The EMBO Journal vol.4 no.6 pp.1373-1380, 1985

Structure of the sucrose synthase gene on chromosome 9 of *Zea mays* L.

W.Werr, W.-B.Frommer, C.Maas and P.Starlinger

Institut für Genetik der Universität Köln, 5000 Köln 41, FRG

Communicated by P.Starlinger

The structure of the shrunken gene of *Zea mays* encoding sucrose synthase (EC 2.4.1.13) was determined by (i) sequencing the transcription unit and ~1.2 kb of 5'-upstream sequences from a genomic clone, (ii) by sequencing a nearly full length cDNA clone and (iii) by determining the transcription start site by a combination of primer extension experiments with synthetic oligodeoxynucleotide primers and S1 mapping. The sucrose synthase gene is 5.4 kb long, of which 2746 bp are found in the mature mRNA. The gene is interrupted by 15 introns. The first two introns are ~1 kb and ~0.5 kb in length, respectively, while the other introns are much smaller. A TATA box is located 30 bp upstream from the transcription start site. Approximately 610 bp upstream of the transcription start site a direct repeat of 16 nucleotides, separated by a 4-fold repetition of the sequence GGTGG is detected. The 16-bp sequence has similarities to a sequence repeat found between two promoters of a maize zein gene also expressed in the endosperm tissue. The transposable element *Ds* in the mutant *sh-m5933* and *sh-m6233* alleles is inserted in the seventh and first intron, respectively. The genomic and cDNA clones were obtained from different maize lines. This allows the determination of polymorphic sites which are frequent in 3rd codon position and absent in 1st and 2nd codon positions. In addition, the 3'-untranslated sequence shows two duplications that may have arisen by the insertion and subsequent excision of transposable elements.

Key words: sucrose synthase/transcription signals/sequence polymorphism/transposon footprints

Introduction

The *Shrunken* (*Sh*) locus on chromosome 9 of *Zea mays* encodes the enzyme sucrose synthase (EC 2.4.1.13) which catalyzes the cleavage reaction of sucrose to UDP-glucose and fructose in both directions (Chourey and Nelson, 1976). The enzyme is involved in starch metabolism of the developing endosperm. In maize strains homozygous for recessive *sh* mutations, sucrose synthase activity is decreased to 2-6% and starch content in the mature kernels is decreased to 60%. This causes the *shrunken* phenotype (Chourey and Nelson, 1976; Chourey, 1981). In wild-type, the protein encoded by the *Sh* locus amounts up to 3% of the total protein. The active form of the enzyme consists of four identical subunits of ~88 kd (Tsai, 1974). It is not known whether the high residual starch content in the mutants indicates that the enzyme is present in large excess in the wild-type or whether alternative pathways of starch biosynthesis are used.

A residual sucrose synthase (B) activity is detected in *sh* mutants (Chourey, 1981; Chourey and Nelson, 1976), including a deletion of the *Sh* gene (Burr and Burr, 1981; Chaleff *et al.*, 1981; Döring *et al.*, 1981). This indicates the presence of a second gene. In support of this idea, McCormick *et al.* (1982) detected a mRNA species and DNA fragments hybridizing weakly to a cDNA clone of the *Sh* gene in RNA and DNA isolated from maize strains homozygous for *Sh* deletions. The *in vitro* translation product of hybrid-selected mRNA transcribed from the sucrose synthase (B) gene has a similar, slightly faster electrophoretic mobility than the *in vitro* synthesized sucrose synthase (A) subunit encoded by the *Sh* locus and is precipitated by antiserum against sucrose synthase (A) protein (McCormick *et al.*, 1982).

The two genes seem to be regulated differently. In wild-type strains sucrose synthase A activity increases 40-fold during kernel development in the endosperm tissue, starting at day 5-8 after pollination. The enzyme activity reaches a maximum 40 days after fertilization and drops afterwards (Chourey, 1981). In strains in which the *Sh* gene is deleted the residual sucrose synthase B activity stays at a low level. Independent of the genotype (*Sh* or *sh*), only sucrose synthase B is detected in the embyro of the developing kernel (Chourey and Nelson, 1976).

A second reason for the interest in studying expression of sucrose synthase genes is the reaction catalyzed by the enzyme. Sucrose is the major transport form of assimilate on its way from the photosynthesizing leaves to the energy-consuming tissues, e.g., to the kernels. The observations from several plants are in agreement with the hypothesis that synthesis of sucrose is mainly catalyzed by sucrose-6-phosphate synthase (EC 2.4.1.14) and sucrose phosphatase (EC 3.1.3.24) and that sucrose synthase has its major role in sucrose breakdown for respiration or starch biosynthesis (Hawker, 1971; Downton and Hawker, 1973; Vieweg, 1974; Preiss and Levi, 1980). The enzyme therefore may be involved in the utilization of sucrose within different plant tissues, and may thus influence where the sucrose is stored or broken down, respectively.

As a step towards the study of gene regulation during plant development, we present here the exon-intron structure of the *Sh* gene. In addition, we discuss some evolutionary aspects deduced from the comparison of different alleles of the *Shrunken* gene and DNA sequences upstream of the transcription unit which might be important for regulation.

Results

DNA sequence studies of the Shrunken gene

For our sequence studies we used one genomic and two cDNA clones. A 600-bp 3'-terminal cDNA clone and a 16.3-kb genomic clone have been described previously (Geiser *et al.*, 1980, 1982). A 2.572-bp cDNA clone (pWW110/1) was isolated from a cDNA library kindly provided by A.Gierl and Zs.Schwarz-Sommer, Max-Planck-Institut für Züchtungsforschung, Cologne. On the genomic clone, the whole transcription unit and a DNA segment of 1140 bp located in front of the transcription unit were sequenced from both strands. The small cDNA clone was also sequenced from both strands, while from the larger cDNA clone only one strand was sequenced. Comparison of the two clones yielded the structure shown in Figure 1, where the cDNA clone

