Evolution of Retroposons by Acquisition or Deletion of Retrovirus-like Genes[1]

Marcella A. McClure

Department of Ecology and Evolutionary Biology, University of California, Irvine

The retroid family consists of all genetic elements that encode a potential reverse transcriptase (RT). Members of this family include a diversity of eukaryotic genetic elements (viruses, transposable elements, organelle introns, and plasmids) and the retrons of prokaryotes. Some retroid elements have, in addition to the RT gene, other genes in common with the retroviruses. On the basis of RT sequence similarity, the retroposon group is defined as the eukaryotic long interspersed nuclear elements, the transposable elements of (1) *Drosophila melanogaster* (I and F factors), (2) *Trypanosoma brucei* (ingi element), (3) *Zea mays* (Cin4), (4) *Bombyx mori* (R2Bm), and members of the group II introns and plasmids of yeast mitochondria. The data presented here elucidate the extent of the relationships between the retroposons and other retroid-family members. Protein-sequence alignment data demonstrate that subsets of the retroposons contain different assortments of retroviral-like genes. Sequence similarities can be detected between the capsid, protease, ribonuclease H, and integrase proteins of retroviruses and several retroposon sequences. The relationships among the retroposon capsid-like sequences are congruent with the RT sequence phylogeny. In contrast, the similarity between ribonuclease H sequences varies in different subbranchs of the retroposon lineage. These data suggest that xenologous recombination (i.e., the replacement of a homologous resident gene by a homologous foreign gene) and/or independent gene assortment have played a role in the evolution of the retroposons.

Introduction

The protovirus hypothesis, first put forth by Temin in 1970, suggests that retroviruses evolved from cellular transposable elements (Temin 1970, 1980). Many classes of genetic elements, including cellular transposable elements from diverse organisms, are now known to contain amino acid sequences with similarity to the reverse transcriptase (RT) of retroviruses (Xiong and Eickbush 1988*b*; Doolittle et al. 1989). The retroid family [all RT containing genetic elements (Fuetterer and Hohn 1987; Boeke and Corces 1989)] is composed of retroviruses, retrotransposons, two different classes of DNA viruses (caulimoviruses of plants and hepadnaviruses of animals), non-long-terminal-repeat (non-LTR) transposable elements (non-LTR-TEs), and members of the group II introns and plasmids of yeast mitochondria. There has been much speculation as to both the origin of the retroid family and the possible role(s) its various members may play in the eukaryotic genome (Finnegan 1983; Temin and Engels 1984; Baltimore 1985; Temin 1985; Rogers 1986). The bacterial RT sequences (Inouye et al. 1989; Lampson et al. 1989; Lim and Maas 1989), now dubbed "retrons"

1. Key words: evolution, retroid elements, protein sequence analysis.

Address for correspondence and reprints: Marcella A. McClure, Department of Ecology and Evolutionary Biology, University of California, Irvine, Irvine, California 92717.

Mol. Biol. Evol. 8(6):835–856. 1991.

0737-4038/91/0806-0007$02.00

(Temin 1989), are recent additions to the retroid family. The retron sequences will not be considered further in the present study on the evolution of eukaryotic retroid elements. Finding RT sequences and activity in both eukaryotes and prokaryotes, however, supports the suggestion that an ancient RT activity was necessary for the conversion of an RNA-based genetic system to a DNA-based one (Eigen and Schuster 1982; Darnell and Doolittle 1986).

On the basis of the previously published analysis of the RT sequences, the non-LTR-TEs and the group II introns and plasmids of yeast mitochondria are more closely related to one another than to any other retroid element (Michel and Lang 1985; Di Nocera 1988; Xiong and Eickbush 1988*b;* Doolittle et al. 1989). Neither the non-LTR-TEs nor the yeast mitochondrial introns and plasmids have sequences analogous to LTRs. For the sake of simplicity the non-LTR-TEs and mitochondrial RT elements will all be referred to as the retroposons throughout the present paper (Rogers 1985; Hutchison et al. 1989). The multiple protein-sequence alignment data and phylogenetic trees presented here illustrate both the relationship among the retroposons and that between the retroposons and other retroid elements. The eukaryotic retroid-family subgroups are first briefly described, to provide the necessary context for the discussion of retroposon evolution.

Retroviruses generally code for structural proteins [i.e., a matrix protein (MA), a capsid protein (CA), a nucleocapsid protein (NC), and envelope proteins) and enzymatic proteins [i.e., protease (PR), RT, and integrase (IN)] (Varmus and Brown 1989). Flanking LTRs are a hallmark of all retroviral genomes. Retrotransposons are nuclear sequences also bounded by LTRs and capable of transposing via an RNA intermediate and are not known to be infectious (Weiner et al. 1986). There are two distinct lineages of retrotransposons based on RT sequence relationships (Xiong and Eickbush 1988*b;* McClure, accepted). One group shares common ancestry with the caulimoviruses and Dirs-1 element of *Dictyostelium discoideum,* while the other's closest relatives are the hepadnaviruses (retrotransposons versus copia-like retrotransposons; fig. 1). Comparisons of deduced protein sequences in these lineages have shown that more than the RT gene is found in common with retroviruses (Covey 1986; Doolittle et al. 1989). Retrotransposons share the same gene order as present-day infectious retroviruses. In contrast, copia-like retrotransposon integration proteins are found between the protease and polymerase genes—rather than after the RT, as in the retroviruses. Not all members of the retroid family, however, contain a full complement of retroviral genes (fig. 1). The DNA viruses, for example, replicate through an RNA intermediate, but their life cycles have little else in common with retroviruses.

The mitochondrial introns and plasmids, constituting one of the two major lineages of the retroposons, are found neither in all species of yeast nor in all strains of a given species of yeast (Natvig et al. 1984; Lang and Ahne 1985) (fig. 2). In addition, these types of elements have not been found in the mitochondria of any other organisms. The Mauriceville-Ic plasmid RT protein is known to be functional (Kuiper and Lambowitz 1988) and is most closely related to group II mitochondrial intron RT sequences (Michel and Lang 1985). Introns 1 and 2 of the *Saccharomyces cerevisiae* cytochrome oxidase subunit I gene are closely related to the *Schizosaccharomyces pombe* apocytochrome b gene intron (Lang and Ahne 1985), as discussed in the present paper.

The other major lineage of the retroposons, the non-LTR-TEs, can be further classified into those that integrate into the host genome in a non-sequence-specific

