Volume 16 Number 13 1988 Nucleic Acids Research

Primary structure and functional organization of *Drosophila* 1731 retrotransposon

F.Fourcade-Peronnet, L.d'Auriol[1], J.Becker, F.Galibert[1] and M.Best-Belpomme

UA CNRS 1135, Université P.et M.Curie, 7 quai St-Bernard, 75005 Paris and [1]Centre G.Hayem Hôpital St-Louis, 2 place du Docteur Fournier, 75010 Paris, France

Received April 22, 1988; Accepted June 8, 1988 Accession no.X07656

ABSTRACT

We have determined the nucleotide sequence of the Drosophila retrotransposon 1731. 1731 is 4648 bp long and is flanked by 336 bp terminal repeats (LTRs) previously described as being reminiscent of provirus LTRs. The 1731 genome consists of two long open reading frames (ORFs 1 and 2) which slightly overlap each other. The ORF1 and 2 present similarities with retroviral gag and pol genes respectively as shown by computer analysis. The pol gene exhibits several enzymatic activities in the following order: protease, endonuclease and reverse transcriptase. It is possible that 1731 also encompasses a ribonuclease H activity located between the endonuclease and reverse transcriptase domains. Moreover, comparison of the 1731 pol gene with the pol region of copia shows similarities extending over the protease, endonuclease and reverse transcriptase domains. We show that codon usage in the two retrotransposons is different. Finally, no ORF able to encode an env gene is detected in 1731.

INTRODUCTION

The retrotransposons constitute a broad class of moderately repeated mobile genetic elements abundantly represented among Eukaryotes, as for example the copia-like elements of Drosophila (for a review see 1), the Ty element of yeast (2), the intracisternal type A particles (IAP) of rodents (3) and the THE1 repeats of humans (4). These elements share structural features analogous to the proviral form of Vertebrate retroviruses: the entire element is surrounded by short direct repeats (4/5 bp), corresponding to the duplication of the target insertion site; long terminal repeats (LTRs) about 200 to 600 bp flank the element, each LTR begins by TG... and ends by ...CA; these two nucleotides are themselves part of short inverted repeats which flank the LTRs. Moreover, the retrotransposons contain, just downstream from the 5'LTR, sequences which could serve as tRNA primer-binding sites and oligopurine tracks which act in priming the second strand synthesis during retrovirus replication at the internal boundary of the 3'LTR (for a review of retrovirus replication see 5).The complete sequence of four Drosophila copia-like elements (copia, 412, 17.6 and 297) (6 to 9) as well as the sequence of Ty (10) and IAP (11) show the existence of at least one or two open reading frames (ORFs) which encode proteins showing similarities respectively with the gag and pol genes of retroviruses. The occurence of a third ORF encompassing an env gene will be discussed later. The pol gene of retrotransposons would encode both integrase (endonuclease) and reverse transcriptase activities and it seems likely that the transposable elements mediate their own mobility through encoded enzymatic

activities. These elements are fully transcribed into a polycistronic polyadenylated messenger RNA originating in the 5'LTR and terminating in the 3'LTR. Virus-like particles containing the corresponding RNA genome and reverse transcriptase activity have been described in the case of copia (12) and Ty (13). Circular extrachromosomal forms of copia-like elements (14-15), similar to those seen within cells infected with retroviruses, have been observed and it is widely believed that circular DNAs are intermediates in transposition (14). Moreover, Boeke et al. (16), using a Ty element with an artificially inserted intron, have shown that precise removal of the intron occurs during transposition, demonstrating that Ty transposes through an RNA intermediate. Thus, retrotransposons appear to be retroviruses having an exclusively intracellular life cycle.

The role of retrotransposons has been extensively discussed (17-1-18). They can behave as insertional mutagens and it is now clear that many mutations in Drosophila melanogaster are due to the insertion of transposable elements occasionally leading to inactivation of the recipient locus (19-20). Moreover, their insertion might introduce a new promoter and then might provoke the inappropriate expression of an adjacent gene (see for example 21-22).

We have previously reported the isolation of a new retrotransposon of Drosophila (23) through a search for genes modulated by ecdysterone , a steroid hormone which plays a crucial role in insect development. This element, named 1731, is fully transcribed into a polyadenylated messenger RNA and its transcripts decrease under ecdysterone treatment. 1731 is flanked by long terminal repeats (LTRs), which show similarities with provirus LTRs. Moreover, 1731 is moderately repeated in the genome and extrachromosomic DNA circles occur in Drosophila cells. We report here the complete nucleotide sequence of 1731. It contains two large open reading frames (ORFs) showing similarities in size and location respectively to the gag and pol genes of retroviruses.

MATERIALS AND METHODS

Materials

Restriction enzymes, DNA polymerase I (Klenow fragment) and labelled nucleotides were purchased from Amersham.

