

Highly structured sequence homology between an insertion element and the gene in which it resides

(soybean seed lectin/transposable element/DNA sequence)

PATSY R. RHODES AND LILA O. VODKIN

Plant Genetics and Germplasm Institute, Agricultural Research Service, U.S. Department of Agriculture, Beltsville, MD 20705

Communicated by T. O. Diener, September 25, 1984

ABSTRACT The recessive allele for soybean seed lectin results from the insertion of a DNA segment (designated Tgm1) into the coding region of the gene. The termini of Tgm1 display structural features characteristic of a transposable element. The complete sequence of Tgm1 contains 3550 base pairs (bp) and can be divided into three regions (left arm, midsection, and right arm). No large open reading frames were found, but an extensive, highly structured border with homology to the lectin gene was revealed. The left border (726 bp) comprising most of the left arm and extreme right border (144 bp) of the right arm consist of various forms of a basic 54-bp repeating unit. This 54-bp unit is comprised of a stem-loop structure and interhairpin sequence that occurs 13 times in the left arm and 2 times in the right arm of Tgm1. Progressively degenerate forms of this repeating unit appear toward the termini of Tgm1, but the dyad symmetry remains highly conserved. Seven nucleotides (A-C-A-T-C-G-G and its complement) maintained within the stem also appear as a subset of inverted repeats found at nearly equal distances from the target site in the lectin gene. Together with the inverted repeat termini and a duplication in the left arm, this 7-bp sequence occurs a total of 33 times in Tgm1. We infer that the dyad symmetries containing this sequence are involved in target gene selection. The repeating unit format of Tgm1 describes a distinct class of eukaryotic elements that includes representatives known to be mobile in snapdragon and maize.

Mobile genetic elements do not have fixed positions in prokaryotic or eukaryotic genomes. Genetic analyses of mutable alleles have made it possible to recognize such elements in maize (1, 2) but few other plants lend themselves to genetic manipulation on the same scale. Recently, Vodkin *et al.* (3) described an element isolated from the soybean seed lectin gene that possesses 30-base-pair (bp) imperfect inverted repeat ends and duplicates 3 bp of the lectin sequence. Inverted repeat termini and small duplications of target site material are structural evidence for transposable element action. These features of the lectin insertion suggest that it arose by a transposition mechanism, although it has not been shown to be genetically unstable. For simplicity, we designate the soybean lectin gene insertion as Tgm1 (transposable element *Glycine max*). Although Tgm1 interrupts the lectin gene within its coding region, the element apparently blocks transcriptional activity (4) of the gene leading to a lectinless phenotype. Just how Tgm1 prevents transcription of sequences located upstream from the insertion site is still not understood, but the promoter and other 5' sequences of both the uninterrupted lectin gene and the mutant allele that carries Tgm1 appear to be normal (3).

Sequences related to Tgm1 occur in the genomes of both lectin-positive and lectin-negative soybean lines (4). Southern hybridization patterns indicate that more sequences

share similarities with the borders of Tgm1 than with its middle region. Among those with homology to the internal region, the distribution of restriction sites analogous to Tgm1 is found only in genomic DNA from the lectin-negative line. This indicates that either Tgm1 does not exist in the lectin-positive genome or that it is present in a rearranged form at another location.

Our analysis of the complete Tgm1 sequence details the structural properties of a distinct class of eukaryotic transposable elements whose members include Tgm1 in soybean, Tam1 in snapdragon (5), and Spm-18 in maize (6). In addition to sequence similarities among the termini of all three elements, we show that an extensive array of repeating dyad symmetries that form the borders of Tgm1 is also reflected in the partial sequence data for Tam1 and suggest that short regions of homology between dyad symmetries of the element and the mutable allele are likely to characterize these plant transposable elements.

