because oncoviruses are much more numerous than lentiviruses and spumaviruses, at least among the retroviruses studied so far. It is important to note that the GC-rich class seems not only to comprise the majority of retroviral genomes, but also to integrate into the very small GC-richest compartments of the mammalian or avian genomes. In contrast, the minor GC-poor class appears to integrate into the very large GC-poor compartments of the host genomes.

(d) Evolutionary implications

We will now discuss the question as to why the compositional distribution of retroviral sequences is not a continuous one. This question has interesting implications for the compositional evolution of retroviral genomes. It should be recalled that GC-rich isochores only appeared in vertebrate evolution after the emergence of warm-blooded vertebrates (Bernardi et al., 1985; 1989; Bernardi and Bernardi, 1990a,b; 1991). At that time, some compartments of the genomes of cold-blooded vertebrates underwent a GC increase by fixation of directional point mutations (Bernardi et al., 1988), leading to the formation of the neogenome of warm-blooded vertebrates, whereas other compartments did not undergo such GC increase and formed the paleogenome (Bernardi, 1989; Fig. 6).

One should now take into account the great evolutionary antiquity of retroviruses (Temin, 1980; 1989) and, in particular, the fact that ancestors of the retroviral genomes under consideration were in all likelihood present in cold-blooded vertebrates. Even if sequence homology evidence is not yet available, information on retroviruses from cold-blooded vertebrates exists. For example, a type C virus is responsible for lymphosarcoma in northern pike, *Esox lucius* (Papas et al., 1976; see Martineau et al., 1992, for other examples).

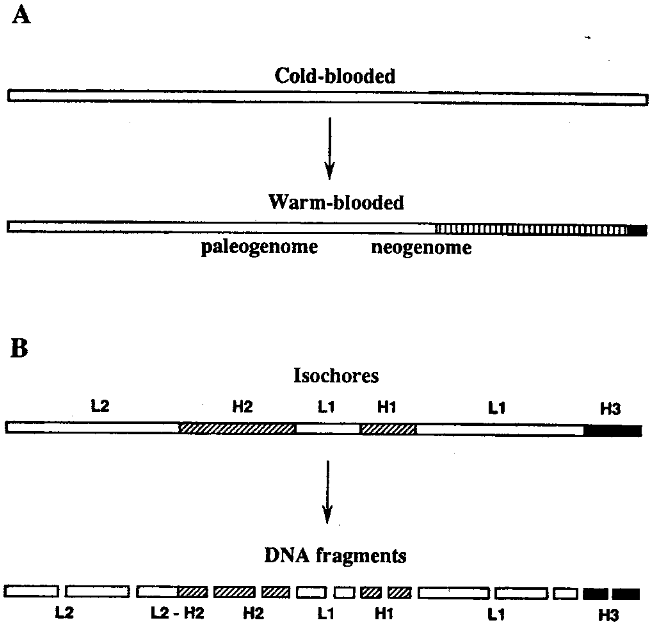

Fig. 6. (A) Scheme of the compositional genome transition accompanying the emergence of warm-blooded from cold-blooded vertebrates. The compositionally homogeneous, GC-poor genomes of cold-blooded vertebrates are changed into the compositionally heterogeneous genomes of warm-blooded vertebrates. The latter comprise a paleogenome (corresponding to about two thirds of the genome) which did not undergo any large compositional change and a neogenome (corresponding to the remaining one third of the genome, with the GC-richest part only representing 3% of the genome). In the scheme, the mosaic structure of the warm-blooded vertebrate genome is neglected; GC-poor isochores (open bar), GC-rich isochores (hatched bar) and GC-richest isochores (blackened bar) are represented as three contiguous regions. (B) Scheme of genome organization in warm-blooded vertebrates. The genome consists of long DNA segments (>300 kb, on the average) that are compositionally homogeneous (above a size of 3 kb) and belong in a small number of families, GC-poor (open boxes; L1 and L2), GC-rich (hatched boxes; H1 and H2) and very GC-rich (blackened boxes; H3). Physical and enzymatic degradation occurring during DNA preparation generates large DNA fragments, routinely in a size range of 50–100 kb.

At this point, two different situations should be considered. In the case of retroviral genomes which are located in the paleogenome of warm-blooded vertebrates, no compositional changes probably occurred relative to their ancestors located in the genome of cold-blooded vertebrates, even if, obviously, a number of mutations and genome rearrangements took place.

In the case of retroviral genomes located in the neogenome of warm-blooded vertebrates, the explanation that these genomes originated from the host genomes after the compositional transition from cold- to warm-blooded vertebrates took place is certainly incorrect because of the considerations made above about the antiquity of retroviruses. An alternative explanation, which appears to be the correct one, is that these retroviral genomes co-evolved compositionally in their integrated form, in concert with the host genome's isochores and genes that underwent the compositional transition.

In other words, what is proposed here is that the compositional evolution of integrated retroviral genomes followed the same pathways as the host genes carried by the same families of isochores. Very interestingly, even regulatory sequences (LTRs) largely followed such compositional evolution in spite of constraints associated with their regulatory sequences. Needless to say that what has just been discussed only applies to the compositional evolution of retroviral sequences. It is obvious that a large number of other mutations (occurring during transcription and retrotranscription) which did not cause compositional changes, accompanied this compositional evolution.

As a final remark, it should be stressed that nonexpressed retroviral sequences exhibit, as a rule, a much broader distribution in the host genome than that found for expressed sequences (Rynditch et al., 1991). Since the initial integration of a retroviral sequence may be targeted towards compositional compartments which are relatively broad in GC (Salinas et al., 1987) this may indicate (*i*) that selection for expressed sequences favors the retroviral integrated sequences which are characterized by a good compositional matching with the host genome isochores; (*ii*) that the non-isopycnic integration may concern largely deleted and/or rearranged sequences.