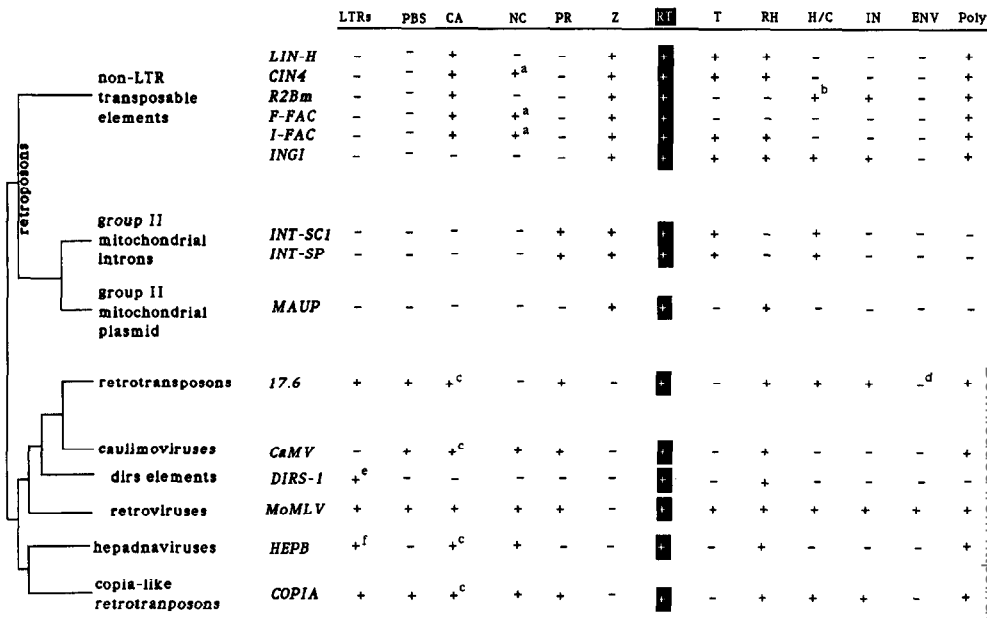


FIG. 1.—Schematic representation of eukaryotic retroid-family phylogeny determined by RT sequence similarity (*left*) and table of retrovirus gene complement found in representatives of major lineages. The distance scores calculated from the multiple alignment of 36 RT sequences were used to generate the topology via the method of Fitch and Margoliash (1967) and are in agreement with the tree published by Xiong and Eickbush (1988*b*). The RT sequence multiple alignment indicates the presence of six highly conserved motifs (McClure, accepted). The earlier alignment by Doolittle et al. (1989) showed only two conserved motifs and generated a tree suggesting that the hepadnaviruses' closest relatives are the retroviruses. Phylogenetic analysis of 36 RH sequences, however, produces a tree supporting the relationship of the hepadnavirus and copia-like retrotransposon lineages (data not shown), in agreement with the RT phylogeny indicated on the left side of the figure. Branch lengths are arbitrary. The gene that defines membership in the retroid family—i.e., the RT gene—is highlighted in the table. A superscript "a" denotes that among the non-LTR-TEs the NC relationship only consists of alignment to the CXXCXXXXHXXXXC motif of retrovirus NC. A superscript "b" denotes that in the *R2Bm* this region is highly divergent (see fig. 7). A superscript "c" denotes presence of CAs but absence of convincing similarity to retrovirus capsid sequences. A superscript "d" denotes presence of a region that could encode an envelope-like protein. A superscript "e" denotes presence of inverted LTRs. A superscript "f" denotes a region suggested to be related to retroviral LTRs (Miller and Robinson 1986). The retroid element key is as follows: *LIN-H* is human LINE; *CIN4* is from *Zea mays; R2Bm* is from *Bombyx mori; F-FAC* and *I-FAC* from *Drosophila melanogaster* and *INGI* from *Trypanosoma brucei* are other transposable elements; *INT-SC1* is the first intron of cytochrome oxidase subunit 1 from *Saccharomyces cerevisiae, INT-SP* is the intron of apocytochrome oxidase subunit from *Schizosaccharomyces pombe; MAUP* is the Mauriceville plasmid-1c strain of *Neurospora crassa; 17.6* is a transposable element from *D. melanogaster; CaMV* is cauliflower mosaic virus; *DIRS-1* is a transposable element from *Dictyostelium discoideum; MoMLV* is Moloney murine leukemia virus; *HEPB* is human hepatitis B virus, ayw strain; and *COPIA* is a transposable element from *D. melanogaster.* The gene key is as follows: PBS is the tRNA primer binding site; T is a region connecting the RT and RH; ENV is the membrane protein; and Poly (A) is the 3′ poly A tract; all other genes shown are as defined in the text.

manner [e.g., the long interspersed nuclear elements (LINEs) of mammals, and the I and F factors of the fruit fly, *Drosophila melanogaster*], one that has been directly demonstrated to integrate into the 28S rRNA gene [i.e., the R2Bm element of the silk moth, *Bombyx mori* (Burke et al. 1987)], and those that are inferred to exhibit target-site specificity [i.e., ingi of the protozoan *Trypanosoma brucei* (Kimmel et al. 1987) and the Cin4 element of corn, *Zea mays* (Schwarz-Sommer et al. 1987)]. It is inter-

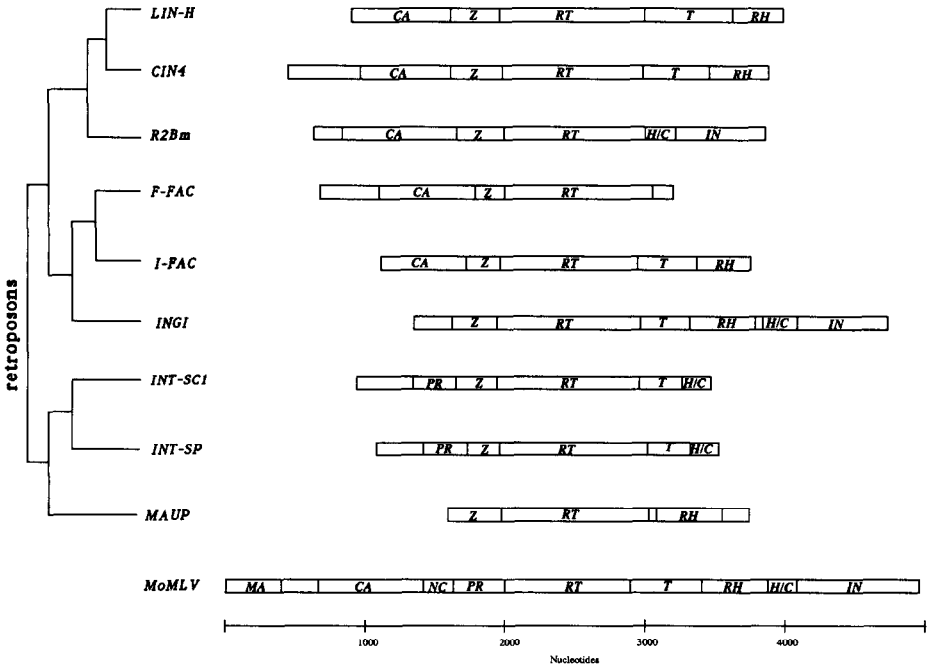


FIG. 2.—Summary maps of retrovirus-like genes in retroposon lineage of retroid family. The retroposon subtree relationships, (*left*) are based on RT similarities (fig. 1). Each map is a schematic of the entire RT-containing reading frame for each of the retroposons and *MoMLV*. *Z* = segment unique to retroposons, as described in Results. *MoMLV* is shown to indicate general retroviral gene order and complement (virus envelope gene not shown). *MA* segment of *MoMLV* = MA. Other abbreviations are as in fig. 1.

esting to note that the ability to integrate in a sequence-specific manner is not congruent with the RT phylogenetic relationship (fig. 2).

The LINEs are found throughout the genomes of mammals, and those present in human teratocarcinoma cells make unit-length transcripts, suggesting that they may code for functional proteins (Skowronski et al. 1988). The I factor is a transposable element involved in the I-R system of hybrid dysgenesis (Fawcett et al. 1986; Finnegan 1989), and the F factor is one of the most mobile in *D. melanogaster* (Di Nocera and Casari 1987; Di Nocera 1988). The ingi element has only been found embedded in a 512-bp ribosomal mobile element (RIME) of *T. brucei* (Kimmel et al. 1987). Cin4 of *Z. mays* is found within the A-1 gene and was the first plant sequence found to have RT coding capacity of presumed nonviral origin (Schwarz-Sommer et al. 1987). R2Bm is a 28S rRNA insertional element of *B. mori* (Burke et al. 1987).

Retroviruses and some retrotransposons encode an NC that contains a repeated cysteine/histidine (C/H) motif. The viral NC motif is necessary for recognition of RNA sequences, although the actual zinc-binding property of this region has not been demonstrated (Gorelick et al. 1988; Jentoft et al. 1988). Several of the non-LTR-TEs (e.g., the Cin4, the I factor, and the F factor) encode a region, in an alternative open-reading-frame (ORF), which is nonoverlapping with the RT reading frame and which can be aligned to the viral NC motif (Fawcett et al. 1986; Di Nocera and Casari 1987; Schwarz-Sommer et al. 1987). This relationship is limited, however, to the C/H motif, and the position and repeat number of the pattern vary among the Cin4, the I factor, and the F factor.

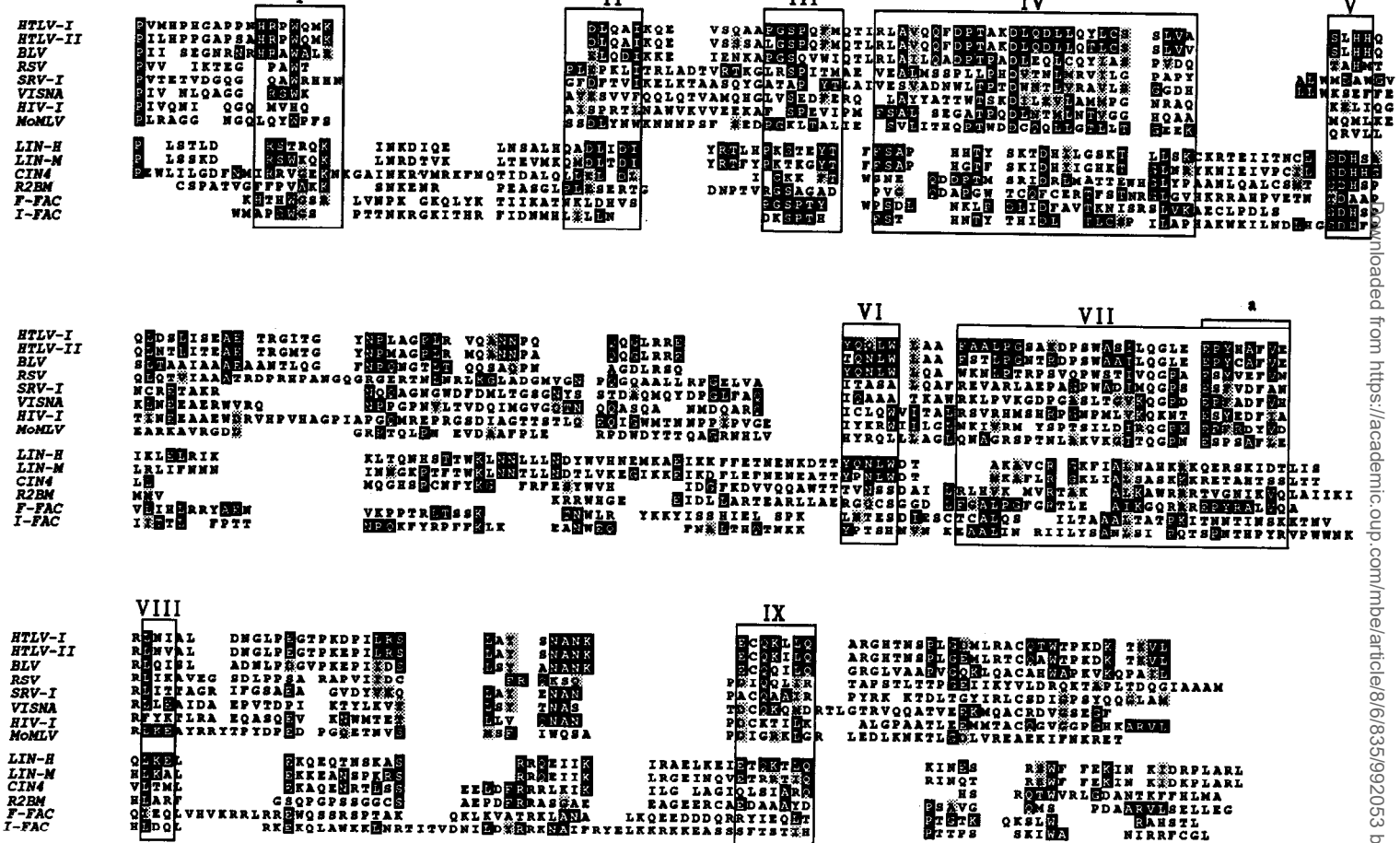
The present paper documents the extent of sequence relationships both among the retroposons and between the retroposons and other members of the retroid family. It was previously suggested, on the basis of three different small motifs found in common among various retroposon sequences and retrovirus proteins [see motif alignment between HTLV-I and human LINEs (fig. 3, VI), motif alignment between INT-SC1 and MoMLV (fig. 4, I), and motif alignment of ingi and HTLV-I) (fig. 5, I)], that perhaps these two groups share other genes in addition to the RT (Doolittle et al. 1989). The CA, PR, ribonuclease H (RH), and IN multiple-alignment data presented here demonstrate that different subsets of retroposons contain different collections of retroviral-like genes, indicating that these genes occasionally assort independently of one another. These findings pose interesting questions regarding the nature of the actual retroposon ancestor and about when and how such elements found their way into the genomes of these diverse organisms and organelles (mammals, insects, plants, parasites, and yeast mitochondria).

