Self-synthesizing DNA transposons in eukaryotes

Vladimir V. Kapitonov* and Jerzy Jurka*

Genetic Information Research Institute, 1925 Landings Drive, Mountain View, CA 94043

Communicated by Margaret G. Kidwell, University of Arizona, Tucson, AZ, January 31, 2006 (received for review November 28, 2005)

Eukaryotes contain numerous transposable or mobile elements capable of parasite-like proliferation in the host genome. All known transposable elements in eukaryotes belong to two types: retrotransposons and DNA transposons. Here we report a previously uncharacterized class of DNA transposons called Polintons that populate genomes of protists, fungi, and animals, including entamoeba, soybean rust, hydra, sea anemone, nematodes, fruit flies, beetle, sea urchin, sea squirt, fish, lizard, frog, and chicken. Polintons from all these species are characterized by a unique set of proteins necessary for their transposition, including a protein-primed DNA polymerase B, retroviral integrase, cysteine protease, and ATPase. In addition, Polintons are characterized by 6-bp target site duplications, terminal-inverted repeats that are several hundred nucleotides long, and 5′-AG and TC-3′ termini. Analogously to known transposable elements, Polintons exist as autonomous and nonautonomous elements. Our data suggest that Polintons have evolved from a linear plasmid that acquired a retroviral integrase at least 1 billion years ago. According to the model of Polinton transposition proposed here, a Polinton DNA molecule excised from the genome serves as a template for extrachromosomal synthesis of its double-stranded DNA copy by the Polinton-encoded DNA polymerase and is inserted back into genome by its integrase.

ATPase | cysteine protease | DNA polymerase | integrase | transposable elements

Genomes of most eukaryotes are populated by DNA copies of parasitic elements known as transposable elements (TEs) capable of reproducing themselves in the host genome in a non-Mendelian fashion (1, 2). Understanding the biology of transposable elements is of great importance because of their increasingly well-documented impact on the host genome (2, 3). Moreover, transposable elements can be used as powerful tools in genetic engineering (4). Despite an enormous diversity of eukaryotic TEs, they belong to only two types, called retrotransposons and DNA transposons. Whereas a retrotransposon is transposed (retroposed) via reverse transcription of its mRNAs, a DNA transposon is transposed via transfer of its genomic copy from one site to another. Each type includes different classes and families of TEs composed of autonomous and nonautonomous elements. Whereas an autonomous element encodes a complete set of enzymes characteristic of its family, a nonautonomous element encodes none, or only some of them, and depends on enzymes encoded by its autonomous relative. Transposition of a retrotransposon is catalyzed by reverse transcriptase and endonuclease (EN) domains of a polyprotein encoded by itself or by other retrotransposons. All retrotransposons can be further divided into two subclasses called LTR and non-LTR retrotransposons (5). In addition to the reverse transcriptase/EN polyprotein, most non-LTR retrotransposons code for a second protein characterized by poorly understood activities, including RNA/DNA binding, chaperone, and esterase. An mRNA molecule expressed during transcription of the genomic non-LTR retrotransposon is reverse transcribed and inserted in the genome (5). LTR retrotransposons, including endogenous retroviruses, represent the most complex TEs in eukaryotes. An LTR retrotransposon may carry three ORFs coding for the gag, env, and pol proteins, the latter is composed of the reverse transcriptase, EN, and aspartyl protease domains (5). The endonuclease domain in LTR retrotransposons is usually called integrase (INT) and is distantly related to the DDE transposases (TPase) encoded by Mariner DNA transposons (6).

DNA transposons identified so far in eukaryotes belong to two classes characterized by the so-called “cut-and-paste” (7) and “rolling-circle” (8) mechanisms of transposition. Unlike retrotransposons, which synthesize their DNA copies by using their own RNA-dependent DNA polymerase (reverse transcriptase), DNA transposons cannot synthesize DNA. Instead, they multiply by using the host replication machinery. A typical autonomous mariner (9), hAT (10), piggyBac (11), P (12), Merlin (13), or Transib (14) DNA transposon encodes only a single protein called transposase, which acts as an endonuclease and catalyzes transfer of transposon DNA strands from one genomic site to another. In the En/Spm (10), MuDR (15), Harbinger (16), and Helitron (8) superfamilies, an autonomous transposon usually encodes a TPase and one DNA-binding protein.