Synthetic oligonucleotides were prepared with the use of an automated DNA synthetizer as previously described (23).

M13 cloning and DNA sequencing

The BamHI/SalI 3.1 Kb fragment of subclone pFP5c (23) was transferred into M13mp8/M13mp9 and sequenced by the chain termination method (24).

The sequence of both strands was obtained. We used the cyclone system of IBI and the 17-mer universal primer for a large part of the sequence. It was completed with 25-mer synthetic oligonucleotides used as specific primers. Regions rich in GC were sequenced using the 7-deazadGTP kit of ABN.

Alignment of protein or nucleic acid sequences were realised using the Goad and Kanehisa method (25)

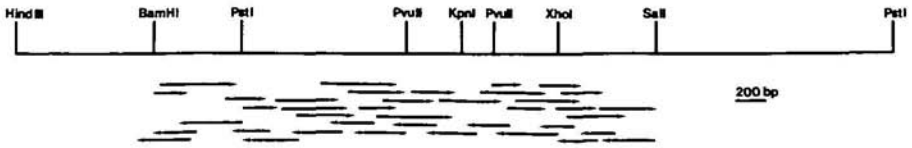


Fig. 1: Sequencing strategy for 1731. The arrows indicate the extend of sequence determined from each template. H: HindIII; B: BamHI; P: PstI; Pu: PvuII; K: KpnI; X: XhoI; S: SalI. Sequence strategy for the HindIII/BamHI and SalI/PstI fragments were previously described (23).

and the mutation matrix of Dayhoff et al. (26). Alignment between 1731 and copia were realised using the DIAGON graphic program of Staden (27).

RNA isolation

RNA was extracted from Kc cells and analysed by Northern blotting as previously described (23).

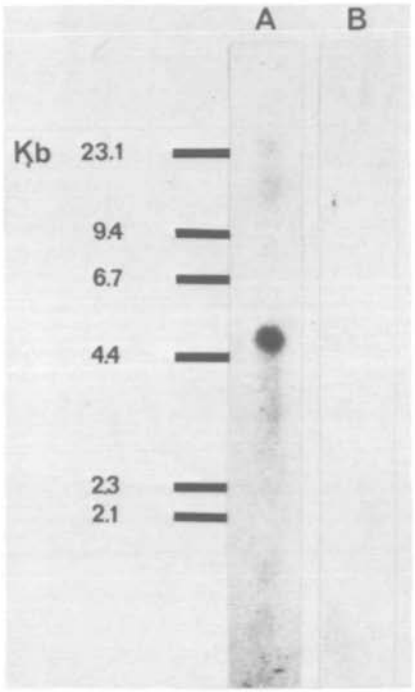
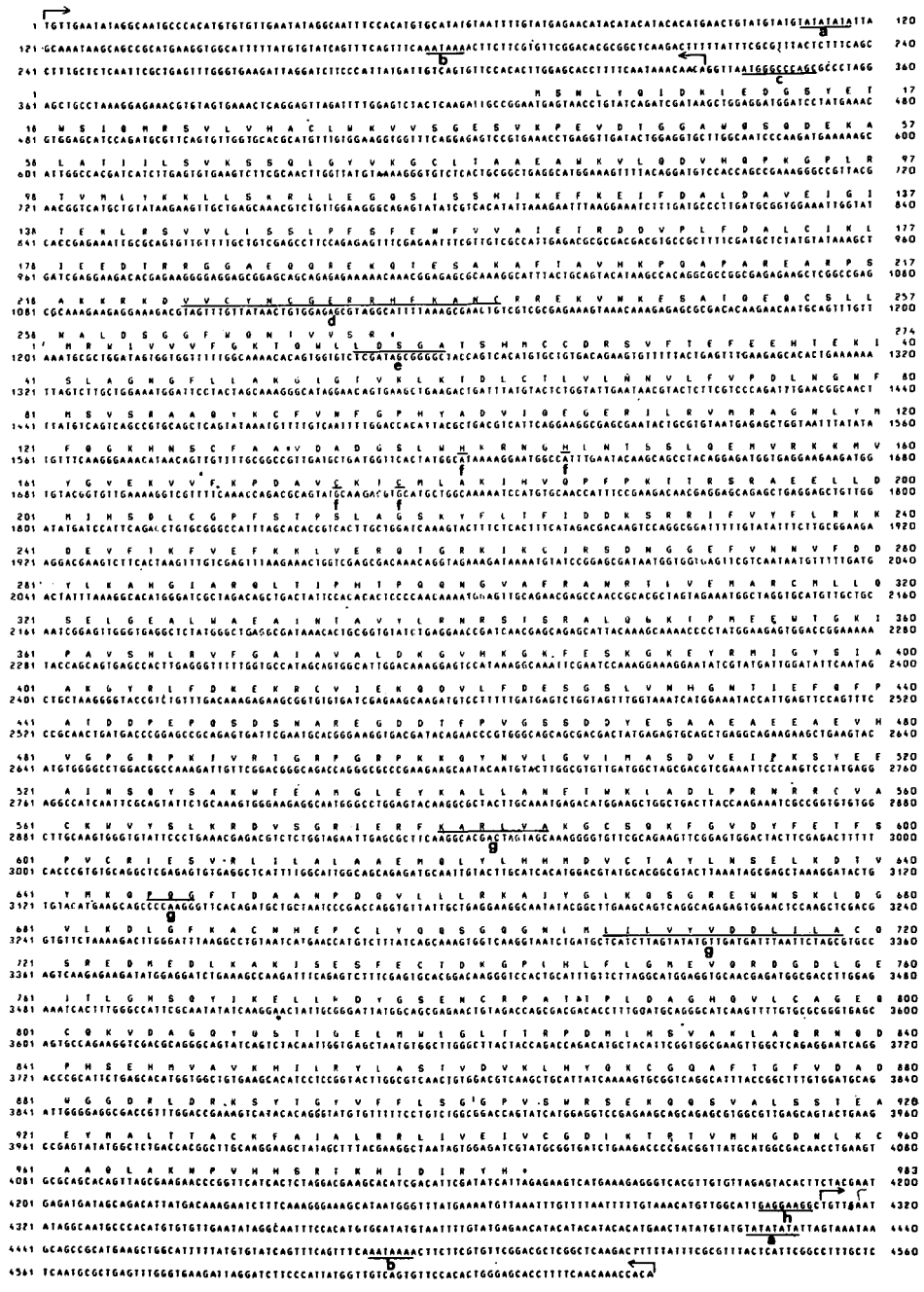


