The EMBO Journal vol.9 no.10 pp.3353 – 3362, 1990

Origin and evolution of retroelements based upon their reverse transcriptase sequences

Yue Xiong and Thomas H.Eickbush

Department of Biology, University of Rochester, Rochester, NY 14627, USA

Communicated by D.J.Finnegan

To study the evolutionary relationship of reverse transcriptase (RT) containing genetic elements, a phylogenetic tree of 82 retroelements from animals, plants, protozoans and bacteria was constructed. The tree was based on seven amino acid domains totalling 178 residues identified in all RTs. We have also identified these seven domains in the RNA-directed RNA polymerases from various plus-strand RNA viruses. The sequence similarity of these RNA polymerases to RT suggests that these two enzymes evolved from a common ancestor, and thus RNA polymerase can be used as an outgroup to root the RT tree. A comparison of the genetic organization of the various RT containing elements and their position on the tree allows several inferences concerning the origin and evolution of these elements. The most probable ancestor of current retroelements was a retrotransposable element with both *gag*-like and *pol*-like genes. On one major branch of the tree, organelle and bacterial sequences (e.g. group II introns and bacterial msDNA) appear to have captured the RT sequences from retrotransposons which lack long terminal repeats (LTRs). On the other major branch, acquisition of LTRs gave rise to two distinct groups of LTR retrotransposons and three groups of viruses: retroviruses, hepadnaviruses and caulimoviruses.

Key words: group II introns/msDNA/retrotransposable elements/RNA-dependent RNA polymerases/viruses

Introduction

RNA-directed DNA polymerase or reverse transcriptase (RT) was first discovered nearly twenty years ago as a retroviral encoded enzyme catalyzing DNA replication from an RNA template (Baltimore, 1970; Temin and Mizutani, 1970). Since then many genetic elements from a wide variety of organisms have been shown to contain open-reading frames (ORFs) encoding proteins that are similar in sequence to retroviral reverse transcriptases (reviewed in Rogers, 1985; Finnegan, 1985; Weiner *et al.*, 1986; Boeke and Corces, 1989). These genetic elements fall into several groups: hepadnaviruses of animals and the caulimoviruses of plants, both DNA viruses (Toh *et al.*, 1983); transposable elements first discovered in yeast and *Drosophila melanogaster* which like retroviruses contain *gag* and *pol* genes and long terminal repeats (LTRs) (Saigo *et al.*, 1984; Clare and Farabaugh, 1985; Mount and Rubin, 1985); certain fungal group II mitochondrial introns and a mitochondrial plasmid (Michel and Lang, 1985); and a group of transposable elements first found in mammals and *D.melanogaster* that also contain retroviral-like *gag* and *pol* genes but do not contain LTRs (Fawcett *et al.*, 1986; Hattori *et al.*, 1986; Loeb *et al.*, 1986). The amino acid sequence similarity detected in the ORFs of these elements suggested a common origin for these many diverse RT sequences.

Although sequence similarity can be detected in other coding regions between certain of these different groups of elements, the RT region is the only region common to all elements and thus can be used for a comprehensive phylogenetic analysis of retroelements. We have previously conducted such an analysis using 37 RT sequences representative of each of these groups (Xiong and Eickbush, 1988a). The retroelements could be divided into two major branches. One branch contained the group II mitochondrial intron sequences and the non-LTR retrotransposable elements. This group of transposable elements has also been called the Line 1-like elements (Singer and Skowronski, 1985) or the poly(A)-type retrotransposons (Boeke and Corces, 1989). The second major branch of the RT tree contained the retroviruses and the LTR containing retrotransposable elements. The hepadnaviruses, copia and Ty1 represented the most distant members of this branch, while the caulimoviruses grouped closer to the retroviruses. An analysis of RT sequences has also been conducted by Doolittle and co-workers (Doolittle *et al.*, 1989) with somewhat different conclusions.

Since our previous report, additional genetic elements from each of these categories have been identified in a broad range of taxa including plants and protozoans. In addition RT containing genetic elements have been found that do not fit into any of the previously defined categories. Most interesting are the RT sequences recently identified in bacteria. These sequences produce multicopy single-stranded DNA (msDNA) containing both DNA and RNA covalently linked by a branched rG residue (Inouye *et al.*, 1989, 1990; Lampson *et al.*, 1989; Lim and Mass, 1989). To address the question of the origin of retroelements we have compared the retroelement RT sequences with the RNA-directed RNA polymerases of various plus-strand RNA viruses from bacteria, plants and animals. These RNA polymerase sequences have previously been shown to be related to RT sequences (Kamer and Argos, 1984; Poch *et al.*, 1989). The phylogenetic tree derived from this analysis provides a framework to evaluate possible models for the origin and evolution of the different categories of retroelements and RNA viruses.

Results

Alignment of RT sequences

In our previous study (Xiong and Eickbush, 1988a) alignment of RT sequences was based upon groups of conserved amino acid residues that could be identified in all RT-like sequences available at that time. The residues used were a modification of those originally identified by Toh *et al.*

