Proc. Natl. Acad. Sci. USA
Vol. 79, pp. 1022–1026, February 1982
Biochemistry

Complete nucleotide sequence of a soybean actin gene

(introns/gene structure/multigene family/evolution)

DILIP M. SHAH, ROBIN C. HIGHTOWER, AND RICHARD B. MEAGHER

Department of Molecular and Population Genetics, University of Georgia, Athens, Georgia 30602

Communicated by Norman H. Giles, October 21, 1981

ABSTRACT Soybean contains a small multigene family of actin-related sequences. We have determined the complete nucleotide sequence of a soybean actin gene carried on the recombinant plasmid pSAc3. As deduced from the nucleotide sequence, this soybean actin is composed of 376 amino acids. Compared to other eukaryotic actins, pSAc3 actin has a deletion of one amino acid between residues 118 and 122. The initiator methionine is followed by alanine, which is not found at this position in other eukaryotic actins. pSAc3 actin differs, in primary sequence, more from fungal and animal actins than any of the known nonplant actins differ from each other. pSAc3 actin appears to be related to both cytoplasmic and muscle specific actins in the location of specific NH_2-terminal amino acids. The coding sequence is interrupted by three small introns, each less than 90 base pairs long. The splice junctions are similar to those found in other eukaryotic genes, suggesting the presence of a similar splicing apparatus in higher plants. Introns 1 and 3 interrupt the reading frame after codons 20 and 355, respectively. Intron 2 splits a glycine codon at position 151. None of these intron positions is conserved relative to the positions of introns in other actin genes examined.

Actin is a ubiquitous protein in eukaryotic cells. Although higher plants are not motile, actin is involved in a number of processes at the cellular level—for example, chromosomal movement and condensation, cytoplasmic streaming, and maintenance of cytoskeletal structure. Actin-like proteins and filaments have been reported in a number of plants (1–4). Recently, soybean has been shown to contain a 45,000-dalton protein which immunoprecipitates with antisera against mammalian cytoplasmic actin (5).

Because of the role actin plays in plant cells, one might expect plant actins to resemble cytoplasmic animal actins. The actins of lower eukaryotes—*Physarum* (6, 7), yeast (8–10), and *Dictyostelium* (11)—resemble the cytoplasmic actins of animal cells in the location of a few critical amino acids which distinguish cytoplasmic actins from muscle actins. Surprisingly, all six *Drosophila* actin genes encode proteins whose amino-terminal sequences are similar to those of cytoplasmic actins (12).

We have reported that a small multigene family of actin-related sequences exists in soybean (13). Heteroduplex studies with a *Drosophila* actin gene demonstrated that the soybean actin genes, contained in the recombinant plasmids pSAc1 and pSAc3, are conserved over most of the polypeptide encoding region.

In this paper we describe the complete nucleotide sequence of the soybean actin gene in pSAc3. Comparison of the pSAc3 actin DNA sequence with yeast and *Drosophila* actin sequences reveals a few unique features of this gene which are pertinent to the function and evolution of actin gene families in higher plants.

MATERIALS AND METHODS

Plasmid. Recombinant plasmid pSAc3, bearing a soybean actin gene on a 3.0-kilobase (kb) *Hin*dIII fragment of soybean genomic DNA has been described; the actin coding sequence on this fragment spans approximately 1.4 kb as determined by Southern blotting and heteroduplex experiments (13). Plasmid DNA was isolated as described by Meagher *et al.* (14).

Fine Structure Restriction Mapping. The 3.0-kb *Hin*dIII fragment was isolated from pSAc3 by preparative agarose gel electrophoresis and further purified by banding in a CsCl-ethidium bromide gradient (15, 16). These *Hin*dIII ends and ends produced by other restriction endonucleases were labeled by using [α-^{32}P]dNTP (New England Nuclear) and the Klenow fragment of *Escherichia coli* DNA polymerase I (New England BioLabs). The 50-μl reaction mixtures contained 5–50 pmol of ends, 2500 pmol of three unlabeled trinucleotides, and 60 pmol (usually 50 μCi, 1 Ci = 3.7×10^{10} becquerels) of the radioactive nucleotide in 50 mM NaCl/10 mM Tris·HCl, pH 7.6/10 mM $MgCl_2$/1 mM dithiothreitol. Reactions were initiated on ice with 2 units of Klenow fragment and carried out at 15°C for 20 min. In order to ensure complete filling in of the ends, an excess of the limiting nucleotide was then added for an additional 20 min. The reaction was stopped by heating to 65°C. In control experiments under these conditions, 30 pmol of *Sau*3A ends were 90% labeled in 15 min. The fragment, labeled at both ends, was cleaved asymmetrically with a restriction enzyme and the resulting fragments were purified by gel electrophoresis, electroelution, and ethanol precipitation. Partial digestion products of the end-labeled fragments were electrophoresed on agarose gels; the gels were fixed in 10% acetic acid and autoradiographed after drying. Restriction sites were mapped by the difference in the molecular weights of the end-labeled partial digestion products as described by Smith and Birnstiel (17).