Fig.2: Northern blot analysis of polyA RNAs from Kc cells with M13 strand specific probes.
A: Probe: BamHI/SalI 3.1 Kb internal fragment of 1731 subcloned in M13mp8 (minus strand)
B: Probe: BamHI/SalI 3.1 Kb internal fragment of 1731 subcloned in M13mp9 (plus strand)
Molecular size markers: HindIII digest of lambda phage DNA.



M13 strand specific probes

The M13mp8 and M13mp9 recombinants which contain the BamHI/SalI 3.1Kb fragment of 1731 were P32-labelled by enzymatic extension of the 17-mer hybridization primer (Amersham) as described by Hu and Messing (28).

RESULTS

Nucleotide sequence of 1731, general features

The nucleotide sequence of the HindIII/PstI Drosophila genomic fragment containing 1731 was obtained by combining the sequences of the HindIII/BamHI and SalI/PstI fragments previously established (23) with the sequence of the BamHI/SalI internal fragment determined according to the strategy shown in figure 1. The whole sequence is 5486 nucleotides long. It contains 1731 which is 4648 bp long, flanked on one side, by 386 bp and on the other side, by 452 bp of Drosophila genomic sequences. 1731 is bracketed by two 5 bp direct repeats (GAATA). Such direct repeats occur in proviruses and retrotransposons and seem to be generated during integration (29-30). The LTRs are 336 bp long and were previously described as being reminiscent of provirus LTRs (23). 1731 is AT rich, containing 37.6% A, 27.2% T, 15.0% C and 20.2% G as the flanking sequences which contain 60.5% AT and share no similitude with any sequences of the GENBANK database.

Translation of 1731 shows two long overlapping open reading frames (ORF1 and ORF2). These two long ORFs reside on the DNA strand (plus strand) which contains the TATA-box, the polyadenylation signal, the possible polyA acceptor sequence, the putative tRNA primer binding site and the polypurine-rich track. The complementary strand (minus strand) does not show any open reading frame of significant size. Northern blots of Drosophila cell RNAs have been probed with the M13mp8 and M13mp9 recombinants containing the BamHI/SalI internal fragment of 1731 (figure 2). The M13mp8 recombinant probe which contains a part of the minus strand reveals the major transcript 4.6 Kb long (23), while the M13mp9 recombinant probe (plus strand) does not reveal any transcript. The presence of ORF1 and ORF2 in messenger RNA strongly suggests their usage by the translational machinery.

Moreover, we compared the sequences flanking the ATG of ORFs 1 and 2 with the consensus sequence flanking translational starts in Drosophila (31). It reveals that the ORF1-ATG (with a G in position -3) is in a better context for initiation of translation than the ORF2-ATG (with a T in position -3).

Fig.3: Nucleotide sequence of 1731 (plus strand).
The arrows indicate the 5' and the 3' LTRs.
a: TATA box; b: polyadenylation signal; c: primer binding site (see figure 4); d: nucleic acid binding site (see figure 5); e: region showing homology with retrovirus and retrotransposon proteases(see figure 6); f: zinc-binding site (beginning of the endonuclease domain) (see figure 7); g: region showing homology with reverse transcriptases (see figure 9); h: polypurine-rich track.

