Volume 10 Number 2 1982 Nucleic Acids Research

The primary structures of two leghemoglobin genes from soybean

Jens Jørgen Hyldig-Nielsen*, Erik Ø.Jensen*, Kirsten Paludan*, Ove Wiborg*, Roger Garrett+, Poul Jørgensen* and Kjeld A.Marcker*

+Division of Biostructural Chemistry, Department of Chemistry, and *Department of Molecular Biology and Plant Physiology, University of Aarhus, 8000 Aarhus C, Denmark

Received 10 November 1981; Accepted 2 December 1981

ABSTRACT

We present the complete nucleotide sequences of two leghemoglobin genes isolated from soybean DNA. Both genes contain three intervening sequences which interrupt the two coding sequences in identical positions. The 5' and 3' flanking sequences in both genes contain conserved sequences similar to those found in corresponding positions in other eukaryotic genes. Thus, the general DNA sequence organization of these plant genes is similar to that of other eukaryotic genes.

INTRODUCTION

Leghemoglobins (Lb) are myoglobin like proteins which have only been found in the nitrogen-fixing root nodules of legumes symbiotically associated with Rhizobia. Recent evidence shows that the Lb genes are encoded in the plant genome[1]. Soybean nodules contain four major species of Lbs called Lba, c_1, c_2 and c_3, respectively[2]. The soybean nodule also contains several minor Lb compounds, but some of these components seem to be posttranslational modification products of some of the major components[3]. The differences in the amino acid sequences among the various Lb compounds are small corresponding to 6-8 amino acid substitutions only[4]. By computer analysis Hunt et al.[5] have compared homologies among the amino acid sequences of globins including several Lbs. The observed structural homology suggests that Lbs and globins have a common evolutionary origin.

We have so far isolated five separate Lb genes from soybean DNA. In this paper we report complete nucleotide sequences of a cloned Lba gene and a cloned Lbc gene. The two genes are very similar. Both coding regions contain three intervening sequences which interrupt at codons 32 (IVS-1), 68-69 (IVS-2) and

0305-1048/82/1002-0689 $2.00/0

103-104 (IVS-3) in both genes. The positions of IVS-1 and IVS-3 in the Lb coding sequences are the same as the positions of the two interruptions found in all other known globin coding sequences[6]. IVS-1 and IVS-3 show a considerable size variation in the two genes. Thus, IVS-1 is 119 bp in the Lba gene and 169 bp in the Lbc gene, while IVS-3 is 680 bp and 285 bp, respectively. The two IVS-1 sequences display a striking homology and some homology is also apparent in the IVS-2 sequences. In contrast the two IVS-3 sequences are widely divergent. The nucleotide sequences upstream from the structural genes contain an ATA box located 30 nucleotides away from a putative cap addition site. In the 3' noncoding regions the sequence GATAAA is located 20-21 bp proximal to the possible polyA addition sites. This sequence probably corresponds to the hexanucleotide AATAAA found in a similar position in most other eukaryotic genes[7]. Thus, the general DNA sequence organization of these plant genes is similar to that of other known eukaryotic genes.

MATERIALS AND METHODS

Restriction endonucleases were purchased from either Biolabs, New England, or Boehringer, Mannheim, DNA polymerase (Klenow fragment) 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 and γ-^{32}P-ATP were from New England Nuclear.

Isolation of genomic Lb-genes. The genomic recombinant molecules containing Lb-sequences were isolated from two different soybean DNA libraries. One library was constructed from a complete EcoRI digest of soybean DNA, using λgtWes/λB as a vector. The other was constructed by R.Goldberg and R.Fisher from a partial EcoRI digest of soybean DNA using Charon 4 as a vector. About 6 x 10^5 recombinant Charon phages and 8 x 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.*[8] The procedures used to construct subclones and to prepare plasmid DNA were according to Lacy *et al.*[9]

M13 Cloning. Appropriate DNA fragments were subcloned in the filamentous bacteriophages M13mp7 [10] and propagated in the host JM 101 in 2 NZY (10 g NaCℓ, 4 g MgCℓ$_2$, 7 H$_2$O, 20 g NZamide-typeA, 10 g yeast extract in 1 ℓ H$_2$O). 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.[11] using a synthetic dodeca deoxy nucleotide as primer, or in some cases by the chemical degradation procedure described by Maxam and Gilbert[12]. Sequencing reaction products were electrophoresed on 6 or 8% 0.3 mm polyacrylamide-urea gels.

