

# Evidence that a Family of Miniature Inverted-Repeat Transposable Elements (MITEs) from the *Arabidopsis thaliana* Genome Has Arisen from a *pogo*-like DNA Transposon

Cédric Feschotte and Claude Mouchès

Laboratoire Ecologie Moléculaire et Faculté Sciences et Techniques Côte-Basque, Université de Pau et des Pays de l'Adour, Pau, France

Sequence similarities exist between terminal inverted repeats (TIRs) of some miniature inverted-repeat transposable element (MITE) families isolated from a wide range of organisms, including plants, insects, and humans, and TIRs of DNA transposons from the *pogo* family. We present here evidence that one of these MITE families, previously described for *Arabidopsis thaliana*, is derived from a larger element encoding a putative transposase. We have named this novel class II transposon *Lem1*. We show that its putative product is related to transposases of the Tc1/*mariner* superfamily, being closer to the *pogo* family. A similar truncated element was found in a tomato DNA sequence, indicating an ancient origin and/or horizontal transfer for this family of elements. These results are reminiscent of those recently reported for the human genome, where other members of the *pogo* family, named *Tiggers*, are believed to be responsible for the generation of abundant MITE-like elements in an early primate ancestor. These results further suggest that some MITE families, which are highly reiterated in plant, insect, and human genomes, could have arisen from a similar mechanism, implicating *pogo*-like elements.

## Introduction

Transposable elements are divided into two major classes according to their mode of transposition (Finnegan 1989). Class I elements (retroelements) transpose by means of an RNA intermediate generated by reverse transcription, while class II elements transpose via a DNA intermediate. Several elements are difficult to classify, mainly because their mechanisms of transposition remain unclear. These include several families of short (100–500-bp) interspersed elements with terminal inverted repeats (TIRs) that have been designated miniature inverted-repeat transposable elements (MITEs). MITEs were first described for grass genomes (Bureau and Wessler 1992, 1994) but have also been found in a wide range of organisms, including fungi (Yeadon and Catchside 1995), mosquitoes (Tu 1997), beetles (Braquart, Royer, and Bouhin 1999), and some vertebrates, like *Xenopus* (Unsal and Morgan 1995), humans (Smit 1996; Smit and Riggs 1996) and teleost fishes (Izsvák et al. 1999). In plants and mosquitoes, MITEs are frequently associated with wild-type genes, indicating a potential role for these elements in gene regulation and genome organization (Wessler, Bureau, and White 1995; Bureau, Ronald, and Wessler 1996; Tu 1997).

To date, no MITEs have been found to encode any product required for their movement, and their transposition mechanism remains unknown. Because they have TIRs and generally generate short specific se-

quence duplication upon insertion, it has been suggested that MITEs could be nonautonomous elements mobilized by transposase activity encoded by class II elements (Bureau and Wessler 1994; Unsal and Morgan 1995; Smit and Riggs 1996). However, MITEs differ from DNA-mediated elements in being present in high copy numbers, which indicates that there may be other processes than the cut-and-paste activity involved in their transposition cycle to explain such a proliferation in genomes.

Since TIR similarities exist between some MITE families and class II transposons, we wondered if either these MITEs are deleted forms of larger “master” elements, encoding a transposase, or homology is restricted to the TIRs because it results from a convergent evolution process due to a common transposition mechanism, using the same type of transposase.

We present here evidence that one of these MITE families, previously described in *Arabidopsis thaliana*, is closely related to a larger element, named *Lem1*, which could encode a putative transposase. As sequence similarity between the MITE and *Lem1* is not restricted to the TIRs, but encompasses the entire MITE consensus sequence, we propose that members of the *Arabidopsis* MITE family are deleted forms of a full-length class II element. We show that *Lem1* potentially encodes a product related to the *pogo* family of transposases. Based on these results, we propose a common model for the origin of some MITE families which are highly reiterated in several distant eukaryote genomes.

## Materials and Methods

Most sequence analysis was done with tools available at the Infobiogen WWW server (<http://www.infobiogen.fr>). Database searches were performed with BLASTN and TBLASTN (Altschul et al. 1990) using default parameters. Multiple-sequence alignments were constructed by CLUSTAL W, version 1.7 (Thompson et al. 1994). Pairwise alignments of amino acid and nucle-

Abbreviations: HTH, helix-turn-helix; MITE, miniature inverted-repeat transposable element; ORF, open reading frame; TIR, terminal inverted repeat.

Key words: miniature inverted-repeat transposable element (MITE), DNA transposon, Tc1/*mariner* superfamily, *pogo*, *Arabidopsis*, evolution.

Address for correspondence and reprints: Claude Mouchès, Laboratoire Ecologie Moléculaire et Faculté Sciences et Techniques Côte-Basque, Université de Pau et des Pays de l'Adour, BP 1155, F-64 013 Pau, France. E-mail: [claudemouches@univ-pau.fr](mailto:claudemouches@univ-pau.fr).

*Mol. Biol. Evol.* 17(5):730–737, 2000

© 2000 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038

Organism	Element	TSD	TIRs	Type	Size, bp	Copy No.
<i>C. pipiens</i>	<b>Mimo</b>	TA	CAGT <b>AGTTGTTCCGCTAA-CTGGGC</b>	MITEs	346	500-1,000
<i>C. pipiens</i>	<b>Nemo1</b>	CA	CAGT <b>GAAA</b> <sub>g</sub> <b>CCCGACCATGGTAA</b> <sub>TT</sub>	MITEs	324	n.d.
<i>A. aegypti</i>	<b>Wujin</b>	TA	CAGT <b>CAAACCTCCATGA-GTTCCA</b>	MITEs	185	2,100
<i>A. thaliana</i>	<b>Emigrant</b>	TA	CAGT <b>AAAACCTCTATAA-ATTAAATA</b>	MITEs	524	500-1,000
<i>H. sapiens</i>	<b>MER(II)</b>	TA	CAGT <b>NGTCCCTCGNTAT-CCGCGG</b>	MITEs	488	>30,000
<i>H. sapiens</i>	<b>Tigger1</b>	TA	CAG <b>GCATACCTCGTTT-ATTGCG</b>	ClassII	2417	3,000
<i>H. sapiens</i>	<b>Tigger2</b>	TA	CAGT <b>GCACCTTCACA-ACaCGG</b>	ClassII	2708	1,000
<i>D. melanogaster</i>	<b>Pogo</b>	TA	CAGT <b>ATAA-TTCCCTA-CTGCATCGA</b>	ClassII	2121	80