DNA Sequence Determination. Fragments to be analyzed were purified by electroelution from 5% acrylamide gels in the buffer described by Carreira *et al.* (16) followed by two ethanol precipitations. Fragments were end labeled as described above and analyzed by the method of Maxam and Gilbert (18).

Computer Analysis of DNA Sequence Data. Sequence data were stored and analyzed on the DNA sequence analysis programs (SEQ) that are part of the Stanford Molgen Project and the National Institutes of Health SUMEX-AIM Facility. The yeast actin sequence (10) was obtained from the DNA sequence bank maintained by the National Biomedical Research Foundation at Georgetown University. An Apple II plus computer in combination with a D. C. Hayes microcoupler was used to interact with the Stanford system via the TYMNET satellite communications system. Files were transferred between the Apple system and these distant computers by using a BITS program (Software Sorcery, McLean, VA).

Abbreviation: kb, kilobase(s).

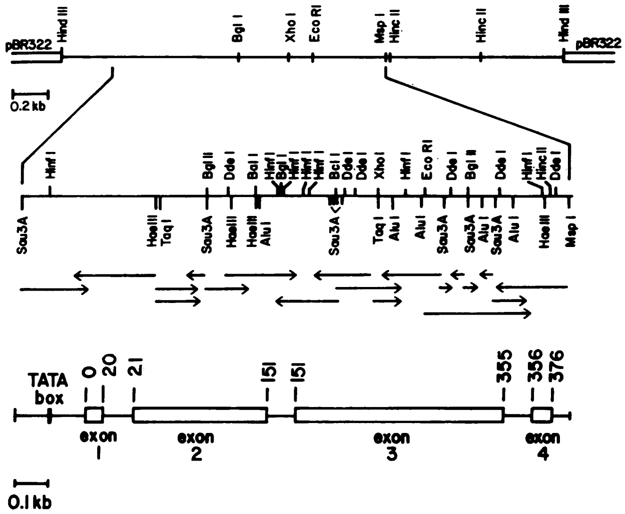

FIG. 1. Physical map of the 3.0-kb *Hin*dIII insert in pSAc3 containing the actin gene. Fine structure restriction map of the actin-encoding sequence including portions of the flanking regions and introns is drawn on an expanded scale. The sites were determined by computer analysis of the nucleotide sequence confirming the map derived by restriction analysis. The DNA sequence strategy is shown just below the fine restriction map. The functional diagram of the pSAc3 actin gene as derived from DNA sequence data is drawn colinear with the fine structure restriction map. The locations of exons (open boxes), introns, flanking sequences, and a presumed T-A-T-A box are shown. Codon numbers are used to indicate junctions between flanking sequences, exons, and introns.

RESULTS AND DISCUSSION

The isolation of two recombinant phages, λSAc1 and λSAc3, containing the actin-related sequences from the genomic library of soybean has been described (13). Using electron microscope heteroduplex mapping, we showed that the actin-related sequences within the subclones pSAc1 and pSAc3 had homology with the entire actin-encoding sequences of *Drosophila* and *Dictyostelium.* The actin-encoding sequence of pSAc1 was also shown to contain a presumed intervening sequence of approximately 300 base pairs located 390 base pairs from the 5′ end of the actin-encoding region. There was no evidence for the presence of intervening sequences in pSAc3. Intervening sequences less than 100 base pairs long could not have been detected by heteroduplex analysis.

A physical map of the 3.0-kb *Hin*dIII insert in pSAc3 showing the cleavage sites for a number of four-base and six-base restriction endonucleases was deduced and used to develop a sequence-determination strategy. The physical map and the strategy are shown in Fig. 1. The complete nucleotide sequence of the pSAc3 actin coding region including its introns and a 195-base-pair 5′ flanking region is shown in Fig. 2.

