Volume 10 Number 11 1982 Nucleic Acids Research

The nucleotide sequences of two leghemoglobin genes from soybean

Ove Wiborg, Jens Jørgen Hyldig-Nielsen, Erik Ø.Jensen, Kirsten Paludan and Kjeld A.Marcker

Department of Molecular Biology and Plant Physiology, University of Aarhus, DK-8000 Aarhus C, Denmark

Received 20 April 1982; Accepted 7 May 1982

ABSTRACT

We present the complete nucleotide sequences of two leghemoglobin genes isolated from soybean DNA. Both genes contain three intervening sequences in identical positions. Comparison of the coding sequences with known amino-acid sequences of soybean leghemoglobins suggest that the two genes correspond to leghemoglobin c_2 and leghemoglobin c_3, respectively.

INTRODUCTION

Leghemoglobins (Lbs) are synthesized exclusively in the nitrogen-fixing root nodules that develop owing to the symbiotic association of Rhizobia with legumes. Soybean nodules contain four major species of Lbs called Lba, Lbc_1, Lbc_2, and Lbc_3, respectively[1]. In addition several minor Lb components are present in the nodule, but some of these components most likely represent post-translational modification products of some of the major components[2]. The difference in the amino-acid sequences among the various Lb components are small corresponding to 6-8 amino-acid substitutions[3].

The Lbs are encoded in the plant genome by a small family of genes [4,5]. We have so far isolated six separate Lb genes from soybean. Recently we have determined the complete nucleotide sequences of two Lb genes, which most likely corresponded to Lba and Lbc_1, respectively[6]. In this paper we report the nucleotide sequences of two other soybean Lb genes which presumably correspond to Lbc_2 and Lbc_3, respectively. The general DNA sequence organization of the Lbc_2 and Lbc_3 genes are very similar to that of the Lba and the Lbc_1 genes. The coding regions contain three intervening sequences which interrupt at codons 32(IVS-1), 68-69

0305-1048/82/1011-3487$ 2.00/0

(IVS-2) and 103-104(IVS-3), respectively. The 5' and 3' flanking sequences contain conserved sequences similar to those found in other eukaryotic genes including the Lba and Lbc_1 genes. In addition the consensus sequences for splicing are present at the three intron/exon boundaries as also found for the corresponding sequences in the Lba and Lbc_1 genes.

MATERIALS AND METHODS

Restriction endonucleases were purchased from either Biolabs, New England, or Boehringer, Mannheim. DNA polymerase (Klenow fragment) was from Boehringer, Mannheim. Polynucleotide kinase and dideoxynucleoside triphosphates were from PL Biochemicals, T4 DNA ligase from Bethesda Research Laboratories, and the dodecadeoxynucleotide primer from Collaborative Research. α-^{32}P-dATP was from New England Nuclear.

Isolation of genomic Lb-genes. The genomic recombinant molecules containing Lb-sequences were isolated from a library which was constructed from a complete EcoRI digest of soybean DNA, using λgtWes/λB as a vector. About 8×10^5 recombinant λgtWes/λB phages were screened with a ^{32}P-labelled Lb cDNA clone according to the method described by Maniatis et al.[7]. The procedures used to construct subclones and to prepare plasmid DNA were according to Lacy et al.[8].

M13 Cloning. Appropriate DNA fragments were subcloned in the filamentous bacteriophages M13mp7[9] and propagated in the host JM101 in 2 NZY (10 g NaCℓ, 4 g $MgC\ell_2$, 7 H_2O, 20 g NZamide-typeA, 10 g yeast extract in 1 ℓ H_2O). DNA was purified from the supernatant by precipitating with 2.5% polyethylene glycol, 0.5 M NaCℓ. The single-stranded DNA was finally purified by extraction with phenol followed by ethanol precipitation.

DNA Sequencing. DNA Sequencing was performed by the dideoxy chain termination method described by Sanger et al.[10] using a synthetic dodeca deoxy nucleotide as primer. Sequencing rection products were electrophoresed on 6 or 8% 0.3 mm polyacrylamide-urea gels.

