Volume 13 Number 22 1985 Nucleic Acids Research

# Genomic cloning and characterization of a ricin gene from *Ricinus communis*

Kevin C.Halling, Amy C.Halling, Elizabeth E.Murray, Beth F.Ladin, L.L.Houston* and Robert F.Weaver

University of Kansas, Department of Biochemistry, Lawrence, KS 66044, USA

Received 26 August 1985; Accepted 15 October 1985

ABSTRACT

A genomic clone that specifies a single polypeptide precursor for ricin, a toxic lectin of Ricinus communis (castor bean), was isolated, sequenced and S1 mapped. The gene encodes a 64 kDa precursor which contains, in the following order: a 24 or 35 amino acid signal peptide, the A chain, a 12 amino acid linker peptide, and the B chain. The 5'-end of the ricin mRNA maps approximately 35 bases upstream from the first methionine codon. Two putative TATA boxes and a possible CAAT box lie in the 5'-flanking region. Two possible polyadenylation signals are found in the 3' flanking region. No introns were found, which is typical of other lectin genes that have been sequenced. Southern blot analysis suggests that the castor bean genome contains approximately six ricin-like genes.

## INTRODUCTION

Ricin and Ricinus communis agglutinin (RCA) are homologous lectins that can be isolated from the castor bean plant (Ricinus communis) (1-4). Ricin ($M_r \cong 65,000$), which contains an A and a B chain linked by a disulfide bond (5,6) is extremely toxic to eukaryotic cells while RCA ($M_r \cong 130,000$) is composed of two A and two B chains and is relatively non-toxic. Ricin kills eukaryotic cells by inhibiting protein synthesis (5). The B chain of ricin binds the toxin to galactose on the cell surface, an event that is necessary for the internalization of ricin. Once inside the cell the A chain enzymatically inactivates 60S ribosomal subunits of some eukaryotes (6). The enzymatic inactivation is so efficient that internalization of a single molecule of ricin is sufficient to kill one cell (7).

Although RCA is relatively non-toxic in vivo, purified RCA A chain inhibits cell-free protein synthesis, indicating some similarity between the A chains of each (4). The entire amino acid sequences for ricin A and B chains have been reported by Funatsu (8,9). The first seventeen amino acids for RCA A chain have been shown to be identical to ricin A chain sequences (4). Whether there are any differences between ricin and RCA A chains remains to be elucidated. Ready et al. (10) recently sequenced the 30 N-terminal amino

acids of the RCA B chain and found that the B chain amino acid sequences for ricin and RCA were identical for the first 23 residues but differed at residues 24-27, 29 and 30.

Butterworth and Lord found that when castor bean total poly(A)$^+$ RNA was translated in vitro, a 60 kDa translation product could be immunoprecipitated with an antibody to either ricin A chain or ricin B chain (11). More recently, Lord has used pulse-chase experiments to show that the 60 kDa polypeptide is synthesized on membrane bound ribosomes, translocated into the lumen of the endoplasmic reticulum and transported to protein bodies in the castor bean cell (12). The 60 kDa polypeptide is proteolytically cleaved to yield ricin A and B chains that have molecular weights of 30 kDa and 34 kDa respectively. Whether this cleavage takes place within the protein bodies or as the precursor moves across the protein body membrane is not clear. Lamb et al. (13) have sequenced two overlapping ricin cDNA clones. Their work confirms that ricin is translated as a large precursor called preproricin, which contains a signal peptide and the A and B chains of ricin separated by a 12 amino acid linker peptide.

In this paper we report the DNA sequence for a genomic ricin clone. Our work confirms that ricin is synthesized as a large precursor which is processed to yield the mature ricin A chain and B chain subunits. The ricin gene contains no introns and possesses consensus sequences found in the 5' and 3' flanking regions of other eukaryotic genes. In addition, we present evidence that ricin and RCA are members of a multigene family.

## MATERIALS AND METHODS

Materials. Restriction endonucleases and Klenow fragment were obtained from Bethesda Research Laboratories (Bethesda, MD), Pharmacia (Piscataway, NJ), or International Biotechnology Inc. (New Haven, CT). S1 nuclease was obtained from BRL. Oligo (dT) cellulose, and deoxynucleoside triphosphates were obtained from Pharmacia; dideoxynucleoside triphosphates were from Sigma; gene screen, $\alpha$-$^{32}$P-dATP, and colony/plaque screen were from New England Nuclear (Boston, MA); acrylamide, urea, TEMED, ammonium persulfate, and phenol were from International Biotechnologies Inc. (New Haven, CT), A lambda packaging kit (Packagene) was purchased from Promega Biotech (Madison, WS).

Plant Material-Castor beans (Ricinus communis) were obtained from ornamental plants in eastern Kansas.

Castor bean mRNA isolation and fractionation-Total RNA was prepared from castor bean seeds after removing seed pods and seed coats. Seeds were ground in liquid nitrogen, then homogenized with four volumes of extraction buffer

(150 mM NaCl; 50 mM Tris-HCl, pH 9.0; 5 mM EDTA) in a Sorvall Omni-mixer. Protein was removed by two extractions with phenol saturated with extraction buffer and two extractions with chloroform:isoamyl alcohol (24:1), followed by an ether extraction. The RNA was ethanol precipitated with one tenth volume of 22% potassium acetate and 2.2 volumes of ethanol. To remove carbohydrate, the RNA was resuspended in TE (10 mM Tris-HCl, pH 7.5; 1 mM EDTA) and precipitated by making the solution 2 M in LiCl. The precipitated RNA was resuspended in TE and ethanol precipitated a final time. Poly(A)$^+$ RNA was prepared using oligo(dT)-cellulose chromatography.