Soybean Actin Is Highly Diverged. As deduced from the nucleotide sequence, pSAc3 actin contains 376 amino acids. There are two differences between pSAc3 actin and other actins which affect the colinearity of the amino acid sequence. First, comparison of the primary sequence of pSAc3 actin with that of other eukaryotic actins reveals a deletion of one amino acid between residues 118 and 122. Second, as shown in Fig. 3, the initiation codon ATG of pSAc3 actin gene is followed by a codon

FIG. 2. Complete nucleotide sequence of the pSAc3 actin gene including its introns and 195 base pairs of the 5′ flanking region. The derived amino acid sequence of pSAc3 actin is shown and numbered separately from the nucleotide sequence. Amino acids at positions 104 and 108 are undetermined.

```
        10         20         30         40         50         60         70
GATCTTGTGT GTGTGTGTGT GTTTTGTGCC ATGTACTTTT GCTATTTCAG GTTCAAAACT GTATGGTCTT

        80         90        100        110        120        130        140
TTAAGTGAAT CCTGGTCTCN GTTGGCTGTA AATGTCCTGG TTTCACGGGT TGTCCATACA ATACAATTTT

       150        160        170        180        190                3
GTTACTTACT GCACTTTACT TCATGACACC CATGCTTTCT GTAGGTTTGT AAAAG ATG GCA GAT
                                                             MET Ala Asp

                                   13                                  20       265
GCC GAG GAT ATT GAA CCC CTT GTT TGC GAC AAT GGA ACT GGA ATG GTC AAG GTTAGTACTT
Ala Glu Asp Ile Glu Pro Leu Val Cys Asp Asn Gly Thr Gly MET Val Lys

       275        285        295        305        315        325        335
CGTGGATTTT TGTTTTGTTG CATGTGCATA TGCAAGTCTT GAGTTCATAA GTGAATAATT TATTTTAATA

          21                          28                          36
TGGAACAG GCA GGA TTT GCT GGT GAC GAT GCT CCT AGG GCT GTC TTT CCC AGT ATT
         Ala Gly Phe Ala Gly Asp Asp Ala Pro Arg Ala Val Phe Pro Ser Ile

                                     46                              54
GTT GGC CGA CCT CGA CAT ACT GGT GTT ATG GTT GGG ATG GGT CAA AAG GAT GNC
Val Gly Arg Pro Arg His Thr Gly Val MET Val Gly MET Gly Gln Lys Asp

                                     64                              72
TAT GTT GGT GAT GAA GCC CAA TCC AAA AGA GGT ATT CTT ACT CTC AAG TAT CCC
Tyr Val Gly Asp Glu Ala Gln Ser Lys Arg Gly Ile Leu Thr Leu Lys Tyr Pro

                                     82                              90
ATT GAG CAT GGT ATA GTC AGC AAC TGG GAT GAC ATG GAA AAG ATC TGG CAT CAC
Ile Glu His Gly Ile Val Ser Asn Trp Asp Asp MET Glu Lys Ile Trp His His

                                    100                             108
ACA TTT TAT AAT GAA CTT CGT GTT GTT TCC GAG GAG CCC CAN GTG CTC TCA NCT
Thr Phe Tyr Asn Glu Leu Arg Val Val Ser Glu Glu Pro     Val Leu Ser

                                    118                             126
GAG GCC CCC CTC AAC CCA AAG GTC AAC AGA GAA AGT GCC CAA ATC ATG TTT GAG
Glu Ala Pro Leu Asn Pro Lys Val Asn Arg Glu Ser Ala Gln Ile MET Phe Glu

                                    136                             144
ACC TTC AAT GTG CCT GCC ATG TAT GTG GCC ATC CAA GCT GTT CTC TCC TTG TAT
Thr Phe Asn Val Pro Ala MET Tyr Val Ala Ile Gln Ala Val Leu Ser Leu Tyr

                  150           746        756        766        776
GCA AGT GGT CGC ACA ACT G GTTTGTATATCN CCATCTGTCT GATTCGATTC TTTTATTAAT
Ala Ser Gly Arg Thr Thr

       786        796        806        151                     157
CACATTGTTC TTTCCCTGAA TCAATTCCTT TTAAAACAG GT ATT GTA TTG GAC TCT GGT
                                        Gly Ile Val Leu Asp Ser Gly

                                    167                             175
GAT GGT GTG AGT CAC ACT GTA CCA ATC TAT GAG GGG GAT GCA CTC CCC CAT GCC
Asp Gly Val Ser His Thr Val Pro Ile Tyr Glu Gly Asp Ala Leu Pro His Ala

                                    185                             193
ATC CTT CGT TTG GAT CTC GCT GGT CGT GAT CTA ACT GAT CAT TTG ATG AAG ATC
Ile Leu Arg Leu Asp Leu Ala Gly Arg Asp Leu Thr Asp His Leu MET Lys Ile

                                    203                             211
CTC ACT GAG AGA GGA TAC ATG TTC ACC ACC TCT GCT GAG CGA GAA ATT GTC CGT
Leu Thr Glu Arg Gly Tyr MET Phe Thr Thr Ser Ala Glu Arg Glu Ile Val Arg

                                    221                             229
GAC ATG AAG GAA AAA CTT GCA TAT GTT GCC CTA GAT TAT GAG CAA GAA CTC GAG
Asp MET Lys Glu Lys Leu Ala Tyr Val Ala Leu Asp Tyr Glu Gln Glu Leu Glu

                                    239                             247
ACT GCA AAA AGC AGT TCA TCA GTT GAG AAA AGC CAT GAG CTT CCT GAT GGA CAA
Thr Ala Lys Ser Ser Ser Ser Val Glu Lys Ser His Glu Leu Pro Asp Gly Gln

                                    257                             265
GTT ATC ACA ATT GGG GCA GAG AGA TTC CGT TGC CCA AAA ATT CTT TTC CAA CCA
Val Ile Thr Ile Gly Ala Glu Arg Phe Arg Cys Pro Lys Ile Leu Phe Gln Pro

                                    275                             283
TCT ATG ATT GAA ATG GAA GCT GCT GGA ATT CAT GAG ACC ACC TAC AAT TCT ATC
Ser MET Ile Glu MET Glu Ala Ala Gly Ile His Glu Thr Thr Tyr Asn Ser Ile

                                    293                             301
ATG AAG TGT GAT GTG GAT ATC AGA AAG GAT CTA TAT GGC AAC ATT GTT CTT AGT
MET Lys Cys Asp Val Asp Ile Arg Lys Asp Leu Tyr Gly Asn Ile Val Leu Ser

                                    311                             319
GGT GGC TCA ACT ATG TTT CTT GGT ATT GCA GAC CGT ATG AGC AAG GAG ATC TCT
Gly Gly Ser Thr MET Phe Leu Gly Ile Ala Asp Arg MET Ser Lys Glu Ile Ser

                                    329                             337
GCT CTT GCT CCT AGC AGC ATG AAG ATT AAG GTT GTA GCT CCA CCA GAA AGA AAA
Ala Leu Ala Pro Ser Ser MET Lys Ile Lys Val Val Ala Pro Pro Glu Arg Lys

                                    347                             355
TAT AGT GTC TGG ATT GGA GGA TCA ATC CTT GCA TCC CTC AGC ACC TTC CAG CAG
Tyr Ser Val Trp Ile Gly Gly Ser Ile Leu Ala Ser Leu Ser Thr Phe Gln Gln

      1439       1449       1459       1469       1479       1489       1499
GTGATTATTT TATCATGAAT TAGCTCTTTC GTGTGTGCCA TTTTCATTGA ATGCTAAAAA AAATGTTGTG

          356                     363                     370
GTTTTGCAG ATG TGG ATT TCT AAA GGT GAA TAT GAT GAG TCT GGC CCA TCC ATT
          MET Trp Ile Ser Lys Gly Glu Tyr Asp Glu Ser Gly Pro Ser Ile

                     376              1584       1594       1604       1614
GTC CAC AGG AAG TGC TTC TAA CTTCTAAGTA CAAAATATGC TTCAATGGTG AATTATTTTC
Val His Arg Lys Cys Phe  .

GCTCGA
```