FIG. 1.—Homologies in terminal inverted repeats (TIRs) and target site duplications between several miniature inverted-repeat transposable element (MITE) families and transposons of the *pogo* family. Alignment of the TIR sequences was done by eye. Conserved (at least 4/8) bases are in white type on a black background. A gap was introduced to maintain the best alignment. TIRs, target site duplication (TSD) sequences, and lengths of the *Mimo*, *Wujin*, and *Emigrant* families were deduced from consensus sequences calculated on alignment of several MITE copies (Tu 1997; Casacuberta 1998; unpublished data). The *Nemo1* element was found nested in a *Mimo* copy (unpublished data); it possesses 24-bp imperfect TIRs (two mismatches, as indicated by lowercase type). The 5' TIR of *Nemo1* is aligned here. This is the only sequenced copy of a family of repetitive sequences to be described elsewhere. MER(II) represents a general consensus for the second group of human MERs based on a simple majority rule in alignment of consensus TIR sequences of MER28, MER8, MER2, MER44, MER46, MER6, and MER7 (Smit and Riggs 1996). "N" indicates a highly variable nucleotide in the alignment. Other information on human transposons are from Smit and Riggs (1996). Data on *pogo* elements are from Tudor et al. (1992).

otide sequences were done with the ALIGN program (Myers and Miller 1988) of the FASTA package. Potential initiation ATG codons were identified by using the NetSart 1.0 program (Pedersen and Nielsen 1997) with settings for *A. thaliana*, and consensus splice sites for *A. thaliana* were predicted by the NetGene2 program (Hebsgaard et al. 1996). Both programs are available at the Center for Biological Sequence Analysis server (<http://www.cbs.dtu.dk>). Predictions for helix-turn-helix motifs were done with the HTH motif prediction method (Dodd and Egan 1990), available through the Network Protein Sequence Analysis at <http://pbil.ibcp.fr/cgi-bin/npsa>.

## Results

### Homologies in TIRs Between Several MITE Families and Class II Transposons

We recently found members of several novel families of MITEs in the genome of *Culex pipiens* mosquitoes (unpublished data). Families have no significant sequence identity to each other or to any other known transposable elements. However, one of these families, named *Mimo*, and an additional MITE-like element, *Nemo1*, possess, respectively, 23- and 25-bp TIRs that show some similarities (fig. 1) to *Wujin*, a MITE family described from the yellow fever mosquito, *Aedes aegypti* (Tu 1997), and with a MITE family from the plant *A. thaliana* described as the *Emigrant* family (Casacuberta et al. 1998) or as *MathE2* elements (Surzycki and Belknap 1999), successively.

TIR similarity (17/23 nucleotides) between *Wujin* and *Emigrant* elements was previously noticed (Casacuberta et al. 1998), and it was suggested that these elements might belong to the same MITE family. Therefore, *Mimo* and *Nemo* elements from the *C. pipiens* genome could be considered members of this same MITE family. What strikes us more significantly is the fact that all of these MITEs have TIRs that begin with CAGT (or CACT), like TIRs of several class II transposons of the Tc1/*mariner* superfamily. Moreover, like most Tc1/*mariner* elements, these MITEs are generally flanked by

the TA dinucleotide, probably resulting from a target site duplication upon integration of the element (van Luenen, Colloms, and Plasterk 1994; Hartl, Lohe, and Lozovskaya 1997; Plasterk, Izsvák, and Ivics 1999). The highest TIR sequence similarities (fig. 1) were found with the *Drosophila pogo* (Tudor et al. 1992) and the human *Tigger* elements (Robertson 1996; Smit and Riggs 1996). We therefore hypothesize that genomes of *C. pipiens*, *A. aegypti*, or *A. thaliana* could also contain ancestral *pogo*-like elements.

### Identification of a *pogo*-like Element Closely Related to a MITE Family

In order to identify a potential source of transposase responsible for the spread of MITEs found in mosquitoes and *Arabidopsis*, we used both *pogo* and *Tigger1* putative products as queries in TBLASTN searches (Altschul et al. 1990) against current DNA databases. No matching mosquito sequences were identified, but significant sequence similarities ( $P_{\text{TBLASTN}} = 2e^{-44}$  with *pogo*,  $P_{\text{TBLASTN}} = 3e^{-24}$  with *Tigger1*) were found within a region of a BAC clone from *A. thaliana* chromosome II (GenBank accession number AC006161). This region (851 bp, from position 85898 to position 86749 in the GenBank sequence) coincided with a predicted open reading frame (ORF) coding for a putative DNA-binding protein, reinforcing the idea that it could correspond to a transposase gene.

