J Mol Evol (1998) 47:463–470

JOURNAL OF MOLECULAR EVOLUTION

Similar Target Site Selection Occurs in Integration of Plant and Mammalian Retroposons

Christophe Tatout, Laurence Lavie, Jean-Marc Deragon

Biomove, UMR6547 CNRS, Université Blaise Pascal Clermont-Ferrand II, 63177 Aubière Cedex, France

Received: 27 February 1998 / Accepted: 5 May 1998

Abstract. The reverse transcription of RNA in DNA is responsible for the generation of large families of repetitive sequences called retroposons or non-LTR retrotransposons. Recent reports established that the integration of mammalian SINE and LINE retroposons occurs at nonrandom staggered breaks, probably resulting from the action of a LINE-encoded endonuclease (Feng et al. 1996; Jurka 1997; Jurka et al. 1998). We report here that plant SINE S1 retroposons also integrate at nonrandom staggered breaks. One of the two nicks involved in S1 integration is associated mainly with the 5′-Y/AAANNNG-3′ motif. The other nick at opposite DNA strand occurs preferably within 14–16 bp, a situation also observed for mammalian retroposons, but is not associated with any specific motif. Further studies on the distribution of dinucleotides surrounding the two nicking sites showed that, as for mammalian retroposons, S1 retroposons integrate at sites rich in TA, CA, and TG dinucleotides. These dinucleotides were reported as specific DNA sites where special DNA structures called ''kinks'' may occur under bending constraints. Nicking sites are preceded by peaks in frequency of di-pyrimidine followed by peaks of di-purine. These results suggest that the general A/T richness of a given DNA region and the presence of short runs of pyrimidines followed by short runs of purines could represent a favorable context for the integration of retroposons. In such a context, an endonuclease upon fixation could be able to generate the kink at the pyrimidine/purine transition and to nick the DNA. The similarities in target site selection observed for plant and mammalian retroposons suggest that retroposition is a surprisingly well conserved process.

Key words: Repetitive sequence — Retrotransposon — Alu element — Transposable element — Retroposition

Introduction

Transposable elements are discrete mobile DNA fragments that can insert into nonhomologous target sites. Diverse patterns of target site selectivity are observed, but although some display little obvious specificity, none appears to be truly random. Retroposons represent a class of transposable elements whose amplification involves the reverse transcription of an RNA intermediate. Retroposons are very abundant in mammals and can also be found in moderate to high copy numbers in other eukaryotes, from fungus to fish and plants (Kachroo et al. 1995; Kido et al. 1993; Yoshioka et al. 1993; Deragon et al. 1994). Studies on the LINE R2Bm and on the SINE Alu led to a retroposition mechanism which could account for LINE and SINE retroposition (Luan et al. 1993; Jurka 1997). In this model, a first cleavage on one strand occurs in the target DNA and produces a priming site for reverse transcriptase. This priming site is believed to be T-rich and allows hybridization with the poly(A) tail of the retroposon RNA. This generates a DNA copy of the element linked to the new insertion site. A second cleavage occurs at the other DNA strand of the target DNA. For most SINE and LINE elements, this second nicking

Correspondence to: J.-M. Deragon

site occurs at a distance from the first one, creating after retroposon integration a target site duplication (TSD) on both sides of the retroposon.

Recently, studies on a large number of human Alu and rodent B1, B2, and ID SINE elements revealed their capacity to integrate in selected target sites (Jurka and Klonowski 1996; Jurka 1997; Jurka et al. 1998). The first nick in the integration process of these elements is strongly associated with the 5′-TT/AAAA-3′ hexanucleotide or, more generally, with the 5′-(Y)n/(R)n-3′ motif (Jurka 1997). This nick is assumed to occur predominantly between the dipyrimidine TT and the following di-purine AA, but the exact position of the nicking sites was not determined in most cases since the A-rich 5′ end of the TSD merged with the poly(A) tail of the retroposon. The second nick is made on the other strand in 3′ of this hexanucleotide, preferably within 15–16 base pairs, at a less conserved motif that can be represented as 5′-TYTN-3′, where Y denotes pyrimidine; T, thymine; and N, any base (Jurka 1997).

Studies on the human L1 LINE element also revealed their capacity to integrate in selected target sites (Feng et al. 1996). L1 integration target sites are made of short runs of purine (often A's) preceded by short runs of pyrimidines. The endonuclease encoded by the L1 element was shown to be capable in vitro of generating nicks between short runs of pyrimidines and short runs of purines as expected from the L1 target site specificity (Feng et al. 1996).

Taken together, these studies suggest that staggered breaks generated during mammalian retroposon integration do not result from random nicking but probably from the action of an endonuclease. The obvious similarities between Alu and L1 integration sites also imply that their integrations probably depend on the enzymatic activity of an endonuclease encoded by the L1 element itself (Feng et al. 1996; Jurka 1997) and that Alu retroposons can be considered as truly parasitic elements of the L1 retroposition process (Boeke 1997).

Analysis of the distribution of dinucleotides at mammalian retroposon target sites revealed that they were highly enriched in TA but also in CA and TG dinucleotides (Jurka et al. 1998). Under bending constraints, these dinucleotides are known to form ''kinks'' which represent abrupt deflections of the double helical structure leading to unstaking of two neighboring base pairs (McNamara et al. 1990). These dinucleotides are preceded at proposed nicking sites by enhanced frequencies of dipyrimidines and followed by peaks of di-purines. These characteristic features have been shown to be sites for endonuclease cleavage such as the EcoRV restriction enzyme (Winkler et al. 1993). The L1 endonuclease could therefore be able to generate sequence dependent DNA kinks followed by nicks upon fixation at Alu and L1 integration sites (Jurka et al. 1998). The occurrence of (Py)/(Pu) tracts interrupt by kink dinucleotides could represent a primary signal for LINE endonuclease and hot spots for SINE and LINE integration events.

