The EMBO Journal vol.6 no.6 pp.1555–1563, 1987

Transcription of transposable element *Activator* (*Ac*) of *Zea mays* L.

R.Kunze, U.Stochaj[1], J.Laufs and P.Starlinger

Institut für Genetik, Universität zu Köln, D-5000 Köln 41, FRG

[1]Present address: Institut für Anatomie, Universität Marburg, D-3550 Marburg, FRG

Communicated by P.Starlinger

Transcripts of various sizes hybridize to the transposable element *Ac* of *Zea mays* in most maize lines. A 3.5-kb mRNA with an abundance of $1-3 \times 10^7$ of the poly(A) RNA, however, is found exclusively in those lines that carry an active *Ac*. Plants with two *Ac* elements contain slightly more 3.5-kb *Ac* transcript than those with only one *Ac*. Overlapping cDNA clones spanning most of the message have been isolated and sequenced. The 5′-end of the transcript was determined by Northern hybridization and S1 mapping. It starts at several sites over a distance of nearly 100 bases, contains an AUG-free leader 600–700 nucleotides long, has a long open reading frame encoding 807 amino acids and an untranslated 3′-sequence of 239 nucleotides. Four introns with a combined length of 654 bases are removed from the primary transcript. Radiosequencing of *in vitro* translation products shows that translation of the long open reading frame begins at the first AUG, even though it is located in an unfavourable sequence context. The transcript is found in all organs investigated, provided an active *Ac* is present in the stock.

Key words: Activator element/expression/transposable elements/*Zea mays*

Introduction

The transposable element *Activator* (*Ac*) is the autonomous member of a transposable element family that has also non-autonomous members called *Dissociation* (*Ds*). *Ds* elements can transpose only in the presence of *Ac*. *Ac* was initially described in 1947 and was the first autonomous transposable element to be identified (McClintock, 1947, 1951). All maize lines tested so far, even if devoid of genetically characterized *Ac* or *Ds* elements, carry >30 DNA sequences which hybridize to *Ac* and *Ds* sequences (Geiser *et al.*, 1982; Fedoroff *et al.*, 1983; Behrens *et al.*, 1984).

Two *Ac* elements in different loci of the maize genome were cloned and sequenced (Pohlman *et al.*, 1984; Müller-Neumann *et al.*, 1984) and shown to be identical. The molecular structure of some *Ds* elements was also determined (Döring and Starlinger, 1984; Döring *et al.*, 1984; Merckelbach *et al.*, 1986). *Ac* fails both to transpose autonomously and to transactivate *Ds* elements, if certain regions of its sequence are deleted. Two examples are the *wx-m9::Ds* element where only 194 bp from the full length *Ac* element are missing (Fedoroff *et al.*, 1983) and the *Ds* element *bz-m2* (*DI*) which carries a 1312-bp deletion in a different region of the *Ac* element (Dooner *et al.*, 1986). The frequency of transposition correlates negatively with the number of *Ac* copies present in a cell ('negative dosage effect') (McClintock, 1951). Furthermore, when the *Ac* dosage is increased, *Ac*-dependent transposition events occur at a later time during endosperm development (McClintock, 1948). These functions must also be encoded for by *Ac*, because the 1312-bp deletion in *bz-m2* (*DI*) also abolishes the ability of the element to contribute to the negative dosage effect (Dooner *et al.*, 1986). The delayed appearance of effects caused by two *Ac* elements during endosperm development indicates an interaction with cellular functions changing during development, e.g. the rate of cell division (Schwartz, 1984).

Since these *Ac* activities are not obviously related to each other, it is conceivable that *Ac* encodes several functions. The loss of all of them by the same deletion is not, in itself, incompatible with several functions; a deletion could potentially inactivate more than one gene. This is unlikely, however, because the deletions in *wx-m9::Ds* and *bz-m2* (*DI*), are located at different positions within the open reading frames (Pohlman *et al.*, 1984; Müller-Neumann *et al.*, 1984). Are there several gene functions expressed by the same DNA segment? In order to study these questions, we have begun an investigation of the transcript(s) of transposable element *Ac* and report here on the major transcript of this element.

Results

Northern blot analysis of Ac transcripts

Since the number of *Ac* transcripts cannot be predicted and because nearly all maize lines contain many *Ds* copies in addition to *Ac*, a transcript hybridizing to an *Ac* probe need not necessarily be an *Ac* product. Therefore, in order to identify authentic *Ac* transcripts, we looked for the correlation between

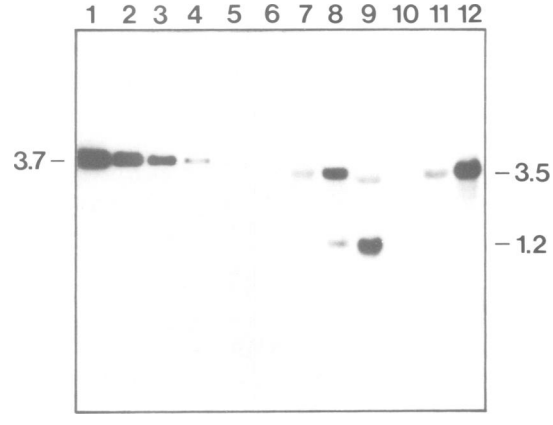

Fig. 1. *Ac*-homologous transcripts. Poly(A)$^+$ RNA from various sources was hybridized to probe K. The location of the probe on the *Ac* sequence is shown in Figure 3. **Lanes 1–5** contain defined amounts of a 3.7-kb plasmid fragment homologous to the whole length of probe K: **lane 1**: 7 pg; **lane 2**: 3.5 pg; **lane 3**: 1.6 pg; **lane 4**: 0.8 pg; **lane 5**: 0.4 pg. **Lanes 6–11** contain 3 μg poly(A) RNA each (except lane 9, which contains 12 μg poly(A) RNA) from varying maize lines: **lane 6**: P^{rr} (kernels); **lane 7**: P^{vv}::Ac (kernels); **lane 8**: *wx-m9::Ac* (seedlings); **lane 9**: *wx-m9::Ds* (seedlings); **lane 10**: *wx-m7::Ac* (seedlings); **lane 11**: *adh1-2F11, bz2-m, Ac* (kernels). **Lane 12** contains 4.5 μg poly(A) RNA from tobacco leaves transformed with *Ac*. (Sizes in kb.)

