The EMBO Journal vol.4 no.5 pp.1267-1272, 1985

Close structural resemblance between putative polymerase of a *Drosophila* transposable genetic element 17.6 and *pol* gene product of Moloney murine leukaemia virus

Hiroyuki Toh, Reiko Kikuno, Hidenori Hayashida, Takashi Miyata, Wataru Kugimiya[1,2], Satoshi Inouye[1], Shunji Yuki[1] and Kaoru Saigo[1]

Department of Biology, Faculty of Science, Kyushu University, and [1]Department of Biochemistry, Kyushu University School of Medicine, Fukuoka 812, Japan

[2]Present address: Fuji Oil Co. Ltd., Izumi-Sano, Osaka 596, Japan

Communicated by M.L.Birnstiel

We have made a computer-assisted search for homology among polymerases or putative polymerases of various viruses and a transposable element, the *Drosophila copia*-like element 17.6. The search revealed that the putative polymerase (second open reading frame) of the *copia*-like element 17.6 bears close resemblance in overall structural organization to the *pol* gene product of Moloney murine leukaemia virus (M-MuLV): they show significant homology to each other at both the N- and C-terminal portions, suggesting that the 17.6 putative polymerase carries two enzymatic activities, related to reverse transcriptase and DNA endonuclease. The putative polymerase of cauliflower mosaic virus (CaMV) shows striking homology with the putative polymerase of 17.6 over almost its entire length, but it lacks the DNA endonuclease-related sequence. Furthermore, it was shown that the N-terminal ends of the M-MuLV *pol* product and the CaMV and 17.6 putative polymerases exhibit strong sequence homology with the *gag*-specific protease (p15) of Rous sarcoma virus (RSV) as well as the amino acid sequence predicted from the *gag/pol* spacer sequence of human adult T-cell leukaemia virus (HTLV). These p15-related sequences contain a highly conserved stretch of amino acids which show a close similarity with sequences around the active site amino acids Asp-Thr-Gly of the acid protease family, suggesting that they have an activity similar to acid protease. On the basis of the alignment of reverse transcriptase-related sequences, a dendrogram representing phylogenetic relationships among all the viruses compared together with 17.6 was constructed and its evolutionary implication is discussed.

Key words: Drosophila/copia-like element 17.6/*pol* gene/Moloney murine leukaemia virus/homology

Introduction

The *pol* gene of retroviruses codes for at least three enzymes with different activities in a single transcription unit: DNA polymerase and RNase H of reverse transcriptase located on the N-terminal region of the gene product and DNA endonuclease located on the C-terminal region (Varmus and Swanstrom, 1982; Dickson *et al.*, 1982). Comparison of amino acid sequences of the *pol* gene products between such distantly related viruses as Rous sarcoma virus (RSV), Moloney murine leukaemia virus (M-MuLV) and human adult T-cell leukaemia virus (HTLV) revealed strong homology at both the N- and C-terminal regions (Toh *et al.*, 1983). The N-terminal highly conserved region shows obvious homology in amino acid sequence with putative polymerase of hepatitis B virus (HBV) and cauliflower mosaic virus (CaMV) (Toh *et al.*, 1983), known to be double-stranded DNA viruses that include a reverse transcription step in their life cycle (Summers and Mason, 1982; Pfeiffer and Hohn, 1983; Hull and Covey, 1983; Guilley *et al.*, 1983; Varmus, 1983 for review). Furthermore it has recently been shown that a *Drosophila* transposable element 17.6, a member of the groups of genetic elements including *copia*, whose particle involves a reverse transcriptase activity (Shiba and Saigo, 1983; Flavell, 1984), contains a conserved region in the putative polymerase gene that shows apparent homology in sequence with the N-terminal highly conserved region of the *pol* gene product (Saigo *et al.*, 1984). This evidence strongly suggests that the conserved sequence shared in common among these viruses and an insect transposable element is a functionally important constituent of reverse transcriptase (Toh *et al.*, 1983; Sigo *et al.*, 1984). On the other hand, no evidence for the presence of homology across such wide evolutionary distance has so far been reported for the C-terminal region carrying DNA endonuclease activity that is thought to play an important role in the provirus integration into host DNA (Varmus and Swanstrom, 1982). It may be of particular importance to know whether or not transposable genetic elements encode DNA endonuclease.

We report here that the putative polymerase of the *copia*-like element 17.6 contains a segment that shows marked sequence homology to the C-terminal region of the *pol* gene product, suggesting the involvement of DNA endonuclease. In contrast, no such sequences exist in the CaMV and HBV DNAs. Furthermore, near the N-terminal end of the putative polymerases of 17.6 and CaMV and the polymerase of M-MuLV, and also within the spacer region between *gag* and *pol* genes of HTLV, we found sequences homologous to that of *gag*-specific protease (p15) (Dickson *et al.*, 1982) of RSV. These p15-related sequences contain highly conserved amino acids similar to those around active sites of acid proteases. Thus the putative polymerase of 17.6 bears close resemblance in overall structural organization to the *pol* gene product of M-MuLV. We also discuss the evolution of reverse transcriptase-related sequences.