In this paper, we report the characterization of integration sites of a plant SINE retroposon called S1 (Deragon et al. 1994). S1 elements are short repeats (~180 bp) that occupy 500 to several thousand loci by haploid genome in the different crucifer species studied (Lenoir et al. 1997). They present all structural features found in SINE retroposons. S1 elements present a primary and secondary sequence homology to several tRNA species (Deragon et al. 1994, 1996). They possess a 3′-terminal A-rich region composed of a poly(A) followed in a few cases by a small number of (TA) or (TAA) repeats. Most S1 elements possess two conserved polymerase III motifs (box A and B) that could potentially be used to direct transcription (Deragon et al. 1996). S1 insertion events usually generate TSD. While S1 elements are GC-rich (around 60%), the direct repeats and the flanking regions are usually AT-rich. We show here that S1 element TSD exhibit many similarities with mammalian ones, suggesting that plant and mammalian retroposition are highly conserved processes and integrate at selected target sites.

Materials and Methods

Biological Materials

DNAs from 11 crucifer species used in this study (i.e., *Brassica napus, Brassica nigra, Brassica juncea, Brassica montana, Brassica hilarionis, Brassica incana, Brassica macrocarpa, Brassica villosa, Brassica cretica, Sinapis arvensis*, and *Brassica rupestris*) were kindly provided by Suzanne Warwick (seed sources given by Warwick and Black 1991). These DNAs were extracted and purified as by Warwick and Black (1991).

Reverse PCR, Cloning, and Sequencing

DNA was first digested with several restriction enzymes that do not cut in consensus S1 sequences. The restriction enzymes used were *Mbo*I, *Tai*I, *Nla*III, and *Ase*I (all from New England Biolabs). The DNA was then ligated at low concentration (1 μg in 250 μl) and double (nested) PCR reactions using Gold Star Taq polymerase from Eurogentec in a Statagene thermocycler were done in standard conditions. Oligonucleotides used for the first PCR were 5′-CTGGRCACGCCTCCCCC-3′ and 5′-GGTACAKRCAMARGYTGRCGCCGG-3′ and for the second PCR 5′-CCACTGGACTACGAGGTCC-3′ and 5′-GGTCAACACCTGGTTAAT-3′ or 5′-GCTGGCGCCGGGCCTAGG-3′. pGEM-T from Promega and T7 RNA polymerase from Pharmacia were used as recommended by the manufacturers to respectively clone and sequence the PCR products. Most of the insertion sites from *Brassica napus* were obtained by screening a genomic library as described by Deragon et al. 1996.

Amplification of Empty Insertion Sites

DNA from *Brassica hilarionis* was used to amplify most empty sites orthologous to S1 containing sites except for na2, na18, cr11, cr21,

A

```
ni2     ccattgtatcacagtctttt----------------------ACCC~AAAA----------------------gaatatgaagagcaagctttt
ni9     ttcaaaaatgaaaagtaaat----------------------ACCC~AAAA----------------------gtaatttttttgtttaaca
ni13    agataagagtggtgttttgat---------------------ACCC~AAAA---------------------catggaactttgatttctct
na14b   cttgcatgcctgcaggaacc----------------------ACCC~AAAA----------------------ctgcaggcatgcaagcttgg
na18    ggtgataaatgttattattt----------------------ACCC~AAAA----------------------ctatgagaaactttaccatg
na27    ggaaaaatttgtgaaatttg----------------------ACCC~CAAA----------------------atgcatacactccttgtgcg
hi3     gtttacaacttncatttcgt----------------------ACCC~AAAA----------------------tttcatttcatgctatagga
cr11    gcatcataaatattttcggt----------------------ACCC~AAAA----------------------ttaggtataaaggtcacccc
cr12    tcgtcatcaaatatccattt----------------------ACCC~AAAA----------------------cattgaaaaaacttgtcatg
inc20   agcattttgggttttagagt----------------------ACCC~AAAA----------------------catataataagagtgataat
vil35   tatgctcgaagtgtagcttg----------------------ACCC~AAAA----------------------catgatataagatgtaaaac
ma43b   gaaaacgaattgaggcaaat----------------------ACCC~AAAA----------------------tacccgtttatgatttctat
jun4-5  agggaaaaaaaattgatata----------------------ACCC~AAAA----------------------ccgaacatatactagcagta
```

