Nucleic Acids Research, 1993, Vol. 21, No. 9 2117–2123

Pao, a highly divergent retrotransposable element from *Bombyx mori* containing long terminal repeats with tandem copies of the putative R region

Yue Xiong+, William D.Burke and Thomas H.Eickbush*
Department of Biology, University of Rochester, Rochester, NY 14627, USA

Received January 28, 1993; Revised and Accepted March 15, 1993 GenBank accession no. L09635

ABSTRACT

Analysis of aberrant ribosomal DNA (rDNA) repeats of *Bombyx mori* resulted in the discovery of a 4.8 kilobase retrotransposable element, *Pao*. Approximately 40 copies of *Pao* are present in the genome with most located outside the rDNA units. The complete sequence of one *Pao* element and partial sequence of four other copies indicated that *Pao* encodes an 1158 amino acid open-reading frame (ORF). Located within this ORF are domains with sequence similarity to retroviral *gag* genes, aspartic protease and reverse transcriptase. RNase H and integrase domains were not identified suggesting that the cloned copies were not full-length elements. *Pao* elements contain long terminal repeats (LTRs) with a central region composed of variable numbers of 46 bp tandem repeats. The variable region appears to correspond to the R region of retroviral LTRs, the region responsible for strand transfer during reverse transcription. Based on a sequence analysis of its reverse transcriptase domain, *Pao* is most similar to *TAS* of *Ascaris lumbricoides*. *Pao* and *TAS* represent a subgroup of LTR retrotransposons distinct from the *Copia-Ty1* and *Gypsy-Ty3* subgroups.

INTRODUCTION

Over 50 retrotransposable elements have now been identified in higher plants and animals, protozoans and fungi. These retrotransposable elements can be divided into two major classes based on the structural features of their genomes, mechanisms of integration, and phylogenetic analysis using their reverse transcriptase sequences (1–5). One class of retrotransposable elements is structurally and functionally analogous to the integrated form of retroviruses. Members of this class are flanked by long terminal direct repeats (LTRs) and encode proteins including nucleic acid binding proteins (*gag*-like proteins), aspartic protease, reverse transcriptase, RNase H and integrase enzymes that are similar to those encoded by retroviruses (6,7). The second class of retrotransposable elements is less related to retroviruses both structurally and functionally. They do not contain LTRs and most do not encode enzymes similar to retroviral protease, RNase H and integrase (8–11). Reverse transcription and integration of the non-LTR retrotransposable elements appear to be different from that of retroviruses (12–14).

Previous studies from our laboratory have reported retrotransposable elements inserted in the ribosomal DNA (rDNA) units of *Bombyx mori* (15,16). *R1* and *R2* were found inserted in specific sequences of the 28S ribosomal RNA genes (17,18). *R1* and *R2* are highly adapted for sequence-specific integration into these 28S gene sites (19–21). A third retrotransposable element, *Dong*, was found in the non-transcribed spacer region of an rDNA unit (22). Most copies of *Dong* are located outside the rDNA units. *R1*, *R2* and *Dong* are members of the non-LTR class of retrotransposable elements. In this report we describe a fourth retrotransposable element originally found inserted into the rDNA repeats of *B. mori*. The element has been named *Pao*, which is a Chinese word for 'running'. *Pao* is a member of the LTR class of elements, however the LTRs of *Pao* elements are unusual in that they contain a tandemly repeated segment. Based on the sequence of its reverse transcriptase, *Pao* does not belong to either of the previously identified subgroups of LTR retrotransposable elements, the *Copia-Ty1* or *Gypsy-Ty3* groups.

MATERIALS AND METHODS

Lambda clones B98 and B113 were isolated from a charon 4 library of partially *Eco*R1-digested fragments of *B. mori* DNA (15). A 2.4 kb *Eco*R1 fragment from the Pao element in clone B113 was used as hybridization probe to isolate lambda clones P101, P103, P104, P105 and P106 from a charon 35 library made of partially *Sau*3A-digested genomic DNA (23). Lambda genomic library screens and Southern analyses were conducted at high hybridization stringency (68°C, 0.3 M NaCl with final washes of the filters at 68°C, 0.015 M NaCl) as previously described (24). Appropriate restriction fragments containing the *Pao* element from each lambda clone were subcloned in pUC18 plasmids. Detailed restriction maps of the subclones were constructed, and specific restriction fragments were recloned into

* To whom correspondence should be addressed

+ Present address: PO Box 100, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA

M13mp18/19 vectors for sequencing (25). Polymerase chain reaction (PCR) amplification from genomic DNA of Pao internal segments was conducted with one primer (5'-TTACGTAGACG-ACTACTTGG-3') complementary to the sequence encoding the YVDD conserved motif of the reverse transcriptase domain, and the second primer (5'-AACCCTATTTCGTCCTGCAC-3') complementary to a segment starting within the right LTR. The PCR products were cloned into ddT-tailed M13mp18 vector for sequence analysis (26).

DNA and protein sequence analysis were analyzed with the help of the computer program package, MacVector, available through International Biotechnologies, Inc. A search of GenEMBL data base for potential tRNA primers was conducted with the GCG program WordSearch on the University of Rochester VAX computer system. Procedures for amino acid sequence alignment of the reverse transcriptase domain and phylogenetic tree construction using the Neighbor-Joining method (27) were done as previously described (3,5).