Results and Discussion

Homology of the 17.6 putative polymerase with the C-terminal region of retroviral pol gene products

Homology matrix comparison of the amino acid sequence encoded for by the second open reading frame (ORF 2 or putative polymerase) of the *copia*-like element 17.6 with that encoded for the *pol* gene of M-MuLV (Shinnick *et al.*, 1981) revealed the presence of two regions of significant homology at the N- and C-terminal regions of these proteins (Figure 1a). The two regions correspond to the highly conserved regions of retroviral *pol* gene products (Toh *et al.*, 1983) and the ORF 2 product of *copia*-like elements (Saigo *et al.*, 1984). When the 17.6 sequence was compared with RSV and HTLV, apparent homologies were also detected at the same regions, although less extensive (data not shown). A statistical test showed that the observed homologies

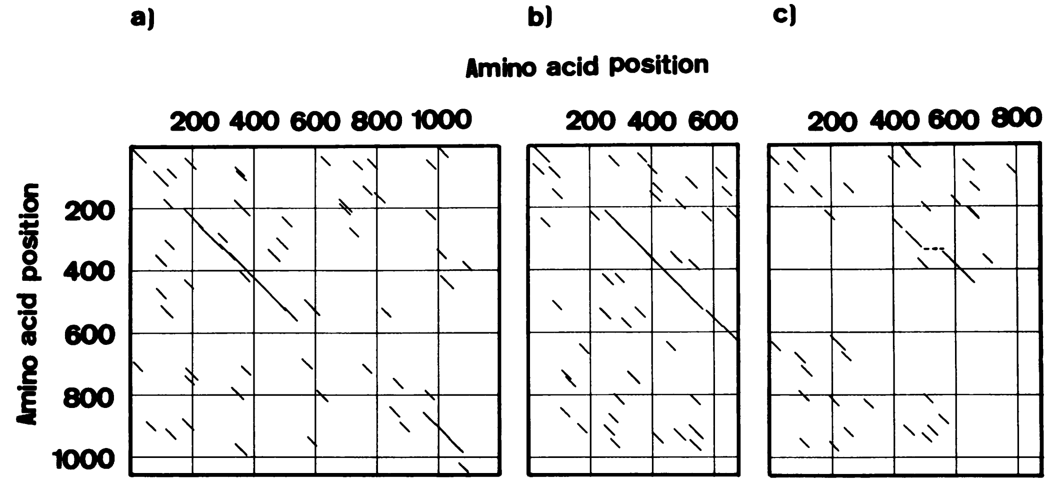

Fig. 1. Homology matrix comparisons of the amino acid sequence encoded by the ORF 2 of *copia*-like element 17.6 (ordinate) with those encoded by (**a**), *pol* of M-MuLV (abscissa), (**b**), ORF 5 of CaMV and (**c**), ORF 6 of WHV. A computer program (Toh *et al.*, 1983) was used to generate diagnonal lines indicating segments of 30 residues that show homology with a probability of occurrence by chance of $<4 \times 10^{-4}$. A horizontal dotted line in (**c**) indicates an insertion presented in WHV.

Table I. Amino acid sequence comparisons of the 17.6 putative polymerase with the M-MuLV *pol* gene product and the CaMV and WHV putative polymerases

	Region[a]	Positions[b]	% Homology	Probability[c]
(i) 17.6 *versus* M-MuLV	I	1–109	23	3.2×10^{-6}
	II	1–325	28	$<10^{-9}$
	III	1–128	25	5.3×10^{-7}
	III	191–229	36	1.2×10^{-4}
	IV	1–200	23	3.7×10^{-7}
(ii) 17.6 *versus* CaMV	I	1–109	19	2.2×10^{-4}
	II	1–325	35	$<10^{-9}$
	III	1–187	33	$<10^{-9}$
(iii) 17.6 *versus* WHV	II	28–228	16	1.3×10^{-6}

[a] The region defined in Figure 1.
[b] Amino acid positions based on the alignment shown in Figure 3.
[c] The probability that the observed sequence similarity is realized by chance, which was evaluated as described in Materials and methods.