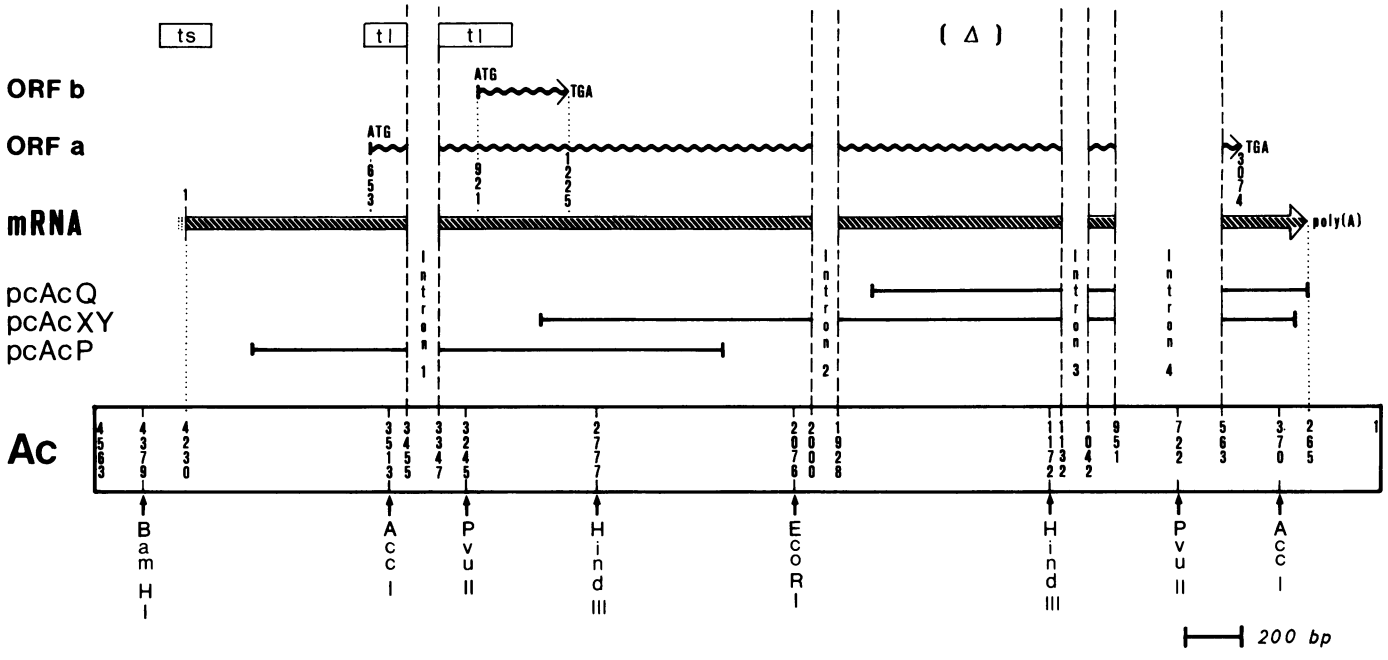

Fig. 2. Schematic representation of the structural organization. **Ac**: the lowest bar shows a restriction map of *Ac*, the corresponding positions of the restriction sites and the positions of the first and last nucleotide of the exons. **pcAcP, pcAcXY** and **pcAcQ**: these lines represent the sequences of the three overlapping cDNA clones. **mRNA**: the bar shows the alignment of the 3.5-kb transcript. The poly(A) tail is indicated at its right end. The numbering starts with position 1 at the 5'-end of the transcript (corresponding to position 4230 on the *Ac* sequence) and ends with position 3312 (corresponding to position 265 on the *Ac* sequence). The positions of the AUG and the UGA codons defining the open reading frames ORFa (807 amino acids long) and ORFb (102 amino acids long) are named. ORFa and ORFb are diagrammed above the mRNA. **ts**: this segment of the *Ac* sequence is shown in Figure 5. **tl**: this segment of the cDNA sequence is shown in Figure 6. (Δ): the 194-bp deletion in *wx-m9::Ds*.

the presence of an active *Ac* element in different strains and a particular transcript.

Poly(A) RNA was isolated from *Ac*-containing and *Ac*-free maize lines and analysed by Northern blotting experiments. At least four of seven *Ac*-containing maize lines analysed have active elements at different positions within the genome. Four maize lines studied have no genetically detectable *Ac* activity. In most RNA preparations of both Ac^+ and ac^- lines several bands are detected. The patterns vary between different maize lines. In addition, they alter with different probes. Figure 1 shows a Northern blot hybridized with probe K (see Figure 3).

The only transcript present in all *Ac*-containing and absent from all *Ac*-free lines tested is ~3.5 kb long. This observation suggests that the 3.5-kb RNA is an *Ac* transcript. It is lacking in mutant *wx-m9::Ds* and a 3.3-kb band which is absent in *wx-m9::Ac* is detected instead (lane 9 in Figure 1). This is expected, since *Ds9* differs from *Ac* by a 194-bp deletion (indicated in Figure 2).

We conclude that the 3.5-kb RNA is the only transcript correlated with an 'active' *Ac*. Additional support for this hypothesis is obtained by analysis of the mRNA from a tobacco line which is stably transformed with *Ac* (Baker *et al.*, 1987). Lane 12 of Figure 1 shows that in this material the 3.5-kb transcript is also found. If there exist other *Ac*-specific transcripts, their abundance is below our detection limit, which we estimated to be $\sim 10^{-8}$ of the poly(A) RNA.

Isolation and characterization of cDNA clones

cDNA was prepared from *wx-m7::Ac* mRNA and cloned in the λ vector NM1149. In Northern blotting experiments, the RNA employed shows transcripts 0.9 and 1.2 kb long in addition to the 3.5 kb one. We therefore analysed only three clones with an insert >1.2 kb. The inserts of these clones have lengths of 1.3, 1.5 and 2.1 kb respectively.

The inserts were subcloned into pUC9 and sequenced. With the exception of three single base substitutions all three cDNA inserts are perfectly homologous to the genomic *Ac* sequence. These sequence heterogeneities may be due to either misincorporation during cDNA synthesis or mutations in the *Ac* sequence. Two of the substitutions are within the protein coding region but do not alter the corresponding amino acids. The third is at the very last nucleotide preceding the poly(A) tail.

The three clones overlap extensively and span, altogether, 3.1 kb. Only one of them carries a poly(A) tail. The alignment of the three cDNA clones with *Ac* is shown in Figure 2.

The common sequence deduced from the three cDNA clones differs from the *Ac* sequence by the presence of a poly(A) tail at its 3'-end, and by the absence of four introns with a combined length of 654 bases. The combined cDNA sequence without its poly(A) tail is 3079 bp long. Together with the introns 3733 bases of the 4563 bases-long *Ac* would be covered by the cDNA. Note that the direction of transcription is opposite to the numbering of *Ac* (Pohlman *et al.*, 1984; Müller-Neumann *et al.*, 1984). The cDNA thus starts at position 3997 of *Ac* and ends at position 265. Its complete sequence is available from the authors upon request.

Mapping of the 5'-end of the 3.5-kb transcript

Using denatured plasmid fragments and a mRNA ladder as size standards on Northern blots, we estimate the length of the *Ac* transcript as 3.5 kb. Judging from this estimation, ~300 nucleotides from the 5'-end of the transcript were not cloned as cDNA. Therefore, we had to determine the 5'-end by indirect methods. For this purpose, we performed both sensitive Northern hybridization experiments with various single-stranded DNA probes labelled to a high specific activity, and an S1-nuclease protection assay for the precise determination of transcription start sites.