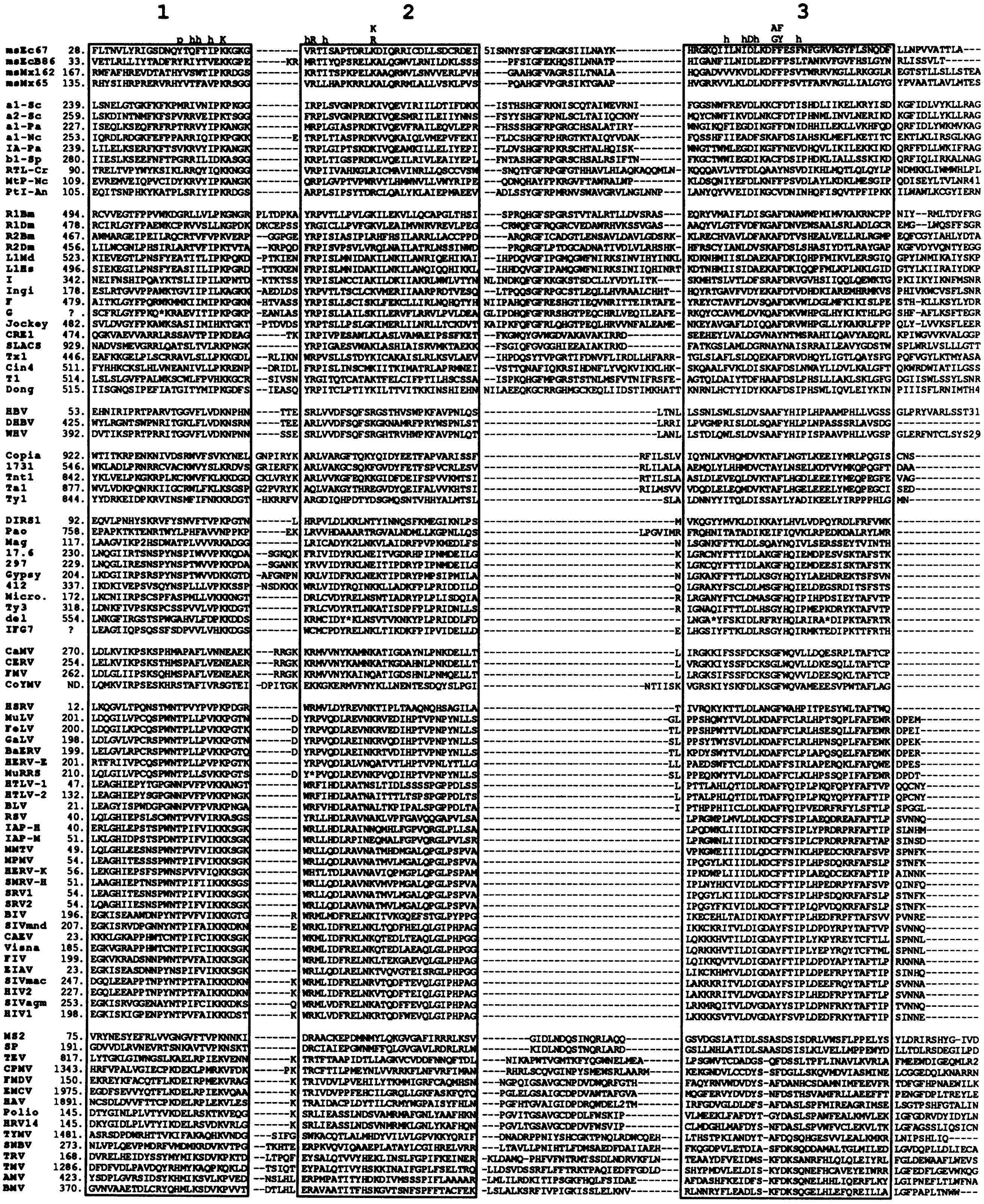
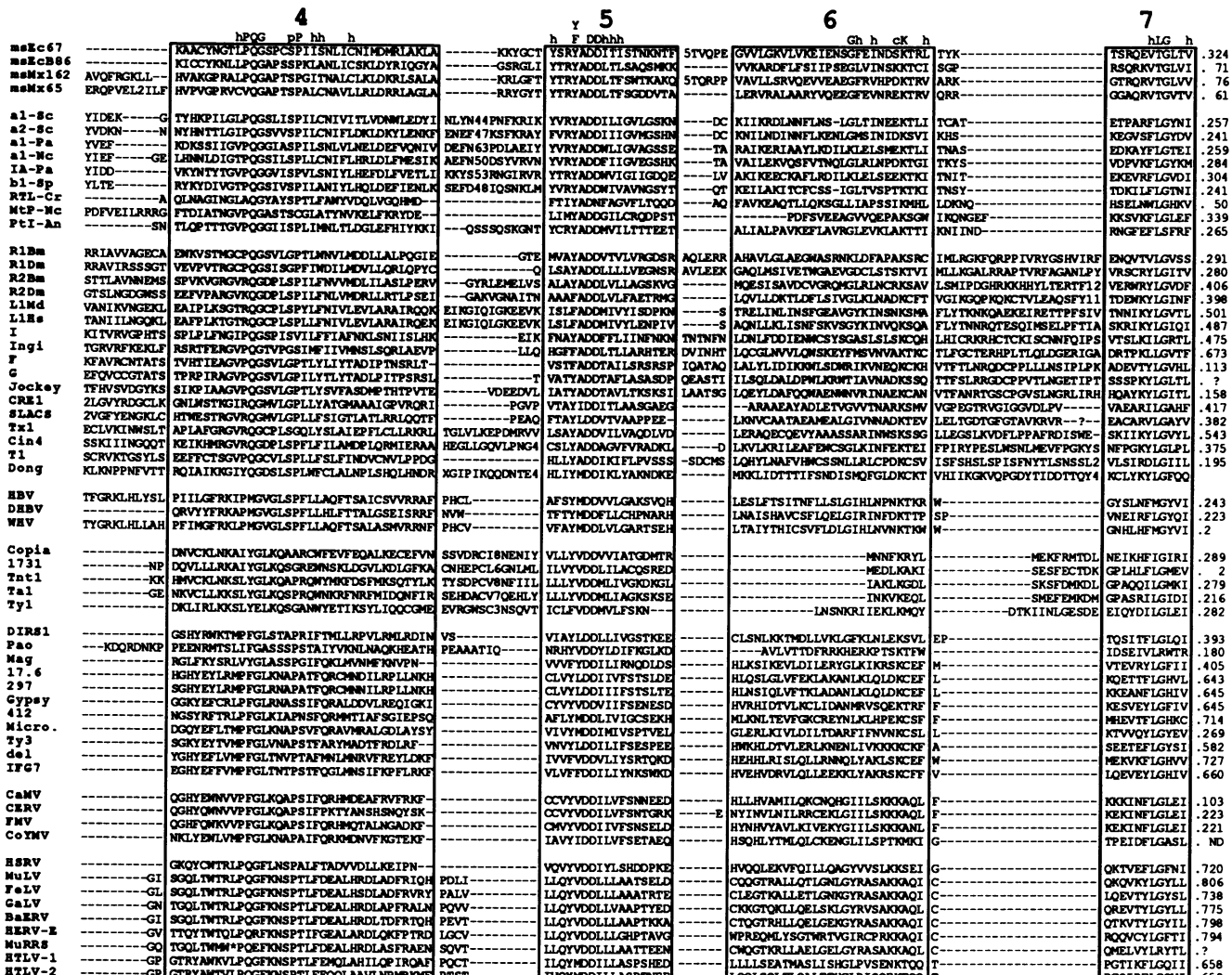


Fig. 1. Amino acid sequence alignment of RT-related and RNA-directed RNA polymerase sequences. Sources and abbreviations for each element are described in Xiong and Eickbush (1988a), Poch *et al.* (1989) or in Materials and methods (see Figure 3). The numbers at the beginning and end of each sequence indicate the number of residues from either the 5′ and 3′ end of the total ORF containing the polymerase sequence. Numbers within the sequence represent the number of amino acids present but omitted from the figure. Question marks indicate those instances in which the ends of



this ORF have not been determined. An asterisk (*) indicates a stop codon at that position. A question mark between domains 6 and 7 of the SLACS element indicates that a change in frame was necessary to maintain sequence similarities. Largely unvaried or chemically similar residues are shown at the top of the alignment. See text for a description of the criteria used in this assignment. h, hydrophobic residue; p, small polar residues; c, charged residue.

(1983, 1985). Alignment of residues between these fixed sites was conducted using algorithms that confer a substantial penalty for the insertion of gaps (Feng *et al.*, 1985). Seven peptide regions (domains 1–7) containing 178 amino acids were found to be common to all elements. This alignment differed from that subsequently reported by Doolittle *et al.* (1989), which was based upon a progressive alignment scheme (Feng and Doolittle, 1987). There was no disagreement between these two alignment procedures in comparisons where the level of amino acid identity was high and only a few small gaps were needed to maintain identity, as for example for different RT sequences from the same category of elements. However, the 'conserved residues' and the 'progressive alignment' methods differed substantially when the sequences compared were from different categories of RT elements. In these comparisons the number of identical residues was lower and the insertion of larger gaps was needed to maintain the alignment. For example, when comparing sequences from retroviruses and non-LTR retrotransposable elements only the domains we have labelled 4, 5 and 6 were identically aligned by the two methods. When comparing the group II introns and retroviruses only domains 5 and 6 are identically aligned by the two methods. These differences resulted from the inability of the progressive alignment algorithms to detect protein segments or domains in the various groups of retroelements that are not present in the retroviruses. These additional segments are as large as 70 amino acids (see Figure 1).

We believe the alignment based upon conserved residues as shown in Figure 1 is to be preferred for the following reasons. First, our original alignment utilizing 37 RT sequences required no major adjustments when 45 newly discovered RT sequences were added. The only change from our previous alignment is in domains 1 and 6 of the copia and Ty1 elements. This adjustment in alignment was due to the addition of three new retrotransposable elements (1731, TNT1 and Ta1) with substantial similarity to the copia and Ty1 elements, allowing a better identification of conserved residues within this group.

