J Mol Evol
DOI 10.1007/s00239-008-9178-1

Analysis of the DDE Motif in the *Mutator* Superfamily

Aurélie Hua-Van · Pierre Capy

Received: 16 October 2008 / Accepted: 16 October 2008

Abstract The eukaryotic *Mutator* family of transposable elements is widespread in plants. Active or potentially active copies are also found in fungi and protozoans, and sequences related to this family have been detected in metazoans as well. Members of this family are called *Mutator*-like elements (*MULEs*). They encode transposases, which contain a region conserved with transposases of the *IS256* prokaryotic family, known to harbor a DDE catalytic domain. Different DDE or D34E motifs have been proposed in some groups of eukaryotic *MULEs* based on primary sequence conservation. On a large number of protein sequences related to, and representative of, all *MULE* families, we analyzed global conservation, the close environment of different acidic residues and the secondary structure. This allowed us to identify a potential DDE motif that is likely to be homologous to the one in *IS256*-like transposases. The characteristics of this motif are depicted in each known family of *MULEs*. Different hypotheses about the evolution of this triad are discussed.

Keywords Transposase · DDE catalytic core · Mutator-like elements · Protein domain · Transposable element

Electronic supplementary material The online version of this article (doi:10.1007/s00239-008-9178-1) contains supplementary material, which is available to authorized users.

A. Hua-Van (✉) · P. Capy
Laboratoire Evolution, Génomes et Spéciation UPR9034, Centre National de la Recherche Scientifique, 91198 Gif-sur-Yvette, France
e-mail: ahuavan@legs.cnrs-gif.fr

A. Hua-Van · P. Capy
Université Paris-Sud XI, Orsay 91405, France

The *Mutator* superfamily is a large and heterogeneous eukaryotic class II transposable element (TE) family. *Mutator*-like elements (*MULEs*) are characterized by long terminal inverted repeats (TIRs) that surround one or several open reading frames (ORFs), one of which encodes the transposase. Full-length *MULEs* can be divided in several clades based on sequence similarity at the protein level. The most characterized group is the *MuDR* clade (*MuDR*-like elements). Its representative is the active *MuDR* element (See Walbot and Rudenko [2002] for review); this has been widely studied in maize and serves as a model. Unlike the other *MuDR*-like elements, the *MuDR* element contains two tail-to-tail ORFs, encoding MURA and MURB proteins (Lisch et al. 1999). The first corresponds to the transposase, whereas the second has no clear function. *MuDR*-like elements have so far been identified only in plants. The active element *Jittery*, which defines another group of plant *Mutator* elements, was identified more recently in maize (Xu et al. 2004) and is characterized by one transposase ORF only. *MULEs* have also been described in fungi (Chalvet et al. 2003) and protozoans (Pritham et al. 2005), which represent two new families. Finally, *MULEs* sequences have been reported in metazoans as well (Pritham et al. 2005).

In addition to these families of simple full-length copies, a large population of nonautonomous copies exists, at least in plants. Some are inactive deletion derivatives or elements only sharing TIRs with the corresponding full-length elements (Lisch 2002). Arabidopsis and rice genome surveys have showed several groups of *MULEs* based on nucleotide similarities (Turcotte et al. 2001; Yu et al. 2000). Other nonautonomous elements contain host sequences or genes, illustrating the capacity of this family to capture and amplify host sequences. This phenomenon has been observed in every plant studied thus far but is

especially important in rice. Elements containing such extra sequences trans-duplicated from the genome are called *Pack-MULEs* (Jiang et al. 2004). Elements with extra host sequences that also contain a MURA-related ORF have also been described. In Arabidopsis, some groups, so-called *non-TIR MULEs*, appear to lack TIRs (Yu et al. 2000). Finally, several examples of domestication have also been described for this group, e.g., the FHY3 and FAR1 proteins involved in gene-expression regulation (Hudson et al. 2003) and the MUSTANG protein family of yet unknown function (Cowan et al. 2005), which all derive from *MULE* transposases.

Although distantly related, all *MULE* transposases share a conserved domain (domain 1) characterized by several acidic residues (D or E) preceding a CH motif (Chalvet et al. 2003). Weak sequence similarities in this region link the *Mutator* superfamily and the prokaryotic *IS256* family, which is present in diverse Prokaryotes (Eisen et al. 1994).

Some *Mutator* transposases (mainly *MuDR*-like element and the fungal element *Hop*) display a CCHC motif in the C-terminus, referred to as domain 2, which is absent in the *IS256* family (Chalvet et al. 2003; Yu et al. 2000). Another C terminal Zn finger domain of the form $CX_2CX_nCX_2H$, referred to as the SWIM domain, has been found in some *MULE* transposases and also in other nontransposase proteins (Makarova et al. 2002). Finally, another Zn finger motif, WRKY-GCM1, belonging to a large family present in some transcription factors, has been found in the N-terminal portion of all *MULE* transposases (Babu et al. 2006). Both of these Zn finger motifs are specific to eukaryotic proteins.

The transposases of the related *IS256* family harbor a DDE domain, which is also present in several other prokaryotic and eukaryotic transposases and in LTR-retroelement integrases (Haren et al. 1999). The DDE catalytic site is the most common site involved in transposition reaction. The three acidic residues of this motif are dispersed over the primary sequence but form, in the tertiary structure, a pocket able to bind a divalent ion, which is required for the transposition reaction (Davies et al. 1999). In the *IS256* transposase, it has been shown that each residue of the predicted DDE motif is necessary for the formation of the circular forms of the element, which are assumed to correspond to transposition intermediates (Loessner et al. 2002). Mutagenesis of any of the three residues completely abolishes the formation of the *IS256* circles. The first two Ds of the motif lie in the conserved region and are themselves conserved between the eukaryotic and prokaryotic sequences (domain 1). However, the bacterial E residue is located outside the region conserved between Prokaryotes and Eukaryotes, calling into question the existence of a DDE motif in *Mutator* transposases (Fig. 1).