RESULTS

Cloning of Pao elements

We have previously reported a series of lambda clones containing *B. mori* rDNA units with DNA insertions in either the 28S gene or within the non-transcribed spacer region (15,16). One of the clones, B98, which contained part of an *R2* element inserted in the 28S gene (17) also contained part of a second putative mobile element inserted within the non-transcribed spacer region of the rDNA unit (Figure 1). It was postulated (15) that the other end of this element might correspond to the insertion found in the lambda clone, B113, because both insertions were located between the *Kpn*I and *Xba*I restriction sites of the rDNA spacer (Figure 1). To determine whether these insertions represented a single mobile element, we conducted a series of hybridizations with genomic DNA isolated from strain 703 (the strain used in the construction of the library) and a series of geographical races of *B. mori* (28). Using insertion fragments from B113 as probe, it was found that the insertions in B113 and B98 did represent the two ends of a single DNA element, which we have named *Pao*. Strain 703 contained approximately 40 copies of the element with most copies located outside the rDNA units (data not shown). The genomic hybridization pattern varied in the different races of *B. mori* consistent with *Pao* being a mobile element.

To isolate a series of independent copies of *Pao*, the 2.4 kb *Eco*R1 fragment from B113 was used to screen a second lambda library (23). Of the five positive clones obtained (Figure 1A) P101 and P105 contained complete *Pao* elements, P103 and P106 contained 3' truncated elements, and P104 contained an element that was truncated by cloning into the lambda vector. Because 3' truncations are not usually associated with the integration of retrotransposable elements, the truncations in P103 and P106 are presumably the result of deletions or insertions subsequent to the *Pao* integration events. The P105 *Pao* element was located in the non-transcribed spacer region of an rDNA unit representing a complete version of the *Pao* fragments present in clones B98 and B113.

Nucleotide sequence of *Pao* elements

The complete sequence of the 4,791 bp *Pao* element from P105 and its flanking rDNA sequences is shown in Figure 2. The element was inserted between the C and D clusters of short direct repeats in the non-transcribed spacer region of the rDNA unit (29). The *Pao* element was inserted in a transcriptional orientation opposite to that of the rDNA units. The P105 *Pao* was flanked by a 634 bp terminal repeat that had many of the properties of a retroviral LTR (see below). The sequence also revealed a single 1158 aa ORF which contained a 200 aa region with similarity to reverse transcriptase. These features suggested *Pao* was a member of the LTR class of retrotransposable elements. To insure that we have determined the general features of *Pao* elements in *B. mori*, and have not been misled by idiosyncratic features

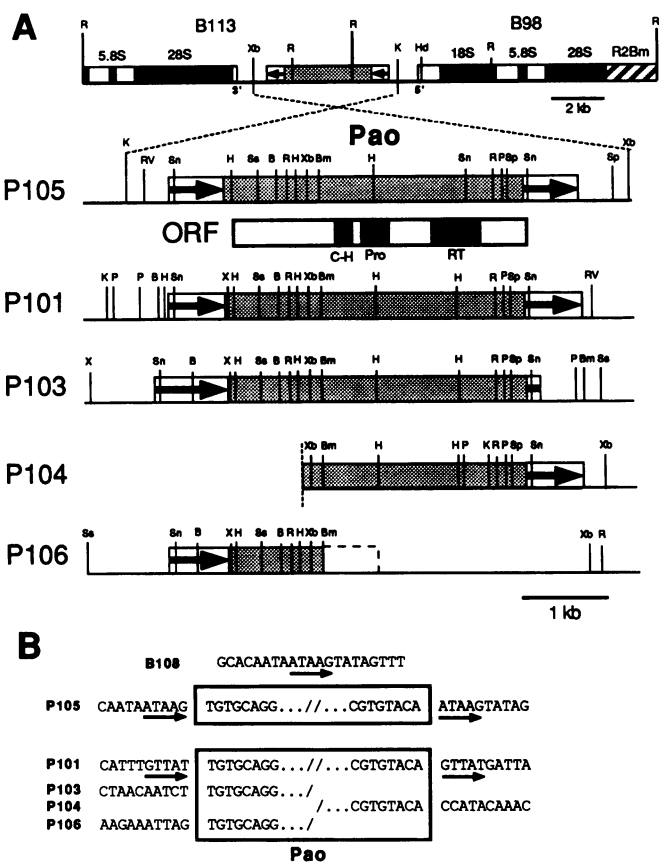

Figure 1. Summary of the cloned *Pao* elements from *B. mori*. (**A**) Organization and restriction map of the elements. Shown at the top is a schematic diagram of lambda clones B113 and B98 drawn as two consecutive rDNA units. Ribosomal RNA genes are indicated by solid boxes. The R2Bm element inserted in the 28S gene of B98 is shown as a diagonally shaded box. The *Pao* element located in the non-transcribed spacer region is shown as a stippled box (internal region) flanked by its LTRs (boxes with arrows). The transcriptional orientation of the *Pao* elements is opposite to that of the rDNA units. In the lower portion of the panel, more detailed restriction maps of the five cloned Pao elements are shown along with a small segment of flanking chromosomal DNA. Clone P105 contains a *Pao* element inserted in an rDNA unit, and is present in the opposite orientation to that shown in the upper portion of the figure. The Pao elements in clones P103 and P106 are truncated at their 3' ends. The exact location of this truncation in P106 is not known. The element in P104 has been truncated during the cloning process. The location of the 1158 aa ORF in the P105 *Pao* element is shown below the restriction map of the element. The location of the three identified domains in the ORF are indicated. C-H, three cysteine-histidine nucleic acids binding motifs, Pro, aspartic protease domain, RT, reverse transcriptase domain. Restriction enzymes: B, *Bgl*II; Bm, *Bam*HI; H, *Hin*cII; Hd, *Hin*dIII; K, *Kpn*I; P, *Pst*I; R, *Eco*RI; RV, *Eco*RV; Sn, *Sna*BI; Sp, *Sph*I; Ss, *Sst*I; X, *Xho*I; Xb, *Xba*I. (**B**) Junction sequences of the cloned *Pao* elements. Horizontal arrows, 5 bp target site duplications seen in P101 and P105. B108, sequence of the target DNA for the *Pao* insertion in P105, obtained from uninserted rDNA unit in clone B108 (16). Sequences within the boxes correspond to *Pao*.

of the element in P105, we have also determined most of the nucleotide sequence of the *Pao* element in P101 and the LTR regions from P103, P104 and P106 (Figures 3 and 4). Nucleotide sequence divergence between the internal region of *Pao* in P105 and P101 was 4.0% (134 nucleotide differences in 3360 bp). Nucleotide sequence divergence between the LTR regions from the five copies of *Pao* averages 4.8% (range 2.1%−6.6%). Sequence divergence between the left and right LTRs of the same *Pao* element was nearly absent (0.0% for P103 and P105, 0.4% for P101).