Further support for the alignment of RT sequences shown in Figure 1 has come from Webster *et al.* (1989) and Poch *et al.* (1989). Using methods which combine sequence similarities with predicted structures of the peptides these authors have independently identified common blocks of structural similarity among RT sequences. In the case of Webster *et al.* (1989) the four blocks identified correspond to our domains 2–5. In the case of Poch *et al.* (1989) the five motifs identified (regions a–e) correspond to our domains 3–7. Thus of the seven shared domains shown in Figure 1 only domain 1 has not been independently confirmed.

Within the seven conserved domains present in all RT sequences we have identified 42 conserved positions that contain identical or chemically similar residues in the majority of the 82 RT sequences analyzed in this report. These conserved positions are shown at the top of the alignment in Figure 1. To be classified as a conserved position the residues had to be present in over 50% of the RT elements from three of the four most abundant groups of RT-containing elements: retroviruses, LTR containing retrotransposons, non-LTR retrotransposons and group II mitochondrial introns. The 42 positions identified by this criterion were found to be present on average in 88% of all RT sequences (range 55–100%). Two sets of highly conserved residues characteristic of the LTR group (retroviruses, hepadnaviruses, caulimoviruses, LTR-containing retrotransposons) and the non-LTR group (group II introns and non-LTR retrotransposons) have been previously described, and are essentially unchanged from our original report (Xiong and Eickbush, 1988a; Figure 1). The 42 conserved positions identified in this report are representative of all RT-containing sequences. The LTR group had on average 88.9% of these conserved positions, while the non-LTR group had on average 86.4% of these conserved positions. The newly identified msDNAs contained 85.2% of these sites.

Comparison with RNA-directed RNA polymerases

A number of studies have attempted to find the relationship between various RNA and DNA polymerases by sequence comparisons (Kamer and Argos, 1984; Argos, 1988; Poch *et al.*, 1989). As expected, the enzymes most related to RTs are the RNA-directed RNA polymerases identified in various RNA viruses, in particular the polio-like group. To determine whether the RNA-directed RNA polymerases can be sufficiently aligned with the RT sequences to serve as an outgroup to root the RT tree, we have compared 15 RNA polymerase sequences from bacteria, plants and animals. Alignment of these polymerase sequences from representative viruses is shown in Figure 1. The alignment we obtained between the RNA polymerases and RT sequences in domains 3–6 agrees with that of Poch *et al.* (1989). The conserved nature of the RNA polymerase sequences in the regions corresponding to domains 3, 4, 5 and 7 have also been reported by Halibi and Symons (1989). Most of the RNA viruses did not have a complete domain 6, a situation similar to the RT elements of the copia/Ty1 group. Using the sets of conserved residues in the RT sequences we were also able to identify domains 1 and 2. The 15–30 amino acid residue segment between domains 2 and 3 of the RNA polymerases was similar to that found in the non-LTR elements (non-LTR retrotransposons, group II introns and msDNA). Of the 42 conserved positions in the RT sequences, on average 25.2 or 60.0% are also conserved in the RNA polymerases. These residues were most conserved in the animal RNA viruses of the polio group (72.1%) and least conserved in the Sindbis-like plant viruses (46.0%). While the total level of sequence identity between the RT and RNA polymerase sequences is low (average 12.2%) this value is similar to the level of identity detected between the most divergent groups of RT elements (the copia/Ty1 related elements and the msDNA elements have on average only 12.5% amino acid identity).

Generation of a phylogenetic tree

Using the number of identical residues scored in the 178 positions in the 7 domains shown in Figure 1, the neighbor-joining (NJ) method (Saitou and Nei, 1987) was used to generate a phylogenetic tree of the 82 RT and 15 RNA polymerase sequences. A simplified version of the unrooted tree generated by the NJ method is shown in Figure 2. To make it easier to visualize the topology of this tree, elements that are of the same structure and are localized on the same branch of the tree, are indicated with a box. Only seven elements do not fall within the eleven major categories of elements, two of which are shown in Figure 2. Branch points of all elements are shown in the rooted tree in Figure 3.

Since total sequence identity between the different

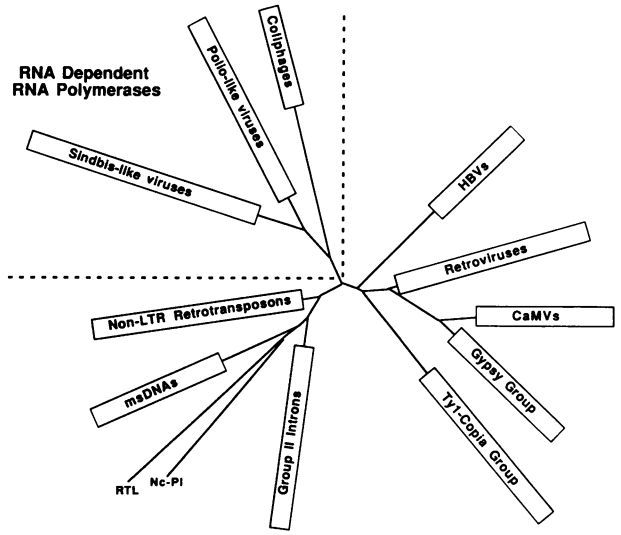


Fig. 2. Unrooted phylogenetic tree of the RT and RNA polymerase sequence constructed by the NJ method (Saitou and Nei, 1987). While data from all 82 RT and 15 RNA polymerase sequences shown in Figure 1 were used to generate the tree, to simplify visual comparison of the major topologies of the tree, elements from the same class that are located on the same branch of the tree are indicated by a box. The length of the boxes correspond to the most divergent element within that box.

categories of elements is low, we have also generated a phylogenetic tree of the different elements using only the sequence data from domains 3–7. These domains contain the highest levels of sequence identity, and are the only domains previously recognized in the RNA polymerases (Poch *et al.*, 1989; Halibi and Symons, 1989). These five domains contain 123 residues, or 69% of the number used in the full alignment. Using this reduced data set the relationship of each of the major categories of elements remains the same, i.e. the topology of this tree is identical to that shown in Figure 2. There are however minor differences in the order of certain branches within the non-LTR retrotransposons, Copia/Ty1 and retroviral branches (data not shown). Since in these instances, the differences in topology concern only elements with relatively high levels of sequence identity, the topology derived from the larger data set (178 amino acids) is presented in this report.

As shown in Figure 2 the RNA polymerases are all located on one branch which joins the RT branches on the segment connecting the non-LTR retrotransposons with the hepadnaviruses. The coliphages, MS2 and SP, are the most distant members of this branch. The eukaryotic viruses clearly fall into the polio-like and the Sindbis-like groups. When these viral RNA polymerase sequences are used to root the tree, all RT containing elements fall into two major branches. One branch contains the bacterial msDNAs, group II introns and non-LTR retrotransposons while the second branch contains the three types of viruses (hepadnaviruses, caulimoviruses and retroviruses) and the LTR-containing retrotransposons. This is the same rooting of the RT tree suggested in our original report on the basis of other considerations (Xiong and Eickbush, 1988a). As in that report these two major branches will be called the LTR branch and the non-LTR branch.