MATERIALS AND METHODS

A 16.4-kilobase pair (kb) *EcoRI* λ clone, pS-5, which contains the seed lectin gene and intervening Tgm1 element from the lectin-negative cultivar Sooty, was described recently (4). Three *EcoRI* and *HindIII* restriction fragments of pS-5 that contain the Tgm1 element and flanking lectin gene sequences (pS-5EH4.5, pS-5H1.3, and pS-5H1.5) are present as subclones in pBR322 (host strain HB 101). These three were further subcloned by standard methods into M13 mp8-11 vectors (host strain JM 103) for subsequent DNA sequence analysis by the dideoxynucleotide chain-termination method as described (3). Subclones in M13 were also obtained for *Hpa* I fragments from the 1.0-kb *Bam*HI/*Hind*III region of pS-5EH4.5, which contains the 5' section of Tgm1, *Xba* I and *Hinf*I fragments from pS-5H1.3, which contains the midsection of Tgm1, *Sau*3A fragments from pS-5H1.5, which contains the 3' region of Tgm1, and a *Bam*HI-*Bgl* II fragment, which spans both of the internal *Hind*III sites of Tgm1, from pS-5p1 (the original pS-5 subcloned into pBR325).

Instability was frequently encountered when subcloning certain regions of the Tgm1 element into M13 vectors. This instability and a paucity of restriction sites necessitated the use of synthetic oligonucleotide primers to extend the reading of large but more stably cloned fragments. To reduce the likelihood of multiple priming, new primer sites were selected where possible from regions that had <60% homology with any other area in the known Tgm1 and M13 data bases by using the homology search programs of Larson and Messing (7). Oligomers at least 14 nucleotides (nt) long were then prepared with the aid of an Applied Biosystems model 380A automated DNA synthesizer (Foster City, CA).

A succinct description of the synthesis cycle and subsequent purification of the crude product by HPLC has been

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: bp, base pair(s); kb, kilobase pair(s); nt, nucleotide(s).

presented by Broido *et al.* (8). Typically, half of the lyophilized crude product was dissolved in 500 μ l of 0.01 M triethylammonium acetate (pH 7.1) and centrifuged for 15 min at $12,000 \times g$ immediately before injection to remove debris. Separation of incomplete sequences, which elute with the void volume, from the finished product was achieved on a μ Bondapak C_{18} column (0.39×30 cm, Waters Associates) equipped with Bio-Rad reversed phase micro-guard cartridges at a flow rate of 2 ml/min. The major 5'-dimethoxytrityl-bearing component eluted between 7.5 and 10 min with 0.1 M triethylammonium acetate (pH 7.1) and a linear 20–30% acetonitrile gradient over 10 min. Normally, 200–300 μ g of purified oligomer was obtained from one 200- μ l aliquot of crude product, of which 3–4 ng was used to prime a dideoxy sequencing reaction.

RESULTS

Tgm1 is integrated into the single open reading frame of the soybean seed lectin gene (Fig. 1). In contrast to the simple structure of the target gene, Tgm1 is a complex and highly structured element. First, nearly a quarter of the entire length of Tgm1 is comprised of related border sequences that extend inward from the 5' and 3' termini of the element. An extensive array of border sequences on the 5' end of Tgm1 are separated from a lesser assemblage at the 3' end by 2.68 kb of internal sequence unrelated to the border. Second, the base composition of this 2.68-kb internal material varies in a distinctly nonrandom manner. The A+T base composition of Tgm1 (74%) is high overall, but certain areas are even richer in A+T. Areas ranging from 80% to 90% A+T are scored boldly in Fig. 1B. Interspersed among the A+T-rich areas are sequences that range from 60% to 70% A+T. From a topographical point of view, Tgm1 is organized into three segments: a left arm composed mostly of border sequence, a midsection with extensive areas of 60–70% A+T, and a right arm that is predominantly 80–90% A+T with a very short border sequence at its 3' end. The *Hind*III restriction sites (Fig. 1C) in Tgm1 roughly separate the midsection from the left and right arms.