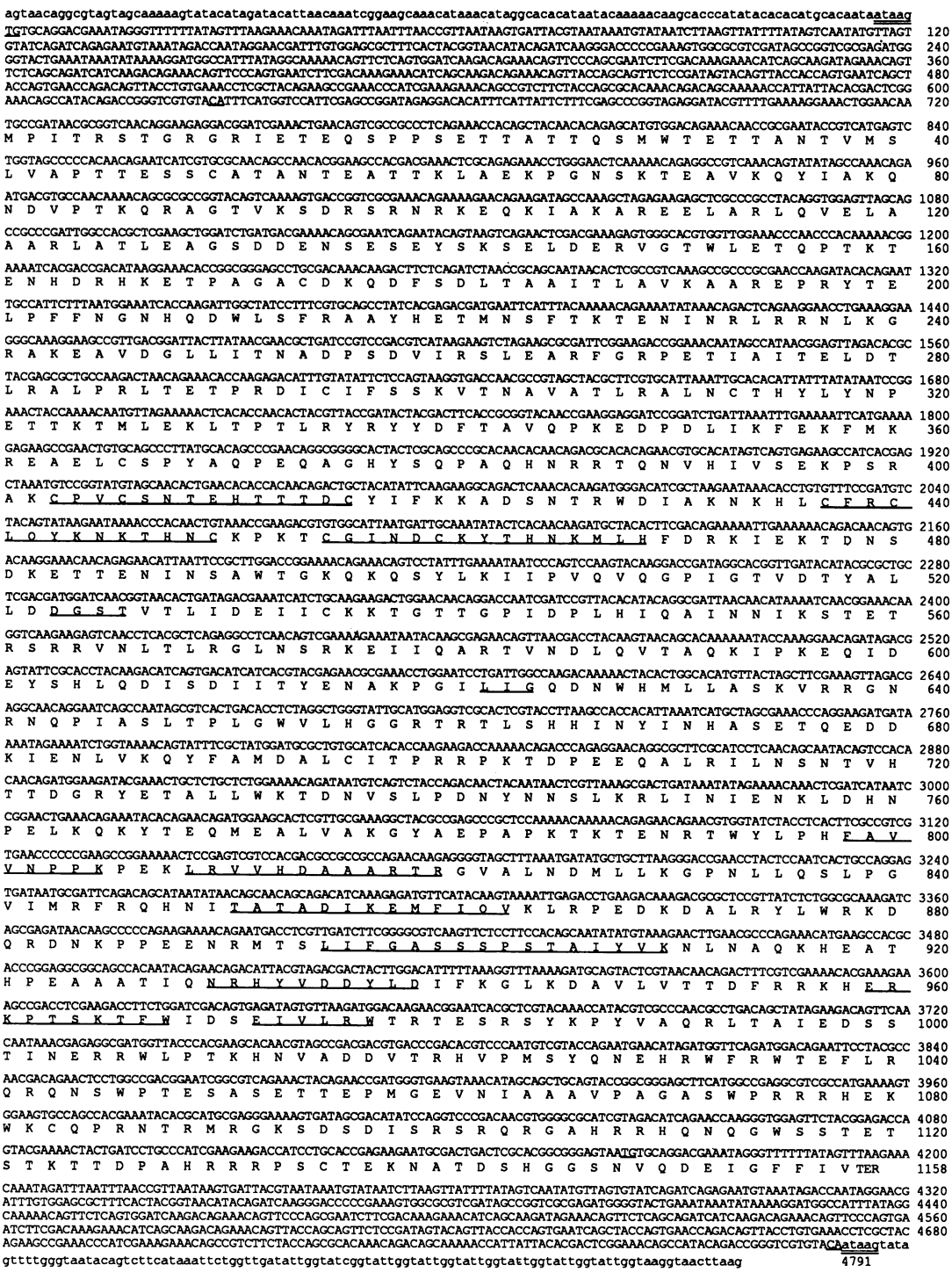

Figure 2. Complete nucleotide sequence of the *Pao* element from clone P105. Lower case letters correspond to flanking sequences from the non-transcribed spacer region of the rDNA unit with the target duplication double underlined. The LTRs begin and end with the underlined nucleotides TG and CA, respectively. The deduced protein coding sequence of the ORF is shown below the nucleotide sequence. This ORF ends 31 nucleotides within the downstream LTR. Underlined amino acid residues correspond to conserved sequence motifs identified in other retroelements: three cysteine-histidine motifs of *gag*-like genes (34), DGST and LIG conserved motifs of aspartic proteases (35), and the seven motifs detected in reverse transcriptase sequences (5).

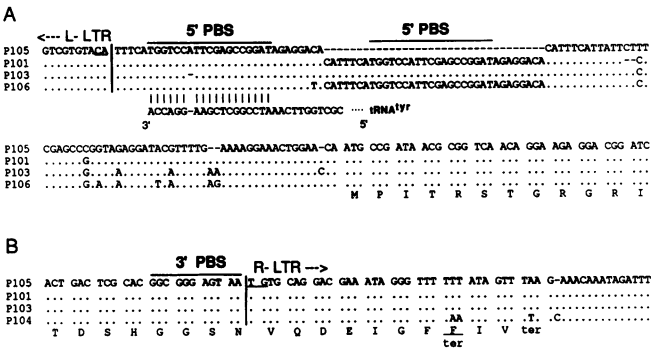

Figure 3. Nucleotide sequence variation at the beginning (panel A) and end (panel B) of the ORF in *Pao* elements. All sequences have been compared to the P105 sequence with identity to the P105 sequence indicated by a (.), and deletions of a nucleotide indicated with a (−). The two nucleotide differences within the Phe codon near the termination codon give rise to a new termination codon. Boundaries of the left (L) and right (R) LTRs are marked by a vertical line. Sequences corresponding to the putative 5′ and 3′ primer binding sites (PBS) are overlined. The 3′ end of the tRNA^{tyr} which can serve as a primer for first strand synthesis is shown below the 5′ PBS.