Here we report a third class of DNA transposons called Polintons that are widespread in protists, fungi, and animals, including entamoeba, trichomonas, soybean rust, sea urchin, sea anemone, sea squirt, fishes, chicken, lizard, frog, insects, and worms. Autonomous Polintons are typically 15–20 kb long and encode up to 10 different proteins, including DNA polymerase B (POLB), retroviral-like integrase, adenosiviral-like protease (PRO), and putative ATPase (ATP). They are the most complex DNA transposons in eukaryotes. Based on reported structural and evolutionary characteristics, we propose a model of Polinton transposition. We discuss implications of our findings, including likely origin of Polintons from a linear plasmid and evolution of adenoviruses from an ancient Polinton.

Results and Discussion

Polintons in Vertebrates, Tunicates, and Echinoderms. Analysis of the recently reported eukaryotic gene family related to retroviral integrases called c-integrases (17), led us to identification of two families of 360- and 369-aa zebrafish c-integrases (INT-1,DR and INT-2,DR). Each family is composed of several copies dispersed in the genome and characterized by 98% intrafamily and 72% interfamily sequence identities. After expansions of the integrate-encoding regions and multiple alignments of the expanded sequences, we derived 18,485-bp and 16,276-bp consensus sequences (called Polinton-1,DR and Polinton-2,DR). Both elements show hallmarks of similarly structured TEs: (i) 344- and 355-bp terminal inverted repeats (TIRs) and (ii) each individual clone is flanked by different 6-bp target-site duplications (TSDs). Although the internal portions of the consensus sequences (958–12,713 in Polinton-1,DR and 621–12665 in Polinton2,DR) are 65% identical to each other, the remaining terminal portions are different. Given that both Polinton families include several

Conflict of interest statement: No conflicts declared.

Abbreviations: INT, integrase; POLB, family B DNA polymerase; PRO, protease; TE, transposable element; TIR, terminal inverted repeat; TPase, transposase; TSD, target-site duplication.

*To whom correspondence may be addressed. E-mail: vladimir@girinst.org or jurka@girinst.org.

© 2006 by The National Academy of Sciences of the USA
proteins are the same (Fig. 1; INT-PX-PW-PY-PRO-PZ-ATP-site). In both which is published as supporting information on the PNAS web 2

In addition to fish and amphibians, we identified one set of seven intronless ORFs coding for DNA polymerase B (1383-aa POLB-1 DR and 1392-aa POLB-2 DR, see DNA polymerase), adenoviral protease (178-aa PRO-1 DR and PRO-2 DR; see Cysteine protease), ATPase (222-aa ATP-1 DR and 231-aa ATP-2 DR; see ATPase in Polintons), and four unclassified proteins (280-aa PX-1 DR and 282-aa PX-2 DR, 435-aa PW-1 DR and PW-2 DR, 151-aa PW-1 DR and 156-aa PW-2 DR, and 256-aa PZ-1 DR and 260-aa PZ-2 DR; see Table 1, which is published as supporting information on the PNAS web site). In both Polintons, order and orientation of all eight proteins are the same (Fig. 1; INT-PX-PW-PY-PRO-PZ-ATP-POLB; proteins encoded by the second strand are underlined). The same orientation and identity of these eight proteins was also observed in other fish genomes, including fugu and pufferfish (data not shown). However, we did not find Polintons in invertebrate DNA sequences, we identified various 5- to 15-kb

In addition to fish and amphibians, Polintons were also found in reptiles and birds. We identified one Polinton-1 SPU in the tuatara (Sphenodon punctatus) genome (GenBank accession no. AC153757, position 112,530–100,591). It is flanked by the ATGGCA 6-bp TSD, has ~800-bp TIRs, and contains remnants of the PY-PRO-PZ-ATP-POLB coding block with order and orientation identical to those in the zebrafish Polintons (Fig. 1). Because the coding regions contain several stop codons, the identified immobile copy of Polinton-1 SPU was inactivated by mutations after its insertion. The missing INT-PX-PRO block likely has been deleted from the originally intact element. In the chicken genome, we also detected mutated remnants of Polintons (data not shown). However, we did not find Polintons in mammalian genomes.