Isolation of castor bean DNA-Castor bean seeds were sprouted in vermiculite for 4-6 days in the absence of light. DNA was extracted by the method of Taylor and Powell (14). Seedlings were crushed in liquid nitrogen and then homogenized in a pre-cooled Sorvall Omni-mixer. A volume of 2-mercaptoethanol equal to 2% of the seed volume and one volume of boiling extraction buffer (2% cetyltrimethyl-ammonium bromide (CTAB); 100 mM Tris-HCl, pH 8.0; 20 mM EDTA; 1.4 M NaCl) were added, and the mix was then transferred to a 55°C water bath and stirred until the temperature reached 50°C. The homogenate was extracted with an equal volume of chloroform:isoamyl alcohol (24:1), the aqueous phase was removed and 0.1 volume of 10% CTAB was added, and the aqueous phase was extracted again with chloroform:isoamyl alcohol. The aqueous phase was removed and a two-fold reduction in salt concentration was obtained by adding one volume precipitation buffer (1% CTAB; 50 mM Tris-HCl, pH 8.0; 10 mM EDTA; 1% 2-mercaptoethanol), which causes the formation of a CTAB/nucleic acid precipitate. The precipitate was pelleted by centrifugation in a Beckman J13 rotor at 4,000xg for 5 min. The pellet was resuspended in a 1.0 M CsCl solution and layered on top of a 5.7 M CsCl cushion in Beckman SW 41 polayllomer centrifuge tubes, and spun for 17 hours at 40,000 rpm. The DNA was collected from the 1 M/5.7 M CsCl solution interface and ethanol precipitated.

Construction of genomic clone bank-Castor bean genomic DNA was partially digested with Eco R1 and size fractionated on a 10-40% sucrose gradient. DNA fragments having a size of 15-20 kilobases (kb) were ligated to lambda Charon 4 arms and packaged with a commercial packaging extract. E. coli, strain K802, was infected with the packaged DNA and plated on Luria-Bertani medium, supplemented with $Mg^{++}$ and maltose.

Screening of a castor bean genomic library-Lambda clones were plated on large petri dishes and transferred to Gene Screen and amplified overnight on the filter according to the method of Benton and Davis (15). The filters were

pre-washed with 50 mM Tris-HCl, pH 8.0; 1 M NaCl; 1 mM EDTA; 0.1% SDS at 42°C for 1 hour. The filters were pre-hybridized for 6 hours in 50% formamide, 5X Denhardt's solution (1% Ficoll, 1% polyvinylpyrrolidone, 1% BSA); 5X SSPE (0.75 N NaCl, 0.05 M $NaH_2PO_2 \cdot H_2O$, 0.005 M EDTA); 0.1% SDS; 100 µg/ml calf thymus DNA at 42° C. A ricin cDNA clone (Ladin et al., manuscript in preparation) was $^{32}P$-labeled by nick translation and added to the prehybridization mixture, and allowed to hybridize at 42°C for 6-12 hours. The filters were then washed three times with 2X SSC (0.3 M NaCl, 0.03 M Na Citrate), 0.1% SDS, at 25°C for 5 minutes and two times with 1X SSC at 68° for 30 minutes. The filters were blotted dry and autoradiographed overnight. Positive clones were picked and stored in SM media (100 mM NaCl; 8 mM $MgSO_4 \cdot 7H_2O$; 50 mM Tris, pH 7.5; 0.01% gelatin).

DNA Sequencing-Nucleotide sequencing was done according to methods outlined by Sanger (16). The Messing M13 vectors were used for cloning and sequencing (17).

S1 Mapping-The transcription start site for the ricin gene was determined using the Weaver and Weissman modification (18) of the S1 nuclease mapping procedure of Berk and Sharp (19). A 2200 base-pair Bam HI/Eco RI DNA fragment spanning the 5' end of the ricin gene (see Figure 1) was used for nuclease S1 mapping. The Eco RI/Bam HI fragment was labeled at the Bam HI end with $T_4$ polynucleotide kinase and $\gamma$-$^{32}P$-ATP. A probe with an activity of 12,000 cpm was hybridized with 5 micrograms of castor bean poly(A)$^+$ RNA overnight at 39°C. The DNA-RNA hybrids were then digested one hour with 60 units of S1 nuclease at 20°C and run on a 6% denaturing polyacrylamide gel (7.8 M urea, in buffer containing 45 mM Tris·HCl, 45 mM boric acid, 1 mM EDTA). Sequencing ladders from dideoxy sequencing of the probe were used as size markers.

Southern Blot Analysis-Total chromosomal DNA was extracted from Ricinus communis plantlets 5 days after germination. After digestion with restriction endonucleases, the DNA was separated by electrophoresis on a 0.8% agarose gel and transferred to nitrocellulose (20). Hybridization was with a $^{32}P$-labeled ricin DNA fragment spanning nucleotides 1946-2164 (see Figure 2).

