Similarity of Reverse Transcriptase–like Sequences of Viruses, Transposable Elements, and Mitochondrial Introns[1]

Yue Xiong and Thomas H. Eickbush

Department of Biology, University of Rochester, River Campus

Sequences similar to reverse transcriptase (RT) of retroviruses have been found in certain DNA viruses, mitochondrial intron sequences, and a wide variety of transposable elements. While total amino acid similarity between these diverse elements is quite low, we have identified seven regions, consisting of 182 amino acids, that are common to all elements. Highly conserved residues identified in each of these regions are diagnostic for the identification and alignment of these and for future RT-like sequences. Using both the neighbor-joining and the unweighted-pair-group methods, we have derived a probable phylogenetic tree for all RT-containing elements. These elements can be divided into two major groups. Retroviruses and DNA viruses whose propagation involves an RNA intermediate are grouped with a series of transposable elements containing long terminal repeats (LTRs). The second group is made up of RT-containing sequences of fungal mitochondrial introns and a series of transposable elements that lack LTRs. The transposable elements, copia and Ty, were found to be the most difficult to position on the phylogenetic tree, as a result of their higher rate of sequence divergence. The data are most consistent with their being distant members of the LTR group (retroviruses/LTR retrotransposons).

Introduction

Retroviruses comprise a large family of animal viruses distinguished by the property that their single-strand RNA genomes are replicated through a DNA intermediate by the virus-encoded enzyme reverse transcriptase (RT). Although retroviruses differ from one another in detail, they all share a common structure in which three essential genes, *gag, pol,* and *env,* are flanked by two long terminal repeats (LTRs) (reviewed in Varmus 1983). The *pol* gene encodes several different enzymatic activities including RNase H, RT, and integrase (Varmus 1983). The RT region of the *pol* gene is the most highly conserved sequence of the retroviral genome and has been used to determine the phylogenetic relationship among retroviruses (Chiu et al. 1985; Sagata et al. 1985; Sonigo et al. 1986; McClure et al. 1988; Yokoyama et al. 1988) as well as between these viruses and certain *Drosophila* retrotransposons (Toh et al. 1985; Yuki et al. 1986*b*). RT-like sequences have been identified in many other widely different elements, including plant and animal DNA viruses (Toh et al. 1983), the mammalian L1 repetitive-DNA family (Hattori et al. 1986; Loeb et al. 1986), transposable elements found in fruit flies, yeast, trypanosomes, and slime mold (Saigo et al. 1984; Cappello et al. 1985; Clare and Farabaugh 1985; Mount and Rubin 1985; Fawcett et al. 1986;

[1] Key words/abbreviations: RT = reverse transcriptase; LTR = long terminal repeat; ORF = open reading frame.

Address for correspondence and reprints: Dr. Thomas H. Eickbush, Department of Biology, University of Rochester, Rochester, New York 14627; Telephone (716) 275-7247.

Mol. Biol. Evol. 5(6):675–690. 1988.

0737-4038/88/0506-0005$02.00

Inouye et al. 1986; Marlor et al. 1986; Yuki et al. 1986*a*; Di Nocera and Casari 1987; Kimmel et al. 1987), and even in mitochondrial plasmid and intron sequences of fungi (Michel and Lang 1985; Matsuura et al. 1986). While the role of reverse transcription has not been established in the life cycle of each of these elements, evidence for reverse transcription has been obtained in the propagation of elements from most major groups, including DNA viruses (Summers and Mason 1982; Pfeiffer and Hohn 1983), transposable elements (Shiba and Saigo 1983; Boeke et al. 1985; Garfinkel et al. 1985), and a mitochondrial plasmid (Akins et al. 1986).

We have recently identified two insertion elements in the 28S ribosomal genes of the silkmoth *Bombyx mori* that contain open reading frames with similarities to RT (Burke et al. 1987; Xiong and Eickbush 1988). The similarity was greatest to five other recently described retrotransposable elements, the I, F, and G elements of *D. melanogaster* (Fawcett et al. 1986; Di Nocera and Casari 1987; Di Nocera 1988), L1 repetitive DNA of mammalia (Hattori et al. 1986; Loeb et al. 1986), and ingi of *Trypanosoma brucei* (Kimmel et al. 1987). These seven elements share the property of not containing LTRs. Since LTRs are common to all retroviruses and other retrotransposons, we have previously suggested that these elements form a distinct group, termed the non-LTR retrotransposons (Xiong and Eickbush 1988). To examine how the elements of this group and all other RT-containing elements might be evolutionarily related, we have conducted a comprehensive amino acid sequence comparison of RT-like sequences.

Sequence Sources

Sources and full names of the elements compared are as follows: R1Bm, type I ribosomal insert from *Bombyx mori* (Xiong and Eickbush 1988); R2Bm, type II ribosomal insert from *B. mori* (Burke et al. 1987); L1Md, LINE 1 repetitive element from *Mus domesticus* (Loeb et al. 1986; Shehee et al. 1987); ingi, mobile element from *Trypanosoma brucei* (Kimmel et al. 1987); Sc-a1 and Sc-a2, introns a1 and a2 of *Saccharomyces cerevisiae* mitochondrial cytochrome oxidase subunit I gene (Bonitz et al. 1980); Pa-a1, intron of *Podospora anserina* mitochondrial cytochrome oxidase subunit I gene (Osierwacz and Esser 1984); Pa-IA, intron A of *P. anserina* mitochondrial cytochrome oxidase subunit I gene (Matsuura et al. 1986), Sp-b1, intron of *Schizosaccharomyces pombe* mitochondrial cytochrome b gene (Lang et al. 1985); Nc-p1, mitochondrial plasmid of the mauriceville-1c strain of *Neurospora crassa* (Nargang et al. 1984); Ty, mobile element of *S. cerevisiae* (Clare and Farabaugh 1985); 17.6; copia, 297, gypsy, 412, I, F, and G elements from *D. melanogaster* (Saigo et al. 1984; Mount and Rubin 1985; Fawcett et al. 1986; Inouye et al. 1986; Marlor et al. 1986; Yuki et al. 1986*a;* Di Nocera and Casari 1987; Di Nocera 1988); CaMV, cauliflower mosaic virus (Gardner et al. 1981); DIRS-1, mobile element from *Dictyostelium discoideum* (Cappello et al. 1985); HBV, human hepatitis B virus (Galibert et al. 1979); DHBV, duck hepatitis B virus (Mandart et al. 1984); MuLV, Moloney murine leukemia virus (Shinnick et al. 1981); HERV, human endogenous retroviral DNA (Repaske et al. 1985); RSV, Rous sarcoma virus (Schwartz et al. 1983); HSRV, human spumaretrovirus (Maurer et al. 1988); MMTV, mouse mammary tumor virus (Chiu et al. 1985); IAP, Syrian hamster intracisternal A-particle (Ono et al. 1985); HTLV-I, human adult T-cell leukemia virus (Seiki et al. 1983); HTLV-II, human T-cell leukemia virus type II (Shimotohno et al. 1985); BLV, bovine leukemia virus (Sagata et al. 1985); HIV-I, human immunodeficiency virus (Wain-Hobson et al. 1985); HIV-II, human immunodeficiency virus type 2 (Guyader et al. 1987); Visna, ovine visna virus (Sonigo et al. 1985); SRV, simian acquired immune deficiency syndrome virus (Power et al. 1986); and EIAV, equine infectious anemia virus (Chiu et al. 1985).

Sequence Alignment

The level of sequence similarity between the different RT-like sequences is quite low. This is true not only between different classes of elements (e.g., retroviruses vs. transposable elements) but even between elements of the same class. For example, amino acid similarity of the RT region between the mammalian retroviruses MuLV and Visna is only 25%. Because of this low sequence similarity and the absence of computer programs capable of the simultaneous alignment of such a large number of sequences, two assumptions were necessary. The first assumption was that initial sequence alignments could be based upon groups of conserved amino acid residues that have been identified to various degrees in all published RT-like sequences. The residues used were a modification of those originally identified by Toh et al. (1983, 1985) in retroviral, HBV, CaMV, and several *Drosophila melanogaster* retrotransposon sequences. With the exception of copia and Ty (discussed in greater detail below), identification of this original set of conserved residues in the 37 RT-like sequences discussed in the present report were unambiguous—and in most cases were in agreement with similar identifications made to various degrees in previous reports. The number of amino acids separating these groups of conserved residues, while similar within the same class of element, varied considerably between the different classes. This variation in the number of residues between the fixed amino acid positions led to the second assumption used in sequence alignment; that is, the alignment between the fixed residues was adequately conducted using algorithms that confer a substantial penalty for the insertion of gaps. The method used, the unitary matrix (UM) method (Doolittle 1981; Feng et al. 1985), assigns a score of 2.0, 1.0, and 0 to matched cysteines, other amino acids matches, and mismatches, respectively. A −2.5 penalty is incorporated when a gap, regardless of size, is introduced. Since the value of this penalty is somewhat arbitrary, we also conducted the alignment with a penalty of −1.5/gap. With either penalty the UM method resulted in seven major regions that are common to all sequences, with virtually all gaps between these regions localized to the same positions in different elements (fig. 1). If the penalty for the insertion of gaps was reduced to a value of −1.0, it was necessary to introduce multiple gaps between the groups of fixed residues. Thus, in the absence of computer programs that can simultaneously analyze all sequences, it was not possible to obtain an optimum alignment by using this lower gap penalty.