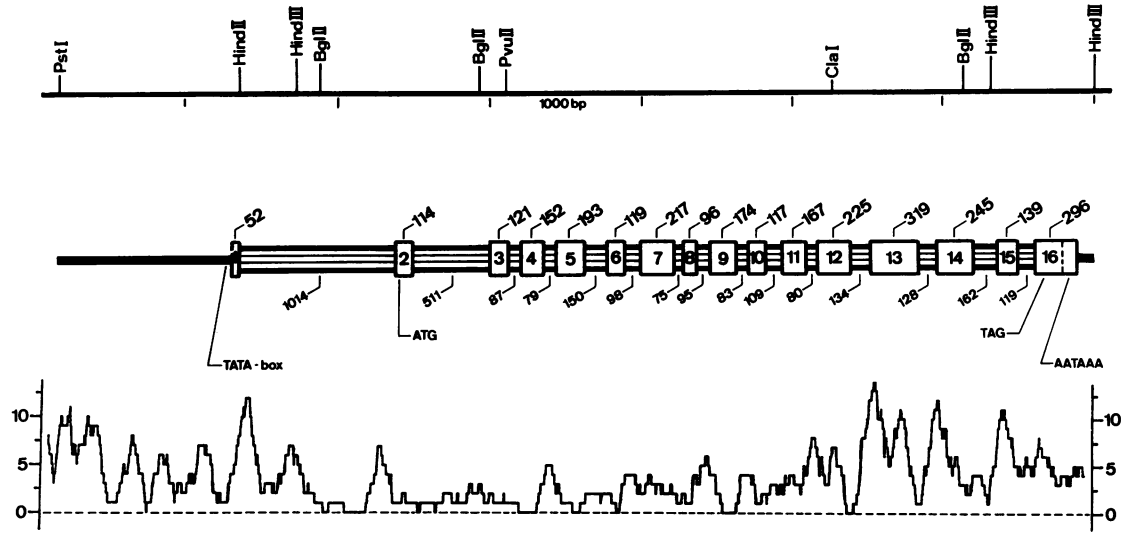

Fig. 1. Exon/intron structure of the *shrunken* gene. The exons are marked by the numbered boxes. The size of the exons is given above the drawing, intron sizes below. The total length of the drawing indicates the region which has been sequenced. A partial restriction map of the genomic DNA is shown in the upper part. The distribution of the dinucleotide CpG is given in absolute numbers for consecutive segments of 100 bp each.

(pWW110/1) ends within exon 3. The full sequence of both clones is entered in the EMBL DNA library and is available from the authors upon request.

Determination of the transcription start site

The large cDNA clone is not full length and terminates in exon 3. This was determined by S1 mapping experiments which were carried out in order to determine the approximate size and the order of the transcription unit and the size of the exons. For these experiments, the 7.1-kb *Cla*I fragment was labeled at the 5′ termini. This fragment has one end in the central part of the transcribed region, extends beyond the transcription start and contains all upstream sequences present in our genomic wild-type clone. Aliquots of the labeled fragments were cleaved with restriction endonucleases *Bam*HI, *Bgl*I, *Xba*I, *Hind*III, *Sph*I and *Bgl*II. The fragments located towards the 3′ end of the gene were purified, hybridized to mRNA, digested with S1 endonuclease, separated on agarose gels and subjected to autoradiography. The result of such an experiment is shown in Figure 2. As described in the sketch in Figure 2b, the longest band in each lane represents the full-length protected fragment, while the shorter bands are formed by partial S1 cleavage opposite to digested intron loops. The successive shortening of the labeled fragments by restriction digestion should have no influence on the length of the longest bands, until the first exon has been (partially) removed from the DNA before hybridization. By this method, the first exon is located in the 700-bp interval between the *Xba*I and the *Hind*III site (Figure 1), while the second exon is located between the *Hind*III and *Sph*I sites. The third exon is thus located within the central 3.3-kb *Bgl*II fragment. It contains a *Pvu*II site which is also present in the cDNA clone pWW110/1. The position of the intron 2/exon 3 boundary was determined by a S1 mapping experiment with DNA fragments labeled at the 5′ terminus of this *Pvu*II site. The resulting S1-resistant DNA fragment was 40 bp long, thus the *Pvu*II site is located 40 bp downstream of the intron/exon boundary (indicated in Figure 3c). The first three exons are thus spread over a distance of >1.5 kb. As we had no cDNA clone in this region available, we decided to perform primer extension experiments with reverse transcriptase.

Poly(A)$^+$ RNA isolated from developing endosperm was reverse transcribed in the presence of the synthetic oligodeoxynucleotide AAGCATTCCCTTGCCC as primer. The position of the primer sequence within exon 3 is indicated in Figure 3c. Reverse transcription was done on poly(A)$^+$ RNA isolated either from maize strains carrying the *Sh* allele, or as a control from maize carrying the deletion *sh bz-m4*. Reverse transcripts were electrophoresed on 6% denaturing polyacrylamide gels. As shown in Figure 3, only the wild-type RNA gives rise to a prominent double band at position 191/192, while many faint bands are also present in the control. Most likely, this strong band is a reverse transcript extending to the 5′ end of the mRNA. The size of exon 2 was estimated accurately from S1 mapping experiments to be 115 bp. In these experiments S1-digested RNA/DNA hybrids were denatured and electrophoresed on polyacrylamide gels containing 7 M urea. After transfer to nitrocellulose and hybridization to radioactively labeled DNA fragments, the length of the individual exons could be determined precisely.

If the known size of exon 2 and the residual bases of exon 3 towards the 5′ end of the synthetic oligonucleotides are added and subtracted from the length of the reverse transcript (191/192 bases), there remain 50 or 51 nucleotides for the size of exon 1, which is in agreement with the size estimated from S1 mapping experiments (Figure 2).

To determine the position of exons 1 and 2, the primer extension experiment was repeated in the presence of dideoxynucleotides to determine the sequence of the reverse transcript. The autoradiographs obtained from this experiment showed a high background most likely due to unspecific priming of reverse transcriptase as already visible in the *sh bz-m4* lane of Figure 3a. Therefore only parts of the sequence could be unambiguously determined. In general, the sequence was less clear at the higher transcript sizes.

Comparing this limited cDNA sequence information with the known genomic DNA sequence, we could recognize ~70% of exon 2 sequence. The sequence information obtained at the boundary between exons 1 and 2 is shown in Figure 3c. The unambiguously determined nucleotides extending further than exon 2 were screened in a computer search against all DNA sequences which were available upstream of exon 2. Even with one mismatch allowed, homology was found only once as indi-

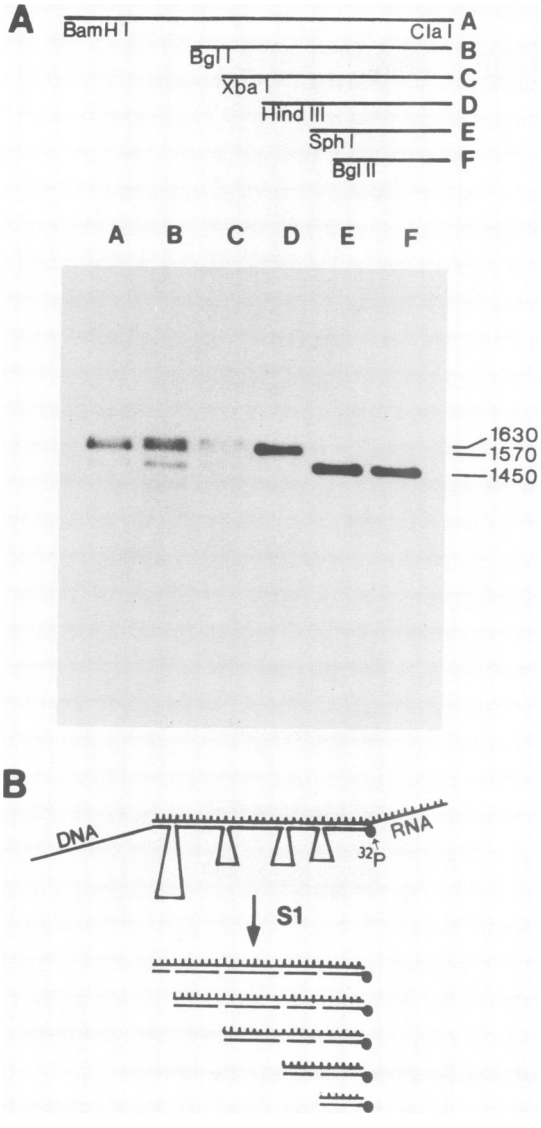

Fig. 2. **(A)** The DNA fragments used for hybridization to poly(A)$^+$ RNA are indicated above the autoradiogram. The DNA fragments carried the label at the 5′ terminus of the left end. S1-resistant RNA/DNA hybrids were electrophoresed on 1.2% agarose gels. **(B)** A schematic drawing: formation of RNA/DNA hybrids, digestion with endonuclease S1 and possible S1-resistant hybrid molecules.