Methods

All computer analysis were carried out on a SUN workstation model 3/50 running SunOS 4.03. Sequences were taken from GenBank 60.0 or PIR 21.0. The retroposon sequences used in the present study are representative of their respective subgroups. Addition of other available sequences (e.g., the jockey or G element of *Drosophila melanogaster* or the R1Bm of *Bombyx mori*) for any of the subgroups does not alter the conclusions presented here (data not shown). For example, preliminary analysis of the jockey element clearly indicates the presence of a capsid-like sequence in the same position in which it is found for the I and F factors.

The strategy utilized to find the potential relationships described in Results consists of four stages: a very sensitive data-base search, an initial pairwise alignment, a subsequent multiple alignment, and, in some cases, a final manual refinement. Each stage is described in more detail below. This method has been successful in delineating other distant relationships (Johnson et al. 1986; McClure et al. 1987; McClure and Perrault 1989; Doolittle et al. 1989).

Initially all retroposon ORFs were compared with one another. When similarities were found, multiple alignment of the various segments [CA, Z, PR, RT, RH, H/C, and IN] was carried out. Subsequently, all residues exclusive of the RT segment (fig. 2), including other ORFs, were compared both with all other retroviral protein sequences and with the PIR data base by a method that detects small regions of similarity between much larger proteins (Doolittle 1987). When small regions were found [e.g., identity between HTLV-like viruses and the F factor (fig. 3, III) or identity between HTLV-like viruses and the LINEs (fig. 3, VI)], then a segment equivalent in size to the retroviral gene containing the match was aligned to the sequence of interest. Such pairwise alignments were made using the Feng et al. (1985) implementation of the Needleman and Wunsch (1970) algorithm. Two different scoring methods were employed in this analysis: the unitary matrix (Schwartz and Dayhoff 1978) and the minimum mutation matrix of Dayhoff (1978). When comparing sequences that are suspected to be highly divergent, the unitary matrix has proved invaluable in detection of local regions of identity, and it is useful for indicating regions of duplication, insertion, and deletion (author's unpublished observation). Initial multiple alignments (including prealignment of more closely related subsets) were generated by the program of Feng and Doolittle (1987), with the user-specified "weighting option" set at a value of 2. Manual refinements were introduced either when obvious regions of identity or



similarity were not detected by the program or when alternative gapping would either produce more consistent local region relationships or minimize the mutational events required to align one set of sequences to another.

It should be noted that the last stage of this analysis, manual refinement, is recognized as part of the state of the art in multiple alignment of distantly related sequences. In addition, there is not a published statistical method, to date, that can adequately evaluate distant relationships based on multiple alignments. Furthermore, with current methods, statistical confidence cannot be generated for any of the gene relationships between the retroviruses and any other retroid lineage. It is assumed that an ordered series of motifs, as indicated in any of the multiple alignments presented here, does not arise time and again by chance [for a discussion of these concerns, see the work of McClure (accepted)].

The multiple-alignment data for the CA, Z, RT, and RH sequences were used to generate pairwise difference matrices (Feng and Doolittle 1987). Phylogenetic trees were constructed from each matrix by an unweighted pair-group method of clustering (Fitch and Margoliash 1967).

Results

Summary Maps of Retrovirus-like Genes in Retroposons

At a minimum, infectious retroviruses code for an MA, a CA, an NC which associates with the viral RNA, a PR which cleaves the viral polypeptide, an RNA-dependent DNA polymerase (i.e., RT) connected by a tether region (T) to a segment with RH activity, an IN, and envelope proteins which insert into the host membrane to provide the virus coat (Varmus and Brown 1989). The results of the study determining the presence or absence of putative homologues of these genes in retroposons are summarized in figure 2. The phylogenetic relationship among the retroposons as deduced from RT sequences (Xiong and Eickbush 1988*b*; Doolittle et al. 1989; McClure, accepted) is indicated on the left side of the figure. The gene maps were determined by using the position of the RT sequences as a starting point for comparing all surrounding regions, first to all other regions among the retroposons and then to

FIG. 3.—Multiple alignment of retrovirus capsid and retroposon caspid-like sequences. A motif indicates a local region of identity defined by one of two rules: (1) one or more of the retroposons has either at least three contiguous residues or four (of five) ungapped residues that are identical to one or more of the representative retrovirus sequences, provided that remnants of this motif (including similar amino acids) are present in at least half of each of the two sets of sequences being compared; and (2) most of the retroposons share identical residues (i.e., fewer than three residues which may not be contiguous) with most of the retroviruses or with a specific subset of retroviruses. Regions I–IV and VI–IX follow rule 1, and region V follows rule 2. The color reversal of amino acid residues in each column indicates the global pattern of relationship between the two sets of sequences, and a pattern is determined by the following rules: Initial matches are defined as two or more identical residues in the retroposon set that are common to any other retroid member. Secondary matches are defined either as any two identical residues common to both sets and within plus or minus three positions of an initial match or as those that form a consistent pattern without gaps. If more than one set of matches occurs within a column of the alignment, the set with the majority of matches and conservative replacements is color reversed. Shaded residues indicate conservative replacements relative to color-reversed residues found in common between the two sets, on the basis of the similarity scheme (F,Y), (M,L,I,V), (A,G), (T,S), (Q,N), (K,R), and (E,D). By the criteria defined above, motifs are local islands of identity within a broader global pattern of relationship between the retroposons and the other retroid elements. *HTLV I* and *HTLV II* = human T-cell leukemia virus, types I and II; *BLV* = bovine leukemia virus; *RSV* = Rous sarcoma virus; *SRV-1* = simian retrovirus, type I; *VISNA* = visna lentivirus; *HIV-1* = human immunodeficiency virus, type 1; and *LIN-M* = mouse LINE. All other abbreviations are as in fig. 1. A blank line separates the retroposons (*below*) from the other retroid sequences (*above*).

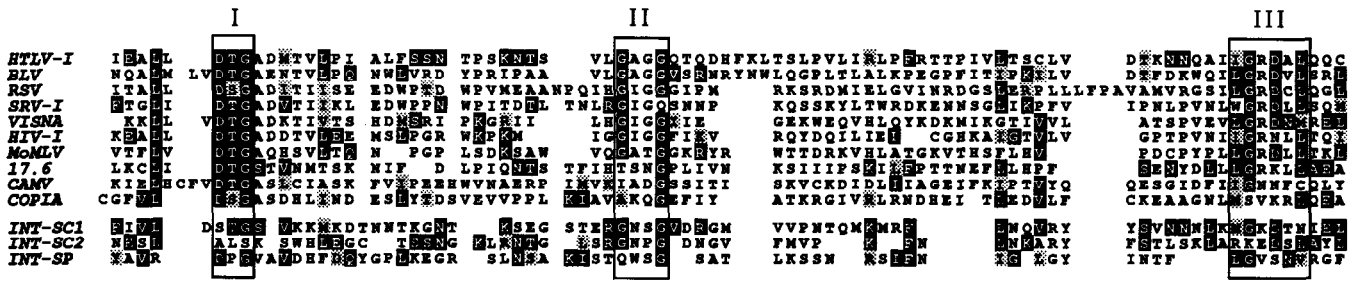


FIG. 4.—Multiple alignment of representative protease sequences from retroviruses, retrotransposons, caulimoviruses, and copia-like retrotransposons with group II mitochondrial intron protease-like sequences. Motifs I–III indicate the highly conserved residues of the viral PR that comprise the active site of the protein. Abbreviations are as in fig. 1 and 3. Boxing, color reversal, and shading of residue columns are as in fig. 3.