The alignment in figure 1 contains seven regions that are common to all elements (boxed regions numbered 1–7). The amino terminal border of box 1 and the carboxyl-terminal border of box 7 were defined by a lack of similarity between different types of elements. All other borders of the boxed regions are defined by gaps in the sequence of one or more groups of elements. Each of the boxed regions is defined by a series of conserved amino acid positions. Conserved residues found in the non-LTR retrotransposable elements and mitochondrial sequences are shown at the top of the sequences, while conserved residues found in retroviruses and LTR retrotransposable elements (but not always by copia or Ty) are indicated at the bottom of the sequences. Approximately one-half of these residues (indicated with an asterisk) are shared in all four classes of elements. This newly defined set of conserved residues should be useful

```
                             1                                                  2                                                                                  3
                                      + + K+                  RP+++      K+          +                GF     +        +                                    FD +     +   +
R1Bm    494. RCVVEGTFPPVWKDGRLLVLPKGNGR PLTDPKA YRPVTLLPVLGKILEKVLLQCAPGLTHSI ---SPRQHGFSPGRSTVTALRTLLDVSRAS--- EQRYVMAIFL[illegible]AFDNAWWPMIMVKAKRNCPP NIY--RMLTDYFRG
R2Bm    467. AWMARGEIPEILRQCRTVFVPKVERP --GGPGE YRPISIASIPLRHFHSILARRLLACCPPD ----ARQRGFICADGTLENSAVLDAVLGDSRK- KLRECHVAVL[illegible]A[illegible]AFDTVSHEALVELLRLRGMP EQFCGYIAHLYDTA
L1Md    523. KIEVEGTLPNSFYEATITLIPKPQKD -PTKIEN FRPISIMNIDAKILNKILANRIQEHIKAI --IHPDQVGF     MQGWFNIRKSINVIHYINKL KDKNHMIISL[illegible]A[illegible]AFDKIQHPFMIKVLERSGIQ GPYLNMIKAIYSKP
I       342. NEIFNSHIPQAYKTSLIIPILKPNTD -KTKTSS YRPISLNCCIAKILDKIIAKRLWWLVTYN NLINDKQFGF[illegible]TSDCLLYVDYLITK---- SKMHTSLVTLDFSRAFDRVGVHSIIQQLQEWKTG PKIIKYIKNFMSNR
Ingi    178. ESLRTGVVPPAWKTGVIIPILKAGKK -AEDLDS YRPVTLTSCLCKVMERIIAARPRDTVESQ --LTPQQSGF     TLEQLLHVRAALCHHT-- HQYRTGAVFVDYEKAFDTVDHDKIAREMHRMKVS PHIVKWCVSFLSNR
F       479. AITKLGYFPQRWKMMKIIMIPKPGKN -HTVASS YRPISLLSCISKLFEKCLLIRLNQHQTYH NIIPAHQFGF     TIEQVNRITTEIRTAFE- YREYCTAVFLDVSQAFDKVWLDGLMFKIKISLPE STH-KLLKSYLYDR
G         ? . SCFRLGYFPKQ*KRAEVITIPKPGKP -EANLAS YRPISLLAILSKILERVFLRRVLPVLDEA GLIPDHQFGF     TPEQCHRLVEQILEAFE- RKQYCCAVMLDVKQAFDKVWHPGLHYKIKTHLPG SHF-AFLKSFTEGR

Sc-a1   239. LSNELGTGKFKFKPMRIVNIPKPKGG ------- IRPLSVGNPRDKIV[illegible]VIRIILDTIFDKK --ISTHSHGFRKNISCQTAIWEVRNI------- FGGSNWF[illegible]LKKCFDTISHDLIIKELKRYISD KGFIDLVYKLLRAG
Sc-a2   259. LSKDINTNMFKFSPVRRVEIPKTSGG ------- FRPLSVGNPREKIV[illegible]SMRIILEIIYNNS --FSYYSHGFRPNLSCLTAIIQCKNY------- MQYCNWF[illegible]LNKCFDTIPHNMLINVLNERIKD KGFIDLLYKLLRAG
Pa-a1   227. ISEQLKSEQFRFRPTRRVYIPKANGK ------- MRPLGIASPRDKIV[illegible]VFRAILEQVLEPR --FHSSSHGFRPGRGCHSALATIRY-------- WNGIKQF[illegible]IKGFFDNIDHHILEKLLVKHFQD QRFIDLYWKMVKAG
Pa-IA   239. LILELKSERFKFTSVKRVYIPKANGK ------- TRPLGIPTSKDKIV[illegible]AMKILLELIYEPI --FLDVSHGFRPKRSCHTALHQISK-------- WNGTTWM[illegible]IKGFFNEVDHQVLIKILEKKIKD QRFFDLLWKIFRAG
Sp-b1   280. IIESLKSEEFNFTPGRRILIDKASGG ------- KRPLTIGSPRDKLV[illegible]ILRIVLEAIYEPL --FNTASHGFRPGRSCHSALRSIFTN------- FKGCTWW[illegible]IKACFDSIPHDKLIALLSSKIKD QRFIQLIRKALNAG
Nc-p1   109. EVREMVEIQPVCIDYKRVYIPKANGK ------- QRPLGVPTVPWRV[illegible]MWNVLLVWYRIPE ---QDNQHAYFPKRGVFTAWRALWP-------- KLDSQNI[illegible]LKNFFPSVDLAYLKDKLMESGIP QDISEYLTVL.44.

Copia   927. RPENKNIVDSRWVFSVKYNELGNPIR -----YK ARLVARGFTQKYQIDYEETFAPVARISSF --RFILSLV------------------------ IQYNLKVHQMDVKTAFLNGTLKEEIYMRLPQGIS CNS-----------
Ty      844. YYDRKEIDPKRVINSMFIFNKKRDGT -----HK RRFVARGDIQHPDTYDSGMQSNTVHHYAL --MTSLSLA------------------------ LDNNYYITQLDISSAYLYADIKEELYIRPPPHLG MN------------

17.6    230. LNQGIIRTSNSPYNSPIWVVPKKQDA --SGKQK FRIVIDYRKLNEITVGDRHPIPNMDEILG --------------------------------K LGRCNYFTTIDLAKGFHQIEMDPESVSKTAFSTK --------------
297     229. LNQGLIRESNSPYNSPTWVVPKKPDA --SGANK YRVVIDYRKLNEITIPDRYPIPNMDEILG --------------------------------K LGKCQYFTTIDLAKGFHQIEMDEESISKTAFSTK --------------
Gypsy   204. LKDGIIRPSRSPYNSPTWVVDKKGTD -AFGNPN KRLVIDFRKLNEKTIPDRYPMPSIPMILA --------------------------------N LGKAKFFTTLDLKSGYHQIYLAEHDREKTSFSVN --------------
412     337. IKDKIVEPSVSQYNSPLLLVPKKSSP -NSDKKK WRLVIDYRQINKKLLADKFPLPRIDDILD --------------------------------Q LGRAKYFSCLDLMSGFHQIELDEGSRDITSFSTS --------------
CaMV    270. LDLKVIKPSKSPHMAPAFLVNNEAEK ---RRGK KRMVVNYKAMNKATIGDAYNLPNKDELLT --------------------------------L IRGKKIFSSFDCKSGFWQVLLDQESRPLTAFTCP --------------
DIRS1    92. EQVLPNHYSKRVFYSNVFTVPKPGTN ------L HRPVLDLKRLNTYINNQSFKMEGIKNLPS --------------------------------M VKQGYYMVKLDIKKAYLHVLVDPQYRDLFRFVWK --------------

HBV      53. EHNIRIPRTPARVTGGVFLVDKNPHN ----TTE SRLVVDFSQFSRGSTHVSWPKFAVPNLQS -----------------------------LTNL LSSNLSWLSLDVSAAFYHIPLHPAAMPHLLVGSS GLPRYVARLS.47.
DHBV    425. WYLRGNTSWPNRITGKLFLVDKNSRN ----TEE ARLVVDFSQFSKGKNAMRFPRYWSPNLST -----------------------------LRRI LPVGMPRISLDLSQAFYHLPLNPASSSRLAVSDG --------------
MuLV    201. LDQGILVPCQSPWNTPLLPVKKPGTN ------D YRPVQDLREVNKRVEDIHPTVPNPYNLLS -------------------------------GL PPSHQWYTVLDLKDAFFCLRLHPTSQPLFAFEWR DPEM----------
HERV    201. RTFRIIVPCQSPWNTPLLPVPKPGTK ------- YRPVQDLRLVNQATVTLHPTVPNLYTLLG -------------------------------LL PAEDSWFTCLDLKDAFFSIRLAPERQKLFAFQWE DPES----------
RSV      40. LQLGHIEPSLSCWNTPVFVIRKASGS ------- YRLLHDLRAVNAKLVPFGAVQQGAPVLSA --------------------------------- LPRGWPLMVLDLKDCFFSIPLAEQDREAFAFTLP SVNNQ---------
MMTV     49. LQLGHLEESNSPWNTPVFVIKKKSGK ------- WRLLQDLRAVNATMHDMGALQPGLPSPVA --------------------------------- VPKGWEIIIIDLQDCFFNIKLHPEDCKRFAFSVP SPNFK---------
HTLV1    47. LEAGHIEPYTGPGNNPVFPVKKANGT ------- WRFIHDLRATNSLTIDLSSSSPGPPDLSS --------------------------------L PTTLAHLQTIDLRDAFFQIPLPKQFQPYFAFTVP QQCNY---------
HTLV2   132. LEAGHIEPYSGPGNNPVFPVKKPNGK ------- WRFIHDLRATNAITTTLTSPSPGPPDLTS --------------------------------L PTALPHLQTIDLTDAFFQIPLPKQYQPYFAFTIP QPCNY---------
BLV      21. LEAGYISPWDGPGNNPVFPVRKPNGA ------- WRFVHDLRATNALTKPIPALSPGPPDLTA --------------------------------I PTHPPHIICLDLKDAFFQIPVEDRFRFYLSFTLP SPGGL---------
HIV1    198. EGKISKIGPENPYNTPVFAIKKKDST ------K WRKLVDFRELNKRTQDFWEVQLGIPHPAG --------------------------------- LKKKKSVTVLDVGDAYFSVPLDEDFRKYTAFTIP SINNE---------
HIV2    227. EGQLEEAPPTNPYNTPTFAIKKKDKN ------K WRMLIDFRELNKVTQDFTEIQLGIPHPAG --------------------------------- LAKKRRITVLDVGDAYFSIPLHEDFRPYTAFTLP SVNNA---------
Visna   185. EGKVGRAPPHWTCNTPIFCIKKKSGK ------- WRMLIDFRELNKQTEDLAEAQLGLPHPGG --------------------------------- LQRKKHVTILDIGDAYFTIPLYEPYRQYTCFTML SPNNL---------
EIAV     23. EGKISEASDNNPYNSPIFVIKKKSGK ------- WRLLQDLRELNKTVQVGTEISRGLPHPGG --------------------------------- LIKC[illegible]VVLDIGDAYFTIPLDPEFRPYTAFTIP SINHQ---------
IAP      40. ERLGHLEPSTSPWNTPIFVIKKKSGK ------- WRLLHDLRAINNQMHLFGPVQRGLPLLSA --------------------------------- LPQ[illegible]IIDIKDCFFSIPLYPRDRPRFAFTIP SLNHM---------