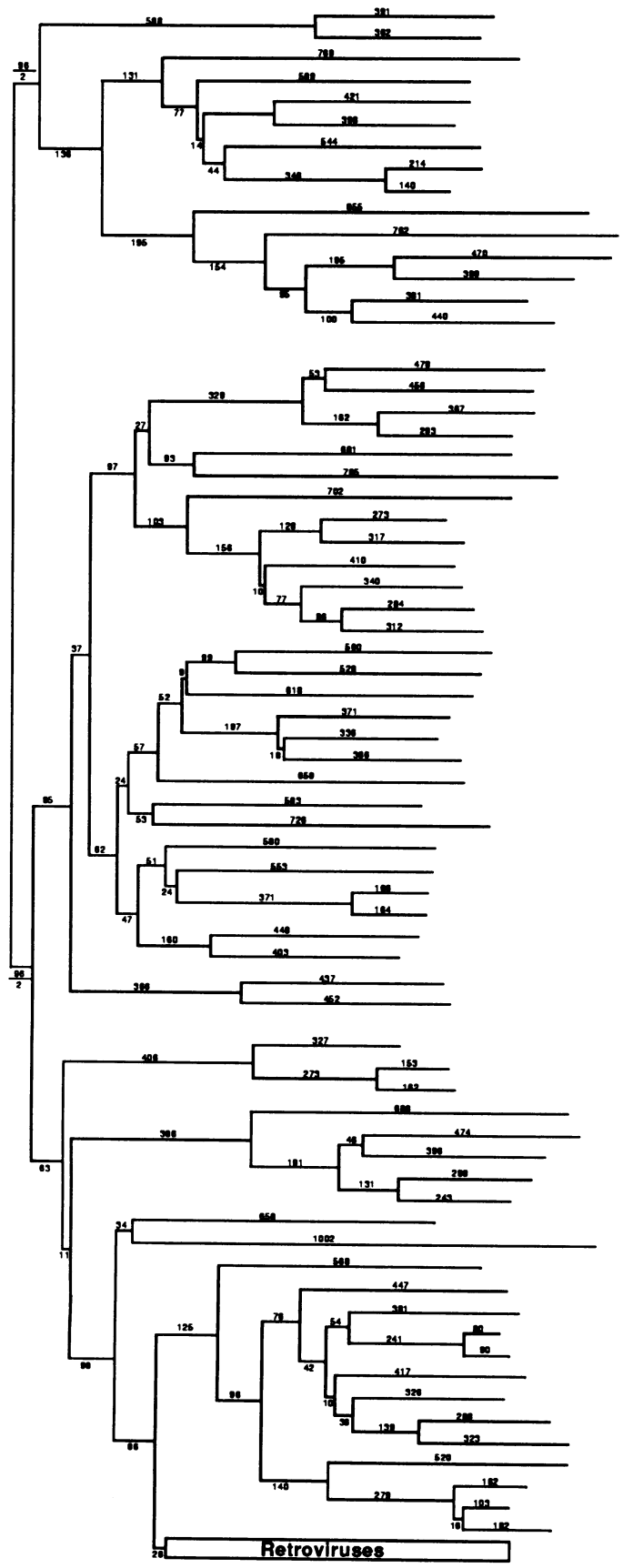
The newly discovered msDNA elements are grouped together with a series of retroelements found in organelle genomes. These include the Mauriceville mitochondrial plasmid sequence, Nc-P1, from *Neurospora crassa* (Nargang *et al.*, 1984), and an ORF containing RT sequences (RTL) linked to a mitochondrial ribosomal RNA gene fragment in *Chlamydomonas reinhardtii* (Boer and Gray, 1988). These elements in turn are most related to the group II mitochondrial and plastid introns on the non-LTR branch.

It has been suggested that the discovery of bacterial RT demonstrates that this enzyme is more ancient than the separation of prokaryotes and eukaryotes with msDNA elements being the possible ancestor of all currently identified retroelements (Temin, 1989). Two features of these elements have been used to support this idea. First, the codon usage analysis of msDNA-associated RT of *Myxococcus* indicated that the msDNA-Mx162 was not recently acquired into the genome (Inouye *et al.*, 1989). However, both GC content and codon usage analysis of msDNA-associated RT of *E.coli* indicated that it was recently acquired in this species (Lampson *et al.*, 1989). Second, the ORFs giving rise to msDNAs are the shortest and simplest of the retroelements containing only RNase H activity in addition to the RT activity.

The question of the relative antiquity of msDNA elements is an important issue as it bears directly on how the tree of RT sequences is to be rooted. The position of the viral RNA polymerases on the tree argues against rooting the tree with the msDNA elements. If the msDNA elements are assumed to be the progenitors of all retroelements, then the RNA viruses appear to be a category of retroelements whose polymerases have undergone a substantial change from synthesizing DNA to synthesizing RNA. This radical change would appear to have taken place sometime after the non-LTR retrotransposons diverged and before the hepadnaviruses. This disjunction in the tree is avoided if the RNA polymerases are assumed to be an outgroup. In addition there are reasons to believe that the RNA viruses are older than other retroelements. First, they have greater diversity in their genomic organization and sequences than any other branch of the tree. Second, RNA viruses are present in a wider diversity of prokaryotic and eukaryotic organisms than are the elements of any other branch of the tree. For these reasons we suggest, as have others (Lazcano *et al.*, 1988; Poch *et al.*, 1989), that the RNA viruses are as old or older than retroelements and are therefore the most reasonable branch on which to root the RT tree. A complete version of this rooted tree is shown in Figure 3. The position of most of the elements is shown in panel A. Retroviruses, which are all located on one branch of this tree, are presented in panel B.

The tree shown in Figure 3 has essentially the same topology as that in our previous report even though it contains an additional 45 new RT sequences as well as 15 RNA polymerase sequences (compare with Xiong and Eickbush, 1988a; Figure 4). Only two differences are observed, one is in the location of the branch containing the copia and Ty1 elements, and the second is the location of HSRV within the retroviral branch. Copia and Ty1 were originally branched together with hepadnaviruses, but in the current tree they exhibit a closer relationship to the other LTR-retrotransposons and retroviruses. This difference in topology is due to a change in the sequence alignment of these elements in domains 1 and 6, as was noted above. In the case of HSRV, our original report based upon 14 retroviruses, placed HSRV as a separate branch of retroviruses. In the current tree, which is based upon 29 retroviral