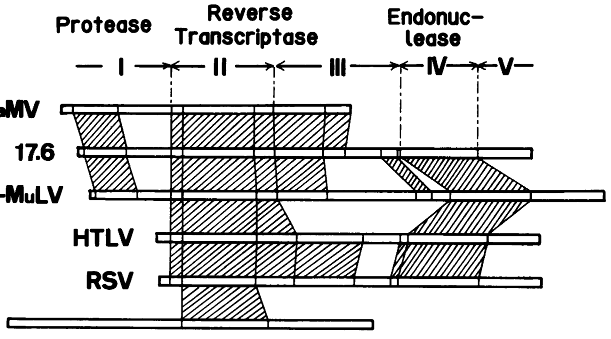

Fig. 2. Homology profiles along the length of the polymerases or putative polymerases. Homologous regions are indicated by diagonal lines. These sequences are tentatively divided into five regions, on the basis of the positions of highly conserved regions among the members (regions II and IV). Note that CaMV and WHV contain no sequences homologous to region IV. A proposed assignment of functional domains is also shown (see text): the N-terminal portion of the region I, protease; the region II and the N-terminal half of the region III, reverse transcriptase; the region IV, endonuclease.

are highly significant (Table I). Since the N- and C-terminal regions of the retroviral *pol* gene product are known to carry reverse transcriptase and DNA endonuclease activities, respectively (Varmus and Swanstrom, 1982; Dickson *et al.*, 1982), these results strongly suggest the presence of two enzymes related to reverse transcriptase and DNA endonuclease in the *copia*-like element 17.6. Saigo *et al.* (1984) have detected homology at the N-terminal region of the 17.6 sequence and retroviral sequences and suggested the presence of reverse transcriptase in the 17.6. On the other hand, homology was limited only to the N-terminal region, when the ORF 2 product of 17.6 was compared with the ORF 5 product (putative polymerase) of CaMV (CM 1841 strain) and the ORF 6 product (putative polymerase) of woodchuck hepatitis virus (WHV), a member of the hepatitis B virus group (Figure 1b,c). The observed homologies of the N-terminal region are statistically significant (Table I). Further comparisons of amino acid sequences between the C-terminal region of 17.6 and all the ORF products of CaMV and WHV revealed no such obvious homology, suggesting that a DNA endonuclease similar to that of retroviruses is not encoded for by the CaMV and WHV (HBV) DNAs. This evidence may have an important implication for different replication strategies of the RNA and DNA viruses that undergo reverse transcription: for the replication of the viral genome, the former activity is required to integrate the copied DNA into the host genome, whereas the latter is not.

The polymerase and putative polymerase sequences of retroviruses, CaMV, HBV and *copia*-like element 17.6 were aligned for each of the four blocks (I–IV) which were tentatively defined on the basis of the positions of conserved regions II and IV (Figures 2 and 3) [comparison of putative polymerase sequences between 17.6 and WHV revealed a new conserved region (positions 28–125 in the alignment of region II) shared among all the members, which were not detected in our previous analysis (Toh *et al.*, 1983)]. The alignments revealed several invariant positions and highly conserved clusters of amino acids (positions 14–18 in region I; 45–50, 87–99, 137–148, 176–182 and 223–228 in II; 101–105, 133–150 and 174–180 in IV), which may play critical roles in function. The extensive homology found between the 17.6 and CaMV sequences is striking, despite the lack of taxonomic relatedness.