HSRV     12. LKQGVLTPQNSTMNTPVYPVPKPDGR ------- WRMVLDYREVNKTIPLTAAQNQHSAGILA --------------------------------T IVRQ[illegible]TLDLANGFWAHPITPESYWLTAFTWQ --------------
SRV      54. LEAGHITESNSPWNTPIFVIKKKSGK ------- WRLLQDLRAVNATMVLMGALQPGLPSPVA --------------------------------- IPQ[illegible]IIDLKDCFFSIPLHPSDQKRFAFSLP STNFK---------
                             N+P++ + K            R + D R  N              +                                  +         +D+  ++  + +          +F+
                                   * *            * *                                                                           *    *   *
                                   4                                        5                             6                                          7
                          G PQG   ++ L  +    +                          +ADD  +                               + +   K+                               +LG
R1Bm   RRIAVVAGECA EWKVSTMGCPQGSVLGPTLWNVLMDDLLALPQGIE ----------GTE MVAYADDVTVLVRGDSR AQLERR AHAVLGLAEGWASRNKLDFAPAKSRC IMLRGKFQRPPIVRYGSHVIRF ENQVTVLGVSS .291
R2Bm   STTLAVNNEMS SPVKVGRGVRQGDPLSPILFNVVMDLILASLPERV ---GYRLEMELVS ALAYADDLVLLAGSKVG ------ MQESISAVDCVGRQMGLRLNCRKSAV LSMIPDGHRKKHHYLTER.14. VERWRYLGVDF .406
L1Md   VANIKVNGEKL EAIPLKSGTRQGCPLSPYLFNIVLEVLARAIRQQK EIKGIQIGKEEVK ISLFADDMIVYISDPKN -----S TRELINLINSFGEAVGYKINSNKSMA FLYTKNKQAEKEIRETTPFSIV TNNIKYLGVTL .501
I      KITVRVGPHTS SPLPLFNGIPQGSPISVILFFIAFNKLSNIISLHK ----------EIK FNAYADDFFLIINFNKN TNTNFN LDNLFDDIENWCSYSGASLSLSKCQH LHICRKRHCTCKISCNNFQIPS VTSLKILGRTL .475
Ingi   TGRVRFKEKLF RSRTFERGVPQGTVPGSIMFIIVMNSLSQRLAEVP ----------LLQ HGFFADDLTLLARHTER DVINHT LQCGLNVVLQWSKEYFMSVNVAKTKC TLFGCTERHPLTLQLDGERIGA DRTPKLLGVTF .673
F      KFAVRCNTATS TVHTIEAGVPQGSVLGPTLYLIYTADIPTNSRLTV ------------- -STFADDTAILSRSRSP IQATAQ LALYLIDIKKWLSDWRIKVNEQKCKH VTFTLNRQDCPPLLLNSIPLPK ADEVTYLGVHL .113
G      EFQVCCGTATS TPRPIRAGVPQGSVLGPILYTLYTADLPITPSRSL ------------T VATYADDTAFLASASDP QEASTI ILSQLDALDPWLKRWTIAVNADKSSQ TTFSLRRGDCPPVTLNGETIPT SSSPKYLGLTL . ?

Sc-a1  YIDEK-----G TYHKPILGLPQGSLISPILCNIVITLVDNWLEDYI NLYN.46.FKRIK YVRYADDILIGVL     ----DC KIIKRDLNNFLNS-LGLTINEEKTLI TCAT------------------ ETPARFLGYNI .257
Sc-a2  YVDKN-----N NYHNTTLGIPQGSVVSPILCNIFLDKLDKYLENKF ENEF.49.FKRAY FVRYADDIIIGVM     ----DC KNILNDINNFLKENLGMSINIDKSVI KHS------------------- KEGVSFLGYDV .241
Pa-a1  YVEF------- KDKSSIIGVPQGGIASPILSNLVLNELDEFVQNIV DEFN.65.LAEIY YVRYADDWLIGVA     ----TA RAIKERIAAYLKDILKLELSMEKTLI TNAS------------------ EDKAYFLGTEI .259
Pa-IA  YIDD------- VKYNTYTGVPQGGVISPVLSNIYLHEFDLFVETLI KKYS.55.GIRVR YTRYADDWVIGII     ----LV AKIKEECKAFLRDILKLELSEEKTKI TNIT------------------ EKEVRFLGVDI .304
Sp-b1  YLTE------- RYKYDIVGTPQGSIVSPILANIYLHQLDEFIENLK SEFD.50.SNKLM YVRYADDWIVAVN     ----QT KEILAKITCFCSS-IGLTVSPTKTKI TNSY------------------ TDKILFLGTNI .241
Nc-p1  PDFVEILRRRG FTDIATNGVPQGASTSCGLATYNVKELFKRYDELI ------------- --MYADDGILCRQ     ------ -------PDFSVEEAGVVQEPAKSGW IKQNGEF--------------- KKSVKFLGLEF .339

Copia  ----------- DNVCKLNKAIYGLKQAARCWFEVFEQALKECEFVN SSVD.11.NENIY VLLYVDDVVIATGDMTR MN---- NFKRYLMEKFRMTDL----------- ---------------------- NEIKHFIGIRI .289
Ty     ----------- DKLIRLKKSLYELKQSGANWYETIKSYLIQQCGME EVRG..6.NSQVT ICLFVDDMVLFSKNL-- ------ NSNKRIIEKLKMQYDTKIINLGESDE ---------------------- EIQYDILGLEI .282

17.6   ----------- HGHYEYLRMPFGLKNAPATFQRCMNDILRPLLNKH ------------- CLVYLDDIIVF       ------ HLQSLGLVFEKLAKANLKLQLDKCEF L--------------------- KQETTFLGHVL .643
297    ----------- SGHYEYLRMPFGLRNAPATFQRCMNNILRPLLNKH ------------- CLVYLDDIIIF       ------ HLNSIQLVFTKLADANLKLQLDKCEF L--------------------- KKEANFLGHIV .645
Gypsy  ----------- GGKYEFCRLPFGLRNASSIFQRALDDVLREQIGKI ------------- CYVYVDDVIIF       ------ HVRHIDTVLKCLIDANMRVSQEKTRF F--------------------- KESVEYLGFIV .645
412    ----------- NGSYRFTRLPFGLKIAPNSFQRMMTIAFSGIEPSQ ------------- AFLYMDDLIVI       ------ MLKNLTEVFGKCREYNLKLHPEKCSF F--------------------- MHEVTFLGHKC .714
CaMV   ----------- QGHYEWNVVPFGLKQAPSIFQRHMDEAFRVFRKF- ------------- CCVYVDDILVF       ------ HLLHVAMILQKCNQHGIILSKKKAQL F--------------------- KKKINFLGLEI .103
DIRS1  ----------- GSHYRWKTMPFGLSTAPRIFTMLLRPVLRMLRDIN VS----------- VIAYLDDLLIV       ------ CLSNLKKTMDLLVKLGFKLNLEKSVL EP-------------------- TQSITFLGLQI .393

HBV    TFGRKLHLYSL PIILGFRKIPMGVGLSPFLLAQFTSAICSVVRRAF PHCL--------- AFSYMDDVVLGAKSVQH ------ LESLFTSITNFLLSLGIHLNPNKTKR W--------------------- GYSLNFMGYVI .243
DHBV   ----------- QRVYYFRKAPMGVGLSPFLLHLFTTALGSEISRRF NVW---------- TFTYMDDFLLCHPNARH ------ LNAISHAVCSFLQELGIRINFDKTTP SP-------------------- VNEIRFLGYQI .223
MuLV   ---------GI SGQLTWTRLPQGFKNSPTLFDEALHRDLADFRIQH PDLI--------- LLQYVDDLLLAATSELD ------ CQQGTRALLQTLGNLGYRASAKKAQI C--------------------- QKQVKYLGYLL .806
HERV   ---------GV TTQYTWTQLPQRFKNSPTIFGEALARDLQKFPTRD LGCV--------- LLQYVDDLLLGHPTAVG ------ WPREQMLYSGTWRTVGIRCPRKKAQI C--------------------- RQQVCYLGFTI .794
RSV    ---------AP ARRFQWKVLPQGMTCSPTICQLVVGQVLEPLRLKH PSLC--------- MLHYMDDLLLAASSHDG ------ LEAAGEEVISTLERAGFTISPDKVQR ---------------------- EPGVQYLGYKL .665
MMTV   ---------RP YQRFQWKVLPQGMKNSPTLCQKFVDKAILTVRDKY QDSY--------- IVHYMDDILLAHPSRSI ------ VDEILTSMIQALNKHGLVVSTEKIQK ---------------------- YDNLKYLGTHI . ?
HTLV1  ---------GP GTRYAWKVLPQGFKNSPTLFEMQLAHILQPIRQAF PQCT--------- ILQYMDDILLASPSHED ------ LLLLSEATMASLISHGLPVSENKTQQ T--------------------- PGTIKFLGQII .658
HTLV2  ---------GP GTRYAWTVLPQGFKNSPTLFEQQLAAVLNPMRKMF PTST--------- IVQYMDDILLASPTNEE ------ LQQLSQLTLQALTTHGLPISQEKTQQ T--------------------- PGQIRFLGQVI .659
BLV    ---------QP HRRFAWRVLPQGFINSPALFERALQEPLRQVSAAF SQSL--------- LVSYMDDILYASPTEEQ ------ RSQCYQALAARLRDLGFQVASEKTSQ T--------------------- PSPVPFLGQMV .640
HIV1   ---------TP GIRYQYNVLPQGWKGSPAIFQSSMTKILEPFRKQN PDIV--------- IYQYMDDLYVGSHLEIG -----Q HRTKIEELRQHLLRWGLTTPDKKHQK ---------------------- EPPFLWMGYEL .614
HIV2   ---------EP GKRYIYKVLPQGWKGSPAIFQHTMRQVLEPFRKAN KDVI--------- IIQYMDDILIASDRTDL -----E HDRVVLQLKELLNGLGFSTPDEKFQK ---------------------- DPPYHWMGYEL .618
Visna  ---------GP CVRYYWKVLPQGWKLSPAVYQFTMQKILRGWIEEH PMIQ--------- FGIYMDDIYIGSDLGLE -----E HRGIVNELASYIAQYGFMLPEDKRQE ---------------------- GYPAKWLGFEL .730
EIAV   ---------EP DKRYVWNCLPQGFVLSPYIYQKTLQEILQPFRERY PEVQ--------- LYQYMDDLFVGSNGSKK -----Q HKELIIELRAILLEKGFETPDDKLQE ---------------------- VPPYSWLGYQL .727
IAP    ---------EP DKRFQWKVLPQGMANSPTICQLYVQEALEPIRKQF TSLI--------- VIHYMDDILICHKELDV ------ LQKAFPMLVAELKQWGLEIASEKVQI ---------------------- ADTGLFLGSKI .170
HSRV   ----------- GKQYCWTRLPQGFLNSPALFTADVVDLLKEIPN-- ------------- VQVYVDDIYLSHDDPKE ------ HVQQLEKVFQILLQAGYVVSLKKSEI G--------------------- QKTVEFLGFNI .720
SRV    ---------EP MQRFQWKVLPQRMANSPTLCQKYVATAIHKVRHAW KQMY--------- IIHYMDDILIAGKDGQQ ------ VLQCFDQLKQELTIAGLHIAPEKIQL ---------------------- QDPYTYLGFEL .624
                      + +  +P G   +P +    +   +                         Y+DD+++ +                    +   +   G      K                                +LG  +
                            * *   ** *                                  * **  *                                     *                                ***
```