A



name	full name & host	
MS2	group I RNA coliphage (b)	Plus-strand RNA Viruses
SP	group IV RNA coliphage (b)	
TEV	tobacco etch virus (b)	
CPMV	cowpea mosaic virus (b)	
FMDV	foot and mouth disease virus (b)	
EMCV	encephalomyocarditis virus (b)	
HAV	human hepatitis A virus (b)	
Polio	poliovirus (b)	
HRV14	human rhinovirus 14 (b)	
TYMV	turnip yellow mosaic virus (b)	
SNBV	Sindbis virus of chicken (b)	
TMV	tobacco mosaic virus (b)	
TRV	tobacco rattle virus	
AMV	alfalfa mosaic virus (b)	
BMV	brome mosaic virus (b)	
Ec67	E.coli	msDNA-Associated RT
EcB86	E.coli B	
Mx162	M.xanthus	
Mx65	M.xanthus	
Nc-Pl	N.crassa mt. plasmid (a)	
RTL	RT-like protein (protozoan)	
Ptl	plastid petD (green alga)	Group II Introns
a1-Sc	S.cerevisiae mt. (a)	
a2-Sc	S.cerevisiae mt. (a)	
a1-Nc	N.crassa mt.	
b1-Sp	S.pombe mt. (a)	
a1-Pa	P.anserina mt. (a)	
a2-Pa	P.anserina mt. (a)	
R1Bm	B.mori (silkmoth) (a)	Non-LTR Retrotransposons
R1Dm	D.melanogaster	
Ingi	T.brucei (protozoan) (a)	
F	D.melanogaster (a)	
G	D.melanogaster (a)	
Jockey	D.melanogaster	
T1	A.gambiae (mosquito)	
I	D.melanogaster (a)	
Dong	B.mori (silkmoth)	
Tx1	X.laevis (frog)	
Cin4	Z.may (corn)	
L1Md	mouse LINE 1 (a)	
L1Hs	human LINE 1	
R2Bm	B.mori (silkmoth) (a)	
R2Dm	D.melanogaster	
CRE1	C.fasciculata (protozoa)	
SLACS	T. brucei (protozoan)	
DHBV	duck hepatitis B virus (a)	Hepadnaviruses
HBV	human hepatitis B virus (a)	
WHV	woodchuck hepatitis B virus	
Ty1	S.cerevisiae (a)	LTR Retrotransposons
Copia	D.melanogaster (a)	
1731	D.melanogaster	
Ta1	A.thaliana (flowering plant)	
Tnt1	N.tabacum (tobacco)	
DIRS1	D.discoideum (slime mold) (a)	
Pao	B.mori (silkmoth)	
Mag	B.mori (silkmoth)	
412	D.melanogaster (a)	
Gypsy	D.melanogaster (a)	
17.6	D.melanogaster (a)	
297	D.melanogaster (a)	
Micropia	D.melanogaster	
Ty3	S.cerevisiae	
IFG7	P. radiata (pine tree)	
del	L.henryi (lily)	
CoYMV	commelina yellow mottle virus	Caulimoviruses
FMV	figwort mosaic virus	
CaMV	cauliflower mosaic virus (a)	
CERV	carnation etched ring virus	

sequences, HSRV was found as the most distant member of the MuLV group. All retroviruses fall into four major groups; the MuLV group, the MMTV/RSV group, the HTLV group and the lentiviral group. The preservation of essentially the same topology for the various retroelements with the use of more than twice the number of sequences, suggests that the addition of many elements yet to be discovered will not significantly change this tree.

Sixteen new retrotransposable elements have been identified since our previous report. Seven of these retrotransposons lack terminal repeats. Five of these non-LTR elements; Cin4 of *Zea mays*; Jockey, T1 and Dong of insects, and Tx1 of *Xenopus laevis* are located on the major non-LTR retrotransposon branch of the tree. The two remaining non-LTR elements, CRE1 and SLACS, are located as a separate branch on the tree somewhat closer to the LTR containing elements, thus are the most distant members of the non-LTR branch. These unusual elements are found exclusively in the spliced leader (miniexon) genes of trypanosomatids (Aksoy *et al.*, 1990; Gabriel *et al.*, 1990).

The nine remaining retrotransposable elements that are new to the tree all contain terminal repeats; seven of these contain typical LTR structures. Three of these elements; 1731 from *D.melanogaster* and Ta1 and TNT1 from plants are closely related to copia and Ty1, and like these elements have their integrase domains located amino-terminal to the RT domain. The four remaining elements; micropia from *D.melanogaster*, Ty3 from *S.cerevisiae* and IFG7 and del from plants are clustered within the gypsy branch, and have their integrase domain located carboxylterminal of the RT domain. The two remaining elements are both from *Bombyx mori* and contain unusual LTRs. MAG contains terminal repeats only 70 bp in length, considerably shorter than any previously identified LTR (Michaille *et al.*, 1990). POA contains 600–800 bp terminal repeats, which have many of the characteristics of LTRs, but contain a central 300–500 bp region composed of a tandemly repeated DNA sequence (Y.Xiong and T.H.Eickbush, unpublished). The only other retrotransposable element which does not fit within either the copia/Ty1 or gypsy groups is the DIRS element from *Dictyostelium discoideum*. This element also contains an unusual terminal repeat, in this case the terminal repeats are in an inverted orientation (Cappello *et al.*, 1985).

Clearly there is a strong correlation between RT sequence and the terminal structure of the element. There are no examples of an LTR-containing retrotransposon whose RT sequences fall on the non-LTR branch, or of a non-LTR retrotransposon whose RT sequences fall on the LTR branch. Thus there is no evidence for sequence exchange between members of the different groups of retrotransposable elements present in the same species (e.g. *S.cerevisiae*, *D.melanogaster* and *B.mori*).

Discussion

Similarities of RNase, protease and integrase sequences between elements of the LTR branch have been reported by Doolittle and coworkers (Doolittle *et al.*, 1989). These sequence similarities, however, are not consistently detected within the elements of the non-LTR branch (Y.Xiong and

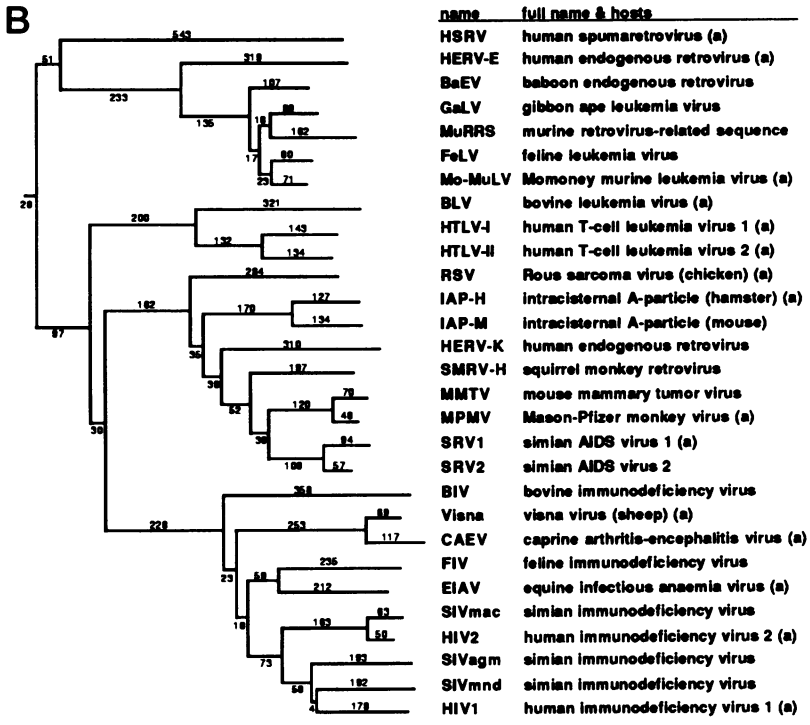


Fig. 3. Complete phylogenetic tree of RT elements rooted at the connection between the RT elements and the RNA-directed RNA polymerases. (**A**) All portions of the complete tree except for the retroviruses. (**B**) Retroviral branch of the tree. The number above or below each horizontal line indicates the branch length. The branch length between the node connecting all RT sequences and the RNA polymerase sequences was divided equally. Functional classification of each element, its full name and the host are presented to the right of the tree. References to the sequences used can be found in (a) Xiong and Eickbush (1988a), (b) Poch *et al.* (1989) or listed in Materials and methods.