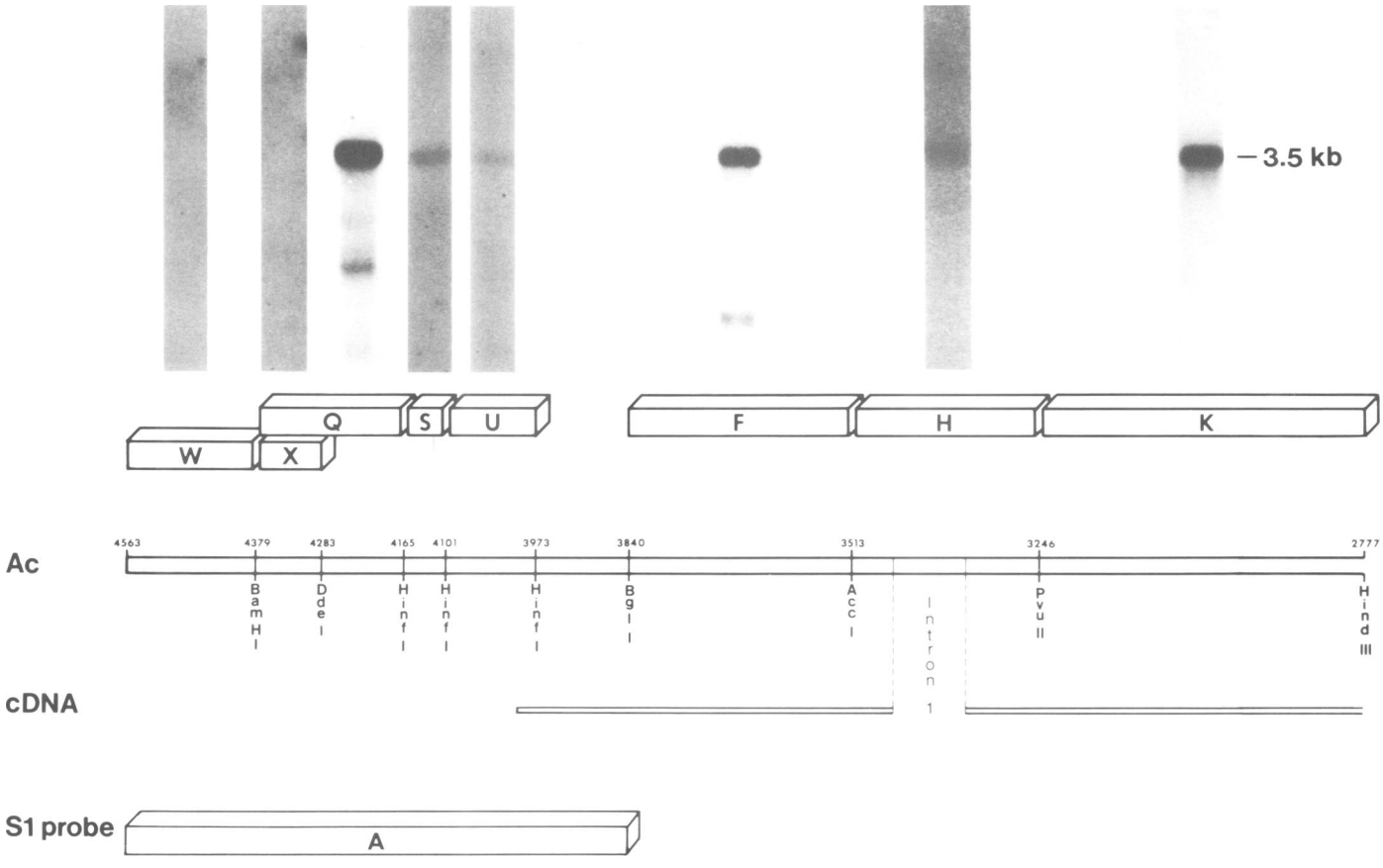

Fig. 3. Northern analysis of the transcription start region. 5 μg *wx-m7::Ac* poly(A)$^+$ RNA was electrophoresed and blotted as described. The position of the 3.5-kb transcript on the filters is indicated. The bars represent the single-stranded M13 probes used for hybridization of the filters shown above. The *Ac* sequence is aligned below. The next lower line delineates the 5′-terminal segment of the cDNA clone pcAcP. **A** represents the longer one of the single-stranded M13 probes which were used in the S1 protection assays.

The probes used for Northern hybridization experiments are shown in Figure 3. From the abundance of the 3.5-kb transcript in our poly(A) RNA preparation (see below) and calibration experiments with plasmid DNA as a marker, we conclude that regions of homology of ≥30 bp should be detected.

Probes Q, S, U, F, H and K hybridize to the 3.5-kb transcript, whereas probes W and X do not (Figure 3). We conclude that transcription starts within the 3′-half of segment Q or that an exon shorter than 30 bp is located within the region covered by probes W and X.

For S1 protection experiments, we hybridized poly(A) RNA from several maize lines to single-stranded DNA fragments extending from a *BglI* site outside of *Ac* to either the *BglI* site at position 3839 of the *Ac* sequence or to the *Hin*fI site at position 3972 of *Ac* (values in parentheses). Subcloning of these fragments into an M13 vector allowed the synthesis of single-stranded DNA probes with a high specific activity. These probes carry 724 (591) bases of *Ac* sequence and overlap on a length of 158 (25) nucleotides with the cDNA clone pcAcP (Figure 3). Control experiments were performed with similarly prepared probes spanning introns 1 and 2, respectively. In these cases, the expected bands were seen and no other bands were visible (data not shown). Upon S1 nuclease digestion of the longer one of the above-mentioned terminal probes which was annealed to mRNA containing the 3.5-kb transcript, a cluster of bands spanning ~100 bases is seen (Figure 4). As expected, the shorter terminal probe generates a similar pattern, but each protected band is 133 nucleotides shorter than with the longer probe (data not shown). By varying the incubation temperature for the S1 nuclease digestion between 20 and 37°C, no difference in the pattern could be achieved. Prolonging the S1 nuclease incubation from the point where the probe was not yet completely degraded to 15× this time did not alter the observed pattern either. The most 5′-located, rather weak signal corresponds to position 4261 of the *Ac* sequence. The two strongest bands coincide with positions 4230 and 4206 of the *Ac* sequence, respectively. Figure 5 shows the corresponding *Ac* segment. The sizes of the arrows indicate the different intensities of the five more prominent bands obtained in the S1-protection assay. Henceforth, position 4230 of *Ac* will be designated at position 1 of the mRNA.

The pattern of bands is characteristic, it is seen in all *Ac*-carrying lines investigated, including the P^{vv} line. The latter observation is important, because the P^{rr} line, which is nearly isogenic with P^{vv}, but gentically *Ac* free, does not show any protected fragments in the S1 experiment. In addition to this banding pattern, the RNA of the P^{vv} line contains a band, the position of which coincides with the first nucleotide of *Ac*. The exact size of this signal was determined in an additional S1 experiment designed to give a higher resolution with probe W (data not shown). This extra band is probably due to a chimeric transcript beginning outside of *Ac* and extending into the *Ac* element to a position beyond the end of the DNA probe. On Northern blots with P^{vv} RNA a very long transcript is indeed detected in addition to the 3.5-kb mRNA with an *Ac* probe (not visible in Figure 1).

Interestingly, the characteristic S1-protection pattern seen with

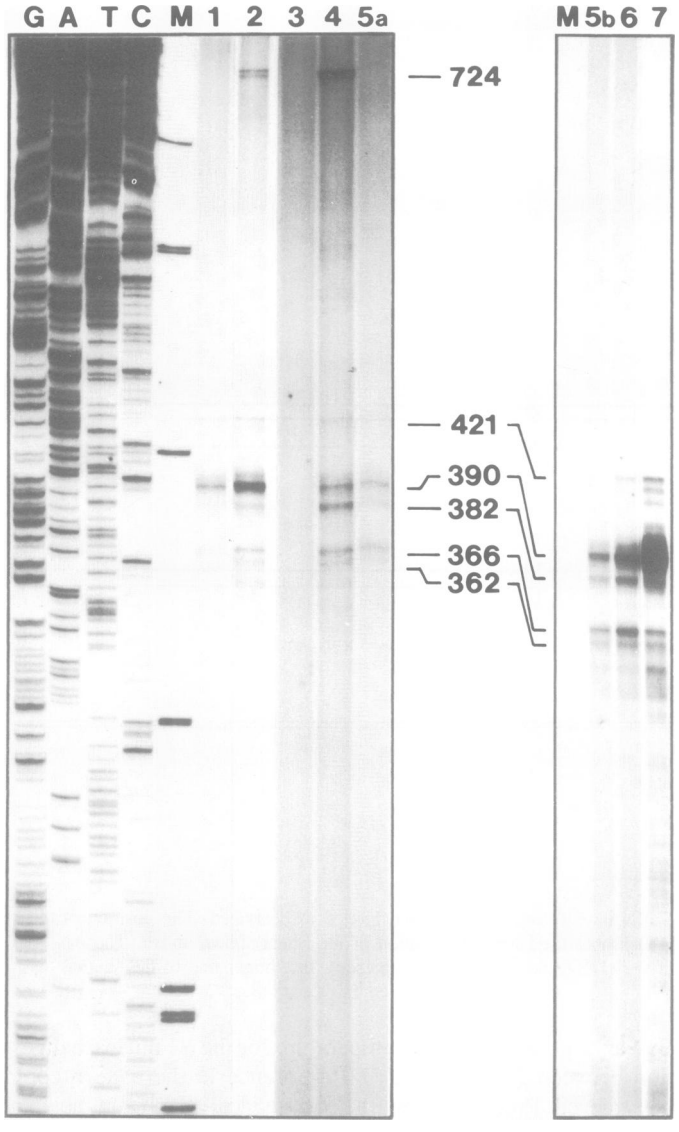

Fig. 4. Mapping of the transcription start by S1 protection assay. **G, A, T, C**: chain termination sequencing ladder of the probe A used in the S1 experiments. The lengths of the protected fragments do not coincide with the positions on the *Ac* sequence read from the sequencing ladder, but differ by the 55 M13-derived bases at the 5′-end of probe A which are clipped off during S1 nuclease incubation. **M**: pBR322 *Hpa*II-fragments. Sizes in bp: 622, 527, 404, 309, 242/238, 217, 201. **Lanes 1–7** show the protected fragments generated by S1 digestion of probe A hybridized with 10 μg of poly(A)$^+$ RNA. Between the two panels the sizes in nucleotides of the more prominent bands are printed. **Lane 1**: *adh1-2F11, bz2-m, Ac;* **lane 2**: *sh bz-m4, Ac;* **lane 3**: P^{rr}; **lane 4**: P^{vv}*::Ac;* **lane 5**: *wx-m7::Ac;* **lane 6**: *wx-m9::Ac;* **lane 7**: *wx-m9::Ds.*

poly(A) RNA from *Ac*-containing lines is also observed with RNA from most *Ac*-free lines: lane 7 in Figure 4 shows the protected fragments of *wx-m9::Ds* RNA. They are identical with those of *wx-m9::Ac* RNA in size, but differ in their relative intensities. There may be even several *Ds* elements giving rise to transcripts shorter than the 3.5-kb species, but containing the 5′-portion of the *Ac* transcript which is protected by the probe. Some of the smaller bands detected on Northern blots by *Ac* probes might be the products of such *Ds* elements.