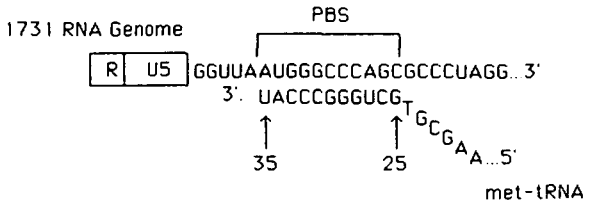


Fig.4: Alignment between the 1731 genome and a Drosophila melanogaster met-tRNA (34). The sequence of the 1731 RNA genome was deduced from the DNA sequence of the plus strand (figure 3). In our model, the Drosophila met-tRNA should have lost the nucleotides located downstream position 35. PBS: primer binding site.

This would argue for the production of an ORF1-ORF2 fusion protein obtained by a specific frame-shifting event as was proposed for HIV (32), MMTV (33), 17.6 (9) and 412 (7).

Figure 3 shows the nucleotide sequence of the plus strand of 1731; the amino-acid sequence of ORF1 and 2 are represented beneath the corresponding nucleotide sequences. As in the case of the other retrotransposons sequenced so far, ORF1 and ORF2 show similarity to the gag and pol genes of retroviruses respectively.

Binding site for tRNA primer

Retroviruses replicate through a DNA intermediate. Reverse transcription of their RNA genome is primed by a tRNA whose binding site is located just downstream from the 5'LTR (5). In order to find the tRNA primer binding site of 1731, the 30 nucleotides located downstream the 5'LTR (position 337 to 366) were compared with a tRNA sub-bank from GENBANK. Alignment between this 1731 sequence and a Drosophila melanogaster initiator met-tRNA (34) was obtained. This alignment involves an 11 nucleotide homologous region (ATGGGCCCAGC) corresponding to position 342 to 352 of 1731 and to position 25 to 35 of the met-tRNA (figure 4). It has been previously reported that the copia retrotransposon is primed by a truncated met-tRNA as shown by primer extension analysis of the virus-like particles RNAs (35). Similarly, we propose that the first strand synthesis of 1731 is primed by a met-tRNA modified by the loss of the nucleotides located downstream from the 11 nucleotide sequence, therefore localizing the primer binding site of 1731 between residues 342 and 352.

Gag similarities

The ORF at the 5' side of 1731 (ORF1) begins with the ATG in position 431 and ends with the TAG in position 1250 (figure 3). This ORF could encode a protein with a molecular weight of 30 000 daltons. Computer searches through the GENPRO and NBRF databases show similarities between this putative protein and the gag polyprotein of retroviruses; it should be noted that the whole ORF1-encoded protein presents 23% homology with a corresponding region of the HIV/ARV2 gag gene (36). Figure 5

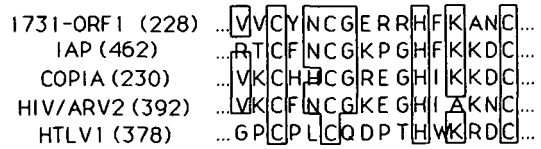


Fig.5: Nucleic acid binding site of 1731, IAP (11), copia (6), HIV/ARV2 (36), and HTLV I (37). Numbers in parentheses indicate the position of the first amino acid in the appropriate protein or ORF.

shows a 16 amino acid conserved region corresponding to the nucleic acid binding domain of some retroviruses (HIV/ARV2, HTLV) (36-37) and retrotransposons (IAP, copia) (11-6). These similarities lead one to conclude that the 1731 ORF1 encodes a "gag-like" protein.

Protease, endonuclease (integrase), RNAse H and reverse transcriptase similarities

The ORF2 (ATG in position 1203; TAG in position 4149) slightly overlaps the ORF1 and could encode a protein with a molecular weight of 110000daltons. Computer analysis shows similarities between this large ORF and the pol gene of retroviruses and suggests the existence of three enzymatic domains in the 1731 ORF2.

The first domain of about 140 amino acids is characterized by the five amino acid long stretch Leu-Asp-Ser-Gly-Ala. This domain is highly conserved among retroviruses and retrotransposons. It occurs for example in RSV (38), copia (6) and Ty (10) (figure 6) and is described as being the retroviral protease domain. Moreover, this domain could encode a polypeptide with a molecular weight of 15000 daltons which recalls the P15 RSV protease involved in the cleavage of viral polyproteins.

