Phylogenetic Relationships of Reverse Transcriptase and RNase H Sequences and Aspects of Genome Structure in the Gypsy Group of Retrotransposons[1]

Mark S. Springer, and Roy J. Britten†*

*Department of Biology, University of California; and †Division of Biology, California Institute of Technology

The *gypsy* group of long-terminal-repeat retrotransposons contains elements having the same order of enzyme domains in the *pol* gene as do retroviruses. Elements in the *gypsy* group are now known from yeast, filamentous fungi, plants, insects, and echinoids. Reverse transcriptase and RNase H amino acid sequences from elements in the *gypsy* group—including the recently described *SURL* elements, *TED, Cft1,* and *Ulysses,*—were aligned and analyzed by using parsimony and bootstrapping methods, with plant caulimoviruses and/or retroviruses as outgroups. Clades supported at the 95% level after bootstrapping include (1) *17.6* with *297* and (2) all of the *SURL* elements together. Other likely relationships supported at lower bootstrap confidence intervals include (1) *SURL* elements with *mag,* (2) *17.6* and *297* with *TED,* and this collective group with *412* and *gypsy,* (3) *Tf1* with *Cft1,* (4) *IFG7* with *Del,* and (5) all of the retrotransposons in the *gypsy* group together, to the exclusion of *Ty3.* In contrast with an earlier analysis, our results place *mag* within the *gypsy* group rather than outside of a cluster that contains *gypsy* group retrotransposons and plant caulimoviruses. Several features of retrotransposon genomes provide further support for some of the aforementioned relationships. The union of *SURL* elements with *mag* is supported by the presence of two RNA binding sites in the nucleocapsid protein. Location of the tRNA primer binding site and the presence of a long open reading frame 3′ to the *pol* gene support the *17.6-297-TED-412-gypsy* cluster.

Introduction

Retrotransposons containing long terminal repeats (LTRs) have now been identified in the genomes of a number of organisms and can be divided into two groups on the basis of both phylogenetic analysis of amino acid sequences and structural features of the genome (Xiong and Eickbush 1988, 1990; Doolittle et al. 1989). In the *copia* group, with representatives from *Drosophila* (*copia* and *1731*), yeast (*Ty1*), plants (*Tnt1, Ta1-3, Tst1, Wis,* and *Bis*), and *Physarum* (*Tp1*), the integrase gene is located between the protease and reverse transcriptase genes. In the *gypsy* group, with representatives from insects (*gypsy, 412, 17.6, 297, mag, micropia,* and *Ulysses*), yeast (*Ty3* and *Tf1*), filamentous fungi (*Cft1*), echinoids (*SURL* elements), and plants (*IFG7* and *Del*), the integrase gene is located 3′ to the RNase H gene. The *gypsy* group of LTR retrotransposons is related to plant caulimoviruses and to retroviruses, on the basis of reverse transcriptase sequences (Xiong and Eickbush 1990).

1. Key words: retrotransposon, reverse transcriptase, RNase H.

Address for correspondence and reprints: Mark S. Springer, Division of Biology, California Institute of Technology, Pasadena, California 91125.

Mol. Biol. Evol. 10(6):1370–1379. 1993.

0737-4038/93/1006-0017$02.00

Here, we examine phylogenetic relationships among members of the *gypsy* group by using amino acid sequences from the reverse transcriptase and RNase H proteins. Previous phylogenetic analyses of the *gypsy* group include Doolittle et al. (1989) and Xiong and Eickbush (1990). Xiong and Eickbush (1990) included 10 elements from the *gypsy* group in their analysis of reverse transcriptase sequences. Since that time, sequences for *SURL* elements, *TED, Tf1, Cft1,* and *Ulysses* have become available. We also evaluate the distribution and evolution of structural features in these retrotransposons in the light of amino acid–based phylogenies. Several structural features corroborate phylogenetic analysis on the basis of amino acid sequences.

Methods

Amino acid sequences and features of retrotransposons were obtained from GenBank and from references given in figure 1. Sequences of representative plant caulimoviruses were also obtained from GenBank. Delineation of boundaries for the reverse transcriptase protein correspond to that used by Xiong and Eickbush (1990). Delineation of RNase H sequence boundaries roughly corresponds to the region identified by McClure (1991). Multiple alignments were made by using CLUSTAL (Higgins and Sharp 1988), and adjustments were made by eye when conserved residues defined in Xiong and Eickbush (1990) and McClure (1991) were not aligned. Maximum parsimony and bootstrapping were performed by using PAUP, version 3.0s (Swofford 1991), with gaps counted as missing data. Plant caulimoviruses and/or retroviruses were used as outgroups. Each step on a parsimony tree corresponds to a single amino acid replacement. Because exact methods of finding minimum-length trees could not be used for the complete set of sequences, a heuristic approach using 100 replications with random input orders was employed. We also used a starting tree consistent with the tree given in Xiong and Eickbush (1990) as a baseline for searching for shorter trees. A distance matrix based on the aligned amino acid sequences was constructed by using the Kimura (1983, p. 175) option of the PROTDIST program on PHYLIP and was analyzed by using the neighbor-joining method (Saitou and Nei 1987).

Results

Alignments

Figure 1 shows a multiple alignment of amino acid sequences from the reverse transcriptase region. Overall, this alignment is similar to that of Xiong and Eickbush (1990), and most of the conserved blocks in their alignment are retained in the present alignment. Figure 2 shows an alignment of sequences from the RNase H region.