## RESULTS AND DISCUSSION

### Cloning and sequence of a ricin gene.

Hybridization of a $^{32}P$-labeled ricin cDNA to a castor bean genomic library allowed us to identify a ricin genomic clone. This clone, lambda-Ric1, contained a 16 kb insert between the Eco RI sites of the two lambda arms. When this clone was cleaved with Eco RI, it yielded four subfragments from the insert, (5.7, 4.5, 3.0, and 2.2 kb). The 5.7 kb

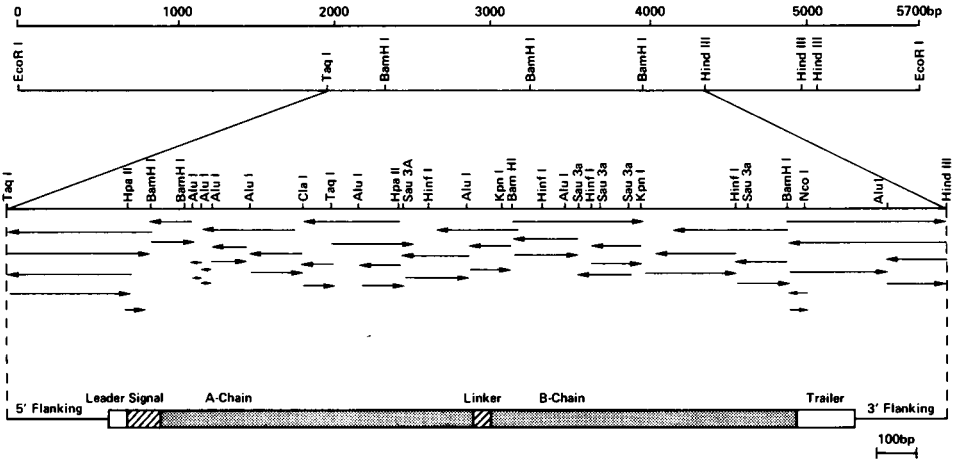


Fig. 1. Structural map of pAKG and sequencing strategy. Top: the 5.7 kb Eco R1 fragment containing the ricin gene. Middle: detailed restriction map of the 2.4 kb region bordered by a Taq I site on the left and a Hind III site on the right, which contains the entire ricin coding region plus approximately 250 bases of 5' and 230 bases of 3' flanking sequences. Not all restriction sites are included. Sequencing was performed as indicated by the arrows. Bottom: schematic representation of the ricin gene. The transcriptional unit is represented by boxed areas. The mature protein subunits are indicated by stippled boxes and the signal and linker peptides are represented by crosshatched boxes.

fragment was subcloned into pUC12. This subclone, pAKG, has the structure shown in Figure 1 and contains the entire ricin gene and approximately 2300 bases of 5' flanking sequence and 1500 bases of 3' flanking sequence. Hence, the ricin gene is encoded in a 1900 bp stretch.

We have sequenced approximately 2400 bases of pAKG (see Figure 1). As reported by Lamb et al. (13), the A and B chains are encoded on a single mRNA that contains the information for a precursor polypeptide of approximately 64 kDa. The precursor encoded by pAKG includes in the following order: a 24 or 35 amino acid signal peptide, the A chain, a linker peptide of 12 amino acids, and the B chain. The ambiguity about the length of the signal peptide comes from the fact that there are two possible initiation codons in phase with the rest of the protein. Lamb et al. reported a 24 amino acid signal peptide; however, their sequence did not extend far enough to detect the upstream initiation codon. Assuming a 35 amino acid signal peptide, the unglycosylated molecular weight of the precursor is 64,098. The unglycosylated A and B chains have molecular weights of 29,876 and 28,950, respectively.