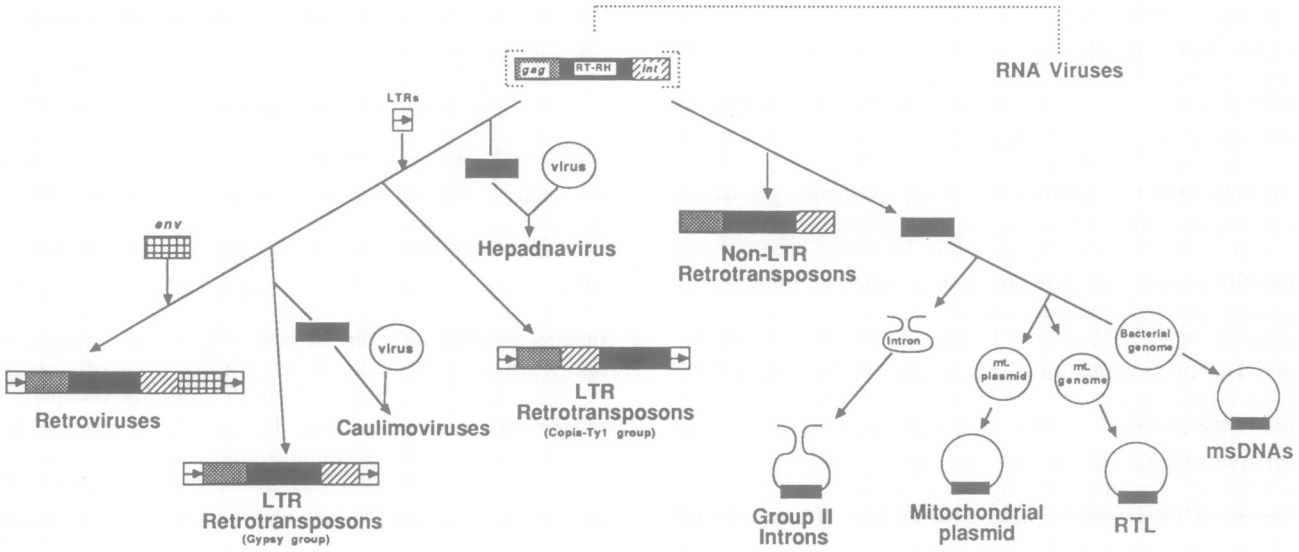


Fig. 4. A scheme for the origin of retroelements. Important structural features of each category of RT are presented. Shaded boxes correspond to domains of the ORF of the element, solid shading, RT region; stippled, *gag* region; diagonal shading, integrase domain; cross-hatched, envelope gene. LTRs present in certain elements are diagrammed as an open box with an arrow. Structural features of the ancestral retrotransposable elements (shown in brackets) are assumed based on structures present on both major branches of the tree. For the hepadnaviruses, caulimoviruses, group II introns, the Mauriceville mitochondrial plasmid, RTL sequences and msDNA it is assumed that only a portion of the *pol* gene, containing the RT domain, entered an already existing element.

T.H.Eickbush, unpublished data). In this report we have used the RT domain, the only domain found in all retroelements, to determine the phylogenetic relationships of the many diverse RT elements. In an effort to root the tree and determine in what order the different types of retroelements arose, we have used as an outgroup the sequences of another polymerase which shows the greatest similarity to RT, the RNA-directed RNA polymerase of RNA viruses. This does not imply that all retroelements evolved directly from the RNA viruses, only that the RNA viruses and retroelements share a common ancestor. An alternative model in which all retroelements evolved from bacterial msDNA, requires that the RNA viruses be relatively young, evolving from retroelements at about the time of the hepadnaviruses. This alternative model appears inconsistent with both the variety of structures and distribution of the RNA viruses and the substantial differences in enzymatic properties.

An order for the evolutionary acquisition of retroelement functions

Figure 4 shows a summary drawing of the phylogenetic tree obtained from our analysis to which we have added schematic drawings of the important structural features of each group of RT elements. Because retrotransposons are the only elements common to both the LTR and non-LTR branches, their structure is shown as the most likely progenitor of all current retroelements. Since there is no evidence for LTRs in the RNA viruses or hepadnaviruses, we have assumed that the progenitor element did not contain LTRs. The diversity of non-LTR retrotransposons, revealed by the deep branches on the tree in this group, and their wider distribution than any other class of retroelements (see Figure 3) supports the suggestion that the non-LTR retrotransposons are the oldest group of retroelements. The coding capacity of the ancestral retrotransposable elements, based upon the presence of features in both the LTR and non-LTR branches, are a *gag* gene and a *pol* gene either as two separate ORFs or one larger ORF. The only similarity in sequence of the *gag* gene between the two major branches is a series of Cys-motifs with similar spacing of cysteine and histidine residues (Covey, 1986; Fawcett *et al.*, 1986; Xiong and Eickbush, 1988b). Those elements that lack these Cys-motifs (e.g. L1 elements) still retain this second small ORF at their 5′ ends. In the case of retroviruses, the Cys motif resides in the nucleocapsid protein which is believed to be essential for efficient reverse transcription (see review by Varmus and Brown, 1989). No examples of retroelements with the ability to integrate but without a *gag* gene have been found, suggesting that these *gag* genes may be essential to sequester the retroelement's RNA.

The *pol* gene of the progenitor element is shown containing an RT domain and an integrase domain. While actual sequence similarity within the integrase domain has not been detected between the LTR and non-LTR elements, preliminary data suggest that such a domain is also located downstream of the RT domain in the non-LTR elements. A conserved Cys-motif has been detected downstream of the RT region in many of the non-LTR retrotransposons, while very little sequence similarity is detected upstream of the RT domain even between the same non-LTR elements present in different species (Jakubczak *et al.*, 1990). In the case of the copia/Ty1 group of retroelements, the integrase domain is located upstream of the RT domain, representing the rearrangement unique to this branch.

Given this core structure for the progenitor retroelements, the remaining categories of retroelements can be explained by a gain or loss of various functions. In the case of the three types of viruses on the LTR branch, the retroviruses are the easiest to explain. These elements are most similar to the gypsy group of retrotransposons and may represent a retrotransposable element which has acquired an envelope (*env*) gene making it possible for them to leave the cell. This model has been proposed for a number of years based on other lines of evidence (Temin, 1980; Finnegan, 1983). The origins of the hepadnaviruses and caulimoviruses are more difficult to explain, because the genomic structure of these

viruses is so different from that of the retrotransposons. Either many different functions were acquired or modified in two branches of retrotransposons or as would seem more likely, segments of the *pol* gene were acquired by pre-existing viruses. This segment would have also included the RNase H domain in the case of the hepadnaviruses, and the RNase H and protease domains in the case of the caulimoviruses (Doolittle *et al.*, 1989).

A transfer of at least the RT domain of the *pol* gene from a retrotransposon also appears to have occurred with the various retroelements of organelle and bacterial genomes. The best studied are the group II introns of mitochondria and plastids (see reviews by Lambowitz, 1989; Perlman *et al.*, 1989). Most group II introns do not contain RT ORFs and in those that do, it is located in a domain that has no effect on the splicing of the RNA. Thus it is not clear whether the RT containing group II introns are the progenitors, and many of these elements have lost their ORF or as appears more likely the RT ORF has become associated with an already functional intron. In the other three elements a similar event may also have occurred, in which the RT region was captured by a mitochondrial plasmid (the *N.crassa* Mauriceville plasmid), by the mitochondrial genome itself (RTL sequence of *C.reinhardtii*), or by the bacterial genome (msDNAs).

The distribution of retrotransposable elements and viruses

Each of the major groups of retrotransposable elements can be found in animals, plants and either protozoans or fungi. In certain cases even closely related retrotransposons can be found in widely different organisms. [Note in Figure 3 the location of copia (animal) and Ta1 (plant), or of micropia (animal) Ty3 (yeast) and del (plant).] On the other hand, the three types of viruses are each localized to particular taxa, hepadnaviruses and retroviruses to vertebrates and the caulimoviruses to plants. The RT tree does not support the simplest explanation for this difference in distribution: that the retrotransposable elements are more widespread because they are older. Both retroviruses and caulimoviruses predate the divergence of the retrotransposable gypsy, Ty3 and del.