Phylogenetic Trees

Two minimum-length trees containing 2,219 amino acid replacements were found for the combined reverse transcriptase/RNase H sequences. One of these trees, rooted by using plant caulimoviruses, is shown in figure 3. On the second tree (not shown) the *Tf1-Cft1* group and *IFG7-Del* groups switch positions, and *micropia* is closer to *Ulysses* than to the *SURL-mag* group. Also shown on the tree in figure 3 are the consensus results of 500 bootstrap replications. Results summarized in figure 3 show (1) a likely sister-group relationship (86%) of *TED* (from the cabbage looper *Trichoplusia ni*) with *17.6* plus *297* (from *Drosophila*), (2) a likely sister-group relationship (73%) between the plant retrotransposons *IFG7* and *Del,* (3) a likely sister-group relationship (80%) between *SURL* elements and *mag,* (4) a likely sister-group rela-

```
17.6          LNQGIIRTSNSP--YNSPIWVVPKKQDA------SGKQKFRIVIDYR-KLNEITVGDRHPIPNMDEILGKLGRC-NYFTTIDLAKGFHQIEMDPESVSKTAFST
297           LNQGLIRESNSP--YNSPTWVVPKKPDA------SGANKYRVVIDYR-KLNEITIPDRYPIPNMDEILGKLGKC-QYFTTIDLAKGFHQIEMDEESISKTAFST
TED           LDQGIIRPSDSA--WSSPIWVVPKKIDA------SGKQKWRLVVDFR-KLNEKTIDDKYPIPNISDVLDKLGKC-QYFTTLDLASGFYQVEMDPQDISKTAFNV
412           IKDKIVEPSVSQ--YNSPLLLVPKKSSP-----NSDKKKWRLVIDYR-QINKKLLADKFPLPRIDDILDQLGRA-KYFSCLDLMSGFHQIELDEGSRDITSFST
gypsy (Dm)    LKDGIIRPSRSP--YNSPTWVVDKKGTD-----AFGNPNKRLVIDFR-KLNEKTIPDRYPMPSIPMILANLGKA-KFFTTLDLKSGYHQIYLAEHDREKTSFSV
gypsy (Dv)    LDDGIIRPSRSP--YNSPTWVVDKKGTD-----SYGNPKKRLVIDFR-KLNEKTIPDRYPMPRIPMILANLGKA-KYFTTLDLKSGYHQIYLAEHDREKTSFSV
SURL (Sp)     M--DVITRVDEPTDWVSSLVVVMKKNGQ-----------LRVCLDPR-DLNRAIKREHYQLPSRAEITAHFAGA-KYFSKLDASSGFWQIQLDDESSKLCTFIT
SURL (Tg)     L--DVITPVDEPTDWVSSLVVVMKKNGQ-----------LRVCLDPR-DLNRAIKREHYQLPSRAEITAHFAGA-KYFSKLDASSGFWQIQLDDESSKLCTFIT
SURL (Lv)     M--DVIKQVDEPTDWVSSLVVVKKKNGQ-----------LRVCLDPR-DLNRAIKREHYKLPSRAEITSQFAGA-KYFSKLDASSGFWQIQLDEDSSKLCTFIT
mag           LAAGVIKPVDHS-DWATPLVVVRKADGG-----------LRICADYKVTLNKVLAIDRFPVPKMEDLFSNLSGN-KFFTKLDLSQAYNQIVLSERSSEYTVINT
micropia      IRCNIVRPSCSP--FASPMLLVKKKNGT-----------DRLCVDFR-ELNSNTISDKYPLPLISDQIARLRGA-NYFTCLDMASGFHQIPIHPESVEYTAF--
Ulysses       LKLGIIEESDSP--WSNRTTVVMRPGKN------------RFCLDAR-KLNSVTVKDAYPLPCIEGILSRSTRL-ILSLASTLSSRSGNRDGGEEQGVYGVYCT
IFG7          LEAGIIQPSQSS--FSDPVVLVHKKDGS-----------WCMCPDYR-ELNKLTIKDKFPIPVIDELLDELHGS-IYFTKLDLRSGYHQIRMKTEDIPKTTFRT
Del           LNKGFIRGSTSP--WGAHVLFDPKKDDS-----------KRMCIDY*-KLNSVTVKNKYPLPRIDDLFDQLNGA-*YFSKIDLRFRYHQLRIRA*DIPKTAFRT
Tf1           LKSGIIRESKAI--NACPVMFVPKKEGT-----------LRMVVDYK-PLNKYVKPNIYPLPLIEQLLAKIQGS-TIFTKLDLKSAYHLIRVRKGDEHKLAFRC
Cft1          LAKGWIRRSTSS--AGTPCMFVPKANGK-----------LRLVQDYR-KLNEITIKNRYPLPNIEEAQDRLTGS-DWYTKIDLRDAFYAIRMAEGEEWKTAFRT
Ty3           LDNKFIVPSKSP--CSSPVVLVPKKDGT-----------FRLCVDYR-TLNKATISDPFPLPRIDNLLSRIGNA-QIFTTLDLHSGYHQIPMEPKDRYKTAFVT
COYMV         LQMKVIRPSESK--HRSTAFIVRSGTEIDPITGKEKKGKERMVFNYK-LLNENTESDQYSLPGINTIISKVGRS-KIYSKFDLKSGFWQVAMEEESVPWTAFLA
CERV          LELKVIKPSKST--HMSPAFLVENEAER-------RRGKKRMVVNYK-AMNKATKGDAHNLPNKDELLTLVRGK-KIYSSFDCKSGLWQVLLDKESQLLTAFTC
FIGWORT       LDLGLIIPSKSQ--HMSPAFLVENEAER-------RRGKKRMVVNYK-AINQATIGDSHNLPNMQELLTLLRGK-SIFSSFDCKSGFWQVVLDEESQKLTAFTC
CMV           LDLKVIKPSKSP--HMAPAFLVNNEAEK-------RRGKKRMVVNYK-AMNKATVGDAYNLPNKDELLTLIRGK-KIFSSFDCKSGFWQVLLDQESRPLTAFTC
HIV-I         EGKISKIGPENP--YNTPVFAIKKKDST----------KWRKLVDFR-ELNKRTQDFWEVQLGI-PHPAGLKKK-KSVTVLDVGDAYFSVPLDKDFRKYTAFTI
HUER          LEKGHIEPSFSP--WNSPVFVIQKKSGK-----------WHTLTDLR-AVNAVIQPMGPLQPGL-PSPAMIPKD-WPLIIIDLKDCFFTIPLAEQDCEKFAFTI
MOUSE IAP     LKLGHIDPSTSP--WNTPIFVIKKKSGK-----------WRLLHDLR-PINEQMNLFGPVQRGL-PVLSALPRG-WNLIIIDIKDCFFSIPLCPRDRPRFAFTI