When the putative protein encoded by the 1731 ORF2 was submitted to computer analysis, similitudes were also detected in the region coding for the endonuclease (integrase) activity of retroviruses. The existence of an amino acid configuration which constitutes a zinc-binding domain interacting with DNA, could define the amino-terminal region of the 1731 endonuclease domain. This zinc-binding region, described in retroviruses by Johnson et al. (39), involves a pair of histidines separated by 20 to 30 residues from a pair of cysteines. 1731 contains such a zinc-binding site at position 140 of the ORF2 protein. This site is made by a pair of histidines and a pair of cysteines, each pair being separated by 28 amino acids from the other (figure 7). Moreover, comparing the sequence of copia, 412, 17.6 (6 to

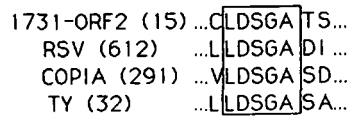
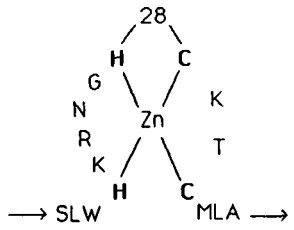


Fig.6: Homologies between 1731 and the protease region of RSV (38), copia (6) and Ty (10). Numbers in parentheses indicate the position of the first amino acid in the appropriate ORF.

A-



B-

1731	(140)	···H	KRNG	H	28	C KT C ···
MoMLV	(848)	···H	QLT	H	32	C KA C ···
COPIA	(419)	···H	ERFG	H	27	C EI C ···
412	(901)	···H	DDPIQGG	H	29	C QK C ···
17.6	(758)	···H	EKLL	H	29	C SI C ···
GYPSY	(726)	···H	NRA	H	29	C RV C ···

Fig.7: A- Representation of the putative zinc-binding site of 1731.
B- Alignment of the putative zinc-binding sites of 1731, MoMLV (46), copia (6), 412 (7), 17.6 (8) and gypsy (40) endonucleases. Numbers in parentheses indicate the position of the first histidine residue in the appropriate ORF.

8) and gypsy (40) endonucleases, we noticed the existence of this binding site in their amino-terminal portion (fig.7). The 1731 endonuclease region presents 2 short conserved amino acid sequences Asp-Asn-Gly and Glu-Arg-Ala-Asp-Arg- Thr-Leu, whose successive order and relative localisations recall those found in retroviral endonucleases (figure 8). Alignment of retroviral endonucleases (39) also involves the short amino acid stretch Try-X-Gly. The appearance of the same amino acid sequence in 1731 ORF2 suggests, by analogy with retroviruses, that the end of the 1731 endonuclease domain may be about 50 residues after this stretch. The approximate size of the entire domain is 270

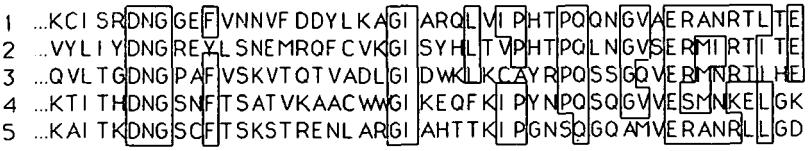


Fig.8: Alignment of the putative endonucleases of 1731 copia (6), MoMLV (46), HIV/ARV2 (36), RSV (38).
1- 1731-ORF2; position of the first amino acid in the ORF2: 263.
2- copia; position of the first amino acid in the appropriate domain: 545.
3- MoMLV; position of the first amino acid in the appropriate domain: 970.
4- HIV/ARV2; position of the first amino acid in the appropriate domain: 838.
5- RSV; position of the first amino acid in the appropriate domain: 688.

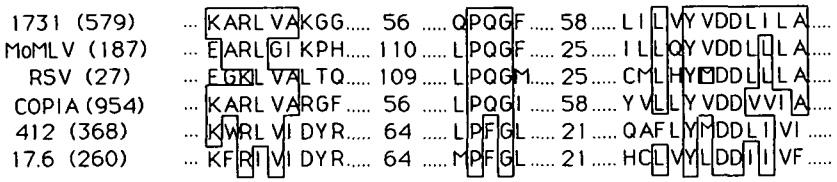


Fig.9: Alignment of the putative reverse transcriptases of 1731, MoMLV (46), RSV (38), copia (6), 412 (7) and 17.6 (9). Numbers in parentheses indicate the position of the first amino acid in the appropriate ORF.

residues, which is in agreement with the length of corresponding retroviral polypeptides previously described.

Retroviral reverse transcriptases are encoded by a portion of the pol gene about 240 residues long and show two highly conserved sequences among them (39). 1731 also contains these two conserved sequences (figure 9) at positions 645 and 711 of ORF2. This suggests that the carboxyl-terminal portion of the 1731 ORF2 encodes a reverse transcriptase-like activity. Moreover, the beginning of this domain could be located in position 579 by alignment of the Lys-Ala-Arg-Leu-Val-Ala stretch with the amino-terminal portion of retroviral reverse transcriptase, as shown in the figure 9.