Soybean pSAc3	Met	Ala	Asp	Ala	Glu	Asp	Ile	Gly	Pro	Leu	Val	Cys	Asp	Asn	Gly	Thr	Gly	Met	Val	Lys
Yeast			Met	Asp	Ser	Glu	Val	Ala	Ala	Leu	Val	Ile	Asp	Asn	Gly	Ser	Gly	Met	Cys	Lys
Dictyostelium		X -	Glu Asp	Gly	Glu	Asp	Val	Gln	Ala	Leu	Val	Ile	Asp	Asn	Gly	Ser	Gly	Met	Cys	Lys
Drosophila-DmA-1		Met	Cys	Asp	Asp	Asp	Ala	Gly	Ala	Leu	Val	Ile	Asp	Asn	Gly	Ser	Gly	Met	Cys	Lys
Drosophila-DmA-2		Met	Cys	Asp	Glu	Glu	Val	Ala	Ala	Leu	Val	Val	Asp	Asn	Gly	Ser	Gly	Met	Cys	Lys
Muscle Specific			Asp	Glu	Asp	Glu	Ser Thr	Thr	Ala	Leu	Val	Cys	Asp	Asn	Gly	Ser	Gly	Leu	Val Cys	Lys
Cytoplasmic				Asp Glu	Asp Glu	Asp Glu	Ile	Ala	Ala	Leu	Val	Val Ile	Asp	Asn	Gly	Ser	Gly	Met	Cys	Lys