RESULTS AND DISCUSSION

Sequencing strategy and procedures. Southern blotting ana-

lysis of EcoRI digests of soybean DNA revealed the presence of seven hybridizing fragments of lengths 1.4, 4.2, 5.5, 6.0, 7.5, 12 and 13 kb, respectively[5]. We have isolated six clones carrying chromosomal Lb genes from a soybean DNA library which was constructed from a complete EcoRI digest of soybean DNA. The sizes of the cloned DNA fragments were 1.4, 4.2, 6.0, 7.5, 12 and 13 kb, respectively. Figure 1 records the restriction nuclease cleavage maps of the cloned 13, 12 and 6.0 kb fragments. The 13 and 12 kb fragments are very similar, the major difference being that in the 12 kb fragment a region of about 1.5 kb between a BglII and a SacI site has been deleted when compared to the 13 kb fragment. Otherwise almost all restriction endonucleases tried cleave in identical positions in the two fragments. The observed deletion in the 12 kb fragment is not due to a cloning error since both genes have been detected in soybean DNA by Southern blotting analysis. Both fragments contain a complete Lb gene.

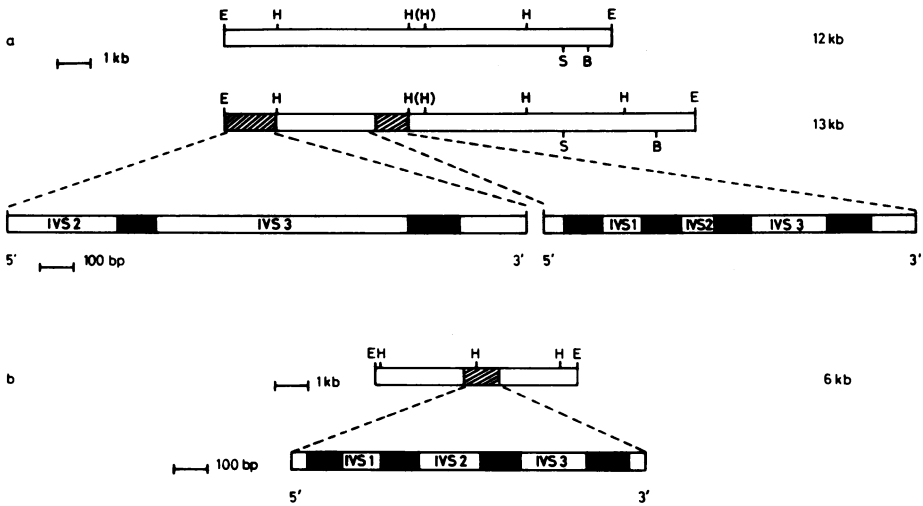

Figure 1. a) Restriction cleavage maps of the cloned 13 and 12 kb fragments and b) restriction cleavage map of the 6.0 kb fragment Regions containing Lb gene sequences (crosshatched) are enlarged in the lower part of each figure showing coding sequences (solid boxes) and non-coding sequences (open boxes). IVS-1, IVS-2, and IVS-3 denote intervening sequences 1, 2, and 3. H: HindIII, E: EcoRI, B: BglII, and S: SacI. (H): This HindIII site generates a 0.5 kb fragment, which may be located as indicated or 3 kb further downstream.

About 2.6 kb upstream from the complete gene, DNA sequence analysis has revealed the presence of a Lb gene consisting of a 3' noncoding end, exons 4 and 3, the entire IVS-3 (778 bp) and part of IVS-2. The complete gene and the incomplete gene are oriented in the same direction. We have recently shown that the 7.5 kb fragment containing the Lbc_1 gene is directly linked to the 13 kb fragment such that the Lbc_1 gene is 5' to the incomplete gene on the 13 kb fragment. DNA sequence analysis has shown that the missing 5' parts of the incomplete gene on the 13 kb fragment is present on the 7.5 kb fragment. This particular gene is about 3 kb long and DNA sequence data have revealed many irregularities in its sequence which suggest that it is nonfunctional and most probably is a pseudo gene. The complete nucleotide sequence of this gene will be the subject of another publication.

