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 Lemi1. We show that its putative product is related to transposases of the Tcl/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 Catch-aside 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 sequences 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 Lemi1, which could encode a putative transposase. As sequence similarity between the MITE and Lemi1 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 Lemi1 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-
the Tc1 (or CACT), like TIRs of several class II transposons of all of these MITEs have TIRs that begin with CAGT family. What strikes us more significantly is the fact that no MITE could be considered members of this same MITE family. These MITE elements, these MITEs are generally flanked by imperfect TIRs (two mismatches, as indicated by lowercase type). The 5' TIR of Nemol 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 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 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.

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, Nemol, 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 Nemol 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 superfamly. 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, andlovskaya 1997; Plasterk, Izsáv, 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 ($P_{\text{TBLASTN}} = 2 \times 10^{-44}$ with pogo, $P_{\text{TBLASTN}} = 3 \times 10^{-24}$ 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 andBelknap 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,
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 Lemi1 (larger Emigrant). As sequence homology between Lemi1 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 Lemi1.

The TIRs of Lemi1 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, Lemi1 is flanked by a TA dinucleotide, indicating a putative TA target site duplication, a hallmark of the mariner superfamily of transposable elements, being closer to the pogo family.

By using the DNA sequence of Lemi1 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; \( P_{BLAST} = 2 \times 10^{-12} \)). Because of a severe truncation at the 3' end, this element, named Lemi2, is only 1,008 bp long. The 5' end of Lemi2 is defined by a putative TIR which shares high homology with those of Lemi1 (2024 bp) and is flanked by a TA dinucleotide, reminiscent of the target site duplication. Despite relatively good conservation at the nucleotide level between Lemi2 and Lemi1 (68.3%), it is very difficult to align Lemi2 truncated product with transposases of other pogo-like elements, because some frameshifts are needed to maintain a significant amino acid alignment (data not shown). BLAST searches using Lemi2 sequence as a query did not reveal any other member of this family in the tomato sequences available in databases. It is likely that Lemi2 is a "molecular fossil" of an ancestral pogo-like element of the Solanaceous genome. Interestingly, Lemi2 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 euakryote genes (McDonald 1995; Britten 1996; Kidwell and Lisch 1997).

Coding Capacity of Lemi1

We carefully analyzed the DNA sequence of Lemi1 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 Lemi1 has suffered at least one single mutation, since a stop codon occurs in this frame at position 85714. This suggests that Lemi1 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-
This suggests that ORF1 and ORF2 of the Tc1/mariner superfamily of transposable elements are joined by splicing, like those of the Tc1/mariner superfamily. A search for 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 Fov1 group (Daboussi, Langin, and Bryggo 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 Lemi1 putative product with the D,D35E domains of several transposases from pogo family members (fig. 4). According to this alignment, Lemi1 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, Lemi1 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 Lemi1, 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 (Pietrokovski and Henikoff 1997; Wang, Hartwood, and Finnegan 1999). The presence of a putative HTH motif in the N-terminal region of Lemi1 was also indicated by the Dodd and Egan (1990) method, despite low statistical significance (data not shown). Nevertheless, it is an additional indication that Lemi1 could encode a pogo-like transposase.

Discussion

Lemi1: 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, Lemi1, which has coding capacity for a putative transposase. We show that Lemi1 belongs to the Tc1/mariner superfamily of transposable elements, being closer to the Drosophila pogo and the human Tigger elements.

To our knowledge, Lemi1 is 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
display significant sequence similarity with *Lemi1*. A remnant of another ancient *Lemi* 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 *Lemi1* Responsible for the Mobility of Emigrant MITEs?

There are at least 250 *Lemi1*-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, *Lemi1* 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 *Lemi1* suggests that this copy might be no longer active; also, it remains uncertain that *Lemi1* 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 *Lemi1* copy that could provide a source of transposase for *Emigrant* elements. Hybridizations of *Arabidopsis* DNA with a large internal coding fragment of *Lemi1* 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 *Lemi* 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-

---

**Fig. 4.** Amino acid alignment of the central region of the putative product of *Lemi1* 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* (Z08500) and the distantly related *Tc1* (X01005), and also members of the fungal Fot1 group: *Magnaporthe grisea* Pot2 (U33638), *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 assign a leader). Residues of the DDD (or DDE) motifs are indicated by crosses.
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 Tcl/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 (Pietrokovski and Henikoff 1997; Wang, Hartwood, 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 (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. Mourguier, N. Pourtau, and J.-C. Salvador 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


MITEs Originated from pogo-like DNA Transposons


Pierre Capy, reviewing editor

Accepted January 4, 2000