Previous investigators have identified DDE-like motifs from the alignments of several plant *MULEs*. A D(34)E motif, assumed to correspond to the last two residues, was identified in all *MULEs* (Lisch 2002; Fig. 1). Indeed, the most common form of the DDE signature is characterized by an almost invariant 34- or 35-amino acid (AA) spacer, located between the second D and the E (Haren et al. 1999). More recently, an analysis of sugarcane *MULEs* of the *MuDR*-like clade identified a putative DDE triad of the type $DX_{18}DX_{15}E$ (Rossi et al. 2004; Fig. 1). In this case, the first D and the last E correspond to the previously found D(34)E motif. All of these residues, including the last E, lie in the region conserved between eukaryotic and bacterial sequences, i.e., domain 1 (see Fig. 1). The first D corresponds to the first D of the bacterial DDE, but the other two residues are located upstream of the second bacterial D motif. Given that the spacer between the second D and the

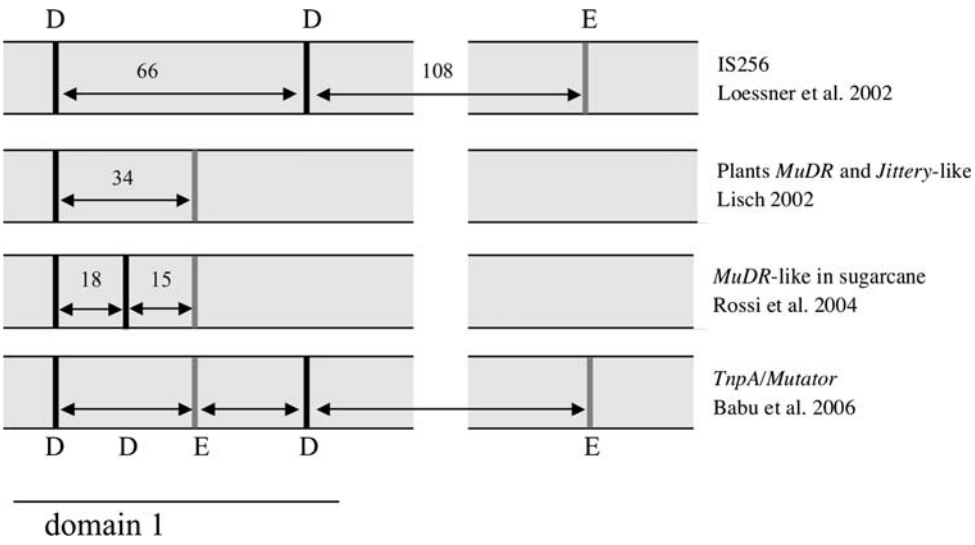


Fig. 1 Structure of the different conserved motifs reported in the literature. Conserved Ds are indicated by vertical black bars and conserved Es by vertical grey bars. The distance in AA between two residues is noted above the double arrows. Type of elements and references are on the right-hand side

E in *IS256* is not 35-AA but rather 108-AA long (Chandler and Mahillon 2002), it appears unlikely that this motif, although quite well conserved, corresponds to a bona fide DDE motif.

Recently, while analyzing the Zn finger present in the N terminal part of *Mutator* and numerous other proteins, Babu et al. (2006) briefly described in some bacterial and *Mutator* transposases a catalytic core composed of several acidic residues (DEDE) with a secondary structure similar to that of other DDE sequences (Fig. 1). However, no more details were given.

In the work presented here we tried to identify the DDE motif in the eukaryotic *Mutator* superfamily by analyzing a pool of eukaryotic sequences representative of the different *MULEs* families and their host species. We took into account the similarity found with bacterial sequences and used methods differing from primary sequence comparison to circumvent the wide sequence divergence observed within this superfamily. Detailed analysis of the protein sequences confirmed the existence of a core catalytic domain of DDE type succinctly proposed by Babu et al. (2006). Differences can be noticed between the different *MULE* families, raising questions about how they evolved.

Materials and Methods

Mutator-*Like Sequences Used as Queries in PSIBLAST*

Nine transposase sequences from the *Mutator* family, corresponding to both known representatives and various divergent sequences (from metazoans, identified after a database homology search), as well as six sequences from the *IS256* family, representing the different clades described by Chandler and Mahillon (2002), were chosen as query sequences. They were first aligned and the conserved 1 domains PSIBLASTed in a batch against the UniProt Database (http://www.expasy.org; version 10). The accession numbers of the sequences and the domain limits are indicated in Table 1. A cut-off value of 0.001 and a maximum of five iterations were used as PSIBLAST parameters (five iterations were enough to reach a plateau, if not convergence).

Filtering Processes

Different protein sequences ($n = 2107$) were retrieved and first filtered for size (150 to 1500 AA), resulting in 1940 sequences that were submitted to a cluster analysis with a threshold of 75% and a coverage length of 0.5 (BLASTCLUST from the NCBI BLAST package).

Alignments and Clade/Phylogeny Analysis

Sequences ($n = 642$) were retained from the cluster analysis, 422 of which were unclustered, with 465 from Eukaryotes (including 2 from viruses; UNIPROT accessions O92526_9VIRU and Q8UZC5_9VIRU) and 177 from Prokaryotes (of which 9 were from Archaea). The prokaryotic sequences were aligned with MUSCLE (Edgar 2004) using manual adjustments. For eukaryotic sequences, the alignment did not extend beyond conserved domain 1, except in a clade-by-clade analysis. Clade analysis of all eukaryotic sequences was performed using a neighbor-joining (NJ) method from PAUP4.0b10 (Swofford 2002).

Table 1 Transposase sequences used for PSIBLASTs[a]

Element	UNIPROT ID	Start	End	Species
IS1354	Q57363_METDI	171	293	*Methylobacterium dichloromethanicum*
	P94292_BURCE	148	272	*Burkholderia cepacia*
IS256	Q7DHL0_STAAU	140	272	*Staphylococcus aureus*
IS406	TRA6_BURCE	141	263	*B. cepacia*
	Q9RMI8_BRELN	142	281	*Brevibacterium linens*
	Q932H1_STAAM	137	267	*S. aureus (strain Mu50)*
Jittery	Q9M4X4_MAIZE	222	344	*Zea mays*
	Q9C9S3_ARATH	413	539	*Arabidopsis thaliana*
MuDR	Q42419_MAIZE	304	427	*Z. mays*
Hop	Q870E2_FUSOX	207	332	*F. oxysporum*
	Q5GIT1_CUCME	407	530	*Cucumis melo*
	Q8S7N0_ORYSA	258	380	*Orysa sativa*
	Q1RLC1_CIOIN	272	396	*Ciona intestinalis*
	O44823_CAEEL	251	378	*C. elegans*
	Q966Z9_CHIPA	102	234	*Chironomus pallidivittatus*

[a] The beginning and the end of the region used are indicated for each sequence

Aligned domains of a subset of 100 sequences were submitted to phylogenetic analysis by maximum parsimony (MP) using PAUP4.0b10, and 100 bootstrap replicates were performed.

Similarity and Consensus Analysis

Consensus sequences were obtained using WebLogo (http://weblogo.berkeley.edu/; Crooks et al. 2004). The similarity analysis within each group was performed on the MUSCLE alignment using the plotcon program, including in the EMBOSS package (Rice et al. 2000). This program evaluated the average similarity using the following formula (Equation 1):

$$\text{Av. Sim.} = \text{sum}(\text{Mij} * \text{wi} + \text{Mji} * \text{wj})/ [(\text{Nseq} * \text{Wsize}) * ((\text{Nseq} - 1) * \text{Wsize})],$$

where w is sequence weighting, M is matrix comparison table (EBLOSUM62 is the substitution matrix used by default), i,j are residues i or j, respectively, Nseq is number of sequences in the alignment, and Wsize is window size. Conserved regions were arbitrarily defined as block of residues with a similarity value > 0.008 when window size is set to 50 (see Fig. 2).

