The EMBO Journal vol.6 no.13 pp.3873–3880, 1987

Cin4, an insert altering the structure of the *A1* gene in *Zea mays*, exhibits properties of nonviral retrotransposons

Zsuzsanna Schwarz-Sommer, Lise Leclercq, Elke Göbel and Heinz Saedler

Max-Planck-Institut für Züchtungsforschung, 5000 Köln 30, FRG

Communicated by H.Saedler

A wild-type allele of the *A1* gene of *Zea mays* contains a 1.1-kb-long insert termed Cin4-1, which alters the structure of the transcription unit compared to other *A1* alleles. The Cin4-1 element is a member of a family of elements occurring in 50–100 copies in the maize genome. Genomic cloning and sequence analysis of several family members and their flanking regions allowed classification of Cin4 as a nonviral retrotransposon. Individual Cin4 elements terminate in an oligo(A) track of variable size (6–11 residues) at their 3′-end. The 5′-ends of family members are heterogeneously truncated with respect to the longest Cin4 element. Cin4 elements are flanked by small direct duplications, the size of which varies between 3 and 16 bp. On the basis of a comparison of the target sequence and the sequence of Cin4 we suggest and discuss a model of the mechanism of Cin4 integration via *in situ* cDNA synthesis on an RNA template. The longest Cin4 element analysed so far has two non-overlapping open reading frames (ORFs) comprising 2793 nucleotides (ORF1) and 3489 nucleotides (ORF2). The putative 1163 amino acid long Cin4 protein derived from the sequence of ORF2 has the capacity to encode a reverse transcriptase-like protein and a DNA-binding domain. The conservation pattern of these two domains and the overall organisation of Cin4 is similar to that detected in nonviral retrotransposons in animals. The origin and function of Cin4 are discussed.

Key words: plant/retrotransposition/mechanism of integration

Introduction

Transposition via a DNA intermediate is a universal phenomenon in prokaryotes and in eukaryotes allowing rapid reorganization of their genomes. In contrast, RNA-mediated mobility of genetic information seemed for a long time to be an attribute to higher eukaryotes since it was restricted mainly to mammals (Rogers, 1985). The detection of reverse transcriptase activity or homology to reverse transcriptase-like proteins in so-called viral retrotransposons such as the *Ty* elements of yeast (Boeke *et al.*, 1985), the *copia* element of *Drosophila* (Saigo *et al.*, 1984) and *CaMV*, a retroid DNA virus in plants (Pfeiffer and Hohn, 1983; Fuetterer and Hohn, 1987), indicates that reverse transcriptase-governed transposition via an RNA intermediate may be a more general phenomenon. In fact, dispersed repetitive sequences in wheat (Harris and Flavell, 1986; Harberd *et al.*, 1987) and insertion elements in maize (Blumberg, 1985; Johns *et al.*, 1985) show some characteristics of viral retrotransposons.

Until recently, other potential nonviral retrotransposons, in particular processed pseudogenes and similar dispersed repetitive, intronless sequences terminating in a poly(A) track (for terminology of retroid elements see Weiner *et al.*, 1986) have only been found in mammals. There is, however, circumstantial evidence for the occurrence of pseudogenes in yeast (Fink, 1987) and a single-copy processed actin pseudogene in potato has also been reported (Drouin and Dover, 1987).

One class of such ‘retroid’ sequences, termed LINEs (or L1 family), deserves special attention. These elements are highly reiterated in the genome of mammals (for review see Rogers, 1985). DNA sequence analysis of L1 elements from human (Hattori *et al.*, 1986), mouse (Loeb *et al.*, 1986), cat (Fanning and Singer, 1987) and rabbit (Demers *et al.*, 1986) genomes indicates general principles in the organization of these sequences. The most striking feature of LINEs and related sequences is the conservation of a region within a long open reading frame (ORF) with homology to retroviral reverse transcriptases (Hattori *et al.*, 1986). ‘Retroid’ elements of this type belong to the class of nonviral retrotransposons (Weiner *et al.*, 1986). A similar class of elements has recently been identified in non-mammals, the I elements of *Drosophila* (Fawcett *et al.*, 1986) and the Ingi elements of *Trypanosoma* (Kimmel *et al.*, 1987). L1, Ingi and I elements share no sequence homology but display similarity in their overall structural organization and in the presence of the highly conserved region within a long ORF mentioned above. These data suggest that such elements are functionally related and that they originated from a common progenitor during evolution.

Here we provide evidence that the Cin4 element in maize is analogous to L1, Ingi and I elements in all characteristic features. This homology defines Cin4 as the first nonviral retrotransposon with coding capacity for reverse transcriptase identified in plants.

Results

Detection of Cin4 as an insert within the A1 transcription unit

We previously reported that the wild-type *A1* gene of maize has two allelic forms termed type 1 and type 2 (Schwarz-Sommer *et al.*, 1987). The type 2 allele differs from type 1 by a 1092-bp-long insert called the Cin4-1 element present in the 3′-untranslated region of the former (Figure 1). As determined by sequence analysis of type 2 cDNAs, termination of transcription in the type 2 allele occurs within the Cin4-1 sequence. The two *A1* alleles, therefore, contain unrelated sequences within the last 200-bp of their 3′-end.

DNA sequence analysis of Cin4-1 revealed several intriguing features: in the type 2 allele the insert is flanked by a direct duplication of 7 bp occurring only once in the type 1 allele (Figure 1, bottom line). This indicates that integration of Cin4 was associated with generation of a 7-bp-long staggered nick within the type 1 allele. Integration by generation of staggered nicks is characteristic for transposons which, in addition, often show other common structural features like terminal repeats (for review see Nevers *et al.*, 1986; Schwarz-Sommer, 1987). Cin4-1, however, does not possess structural termini (sequence data not shown). We found instead 11 A residues at one end. A potential

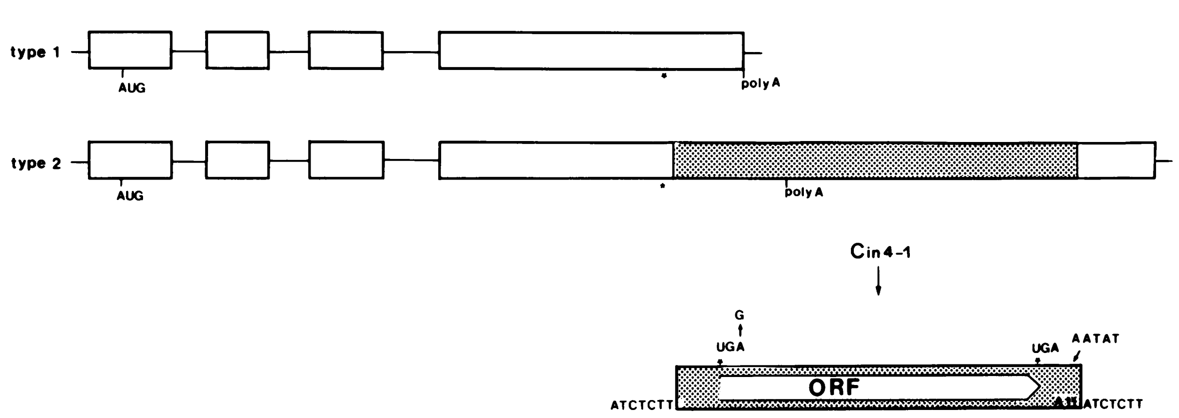

Fig. 1. The effect of the 1.1-kb-long Cin4-1 insert on the structure of a wild type *A1* gene of *Zea mays*. Exons within the type 1 and type 2 alleles are represented by boxes and introns by horizontal lines (for details on the structure of the *A1* gene see Schwarz-Sommer *et al.*, 1987). Start (AUG) and stop (*) of translation and the position of the polyadenylation site are indicated below the scheme. The Cin4-1 element is shown as a shaded box within the fourth exon. The overall structure of the insert is depicted at the bottom of the figure. Sequence data are available upon request. Translation stops bordering the long ORF are indicated above the scheme. A point mutation (G−A) which introduces the translation stop at the beginning of the ORF of Cin4-1 distinguishes Cin4-1 from all other cloned Cin4 elements. The short duplication flanking the insert in the *A1* gene, the position of a putative polyadenylation signal and the size of the oligo(A) track at the 3-end of Cin4-1 are also depicted.