cated in Figure 3c. This position is within the interval between the *Xba*I and *Hind*III site which was determined above (Figure 2). The number of nucleotides between the last nucleotide of the cDNA that had been determined with certainty and the 3′ end of the reverse transcript could be determined accurately, because random termination of the reverse transcriptase reaction caused a ladder of nucleotides in each lane. The distance was found to be 34 and 35 nucleotides. No dinucleotide AG, which could serve as an intron/exon boundary, is found in this interval within the DNA sequence. Thus the experiment places the transcription start tentatively, as shown in Figure 3.

A *Hind*III site is located within the first exon, as determined by the primer extension experiment. DNA was labeled at this *Hind*III site and used for a S1 protection experiment. Poly(A)$^+$ RNA obtained from wild-type endosperm protects DNA fragments of a length between 32 and 46 bp (Figure 3b). The four most pronounced RNA species are located at the positions indicated in Figure 3c. This confirms the transcription start site determined by the reverse transcription experiment in general, but places the major transcription start sites 3 – 6 bp in front of the terminus of the reverse transcript. The reason for this is unknown. The most prominent signal detected in the S1 mapping experiment is defined as transcription start and numbered +1 in our sequence data. The boundary between exon 2 and intron 2 was placed from the known size of exon 2 (115 bp) and the GT/AG rule (Breathnach and Chambon, 1981). Approximately 40 bp upstream and downstream of this splice site (shown in Figure 3c) no other GT dinucleotide is found in the genomic sequence.

The insertion sites of Ds in the sh-m5933 and sh-m6233 alleles

The 4.1-kb insertion in the *sh-m6233* allele is flanked by the duplication of the 8-bp CTTGTCCC. This 8-bp sequence is found at positions 193 – 220 in intron 1. In the *sh-m5933* allele (Courage-Tebbe *et al.*, 1983), only the 3′-adjacent sequence of the 30-kb insertion is known. The 8-bp sequence GTCGCTTT located adjacent to the insertion site on the 3′ side is located at positions 2970 – 2977 in intron 7. This insertion site is 12 bp upstream of the boundary to exon 8.

Protein coding capacity of the Shrunken mRNA

The first AUG is found in exon 2 and starts an open reading frame of 802 amino acid codons which is terminated by a TAG stop codon in exon 16. The predicted mol. wt. of the sucrose synthase monomer is 91.731 daltons. The predicted amino acid composition has been confirmed by a total hydrolysis of purified sucrose synthase protein.

Discussion

We have analyzed the transcription unit of the sucrose synthase gene and ~1.1 kb upstream of the transcription start by DNA sequencing. The results are discussed from 5′ to 3′ following the direction of transcription.

The 5′-flanking region of the sucrose synthase gene

We have examined the DNA sequence which extends ~1 kb upstream of the starting point of transcription, for sequences that might play a role in expression of the gene. Around position −610 we find a 16-bp direct repeat

TGGCGGGGA.GGAAATA
TGG.GGGGAGGGAAATA

which is separated by 22 nucleotides. A tandem duplication of 15 bp TTTAGGAAAAA$^{T}_{TT}$TAG is found in front of a zein C gene, which is also expressed in the endosperm tissue (Langridge and Feix, 1983). This duplication is located between the two promoters used for transcription of the gene and oriented opposite to the direction of transcription. Comparison of the duplicated sequences reveals a conserved part GGAAA$^{T}_{A}$A found in front of both endosperm genes. The sequence GGAAA found in all four repeats is also part of a consensus sequence found in animal enhancer elements TGGAAA (Weiher *et al.*, 1983).

The 22 nucleotides separating the duplicated 16 mers in front of the sucrose synthase gene contain a 4-fold repetition of the pentanucleotide GGTGG, a fifth copy overlaps the 16 mer duplicated downstream. It is interesting to note that the trinucleotide GTG is present five times within this region. This trinucleotide is reported to be frequent in procaryotic and eucaryotic DNA sites where interactions with proteins are anticipated for regulation or recombination (Cheng *et al.*, 1984).

Short nucleotide sequences have been implicated in the coordinate expression of eucaryotic genes (Davidson *et al.*, 1983). Therefore a possible role of these sequences in the expression