Polintons also populate genomes of tunicates and echinoderms. We identified two Polintons in the sea squirt (Ciona intestinalis) genome: 15,061-bp Polinton-1 CI (scaffold.157, position 54,569–39509; TSD: CACAAG) and 13,695-bp Polinton-2 CI (scaffold.257, position 48,367–62,061; TSD: CTCGAC). Although these elements represent two distinct families (they are ~65% identical to each other), they are characterized by the same unusual order of the Polinton proteins: PRO-POLB-ATP-PZ-INT-PX-PW-PY (Fig. 1). We identified four families of full-sized Polintons in the echinoderm sea urchin (Strongylocentrotus purpuratus) genome with order and orientation of all eight proteins resembling those in fish Polintons (Fig. 1). None of the sea urchin Polintons is flanked by target site duplications. However, given the small number of copies of these elements and the preliminary assembly of the genome from ~1,000-bp short shotgun sequences, we cannot rule out a misassembly of highly identical Polinton copies, which could prevent us from observing the TSDs.

Polintons in invertebrates. Using the protein sequences encoded by vertebrate Polintons as seeds in TBLASTN searches against invertebrate DNA sequences, we identified various 5- to 15-kb
regions coding for the Polinton proteins. Based on studies of these regions and their flanks, we identified them as internal parts of Polintons with 200- to 1200-bp TIRs and 6-bp TSDs. We derived consensus sequences of complete Polintons in the nematode (16,653-bp Polinton-1.LCB, Caenorhabditis briggsae), red flour beetle (13,486-bp Polinton-1.TC, Tribolium castaneum), and fruit fly (14,782-bp Polinton-1.DY, Drosophila yakuba).

The structure of Polintons in vertebrates, tunicates, and echi-

noderms is notably closer to those in insects than in nematodes (Fig. 1). Whereas insect Polintons encode the same set of eight proteins found in the vertebrate TEs, although in a different insect-specific order and orientation (ATP-PZ-INT-PX-PW-

PY-PRO-POLB), the nematode Polintons code for nematode-

specific PCI and PC2 proteins, instead of PX, PZ and PW, and are characterized by a unique order and orientation of all proteins: INT-PY-PCI-ATP-POLB-PC2-PRO (Fig. 1).

Based on TBLASTN searches against Trace archives, we also found that the Polinton-specific proteins are encoded in multiple copies by the cnidarian genome (sea anemone Nematostella vectensis) ancestral to insects, worms, and vertebrates.

Polintons in Protists and Fungi. Although Polintons were not found in plants, we identified them in several other eukaryotic kingdoms, including fungi (soybean rust Phakospora pachyrhizi and Glomus intraradices) and such protists as parabasilids (Trichomonas vaginalis), entamoeba (Entamoeba invadens, E. histolytica, and E. dispar), and heterokonts (Phytophthora infestans).

The soybean rust genome is not assembled, but we identified numerous ~1-kb shotgun sequences encoding Polinton proteins. After analysis of short fragments encoding parts of two different proteins, we conclude that the soybean rust Polintons are characterized by a standard set of eight Polinton proteins: [POLB-ATP], [PX-PW-PY-PRO], and [PZ-INT], where three groups of proteins with known order and orientation are delimited by brackets. Even this incomplete information indicates that the order and orientation of these proteins is unique, although very similar to those in vertebrate and insect Polintons (Fig. 1). We detected one copy of Polinton-1.GI encoding INT and POLB in the G. intraradices fungal genome (GenBank accession no. AC163889, position 5,506–17,459; the CACCTT TSD).