ACKNOWLEDGEMENTS

A.R. and S.Z. thank the French Ministry for Research and Technology (MRT) for Fellowships. S.Z. was also the recipient of a short-term Fellowship from the Federation of European Biochemical Societies (FEBS). This work received the financial support of the Association Française pour la Recherche contre le Cancer (ARC).

REFERENCES

Aïssani, B. and Bernardi, G.: CpG islands: features and distribution in the genome of vertebrates. Gene 106 (1991a) 173–183.

Aïssani, B. and Bernardi, G.: CpG islands, genes and isochores in the genome of vertebrates. Gene 106 (1991b) 185–195.

Bernardi, G.: The isochore organization of the human genome. Annu. Rev. Genet. 23 (1989) 637–661.

Bernardi, G.: Genome instability and species formation in vertebrates. J. Mol. Evol. (1992) in press.

Bernardi, G. and Bernardi, G.: Compositional patterns in the nuclear genomes of cold-blooded vertebrates. J. Mol. Evol. 31 (1990a) 265–281.

Bernardi, G. and Bernardi, G.: Compositional transitions in the nuclear genomes of cold-blooded vertebrates. J. Mol. Evol. 31 (1990b) 282–293.

Bernardi, G. and Bernardi, G.: Compositional properties of nuclear genes from cold-blooded vertebrates. J. Mol. Evol. 31 (1991) 57–67.

Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M. and Rodier, F.: The mosaic genome of warm-blooded vertebrates. Sciences 228 (1985) 953–958.

Bernardi, G., Mouchiroud, D., Gautier, C. and Bernardi, G.: Compositional patterns in vertebrate genomes: conservation and change in evolution. J. Mol. Evol., 28 (1988) 7–18.

Coffin, J.M.: Retroviridae and their replication. In: Fields, B.N., Knipe, D.M., Chanock, R.M., Melnick, J.L., Hirsch, M.S., Monath, T.P. and Roizman, B. (Eds.), Virology, 2nd ed., Raven Press, New York, 1990, pp. 1437–1500.

D'Onofrio, G. and Bernardi, G.: A universal compositional correlation among codon positions. Gene 110 (1992) 81–88.

Gardiner-Garden, M. and Frommer, M.: CpG islands in vertebrate genomes. J. Biol. Chem. 196 (1987) 261–282.

Gouy, M., Milleret, F., Mugnier, Jacobzone, M. and Gautier, C.: ACNUC: a nucleic acid sequence data base and analysis system. Nucleic Acids Res. 12 (1984) 121–127.

Kato, S., Matsuo, K., Nishimura, N., Takahashi, N. and Takano, T.: The entire nucleotide sequence of baboon endogenous virus DNA: chimeric genome structure of murine type C and simian type D retroviruses. Jap. J. Genet. 62 (1987) 127–137.

Kettmann, R., Meunier-Rotival, M., Cortadas, J., Cuny, G., Ghysdael, J., Mammerickx, M., Burny, A. and Bernardi, G.: Integration of bovine leukemia virus DNA in the bovine genome. Proc. Natl. Acad. Sci. USA 76 (1979) 4822–4826.

Kettmann, R., Cleuter, Y., Mammerickx, M., Meunier-Rotival, M., Bernardi, G., Burny, A. and Chantrenne, H.: Genomic integration of bovine leukemia provirus: comparison of persistent lymphocytosis with lymph node tumor form of enzootic bovine leukosis. Proc. Natl. Acad. Sci. USA 77 (1980) 2577–2581.

Martineau, D., Bowser, P.R., Renshaw, R.R. and Casey, J.W.: Molecular characterization of a unique retrovirus associated with a fish tumor. J. Virol. 66 (1992) 596–599.

Mouchiroud, D., D'Onofrio, G., Aïssani, B., Macaya, G., Gautier, C. and Bernardi, G.: The distribution of genes in the human genome. Gene 100 (1991) 181–187.

Papas, T.S., Dahlberg, J.E. and Sonstegard, R.A.: Type C virus in lymphosarcoma in northern pike (*Esox lucius*). Nature 261 (1976) 506–508.

Rynditch, A., Kadi, F., Geryk, J., Zoubak, S., Svoboda, J. and Bernardi, G.: The isopycnic, compartmentalized integration of Rous sarcoma virus sequences. Gene 106 (1991) 165–172.

Salinas, J., Zerial, M., Filipski, J., Crepin, M. and Bernardi, G.: Nonrandom distribution of MMTV proviral sequences in the mouse genome. Nucleic Acids Res. 15 (1987) 3009–3022.

Singer, M.F.: SINES and LINES: highly repeated short and long interspersed sequences in mammalian genomes. Cell 28 (1982) 433–434.

Soriano, P., Meunier-Rotival, M. and Bernardi, G.: The distribution of interspersed repeats is non-uniform and conserved in the mouse and human genomes. Proc. Natl. Acad. Sci. USA 80 (1983) 1816–1820.

Tazi, J. and Bird, A.P.: Alternative chromatin structure at CpG islands. Cell 60 (1990) 909–920.

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

Temin, H.M.: Retrons in bacteria. Nature 339 (1989) 254–255.

Weinberg, R.A.: Integrated genomes of animal viruses. Annu. Rev. Biochem. 49 (1980) 197–226.

Zerial, M., Salinas, J., Filipski, J. and Bernardi, G.: Genomic localization of hepatitis B virus in a human hepatoma cell line. Nucleic Acids Res. 14 (1986) 8373–8386.