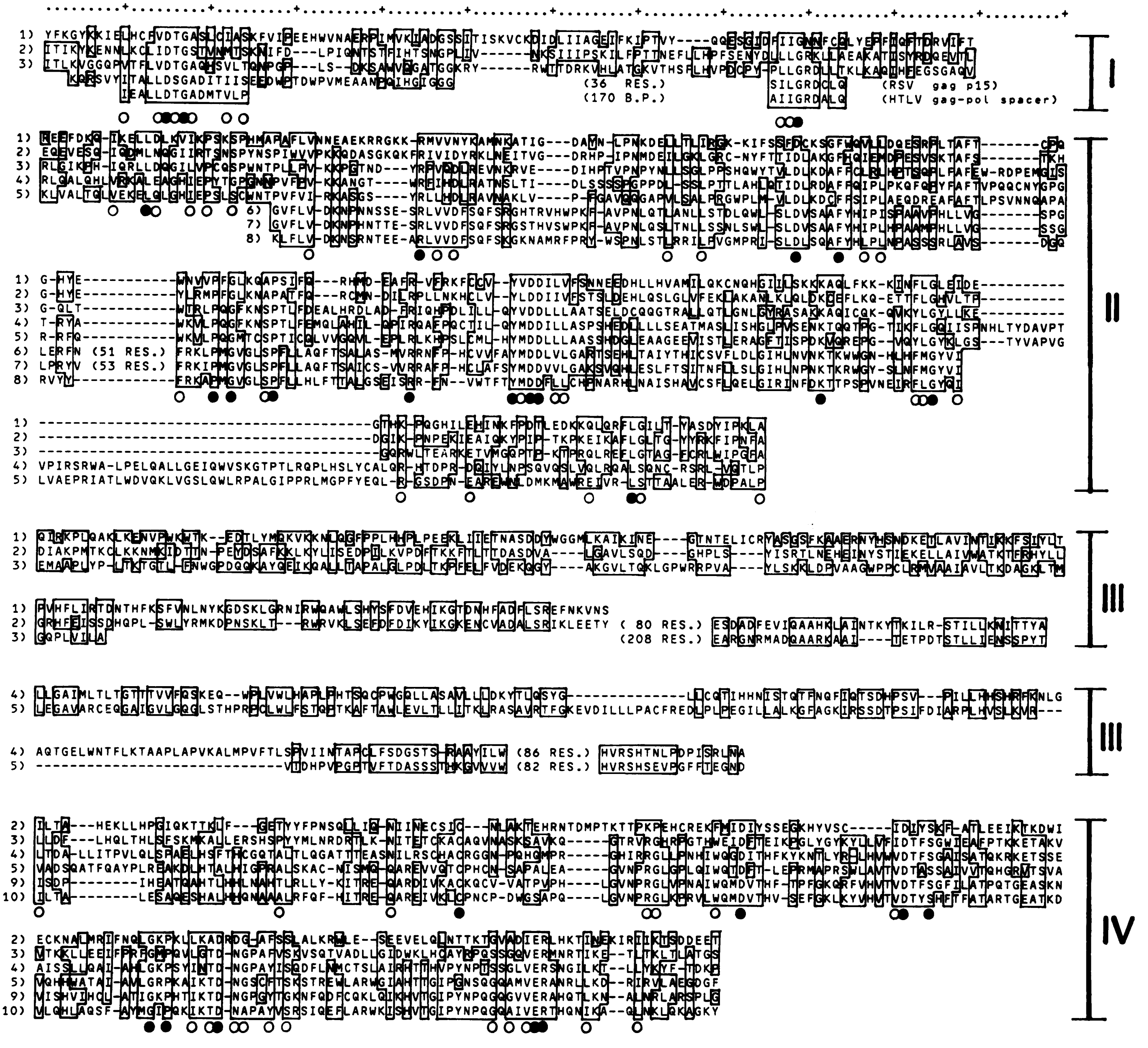

Fig. 3. Alignments of amino acid sequences of the polymerases and putative polymerase. **(1)**, CaMV (CM 1841 strain); **(2)**, *copia*-like element 17.6; **(3)**, M-MuLV; **(4)**, HTLV; **(5)**, RSV; **(6)**, HBV; **(7)**, WHV; **(8)**, DHBV (duck hepatitis B virus); **(9)**, MMTV (mouse mammary tumor virus); **(10)**, SMRV (squirrel monkey retrovirus). The alignments of amino acid sequences of the regions indicated by diagonal lines in Figure 2 were shown. The aligned regions of 17.6 correspond to nucleotide positions (Figure 1 of Saigo *et al.*, 1984) 2390–2689 for region I, 3002–3724 for II, 3725–4234 and 4475–4588 for III and 4604–5146 for IV. In region I, the amino acid sequences of CaMV, 17.6 and M-MuLV exhibit apparent homologies with that of the *gag*-specific protease p15 of RSV in part. Thus the alignment of these sequences was shown. A statistical test showed that the observed homology between the sequences of RSV p15 and 17.6 putative polymerase is highly significant, with the probability of occurrence by chance of 3.2 x 10^{-6}. The amino acid sequences predicted from the different ORFs of *gag*/*pol* spacer sequence of HTLV show marked homology with the p15 sequence in part (both reading frames differ from those of *gag* and *pol*). Thus these sequences were included in the alignment. Gaps (–) were inserted to increase sequence similarity. Positions at which more than half of the aligned sequences share identical or chemically similar amino acids were boxed. ● and ○, positions that are occupied by identical and chemically similar amino acids among all the aligned sequences, respectively. Chemically similar amino acids are defined as pairs of residues, both of which belong to the same group (Schwartz and Dayhoff, 1978), the groups being as follows: A, S, T, P and G; N, D, E and Q; H, R and K; M, L, I and V; F, Y and W.

The region II mentioned above contains two blocks (positions 28–110 and 137–228 in the alignment) that share homology in common among all viruses and 17.6. A highly conserved amino acid sequence Tyr-hydrophobic residues-Asp-Asp flanked by three hydrophobic residues (positions 176–182) presents in the C-terminal block. A similar amino acid sequence, Tyr-Gly-Asp-Asp, flanked by hydrophobic residues also exists in the polymerases of the picornavirus group (Kamer and Argos, 1984) as well as bacteriophages (Kamer and Argos, 1984; Inokuchi and Hiroshima, personal communication). Such widespread occurrence of the sequence Tyr-X-Asp-Asp flanked by hydrophobic residues in various viral-coded polymerases suggests that these amino acids play a critical role in polymerase function (Kamer and Argos, 1984). It is therefore likely that the amino acid se-