In the T. vaginalis genome, we found a very abundant Polinton-1.TV family characterized by a 20,724-bp consensus sequence (Fig. 1). This family, together with its nonautonomous deriva-

tives, constitutes ~5% of the genome, which makes T. vaginalis an incubator of Polintons, unlike the other genomes where Polintons constitute only a minor component (~<0.2%; data not shown). General properties of Polinton-1.TV are the same as observed in metazoan Polintons: 6-bp TSDs and 160-bp TIR. Although the Polinton-1.TV encodes POLB and INT similar to those in metazoan Polintons, it encodes two different ATPases (ATP and ATP1 in Fig. 1), which are not significantly similar to ATPases from the metazoan Polintons (BLAST E values >1). Excluding POLB, INT, and ATPases, T. vaginalis Polintons do not code for other proteins found in the metazoan TEs. However, we encode six additional proteins, called PTVP1–PTVP6 (Fig. 1). Only PTVP2 is similar to other known proteins; it contains a 200-aa domain similar to the C-terminal domain of structural proteins involved in assembly of phage tails (Table 1).

The entamoeba genome is the only other sequenced genome coding for PTVPs. This protein is encoded by Polinton-1.EI TEs in E. invadens. Although the E. invadens genome was not assembled, we reconstructed the 16,504-bp Polinton-1.EI consensus sequence from ~1,000-bp short shotgun sequences. This transposon is characterized by the 597-bp TIR and the POLB-ATP1-ATP-PVT6-INT pattern of its proteins.

A 18,398-bp single copy of Polinton-1.PI found in P. infestans is characterized by the POLB-INT-ATP-PRO pattern, 6-bp TSD, and 113-bp TIRs.

The Tetrahymena thermophila ciliate genome also contains Polintons. The genome is not assembled, and we identified several 2- to 3-kb whole genome shotgun sequences coding for Polinton INT and POLB (GenBank accession nos. AAGF01001397, AAGF01001309, and AAGF01001471). INTs encoded by these sequences are ~50% identical to INT encoded by TIR transposable elements (18). Moreover, some copies of TIR encode ATP (18), which is similar to the Polinton ATP. TIR is characterized by 6-bp TSDs, >100-bp TIRs with the 5’-AGAGA terminus similar to that conserved in Polintons (ref. 18; Fig. 2). Therefore, TIR elements are nonautonomous derivatives of Polintons.

Nonautonomous Polintons. The best examples of nonautonomous Polintons were found in zebrafish. For instance, a Polinton-1.NI.DR family consists of ~20 recently transposed TEs with ~97% pairwise identity, including five 11- to 17-kb copies 98% identical to each other and flanked by different 6-bp TSDs (AL953910, position 5,323–19,910; BX897696, 116,625–128,256; AL935200, 105,042–120,037; BX005307, 157,480–142,142; BX842670, 125,298–109,851; BX294179, 10.1073/pnas.0600833103 Kapitonov and Jurka
dromes by the host replication machinery, after accidental DNA annealing triggered by single-stranded breaks at the stem-loop border (Fig. 5B). Given presence of nonautonomous Polintons with long TIRs in the sea squirt (chr08q, position 910,025–916,953; 2796-bp TIR), and sea urchin genomes (Spur20030922-genome: Contig884, 484–9,563; 3,900-bp TIR), the proposed “palindromization” of internal portions of Polintons may be quite common and may represent a more general model for palindromization of stem-loop structures.

Structural Hallmarks of TIRs. All identified Polintons are characterized by the universally conserved 5’-AG and CT-3’, in addition to already described TIRs that are several hundred nucleotides long, 6-bp TSDs, and specific sets of proteins, including universally present POLB, INT, and ATP (Fig. 2). This specificity of termini is linked to a structural similarity between the Polinton and retroviral integrases (17). Moreover, LTR retrotransposons that belong to the group I of the Gypsy superfamily are also characterized by the conserved 5’-AG and YT-3’ termini (19). Although there is no significant sequence identity between TIRs from Polintons found in different species, all of them appear to share short simple microsatellite-like terminal repetitions. The 5’- (AGA), 5’- (AGC), and 5’- (AGA)2 microsatellites represent less common types of terminal repetitions (Fig. 2).