A detailed examination of the left arm border (nt 1–726) and right arm border (nt 3406–3550) reveals an intricate array

of related sequences (Fig. 2). These extensive border sequences are variations based upon a prototype 54-bp repeating unit found at nt 673–726. Because of internal complementarity, the repeating unit can be organized into a hairpin structure that features an *Hpa* I site (G-T-T-A-A-C) at the base of a 16-bp stem (Fig. 2). Display of the repeating unit as a hairpin effectively conveys both the position and character of each repeat and should not be construed to represent biologically significant arrangements of Tgm1 borders. Thus Fig. 2 shows that there are 13 repeating units in the left arm and only 2 in the right arm. Certainly, more intricate configurations among the repeating units could be devised by pairing nonadjacent dyad symmetries.

A prominent feature of the repeating unit is a 7-nt sequence, A-C-A-T-C-G-G, and its complement, C-C-G-A-T-G-T. These sequences are a part of the 16- to 18-bp inverted repeats that comprise the stem of each hairpin shown in Fig. 2. We have previously reported the occurrence of the same 7 nt (3) within the 30-bp imperfect inverted repeat termini of Tgm1 (see also Fig. 2) as well as within 17-bp inverted repeats in the lectin gene coding region (C-A-A-A-T-C-C-A-C-A-C-A-T-C-G-G-A and T-C-C-G-A-T-G-T-G-G-T-C-G-A-T-T-T-G found 80 nt 5' and 67 nt 3' of the insertion target site, respectively; see ref. 3 for complete lectin gene sequence). It was surprising to find these 7 bp conserved within the Tgm1 repeating unit. Including all derivative forms, this sequence appears a total of 33 times in the element. Because of the prevalence of these 7 bp throughout the borders, we consider this sequence to be a molecular signature for Tgm1 in the sense that it is an identifying characteristic of the element.

The basic 54-bp repeat unit, consisting of two *Hpa* I sites and two 7-bp signature sequences amid loop and interhairpin material, appears as five near-perfect tandem duplications in the left arm extending from nt 459–726 (Fig. 2; shading and arrows). Between these and the 5' terminus of Tgm1 are 8 more repeat units (hairpins). These 8 copies differ from the basic repeat unit primarily in the interhairpin material and to some extent in the sequences that comprise the loop. The *Hpa* I site and signature sequence are conserved within the 16-bp stem. However, neither the first (extreme 5') hairpin nor the third hairpin has remnants of an *Hpa* I site. A further