polyadenylation signal AATAT precedes the poly(A) track of Cin4-1 by 38 bp (Figure 1, Proudfoot and Brownley, 1976) suggesting that it originated from a polyadenylated transcript.

These observations and also the fact that Cin4-1 contains an 0.8-kb-long ORF (Figure 1) prompted us to ask whether Cin4-1 represents a family of elements in plants which move by reverse transcription of an RNA intermediate.

An 0.6-kb-long segment of Cin4-1 is conserved in 80% of its 50−100 copies in the maize genome

The copy number of Cin4-1-related sequences is about 50−100 per diploid maize genome (Figure 2). This rough estimate is based on the number (and intensity) of bands of varying size which light up in Southern blots using the central portion of Cin4-1 as a probe, and when genomic maize DNA is digested with an enzyme not restricting Cin4-1 (Figure 2C, left panel). The number of copies within a single band cannot be estimated with accuracy. As also documented in Figure 2, the hybridization pattern of *Bam*HI fragments carrying Cin4-1-homologous sequences varies when comparing different maize lines. This might indicate frequent changes in the chromosomal location of the element. Whether these are due to transposition of Cin4 or reflect restriction fragment length polymorphism of chromosomal *Bam*HI sites cannot be deduced from such experiments.

The most striking result of Southern blot experiments is that a 0.6-kb-long *Hinf* fragment from Cin4-1 is conserved in at least 80% of the genomic copies (Figure 2C, right panel). This portion of Cin4-1 is part of the long ORF within the element (Figure 1, discussed in the following section).

Independent Cin4 copies are truncated at their 5′-ends

We have cloned and analysed 30 genomic copies of Cin4. Using a probe derived from the 5′-portion of Cin4-1 one can identify longer elements with homology extending further 5′ than the endpoint of Cin4-1. These elements are identical in structure at their 3′-ends, as indicated by restriction analysis and in some cases also by DNA sequence analysis. Cin4-1 therefore is a 5′-truncated copy of a larger element.

By deriving new '5′-prime' probes from these longer copies, still longer elements can be identified. As far as we can tell from restriction analysis and limited sequence analysis the variously truncated Cin4 copies are co-linear with each other (see Figure 2). In some copies we detected small duplications but large internal inversions and duplications also occur (data not shown). In addition, some of the genomic clones contain Cin4 elements at distinguishable chromosomal positions which arose by duplication of regions extending several kb upstream and downstream of Cin4-homologous sequences. Outside of these homologous regions these clones contain unrelated sequences.

During analysis of truncated Cin4 copies we found no evidence for 3′-truncation. Their occurrence in the maize genome seems to be unlikely since several Cin4 copies which were identified by hybridization to a region 400 bp upstream of the 3′-end contained the same 3′-terminus ending in a poly(A) track as proven by sequence analysis. Recombinant clones which contain regions upstream of Cin4-1 but do not hybridize to the actual 3′-end of the element were incomplete genomic clones which arose during partial digestion of the plant DNA with *Mbo*I.

In summary, we assume that the majority of Cin4 copies occur in a 5′-truncated form. This assumption is further supported by Southern blot experiments. As shown in panels A and B of Figure 2, the number of bands decreases when fragments derived from the 5′-region of Cin4-15 are used as probes.

5′-Truncated Cin4 elements are flanked by small direct repeats

As compiled in Figure 3, the six variously truncated Cin4 elements analysed so far are all flanked by small duplications (as reference for the position of the 5′-end of the analysed elements with respect to each other, see vertical arrows in Figure 2). The size of the duplicated region varies between 3 and 16 bp as does the size of the oligo(A) track upstream of the duplication at the 3′-end. These analyses indicate that integration of Cin4 is preceded by generation of staggered nicks of variable size. The data also suggest that the truncation process occurs *prior* to (or during) integration and is not due to deletions affecting the 5′-region of an already integrated full size element.

Closer inspection of the duplicated sequences revealed that the targets of Cin4 integrations are somewhat homologous to sequences of the element present in the contiguous 5′-region of longer copies (bases in shaded boxes below the target sequence in Figure 3). In some integrated copies this homology extends further than the target duplication. This feature suggests target site specificity of Cin4 integrations. It also prevents determination of the exact size of the target duplications since part of the flanking sequences can either belong to Cin4 or to the target.