The presence of related retrotransposable elements in very different taxa indicates either that the retrotransposons have spread horizontally, or that most of the major branch points on the RT tree are older than the evolution of metazoans. This latter possibility seems unlikely since it would mean that the branches giving rise to the viruses also predate metazoans. Indeed, it has been suggested that retroviruses evolved at about the time of the mammals (Doolittle *et al.*, 1989; Temin, 1989). The current distribution of retroelements can be explained if one assumes retrotransposons have been horizontally transferred across major taxonomic groups of organisms. Once functional retrotransposons were within a new taxa, new types of viruses evolved either by the capture of RT sequences from these transposons by pre-existing viruses, or by these transposable elements acquiring additional genes and becoming a virus.

Materials and methods

Sequence sources

TRV (Hamilton *et al.*, 1987), Ec67 (Lampson *et al.*, 1989), EcB86 (Lim and Mass, 1989), Mx162 (Inouye *et al.*, 1989), Mx65 (Inouye *et al.*, 1990), RTL-Cr (Boer and Gray, 1988), Ptl (Kuck, 1989), a1-Nc (Field *et al.*, 1989), R1Dm and R2Dm (Jakubczak *et al.*, 1990), Jockey (Priimagi *et al.*, 1988), T1 (Besansky, 1990), Cin4 (Schwarz-Sommer *et al.*, 1987), L1Hs (Hattori *et al.*, 1986), Dong (Y.Xiong and T.H.Eickbush, unpublished), Tx1 (Garrett *et al.*, 1989), CRE1 (Gabriel *et al.*, 1990), SLACS (Aksoy *et al.*, 1990), WHV (Giroens *et al.*, 1989), FMV (Richins *et al.*, 1987), CERV (Hull *et al.*, 1986), CoYMV (N.E.Oiszewski, unpublished), Ta1 (Voytas and Ausubel, 1988), TNT1 (Grandbastien *et al.*, 1989), 1731 (Fourcade-Peronnet *et al.*, 1988), micropia (Huijser *et al.*, 1988), Ty3 (Hansen *et al.*, 1988), IFG7 (Kossack,D., unpublished), del (Smyth *et al.*, 1989), BaEV (Kato *et al.*, 1987), GaLV (Delassus *et al.*, 1989), MuRRS (Schmidt *et al.*, 1985), FeLV (Donahue *et al.*, 1988), SMRV-H (Oda *et al.*, 1988), IAP-M (Mietz *et al.*, 1987), HERV-K (Ono *et al.*, 1986), SRV2 (Thayer *et al.*, 1987), MPMV (Sonigo *et al.*, 1986). BIV (Garvey *et al.*, 1990), SIVagm (Fukasawa *et al.*, 1988), FIV (Olmsted *et al.*, 1989), SIVmac (Chakrabarti *et al.*, 1987), SIVmnd (Tsujimoto *et al.*, 1989). Descriptions and full names of the elements are given in Figure 3.

Sequence alignment and formation of a phylogenetic tree

The procedure for the sequence alignment and phylogenetic tree construction has been previously described (Xiong and Eickbush, 1988a). Briefly, conserved residues present in each RT sequence were used to identify the seven domains. Alignment of residues between these fixed positions was by the Unitary Matrix (UM) method (Feng *et al.*, 1985). In the case of the RNA-directed RNA polymerases domains 3, 4, 5 and 6 have been previously identified (Poch *et al.*, 1989). Domains 1, 2 and 7 were found by first identifying conserved residues among the three major groups of viruses (Sindbis-like, Polio-like and Coliphage), followed by identifying similarities in these RNA polymerase residues with the conserved RT residues. The percent divergence for all pairwise comparisons of the 97 aligned sequences was calculated by dividing the number of different residues by the total number of compared residues. Before tree construction all values were changed to distances with Poisson correction, $d = -\log_e S$, where S = sequence similarity (Nei, 1987). These corrected values were then used to construct phylogenetic trees by the neighbor-joining (NJ) method (Saitou and Nei, 1987).

Acknowledgements

We thank Lynne Rosen for computer assistance and Bill Burke and John Jakubczak for comments on the manuscript. This work was supported by ACS grant NP number 691.

References

Aksoy,S., Williams,S., Chang,S. and Richards,F.F. (1990) *Nucleic Acids Res.*, **18**, 785–792.
Argos,P. (1988) *Nucleic Acids Res.*, **16**, 9909–9916.
Baltimore,D. (1970) *Nature*, **226**, 1209–1211.
Besansky,N.J. (1990) *Mol. Cell. Biol.*, **10**, 863–871.
Boeke,J.D. and Corces,V. (1989) *Annu. Rev. Microbiol.*, **43**, 403–434.
Boer,P.H. and Gray,M.W. (1988) *EMBO J.*, **7**, 3501–3508.
Cappello,J., Handelsman,K. and Lodish,H.F. (1985) *Cell*, **43**, 105–115.
Chakrabarti,L., Guyader,M., Alizon,M., Daniel,M.D., Desrosiers,R.C., Tiollais,P. and Sonigo,P. (1987) *Nature*, **328**, 543–547.
Clare,J. and Farabaugh,P. (1985) *Proc. Natl. Acad. Sci. USA*, **82**, 2829–2833.
Covey,S.N. (1986) *Nucleic Acids Res.*, **14**, 623–633.
Delassus,S., Sonigo,P. and Wain-Hobson,S. (1989) *Virology*, **173**, 205–213.
Donahue,P.R., Hoover,E.A., Beltz,G.A., Riedel,N., Hirsch,V.M., Overbaugh,J. and Mullins,J.I. (1988) *J. Virol.*, **62**, 722–731.
Doolittle,R.F., Feng,D.-F., Johnson,M.S. and McClure,M.A. (1989) *Quart. Rev. Biol.*, **64**, 1–30.
Fawcett,D.H., Lister,C.K., Kellett,E. and Finnegan,D.J. (1986) *Cell*, **47**, 1007–1015.
Feng,D.-F., Johnson,M. and Doolittle,R.F. (1985) *J. Mol. Evol.*, **2**, 112–125.
Feng,D.-F. and Doolittle,R.F. (1987) *J. Mol. Evol.*, **25**, 351–360.
Field,D.J., Sommerfield,A., Saville,B.J. and Collins,R.A. (1989) *Nucleic Acid Res.*, **17**, 9087–9099.
Finnegan,D.J. (1983) *Nature*, **302**, 105–106.
Finnegan,D.J. (1985) *Int. Rev. Cytol.*, **93**, 281–326.
Fourcade-Peronnet,F., d'Auriol,L., Becker,J., Bailibert,F. and Best-Belpomme,M. (1988) *Nucleic Acids Res.*, **16**, 6113–6125.
Fukasawa,M., Miura,T., Hasegawa,A., Morikawa,S., Tsujimoto,H., Miki,K., Kitamura,T. and Hayamai,M. (1988) *Nature*, **333**, 457–461.