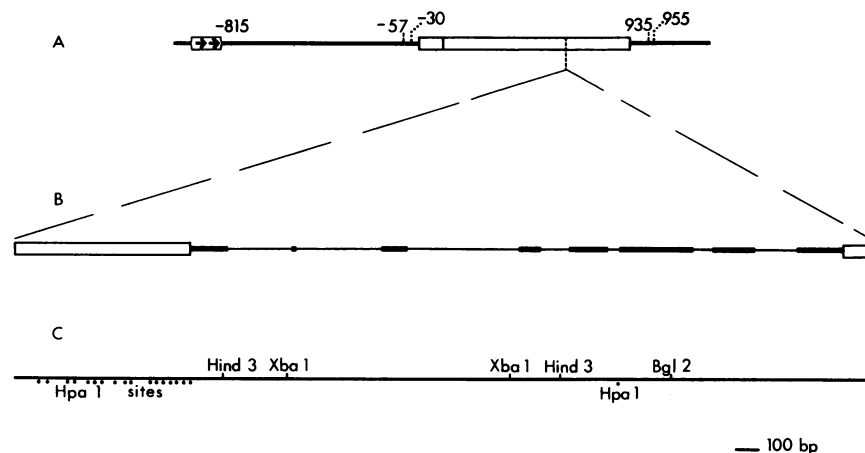


FIG. 1. Schematic drawing of the soybean seed lectin gene showing the location of the insertion target site (A) structural features of the 3.5-kb Tgm1 insertion (B) and selected restriction sites within the Tgm1 element (C). (A) The segment presented encompasses 2100 bp of coding and flanking sequence with the numbers showing the distance in nucleotides from the initiation codon (3). Two tandem repeats of 58 bp each are designated by arrows enclosed by the box at -815. The promoter (T-A-T-A-A-A-T-A) at -57, the cap site at -30, a poly(A) signal (A-A-T-A-A-T) at +935, and the major poly(A) addition site at +955 are marked with dashed (signals) or dotted (ends of mRNA) lines. The 855-bp coding region (open box), which contains no introns, is partitioned into a short initial section coding for a 32-amino acid signal peptide followed by the coding region for the mature protein. The dashed line through this segment represents the location of the insertion target site at +600. (B) The open boxes delineate 726 bp of the left arm and 144 bp of the right arm, which consist of tandem repeating units. The relative position of 80–90% A+T-rich areas (bold lines) and 60–70% A+T areas (thin lines) are also presented. (C) Restriction sites for the Tgm1 element shown in B are indicated. Clustered *Hpa* I sites associated with the repeating unit in the left arm are designated with a dot only. All figures are drawn to scale.