Variable Regions

FIG. 3. Amino-terminal sequences of pSAc3 actin and yeast (9, 10), *Dictyostelium* (11), *Drosophila* (12), and vertebrate actins (7). Variable regions are underlined.

that specifies alanine. With the exception of these two codons, the amino acid sequence is colinear with the amino acid sequence of other actins. The derived amino-terminal sequence of pSAc1 actin is identical to that of pSAc3 actin. An analogous situation also exists in *Drosophila* actin genes where the initiation codon ATG is followed by a cysteine codon (12). It remains to be established if methionine and alanine are cleaved off the mature polypeptide post-translationally in soybean.

The amino acid sequence of actin from evolutionarily diverged organisms is highly conserved (6). Only 17 amino acid replacements separate *Physarum* actin from mammalian cytoplasmic actin. The greatest divergence yet reported is between yeast and mammalian actins—41 (10.9%) amino acid replacements (10). However, recent data show that *Drosophila* actin differs from yeast actin by 51 (13.6%) amino acid substitutions (F. Sanchez and B. McCarthy, personal communication). The primary sequence of pSAc3 actin reveals 70 (19.0%) and 59 (16.0%) amino acid substitutions compared to yeast and *Drosophila* actins, respectively (ref. 10; F. Sanchez and B. McCarthy, personal communication).

However, nucleotide sequence comparisons show 30.1% divergence between the coding regions of pSAc3 and yeast actin genes and 29.7% divergence between the coding regions of pSAc3 and *Drosophila* actin genes. *Drosophila* and yeast actin-encoding sequences differ in 27% of their nucleotide sequences. Thus, this soybean actin gene appears to be the most highly diverged actin gene yet examined.

The partial nucleotide sequence of a different soybean actin gene in pSAc1 indicates that these two soybean actin genes have diverged in approximately 15–20% of their nucleotide sequences. This shows that there is a high degree of divergence between two members of the soybean actin gene family. This is further supported by the observation that there is little if any similarity in the high-resolution restriction maps of these two genes (unpublished data).

The majority of base substitutions in the pSAc3 actin gene are in the third position of a codon and most have not resulted in amino acid replacements. Many of the amino acid substitutions that have occurred between yeast and pSAc3 actins represent changes to functionally related amino acids. For example, there are 11 threonine/serine interchanges and 6 isoleucine/valine interchanges. It is unlikely that these conservative amino acid replacements are functionally significant.

The amino acid sequence in pSAc3 actin showing the greatest divergence lies in the amino-terminal 18 amino acids. This is the region showing the greatest amino acid variation in other eukaryotic actins (4, 7). Analysis of the amino-terminal tryptic peptides of several vertebrate actins revealed a large number of amino acid substitutions in these peptides (7). These amino acid substitutions were found to be more tissue specific than species specific. For example, muscle actins differ from cytoplasmic actins in a number of amino acid exchanges limited to the amino-terminal tryptic peptide.