RESULTS AND DISCUSSION

Sequencing strategy and procedures. Southern blotting analysis 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[6]. 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 using λgtWesλB as a vector. The sizes of the cloned DNA fragments were 1.4, 4.2, 6.0, 7.5, 12 and 13 kb, respectively. The 7.5 kb EcoRI fragment was previously characterized by restriction enzyme analysis and partial sequencing[6]. Sequence analysis of the 4.2 kb and the 1.4 kb EcoRI fragments revealed that both contained incomplete Lb genes, the 4.2 kb fragment carrying a 5' end and the 1.4 kb fragment a 3' end. The combined 4.2 kb and 1.4 kb EcoRI fragments were isolated from a Charon 4 library which was constructed from a limited EcoRI digest of soybean DNA. The DNA sequences of both genes were determined largely by the chain terminator method after cloning appropriate fragments into the single-stranded phage M13mp7 [10]. In a few instances the DNA sequences were determined using the chemical degradation procedure of Maxam and Gilbert[12]. Figs 1a and 1b outline the strategy used for the determination of the DNA sequences of both genes.

Nucleotide sequence of the two Lb genes. The entire nucleo-

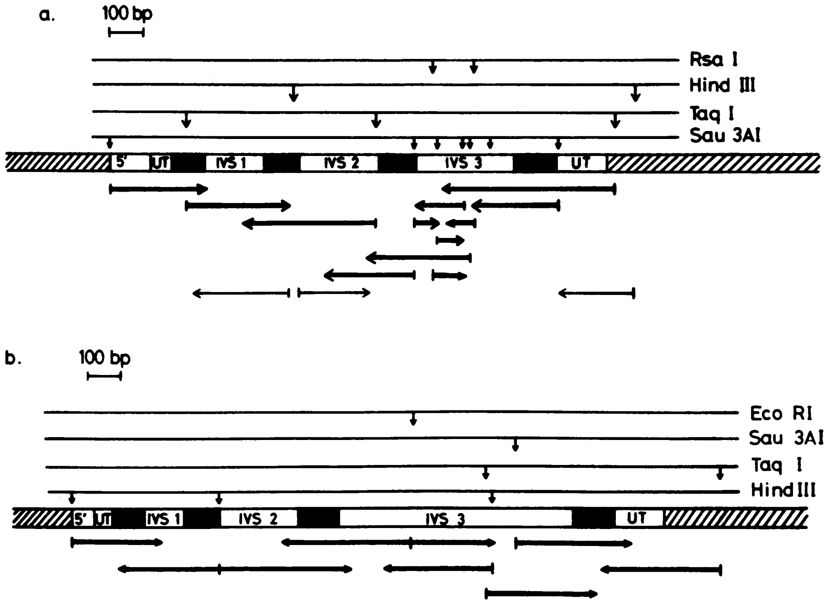

Figure 1. Strategy for determining the nucleotide sequence of the Lbc gene (a) and the Lba gene (b)
A detailed restriction nuclease map for those restriction nuclease sites (vertical arrows) used in deriving the sequence. The extent and direction of each sequence reading are indicated by horizontal arrows. Thick horizontal arrows indicate sequences determined by the dideoxy sequencing technique[11]. Thin horizontal arrows indicate sequences determined by the chemical degradation procedure[12]. UT represent sequences corresponding to the non-translated 5' and 3' regions of the Lb mRNA. IVS-1, IVS-2, and IVS-3 denote intervening sequences.