Secondary and Tertiary Structure Analyses

Secondary structures were determined using PSIPRED (http://bioinf.cs.ucl.ac.uk/psipred/; Bryson et al. 2005), and JPRED (http://www.compbio.dundee.ac.uk/approximately www-jpred/; Cuff et al. 1998) on a sample of approximately 20 sequences for each clade. The two methods yielded similar results, although the confidence level was sometimes different.

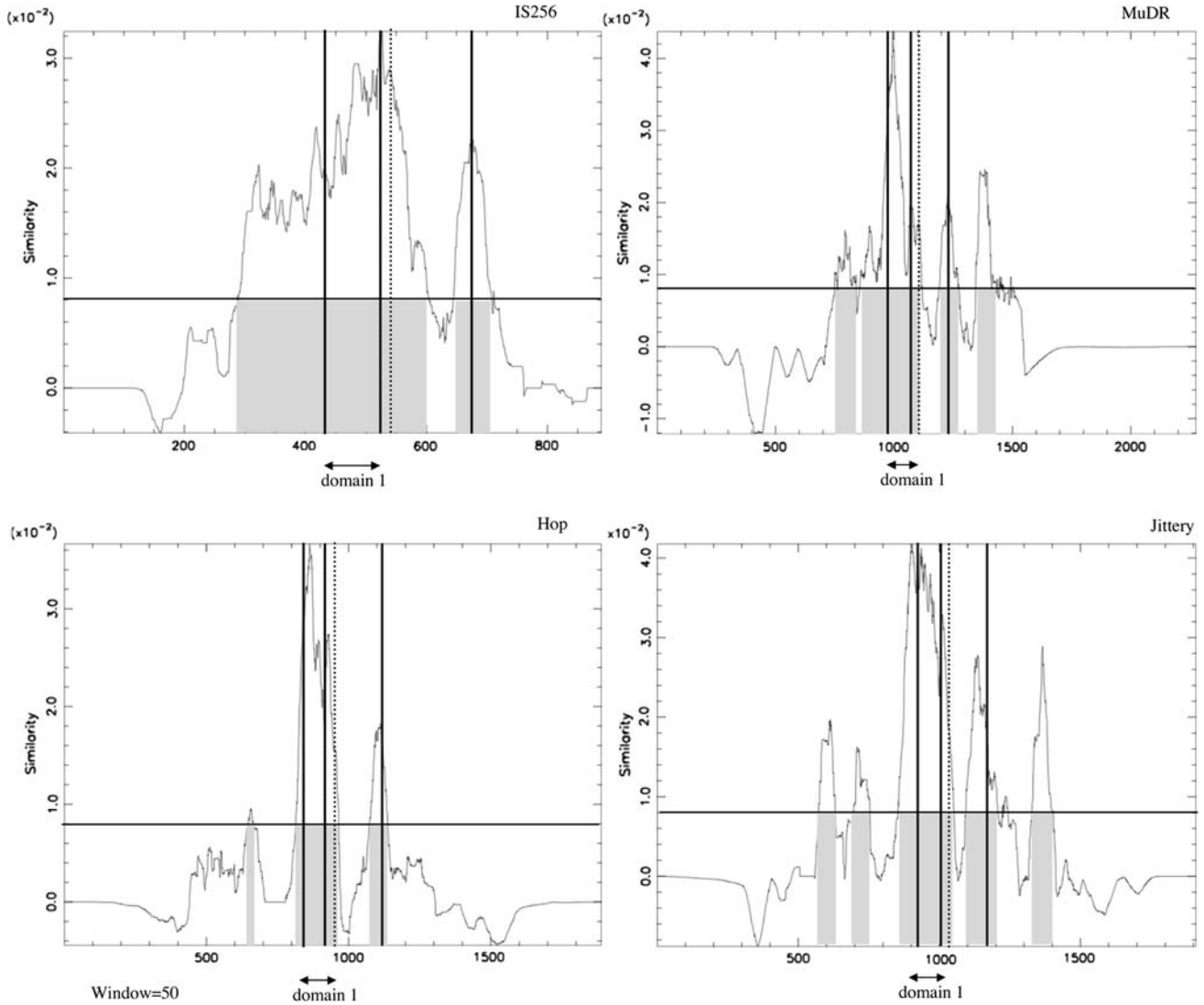


Fig. 2 Conservation plot of the IS256/*Mutator* groups. DDE residues are indicated by vertical bars and conserved CH by a vertical dotted line. Conserved regions (>0.8 10^{-2}) are in grey

Tertiary structures (fold modeling) were assayed on four reference sequences using different methods available by way of various Web and meta servers: 3D-Jury (http://meta.bioinfo.pl/submit_wizard.pl; Ginalski et al. 2003); Pcons (http://pcons.net/; Wallner and Elofsson 2005); Phyre (http://www.sbg.bio.ic.ac.uk/phyre/html/index.html; Bennett-Lovsey et al. 2008), which uses threading methods and PDB or SCOP data banks; and I-TASSER (http://zhang.bioinformatics.ku.edu/I-TASSER/; Zhang 2008), which uses combined threading and ab initio methods.

Results

Sequence Recovery and Alignment Methods

To investigate whether a DDE motif could be detected in the eukaryotic transposases, we collected a large number of proteins that show similarities to *MULE* transposases and compared them with bacterial *IS256* transposases. Those proteins were collected with a PSI-BLAST search using domain 1 of 15 transposases sequences as queries (see Materials and Methods). A total of 2107 sequences were retrieved. More than 1194 were from Eukaryotes: Among them, at least one third were from plants, and >300 were from the recently sequenced genome of the protozoan *Trichomonas vaginalis* (Carlton et al. 2007). The pool of sequences was then filtered using diverse criteria (see Materials and Methods) resulting in 465 eukaryotic/viral and 177 prokaryotic sequences (Eubacteria/Archaea).

These sequences presented so much diversity that only domain 1 could be correctly aligned from a global analysis. We chose to separate the prokaryotic sequences from the eukaryotic ones and to split the eukaryotic sequences into distinct clades according to families. An NJ analysis of the conserved domain 1 (AA 305 to 425 of *MuDR* transposase) of the eukaryotic sequences resulted in four groups. Fourteen sequences that did not align over domain 1 were excluded. Three clades corresponded to the known families: Hop (48 *Hop*-like sequences from fungi and plants), Jittery (132 *Jittery*-like sequences from plants), and MuDR (205 *MuDR*-like sequences from plants). The 66 remaining sequences were put into an artificial clade named "others." Sequences from several kingdoms were present in this last group. Sequences from each clade were first aligned, and then a global alignment of the alignments from each clade was performed (a subset of 100 aligned sequences is available as supplementary data).