DNA Polymerase B. Each autonomous Polinton encodes a 1,000- to 1,400-aa protein whose ~700-aa C-terminal portion is significantly similar to various protein-prime DNA polymerases that belong to the B family of DNA-dependent DNA polymerases.

Protein-prime POLBs constitute a distinctive group of eukaryotic and prokaryotic DNA polymerases encoded by various phages, vertebrate adenoviruses, and linear plasmids from plant and fungal mitochondria. These polymerases display both 3’ to 5’ exonucleolytic and 5’ to 3’ synthetic activities defined by two structurally independent N- and C-terminal domains (20). After inspection of a multiple alignment of the Polinton and POLB polymerases, we found that all known motifs conserved in the protein-prime POLB polymerases are also conserved in the Polinton polymerases (Fig. 6, which is published as supporting information on the PNAS web site). These motifs include the Exo I (defined by the DXE consensus), Exo II (Nx3F/YD), and Exo III (Yx3D) motifs that constitute a catalytic core of the exonuclease conserved in all proofreading POLBs (21–23). The catalytic core of VSR consists of Asp-51 and His-69 (26, 27), which are also conserved in the Polinton POLBs (Fig. 6).

Although all Polinton POLBs cocluster with protein-prime POLBs, they constitute two distinctive clades (Polinton POLB1 and POLB2; Fig. 3). Based on the protein identity, members of the same clade are closer to each other than to members of the other clade. The POLB1 clade includes polymerases encoded by the vertebrate, sea urchin, sea squirt, sea anemone, and some insect (Polinton-1-TC) Polintons. The POLB2 clade includes proteases encoded by the insect, nematode, fungus, and protozoan transposons.

All POLB1s contain a unique ~145-aa insertion at the same position between the conserved Exo III and “YxGG” motifs (Fig. 6). The insertion position is ~15 bp upstream of the YxGG motif (the exact position is not clear because of low identities of POLB1s to other POLBs). The insertion sequence is significantly similar to bacterial ~140-aa “very-short-patch” DNA repair endonucleases (VSR), which initiate nuclease excision repair of G:T mismatches introduced by deamination of 5’-methylcytosines (25). In PSI-BLAST searches against GenBank proteins concatenated with 15 different POLB1 insertions, the VSR proteins were the only proteins, excluding the POLB1 insertions, which were similar to a profile derived from the POLB1 insertions (E < 10–5). The catalytic core of VSR consists of Asp-51 and His-69 (26, 27), which are also conserved in the POLB1 insertions (Fig. 6). Although most members of the POLB2 clade do not contain any long insertions between the Exo III and YxGG motifs, we identified several families of Polintons in different species that contain an ~140-aa specific insertion at a position nearly identical to that in POLB1. Although there is no sequence similarity between these two insertions, the POLB2 insertion is similar to HNH homing nucleases (data not shown). Presence of the VSR and HNH nucleases inserted at the same position between the exonuclease and polymerase domains of the Polinton POLBs appears to be very unusual. We did not find any other POLBs with insertions of the nucleases, except for several known bacterial POLBs containing the homing nuclease...
encoded by inteins: analogues of “amino acid self-spliced” inteins. In eukaryotes, only a few inteins have been found, in fungi. Although termini of the POLB1 and POLB2 insertions are not perfectly defined and not similar to conserved termini of known bacterial inteins, one cannot rule out an intein-like nature of the POLB1 and POLB2 insertions.