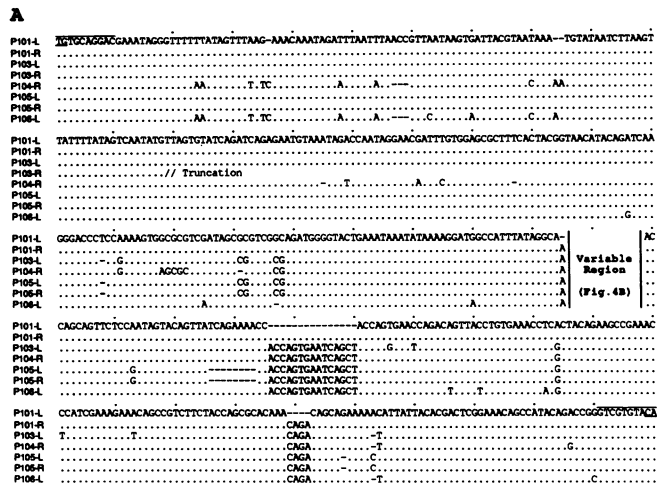

Figure 4. LTR sequences from cloned *Pao* elements. (**A**) Nucleotide sequences upstream and downstream of the variable region. All sequences have been compared to the left LTR of the *Pao* element in clone P101. An imperfect (8 of 10 bp) inverted repeat at the beginning and end of the LTR are overlined. The right LTR of the P103 Pao element is truncated after 119 bp. (**B**) Sequence of the variable region of the LTRs showing the 46 bp repeat. The variable segment of the left and right LTRs of clones P101 and P105 are identical, therefore only one sequence from each element is shown. Many of the 46 bp repeats contain internal deletions. Most of the pyrimidine residues in the RNA synonymous strand present are clustered in three regions which have been underlined.

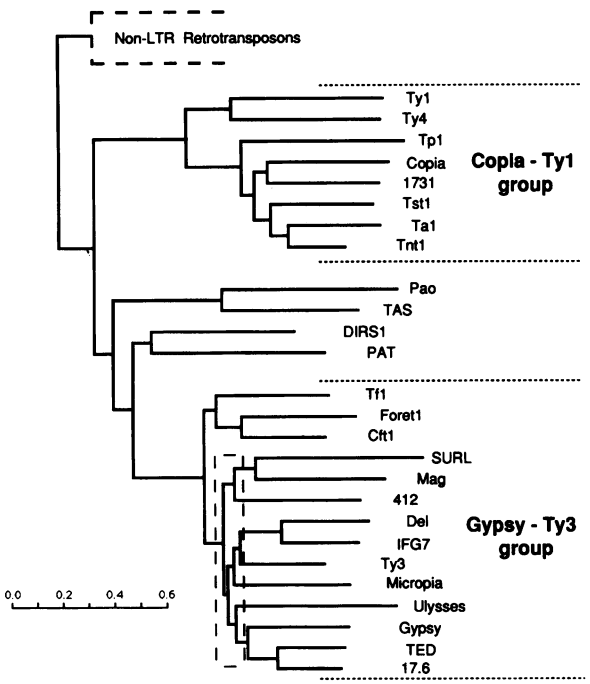

Figure 5. Phylogenetic tree of the LTR retrotransposable elements based on the amino acid sequences of their reverse transcriptase domain using the Neighbor Joining method of tree construction (27). The extent of the sequence divergence (shown as substitutions per amino acid position) are indicated by the horizontal branch lengths (vertical lengths have no significance). The tree is rooted on the branch leading to the non-LTR retrotransposable elements (5). Retrotransposable elements that correspond to the previously defined subgroups *Copia-Ty1* and *Gypsy-Ty3* are indicated by dotted lines. References to the published sequence of each element and the class of organism from which it was isolated can be found in the text.

The P105 *Pao* element was flanked by a duplication of the sequence ATAAG (Fig.1B). This sequence represented a 5 bp target site duplication based on the sequence of the uninserted rDNA repeat from clone B108 (15). An apparent 5 bp target site duplication was also seen flanking the *Pao* element in P101 (Fig. 1B). The nucleotide sequence of the junctions from P105 and P101, as well as the single junctions available from the *Pao* elements in clone P103, P104 and P106 suggested that *Pao* inserted preferentially into AT rich regions but with no other obvious sequence specificity. A number of retrotransposable elements are known to preferentially insert into AT-rich sequences (7).

The nucleotide sequence of *Pao* suggested that its mechanism of first and second strand DNA synthesis from an RNA transcript and its mechanism of integration was likely to be similar to that of retroviruses (30). First, immediately adjacent to the 5′ LTR of Pao (Figure 3A) was a region complementary to the first 18 nucleotides at the 3′ end of the tRNA^{tyr} found in *D.melanogaster* (31). It is likely that the *B. mori* tRNA^{tyr} is identical to the *D.melanogaster* tRNA^{tyr} in this 3′ region, because this sequence is conserved between insect and mammalian tyrosine tRNAs (32,33). In retroviruses, such tRNA complementary sequences, termed the 5′ primer binding site (5′ PBS), have been shown to be the site for the initiation of first-strand DNA synthesis (cDNA synthesis) from a bound tRNA molecule (30). Second,

Pao elements contained a polypurine tract 11 bp in length (9/11 nucleotides) adjacent to the 3′ LTR (Figure 3B). Second-strand synthesis in retroviruses is initiated from a similarly located polypurine tract (30). Finally, similar to the LTRs of retroviruses and most retrotransposable elements, the terminal repeats of Pao contained TG at their 5′ end and CA at their 3′ end (Figure 4A). These sequences were part of a larger, imperfect (8 of 10 nucleotides) inverted repeat located at the termini. Inverted repeats at the ends of retroviral LTRs are part of the *att* sites which are involved in the integration of the double stranded-DNA intermediate into host DNA (30).

