Retrotransposons and Retroviruses: Analysis of the Envelope Gene

Emmanuelle Lerat and Pierre Capy
Laboratoire Populations, Génétique et Evolution, Centre National de la Recherche Scientifique, Gif-sur-Yvette, France

Retroviruses and long terminal repeat (LTR) retrotransposons share a common structural organization. The main difference between these retroelements is the presence of a functional envelope (env) gene in retroviruses, which is absent or nonfunctional in LTR retrotransposons. Several similarities between these two groups of retroelements have been detected for the reverse transcriptase, gag, and integrase domains. Assuming that each of these domains shares a common ancestral sequence, several hypotheses could account for the emergence of retroviruses from LTR retrotransposons. In this context, the positions of elements such as gypsy and the members of the Ty3 subfamily are not clear, since they are classified as retroviruses but phylogenetically they are assigned to the LTR retrotransposon group. We compared the env gene products of these retroelements and identified two similar motifs in retroviruses and LTR retrotransposons. These two regions do not occur in the same order. If we assume that they are derived from the same ancestral sequence, this could result from independent acquisition of the various domains rather than the single acquisition of the whole env gene. However, we cannot exclude the possibility that the env gene was reorganized after being acquired. Trees based on these regions show that these two groups of elements are clearly distinguished. These trees are similar to those obtained from reverse transcriptase or integrase. In trees based on reverse transcriptase, the retroviruses with complete or partial env genes can be distinguished from the other LTR retrotransposons.

Introduction

Transposable elements (TEs) are divided into two main classes, depending on whether their transposition mechanisms use RNA or DNA intermediates (Finnegan 1989), and into several subclasses based on the presence/absence or order of domains (Capy et al. 1997b). In spite of clear structural differences, some similarities between retroviruses and retrotransposons can be detected at the protein level for domains in gag and pol genes like reverse transcriptase, Rnase H, protease, and integrase (see Capy et al. 1997a and references therein). If these similarities are not due to convergence, they must reflect common ancestral sequences. Assuming this, they can be used to infer the evolutionary history of different domains (Xiong and Eickbush 1990; McClure 1991; Capy et al. 1996, 1997a).

One of the main questions raised by the evolution of TEs is that of their relationships to retroviruses. The possibility of such a connection was initially discussed by Temin (1980) and Flavell (1981). Long terminal repeat (LTR) retrotransposons, and, more specifically, those of the Ty3 superfamily, share some common structures with retroviruses, such as LTR, gag, and pol genes and, in some cases, incomplete and nonfunctional env genes. Several evolutionary scenarios have already been suggested on the basis of these similarities. For instance, Xiong and Eickbush (1990) and McClure (1991) have studied reverse transcriptase similarities in several retroelements, including retrotransposons and retroviruses. These authors suggest that retroviruses may evolve from LTR retrotransposons by acquiring a functional env gene. Capy et al. (1996) have also suggested this on the basis of comparison of DDE integrase/transposase.

Key words: transposable elements, retrotransposons, retroviruses, envelope gene, evolution.

Although the transition from LTR retrotransposons to retroviruses seems to predominate, some oscillation between them cannot be excluded. A retrovirus, which loses the activity of its envelope gene (env), becomes an LTR retrotransposon. A clear definition of both is required before we can analyze conversions between LTR retrotransposons and retroviruses. In this paper, we will consider two distinct retroelements: the retroviruses and the errantiviruses. Following the nomenclature proposed by Boeke et al. (1998), the errantiviruses are LTR retrotransposons with env-like genes. In this context, the gypsy element, first reported as an LTR-retrotransposon (Bayev et al. 1984), was then considered an insect retrovirus (Kim et al. 1994; Song et al. 1994). Phylogenies based on reverse transcriptase or integrase clearly show that this element is always classified as an LTR retrotransposon (Xiong and Eickbush 1990; Capy et al. 1996, 1997a). Therefore, in the present work, this element will be considered as an errantivirus.

Several domains of these two entities have already been compared, including reverse transcriptase, integrase, Rnase H, and the gag gene, but no comparisons of their env genes had been done. This was probably because of the rapid evolution and the high variability of this gene. In this study, we attempted to compare the env-like gene of LTR retrotransposons (errantiviruses), with the env gene of retroviruses.