B

```
ni1     ctcaactgattttag AAGTAAGAATATTTTA----------ACCC~AAAA----------AAGTAAGAATATTTTA actagtactgtaaaa
ni7     actaaataaatatat AAAGTAGCAATAcAA----------ACCC~AAAG----------AAAGTAGCAATAAAA aagaaccctcaatta
ni17    tattgatgcaaatat ATGCAAATT----------ACCC~AAAA----------ATGCAAATT ggcgatttaagttaa
ni20    caaacaatatcgaac AAAACAAAATAATTAT----------ACCC~AAAA----------AAAACAAAATAATTAT atgatagcgaatgga
ni36    gaaatatgaacaagc AACTGTGCAAGGATTT----------ACCC~AAAA----------AACTGTGCAAGGATTT ctgttacaataaagc
ni45    gcagaacagcagaag ATATAATAAACT----------ACCC~ATAT----------ATATAATAAACT ataaagctttgatgt
ni57    tgaaaataaacattg AAACAAGTAAAAAGAT----------ACCC~AAAA----------AAACAAGTAAAAAGAT aaatagttggttggt
ni59    taaatttcagtctcc AAACTATCGATTTGTATT----------ACCC~AAAA----------AAACTATCGATTTGTATT aagatgaacataaaa
na1     tgcggaatgattgag AAAGAAACCATTGGT----------ACCC~AAAA----------AAAGAAACCATTGGT tttaattaaacatta
na2     aaaacaaaatctttg AATAATGTAATTATT----------ACCC~AAAT----------AATAATGTAATTATT aattaaataaaatta
na3     taaaacttatataag AAGTTAAAAGCATC----------ACCC~AAAA----------AAGTTAAAAGCATC aaaatttatactgta
na4     acttaaaaaacaaac AACAATTATGAAAAC----------ACCC~AAAA----------AACAATTATGAAAAC attgctaagtctctt
na5     cgttttctcgtagca AATTTATGGACACA----------ACCC~AAAA----------AATTTATGGACACA aacctgtaatattca
na6     aataaaagtaagact AAAGAAGCTATATTTT----------ACCC~AAAA----------AAAGAAGCTATATTTT tttttccttcttata
na7     gatattgaacatacc AATCAAGACAAAACA----------ACCC~AAAA----------AATCAAGACAAAACA tttgttttggtctta
na8     caataagtgtgttct AAGATGGT----------ACCC~AAAA----------AAGATGGT ggtggaaaaaattgg
na9     -----gatctgaacc AAAGTCACTCTCTTTT----------ACCC~AAAA----------AAAGTCACTCTCTTTT aatacaaacaactaa
na10    aaaagaacaaaaaag AACAAAAATAATACTTTT----------ACCC~AAAA----------AACAAAAATAATACTTTT caaggtatacaccta
na11    tatatttagctactc AAATTGGTAAAAAA----------ACCC~AAAA----------AAATTGGTAAAAAA ttcgtctcataatac
na12    ttccatgcactatgt AAGATGAAAATAGA----------ACCC~AAAA----------AAGATGAAAATAGA atatggatatcagtt
na13    ggtaggtttgaaact ATATAGAAGACTAGTA----------ACCC~AAAA----------ATATAGAAGACTAGTA tggctcctatgcaaa
na14    acactagtggatccc AAACTGTGTCATGT----------ACCC~AAAA----------AAACTGTGTCATGT taaagaagctttatt
na15    ------gtatgaagt TTATGAGTGAT----------ACCC~AAAA----------TTATGAGTGAT taatatatattaatt
na16    ctacatgaaggatcca ATATATAGCCATATT----------ACCC~ATAT----------ATATATAGCCATATT acatgaaacttttt
na17    tatatatctatatag AAGCTAGTTATT(A)13TGGGCATGTT-ACCC~AAAA----AAGCTAGTTAT(A)8TGGGCATGAA gctggctccaactaa
na19    gaatttgattcctc AAACTTTATAATTAC----------ACCC~AAAA----------AAACTTTATAATTAC aatgattgaaacacg
na30    aaaagctcttctttc AAAACTAATTGAAAAG----------ACCC~AAAA----------AAAACTAATTGAAAAG aatgaagctactctt
na31    ttaggatctgttaac AACAAAGCCTAAA----------ACCC~AAAA----------AACGAAGCCTAATT aaagaaaatttacgt
na32    tatattctttgttgt AAGTATGTTTTGTTGGTTT---------ACCC~AAAA----------AAGTATGTTTTGTTGGATG actttattcacgaa
na34    aaaaacttatgtgac AAAACTGAGATTA----------ACCC~AAAA----------AAAACTGAGATTA ttttgcataatttgc
hi2     acatatatgtcgacg AAATTTATGTCAATAT----------ACCC~AAAA----------AAATTTATGTCAATAT ccttgaatttgaagt
hi6     ----------gatc AGAAGAATATAAACC----------ACCC~AAAA----------AGAAGAATATATAAACC agaaagagatc----
cr8     ----------gatc AACAAATGATGAAAAA----------ACCC~AAAA----------AACAAATGATGAAAAA ttaaagtatttcact
mo13    -----gatcggtatc AACAAATGACCCTGAAG----------ACCC~AAAA----------AACAAATGACCCTGAAG tccaagaggcagcaa
mo13b   ataaaatcgtatgtg AAAAAATTGTAAGGCA----------ACCC~AAAA----------AAAAAATTGTAAGGCA tctttttgtgccaaat
mo14    atcgacgttcttttta ATATTACTCATCGA----------ACCC~AAAT----------ATATTACTCATCGA cgggttacgttatta
inc19   acattattaagtatt AACTATGGAAAtgaaaGGAT----------ACCC~AAAA----------AACTATGGAAAGGAT gatc----------
inc21   aaatataacaatccg AACAATCATGTGACG----------ACCC~AAAA----------AACAATCATGTGACG caaatttgtctgatg
vil26   acaatgttttgtagt AACAAACGTTTATTC----------ACCC~AAAA----------AACAACGTTTATTC aatcatatataatct
vil26b  gttaatttgttattc AAGTCTATAAGGTGTA----------ACCC~AAAA----------AAGTCTATAAGGTGTA ggtaaacaaaggaaa
vil30   ----------gatc ATAACCAATAATTAC----------ACCC~AAAA----------ATAACCAATAATTAC ataacaaactaagaa
vil30b  ttttaccctgatgtt AAATCTCATATAtAG----------ACCC~AAAA----------AAATCTCATATAaAG gcttatttgtttggt
rup41   atattgctcgtggat AACATTGTTGTGAGT----------ACCC~AAAA----------AAcCATTGTTGTGAGTC aagtatttctttc--
ma44    ttgaatatgtaatac AATTCTTCAGAATc----------ACCC~AAAA----------AATTCTTCAGAAT ttttgtggatgtcaa
ma45    agaaccggtggttcc AATATTTGAAGAAATT----------ACCC~AAAA----------AATATTTGAAGAAATT gacaatcctttaatt
ju9-8   -caactaatggttac AAAATTCTACATTTGc----------ACCC~AAAA----------AAAATTCTACATTTG tagtgttccaaacat
ju4-8   tgagttttgcttgat AACTTTGCGGTTT----------ACCC~AAAA----------AACTTTGCGGTTT gaggtccggcctaa-
ju10-4  tcattaagatgggag AAAACAGAAGTTCGGAATT---------ACCC~AAAA----------AAAACAGAAGTTCGGAATT tattagggaagctag
jun4-2  tgagttttgcttgat AACTTTGC----------ACCC~AAAA----------AACTTTGC ggtttgaggccggcc
jun5-3  --tattgtgagtttc ATAGTTGAATAACTT----------ACCC~AAAA----------ATAGTTGAATAACTT agctcaatataggct
rapa30  ttgaaccaaaatatg TCCCCTATA----------ACCC~AAAA----------TCCCCTAT tattattttaggagc
```