tide sequence of the gene contained in the 7.5 kb fragment is shown in Fig.2a while the sequence of the gene contained in the combined 4.2, 1.4 kb fragment is shown in Fig.2b. Comparison of these genes with known amino acid sequences of soybean Lbs indicatesthat the sequence represented in Fig.2b corresponds to Lba. There are a few discrepancies between the DNA sequences determined here and the published amino acid sequence[4]. Thus at positions 102-105 in the amino acid sequence a -Phe-Val-Val-Lys- sequence was determined, while the corresponding DNA sequence indicates a -Phe-Val-Val-Val-Lys- sequence. In addition the carboxy terminal sequence was reported as -Lys-Ala-Lys- while the

DNA sequence corresponds to -Lys-Lys-Ala. The DNA sequence shown in Fig.2a indicates that in this case the gene corresponds to one of the Lbc varieties. Available amino acid sequence data[4] suggest that in this case the determined DNA sequence corresponds to Lbc_1. 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).

DNA sequences in both genes corresponding to restriction enzyme cleavage sites were verified by cleavage with the appropriate restriction enzyme with one exception. In the Lbc gene a DNA sequence corresponding to a ClaI cleavage site was determined at nucleotide positions 1024-1029. This sequence has been read on both strands from several different clones. However, despite repeated attempts neither ClaI nor TaqI cleave in this position. At present we have no explanation for this discrepancy.

Flanking and non-coding regions. In both genes the initiation codon ATG immediately precedes the N-terminal codon. Thus, there is no leader sequence coding for a signal peptide and consequently there is no indication for transport of Lb through membranes in the nodule. The sequences determined include a 174 bp (Lbc) and a 114 bp (Lba) region 5' to the ATG initiator codon. The sequence of the 5' non-coding end of Lb cDNA has not been determined. Thus, cap addition sites and ATA boxes can only be inferred by homology with the 5' non-coding regions of other eukaryotic genes. Unfortunately, this comparison does not give an unambiguous answer. In both genes there are two potential cap addition sites. In the Lba gene these sites correspond to nucleotide positions 64 and 73 and in the Lbc gene to positions 117 and 126, respectively. Depending upon the choice of position for the cap addition site the ATA box in both genes is located 30 or 39 nucleotides upstream corresponding to positions 34 (Lba) and 87 (Lbc).