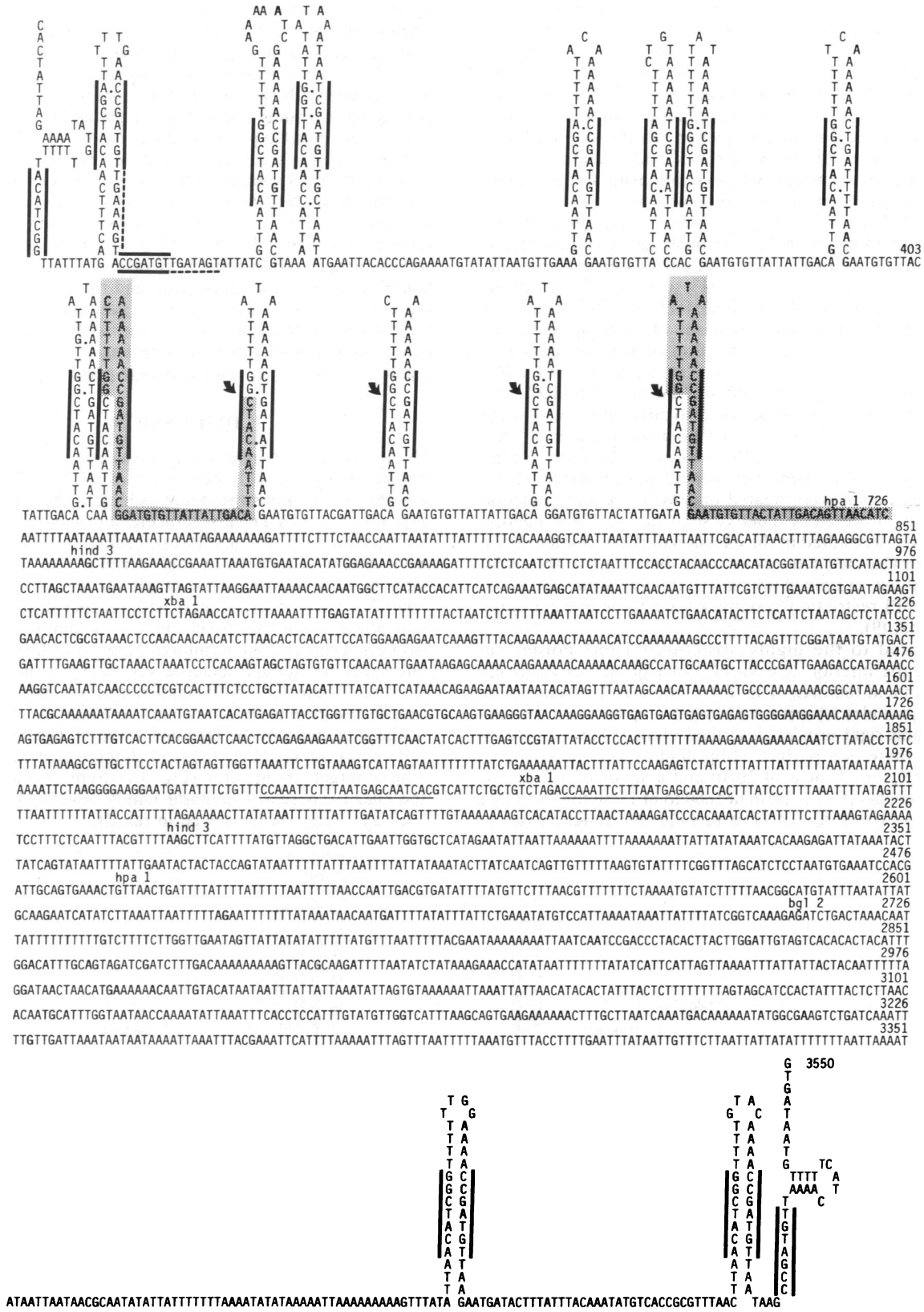


FIG. 2. Nucleotide sequence of the Tgm1 element. The duplication of target site material (C-T-A) that directly flanks Tgm1 is not shown (3). The nucleotides are numbered from the 5' border and are presented on the right side of the figure. The signature sequence (A-C-A-T-C-G-G) and its complement are denoted by bold lines; dots in the hairpins indicate noncomplementary nucleotides; a tandem duplication of one side of the stem in the first hairpin is marked with a dashed line; and a perfect 24-bp duplication bracketing the *Xba* I site at nt 2046 is underlined. *Hpa* I sites at the base of many of the hairpins are not designated; otherwise, all restriction sites corresponding to those in Fig. 1 are shown. Five near-perfect tandem duplications of the 54-bp repeating unit extending from nt 459 to nt 726 are indicated with shading and arrows.

distinction of the first hairpin in the left arm is a near-perfect 15-bp tandem duplication of one side of the stem. The significance of this duplication is also not known but it occurs only once in Tgm1. Two more copies of the repeat unit are found at the 3' border of Tgm1, and, like their counterparts at the 5' border, these differ from the basic repeat unit in loop and interhairpin sequence while maintaining a remnant of the *Hpa* I site and a signature sequence. Evidently, the basic repeat unit degenerates toward the 5' and 3' ends of Tgm1 into a variety of sequences whose relationship to each other is apparent only through the conserved stem area.

We do not know whether this complex arrangement of tandem repeating units with dyad symmetries has a function *in vivo*; however, we have observed that the borders cause instability in the propagation of M13 subclones from certain regions of Tgm1. A 1.0-kb *Bam*HI-*Hind*III fragment that contains the left arm and 104 bases of flanking lectin sequence resulted in frequent deletion of part or all of the cloned fragment from the M13 phage. A 2.8-kb *Bam*HI-*Bgl*II fragment consisting of the left arm and an additional 1.8 kb from the midsection of Tgm1 was also difficult to maintain in culture. On the other hand, a 1.5-kb *Hind*III fragment that contains the right arm and 238 bases of flanking lectin gene sequence was stable. However, when *Sau*3A fragments from within this region were ligated into M13, frequent deletion events were detected by dideoxy sequencing. These deletion events always resulted in the excision of the 3' border region along with flanking sequences some distance to either side. None of those examined appears to involve either the *Eco*K or the *Eco*PI restriction systems known to be present in the JM103 host (9).