The retroviral env polypeptide is composed by subunit: a transmembrane protein (TM) and a larger peptide (SU) containing the receptor-binding function and antigenic sites able to elicit neutralizing antibodies in the infected host. In the human immunodeficiency virus (HIV), the polyprotein gp160 shares particular characteristic of env proteins: a signal peptide at the N-terminal end, a proteolytic cleavage site leading to glycoproteins gp41 (TM) and gp120 (SU), a transmembrane domain near the C-terminal end, a leucine zipper motif near the N-terminal end of gp41 corresponding to a fusion peptide involved in membrane fusion, and several glycosylated sites. The glycoprotein gp120 contains con-
Table 1
Accession Numbers of the Different Elements Used in the Comparison of the env Proteins

<table>
<thead>
<tr>
<th>Element</th>
<th>Host</th>
<th>Type</th>
<th>Accession No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>tirant</td>
<td>Drosophila melanogaster</td>
<td>LTR retrotransposon</td>
<td>X93507a</td>
</tr>
<tr>
<td>tom</td>
<td>Drosophila ananassae</td>
<td>LTR retrotransposon</td>
<td>Z24451a</td>
</tr>
<tr>
<td>TED</td>
<td>Trichoplusia ni</td>
<td>LTR retrotransposon</td>
<td>C36329b</td>
</tr>
<tr>
<td>297</td>
<td>D. melanogaster</td>
<td>LTR retrotransposon</td>
<td>C24872b</td>
</tr>
<tr>
<td>17.6</td>
<td>D. melanogaster</td>
<td>LTR retrotransposon</td>
<td>P04283c</td>
</tr>
<tr>
<td>gypsy-sub</td>
<td>Drosophila subobscura</td>
<td>LTR retrotransposon</td>
<td>X72370a</td>
</tr>
<tr>
<td>gypsy-vir</td>
<td>Drosophila virilis</td>
<td>LTR retrotransposon</td>
<td>M31325a</td>
</tr>
<tr>
<td>gypsy-mel</td>
<td>D. melanogaster</td>
<td>LTR retrotransposon</td>
<td>M38438a</td>
</tr>
<tr>
<td>yoyo</td>
<td>Ceratitis capitata</td>
<td>LTR retrotransposon</td>
<td>S52576b</td>
</tr>
<tr>
<td>B104</td>
<td>D. melanogaster</td>
<td>LTR retrotransposon</td>
<td>U60529a</td>
</tr>
<tr>
<td>ZAM</td>
<td>D. melanogaster</td>
<td>LTR retrotransposon</td>
<td>Z48503a</td>
</tr>
<tr>
<td>FIV</td>
<td>Feline immunodeficiency virus</td>
<td>Retrovirus</td>
<td>U56928a</td>
</tr>
<tr>
<td>Visna</td>
<td>Ovine/caprine lentivirus</td>
<td>Retrovirus</td>
<td>M60609a</td>
</tr>
<tr>
<td>PLV-14</td>
<td>Puma lentivirus</td>
<td>Retrovirus</td>
<td>U03982a</td>
</tr>
<tr>
<td>HIV1</td>
<td>Human immunodeficiency virus</td>
<td>Retrovirus</td>
<td>AF063223a</td>
</tr>
<tr>
<td>HIV2</td>
<td>Human immunodeficiency virus</td>
<td>Retrovirus</td>
<td>M15390a</td>
</tr>
<tr>
<td>SIVcpg</td>
<td>Simian immunodeficiency virus</td>
<td>Retrovirus</td>
<td>U84097a</td>
</tr>
<tr>
<td>SIVmnd</td>
<td>Simian immunodeficiency virus</td>
<td>Retrovirus</td>
<td>P22380f</td>
</tr>
<tr>
<td>SIVsm</td>
<td>Simian immunodeficiency virus</td>
<td>Retrovirus</td>
<td>M31325a</td>
</tr>
<tr>
<td>AGMtom</td>
<td>African green monkey (SIV→mangabey)</td>
<td>Retrovirus</td>
<td>Q02837b</td>
</tr>
<tr>
<td>AGMub</td>
<td>African green monkey (SIV→mangabey)</td>
<td>Retrovirus</td>
<td>S46352b</td>
</tr>
<tr>
<td>AGMman</td>
<td>African green monkey (SIV→mangabey)</td>
<td>Retrovirus</td>
<td>U589991</td>
</tr>
</tbody>
</table>

^a^ EMBL/GenBank.
^b^ NBRF.
^c^ SwissProt.

served cysteine residues involved in gp41–gp120 interaction (Varma and Brown 1989; Coffin 1990). The evolution rate of the env genes can be very high to escape neutralizing antibody recognition. However, the env genes can be represented as a succession of variable domains flanked by conserved regions. In silico hybridization with low stringency parameters was used to detect these conserved regions. Two motifs were found and used to infer trees.