DNA sequencing leads us to believe that pAKG is a genomic clone for

```
1
TCGACATTATATGATTTTAAATCAATTCCGTTTCTAATTTATAATTATTTCGTTAAACCAATCAA
66
TTCCCTTTAAACACTGCTTATGCATATTCTGTCTCAATTTATATATGGCATTGCATTCTTCCGTAT
132
TAATTTATAAGTTCACTTTTTATTGATCAAGTATTTGTGGTTTTCTTTATATAAAAAAATGTATTA
198                                                          VV
GTGTTTTTCTGTATTAATTTTATAAGTTCATCTTTATGAGAATGCTAATGTATTTGGACAGCCAAT
264
                                    M  K  P  G  G  N  T  I  V  I  W  M  Y
AAAATTCCAGAATTGCTGCAATCAAGGATGAAACCGGGAGGAAATACTATTGTAATATGGATGTAT
330                         |  Signal Peptide
A  V  A  T  W  L  C  F  G  S  T  S  G  W  S  F  T  L  E  D  N  N
GCAGTGGCAACATGGCTTTGTTTTGGATCCACCTCAGGGTGGTCTTTCACATTAGAGGATAACAAC
396                                                  Signal Peptide|
I  F  P  K  Q  Y  P  I  I  N  F  T  T  A  G  A  T  V  Q  S  Y  T
ATATTCCCCAAACAATACCCAATTATAAACTTTACCACAGCGGGTGCCACTGTGCAAAGCTACACA
|A-chain
N  F  I  R  A  V  R  G  R  L  T  T  G  A  D  V  R  H  E  I  P  V
AACTTTATCAGAGCTGTTCGCGGTCGTTTAACAACTGGAGCTGATGTGAGACATGAAATACCAGTG
519
L  P  N  R  V  G  L  P  I  N  Q  R  F  I  L  V  E  L  S  N  H  A
TTGCCAAACAGAGTTGGTTTGCCTATAAACCAACGGTTTATTTTAGTTGAACTCTCAAATCATGCA
594
E  L  S  V  T  L  A  L  D  V  T  N  A  Y  V  V  G  Y  R  A  G  N
GAGCTTTCTGTTACATTAGCGCTGGATGTCACCAATGCATATGTGGTCGGCTACCGTGCTGGAAAT
660
S  A  Y  F  F  H  P  D  N  Q  E  D  A  E  A  I  T  H  L  F  T  D
AGCGCATATTTCTTTCATCCTGACAATCAGGAAGATGCAGAAGCAATCACTCATCTTTTCACTGAT
726
V  Q  N  R  Y  T  F  A  F  G  G  N  Y  D  R  L  E  Q  L  A  G  N
GTTCAAAATCGATATACATTCGCCTTTGGTGGTAATTATGATAGACTTGAACAACTTGCTGGTAAT
792
L  R  E  N  I  E  L  G  N  G  P  L  E  E  A  I  S  A  L  Y  Y  Y
CTGAGAGAAAATATCGAGTTGGGAAATGGTCCACTAGAGGAGGCTATCTCAGCGCTTTATTATTAC
858
S  T  G  G  T  Q  L  P  T  L  A  R  S  F  I  I  C  I  Q  M  I  S
AGTACTGGTGGCACTCAGCTTCCAACTCTGGCTCGTTCCTTTATAATTTGCATCCAAATGATTTCA
924
E  A  A  R  F  Q  Y  I  E  G  E  M  R  T  R  I  R  Y  N  R  R  S
GAAGCAGCAAGATTCCAATATATTGAGGGAGAAATGCGCACGAGAATTAGGTACAACCGGAGATCT
990
A  P  D  P  S  V  I  T  L  E  N  S  W  G  R  L  S  T  A  I  Q  E
GCACCAGATCCTAGCGTAATTACACTTGAGAATAGTTGGGGGAGACTTTCAACTGCAATTCAAGAG
1056
S  N  Q  G  A  F  A  S  P  I  Q  L  Q  R  R  N  G  S  K  F  S  V
TCTAACCAAGGAGCCTTTGCTAGTCCAATTCAACTGCAAAGACGTAATGGTTCCAAATTCAGTGTG
1122
Y  D  V  S  I  L  I  P  I  I  A  L  M  V  Y  R  C  A  P  P  P  S
TACGATGTGAGTATATTAATCCCTATCATAGCTCTCATGGTGTATAGATGCGCACCTCCACCATCG
1188
S  Q  F  S  L  L  I  R  P  V  V  P  N  F  N  A  D  V  C  M  D  P
TCACAGTTTTCTTTGCTTATAAGGCCAGTGGTACCAAATTTTAATGCTGATGTTTGTATGGATCCT
 A-chain| Linker Peptide                          |B-chain
1254
E  P  I  V  R  I  V  G  R  N  G  L  C  V  D  V  R  D  G  R  F  H
GAGCCCATAGTGCGTATCGTAGGTCGAAATGGTCTATGTGTTGATGTTAGGGATGGAAGATTCCAC
1320
N  G  N  A  I  Q  L  W  P  C  K  S  N  T  D  A  N  Q  L  W  T  L
AACGGAAACGCAATACAGTTGTGGCCATGCAAGTCTAATACAGATGCAAATCAGCTCTGGACTTTG
1386
K  R  D  N  T  I  R  S  N  G  K  C  L  T  T  Y  G  Y  S  P  G  V
AAAAGAGACAATACTATTCGATCTAATGGAAAGTGTTTAACTACTTACGGGTACAGTCCGGGAGTC
1452
Y  V  M  I  Y  D  C  N  T  A  A  T  D  A  T  R  W  Q  I  W  D  N
TATGTGATGATCTATGATTGCAATACTGCTGCAACTGATGCCACCCGCTGGCAAATATGGGATAAT
1518
G  T  I  I  N  P  R  S  S  L  V  L  A  A  T  S  G  N  S  G  T  T
GGAACCATCATAAATCCCAGATCTAGTCTAGTTTTAGCAGCGACATCAGGGAACAGTGGTACCACA
```

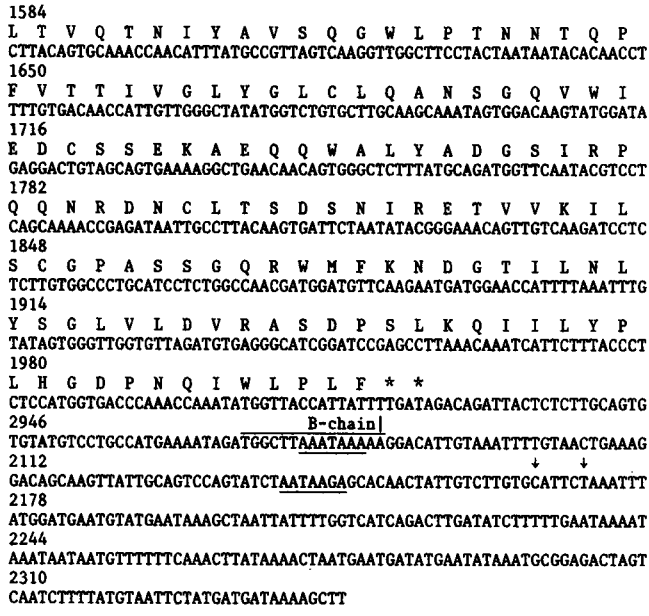


Fig. 2. Nucleotide sequence of the ricin gene and its flanking regions, and the amino acids deduced from this sequence. Nucleotides are numbered starting with the first base of the 5' Taq I site. VV indicates transcription start sites. The deduced amino acids are written above the first base of each codon, starting with the first ATG triplet in the transcribed region. Each translated region is bracketed below the nucleotide sequence, identifying the signal peptide, A-chain, linker peptide and B-chain. Each of the four CAAT boxes are underlined. The two repeats of the dual TATAA sequence are overlined. The two putative polyadenylation signals are underlined. Two possible polyadenylation sites, determined from two different cDNA clones, are each marked by an arrow.