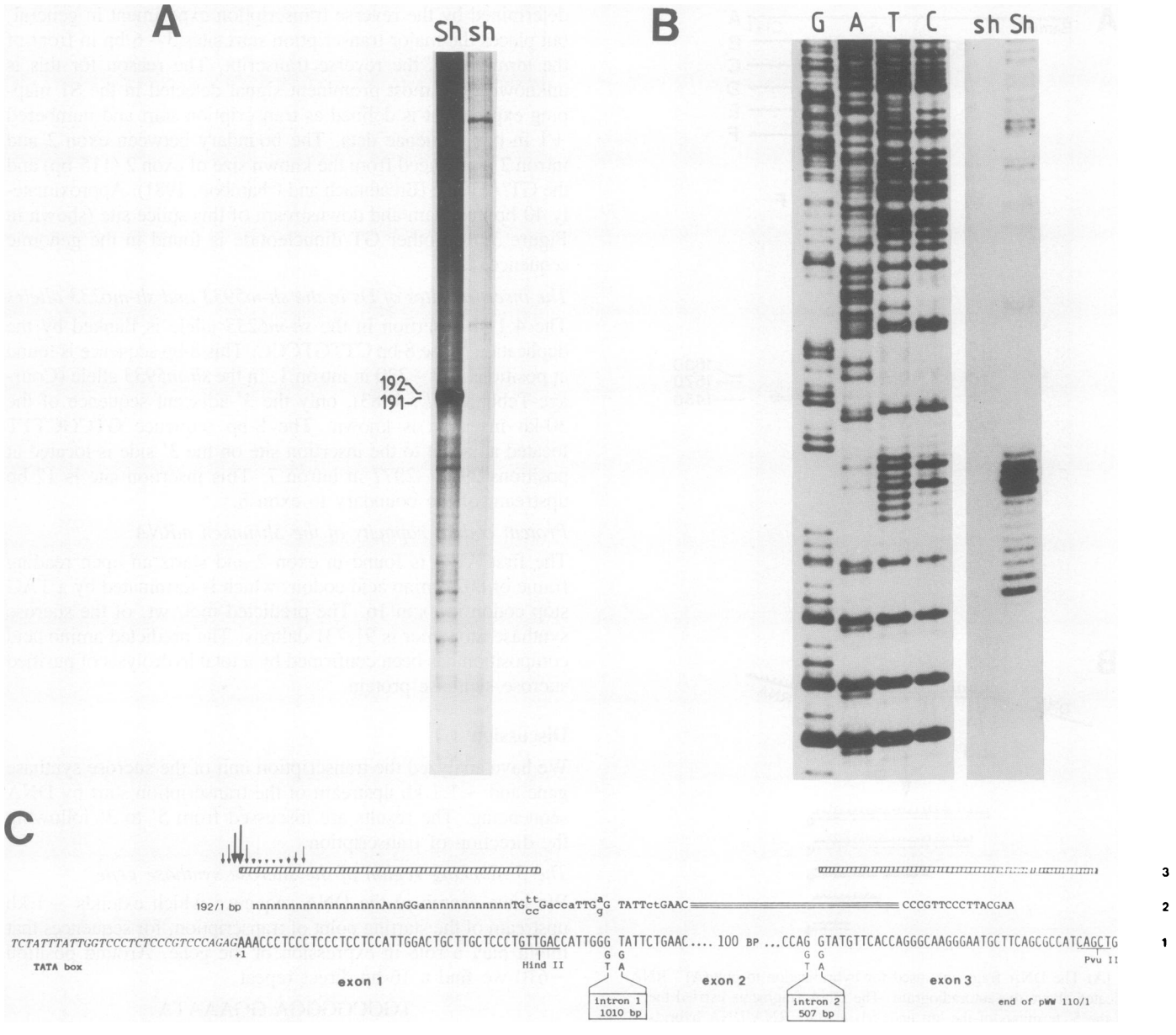

Fig. 3. **(A)** Primer extension experiment using the synthetic oligonucleotide from exon 3 as a primer for AMV reverse transcriptase [see **(C)**]. **(B)** Estimation of the transcription start by a S1 mapping experiment. The *Hind*III site shown in **C** was labeled at the 5′ terminus and used for DNA sequencing and the RNA protection experiment. The RNA sources used in the experiments were isolated from kernels with *Sh* phenotype or from kernels homozygous for the deletion allele *sh bz-m4*. **(C)** Summary of the experiments which determine the transcription start. The DNA sequence (1) shows exon 1 and 2 and parts of exon 3, as well as some sequences upstream of the transcription start. Above in line 2 the sequence of the synthetic oligonucleotide used for priming of the reverse transcriptase, its position within exon 3 and the end of the reverse transcript shown in part **A** is demonstrated. The cDNA sequence information obtained at the exon 1/2 boundary is shown. The size of the letters is a measure of how unambiguously the base could be determined. Line 3 shows the size of DNA fragments obtained after S1 digestion of RNA/DNA hybrids. The main signal in the S1 mapping experiment is designated as major transcription start (+1).

of the sucrose synthase gene is presently investigated in a heterologous plant system (*Nicotiana tabacum*) as well as a transient expression experiment in maize protoplasts.

At position −124 upstream of the main transcription start the sequence CCAAGCT has similarity to the CAAT-box (GG$^{C}_{T}$CAATCT, Breathnach and Chambon, 1981). The sequence (TATTTATT) located at position −29 resembles a TATA box. Both positions are marked in Figure 4.

The structure of the transcription unit

Most of the gene structure was deduced from comparison of a nearly full length cDNA clone to the genomic DNA sequence. The DNA sequence of the genomic clone has been determined by sequencing both strands, while only one strand of the cDNA clone was analyzed. As the cDNA clone pWW110/1 ends in exon 3 (Figure 1), the position of exon 1 and 2 within the genomic sequence and the 5′ end of exon 3 were determined by a combination of primer extension and S1 mapping experiments. At the transcription start site the results from these two types of experiments do not agree completely. The primer extension experiment places the transcription start at position +3 and +4. In the S1 experiment, start sites are found within a 13-bp region, extending from −3 to +10. Since these bands are found even after extended incubation with S1 (data not shown), they seem to really reflect different start sites rather than exonucleolytic digestion by S1. The two strongest start sites are located two nucleotides