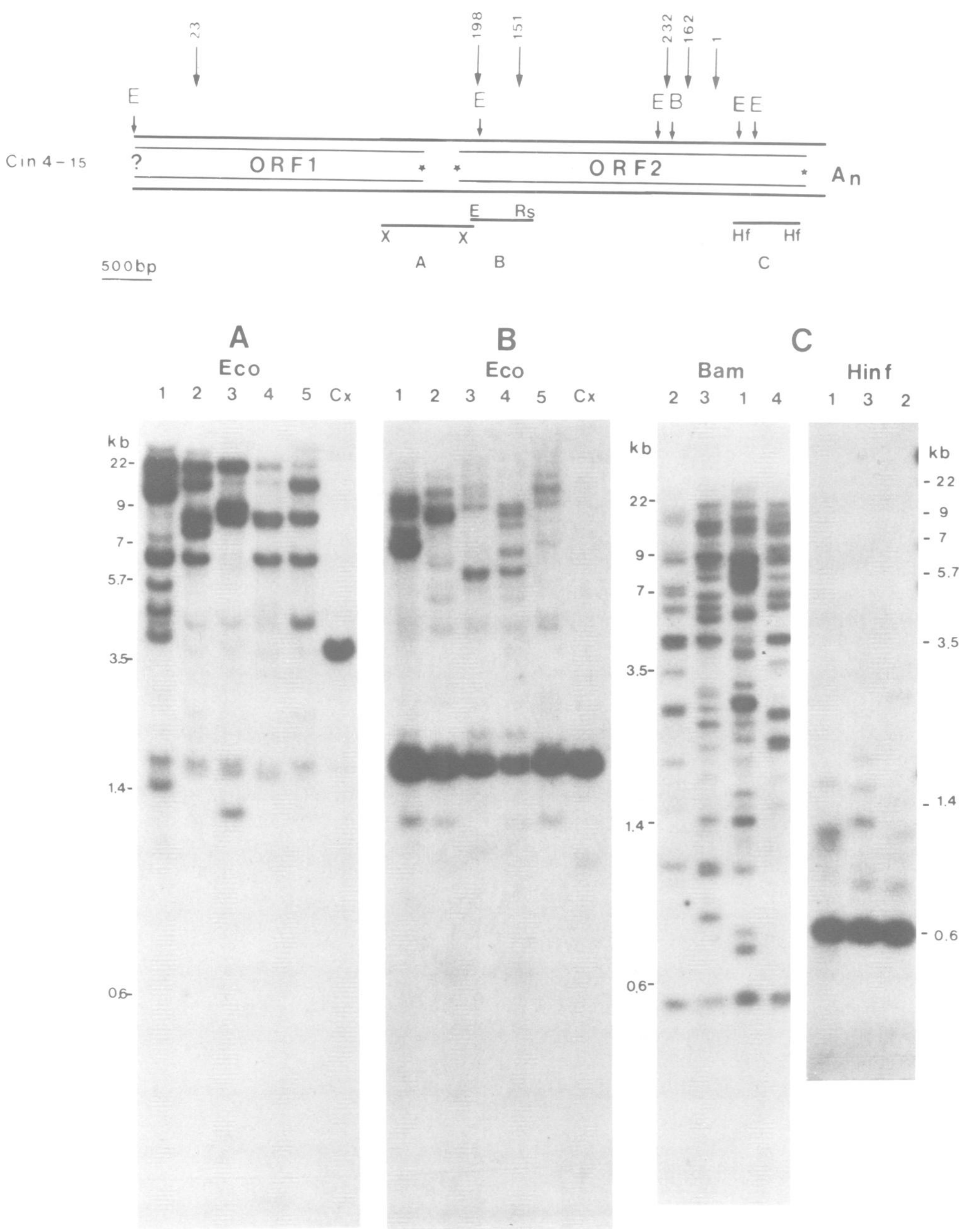

Fig. 2. Conservation of Cin4-1-related sequences in the maize genome. The scheme at the top represents the Cin4-15 element, which is the largest Cin4 copy analysed so far. Note that the *Eco*RI site (E) at the left was introduced during cloning and hence it is unique to the Cin4-15 clone. All other *Eco*RI sites are either still present in Cin4-elements or are absent due to truncation. The same holds for the *Bam*HI (B) site. The truncation endpoints of other copies are indicated by large vertical arrows, and the designation of the cloned copy is shown. For further details see Results and Figure 3. The scheme also indicates the position of the ORF. Asterisks illustrate translation stops. The question mark indicates that the Cin4-15 clone ends at this position and that we have not yet analysed longer Cin4 elements. A_n stands for the oligo(A) track defining the 3′-end of Cin4. The size of the radioactive probes used for Southern experiments and their positions within the Cin4 element are shown below the scheme. Their designation as **A, B** and **C** corresponds to their use as probes in different experiments as indicated above the panels. Capital letters stand for the restriction enzymes used to prepare the probes (X = *Xba*; E = *Eco*RI; Rs = *Rsa*; Hf = *Hinf*). For the experiments documented in this Figure, 7 μg genomic DNA were digested with *Eco*RI (panels **A** and **B**), *Bam*HI (panel **C**, left) or *Hinf* (panel **C**, right). **Lane 1**: *Zea mays* ssp. *parviglumis* (teosinte Guerrero), **lanes 2–5**: *Zea mays* ssp. *mays* (2: Line C, 3: *a1-m1* 5179, 4: *a1-mdt*, 5: *a1-pale mr*, see Materials and methods). The control lanes (Cx) contain 7 μg of *Eco*RI-digested genomic *Antirrhinum majus* DNA (which does not contain Cin4-homologous sequences) together with *Eco*RI-digested DNA from the recombinant phage carrying Cin4-15. The amount of phage DNA corresponds to approximately 20 copies. After ethanol precipitation the DNA was loaded onto agarose gels [1% for panel **A** and **B**, 0.8% for panel **C** (left) and 1.5% for panel **C** (right)]. To avoid smearing of bands the autoradiographs presented in panel **C** are underexposed as compared to those presented in panels **A** and **B**. For copy number estimation we assumed that the weak bands in panel **C** (left lane) represent single copies. For further technical details see references quoted in Materials and methods.

Cin4 contains two long ORFs with homology to conserved domains in nonviral retrotransposons

During sequence analysis of several truncated Cin4 elements we realized that the ORF detected in Cin4-1 extends even further upstream in larger elements (see ORF2 in Figure 2). Cin4-1 contains a point mutation introducing a stop codon not present in other copies (Figure 1). Combining overlapping DNA sequences derived from the analysis of seven independent copies revealed that Cin4 contains two long non-overlapping ORFs of 2793 bp

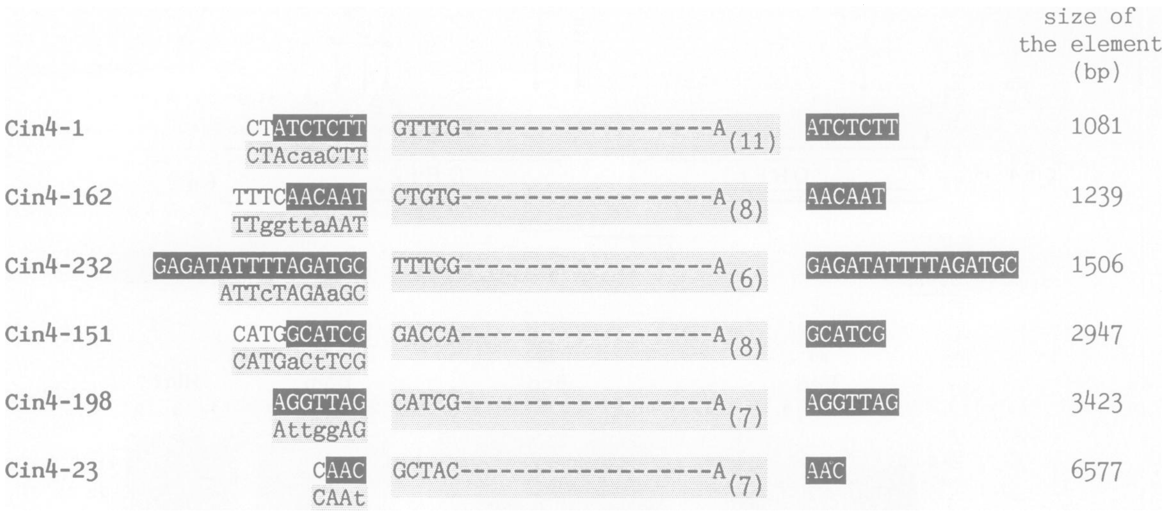
size of
the element
(bp)
Cin4-1 CTATCTCTT CTAcaaCTT GTTTG-------------------A(11) ATCTCTT 1081
Cin4-162 TTTCAACAAT TTggttaAAT CTGTG-------------------A(8) AACAAT 1239
Cin4-232 GAGATATTTTAGATGC ATTcTAGAaGC TTTCG-------------------A(6) GAGATATTTTAGATGC 1506
Cin4-151 CATGGCATCG CATGaCtTCG GACCA-------------------A(8) GCATCG 2947
Cin4-198 AGGTTAG AttggAG CATCG-------------------A(7) AGGTTAG 3423
Cin4-23 CAAC CAAt GCTAC-------------------A(7) AAC 6577

Fig. 3. Target site duplications flanking individual truncated Cin4 elements. The 5′-ends of individual Cin4 elements were determined by sequence comparison with the truncated element Cin4-23 the 5′-end of which was defined by sequence comparison with the further upstream extending Cin4-15 element (for positions of the truncation sites with respect to Cin4-15, see Figure 2). The total lengths of the truncated elements in bp are indicated at the right of the Figure. The flanking short duplications are shown in black boxes. Shaded boxes represent sequences contained in Cin4. Capital letters in these boxes indicate identical nucleotides present within the chromosomal target and within the contiguous sequence in longer Cin4 elements. The sequence at the 5′-end of individual elements and the number of A residues (in brackets) at the 3′-end of the elements are also indicated by capital letters.