We were interested in determining if pSAc3 actin resembled the cytoplasmic actins of other eukaryotic cells. Fig. 3 shows the comparison of the first 20 amino acids of pSAc3 actin with those of other eukaryotic actins. The pSAc3 amino acid sequence -Glu-Asp-Ile- at position 5–7 is characteristic of the cytoplasmic actins of mammalian cells. Surprisingly, however, pSAc3 actin contains cysteine at position 12 and valine at position 19. These amino acids are found at these positions only in muscle specific actins. Based on this comparison, pSAc3 actin

Table 1. Exon/intron junctions

	5′ exon		exon 3′
Gene	Left	intron	Right
pSAc3, 1	AAGGUUAGUACU		AUCGAACAGGCA
pSAc3, 2	GUGGUUUGUAUA		UUAAAACAGCUA
pSAc3, 3	CAGGUGAUUAUU		GUUUUGCAGAUG
Consensus	A/CAGGUAAGU		UYUYYYUXCAGG
Yeast actin	GUGGUAUGUUCU		UAUGUUUAGAGG
Phaseolin A	CAUGUACUG		UUGUCCUGUAGG
Phaseolin B	AAUGUAAGA		GAUUUUUAUAGA
Phaseolin C	GAGGUAAAU		GGGGGAUUUAGG
Lb, 1	AUUCGUAAGU		AAAUAGG
Lb, 2	AUUGGUAAGU		UUGUAGG
Lb, 3	CGUGGUAAGU		UGUAGG

Comparison of the exon/intron junctions of the three introns in pSAc3 actin gene with the consensus sequence (19) and with the exon/intron junctions found in the genes encoding yeast actin (9, 10), French bean phaseolin (20) and soybean leghemoglobin (Lb) (21).

Table 2. Codon usage* for the translated sequence of pSAc3 actin

Phe	TTT	5	Ser	TCT	7	Tyr	TAT	11	Cys	TGT	1
	TTC	6		TCC	5		TAC	2		TGC	3
Leu	TTA	0		TCA	5	?	TAA	1	?	TGA	0
	TTG	4		TCG	0	?	TAG	0	Trp	TGG	4
Leu	CTT	11	Pro	CCT	5	His	CAT	7	Arg	CGT	6
	CTC	9		CCC	6		CAC	3		CGC	1
	CTA	3		CCA	7	Gln	CAA	7		CGA	3
	CTG	0		CCG	0		CAG	2		CGG	0
Ile	ATT	16	Thr	ACT	9	Asn	AAT	4	Ser	AGT	7
	ATC	11		ACC	6		AAC	4		AGC	7
	ATA	1		ACA	3	Lys	AAA	7	Arg	AGA	6
Met	ATG	17		ACG	0		AAG	13		AGG	2
Val	GTT	13	Ala	GCT	11	Asp	GAT	17	Gly	GGT	14
	GTC	7		GCC	8		GAC	6		GGC	4
	GTA	3		GCA	9	Glu	GAA	12		GGA	8
	GTG	5		GCG	0		GAG	17		GGG	3

* Number of total occurrences of each codon in actin sequence. Three codons were undetermined.

appears to be related to both cytoplasmic and muscle specific actins. The amino-terminal sequence of pSAc1 actin is identical to that of pSAc3 and thus also appears to resemble both cytoplasmic and muscle specific actins (unpublished data). This finding suggests that the specific amino acid substitutions found within the amino-terminal peptides of tissue specific actins lack functional significance. This is further supported by the observation by Fyrberg *et al.* (12) that all six *Drosophila* actin genes encode proteins that resemble the cytoplasmic actins in their amino-terminal sequence.

The second region of sequence variation found in other actins is between residues 259 and 298 (7, 11). In pSAc3 actin, large variation is found between amino acid residues 228 and 277, showing 37% divergence when compared to the corresponding region of yeast actin. The remaining amino acid substitutions appear to be distributed randomly throughout the coding sequence.

pSAc3 Actin Gene Contains Three Small Introns. The polypeptide-encoding region of the pSAc3 actin gene is interrupted by three small intervening sequences (Fig. 1). Intron 1 interrupts the reading frame between codons 20 and 21 and is 88 base pairs long. Intron 2 splits the glycine codon at position 151 and consists of 81 base pairs. Intron 3 is located after codon 355 and is 79 base pairs long. All three introns start with 5′ G-T and end with 3′ A-G (Table 1). In a number of nucleotide positions the splice junctions are similar to the eukaryotic consensus sequences for exon/intron junctions (19) and to those exon/intron junctions found in genes encoding French bean phaseolin (20) and soybean leghemoglobin (21). Furthermore, all three introns are A+T-rich and encode stop codons in all three reading frames. This evidence strongly argues that these are *bona fide* introns.