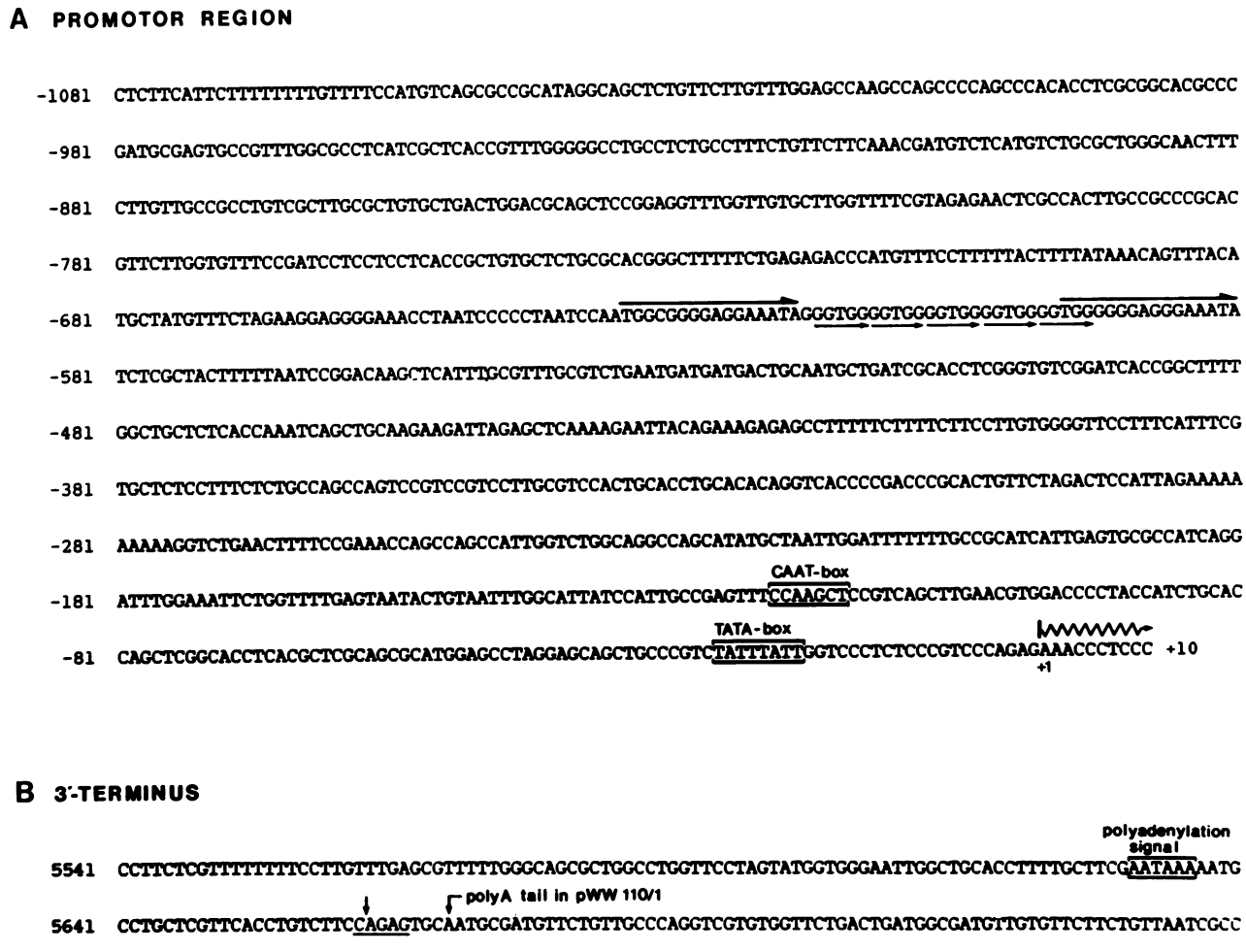

Fig. 4. Selective DNA sequences. **(A)** DNA located in front of the transcription start. **(B)** 3′ end of the sucrose synthase gene. The arrow 7 bp in front of the poly(A) tail in pWW110/1 indicates the position where polyadenylation starts in the cDNA clone isolated from W23 x K55.

upstream from those determined by the primer extension experiments. This difference may be explained by incomplete extension due to the cap structure at the 5′ end of the mRNA (Banerjee, 1980; Nathans and Hogness, 1983). It is not clear why the microheterogeneity observed in the S1 protection experiment is not also seen in the primer extension experiment. It is possible that the minor bands have not been detected in the primer extension experiments due to the high background found in these autoradiographs. Microheterogeneity at the 5′ end of the mRNA has also been determined for a soybean lectin gene (Vodkin *et al.*, 1983) as well as for the maize alcohol dehydrogenase gene (Dennis *et al.*, 1984).

The first ATG is found 72 nucleotides downstream of the main transcription start within exon 2. It is located within the sequence GAGCCATGG, which has extensive homology with other translation start sites of eucaryotic genes (CC$^{A}_{G}$CCATGG, Kozak, 1984).

This ATG starts an open reading frame of 2406 bp or 802 amino acids which predicts a mol. wt. for the protein of 91.731 daltons. The open reading frame is followed by an untranslated region of 269 nucleotides between the terminal TAG and the poly(A) addition in our cDNA clone pWW110/1. The exact polyadenylation site cannot be determined, because genomic clone and cDNA diverge at the position of two A residues, as is often found in polyadenylation sites. 31 nucleotides in front of the poly(A) tail the sequence AATAAA is found. This signal is usually found in front of the poly(A) addition site (Proudfoot and Brownlee, 1976; Proudfoot, 1984). The sequence CAYUG is often observed between the AATAAA signal and the polyadenylation site (Berget, 1984). The only sequence slightly resembling this signal is the pentanucleotide CAGAG located three nucleotides in front of the poly(A) tail.

A cDNA clone isolated from maize line W23 x K55 by Chaleff *et al.* (1981) has a poly(A) tail starting 7 bp upstream of the poly(A) addition site in pWW110/1 (Sheldon *et al.*, 1983; C.Hannah, personal communication). The reason for the utilization of different polyadenylation sites remains unclear because the corresponding genomic sequences of W23 x K55 and line C are not known and might contain differences, which could explain this result. The sequence AATAAA is also found three times within the primary transcript, in introns 2, 6 and 11 (Table I). This confirms the observation by Proudfoot (1984) that AATAAA is not sufficient for poly(A) addition to occur or else that some poly(A) addition products are unstable and escape detection.

Exons and introns

The sizes of the exons and introns are shown in Figure 1. The exon-intron borders are listed in Table I. Only intron 1 (1114 bp) and 2 (511 bp) are larger than 162 bp, seven of the other introns are even shorter than 100 bp.

Table I.

Selected sequences in the transcription unit

1080	GAGCCATGG translation start
1585	CTTCAGATCTAATAAAAAGGATATGAGATGTCCATC
2592	GTGAATGCTCAATAAAACGTTCTGTACTTGCTATGG
3852	TTTAGTAGTAAATAAAACTAGTATGTGATGTTTTCT
5401	TAG end of the open reading frame

Exon/intron boundaries

49	GGG	GTATGCTT	intron 1	AGCTCGAATTGCAG	TAT
1177	CAG	GTGGGCTT	intron 2	TACCACTTCTACAG	GTA
1809	CAG	GTAACACT	intron 3	TTGTCTGCATATAG	GAA
2048	ACA	GTAAGTTC	intron 4	TCCTTTTTTACCAG	ATC
2320	ACG	GTGAGCTT	intron 5	GTTTTCTGTTACAG	ACG
2589	TAG	GTGAATGC	intron 6	ATGATCTGTGTTAG	GTT
2904	CAG	GTACAAAA	intron 7	CAGTCGCTTTGCAG	GTT
3075	ATT	GTATGTTT	intron 8	CTTATTTGTTGCAG	GTT
3344	GAG	GTATACAG	intron 9	ATTCTGTGCTGCAG	GAT
3544	CAG	GTCTGTTT	intron 10	GTACATACTTGCAG	TGT
3820	AAG	GTAGAATT	intron 11	TGTTGTTTCTGCAG	CAA
4125	CAA	GTGAGTAT	intron 12	TTACTTGCTTCCAG	GTT
4578	CAG	GTATATGC	intron 13	TTTTGTGTGGGTAG	CCT
4951	GAA	GTATGCAT	intron 14	TTTGGATTGCTCAG	GTA
5252	CTG	GTAAGCCG	intron 15	TTTCTGGAATCCAG	GCA

The numbers in front of the short DNA sequences shown correspond to the number of the first nucleotide within our complete genomic sequence information. The positions of the potential polyadenylation signals AATAAA within the transcribed region are located in introns 2, 6 and 11, respectively.

The borders between exons and introns are in full agreement with the GT-AG rule (Breathnach and Chambon, 1981). The consensus sequence of introns at the donor site (Mount, 1982) is well conserved with the exception of position +6, where we observe a preference (seven out of 15) of C instead of T. The generally less-conserved acceptor site agrees well with the consensus sequence of published introns of animal genes.

Introns 1 – 14 carry a stop codon in frame, thus preventing translation of the unspliced RNA. Intron 15 is an exception. It could be translated in-frame with exons 15 and 16, yielding a translation product ending at a stop codon 67 bp in front of the AATAAA in a region that is not translated from the mature mRNA. Whether this is of biological significance is not known.