FIG. 1.—Amino acid sequence alignment of RT-related sequences. Sources and abbreviations for each element and the significance of the boxed regions are described in the text. The number at the beginning and end of each sequence indicates from either the 5′ and 3′ end of *pol* genes or the RT-containing open reading frames to the first and last residue in the figure, respectively. The question mark indicates those instances in which the exact ends of the ORF have not been unambiguously determined. While groups of elements are similar beyond these regions, no sequences conserved by all elements could be detected. Other numbers within certain sequences indicate the number of amino acids omitted from the sequence. The asterisk (*) in G element indicates the stop codon at that position. The letters and plus (+) symbols on the top of the alignment indicate the largely unvaried and chemically similar results, respectively, found in at least 11 of the 13 mitochondrial sequences and non-LTR retrotransposons. The letters and plus (+) symbols at the bottom of the alignment indicate the largely unvaried and chemically similar residues found in at least 18 of the 22 viruses and LTR-containing retrotransposons, not including copia and Ty. Asterisks (*) indicate those residues conserved in all four major groups of elements.

for the identification and classification of additional RT-containing elements, as well as aid in the functional analysis of enzymatic activity.

Virtually all gaps in the sequences are localized to the six segments between the boxed regions. The number of residues in these gap regions is quite low (2–14 amino acids) in the case of the retroviruses and the LTR-containing transposable elements, while the non-LTR transposons and mitochondrial sequences contain in these gaps ~90 amino acids and ~105 amino acids, respectively. In some instances, particularly between boxes 2 and 3, sequence similarity between the mitochondrial and non-LTR retrotransposon sequences could be detected in these gap regions. The seven boxed regions in figure 1 should not be confused with the smaller regions we have previously identified (Xiong and Eickbush 1988) as containing maximum similarity within the non-LTR group of retrotransposons.

By far the most difficult elements to align were copia and Ty. While it was relatively easy to align these two sequences versus each other, the conserved residues found in all other RT-containing elements could only be unambiguously identified in copia and Ty for boxes 3, 5, and 7. Indeed, if we were to eliminate from similarity calculations the conserved residues we have used to align each sequence, average sequence similarity of copia and Ty to all other elements would only be 7%–8%, not above the similarity of random sequences. This raises the question of whether these two elements are homologous to the other elements. There is little doubt that these elements contain RT-like activity. Full-length copia and Ty transcripts packaged into virus-like particles have been reported (Shiba and Saigo 1983; Garfinkel et al. 1985), and RNA intermediates have been clearly demonstrated to be involved in the propagation of the Ty element (Boeke et al. 1985). Finally, the similarity of the overall structure of these elements to that of other transposable elements and retroviruses suggests a common origin. Thus we have included these two elements in our phylogenetic comparisons, even though their sequence similarity to the other elements is minimal.

Formation of Phylogenetic Trees

The proportion of identical amino acid positions was calculated for the seven regions common to all sequences (boxed sequences in fig. 1). These regions contained 182 amino acid residues for most of the elements. While the residues between the boxed regions are not included in our calculations, the sizes of these insertions/gaps as well as their sequences are consistent with the major branch points of the trees derived from the sequence of the common regions.

The percent divergence for all pairwise comparisons of the 37 aligned sequences was calculated by dividing the number of different residues by the total number of compared residues (fig. 2). Before tree construction all values were changed to distances with Poisson correction, $d = -\log_e S$, where S = sequence similarity (Nei 1987, p. 41). These corrected values were then used to construct phylogenetic trees by the unweighted-pair-group method (UPGMA) (Sneath and Sokal 1973, pp. 230–234) and the neighbor-joining (NJ) method (Saitou and Nei 1987). UPGMA is known to reliably give the correct topology for a phylogenetic tree when the rate of substitutions is approximately constant for each element (Nei 1987, pp. 287–326). This is not likely to be true for the RT sequences since the different mechanisms of propagation for the elements in this comparison (viruses, transposable elements, and introns) could easily give rise to different rates of sequence change. The NJ method, on the other hand, has been shown to be a reliable method of determining the correct topology when elements have different rates of sequence divergence (Saitou and Nei 1987). We believe,