Conserved Domains in Each Clade

The clade-by-clade alignment allowed the removal of sequences according to different criteria, such as suspect alignment and sequences that were too short or contained internal gaps or insertions that were too large. The final filtered alignments were submitted to a conservation quality analysis, resulting in the plots shown in Fig. 2. For the divergent sequences in the "others category," alignment was made manually after having eliminated numerous sequences that were clearly truncated in domain 1 or structurally too different. The similarity plot made on the remaining 25 sequences showed a general form of the curve that was quite similar to other groups, i.e., two domains of higher conservation. However, conservation was globally lower in this group (data not shown).

The prokaryotic sequences (*IS256*-like) displayed strong conservation compared with the three eukaryotic groups. Two domains were observed, the first containing the first two Ds of the DDE motif and the second, shorter one containing the last E (Fig. 2 upper left panel).

A variable number of conserved domains were observed in eukaryotic groups (Fig. 2). However, the largest domain found was always the domain 1, containing the two conserved Ds (vertical lines). This domain was always preceded by one or two short conserved regions that contained a Zn finger motif present as two variant forms: $C(X_{>10})C(X)H(2)H$ for Jittery and Hop or $C(X_{<10})$-$C(X)H(2)C$ for MuDR. Both motifs belong to the WRKY-GCM1 DNA-binding domains superfamily (Babu et al. 2006).

Downstream from domain 1, two other domains were identified in MuDR and Jittery, but only one domain was identified in Hop. The domain found only in MuDR and Jittery contained the conserved SWIM motif (CCCH) described by Makarova et al. (2002). In the *Hop* clade this motif was often altered, and several sequences did not go beyond this domain. Yu et al. (2000) and Chalvet et al. (2003) previously described another Zn finger motif of the CCHC type in Arabidopsis *MULE* and *Hop* sequences. This domain appears to be conserved only within the *MuDR*-like group and is also found in some *Hop*- and *Jittery*-like sequences. It is located downstream of the SWIM motif, but nothing is known of its functionality.

The Conserved D and E Residues

In Prokaryotes, the conserved second domain, downstream from the large domain 1, contains the E of the DDE motif. Hence, the DDE triad is spread over two different domains. The E residue lies in the most conserved part of this domain. We searched for any conserved D and E residues in the three eukaryotic groups and in the *IS256* family as well. Only sequences extending beyond the second conserved domain, and aligning correctly with other sequences over the suspected areas, were retained. All of the residues

conserved at >70% in the alignments are shown in Table 2.

In the prokaryotic sequences, only two conserved Ds and two conserved Es were detected, although global conservation was high. The 2 Ds corresponded to the D of the DDE motif and surrounded the first conserved E. The second E was located an average of 108 AA downstream of the second D and is the E of the DDE motif (see Table 2).

In the three eukaryotic groups, the two Ds and the E in between were also conserved. In Hop, as in IS256, only two Ds and two Es appeared to be conserved at >70%, whereas some other positions were less conserved. The average distance between the last D and E was 131 AA, higher than for Prokaryotes, because of the presence of an insert of variable size and sequence in most *Hop*-like proteins. For the 88 *MuDR*-like sequences used in this analysis, a conserved extra D was observed just 19 AA downstream from the first one, but it did not correspond to a residue of the triad. The average distance was slightly lower than for Prokaryotes. In Jittery, a conserved D was present just upstream from the first conserved E, and three other Es were detected in the conserved region downstream from domain 1. The average distance between the second D and the next E was 108 AA, as found in Prokaryotes.

Hence, each group is characterized by a conserved E, which is at a distance comparable with the Prokaryote reference group. Often, this E corresponded to the most conserved one (>70%).

The AA Environment

A high conservation level and spacing suggest that the proposed E corresponds to the E of the DDE signature. When the different alignments were combined, the suspected Es came together. This was not the case when we tried to align all of the sequence at once, probably because the divergence between sequences was too great, particularly in the region between the two conserved domains.

Fig. 3 Primary and secondary structure analyses. **a** AA consensus around the conserved acidic residues as determined by WebLogo analysis. The position indicated below each column is relative to the alignment presented in panel B. **b** Schematic alignment of secondary structures (PSIPRED) in the region encompassing the eukaryotic Zn finger motifs and the DDE triad. β-sheets are in blue, and α-helices are in red. Coil-coiled regions are in grey. The conserved residues are shown below the alignment. Residues highlighted in panel A are in grey background. The numbering corresponds to that of the alignment of the 100 sequences used. The grey curve above the alignment represents the confidence value provided by PSIPRED and were averaged for each position. The alignment was anchored on conserved residues and manually adjusted for secondary structures. Numbering is relative to the entire alignment

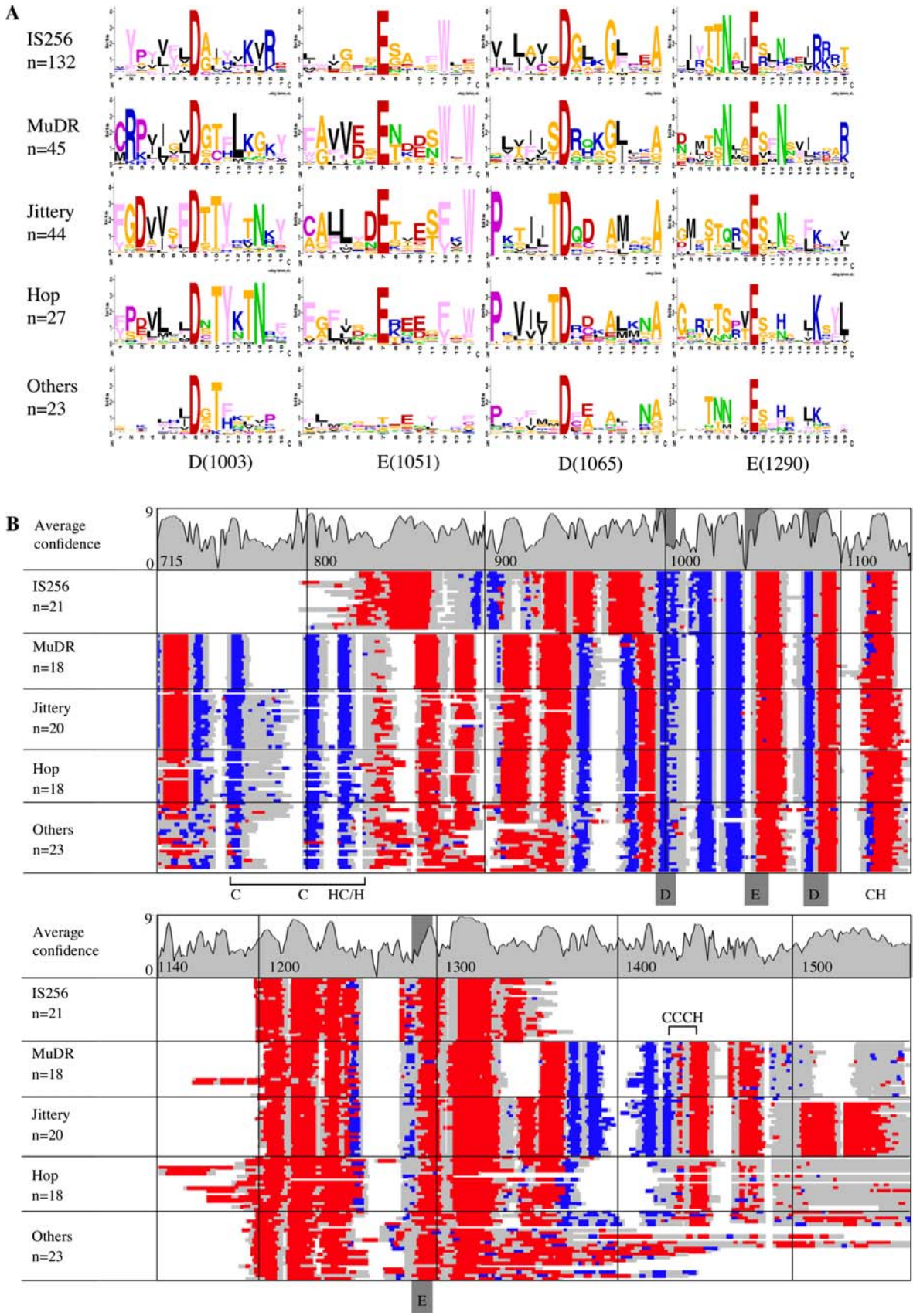
However, to verify that the Es did not align by chance, we compared their environment using a WebLogo analysis. This analysis, carried out on the conserved acidic residues, including the intervening E, is presented in Fig. 3a. Only sequences containing a perfect triad were kept. This analysis also included the group "others", which was comprised of disparate sequences. The conserved residues, including the triad, could be readily aligned for some of them, although global conservation was low. In these 23 sequences, E residues could be detected between the two Ds, but they were in different positions. The homology between prokaryotic and *MuDR*-like sequences was evident for the last D and E and between Jittery and Hop for the two Ds. Moreover, few other positions appeared to be conserved across groups.

