

---

**A small family of nodule specific genes from soybean**

---

Niels N.Sandal, Kirsten Bojsen<sup>1</sup> and Kjeld A.Marcker

---

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

---

Received November 19, 1986; Revised and Accepted January 29, 1987

---

**ABSTRACT**

The primary structure of two nodule specific soybean genes are presented. The two genes code for primary products of 20.0 (nodulin 20) and 22.7 (nodulin 22) kdaltons, respectively. Both genes are related to the nodulin 23 and 44 genes. Alignment of the deduced amino acid sequences of all four genes revealed three domains of high homology interrupted by highly diverged regions due to numerous duplication and insertion events. The first conserved domain codes for a putative signal peptide, while the two others each contain four Cys residues that can be arranged in a way reminiscent of the metal binding domains present in some enzymes and in several DNA binding proteins.

**INTRODUCTION**

During the soybean - *Bradyrhizobium japonicum* symbiosis several host-encoded (nodulin) genes are specifically expressed in the nodule in addition to the leghemoglobin (Lb) genes<sup>1</sup>. Nodulins include nodule specific forms of uricase<sup>2,3</sup>, glutamine synthetase<sup>4</sup> and sucrose synthetase<sup>5,6</sup>, but the functions of the majority of the nodulins are at present unknown. During the differentiation of nodules one of the major changes that occurs inside infected cells is the formation of a subcellular membrane compartment in which bacteria reside. The membrane enclosing the bacteroids, the peribacteroid membrane (pbm), originates from the plasma membrane of the host<sup>7</sup>, but is modified during the progression of the symbiosis. The pbm serves an important function in infected cells. It is required not only as a structural barrier to segregate bacteroids from the host cytoplasm but also to transport substrates required by both the plant and the bacteroid for efficient nitrogen fixation. New polypeptides are integrated into the pbm and possibly have specific functional or structural roles in order to support the

requirements of the symbiosis. Some of the nodulins may therefore be involved in pbm structure and function.

Recently it was found that nodulins 23 and 24 are pbm proteins<sup>8,9</sup>. It is shown here that the nodulin 23 gene from soybean is a member of a small family of genes which consists of at least four members. It is suggested that the gene products of this gene family are involved in pbm structure or function. The gene products contain putative metal binding domains and it is therefore further suggested that metal binding is important for their function.

#### MATERIALS AND METHODS

Screening of phage libraries. Using two related nodule-specific cDNAs as probes three genomic soybean DNA libraries were screened. These were a partial AluI/HaeIII, a partial EcoRI and a partial Sau3A library. Phages from plaques corresponding to positive signals were purified and the DNA prepared according to Maniatis *et al.*<sup>10</sup>.

Preparation of soybean DNA. Chromosomal soybean DNA was prepared from 4 day-old seedlings as previously described (Marcker *et al.*<sup>11</sup>).

Subcloning into pBR322 and 328. Recovery of DNA fragments from agarose gels for subcloning was performed according to Dretzen *et al.*<sup>12</sup>. The procedures used for subcloning and plasmid preparation were according to Maniatis *et al.*<sup>10</sup>

DNA Hybridization. Hybridization of filters was as described in Bojsen *et al.*<sup>13</sup>.

RNA preparation and hybridization. RNA preparation, Northern blotting and RNA hybridization were as described in Marcker *et al.*<sup>14</sup>.