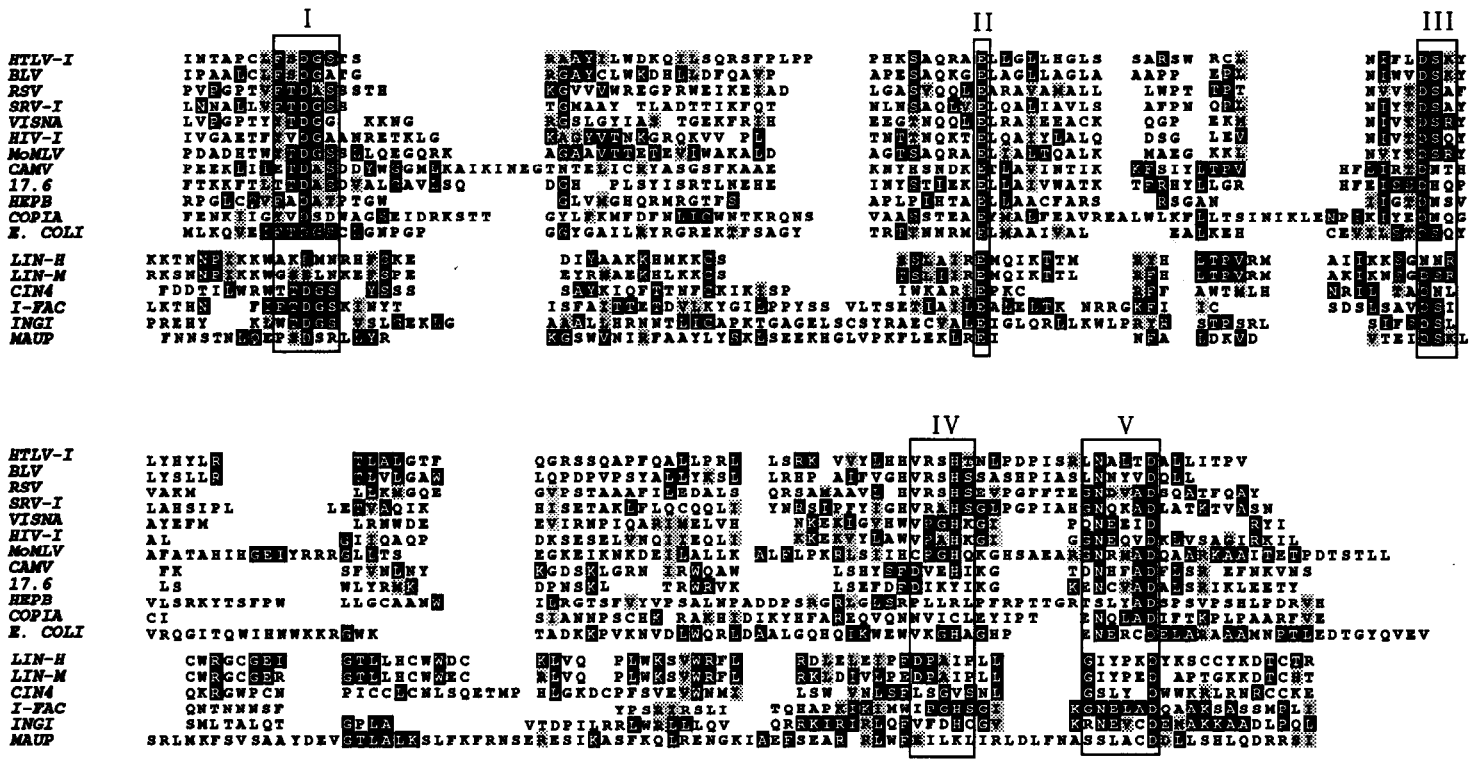


FIG. 5.—Multiple alignment of representative RH sequences from retroviruses, retrotransposons, caulimoviruses, copia-like retrotransposons, and hepadnaviruses and of *Escherichia coli* RH sequence with retroposon RH-like sequences. Remnants of regions I–V are found in all RH sequences. Abbreviations are as in fig. 1 and 3. Boxing, color reversal, and shading of residue columns are as in fig. 3.

other retroviral protein sequences and to the PIR protein data base. Once a potential relationship was found, pairwise sequence comparisons were carried out as described above (see Methods). Within the retroposon group the order of the various pairwise relationships is the same as that of the RT sequences. New potential homologues were therefore prealigned with their nearest neighbors, prior to subsequent multiple alignment (Feng and Doolittle 1987) of the entire set (see below).

By the strategy outlined above, the segment ("Z" in fig. 2) immediately amino terminal to the RT sequences was found to be unique to the retroposons and thus serves as a second marker (the similarity of the RT-like sequences is the first) for this group. The Z sequences were found to be as divergent from one another as were their respective RT genes, and their similarities produce the same tree topology for the retroposons as does the RT-based phylogeny (fig. 6*A* and *B*). Whether this segment, averaging 100 residues in length, is actually part of the RT gene will have to be determined experimentally. However, I found no similar region either within other members of the retroid family or in the PIR data base. The Z segment includes a previously described region, averaging 22 residues in length, found in the LINEs, R2Bm, and the I and F factors and ingi (Xiong and Eickbush 1988*c*).

The following section describes in detail the sequence relationships, exclusive of the Z/RT region, both among the retroposons themselves and between this group and other retroid elements. The multiple alignments presented were derived from larger data sets, but for the sake of brevity only representative sequences of the major retroid lineages and retroviruses are shown. Not all retroposons possess the same set of retroviral-like sequences. The LINEs and Cin4, for example, have both T and RH sequences, while their next nearest neighbor, R2Bm, has an H/C and an IN sequence in the analogous position (see fig. 2). In addition, each set of retroposon homologues should be most similar to their nearest neighbors, thereby producing the same phylogenetic tree topology. This is the case when a subtree containing only retroposons sequences is constructed. This is not the case, however, when the retroposon group is compared with other members of the retroid family, suggesting that xenologous recombination and/or independent gene assortment has played a role in the evolution of the retroposons. Xenologous recombination is defined as the replacement of a homologous resident gene (e.g., the I factor RH sequence) by a homologous foreign gene (e.g., a retroviral RH sequence).

An alternative, albeit unlikely, explanation for lack of congruent tree topology within a subgroup could be vastly unequal rates of change among viral protein homologues. This would be inconsistent, however, with the evidence thus far accumulated for the retroviral genes, which suggests that, like cellular proteins, homologous viral proteins within a subgroup exhibit relatively constant rates of change (McClure et al. 1988; Xiong and Eickbush 1988*b;* Doolittle et al. 1989). Of course, some of the discrepancies in the expected sequence similarities could be due to convergence, but the conservation of a specific order of motifs in a protein argues against convergence as a major force. Functional constraints seem a more likely explanation for the maintenance of a specific motif order among distantly related proteins. In addition, when a specific subset of the retroposons contains the same gene complement in the same order, it is difficult to attribute such consistency to convergence.

As mentioned above, only the Mauriceville plasmid RT protein has, been shown to be functionally active to date, (Kuiper and Lambowitz 1988). It is possible that the sequence relationships described here are merely remnants of inactive proteins. As indicated in the distance data for both the CA and RH relationships, however,