	R1Bm	R2Bm	L1Md	I	Ingl	F	G	Sc-a1	Sc-a2	Pa-a1	Pa-IA	Sp-b1	Nc-pl	Copia	Ty	17.6	297	Gypsy	412	CaMV	DIRS1
R1Bm	–																				
R2Bm	.775	–																			
L1Md	.747	.725	–																		
I	.787	.764	.697	–																	
Ingl	.725	.753	.775	.708	–																
F	.672	.757	.729	.706	.695	–															
G	.684	.740	.734	.723	.706	.511	–														
Sc-a1	.774	.774	.740	.763	.791	.773	.778	–													
Sc-a2	.781	.781	.736	.781	.815	.819	.802	.446	–												
Pa-a1	.753	.781	.815	.826	.820	.814	.791	.554	.612	–											
Pa-IA	.798	.803	.809	.820	.826	.791	.831	.582	.612	.449	–										
Sp-b1	.785	.802	.751	.780	.791	.790	.813	.548	.542	.492	.537	–									
Nc-pl	.805	.769	.811	.811	.811	.834	.815	.792	.757	.740	.698	.768	–								
Copia	.880	.886	.892	.904	.874	.910	.922	.886	.880	.898	.898	.886	.886	–							
Ty	.869	.898	.864	.881	.903	.891	.880	.903	.915	.903	.903	.914	.892	.745	–						
17.6	.837	.843	.843	.871	.871	.842	.870	.819	.803	.865	.876	.876	.888	.892	.864	–					
297	.871	.826	.860	.865	.876	.847	.881	.814	.809	.843	.854	.870	.876	.868	.858	.157	–				
Gypsy	.865	.882	.882	.876	.865	.870	.876	.859	.843	.876	.888	.864	.864	.862	.852	.534	.478	–			
412	.837	.837	.854	.848	.860	.847	.870	.842	.854	.871	.860	.842	.846	.862	.881	.545	.573	.590	–		
CaMV	.898	.876	.864	.870	.864	.852	.881	.858	.876	.842	.842	.841	.827	.855	.863	.638	.633	.644	.667	–	
DIRS1	.837	.831	.831	.831	.837	.825	.780	.814	.826	.831	.837	.842	.846	.850	.841	.730	.753	.764	.758	.740	–
HBV	.871	.820	.843	.820	.848	.825	.842	.853	.848	.860	.860	.831	.858	.838	.886	.854	.854	.843	.803	.831	.798
DHBV	.837	.820	.820	.831	.831	.819	.808	.836	.809	.860	.865	.814	.846	.850	.858	.826	.831	.815	.758	.808	.792
MuLV	.831	.792	.826	.787	.820	.808	.808	.825	.815	.837	.820	.831	.840	.880	.881	.730	.736	.713	.758	.712	.753
HERV	.826	.815	.848	.837	.854	.819	.831	.825	.843	.848	.837	.853	.834	.880	.869	.742	.764	.719	.758	.734	.764
RSV	.820	.809	.843	.843	.837	.847	.819	.802	.809	.843	.854	.808	.799	.886	.886	.747	.753	.708	.747	.780	.770
MMTV	.843	.848	.837	.854	.860	.836	.847	.814	.815	.798	.831	.814	.811	.892	.881	.747	.758	.742	.770	.751	.798
HTLV1	.860	.837	.837	.826	.843	.825	.836	.831	.837	.843	.837	.797	.834	.886	.869	.736	.725	.725	.781	.706	.742
HTLV2	.843	.843	.831	.809	.854	.859	.847	.791	.826	.826	.826	.814	.834	.886	.864	.725	.736	.736	.781	.706	.742
BLV	.848	.837	.848	.843	.860	.864	.819	.819	.826	.809	.831	.814	.846	.880	.869	.758	.747	.702	.764	.751	.764
HIV1	.860	.826	.837	.848	.865	.842	.836	.831	.860	.848	.854	.825	.840	.892	.824	.753	.753	.736	.753	.729	.792
HIV2	.876	.865	.843	.860	.860	.836	.853	.825	.837	.843	.860	.842	.864	.874	.801	.742	.742	.713	.781	.740	.753
Visna	.848	.826	.843	.843	.826	.836	.847	.831	.820	.820	.843	.859	.858	.904	.830	.742	.747	.753	.758	.757	.803
EIAV	.848	.826	.815	.843	.848	.831	.842	.814	.792	.831	.815	.797	.840	.898	.835	.742	.753	.747	.764	.763	.775
IAP	.860	.876	.860	.871	.860	.847	.836	.808	.826	.792	.815	.797	.817	.850	.869	.770	.781	.770	.764	.734	.770
HSRV	.830	.830	.813	.830	.864	.869	.823	.811	.813	.830	.824	.823	.784	.861	.845	.642	.665	.688	.716	.710	.756
SRV	.854	.888	.843	.860	.893	.836	.836	.797	.820	.815	.831	.819	.817	.886	.875	.742	.764	.753	.781	.768	.792

	HBV	DHBV	MuLV	HERV	RSV	MMTV	HTLV1	HTLV2	BLV	HIV1	HIV2	Visna	EIAV	IAP	HSRV	SRV
HBV	–															
DHBV	.545	–														
MuLV	.792	.775	–													
HERV	.848	.798	.427	–												
RSV	.792	.747	.646	.685	–											
MMTV	.803	.770	.657	.702	.506	–										
HTLV1	.758	.770	.624	.663	.601	.590	–									
HTLV2	.758	.758	.612	.624	.618	.584	.242	–								
BLV	.787	.770	.657	.691	.640	.652	.455	.438	–							
HIV1	.792	.826	.708	.742	.629	.685	.669	.680	.725	–						
HIV2	.803	.820	.697	.742	.629	.663	.640	.652	.674	.343	–					
Visna	.826	.815	.753	.758	.680	.691	.685	.702	.702	.494	.483	–				
EIAV	.798	.781	.708	.713	.657	.657	.657	.663	.680	.444	.478	.494	–			
IAP	.803	.758	.691	.708	.455	.466	.579	.584	.635	.640	.612	.640	.596	–		
HSRV	.847	.784	.619	.676	.716	.693	.699	.688	.705	.739	.744	.722	.739	.710	–	
SRV	.809	.798	.669	.685	.511	.343	.612	.596	.612	.691	.635	.674	.652	.421	.699	–

FIG. 2.—Amino acid divergence of the RT-related sequences. Divergence was calculated by dividing the number of different residues from the seven boxed regions in fig. 1 by the total number of compared residues. The total number of compared residues was 182 amino acids for most elements; 181 residues for CaMV, F, G, Sc-a1, and Sp-b1; 180 residues for Ty and HSRV; 173 for Nc-p1; and 171 residues for copia.

however, that it is important to include the UPGMA tree, since genetic-distance estimates are known to be subject to stochastic errors and since the procedure for distance averaging in UPGMA reduces the effects of this error on the estimation of branch length (Nei 1987, pp. 287–326). Thus, in instances where the UPGMA and NJ methods give the same topology, the reliability of that topology can be considered quite high.

A comparison of the phylogenetic trees derived from the two methods is shown in figure 3. Both methods result in the clustering of the non-LTR retrotransposable elements with the class II mitochondrial introns and of the retroviruses with the LTR retrotransposable elements (except for copia and Ty). The topology within each of these groups is nearly identical by these two methods and will be discussed in greater detail below. Thus, at this time these groups are indicated by boxes in figure 3, and only the major branch points of the trees are discussed. One difference exists between the topology of the two trees in figure 3. In the UPGMA tree, copia and Ty are the most divergent branch, while in the NJ tree, copia and Ty branch from the hepatitis

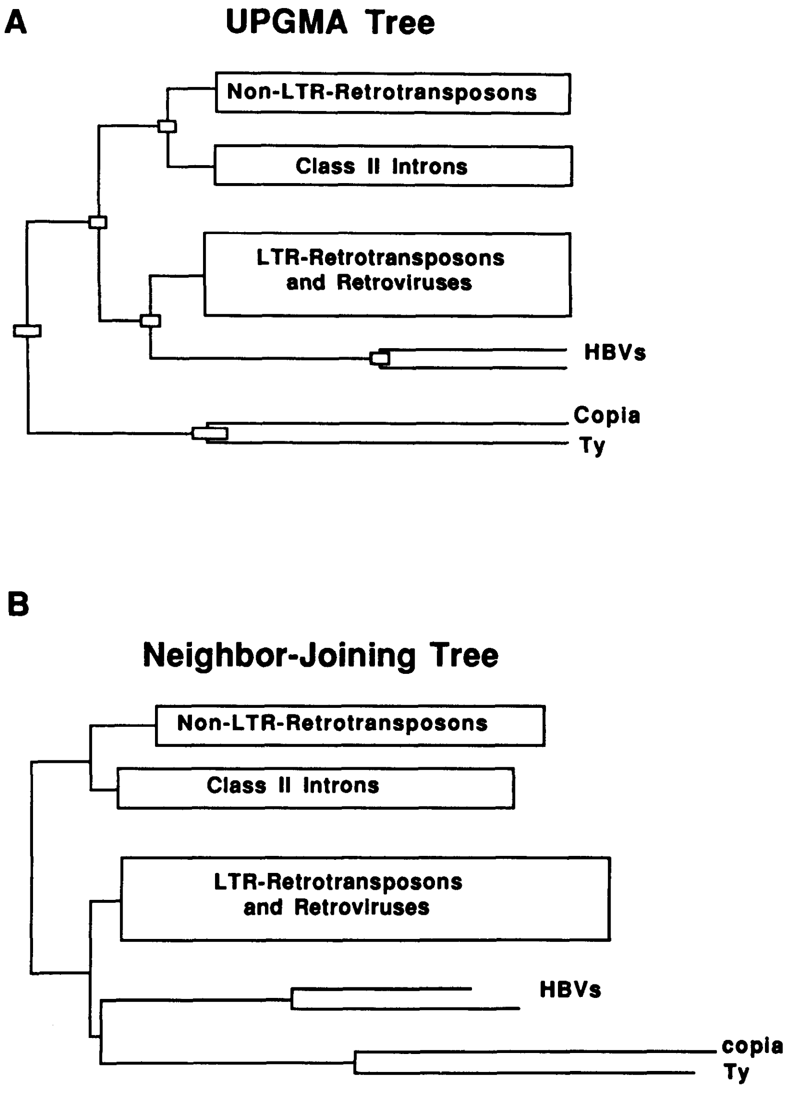

FIG. 3.—Comparison of the phylogenetic trees constructed by (A) the UPGMA (Sneath and Sokal 1973, pp. 230–234) and (B) the NJ method (Saitou and Nei 1987). The open boxes in the UPGMA tree correspond to the SE of branch point as calculated by Nei et al. (1984). The length of horizontal lines in the NJ tree correspond to the branch length. To simplify visual comparison of the major topologies of these two trees, elements of the same class that are located on the same branch of the tree are indicated by a box. However, data from all 37 elements were used in the generation of these trees. The lengths of the boxes in the NJ tree correspond to the longest total branch length of any element within the box.