C

```
Ta11-1  cattatattatttgt AACTATATTCACTATGG----------ACAA~AAAA----------AACTATATTCACTATGG gtttggaagaccaat
```

Fig. 1. Alignment of S1 integration sites from 10 *Brassica* species. The corresponding *Brassica* species from which each element has been isolated is reported at the *left: B. nigra* (ni), *B. napus* (na), *B. hilarionis* (hi), *B. cretica* (cr), *B. montana* (mo), *B. rupestris* (rup), *B. incana* (inc), *B. macrocarpa* (ma), *B. villosa* (vil), *B. rapa* (rapa), and *B. juncea* (jun). Names in *boldface* identify S1 sites where the orthologous empty sites were cloned and sequenced (see text and Fig. 2). S1 sequence and poly(A) tract are denoted by ACCA–AAAA. On each side of the SINE, TSDs were adjusted to the left so that they all started at the same position and to the right so that they all ended at the same position. TSDs are in *uppercase letters* and 5′ and 3′ flanking sequences are in *lowercase letters.* Mismatches in the TSDs are also in *lowercase letters.* **A** S1 integration sites without TSD. **B** S1 integration sites with TSDs. **C** Integration site reported by Wright et al. (1996) for a LINE element called Ta11-1 found in *Arabidopsis thaliana.* The Ta11-1 sequence is denoted by ACAA–AAAA.

vil30b, and ma44 sites where DNA from *Brassica oleracea, Brassica incana, Brassica cretica, Brassica villosa,* and *Brassica macrocarpa* was used instead. Empty sites orthologous to *Brassica nigra* S1 sites were amplified using DNA from *Sinapis arvensis.* All empty sites were cloned and sequenced as described above.

χ^2 *Analysis*

The analyses were done either on individual bases or on dinucleotides essentially as described by Jurka (1997). Briefly, $\chi^2 = \Sigma^4 i = 1 (O_i - E_i)^2/E_i$, where O_i is the individual base or dinucleotide occurrences, E_i

is the total number of bases or dinucleotides at a given position × base composition. We used a significance level of $P < 0.01$ for 3 df.

Results

Characterization of S1 Insertion Sites

To study the target site specificity of S1 retroposons, we characterized 64 S1 integration sites from 10 closely related species of crucifers. A preliminary analysis led us to classify these insertion sites in two groups considering the absence (Fig. 1A) or the presence (Fig. 1B) of target site duplication (TSD). Fifty-one of the 64 S1 integration sites have TSD. The number of bases of these TSD seems to be predetermined largely since in most cases (35/51) they range from 14 to 16 bp, with a maximum at 16 bp (see Fig. 1B). We found only one case with an exceptionally long TSD of 35 bp (na17). Only 13/64 (20.3%) of the insertions lack TSD.

In order to get a better understanding of the S1 integration process, we amplified, cloned, and sequenced S1 target sites in closely related *Brassica* species (i.e., orthologous sites). Since S1 retroposition in crucifer is a highly dynamic process leading to a large number of species-specific insertion events (Lenoir et al. 1997), we can easily obtain from closely related species sequence information on target sites before S1 integration (i.e., ''empty'' sites). First, 16 empty sites orthologous to S1 sites with TSD (names in boldface in Fig. 1B) were sequenced. As expected, we did not detect any duplication of the target sequence in these empty sites, suggesting that TSD were generated as a consequence of S1 integration events (not shown). Apart from the TSD, the 16 S1 target sites were unmodified by the integration events. Six empty sites orthologous to S1 sites lacking TSD (names in boldface in Fig. 1A) were next sequenced. Surprisingly, all these empty sites either had an additional sequence and/or lacked a short sequence compared to orthologous S1 sites. The nature and sizes of the deleted or inserted fragments generated upon S1 integration at these sites are given in Fig. 2. We conclude that while insertion generating TSD leaved the target site unaltered, most of the S1 insertions without TSD were associated with small deletions and/or small insertions at the target site.

Target Site Features

We next analyzed in more detail the sites with TSD. We aligned the 5′ and 3′ direct repeats and adjusted them so that all 5′ repeats start or all 3′ repeats end at the same position (Fig. 1B). Since the 16 sites analyzed resulted from an accurate integration process (see above), we assumed that the remaining sites followed this rule. Therefore, in each case, one copy of the TSD was fused to

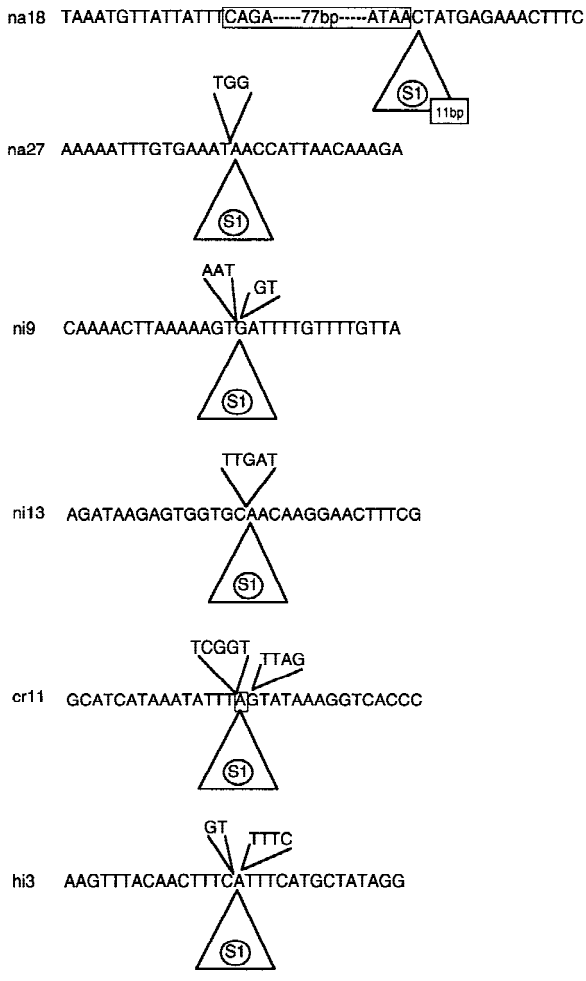

Fig. 2. Characteristics of S1 integration events that do not generate target site duplications. The partial sequence of six ''empty'' sites is shown. Sites of S1 integration are pointed to by the *triangles.* Sequences that were deleted in the integration events are in *boxes* and supplementary sequences that were inserted either 5′ or 3′ of the S1 elements are written *above* each site. The integration event that generated the na18 sites was described previously by Gilbert et al. (1997).