The Lbc sequence includes a longer 5' flanking region than that determined for the Lba gene. It is noteworthy that 40 nucleotides upstream from the ATA box the sequence -CCAAG- occurs at positions 43-47. In most eukaryotic genes a homologous sequence occurs at or close to this position[14].

```
                                     30                                      60
GAT CAT TGG CTC TXX GTC ATG CCG ATT GAC ACC CTC CAC AAG CCA AGA GAA ACT TAA GTT

                                     90                                     120
GTA AAC TTT CTC ACT CCA GCC TTC TAT ATA ACA TGT ATT GGA TGT GAA GTT ATT GCA TAA

                                    150                                     180
                                                                            GLY
CTT GCA TTG AAC AAT AGA AAA TAA CAA AAA AAA GTA AAA AAG TAG AAA AGA AAT ATG\GGT

                                    210                                     240
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

                                    270                                     300
ILE PRO GLN TYR SER VAL VAL PHE TYR ASN SER
ATT CCT CAA TAC AGC GTT GTG TTC TAC AAT TC/GTAA GTT TTC TCT ATA AGC ATG TGT CTT

                                    330                                     360
TCA TTC TAT GTT TTT CTT CTG GAA ATT TTT TGT GTT TGA AAA AAG ATA TAT ATA TAT ATA

                                    390                                     420
TAT ATA TAT ATA TAT ATA TAT ATA TAT ATA TAT ATA TAT TTT GTT AAT GTG AGT GGT TTT

                                    450                                     480
                             ILE LEU GLU LYS ALA PRO ALA ALA LYS ASP LEU PHE SER
GGT TTG ATT AAA AAT AAA TAG/GATT CTG GAG AAA GCA CCT GCA GCA AAG GAC TTG TTC TCA

                                   510                                     549
PHE LEU ALA ASN GLY VAL ASP PRO THR ASN PRO LYS LEU THR GLY HIS ALA GLU LYS LEU
TTT CTA GCA AAT GGA GTA GAC CCC ACT AAT CCT AAG CTC ACG GGC CAT GCT GAA AAG CTT

                                  570                                      600
PHE ALA LEU
TTT GCA TTG\GT AAG TAT CAG CCA ACT AAA ATT ATA ACT ATT TTA TGT GAT TAA TTT TAA

                                    630                                     660
GAT TAA ACA TCA TGT ATT TTA ACA CTC TTA AAA TAT CAA TGA ACA TTA ATT TTT TGA ATT

                                    690                                     720
GTA TTT TAT ATT TTT ACC ATA TCT TGA ACT AGG AAT AAT ATA TAA ATT TCT ATT AGT ATT

                                    750                                     780
TCT TGG TAA TTA CAT ATA TAT ATA TAT ATA TAA TCC TTG TGA TAA TTA TTT TTC GAA TTT
                                    810                                     840
     VAL ARG ASP SER ALA GLY GLN LEU LYS THR ASN GLY THR VAL VAL ALA ASP ALA ALA
GTAG/GTG CGT GAC TCA GCT GGT CAA CTT AAA ACA AAT GGA ACA GTG GTG GCT GAT GCT GCA

                                   870                                     900
LEU VAL SER ILE HIS ALA GLN LYS ALA VAL THR ASP PRO GLN PHE VAL
CTT GTT TCT ATC CAT GCC CAA AAA GCA GTC ACT GAT CCT CAG TTC GTG/GT ATG ATA AAT

                                    930                                     960
AAT ACT AGT AAA ATG TTA CAA TAA ATG CAA ACT TAA GTT TTA CGT ACA TAG TGA TCA TGA

                                    990                                    1020
CTT CAT GCA TGG CTA TTA TTT TTT CAT ATT TAT TGA AGT CAA CTT AAA ATT TTG TAA ATA

                                   1050                                    1080
CAG ATC GAT GCT AGT AAT TTG TTG AGA TCA TGA GAA AAC GTA CCA CTA CTC CAA TAG CAT
```

```
                                   1110                                    1140
TAC TCA TTT TGA AAA TTG TAT AAC TGT GAT CTA ATT ATA AGG AAA AAG TGT ATA TAA GAG

                                   1170                                     1200
                                             VAL VAL LYS GLU ALA LEU LEU LYS THR
CTA ATC CAT TAT TAA TGT TTT TTA TAT TTT GTAG/GTG GTT AAA GAA GCA CTG CTG AAA ACA

                                   1230                                    1260
ILE LYS GLU ALA VAL GLY GLY ASN TRP SER ASP GLU LEU SER SER ALA TRP GLU VAL ALA
ATA AAG GAA GCT GTT GGC GGC AAT TGG AGT GAC GAA TTG AGC AGT GCT TGG GAA GTA GCC

                                   1290                                    1320
TYR ASP GLU LEU ALA ALA ALA ILE LYS LYS ALA ***
TAT GAT GAA TTG GCA GCA GCA ATT AAA AAG GCA TAA/TT AGG ATC TAC TGC ATT GCC GTA

                                   1350                                    1380
AAG TGT AAT AAA TAA ATC TTG TTT CAA CTA AAA CTT GTT ATT AAA CAA GTT CCC TAT ATA

                                   1410                                    1440
AAT GTT GTT TAA AAT AAG TAA ATT TCA TTG TAT TGG ATA AAC ACT TTT AAG TTA TAT ATT

                                   1470                                    1500
TCC ATA TAT TTA CGT TTG TGA ATC ATA ATC GAT ACT TTA TAA AAA TAA ATT CCA AAT AAT

TTA TAC GTT TTA AAA ATT ATT TT
```

Figure 2a. The nucleotide sequence of a soybean Lbc gene.

We have recently determined the DNA sequence of the 3' noncoding region of a Lb cDNA clone. This region consists of 142 nucleotides. However, the cDNA sequence is not identical to the 3' ends of any of the two Lb genes presented here, although a very strong homology is apparent. By comparison of the 3' noncoding regions in the two Lb genes with the corresponding region in the cDNA sequence the polyA addition site is located in nucleotide positions 1725 (Lba) and 1441 (Lbc), respectively.