Although conservation around the E of the DDE triad was globally weaker than that of residues in domain 1, several trends could be observed. According to Chandler and Mahillon (2002), AA +7 after the E (two helical turns) in DDE-containing proteins is usually a basic residue (R or K). Residues located to approximately one helical turn are also often conserved (residues 3 or 4 upstream and

Table 2 Positions of conserved Ds and Es (>70%) in each clade[a]

Conserved D/E		IS (n = 125)		Hop (n = 26)		MuDR (n = 88)		Jittery (n = 88)	
		Position	%	Position	%	Position	%	Position	%
D	First D of DDE	433	99.2	845	96.1	976	94.3	919	94.3
D						995	94.3	962	71.6
E		487	98.4	880	96.1	1011	93.1	963	84.1
D	Second D of DDE	521	100	912	92.3	1044	87.5	1003	75
E	E of DDE	675	100	1112	96.1	1223	95.5	1178	79.6
E								1214	80.7
E								1257	70.5
Average DE spacer length		108 AA		131 AA		102 AA		108 AA	

[a] For each residue, the position is given relative to the alignment. The DE spacer length (in AA) represents the average distance between D and E and is therefore lower than the maximum distance deduced from the positions in the alignment


A
IS256
n=132
MuDR
n=45
Jittery
n=44
Hop
n=27
Others
n=23
D(1003)
E(1051)
D(1065)
E(1290)
B
Average
confidence
9
0
715
800
900
1000
1100
IS256
n=21
MuDR
n=18
Jittery
n=20
Hop
n=18
Others
n=23
C
C
HC/H
D
E
D
CH
Average
confidence
9
0
1140
1200
1300
1400
1500
IS256
n=21
CCCH
MuDR
n=18
Jittery
n=20
Hop
n=18
Others
n=23
E

downstream from the E). In eukaryotic *Mutator* proteins, AAs −5 to −3 relative to the E were often S, T, N or Q and AA +1 was often S and AA +3, N, or H. These match quite well the prokaryotic consensus described by Chandler and Mahillon (2002) as well as our own, which was obtained from 132 prokaryotic sequences. In retroviral integrases, the +7 residue is important because it interacts with the DNA. R or K residues in the +7 position were observed in IS256, "others", and Hop, but they appeared less obvious for MuDR and Jittery because they were only present in two thirds and three fourths of the sequences, respectively. However, in MuDR, a basic residue was conserved in +10. Considering that the +10 AA would be on the same face of the DNA helix as the +7 AA, it could, in theory, functionally replace it.

The Conserved Secondary Structure

Because the homology of the E in different clades was not obvious at first glance, because of divergence between the groups, we hypothesized that it might be more evident at the secondary structure level. An analysis of the secondary structure of 100 sequences was carried out using the PSIPRED and JPRED algorithms. The secondary structures obtained by both methods were quite similar. They were first aligned on the basis of the AA alignment and adjusted by hand by aligning α-helices and β-sheets (see Fig. 3b for PSIPRED alignment). For eukaryotic sequences, a good structural similarity was observed over a large region, starting from the N-terminal Zn finger motif up to a large α-helix located downstream from the E-containing helix. Downward, structural unity was observed only for MuDR and Jittery on the region containing the SWIM Zn finger. Overall, these two families appear highly similar at the secondary structure level compared with the somewhat divergent primary sequences. In N terminal regions, the Zn finger has been shown to be of the WRKY type (Babu et al. 2006), which is typically composed of four β-sheets (Yamasaki et al. 2005). This structure was predicted with a good confidence level for most sequences. *Jittery* and *Hop* groups exhibited the insert-containing subtype, whereas the remaining sequences (mainly *MuDR*-like) were characterized by the classical insert-free subtype ($CX_4CX_nHX_1H$). Another suite of four β-sheets was found in Jittery and MuDR just upstream from the SWIM motif CCCH. This structure was absent from half of the sequences in the "others" category and from all *Hop*-like sequences. Accordingly, the SWIM CCCH motif is less conserved in *Hop*-like sequences and is absent in most sequences of the "others" clade. The predicted structure of the SWIM motif differs slightly from the one previously predicted (Makarova et al. 2002) because of a short β-sheet predicted on the two first Cs. In JPRED as in PSIPRED analyses, this β-sheet was present but was supported by a good confidence level only in *Jittery*-like sequences. This difference with Makarova et al.'s structure can be imputed to the difference in the methodology used and may be solved by domain crystallization. In the group "others," two subgroups became apparent, distinguishable by the type of WRKY motif (without insert for sequences from *T. vaginalis* and with insert of 9 to 26 AA for other sequences) and the presence or absence of the stretch of β-sheets associated with the SWIM motif in the C terminus part (clearly present in five sequences only).