B viruses. The NJ method gives an unrooted tree, and there are two possible positions for the root of this tree. The first position is as shown in figure 3B, with the tree rooted between the class II intron / non-LTR retrotransposon branch and the retrovirus / LTR retrotransposon branch. The second position is with copia and Ty as the most distant branch. While this second position for the root would make the tree more consistent with the UPGMA tree, we suggest that this is the incorrect root, for the following reasons: First, if copia and Ty are the most divergent branch of the tree, then the

HBVs should be equally distant from all other elements in the trees. This is not consistent with the data, which clearly indicate that the HBVs are more closely related to the retroviruses and LTR retrotransposons. Second, while copia and Ty clearly have the least similarity to all other elements, they consistently have greater similarity to elements in the retrovirus/LTR transposon branch than to elements in the class II intron/non-LTR transposon branch. This finding is consistent with a faster rate of evolution for copia and Ty. This faster rate of divergence was also detected by the NJ method and is revealed in the NJ tree by the longer branch lengths of these two elements. Thus the NJ tree appears to be a more reliable estimate of the real phylogenetic relationship of the various RT elements. The complete NJ tree for all 37 RT-containing elements is presented in figure 4.

Discussion

Two striking features can be seen in the phylogenetic tree presented in figure 4. First, elements present in very diverse species sometimes occupy close positions on the tree. For example, CaMV (a plant virus) is closely related to a series of insect transposable elements; the non-LTR retrotransposons are present in a trypanosome, insects, and mammals; and an insect transposable element (copia) is more closely related to a yeast transposable element (Ty) than to any of the other insect transposable elements. This indicates that either the progenitors of these various groups of elements are very ancient (i.e., before the separation of fungi from other eukaryotes in the case of copia and Ty) or, as is more likely, there has been extensive horizontal transfer of elements between species. While our comparison is suggestive, resolution of this issue must await the identification of the same element in several distant species.

The second striking feature of the tree in figure 2 is that all RT-containing elements can be divided into two major branches: a virus and LTR-containing retrotransposon branch and a class II intron and non-LTR retrotransposon branch. We will discuss each of these major branches separately.

Viruses and Retrotransposons with LTRs

The largest branch of the phylogenetic tree shown in figure 4 is composed of the retroviruses, several DNA viruses that propagate via an RNA intermediate, and seven retrotransposable elements. While an even larger number of retroviral sequences have been published, in the present report we have utilized only 14 of these sequences, selected as being representative of the full diversity present among currently known retroviruses. The phylogenetic relationships among retroviruses, relationships based upon their RT sequences, have been analyzed elsewhere, with somewhat varying conclusions (Chiu et al. 1985; Sagata et al. 1985; Sonigo et al. 1985, 1986; Toh et al. 1985; McClure et al. 1988). Part of this controversy is a result of the different rates of sequence divergence in different groups of viruses. Indeed, the only major difference we detected between the UPGMA and NJ trees (other than the location of copia and Ty as described above) was in the retroviral branch. The UPGMA tree placed the lentiviruses (HIV, Visna, and EIAV) as equally distant from the RSV/MMTV and the HTLV/BLV branches. The NJ tree as shown in figure 4 suggests that the lentiviruses have diverged at a more rapid rate and are closer to the RSV/MMTV group. Recently, using the NJ method, Yokoyama et al. (1988) have obtained the same phylogenetic relationship for the retroviruses, on the basis of the nucleotide sequences of the RT and *env* regions. The topology we obtained for the retroviruses by using the NJ method is identical to that of Yokoyama et al., except that their exclusively retroviral tree is

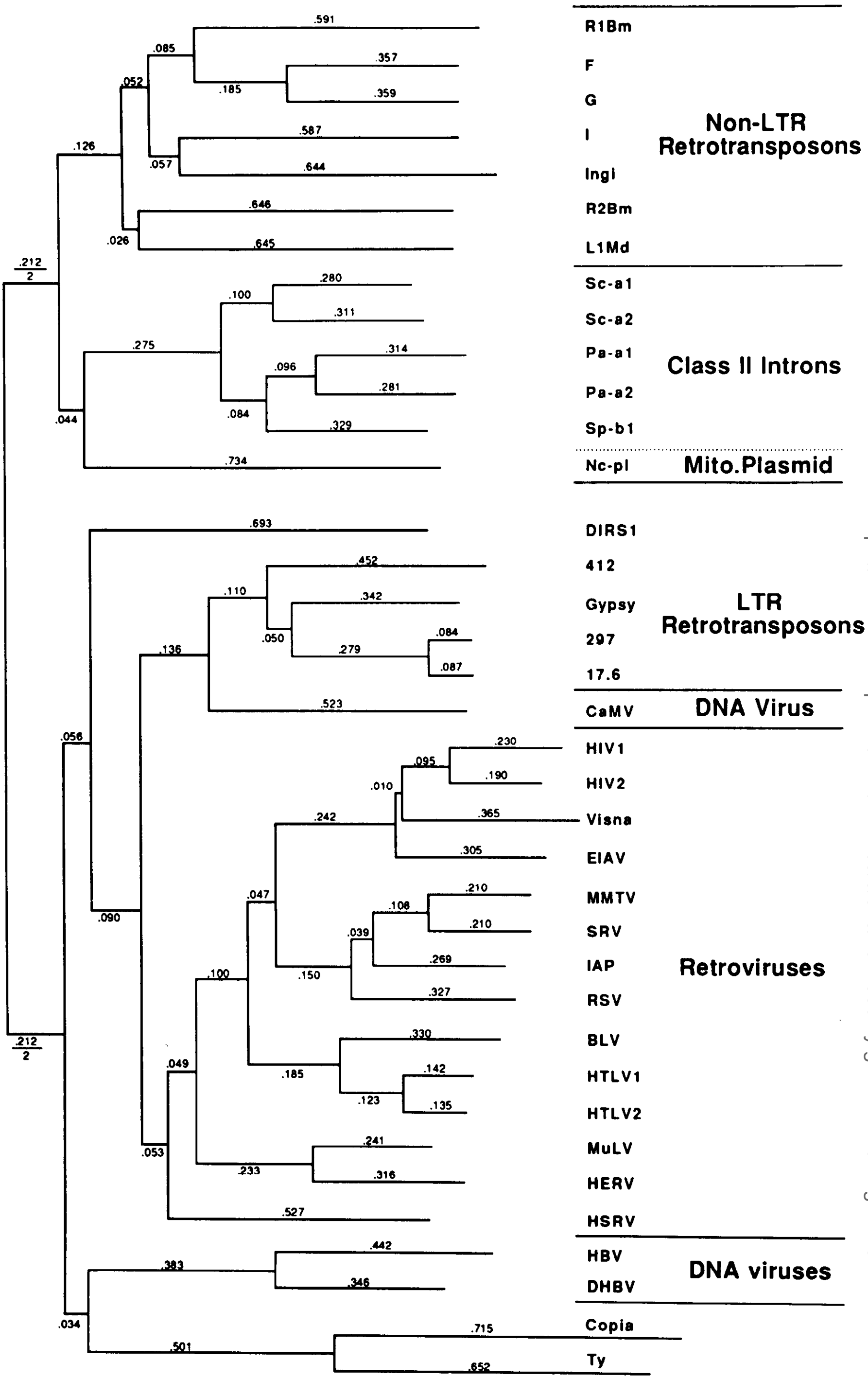

FIG. 4.—Phylogenetic tree of RT-related sequences. The tree was constructed using the percentage divergence presented in fig. 2 after Poisson correction and application of the NJ method (Saitou and Nei 1987). The number above or below each horizontal line indicates the branch length. The branch length between the node connecting the LTR retrotransposons and viruses with the node connecting class II introns and non-LTR retrotransposons, which contains the root of the tree, was divided equally between the two major branches (see text for the discussion of rooting of the tree). Functional classification for the various groups of elements is presented to the right of the tree.

rooted in a different location. We feel that using the retroviruses as only a part of a comprehensive analysis of all RT sequences is the most reliable method of rooting the retroviral branch. With the branch rooted as in figure 4, the human foamy virus (HSRV) is the most ancient member of the retroviral family, while the lentiviruses appear to be derived from the RNA tumor viruses (oncoviruses), specifically the branch leading to RSV and MMTV.

With all retroviruses located on one subbranch, an intermingling of transposable elements and DNA viruses occurs throughout the remainder of this major branch of the RT tree. The hepatitis B viruses, copia, and Ty represent the most distant elements of this branch, followed by the *Dictyostelium discoideum* transposable element DIRS I. Perhaps the most surprising conclusion is that four *D. melanogaster* retrotransposable elements (17.6, 297, Gypsy, and 412) are most closely related to CaMV. The greater similarity of the CaMV RT region to these *D. melanogaster* transposable elements than to retroviruses had been detected previously by Yuki et al. (1986*b*). A common structural feature of all elements from this major branch of the RT tree is the presence of LTRs. The retroviruses and the six transposable elements from *D. melanogaster* and *Saccharomyces cerevisiae* are believed to contain LTRs of similar structure and function (reviewed in Varmus 1983). The DNA viruses, hepatitis B virus, and cauliflower mosaic virus, although lacking complete LTRs, contain certain features of LTRs that are critical to their life cycle (Summers and Mason 1982; Pfeiffer and Hohn 1983; Miller and Robinson 1986; see also Ganem and Varmus 1987). Finally, the *D. discoideum* transposable element, DIRS I, contains inverted long terminal repeats (Cappello et al. 1985). Because of the critical role played by LTRs in the propagation of these elements, we have termed this branch of the tree the LTR branch.