A sequence GATAAA is found 20 or 21 bp upstream from the presumed polyA addition site in both genes. An identical sequence is present in a similar position in the cDNA sequence. One feature common to almost all eukaryotic poly adenylated mRNAs is the occurrence of the sequence AAUAAA 12-33 nucleotides in front of the polyA addition site. This hexanucleotide probably functions as a polyA addition signal. Because of the location of the GATAAA sequence and its obvious homology with AAUAAA we suggest that the hexanucleotide GAUAAA functions as

```
                                     30                                      60
AAG CTT TGG TTT TCT CAC TCT CCA AGC CCT CTA TAT AAA CAA ATA TTG GAG TGA AGT TGT

                                     90                                     120
                                                                            VAL
TGC ATA ACT TGC ATC GAA CAA TTA ATA GAA ATA ACA GAA AAT TAA AAA AGA AAT ATG/GTT

                                    150                                     180
ALA PHE THR GLU LYS GLN ASP ALA LEU VAL SER SER SER PHE GLU ALA PHE LYS ALA ASN
GCT TTC ACT GAG AAG CAA GAT 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\G TAA GTT TTC TCT CTA AGC ATG TGT CTT

                                    270                                     300
CCA TTC TAT GTT TTT CTT TTG GAA ATT TGT TGT GTT TGA AAA AAG ATA TAT TGT TAA TGT

                                    330                                     360
                                           ILE LEU GLU LYS ALA PRO ALA ALA LYS ASP
GAG TGG TTT TGG TTT GAT TAA AAA TGA ATAG/G ATA CTG GAG AAA GCA CCT GCA GCA AAG GAC

                                390                                     420
LEU PHE SER PHE LEU ALA ASN GLY VAL ASP PRO THR ASN PRO LYS LEU THR GLY HIS ALA
TTG TTC TCA TTT CTA GCA AAT GGA GTA GAC CCC ACT AAT CCT AAG CTC ACG GGC CAT GCT


                                   450                                     480
GLU LYS LEU PHE ALA LEU
GAA AAG CTT TTT GCA TTG/GTAA GTA TCA CCC AAC TAA AAT TAT AAC TAT TTT ATG TGA


                                    510                                     540
TTA ATT TTA AGA TTA AGC ATC ATG TAT TTT AAC ACT CTT AAA ACA TCA ATG AAC ATT AAT

                                    570                                     600
TGT TTG AAT TGT ATT TTA TAT TTT TGC CAT ATC TTG AAC TAG GAA TAG TAT ATA AAT TTC

                                    630                                     660
TAT TAG TAT TTG TTG ATA ATT ATT TTT CTT TCA TAA CTA TCT TGT CAC ATA TTA TAT ATT

                                       690                                     720
                 VAL ARG ASP SER ALA CLY GLN LEU LYS ALA SER GLY THR VAL VAL ALA
TTT TGA ATT GTAG/GTG CGT GAC TCA GCT GGT CAA CTT AAA GCA AGT GGA ACA GTG GTG GCT

                                  750                                     780
ASP ALA ALA LEU GLY SER VAL HIS ALA GLN LYS ALA VAL THR ASP PRO GLN PHE VAL
GAT GCC GCA CTT GGT TCT GTT CAT GCC CAA AAA GCA GTC ACT GAT CCT CAG TTC GTG/GT

                                    810                                     840
ATG ATA AAT AAT GAA ATG TTA TAA TAA ATT ATG CAT ACT TCA ATT TTT CAT GGA GCA GTA

                                    870                                     900
TAA TGA TCA ACA CAC ACT TCT TTT GTT TCA TGC ATT TGA TAA CTA CAA TCT TAA AAT GTT

                                    930                                     960
GCA ATC TTA AAA ATA GTA TTA AAA ATA TAA CAT TTA ATT AGC TCA TCA ATA TTT TTC TGT

                                    990                                    1020
TGC AAT TTT TTA TGA AAA AAT TAT AAT TAT GAA TTC TTT GAG CAA TGT TTA ATT AAA AAA
```

```
                                   1050                                    1080
TTG ATT TAA TAA TGA AAT AAC TAA GCT ACC TCT GTC TCG TTT TTC ATT TAA ACT ATG ACA

                                   1110                                    1140
TAA ACA ATG AAT AAA GTA AAC TAA ACC ATG ACA TGT TTA TTT TTG AAT GAG GTT ATT AAT

                                   1170                                    1200
AAT TTT TTT TCA CTA TCT ATT GCA ATG TTC ATT GAT TAT CAA TTA TCT TGG TTG CAT TGA

                                   1230                                    1260
TTC TCT CGA TTT TTT TCT TGA GGT TAA GCT TCA GTT CAA TAT ATA TTC ATT TTT TGA TAA

                                   1290                                    1320
AAA AAA ATA GTA CAA TAT ATT TTC ATT TAG CTG ATC ATA TTT ATT TAA GTT CAA CTT AAA

                                   1350                                    1380
ATT TTA TAG ATG TTA ATT GAT ATA ATT TGT TGA GAT GAT GAG AAG ACC AAT ACC ATT ACG

                                   1410                                    1440
TAC TCT TTT GAA AGT GTT ATA TGG ATT TTA ATT ATA AGG AAA AAT GTA AGA GCT AAA CCA

                                   1470                                    1500
                        VAL VAL LYS GLU ALA LEU LEU LYS THR ILE LYS ALA ALA VAL
TTG CTG ATG ATT TTG AAG/GTG GTT AAA GAA GCA CTG CTG AAA ACA ATA AAG GCA GCA GTT

                                   1530                                    1560
GLY ASP LYS TRP SER ASP GLU LEU SER ARG ALA TRP GLU VAL ALA TYR ASP GLU LEU ALA
GGG GAC AAA TGG AGT GAC GAG TTG AGC CGT GCT TGG GAA GTA GCC TAC GAT GAA TTG GCA

                                   1590                                    1620
ALA ALA ILE LYS LYS ALA ***
GCA GCT ATT AAG AAG GCA TAA TTA GTA TCT ATT GCA GTA AAG TGT AAT AAA TAA ATC TTG

                                   1650                                    1680
TTT CAC TAT AAA ACT TGT TAC TAT TAG ACA AGG GCC TGA TAC AAA ATG TTG GTT AAA ATA

                                   1710                                    1740
ATG GAA TTA TAT AGT ATT GGA TAA AAA TCT TAA GGT TAA TAT TCT ATA TTT GCG TAG GTT

                                   1770                                    1800
TAT GCT TGT GAA TCA TTA TCG GTA TTT TTT TTC CTT TCT GAT AAT TAA TCG GTA AAT TAT

                                   1830                                    1860
ACA AAT AAG TTC AAA ATG ATT TAT ATG TTT CAA AAT TAT TTT AAC AGC AGG TAA AAT GTT

ATT TGG TAC GAA AGC TAA TTC GTC GA
```