To our surprise, BLAST searches (Altschul et al. 1990) in databases using the *Arabidopsis* DNA surrounding this ORF revealed that the putative coding region is flanked by sequences highly similar to members of the *Emigrant/MathE2* MITE family described from *A. thaliana* (Casacuberta et al. 1998; Surzycki and Belknap 1999). As shown in figure 2, the BAC clone in fact contains an entire *Emigrant* element (72.2% identity with a consensus nucleotide sequence for the *Emigrant* family) with a greatly enlarged central region. The overall size of this novel copy would then be 2,114 bp. Such a size is not expected for a so-called 'miniature' element, so we wondered if the 2,114-bp element could,

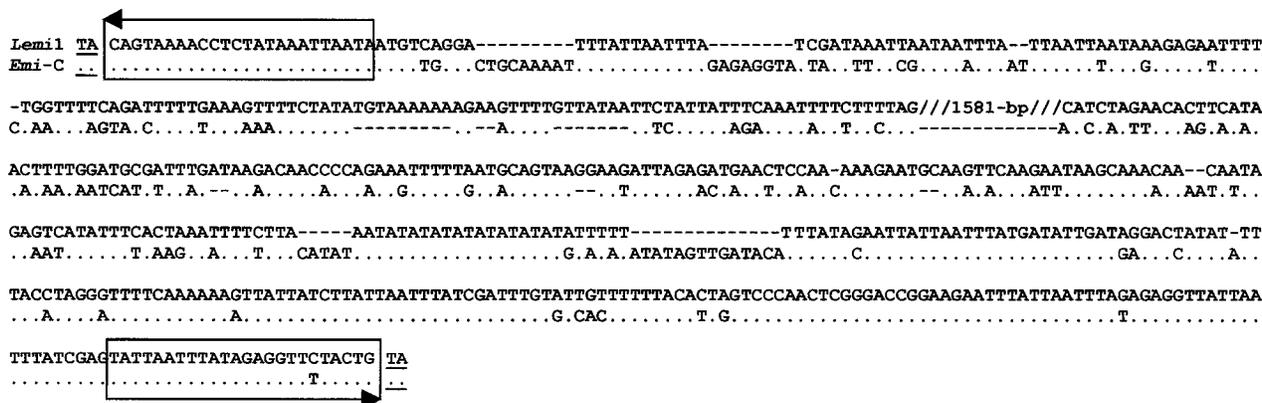


FIG. 2.—Sequence alignment between a consensus for the *Emigrant* miniature inverted-repeat transposable element (MITE) family and *Lem1*. The *Emigrant* consensus (*Emi-C*) was based on sequence alignment of 11 complete *Emigrant* elements (Casacuberta et al. 1998). Dots denote identity, and dashed lines indicate gaps. Terminal inverted repeats are boxed, and TA duplications are underlined.

rather, be a composite one, resulting from a secondary insertion in an *Emigrant* element. There are no sequence features (like TIRs or target site duplications) that further support this hypothesis. In addition, TIR similarities between *Emigrant* and *pogo*-like transposons (fig. 1) suggest, rather, that this element is an *Emigrant* copy with coding capacity. To distinguish this longer element from shorter copies (i.e., MITEs), we named it *Lem1* (larger *Emigrant*). As sequence homology between *Lem1* and *Emigrant* MITEs is not restricted to the TIRs, but encompasses all of the consensus *Emigrant* sequence (fig. 2), we conclude that these MITEs are derived from the larger element *Lem1*.

The TIRs of *Lem1* are the same as the 24 bp defined for *Emigrant* by previous work (Casacuberta et al. 1998; Surzycki and Belknap 1999), except for one mismatch in the 3' TIR (fig. 2). Like most *Emigrant* elements, *Lem1* is flanked by a TA dinucleotide, indicating a putative TA target site duplication, a hallmark of the *pogo*/Tc1/*mariner* group (Tudor et al. 1992; Doak et al. 1994; van Luenen, Colloms, and Plasterk 1994; Smit and Riggs 1996). Given TIR and target site similarities, as well as a coding capacity for a product closely related to *pogo* and *Tigger* transposases (see below), we therefore propose that *Lem1* is a novel member of the Tc1/*mariner* superfamily of transposable elements, being closer to the *pogo* family.

By using the DNA sequence of *Lem1* as a query in BLAST searches, we were able to identify an additional truncated *pogo*-like element in a *Lycopersicon esculentum* (tomato) DNA sequence (GenBank accession number Z12833;  $P_{\text{BLAST}} = 2 e^{-12}$ ). Because of a severe truncation at the 3' end, this element, named *Lem2*, is only 1,008 bp long. The 5' end of *Lem2* is defined by a putative TIR which shares high homology with those of *Lem1* (20/24 bp) and is flanked by a TA dinucleotide, reminiscent of the target site duplication. Despite relatively good conservation at the nucleotide level between *Lem2* and *Lem1* (68.3%), it is very difficult to align *Lem2* truncated product with transposases of other *pogo*-like elements, because several frameshifts are needed to maintain a significant amino acid alignment (data not shown). BLAST searches using *Lem2* se-

quence as a query did not reveal any other member of this family in the tomato sequences available in databases. It is likely that *Lem2* is a "molecular fossil" of an ancestral *pogo*-like element of the Solanaceae genome. Interestingly, *Lem2* is inserted in the 5' regulatory region of the polyphenol oxidase A gene (Newmann et al. 1993), suggesting that its sequence could now play a role in gene regulation, as was strongly indicated for some MITEs associated with several plant genes (Bureau and Wessler 1994; Wessler, Bureau, and White 1995; Bureau, Ronald, and Wessler 1996) and for other repetitive sequences inserted in, or close to, many eukaryote genes (McDonald 1995; Britten 1996; Kidwell and Lisch 1997).