Class II Introns and Non-LTR Retrotransposons

The second major grouping of RT-related sequences is composed of a series of mitochondrial intron sequences from fungi, a mitochondrial plasmid, and a number of transposable elements from widely different species. The topology of this branch of the UPGMA tree is identical to that of the NJ tree shown in figure 4. Particularly diagnostic of this branch of the tree is the highly conserved YXDD box, flanked by several hydrophobic residues, found in segment 5. All elements from this branch have an alanine at position X, while all members of the LTR branch have a hydrophobic residue at this position. Such high constraint for a residue located in a region known to be important for the RT activity (Larder et al. 1987) raises the interesting question of whether the conservation at this (and some other) amino acid positions is the reflection of different properties of the RT enzyme encoded by the elements from this branch of the tree.

The higher RT similarity between various members of the non-LTR retrotransposon group has been noted elsewhere (Fawcett et al. 1986; Burke et al. 1987; Di Nocera and Casari 1987; Kimmel et al. 1987; Di Nocera 1988; Xiong and Eickbush 1988). Our results support this conclusion and also indicate that R2 and L1 are the most distant members of this group. On the basis of its genetic organization and insertion properties, R2 appears to be least similar to the other elements because it does not contain a second ORF with similarity to the *gag* ORF of retroviruses and does not give rise to a target-site duplication at its specific insertion site (Burke et al. 1987). These properties are shared by the class II mitochondrial introns.

The mitochondrial intron sequences shown in figure 4 are from three species of fungi and are all classified as class II introns on the basis of predictions of their secondary structure and short conserved sequence elements (Michel and Dujon 1983). On the basis of the similarity of its RT-like sequence, the *Neurospora* plasmid (Nc-pl), is closely related to this intron group, confirming previous suggestions based on its codon usage and conserved DNA sequence elements (Nargang et al. 1984). The position of these mitochondrial sequences on a common branch of the RT tree with a distinct class of transposable elements that do not have LTRs and whose members can also occupy specific positions in genes supports suggestions that these mitochondrial sequences are or at one time were mobile elements (Borst and Grivell 1981; Nargang et al. 1984; Michel and Lang 1985).

We have termed this second major branch of the RT tree the non-LTR branch. One of the fundamental roles of LTRs is to enable replication, via an RNA intermediate, with no net loss of sequence information (Temin 1982; Varmus 1983). The absence of LTRs in the elements of this branch implies that they must employ alternative mechanisms in their propagation. In the case of the *Neurospora* plasmid, full-length transcripts can be easily achieved because the genome is circular (Nargang et al. 1984). For I and L1 it has been suggested that full-length progeny are made by utilizing an internal promoter and 5′-end tandem repeats, respectively (Fawcett et al. 1986; Loeb et al. 1986). A quite different mechanism, utilization of an exogenous (host) gene promoter, has apparently been adopted by other members of the non-LTR group. Ingi has been suggested to rely on its fortuitous location next to an external promoter (Kimmel et al. 1987). However, most of the remaining members appear to have acquired an external promoter by their ability to insert into specific sites within a particular host gene. In the case of R1 and R2, the elements have inserted in a fraction of the members of the ribosomal RNA multigene family (Burke et al. 1987; Xiong and Eickbush 1988). Although the transcription or processing of these genes is severely affected (Long and Dawid 1979), a fraction of the ribosomal genes can evidently be interrupted at a tolerable cost to the host. Mitochondrial intron sequences, on the other hand, insert within a unique gene. In this case, utilization of these mitochondrial genes by the host requires that these introns be efficiently spliced from the transcripts. Other members of the class II intron family have been shown to be self-splicing (reviewed in Cech 1986). This raises the question of whether self-splicing introns are descendents of non-LTR retrotransposons—and of whether these RT-containing introns, as well as non-LTR retrotransposons such as R1 and R2, can undergo self-cleavage if not self-splicing.

Concluding Comments

We have compared the RT-like sequences from a variety of sources and identified conserved regions that will be useful for the identification and analysis of RT sequences in other elements. These conserved regions also have been used to construct a possible phylogenetic tree. It has been suggested by Temin (1980) that the origin of retroviruses was from cellular transposable elements. Since all retroviruses are located on a minor branch of the tree, while retrotransposons are found throughout the tree, our data support this hypothesis. Indeed, the location of transposable elements on each of the major branches of the tree further suggests that transposable elements may have been the precursors of all present-day RT-containing elements. Two major divisions of the RT-containing elements are apparent, one with and one without LTRs. The tree leaves unresolved the issue of whether the LTR branch gained LTRs or whether the non-LTR branch lost its LTRs. We prefer the former hypothesis, because the acquisition of LTRs—and thus of their own endogenous promoter—would have made the transposable elements more independent, perhaps facilitating their evolution into DNA or RNA viruses. Those elements that did not gain (or, alternatively, lost) their LTRs must rely on exogenous promotion of transcription. Perhaps the most interesting mechanism adopted by these non-LTR elements for this purpose is the specialization for insertion into specific host genes.

Acknowledgments

We thank Bill Burke, John Jakubczak, and Uzi Nur for comments on the manuscript. We thank Masatoshi Nei for supplying us with the NJ and SE programs, and Chenyang Shi and Yijing Yan for assistance in establishing these programs on the University of Rochester VAX computer system. We thank Chung-I Wu for computer assistance and for comments on the manuscript. This work was supported by a grant from the National Institute of Health and was aided by grant IN-18 from the American Cancer Society.

LITERATURE CITED

AKINS, R. A., R. L. KELLEY, and A. M. LAMBOWITZ. 1986. Mitochondrial plasmids of *Neurospora:* integration into mitochondrial DNA and evidence for reverse transcription in mitochondria. Cell **47:**505–516.

BOEKE, J. D., C. A. GARFINKEL, C. A. STYLES, and G. R. FINK. 1985. Ty elements transpose through an RNA intermediate. Cell **40:**491–500.

BONITZ, S. G., G. CORUZZI, B. E. THALENFELD, A. TZAGOLOFF, and G. MACINO. 1980. Assembly of the mitochondrial membrane system: structure and nucleotide sequence of the gene coding for subunit 1 of yeast cytochrome oxidase. J. Biol. Chem. **255:**11927–11941.

BORST, P., and L. A. GRIVELL. 1981. One gene's intron is another gene's exon. Nature **289:** 439–440.

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.

CAPPELLO, J., K. HANDELSMAN, and H. LODISH. 1985. Sequence of *Dictyostelium* DIRS-1: an apparent retrotransposon with inverted terminal repeats and an internal circle junction sequence. Cell **43:**105–115.

CECH, T. R. 1986. The generality of self-splicing RNA: relationship to nuclear mRNA splicing. Cell **44:**207–210.

CHIU, I.-M., A. YANIV, J. E. DAHLBERG, A. GAZIT, S. F. SKUNTZ, S. R. TRONICK, and S. A. AARONSON. 1985. Nucleotide sequence evidence for relationship of AIDS retrovirus to lentiviruses. Nature **317:**366–368.

CLARE, J., and P. FARABAUGH. 1985. Nucleotide sequence of a yeast Ty element: evidence for a novel mechanism of gene expression. Proc. Natl. Acad. Sci. USA **82:**2829–2833.

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

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

DOOLITTLE, R. F. 1981. Similar amino acid sequences: chance or common ancestry? Science **214:**149–159.

EICKBUSH, T. H., and B. ROBINS. 1985. *Bombyx mori* 28S ribosomal genes contain insertion elements similar to the type I and II elements of *Drosophila melanogaster.* EMBO J. **4:**2281–2285.

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., M. S. Johnson, and R. F. Doolittle. 1985. Aligning amino acid sequences: comparison of commonly used methods. J. Mol. Evol. **21**:112–125.

Galibert, F., E. Mandart, F. Fitoussi, P. Tiollais, and P. Charnay. 1979. Nucleotide sequence of the hepatitis B virus genome (subtype ayw) cloned in *E. coli.* Nature **281**:646–650.

Ganem, D., and H. E. Varmus. 1987. The molecular biology of the hepatitis B viruses. Annu. Rev. Biochem. **56**:651–693.

Gardner, R. C., A. J. Howarth, P. Hahn, M. Brown-Luedi, R. J. Shepherd, and J. Messing. 1981. The complete nucleotide sequence of an infectious clone of cauliflower mosaic virus by M13mp7 shotgun sequencing. Nucleic Acids Res. **9**:2871–2887.

Garfinkel, D. J., J. D. Boeke, and G. R. Fink. 1985. Ty element transposition: reverse transcriptase and virus-like particles. Cell **42**:507–517.

Guyader, M., M. Emerman, P. Sonigo, F. Clavel, L. Montagnier, and M. Alizon. 1987. Genome organization and transactivation of the human immunodeficiency virus type 2. Nature **326**:662–669.

Hattori, M., S. Kuhara, O. Takenaka, and Y. Sakaki. 1986. L1 family of repetitive sequences in primates may be derived from a sequence encoding a reverse transcriptase–related protein. Nature **321**:625–627.

Inouye, S., S. Yuki, and K. Saigo. 1986. Complete nucleotide sequence and genome organization of a *Drosophila* transposable element, 297. Eur. J. Biochem. **154**:417–425.