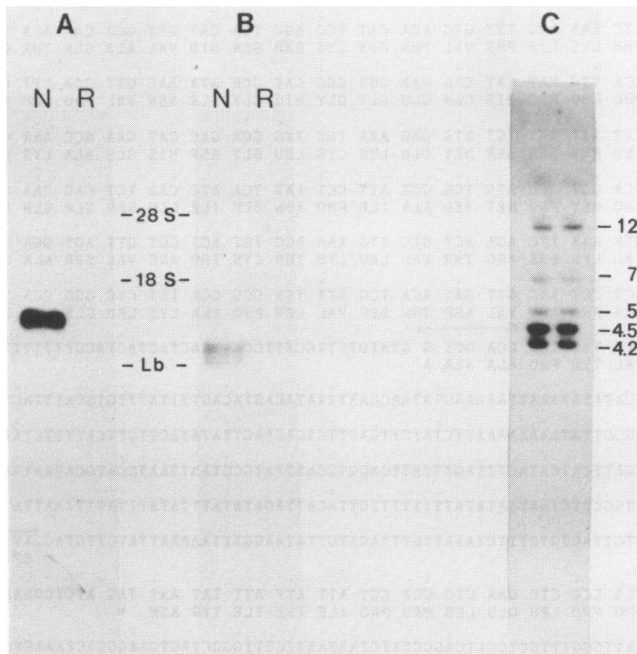
DNA sequencing. DNA sequencing was performed by the dideoxy chain termination method (Sanger *et al.*<sup>15</sup>). Appropriate restriction fragments were cloned into the M13 phages Mp8, Mp9, Mp18 or Mp19.

Primer extension analysis. Suitable primers were synthesized and used for primer extension analysis according to Boel *et al.*<sup>16</sup>. The labelled primers were annealed to 10 µg poly(A)<sup>+</sup> RNA from 20 days old nodules. The extended oligodeoxynucleo-

tides were eluted from polyacrylamide urea gels and sequenced according to Maxam and Gilbert<sup>17</sup>.

### RESULTS

From a soybean nodule cDNA library two cDNA clones were isolated and characterized. The two clones ( $N_{10,8}$  and  $N_{12}$ ) are related because the sequence analysis revealed an extensive homology between the two clones.  $N_{10,8}$  hybridized to two RNAs of 0.8 and 1.0 kb respectively, while  $N_{12}$  hybridized to a 1.2 kb RNA (Fig. 1). When compared to similar experiments using Lb cDNA as a probe the Northern blot analysis suggests that these