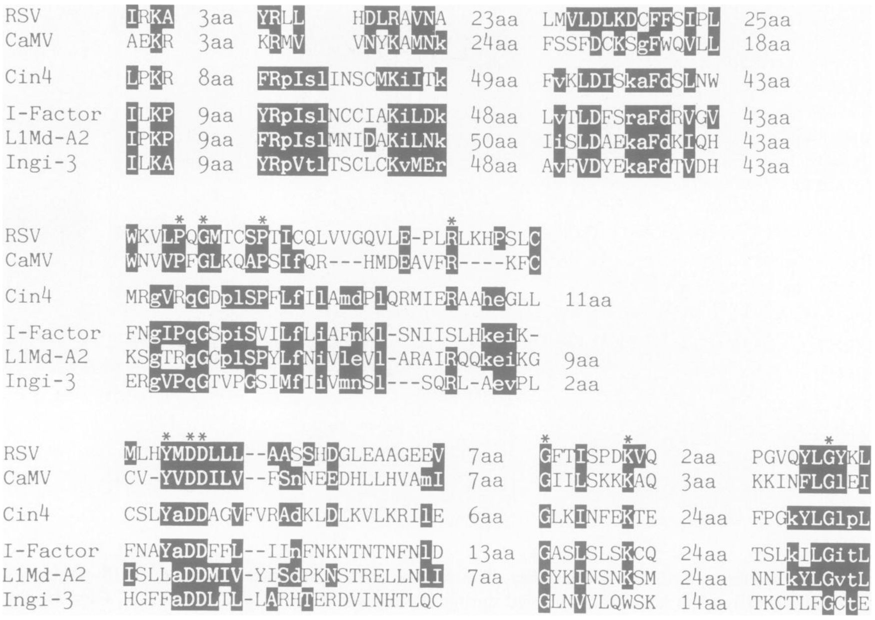
RSV IRKA 3aa YRLL HDLRAVNA 23aa LMVLDLKDCFFSIPL 25aa
CaMV AEKR 3aa KRMV VNYKAMNK 24aa FSSFDCKSgFWQVLL 18aa
Cin4 LPKR 8aa FRpIslINSCMKiITk 49aa FvKLDISkaFdSLNW 43aa
I-Factor ILKP 9aa YRpIslNCCIAKiLDk 48aa LvILDFSraFdRVGV 43aa
L1Md-A2 IPKP 9aa FRpIslMNIDAKiLNk 50aa IiSLDAEkaFdKIQH 43aa
Ingi-3 ILKA 9aa YRpVtlTSCLCKvMEr 48aa AvFVDYEkaFdTVDH 43aa
* * * *
RSV WKVLPQGMTCSPTICQLVVGQVLE-PLRLKHPSLC
CaMV WNVVPFGLKQAPSIfQR---HMDEAVFR----KFC
Cin4 MRgVRqGDplSPFLfIlAmdPlQRMIERAAheGLL 11aa
I-Factor FNgIPqGSpiSVILfLiAFnKl-SNIISLHkeiK-
L1Md-A2 KSgTRqGCplSPYLfNiVleVl-ARAIRQQkeiKG 9aa
Ingi-3 ERgVPqGTVPGSIMfIiVmnSl---SQRL-AevPL 2aa
* ** * * *
RSV MLHYMDDLLL--AASSHDGLEAAGEEV 7aa GFTISPDKVQ 2aa PGVQYLGYKL
CaMV CV-YVDDILV--FSnNEEDHLLHVAmI 7aa GIILSKKKAQ 3aa KKINFLGlEI
Cin4 CSLYaDDAGVFVRAdKLDLKVLKRIlE 6aa GLKINFEKTE 24aa FPGkYLGlpL
I-Factor FNAYaDDFFL--IInFNKNTNTNFNlD 13aa GASLSLSKCQ 24aa TSLkILGitL
L1Md-A2 ISLLaDDMIV-YISdPKNSTRELLNlI 7aa GYKINSNKSM 24aa NNIkYLGvtL
Ingi-3 HGFFaDDLIL-LARHTERDVINHTLQC GLNVVLQWSK 14aa TKCTLFGCtE

Fig. 4. Amino acid sequence homology within the conserved region of known and putative reverse transcriptases. The amino acid sequence of the conserved domain within the ORF of Cin4 is aligned with the conserved domain of the polymerase gene product of Rous sarcoma virus (RSV, Schwartz *et al.* 1983) and with the putative polymerase gene product of cauliflower mosaic virus (CaMV, Gardner *et al.*, 1981). A similar alignment was first pointed out by Hattori *et al.* (1986) for the ORF2 of L1 elements in primates. The figure includes the homology boxes found within the ORF2 of other nonviral retrotransposons (see Fawcett *et al.*, 1986 for I elements, Loeb *et al.*, 1986 for L1Md-A2 elements and Kimmel *et al.*, 1987 for Ingi-3 elements). Conservative positions are indicated by dark boxes and correspond to the conservation first detected by Toh *et al.* (1983). The ten invariant amino acids found by these authors are indicated by asterisks. Positions conserved in at least three of the four nonviral retrotransposons are shown in dark boxes by small letters. For comparing sequences the same groups of residues have been used as by Fawcett *et al.* (1986): P, A, G, S, T (neutral or weakly hydrophobic); Q, N, E, D (hydrophylic, acid amine); H, K, R (hydrophilic, basic); L, I, V, M (hydrophobic); F, Y, W (hydrophobic, aromatic); C (cross-link forming). Gaps were introduced to increase similarity and numbers give the distance in bp between homology blocks. For the location of this conserved domain within the ORFs of Cin4 and of other nonviral retrotransposons see Figure 6.

(ORF1) and 3489-bp (ORF2) lengths (sequence data available upon request, for compilation see Figures 2 and 6). The question mark at the beginning of ORF1 in Figure 2 (and Figure 6) reflects our uncertainty about the length of the authentic Cin4 ORF1.

Sequence comparison of large overlapping areas obtained from independent Cin4 clones revealed that the copies show more than 90% homology to each other. According to these data the Cin4 copies represent pseudogenes.