It has been reported that the pol gene of retroviruses also encompasses a ribonuclease H (RNAse H) activity. Alignment between the RNAse H of some retroviruses and the RNAse H of E. coli showed that the RNAse H activity is associated with the reverse transcriptase domain (39). Computer analysis using similar techniques was undertaken for the entire 1731 sequence. There was weak homology (18%) between 1731 and the E.coli RNAse H from position 420 to 530 of the ORF2. This alignment is located between the endonuclease and reverse transcriptase domains previously described, without overlapping any of these two regions. These two points argue for the existence of a RNAse H domain in 1731.

Absence of env gene

No ORF able to encode for the totality or a part of an env gene was detected in the 1731 internal nucleotide sequence.

Differences and similarities between 1731 and copia; codon usage

The nucleotide sequence of copia (6) shows the existence of one large ORF which presents similarities to the gag gene of retroviruses in its amino-terminal region, and to the pol gene of retroviruses in its carboxyl-terminal region. Comparison between 1731 and copia was performed by various techniques including the DIAGON program of Staden (27). Amino acids similarities between 1731-ORF1 and the gag region of copia reach 24% and extend over the entire gag gene of 1731 (not shown). A higher score was obtained when 1731-ORF2 was compared to the pol region of copia. The matrix obtained (figure 10) reveals two large similar regions which correspond to the protease and endonuclease domain (31% homology) and to the reverse transcriptase domain (36% homology) respectively. Apart from these

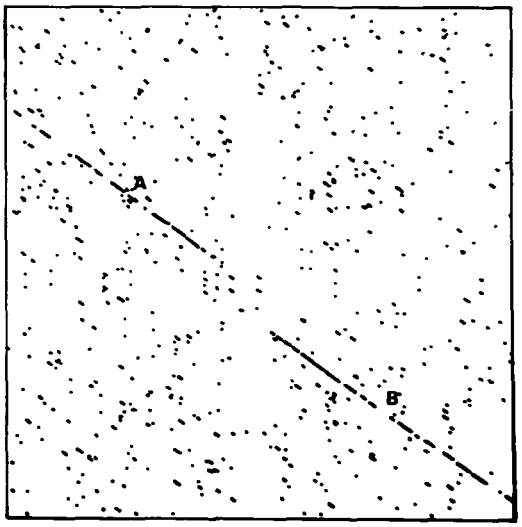


Fig.10: Alignment of 1731 and copia (6) using the DIAGON program of Staden (27). Abscissa: 1731-ORF2; ordinate: copia-ORF. The first alignment (A) corresponds to the protease and endonuclease domain. The second alignment (B) corresponds to the reverse transcriptase domain.

amino acid homologies, no significant similarities were observed at the nucleic acid level. We checked codon usage in copia and 1731 and showed that these two retrotransposons are very different with regard to the choice of codon usage. For example, CCT was used in 16 out of the 39 proline codons in copia and in only 4 out of the 38 proline codons in 1731; conversely, AAG was used in 60 out of the 92 lysine codons in 1731 and in only 43 out of the 138 lysine codons in copia. In the reverse transcriptase domain previously described, 96 of the 157 identical amino acids (except the methionine and tryptophan residues) are encoded by different codons.

Moreover, codon usage in copia and 1731 differs from codon usage in Drosophila as revealed by examining a Drosophila sub-bank of GENBANK. Lastly, we showed that codon usage in 1731 and copia also differs from codon usage in E.coli, yeast, rat and man.

DISCUSSION

We report here the complete nucleotide sequence of 1731, a retrotransposon of Drosophila (23). As it has been previously described for integrated viral sequences (41), 1731 shows a base composition (GC: 35.2%) very close to the base composition of the region where it is inserted (GC: 39.5%).

The overall structure of 1731 is reminiscent of vertebrate proviral structure. 1731 is flanked by 336 bp LTRs and presents two long open reading frames (ORF1 and ORF2) respectively corresponding

to the gag and pol genes of retroviruses. The putative protein encoded by the gag gene of 1731 shows similarities with the gag protein of HIV/ARV2 (36). Moreover, the ORF2 would present three specific domains respectively assigned to protease, endonuclease (integrase) and reverse transcriptase activities. This can be deduced from sequence homologies noticeable in the inner part as well as at the limits of each domain. In addition, the size of these domains is similar to their retroviral counterparts. The first domain, of about 140 residues, would encode a protease activity involved in cleaving a polyprotein into smaller functional proteins. The second domain, corresponding to the endonuclease (integrase), covers about 270 residues and presents short nucleotide stretches that are conserved among retroviral endonuclease genes. It should be noted that the amino-terminal portion of this domain has a zinc-binding site also found in the endonuclease domain of retroviruses (39) and which contains a pair of histidines at the amino-terminal end and a pair of cysteines at the carboxyl-terminal end. Zinc-finger structures, with two cysteines at the amino-terminal end and two histidines at the carboxyl-terminal end, have been described as protein motifs implicated in nucleic acid recognition (42). Thus, it is possible that "zinc-fingers" of retroviral and retrotransposon endonucleases are also involved in a preliminary step of the integration process. Lastly, we assign a RNA-dependent DNA polymerase activity to the third domain of the 1731 ORF2. Similarities with retroviral reverse transcriptases extend over the whole domain and include the highly conserved sequence among retroviral reverse transcriptases: Tyr-hydrophobic residue-Asp-Asp. This sequence has also been shown to be conserved among hepadnavirus polymerases (43) and has been proposed to play a critical role in polymerase function (44). Moreover, it is possible, with regard to amino acid similarities, length and localization, that 1731 also encompasses a RNAse H activity, as is the case for retroviruses (39). Taken together, these results allow us to define the ORF2 as the pol gene of 1731. The organisation of this gene (5'-protease-endonuclease-reverse transcriptase-3') is reminiscent of those of copia and Ty, although the order in retroviral pol gene is 5'-protease-reverse transcriptase-endonuclease-3', as is also the case for the retrotransposons 17.6 and 297.