RSV           LQLGHIEPSLSC--WNTPVFVIRKASGS-----------YRLLHDLR-AVNAKLVPFGAVQQGA-PVLSALPRG-WPLMVLDLKDCFFSIPLAEQDREAFAFTL
BLV           LEAGYISPWDGP--GNNPVFPVRKPNGA-----------WRFVHDLR-ATNALTKPIPALSPGP-PDLTAIPTHPPHIICLDLKDAFFQIPVEDRFRFYLSFTL
MOMLV         LDQGILVPCQSP--WNTPLLPVKKPGTN----------DYRPVQDLR-EVNKRVEDIHPTVPNPYNLLSGLPPSHQWYTVLDLKDAFFCLRLHPTSQPLFAFEW
HTLV-I        LEAGHIEPYTGP--GNNPVFPVKKANGT-----------WRFIHDLR-ATNSLTIDLSSSSPGP-PDLSSLPTTLAHLQTIDLKDAFFQIPLPKQFQPYFAFTV
```

```
17.6          KH-------GHYEYLRMPFGLKNAPATFQRCMNDILRPLLNKH----CLVYLDDIIVFSTSLDEHLQSLGLVFEKLAKANLKLQLDKCEFLKQETTFLGHVL
297           KS-------GHYEYLRMPFGLRNAPATFQRCMNNILRPLLNKH----CLVYLDDIIIFSTSLTEHLNSIQLVFTKLADANLKLQLDKCEFLKKEANFLGHIV
TED           EH-------GHFEFLRMPMGLKNSPSTFQRVMDNVLRGLQNNI----CLVYLDDIIVYSTSLQEHLENLERVFQRLRESNFKIQMDKSEFLKLETAYLGHII
412           SN-------GSYRFTRLPFGLKIAPNSFQRMMTIAFSGIEPSQ----AFLYMDDLIVIGCSEKHMLKNLTEVFGKCREYNLKLHPEKCSFFMHEVTFLGHKC
gypsy (Dm)    NG-------GKYEFCRLPFGLRNASSIFQRALDDVLREQIGKI----CYVYVDDVIIFSENESDHVRHIDTVLKCLIDANMRVSQEKTRFFKESVEYLGFIV
gypsy (Dv)    SS-------GKYEFCRLPFGLRNASSIFQRAIDDILREHIGKI----CFVYVDDVIIFSKNETEHLQHINIVLKCLIDANMRVGPEKTRFFKESIEFLGFIV
SURL (Sp)     PY-------GRYKFLRLPFGICSAPEVYPKIVHQLFAHIPG------VNTMMDDVIVWRTTQQERDDRLRKVLSIVRKMNLKLNKDKCEFNVKKLTFIGDLI
SURL (Tg)     PY-------GRYKFLRLPFGICSAPEVYHKIVHQMFAHIPG------VNTMMDDVIVWGTTQQEHDNRLREVLSIARRMNLKLNKDKCEFSVKKLTFIGDLI
SURL (Lv)     PY-------GRYRFLRLPFGICSAPEVFHKIIHNLFVDIPG------VNTMMDDIVVWGSTQEEHDDRLRKVLDIAQKSNLKLNRDKCEFNVNQMTFIGDLI
mag           HR-------GLFKYSRLVYGLASSPGIFQKLMVNMFKNVPN------VVVFYDDILIRNQDLDSHLKSIKEVLDILERYGLKIKRSKCEFMVTEVRYLGFII
micropia      -----------------VPDGLKNAPSVFQRTVINALGDLANSF----VIVYMDDIMVVSPTKELALERLKTVLNVLTKAGFTFNLAKCSFLKTTVQYLGYEV
Ulysses       RR-------PLYQFRHMPFGLCNAAQHFEA-HDKVIPANLRSN----VFVYLDDLLIISADFPTHLKYLELVAECLRNANLTIGMAKSKFLFRNLNYLGFIQ
IFG7          HE-------GHYEFFVMPFGLTNTPSTFQGLMNSIFKPFLRKF----VLVFFDDILIYNKSWKDHVEHVDRVLQLLEEKKLYAKRSKCEFVLQEVEYLGHIV
Del           RY-------GHYEFLVMPFGLTNVPTAFMNLMNRVFREYLDKF----IVVFVDYVLIYSRTQKDHEHHLRISLQLLRNNQLYAKLSKCEFWMEKVKFLGHVV
Tf1           PR-------GVFEYLVMPYGISTAPAHFQYFINTILGEAKESH----VVCYMDDILIHSKSESEHVKHVKDVLQKLKNANLIINQAKCEFHQSQVKFIGYHI
Cft1          RY-------GLYEFLVMPMGLTNAPASCQDLVNETLRDLLDVC----VVAYMDDILVYTKGSLQEHTKQVQDVFERLTKSGFKTAPEKCEFHKKEVKFLGFII
Ty3           PS-------GKYEYTVMPFGLVNAPSTFARYMADTFRDLRF------VNVYLDDILIFSESPEEHWKHLDTVLERLKNENLIVKKKKCKFASEETEFLGYSI
COYMV         GN-------KLYEWLVMPFGLKNAPAIFQR-KMDNVFKGTEKF----IAVYIDDILVFSETAEQHSQHLYTMLQLCKENGLILSPTKMKIGTPEIDFLGASL
CERV          PQ-------GHYQWNVVPFGLKQAPSIFPKTYANSHSNQYSKY----CCVYVDDILVFSNTGRKEHYIHVLNILRRCEKLGIILSKKKAQLFKEKINFLGLEI
FIGWORT       PQ-------GHFQWKVVPFGLKQAPSIFQR-HMQTALNGADKF----CMVYVDDIIVFSNSELDHYNHVYAVLKIVEKYGIILSKKKANLFKEKINFLGLEI
CMV           PQ-------GHYEWNVVPFGLKQAPSIFQR-HMDEAFRVFRKF----CCVYVDDILVFSNNEEDHLLHVAMILQKCNQHGIILSKKKAQLFKKKINFLGLEI
HIV-I         PSINNETPGIRYQYNVLPQGWKGSPAIFQSSMTKILEPFRKQNPDIVIYQYMDDLYVGSDLEIGQHRTKIEELRQHLLRWGFTTPDKKHQK-EPPFLWMGYEL
HUER          PAINNKEPATRFQWKVLPQGMLNSPTICQTFVGRALQPVREKFSDCYIIHYIDDILCAAETKDK-LIDCYTFLQAEVANAGLAIASDKIQT-STPFHYLGMQI
MOUSE IAP     PSINSDEPDNRYQWKVLPQGMSNSPTMCQLYVQEALLPVREQFPSLILLLYMDDILLCHKELTM-LQKAYPFLLKTLSQWGLQIATEKVQI-SDTGQFLGSVV
RSV           PSVNNQAPARRFQWKVLPQGMTCSPTICQLVVGQVLEPLRLKHPSLCMLHYMDDLLLAASSHDG-LEAAGEEVISTLERAGFTISPDKVQR-EPGVQYLGYKL
BLV           PSPGGLQPHRRFAWRVLPQGFINSPALFERALQEPLRQVSAAFSQSLLVSYMDDILYASPTEEQ-RSQCYQALAARLRDLGFQVASEKTSQTPSPVPFLGQMV
MOMLV         RDPE-MGISGQLTWTRLPQGFKNSPTLFDEALHRDLADFRIQHPDLILLQYVDDLLLAATSELD-CQQGTRALLQTLGNLGYRASAKKAQICQKQVKYLGYLL
HTLV-I        PQQCNYGPGTRYAWRVLPQGFKNSPTLFEMQLAHILQPIRQAFPQCTILQYMDDILLASPSHAD-LQLLSEATMASLISHGLPVSENKTQQTPGTIKFLGQII
```