Figure 2b. The nucleotide sequence of a soybean Lba gene.

the polyA addition signal in this particular plant system.

Intervening sequences. Inspection of the DNA sequences reveals that the coding sequences in both genes are interrupted at codon 32, 68-69 and 103-104. The length of the intervening sequences are for the Lba gene: IVS-1 119 bp, IVS-2 233 bp and IVS-3 680 bp. The corresponding lengths in the Lbc gene are 169, 234 and 285 bp, respectively. Thus, a considerable size variation is noted for in particular the two IVS-3 sequences. A curious feature of the Lbc IVS-1 sequence is the presence of an AT repeat consisting of 52 nucleotides which is almost absent from the corresponding Lba sequence. Breathnach et al.[15] have noted the tendency for intervening sequences to begin with the dinucleotide GT and end with the dinucleotide AG. For the two Lb genes all sequences around splicing junctions can be aligned such that intervening sequences start with GT and terminate with AG. Comparison of the two IVS-1 sequences reveals very little divergence apart from the presence of the AT repeat in the Lbc sequence. In fact this divergence is of the same order of magnitude as that observed for the coding sequences. Similarly, the two IVS-2 sequences display a considerably homology up to nucleotide positions 623 (Lba) and 732 (Lbc) whereafter a considerable divergence is noted, until shortly before the splicing junctions. In contrast no homology is observed in the two IVS-3 sequences except for sequences located around the splicing junctions. Thus sequences located towards the 5' end of the gene seem to be much more conserved than sequences located towards the 3' end. Similar observations have also been reported for the sequence divergence of the two intervening sequences present in β-like globin genes[14]. Analysis of these genes showed that in recently diverged pairs of genes the small intervening sequence which interrupts at codons 30-31 displays much less sequence divergence than the larger intervening sequence which interrupts at codons 104-105.