Compared to the highly structured Tgm1 border sequences, the interior 2.68 kb of Tgm1 possesses relatively few prominent features. The distribution of A+T base composition provides a basis for distinguishing between the midsection and right arm as described earlier (Fig. 1B). Additionally, there is a perfect 24-bp duplication bracketing an *Xba* I site (Fig. 2; near nt 2050) and a series of 4 T-G-A-G tandem repeats (Fig. 2; from nt 1680 to nt 1710) followed by a short G-A-G-T-G-A-G-A-G at nt 1726. No function can be

assigned to these sequences at present. There are open reading frames in both orientations but these are quite short, ranging from 40 to 63 codons in length. Except for some frames 60 codons in length throughout the 5' border region, all of these open frames are associated with the 60–70% A+T regions of the midsection and right arm. In the absence of large open reading frames, splicing would be the only way to code for a protein product of substantial size. Perhaps coincidentally, soybean donor and acceptor consensus sequences for RNA splicing (10) are contiguous to one another at nt 2815 to nt 2825 (G-T-A-A-G-T-G-T-A-G-G on the opposite strand), but we can find no other evidence of consensus sequences in Tgm1. In spite of this apparent lack of protein coding capacity, transcripts have been detected (4) that were homologous with the midsection and right arm of Tgm1. If these transcripts are derived from Tgm1 and not from one of the closely related sequences in the soybean genome, it is conceivable that the RNA transcript may play a structural or regulatory role rather than encoding protein information.

DISCUSSION

The importance of insertion element ends to the transposition process is a widely held assumption (11, 12). In fact, it was recently demonstrated that as long as the transposase is available, the only Tn5 insertion element sequences essential for transposition are the 19-bp termini (13, 14). Elements with similar sequences at their termini also duplicate the same number of bases at the target site (15, 16) and thus are related to each other, at least with respect to the transposition process. On this basis, there is a family resemblance between Tgm1 and the termini of Tam1, a 17-kb element that interrupts the mutable chalcone synthase gene in snapdragon (5), and of Spm-18, a 2-kb element that interrupts a glucosyl transferase gene at the *waxy* locus in maize (6). At least 75% of the sequences in the terminal 13 bp of Tgm1, Tam1, and Spm-18 are similar (Fig. 3A), and all three elements duplicate 3 bp of target site material (Fig. 3A). Spm-18 is a member of the maize suppressor-mutator family described by McClintock and Peterson (reviewed in ref. 2).