FIG. 1.—Alignment of amino acid sequences from the reverse transcriptase region for the *gypsy* group of retrotransposons, several plant caulimoviruses, and several retroviruses. Abbreviations for plant caulimoviruses and retroviruses are as follows: COYMV = *Commelina* yellow mottle virus; CERV = carnation etched-ring virus; FIGWORT = figwort mosaic virus; CMV = cauliflower mosaic virus; HIV-I = human immunodeficiency virus I; HUER = human endogenous retrovirus K; MOUSE IAP = mouse intracisternal A particle; RSV = Rous sarcoma virus; BLV = bovine leukemia virus; MOMLV = Moloney murine leukemia virus; and HTLV-I = human T-cell leukemia virus I. An asterisk (*) denotes a stop codon. References for the *gypsy* retrotransposons are as follows: *17.6* (Saigo et al. 1984), *297* (Inouye et al. 1986), *412* (Yuki et al. 1986), *gypsy* (Marlor et al. 1986; Mizrokhi and Mazo 1991), *micropia* (Lankeneu et al. 1988), and *Ulysses* (Scheinker et al. 1990) are from *Drosophila; TED* (Friesen and Nissen 1990) is from the cabbage looper (*Trichoplusia ni*); *SURL* elements (Springer et al. 1991) are from echinoids; *mag* (Michaille et al. 1990) is from the silk moth (*Bombyx mori*); *IFG7* (Kossack 1989) is from *Pinus radiata; Del* (Sentry and Smyth 1989; Smyth et al. 1989) is from the lily (*Lilium henryi*); *Tf1* (Levin et al. 1990) is from fission yeast (*Schizosaccharomyces pombe*); *Cft1* (McHale et al. 1992) is from *Cladosporium fulvum;* and *Ty3* (Hansen et al. 1988) is from *Saccharomyces cerevisiae.*

```
17.6          DFTKKFTLTTDASDVALGAVLSQDGH---------PLSYISRTLNEHEINYSTIEKELLAIVWATKTFRHYLLGR
297           DFEKKFVLTTDASNLALGAVLSQNGH---------PISFISRTLNDHELNVSAIEKELLAIVWATKTFRHYLLGR
TED           DFTREFNLTTDASNFAIGAVLSQGPIGSDK-----PVCVASRTLNESELNYSTIEKELLAIVWATKVFRPVLFGR
412           DFSKEFCITTDASKQACGAVLTQNHNGHQL-----PVAYASRAFTKGESNKSTTEQELAAIHWAIIHFRPVIYGK
gypsy (Dm)    DFKKPFDLTTDASASGIGAVLSQEGR---------PITMISRTLKQPEQNVATNERELLAIVWALGKLQNFLYGS
gypsy (Dv)    NFQKPFDLTTDASASGIGAVLSQGNR---------PITMISRALKQAEQNVATNERELLAIVWALGRLQNFLYGS
SURL (Sp)     DCKKLTKLSADASKDGIGAVLLQQYDQDWV-----PIAYASRSMTDAETRYAQIEKELLAITYACERFHQYIYGQ
SURL (Tg)     DCNKPTKLSADASKNGIGAVLLQQHDENWV-----PIAYASRSMTDAETRYAQIEKELLAITYACERFHQYIYGQ
SURL (Lv)     DCTKPTKISADASKNGLGAVLLQQHEQNWH-----PIAYASRAMTDAETRYAQIEKELLAITYGCEKFHQYIYGQ
mag           DMSLESVLTVDASARGLGAVLAQRGPGCQER----VVAYASRALTTHELHYSQIHKEALAIVFAVEKFHQYLYGR
micropia      DPQVPIELHTDASACGVGAILLHRIESKPH-----VIEVFSKTTTSVESRVHSVELETLAVVKAVKHFRHYLIGR
Ulysses       DFRRPFFIQCDASHYGVGAVLFQLDDEQQER----PIAFFSAKLNKHQINYSVTEKECLAAKLAIHRFRPYVEMM