Materials and Methods

All the sequences used in this work were extracted from EMBL, GenBank, NBRF, and SwissProt, with SRS 5.0 on the INFOBIOGEN site. Their names and accession numbers are given in tables 1–3. Similar motifs (adjacent amino acids) or signatures (nonadjacent amino acids) were detected using BLASTP 2.0, which can cope with gaps (Altschul et al. 1997). The parameters were chosen to obtain as many sequences as possible, even with a low percentage of similarity. For instance, the following settings were used: EXPECT = 1,000 instead of 10, DESCRIPTION = 500 instead of 100, and ALIGNMENT = 500 instead of 100. The alignments between all the sequences were then performed manually using the SeqPup editor, version 0.8c (Gilbert 1998). Using the MEME program (Bailey and Elkan 1994) with all the sequences used in the alignments, the only motifs found between retroviruses and errantiviruses correspond to the two motifs first detected with BLASTP.

To estimate the validity of the alignment, a program in C (available upon request) was used. This program compares the distance distributions between the manually aligned sequence set and a randomly aligned sequence set. Distances were computed using the amino acid classification proposed by Hall (see the PHYLIP package, version 3.5c; Felsenstein 1993). Distributions were then compared using a Kolmogorov and Smirnov test (Sokal and Rolf 1995, p. 887).

In order to assess the amount of the phylogenetic signal in our data, a PTP test (Swofford et al. 1996) was performed. The tests were highly significant ($P < 0.001$ with 1,000 replicates) for all the data sets used. This includes the alignments based upon the env, integrase, and reverse transcriptase genes.

Trees were inferred using PAUP, version 3.1.1 (Swofford 1993), which uses the parsimony method; Puzzle, version 3.1 (Strimmer and von Haeseler 1996), which carries out an analysis using the maximum-likelihood method based on the “quartet puzzling” algorithm; and the neighbor-joining method in the PHYLIP package. Three models of amino acid substitution were used by PHYLIP: the Dayhoff, Schwartz, and Orcutt (1978) PAM matrix, based on the probability of the switch from one amino acid to another, and the George-Hunt-Barker matrix (George, Hunt, and Barker 1988) and the Hall matrix, which are both based on amino acid classification. Since all methods give similar topologies, only the results obtained with PAUP are reported. The heuristic search was used with stepwise addition (random addition and 10 replicates) and tree bisection-reconnection branch-swapping options.

Results

The Motifs

Two conserved motifs were determined among the LTR retrotransposons. The KRG motif was initially detected in Tirant (Drosophila melanogaster), tom (Dro-
sophila ananassae), and TED (Trichoplusia ni). Then, 25 amino acids were aligned around this motif and used in a BLASTP search. This allowed us to detect several errantiviruses and retroviruses, including HIV and SIV (Fig. 1A).

Using the complete sequence of the env gene product of 17.6 (D. melanogaster) in a BLASTP search, a second motif, LTPL, was defined. This motif was found in both LTR retrotransposons and retroviruses. The alignment for 46 amino acids found for seven retrotransposons is shown figure 1B. Each sequence of this alignment was then used in a BLASTP search. Using yoyo (Ceratitis capitata), a puma lentivirus (PLV-14) env protein was detected. Four new sequences were obtained, including FIV, SIV-cpz (chimpanzee), and HIV1, using the complete sequence of this protein.

**Positions of the Motifs**

In this paper, the env gene of HIV is used as a reference for the positions of the various motifs. In HIV, this gene encodes for a polyprotein, which, after proteolysis, yields two proteins (Fig. 2): gp120 (a surface glycoprotein) and gp41 (a transmembrane glycoprotein). The two proteins are involved in the adsorption and pen-
A. Motif KRG

**FIG. 1.** Alignments of LTR retrotransposons and retroviruses for the two motifs. In the trees of figure 3, the gaps were removed.

The two motifs were found in LTR retrotransposons, gypsy-like elements, and retroviruses, with the exception of the visna retroviruses (ovine/caprine lentivirus), which only contain the KRG motif. In HIV and SIV, the KRG motif is localized between the two glycoproteins at position 500 on the polyprotein (fig. 2). This corresponds to the cleavage site. In the LTR retrotransposons, this motif is found around position 100, in the N-terminal part of the protein. This does not correspond to the cleavage site of gypsy.