Between Prokaryotes and Eukaryotes, the structural similarity extended from AAs 999 to 1320, thus encompassing the two conserved domains and the three catalytic residues. The conserved domain 1, made of three β-sheets, followed by a short α-helix and another β-sheet, has the typical structure found in other prokaryotic or eukaryotic DDE domains (*HIV*, *Tn5*, *Hermes*, or *Mos1* [Davies et al. 2000; Dyda et al. 1994; Hickman et al. 2005; Richardson et al. 2006]). As expected, the two Ds of the triad fell at the end of β-sheets. In all sequences the E was located at the beginning of an α-helix, which is also the case for the E of every other DDE motif. Between the last D and the E, a suite of α-helices is predicted for any sequence, forming an inserted sequence that is absent in typical catalytic DD(35)E domains. In this region, structural homogeneity was more obvious for *IS256*, *Jittery*, or *MuDR* groups, whereas Hop and the "others" groups displayed higher diversity.

Fold Modeling

The full DDE domains of *IS256*, *MuDR*, *Jittery*, and *Hop* transposases were analyzed using different Web-based three-dimensional (3-D) prediction programs (see Materials and Methods). All of them compare the query sequence with proteins in Protein Data Bank (PDB). With a few exceptions, results were obtained with a low score, probably reflecting the absence of crystallized *Mutator*-like proteins. However, various crystallized DDE proteins constantly showed up among the 10 or 20 best models listed by each server (supplementary Table 1). Globally, methods detected homology with every crystallized DDE protein present in data banks (retroviral, such as *HIV* or *RSV* integrases, bacteriophage *Mu* transposase, and *Mos1* and *Hermes* transposases), with the notable exception of the *Tn5* transposase. Although the full domain was always sent as query, few comparisons (mainly with *HIV* integrases) succeeded in aligning the third residue, explaining in part the low score observed. As a consequence, the different fold modeling obtained for the four sequences usually placed the two Ds in close proximity to each other, but only in a few of them could the E be placed next to the

Ds (two such 3-D models obtained for *Jittery* are given in the supplementary data). The likely reason for this is the *Mutator* intervening region, which does not fit with any such region in structurally known DDE proteins.

Phylogenetic Relations

Phylogenetic analysis was performed on the catalytic core (AA 325 to 502 relative to the *MuDR* transposase) of 100 sequences. Six different clades were observed, with "others" being split into two groups (Other1 and Other2). However, bootstrap support was only obtained for IS256, MuDR, and Jittery and for some subsets in Hop and Other1 (Fig. 4). The two groups of sequences from "others" did not correlate either with the presence or absence of the SWIM motif or with the type of WRKY motif. For example, the *T. vaginalis* sequences, which all have a WRKY-type Zn finger without an insert, were found in the two clades Other1 and Other2, whereas the SWIM motif was only found in a subset of sequences in the Other1 clade.

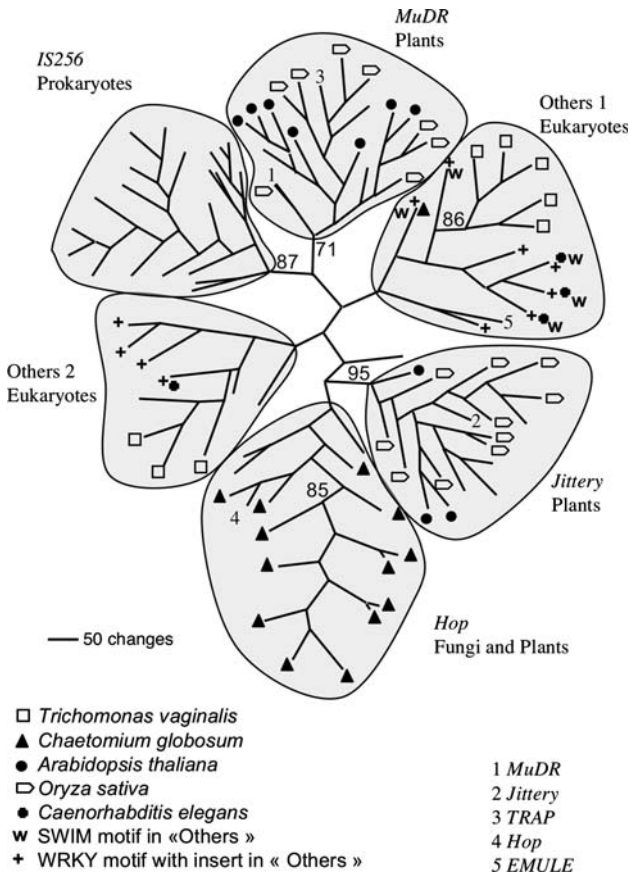


Fig. 4 Phylogenetic tree based on the DDE region constructed by MP (default parameters) using PAUP4.0b*. Bootstrap numbers at the nodes were obtained after 100 replicates

Discussion

The DDE catalytic domain is the most common domain involved in transposition of class I and II elements. It has been searched for in every transposase of every TE superfamily. It has been found in a number of new superfamilies such as *PIF/Harbinger*, *Transib*, and *Merlin* (Feschotte 2004; Kapitonov and Jurka 2003; Zhang et al. 2001) and has been recently identified in the tertiary structure of a *hAT* transposase (Hickman et al. 2005). However, the catalytic site of other superfamilies is still elusive, e.g., in *P*. In the *Mutator* superfamily, several investigators previously described a DDE catalytic triad or at least a D(34)E block, which would mean that MULEs belong to the DDE transposase club. This would not be surprising because MULE elements are clearly related to the *IS256* family, a prokaryotic family known to contain a DDE catalytic core. However, by comparing the position of the *IS256* DDE and the DDE motif proposed in MULEs, it seems clear that the proposed catalytic domain was not the right one. More recently, Babu et al. (2006) described a catalytic core common to both eukaryotic and prokaryotic transposases. However, they gave few details about the structure and conservation of the domain in the different families.

In this article, we have described the detailed analysis of numerous sequences related to the *Mutator* and *IS256* families. Using several approaches including similarity, environmental consensus, and secondary and tertiary structures we were able to identify without ambiguity a DDE motif in all *Mutator* families located in the position suspected by Babu et al.(2006), which fits with what we would expect from the *IS256* DDE location. This motif extends across two domains in *IS256* proteins as well as in eukaryotic transposases. Each time, both domains were separated by a spacer of variable length and sequence, even between sequences from the same family. Although the two first Ds could be detected without any difficulty because they lie in the region conserved between prokaryotic *IS256* and *MULE* transposases, the last residue was the most problematic because it lies in a region that does not present obvious conservation at the primary sequence level compared with Prokaryotes and even between eukaryotic groups. However a family-per-family analysis uncovered a second conserved domain containing a conserved E, although conservation was systematically lower than for domain 1. This is surprising because the E residue, as well as the α-helix containing it, are usually extremely well conserved. At the secondary structure level, however, both domains are well conserved, whereas the intervening sequence is variable in size and structure, albeit always composed of α-helices. The structure of the widespread DD(35)E motif suggested an invariant structure for

the two last residues. However, it appears from our data and recent data for *hAT* (Hickman et al. 2005) that the constant part of the motif is the block of two Ds.