Moreover, we demonstrate the existence of homologies between 1731 and copia in the protease, endonuclease and reverse transcriptase domains. These homologies are found despite a different codon usage in the two retrotransposons. For the time being, we cannot determine whether similarities between 1731 and copia reflect evolution from a common ancestor or convergence towards a similar function. Neverthless, these results reinforce the likelihood that these enzymatic functions exist in 1731.

Finally, the analysis of 1731 sequence shows the absence of an env gene. Although this is also the case for copia (6) and Ty (10),the absence of env gene is not a common characteristic of retrotransposons. The retrotransposons 17.6 and 297 of Drosophila (8-9) show a large 3' ORF described as an env-like gene. However, no similarities between these large ORFs and the conserved sequence of retroviral env genes were detected. Whether the product of this env-like gene is functional remains a matter of

speculation. The nucleotide sequences of some mouse IAP elements show a region of about 1100 nucleotides located upstream to the 3'LTR (11). This region, long enough to encode an env gene, shows multiple stop codons and no homology to env genes. Neverthless, the nucleotide sequences of other mouse IAP elements show a 3' ORF sharing homology with the env region of several retroviruses such as MMTV and MAV (Myeloblastosis Associated Virus) (45).

The overall organisation of 1731, the existence in Drosophila cells of 1731 extrachromosomic DNA circles (23) and the presence of virus-like particles containing the 1731 RNA genome (J.L.Becker, personal communication) recall the characteristic features of retroviruses. The main difference between 1731 and retroviruses would be the absence of an env gene in 1731, implying the absence of an extracellular phase during the 1731 life cycle. Moreover, the existence of 1731 transcripts and 1731 DNA circles indicates that 1731 might transpose by a mechanism analogous to that described for Ty (16). The decrease of 1731 transcripts in the presence of ecdysterone could provide a negative control of 1731 spread out.

ACKNOWLEDGEMENTS

The authors wish to thank Dr.E.Mac Intyre for careful reading of the manuscript and Dr.C.Mugnier for assistance in computer analysis. This work was supported by grants from CNRS to F.G. and Association pour la Recherche sur le Cancer to M.B.B.

REFERENCES

1- Finnegan,D.J. and Fawcett,D.H. (1986) in *Oxford surveys in Eukaryotic genes*, **3**, 1-62
2- Cameron,J.R., Loh,E.Y. and Davis,R.W. (1979) *Cell*, **6**, 739-751
3- Lueders,K.K and Kuff,E.L. (1977) *Cell*, **12**, 963-972
4- Paulson,K.E., Deka,N., Schmid,C.W., Misra,R., Schindler,C.W., Rush,M.G., Kadyk,L. and Leinwand,L. (1985) *Nature*, **316**, 359-361
5- Varmus,H.E. (1982) *Science*, **216**, 812-820
6- Mount,S.M. and Rubin,G.M. (1985) *Mol.Cel.Bio.*, **5**, 1630-1638
7- Yuki,S., Inouye,S., Ishimaru,S. and Saigo,K. (1986) *Eur.J.Bioch.*, **158**, 403-410
8- Saigo,K., Kugimiya,W., Matsuo,Y., Inouye,S., Yoshioka,K. and Yuki,S. (1984) *Nature*, **312**, 659-661
9- Inouye,S., Yuki,S. and Saigo,K. (1986) *Eur.J.Bioch.*, **154**, 417-425
10- Clare,J. and Farabaugh,P. (1985) *Proc.Natl.Acad.Science USA*, **82**, 2829-2833
11- Mietz,J.A., Grossman,Z., Lueders,K.K. and Kuff,E.L. (1987) *J.of virology*, **61**, 3021-3029
12- Shiba,T. and Saigo,K. (1983) *Nature*, **302**, 119-124
13- Garfinkel,D.J., Boeke,J.D. and Fink,G.R. (1985) *Cell*, **42**, 507-517
14- Flavell,A.J. and Ish-Horowicz,D. (1983) *Cell*, **34**, 415-419
15- Junakovic,N. and Ballario,P. (1984) *Plasmid*, **11**, 109-115
16 -Boeke,J.D., Garfinkel,D.J., Styles,C., Fink,G.R. (1985), *Cell*, **40**, 491-500
17 -Baltimore,D. (1985) *Cell*, **40**, 481-482
18- Shapiro,J.A. and Cordell,B. (1982) *Biol.Cell.*, **43**, 31-54
19 -Bingham,P.M. and Judd, B.H. (1981) *Cell*, **25**, 705-711
20- Levis,R., O'Hare,K. and Rubin,G.M. (1984) *Cell*, **38**, 471-481
21- Degols,G., Jauniaux,J.C. and Wiame,J.M. (1987) *Eur.J.Bioch.*, **165**, 289-296