Base composition

Three contiguous exons, no. 13, 14 and 15, are significantly more GC-rich (57, 58 and 60%) than the other exons (51%). No such deviation is seen in the introns, which are in general more AT-rich than the exons. Much of the GC excess arises in third codon position and is thus not ascribable to the amino acid sequence. It will be interesting to see whether the 3-dimensional structure of the protein indicates that exons 13 – 15 form a separate domain. In this case, it could be discussed whether the gene is composed of two subgenes that have evolved separately. The lack of GC excess in the introns might then indicate that these have been added to the gene after the evolution of the coding domain. The dinucleotide CpG is under-represented in eucaryotic DNA (Bird, 1980). The distribution of CpG in the exons is shown in Figure 1. Again, exons 13, 14 and 15 show an excess of CpG, which is even higher than expected on the basis of the CG-content (46 expected/64 observed).

Codon preferences

Codon usage in the sucrose synthase gene has been tabulated and is available from the authors on request. An interesting deviation from a random usage is found for those codons where the second base is a T, and where A and G are used synonymously in the third position. In these cases, G is preferred strongly over A (valine GTG/A: 17/2; leucine TTG/A: 16/0; CTG/A: 34/3). A preference for the TG over TA in second and third codon position is also found in other eucaryotic genes (Beyreuther *et al.*, 1983) but to a lesser extent. The codon preference described results in a measurable though numerically smaller preference for TG over TA dinucleotides in the exons. No such preference is seen in the introns. This might indicate that the selection leading to this preference is exerted at the level of translation. It has not been reported, however, that these codons are served by different tRNA molecules (Sprinzl and Gauss, 1984).

Table II.

Transitions: 11	Transversions: 6
C→T (5), T→C (4)	A→C, G→T, C→G
A→G, G→A	C→A, T→G, T→A

Deviations between the DNA sequence of the genomic clone and the cDNA clone pWW110/1 within the protein coding region. The G to A transition is a difference between the genomic clone and the cDNA clone pKS500.

Evolutionary aspects

The genomic clone sequenced by us was isolated from one of McClintock's strains. Our long cDNA clone pWW110/1 was isolated from line C which is a corn belt dent. Our short cDNA clone pKS500 is derived from the German commercial hybrid EDO. The genomic clone described by Sheldon *et al.* (1983) was isolated from Black Mexican Sweet. Some sequence information from the 3′ end of the gene of this clone is available. We could thus search for sequence heterogeneities between these maize lines.

A comparison between the genomic and the cDNA sequence yields information about the exons. Since our cDNA clone (pWW110/1) is not complete at the 5′ end and since we have not sequenced that part of it which was already known from the cDNA clone pKS500, we can compare a length of 2100 bp of exon DNA only. In the exon sequences investigated, we find 16 base substitutions, both transitions and transversions (see Table II). Though the cDNA clone has been sequenced from one strand only, we have inspected our gels carefully and could read the positions in question unambiguously. Whether the difference arises from mistakes of the reverse transcriptase cannot be excluded without sequencing another cDNA clone. It is noteworthy, however, that all of these differences are located in third codon positions and none of them leads to an amino acid substitution. This can hardly be explained by random mistakes of the reverse transcriptase reaction. Within the 270-bp protein coding region of our previously isolated cDNA clone pKS500 (Geiser *et al.*, 1980) we find one transition in the 3rd codon position, which is also silent on the protein level. Miyata *et al.* (1982) reported an accumulation of silent base substitutions in several animal genes of 5.37×10^{-9}/year.

We have compared a sequence of 2100 bp length between two maize lines, of which 540 positions potentially lead to silent substitutions. As we found 16 differences between the two sequences, we arrived at an evolutionary distance of 6×10^6 years between the two maize lines implying a separation time of 3×10^6 years. If we compare the 3′-untranslated region of the two sucrose synthase isolates mentioned above, we find within 270 bp 10 base substitutions (in addition to two deletion/insertion mutations). Since the 3′-untranslated region does not allow the distinction

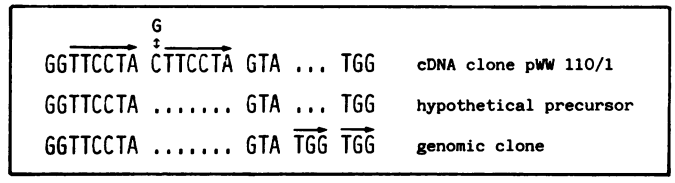

Fig. 5. Comparison of DNA sequences found in the 3′-untranslated region of different alleles of the *shrunken* locus, which could reflect transposon footprints. A possible common precursor allele is shown.

between silent and other codon positions, we compare all substitutions in this sequence of 270 bp with the number of substitutions in the 540 silent positions within coding sequences. The numbers found (16/540 and 10/270) are so similar that we must assume that the nucleotide exchanges in 3′-untranslated regions are accumulated at similar rates and are thus no more or less conserved than the silent positions within the coding region.

The calculated distance between the two alleles would mean that the two alleles have been separated 3 x 10^6 years ago. This is two orders of magnitude longer than the time that has elapsed since the domestication of maize (Galinat, 1977). This may either mean that *Z. mays* has not gone through a very small population size upon domestication and that thus much of the diversity of its progenitors is reflected in its present genotype, or else that *Z. mays* evolves two orders of magnitude more rapidly than other species.

Comparison between these two sequences and those available from some other lines (Black Mexican Sweet, Sheldo *et al.*, 1983; pKS500, this study; W23 x K55, Chaleff *et al.*, 1981) detects a varying degree of divergence ranging from no substitutions between pKS500 and isolated from the German line EDO and the genomic clone sequenced here to 17 deviations between Black Mexican Sweet and line C. These results may indicate a lack of selection for the sequence of the 3′-untranslated region of the sucrose synthase gene.

In any case, it is interesting to note that we find no amino acid polymorphisms. It is surprising that a protein as large as sucrose synthase (mol. wt. 92 000) does not have amino acids, the replacement of which is neutral enough to be detected as a sequence polymorphism. This result is similar, however, to findings with the alcohol dehydrogenase gene of *Drosophilia melanogaster* (Kreitmann, 1983) and supports his suggestion that most amino acid replacements are not neutral, but are rather selected against.

Spontaneous mutations in maize are often caused by the insertion of transposable elements. The integration of a transposon creates a short sequence duplication in the host sequence. In contrast to transposons of other organisms, these duplications persist upon excision of the transposon. They are, however, often mutated during the excision process (Sachs *et al.*, 1983; Bonas *et al.*, 1984; Fedoroff *et al.*, 1983; Weck *et al.*, 1984). Mutations occurring during the excision process are explained by a hypothesis advanced by Nevers *et al.* (1985).

If the integration of the transposon has occurred in an exon, the remaining duplication alters the predicted sequence of the revertant protein, if the number of remaining nucleotides is three or a multiple thereof. Two such instances have been described by Schwarz-Sommer *et al.* (1985). Proteins found in maize strains after the reversion of a transposon-induced mutation are sometimes altered in their properties (Echt and Schwartz, 1981; Dooner and Nelson, 1979; Shure *et al.*, 1983). These alterations may be due to the presumed insertions of one or a few additional amino acids.