ricin, not RCA. A comparison (Figure 3) of the B chain N-terminal amino acid sequence of pAKG to the reported amino acid sequences of RCA and ricin (10) reveals 6 amino acid differences between RCA and pAKG but only three differences between pAKG and ricin. Moreover, each of the differences between pAKG and ricin could be explained by protein sequencing artifacts. The first difference occurs at residue 22 [aspartic acid (pAKG)/asparagine (ricin)] and the second and third differences come from switching the order of a histidine and an asparagine residue at positions 29 and 30. The A-chain sequences for ricin, RCA, and pAKG are identical up to residue 17, which is as far as RCA A chain has been sequenced.

The ricin coding sequences are located within the middle one-third of the 5.7 kb Eco R1 fragment. Hybridization studies (results not shown) indicate that regions extending to 2000 bases upstream and 1500 bases downstream from

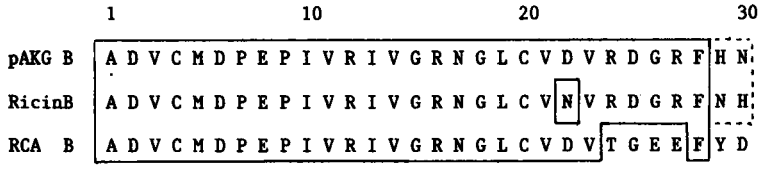


Fig. 3. Comparison of pAKG deduced amino acid sequence to the amino acid sequences of ricin and RCA B chains. The 30 N-terminal amino acids of pAKG, ricin and RCA B chains are shown. Boxes enclose identical residues. The dotted box indicates an apparent inversion of two amino acids.

the gene do not hybridize to ricin-specific probes. Therefore, there is no other member of the ricin-RCA gene family within 2 kb on the 5'-side or within 1.5 kb on the 3'-side of this ricin gene. In addition, no other ricin-like sequences are found on the original 16 kb insert in lambda-Ric1. However, because we do not know where the 5.7 kb fragment lies within the 16 kb fragment we cannot draw further inferences from these data.

Comparison of pAKG to two cDNA clones for ricin or RCA (Ladin et al., manuscript in preparation), and to the amino acid sequence of ricin (8,9), reveal that no introns are present in the genomic clone. This is also true of other lectin genes that have been sequenced. The soybean lectin gene, (21) and Phaseolus vulgaris lectin genes (22,23) do not have introns. Furthermore, certain other plant genes lack introns. These include: all of the genes for the maize storage protein (zein) that have been sequenced (24,25), the soybean 15 kDa protein (26) and the soybean Bowman-Birk protease inhibitor gene (27).

Locating the probable transcription start site.

S1 mapping reveals that transcription starts at either base number 256 or 257 in Figure 2, assuming no 5'-end post-transcriptional processing. The two strong bands seen in Figure 4 may indicate that there are two equal start sites. On the other hand, the lower band could be due to "S1 nibbling". Assuming a 35 amino acid signal peptide, the length of the 5'-untranslated region is 34 or 35 bases. This agrees with the reported length of some other plant leader sequences (28).

Interestingly, there are several ATG triplets following the transcription start site. Most eukaryotic translation start sites are at the 5'-most AUG codon on the message. This has led to the proposal of a scanning model (29) for translation initiation. According to this model, ribosomes bind the 5' end of the message, then migrate 5' to 3', scanning for the first AUG triplet, which is the initiation codon. However, not all messages initiate at the first AUG triplet on the message. Kozak (30,31) has proposed that the