The DNA sequences of the complete gene present in the 13 kb fragment and of the gene present in the 6.0 kb fragment were determined by the chain termination method after cloning appropriate fragments into the single-stranded phage M13mp7 [10]. Figures 2a and b outline the strategy used for the determination of the DNA sequences of both genes.

Nucleotide sequence of the two Lb genes. The entire nucleotide sequence of the complete gene contained in the 13 kb fragment is shown in Figure 3a while the sequence of the gene contained in the 6.0 kb fragment is shown in Figure 3b. Comparison of these genes with known amino-acid sequences indicates that the gene represented in Figure 3a most likely corresponds to Lbc_3 while the gene represented in Figure 3b most likely corresponds to Lbc_2. However this assignment is not conclusive, since amino-acid sequence analysis has not yet been completed on homogenous Lbc varieties (Whittaker, R.G., personal communication).

Flanking and non-coding regions. The sequences determined include a 175 bp (Lbc_2) and a 103 bp (Lbc_3) region 5' to the ATG initiator codon. Both genes contain potential cap addition sites, these being located at positions 120 or 129 (Lbc_2) and at positions 53 or 62 (Lbc_3). In addition both genes contain ATA boxes further upstream from the potential cap addition sites. In the Lbc_2 gene the ATA box is located at position 91 while for the Lbc_3 gene it is located at position 24. Identical sequences were

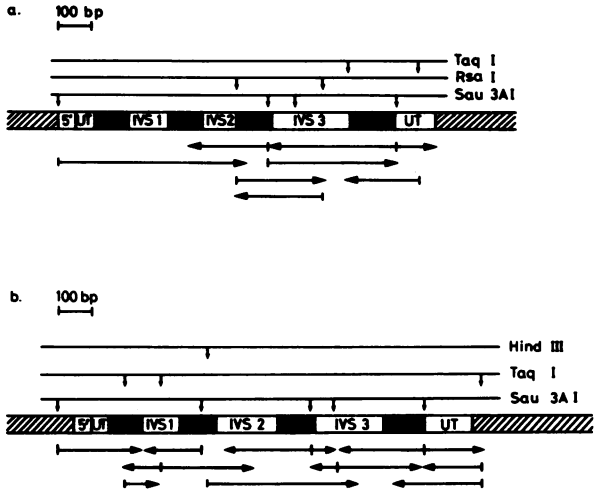

Figure 2. Strategy for determining the nucleotide sequence of the Lbc_3 gene (a) and the Lbc_2 gene (b)
A detailed restriction nuclease map for those restriction nuclease site (vertical arrows) used in deriving the sequence. The extent and direction of each sequence reading are indicated by horizontal arrows. UT represents sequences corresponding to the non-translated 5' and 3' regions of the Lb mRNA. IVS-1, IVS-2, and IVS-3 denote intervening sequences.

also found in similar positions in the Lba and Lbc_1 sequences [6]. The Lbc_2 gene includes a longer 5' flanking region than that determined for the Lbc_3 gene. It is noteworthy that upstream from the ATA box the sequence -CCAAG- occurs at positions 44-48. In most eukaryotic genes including Lba a homologous sequence occurs at or close to this position [11].

The 3' non-coding regions of both genes contain the sequence GATAAA (Lbc_2: 1232-1237, Lbc_3: 1100-1105). We have previously proposed that this sequence may serve as a polyA addition signal in the particular plant gene system [6].