Viral protease	
CaMV ORF5	LHCFV-DTGASLCIAS
17.6 ORF2	LKCLI-DTGSTVNMTS
M-MuLV pol	VTFLV-DTGAQHSVLT
RSV gag p15	ITALL-DSGADITIIS
HTLV (gag-pol)	IEALL-DTGADMTVLP
Acid protease (C-terminal)	*
pepsinogen (human)	CQAIV-DTGTSLLTGP
(porcine)	CQAIV-DTGTSLLTGP
prochymosin (bovine)	CQAIL-DTGTSKLVGP
penicillopepsin	FSGIA-DTGTTLLLLB
renin (mouse)	CEVVV-DTGSSFISAP
(human)	CLALV-DTGASYISGS
Acid protease (N-terminal)	*
pepsinogen (human)	DFTVVFDTGSSNLWVP
(bovine)	DFTVIFDTGSSNLWVP
(porcine)	DFTVIFDTGSSNLWVP
prochymosin (bovine)	EFTVLFDTGSSDFWVP
penicillopepsin	TLNLNFDTGSADLWVF
renin (mouse)	TFKVMFDTGSANLWVP
(human)	TFKVVFDTGSSNVWVP

Fig. 4. Comparison of amino acid sequences of highly conserved segments of viral-coded proteases with those around the active sites of acid proteases. Since pepsin-related acid proteases consist of two topologically similar domains and two active site aspartic acids (Hsu *et al.*, 1977), the amino acid sequences around the two aspartic acids were aligned between the N- and C-terminal halves. Bold-faced letters indicate amino acids that are identical or chemically similar (see Figure 3 legend) among 16 sequences (90%) out of 18 at each position. Gaps (−) were inserted to increase sequence similarity. The aspartic acids at the active sites of acid proteases are marked with asterisks. References for sequence data used: human (Sogawa *et al.*, 1983), porcine (Tang *et al.*, 1973) and bovine (Harboe and Foltmann, 1975) pepsinogen; bovine prochymosin (Foltmann *et al.*, 1977); penicillopepsin (Hsu *et al.*, 1977); renin from mouse (Panthier *et al.*, 1982) and human (Hobart *et al.*, 1984).

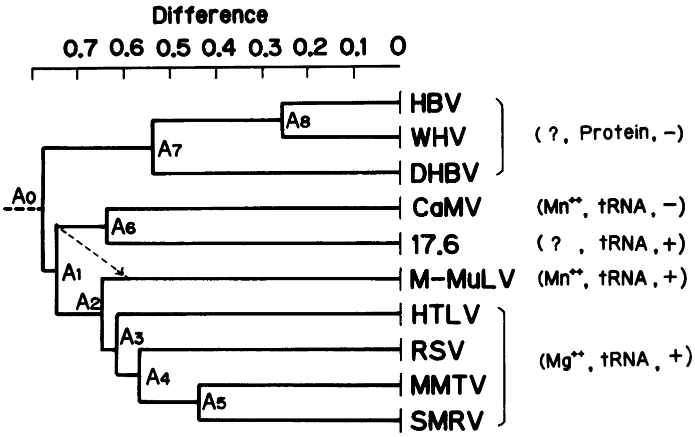

Fig. 5. A dendrogram representing phylogenetic relationships among reverse transcriptase-related sequences. Abscissa represents the average value of sequence differences between the sequence pair connected at each nodal point ($A_1 - A_8$). Identical tree topology was also obtained for the five retroviruses and 17.6, when the data of region IV were used. Since sequence data of region II are not available for MMTV and SMRV, the two lineages leading to these viruses were inferred from a dendogram constructed from region IV as follows: $K_{II}(A_i) = K_{IV}(A_i) \times K_{II}(A_3)/K_{IV}(A_3)$, where i = 4 or 5, or $K_{IV}(A_i)$, for example, is the difference of node A_i in the dendrogram constructed from region IV. Maximum likelihood method (Felstenstein, 1981) reproduced the same tree topology as that presented here, if the HBV group was assumed to be a direct descendant from the most remove ancestor A_0. A slightly different tree topology was obtained when the modified matrix method (Li, 1981) was used: the common ancestor of 17.6 and CaMV (i.e., A_6) was shown to be derived from the lineage leading to M-MuLV (indicated by dotted arrow). The preferred cation for DNA polymerase activity (Chiu *et al.*, 1984; Volovitch *et al.*, 1984), primer molecules required for initiation of DNA synthesis (Hull and Covey, 1983; Guilley *et al.*, 1983; Saigo *et al.*, 1984; Taylor, 1977; Molnar-Kimber *et al.*, 1983), and the presence (+) of DNA endonuclease-related sequence (Varmus and Swanstrom, 1982; Dickson *et al.*, 1982; Toh *et al.*, 1983; Chiu *et al.*, 1983 and the present result) are also shown in parentheses.

quences of the N- and C-terminal blocks carry RNase H and DNA polymerase activities, respectively. This argument is consistent with evidence showing that the sequence carrying RNase H activity is located before that of DNA polymerase (Crouch and Dirksen, 1982). In the reverse transcriptase-containing viruses, the second amino acid position of the Tyr-X-Asp-Asp is occupied by a bulky hydrophobic residue, while it is replaced by a small residue, glycine, in picornaviruses and bacteriophages. Thus it remains possible that this difference may be related to structural difference around active sites between RNA-dependent DNa polymerase and RNA-dependent RNA polymerase.