adjacent sequences to reconstitute the site before S1 insertion, and in most cases, 30 bases surrounding each nicking site (15 bases in 5′ and 15 bases in 3′) were analyzed (see Table 1 and Fig. 3). The general feature of these insertion sites is that they are all associated with AT-rich regions (Table 1). This AT richness seems not linked to the integration of S1 elements in very specific genomic regions since we have shown previously that S1 elements are generally distributed on each *Brassica napus* chromosome (unpublished results). Here we also show that they are not found in tandem repeat, at least for the 65 insertion sites characterized. The most striking result is that S1 elements integrate at nonrandom sites. This is particularly true for one of the two nicking sites (i.e., the strong site) that is clearly associated with the 5′-Y/AAANNNG-3′ consensus motif, where Y denotes pyrimidines (Table 1 and Fig. 3A). The signal at the other nicking site seems not to be associated with any obvious consensus motif (Table 1 and Fig. 3B). Thus, S1

Table 1. Base occurrences at different positions of the 5′ and 3′ regions flanking the two S1 nicking sites[a]

	5′ nicking site						3′ nicking site					
N	P	T	C	A	G	Total	P	T	C	A	G	Total
−15	A/T	16	5	16	6	43	A	3	7	38	3	51
−14	A	11	7	17	7	44	A	9	5	35	2	51
−13	A	13	4	26	2	45	A	9	4	30	8	51
−12	A	11	3	25	6	45	A	15	7	23	6	51
−11	A	16	4	19	6	45	A	20	6	22	3	51
−10	A	16	8	18	5	47	T	27	2	14	8	51
−9	T	20	6	14	8	48	A	13	9	17	12	51
−8	T	22	8	11	7	48	A	17	4	20	10	51
−7	A	11	6	18	13	48	T	18	8	17	8	51
−6	T	17	7	13	11	48	A	10	13	18	10	51
−5	T	15	7	14	12	48	A	11	4	26	10	51
−4	T	25	2	15	9	51	A	19	3	22	7	51
−3	A	17	5	20	9	51	A	16	2	23	10	51
−2	A	16	12	18	5	51	T	22	6	17	6	51
→ −1	C	15	21	2	13	51	T	24	8	13	6	51
1	A	2	0	49	0	51	A	15	5	22	9	51 ←
2	A	8	1	41	1	51	A	13	8	22	8	51
3	A	5	13	25	8	51	T	23	6	13	9	51
4	A	18	8	20	5	51	T	19	4	16	12	51
5	A	19	9	20	3	51	A	15	4	20	11	50
6	A	21	3	22	5	51	T	20	7	13	10	50
7	G	11	4	15	21	51	A	16	8	22	4	50
8	A/T	17	10	17	7	51	T	19	9	13	9	50
9	A	15	5	21	10	51	A	13	8	19	10	50
10	A	12	5	24	10	51	T	22	9	16	3	50
11	A	19	6	20	6	51	A	16	9	17	8	50
12	T	21	4	20	6	51	A	16	9	18	6	49
13	A	18	4	21	8	51	A/T	16	8	16	9	49
14	A	14	7	19	11	51	A	16	6	16	10	48
15	A/T	20	6	20	5	51	A	17	5	21	4	47
COMP (%)		31.2	12.9	40.6	15.2			32.2	12.8	39.6	15.3	
AT%				71.8						71.8		
GC%				28.1						28.1		

[a] Only insertion sites with TSDs were analyzed (see Fig. 1B). The presumed positions of the two nicking sites are given by arrows. For the 5′ nicking site, position 1 correspond to the first base of the TSD, while for the 3′ nicking site position −1 corresponds to the last base of the TSD. N, nucleotide position; P, predominant nucleotide; total, number of bases analyzed at each position; T, C, A, and G, frequency of each nucleotide for a given position; COMP, percentage for each nucleotide for the region analyzed.

weak site seems to be less conserved (weaker) than the mammalian one [represented by the 5′-TYTN-3′ motif (Jurka 1997)]. This could happen if the distance between the two nicks is more often imposed by the enzyme involved in S1 integrations compared to the enzyme involved in the integration of mammalian retroposons. In support of this, we observed that although the two nicks involved in Alu and ID integration are usually made within 14–16 base pairs, only 22% (for Alu) and 31% (for ID) of the integration events show this preferred configuration (data from Jurka 1997). For S1, the two nicks are separated by 14 to 16 base pairs in 68% of the integration events, suggesting that this configuration may often be forced by the enzyme itself even at sites poorly resembling a perfect integration sequence, explaining the lower level of conservation of this second (weak) site.

We next studied the distribution of dinucleotides around both nicking sites (Figs. 4 and 5). We made three groups of dinucleotides as described by Jurka et al. (1998): the purine doublets AA, AG, GA, and GG; the pyrimidine doublets TT, TC, CT, and CC; and dinucleotides associated with kinks TA, CA, and TG. For the strong (i.e., more conserved) site (Fig. 4), the dinucleotide frequencies are significantly higher for di-pyrimidines before the nick (position −1), for dinucleotides associated with kinks at the nicking site (position 0) and for di-purine after the nick (positions +1 and +2). For the weak (i.e., less conserved) site (Fig. 5), the frequency of purine and pyrimidine doublets is not significantly higher before position −1) and after (position +1) the nick. The frequency of TA, CA, and TG dinucleotides reaches a maximum at the weak nicking site (position 0) but this peak is barely significant (Fig. 5). Therefore, although cleavage at the weak site obviously depends on the position of the first nick and probably takes place rarely at optimal sites, it could still be done preferentially

at TA, CA, and TG dinucleotides. The poor statistics observed at this position could be the result of the relatively small number of S1 target sites analyzed. These results are very similar to those obtained for mammalian retroposons (Jurka et al. 1998), suggesting that plant retroposons also integrate at sequence-dependent DNA kinks. However, in our case, the primary determinant for the recognition of the weak site is the distance from the first nicks, not the sequence by itself.