IFG7          DFMKTFIVECDASGNGIGAVLMQDEI---------PIAFEGHPIRGKFLHKALYEKEMLAILHALKKWRPVLMGR
Del           ISG*PFVVVTDASLAGLEGVLMQDGR---------VVAYASRQLKVHENNVPTHDLELAVVIFILKLWRHVLYGE
Tf1           DFSKKILLETDASDVAVGAVLSQKHDDDKVV----PVGVYSAKMSKAQLNYSVSDKEMLAIIKSLKHWRHYLEST
Cft1          DGSKEVHIETDASDMAIGACLTQTHDGKRH-----PVAYVSRKMTTAEQNYDIHDKELLAIVAAMQHWRVYVEGP
Ty3           NNKANVRLTTDASKDGIGAVLEEVDNKNKLVG---VVGVFSKSLESAQKNVPAGELLLGIIKALHHFRVYMLHGK
COYMV         PKDSFIIIETDGCMTGWGAVCKWKMSKHDPRSTERICAYASGSFNPIKS---TIDAEIQAAIHGLDKFKIVYLDK
CERV          EPNDKLVIETDASEEFWGGILKAIHN-----SHEVICRYASGSFKAAERNVHSNEKELLAVIRVIKKFSIYLTPS
FIGWORT       KPEDHLIIETDASDSFWGGVLKARALD----GVELICRYSSGSFKQAEKNVHSNDKELLAVKQVITKFSAYLTPV
CMV           LPEEKLIIETDASDDYWGGMLKAIKIN-EGTNTELICRYASGSFKAAEKNVHSNDKETLAVINTIKKFSIYLTPV

17.6          --HFEISSDHQPLSWLVRMK-----DPNS-KLTRW---RVKLSEFDFDI--KVIKGKENCVADALSRIKLEETY
297           --QFLIASDHQPLRWLHNLK-----EPGA-KLERW---RVALSEYQFKI--DVIKGKENSVADALSRIKIEENH
TED           --KFKILTDHKPLQWMNNLK-----DPNS-RMTRW---RLALSEYDFSV--UVKKGKSNTNADALSRVEIHTTE
412           --HFTVKTDHRPLTYLFSMV-----NPSS-KLTRI---RLELEEYNFTV--EVLKGKDNHVADALSRITIKELK
gypsy (Dm)    -REINIFTDHQPLTFAVADR-----NTNA-KIKRW---KSVIDQHNAKV--FVKPGKENFVADALSRQNLNALQ
gypsy (Dv)    -REINIFTDHQPLTFAVSDK-----NTNS-KIKRW---KSVIDQHNAKM--FVKPGKENLVADALSRQNINALE
SURL (Sp)     --QVEVETDHKPHIPLFVKSLG---DCPL-RIQRL---LIRVQRYDLKV--VVTPGKVMYTADTLSRAVDPKAE
SURL (Tg)     --QVEVETDHKPLIPLFVKSLG---DCPL-RIQRL---LIRVQRYDLKV--MVTPGKVMYTADTLSRAVDPKAD
SURL (Lv)     --KIEVETDHKPLIPLFVKSLA---DCPL-RIQRL---LIRVQRYDLKV--SVTPGKVMFTADTLSRAVDPKAE
mag           --KFILRTDHKPLVSIFGPNIGIPSAAAS-RLQRW---AIKLSAYDFEI--EVVRT-DKNVADALSRLIESQKN
micropia      --EFVVVTDCNSLKASRTKI-----DLTP-RVHRW---WAVLQSFNFEI--QVREGKRMAHVDFLSRNPLSPEH
Ulysses       --PFTVITDHASLQWLMSLK-----DLSG-RLARW---SLELQAFPFSM--QVRKGADNVCRHIVRSVEEVELT
IFG7          --HFNVKTDHDSLKVFLEQR-----LSSE-E*QKW---VTKMLGYDFEI--IVKKGKKNVVANALSRKDEDVEA
Del           --DFELYCDHKSLKYISTQK-----DLNL-RQR*W---IEVLKDFDFSI--FVHPGKANVVADALSRKSQISHL
Tf1           IEPFKILTDHRNLIRITNGE----SEPENKRLARW---QLFLQDFNFEI--NVRPGSANHIADALSRIVDETEP
Cft1          -PKLTILSDHKNLTYFTTTK-----ELTR-RQARW---SELLGQVKFEI--KVTPGTENGPADALSRRSDYMEG
Ty3           --HFTLRTDHISLLSLQNKN-----EPAR-RVQRW---LDDLATVDFTL--EVLAGPKNVVADAISRAIVTITP
COYMV         -KELIIRSDCEAIIKFYN---KTNENKP--SRVRWLTFSDFLTGLGITVTFEHIDGKHNGLADALSRMINFIVE
CERV          --RFLIRTDNKNFTHFVNIMLKGDRKQG--RLVRW---QMWLSQVDFDV--EHIAGTKNVFADFLQENTLTNYV
FIGWORT       --RFTVRTDNKNFTVFLRINLKGDSKQG--RLVRW---QNWFSKVQFDV--EHLEGVKNVLADCLTRDFNA---
CMV           --HFLIRTDNTHFKSFVNLNVKGDSKLG--RNIRW---QAWLSHVSFDV--EHIKGTDNHFADFLSREFNKVNS
```