Intervening sequences. The two Lb genes are both interrupted at three places by relatively small intervening sequences. The positions of these correspond precisely to the ones determined for the Lba and Lbc_1 genes [6]. All intron/exon boundaries are in agreement with the consensus sequence [12]. The three introns start with the sequences -GTAAG- or -GTATGA- and terminate with -AG as is typical of other eukaryotic genes. However, we have previous-

```
GAT CAC TCT TCA AGC CTT CTA TAT AAA TAA GTA TTG GAT GTG AAG TTG TTG CAT AAC TTG
                                    30                                      60

                                                                GLY ALA PHE THR ASP
CAT TGA ACA ATT AAT AGA AAT AAC AGA AAA GTA GAA AAG AAA TAT G/GGT GCT TTC ACT GAT
                                    90                                     120

LYS GLN GLU ALA LEU VAL SER SER SER PHE GLU ALA PHE LYS THR ASN ILE PRO GLN TYR
AAG CAA GAG GCT TTG GTG AGT AGC TCA TTT GAA GCA TTC AAG ACA AAC ATT CCT CAA TAC
                                   150                                     180

SER VAL VAL PHE TYR THR SER
AGT GTT GTG TTC TAC ACC TC/GTA AGT ATT CTA TCT AAA TTA TGT GTC TTA TTG TAT GTT
                                   210                                     240

TAA CTT TCG TGG TTT GTT GTG TTT GAA AAA AAG ATA TAT ATT GTT AAT GTG AGT GGT TTT
                                   270                                     300

                                ILE LEU GLU LYS ALA PRO VAL ALA LYS ASP LEU PHE SER
GGT TTG ACT AAA AAT GAA TAG/G ATA CTG GAG AAA GCA CCT GTA GCA AAG GAC TTG TTC TCA
                                   330                                     360

PHE LEU ALA ASN GLY VAL ASP PRO THR ASN PRO LYS LEU THR GLY HIS ALA GLU LYS LEU
TTT CTA GCT AAT GGA GTA GAC CCC ACT AAT CCT AAG CTC ACG GGC CAT GCT GAA AAA CTT
                                   390                                     420

PHE GLY LEU
TTT GGA TTG/GT AAG TAT CCA GCC TAC TAA AAT TAA AAT CCT ATT AGT ATT TTT TAT TAT
                                   450                                     480

                                                                VAL ARG ASP SER
TTT TCT TCC ATG ATT GTC TTG TCA CAT ATT ATA TAT TTT TTG AAT TAT AG/GTA CGT GAT TCA
                                   510                                     540

ALA GLY GLN LEU LYS ALA SER GLY THR VAL VAL ILE ASP ALA ALA LEU GLY SER ILE HIS
GCT GGT CAA CTT AAA GCA AGT GGA ACA GTG GTG ATT GAT GCC GCA CTT GGT TCT ATC CAT
                                   570                                     600

ALA GLN LYS ALA ILE THR ASP PRO GLN PHE VAL
GCC CAA AAA GCA ATC ACT GAT CCT CAA TTT GTG/G TAT GAT AAA TAA TGA AAA GCT ACA
                                   630                                     660

ATA AAT GCA CAA ATA CTT AAT TTT ACA TAG TGC AGT GCT ATA TGA TCA TCA CTT TTG CTT
                                   690                                     720

AGT AAT GAA TTT ACT TTT TTT TTT TAC AGA AGT AAT GGA TTT ACT TAA AAT CTT AAA TTA
                                   750                                     780

TGT ACT TCT TTA AAG AGT TTT GTA TGG AAT TTT AAT TAT AGG AAA AAT GTA AGA GCT AAA
                                   810                                     840

                             VAL VAL LYS GLU ALA LEU LEU LYS THR ILE LYS GLU ALA
CCA TTG CTG ATG ATT TCG AAG/GTG GTT AAA GAA GCA CTG CTG AAA ACA ATA AAG GAG GCA
                                   870                                     900

VAL GLY ASP LYS TRP SER ASP GLU LEU SER SER ALA TRP GLU VAL ALA TYR ASP GLU LEU
GTT GGG GAC AAA TGG AGT GAC GAG TTG AGC AGT GCT TGG GAA GTA GCC TAT GAT GAA TTG
                                   930                                     960

ALA ALA ALA ILE LYS LYS ALA PHE ***
GCA GCA GCT ATT AAG AAG GCA TTT TAG/GAT CTA CAA TTG CCT TAA AGT GTA ATA AAT AAA
                                   990                                    1020

TAT TAT TTC ACT AAA ACT TGT TAT TAA ACC AAG TTC TCG ATA TAA ATG TTG GTT AAA CTA
                                  1050                                    1080

AGT AAA TTA TAT GGT ATT GGA TAA ACA ATC TTA AGC TT
                                  1110
```

Figure 3a. The nucleotide sequence of a soybean Lbc_3 gene

ly noted a considerably size variation for the length of some of the corresponding intervening sequences. Thus the length of IVS-2 varies from 100 nucleotides in Lbc_3 to 190 nucleotides for the corresponding sequence in Lbc_2.