Gabriel,A., Yen,T.J,. Schwartz,D.C., Smith,C.L., Boeke,J.D., Sollner-Webb,B. and Cleveland,D.W. (1990) *Mol. Cell. Biol.*, **10**, 615–624.
Garrett,J., Knutzon,D.S. and Carroll,D. (1989) *Mol. Cell. Biol.*, **9**, 3018–3027.
Garvey,K.J., Oberste,M.S., Elser,J.E., Braun,M.J. and Gonda,M.A. (1990) *Virology*, **175**, 391–409.
Girones,R., Cote,P.J., Hornbuckle,W.E., Tennant,B.C., Gerin,J.L., Purcell,R.H. and Miller,R.H. (1989) *Proc. Natl. Acad. Sci. USA*, **86**, 1846–1849.
Grandbastien,M., Spielmann,A. and Caboche,M. (1989) *Nature*, **337**, 376–380.
Halibi,N. and Symons,R.H. (1989) *Nucleic Acids Res.*, **17**, 9543–9555.
Hamilton,W.D.O., Boccara,M., Robinson,D.J. and Baulcombe,D.C. (1987) *J. Gen. Virol.*, **68**, 2563–2575.
Hanson,L.J,. Chalker,D.L. and Sandmeyer,S.B. (1988) *Mol. Cell. Biol.*, **8**, 5245–5256.
Hattori,M., Kuhara,S., Takenaka,O. and Sakaki,Y. (1986) *Nature*, **321**, 625–627.
Huijser,P., Kirchhoff,C., Lankenau,D.-H. and Hennig,W. (1988) *J. Mol. Biol.*, **203**, 233–246.
Hull,R., Sadler,J. and Longstaff,M. (1986) *EMBO J.*, **5**, 3083–3090.
Inouye,S., Hsu,M.-Y., Eagle,S. and Inouye,M. (1989) *Cell*, **56**, 709–717.
Inouye,S., Herzer,P.J. and Inouye,M. (1990) *Proc. Natl. Acad. Sci. USA*, **87**, 942–945.
Jakubczak,J.L., Xiong,Y. and Eickbush,T.H. (1990) *J. Mol. Biol.*, **212**, 37–52.
Kamer,G. and Argos,P. (1984) *Nucleic Acids Res.*, **12**, 7269–7282.
Kato,S., Matsuo,K., Nishimura,N., Takahshi,N. and Takno,T. (1987) *Jap. J. Genet.*, **62**, 127–137.
Kuck,U. (1989) *Mol. Gen. Genet.*, **218**, 257–265.
Lambowitz,A.M. (1989) *Cell*, **56**, 323–326.
Lampson,R.C., Sun,J., Hsu,M.-Y., Vallejo-Ramirez,J., Inouye,S. and Inouye,M. (1989) *Science*, **243**, 1033–1038.
Lazcano,A., Fastag,J., Gariglio,P., Ramirez,C. and Oro,J. (1988) *J. Mol. Evol.*, **27**, 365–376.
Lim,D. and Mass,W.K. (1989) *Cell*, **56**, 891–904.
Loeb,D.D., Padgett,R.W., Hardies,S.C., Schehee,W.R., Comer,M.B., Edgell,M.H. and Hutchison,C.A. III, (1986) *Mol. Cell. Biol.*, **6**, 168–182.
Michaille,J., Mathavan,S., Gaillard,J. and Garel,A. (1990) *Nucleic Acids Res.*, **18**, 674.
Michel,F. and Lang,B.F. (1985) *Nature*, **316**, 641–643.
Mietz,J.A., Grossman,Z., Luders,K.K. and Kuff,E.L. (1987) *J. Virol.*, **61**, 3020–3029.
Mount,S.M. and Rubin,G.M. (1985) *Mol. Cell. Biol.*, **5**, 1630–1638.
Nargang,F.E., Bell,J.B., Stohl,L.L. and Lambowitz,A.M. (1984) *Cell*, **38**, 441–453.
Nei,M. (1987) *Molecular Evolutionary Genetics*. Columbia University Press, New York.
Oda,T., Ikeda,S., Watanabe,S., Hutsushika,M., Akiyama,K. and Mitsunobu,F. (1988) *Virology*, **167**, 468–476.
Olmsted,R.A., Hirsch,V.M., Purcell,R.H. and Johnson,P.R. (1989) *Proc. Natl. Acad. Sci. USA*, **86**, 8088–8092.
Ono,M., Yasunaga,T., Miyata,T. and Ushikubo,H. (1986) *J. Virol.*, **60**, 589–598.
Perlman,P.S., Peebles,C.L. and Daniels,C. (1989) In Stone,E.M. and Schwartz,R.J. (eds), *Intervening Sequences in Evolution and Development*. Oxford University Press, Oxford, UK.
Poch,O., Sauvaget,I., Delarue,M. and Tordo,N. (1989) *EMBO J.*, **8**, 3867–3874.
Priimagi,A.F., Mizrokhi,L.J. and Ilyin,Y.V. (1988) *Gene*, **70**, 253–262.
Richins,R.D., Scholthof,H.B. and Shepherd,R.J. (1987) *Nucleic Acids Res.*, **15**, 8451–8466.
Rogers,J. (1985) *Int. Rev. Cytol.*, **93**, 187–279.
Saigo,K., Kugimiya,W., Matsuo,Y., Inouye,S., Yoshioka,K. and Yuki,S. (1984) *Nature*, **312**, 659–661.
Saitou,N. and Nei,M. (1987) *Mol. Biol. Evol.*, **4**, 406–425.
Schmidt,M., Wirth,T., Kroger,B. and Horak,I. (1985) *Nucleic Acids Res.*, **13**, 3461–3470.
Schwarz-Sommer,Z., Leclercq,L., Gobel,E. and Saedler,H. (1987) *EMBO J.*, **6**, 3873–3880.
Singer,M.F. and Skowronski,J. (1985) *Trends Biochem. Sci.*, **10**, 119–122.
Smyth,D.R., Kalitis,P., Joseph,J.L. and Sentry,J.W. (1989) *Proc. Natl. Acad. Sci. USA*, **86**, 5015–5019.
Sonigo,P., Barker,C., Hunter,E. and Wain-Hobson,S. (1986) *Cell*, **45**, 375–385.
Temin,H.M. (1980) *Cell*, **21**, 599–600.
Temin,H.M. (1989) *Nature*, **339**, 254–255.
Temin,H.M. and Mizutani,S. (1970) *Nature*, **226**, 1211–1213.
Thayer,R.M., Power,M.D., Bryant,M.L., Gardner,M.B., Barr,P.J. and Luciw,P.A. (1987) *Virology*, **157**, 317–329.
Toh,H., Hayashida,H. and Miyata,T. (1983) *Nature*, **305**, 827–829.
Toh,H., Kikuno,R., Hayashida,H., Miyata,T,. Kugimiya,W., Inouye,S., Yuki,S. and Saigo,K. (1985) *EMBO J.*, **4**, 1267–1272.
Tsujimoto,H., Hasegawa,A,. Maki,N., Fukasawa,M., Miura,T., Speidel,S., Cooper,R.W., Moriyama,E.N., Gojobori,T. and Hayami,M. (1989) *Nature*, **341**, 539–541.
Varmus,H.E. and Brown,P. (1989) In Berg,D. and Howe,M. (eds), *Mobile DNA*. Amercian Society for Microbiology, Washington, DC.
Voytas,D.F. and Ausubel,F.M. (1988) *Nature*, **336**, 242–244.
Webster,T.A., Patarca,R., Lathrop,R.H. and Smith,T.F. (1989) *Mol. Biol. Evol.*, **6**, 317–320.
Weiner,A.M., Deininger,P.L. and Efstratiadis,A. (1986) *Annu. Rev. Biochem.*, **55**, 631–661.
Xiong,Y. and Eickbush,T.H. (1988a) *Mol. Biol. Evol.*, **5**, 675–690.
Xiong,Y. and Eickbush,T.H. (1988b) *Mol. Cell. Biol.*, **8**, 114–123.

Received on June 19, 1990