22- Errede,B., Company,M. and Hutchinson III,C.A. (1987) *Mol.Cel.Biol.*, **7**, 258-265
23- Peronnet,F., Becker,J.L., Becker,J., d'Auriol,L., Galibert,F., and Best-Belpomme,M.(1986) *Nucleic Acids Res.*, **14**, 9017-9033
24- Sanger,F., Nicklens,S., and Coulson,A.R. (1977) *Proc.Natl.Acad.Sci. USA*, **74**, 5463-5467
25- Goad,W. and Kanehisha,M.I. (1982) *Nucleic Acids Res.*, **10**, 247-264
26- Dayhoff,M.O., Schwartz,R.M., and Orcutt,B.C. (1979) in *Atlas of Protein Sequence and Structure*, Vol.5, suppl.3, pp.545-562 (M.O.Dayhoff, ed.) National Biomedical Research Foundation, Washington,D.C.
27- Staden,R. (1982) *Nucleic Acids Res.*, **10**, 2951-2961
28- Hu,N., and Messing,J. (1982) *Gene*, **17**, 271-277
29- Dunsmuir,P., Brorein,W.J.Jr, Simon,M.A., and Rubin,G.M. (1980) *Cell*, **21**, 575-579
30- Majors,J.E., and Varmus,H.E. (1981) *Nature*, **289**, 253-257
31- Cavener,D.R. (1987) *Nucleic Acids Res.*, **15**, 1353-1361
32- Jacks,T., Power,M.D., Masiarz,F.R.,Luciw,P.A., Barr,P.J. and Varmus,H.E. (1987) *Nature*, **331**, 280-283
33- Jacks,T., Townsley,K. and Varmus,H.E. (1987) *Proc.Natl.Acad.Sci.USA*, **82**, 1618-1622
34- Sharp,S., Defranco,D., Silberklang,M., Hosbach,H.A., Schmidt,T., Kubli,E., Gergen,J.P., Wensink,P.C. and Soell,D. (1981) *Nucleic Acids Res.*, **9**, 5867-5882
35- Kikuchi,Y., Ando,Y. and Shiba,T. (1986) *Nature*, **323**, 824-826
36- Sanchez-Pescador,R., Power,M.D., Barr,P.J., Steimer,K.S., Stempien,M.M., Brown-Shimer,S.L., Gee,W.W., Renard,A., Randolph,A., Levy,J.A., Dina,D. and Luciw,P.A. (1985) *Science*, **227**, 484-492
37- Seiki,M., Seisuke,Y., Hirayama,Y. and Yoshida,M. (1983) *Proc.Natl.Acad.Science USA*, **80**, 3618-3622
38- Schwartz,D.E., Tizard,D.R. and Gilbert,W. (1983) *Cell*, **32**, 853-869
39- Johnson,M.S., McClure,M.A., Feng,D.F., Gray,J., and Doolittle,R.F. (1986) *Proc.Natl.Acad.Sci.USA*, **83**, 7648-7652
40- Yuki,S., Ishimaru,S., Inouye,S. and Saigo,K. (1986) *Nucleic Acids Res.*, **14**, 3017-3030
41- Bernardi,G., Olofsson,B., Filipski,J, Zerial,M., Salinas,J., Cuny,G., Meunier-Rotival,M. and Rodier,F. (1985) *Science*, **228**, 953-958
42- Miller,J., Mc Lachlan,A.D. and Klug,A. (1985) *Embo J.*, **4**, 1609-1614
43- Mandart,E., Kay,A. and Galibert,F. (1984) *J.of Virology*, **49**, 782-792
44- Zavriev,S.K. and Borisova,O.V. (1987) *Molecular Biol.*, **21**, 192-202
45- Aota, S.I., Gojobori, T., Shigesada,K., Ozeki,O. and Ikemura,T. (1987) *Gene*, **56**,1-12
46- Shinnick,T.M., Lerner,R.A. and Sutcliffe,J.G. (1981), *Nature*, **293**, 543-548