Computer screening at the protein or at the nucleic acid level did not reveal homology of the Cin4 ORFs with any known sequences in the EMBL or in the NBRF data bank. However, a

```
Nucleic acid binding "fingers"

RSV              C QL--  C ---------NGMG  H NAKQ  C R
MoMSV            C TY--  C ---------EEQG  H WAKD  C P
CaMV             C WI--  C ---------NIEG  H YANE  C P
TFIIIA        YI C ****  C DKRFTKK**LKR  H ****  H

ORF1

I-Factor^a       C KK--  C ---------LRFG  H PTPI  C K
I-Factor^b       C IN--  C ---------SETK  H
I-Factor^c       C LN--  C -----RNNPELD  H
Cin4^a           C YN--  C ---------LSPD  H LAFR  C S
Cin4^b           C WQ--  C ---------LHFG  H RARA  C P

ORF 2

L1Md-A2          C WRG-  C GERGTLL-----  H CWWE  C R
I-Factor         C PF--  C QGDISLN-----  H IFNS  C P
Cin4             C CL--  C NLSQESMP----  H LGKD  C P


Ingi-3^a      TK C **    C DATYQCR**AVT  H ****  H
Ingi-3^b      LH C **    C TSKFAVP**LLH  H ****  H
Ingi-3^c      FQ C **    C EASFGTR**LSL  H ****  H
```

Fig. 5. Amino acid sequence homology within a conserved region of nonviral retrotransposons with putative nucleic acid-binding 'fingers'. This homology domain was first detected within ORF2 of L1 elements by Fanning and Singer (1987) and within ORF1 of I elements by Fawcett *et al.* (1986). We screened for similar regions within the sequence of Cin4 (sequence data available upon request) and the published sequence of Ingi-3 (Kimmel *et al.*, 1987) using the conservation pattern of putative DNA-binding 'fingers' obtained within the *gag* region of retroviruses (Covey, 1986) and repeatedly within the zinc-binding domain of TFIIIA (Miller *et al.*, 1985). The conservative cystein and histidine residues are in dark boxes. Dashes indicate that spacing between conserved positions is not stringently conserved (see Covey *et al.*, 1986). Asterisks stand for conserved spacing. a, b and c are repeated conserved domains in the ORF1 of Cin4 and I elements and in the ORF2 of Ingi-3.

high degree of homology (57%) could be detected by comparing ORF2 with the sequence of conserved amino acids within a domain of the *pol* region of retroviruses, which supposedly corresponds to reverse transcriptase (Toh *et al.*, 1983; see Figure 4). The conservation includes nine out of the ten invariant amino acids also detected in putative polymerases of other viruses. Cin4 shows a larger degree of homology to the reverse transcriptase domain of nonviral retrotransposons than to the viral reverse transcriptase domains, with respect to the spacing between homology blocks and to the presence of additional conserved amino acid positions (Figure 4).

A second conserved domain is located downstream of the reverse transcriptase homology box within the ORF2 of Cin4 which shows homology to putative nucleic-acid-binding 'fingers' (Figure 5). This conserved domain is also present within ORF1 (Figure 5). The ORF1 of Cin4 thus resembles the *gag* region of retroviruses which also contains such 'fingers' (Covey, 1986; Fuetterer and Hohn, 1987).

Expression of Cin4-related sequences and the problem of the 'full-size' Cin4 element

At present we have no clear experimental evidence for the transcription of an 'authentic' Cin4 copy which would initiate at an internal promoter. The following section provides some insights on the expression of Cin4-related sequences without documenting details of the experimental results.

The identification of a full-length Cin4 element as a cDNA suggests both its occurrence in the genome as well as its transcription. In the case of the L1 family in humans, Skowronski and Singer (1985) found large transcripts in an embryonal stage NTera2D1 cell line which were co-linear with large L1 elements. L1 was not at all, or only weakly, expressed in differentiated tissues. Based on this observation, we used RNA from somatic maize embryos to prepare a cDNA library, which was then screened with sequences contained in Cin4. Although we were able to isolate about 200 clones with homology to the 3'-region of Cin4 (with Cin4-1 as a probe), none of the 500 000 recombinant phages hybridized to probes derived from upstream region of longer Cin4 elements. This result indicates that either the full-sized Cin4 element is not expressed in this tissue or that it is not expressed in the developmental stage at which the embryos were harvested. Similar cDNA cloning experiments were also carried out with RNA originating from young maize leaves. The abundance of cDNAs showing homology to Cin4-1 is about ten times lower than in somatic embryos and again none of the 20 positive clones hybridized to sequences located further upstream in large Cin4 elements.

By restriction and hybridization to Cin4-1 we analysed 24 randomly selected phages from the somatic embryo clones and 12 clones selected from the leaf cDNA library. In addition, large areas of the inserts of four somatic embryo DNA clones and six leaf cDNA clones were sequenced. The results can be briefly summarized as follows. The Cin4-homologous regions contained in the cDNAs are transcribed from an external promoter as indicated by additional sequences upstream of the homology region. Cin4 elements can insert in either direction with respect to transcription. The Cin4-homologous regions are co-linear with the sequence of Cin4-1 and only one cDNA contained larger rearrangements shown to be deletions. None of the cDNAs were identical to each other in structure indicating that none of the sequences is abundantly transcribed. In agreement with these observations Northern experiments carried out with the same probe and the same RNA revealed a smear.

The conclusion of these experiments is that integration of Cin4 very likely occurs preferentially into transcribed regions because the number of the different cDNA clones corresponds to the number of integrated Cin4 copies. The experiments also indicate that the full-size Cin4 element is not transcribed in the tissues investigated.

We note here that it is difficult to find criteria other than expression to define the 'master' Cin4 element. We use the presence of target duplications flanking the element and the occurrence of the poly(A) track at the 3'-end to define the ends of an integrated copy. By this definition, the longest Cin4 copy which is flanked by target duplication (Cin4-23 in this report) represents a truncated element because its structure could only be elucidated by comparing its 5'-end with the sequence of a still longer copy (Cin4-15, see Figures 2 and 6). The 5'-end of the longest copy remains undefined. Thus, although these are the only criteria we can follow experimentally at the moment, they are biased by definition. This bias also holds true for defining the ends of other nonviral retrotransposons like the L1 elements or the Ingi-3 element. The I element of *Drosophila* is the only exception. In this case the dysgenic nature of stocks containing independent I elements of identical size within the *white* locus gives unique functional proof for the integrity of the element. Other assumptions necessary in definition of an intact nonviral retrotransposon imply unknown processes which are related to their origin from viral retrotransposons or even from retroviruses. We are presently looking for Cin4 copies which do not terminate in an oligo(A) track and which are bordered by long terminal repeats flanked by small target duplications outside these repeats.