The position of the LTPL motif is more variable in all sequences. In HIV and SIV, this motif is found at the beginning of glycoprotein gp120, around position 100, in a region known as the “variable region,” or, more precisely, in the V2 loop. In the puma retrovirus (PLV-14) and FIV, the LTPL and KRG regions are close together. In the retrotransposons and gypsy-like elements, the LTPL motif is localized in the C-terminal end of the protein, around position 400.

Other similarities were detected between the env gene of the sequences previously mentioned. These similarities were generally limited to a few retrotransposons, and most of them were localized in HIV gp120. The LTPL and KRG motifs were therefore used both separately and together to infer unrooted trees.

Trees Based on the Motifs

Using PAUP, four trees based on the KRG motif were obtained with a consistency index of 0.825. There is no topological difference between these four trees. Three distinct clusters were observed (fig. 3A): a cluster of retrotransposons including ZAM, Tirant, TED, tom, and 297; a cluster of retroviruses including HIV1, HIV2, and SIVcpz; and a cluster of gypsy-like elements including gypsy-snl, gypsy-sub, and gypsy-vir. The trees of figure 3, the gaps were removed.

B. Motif LTPL

**FIG. 1.** Alignments of LTR retrotransposons and retroviruses for the two motifs. In the trees of figure 3, the gaps were removed.
We found some changes inside the clusters previously defined. Two subgroups of retroelements can be distinguished, one consisting of ZAM, Tirant, and TED, and one consisting of tom, 17.6 and 297 (fig. 3B). Yoyo occupies an intermediate position between these two groups. The last retroelement, gypsy, is associated with feline retroviruses (FIV and PLV-14). Similarly, some subgroups of retroviruses can be observed, such as HIV2, associated with SIVsm, and HIV1, associated with...
green monkey retroviruses (AGMgri, AGMsab, AGMtan) and SIV (SIVmnd and SIVcpz).

Figure 3C shows the tree based on the two motifs. Clear segregation is again observed between the different types of elements. Among the retrotransposons, two groups can be identified: one consisting of tom, 17.6, and 297, and one consisting of ZAM, Tirant, and TED. The yoyo and gypsy elements are grouped together. The feline retroviruses are closely related to the retrotransposons and to the gypsy-like elements, whereas human and monkey retroviruses are more distant.

Comparison with Other Trees

The tree based on the env gene and including both motifs was then compared with the trees obtained from reverse transcriptase and integrase (fig. 4). Whichever domain was used, the retroviruses were separated from the other elements. From the nonretroviruses, the gypsy/
**Env (KRG + LTPL)**

**Reverse transcriptase (region 4)**

**Integrate (DDE signature)**

---

**Fig. 4.**—Trees obtained using the PAUP program. Trees obtained using the neighbor-noining method and those obtained using the the maximum-likelihood method are similar. Numbers given along the branches are the bootstrap values after 100 repetitions. See tables 1–3 for the accession numbers of the sequences. Gypsy is the *D. melanogaster* element.

*Downloaded from https://academic.oup.com/mbe/article-abstract/16/9/1198/2925518 by guest on 28 April 2019*
yoyo and toml/297/17.6 groups were clearly distinguished, whereas the positions of ZAM and TED were not stable. The bootstrap values were relatively weak, but the same classification of the elements was obtained regardless of the method and domain used.

Discussion

In this study, we investigated the relationships between nonretroviral transposable elements and retroviruses by comparing their env proteins. Two motifs were detected. Their locations within the env gene are clearly different. However, it must be stressed that the organization of the insect retroviral element gypsy is similar to that of the LTR retrotransposons. Whatever the methods used, trees obtained from an alignment of these motifs are similar to those deduced from other domains, such as reverse transcriptase and integrase.

It is puzzling that no endogenous retrovirus or avian retrovirus was detected. It is possible that the env gene of endogenous retroviruses is too degenerate to present detectable homology. This could be due to the high mutation rate and reorganization of the env gene. Furthermore, the two motifs were not detected in the envelope gene of the plant retrotransposon SIRE-1 (Laten, Majumdar, and Gaucher 1998). This element, which has been detected in the soybean genome, is a member of the copia/Ty1 superfamily.