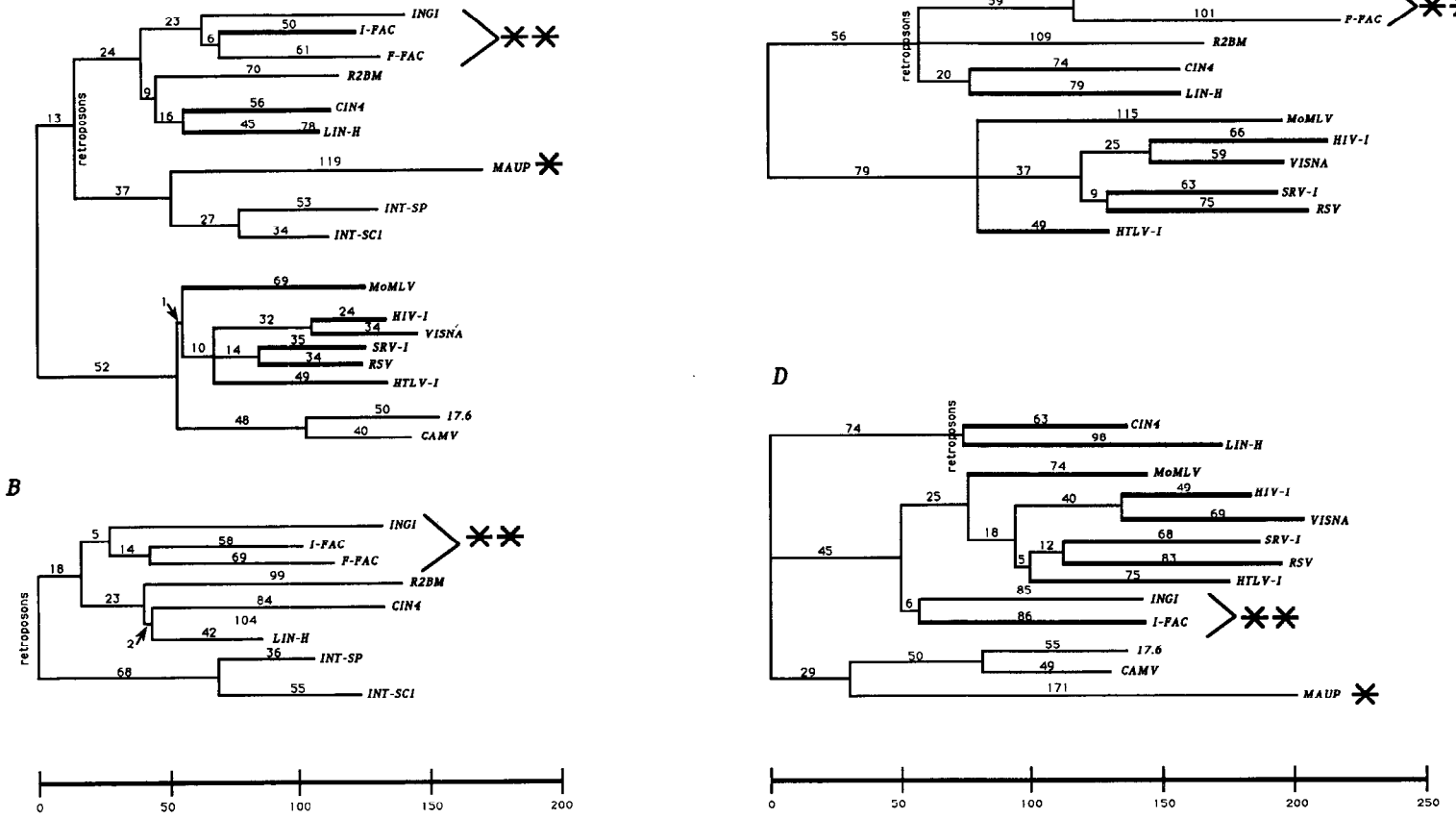


FIG. 6.—Phylogenetic trees for the RT (*A*), Z (*B*), CA (*C*), and RH (*D*) relationships. The topologies are congruent for the RT, Z, and CA trees. The RH tree indicates that the ingi and I factor RH sequences shared a common ancestor with the retroviruses (double asterisks indicate the limb change in *D* vs. *A–C*). The Mauriceville plasmid RH sequence descended from the ancestor of the *17.6* and *CAMV* RH sequences (limb position change is denoted by an asterisk in *D* vs. *A*). Thicker lines indicate sequences present in the each of the RT, CA, and RH trees. *Z* is unique to the retroposons. The Mauriceville plasmid (*MAUP*) contains a Z segment most similar to the *INT-SP* and *INT-SCI* Z; however, it is so distantly related that it cannot be included in the tree. The trees are scaled to the same arbitrary distance measure. Abbreviations are as in fig. 1 and 3.

similarity can be detected for specific subsets of the retroviruses (tables 1 and 2). If all these sequence relationships represent inactive proteins, then all resemblances to any functional protein should be lost. The mode and timing of entry of the variety of potential genes into the retroposon lineage is discussed in light of two proposed models (see Discussion).

Multiple Alignments

Capsid-like Region

An interesting similarity is observed between amino acid segments found in the non-LTR-TEs (LINEs, the Cin4 and R2Bm elements, and the I and F factors) and the viral capsid sequences of the retroviruses (fig. 3). As expected, the CAs of the retroviruses are more closely related among themselves. We have previously suggested that a recombination event accounts for the unexpectedly close relationship between the CAs of HTLV I and II (McClure et al. 1988). If such recombinants are excluded, viral CAs are quite divergent, and remnants of only three motifs can be found in common among these proteins (fig. 3, III, VIIa, and IX). Remnants of the viral capsid motifs were detected in some of the retroposon sequences (fig. 3, III, VIIa, and IX). The capsid-like sequences of the non-LTR-TEs were found to be more similar to one another than to any of the retroviral capsid sequences, and the topologies of the phylogenetic trees for the RT and CA genes are congruent (fig. 6*A* and *C*). A small hallmark motif was found among the non-LTR-TE sequences which can still be detected in the capsids of the HTLV-like viruses (fig. 3, V). It is surprising that, by the criteria of distance scores and percent identity, the sequences of non-LTR-TEs were found to be more similar to HTLV I–like sequences than to any other of the representative retroviruses (table 1).

Table 1
Pairwise Capsid Protein Comparisons

	DISTANCE VALUE *(% identical residues from representative retroposons and retroviruses)* FOR[a]		
	H-LIN	*CIN4*	*I-FAC*
HTLV-I	219 (15.9)	249 (15.0)	256 (13.1)
SRV-I	313 (5.3)	330 (6.2)	312 (6.2)
RSV	420 (6.5)	415 (5.1)	331 (6.1)
VISNA	305 (9.1)	366 (6.8)	375 (5.0)
HIV-I	316 (6.9)	413 (4.8)	386 (7.0)
MoMLV	317 (7.7)	331 (7.7)	364 (5.9)
LIN-H	...	154 (25.0)	196 (18.3)
CIN4	154 (25.0)	...	181 (17.8)
I-FAC	196 (18.3)	181 (17.8)	...

NOTE.—Abbreviations are as in figs. 1 and 3.

[a] Distance values were calculated from the relationship $-\ln[(S_{real} - S_{random})/(S_{identical} - S_{random})] \times 100$, where S_{real} = actual S for pair of protein sequences; $S_{identical}$ = average S obtained when each of two sequences of interest is aligned with itself; S_{random} = mean S for randomly scrambled and aligned pair. The *S* values were calculated from multiple sequence alignments (Feng and Doolittle 1987) by using the modified mutation matrix (Feng et al. 1985). Each gapped position opposite an amino acid was assessed a penality of 8 points, which is the default value set by the program. % Identical residues was calculated for the shorter of the two seuqences.

Table 2
RT and RH Pairwise Distance Values

	Distance Value for[a]			
	I-FAC		*MAUP*	
	RT	RH	RT	RH
HTLV-I	228	199	294	351
SRV-I	253	177	298	544
RSV	235	223	297	352
VISNA	237	248	295	419
HIV-I	229	200	293	317
MoMLV	240	184	272	317
CAMV	247	300	300	265
17.6	257	271	315	279
LIN-H	143	316	255	392
CIN4	159	263	268	324
INGI	136	171	310	336
MAUP	247	373	...	...
INT-SC1	207	NP	181	NP
INT-SP	209	NP	200	NP

NOTE.—NP = gene not present. All other abbreviations are as in figs. 1 and 3.
[a] From comparison of the I factor and Mauriceville plasmid to representative retroposons and retroviruses.

All distance scores and percent identities were calculated for the entire length of each pair of sequences and therefore tend to obscure any conserved local regions (table 1). The conservation of small motifs in a specific order among distantly related proteins is well known, and these can serve as indicators of potential functions (Hanks et al. 1988). Distance values and percent identities for each of seven local regions (I, III–VII, and IX; data not shown) support the suggestion that the capsid-like amino acid segments found in the retroposons are more closely related to the HTLV-type retroviruses (table 1). Furthermore, the degree of sequence similarity within motifs is greater than that in other regions of the multiple alignment, indicating that the residues conserved between the retroviral capsids and the retroposons are not randomly dispersed throughout the length of the sequences but are found in discrete, local regions [for a discussion of approaches to statistical analysis of such alignments, see the work of McClure (accepted)].

Protease-like Region

Remnants of the three motifs characteristic of retroviral protease sequences (Sagata et al. 1984; Toh et al. 1985) can be found in the group II mitochondrial introns (fig. 4, I–III). The viral protease is considered to be an ancient member of the pepsin family, dating back prior to the duplication event that gave rise to the present-day double-domain aspartic acid protease family (Peral and Taylor 1987; Doolittle et al. 1989; Miller et al. 1989). In contrast to the capsid-like sequences of the non-LTR-TEs, the group II intron protease-like sequences are not more closely related either to any particular retrovirus or retrotransposon protease or to any member of the aspartic acid protease family (author's unpublished data). Experimental laboratory work will be necessary to determine whether these highly divergent protease-like sequences are

FIG. 7.—Multiple alignment of representative retrovirus IN sequences with retroposon sequences. Asterisks denote the conserved H/C motif, and regions I–IV denote additional conserved residues. Sequences from *INGI* and *R2BM* can be aligned with the complete viral IN. The H/C motif of *R2Bm* is not conserved; the first position of the first H is occupied by a P; and the two C residues have been replaced by S's. The group II yeast mitochondrial intron sequences terminate just 14–15 residues beyond the H/C motif. As there are only two sequences beyond the motif region, the boxing, color reversal, and shading rules described in fig. 3 are applied to matches in either of the two sequences.

functional. However, if this region is a protease, it is of the single-domain variety, consistent with its being related to the retroviral PR.

RH

The position of the residues responsible for the RH activity within the retroviral polymerase sequence was predicted via primary sequence comparisons, and, although these sequences are quite divergent, distinct motifs can still be found in common with the *Escherichia coli* RH (Johnson et al. 1986). Subsequent experimental studies confirmed the predicted position of the retroviral RT and RH sequences (Tanese and Goff 1988). Although the retroposon sequences are also quite divergent, remnants of the five conserved motifs of the RH are still evident in some members (fig. 5, I–V). As expected, the RH sequences of the I factor and ingi are closest to one another, consistent with their Z-, RT-, and CA-region relationships. On the basis of the same rationale, one of the other retroposons with an RH sequence should be the next nearest neighbor to the I factor/ingi pair. As illustrated by the pairwise distance scores for the I factor, the next closest relatives are unexpectedly found among the retroviruses (table 2). A second case of noncongruence between RT and RH relationships is found for the group II plasmid RH segment. It is more similar to the retrotransposons and caulimoviruses than to any other retroid RH sequence (table 2). The phylogenetic tree topologies for the RT, Z, and CA relationships are congruent (fig. 6*A–C*). In contrast, the RH tree topology reflects the values given in table 2. The I factor and ingi RH sequences appear to have shared a common ancestor with the retroviruses rather than with their closest relatives, other retroposons (fig. 6*A–C* vs. fig. 6*D*). Likewise, the Mauriceville plasmid RH sequence is significantly closer to the retrotransposons and DNA plant-virus RH sequences (fig. 6*A* vs. fig. 6*D*).