Discussion

In previous studies, we observed that S1 elements and mammalian SINEs not only share structural similarities but also have similar patterns of evolution (Deragon et al. 1994; Lenoir et al. 1997; Gilbert et al. 1997). Here we show that, for S1 sites presenting TSD, these similarities extend to the molecular mechanism of integration. The most conserved feature of eukaryotic retroposon integration may be the capacity for the target sequence to form a special DNA structure called kinks. The general A/T richness of a given DNA region and the presence of short runs of pyrimidines followed by short runs of purines could represent a favorable context for retroposition events. In such a context, an endonuclease upon fixation could be able to generate the kink at the pyrimidine/purine transition and to nick the DNA. We observed that cleavage at the weak site occurs at a relatively fixed distance from the strong site. This distance may be imposed by the properties of the endonuclease fixed at the strong nicking site. However, although the second nicking site may represent a weaker recognition site for the same enzyme, it could also be a target for other enzymes linked, for example, to the DNA recombination/repair machinery. This model is compatible with the enzymatic properties of the human L1 endonuclease (Feng et al. 1996) but awaits validation in plant. To evaluate the potential implication of a LINE endonuclease in S1 integration, we intend to purify a *Brassica napus* LINE endonuclease and to test its enzymatic properties on ''empty'' S1 insertion sites. This way, we should be able to determine if the similarities in target site selection observed for plant and mammalian retroposons are linked to the functional conservation of the plant and mammalian LINE endonuclease domain. It is interesting to note that the only integration site reported for a LINE element in crucifer [the Ta11-1 element from *Arabidopsis thaliana* (Wright et al. 1996)] shows the same characteristics as S1 integration sites do (Fig. 1C).

We observed that a small proportion of S1 insertion events was not associated with TSD. These integration events were always associated with small deletions or more often to small insertions. These modifications of the target site have already been reported in several cases of mammalian retroposon integrations (Jurka et al. 1997; Maestre et al. 1995). It is not clear if these integrations result from the action of the endonuclease implicated in ''normal'' (TSD generating) events and if they all can be accounted for by a single mechanism. Small deletions (less than 20 bp) could result from the degradation by an exonuclease of the single-strand regions formed in the ''normal'' integration process. However, deletions at the integration sites (especially the longer ones) could also result from the integration of S1 elements at random double-strand breaks (DSBs) further opened by exonucleases associated with the recombination/repair ma-

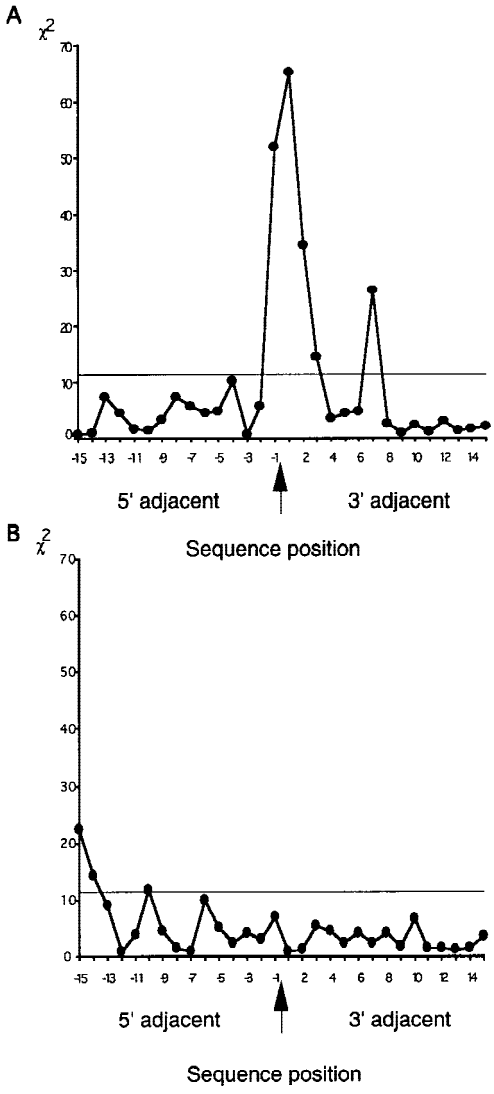

Fig. 3. The χ^2 values for individual positions surrounding S1 nicking sites. From Table 1, χ^2 values were calculated (see Materials and Methods) for **(A)** the region surrounding the 5′ nicking site and **(B)** the region surrounding the 3′ nicking site. The *horizontal lines* correspond to significance levels of $P < 0.01$ for 3 df. Significant values are found only for the region surrounding the 5′ nicking site. Positions −1, 1, 2, 3, and 7 have very high χ^2 values, indicating a nonrandomless distribution of nucleotides at these positions. This can be summarized by the consensus sequence 5′-Y/AAANNNG-3′, or 5′-CNNNTTT/R-3′ on the other DNA strand, which is, according to the retroposition model (Luan et al. 1993; Jurka 1997), the strand that is cleaved. The *arrows* indicate the positions of the hypothetical nicking sites (between nucleotide −1 and nucleotide +1).