**Figure 1.** Northern blotting analysis of PolyA<sup>+</sup> RNA from nodules (N) harvested 17 days post infection and polyA<sup>+</sup> RNA extracted from uninfected roots (R). A: Hybridization with the  $N_{12}$  cDNA clone. B: Hybridization with the  $N_{10,8}$  cDNA clone. 28S, 18S and Lb refer to the positions of the two ribosomal RNAs and Lb mRNA, respectively. C: Corresponding Southern blotting analysis of soybean DNA digested with EcoRI. Both lanes contain the same amount of soybean DNA. The hybridization was performed with a M13 probe generated from clone  $N_{10,8}$ , which contains sequences homologous to  $N_{12}$ . The numbers refer to the sizes of the hybridizing fragments.

```

GAATTCCTCCAAACGCTCAGAGAGACGGAGAAGAGATTGAAACTTCTACTTGTACTGTCTTCATGCGATTCTTTTTTCTCCC -384
ACCACGAATACTATCTCGCAAATCCCAACGGTGGAAAGGGTGAGAAATTTGAATTTCCGAACAATATATCCAAATTTTCATGAA -304
AATCCAAACGGTTAAACGAAATCGGGATCTTAAAAAATATTTAATAAAAAGAAAATGGAAGCAATAAATATGAAAATTGCA -224
ATAGTAATTAAGTGTGATGATAAAAATATATCTATCTTTTTATTTATCTATATTTTCATTTTAATGTTAGAAAGACAAATATA -144
TATATATAGCTCTTGGAAAGCAAACGAGTCAGGATATATGACAAAGATCAAAGCTCGTAATAATTTTTTTTTATAATAAT -64
AAACATAGATGTTGAGTTCCTAGTTTCTCTTGTATATATATTGCAAAATTCACACATACAGAGCACCAGTCCCAATAAAG 17
TTCTCAAAGTGCAAAGTCTTATCACTAAGAGAACCAATTA ATG GAG AAA ATG AGA GTG GTA CTA ATT 87
(MET GLU LYS)MET ARG VAL VAL LEU ILE
ACT CTA TTG TTG TTT ATA GGT GCA GCA GTT GCA GAA AAA GCT GGT AAT GGC LAA GCT GCC 147
THR LEU LEU LEU PHE ILE GLY ALA ALA VAL ALA GLU LYS ALA GLN ASN GLY LYS ALA ALA
AAT AAT CCT GCA GAA GAT GCT AGT GAT GGC GAA GCC ATT AAT CTT GTA GAA GAA GCT GGT 207
ASN ASN PRO ALA GLU ASP ALA SER ASP GLY GLU ALA ILE ASN LEU VAL GLU GLU ALA GLY
GGT ATT GGT GAT GCC ATT ACT CCT GCA GAA GGC AAA GCC ACT AAT CTT CAA CGC TAT GAG 267
GLY ILE GLY ASP ALA ILE THR PRO ALA GLU GLY LYS ALA THR ASN LEU GLN ALA TYR GLU
TCA GCT AGA TTC AAA AAG TTT GTG ACA CAT TGC AGC TCA CAT GTT GCT CAA ACA TGC AGT 327
SER ALA ARG PHE LYS LYS PHE VAL THR HIS CYS SER SER HIS VAL ALA GLN THR CYS SER
GGA AAT GAT CCA TTG CAT CAT CAG GAA GGT GGC CAT GGA ATA AAC GTT CCA CTT GGG TTG 387
GLY ASN ASP PRO LEU HIS HIS GLN GLU GLY GLY HIS GLY ILE ASN VAL PRO LEU GLY LEU
TCA TTT TGC CTT TTT GAT TCT ATG GAG AAA TGC TTG GGA GAC CAT GAA GCC AAA CTT ATA 447
SER PHE CYS LEU PHE ASP SER MET GLU LYS CYS LEU GLY ASP HIS GLU ALA LYS LEU ILE
GAT CCC AAC CCA GGT CCC ATG TCG GCT ATT CCT AAT TCA ATC CAA TCT CAG CAA CTC CTC 507
ASP PRO ASN PRO GLY PRO MET SER ALA ILE PRO ASN SER ILE GLN SER GLN GLN LEU LEU
ATT GAG ACT GTA AAA TTC AGA ACT GTC TTG AAA ACC TGT ACT CGT GTT AGT GCA CAA TTT 567
ILE GLU THR VAL LYS PHE ARG THR VAL LEU LYS THR CYS THR ARG VAL SER ALA GLN PHE
TGT TTA ACT GCT CCT AAC GTT GAT ACA TCG GTT TTA CCG GCA TGT CTC GGG CCA TCT CTC 627
CYS LEU THR ALA PRO ASN VAL ASP THR SER VAL LEU PRO ALA CYS LEU GLY PRO SER LEU
AAT CAA TGT GTT TAT CCT GCA GCT G STATGCTTGCCTTCCAATTACTACTACTACGTTTTTTTTAATGGT 698
ASN GLN CYS VAL TYR PRO ALA ALA A
TAAATATATATATCATATATAAAATTAAGAAGTATAACCAATTAATACAGTACAGTTTTATTTGTGCATTACTCAACAAC 778
TCGTTTCAGTTGAGCCTTATAAAAAAATCTATCTTGACTTCTCACTACTTATATACTTGTTCATTTTCTAANTGATAA 858
AAAATTAAGTACATGATTCATCACTTTAGTTTATTCAGGTGCATTTATGCGTAATTAATCCATGCATAATATATTTTT 938
TGCAAGTATAGTATGGCTTCTAATGATATATTTTTTTCTTACATTAGATATATTTATATTTATTTAATTTATGTGATAC 1018
GCTTCACAAGGTTTGTACTGTCTTCTAAATTTGTTAGATGTTATAAGGTTAAAAAATATCTTGTAC AT GCA TTT 1095
SP ALA PHE
ACA CCT GGC CCG CCG CTC GAA CTC CCA CCT ATT ATT ATT TAT AAT TAG ATCTCGGAAGTACTGC 1159
THR PRO GLY PRO PRO LEU GLU LEU PRO PRO ILE ILE ILE TYR ASN *
AAACAGAAGAATAAATGGGTTTGCATCAGCCTATCTAATAATCCTTTGGCCTAGTGAAGGACTAAAGTCTTTATCT 1239
CTCTTGAACGTTCCAGTGATATGCAATATACATGATTGCCTTGCANTTAACCTTGTATATTATGATGAAGGATCTGTTA 1319
TACCTTCATTATCCAAATTTGCTACATCTGGTCTGTATAAGAAATCACTCATGTGATGAATCATGATCGAATAATTACATT 1399
CTACTAATAGTAATTACTAAAACATGTTTACATCTATTTTTGGTTCAAATTTGCATTGCCACACATAATCA