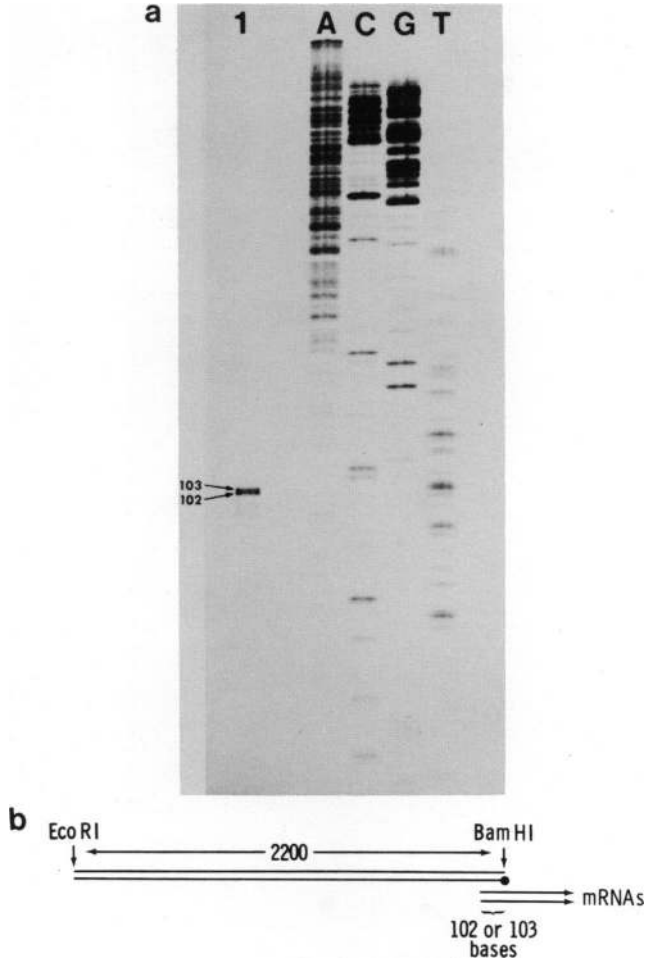


Fig. 4. S1 mapping analysis of pAKG. a. A 2200 bp Bam HI/Eco RI DNA fragment spanning the 5' end of the ricin gene (see Figure 1) was used for nuclease S1 mapping. The Eco RI/Bam HI fragment was labeled at the Bam HI end with T4 polynucleotide kinase and $\gamma$-$^{32}$P-ATP. This fragment was hybridized with castor bean seed poly(A)+ RNA and then digested with S1 nuclease. The digestion products were run on a 5% denaturing polyacrylamide gel as described in Materials and Methods. Lane 1, S1 digestion products; lanes 3-6, A, C, G and T lanes of a sequencing reaction used as a size marker. The sizes of the protected fragments are indicated. b. The 2200 bp probe is shown along with the two mRNAs that would protect 102 and 103 bases, respectively, of the probe.

sequence context of an AUG triplet is important for the AUG to initiate translation, and that some AUG triplets may be bypassed if they do not have the correct sequence. The consensus sequence for eukaryotic initiation is

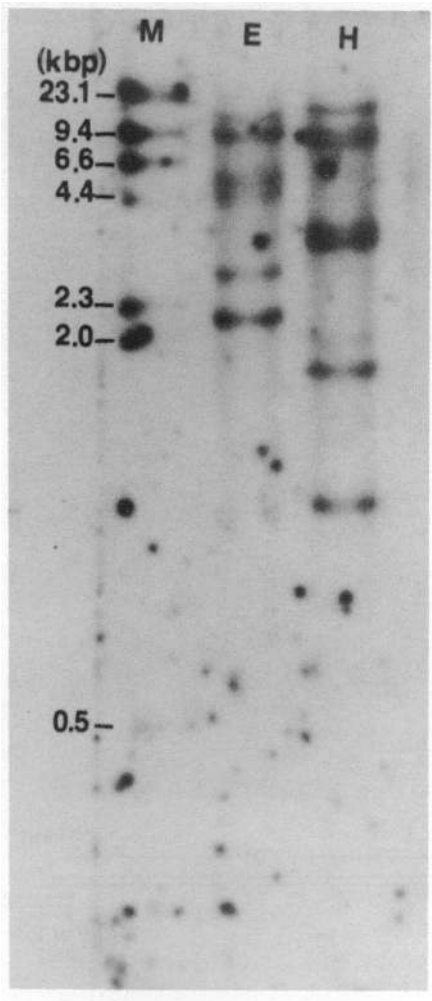


Fig. 5. DNA Southern blot analysis of the Ricinis communis genome. Ricinis communis genomic DNA was digested with either Eco R1 (lane E) or Hind III (lane H), electrophoretically separated on an agarose gel and transferred to nitrocellulose. A nick translated cDNA clone for ricin was used to probe the Southern blot. Labeled Hind III cut lambda fragments were used as markers (lane M).

$^{A}_{G}$NNAUGG. Messing (28) has extended this sequence analysis to plant genes and finds the consensus: $^{C}_{G}$AANNAUGG.

Of the multiple ATG codons 3' from the transcription start in the ricin gene, the first and third are in the correct reading frame. It seems likely that the signal peptide starts with the first methionine on the transcript, 35

amino acids upstream from the first amino acid on the mature protein. The sequence around this methionine codon matches Messing's plant consensus sequence in 6 out of 7 bases, and Kozak's consensus in 4 out of 5 bases. If this methionine codon were bypassed, the next potential start codon would be out of phase with the protein, and would have to be bypassed also. However, until the protein sequence of the precursor form of the protein is known, we cannot unambiguously identify the translation start.

5' Gene flanking region.

The 5' flanking region of the ricin gene is 75% A+T. This is commonly found in other plant genes (21,22,24,36). A TATAA sequence is found 33 bases upstream from the start of transcription, and matches the consensus sequence for the TATA box that has been shown to be important for eukaryotic transcription (32,33). In addition, there is a TATTAA sequence four bases upstream from the first TATAA sequence, which may be important to transcriptional control as well. The dual TATA sequence reads TATTAATTTTATAA. Interestingly, a near perfect 19 bp direct repeat of this double TATA box region (TATTAATTTATAA) is found 61 bp farther upstream. We do not know the significance of this, but it could represent dual promoters.

A second putative control element has been located 93 bases upstream from the start of transcription. The sequence CAAGT, may correspond to the CAAT box that is sometimes important for the efficiency of eukaryotic transcription (34). Usually, CAAT boxes are found 77±10 bases upstream from the start of transcription, although longer intervals have been found (35). Three other CAAT sequences have been located at 153, 190 and 194 bases upstream from the start of transcription. These are farther upstream than most CAAT boxes, but they match the consensus sequence better. Messing (28) has reported another possible plant consensus control element in zein genes and other plant gene promoters:

$$\overset{C}{T}\overset{A}{A}_{(2\text{-}5)}\overset{G}{T}\overset{}{N}\overset{}{G}\overset{}{A}_{(2\text{-}4)}\overset{CC}{TT}$$

We find nothing that resembles this consensus sequence in the area that we have sequenced.

3'-Untranslated region and 3'-flanking region.