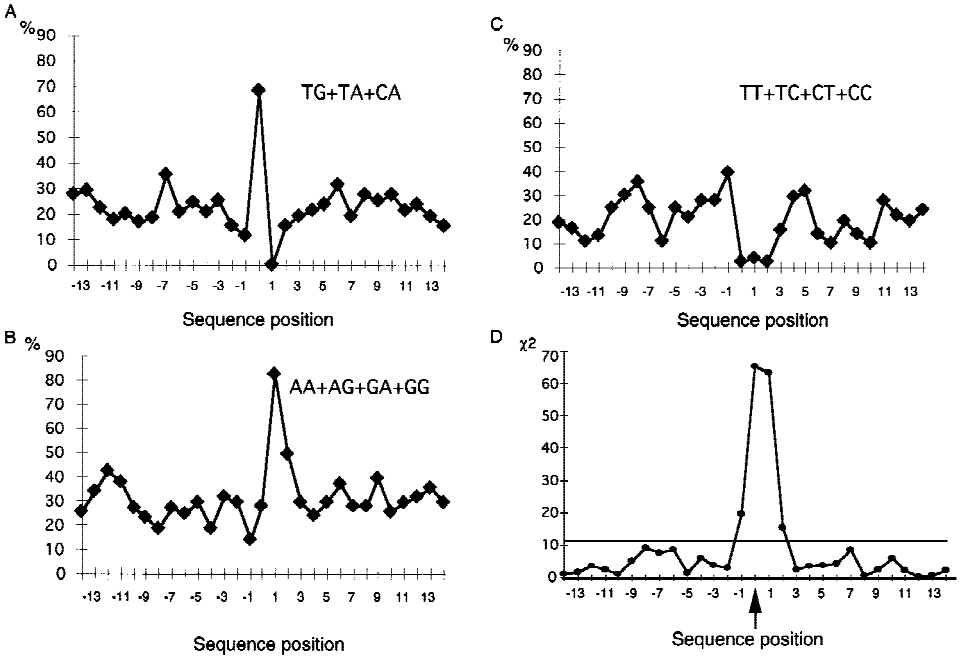

Fig. 4. Dinucleotide distributions surrounding the S1 5′ nicking site. Three groups of dinucleotides were formed: **(A)** the dinucleotides associated with kinks, TG + TA + CA (K); **(B)** the purine dinucleotides AA + AG + GA + GG (dR); and **(C)** the pyrimidine dinucleotides TT + TC + CT + CC (dY). The χ^2 values of the dinucleotide distributions for each position are presented in **D.** The *horizontal line* corresponds to significance levels of $P < 0.01$ for 3 df. The same 30 nucleotides listed in Table 1 were analyzed. χ^2 analysis indicates that positions −1, 0, 1, and 2 are significant. The 5′ nicking site is therefore composed at the dinucleotide level of a peak of di-pyrimidine followed by a kinkable dinucleotide and a strong peak of di-purine, which can be summarized as (dY)(K)(dR)2.

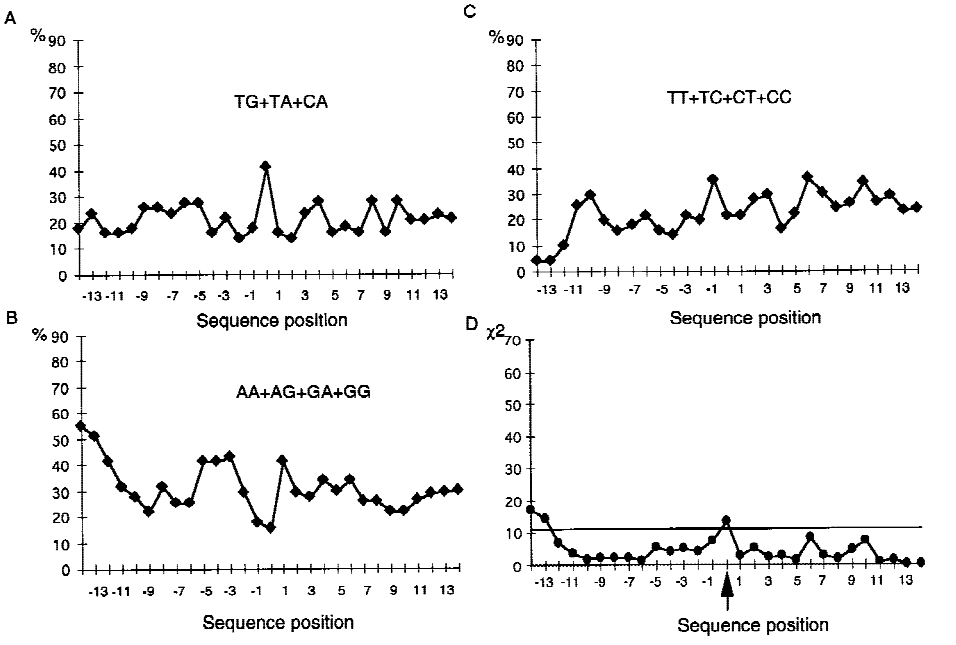

Fig. 5. Dinucleotide distributions surrounding the S1 3′ nicking site. Three groups of dinucleotides were formed as in Fig. 4. **(A)** Dinucleotides TG + TA + CA, **(B)** dinucleotides AA + AG + GA + GG, and **(C)** dinucleotides TT + TC + CT + CC. **D** The χ^2 values of the dinucleotide distributions as in Fig. 4. χ^2 analysis indicates a single position (position 0) slightly above the significance level of $P < 0.01$.

chinery (Haber 1995). The origin of the small insertions observed in several cases is also not clear. Since these supplementary nucleotides were not observed in the 16 empty sites orthologous to S1 sites with TSD, they probably do not result from the initiation of reverse transcription 3′ of the poly(A) tail of a S1 transcript or from the reverse transcription of a longer S1 transcript initiated a few bases 5′ of its normal site. In support of this, we have shown previously that major S1 transcripts initiate only in position +1 and (less frequently) in position −1, sug-

gesting that initiations a few bases 5′ of the normal site are rare events (Deragon et al. 1996). We therefore suggest that these small insertions are strictly associated with particular S1 insertion events that do not generate TSD and are not related to the nature of the S1 transcript. The mechanism responsible for these small insertions is unknown at this time.