The positions of introns within a coding sequence are highly conserved in several eukaryotic gene families. For example, intron positions are highly conserved in genes encoding globin (22), immunoglobulin (23, 24), and ovalbumin (25). It is particularly striking that the gene for a plant globin, soybean leghemoglobin, contains introns in the same locations as do the globin genes of distantly related animals (21). In contrast, the placement of introns in the actin gene families so far examined is highly diverged. The locations of the introns within the pSAc3 actin gene are unique. None of the other eukaryotic actin genes contains introns in these positions. For example, the yeast actin gene contains an intron in codon 4 (9, 10). Among six *Drosophila* actin genes, one gene contains an intron in the 5′ flanking sequence 8 nucleotides upstream from the AUG initiation codon, one gene contains an intron in codon 13, and two other genes each have an intron in codon 307 (12). Of the two sea urchin actin genes examined, one contains an intron between codons 121 and 122 and the other contains an intron in codon 204 (26). In contrast, none of the several *Dictyostelium* actin genes examined contains introns (27). It will be of interest to examine additional soybean actin genes for the placement of introns.

The partial sequence analysis of the soybean actin gene in pSAc1 indicates that the positions of introns 1, 2, and 3 in pSAc3 are also conserved in the pSAc1 actin gene. However, the sizes of introns and the nucleotide sequences within these introns in pSAc1 are different from those in pSAc3. Fyrberg *et al.* (12) discussed the evolutionary implications of the divergent placement of introns in various actin genes. Gilbert (28) has suggested that the rapid evolution of a gene is facilitated by the presence of introns within a polypeptide coding sequence. It follows from his argument that the older form of a gene would have more introns. The coding regions of individual actin genes in yeast, *Drosophila*, and sea urchin contain only one intron. Because each of the pSAc3 and pSAc1 actin genes is split by three introns, it is likely that soybean actin genes represent an older form of the actin gene. This would suggest that plant actin genes diverged from the animal actin genes much earlier in evolution than the animal actin genes diverged from each other.

The fact that the pSAc3 actin sequence is highly divergent from animal and fungal actin sequences lends further credence to this idea. Comparisons of the numbers and positions of introns within the actin genes of various higher plants will be most informative in this regard. In some genes, introns are thought to divide the encoding regions of separate functional domains of the polypeptide (24, 29, 30). In view of the divergent placement of introns in actin genes we think it unlikely that introns separate the functional domains of the actin polypeptide.

Codon Usage in pSAc3 Actin Gene. The yeast actin gene shows an extreme bias in codon usage for at least eight amino acids (10)—for example, all 28 glycine residues are encoded by the codon GGU. In contrast to yeast actin the codons appear to be used randomly in soybean. Table 2 shows the codon usage for this soybean actin gene.

Potential Regulatory Signals for Soybean Actin Gene Expression. The sequence 195 nucleotides upstream from the AUG start codon for this actin gene has been determined. Attempts to match the 5′ flanking sequence to the flanking region of the yeast actin gene show no apparent homology, with one surprising exception. A sequence 98 base pairs upstream from the soybean AUG start, beginning at nucleotide 97, shows striking homology with the potential T-A-T-A sequence from the yeast actin gene (10). This soybean sequence *T-G-T-A-A-A-T-*

G-T-C-C-*T*-G-G-*T*-T-*T*-*C*-*A*-*C*-*G* shares the italic residues with the yeast actin sequence. In yeast, the 5′ end of the actin transcript has been mapped 44 nucleotides downstream from this sequence (31). The soybean sequence T-G-T-A-A-A-T-G is a potential promoter recognition sequence similar to that seen in other eukaryotic genes. A potential cap sequence 27 nucleotides downstream from the start of this T-A-T-A sequence, *C*-*C*-*A*-*T*-A-*C*-*A*, resembles the consensus cap sequence (32) in those residues that are italic. More experimental work is required to find out if these sequences play a functional role in the transcriptional control of this gene.

We thank Federico Sanchez and Brian McCarthy for supplying their unpublished DNA sequences of *Drosophila* actin genes. Computer analysis of our DNA sequence data was carried out via the Stanford Molgen Project and the National Institutes of Health SUMEX-AIM Facility and with the generous assistance of Jan Clayton. This work was supported by the Science and Education Administration of the U.S. Department of Agriculture under Grant 5901-0410-8-0068-0 from the Competitive Research Grants Office and by the National Science Foundation under Grant PCM-7900833 from the Genetic Biology section. D.M.S. was supported by a postdoctoral fellowship from the Corporate Research and Development section of Monsanto Co.