The ricin coding region is followed by a 3'-untranslated region that is probably 130 to 160 bases long. The cDNA clones we have sequenced are polyadenylated at two different positions, giving rise to 3' untranslated regions of 148 and 153 bases. We assign the C and T, respectively, as the polyadenylation sites because this fits the C(A) or T(A) consensus sequence for such sites. Lamb et al. (13) report a 3' untranslated region of 159

bases. Since none of the cDNA sequences exactly matches the genomic sequence presented in this paper, we cannot unequivocally identify the polyadenylation site of the genomic sequence. However, one of the cDNA sequences differs in only one base from the genomic sequence, and its 3' untranslated region is 148 bases long. We believe that this is the most likely assignment for the length of the 3' untranslated region of pAKG. There are two sequences in the 3' untranslated region of the ricin gene that resemble the dual plant gene polyadenylation signals reported by Messing (28). The canonical polyadenylation signal AATAAA (37), is not always found 10-30 bases upstream from the polyadenylation site in plant genes, but instead is found 25-44 bases downstream from the stop codon. The second polyadenylation signal is found 16-35 bases upstream from the polyadenylation site and has a slightly different consensus sequence, $AATAA_{1-3}$. The two sequences we have found are 55 bases downstream from the second stop codon, and 20 or 25 bases upstream from the polyadenylation sites found in the cDNA clones.

The 3'-flanking region of the gene is also A+T-rich (75% A+T), a characteristic shared by other plant genes, including a Phaseolus vulgaris lectin gene (22), a phaseolin gene (36), a maize zein gene (24), and the soybean Le1 gene (21).

Signal and linker peptides.

The signal peptide is 35 or 24 amino acids long, as discussed above. A 35 amino acid signal peptide is longer than most that have been reported (38). In spite of this, it seems likely that the first methionine at amino acid residue -35, is the start of translation for reasons outlined above.

As in other signal peptides, a lysine residue is found near the N-terminus. The ricin signal peptide is not as hydrophobic as some signals, although at least one other plant signal peptide is relatively hydrophilic (27). In addition, two asparagine residues are found at the clipping site, whereas most signal peptides are clipped after an alanine, serine or glycine (38).

The ricin precursor contains a 12 amino acid linker peptide. Other plant proteins, such as glycinin (39) and pea seed lectin (40) also contain linker peptides. Unlike the other linker peptides, which are hydrophilic, the ricin linker peptide is relatively hydrophobic. The hydrophilicity has been proposed to be important for exposing this portion of the protein to protease attack (30). Instead, the ricin linker contains two proline residues within the linker and an additional five proline residues within ten amino acids on either side of the linker, which could be expected to prevent secondary

structure in this region, and therefore aid in the removal of the linker peptide. As in both of the other linker peptides there is an asparagine residue at the C-terminal end of the ricin linker. The structure of the linking peptide must allow the formation of a disulfide bond between a cysteine nine residues from the C-terminus of the A chain and a cysteine four residues from the N-terminus of the B chain.

Ricin is a member of a multigene family.

Our predicted A and B chain amino acid sequences differ in 6 and 28 places, respectively, from the amino acid sequences for A and B chains published by Funatsu (8,9). Furthermore, our A and B chains contain 267 and 262 amino acids, respectively, while the Funatsu A and B chains contain 265 and 260 amino acids, respectively. The region that differs most from the Funatsu sequence is the C-terminal end of the B chain. Some of this variation we attribute to the fact that ricin-like proteins are encoded by a family of genes. Cawley et al. (4) found that they could isolate 3 forms of ricin ($ricin_1$, $ricin_2$, and $ricin_3$) and two forms of RCA ($RCA_1$ and $RCA_2$) hinting at the existence of a gene family. Furthermore, we have sequenced two distinct cDNA clones whose differences occur predominantly near the C-termini of the B chains. We provide further evidence below that ricin and RCA are products of a multigene family which encodes several ricin-like proteins.

We find 10 nucleotide differences between the cDNA sequence reported by Lamb et al. (13) and pAKG, two of which change an amino acid, seven of which are in the third bases of codons, and one of which is in the 3' untranslated region. Again, it is likely that many of these differences arise from real variations between different genes in the family. In light of these differences, the possibility exists that the sequence presented by Lamb et al., for two overlapping cDNAs, is a hybrid of two variant genes for ricin.

Cawley et al. (4) have suggested that ricin and RCA might be derived from the alternative processing of a precursor at its C-terminal end to give both RCA and ricin. This seems unlikely in light of the recent finding by Ready et al., of differences between the N-terminal ricin and RCA B chain sequences (10). It appears that separate genes encode ricin and RCA. The sequence we obtain for this region coincides with that for ricin, not RCA.

To provide additional evidence that pAKG is a member of a multigene family, we characterized the ricin-like genes in total chromosomal DNA. Southern blot analysis of Ricinus communis DNA digested with either Hind III or Eco RI (Figure 5), shows multiple bands of hybridization as follows. Digestion with Hind III produces eight major bands at: 17.8, 10.6, 9.1, 3.8,

3.4, 2.0, 1.7, 1.0 kb. Digestion with Eco RI produces six major bands at 14.1, 10.0, 5.7, 5.0, 2.8, 2.2 kb. The 5.7 kb Eco RI band probably corresponds to pAKG. Hind III and Eco RI were chosen because the sequence of pAKG revealed no sites for these enzymes. In the Eco RI digest all the restriction fragments are large enough to encode a full length gene. Because the cDNA probe was only 218 bases in length, there is a relatively low probability of its spanning an Eco RI site in any of the genes to which it hybridizes, as long as the number of genes is small. The fact that all the bands are of more-or-less equal intensity strengthens this conclusion. We therefore estimate that there are six ricin-like genes in the castor bean genome. The extra bands in the Hind III digest, of less than full gene size, probably result from Hind III sites within one or more of the ricin-like genes.