Homology of the 17.6 putative polymerase with avian retroviral gag-specific protease p15

As shown in Figure 2, the CaMV, 17.6 and M-MuLV trio share a segment of ~90 residues located at the N-terminal end of the polymerase or putative polymerases (region I of Figure 2), but the corresponding segment is not present in the *pol* gene products of HTLV and RSV. It has recently been noted that the N-terminal sequence of the M-MuLV *pol* gene product is homologous to the sequence of avian retroviral *gag*-specific protease p15 (Levin *et al.*, 1984). Comparison of the amino acid sequences between the segments of the N-terminal end unique to the CaMV, 17.6 and M-MuLV trio and the p15 protein of RSV revealed an apparent homology, although less extensive, suggesting that these segments involve protease activity. Thus the arrangement of functional domains of the 17.6 putative polymerase bears close resemblance to that of the M-MuLV *pol* gene product, which is organized as NH_2 - protease - reverse transcriptase (RNase H and DNA polymerase) - DNA endonuclease - COOH. The putative polymerase of CaMV is also similar in organization to that of 17.6 and the *pol* gene product of M-MuLV, but lacks a sequence corresponding to the endonuclease (Figure 2).

Comparisons of all the amino acid sequences predicted from the three open reading frames of *gag/pol* spacer sequence of HTLV with the amino acid sequence of p15 revealed the presence of two segments of marked homology (Figure 3), although no long open reading frames exist in this region due to the appearance of several termination codons. Sagata *et al.* (1984) also detected these homologous segments in the same region of HTLV. These p15-related sequences include a stretch of amino acids-(hydrophobic residue)$_2$-Asp-Thr(or Ser)-Gly-Ala(or Ser)-, which is conserved among all the sequences compared. The amino acid sequence of this hexapeptide was subjected to a computer-assisted search for homology with known sequences compiled in our data base and was found to be homologous to sequences of acid proteases around their active sites (Figure 4). The strong conservation of three amino acids Asp-Thr-Gly is remarkable. This result suggests that the p15-related sequences carry an enzymatic activity similar to acid proteases and the conserved hexapeptide forms an active center of these enzymes.

Evolution of reverse transcriptase-related sequences

Using the aligned amino acid sequences of the highly conserved region of polymerases and putative polymerases (positions 28−228 in region II of Figure 3), a dendrogram representing phylogenetic relationships among the reverse transcriptase-related sequences from retroviruses, members of the HBV group, CaMV and *copia*-like element 17.6 was constructed by a simple clustering method (Sokal and Sneath, 1963) (Figure 5). The deduced phylogenetic tree revealed a remote divergence of the HBV group and the other members. This result is consistent with evidence that DNA synthesis is initiated with a tRNA primer in retroviruses (Taylor, 1977) and possibly in CaMV (Hull and Covey, 1983; Guilley *et al.*, 1983) and 17.6 (Saigo *et al.*, 1984), whereas protein priming was suggested in HBV (Molnar-Kimber *et al.*, 1983). Further analysis by more elaborate methods (Felsenstein, 1981; Li, 1981) for constructing phylogenetic trees reproduced an identical tree topology, except for a slight difference in the branching order of M-MuLV and CaMV-17.6 pair (dotted line) in one of the methods (see Figure 5 legend).

The phylogenetic tree suggests Mn^{2+} preference for the polymerase activity of 17.6. The phylogenetic position of CaMV is of particular interest: according to the phylogenetic tree, the common ancestral polymerase gene of the M-MuLV-17.6-CaMV trio possibly contained a region coding for DNA endonuclease. Thus it seems likely that the putative polymerase of CaMV was derived from a transposable genetic element homologous to 17.6 or an ancestral M-MuLV. The present-day CaMV may have evolved by integrating a foreign piece of DNA containing genes for the reverse transcriptase and its related enzymes: a whole set of copied DNA of a retroviral provirus or a *copia*-like element might have been integrated into an ancestral CaMV genome which involved no reverse transcriptase activity, followed by deletions of several genes coded for by the copied DNA, and thereby acquired a new function (reverse transcriptase) that had hitherto not existed. Alternatively the whole genome of CaMV may have descended from that of an ancestral retrovirus or *copia*-like element, by acquiring several mutations that germinated termination codons on the encoded multifunctional genes to form a distinct structural organization from the ancestral genome. *Drosophila copia*-like elements are composed of several groups, each having dozens of copies dispersed in the host genome (for review, see Rubin, 1983). Such multi-copied genomes may allow accumulation of a considerable amount of mutations as in the case of duplicated genes, thereby generating diverse functions. Thus movable genetic elements may serve as a resource for the emergence of new viruses containing new functions.