#### Coding Capacity of *Lem1*

We carefully analyzed the DNA sequence of *Lem1* for protein coding regions. Two main ORFs were clearly identified (fig. 3). ORF1 (from position 85583 to position 86749 in GenBank AC006161) coincides with a predicted gene encoding a putative DNA-binding protein. The ATG initiation codon for this gene was initially predicted at position 85898, but, as suggested by amino acid alignment to other *pogo*-like transposases (not shown), the start codon would, rather, be upstream. Furthermore, the ATG at position 85898 did not fit with the consensus proposed for translation start sites in *A. thaliana* (Pedersen and Nielsen 1997). Based on the NetStart 1.0 prediction server for translation start sites in this plant (Pedersen and Nielsen 1997) and on amino acid alignments with other transposases, we propose that the initiation codon is in the same reading frame but, rather, at position 85595 or 85657 (ORF0). If the initiation codon is indeed one of these two ATGs, then it is obvious that *Lem1* has suffered at least one single mutation, since a stop codon occurs in this frame at position 85714. This suggests that *Lem1* is probably an ancient copy that may no longer be active. For these reasons, it is not possible to conclusively determine the length of ORF1. Nevertheless, as mentioned above and analyzed further in detail, the putative product encoded by ORF1 displays some striking similarity to several DNA-bind-

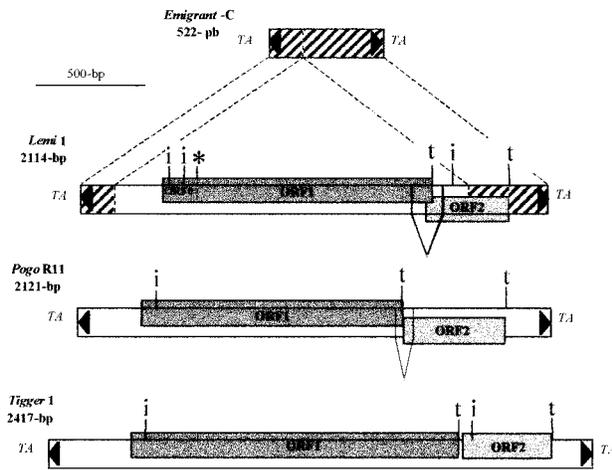


FIG. 3.—Genetic organization of *Emigrant*, *Lem1*, *Pogo*, and *Tigger1* elements. Terminal inverted repeats (black triangles) and TA target site duplications are indicated. Genetic maps for *Pogo*R11 (the only full-length *pogo* copy sequenced) and *Tigger1*, respectively, are drawn from their description in Tudor et al. (1992) (GenBank accession number X59837) and Robertson (1996) (GenBank accession number U49973). Full-length elements possess two large open reading frames (ORFs) (numbered 1 and 2 and boxed in gray) which are in the same frame (*Tigger1*) or in different frames (*Lem1* and *Pogo*R11). In *Lem1*, ORF0 and ORF1 could be translated continuously if a stop codon (represented with an asterisk) is bypassed. Stop codons defining the end of an ORF are represented by the letter “t.” Putative initiation start sites are indicated by the letter “i.” As shown by the lines joined in an open triangle, ORF1 and ORF2 of *Lem1* and *Pogo*R11 can be joined by splicing of a short intron encompassing the ORF1 stop codon. No consensus splice sites have been identified in *Tigger1* (Robertson 1996).

ing proteins, including transposases of the Tc1/*mariner* superfamily of elements.

ORF2 (positions 86750–87101) contains an ATG at position 86784, which fits well with the consensus for the initiation codon in *A. thaliana* (Pedersen and Nielsen 1997), so it could potentially encode a 106-amino-acid polypeptide. No significant similarity was found between this predicted product and any sequences available in databases.

As shown in figure 3, the genetic organization of *Lem1* is very close to those of other members of the *pogo* family. Analysis of *Drosophila melanogaster pogo* cDNAs (Tudor et al. 1992) indicated that full-length *pogo* elements possess two ORFs that can be joined by RNA splicing to encode a single protein of 499 amino acids, namely, the *pogo* transposase. The human *Tigger1* element also has two long ORFs, encoding 454 and 138 residues, respectively, but there is no evidence that the two ORFs are joined by splicing, like those of the *Drosophila pogo* element (Robertson 1996). A search for *A. thaliana* consensus splice sites (NetGene 2; Hebsgaard et al. 1996) in the DNA sequence of *Lem1* revealed the presence of highly significant donor (between positions 86636 and 86637) and acceptor (between positions 86769 and 86770) sites, as well as a branch point between these two splice sites (positions 86739–86754). This suggests that ORF1 and ORF2 of *Lem1* could be joined by splicing of a 133-bp intron in a manner similar to those of the *Drosophila pogo* element (fig. 3). As-

suming that the initiation start point is at position 88595 and that the stop codon between ORF0 and ORF1 is due to a recent mutation or is bypassed, the mature *Lem1* mRNA would then encode a 457-amino-acid peptide which is in the range of other products encoded by *pogo*-like elements.

#### *Lem1* Encodes a Putative Product Closely Related to Transposases of the Tc1/*mariner* Superfamily of Transposable Elements

The central D,D35E region of Tc1/*mariner* transposases is supposed to contain the catalytic domain for transposition (Doak et al. 1994; Capy et al. 1996; Pasterk, Izsvák, and Ivics 1999). Based on amino acid similarities in this region (approximately 160 residues), *pogo* and *Tigger* were recognized as members of the Tc1/*mariner* superfamily (Robertson 1996; Smit and Riggs 1996; Capy et al. 1998), being closer to fungal transposons of the Fot1 group (Daboussi, Langin, and Brygoo 1992) and to Tc2, Tc4, and Tc5 from the nematode *C. elegans* (Yuan et al. 1991; Ruvolo, Hill, and Levitt 1992; Collins and Anderson 1994). We aligned the central region of *Lem1* putative product with the D,D35E domains of several transposases from *pogo* family members (fig. 4). According to this alignment, *Lem1* is closer to *pogo* (41% identity, 75% similarity) and to *Tigger1* and *Tigger2* (32% identity), i.e., with scores in the range of those shown between *pogo* and *Tigger* (41% identity between *pogo* and *Tigger1*). In addition, *Lem1* putative product possesses a D,D32D signature, rather than the D,D35E signature, and thus resembles those of *pogo* (D,D30D) and *Tigger* (D,D33D) transposases. Therefore, we conclude that *Lem1*, *pogo*, and *Tigger* elements are monophyletic.