```

Figure 2. Primary structure of the nodulin 22 gene. The TATA box and a potential polyA addition sequence is underlined. Cap sites are indicated by thin vertical arrows. The putative cleavage site of the signal peptide is indicated by a heavy vertical arrow. Horizontal arrows indicate the length of the N<sub>12</sub> cDNA clone. Underlined sequence: The complementary sequence was used for primer extension analysis.

RNAs are major transcripts within the nodule. Subsequent Southern analysis of genomic soybean DNA revealed the presence of five hybridizing EcoRI fragments of 4.2, 4.5, 5.0, 7 and 12 kb, respectively (Fig.1). Several clones were isolated from three different soybean genomic libraries and further characterized by DNA sequence analysis using the M13 dideoxy chain termination method. In this way two complete genes were characterized with the results shown in Figs 2 and 3. The positions of the cap sites were determined by primer extension analysis using appropriate labelled oligodeoxynucleotides as primers followed by Maxam Gilbert sequence analysis. Both genes are interrupted by an intron near the 3' non-translated end. The gene shown in Fig.2 is contained within the 5.0 kb EcoRI fragment and corresponds to the 1.2 kb mRNA (cDNA clone N<sub>12</sub>), while the gene shown in Fig.3 is contained within the 4.5 kb EcoRI fragment and corresponds to the 1.0 kb mRNA (cDNA clone N<sub>10,8</sub>). The gene corresponding to the 0.8 kb mRNA has not been isolated, but it must be rather homologous to the gene corresponding to the 1.0 kb mRNA, because coding sequences from the first exon of the latter gene cross-hybridize strongly to the 0.8 kb mRNA. However, the smaller mRNA does not cross-hybridize to sequences from the 3' noncoding end and the second exon suggesting a considerable sequence divergence in these regions between the two genes.