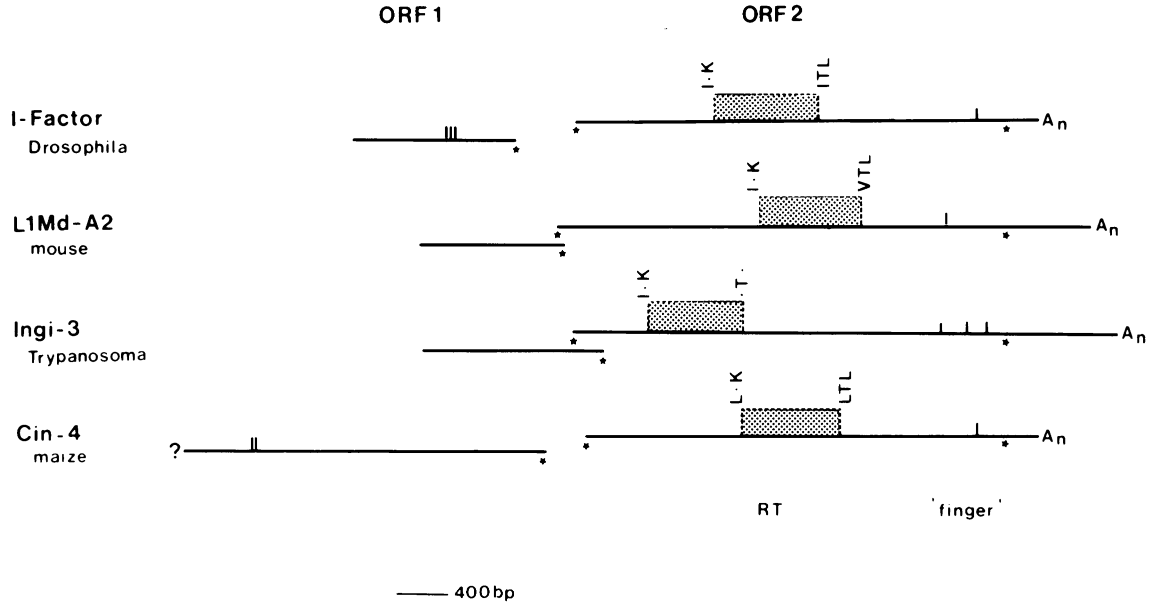

Fig. 6. Similarity in the overall organization of nonviral retrotransposons found in different organisms. The published sequences of nonviral retrotransposons (quoted in Figures 4 and 5) are depicted in linear form reflecting true size and distance relations. For comparison the translation stop (*) at the end of the ORF2 of the elements is aligned to an identical position. The conserved domain with homology to retroviral reverse transcriptase (RT, see also Figure 4) is represented by a shaded box. The letters at the ends of the shaded boxes indicate the amino acids at the ends of the corresponding regions of reverse transcriptase homology (see Figure 4). Vertical lines show the position of putative DNA-binding 'fingers' (see Figure 5). The question mark indicates that the structure of the non-truncated Cin4 element is not yet determined. This figure shows the structure of the longest analysed Cin4-element (Cin4-15). We should note here that the depicted Cin4 structure is based on the compilation of overlapping sequences obtained from various cloned Cin4 elements. Although the sequenced regions do not contain translation stops we do not know whether the entire reading frame in an individual cloned element is open.

Discussion

The Cin4 element in maize is a nonviral retrotransposon

The Cin4-1 element is an insert which alters the structure of the *A1* transcription unit in a wild-type allele of the *A1* locus in *Zea mays*. Sequence analysis of the insert and flanking sequences indicates that the insert arrived at its new location by transposition since it is flanked by a direct duplication of target sequences. Only one copy of this duplication is present in other wild-type *A1* alleles which do not contain the Cin4-1 insert.

There are several lines of evidence which strongly suggest that insertion of Cin4-1 into the *A1* gene occurred after reverse transcription of an RNA intermediate and that other Cin4 copies are dispersed by the same mechanism. Firstly, the maize genome contains many Cin4 copies which are sequence-related to each other at their 3′-end. These copies are truncated at their 5′-end as determined by sequence analysis. Secondly, the Cin4-1 element terminates in an oligo(A) track as do all other truncated Cin4 elements analysed so far. A polyadenylation signal is located upstream of the oligo(A) track. The length of the oligo(A) track varies among independent copies. Thirdly, the Cin4 elements possess no terminal repeats. Fourthly, each element is flanked by a direct duplication of nucleotides, but the number of duplicated bases differs at each independent location of the element. By these criteria Cin4 qualifies as a nonviral retrotransposon, in particular as a processed pseudogene (for review see Weiner *et al.*, 1986). The dispersion of such squences in the eukaryotic genome occurs by reverse transcription of a polyadenylated message transcribed by polymerase II from the promoter of a functional gene.

Strikingly the overall organization of Cin4 corresponds in several aspects to that of other nonviral retrotransposons (see Figure 6) like the L1 element of primates (for compilation see Sakaki *et al.*, 1986) and its relatives found in mouse (Loeb *et al.*, 1986), cat (Fanning and Singer, 1986) and rabbit (Demers *et al.*, 1986) or the I element in *Drosophila* (Fawcett *et al.*, 1986) and the Ingi-3 element in *Trypanosoma* (Kimmel *et al.*, 1987). The type of conservation within the reverse transcriptase domain and the position of the putative DNA-binding 'finger' downstream of this region within ORF2 clearly allow the classification of Cin4 elements as pseudogenes of a nonviral retrotransposon in maize.

As discussed by others (Fawcett *et al.*, 1986; Fuetterer and Hohn, 1987) a potential nucleic-acid-binding domain may be involved in the regulation of activity or the transposition of retroid elements. Interestingly the ORF1 of the I-factor and the ORF1 of Cin4 contains a conserved domain resembling the putative RNA-binding region within the *gag* polypeptide of retroviruses (Figures 5 and 6). The relevance of this observation is not clear but it may indicate functional homology between these two types of elements distinguishing them from other nonviral retrotransposons.

A model of the mechanism of Cin4 integration

Except for the idea that processed pseudogenes and nonviral retrotransposons integrate in a random manner into preformed chromosomal staggered nicks, details of the mechanism of integration are a subject of speculation (Rogers, 1985; Weiner *et al.*, 1986; Wilde, 1986). The nature of the intermediate that integrates and how and where reverse transcription occurs remain open questions. In addition, not all retroid elements need to integrate by the same mechanism.

The experimental approach to answering these questions is limited to the analysis of the elements that are already integrated. Our analysis of six independent Cin4 elements in maize provides some insights on the process of integration. The data compiled in Figure 3 indicate that sequence homology between the target and the 5′-region of the element could play a role in the integration process because the sequence of each target shows homology to sequences present in longer Cin4 elements upstream of the truncation endpoint of each inserted element. The flanking, direct