It was previously predicted that *pogo* and *Tigger* putative transposases bind DNA by a helix-turn-helix (HTH) DNA-binding motif identified in their N-terminal domains (Pietrovski and Henikoff 1997; Wang, Hartswood, and Finnegan 1999). The presence of a putative HTH motif in the N-terminal region of *Lem1* was also indicated by the Dodd and Egan (1990) method, despite low statistical significance (data not shown). Nevertheless, it is an additional indication that *Lem1* could encode a *pogo*-like transposase.

#### Discussion

##### *Lem1* Is a *pogo*-like Element from a Plant Genome that Gave Rise to the *Emigrant* Family of MITEs

In the present work, we present evidence that a family of MITEs from the *A. thaliana* genome, *Emigrant*, derives from a larger element, *Lem1*, which has coding capacity for a putative transposase. We show that *Lem1* belongs to the Tc1/*mariner* superfamily of transposable elements, being closer to the *Drosophila pogo* and the human *Tigger* elements.

To our knowledge, *Lem1* is the the first *pogo* family member to be described in a plant genome and the second that belongs to the Tc1/*mariner* superfamily. A *mariner*-like element, *Soymar1*, has been recently described in soybeans (Jarvik and Lark 1998) but does not

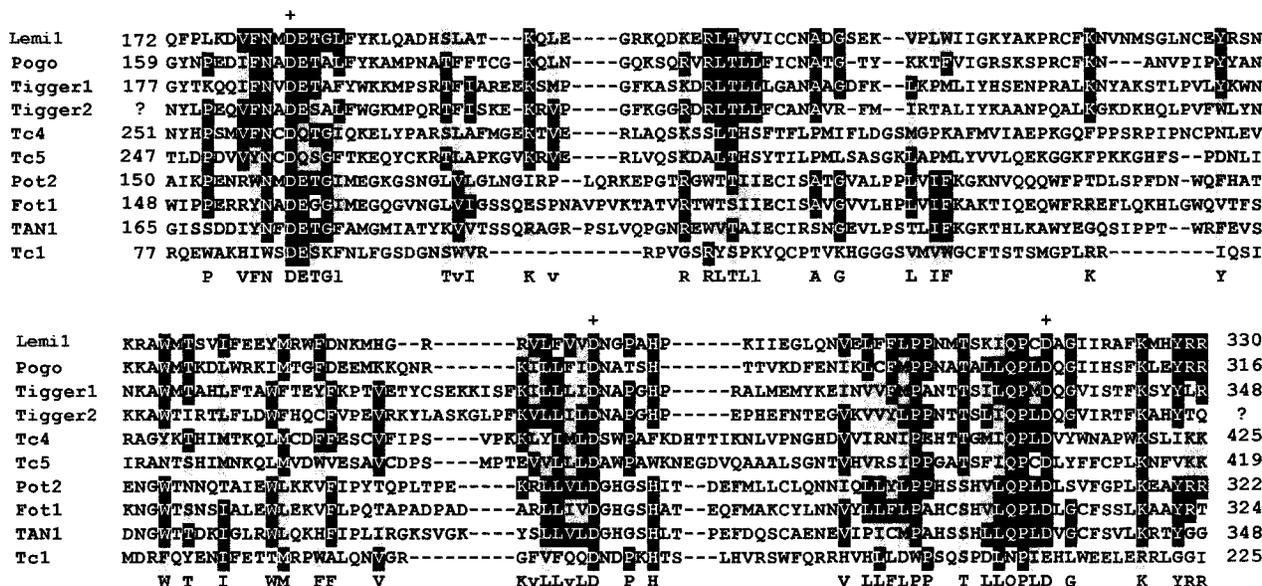


FIG. 4.—Amino acid alignment of the central region of the putative product of *Lem1* with the several conserved D,D35E catalytic domains of Tc1/mariner transposases. This alignment is based on those previously reported by Doak et al. (1994), Smit and Riggs (1996), and Robertson (1996). Alignment was done with CLUSTAL W (Thompson et al. 1994) using default parameters. Amino acid sequences are from *Drosophila melanogaster pogo* (GenBank accession number X59837), *Homo sapiens Tigger1* (U49973) and *Tigger2* (S72489), *Caenorhabditis elegans Tc4* (L00665), Tc5 (Z35400) and the distantly related Tc1 (X01005), and also members of the fungal Fot1 group: *Magnaporthe grisea* Pot2 (Z33638), *Fusarium oxysporum* Fot1 (X70186), and the *Aspergillus awamori* TAN1 element (U58946). Each sequence segment is flanked by coordinates of its first and last residues, except *Tigger2*, for which the ends are not known. Conserved residues in at least 6 of the 10 proteins are marked in white type on a black background for the prominent residue or in gray for other evolutionarily related residues. Dashes indicate gaps introduced for the alignment. Letters below the alignment indicate consensus residues (letters are lowercase when we cannot assigned a leader). Residues of the DDD (or DDE) motifs are indicated by crosses.

display significant sequence similarity with *Lem1*. A remnant of another ancient *Lem1* element with partial coding capacity is present in the tomato genome, indicating an ancient origin and/or horizontal transfer for this family of elements. It also suggests that *pogo*-like elements, albeit rare as full-length copies, might be widespread in eukaryotes.

Is *Lem1* Responsible for the Mobility of Emigrant MITEs?

There are at least 250 *Lem1*-derived MITEs (i.e., *Emigrant* copies) in the available nonredundant *A. thaliana* database (AtDB at <http://genome-www.stanford.edu>). These derivatives are remarkably homogeneous both in size (ranging from 400 to 600 bp) and in sequence, which fits well with the consensus established previously with only 11 *Emigrant* copies (Casacuberta et al. 1998). Strikingly, *Lem1* is the only longer element with coding capacity in the current *Arabidopsis* database which contains, to date, 80% of the total genome. Analysis for coding capacity of *Lem1* suggests that this copy might be no longer active; also, it remains uncertain that *Lem1* was responsible for the recent mobility of *Emigrant* in this plant, as revealed by insertion polymorphisms among *Arabidopsis* ecotypes (Casacuberta et al. 1998). We cannot exclude the possibility that there is, elsewhere in the *Arabidopsis* genome, a functional *Lem1* copy that could provide a source of transposase for *Emigrant* elements. Hybridizations of *Arabidopsis* DNA with a large internal coding fragment of *Lem1* are

needed to assess this possibility. In any case, this will be clarified as soon as the entire *Arabidopsis* sequence is available.