The multiple signals revealed by the S1 experiments are all located between the 5′-ends of probes Q and X used in the hybridization experiments, thus supporting the conclusion that

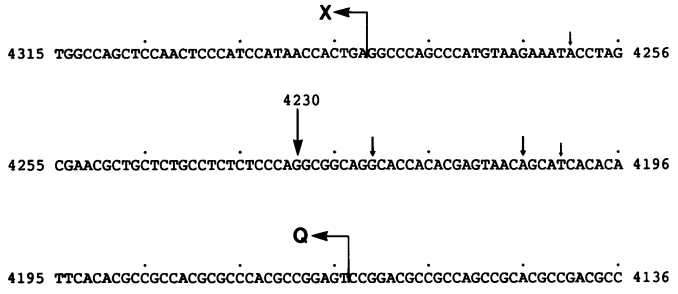

Fig. 5. Transcription start sites of the 3.5-kb mRNA. The location of this sequence region of *Ac* is indicated by the bar 'ts' in Figure 2. The numbering is that of the *Ac* sequence. The arrows mark the positions of the S1-protected bands in Figure 4. The lengths of the arrows correspond roughly with the intensities of the bands. **X** and **Q** indicate the 5′-ends of probes X and Q in Figure 3.

the transcription start is located within this DNA segment (Figure 5).

In vitro translation of the cDNA

Following the leader sequence, the mRNA contains a 2421-nucleotide long open reading frame beginning at the first AUG (ORFa in Figure 2). This AUG codon, however, is not located within the sequence context frequently observed for initiation codons of plants (Lütcke *et al.*, 1987) or eucaryotes in general (Kozak, 1986a) (Figure 6). It lacks a G in position +4 and a purine in position −3. The same is true for the second AUG codon. The next methionine codon within the long reading frame is the 10th AUG (see Figure 6). We were therefore interested to see to what extent translation of the mRNA could be initiated at the first AUG codon. In the absence of *in vivo* data, we decided to utilize the cDNA for *in vitro* translation. For this purpose, we cloned a fragment of the cDNA clone pcAcP extending from position 233 to position 1342 of the mRNA behind the SP6 promoter in plasmid pGem2. Linearization of this clone pcAcN at the *Hin*dIII site which is at the 3′-end of the *Ac* insert followed by *in vitro* transcription and *in vitro* translation of the products should yield a truncated protein. The *in vitro* synthesized mRNA was used in both wheat germ and rabbit reticulocyte cell-free systems for protein synthesis. Both systems yielded virtually identical protein products. The result of an experiment with a wheat germ extract is shown in Figure 7. The truncated, unmodified polypeptide starting at AUG no. 1 is predicted to have a mol. wt of 25 kd. Instead, a larger band of ~35 kd is seen in addition to several less intense bands. The N terminus of the protein yielding the most intense band was deduced as follows. The protein, labeled with [^{35}S]methionine and [^{3}H]proline, was electroeluted from the gel, subjected to automated Edman degradation and the collected fractions were analysed for ^{35}S and ^{3}H. As shown in Figure 8, ^{35}S label is detected only in fraction no. 20, whereas the tritium label occurs in fractions 2, 3, 8 and 9. Proline residues in these positions are only compatible with a translation start at the first AUG, with subsequent removal of the N-terminal methionine. None of the other AUG codons found in the N-terminal part of the presumed coding region of the cDNA would yield a similar amino acid sequence if used as an initiator codon.

The abundance of the Ac transcript

To estimate the abundance of the 3.5-kb *Ac* transcript in the poly(A) RNA fraction in seedlings from several lines, we compared the hybridization intensities of this transcript with that of

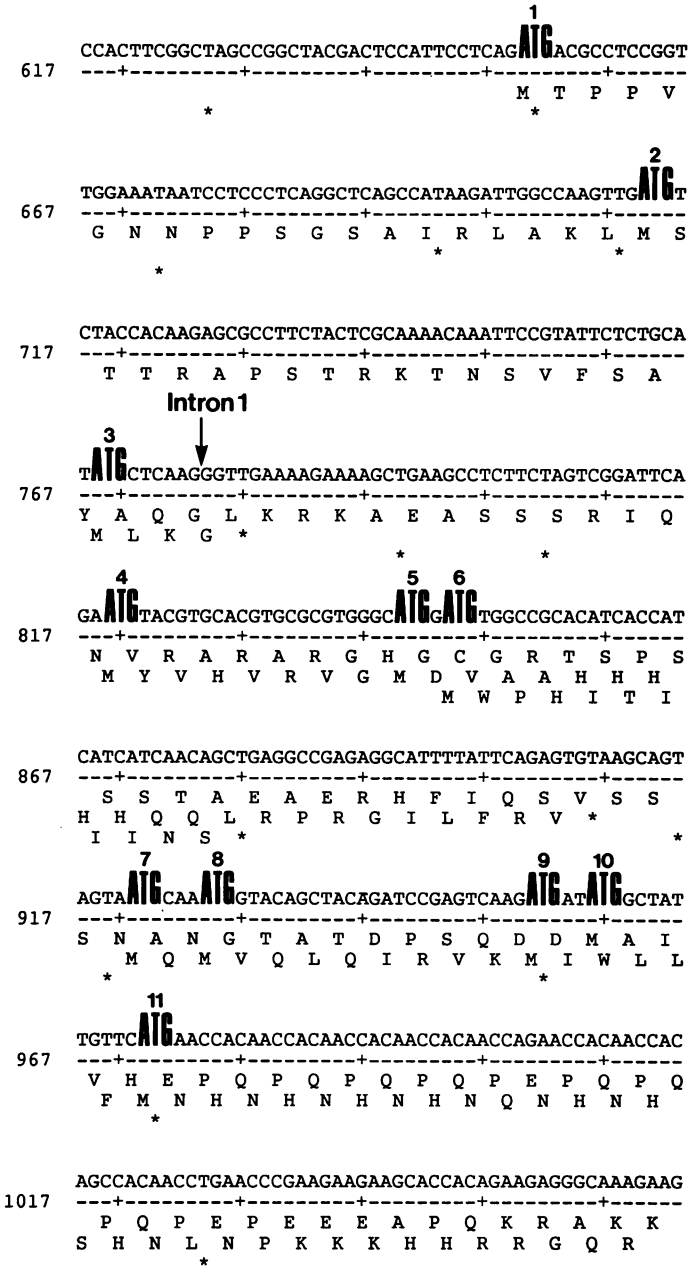

Fig. 6. The N-terminal region of ORFa. This sequence block is represented by the bar 'tl' in Figure 2. The numbers are the positions on the mRNA; position 1 corresponds to position 4230 on the *Ac* sequence. The ATGs are numbered and shown in bold letters. ATG1 opens ORFa, ATGs 2 and 10 lie in the same frame. ATGs 3–6 open only very short coding units. ATGs 7–9 and 11 lie at the beginning of ORFb. The location of intron 1 is indicated by the arrow. The (Pro-Gln) repeat (see text) starts at position 977.