Evolutionary considerations. Analysis of the amino acid sequences of globin including several Lbs revealed sufficient structural homology to suggest that all globins have a common evolutionary origin[5]. The coding sequences of all globin genes so far analyzed are interrupted by two intervening sequences. When the amino acid sequences of all globins and Lbs are aligned to maximize the structural homologies the splicing points of IVS-1 and IVS-3 in the Lbs coincide with the two splicing points found in the globins[6]. This finding supports the notion that indeed all globins share a common ancestor.

Gō[16] has recently defined four regions of the globin polypeptide chain that are distant from one another in the globin fold. This analysis revealed that the splicing points in the coding sequence divide off these compact structures from one another except for the central coding sequence which consists of two such regions. Consequently Gō suggested that the central exon in globins might be the result of a fusion between two exons with a division somewhere between amino acid residues 66-71. The location of IVS-2 between codons 68-69 in the Lb gene is in excellent agreement with Gō's proposal. The correspondance of exons in the Lb genes with compact polypeptide regions in globin protein structure add additional support to the idea that exons correspond to such compact structures[17]. Thus the Lb gene has all the appearance of a primitive globin gene. Accordingly somewhere along the line of globin development the two middle exons fused creating a single exon. It is unclear when and why this happened, but it is not unreasonable to suppose that the fusion occurred when structures having allosteric oxygen binding properties began to evolve. Lb and myoglobin are functional monomers. Thus if this hypothesis is correct the gene structure of myoglobin should have an intervening sequence corresponding to IVS-2 in the Lb gene. Unfortunately the gene structure of myoglobin is not known, so this prediction cannot yet be verified.

The central exonic region in globin binds haem tightly and specifically[18,19]. In Lb IVS-2 separates the proximal and distal haem contacts which may suggest that the divided central exonic region in Lb represents a primitive form of a haem binding domain. A structural similarity between globins and certain cytochromes have been noted[20]. It is interesting that in this case the structural homology corresponds rather precisely to the central exonic region in the globin chain. Based on such chemical and structural considerations Blake[21] has recently proposed that globins and cytochromes may have evolved from a common haem binding domain encoded by one or more exons which by combining with other gene elements has generated a multiplicity of haem binding proteins of diverse functions. This proposal for the mechamism of globin and cytochrome evolution is in agreement with Gilbert's suggestion[22] that in eukaryotes exons code for functional protein units which can serve in rapid protein evolution.

Lb has never been detected in plants other than legumes. The first legumes appeared about 200 million years ago[23]. Assuming that the rate of globin evolution has remained constant with time, calculation then indicates that Lb diverged from the common ancestor about 1500 million years ago[24]. The emergence of a functional Lb in legumes is therefore an enigma. Is is, however, possible that the presence of Lb in legumes is the result of convergent evolution, which then would suggest that the Lb gene did evolve by recombination of two haem binding exons present in plants with two other plant gene elements according to the scheme recently proposed by Blake[21].

Appleby[25] has made the interesting suggestion that a Rhizobium globin-like gene was transferred into the genome of a primitive legume host, after which modern Lb evolved. Such a mechanism is reminiscent of Agrobacterium tumefaciens infection of plants during which bacterial genes are inserted into the plant genome[26]. There are indeed many similarities between rhizobial and agrobacterial infections of plants. However, the presence of intervening sequences in the Lb gene does argue against this possibility since such sequences have never been detected in prokaryotes. Thus we consider the possible rhizobial origin of Lb in legumes rather unlikely.