FIG. 2.—Alignment of amino acid sequences from the RNase H region for the *gypsy* group of retrotransposons and several plant caulimoviruses. Abbreviations are given in fig. 1. An asterisk (*) denotes a stop codon.

tionship (81%) between *Tf1* (from fission yeast) and *Cft1* (from the fungal tomato pathogen *Cladosporium fulvum*), and (5) a possible clade (62%) containing *17.6, 297, TED, 412,* and *gypsy*. In addition, *Ty3* is an outgroup to all other retrotransposons in the *gypsy* group on 70% of the bootstrap trees. *Ulysses* and *micropia* group with *SURL* elements and *mag* on both minimum-length trees, but this association does not hold up after bootstrapping. Likewise, the minimum-length tree shown in figure 3 supports a clade containing all of the retrotransposons that occur in metazoans, but this branch does not occur on the second minimum-length tree, nor is it supported by bootstrapping. In contrast to the tree in figure 3, the shortest tree consistent with that of Xiong and Eickbush (1990) is 30 steps longer, at 2,249 steps.

When we converted our sequence alignments to distances by using the Kimura option of PROTDIST (PHYLIP, version 3.5; Felsenstein 1993) and then employed the neighbor-joining method, the resulting tree (not shown) showed some differences from the minimum-length trees, but all of the branches that are supported at the 50% level in figure 3 are also supported on the neighbor-joining tree.

Minimum-length trees (not shown) based on reverse transcriptase versus RNase H sequences exhibit several conflicts; for example, *SURL* elements cluster with *mag* on reverse transcriptase trees but cluster with the two *gypsy* elements on RNase H trees. However, all of the conflicts involve branches that are not supported after boots-

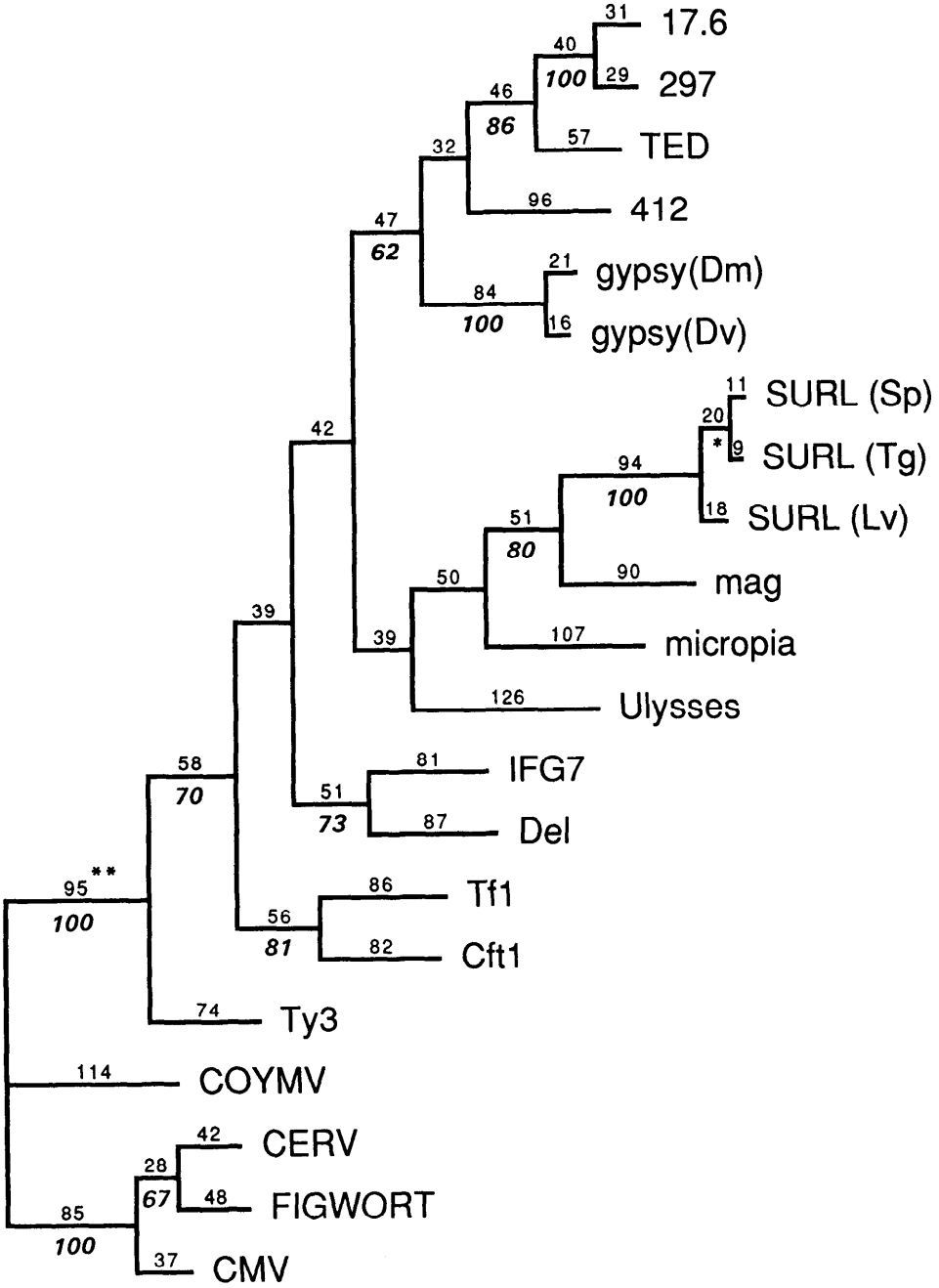

FIG. 3.—One of two minimum-length trees at 2,219 replacements. Numbers above the line are the number of amino acid replacements, and numbers below the line are the percentages, from 500 bootstrap trials, that support the clade. One asterisk (*) denotes a value of 99%, and two asterisks (**) denote the branch of this tree that the root would be on if the reverse transcriptases of the seven retroviruses were used for rooting.

trapping. Bootstrapping the reverse transcriptases and RNase H sequences, respectively, provides support for the following: *17.6* plus *297* (94% and 97%), and for this group with *TED* (85% and 64%); the two *gypsy* elements together (100% and 100%); the three *SURL* elements together (100% and 100%) with *SURL* (Sp) and *SURL* (Tg) as nearest neighbors (96% and 90%); and *Tf1* plus *Cft1* (60% and 73%). In addition, bootstrapping reverse transcriptase sequences provides support for *SURL* elements with *mag* (60%), *IFG7* plus *Del* (71%), and all of the retrotransposons together, except *Ty3* (52%).

Features of *gypsy*-like Elements

Table 1 summarizes the distribution of seven different features of retrotransposons in the *gypsy* group. The phylogenetic significance of these features is discussed below.