defined amounts of denatured plasmid fragments on Northern blots. Since the plasmid fragment is of similar size (3.7 kb) as the 3.5-kb RNA, bands of equal intensity should contain twice as much double-stranded DNA as single-stranded RNA when a single-stranded probe is used that can anneal on its full length both to the marker fragment and to the mRNA. As shown in Figure 1, the hybridization intensities of the 3.5-kb RNA vary in different lines. They are roughly equal when 0.6–2.0 pg of DNA fragment and 3 μg of *wx-m7::Ac* or *wx-m9::Ac* RNA are compared. We conclude that 0.3–1 pg of 3.5-kb *Ac* transcript are contained in 3 μg of poly(A) RNA, that is a fraction of

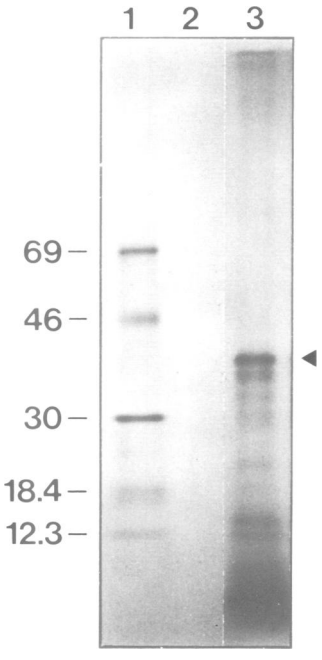

Fig. 7. *In vitro* synthesis of protein Ac-N in a wheat germ cell-free system. Linearized plasmid pcAcN was transcribed with SP6 polymerase in the presence of G (5′)ppp(5′)G. The resulting RNA was translated in a wheat germ cell-free system. Proteins were radioabelled with [^{35}S]methionine and subsequently separated by SDS–PAGE. The translation system was supplemented with no RNA (**lane 2**) and RNA derived from plasmid pcAcN (**lane 3**). The position of protein Ac-N used for radiolabel sequence analysis (Figure 8) is marked by an arrow. The mol. wts of marker proteins (**lane 1**) are indicated on the left margin.

$1-3 \times 10^{-7}$. The concentration of this transcript in poly(A) RNA of stably transformed tobacco leaves is slightly higher (4×10^{-7}). Our estimate rests on the assumption that the transfer efficiencies of DNA and RNA to nitrocellulose are roughly equal, it does not correct for the contamination of poly(A) RNA with residual ribosomal RNA.

We next compared the relative amounts of *Ac* transcript in the poly(A) RNA extracted from various organs (kernels, seedlings, leaves, stalks, roots and anthers). The differences in the intensities of the 3.5-kb band did not exceed a factor of four. We conclude that *Ac* RNA is transcribed in all organs investigated and that there are no pronounced differences in the concentration of *Ac* message among the different organs studied.

We also compared the amount of *Ac* RNA in seedlings, anthers and unfertilized ears containing either one or two copies of *Ac*, and in developing kernels harvested 10 or 12 days after pollination containing either two or three copies of *Ac* in the endosperm. As shown in Figure 9, in the presence of two *Ac* copies (three copies in the endosperm) we find more of the 3.5-kb transcript than with only one *Ac* copy in the genome (two *Ac* copies in the endosperm). We estimate the difference to be <2-fold. Among 14 independent comparison experiments we found two exceptions where the intensity of the 3.5-kb RNA was lower with the higher *Ac* dosage in the genome.

Discussion

Among the various transcripts detected in different maize lines either carrying *Ac* or devoid of it, we identified a 3.5-kb mRNA as the only one that segregates with an active *Ac*. Other transcripts hybridizing to *Ac* probes were detected either in only some of

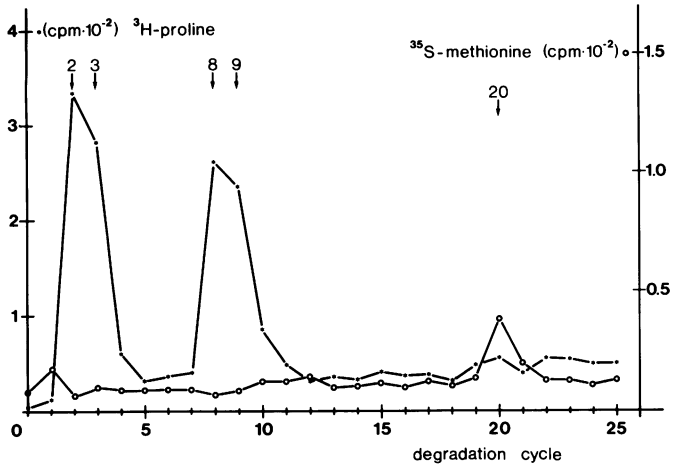

Fig. 8. Partial N-terminal radiolabel sequence analysis of protein Ac-N. Protein Ac-N was radiolabelled biosynthetically in a wheat germ cell-free system with [^{3}H]proline and [^{35}S]methionine. Proteins were separated by SDS−PAGE and blotted onto glass fibre sheets. The radiolabelled protein Ac-N (marked by an arrow in Figure 7) was subjected to automated Edman degradation. The fractions containing the peaks of radioactivity released for [^{3}H]proline (2, 3, 8 and 9) and [^{35}S]methionine (20) are indicated.

the *Ac*-containing strains, or were also found in *Ac*-free strains. They could have their start site within one of the 30−50 DNA sequences which hybridize to an *Ac* probe in most maize lines but are not active *Ac* elements, or they might be initiated from promoters outside of the hybridizing *Ac* or *Ds* sequence, if the element is located within a transcribed region.

Are the cDNA clones isolated derived from the 3.5-kb *Ac* transcript? Three bands were visible on Northern blots with *wx-m7::Ac* poly(A) RNA that was used for the cDNA cloning. Beside the strongest one, the 3.5-kb band, two other, at least 2-fold weaker bands (1.2 and 0.9 kb) were visible. All three cDNA clones selected for the analysis carry inserts longer than the two weaker bands. We therefore think that they are not derived from the two smaller transcripts, but either from the major 3.5-kb mRNA or from a minor messenger which is not detectable on the Northern blot due to its low abundance.

The three cDNA clones overlap extensively; two of them contain the junctions between exons 3 and 4 and exons 4 and 5 (Figure 2). The structural identity between the cDNA and the 3.5-kb transcript was proven for two areas by S1 mapping; a 1.4-kb probe covering the *Ac* sequence between positions 2452 and 1172 including intron 2 protects only the predicted fragments. The same is true for probe H (Figure 3) which spans intron 1 (data not shown). From these results the possibility that one or all three cDNA clones are derived from an alternatively spliced minor mRNA becomes unlikely. We are therefore confident that these cDNA clones are derived from the 3.5-kb transcript.

We have not isolated a complete cDNA clone but have determined the 5′-end(s) of the mRNA by Northern hybridization and by S1 mapping. The hybridization data place the start point of transcription into the 5′-half of fragment Q (Figure 3), and the S1 protection analysis yields a cluster of 5′-ends with the more 5′ located of the two major bands at position 4230. This is in agreement with the 5′-ends being in fragment Q.

The 5′-ends seen in our experiments could be (i) start sites of transcription, (ii) acceptor splice sites, or (iii) products of 5′ exonucleolytic degradation. We do not consider this possibility very likely, as the pattern described has been seen reproducibly in 11 experiments with 10 independent RNA preparations.