There Is a Strong Tendency for *pogo*-like Elements to Give Rise to MITEs

*Emigrant* length homogeneity is in contrast to what is generally reported for nonautonomous elements that derived from full-length class II elements. Most of the time, they have suffered multiple and variable deletion events, leading to length heterogeneity among members of the same family (O'Hare and Rubin 1983; Streck, MacGaffey, and Beckendorf 1986; Feodoroff 1989; Hartl, Lohe, and Lovoskaya 1997). As it seems very unlikely that the same independent deletion event occurred in all *Lem1* elements, we think that the *Emigrant* family of MITEs could have arisen from a subsequent amplification process of a very small number of defective elements. Interestingly, similar processes seem to have occurred in the human genome, in which accumulation of a large number (>100,000) of short inverted-repeat elements (MERs) is attributed to other *pogo*-related elements, the *Tigger* transposons (Smit and Riggs 1996). Similarly, the *D. melanogaster* genome contains many copies of a 190-bp *pogo* internal deletion product but only a few copies of full-sized *pogo* elements (Tudor et al. 1992; Boussy et al. 1993). Since TIR similarities exist between *Emigrant* MITEs from *Arabidopsis* and several MITE families from mosquito genomes (see fig. 1), it is possible that a similar mecha-

nism for the generation of MITEs could also have occurred in these insects. In this case, *pogo*-like elements may have resided, at least at an ancient time, in their genomes. We must now investigate the presence of such elements in mosquito genomes before extending the results reported here for *Arabidopsis* to these insect MITE families. The question of how general the relationship is between MITEs and DNA transposons is a very interesting one. For many MITE families described to date, there is no indication (like TIR similarities) for a filiation to class II transposon families; also, we assume that it is premature to generalize the DNA transposon origin for all MITEs. However, it seems that there is a strong tendency for *pogo*-like elements to give rise to MITEs in several distant eukaryote genomes, i.e., plants, humans, and insects.

### There May Be Some Features in the Transposition Cycle of *pogo*-like Elements that Enhance the Generation of MITE Derivatives

Because the cut-and-paste mechanism of DNA transposition is basically a nonreplicative process, class II elements generally do not reach high copy numbers. So, it is likely that there are some peculiar mechanisms in the transposition cycle of *pogo*-like elements that greatly enhance the generation of a large number of deletion-derived products. Like other transposases of the Tc1/*mariner* superfamily, products encoded by *pogo*-like transposons are organized in several functional domains. These include an N-terminal region with an HTH DNA-binding motif (Petrokovski and Henikoff 1997; Wang, Hartswood, and Finnegan 1999) and a central domain with a DDD motif that is supposed to be equivalent to the catalytic DDE motif of several recombinases (Plasterk, Izsvák, and Ivics 1999). *pogo*-like transposases are distinguished from other transposases by an unusually long C-terminal domain rich in acidic residues (Tudor et al. 1992; Smit and Riggs 1996). This feature is also found in *Lem1*, in which 21 of the last 100 residues are acidic. Interestingly, this feature is also shared by several human and yeast centromeric proteins of the CENP-B group that also possess sequence similarity in both N-terminal and central regions with *pogo*-like transposases, including *Lem1* (Tudor et al. 1992; Smit and Riggs 1996; Lee, Huberman, and Hurwitz 1997; data not shown). It is hypothesized that *pogo*-like transposases and these centromeric proteins could have a common evolutionary origin (Smit and Riggs 1996). Alternatively, it may also result from a convergent evolution process due to constraints imposed by a similar mechanism for binding DNA and by interactions with other common peptides. In the CENP-B family of proteins, the C-terminal acidic domain might be required for protein-protein interaction (Sugimoto, Hagishita, and Himeno 1994; Lee, Huberman, and Hurwitz 1997). This raises many issues concerning the possible involvement of this domain in the transposition of *pogo*-like elements and/or in the generation of *pogo*-derivatives.

It was shown recently that *pogo* and *Tigger* transposases interact with proliferating cell nuclear antigen

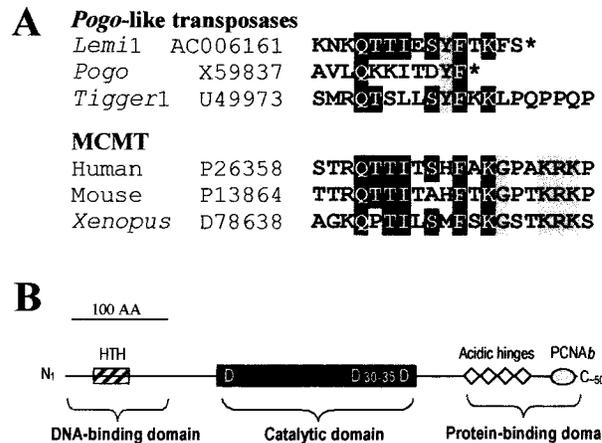


FIG. 5.—A, Putative PCNA-binding motif in *Lem1*. The C-terminus of *Lem1* was aligned by eye to proliferating cell nuclear antigen (PCNA)-binding motifs previously defined for *pogo* and *Tigger1* transposases (Warbrick et al. 1998). PCNA-binding domains for the MCMT (DNA methyltransferase) family of proteins are also aligned, because they share striking similarity with the putative PCNA-binding domain of *Lem1*. Residues conserved in four of the six sequences are highlighted in black. Identical but specific residues for each group of protein are shaded in gray. Asterisks denote the protein termination codon. The accession number for each sequence is given. B, Common putative functional domains of transposases potentially encoded by *pogo*, *Tigger1*, and *Lem1*.