Schwarz-Sommer *et al.* (1985) have also compared intron sequences of the *Waxy* gene of *Z. mays* obtained from different lines. They found perfect or near-perfect sequence duplications that could be ascribed to the insertion and subsequent excision of transposable elements. They advanced the hypothesis that transposon insertions and excisions are frequent events and may thus considerably contribute to the evolution of the maize genome. We have analyzed the 3′-untranslated sequences of five different *Sh* alleles and found two duplications of 7 and 3 bp each that are present in one and missing in the other strain (Figure 5). The two alleles may have evolved from a common precursor. The 7-bp duplication could be the footprint of an as yet unknown transposable element, or it could have been generated from an insertion causing a larger duplication of the insertion site by the deletion of one or a few base pairs (Bennetzen *et al.*, 1984; Döring and Starlinger, 1984). The 3-bp duplication could be the result of an En/Spm insertion and subsequent excision (Schwarz-Sommer *et al.*, 1985).

It must be asked, however, how much these transposon footprints contribute to the evolution of proteins. It is noteworthy in this respect that many of the transposon insertions known have occurred within an intron. This is also true for the three analyzed transposon insertions in the *Shrunken* gene. It is conceivable that many transposon insertions in exons leave footprints that are not comparable with protein functions. Such transposon insertions will not appear to revert frequently and may thus not be recognized as caused by transposable elements. The above-mentioned observations about amino acid replacements caused by single base substitutions supports this assumption. It should be remembered, however, that enzymes found presently may well be adapted to their function and that a more rapid evolution may have occurred in the past.

Materials and methods

Materials

The genomic clone SS1 (Geiser *et al.*, 1982) and the cDNA clone pKS500 (Geiser *et al.*, 1980) were described previously. pWW110/1 was isolated from a cDNA library prepared by Schwarz-Sommer *et al.* (1985) and isolated by standard methods (Maniatis *et al.*, 1982).

Restriction endonucleases were purchased from BRL (Neu-Isenburg), Biolabs (Dreieich) or Boehringer (Mannheim). T4 polynucleotide kinase, DNA polymerase large fragment (Klenow enzyme), alkaline phosphatase and DNase I from Boehringer (Mannheim); T4 ligase from Biolabs (Dreieich). Endonuclease S1 was from Sigma (München). AMV reverse transcriptase was from a preparation of J.Beard (Life Science). ^{32}P-labeled nucleoside triphosphates were purchased from Amersham Buchler (Braunschweig).

Subcloning and plasmid preparation

Vectors were pBR322 (Bolivar *et al.*, 1977), pUC9 (Vieira and Messing, 1982) and pUR250 (Rüther, 1982), which were used for transformation of *Escherichia coli* K12 strains HB101 or RR1ΔM15 (Rüther, 1982). Plasmid DNA was prepared in small amounts by the alkaline lysis method (Maniatis *et al.*, 1982) and larger quantities by the method of Clewell and Helinski (1969).

RNA preparation

Sh RNA was isolated from the German hybrid EDO. The maize line carrying the allele *sh bz-m4* was obtained from B.McClintock. Immature ears were frozen in liquid nitrogen 20–22 days after pollination and stored at −70°C. Frozen kernels were ground, incubated with proteinase K, phenol/chloroform extracted as described by Kloppstech and Schweiger (1976). Poly(A)$^+$ RNA was isolated by oligo(dT)-cellulose affinity chromatography (Collaborative Research). RNA was stored in 70% ethanol at −70°C.

DNA sequencing

The chemical degradation method (Maxm and Gilbert, 1980) was used with modifications described by Garoff and Ansorge (1981). Ordered deletions were created by random DNase I cleavage within the maize insert and a unique restriction site in the vector (Frischauf *et al.*, 190). Deletions with end points at distances of 250-300 bp were used for sequencing both strands after labeling by poly-

nucleotide kinase or fill in reaction by Klenow enzyme. Some restriction sites used for labeling were determined by sequencing DNA fragments spanning these sites.

Preparation of DNA-RNA hybrids and digestion with endonuclease S1

DNA-RNA hybrids were prepared as described by Berk and Sharp (1979). 5 – 10 μg of poly(A)$^+$ RNA or equivalent amount of tRNA were hybridized to 50 ng up to 1 μg of DNA fragments within a total volume of 20 μl. Hybridization was carried out overnight at 43°C. Single-stranded nucleic acids were digested with 200 U S1 endonuclease at 25°C for 30 – 60 min. Reaction was stopped by addition of 5 μg tRNA and 2.5 volumes ethanol.

Primer extension assay and RNA sequencing

The synthetic oligonucleotide AAGCATTCCCTTGCCC was synthesized by a 380 A DNA synthesizer (Applied Biosystems) with subsequent purification of the 16 mer on a 20% acrylamide gel. The reverse transcriptase reaction was carried out as described by Hamlyn *et al.* (1978) with slight modifications. 2.5 pmol oligonucleotide were incubated in 50 mM Tris pH 8.3, 150 mM KCl, 10 mM $MgCl_2$, 20 mM DTT, 40 μM dTTP, dGTP, dCTP each and 6 μCi [α-^{32}P]ATP (400 Ci/mmol) with 2 U AMV reverse transcriptase at 42°C. After 15 min 1 μl of chase mix (250 μM dNTP in 100 mM Tris pH 8.3, 1 U AMV reverse transcriptase/μl) were added and incubation continued for 15 min. 2 μg tRNA were added, nucleic acids were ethanol precipitated and reverse transcription products analyzed on 6% polyacrylamide gels containing 7 M urea. For sequencing four reactions were carried out in the presence of one dideoxynucleotide each.

The computer analyses of DNA sequences were run on a VAX/VMS computer version 3.5. The computer programs used were developed at the University of Wisconsin (Devereux *et al.*, 1984).

Acknowledgements

We are greatly indebted to A.Gierl and Zs.Schwarz-Sommer, Köln-Vogelsang, for the opportunity to use an endosperm cDNA library from line C maize for the isolation of pWW110/1 clone. We thank H.Reinke and K.Beyreuther (Köln) for analysis of the purified sucrose synthase protein. This work was supported by Deutsche Forschungsgemeinschaft through SFB 74.