The main homologies identified in retroviruses and nonretroviral elements have different protein locations in the env gene. They are frequently found in the variable zones of retroviruses, and especially in the V1-V2 loop. These zones are responsible for target cell recognition and retrovirus tropism (Rizzuto et al. 1998; Wyatt et al. 1998). If these motifs have functional roles, they are probably different in LTR retrotransposons and gypsy-like elements. On the one hand, the consensus KRG motif corresponding to a cleavage site in retroviruses (R/K-X-K/R-R, where X is any amino acid in HIV1; Bosch and Pawlita 1990) is not the cleavage site of gypsy. On the other hand, the LTPL motif is not well conserved in the gypsy-like elements (see, for instance, Alberola and de Frutos 1996), and there is no evidence that their functions are similar to those in retrovirus env genes.

Models of TE evolution suggest that LTR retrotransposons could have given rise to retroviruses (Temin 1980; Flavell 1981). More recently, this was also proposed on the basis of reverse transcriptase analysis carried out by Xiong and Eickbush (1990), by Flavell et al. (1995), and by McClure (1991). In this work, the trees obtained using the two motifs (KRG + LTPL) show that the two groups gypsy/yoyo and toml/297/17.6 can be distinguished from retroviruses. Analysis of integrase and reverse transcriptase leads to a similar conclusion. However, in the case of the latter domain, the gypsy/yoyo group seems to be more closely related to the retroviruses (see fig. 4).

Can the env gene be acquired by retrotransposons and lost by retroviruses? In all trees, gypsy is always close to the LTR retrotransposons, and its status as an insect retrovirus is established (Kim et al. 1994; Song et al. 1994). A possible source of bias is that all of the retrotransposons were from arthropods and all of the retroviruses were from vertebrates. This was mainly due to the small number of known vertebrate LTR retrotransposons and the small number of retroviruses detected in invertebrates or plants. In vertebrates, LTR retrotransposons have been reported only from a small domain of the reverse transcriptase. These sequences and those of gypsy, retroviruses, and other LTR retrotransposons were used to infer a tree (fig. 5). This tree clearly shows that there are three groups of sequences: LTR retrotransposons without any ORF3 including those with an env gene without the two motifs detected here, like SIRE-1, those with a partial or complete env gene containing the two motifs, and the retroviruses. Moreover, the branch between retroviruses and the first group of LTR retrotransposons is supported by a bootstrap value of 100, whereas discrimination between retroviruses and the second group of LTR retrotransposons is less robust. This could be due to the evolutionary proximity of these two groups.

Another interesting feature is that gypsy and yoyo seem to be closely related to the spumavirus, HFV. This characteristic is also found when the gag gene is analyzed (data not shown). For this gene, the major homology region (MHR) is present in all retroviruses and in some retrotransposons but not in the spumavirus or arthropod gypsy-like elements.

All of these findings suggest a direct relationship between nonretroviral elements and retroviruses. Evolution between these two entities probably occurs in both directions. On the one hand, inactivation of the env gene of retroviruses could have given rise to LTR retrotransposons. On the other hand, acquisition of a functional env gene by LTR retrotransposons could have given rise to retroviruses. This possibility was discussed by McClure (1991).

The variability of the motif positions in the env gene can be interpreted in at least two ways: it may result from rearrangements of different parts of the gene or from modular acquisition of the gene. The modification of the gypsy position in the LTPL tree, compared with the KRG tree (fig. 3A and B), could argue in favor of a modular acquisition of the two motifs. However, the alternative hypothesis is that the evolution rate of the LTPL motif in gypsy is different from that of the other errantiviruses.

For the moment, there is no decisive argument for or against rearrangements or modular acquisitions. However, there is a contrast between the apparent similarity of motif positions among the retrotransposons and gypsy elements and their variability in retroviruses. According to the modular-acquisition hypothesis, this could be because all of the retroelements with partial env gene and gypsy-like elements have a common ancestral retrovirus. This assumes that no rearrangement has occurred since the emergence of these elements.

We have not been able to answer the question raised by Flavell (1981) about the evolution of retroviruses from transposable elements. It is interesting to note...
The intermediate position of retroviruses (fig. 5) between the LTR retrotransposons with no env genes and those with complete or partial genes. However, the question will probably be answered only when intermediate forms are identified.

Acknowledgments

We thank L. Bailey for the aligned sequences of the env genes of HIV, African green monkey retroviruses, and SIV. This work was supported by the Programme Génome no. 23.

LITERATURE CITED