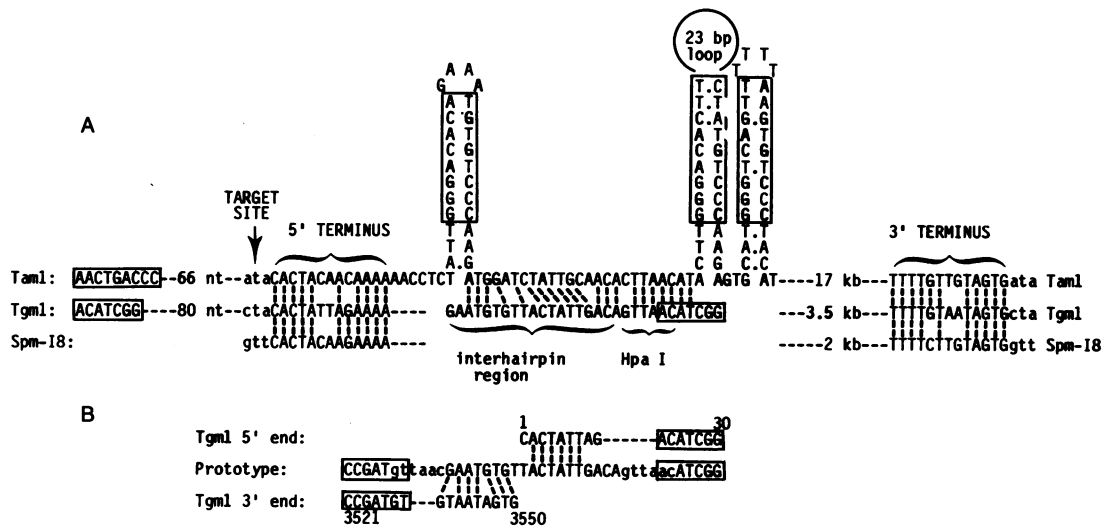


FIG. 3. Familial relationship between Tgm1 and mobile elements from snapdragon (Tam1) and maize (Spm-18). (A) Published Tam1 sequence (5) has been reorganized in a format similar to Tgm1 (Fig. 2). A putative signature sequence and its derivatives in Tam1 and the A-C-A-T-C-G-G signature of Tgm1 are enclosed in boxes. The distances in nucleotides between the occurrence of a signature in the interrupted gene and the Tam1 or Tgm1 element are denoted and the 3-bp duplicated target sites are in lowercase. The appearance of a signature in the lectin gene at the 3' end of Tgm1 is not shown (3). A Tgm1 prototype interhairpin sequence from nt 699 to nt 726 followed by the *Hpa* I and signature sequence is indicated. Because of limited sequence data (6), only the inverted repeat termini of Spm-18 are shown here. Base homologies are connected by vertical lines; dots in the hairpins indicate mismatched bases. (B) A Tgm1 prototype interhairpin flanked by *Hpa* I sites (lowercase) and signature sequences (boxed) is taken from nt 688 to nt 726. Homology between 5' and 3' termini of Tgm1 and the prototype is indicated by vertical lines. Numbers above and below the lines are nucleotide positions for the termini.

This familial relationship between Tgm1 in soybean and two demonstrably mobile elements, Tam1 in snapdragon and Spm-I8 in maize, need not logically extend to the highly structured borders described for Tgm1. However, the Tgm1 termini that are similar in sequence to the termini of these two elements can be viewed as originating from the interhairpin region of a prototype 54-bp repeating unit (Fig. 3B). The signature sequence is also oriented properly for homology with the prototype repeating unit. Viewed in this manner, Tam1 and Spm-I8 may be expected to have borders composed of sequences much like the Tgm1 prototype unit. In fact, scrutiny of the partially sequenced 5' end of Tam1 reveals an interhairpin region that is quite similar to the prototype interhairpin region of Tgm1 (Fig. 3A). Not surprisingly, Schwarz-Sommer *et al.* (6) indicate that Tam1 and Spm-I8 sequence similarities extend to regions beyond the termini. Although there is evidence of extensive secondary structure at the borders of both Tam1 and Spm-I8 (5, 6), there are not enough sequence data available on either element to define a prototype repeat unit or determine whether organizational variation throughout the borders follows a regimented pattern like Tgm1.

Sequence homology among the termini of Tgm1, Tam1, and Spm-I8 suggests the possibility that one transposase may recognize the termini of all three elements. On the other hand, the similarity between Tgm1 termini and the interhairpin region (Fig. 3B) indicates that there must be enough sequence variability in the interhairpin portion of each repeating unit to prevent a transposase from recognizing sequences other than the termini. We do not know what this observation implies with regard to the repeat unit function. However, we do suspect that the signature sequence portion of at least one of the repeating units may play a role in selecting the target gene.

The presence of a signature sequence in the repeating unit dyad symmetry of Tgm1 as well as within inverted repeats in the lectin gene itself argues for the involvement of palindromic structures in target site recognition. As shown in Fig. 3A, it is also possible to rearrange the 5' border of Tam1 into hairpin structures. A perfect complement of a 5' flanking gene sequence is found in the stem of the third hairpin and degenerate forms of this putative signature sequence appear within each stem. Similar analyses at the 3' end of Tam1 are possible, but only imperfect homology to the flanking gene region can be seen with the available data. The only other analogous symmetrical arrangement of palindromic sequences suggestive of site-specific integration is a partial homology reported between inverted repeat termini of ISH1 and an interrupted inverted repeat immediately flanking the duplicated target site in the bacteriorhodopsin gene (17).