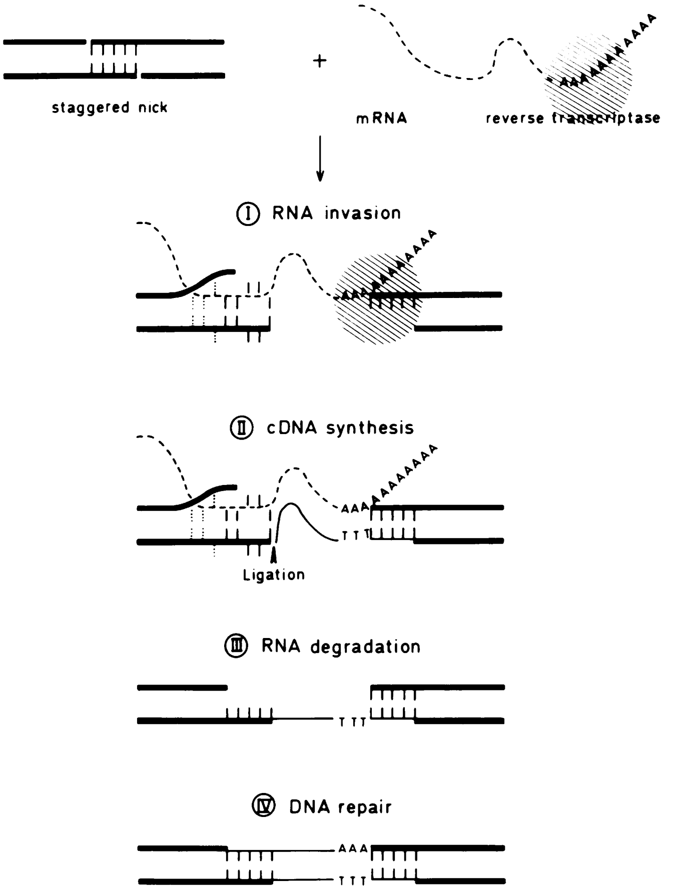

Fig. 7. A molecular model of the mechanism of Cin4 integration. The model is based on the analysis of the structure of integrated Cin4 elements and it is explained in detail in the Discussion. The Cin4 mRNA is depicted by staggered lines. Bold horizontal lines represent chromosomal DNA and thin horizontal lines represent new DNA strands synthesized either during reverse transcription (Step II, lower strand) or during DNA repair (Step IV, upper strand). The 5′ to 3′ orientation of the upper DNA strand is left to the right. The reverse transcriptase protein is symbolized by a shaded circle. Bases within the staggered chromosomal nick are depicted as vertical short lines pointing towards each other. Note that the intermediate hybrid between the 5′-region of the mRNA and the protruding end of the chromosomal DNA in Step I is not perfect (vertical short lines not pointing to each other) and that the hybrid may include bases upstream of the staggered nick (dotted vertical lines).

duplications, on the other hand, indicate that these molecules integrate into staggered nicks. The variability in size of the flanking duplications may indicate that these nicks are preformed as postulated for the integration of processed pseudogenes (see citations above). But it also may be related to element-encoded functions as has been observed in certain cases of bacterial (Haberman *et al.*, 1979) and plant (Sommer *et al.*, 1985; Coen *et al.*, 1986) transposons.

In addition to integration into staggered nicks of variable size, and to the kind of 'site selection' described above, the mechanism of Cin4 integration must be compatible with the following observations: (i) the duplications flanking an individual element are always identical, indicating that chromosomal staggered nicks did not undergo changes during integration; (ii) the oligo(A) track at the 3′-end of the element is shorter than the poly(A) tail of a eukaryotic message; (iii) there is no evidence for the truncation of Cin4 at its 3′-end and (iv) the element encodes a protein containing a putative 'reverse transcriptase' domain.

The model we propose for the mechanism of Cin4 integration combines several aspects of retrotransposon integration suggested by others (see reviews cited above) and the data obtained with Cin4. In the model depicted in Figure 7 insertion is initiated by attachment of the poly(A) tail of the putative Cin4 message to the protruding 5′-end of a randomly selected, preformed, staggered nick of variable size (Step I, right side of the molecule in Figure 7). The short oligo(A) sequence remaining after endonucleolytic or exonucleolytic cleavage of the poly(A) tail, can either be ligated to the chromosomal DNA or fixed without covalent linkage. Although the details of this linkage remain unclear, one can assume that the Cin4 encoded protein might exert all the necessary specificities. The free 3′-end of the chromosomal DNA at this side of the staggered nick can serve as the primer for reverse transcription toward the 5′-end of the message (Step II, lower strand in Figure 7). We propose that at the still unoccupied side of the staggered nick hybridization between the mRNA and sequences within the 5′-protruding end of the staggered nick occurs (Steps I and II, left side of the molecule in Figure 7). As shown schematically in Figure 7 the short, imperfect RNA−DNA hybrid includes several bases upstream of the staggered nick. This invasion of the DNA duplex by the Cin4 RNA may stabilize the duplex but it is not obligatory for integration since the structures of the integrated Cin4-232 and Cin4-198 elements do not reveal this extended homology (see Figure 3). The RNA−DNA hybrid then blocks reverse transcription across RNA sequences located further upstream. The 3′-OH-group at the end of the incoming cDNA strand can be ligated to the free 5′-end of the chromosomal staggered nick (Step II in Figure 7, see arrow), and integration is completed by repair processes. Repair includes degradation of RNA within the RNA−DNA hybrid by RNase H activity (Figure 7, Step III) which is encoded in the *pol* region of retroviruses, and perhaps also in the long ORFs of nonviral retrotransposons. All other enzymes like DNA polymerase, ligase and exonuclease can be recruited from the repair machinery of the plant cell. Due to the repair process the staggered nick is filled and the flanking duplication perfectly matches the original chromosomal sequence (Figure 7, Step IV).

The actual problem with this model is to explain how the linking of the 3′-end of the Cin4 message and the 5′-end of the staggered DNA nick occurs. As argued in the Results section, the occurrence of Cin4 copies which are truncated at their 3′-end is unlikely. One could therefore assume that the reverse transcriptase encoded by Cin4 is template specific and that this specificity is related to some sequences located within the authentic 3′-end of the Cin4 message. Whether this is true could be tested by characterizing the activities and the substrates of the Cin4 encoded protein *in vitro*.

In summary the model of Cin4 integration described above favours *in situ* cDNA synthesis initiated by random attachment of mRNA into staggered chromosomal nicks with protruding 5′-ends. A similar mechanism can be envisaged for the integration of a cDNA into preformed staggered nicks with protruding 3′-ends. We favour integration initiated by the mRNA only because this eliminates the problem of the primer needed for cDNA synthesis. Truncation of Cin4 elements occurs upon integration (and not necessarily during cDNA synthesis), and the basis for selection of the site of truncation is the homology between the mRNA and the target sequence. However, if this is true, then all Cin4 elements in the maize genome are truncated unless full-size elements integrate by a different mechanism.

The origin of Cin4: how recent are transposition events?

There is no doubt that transposition of Cin4 occurred in the past. The two alleles of the *A1* gene described in this and in a previous report (Schwarz-Sommer *et al.*, 1987) could only arise by insertion of Cin4-1 into one of the alleles. The differences in the chromosomal location of Cin4-related sequences in different maize lines (Figure 2) also reflects Cin4 mobility. Our data strongly support the idea that retrotransposition rather than DNA-mediated transposition is responsible for the mobility of members of the Cin4 element family. Whether Cin4-related retrotransposition currently occurs in maize is still a matter of speculation.

The copy number of Cin4 in maize is several orders of magnitude lower than the copy number of L1 elements in mammals (for review see Rogers, 1985). We also outlined in the Results that no direct evidence for the activity of Cin4 could be obtained since no long Cin4 mRNAs were found. This could mean that the mobility of the Cin4 element in maize is suppressed. The limited mobilty could in part be due to inefficiency in any of the enzymatic steps involved in the integration of Cin4 copies. But suppression could also occur if a suppressor is actively synthesized or if an activator is missing. Since the Cin4 element shows structural relation to the I element of *Drosophila* (Fawcett *et al.*, 1986 and see Figure 6) one can speculate that Cin4 was activated by an early dysgenic cross as I elements are mobilized by crosses with R strains (Finnegan, 1985). Alternatively, the apparent immobility of Cin4 can also reflect that Cin4 was only active for a short period of time and that the element is a remnant of an 'infection'. What type of infectious agent was involved remains a puzzle.

The functional relevance of Cin4 for maize (and perhaps the relevance of similar, not yet detected elements in other plants) may be related to its homology to nonviral retrotransposons in other organisms. L1, Ingi, I and Cin4 elements differ in nucleotide sequence. However, certain regions are conserved in all elements at the amino acid level (Figures 4, 5 and 6) suggesting that functional constraint conserves the integrity of these domains. Whether this conservation is indicative of a common functional relevance of all nonviral retrotransposons for the organisms or whether it indicates their common origin during evolution remains open.