The T region, which connects the two functional domains of the polymerase (RT/T/RH), was found in several of the retroposons. Even among the retroviruses the T segment is highly divergent (author's unpublished data). Little T sequence similarity is discernible either among the retroposons of the nucleus (LINEs, Cin4, and I factor and ingi) or between this set and the group II intron segments.

IN

An H/C motif is found near the amino terminus of all known retroviral integration proteins and has been suggested to be a potential zinc-binding finger (Johnson et al. 1986). It is clearly present in the group II mitochondrial introns and in the ingi element (fig. 7). The R2Bm element can be aligned through this region, although the first H of the motif is a proline (P), the second H is present, and the 2 C residues have been replaced by serines (S) (fig. 7, asterisks). Both the ingi element and R2Bm elements have additional residues that can be aligned to the remaining portion of the viral IN and are no more similar to one another than they are to any other retroid IN (data not shown). However, several conserved clusters in addition to the H/C motif are found (fig. 7, I–IV). Mutation of the conserved P residue (fig. 7, III) in the RSV IN generates integration-deficient mutants (Quinn and Grandgenett 1989). Consideration of these experimental data and the conservation of the P-containing motif in both the ingi and R2Bm sequences supports the designation of these two regions as potential integration proteins. The R2Bm element has been shown to encode an endonuclease activity that exhibits integration site specificity for the 28S rRNA gene (Xiong and Eickbush 1988*a*). Another class of 28S insertional elements, R1Bm, has two ORFs. The first has been shown to contain an H/C region that can be aligned

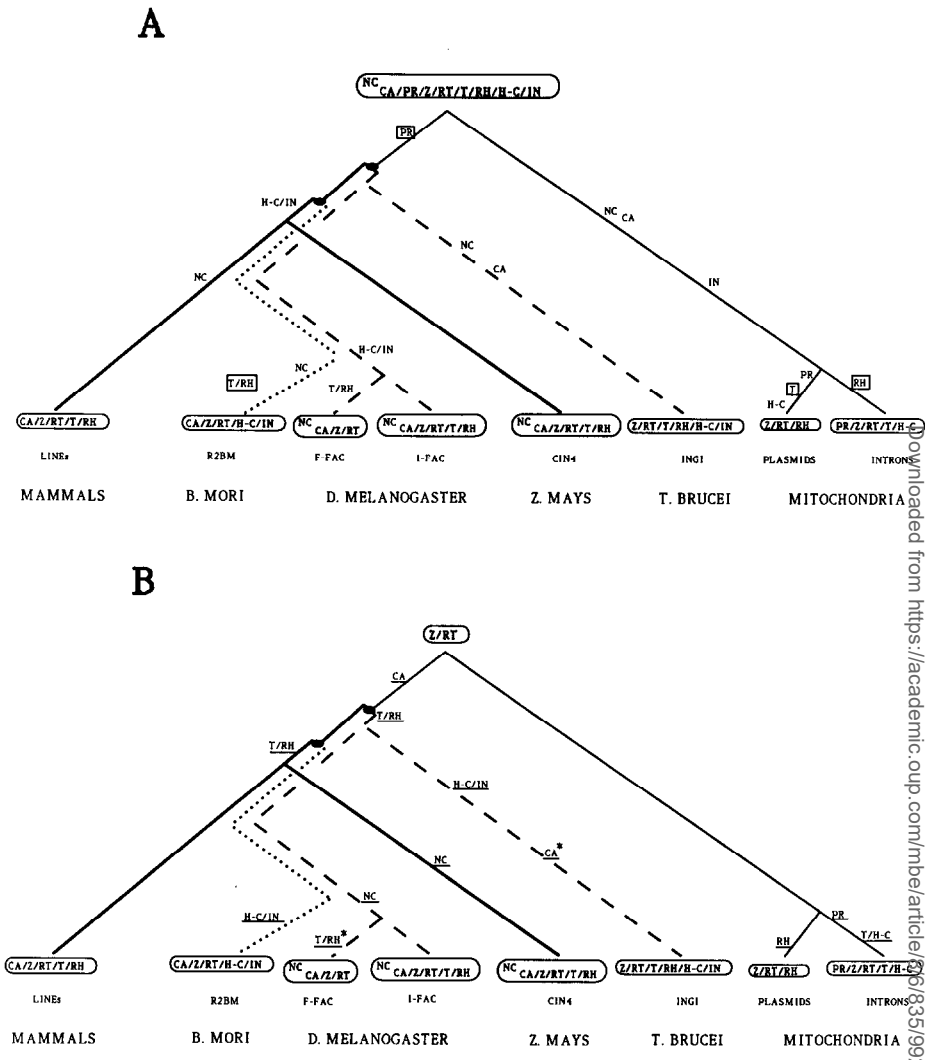


FIG. 8.—Schematic depiction of two possible modes by which present-day retroposons may have arisen. The order of the tips of the modular-gene-deletion (A) and modular-gene-acquisition (B) trees are based on species phylogeny. The RT gene phylogeny does not follow species phylogeny and is indicated by the branches leading to the tree tips. Two duplication events are invoked to account for the various retroposon RT sequence similarities. The thicker broken lines indicate the result of the duplication (black oval) that accounts for the INGI and F-FAC/I-FAC relationships. The thicker unbroken lines depict the result of the duplication (black oval) that accounts for the LINEs and CIN4 relationship. The dotted line indicates the fate of the duplication (black oval) of the R2BM RT sequence prior to the LINEs and CIN4 divergence. Branch lengths are arbitrary. The least number of events (deletion or acquisition of genes) required to generate the tree tips are indicated on each branch. To account for all the data, we must assume that some genes must have exited or entered the tree on more then one occasion. Tree A assumes that the ancestor was a virus-like element that progressively lost genes, thereby generating the descendant tips. This virus-like element differed from known retroviruses in that an NC is encoded in an alternative, nonoverlapping reading frame, to account for the C/H motif of the I and F factors and for the Cin4, as mentioned in Introduction. In addition, there is no envelope protein gene. Boxed segments indicate internal deletions. Tree B assumes that the ancestor was a small unit that progressively acquired the underlined retroviral-like genes, giving rise to the various gene arrangements at the tree tips. Asterisks denote the two end-deletion events that are necessary to account for the data. This model does not allow internal deletions, one functional segment to be swapped for an unrelated one (i.e., nonhomologous recombination), or xenologous recombination and would require several independent gene "capture" events, as illustrated. Of course, the actual ancestor could have been intermediate to these two extremes of possible scenarios.

with the viral NC motif. The second has an RT-like sequence (Xiong and Eickbush 1988*c*). Preliminary analysis of the R1Bm ORF2 indicates the presence of both a highly divergent CA-like sequence and a Z/RT segment. In contrast to my experience with R2Bm, I have been unable to detect an IN-like gene in R1Bm.

Discussion

The data presented here demonstrate that various retroposon lineages have conserved different sets of retroviral-like genes and characteristically contain a unique region, Z, upstream of their RT sequences (fig. 2). Previous studies of RT gene relationships indicate that the retroposons (as defined in the present study) form a discrete branch within the eukaryotic retroid family ((Xiong and Eickbush 1988*b*; Doolittle et al. 1989; McClure, accepted) and fig. 6*A*). My study shows that this topology is not only supported by the RT relationship but also by the presence of the Z region. In addition, the LINEs, R2Bm, Cin4, and I and F factors each contain a capsid-like sequence. The tree topologies generated from the multiple-alignment data for both the Z and CA regions are congruent with that of the RT gene phylogeny (fig. 6*A–C*).

At this point, there are two important measures of gene variability among the retroposons to consider; one is the absence or presence of a given segment, while the other is the degree of similarity among all homologues present. As mentioned above, while not all retroposons contain a CA-like region, those which do so corroborate the validity of the RT-based relationship. In contrast, the RH- and IN-like segments are highly variable in their absence or presence, as well as in their degree of relatedness among the retroposons. The RH sequences of the I factor and ingi are most closely related to one another. Their nearest neighbor, RH, should be among the other retroposons, as is the case for the RT and CA relationships, but they are more similar to several retrovirus RH sequences (table 2 and fig. 6*D*). In addition, the group II plasmid RH segment is more similar to the retrotransposons and caulimoviruses than to any other retroid RH sequences (table 2). This discrepancy can be accounted for in one of two ways: either the RH recombined with a nonretroposon RH, or the plasmid ancestor RT recombined with the intron RT. In the latter case, the plasmid RH would not be expected to be closer to the other retroposon RH sequences, and the plasmid RT gene would therefore be included in the retroposon lineage, because of recombination rather than by direct descent. The RH phylogenetic tree suggests that, for this gene, the group II plasmid shares a common ancestor with the retrotransposon and caulimovirus lineage (fig. 6*D*). The H-C/IN-like sequences are found in both R2Bm and ingi but not in the latters' respective nearest neighbors—LINEs and Cin4, and I and F factors. Furthermore, these two H-C/IN-like sequences are as distant from one another as they are from all other retroid integration proteins. These observations suggest that the evolutionary history of the retroposons involved genes that were acquired and/or lost, by xenologous recombination and/or independent gene assortment.