In terms of general structural organization, the Tgm1 repeat unit format is reminiscent of the *Drosophila* foldback family of transposable elements, which also have extensive borders composed of tandem direct repeats. However, the foldback direct repeat unit is not based upon a dyad symmetry and has no reported sequence homology with the genes in which it is found (18, 19). Foldback termini show no homology with the termini of Tgm1 and members of the foldback family duplicate 9 bp of target site material during transposi-

tion. There are parallels, however, in that the direct repeat unit of foldback varies in a periodic manner along the length of each border much like the degenerate to prototype forms in the left arm of Tgm1. The borders of both also terminate abruptly at the junction with unrelated internal regions.

The repeating unit format of Tgm1 borders and homology between the palindromic signature sequence and the lectin gene are unlike any prokaryotic or eukaryotic transposable element described thus far. All of these features are reflected in our analysis of the published border sequences from Tam1 (Fig. 3A) and, to the extent that data are available, of border sequences from Spm-I8. We infer from these data that members of this structural class of transposable element are characterized by borders that consist of a tandem repeating prototype unit that shows homology to at least the interhairpin portion of the Tgm1 repeat unit. Furthermore, we predict that the repeat unit will be individualized for each member by containing within it a signature sequence (likely to be part of a dyad symmetry) that shows some homology to flanking regions of the genes in which these elements are located.

We thank Dr. Gerald Zon of the Food and Drug Administration, Bethesda, MD, for his advice on the preparation and purification of synthetic oligomers. We thank Mr. Robert Goeken for his excellent technical assistance. P.R.R. was supported by the Agricultural Research Service Research Associate Program. This work was partially funded by U.S. Department of Agriculture Competitive Grant 83-CRCR-1-1272.

1. McClintock, B. (1951) *Cold Spring Harbor Symp. Quant. Biol.* **16**, 13-47.
2. Fedoroff, N. V. (1983) in *Mobile Genetic Elements*, ed. Shapiro, J. A. (Academic, New York), pp. 1-63.
3. Vodkin, L. O., Rhodes, P. R. & Goldberg, R. B. (1983) *Cell* **34**, 1023-1031.
4. Goldberg, R. B., Hoschek, G. & Vodkin, L. O. (1983) *Cell* **33**, 465-475.
5. Bonas, U., Sommer, H. & Saedler, H. (1984) *EMBO J.* **3**, 1015-1019.
6. Schwarz-Sommer, Zs., Gierl, A., Klosgen, R. B., Wienand, U., Peterson, P. A. & Saedler, H. (1984) *EMBO J.* **3**, 1021-1028.
7. Larson, R. & Messing, J. (1982) *Nucleic Acids Res.* **10**, 39-49.
8. Broido, M. S., Zon, G. & James, T. L. (1984) *Biochem. Biophys. Res. Commun.* **119**, 663-670.
9. Felton, J. (1983) *BioTechniques* **1**, 42-43.
10. Mount, S. M. (1982) *Nucleic Acids Res.* **10**, 459-472.
11. Iida, S., Meyer, J. & Arber, W. (1983) in *Mobile Genetic Elements*, ed. Shapiro, J. A. (Academic, New York), pp. 159-221.
12. Berg, D. E. & Berg, C. M. (1983) *Bio/technology* **1**, 417-435.
13. Johnson, R. C. & Reznikoff, W. S. (1983) *Nature (London)* **304**, 280-282.
14. Sasakawa, C., Carle, G. F. & Berg, D. E. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 7293-7297.
15. Calos, M. P. & Miller, J. H. (1980) *Cell* **20**, 579-595.
16. Reed, R. R., Young, R. A., Steitz, J. A., Grindley, N. D. F. & Guyer, M. S. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 4882-4886.
17. Simsek, M., DasSarma, S., RajBhandary, U. L. & Khorana, H. G. (1982) *Proc. Natl. Acad. Sci. USA* **79**, 7268-7272.
18. Potter, S. S. (1982) *Nature (London)* **297**, 201-204.
19. Paro, R., Goldberg, M. L. & Gehring, W. J. (1983) *EMBO J.* **2**, 853-860.