The only feature of *Pao* LTRs that was not typically found in retroviruses and retrotransposable elements is shown in Figure 4. *Pao* LTRs could be divided into three sections: a 280−282 bp conserved region starting with the 5′ TG, a 183−201 bp conserved region ending with the 3′ CA, and a central variable region which was from 155 to 378 bp in length. The central region contained from four (P105) to ten (P103) tandem copies of a 46 bp repeat. The 46 bp repeats had an unusual separation of purine and pyrimidine residues. On the RNA synonymous strand shown in Figure 4B most of the purine residues were clustered in three regions each 5 nucleotides in length. Many copies of the 46 bp repeat contained internal deletions which appeared to be the result of recombinations between these purine rich regions.

The ORF of *Pao* elements

The P105 *Pao* element encoded a single ORF which occupied nearly the entire region between the LTRs. Based on the sequence of four independent copies of *Pao* the only significant sequence variation at the 5′ end of this ORF was the duplication in P101 and P106 of the 34 bp immediately adjacent to the left LTR (Figure 3A). This duplication included the 5′ PBS sequence. Other smaller insertions and deletions were present in the short region upstream of this first ATG, but segmental changes were completely absent in the 150 bp downstream of this codon, consistent with this codon being the initiation codon for the *Pao* ORF. Starting with this methionine codon the total length of the ORF in the P105 *Pao* element was 1158 amino acids. Surprisingly this ORF ended 31 bp into the downstream or right LTR (Figure 3B). Based again on the sequence of four independent *Pao* copies the only sequence variation detected in this region was two T to A transversions in P104 that generated a new termination codon three codons earlier than the one found in the other three copies of *Pao*.

The *Pao* ORF contained three regions with sequence similarity to functional domains found in the ORFs of retroviruses and LTR retrotransposable elements. The boundaries of these domains are shown in Figure 1 and the conserved amino acid motifs of the ORF are underlined in Figure 2. The first domain was composed of three putative cysteine and histidine nucleic acid binding motifs. The first of these motifs had the same spacing of cysteine-histidine residues, C-X_2-C-X_4-H-X_4-C, found in retroviral *gag* genes (34), while the two remaining motifs had a spacing of cysteine and histidine residues not previously noted in *gag* proteins. The second identifiable domain within the *Pao* ORF was that of an aspartic protease located immediately downstream of the *gag*-like domain. The conserved residues of this domain, DGST followed 95 residues latter by the sequence LIG, are again similar to those identified in retroviruses and LTR retrotransposable elements (35). The final identifiable domain in the *Pao* ORF was a reverse transcriptase domain, containing the seven motifs identified in all reverse transcriptase sequences (5).

Two domains associated with the *pol* gene of retroviruses and LTR retrotransposable elements could not be found in *Pao*. They are the RNase H domain located immediately downstream of the reverse transcriptase domain, and an integrase domain, located either upstream or downstream of the reverse transcriptase-RNase H domain (36). A number of LTR retrotransposons are known to encode a single ORF of approximately the same length as the *Pao* ORF (37−39). The *Pao* sequence was unusual however in that a large fraction of this ORF was located upstream of the nucleic acid binding motifs, such that only 680 residues were downstream of these motifs. Thus it was possible that the P105 copy of *Pao* represented a defective copy missing the region encoding the carboxyl end of the ORF. Such a deletion would also be present in the *Pao* elements present in clones B113, P101, P103 and P104 based on their identical restriction map to the P105 *Pao*. In an attempt to clone longer versions of the *Pao* element from *B. mori* the region encoding the carboxyl terminal end of *Pao* ORF was PCR amplified from genomic DNA (see Material and Methods). The PCR fragments generated were 0.65−0.7 kb in length (data not shown) consistent with that predicted from the P105 *Pao* element (nucleotide 3515−4183). Thus no evidence for a longer version of a *Pao* element was obtained by this amplification. Sequence analysis of six cloned PCR products revealed that they were all similar to the sequence in Figure 2. The copies exhibited on average 5.4% nucleotide divergence (range 0.3−8.5%).

Relationship of *Pao* to other retrotransposable elements

The only extensive region of sequence similarity between all retrotransposable elements is the amino acid sequence of the reverse transcriptase domain. We have previously used the amino acid sequence divergence of this domain to examine the phylogenetic relationship of the LTR and non-LTR retrotransposable elements, retroviruses, pararetroviruses and other reverse transcriptase-containing elements (3,5). Figure 5 presents an updated version of this analysis showing only the LTR retrotransposable elements, including eleven newly sequenced elements not available in our previous report (5). The analysis does not include similar elements from closely related species, or different families of the same element within a species. The analysis is based upon the amino acid divergences at 178 positions using the Neighbor-Joining method for tree construction (27). The tree has been rooted using the non-LTR retrotransposable elements (see ref. 5 for a discussion of the possible rooting of the reverse transcriptase tree). Analysis of this same data set using a maximum parsimony approach, PAUP (40), gave a similar tree to that in Figure 5 except for the order of the branches in the dashed box (W. Lathe, unpublished data).