ACKNOWLEDGMENTS

This work was supported by a grant from Eli Lilly & Company, Indianapolis, Indiana.

*Present address: Cetus Corporation, Emeryville, CA, USA

REFERENCES

1. Olsnes, S., Refsnes, K., Christiansen, T.B. and Pihl, A. (1975) Biochem. Biophys. Acta 405, 1-10
2. Funatsu, G. and Funatsu, M. (1977) Agric. Biol. Chem. 41, 1211-1215
3. Nicolson, G.L., Blaustein, J. and Etzler, M.E. (1974) Biochem. 13, 196-204.
4. Cawley, D.B., Hedblom, M.L., and Houston, L.L. (1978) Arch. Biochem. Biophys. 190, 744-755
5. Olsnes, S., Refsnes, K., and Pihl, A. (1974) Nature 249, 627-631
6. Cawley, D.B., Hedblom, M.L., Hoffman, E.J., and Houston, L.L. (1977) Arch. Biochem. Biophys. 182, 690-695
7. Eiklid, K., Olsnes, S., and Pihl, A. (1980) Exp. Cell Res. 126, 321-326
8. Funatsu, G., Yoshitake, S., and Funatsu, M. (1978) Agric. Biol. Chem. 42, 501-503
9. Funatsu, G., Kimura, M., and Funatsu, M. (1979) Agric. Biol. Chem. 43, 2221-2224
10. Ready, M., Wilson, K., Piatak, M., and Robertus, J.D. (1984) J. Biol. Chem. 259, 15252-15256
11. Butterworth, A.G., and Lord, J.M. (1983) Eur. J. Biochem. 137, 57-65
12. Lord, J.M. (1985) Eur. J. Biochem. 146, 403-409
13. Lamb, F.I., Roberts, L.M., and Lord, J.M. (1985) Eur. J. Biochem. 148, 265-270.
14. Taylor, B., and Powell A. (1983) Focus 4:3, 4-6
15. Benton, W.D., and Davis, R.W. (1977) Science 196, 180-182
16. Sanger, F., Nicklen, S., and Coulson, A.R. (1977) Proc. Natl. Acad. Sci. 74, 5463-5467

17. Messing, J., and Vieira, J. (1982) Gene 19, 269-276
18. Weaver, R.F. and Weissmann, C. (1979) Nuc. Acids Res. 7, 1175-1192
19. Berk, A.J., and Sharp, P.A. (1977) Cell 12, 721-732
20. Southern, E. (1980) Methods Enzymol. 69, 152-175
21. Vodkin, L.O., Rhodes, P.R., and Goldberg, R.B. (1983) Cell 34, 1023-1031
22. Hoffman, L.M. (1984) J. Mol. Appl. Genet. 2, 447-453
23. Hoffman, L.M. (1985) EMBO J., in press
24. Pederson, K., Devereux, J., Wilson, D.R., Sheldon, E. and Larkins, B.A. (1982) Cell 29, 1015-1026
25. Hu, N.T., Peifer, M.A., Heidecker, G., Messing, J. and Rubenstein, I. (1982) EMBO J. 1, 1337-1342
26. Fischer, R.L., and Goldberg, R.B. (1982) Cell 29, 651-660
27. Hammond, R.W., Foard, D.E. and Larkins, B.A. (1984) J. Biol. Chem. 259, 9883-9890
28. Messing, J., Geraghty, D., Heidecker, G., Hu, N.T., Kridl, J. and Rubenstein, I. (1983) in Genetic Engineering of Plants, Kosuge, T., Meredith, C.P., and Hollaender, A., Ed., Vol. 26, pp. 211-227, New York, N.Y.
29. Kozak, M. (1980) Cell 22, 7-8
30. Kozak, M. (1981) in Current Topics in Microbiology and Immunology, Shatkin, A.J., Ed., Vol. 93, pp. 81-123, Springer-Verlag, Berlin
31. Kozak, M. (1981) Nuc. Acids Res. 9, 5233-5252
32. Benoist, C., and Chambon, P. (1981) Nature 290, 304-310
33. Grosschedl, R., and Birnstiel, M.L. (1980) Proc. Natl. Acad. Sci. 77, 1432-1436
34. Benoist, C., O'Hare, K., Breathnach, R., and Chambon, P. (1980) Nuc. Acids Res. 8, 127-142
35. 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, O.O., and Proudfoot, N.J. (1980) Cell 21, 653-668
36. Slightom, J.L., Sun, S.M., and Hall, T.O. (1983) Proc. Natl. Acad. Sci. USA 80, 1897-1901
37. Proudfoot, N.J. and Brownlee, G.G. (1976) Nature 263, 211-214
38. Inouye, M., and Halegoua, S. (1980) Crit. Rev. Biochem. 7, 339-371
39. Marco, Y.A., Thanh, V.H., Tumer, N.E., Scallon, B.J., and Nielsen, N.O. (1984) Jour. Biol. Chem. 259, 13436-13441
40. Higgins, T.J.V., Chandler, P.N., Zurawski, G., Button, S.O. and Spencer, D. (1983) Jour. Biol. Chem. 258, 9544-9549