The gene corresponding to the 1.0 kb mRNA codes for a primary product of 20.0 kdaltons, while the gene corresponding to the 1.2 kb mRNA codes for a primary product of 22.7 kdaltons. In accordance with the nomenclature for plant specific nodule proteins<sup>18</sup> the terms nodulin 20 and 22 are proposed for these primary products. However, it is not known whether the primary products are subject to posttranslational modifications in which case the apparent molecular weights would be different. Inspection of other known nodulin gene sequences revealed that nodulins 20 and 22 are related to nodulins 23<sup>19,20</sup> and 44<sup>20,21</sup>. Two different versions of the nodulin 23 sequence have been published<sup>19,20</sup>. Limited DNA sequence analysis in our laboratory of a nodulin 23 gene supports the sequence proposed by Sengupta-Gopalan *et al.*<sup>20</sup> and consequently this sequence is

```

GATCTTATTTCCGATTAGTGAATATTAATCCCTTTTATAATAATTTAAAAACTTCTCCTATTAAGTCCAATTACAATTTA -452
ACTCTTAATTAATTTATTATTATTATTATTAGTCGTTTCATTAGTCACATGTCTTTTCACATGAGACACTAATTCTCAACTC -372
TCTCACTTAGTTCATATGACATTAATAAATATATAGATTAATTTATTATTATTATTATTAATCCTTTCATAAATCAG -292
ATATCTCACTCATAACTCATTAAATCTAACAACCTTTAACCAAGAGGAAAAAGGAAAAAGGAAAAATGAATACTGAGTAA -212
TTAGTGTAAATGGTAAATATATCTTATTGTTTTGTTATCTGTATTCCATTTTAAATGTTAGAAAGGCAAATGTATATATAG -132
CTCTTGAAAGCAAATGAGTCAGGATATATGACAAAGATCGCGCAGGGTCGTTATAAATTTTTATAATAAGATATACC -52
TTGAGCTCTAGTTTATCTTGTATATATATATGCAAATTCACATACATGAGCACCAAAGCAAATAACATTTCCAA ATG 27
NET
AGA GTG GTA TTA ATT ACT TTA TTCCTG TTT ATA GGT GCA GCA GTT GCAGAA GAC GCT GGT 87
ARG VAL VAL LEU ILE THR LEU PHE LEU PHE ILE GLY ALA ALA VAL ALA GLU ASP ALA GLY
ATT GAT GCC ATT ACT CCT GAA GAA GGC AAA GCC AAT AAT ATT ATT GAG GCG TAT GAG TCA 147
ILE ASP ALA ILE THR PRO GLU GLU GLY LYS ALA ASN ASN ILE ILE GLU ALA TYR GLU SER
CCT AGA TTC CAA AAG TTT GTG ACA CAT TGC AGC TCA CAT GTT ACT CAA ACA TGC AGT GGA 207
PRO ARG PHE GLN LYS PHE VAL THR HIS CYS SER SER HIS VAL THR THR CYS SER GLY
AAT GAT CCA TTA AAT AAT CAG GAG GCC AGT AGA ATG AAT AGT CCA TTT GGG TIG TCT TTT 267
ASN ASP PRO LEU ASN ASN GLN GLU ALA SER ARG MET ASN SER PRO PHE GLY LEU SER PHE
TGC CTT TTT GAT TCT ATG GAG AAG TGC TTG GCA GAC CAT AAA GCC TCA CTT AAA GAT CCC 327
CYS LEU PHE ASP SER MET GLU LYS CYS LEU ALA ASP HIS LYS ALA SER LEU LYS ASP PRO
CAA GAT AAC AAC AAC CTA GCT TCA ATG TCG TCT CTT CCT GGC TCA ATC CAA AAT CAG CCA 387
GLN ASP ASN ASN ASN LEU ALA SER MET SER SER LEU PRO GLY SER ILE GLN ASN GLN PRO
CTC CTC ATT GAG ACT GTA AAA TTC AGA GCC GTC TTG AAA ACC TGT TCC CAT GTC AGT GCA 447
LEU LEU ILE GLU THR VAL LYS PHE ARG ALA VAL LEU LYS THR CYS SER HIS VAL SER ALA
CGA TAT TGT TTC ACT AAT CCT AAC GTT GCT ACA TCG GCT TTA GCG GAT TGT CTC ATG CCA 507
ARG TYR CYS PHE THR ASN PRO ASN VAL ALA THR SER ALA LEU ALA ASP CYS LEU MET PRO
TCT CTC ACT CAT TGT GTT TAT CTT TCT A GTATGTTTTGTTTTGCCAATTACTACTACTGCTATTTTTTTTC 577
SER LEU THR HIS CYS VAL TYR PRO SER S
AAATGGTTAAATATATAAATCATATATAAAATTAAGAAATATAATTAACTAATACAATTTTTATTTTTGCATTTCAAAAA 657
CTCATTTCAGTTGAGCCTTATCAAAAAATAAATCTATTTGACTTCTCACTACTTTTTATTCACTTTCTAATTGATAAAAAA 737
AAATTTGCACAATCGACTAACAAAACATGTAAAAAAATGTTTTTTTTTCTCCATAGATTTTTAATATATAGCTGGTGT 817
TTATCAAATTTTATACTCTTCGTAATAAAGAGAGATAAATACAATTTTTTAAATGAAAAAAATATGTAAAAATCT 897
AACACGAGAGAATTCATTTTACACAAATACATATACTCAAATGCATTTATGTGTAAATTAGCCCATGTATAATAATACTTT 977
TTGCAACTATAGTCTGGCTTCTAATGATTTTTTTCTTACATTAGATATATTTTATTTTATTTAATTATATAACAGCGCT 1057
CACATACATGGTTTATTATTGTATTTCTCAAATGTTTAGATGTTAAGGTTTAAAAATTAATCTTTTTTATGCGA GT TCT 1135
ER SER
ATC TTA TTA CCA CCC CCG CCA CCA CCC CCG CCA CTT ATT TAG ATCTTTGAAGTACTTGCTATAGAA 1201
ILE LEU LEU PRO PRO PRO PRO PRO PRO PRO LEU ILE *
AAATATAACTGGGTTTGCTCAATCAGCCTATGTAATAATCCTTTGCCTAGTCAAGGGACAAGTCTTAATCTATCTGAAA 1281
CATCCATAGTTATATGTCAAATAACATATGATTGACTTGAAGCTTGTTTTATGATGAAAGGATTTATTATACCTTCAAT 1361
ATTCCAATTAGCTACATCTGTTTGTGTGAAGAATCACTTGTGCATGAATCATGATCGTTAATGGCTTTCCTACTTAATAG 1441
CTAGTAATTAAAAAACATGTTTACATCTATTTTTTGTGTTGCCATTTCCATTGCCACACATAATCA