Materials and methods

Somatic embryos were obtained from 12-day-old fertilized embryos of greenhouse grown maize (line A 188) by using the cultivation methods described by Potrykus *et al.* (1979). Compact embryogenic calli developing within 4 weeks were harvested and frozen in liquid nitrogen. Maize leaves for the preparation of RNA and DNA used in subsequent cloning experiments were obtained from 4- to 6-week-old greenhouse plants.

The genetic stocks used in the experiments were as follows. Line C (R.A.Brink, Madison, WI) is a colour-converted W22 inbred line. *al-dt* is a maize line obtained from M.M.Rhoades (Bloomington, IN), *a2-m5* 5064 and *a1-pale mr* were obtained from P.A.Peterson (Ames, IA), *a1-ml* 5719A-1 was obtained from B.McClintock (Cold Spring Harbor, NY). The maize lines originating from the three geneticists differ in genetic background and are characterized by mutations in certain genes. In this report we use these mutations to designate different maize lines. *Zea mays* spp. *parviglumis* (teosinte Guerrero) originated from the collection of W.C.Gallinat (Waltham, MA).

All methods applied in the studies reported in this paper (DNA and RNA preparation, cloning of genomic and cDNA, sequence analysis, etc.) were performed as previously described (Schwarz-Sommer *et al.*, 1984, 1985, 1987). Recombinant phages containing truncated Cin4 elements were isolated from a library prepared with *a1-pale mr* DNA partially digested with *Mbo*I. The recombinant phage carrying Cin4-15 originates from a similar library with *a2-m5* 5064 DNA (kindly provided by W.Martin). The leaf cDNA library was prepared with RNA from leaves of *a1-ml* 5719A1 plants.

Acknowledgements

We greatly acknowledge the skilful technical assistance of Rita Berndtgen and Christiane Horst. We also thank Drs Patricia Nevers, Peter Starlinger and Jeff Dangl for their helpful criticism during the preparation of this manuscript.

References

Blumberg vel Spalve,J. (1985) PhD Thesis, University of Cologne.

Boeke,J.D., Garfinkel,D.J., Styles,C.A. and Fink,G.R. (1985) *Cell*, **40**, 491–500.

Coen,S.C., Carpenter,R. and Nartin,C. (1986) *Cell*, **47**, 285–296.

Covey,S. (1986) *Nucleic Acids Res.*, **14**, 623–633.

Demers,G.W., Brech,K. and Hardison,R.C. (1986) *Mol. Biol. Evol.*, **3**, 179–190.

Drouin,G. and Dover,S.A. (1987) *Nature*, **328**, 557–558.

Fanning,T. and Singer,M. (1987) *Nucleic Acids Res.*, **15**, 2251–2260.

Fawcett,D.H., Lister,C.K., Kellett,E. and Finnegan,D.J. (1986) *Cell*, **47**, 1007–1015.

Fink,G. (1987) *Cell*, **49**, 5–6.

Finnegan,D.J. (1985) *Int. Rev. Cytol.*, **93**, 281–326.

Fuetterer,J. and Hohn,T. (1987) *Trends Biochem. Sci.*, **12**, 92–95.

Gardner,R.C., Howarth,A.J., Hahn,P., Brown-Luedi,M., Shepherd,R.J. and Messing,J. (1981) *Nucleic Acids Res.*, **9**, 2871–2887.

Habermann,P., Klaer,R., Kühn,S. and Starlinger,P. (1979) *Mol. Gen. Genet.*, **175**, 369–373.

Harberd,N.P., Flavell,R.B. and Thompson,R.D. (1987) *Mol. Gen. Genet.*, in press.

Harris,N. and Flavell,R.B. (1986) *Heredity*, **57**, 276–277.

Hattori,M., Kuhara,S., Takenaka,O. and Sakaki,Y. (1986) *Nature*, **321**, 625–628.

Johns,M.A., Mottinger,J. and Freeling,M. (1985) *EMBO J.*, **4**, 1093–1102.

Kimmel,B.E., Ole-Moiyoi,O.K. and Young,J.R. (1987) *Mol. Cell. Biol.*, **7**, 1465–1475.

Loeb,D.D., Padgett,R.W., Hardies,S.C., Shehee,W.R., Comer,M.B., Edgell, M.B. and Hutchison,C.A. (1986) *Mol. Cell. Biol.*, **6**, 168–182.

Miller,J., McLachlan,A.D. and Klug,A. (1985) *EMBO J.*, **6**, 1609–1614.

Nevers,P., Shepherd,N.S. and Saedler,H. (1986) *Adv. Bot. Res.*, **12**, 103–203.

Pfeiffer,P. and Hohn,T. (1983) *Cell*, **33**, 781–789.

Potrykus,I., Harms,C.T. and Lörz,H. (19797 *Theor. Appl. Genet.*, **54**, 209–214.

Proudfoot,N.J. and Brownley,G.G. (1976) *Nature*, **263**, 211–214.

Rogers,J.H. (1985) *Int. Rev. Cytol.*, **93**, 188–279.

Saigo,K., Kigimiya,W., Matsuo,A., Inouye,S. and Yoshioka,K. (1984) *Nature*, **312**, 659–661.

Sakaki,Y., Hattori,M., Fujita,A., Yoshioka,K., Kuhara,S. and Takenaka,O. (1986) *Cold Spring Harbor Symp. Quant. Biol.*, **51**, 465–469.

Schwartz,D.E., Tizard,R. and Gilbert,W. (1983) *Cell*, **32**, 853–869.

Schwarz-Sommer,Z. (1987) In Hennig,W. (ed.), *Results and Problems in Cell Differentiation. Vol. 14. Structure and Function of Eukaryotic Chromosomes*. Springer Verlag, Heidelberg, pp. 213–221.

Schwarz-Sommer,Zs., Gierl,A., Klösgen,R.-B., Wienand,U., Peterson,P.A. and Saedler,H. (1984) *EMBO J.*, **3**, 1021–1028.

Schwarz-Sommer,Zs., Gierl,A., Cuypers,H., Peterson,P.A. and Saedler,H. (1985) *EMBO J.*, **4**, 591–597.

Schwarz-Sommer,Zs., Shepherd,N., Tacke,E., Gierl,A., Rohde,W., Leclercq,L., Mattes,M., Berndtgen,R., Peterson,P.A. and Saedler,H. (1987) *EMBO J.*, **6**, 287–294.

Skowronski,J. and Singer,M.F. (1986) *Cold Spring Harbor Symp. Quant. Biol.*, **51**, 457–464.

Sommer,H., Carpenter,R., Harrison,B.J. and Saedler,H. (1985) *Mol. Gen. Genet.*, **199**, 225–231.

Toh,H., Hayashida,H. and Miyata,T. (1983) *Nature*, **305**, 827–829.

Weiner,A.M., Deininger,P.L. and Efstratiadis,A. (1986) *Annu. Rev. Biochem.*, **55**, 631–661.

Wilde,C.D. (1986) *CRC Crit. Rev. Biochem.*, **19**, 323–352.

Received on August 27, 1987; revised on September 28, 1987