The variability of gene content and the data suggesting xenologous recombination or independent assortment of genes found in extant retroposons, as described above, can be used to address several questions: (1) At what point in evolutionary time and through what mode did these elements appear in the cellular genomes of different organisms (human, mouse, plant, insect, and parasite), on the one hand, and in the mitochondrial genome and plasmids of yeast, on the other hand? (2) What was the nature of the retroposon ancestor, and where did it originate? (3) Why are these sequences maintained in genomes?

Until recently, retroposons had only been found in either the yeast mitochondrion or the nuclear genomes of organisms, suggesting the possibility that their appearance in these two compartments occurred independently. However, the finding of an intron in a green-alga plastid gene with similarity to the RT of the group II yeast mitochondrial introns (Kuck 1989) suggests that retroposon RT sequences may be quite ancient.

There are only two ways that retroposons could have appeared in such a diversity of eukaryotes (mammals, insects, plants, and parasites) and organelles (mitochondria and plastids): either an initial copy was present in the ancestral genome prior to the divergence of all these organisms, or such elements have been spread throughout the eukaryotic world as a transmissible agent at various times in evolutionary history. Of course, these two modes are not mutually exclusive. There could have been an initial genomic retroposon with transmissible relatives that have entered and exited the genome at various points. The LINEs are present in all mammals, accounting for ~5% of the genomic DNA, suggesting that they were present in the ancestral mammalian genome (Fanning and Singer 1987).

Models for the descent of two hypothetical ancestors to the present-day retroposons are presented in figure 8. The topology of the tips of each tree assumes the presence of an ancestral retroposon (regardless of its composition) prior to eukaryotic and organelle divergence. Two duplications, either paralogous (i.e., duplication of genomic sequences) or xenologous (i.e., introduction of homologous foreign genes) (Fitch 1970; Gray and Fitch 1983), are invoked to account for the retroposon RT relationship which does not follow species phylogeny. These models assume that several classes of retroposon elements, some of which may have been lost, can exist simultaneously in the same genome. Consistent with such a notion is the presence of R2Bm- and R1Bm-like 28S insertional elements in other insects, including *Drosophila melanogaster* (Xiong and Eickbush 1988*c*).

The modular-gene-deletion tree illustrates the fate of a retrovirus-like element as it evolved to the extant retroposons but does not address the issue of how this ancestor initially arose (fig. 8A). This pathway assumes that a mechanism exists by which deletion events can occur, gene by gene, in a fairly precise manner because the various gene products arise by proteolytic cleavage of a single polyprotein. To date, there is no evidence for a cellular mechanism of this type. An internal deletion of PR initiates the first bifurcation, giving rise to the retroposons of the nucleus and to the yeast mitochondrial plasmids and introns. A series of end (10 total) and internal deletions (four total) as shown in figure 8A is sufficient to generate the retroposons from this virus-like ancestor. To account for the variable degree of relatedness mentioned above, the RH of the ancestral I factor and ingi must have originated by xenologous recombination with a member of the retrovirus lineage, and that of the group II plasmid could have originated from the retrotransposon and caulimovirus ancestor. Similarly, the H-C/IN sequences of ingi and R2Bm must have undergone independent xenologous recombinations with unidentified sources of integration proteins. The sequence similarity observed between the retroposon CA-like sequences and the HTLV-I CA suggests that the ancestor presented in this model either was related to an HTLV-like virus or recombined with one (table 1).

The modular-gene-acquisition tree presumes that the only segment common to all members of this group, the Z/RT region, was the ancestral unit which acquired retroviral-like genes to generate the present-day retroposons (fig. 8B). This model assumes that a mechanism exists by which genes may be appropriated in a modular fashion (i.e., as an intact gene unit). The ability of retroviruses to acquire intact genes

of cellular origin, the oncogenes, is well known. An initial acquisition of a CA segment, most likely from an HTLV-I/II ancestor after its divergence from other retroviruses, differentiates between the Z/RT lineage leading to the nuclear retroposons and that of the yeast mitochondrial elements. Subsequent accumulation of retroviral-like sequences around the hypothetical Z/RT unit requires 10 gene captures and two end-deletion events to generate the descendant tree tips (fig. 8B). Each of the retroposon sequence segments has been searched against the PIR data base, and the only significant similarities found are to retrovirus genes. This implies either that construction of the retroposons occurred via a retrovirus gene–specific mechanism or that a limited common gene pool was the only source of modules available to both retroposons and retroviruses.

The RT gene phylogeny of the eukaryotic retroid family clearly indicates that the retroposons last shared a common ancestor with the other retroid elements prior to the retroposons' divergence from one another (figs. 1 and 6*A*). If one considers these other retroid lineages, however, variability of gene order appears to give rise to a new lineage. Within each lineage, maintenance of gene order and homologue similarity appear to be the norm. For example, within each of the two retrotransposable-element groups, gene content and gene order are conserved (fig. 1). Furthermore, the degree of similarity calculated from multiple alignment of homologous proteins from any of these lineages does not reveal evidence for recombination (Doolittle et al. 1989). Among the retroviruses the general gene order is highly conserved, and recombination between distantly related retroviruses is rare (Clark and Mak 1984; McClure, et al. 1987, 1988). The lack of rampant gene variation and recombination in the other retroid-family lineages, as compared with that of the retroposon branch, is consistent with the evolution of a simple Z/RT unit through modular gene acquisition to the extant retroposons.

The simplest idea, consistent with all the data thus far accumulated, would be that the retroid family has coevolved with eukaryotic systems and that different retroid lineages have evolved to become viruses (RNA and DNA), various transposable elements (both LTR and non-LTR containing), and group II introns and plasmids of organelles (fig. 1). It has been suggested that LINEs may play a role in differentiation and development, perhaps by way of gene inactivation/activation (Fanning and Singer 1987). The more ancient members of the retroid family, therefore, may be intimately associated with basic cellular processes while also providing a recombinational pool through which retroid elements can cycle and new elements can arise, the most successful of which happen to have a retroviral-like gene order. The retroposons could be descendants of one lineage of this combinatoric pool, descendants which by chance have survived at various stages of evolution. The retroid-family descendants may reflect the RNA genetic pool that was instrumental in the conversion of RNA-based systems to DNA-based ones. The relative lack of retroid elements in the prokaryotic world would be analogous to the lack of introns among prokaryotes, and such differential distribution may be a feature of those cellular elements of ancient origin.

Acknowledgments

I would like to thank R. Doolittle and M. Waterman for providing the computer facilities necessary for this work. I am grateful to J. Perrault, R. Hudson and W. Fitch for constructive criticisms on the manuscript. Support was provided by NIH grant AI 28309.

LITERATURE CITED

BALTIMORE, D. 1985. Retroviruses and retrotransposons: the role of reverse transcriptase in shaping the eukaryotic genome. Cell **40**:481–482.

BOEKE, J. D., and V. G. CORCES. 1989. Transcription and reverse transcription of retrotransposons. Annu. Rev. Microbiol. **43**:403–434.

BURKE, W. D., C. C. CALALANG, and T. H. EICKBUSH. 1987. The site-specific ribosomal insertion element type II of *Bombyx mori* (R2Bm) contains the coding sequence for a reverse transcriptase-like enzyme. Mol. Cell. Biol. **7**:2221–2230.

CLARK, S. P., and T. W. MAK. 1984. Fluidity of a retrovirus genome. J. Virol. **50**:759–765.

COVEY, S. N. 1986. Amino acid sequence homology in gag region of reverse transcribing elements and the coat protein gene of cauliflower mosaic virus. Nucleic Acids Res. **14**:623–633.

DARNELL, J. E., and F. W. DOOLITTLE. 1986. Speculations on the early course of evolution. Proc. Natl. Acad. Sci. USA **83**:1271–1275.

DAYHOFF, M. O. 1978. A model of evolutionary change in proteins: detecting distant relationships. Pp. 345–358 *in* M. O. DAYHOFF. Atlas of protein sequence and structure. National Biomedical Research Foundation. Washington, D.C.

DI NOCERA, P. P. 1988. Close relationships between non-viral retroposons in *Drosophila melanogaster*. Nucleic Acids Res. **16**:4041–4052.

DI NOCERA, P. P., and G. CASARI. 1987. Related polypeptides are encoded by *Drosophila* P elements, I factors, and mammalian L1 sequences. Proc. Natl. Acad. Sci. USA **84**:5843–5847.

DOOLITTLE, R. F. 1987. Of urfs and orfs: a primer on how to analyze derived amino acid sequences. University Science, Mill Valley, Calif.

DOOLITTLE, R. F., D.-F. FENG, M. S. JOHNSON, and M. A. MCCLURE. 1989. Origins and evolutionary relationships of retroviruses. Q. Rev. Biol. **64**:1–30.

EIGEN, M., and P. SCHUSTER. 1982. Stages of emerging life—five principles of early organization. J. Mol. Evol. **19**:47–61.

FANNING, T. G., and M. F. SINGER. 1987. Line-1: a mammalian transposable element. Biochim. Biophys. Acta **910**:203–212.

FAWCETT, D. H., C. K. LISTER, E. KELLETT, and D. J. FINNEGAN. 1986. Transposable elements controlling I-R hybrid dysgenesis in D. melanogaster are similar to mammalian LINEs. Cell **47**:1007–1015.

FENG, D.-F., and R. F. DOOLITTLE. 1987. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. J. Mol. Evol. **25**:351–360.

FENG, D.-F., M. S. JOHNSON, and R. F. DOOLITTLE. 1985. Aligning amino acid sequences: comparison of commonly used methods. J. Mol. Evol. **21**:112–125.