There is growing evidence that viral gene products often exhibit sequence homologies across a wide evolutionary distance. In addition to viruses containing a reverse transcriptase activity, there are examples for this: two plant virus groups, brome (and alfalfa) mosaic virus and tobacco mosaic virus show an obvious homology to each other at a region thought to involve a protein related to RNA replication, as well as to the corresponding region of animal alphaviruses (Haseloff *et al.*, 1984). Sequence homologies were also reported between two non-structural proteins encoded by cowpea mosaic virus and the corresponding proteins encoded by picornavirus (Franssen *et al.*, 1984). The unassigned ORF 1 product of CaMV was shown to be homologous in sequence to the M-protein of cowpea mosaic virus in part (Toh *et al.* in preparation). Toh and Miyata (in preparation) have detected a regional yet significant homology between the terminal protein of type II adenovirus and the core protein of HBV. This result appears to reconcile well with evidence that both viruses are unique in that they use a primer protein in replication (Molnar-Kimber *et al.*, 1983; Rekosh *et al.*, 1977). These observations, together with the results presented here, suggest that during viral evolution a piece of DNA carrying a certain function might be able to travel from virus to virus across wide taxonomic differences.

Materials and methods

Amino acid sequence data

Amino acid sequence data were taken from the following papers: *copia*-like element 17.6, Saigo *et al.* (1984); M-MuLV, Shinnick *et al.* (1981); RSV, Schwartz *et al.* (1983); HTLV, Seiki *et al.* (1983); MMTV and SMRV, Chiu *et al.* (1984); CaMV (CM 1841 strain), Gardner *et al.* (1981); HBV, Ono *et al.* (1983); WHV, Galibert *et al.* (1982); DHBV, Mandart *et al.* (1984).

Homology matrix comparison and sequence alignment

Homology searches between amino acid sequences were made using the graphical matrix method, as described previously (Toh *et al.*, 1983). The computer program was used to generate diagonal lines indicating segments of 30 residues long which show homology with a probability of occurrence by chance of $<4 \times 10^{-4}$. Further detailed information for locating gaps was printed out for each of the locally divergent segments. With the aid of these matrices, the sequences were aligned by manual inspection or by the computer-assisted method of Sankoff (1972).

Statistical test of observed homology

To assess the statistical significance of observed sequence homology, the probability *P* that the homology is realized by chance was evaluated; by generating 100 pairs of randomized sequences each having the same amino acid compositions as those of the real sequences which were aligned, the probability *P* was calculated as described previously (Toh *et al.*, 1983).

Construction of phylogenetic tree among homologous sequences

Sequence differences per residue (*K*) were calculated pairwise between the aligned sequences for a segment (positions 28–228) of region II in Figure 3. On the basis of the *K* values, a dendrogram representing phylogenetic relationships among reverse transcriptase-related sequences was constructed by a simple clustering method (Sokal and Sneath, 1963). The phylogenetic tree of these sequences was also constructed by more elaborate methods: maximum likelihood method (Felstenstein, 1981) and modified matrix method (Li, 1981). Note that, in both methods, no assumption is made for the constant rate of evolution for different lineages.

Acknowledgements

We thank Drs Y.Inokuchi, M.Hasegawa and T.Yasunaga for discussions and Miss K.Mitsuyasu for technical assistance. This work was supported in part by grants from the Ministry of Education, Science and Culture of Japan.

References

Chiu,I.-M., Callahan,R., Tronick,S.R., Schlom,J. and Aaronson,S.A. (1984) *Science (Wash.)*, **223**, 364-370.

Crouch,R.J. and Dirksen,M.-L. (1982) in Linn,S. and Roberts,R. (eds.), *Nucleases*, Cold Spring Harbor Laboratory Press, NY, pp. 211-241.

Dickson,C., Eisenman,R., Fan,H., Hunter,E. and Teich,N. (1982) in Weiss,R., Teich,N., Varmus,H. and Coffin,J. (eds.), *RNA Tumor Viruses, Molecular Biology of Tumor Viruses*, ed. 2, Cold Spring Harbor Laboratory Press, NY, pp. 513-648.

Felstenstein,J. (1981) *J. Mol. Evol.*, **17**, 368-376.

Flavell,A.J. (1984) *Nature*, **310**, 514-516.

Foltmann,B., Pedersen,V.B., Jacobsen,H., Kauffman,D. and Wybrandt,G. (1977) *Proc. Natl. Acad. Sci. USA*, **74**, 2321-2324.

Franssen,H., Leunissen,J., Goldbach,R., Lomonossoff,G. and Zimmern,D. (1984) *EMBO J.*, **3**, 855-861.

Galibert,F., Nan Chen,T. and Mandart,E. (1982) *J. Virol.*, **41**, 51-65.

Gardner,R.C., Howarth,A.J., Hahn,P., Brown-Luedi,M., Shepherd,R.J. and Messing,J. (1981) *Nucleic Acids Res.*, **9**, 2871-2888.