References

Banerjee,A.K. (1980) *Microbiol. Rev.*, **44**, 175-205.
Bennetzen,J., Swanson,J., Taylor,W.C. and Freeling,M. (1984) *Proc. Natl. Acad. Sci. USA*, **83**, 4125-4129.
Berget,S.M. (1984) *Nature*, **309**, 179-181.
Berk,A.J. and Sharp,P.A. (1977) *Cell*, **12**, 721-732.
Beyreuther,K., Stüber,K., Bieseler,B., Bovens,J., Dildrop,R., Geske,T., Triesch, I., Trinks,K., Zaiss,S. and Ehring,R. (1983) in Jensen,U. and Fairbrothers,D.E. (eds.), *Proteins and Nucleic Acids in Plant Systematics*, Springer-Verlag, Berlin/ Heidelberg.
Bird,A.P. (1980) *Nucleic Acids Res.*, **8**, 1499-1505.
Bolivar,F., Rodriguez,R.L., Green,P.J., Betlach,M.C., Heynecker,H.L., Boyer, H.W., Crosa,H.J. and Falkow,S. (1977) *Gene*, **2**, 95-113.
Bonas,U., Sommer,H. and Saedler,H. (1984) *EMBO J.*, **3**, 1015-1019.
Breathnach,R. and Chambon,P. (1981) *Annu. Rev. Biochem.*, **50**, 349-383.
Burr,B. and Burr,A.F. (1981) *Genetics*, **98**, 143-156.
Chaleff,D., Mauvais,J., McCormick,S., Shure,M., Wessler,S. and Fedoroff,N. (1981) *Carnegie Inst. Wash. Yearbook*, **80**, 158-174.
Cheng,S., Arnd,K. and Lu,P. (1984) *Proc. Natl. Acad. Sci. USA*, **81**, 3665-3669.
Chourey,P.S. (1981) *Mol. Gen. Genet.*, **184**, 372-376.
Chourey,P.S. and Nelson,O.E. (1976) *Biochem. Genet.*, **14**, 1041-1055.
Clewell,D.B. and Helinski,D.R. (1969) *Proc. Natl. Acad. Sci. USA*, **62**, 1159-1166.
Courage-Tebbe,U., Döring,H.P., Fedoroff,N. and Starlinger,P. (1983) *Cell*, **34**, 383-393.
Davidson,E.H., Jacobs,H.T. and Britten,R.J. (1983) *Nature*, **301**, 468-470.
Dennis,E.S., Gerlach,W.L., Pryor,A.J., Bennetzen,J.L., Inglis,A., Llewellyn,D., Sachs,M.M., Ferl,R.J. and Peacock,W.J. (1984) *Nucleic Acids Res.*, **12**, 3983-4000.
Devereux,J., Haeberli,P. and Smithies,O. (1984) *Nucleic Acids Res.*, **12**, 387-395.
Döring,H.P. and Starlinger,P. (1984) *Cell*, **39**, 253-259.
Döring,H.P., Geiser,M. and Starlinger,P. (1981) *Mol. Gen. Genet.*, **184**, 377-380.
Dooner,H.K. and Nelson,O.E. (1979) *Proc. Natl. Acad. Sci. USA*, **76**, 2369-2371.
Downton,W.J.S. and Hawker,J.S. (1973) *Phytochemistry*, **12**, 1551-1556.
Echt,C.S. and Schwartz,D. (1981) *Genetics*, **99**, 275-284.
Fedoroff,N., Mauvais,J. and Chaleff,D. (1983) *J. Mol. Appl. Genet.*, **2**, 11-30.
Frischauf,A.M., Garoff,H. and Lehrach,H. (1980) *Nucleic Acids Res.*, **8**, 5541-5549.
Galinat,W.C. (1977) in Sprague,G.F. (ed.), *Corn and Corn Improvement*, American Society of Agronomy Inc., Madison, WI, pp. 1-47.
Garoff,H. and Ansorge,W. (1981) *Anal. Biochem.*, **115**, 450-457.
Geiser,M., Döring,H.P., Wöstemeyer,J., Behrens,U., Tillmann,E. and Starlinger, P. (1980) *Nucleic Acids Res.*, **8**, 6175-6188.
Geiser,M., Weck,E., Döring,H.P., Werr,W., Courage-Tebbe,U., Tillmann,E. and Starlinger,P. (1982) *EMBO J.*, **1**, 1455-1460.
Hamlyn,P.H., Brownlee,G.G., Chen-Chi Cheng, Gait,M.J. and Milstein,C. (1978) *Cell*, **15**, 1067-1075.
Hawker,J.S. (1971) *Phytochemistry*, **10**, 2313-2322.
Klopptech,K. and Schweiger,G. (1976) *Cytobiologie*, **13**, 394-400.
Kozak,M. (1984) *Nucleic Acids Res.*, **12**, 857-872.
Kreitman,M. (1983) *Nature*, **304**, 412-417.
Langridge,P. and Feix,G. (1983) *Cell*, **34**, 1015-1022.
Maniatis,T., Fritsch,E.F. and Sambrook,J. (1982) *Molecular Cloning: A Laboratory Manual*, published by Cold Spring Harbor Laboratory Press, NY.
Maxam,A. and Gilbert,W. (1980) *Methods Enzymol.*, **65**, 499-560.
McCormick,J., Mauvais,J. and Fedoroff,N. (1982) *Mol. Gen. Genet.*, **187**, 494-500.
Miyata,T., Hayashida,H., Kikuno,R., Hasegawa,M., Kobayashi,M. and Koike,K. (1982) *J. Mol. Evol.*, **19**, 28-35.
Mount,S.M. (1982) *Nucleic Acids Res.*, **10**, 459-472.
Nathans,J. and Hogness,D.S. (1983) *Cell*, **34**, 807-814.
Nevers,P., Shepherd,N. and Saedler,H. (1985) *Adv. Bot. Res.*, in press.
Preiss,J. and Levi,C. (1980) in Preiss,J. (ed.), *The Biochemistry of Plants*, Vol. 3, Academic Press, NY, pp. 371-424.
Proudfoot,N. (1984) *Nature*, **307**, 412-413.
Proudfoot,N. and Brownlee,G.G. (1976) *Nature*, **263**, 211-214.
Rüther,U. (1982) *Nucleic Acids Res.*, **10**, 5765-5772.
Sachs,M.M., Peacock,W.J., Dennis,E.S. and Gerlach,W.L. (1983) *Maydica*, **28**, 289-302.
Schwarz-Sommer,Z., Gierl,A., Coyper,H., Peterson,P. and Saedler,H. (1985) *EMBO J.*, **4**, 579-583.
Sheldon,E., Ferl,R., Fedoroff,N. and Hannah,L.C. (1983) *Mol. Gen. Genet.*, **190**, 421-426.
Shure,M., Wessler,S. and Fedoroff,N. (1983) *Cell*, **35**, 225-233.
Sprinzel,M. and Gauss,D.H. (1984) *Nucleic Acids Res.*, **12**, 1-58.
Tsai,C.Y. (1974) *Phytochemistry*, **13**, 885-891.
Vieira,J. and Messing,J. (1982) *Gene*, **19**, 259-268.
Vieweg,G.H. (1974) *Planta*, **116**, 347-359.
Vodkin,L.O., Rhodes,R.R. and Goldberg,R.B. (1983) *Cell*, **34**, 1023-1031.
Weck,E., Courage-Tebbe,U., Döring,H.P., Fedoroff,N. and Starlinger,P. (1984) *EMBO J.*, **3**, 1713-1716.
Weiher,H., König,M. and Gruss,P. (1983) *Science (Wash.)*, **219**, 626-631.

Received on 13 March 1985