Although the E domain is shorter and less conserved than domain 1, the residue environment analysis showed several conserved positions around the E, similar to what we observed for the other residues of the motif. The alignment of secondary structures also illustrates the modularity previously proposed for the evolution of TEs (Capy and Maisonhaute 2002). The N termini of prokaryotic and eukaryotic sequences are completely different. They are most likely to have the same function in recognition and binding DNA but have been independently associated with the same catalytic core. Apart from modularity examples, progressive loss or evolution of other domains also occurred. The SWIM motif region is well conserved in *Jittery*- and *MuDR*-like sequences, but it is altered to various degrees in *Hop* sequences. Hence, this group of protein no longer needs the Zn finger, whereas it seems functionally important for Jittery and MuDR. With regard to the different Zn finger, the situation of *Jittery*-like group is somewhat curious. Indeed, Jittery resembles Hop for the N terminal WRKY Zn finger motif but resembles MuDR for the C terminal SWIM motif.

The relatively weak conservation in the domain containing the E residue, along with the high variability seen in the region located between the last D and the E, raises questions about the origin of this residue. Accordingly, and considering that only the E residue, the α-helix that contains it, and its location in the 3D structure are important for function, we could hypothesize that the last part of the triad was recruited several times independently, leading to functional convergence. This hypothesis would explain the variable insert as well as the limited conservation of the second domain between groups. However, the environment close to the E exhibits characteristic residues found in several distantly related sequences and in this way resembles the environment of the two Ds, for which a common origin is beyond doubt. Moreover, the variable domains between the two D and the E are structurally well conserved, notably between Jittery, MuDR, and the prokaryotic *IS256*-like sequences. The downstream sequences (SWIM motif) also seem to have a unique origin. Therefore, the whole catalytic domain may be ancestral, and the high divergence observed may be the result of a relaxed selection pressure for the reasons previously described.

The 3D analyses also provided useful information. *Mutator*-like elements' membership in the DDE superfamily was formerly inferred only from multiple alignments of bacterial relatives and identification of conserved DDE residues that were shown to be essential for transposition of the *Staphylococcus IS256* element (Haren et al. 1999; Loessner et al. 2002). In our analysis, 3D prediction servers constantly found PDB models corresponding to known DDE transposases or integrases, hence supporting the conserved secondary structure observed (at least across domain 1). As expected, the predicted folds for *Mutator* domain 1 usually displayed the two Ds in close proximity (not shown). Alignments between model and query sequences usually failed to align together the Es of the DDE motifs. Such results were probably caused by the different DE distances observed between *Mutator* transposases (107 to 134 AA) and crystallized DDE proteins (34 to 35 AA for retrovirus *Mos1* or bacteriophage *Mu* and > 300 AA for *Hermes*). The overall organization of *Hermes*, including the large inserted domain, resembles that predicted for the RAG1 protein involved in V(D)J recombination and known to harbor a DDE domain (Hickman et al. 2005). The *Tn5* transposase is another example of DDE domain split into two parts by an inserted domain. However, unlike *Hermes*, RAG1, or *Mutator*, which are made exclusively of α-helices, the *Tn5* insert is composed of β-sheets (Davies et al. 2000). These different examples show that extra domains are frequently found inserted between the second D and the E of the catalytic domain and may have a role in establishing some interactions within the catalytic site.

Comparing the groups with each other showed different situations. The variability in structure outside the most conserved regions was striking for the transposases of the *Hop*-like and "others" groups. In contrast, for MuDR and Jittery, the entire sequence, including the DE spacer, was structurally well conserved. This strong conservation suggests recent amplification within these two groups. It is noticeable that both are exclusively found in plants, and both are present within the same genome. Hence, several representatives of both families were probably present in the plant ancestor, with some being recently amplified. The *Hop*-like family is structurally closer to Jittery, by the Zn-finger type in N terminus and the sequence around the triad, although the spacer was variable in length and sequence. This group contains sequences mainly from the fungus *Chaetomium globosum* as well as some sequences from Fabaceae, such as *Medicago trunculata* and *Lotus japonicum* (Holligan et al. 2006). The monophyly of most Chaetomium sequences was supported by the bootstrap value, suggesting species-specific amplification. However, the high divergence level within this group suggests a more ancient diversification. This suggests that sequences closely related to Chaetomium sequences may be present in other (fungal) species but not yet identified. It should be noted that one Chaetomium protein sequence is more closely related to the *Hop* transposase of the ascomycete *Fusarium oxysporum* than to other Chaetomium transposases. The remaining sequences, which exhibit the highest

diversity, come from metazoans, fungi, or protozoans. They do not form a monophyletic clade, and, at least in some species, such as *T. vaginalis* or *Caenorhabditis elegans*, two distinct groups coexist. They may be representatives of old families that have not amplified to the same extent as Jittery and MuDR, for which a large number of sequences is available. However, moderate species-specific amplifications seem to have occurred.

Although we were able to detect a full DDE motif in each protein group, the requirement of this domain for transposition has yet to be biologically demonstrated, which has been done for *IS256* and other DDE transposases and integrases. The proteins used here come from the UniProt database and therefore frequently correspond to conceptual translations of DNA sequences. Hence, we have no data on the expression of these proteins or their activity. Some of them may no longer be transposases because several examples of domesticated transposases are known for MULEs in plants (see Feschotte and Pritham [2007] for a review). Domestication and emergence of new functions may have an effect on the evolution of sequences because selection pressure may have a different impact on different domains, leading to variable evolution rates.

Acknowledgements This work was supported by GDR 2157 Les éléments transposables: du génome aux populations, IFR115 Génome: Structure, Fonction, Evolution, PPF2: Bioinformatique et Biomathématiques. We thank Cushla J. Metcalfe and Malcolm Eden for reviewing the English text.