Discussion

Xiong and Eickbush (1988, 1990) previously examined relationships among retroid elements, including retrotransposons in the *gypsy* group, on the basis of reverse transcriptase sequences. One of the differences on the Xiong and Eickbush (1990) tree is that *mag* is outside of a cluster containing other retrotransposons in the *gypsy* group as well as plant caulimoviruses. To test this hypothesis with our data, it was necessary to include retroviruses as an outgroup to the collective group. We limited this analysis to reverse transcriptase sequences because of the difficulty in aligning RNase H sequences. Retroviruses clearly root the tree (not shown) such that the plant-caulimovirus and retrotransposon groups (including *mag*) are each monophyletic.

Two other differences on the Xiong and Eickbush (1990) tree are as follows: (1) *Ty3* is not peripheral to other *gypsy* retrotransposons but occupies a position close to *IFG7* and *Del*, and (2) *412* is the most peripheral member of the *gypsy* cluster, except *mag*. Whether we (1) use parsimony or neighbor-joining methods, (2) include RNase H and reverse transcriptase or just reverse transcriptase sequences, or (3) restrict our analysis to the reverse transcriptase sequences available to Xiong and Eickbush (1990), *Ty3* occupies the most peripheral position among retrotransposons in the *gypsy* group, and *412* clusters with the insect elements *gypsy, 297, 17.6,* and *TED*.

The overall congruence between reverse transcriptase and RNase H bootstrap trees indicates that a similar phylogenetic signal is present in both, although, when taken separately, each of these proteins provides less resolution than they do in combination with each other. One of the implications of the overall congruence between bootstrap trees is that reverse transcriptase and RNase H have similar evolutionary histories without any interelement recombination that might cause striking differences.

If *Ty3* is taken as an outgroup to all of the other retrotransposons, then the implied primitive character states for the characters in table 1 are +1 ribosomal frameshifting, one RNA binding site, tRNA methionine, a +2 location of the tRNA primer binding site (PBS), and lack of a long open reading frame (ORF) 3′ to the *pol* gene. On the basis of these designations of primitive character states, several of the aspects of genome structure given in table 1 offer additional support for some of the branches on the tree in figure 3. First, *17.6, 297,* and *TED* are united by the putative shared derived character of tRNA serine, although a putative tRNA serine also occurs in *Cft1* (McHale et al. 1992). Second, *17.6, 297, TED, 412,* and *gypsy* share a number of putative derived characters, including a long ORF 3′ to the *pol* gene, a 1-bp overlap of the 5′ LTR and the tRNA PBS, and −1 frameshifting of the *pol* gene relative to the *gag* gene, as well as the absence of RNA binding sites in the nucleocapsid protein. While two of these derived characters have evolved elsewhere on the tree (i.e., −1

Table 1
Select Features of Retrotransposons Related to the *gypsy* Element of *Drosophila*

Feature	Retrotransposon[a]													
	17.6	*297*	*TED*	*412*	*gypsy*[b]	*SURL*[c]	*Mag*	*micropia*	*Ulysses*	*IFG7*	*Del*	*Tf1*	*Cft1*	*Ty3*
1. Ribosomal frameshifting	Yes	Yes	Yes	Yes	Yes	No	Yes	No	Yes	No	No	No	Yes	Yes
No. of nucleotides[d]	−1	−1	−1	−1	−1	0	−1		+1				−1	+1
2. No. of RNA binding sites	0	0	0	0	0	2	2	2		1	1	0 (?)	1	1
3. tRNA PBS	Serine	Serine	Serine	Arginine	Lysine	Methionine	Arginine	Leucine	Lysine	Methionine	Methionine	?	Serine (?)	Methionine
4. Location of tRNA PBS[e]	−1	−1	−1	−1	−1	0	+1	0	+1	+2	+3	?	0	+2
5. Envelope (?) ORF[f]	Yes	Yes	Yes	Yes	Yes	No	No	No	No	No	No	No	No	No
6. LTR length (in bp)	512	415	273	481–571	479–482	254	77	474–504	2,156	333	2,406–2,415	358	427	340
7. Element length (in bp)	7,439	6,995	7,510	7,400	7,469	5,266	4,564	5,500	10,653	5,944	9,345	4,941	6,968	5,428

[a] Boxes enclose clades of fig. 3.
[b] Dm strain.
[c] Tg strain.
[d] No. of nucleotides that must "read" to change from the frame reading "xxx" to that reading "yyy."
[e] Relative to the 5′ LTR. Minus (−) values are base pair of overlap; plus (+) values are the distance between.
[f] A number of retrotransposons contain long (~1,500-bp) ORFs located between the *pol* gene and the 3′ LTR. In retroviruses, the ORF in this region codes for proteins that are essential in the extracellular stage of the life cycle; however, there is as yet no direct evidence that the similarly placed ORF that occurs in some retrotransposons codes for homologous proteins.

frameshifting also occurs in *Cft1* and *mag,* and RNA binding sites are absent in *Ulysses*), the presence of a long ORF 3′ to *pol* and a −1 location of the tRNA PBS are unique to this subset of the *gypsy* group. Third, the putative relationship between *mag, SURL* elements, and possibly *micropia* is potentially strengthened by the exclusive occurrence of two RNA binding sites in the nucleocapsid protein in all of these elements. Most retroviruses also possess two RNA binding sites, but in the somewhat more closely related plant caulimoviruses there is only a single site. Further support for the alliance between *mag* and *SURL* elements comes from the observation that the number of amino acids separating the two RNA binding sites is identical in these elements. *Micropia,* in turn, has 14 additional amino acids that separate the first and second RNA binding sites. The plant elements *Del* and *IFG7* share a number of features, such as a single RNA binding site, a single ORF containing the *gag* and *pol* genes, and a tRNA methionine PBS, but these features appear primitive on the basis of their occurrence in *Ty3.*

The long LTRs in *Ulysses* and *Del* appear homoplastic on the basis of other evidence discussed above, whereas the short LTRs in *mag* are unique to this element. Element length ranges from 4,564 bp in *mag* to 10,653 bp in *Ulysses* and reflects the differences in LTR length. Among other elements, most of the variation results not from differences in LTR length but rather from the additional ORF 3′ to the *pol* gene.