In conclusion the DNA sequence organization of the two Lb genes presented here is very similar to that of two other Lb

```
GAT CAT TTG GCT CTT CAT CAT GCC GAT TGA CAC CCT CCA CAA GCC AAG AGA AAC TTA AGT
                                    30                                      60

TGT AAT TTT TCT AAC TCC AAG CCT TCT ATA TAA ACA CGT ATT GGA TGT GAA GTT GTT GCA
                                    90                                     120

                                                                       GLY
TAA CTT GCA TTG AAC AAT AGA AAT AAC AAC AAA GAA AAT AAG TGA AAA AAG AAA TAT G/GGT
                                   150                                     180

ALA PHE THR GLU LYS GLN GLU ALA LEU VAL SER SER SER PHE GLU ALA PHE LYS ALA ASN
GCT TTC ACT GAG AAG CAA GAG GCT TTG GTG AGT AGC TCA TTC GAA GCA TTC AAG GCA AAC
                                   210                                     240

ILE PRO GLN TYR SER VAL VAL PHE TYR THR SER
ATT CCT CAA TAC AGC GTT GTG TTC TAC ACT TC/GTA AGT TTT CTC TTA AAG CAT GTA TCT
                                   270                                     300

TTC ATT CTC TGT TTT TCC TTT CGA CAT TTT TTG TGT TTG AAA AGA GAT AGT GTC AAT GTG
                                   330                                     360

                                                    ILE LEU GLU LYS ALA PRO ALA ALA LYS
AGT GGG TAT TTT TTT TTA TTA AAA ATT AAC AG/G ATA CTG GAG AAA GCA CCC GCA GCA AAG
                                   390                                     420

ASP LEU PHE SER PHE LEU SER ASN GLY VAL ASP PRO SER ASN PRO LYS LEU THR GLY HIS
GAC TTG TTC TCG TTT CTA TCT AAT GGA GTA GAT CCT AGT AAT CCT AAG CTC ACG GGC CAT
                                   450                                     480

ALA GLU LYS LEU PHE GLY LEU
GCT GAA AAG CTT TTT GGA TTG/GTA AGT ATC ATC CAA CTA AAA TTA TAG CTA TTT TAT GTG
                                   510                                     540

ATT AAT TTT AAG ATT AAA CAT GTA TTT AAC ACT CTT AAA CAT GTA TTT AAC ACT CTT AAG
                                   570                                     600

ATT AAA CAT GTA TTT AAC TAA AAC ATG TAT TTG CTG ATT ATT TTT TTT TTA TAA TTA TCT
                                   630                                     660

                                                  VAL ARG ASP SER ALA GLY GLN LEU LYS ALA
TGT CAC ATA TTA TAT ATT TTT TGA ATT GTA G/GTG CGT GAC TCA GCT GGT CAA CTT AAA GCA
                                   690                                     720

ASN GLY THR VAL VAL ALA ASP ALA ALA LEU GLY SER ILE HIS ALA GLN LYS ALA ILE THR
AAT GGA ACA GTA GTG GCT GAT GCC GCA CTT GGT TCT ATC CAT GCC CAA AAA GCA ATC ACT
                                   750                                     780

ASP PRO GLN PHE VAL
GAT CCT CAG TTC GTG/GT ATG ATA AAT AAT AAA ATG TTA CAA TAA ATG CAC ATA TAC TTA
                                   810                                     840

AAT TTT ACA TGG TGC AGT GTT ATG ATC ATC ATT TTT GTT TAG TAA TGA ATT TAC TTA AAA
                                   870                                     900

TCT TAA ATT ATG TAC TTT TTG AAA GTT TTA TAT GGA ATT TTA ATT ATA GGG AAA AAT GTA
                                   930                                     960

                                                  VAL VAL LYS GLU ALA LEU LEU LYS THR
AGA GCT AAT CCA TTA GTG ATG TTT TGT CTG TAG/GTG GTT AAA GAA GCA CTG CTG AAA ACA
                                   990                                    1020

ILE LYS GLU ALA VAL GLY ASP LYS TRP SER ASP GLU LEU SER SER ALA TRP GLU VAL ALA
ATA AAG GAG GCA GTT GGG GAC AAA TGG AGT GAT GAA TTG AGC AGT GCT TGG GAA GTA GCC
                                  1050                                    1080

TYR ASP GLU LEU ALA ALA ALA ILE LYS LYS ALA PHE ***
TAT GAT GAA TTG GCA GCA GCT ATT AAG AAG GCA TTT TAG/GAT CTA CTA TTG CCG TCA AGT
                                  1110                                    1140

GTA ATA AAT AAA TTT TGT TTC ACT AAA ACT TGT TAT TAA ACA AGT CCC CGA TAT ATA AAT
                                  1170                                    1200

GTT GGT TAA AAT AAG TAA ATT ATA CGG TAT TGA TAA ACA ATC TTA AGT TTT ATA TAT AGT
                                  1230                                    1260

TCC ATA TAC TAA AGT TTG TGA ATC ATA ATC GA
                                  1290
```