References

Babu MM, Iyer LM, Balaji S, Aravind L (2006) The natural history of the WRKY-GCM1 zinc fingers and the relationship between transcription factors and transposons. Nucleic Acids Res 34:6505–6520

Bennett-Lovsey RM, Herbert AD, Sternberg MJ, Kelley LA (2008) Exploring the extremes of sequence/structure space with ensemble fold recognition in the program Phyre. Proteins 70:611–625

Bryson K, McGuffin LJ, Marsden RL, Ward JJ, Sodhi JS, Jones DT (2005) Protein structure prediction servers at University College London. Nucleic Acids Res 33:W36–W38

Capy P, Maisonhaute C (2002) Acquisition and loss of modules: the construction set of transposable elements. Genetika 38:719–726

Carlton JM, Hirt RP, Silva JC, Delcher AL, Schatz M, Zhao Q, Wortman JR, Bidwell SL, Alsmark UC, Besteiro S et al (2007) Draft genome sequence of the sexually transmitted pathogen *Trichomonas vaginalis*. Science 315:207–212

Chalvet F, Grimaldi C, Kaper F, Langin T, Daboussi MJ (2003) *Hop*, an active *Mutator*-like element in the genome of the fungus *Fusarium oxysporum*. Mol Biol Evol 20:1362–1375

Chandler M, Mahillon J (2002) Insertion sequences revisited. In: Craig NL, Craigie R, Gellert M, Lambowitz AM (eds) Mobile DNA II. ASM Press, Washington, DC, pp 305–360

Cowan RK, Hoen DR, Schoen DJ, Bureau TE (2005) *MUSTANG* is a novel family of domesticated transposase genes found in diverse angiosperms. Mol Biol Evol 22:2084–2089

Crooks GE, Hon G, Chandonia JM, Brenner SE (2004) WebLogo: a sequence logo generator. Genome Res 14:1188–1190

Cuff JA, Clamp ME, Siddiqui AS, Finlay M, Barton GJ (1998) JPred: a consensus secondary structure prediction server. Bioinformatics 14:892–893

Davies DR, Goryshin IY, Reznikoff WS, Rayment I (2000) Three-dimensional structure of the *Tn5* synaptic complex transposition intermediate. Science 289:77–85

Davies DR, Mahnke Braam L, Reznikoff WS, Rayment I (1999) The three-dimensional structure of a *Tn5* transposase-related protein determined to 2.9̄A resolution. J Biol Chem 274:11904–11913

Dyda F, Hickman AB, Jenkins TM, Engelman A, Craigie R, Davies DR (1994) Crystal structure of the catalytic domain of *HIV-1* integrase: similarity to other polynucleotidyl transferases. Science 266:1981–1986

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797

Eisen JA, Benito MI, Walbot V (1994) Sequence similarity of putative transposases links the maize *Mutator* autonomous element and a group of bacterial insertion sequences. Nucleic Acids Res 22:2634–2636

Feschotte C (2004) *Merlin*, a new superfamily of DNA transposons identified in diverse animal genomes and related to bacterial IS1016 insertion sequences. Mol Biol Evol 21:1769–1780

Feschotte C, Pritham EJ (2007) DNA transposons and the evolution of eukaryotic genomes. Annu Rev Genet 41:331–368

Ginalski K, Elofsson A, Fischer D, Rychlewski L (2003) 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics 19:1015–1018

Haren L, Ton-Hoang B, Chandler M (1999) Integrating DNA: transposases and retroviral integrases. Annu Rev Microbiol 53:245–281

Hickman AB, Perez ZN, Zhou L, Musingarimi P, Ghirlando R, Hinshaw JE, Craig NL, Dyda F (2005) Molecular architecture of a eukaryotic DNA transposase. Nat Struct Mol Biol 12:715–721

Holligan D, Zhang X, Jiang N, Pritham EJ, Wessler SR (2006) The transposable element landscape of the model legume *Lotus japonicus*. Genetics 174:2215–2228

Hudson ME, Lisch DR, Quail PH (2003) The FHY3 and FAR1 genes encode transposase-related proteins involved in regulation of gene expression by the phytochrome A-signaling pathway. Plant J 34:453–471

Jiang N, Bao Z, Zhang X, Eddy SR, Wessler SR (2004) *Pack-MULE* transposable elements mediate gene evolution in plants. Nature 431:569–573

Kapitonov VV, Jurka J (2003) Molecular paleontology of transposable elements in the *Drosophila melanogaster* genome. Proc Natl Acad Sci USA 100:6569–6574

Lisch D (2002) Mutator transposons. Trends Plant Sci 7:498–504

Lisch D, Girard L, Donlin M, Freeling M (1999) Functional analysis of deletion derivatives of the maize transposon *MuDR* delineates roles for the MURA and MURB proteins. Genetics 151:331–341

Loessner I, Dietrich K, Dittrich D, Hacker J, Ziebuhr W (2002) Transposase-dependent formation of circular IS256 derivatives in *Staphylococcus epidermidis* and *Staphylococcus aureus*. J Bacteriol 184:4709–4714

Makarova KS, Aravind L, Koonin EV (2002) SWIM, a novel Zn-chelating domain present in bacteria, archaea and eukaryotes. Trends Biochem Sci 27:384–386

Pritham EJ, Feschotte C, Wessler SR (2005) Unexpected diversity and differential success of DNA transposons in four species of entamoeba protozoans. Mol Biol Evol 22:1751–1763

Rice P, Longden I, Bleasby A (2000) EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 16:276–277

Richardson JM, Dawson A, O'Hagan N, Taylor P, Finnegan DJ, Walkinshaw MD (2006) Mechanism of *Mos1* transposition: insights from structural analysis. EMBO J 25:1324–1334

Rossi M, Araujo PG, de Jesus EM, Varani AM, Van Sluys MA (2004) Comparative analysis of *Mutator*-like transposases in sugarcane. Mol Genet Genomics 272:194–203

Swofford DL (2002) PAUP*: phylogenetic Analysis Using Parsimony (*and other Methods). Sinauer Associates, Sunderland, MA

Turcotte K, Srinivasan S, Bureau T (2001) Survey of transposable elements from rice genomic sequences. Plant J 25:169–179

Walbot V, Rudenko GN (2002) *MuDR/Mu* transposable elements of maize. In: Craig NL, Craigie R, Gellert M, Lambowitz AM (eds) Mobile DNA II. ASM Press, Washington, DC, pp 533–564

Wallner B, Elofsson A (2005) Pcons5: combining consensus, structural evaluation and fold recognition scores. Bioinformatics 21:4248–4254

Xu Z, Yan X, Maurais S, Fu H, O'Brien DG, Mottinger J, Dooner HK (2004) *Jittery*, a *Mutator* distant relative with a paradoxical mobile behavior: excision without reinsertion. Plant Cell 16:1105–1114

Yamasaki K, Kigawa T, Inoue M, Tateno M, Yamasaki T, Yabuki T, Aoki M, Seki E, Matsuda T, Tomo Y et al (2005) Solution structure of an Arabidopsis WRKY DNA binding domain. Plant Cell 17:944–956

Yu Z, Wright SI, Bureau TE (2000) *Mutator*-like elements in *Arabidopsis thaliana*. Structure, diversity and evolution. Genetics 156:2019–2031

Zhang X, Feschotte C, Zhang Q, Jiang N, Eggleston WB, Wessler SR (2001) P instability factor: an active maize transposon system associated with the amplification of *Tourist*-like MITEs and a new superfamily of transposases. Proc Natl Acad Sci USA 98:12572–12577

Zhang Y (2008) I-TASSER server for protein 3D structure prediction. BMC Bioinformatics 9:40