FINNEGAN, D. J. 1983. Retroviruses and transposable elements—which came first? Nature **302**: 105–106.

———. 1989. The I factor and I-R hybrid dysgenesis in Drosophilia melanogaster. Pp. 503–515 *in* D. E. BERG and M. M. HOWE. Mobile DNA. American Society for Microbiology. Washington, D.C.

FITCH, W. M. 1970. Distinguishing homologous from analogous proteins. Syst. Zool. **19**:99–113.

FITCH, W. M., and E. MARGOLIASH. 1967. Construction of phylogenetic trees. Science **15**:279–284.

FUETTERER, J., and T. HOHN. 1987. Involvement of nucleocapsids in reverse transcription: a general phenomenon? Trends Biochem. Sci. **12**:92–95.

GORELICK, R. J., L. E. HENDERSON, J. P. HANSER, and A. REIN. 1988. Point mutants of Moloney murine leukemia virus that fail to package viral RNA: evidence for specific RNA recognition by a "zinc finger-like" protein sequence. Proc. Natl. Acad. Sci. USA **85**:840–844.

GRAY, G. S., and W. M. FITCH. 1983. Evolution of antibiotic resistance genes: the DNA sequence of a kanamycin resistance gene from *Staphylococcus Aureus*. Mol. Biol. Evol. **1**:57–66.

HANKS, S. K., A. M. QUINN, and T. HUNTER. 1988. The protein kinase family: conserved features and deduced phylogeny of the catalytic domains. Science **241**:42–52.

HUTCHISON, C. A., S. C. HARDIES, D. D. LOEB, W. R. SHEHEE, and M. H. EDGELL. 1989. LINEs and related retrotoposons: long interspersed repeated sequences in the eucaryotic genome. Pp. 593–617 *in* D. E. BERG and M. M. HOWE, eds. Mobile DNA. American Society for Microbiology, Washington, D.C.

INOUYE, S., M.-Y. HSU, S. EAGLE, and M. INOUYE. 1989. Reverse transcriptase associated with the biosynthesis of the branched RNA-linked msDNA in Myxococcus xanthus. Cell **56**:709–717.

JENTOFT, J. E., L. M. SMITH, X. D. FU, M. JOHNSON, and J. LEIS. 1988. Conserved cysteine and histidine residues of the avian myeoblastosis virus nucleocapsid protein are essential for viral replication but are not "zinc-binding fingers". Proc. Natl. Acad. Sci. USA **85**:7094–7098.

JOHNSON, M. S., M. A. MCCLURE, D.-F. FENG, J. GRAY, and R. F. DOOLITTLE. 1986. Computer analysis of retroviral pol genes: assignment of enzymatic functions. Proc. Natl. Acad. Sci. USA **83**:7648–7652.

KIMMEL, B. E., O. K. OLE-MOIYOI, and J. R. YOUNG. 1987. Ingi, a 5.2-kb dispersed sequence element from *Trypanosoma brucei* that carries half of a smaller mobile element at either end and has homology with mammalian LINEs. Mol. Cell. Biol. **7**:1465–1475.

KUCK, U. 1989. The intron of a plastid from a green alga contains an open reading frame for a reverse transcriptase-like enzyme. Mol. Gen. Genet. **218**:257–265.

KUIPER, M. T. R., and A. M. LAMBOWITZ. 1988. A novel reverse transcriptase activity associated with mitochondrial plasmids of Neurospora. Cell **55**:693–704.

LAMPSON, B. C., M. INOUYE, and S. INOUYE. 1989. Reverse transcriptase with concomitant ribonuclease H activity in the cell-free synthesis of branched RNA-linked msDNA of Myxococcus xanthus. Cell **56**:701–707.

LANG, B. F., and F. AHNE. 1985. The mitochondrial genome of the fission yeast *Schizosaccharomyces pombe:* the cytochrome b gene has an intron closely related to the first two introns in the *Saccharomyces cerevisiae* cox1 gene. J. Mol. Biol. **184**:353–366.

LIM, D., and W. K. MAAS. 1989. Reverse transcriptase–dependent synthesis of a covalently linked, branched DNA-RNA compound in E. coli B. Cell **56**:891–904.

MCCLURE, M. A. Sequence analysis of retroid proteins. Adv. Math. Comput. Med. (accepted).

MCCLURE, M. A., M. S. JOHNSON, and R. F. DOOLITTLE. 1987. Relocation of a protease-like gene segment between two retroviruses. Proc. Natl. Acad. Sci. USA **84**:2693–2697.

MCCLURE, M. A., M. S. JOHNSON, D.-F. FENG, and R. F. DOOLITTLE. 1988. Sequence comparisons of retroviral proteins: relative rates of change and general phylogeny. Proc. Natl. Acad. Sci. USA **85**:2469–2473.

MCCLURE, M. A., and J. PERRAULT. 1989. Two domains distantly related to protein-tyrosine kinases in the vesicular stomatitis virus polymerase. Virology **172**:391–397.

MICHEL, F., and B. F. LANG. 1985. Mitochondrial class II introns encode proteins related to the reverse transcriptase of retroviruses. Nature **316**:641–643.

MILLER, M., M. JASKOLSKI, J. K. M. RAO, J. LEIS, and A. WLODAWER. 1989. Crystal structure of a retroviral protease proves relationship to aspartic protease family. Nature **337**:576–579.

MILLER, R. H., and W. S. ROBINSON. 1986. Common evolutionary origin of hepatitis B virus and retroviruses. Proc. Natl. Acad. Sci. USA **83**:2531–2535.

NATVIG, D. O., G. MAY, and J. W. TAYLOR. 1984. Distribution and evolutionary significance of mitochondrial plasmids in *Neurospora* spp. J. Bacteriol. **159**:288–293.

NEEDLEMAN, S. B., and C. D. WUNSCH. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. **48**:443–453.

PERAL, L. H., and W. R. TAYLOR. 1987. A structural model for the retroviral proteases. Nature **329**:351–354.

QUINN, T. P., and D. P. GRANDGENETT. 1989. Avian retrovirus integration protein: structure-functional analysis of viable mutants. Virology **173**:478–488.

ROGERS, J. 1985. The origin and evolution of retroposons. Int. Rev. Cytol. **93**:187–279.

ROGERS, J. 1986. The origins of retroposons. Nature **319**:725.

SAGATA, N., T. YASUNAGE, and Y. IKAWA. 1984. Identification of a potential protease-coding gene in the genomes of bovine leukemia and human T-cell leukemia viruses. FEBS Lett. **178**:79–82.

SCHWARTZ, R. M., and M. O. DAYHOFF. 1978. Matrices for detecting distant relationships. Pp. 353–358 *in* M. O. DAYHOFF, ed. Atlas of protein sequence and structure. National Biomedical Research Foundation, Washington, D.C.

SCHWARZ-SOMMER, Z., L. LECLERCQ, E. GOBEL, and H. SAEDLER. 1987. Cin4, an insert altering the structure of the *A1* gene in *Zea mays*, exhibits properties of nonviral retroposons. EMBO J. **6**:3873–3880.

SKOWRONSKI, J., T. G. FANNING, and M. F. SINGER. 1988. Unit-length LINE-1 transcripts in human teratocarcinoma cells. Mol. Cell. Biol. **8**:1385–1397.

TANESE, N., and S. P. GOFF. 1988. Domain structure of the Moloney murine leukemia virus reverse transcriptase: mutational analysis and separate expression of the DNA polymerase and RNAase H activities. Proc. Natl. Acad. Sci. USA **85**:1777–1781.

TEMIN, H. M. 1970. Malignant transformation of cells by viruses. Perspect. Biol. Med. **14**:11–26.

———. 1980. Origin of retroviruses from cellular moveable genetic elements. Cell **21**:599–600.

———. 1985. Reverse transcription in eukaryotic genome: retroviruses, pararetroviruses, retrotransposons, and retrotranscripts. Mol. Biol. Evol. **2**:455–468.

———. 1989. Retrons in bacteria. Nature **339**:252–255.

TEMIN, H. M., and W. ENGELS. 1984. Movable genetic elements and evolution. Pp. 173–201 *in* J. W. POLLARD, ed. Evolutionary prospects in the 1980s. John Wiley, Chichester.

TOH, H., M. ONO, K. SAIGO, and T. MIYATA. 1985. Retroviral protease-like sequence in the yeast transposon Ty1. Nature **315**:691.

VARMUS, H., and P. BROWN. 1989. Retroviruses. Pp. 53–108 *in* D. E. BERG and M. M. HOWE, eds. Mobile DNA. American Society for Microbiology, Washington, D.C.

WEINER, A. M., P. L. DEININGER, and A. EFSTRATISDIS. 1986. Nonviral retroposons, genes, pseudogenes, and transposable elements generated by the reverse transcriptase flow of genetic information. Annu. Rev. Biochem. **55**:631–661.

XIONG, Y., and T. H. EICKBUSH. 1988*a*. Functional expression of a sequence-specific endonuclease encoded by the retrotransposon R2Bm. Cell **55**:235–246.

———. 1988*b*. Similarity of reverse transcriptase-like sequences of viruses, transposable elements, and mitochondrial introns. Mol. Biol. Evol. **5**:675–690.

———. 1988*c*. The site-specific ribosomal DNA insertion element R1Bm belongs to a class of non-long-terminal-repeat retrotransposons. Mol. Cell. Biol. **8**:114–123.

HOWARD TEMIN, reviewing editor

Received August 22, 1990; revision received March 22, 1991

Accepted March 26, 1991