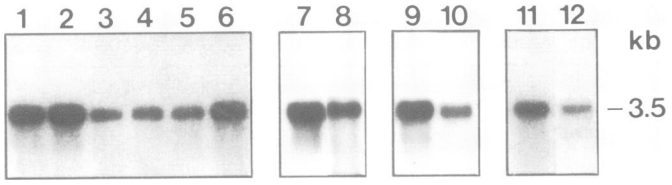

Fig. 9. The dosage dependence of the expression of the 3.5-kb transcript. Equal amounts of poly(A) RNA from maize material containing either one or two doses of *Ac* (two or three doses of *Ac* in the endosperm) were compared for the hybridization intensities of the 3.5-kb band generated with probe F. **Lanes 1−6** contain individually prepared poly(A) RNA from seedlings with either one (lanes 4 and 5) or two (lanes 1, 2, 3 and 6) doses of *Ac*. **Lanes 7** and **8** contain RNA from anthers with either two *Ac* (lane 7) or one *Ac* element (lane 8). **Lanes 9** and **10** contain RNA from unfertilized ears with either two (lane 9) or one (lane 10) *Ac* element in the genome. **Lanes 11** and **12** contain RNA from kernels harvested 10 or 12 days after pollination with three (lane 11) and two (lane 12) doses of *Ac* in the endosperm. **Lane 3** shows one of the exceptional results with RNA from plants containing two copies of *Ac*, but nevertheless generating a weaker 3.5-kb band than RNA from comparable plant material with only one *Ac* copy.

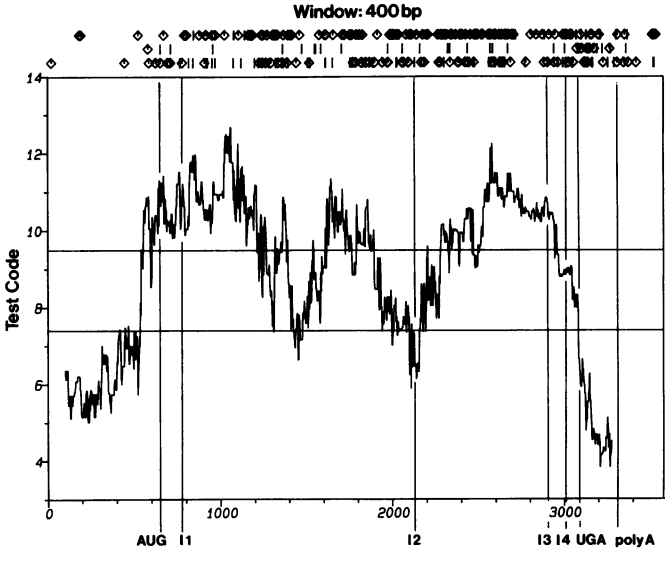

Fig. 10. A plot of the test code program (Devereux *et al.*, 1984) run on the 3312-nucleotide sequence predicted for the *Ac* transcript is shown. Above the plot the positions of AUG codons (short vertical lines) and stop codons (diamonds) are marked. The top region of the plot is supposed to predict coding regions to a 95% level of confidence. The bottom part is supposed to predict non-coding regions to the same confidence level. In the middle region the statistic can make no significant prediction. The marked positions below the plot are: **AUG**: the first AUG that opens ORFa; **I1−I4**: positions of introns 1−4; **UGA**: end of ORFa; **polyA**: poly(A) addition site.

Is it possible that the S1 protection experiment detects intron borders rather than a transcription start and that a small exon located upstream of position 4230 has escaped our detection? If we allow for an uncertainty of one nucleotide in the determination of the first S1-resistant base, position 4230 is very similar to the intervening sequence acceptor splice site which was proposed by Mount (1982). The homology to the consensus sequence for plant genes which was described by Brown *et al.* (1986) is much less pronounced, however. Several lines of evidence argue against an additional exon. First, the S1 protection experiments do not show only one band, but a cluster of them. This is in contrast to the expectation for exon−intron borders and to the observation made with control S1 experiments at introns 1 and 2.

Second, the S1-protected band of the presumed readthrough transcript in the P^{vv} line has exactly the length of the DNA probe and therefore cannot have lost an intron 5′ of position 4230. The easiest explanation is the absence of such an intron. It is also possible, however, that the longer readthrough RNA has a structure which prevents splicing, while the original mRNA does not. Third, the Northern hybridization data also argue against the presence of an additional upstream exon >30 nucleotides, unless it had unusual hybridization properties.

Can the observed multiple bands be artefacts due to RNA secondary structure or other irregularities in the hybrid formation between probe and RNA? The pattern of bands was neither qualitatively altered by changes of the S1 incubation temperature nor by variation of the S1 incubation time. In addition, the shorter S1 probe was hybridized to the mRNA at a different temperature than the longer one (42 versus 52°C). Nonetheless, in both cases the equivalent banding pattern was generated. We therefore assume that the S1 experiments detect a heterogeneity of the mRNA at the 5′-end, possibly due to variable transcription starts with varying efficiencies.

These considerations complete our analysis of the transcript that is processed to a length of 3312 nucleotides plus the poly(A) tail when started at position 4230 of the *Ac* element. It contains an AUG-free leader sequence 652 bp in length, a coding region of 2421 bp (807 amino acids) and an untranslated 3′-sequence of 239 bp. A hexanucleotide ATTAAA is located 44 bp upstream of the poly(A) addition site. The four introns have donor and acceptor splice sites not deviating from those of other plant introns (Brown *et al.*, 1986). Translation of the long open reading frame starts at the first AUG of the transcript and terminates at a UGA codon.

The genomic *Ac* sequence located upstream of the presumed transcription start site does not have motifs common in most promoters. Particularly, it lacks a recognizable TATA box. This might explain the mulitple start points of transcription (Benoist and Chambon, 1981). A number of mammalian housekeeping genes with multiple transcription start sites lack TATA boxes, and have instead GC-rich sequences preceding the transcription start sites. Often, these genes are only weakly expressed on a basal level with little regulation and tissue specificity (McGrogan *et al.*, 1985; Sazer and Schimke, 1986; Dynan, 1986). In all organs investigated, *Ac* is also expressed at a low level. As *Ac* action is a stochastic event, however, we cannot exclude the possibility that *Ac* transcripts are not equally distributed in all cells of a tissue. It is conceivable that only a limited number of cells show an enhanced level of *Ac* transcription and that in these cells transposition events occur. This question will have to be addressed by *in situ* RNA hybridization experiments.

The 3.5-kb message starts with an unusually long (652 bases) untranslated leader devoid of AUG codons, but with nonsense codons in all three reading frames. We do not know whether this leader has a biological function. Examples are known where leader sequences play a role in the regulation of translation in eucaryotic cells. One of them is the GCN4 gene in yeast, where, in contrast to *Ac*, four AUG codons are located upstream of the 'real' AUG and are involved in the repression of translation (Mueller and Hinnebusch, 1986). Secondary structure in the leader can also modulate the initiation of translation by eucaryotic ribosomes (Pelletier and Sonenberg, 1985; Kozak, 1986b). However, no obvious conspicuous structures are found in the *Ac* leader sequence.

It it noteworthy that the *in vitro* translation experiments were carried out with an RNA devoid of the first 232 nucleotides of the leader sequence due to the insertion of only part of the cDNA behind the SP6 promoter. It will be interesting to see whether the choice of initiation codon or the relative frequency of initiation at these codons is altered by the presence of the complete 5′-untranslated sequence.

The initiating AUG in ORFa conforms neither to the eucaryotic (Kozak, 1986a) nor to the plant consensus sequences (Lütcke *et al.*, 1987). The same is true for the second one. The next four AUGs located in other frames are followed shortly by termination codons. It is known that short open reading frames preceding a protein coding sequence do not prevent reinitiation (Kozak, 1986c). The seventh, eighth and ninth AUGs are all located in the reading frame potentially encoding a 102-amino acid frame (ORFb) which is completely covered by ORFa (Figure 2). Only the 10th AUG is both located in ORFa and in the context of the consensus sequence. It was therefore interesting to see that the first AUG is used preferentially *in vitro*. While this does not prove a similar initiation *in vivo*, it is suggestive.

Our methods, however, could not detect the presence of a 102-amino acid protein or of the very short peptides. The shorter protein bands found in the translation experiment may not be the products of alternative translation starts, but rather degradation products. Whether ORFb has a biological role can only be investigated by site-specific mutagenesis.