Another explanation for the presence of a globin gene in the genome of a plant is that it was translocated there recently in evolution as a passenger on a virus. The presence of the Lb gene in legumes would then be the result of a horisontal transmission of a gene. Such a mechanism circumvents the rules of claccical mendelian genetics with rather important implications for our understanding of the mechanism of evolution.

ACKNOWLEDGEMENTS

This work was supported by the Danish Research Council, Novo Industri, A/S, and the Olga and Esper Boel Fond. We thank Drs. R. Goldberg and R.Fisher, UCLA, for making their Charon 4 soybean library available to us, and Mrs. Y.Gregersen and Mrs. A.Marcker for technical assistance.

REFERENCES

1. Baulcombe, D. and Verma, D.P.S. (1978) Nucl.Acids Res. 5, 4141-4153.
2. Fuchsman, W.H. and Appleby, C.A. (1979) Biochem.Biophys.Acta 579, 314-324.
3. Whittaker, R.G., Lennox, S., and Appleby, C.A. (1981) Biochemistry International 3, 117-124.
4. Sievers, S.G., Huhtala, M.-L., and Ellfolk, N. (1978) Acta Chem.Scand. B32, 380-386.
5. Hunt, T.L., Hurst-Calderone, S., and Dayhoff, M.O. (1978). Atlas of Protein Sequence and Structure 5, Suppl.3, 229-251.
6. Jensen, E.Ø., Paludan, K., Hyldig-Nielsen, J.J., Jørgensen, P. and Marcker, K.A. (1981) Nature 291, 677-679.
7. Proudfoot, N.J. and Brownlee, G.G. (1976) Nature 263, 211-214.
8. 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.
9. Lacy, E., Hardison, R.C., Quon, D., and Maniatis, T. (1979) Cell 18, 1273-1283.
10. Messing, J., Crea, R., and Seelong, P.H. (1981) Nucl.Acids Res. 9, 309-322.
11. Sanger, F., Coulson, A.R., Barrell, B.Q., Smith, A.J.H., and Roe, B.A. (1980) J.Mol.Biol. 143, 161-178.
12. Maxam, A.M. and Gilbert, W. (1977) Proc.Natl.Acad.Sci. 74, 560-564.
13. Sanger, F. and Coulson, A.R. (1978) FEBS Lett. 87, 107-110.
14. 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.
15. Breathnach, R., Benoist, C., O'Hare, K., Gannon, F., and Chambon, P. (1978) Proc.Natl.Acad.Sci.USA 75, 4853-4857.
16. Gō, M. (1981) Nature 291, 90-92.
17. Blake, C.C.F. (1979) Nature 277, 598.
18. Eaton, W.A. (1980) Nature 284, 183-185.
19. Craik, C.S., Buchman, S.R., and Beychok, S. (1980) Proc.Natl. Acad.Sci.USA 77, 1384-1388.
20. Argos, P. and Rossmann, M.G. (1979) Biochemistry 18, 4951-4960.
21. Blake, C.C.F. (1981) Nature 291, 616.
22. Gilbert, W. (1978) Nature 271, 501.
23. Ramshaw, J.A.M., Richardson, D.L., Meatyard, B.T., Brown, R.H., Richardson, M., Thompson, E.W., and Boulter, D. (1972) New Phytol. 71, 713.
24. Dayhoff, M.O., Hunt, L.T., McCaughlin, P.J., and Jones, D.D. (1972) in Atlas of Protein Sequence and Structure 1972, Dayhoff, M.O. Ed., National Biomedical Res.Found., Washington, pp.17-30.
25. Appleby (1974) in The Biology of Nitrogen Fixation, Quispel, A. Ed., North-Holland Publishing Company, Amsterdam Oxford, pp.499-554.
26. Zambryski, P., Holsters, M., Kruger, K., Depicker, A., Schell, J., Van Montagu, M., and Goodmann, H.M. (1980) Science 209, 1385-1391.