Over half of the sequenced LTR retrotransposable elements (14 of 26) belong to the *Gypsy-Ty3* subgroup. Seven of these elements are from insects: *17.6* (41), *412* (42), *Gypsy* (43), *Micropia* (44), *Ulysses* (45), *TED* (46), and *Mag* (47). The remaining members of the *Gypsy-Ty3* group were obtained from fungi [*Ty3* (48), *Tf1* (49), *Cft1* (50) and *Foret*I (51)], plants [*Del* (52) and *IFG7* (53)], and a sea urchin [*SURL* (54)]. Three fungal elements (*Tf1*, *Foret1* and *Cft1*) are the most divergent members of the *Gypsy-Ty3* subgroup. The phylogenetic relationships of the members of this subgroup are not well defined, except for

the close relationship of the two plant elements (*Del* and *IFG7*) and the three insect elements on the *Gypsy* branch (*Gypsy*, *TED*, *17.6*). All elements within the *Gypsy-Ty3* subgroup contain *gag*-like and *pol*-like ORFs. The integrase domain is located within their second, *pol*-like ORF, downstream of the reverse transcriptase-RNase H domains. Members of the *Gypsy* branch (*Gypsy*, *TED* and *17.6*) contain a third ORF. This third ORF is similar in location to the *env* gene of retroviruses, and encodes a hydrophobic domain near the carboxyl terminus analogous to the transmembrane portion of retroviral *env* proteins (46).

The second distinct subgroup of retrotransposable elements is the *Copia-Ty1* group. The eight sequenced members of this group are again from widely different organisms including insects [*copia* (37) and *1731* (55)], plants [*Ta1* (38), *Tnt1* (39) and *Tst1* (56)], fungi [*Ty1* (57) and *Ty4* (58)] and a slime mold [*Tp1* (59)]. All members of this group encode their integrase domain upstream of the reverse transcriptase-RNase H domains. The insect and plant elements of this group are unusual in that their *gag*-like and *pol*-like ORFs are fused into a single ORF.

Based on the sequence divergence of their reverse transcriptase domains, the remaining four retrotransposable elements in Figure 5 are not members of either the *Gypsy-Ty3* or *Copia-Ty1* subgroups. These elements are *Pao*, *PAT* (60) and *TAS* (61) isolated from two nematode species and *DIRS1* (62) isolated from *Dictyostelium discoideum*. *Pao* is clearly most related to *TAS* from *Ascaris lumbricoides* and they represent a new subgroup. Additional examples of related retrotransposable elements are needed to determine if *TAS* and *DIRS1* are members of the same subgroup. It is interesting to note that both of these elements have highly unusual LTRs. *DIRS1* elements contain inverted LTRs (62), and *PAT* elements contain an unusual split LTR (61). Thus elements of this subgroup (or subgroups) have evolved somewhat modified methods of replication. The integrase domains in *PAT*, *TAS* and *DIRS1* are located downstream of the reverse transcriptase-RNase H domain suggesting a closer relationship to the *Gypsy-Ty3* subgroup. Based on the divergence of their reverse transcriptase sequences however, they appear to have separated from the *Gypsy-Ty3* group not long after the *Copia-Ty1* group separated from the *Gypsy-Ty3* group.

DISCUSSION

We have described the structure of *Pao*, a retrotransposable element from *B. mori* initially found inserted in the non-transcribed spacer region of an rDNA unit. This represents the fourth retrotransposable element to be discovered in *B. mori* by analysis of aberrant rDNA units (17,18,22). rDNA units are clearly a fruitful area to look for new classes of mobile elements, particularly in species where only limited numbers of genes have been cloned and few spontaneous mutations are available for analysis.

The ORF of the Pao element in clone P105 contained readily identifiable nucleic acid binding motifs similar to those of *gag* proteins, an aspartic protease domain, and a reverse transcriptase domain. However, the conserved amino acid motifs found in RNase H and integrase domains of all retroviruses and other LTR retrotransposons (5,35) could not be detected in this ORF. It is unlikely that we have simply failed to identify the RNase and integrase domain in the *Pao* ORF because of mutations rendering these motifs unrecognizable. The length of the *Pao* ORF downstream of the nucleic acid binding motifs (*gag*-like domain), was only 680 aa in length, considerably less than that found in other retrotransposable elements. Thus this ORF was not large enough to encode both an RNase H and integrase domain in addition to the reverse transcriptase domain. It is also unlikely that Pao elements do not utilize these enzymatic activities, because the LTRs and 5′ and 3′ PBS elements of Pao suggested that they have replicated by a retroviral-like mechanism. In addition, *TAS*, the element most related to *Pao*, contains both RNase H and integrase domains downstream of the reverse transcriptase domain (61). It is likely that all of the cloned copies of *Pao* analyzed in this report contained an internal deletion, presumably downstream of the reverse transcriptase domain, and thus were not autonomous full-length versions of this element. Our attempts to find longer length versions of *Pao* by PCR amplification from genomic DNA were unsuccessful. Thus *Pao* elements may no longer be able to transpose in the genome of *B. mori*. Alternatively, transposition could be possible if the missing functions are supplied *in trans* from another retrotransposable element family.

While not autonomous, the cloned *Pao* elements have clearly undergone retrotransposition to their current location, presumably through the utilization of enzymes supplied *in trans*. The level of sequence variation between the LTRs of different *Pao* elements averaged 4.8% and there was significant variation in length of the LTR central region. Sequence variation between the left and right LTR of the same *Pao* element averaged only 0.2% and their central regions were identical in length. The nearly identical LTR sequences within each *Pao* element indicates that these LTR sequences were regenerated at the duplication event which gave rise to the insertion.