The 807-amino acid protein predicted from the nucleic acid sequence is large enough to fold into several domains which might have different roles. Particularly interesting regarding folding is a 10-fold repetition of proline and glutamine (or glutamic acid) starting at amino acid 108. This rather unstructured region potentially separates an N-terminal domain of 107 amino acids from a longer C-terminal domain. Whether these domains fold separately, and whether they have distinct biological roles, is unknown.

It is noteworthy, however, that a structurally similar though unrelated repetition of the tripeptide Thr-Glu-Pro is found in the hobo transposable element of *Drosophila melanogaster* (Streck *et al.*, 1986). The hobo element has similarities with *Ac* both in the inverted terminal repeat sequence and in the number of host bases duplicated upon insertion.

Repetitions of Pro-Glu are also found in the E1A protein of adenovirus 2 (Moran and Mathews, 1987) and in the protein of *Trypanosoma brucei* (Roditi *et al.*, 1987). In neither of these cases is it known whether an unstructured region of this kind serves as a hinge region to connect more strongly structured protein domains.

If the long open reading frame of *Ac* encodes a transposase, its target sites would be within the nucleus. We therefore scanned the protein sequence for a potential peptide sequence specifying a nuclear localization signal (Smith *et al.*, 1985; Dingwall *et al.*, 1987), but did not find any.

An interesting aspect of *Ac* function is the phenomenon of the negative dosage effect: the frequency of transposition is negatively correlated with the number of *Ac* copies present within a cell. We find, however (in 12 out of 14 comparisons), more 3.5-kb transcript in plants with two copies of *Ac* transcript than in plants with only one copy. This can be explained either by assuming that regulation is exerted post-transcriptionally, or by assuming that in addition to the 3.5-kb transcript there is a yet undetected minority RNA with a different dose dependence.

We can only speculate on the existence of *Ac* transcripts other than the 3.5-kb one. *Ac* does not have sufficient size and open reading frames for genes located outside of the segments spanned by the 3.5-kb transcript. Are there alternative splicing pro-

ducts? Alternative splicing has been proposed for the *P* element of *D. melanogaster* (Laski *et al.*, 1986; Rio *et al.*, 1986). It is noteworthy in this case that one of the alternative splice products is too rare to be found by Northern blot analysis.

A hidden product of alternative *Ac* splicing could lead to the omission of one of the known introns. This would lead to a truncated protein, because all of the introns have nonsense codons in ORFa. Alternatively, a yet undetected intron might be removed from one of the known exon sequences, or one of the known introns could be extended, as is the case in adenovirus E1A RNA (Moran and Mathews, 1987). This would lead to a product lacking an internal peptide. This question can be addressed by systematic mutation studies (Coupland *et al.*, 1987).

We have applied the test code program of Devereux *et al.* (1984), based on Fickett's algorithm (1982), to the nucleotide sequence of the long open reading frame of *Ac* as well as to several plant gene sequences. This program predicts the probability of a nucleotide sequence coding for a protein. While in all other cases the test code rose to a value of high coding probability at the first AUG, as expected, and fell to a low level at the stop codon, a more complicated pattern was seen with *Ac* (Figure 10). Regions with a high score alternate with regions with a low score. A similar pattern is observed with the mRNA of the *D. melanogaster P* element (Laski *et al.*, 1986). We do not know presently whether this is due to an additional selection superimposed on the selection for an amino acid sequence that breaks up periodicities, or whether this pattern indicates alternative splicing. A computer-aided search for particularly well suited splice sites at the end of the low score regions failed to reveal any.

Materials and methods

Genetic materials and RNA isolation

Seven maize lines carrying *Ac* in at least four different locations (*sh-m5933, Ac; sh-m6233, Ac; sh bz-m4, Ac; adh-2F11, bz2-m, Ac; wx-m7::Ac; wx-m9::Ac;* P^{vv}*::Ac*) and five lines devoid of *Ac* activity were used as a source of RNA. Plant material (seedling roots, seedling shoots, developing kernels, leaves, stalks, anthers and husks) was frozen and ground in liquid nitrogen, incubated with proteinase K and hot phenol/chloroform extracted (Kloppstech and Schweiger, 1976; Maniatis *et al.*, 1982). Poly(A)$^+$ RNA was selected by adsorption to oligo(dT) cellulose.

Northern hybridization of RNA

3–8 μg poly(A)$^+$ RNA was separated on 1.2% agarose gels containing 2.2 M formaldehyde. After electrophoresis, RNA was transferred without further treatment of the gel to nitrocellulose in 10 × SSPE. After baking for 30 min at 80°C the filter was pre-incubated in 5 × SSPE, 5 × Denhardt's solution, 10 mM EDTA, 100 μg/ml denatured calf thymus DNA, 0.2% SDS for >1 h at 68°C. Hybridization was overnight at 42°C with 20 μl/cm^2 of 50% formamide, 5 × SSPE, 5 × Denhardt's solution, 10 mM EDTA, 100 μg/ml denatured calf thymus DNA and probe.

Preparation of single-stranded probes

Recombinant M13 DNA was annealed to a 3-fold molar excess of M13 universal sequence primer (17 mer) for 30 min at 60°C. Synthesis was carried out for 15 min at room temperature in the presence of α-^{32}P-labeled dCTP (Amersham, 3000 Ci/mmol) by the Klenow fragment of DNA polymerase I, then a chase with non-radioactive dCTP was added and the incubation continued for 15 min. The mixture was adjusted to 150 mM NaCl, 20 U *Eco*RI were added and incubated for 30 min at 37°C. After denaturing for 10 min at 95°C in 50% formamide, 10 mM EDTA, the reaction mix was electrophoresed in a 1% low melting agarose gel. The position of the single-stranded probe fragment was located by autoradiography of the gel. The band was cut out and eluted. For Northern hybridization the probe was butanol concentrated and added directly to the hybridization solution. For the S1 protection assay the probe was precipitated by 2-propanol together with RNA.

S1 protection assay

The S1 mapping experiment was carried out essentially as described by Burke (1984); 5–15 μg poly(A)$^+$ RNA was hybridized to ~5 ng of labelled single-stranded probe for 3–5 h at 52°C in 5 μl 80% formamide, 0.4 M NaCl, 40 mM Pipes pH 6.5, 1 mM EDTA. 65 U S1 nuclease/μg RNA were added in 200 μl S1 buffer (0.4 M NaCl, 30 mM Na acetate, 4.5 mM Zn acetate, pH 4.5) and incubated for 2–40 min at 30°C. The reaction was stopped by 2-propanol precipitation and analysed on 4–6% sequencing gels.

cDNA cloning

cDNA synthesis was performed according to Schwarz-Sommer *et al.* (1985) with minor modifications. cDNA was cloned into the *Eco*RI site of phage λ NM1149 (Murray, 1983). Approximately 1.3 × 10^6 recombinant phage plaques were screened with nick-translated *Bss*HII fragment from a plasmid containing the whole *Ac* sequence. Ten positive plaques were isolated and three of them were chosen for further analysis. Their inserts were subcloned into the *Eco*RI site of pUC9 (Vieira and Messing, 1982) and subsequently sequenced.

DNA sequence analysis

DNA sequence analysis was performed either by the chemical degradation method (Maxam and Gilbert, 1980) or by the chain termination procedure (Sanger *et al.*, 1977). Remaining gaps in the cDNA sequence were closed by oligodeoxynucleotide-directed chain terminating synthesis directly from double-stranded plasmid DNA. Approximately 70% of the sequence was determined from both strands. Four cDNA sections including the two *Hin*dIII, the *Pvu*II and the *Eco*RI restriction sites were sequenced from a cDNA clone that was prepared from the three overlapping ones and carried their combined length. In this way the integrity of the combination clone around the restriction sites used for the subcloning was confirmed.

In vitro transcription and translation

The plasmid pcAcN was constructed by subcloning a fragment from the cDNA clone pcAcP containing 419 bp of the leader sequence and the first 801 bp of ORFa behind the SP6 promoter of pGem2 (Promega Biotec).

In vitro transcription of linearized plasmid pcAcN followed the protocol of the suppliers (Amersham-Buchler), except that G5′ppp5′G was used for capping.