(PCNA) by their C-terminus (Warbrick et al. 1998). PCNA plays an essential role in replication and repair of DNA by interacting with proteins involved in both processes (Kelman and Hurwitz 1998). We show (fig. 5A) that residues previously defined as consensus for PCNA-binding (Warbrick et al. 1998) are conserved in the C-terminal end of *Lem1*, despite a low amino acid conservation in this region. This feature, as well as the presence of numerous acidic residues, suggests that the C-terminal region of *pogo*-like transposases may play an important role in the transposition process of these elements, perhaps by binding to some proteins involved in DNA replication and repair (fig. 5B). This therefore raises the interesting hypothesis that there might be a close link between the transposition cycle of *pogo*-like elements, replicating DNA, and the proliferation of some MITE families in plant, insect, and human genomes.

### Acknowledgments

We thank F. Brunet, C. Cagnon, R. Duran, Y. Gilbert, S. Karama, B. Lauga, C. MacMahon, P. Mourguiart, N. Pourtau, and J.-C. Salvado for helpful discussions and valuable advice. We are also grateful to J. M. Casacuberta for providing the consensus sequence for *Emigrant* elements. C.F. was supported by a grant from the Ministère de l'Éducation Nationale, de la Recherche et de la Technologie to University of Paris 6.

### LITERATURE CITED

- ALTSCHUL, S. F., W. GISH, W. MILLER, E. W. MYERS, and D. J. LIPMAN. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215**:403–410.