```

Figure 3. Primary structure of the nodulin 20 gene, which corresponds to the N<sub>10,8</sub> cDNA clone. Figure legend as in Figure 2.

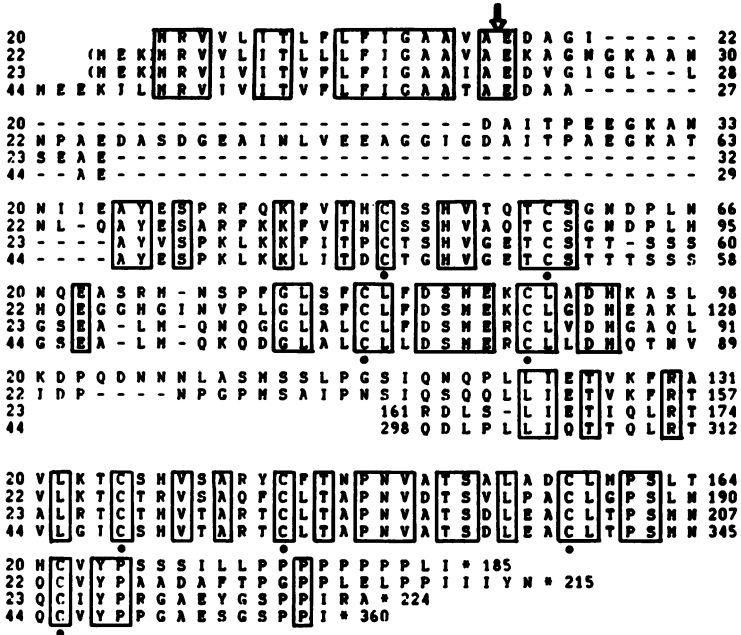


Figure 4. Alignment of the amino acid sequences of nodulins 20, 22, 23 and 44. Conserved residues are boxed, and conserved Cys residues are indicated by dots. The putative cleavage site in the signal peptide is indicated by an arrow. The amino acid residue number is shown on the right. A gap in the nodulin 23 and 44 sequences is introduced, since there is no homology to nodulins 20 and 22 in this region.

used here for comparison with the nodulin 20, 22 and 44 sequences. The derived amino acid sequences of all the genes are shown in Fig.4. In all cases the first ATG is assumed to be the initiator codon.

Three conserved domains are present in all four proteins: The N-terminus of the protein sequences is conserved and is typical of a hydrophobic signal peptide. According to von Heijne <sup>22</sup> the most likely cleavage site of the signal peptide sequence would be after the Ala residue indicated in Fig.4. The two other domains are each centered around four Cys residues.

The sequences between the conserved domains are degenerate due to several deletion, insertion and duplication events. Thus the nodulin 22 gene contains a duplication after the puta-

tive signal peptide sequence (nucleotides 116-192 versus 191-255) creating an extended amino acid sequence in this region, while the nodulin 44 gene contains several duplications in the central divergent region<sup>20</sup>. Finally, the amino acid sequence of the second exon is well conserved only between nodulins 23 and 44.