**Adenoviral Protease.** In addition to the conserved POLB and INT, Polintons encode an ≈170-aa PRO protein (Fig. 1; see also Fig. 7, which is published as supporting information on the PNAS web site) that is significantly similar to proteases in adenoviruses. The latter, also known as adenain or AVP, belong to a superfamily of cysteine proteases and is characterized by the His-Asn/Glu-Cys catalytic triad (28–31). In addition to the catalytic residues, Gln separated by six amino acids from the catalytic Cys, was also reported as a universally conserved amino acid that is necessary for the protease activity (28, 29). The catalytic triad and Gln are also conserved in the Polinton proteases (Fig. 7).

Although Polintons in protozoans do not encode the PRO cysteine protease, it is conserved in metazoan and fungal Polintons. These observations indicate that PRO is necessary for survival of Polintons, which have acquired it in a common ancestor of fungi and animals, after its split from protozoans.

**ATPase.** Most Polintons code also for a ≈200-aa ATP protein (Fig. 1; see also Fig. 8, which is published as supporting information on the PNAS web site) characterized by the Walker A and B motifs conserved in ATPases (32, 33). The Polinton ATPase is most similar to a plethora of hypothetical ATPases from dsDNA viruses. Many of these hypothetical ATPases were mistakenly annotated in GenBank as virion-packaging proteins. However, none of these Polinton-like ATPases contain any C-terminal nuclease domain that is present in real virion packaging proteins (34). The Polinton ATPase probably facilitates DNA synthesis by POLB.

**Mechanism of Transposition.** Given known distant similarities between retroviral INTs and cut-and-paste Tpases, one can expect cut-and-paste transpositions of Polintons catalyzed by their INT. However, some arguments listed below strongly suggest that transposition of Polintons follows a completely different mechanism unseen previously in transposons. First, a perfect conservation of all functional motifs in the extremely diverged POLBs indicates that the DNA-DNA polymerase and proofreading activities are necessary for transposition of Polintons. Second, POLB in Polintons belongs to the group of DNA polymerases that use proteins as primers. This group is composed of polymerases encoded by bacteriophages, linear plasmids, and adenoviruses. Third, all genomes of these objects and Polintons are characterized by TIRs that are usually several hundred base pairs long. Fourth, termini of these genomes and Polintons are composed of short 1- to 3-bp tandem repeats (Fig. 2; figure 6 in ref. 35), which are thought to be necessary for the slide-back mechanism in protein-primed DNA synthesis (35).

Based on all these arguments, we propose that Polintons form a previously uncharacterized class of DNA transposons propagated through protein-primed self-synthesis by their polymerase, according to the model outlined in Fig. 4. First, during host genome replication, the integrase-catalyzed excision of a Polinton element from the host DNA leads to an extrachromosomal single-stranded Polinton that forms a racket-like structure (Fig. 4 A–C). Second, the Polinton POLB replicates the extrachromosomal Polinton (Fig. 4 D–F). Given the arguments listed above, initiation of the replication requires the terminal protein (TP) that binds a free 5′ end of Polinton. It is believed that N-terminal domains of proteins, whose C-terminal parts serve as POLB, encode TP in some linear plasmids (36). Therefore, it is likely the N-terminal 400- to 600-aa domain of the Polinton POLB serves also as TP. After the double-stranded Polinton is synthesized, the INT molecules bind its termini and catalyze its integration in the host genome (Fig. 4 G and H).