It is interesting that, for the tree in figure 3, all of the animal retrotransposons occur on one branch, whereas the two plant elements occur on a second branch. The separate clusters of plant and animal retrotransposons suggest that the host phylogeny imposes a distinct signature on the phylogeny of the retrotransposons; Flavell (1992) previously noted predominantly plant and animal groups for the *copia* group of retrotransposons as well. Flavell (1992) has also characterized the *copia* group as lacking ribosomal frameshifting, whereas in the *gypsy* group the *gag* and *pol* genes are always overlapping. However, the presence or absence of overlapping *gag* and *pol* genes is shown here to exhibit more variation in the *gypsy* group than was previously recognized.

In conclusion, our understanding of the phylogeny of the *gypsy* group of retrotransposons is enhanced by considering not only amino acid sequences but also genetic features of these elements. Some features (e.g., long 3′ ORF) show little or no homoplasy, whereas others (e.g., type of tRNA PBS) are labile and show much more homoplasy.

Acknowledgment

We would like to thank Walter Fitch and two anonymous reviewers for helpful comments.

LITERATURE CITED

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.

FELSENSTEIN, J. 1993. PHYLIP manual, version 3.5. University of Washington, Seattle.

FLAVELL, A. J. 1992. *Ty1-copia* group retrotransposons and the evolution of retroelements in the eukaryotes. Genetica **86**:203–214.

FRIESEN, P. D., and M. S. NISSEN. 1990. Gene organization and transcription of TED, a lepidopteran retrotransposon integrated within the baculovirus genome. Mol. Cell. Biol. **10**: 3067–3077.

HANSEN, L. J., D. L. CHALKER, and S. B. SANDMEYER. 1988. Ty3, a yeast retrotransposon associated with tRNA genes, has homology to animal retroviruses. Mol. Cell. Biol. **8**:5245–5256.

HIGGINS, D. G., and P. M. SHARP. 1988. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. Gene **73**:237–244.

INOUYE, S., S. YUKI, and K. SAIGO. 1986. Complete nucleotide sequence and genome organization of a *Drosophila* transposable genetic element, 297. Eur. J. Biochem. **154**:417–425.

KIMURA, M. 1983. The neutral theory of molecular evolution. Cambridge University Press, Cambridge.

KOSSACK, D. 1989. The IFG copia-like element: characterization of a transposable element present in high copy number in *Pinus* and a history of the pines using IFG as a marker. University Microfilms, University of Michigan, Ann Arbor.

LANKENEU, D.-H., P. HUIJSER, E. JANSEN, K. MIEDEMA, and W. HENNIG. 1988. Micropia: a retrotransposon of *Drosophila* combining structural features of DNA viruses, retroviruses and non-viral transposable elements. J. Mol. Biol. **204**:233–246.

LEVIN, H. L., D. C. WEAVER, and J. D. BOEKE. 1990. Two related families of retrotransposons from *Schizosaccharomyces pombe*. Mol. Cell. Biol. **10**:6791–6798.

MCCLURE, M. A. 1991. Evolution of retroposons by acquisition or deletion of retrovirus-like genes. Mol. Biol. Evol. **8**:835–856.

MCHALE, M. T., I. N. ROBERTS, S. M. NOBLE, C. BEAUMONT, M. P. WHITEHEAD, D. SETH, and R. P. OLIVER. 1992. CfT-I: an LTR retrotransposon in *Cladosporium fulvum*, a fungal pathogen of tomato. Mol. Gen. Genet. **233**:337–347.

MARLOR, R. L., S. M. PARKHURST, and V. G. CORCES. 1986. The *Drosophila melanogaster* gypsy transposable element encodes putative gene products homologous to retroviral proteins. Mol. Cell. Biol. **6**:1129–1134.

MICHAILLE, J.-J., S. MATHAVAN, J. GAILARD, and A. GAREL. 1990. The complete sequence of mag, a new retrotransposon in *Bombyx mori*. Nucleic Acids Res. **18**:674.

MIZROKHI, L. J., and A. M. MAZO. 1991. Cloning and analysis of the mobile element *gypsy* from *D. virilis*. Nucleic Acids Res. **19**:913–916.

SAIGO, K., W. KUGIMIYA, Y. MATSUO, S. INOUYE, K. YOSHIOKA, and S. YUKI. 1984. Identification of the coding sequence for a 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.

SCHEINKER, V. SH., E. R. LOZOVSKAYA, J. G. BISHOP, and M. B. EVGEN'EV. 1990. A long terminal repeat-containing retrotransposon is mobilized during hybrid dysgenesis in *Drosophila virilis*. Proc. Natl. Acad. Sci. USA **87**:9615–9619.

SENTRY, J. W., and D. R. SMYTH. 1989. An element with long terminal repeats and its variant arrangements in the genome of *Lilium henryi*. Mol. Gen. Genet. **215**:349–354.

SMYTH, D. R., P. KALITSIS, J. L. JOSEPH, and J. W. SENTRY. 1989. Plant retrotransposon from *Lilium henryi* is related to Ty3 of yeast and the *gypsy* group of *Drosophila*. Proc. Natl. Acad. Sci. USA **86**:5015–5019.

SPRINGER, M. S., E. H. DAVIDSON, and R. J. BRITTEN. 1991. Retroviral-like element in a marine invertebrate. Proc. Natl. Acad. Sci. USA **88**:8401–8404.

SWOFFORD, D. L. 1991. PAUP: phylogenetic analysis using parsimony, version 3.0s. Illinois Natural History Survey, Champaign.

XIONG, Y., and T. H. EICKBUSH. 1988. Similarity of reverse transcriptase-like sequences of viruses, transposable elements, and mitochondrial introns. Mol. Biol. Evol. **5**:675–690.

XIONG, Y., and T. H. EICKBUSH. 1990. Origin and evolution of retroelements based upon their reverse transcriptase sequences. EMBO J. **9**:3353–3362.

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

WALTER M. FITCH, reviewing editor

Received January 6, 1993; revision received July 5, 1993

Accepted July 23, 1993