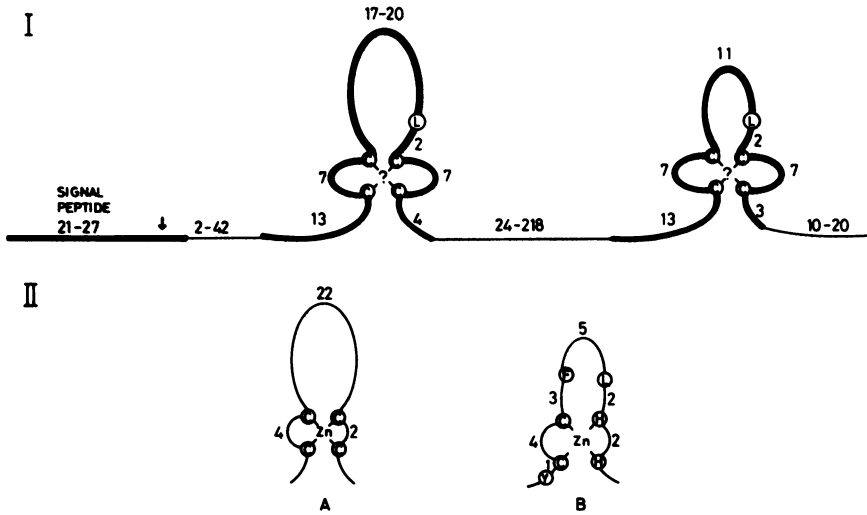
In the conserved regions the amino acid homology between nodulins 20 and 22 is 77%, while the homology between nodulins 23 and 44 is 85%. The amino acid homology between nodulins 20 or 22 versus nodulins 23 or 44 is about 58% in the conserved regions.

#### DISCUSSION

We have characterized a small nodule-specific soybean gene family. The conservation of the amino acid and DNA sequences despite numerous insertion/deletion events implies that the proteins encoded in this gene family evolved from a common ancestor.

The pbm is derived from the plasma membrane of the root cell and serves as the primary interface between the host and the endosymbiont. Some nodulins are integrated in the pbm to allow this membrane to meet the requirements imposed by symbiosis, and a mechanism must therefore exist, which allow specific targetting of these proteins into the pbm<sup>8</sup>. Nodulin 23 is a pbm protein<sup>9</sup>, while nodulin 44 is also located in the peribacteroid space<sup>21</sup>. Since nodulins 23 and 44 are homologous to nodulins 20 and 22 it is tempting to suggest that these proteins also are present in the pbm or in the peribacteroid space. The presence of a conserved putative signal peptide in all these proteins suggests that this sequence may be responsible for targetting them to the pbm or into the peribacteroid space. Apart from the putative signal peptide, two other domains are highly conserved. These domains each contain four Cys residues that are arranged in pairs with the same motif Cys - 7 amino acids - Cys. The domains share an internal homology indicating that the two conserved regions are the result of an ancient duplication of an element containing four Cys residues. The in-





**Figure 5.** I: Proposed schematic structure of two conserved Cys containing regions in nodulins 20, 22, 23 and 44. Heavy lines: conserved regions. Thin lines: divergent regions. Arabic numbers refer to the number of amino acid residues. The question marks indicate that the specific metal ion is unknown. Arrow: cleavage site in signal peptide. IIA: Schematic structure of a zinc binding domain present in *E.coli* aspartate carbamoyl transferase<sup>23,24</sup>. IIB: Schematic structure of a putative zinc binding domain present in transcription factor IIIa from *Xenopus*<sup>25</sup>. Similar domains are found in the Krüppel<sup>26</sup> and Serenidipity<sup>27,28</sup> products from *Drosophila* and ADRI from yeast<sup>29</sup>. For a further discussion of putative metal binding domains see Berg<sup>30</sup>.

ternal homology further implies that the two conserved domains are functionally analogous.