Figure 3b. The nucleotide sequence of a soybean Lbc_2 gene

genes recently determined by us [6]. All Lb genes so far analysed contain putative regulatory sequences identical to or very similar to the corresponding signals found in other eukaryotic genes [13]. The presence of these sequences in the Lb genes suggests that the mechanisms for gene transcription in plants are very similar to those used in other eukaryotes.

ACKNOWLEDGEMENTS

This work was supported by the Danish Research Council, Novo Industri A/S, the Olga and Esper Boel Fond, and De Danske Sukkerfabrikker A/S.

REFERENCES

1. Fuchsman, W.H. and Appleby, C.A. (1979) Biochem.Biophys.Acta 579, 314-324.
2. Whittaker, R.G., Lennox, S., and Appleby, C.A. (1981) Biochemistry International 3, 117-124.
3. Sievers, S.G., Huhtala, M.-L., and Ellfolk, N. (1978) Acta Chem.Scand. B32, 380-386.
4. Sullivan, D., Brisson, N., Goodchild, B., Verma, D.P.S., and Thomas, D.Y. (1981) Nature 289, 516-518.
5. Marcker, K.A., Gausing, K., Jochimsen, B., Jørgensen, P., Paludan, K., and Truelsen, E. (1981) in Genetic Engineering in the Plant Sciences, Panopoulos, N.J., Ed., Praeger Publishers.
6. Hyldig-Nielsen, J.J., Jensen, E.Ø., Paludan, K., Wiborg, O., Garrett, R., Jørgensen, P., and Marcker, K.A. (1982) Nucl. Acids Res. 10, 689-701.
7. Maniatis, T., Hardison, R.C., Lacy, E., Lauer, J., O'Connell, C., Quon, D., Sim, G.K., and Efstratiadis, A. (1978) Cell 15, 687-701.
8. Lacy, E., Hardison, R.C., Quon, D., and Maniatis, T. (1979) Cell 18, 1273-1283.
9. Messing, J., Crea, R., and Seelong, P.H. (1981) Nucl.Acids Res. 9, 309-322.
10. Sanger, F., Coulson, A.R., Barrell, B.Q., Smith, A.J.H., and Roe, B.A. (1980) J.Mol.Biol. 143, 161-178.
11. 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., and Proudfoot, N.J. (1980) Cell 21, 653-668.
12. Breathnach, R., Benoist, C., O'Hare, K., Gannon, F., and Chambon, P. (1978) Proc.Natl.Acad.Sci.USA 75, 4853-4857.
13. Breathnach, R. and Chambon, P. (1981) Ann.Rev.Biochem. 50, 349-383.