In vitro translation in a wheat germ extract was carried out as described by the manufacturers (Amersham-Buchler) with minor modifications: phenylmethylsulphonyl fluoride was present at 2 mM (final concentration) and a cocktail of protease inhibitors was added (antipain, aprotinin, chymostatin, leupeptin and pepstatin, each of them at 0.1 μg/ml).

Protein synthesis was performed in the presence of 833 μCi/ml (30.8 MBq/ml) L-[2,3,3,4,5-^{3}H]proline and 208 μCi/ml (7.7 MBq/ml) L-[^{35}S]methionine.

Proteins shown in Figure 7 were radiolabeled only with L-[^{35}S]methionine.

SDS-PAGE and radiolabel sequence analysis

Proteins of the *in vitro* translation mixture were precipitated with 5% TCA and separated on 11–20% acrylamide gradient gels (Lugtenberg *et al.*, 1975).

For sequence analysis, proteins were blotted after electrophoretic separation onto GF/C glass fibre sheets treated with *N*-trimethoxysilylpropyl-*N,N,N*-trimethylammonium chloride (Aebersold *et al.*, 1986).

Radiolabelled material was cut out of the glass fibre sheets and subjected to automatic Edman degradation without any further treatment. N-terminal sequence analysis followed the procedure of Hunkapiller *et al.* (1983).

Edman degradation was performed twice with 15 and 25 degradation cycles, respectively.

Computer analyses were performed using the programs developed by Devereux *et al.* (1984).

Acknowledgements

We thank H.-P.Döring and C.Lechelt for maize lines, B.Baker and G.Coupland for transformed tobacco material, U.Courage for the plasmid pcAcN, H.Reinke and K.Beyreuther for help with automatic sequencing and several colleagues of the department for critical reading of the manuscript and help with the computer analysis. This research was supported by the Deutsche Forschungsgemeinschaft through SFB 74.

References

Aebersold,R.H., Teplow,D.B., Hood,L.E. and Kent,S.B.H. (1986) *J. Biol. Chem.*, **261**, 4229–4238.

Baker,B., Coupland,G., Fedoroff,N., Starlinger,P. and Schell,J. (1987) *Maize Genet. Coop. News Lett.*, **61**.

Behrens,U., Fedoroff,N., Laird,A., Müller-Neumann,M., Starlinger,P. and Yoder,J. (1984) *Mol. Gen. Genet.*, **194**, 346–347.

Benoist,C. and Chambon,P. (1981) *Nature*, **290**, 304–310.

Brown,J.W.S., Feix,G. and Freudeway,D. (1986) *EMBO J.*, **5**, 2749–2758.

Burke,J.F. (1984) *Gene*, **30**, 63–68.

Coupland,G., Baker,B., Schell,J. and Starlinger,P. (1987) *Maize Genet. Coop. News Lett.*, **61**.

Devereux,H., Haeberli,P., Smithies,O. (1984) *Nucleic Acids Res.*, **12**, 387–395.

Dingwall,C., Dilworth,S.M., Black,S.J., Kearsey,S.E., Cox,L.S. and Laskey,R.A. (1987) *EMBO J.*, **6**, 69–74.

Döring,H.-P. and Starlinger,P. (1984) *Cell*, **39**, 253–259.
Döring,H.-P., Tillmann,E. and Starlinger,P. (1984) *Nature*, **307**, 127–130.
Dooner,H., English,J., Ralston,E. and Weck,E. (1986) *Science*, **234**, 210–211.
Dynan,W.S. (1986) *Trends Genet.*, August 1986, 196–197.
Fedoroff,N., Wessler,S. and Shure,M. (1983) *Cell*, **35**, 235–242.
New York, pp. 1–63.
Fickett,J. (1982) *Nucleic Acids Res.*, **10**, 5303–5318.
Friedemann,P. and Peterson,P.A. (1982) *Mol. Gen. Genet.*, **187**, 19–29.
Geiser,M., Weck,E., Döring,H.-P., Werr,W., Courage-Tebbe,U., Tillmann,E. and Starlinger,P. (1982) *EMBO J.*, **1**, 1455–1460.
Hunkapiller,M.W., Hood,L.E., Dreyer,W.J. and Hewick,R.M. (1983) *Methods Enzymol.*, **91**, 399–413.
Kloppstech,K. and Schweiger,G. (1976) *Cytobiologie*, **13**, 394–400.
Kozak,M. (1986a) *Cell*, **44**, 283–292.
Kozak,M. (1986b) *Proc. Natl. Acad. Sci. USA*, **83**, 2850–2854.
Kozak,M. (1986c) *Cell*, **47**, 481–483.
Laski,F.A., Rio,D.C. and Rubin,G.M. (1986) *Cell*, **44**, 7–19.
Lugtenberg,B., Meijers,J., Peters,R., van der Hoek,P. and van Alphen,L. (1975) *FEBS Lett.*, **58**, 254–258.
Lütcke,H.A., Chow,K.C., Mickels,F.S., Moss,K.A., Kern,H.F. and Scheele, G.A. (1987) *EMBO J.*, **6**, 43–48.
Maniatis,T., Fritsch,E.F. and Sambrook,J. (1982) *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Laboratory Press, NY.
Maxam,A. and Gilbert,W. (1980) *Methods Enzymol.*, **65**, 499–560.
McClintock,B. (1947) *Carnegie Inst. Washington Yearb.*, **46**, 146–152.
McClintock,B. (1948) *Carnegie Inst. Washington Yearb.*, **47**, 155–169.
McClintock,B. (1951) *Carnegie Inst. Washington Yearb.*, **50**, 174–181.
McGrogan,M., Simonsen,C.C., Smouse,D.T., Farnham,P.J. and Schimke,R.T. (1985) *J. Biol. Chem.*, **260**, 2307–2314.
Merckelbach,A., Döring,H.-P. and Starlinger,P. (1986) *Maydica*, **31**, 109–122.
Moran,E. and Mathews,M.B. (1987) *Cell*, **48**, 177–178.
Mount,S. (1982) *Nucleic Acids Res.*, **10**, 459–472.
Mueller,P.P. and Hinnebusch,A.G. (1986) *Cell*, **45**, 201–207.
Müller-Neumann,M., Yoder,J.I. and Starlinger,P. (1984) *Mol. Gen. Genet.*, **198**, 19–24.
Murray,N.E. (1983) In Hendrix,R.W., Roberts,J.W., Stahl,F.W. and Weisberg,S.A. (eds), *Lambda II*. Cold Spring Harbor Laboratory Press, NY, pp. 395–432.
Pelletier,J. and Sonenberg,N. (1985) *Cell*, **40**, 515–526.
Pohlman,R.F., Fedoroff,N. and Messing,J. (1984a) *Cell*, **37**, 635–643.
Rio,D.C., Laski,F.A. and Rubin,G.M. (1986) *Cell*, **44**, 21–32.
Roditi,I., Carrington,M. and Turner,M. (1987) *Nature*, **325**, 272–274.
Sanger,F., Nicklen,S. and Coulson,A.R. (1977) *Proc. Natl. Acad. Sci. USA*, **74**, 5463–5467.
Sazer,S. and Schimke,R.T. (1986) *J. Biol. Chem.*, **261**, 4685–4690.
Schwartz,D. (1984) *Mol. Gen. Genet.*, **196**, 81–84.
Schwarz-Sommer,Zs., Gierl,A., Cuypers,H., Peterson,P.A. and Saedler,H. (1985) *EMBO J.*, **4**, 591–597.
Smith,A.E., Kalderon,D., Roberts,B.L., Colledge,W.H., Edge,M., Gillett,P., Markhaus,A., Pancha,E. and Richardson,W.D. (1985) *Proc. R. Soc. London, Ser. B*, **221**, 43–53.
Streck,R.D., MacGaffey,J.E. and Beckendorf,S.K. (1986) *EMBO J.*, **5**, 3615–3623.
Sutton,W.D., Gerlach,W.L., Schwartz,D. and Peacock,W.J. (1984) *Science*, **223**, 1265–1268.
Vieira,J. and Messing,J. (1982) *Gene*, **19**, 259–268.

Received on March 16, 1987