1. Kersey, Y. M., Helper, P. K., Palevitz, B. A. & Wessells, N. K. (1976) *Proc. Natl. Acad. Sci. USA* **73**, 165–167.
2. Kersey, Y. M. & Wessells, N. K. (1976) *J. Cell Biol.* **68**, 264–275.
3. Vahey, M. & Scordilis, S. (1980) *Can. J. Bot.* **58**, 797–801.
4. Williamson, R. E. (1980) *Can. J. Bot.* **58**, 766–772.
5. Metcalf, T. N., Szabo, L. J., Schubert, K. R. & Wang, J. L. (1980) *Nature (London)* **285**, 171–172.
6. Vandekerckhove, J. & Weber, K. (1978) *Nature (London)* **276**, 720–721.
7. Vandekerckhove, J. & Weber, K. (1978) *J. Mol. Biol.* **126**, 783–802.
8. Gallwitz, D. & Seidel, R. (1980) *Nucleic Acids Res.* **8**, 1043–1059.
9. Gallwitz, D. & Sures, I. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 2546–2550.
10. Ng, R. & Abelson, J. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 3912–3916.
11. Vandekerckhove, J. & Weber, K. (1980) *Nature (London)* **284**, 475–477.
12. Fyrberg, E. A., Bond, B. J., Hershey, N. D., Mixter, K. S. & Davidson, N. (1981) *Cell* **24**, 107–116.
13. Nagao, R. T., Shah, D. M., Eckenrode, V. K. & Meagher, R. B. (1981) *DNA* **1**, 1–9.
14. Meagher, R. B., Shepherd, R. J. & Boyer, H. W. (1977) *Virology* **80**, 362–375.
15. Yang, R. C. A., Lis, J. & Wu, R. (1979) *Methods Enzymol.* **68**, 176–182.
16. Carreira, L. H., Carlton, B. C., Bobbio, S. M., Nagao, R. T. & Meagher, R. B. (1980) *Anal. Biochem.* **106**, 455–468.
17. Smith, H. O. & Birnstiel, M. L. (1976) *Nucleic Acids Res.* **3**, 2387–2398.
18. Maxam, A. M. & Gilbert, W. (1980) *Methods Enzymol.* **65**, 499–559.
19. Lerner, M. R., Boyle, J. A., Mount, S. M., Wolin, S. L. & Steitz, J. A. (1980) *Nature (London)* **283**, 220–224.
20. Sun, S. M., Slightom, J. L. & Hall, T. C. (1981) *Nature (London)* **289**, 37–41.
21. Jensen, E. O., Paludan, K., Hyldig-Nielsen, J. J., Jorgensen, P. & Marcker, K. A. (1981) *Nature (London)* **291**, 677–679.
22. Baralle, F. E., Shoulders, C. C. & Proudfoot, N. J. (1980) *Cell* **21**, 621–626.
23. Efstratiadis, A., Posakony, J. W., Maniatis, T., Lawn, R. M., O'Connell, C., Spritz, R. A., DeRiel, J. K., Forget, B. G., Weissman, S. M., Slightom, J. L., Blechl, A. E., Smithies, O., Baralle, F. E., Shoulders, C. C. & Proudfoot, N. J. (1980) *Cell* **21**, 653–668.
24. Sakano, H., Rogers, J. H., Huppi, K., Brack, C., Traunecker, A., Maki, R., Wall, R. & Tonegawa, S. (1979) *Nature (London)* **277**, 627–633.
25. Stein, J. P., Catterall, J. F., Kristo, P., Means, A. R. & O'Malley, B. W. (1980) *Cell* **21**, 681–687.
26. Durica, D. S., Schloss, J. A. & Crain, W. R. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 5683–5687.
27. Firtel, R. A. (1981) *Cell* **24**, 6–7.
28. Gilbert, W. (1978) *Nature (London)* **271**, 501.
29. Go, M. (1981) *Nature (London)* **291**, 90–92.
30. Tonegawa, S., Maxam, A. M., Tizard, R., Bernard, O. & Gilbert, W. (1978) *Proc. Natl. Acad. Sci. USA* **75**, 1485–1489.
31. Gallwitz, D., Perrin, F. & Seidel, R. (1981) *Nucleic Acids Res.* **9**, 6339–6350.
32. Hentschel, C., Irminger, J.-C., Bucher, P. & Birnstiel, M. L. (1980) *Nature (London)* **285**, 147–151.