Guilley,H., Richards,K.E. and Jonard,G. (1983) *EMBO J.*, **2**, 277-282.

Harboe,M.K. and Foltmann,B. (1975) *FEBS Lett.*, **60**, 133-136.

Haseloff,J., Goelet,P., Zimmern,D., Ahlquist,P., Dasgupta,R. and Kaesberg,P. (1984) *Proc. Natl. Acad. Sci. USA*, **81**, 4358-4362.

Hobart,P.M., Fogliano,M., O'Connor,B.A., Shaefer,I.M. and Chirgwin,J.M. (1984) *Proc. Natl. Acad. Sci. USA*, **81**, 5026-5030.

Hsu,I.-N., Delbaere,L.T.J., James,M.N.G. and Hofmann,T. (1977) *Nature*, **266**, 140-145.

Hull,R. and Covey,S.N. (1983) *Trends Biochem. Sci.*, **8**, 119-121.

Kamer,G. and Argos,P. (1984) *Nucleic Acids Res.*, **12**, 7269-7282.

Levin,J.G., Hu,S.C., Rein,A., Messer,L.I. and Gerwin,B.I. (1984) *J. Virol.*, **51**, 470-478.

Li,W.-H. (1981) *Proc. Natl. Acad. Sci. USA*, **78**, 1085-1089.

Mandart,E., Kay,A. and Galibert,F. (1984) *J. Virol.*, **49**, 782-792.

Molnar-Kimber,K.L., Summers,J., Taylor,J.M. and Mason,W.S. (1983) *J. Virol.*, **45**, 165-172.

Ono,Y., Onda,H., Sasada,R., Igarashi,K., Sugino,Y. and Nishioka,K. (1983) *Nucleic Acids Res.*, **11**, 1747-1757.

Panthier,J.-J., Foote,S., Chambraud,B., Strosberg,A.D., Corvol,P. and Rougeon, F. (1982) *Nature*, **298**, 90-92.

Pfeiffer,P. and Hohn,T. (1983) *Cell*, **33**, 781-789.

Rekosh,D.M.K., Russell,W.C., Bellett,A.J.D. and Robinson,A.J. (1977) *Cell*, **11**, 283-295.

Rubin,G.M. (1983) in Shapiro,J.A. (ed.), *Mobile Genetic Elements*, Academic Press, NY/London, pp. 329-361.

Sagata,N., Yasunaga,T. and Ikawa,Y. (1984) *FEBS Lett.*, **178**, 79-82.

Saigo,K., Kugimiya,W., Matsuo,Y., Inouye,S., Yoshioka,K. and Yuki,S. (1984) *Nature*, **312**, 659-661.

Sankoff,D. (1972) *Proc. Natl. Acad. Sci. USA*, **69**, 4-6.

Schwartz,R.M. and Dayhoff,M.O. (1978) in Dayhoff,M.O. (ed.), *Atlas of Protein Sequence and Structure*, Vol. **5**, National Biomedical Research Foundation, Washington, pp. 353-358.

Schwartz,D.E., Tizard,R. and Gilbert,W. (1983) *Cell*, **32**, 853-869.

Seiki,M., Hattori,S., Hirayama,Y. and Yoshida,M. (1983) *Proc. Natl. Acad. Sci. USA*, **80**, 3618-3622.

Shiba,T. and Saigo,K. (1983) *Nature*, **302**, 119-124.

Shinnick,T.M., Lerner,R.A. and Sutcliffe,J.G. (1981) *Nature*, **293**, 543-548.

Sogawa,K., Fujii-Kuriyama,Y., Mizukami,Y., Ichihara,Y. and Takahashi,K. (1983) *J. Biol. Chem.*, **258**, 5306-5311.

Sokal,R.R. and Sneath,P.H. (1963) *Principles of Numerical Taxonomy*, published by Freeman, San Francisco.

Summers,J. and Mason,W.S. (1982) *Cell*, **29**, 403-415.

Tang,J., Sepulveda,P., Maciniszyn,J., Jr., Chen,K.C.S., Huang,W-Y., Tao,N., Liu,D. and Lanier,J.P. (1973) *Proc. Natl. Acad. Sci. USA*, **70**, 3437-3439.

Taylor,J.M. (1977) *Biochim. Biophys. Acta*, **473**, 57-71.

Toh,H., Hayashida,H. and Miyata,T. (1983) *Nature*, **305**, 827-829.

Varmus,H. (1983) *Nature*, **304**, 116-117.

Varmus,H. and Swanstrom,R. (1982) in Weiss,R., Teich,N., Varmus,H. and Coffin,J. (eds.), *RNA Tumor Viruses, Molecular Biology of Tumor Viruses*, ed. 2, Cold Spring Harbor Laboratory Press, NY, pp. 369-512.

Volovitch,M., Modjtahedi,N., Yot,P. and Brun,G. (1984) *EMBO J.*, **3**, 309-314.

Received on 7 February 1985