The only unusual feature of the *Pao* LTR was a central region composed of 46 bp tandem repeats. The total length of the LTR varies from 620−840 bp for the four sequenced copies which was only slightly longer than that typically seen for LTRs of retroviruses and retrotransposable elements (30). This tandemly repeated region is probably equivalent to the R region of retroviral LTRs. R region sequences correspond to the beginning and end of the RNA transcript and are the sequences necessary for DNA strand transfer during reverse transcription. While the R region is fairly constant in length for any particular retrovirus, it varies considerably (from 10 to 230 bp) between different retroviruses (30). The 46 bp repeat in *Pao* contained an unusual separation of purine and pyrimidine residues. The number of repeats varied from 4 to 10 copies on the five cloned elements with many of the repeats containing internal deletions that appeared to be recombinations between the three pyrimidine rich regions (RNA synonymous strand). It seems likely that this variation in numbers of repeats and internal deletions occurred at the strand transfer step during the reverse transcription of the RNA template. Tandemly repeated sequences may aid the efficiency or rate of this strand transfer, but would also result in length variations between different copies. This method of generating length and sequence variation in the R region of the LTR has previously been suggested for murine intracisternal A-particle genes (63). An abundance of inverted repeats in the LTRs of the *Wis 2−1A* of wheat has also been suggested to be generated during strand transfer of the reverse transcriptase (64).

Finally, based on the sequence of its reverse transcriptase domain, *Pao* is not a member of the previously described *Copia-Ty1* and *Gypsy-Ty3* subgroups of LTR-containing retrotransposable elements. Pao is most closely related to *TAS* elements isolated from *A. lumbricoides* (61). *TAS* contains typical LTRs and encodes all the enzymatic domains typically associated

with the *gag*-like and *pol*-like genes of retrotransposable elements (61,65). The only other LTR retrotransposable elements that are not members of either the *Copia-Ty1* and *Gypsy-Ty3* subgroups are *DIRS1* from *D.discoideum* (62) and PAT from the nematode, *Panagrellus redivivus* (60). These four elements appear to be the first characterized members of one or more additional subgroups of the LTR retrotransposable elements. Based on their sequence divergences these additional subgroups are similar in age to the *Copia-Ty1* and *Gypsy-Ty3* subgroups (Figure 5). Because nearly all retrotransposable elements sequenced to date have been discovered by their insertions in a cloned gene, the members of these new subgroups are either much less abundant in eukaryotes, or retrotranspose less frequently. As more retrotransposable elements are characterized from a greater number of species, it will be interesting to learn whether these new subgroups are as widely distributed in eukaryotes as the *Copia-Ty1* and *Gypsy-Ty3* subgroups.

ACKNOWLEDGEMENTS

We thank Fritz Müller and Y.de Chastonay for communicating the sequence of the *PAT* and *TAS* elements before publication. This work was supported by grants GM31867 from NIH and NP-691 from ACS.