**Evolution of Polintons.** Polintons are present in genomes of species that belong to diverse eukaryotic kingdoms, including opisthokonts (metazoa and fungi), heterokonts (oomycetes), alveolates (ciliates), amoebozoa (entamoeba), and parabasalids. Given the conserved complex structure of Polintons, their monophyletic origin is most likely. Although Polintons are much more complex (up to eight conserved proteins) than known eukaryotic TEs and resemble viruses (adenoviruses and BmDNV-2), we did not find any Polinton protein similar to viral capsid or envelope proteins, which are necessary for the infectious transmission of viruses. Moreover, we are not aware of any viruses capable of spreading over different kingdoms. Most likely, Polintons emerged in a common ancestor of modern species from the eukaryotic crown ≈1 billion years ago. As we reported here, Polintons share their main structural characteristics with “selfish” linear plasmids, bacteriophages, and adenoviruses that multiply by using their protein-primed DNA polymerases. Linear plasmids can be split into two groups: (i) plasmids that exist in mitochondria of plants and fungi and (ii), plasmids that exist in the yeast cytoplasm (37). Although it is likely that mitochondrial linear plasmids evolved from bacteriophages during the evolution of mitochondria from bacteria, understanding the evolution of cytoplasmic plasmids is hampered by different equally plausible scenarios (38, 39). Although Polintons represent a previously unknown link between cytoplasmic plasmids/adenoviruses and mitochondrial plasmids/bacteriophages (Fig. 3), many aspects of evolution of Polintons and cytoplasmic linear plasmids remain unclear. Acquisition of the integrase by a protein-primed replicating genome of an ancient virus or linear plasmid was the most certain stage of the evolution. It has been suggested that the Polinton INT evolved from an INT encoded by an LTR retrotransposon (17). Thus, it might have been acquired after integration of an ancient LTR retrotransposon into the ancestral linear genome. However, we cannot rule out the origin of the Polinton INT from a DNA transposon. For instance, the Tdd-4 transposon from the slime mould Dictyostelium discoideum genome is a DNA transposon characterized by...
its 145-bp TIRs, 5-bp TSDs, and a TPase that is similar to INTs encoded by LTR retrotransposons (40).

Although both clades of the Polinton polymerase are significantly but distinctly coclustered with adenoviruses and cytoplastmics from nematode, fungus, fruit fly, and beetle are closest to the POLB encoded by the BmDNV-2 Bombyx mori densonucleosis virus (Fig. 3). This unique virus encodes several structural virion-related proteins, and its POLB was grouped with the adenoviral POLB (21). It remains to be shown whether the BmDNV-2 virus has evolved from a Polinton or from a virus related to Polintons and adenoviruses.

Polintons are characterized by a highly patchy distribution in different species. In insects, Polintons are present in flies and beetles but absent in mosquitoes and bees. In fungi, they are present in basidiozymates (soybean rust) and glomeromycetes (G. intraradices) but absent in ascomyces (including the completely sequenced yeasts genome). We interpret this patchiness as a frequent loss of Polintons from genomes. Because of the high complexity of Polintons, their transposition may be tightly regulated and may explain their small numbers in most genomes.

Polintons were first reported in preparation, Feschotte and Pritham (17) reported that c-integrases in the zebrafish and nematode genomes are encoded by long TEs, called Maverick, characterized by long TIRs and 6-bp TSDs (41). Moreover, Wuitschick et al. reported in 2002 (18) that the T. thermophila genome contains Tr transposons characterized by the same Maverick-like properties.

Because all these authors have not reported the basic enzymatic machinery or mechanism of transposition of TIR/Maverick TEs, we introduce “Polinton” as a general name of all eukaryotic self-synthesizing DNA transposons.

Materials and Methods

All TEs reported in this work were identified and characterized by using various methods of computational analysis described in our previous papers (8, 14, 16, 19, 42). Significant similarities between distantly related proteins were identified by using PSI-BLAST (43). Multiple alignments of protein sequences were created by using PROBCONS and T_COFFEE (44, 45). Phylogenetic analysis was performed by using MrBAYES 3.0 (46) with the following settings. Rate variation across sites was modeled by using a gamma distribution, with a proportion of sites being invariant (rates = “invgamma”; “mixed” amino acid models). The Markov chain Monte Carlo search was run with four chains for 5 × 10^7 generations, with trees begin sampled every 1,000 generations (the first 1,000 trees were discarded as “burnin”). The MrBAYES input multiple alignment included only the POLB conserved motifs (~400 amino acids, available upon request). The sequences of TEs reported in this work are deposited in Repbase Update (47).

We thank Adam Pavlick and Andrew Gentles for discussions, anonymous reviewers and Margaret Kidwell for helpful suggestions, and Oleksiy Kohany for assistance with computational analysis. This work was supported by National Institutes of Health Grant P41 LM006252-08.