- BOUSSY, I. A., L. CHARLES, M. H. HAMELIN, G. PERIQUET, and D. Y. SHAPIRO. 1993. The occurrence of the transposable element *pogo* in *Drosophila melanogaster*. *Genetica* **88**:1–10.
- BRAQUART, C., V. ROYER, and H. BOUHIN. 1999. DEC: a new miniature inverted-repeat transposable element from the genome of the beetle *Tenebrio molitor*. *Insect Mol. Biol.* **8**: 571–574.
- BRITTEN, R. J. 1996. DNA sequence insertion and evolutionary variation in gene regulation. *Proc. Natl. Acad. Sci. USA* **93**: 9374–9377.
- BUREAU, T. E., P. C. RONALD, and S. R. WESSLER. 1996. A computer-based systematic survey reveals the predominance of small inverted-repeat elements in wild-type rice genes. *Proc. Natl. Acad. Sci. USA* **93**:8524–8529.
- BUREAU, T. E., and S. R. WESSLER. 1992. *Tourist*: a large family of inverted-repeat element frequently associated with maize genes. *Plant Cell* **4**:1283–1294.
- . 1994. *Stowaway*: a new family of inverted-repeat elements associated with genes of both monocotyledonous and dicotyledonous plants. *Plant Cell* **6**:907–916.
- CAPY, P., C. BAZIN, D. HIGUET, and T. LANGIN, eds. 1998. Dynamics and evolution of transposable elements. P. 24 in *Molecular biology intelligence unit*, Springer-Verlag, Austin, Tex.
- CAPY, P., R. VITALIS, T. LANGIN, D. HIGUET, and C. BAZIN. 1996. Relationships between transposable elements based upon the integrase-transposase domains: is there a common ancestor? *J. Mol. Evol.* **42**:359–368.
- CASACUBERTA, E., J. M. CASACUBERTA, P. PUIGDOMENECH, and A. MONFORT. 1998. Presence of miniature inverted-repeat transposable elements (MITEs) in the genome of *Arabidopsis thaliana*: characterisation of the *Emigrant* family of elements. *Plant J.* **16**:79–85.
- COLLINS, J. J., and P. ANDERSON. 1994. The Tc5 family of transposable elements in *Caenorhabditis elegans*. *Genetics* **137**:771–781.
- DABOUSSI, M.-J., T. LANGIN, and Y. BRYGOO. 1992. Fot1, a new family of fungal transposable elements. *Mol. Gen. Genet.* **232**:12–16.
- DOAK, T. G., F. P. DOERDER, C. L. JAHN, and G. HERRICK. 1994. A proposed superfamily of transposase genes: transposon-like elements in ciliated protozoa and a common “D35E” motif. *Proc. Natl. Acad. Sci. USA* **91**:942–946.
- DODD, I. B., and J. B. EGAN. 1990. Improved detection of helix-turn-helix DNA-binding motifs in protein sequences. *Nucleic Acids Res.* **18**:5019–5026.
- FEODOROFF, N. 1989. Maize transposable elements. Pp. 375–411 in D. E. BERG and M. M. HOWE, eds. *Mobile DNA*. American Society for Microbiology, Washington, D.C.
- FINNEGAN, D. J. 1989. Eukaryotic transposable elements and genome evolution. *Trends Genet.* **5**:103–107.
- HARTL, D. L., A. R. LOHE, and E. R. LOZOVSKAYA. 1997. Modern thoughts on an ancient *marinere*: function, evolution, regulation. *Annu. Rev. Genet.* **31**:337–358.
- HEBSGAARD, S. M., P. G. KORNING, N. TOLSTRUP, J. ENGELBRECHT, P. ROUZE, and S. BRUNAK. 1996. Splice site prediction in *Arabidopsis thaliana* DNA by combining local and global sequence information. *Nucleic Acids Res.* **24**: 3439–3452.
- IZSVÁK, Z., Z. IVICS, N. SHIMODA, D. MOHN, H. OKAMOTO, and P. B. HACKETT. 1999. Short inverted-repeat transposable elements in teleost fish and implications for a mechanism of their amplification. *J. Mol. Evol.* **48**:13–21.
- JARVIK, T., and K. G. LARK. 1998. Characterization of *Soy-mar1*, a mariner element in soybean. *Genetics* **149**:1569–1574.
- KELMAN, Z., and J. HURWITZ. 1998. Protein-PCNA interaction: a DNA-scanning mechanism? *Trends. Biochem. Sci.* **23**: 236–238.
- KIDWELL, M. G., and D. LISCH. 1997. Transposable elements as sources of variation in animals and plants. *Proc. Natl. Acad. Sci. USA* **94**:7704–7711.
- LEE, J.-K., J. A. HUBERMAN, and J. HURWITZ. 1997. Purification and characterization of a CENP-B homologue protein that binds to the centromeric K-type repeat DNA of *Schizosaccharomyces pombe*. *Proc. Natl. Acad. Sci. USA* **94**: 8427–8432.
- MCDONALD, J. F. 1995. Transposable elements: possible catalysts of organismic evolution. *Trends. Ecol. Evol.* **10**:123–126.
- MYERS, E. W., and W. MILLER. 1988. Optimal alignments in linear space. *Comput. Appl. Biosci.* **4**:11–17.
- NEWMANN, S. M., N. T. EANNETTA, H. YU, J. P. PRINCE, M. C. DE VICENTE, S. D. TANKSLEY, and J. C. STEFFENS. 1993. Organisation of the tomato phenol oxidase gene family. *Plant Mol. Biol.* **21**:1035–1051.
- O’HARE, K., and G. M. RUBIN. 1983. Structures of P transposable elements and their sites of insertions in the *Drosophila melanogaster* genome. *Cell* **34**:25–35.
- PEDERSEN, A. C., and H. NIELSEN. 1997. Neural network prediction of translation initiation sites in eukaryotes: perspectives for EST and genome analysis. *ISMB* **5**:226–233.
- PIETROKOVSKI, S., and S. HENIKOFF. 1997. A helix-turn-helix DNA-binding motif predicted for transposases of DNA transposons. *Mol. Gen. Genet.* **254**:689–695.
- PLASTERK, R. H. A., Z. IZSVÁK, and Z. IVICS. 1999. Resident aliens: the Tc1/*mariner* superfamily of transposable elements. *Trends Genet.* **15**:326–332.
- ROBERTSON, H. M. 1996. Members of the *pogo* superfamily of DNA-mediated transposons in the human genome. *Mol. Gen. Genet.* **252**:761–766.
- RUVOLO, V., J. E. HILL, and A. LEVITT. 1992. The Tc2 transposon of *Caenorhabditis elegans* has the structure of a self-regulated element. *DNA Cell Biol.* **11**:111–122.
- SMIT, A. F. A. 1996. The origin of interspersed repeats in the human genome. *Curr. Opin. Genet. Dev.* **6**:743–748.
- SMIT, A. F. A., and A. D. RIGGS. 1996. *Tiggers* and DNA transposon fossils in the human genome. *Proc. Natl. Acad. Sci. USA* **93**:1443–1448.
- STRECK, R. D., J. E. MACGAFFEY, and S. K. BECKENDORF. 1986. The structure of hobo transposable elements and their insertion sites. *EMBO J.* **5**:3615–3623.
- SUGIMOTO, K., Y. HAGISHITA, and M. HIMENO. 1994. Functional domain structure of human centromere protein B. *J. Biol. Chem.* **269**:24271–24276.
- SURZYCKI, S. A., and W. R. BELKNAP. 1999. Characterization of repetitive DNA elements in *Arabidopsis*. *J. Mol. Evol.* **48**:684–691.
- THOMPSON, J. D., D. DESMOND, D. G. HIGGINS, and T. J. GIBSON. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**:4673–4680.
- TU, Z. 1997. Three novel families of miniature inverted-repeat transposable elements are associated with genes of the yellow fever mosquito, *Aedes aegypti*. *Proc. Natl. Acad. Sci. USA* **94**:7475–7480.
- TUDOR, M., M. LOBOCKA, M. GOODWELL, J. PETTITT, and K. O’HARE. 1992. The *pogo* transposable element family of *Drosophila melanogaster*. *Mol. Gen. Genet.* **232**:126–134.
- UNSAI, K., and G. T. MORGAN. 1995. A novel group of families of short interspersed repetitive elements (SINES) in *Xenopus*: evidence of a specific target site for DNA-mediated

- ated transposition of inverted-repeat SINEs. *J. Mol. Biol.* **248**:812–823.
- VAN LUENEN, H. G. A. M., S. D. COLLOMS, and R. H. A. PLASTERK. 1994. The mechanism of transposition of Tc3 in *C. elegans*. *Cell* **79**:293–301.
- WANG, H., E. HARTSWOOD, and D. J. FINNEGAN. 1999. *Pogo* transposase contains a putative helix-turn-helix DNA binding domain that recognises a 12 bp sequence within the terminal inverted repeats. *Nucleic Acids Res.* **27**:455–461.
- WARBRICK, E., W. HEATHERINGTON, D. P. LANE, and D. M. GLOVER. 1998. PCNA binding proteins in *Drosophila melanogaster*: the analysis of a conserved PCNA binding domain. *Nucleic Acids Res.* **26**:3925–3932.
- WESSLER, S. R., T. E. BUREAU, and S. E. WHITE. 1995. LTR-retrotransposons and MITEs: important players in the evolution of plant genomes. *Curr. Opin. Genet. Dev.* **5**:814–821.
- YEADON, P. J., and D. E. CATCHESIDE. 1995. *Guest*: a 98 bp inverted repeat transposable element in *Neurospora crassa*. *Mol. Gen. Genet.* **247**:105–109.
- YUAN, J., M. FINNEY, N. TSUNG, and H. R. HORVITZ. 1991. Tc4, a *Caenorhabditis elegans* transposable element with an unusual fold-back structure. *Proc. Natl. Acad. Sci. USA* **88**:3334–3338.

PIERRE CAPY, reviewing editor

Accepted January 4, 2000