REFERENCES

1. Weiner,A.M., Deininger,P.L. and Efstratiadis,A. (1986) *Annual Rev. Biochem.* **55**, 631–661.
2. Boeke,J.D. and Corce,V.G. (1989) *Annual Rev. Microbiol.* **43**, 403–434.
3. Xiong,Y. and Eickbush,T.H. (1988) *Mol. Biol. Evol.* **5**, 675–690.
4. Doolittle,R.F., Feng,D.F., Johnson,M.S. and McClure,M.A. (1989) *The Quarterly Review of Biology* **64**, 1–30.
5. Xiong,Y. and Eickbush,T.H. (1990) *EMBO J.* **9**, 3353–3362.
6. Boeke,J.D. (1989) in Berg,D.E. and Howe,M.M. (eds) *Mobile DNA*, American Society for Microbiology, Washington DC, pp 335–374.
7. Sandmeyer,S.B., Hansen, L.J. and Chalker, D.L. (1990) *Annual Rev. Genet.* **24**, 491–518.
8. Eickbush,T.H. (1992) *New Biologist* **4**, 430–440.
9. Finnegan,D.J. (1989) in Berg,D.E. and Howe,M.M. (eds) *Mobile DNA*, American Society for Microbiology, Washington DC, pp 503–517.
10. Hutchison III,C.A., Hardies,S.C., Loeb,D.D., Shehee,W.R. and Edgell,M.H. (1989) in Berg,D.E. and Howe,M.M. (eds) *Mobile DNA*, American Society for Microbiology, Washington DC, pp 593–617.
11. McClure,M.A. (1991) *Mol. Biol. Evol.* **8**, 835–856.
12. Schwarz-Sommer,Z., Leclercq,L., Gobel,E. and Saedler,H. (1987) *EMBO J.* **6**,3873–3880.
13. Bucheton,A. (1990) *Trends in Genetics* **6**, 16–21.
14. Luan,D.D., Korman, M.H., Jakubczak, J.L. and Eickbush, T.H. (1993) *Cell* **72** (in press).
15. Lecanidou,R., Eickbush,T.H. and Kafatos,F.C. (1984) *Nucleic Acids Res.* **12**, 4703–4713.
16. Eickbush,T.H. and Robins, B. (1985) *EMBO J.* **4**, 2281–2285.
17. Burke,W.D., Calalang,C.C. and Eickbush,T.H. (1987) *Mol. and Cell. Biol.* **7**, 2221–2230.
18. Xiong,Y. and Eickbush,T.E. (1988) *Mol. and Cell. Biol.* **8**, 114–123.
19. Jakubczak,J.L., Burke,W.D. and Eickbush,T.H. (1991) *Proc. Natl. Acad. Sci. USA* **88**, 3295–3299.
20. Xiong,Y., Burke, W.D., Jakubczak, J.L. and Eickbush, T.H. (1988) *Nucleic Acids Res.* **16**, 10561–10573.
21. Xiong, Y. and Eickbush, T.H. (1988) *Cell* **55**, 235–246.
22. Xiong, Y. and Eickbush, T.H. (1993) Nucleic Acids Res. (in press).
23. Hibner,B.L., Burke, W.D. and Eickbush, T.H. (1991) *Genetics* **128**, 595–606.
24. Eickbush,T.H. and Kafatos,F.C. (1982) *Cell* **29**, 633–643.
25. Sanger,F., Nicklen,S. and Coulson,A.R. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 5463–5467.
26. Holton,T.A. and Graham, N.W. (1991) *Nucleic Acids Res.* **19**,1156.
27. Saitou,N. and Nei,M. (1987) *Mol. Biol. Evol.* **4**, 406–425.
28. Xiong,Y., Sakaguchi,B. and Eickbush,T.H. (1988) *Genetics* **120**, 221–231.
29. Fujiwara,H. and Ishikawa,H. (1987) *Nucleic Acids Res.* **15**, 6393–6401.
30. Varmus,H. and Brown,P. (1989) in Berg,D.E. and Howe,M.M. (eds) *Mobile DNA*, American Society for Microbiology, Washington DC, pp 53–108.
31. Suter,B., Altwegg,M., Choffat,Y. and Kubli,E. (1986) *Arch. Biochem. Biophys.* **247**, 233–237.
32. van Tol,H., Stange,N., Gross,H.J. and Beier,H. (1987) *EMBO J.* **6**, 35–41.
33. Johnson,G.D., Pirtle,I.L. and Pirtle,R.M. (1985) *Arch. Biochem. Biophys.* **236**, 448–453.34.
34. Covey,S. (1986) *Nucleic Acids Res.* **14**, 623–633.
35. Pearl,L.H. and Taylor,W.R. (1987) *Nature* **329**, 351–354.
36. Johnson,M.S., McClure,M.A., Feng,D.F., Gray,J. and Doolittle,R.F. (1986) *Proc. Natl. Acad. Sci. USA* **83**, 7648–7652.
37. Mount,S.M. and Rubin,G.M. (1985) *Mol. and Cell. Biol.* **5**, 1630–1638.
38. Voytas,D.F. and Ausubel,F.M. (1988) *Nature* **336**, 242–244.
39. Grandbastien,M., Spielmann,A. and Caboche,M. (1989) *Nature* **337**, 376–380.
40. Swofford, D.L. (1991) PAUP, phylogenetic analysis using parsimony. Illinois Natural History Survey, Champaign.
41. Saigo,K., Kugimiya,W., Matsuo,Y., Inouye,S., Yoshioka,K. and Yuki,S. (1984) *Nature* **312**, 659–661.
42. Yuki, S., Inouye,Y., Ishimaru, S. and Saigo, K. (1986) *Eur. J. Biochem.* **158**, 403–410.
43. Marlor,R., Parkhurst,S. and Corces,V. (1986) *Mol. and Cell. Biol.* **6**, 1129–1134.
44. Huijser, P., Kirchhoff, C., Lankenau, D.-H. and Hennig, W. (1988) *J. Mol. Biol.* **203**, 233–246.
45. Evgen'ev, M.B., Corces, V.G. and Lankenau, D.-H. (1992) *J. Mol. Biol.* **225**, 917–924.
46. Freisen, F.D. and Nissen, M.S. (1990) *Mol. and Cell. Biol.* **10**, 3067–3077.
47. Michaille,J.J., Mathavan,S., Gaillard,J. and Garel,A. (1990) *Nucleic Acids Res.* **18**, 674.
48. Hanson, L.J., Chalker, D.L. and Sandmeyer, S.B. (1988) *Mol. Cell. Biol.* **8**, 5245–5256.
49. Levin,H.L., Weaver,D.C. and Boeke,J.D. (1990) *Mol. and Cell. Biol.* **10**, 6791–6798.
50. McHale, M.T., Roberts, I.N., Noble, S.M., Beaumont, C., Whitehead, M.P., Seth, D. and Oliver, R.P. (1992) *Mol. Gen. Genet.* **233**, 337–347.
51. Julien, J., Poirier-Hamon, S. and Brygoo, Y. (1992) *Nucleic Acids Res.* **20**, 3933–3937.
52. Smyth,D.R., Kalitsis,P., Joseph,J.L. and Sentry,J.W. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 5015–5019.
53. Kossack, D. (1989) Ph.D. Dissertation, University of California, Davis.
54. Springer,M.S., Davidson,E.H. and Britten,R.J. (1991) *Proc. Natl. Acad. Sci. USA* **88**, 8401–8404.
55. Fourcade-Peronnet, F., d'Auriol, L., Becker, J., Bailibert, F. and Best-Belpomme, M. (1988) *Nucleic Acids Res.* **16**, 6113–6125.
56. Camirand,A. and Brisson,N. (1990) *Nucleic Acids Res.* **18**, 4929.
57. Clare, J. and Farabaugh, P. (1985) *Proc. Natl. Acad. Sci. USA* **82**, 2829–2833.
58. Janetzky, B. and Lehle, L. (1992) *J. Biol. Chem.* **267**, 19798–19805.
59. Rothnie,H.M., McCurrach,K.J., Glover,L.A. and Hardman,N. (1991) *Nucleic Acids Res.* **19**, 279–286.
60. de Chastonay, Y., Felder,H., Link,C., Aeby,P., Tobler,H. and Muller,F. (1992) *Nucleic Acids Res.* **20**, 1623–1628.
61. Felder,H., Herzceg,A., de Chastonay,Y., Aeby,P., Tobler,H. and Muller,F. (1992) *Nucleic Acids Res.* (submitted).
62. Cappello,J., Handelsman,K, andf Lodish,H. (1985) *Cell* **43**, 105–115.
63. Christy,R.J., Brown,A.R., Gourlie,B.B. and Huang,R.C.C. (1985) *Nucleic Acids Res.* **13**, 289–302.
64. Lucas,H., Moore,G., Murphy,G. and Flavell,R.B. (1992) *Mol. Biol. Evol.* 9, 716–728.
65. Aeby,P., Spicher,A., de Chastonay,Y., Muller,F. and Tobler, H. (1986) *EMBO J.* **5**, 3353–3360.