A high level of conservation between plants and mammals is unexpected for a ‘‘nonessential’’ process like retroposition. Although a role for reverse transcription in basic biological processes is highly hypothetical at this time, it is worth noting that this enzyme is found in all living organisms from bacteria to higher eukaryotes. Therefore, it is hard to imagine that in all cases this enzyme owes its presence only to parasitic expansions of transposable elements and provides no benefit to the host. One recent finding suggests that reverse transcription could be implicated in double-strand break (DSB) repair in eukaryotes (Moore and Haber 1996; Teng et al. 1996). DSBs are believed to be repaired by several pathways of homologous recombination and by nonhomologous end-joining. In yeast, homologous recombination is a very efficient process which may account for most DSB repair, but recent experiments suggested that when homologous recombination is inhibited, most of the DSBs are repaired by nonhomologous end-joinings (similar to those observed in mammalian cells) and used the Ty1 retrotransposon element (Moore and Haber 1996; Teng et al. 1996). These results suggest that integration of retroposons could be used as one strategy to repair DSBs in eukaryotes. Another interesting finding comes from the implication of the human L1 reverse transcriptase in the generation of high levels of cytoplasmic cDNA molecules expressed in human cells (Dhellin et al. 1997). This work clearly showed that the cDNAs generated by this cytoplasmic reverse transcription are not coupled with integration and are probably not retroposition intermediates. They also showed that L1 RNA and RNA from different cellular genes were reverse transcribed with similar efficiency. These results suggest that cytoplasmic and nuclear (in situ) reverse transcription are distinct processes and that cytoplasmic reverse transcription could be implicated in other biological processes, for example, the regulation of mRNA translation. If these recent findings are really of biological significance, then the coding function of LINE elements should be under positive selection pressure. Furthermore, the amplification mechanism of LINE and SINE elements, which also rely on these proteins, would tend to be conserved among a wild range of eukaryotic species.

Acknowledgments. We thank Chantal Goubely, Philippe Arnaud, Alain Lenoir, and Charles Poncet for their technical help. We also thank Suzanne Warwick for providing us with DNAs from the various *Brassica* species. This work was supported by the CNRS (UMR 6547), by the Université Blaise Pascal, and by a European Community grant (FPIV, Molecular Tools for Biodiversity).

References

Boeke JD (1997) LINEs and Alus: the polyA connection. Nature Genet 16:6–7

Deragon J-M, Landry BS, Pélissier T, Tutois S, Tourmente S, Picard G (1994) An analysis of retroposition in plants based on a family of SINEs from *Brassica napus.* J Mol Evol 39:378–386

Deragon J-M, Gilbert N, Rouquet L, Lenoir L, Arnaud P, Picard G (1996) A transcriptional analysis of the S1Bn (*Brassica napus*) family of SINE retroposons. Plant Mol Biol 32:869–878

Dhellin O, Maestre J, Heidmann T (1997) Functional differences between the human LINE retrotransposon and retroviral reverse transcriptase for in vivo mRNA reverse transcription. EMBO J 16: 6590–6602

Feng Q, Moran JV, Kazazian HH, Boeke JD (1996) Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. Cell 87:905–916

Gilbert N, Arnaud P, Lenoir A, Warwick SI, Picard G, Deragon JM (1997) Plant S1 SINEs as a model to study retroposition. Genetica 100:155–160

Haber JE (1995) In vivo biochemistry: Physical monitoring of recombination induced by site-specific endonucleases. Bioessays 17:609–620

Jurka J (1997) Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. Proc Natl Acad Sci USA 94:1872–1877

Jurka J, Klonowski P (1996) Integration of retroposable elements in mammals: Selection of target sites. J Mol Evol 43:685–689

Jurka J, Klonowski P, Trifonov EN (1998) Mammalian retroposons integrate at Kinkable DNA sites. J Biomol Struct Dyn 15:1–5

Kachroo P, Leong SA, Chattoo BB (1995) Mg-SINE: A short interspersed nuclear element from the rice blast fungus, Magnaporthe grisea. Proc Natl Acad Sci USA 92:11125–11129

Kido Y, Aono M, Yamaki T, Matsumoto K, Murata S, Saneyoshi M, Okada N (1991) Shaping and reshaping of salmonid genomes by amplification of tRNA-derived retroposons during evolution. Proc Natl Acad Sci USA 88:2326–2330

Lenoir A, Cournoyer B, Warwick SI, Picard G, Deragon J-M (1997) Evolution of SINE S1 retroposons in Cruciferae plant species. Mol Biol Evol 14:934–941

Luan DD, Korman MH, Jakubczak JL, Eickbush TH (1993) Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. Cell 72:595–605

Maestre J, Tchénio T, Dhellin O, Heidmann T (1995) mRNA retroposition in human cells: Processed pseudogene formation. EMBO J 14:6333–6338

McNamara PT, Bolshoy A, Trifonov EN, Harrington RE (1990) Sequence-dependent kinks induced in curved DNA. J Biomol Struct Dyn 8:529–538

Moore JK, Haber JE (1996) Capture of retrotransposon DNA at the sites of chromosomal double-strand breaks. Nature 383:644–646

Teng SC, Kim B, Gabriel A (1996) Retrotransposon reverse transcriptase-mediated repair of chromosomal breaks. Nature 383:641–644

Warwick SI, Black LD (1991) Molecular systematics of *Brassica* and allied genera (subtribe Brassicinae; Brassiceae)-chloroplast genome and cytodeme congruence. Theor Appl Genet 82:81–92

Winkler FK, Banner DW, Oefner C, Tsernoglou D, Brown RS, Heathman SP, Bryan RK, Martin PD, Petratos K, Wilson KS (1993) The cristal structure of EcoRV endonuclease and of its complexes with cognate and non-cognate DNA fragments. EMBO J 12:1781–1785

Wright DA, Ke N, Smalle J, Hauge BM, Goodman HM, Voytas DF (1996) Multiple non-LTR retrotransposons in the genome of Arabidopsis thaliana. Genetics 142:569–578

Yoshioka Y, Matsumoto S, Kojima S, Ohshima K, Okada N, Machida Y (1993) Molecular characterization of a short interspersed repetitive element from tobacco that exhibits sequence homology to specific tRNAs. Proc Natl Acad Sci USA 90:6562–6566