Kimmel, B. E., O. K. ole-Moiyoi, and J. R. Young. 1987. Ingi, a 5.2-kilobase dispersed sequence element from *Trypanosoma brucei* that carries half of a smaller mobile element at their end and has homology with mammalian LINEs. Mol. Cell. Biol. **7**:1465–1475.

Lang, B. F., F. Ahne, and L. J. Bonen. 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.

Larder, B. A., D. J. Purifoy, K. L. Powell, and G. Darby. 1987. Site specific mutagenesis of AIDS virus reverse transcriptase. Nature **327**:716–717.

Loeb, D. D., R. W. Padgett, S. C. Hardies, W. R. Shehee, M. B. Comer, M. H. Edgell, and C. A. Hutchison III. 1986. The sequence of a large L1Md element reveals tandemly repeated 5′ end and several features found in retrotransposons. Mol. Cell. Biol. **6**:168–182.

Long, E. O., and I. B. Dawid. 1979. Expression of ribosomal DNA insertions in *Drosophila melanogaster.* Cell **18**:1185–1196.

McClure, M. A., M. S. Johnson, D.-F. Feng, and R. F. Doolittle. 1988. Sequence comparison of retroviral proteins: relative rates of change and general phylogeny. Proc. Natl. Acad. Sci. USA **85**:2469–2473.

Mandart, E., A. Kay, and F. Galibert. 1984. Nucleotide sequence of a cloned duck hepatitis B virus genome: comparison with woodchuck and human hepatitis B virus sequences. J. Virol. **49**:782–792.

Marlor, R., S. Parkhurst, and V. Corces. 1986. The *Drosophila melanogaster* gypsy transposable element encodes putative gene products homologous to retroviral proteins. Mol. Cell. Biol. **6**:1129–1134.

Matsuura, E. T., J. M. Domenico, and D. J. Cummings. 1986. An additional class II intron with homology to reverse transcriptase in rapidly senescing *Podospora anserina.* Curr. Genet. **10**:915–922.

Maurer, B., H. Bannert, G. Darai, and R. M. Flugel. 1988. Analysis of the primary structure of the long terminal repeat and the gag and pol genes of the human spumaretrovirus. J. Virol. **62**:1590–1597.

Michel, F., and B. Dujon. 1983. Conservation of RNA secondary structure in two intron families including mitochondrial-, chloroplast- and nuclear-encoded members. EMBO J. **2**: 33–38.

Michel, F., and F. Lang. 1985. Mitochondrial class II introns encode proteins related to the reverse transcriptases of retroviruses. Nature **316**:641–643.

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.

Mount, S. M., and G. M. Rubin. 1985. Complete nucleotide sequence of the *Drosophila* transposable element copia: homology between copia and retroviral proteins. Mol. Cell. Biol. **5**:1630–1638.

Nargang, F. E., J. B. Bell, L. L. Stohl, and A. M. Lambowitz. 1984. The DNA sequence and genetic organization of a *Neurospora* mitochondrial plasmid suggest a relationship to introns and mobile elements. Cell **38**:441–453.

Nei, M. 1987. Molecular evolutionary genetics. Columbia University Press, New York.

Nei, M., C. Stephens, and N. Saitou. 1984. Methods for computing the standard errors of branching points in an evolutionary tree and their application to molecular data from humans and apes. Mol. Biol. Evol. **2**:66–85.

Ono, M., H. Toh, T. Miyata, and T. Awaya. 1985. Nucleotide sequence of the Syrian hamster intracisternal A-particle gene: close evolutionary relationship of type A particle gene to type B and D oncovirus genes. J. Virol. **55**:387–394.

Osierwacz, H. D., and K. Esser. 1984. The mitochondrial plasmid of *Podospora anserina*: a mobile intron of a mitochondrial gene. Curr. Genet. **8**:299–305.

Pfeiffer, P., and T. Hohn. 1983. Involvement of reverse transcription in the replication of cauliflower mosaic virus: a detailed model and test of some aspects. Cell **33**:781–789.

Power, M. D., P. A. Marx, M. L. Bryant, M. B. Gardner, P. J. Barr, and P. A. Luciw. 1986. Nucleotide sequence of SRV-1, a type D simian acquired immune deficiency syndrome retrovirus. Science **231**:1567–1572.

Repaske, R., P. E. Steele, R. R. O'Neill, A. B. Rabson, and M. A. Martin. 1985. Nucleotide sequence of a full-length human endogenous retroviral segment. J. Virol. **54**:764–772.

Sagata, N., T. Yasunage, J. Tsuzuku-Kawamura, K. Ohishi, Y. Ogawa, and Y. Ikawa. 1985. Complete nucleotide sequence of the genome of bovine leukemia virus: its evolutionary relationship to other retroviruses. Proc. Natl. Acad. Sci. USA **82**:677–681.

Saigo, K., W. Kugimiya, Y. Matsuo, S. Inouye, K. Yoshioka, and S. Yuki. 1984. Identification of the coding sequence for reverse transcriptase-like enzyme in a transposable genetic element in *Drosophila melanogaster*. Nature **312**:659–661.

Saitou, N., and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**:406–425.

Schwartz, D. E., R. Tizard, and W. Gilbert. 1983. Nucleotide sequence of the Rous sarcoma virus. Cell **32**:853–869.

Seiki, M., S. Hattori, Y. Hirayama, and M. Yoshida. 1983. Human adult T-cell leukemia virus: complete nucleotide sequence of the proviral genome integrated in leukemia cell DNA. Proc. Natl. Acad. Sci. USA **80**:3618–3622.

Shehee, W. R., S. F. Chao, D. D. Loeb, M. B. Comer, C. A. Hutchison III, and M. H. Edgell. 1987. Determination of a functional ancestral sequence and definition of the 5′ end of A-type mouse L1 elements. J. Mol. Biol. **196**:757–767.

Shiba, T., and K. Saigo. 1983. Retrovirus-like particles containing RNA homologous to the transposable element copia in *Drosophila melanogaster*. Nature **302**:119–124.

Shimotohno, K., Y. Takahashi, N. Shimizu, T. Gojobori, D. W. Golde, I. S. Chen, M. Miwa, and T. Sugimura. 1985. Complete nucleotide sequence of an infectious clone of human T-cell leukemia virus type II: an open reading frame for the protease gene. Proc. Natl. Acad. Sci. USA **82**:3101–3105.

Shinnick, T. M., T. A. Lerner, and J. G. Sutcliffe. 1981. Nucleotide sequence of the Moloney murine leukemia virus. Nature **293**:543–548.

Sneath, P. H. A., and R. R. Sokal. 1973. Numerical taxonomy. W. H. Freeman, San Francisco.

Sonigo, P., M. Alizon, K. Staskus, D. Klatzmann, S. Cole, O. Danos, E. Retzel, P.

TIOLLAIS, A. HAASE, and S. WAIN-HOBSON. 1985. Nucleotide sequence of the visna lentivirus: relationship to the AIDS virus. Cell **42**:369–382.

SONIGO, P., C. BARKER, E. HUNTER, and S. WAIN-HOBSON. 1986. Nucleotide sequence of Mason-Pfizer monkey virus: an immunosuppressive D-type retrovirus. Cell **45**:375–385.

SUMMERS, J., and W. S. MASON. 1982. Replication of the genome of a hepatitis B–like virus by reverse transcription of an RNA intermediate. Cell **29**:403–415.

TEMIN, H. M. 1980. Origin of retroviruses from cellular moveable genetic elements. Cell **21**: 599–600.

———. 1982. Function of the retrovirus long terminal repeat. Cell **28**:3–5.

TOH, H., H. HAYASHIDA, and T. MIYATA. 1983. Sequence homology between retroviral transcriptase and putative polymerases of hepatitis B virus and cauliflower mosaic virus. Nature **305**:827–829.

TOH, H., R. KIKUNO, H. HAYASHIDA, T. MIYATA, W. KUGIMIYA, S. INOUYE, S. YUKI, and K. SAIGO. 1985. Close structural resemblance between putative polymerase of a *Drosophila* transposable genetic element 17.6 and the pol gene product of Moloney murine leukemia virus. EMBO J. **4**:1267–1272.

VARMUS, H. E. 1983. Retroviruses. Pp. 411–503 *in* J. A. SHAPIRO, ed. Mobile elements. Academic Press, New York.

WAIN-HOBSON, S., P. SONIGO, O. DANOS, S. COLE, and M. ALIZON. 1985. Nucleotide sequence of the AIDS virus, LAV. Cell **40**:9–17.

XIONG, Y., and T. H. EICKBUSH. 1988. The site-specific ribosomal DNA insertion element R1Bm belongs to a class of non-long-terminal repeat retrotransposons. Mol. Cell. Biol. **8**: 114–123.

YOKOYAMA, S., L. CHUNG, and T. GOJOBORI. 1988. Molecular evolution of the human immunodeficiency and related viruses. Mol. Biol. Evol. **5**:237–251.

YUKI, S., Y. INOUYE, S. ISHIMARU, and K. SAIGO. 1986*a*. Nucleotide sequence characterization of a *Drosophila* retrotransposon, 412. Eur. J. Biochem. **158**:403–410.

YUKI, S., S. ISHIMARU, S. INOUYE, and K. SAIGO. 1986*b*. Identification of genes for reverse transcriptase-like enzymes in two *Drosophila* retrotransposons, 412 and gypsy: a rapid detection method of reverse transcriptase genes using YXDD box probes. Nucleic Acids Res. **14**:3017–3030.

MASATOSHI NEI, reviewing editor

Received December 22, 1987; revision received June 21, 1988