Recently it was found that several metal binding proteins such as aspartate carbamoyl transferase from *E.coli*<sup>23,24</sup> and the transcription factor IIIA from *Xenopus*<sup>25</sup> contain pairs of Cys or His residues that can be arranged in so-called 'finger regions' creating metal binding domains (Fig.5). Within these structures there are often other conserved amino acid residues at specific positions such as a Leu residue located two positions away from the third Cys/His residue. In the metal binding domains so far investigated the conserved arrangement is pairs of Cys/His - 2-5 amino acids - Cys/His sequences, while in the nodulins 20, 22, 23 and 44 the arrangement is pairs of Cys - 7

44	TAAATATGAA---TGCACTAGTAATTAGTTTAAATGATAAAAATATATTCTACTGTATTATTTTCTGTA	ACT	-169
23	TAAATATGAA---TGCACTAGTAATTAGTTTAAATGATAAAAATATATTCTA-----		-142
22	TAAATATGAAAATGGCAATAGTAATTAGTGTGATGATAAAAATATATTCTA-----TCTTTTATTATCT		-177
20	AAAATATGAATACTG---AGTAATTAGTCTAATGGTAAAATATATTCTA-----TTGTTTGTATCT		-172
44	ATATTCCATTTTAAATGTTTGAAAGACAGATATATTTCCATATATATCTCTTGGCAA-CTCCTCA-----G		-105
23	-----CAGATATATTTTCTCTTGGCAA-CTCGTGAGAAATG		-104
22	ATATTTTCAATTTTAAATGTTAGAAAAGCAAAATATAT---ATATATAGCTCTTGGAAAGCAAAAGCAGTCA-G		-112
20	GTATTCCATTTTAAATGTTAGAAAAGCAAAATGT-----ATATATAGCTCTTGGAAAGCAAAAGCAGTCA-G		-109
44	AATATATTATAAAGATGA---AAGGTCGTTATAATTTTTTTT---AGAATAAAATATTTATATACAGTTT		-42
23	AATATATTATAAAGATGA---AAGGTCGTTACAATTTTTTTT---AGAATAAAATATTTATATACAA-TT		-42
22	GATATATGCAAAAGATCA---AAGGTCGTAATAATTTTTTTTTTATAATAAAACATAGATGTTGAG-TT		-46
20	GATATATGCAAAAGATCGCGCAGGTCGTTATAAAATTTTTT---ATAATAAAGATATACCTTGAG-CT		-45
44	CCTAGATTTCTGTTAT-----AAAATTCACATATTGAATGAGTATAAAATACATGAGCACCCA-CC		17
23	CCTAGATTTCTGTTAT-----AAAATTCACATATTGTATGAGTATAAAATACATGAGCACACA-CC		17
22	CCTAGTTTCTCTTGTATATATATTGCAAAATTCACAC-----ATACAAGAGCAGCGATCC		10
20	CCTAGTTTATCTTGTATATATATTGCAAAATTCAC-----ATACATGAGCACCAAGCA		9
44	AA-ATTAGTCTCAAATTAAGTAAG-----AAAATGGA		48
23	AA-CTAGTCTCAAATTAAGTAAGGTCGTAATTATTACCGGTAGCTAAGT-AACCAAGTAATTAATGGA		85
22	AATAAAGTTCTCAAAA-GTGCAAA---CTAGTTTATCA-----CTAAGAGAACCA---ATTAATGGA		65
20	AATAACATTCTCAA-----		23
44	GGAGAAAAATATTAATGAGAGTGATAGTAATTACCGTATTCCCTATTTTATAGGTGCAGCAACTGCAGAAGAT		118
23	GAA-----AATGAGGTCATAGTAATTACTGTATTCCCTATTTATAGGTGCAGCAATGCAGAAGAT		146
22	GAA-----AATGAGAGTGGTACTAATTACTCTATTGTTGTTTATAGGTGCAGCAGTTGCAGAAAA		126
20	-----AATGAGAGTGGTATTAATTACTTTATTCCCTGTTTATAGGTGCAGCAGTTGCAGAAGC		81

Figure 6. Alignment of the 5' end sequences of the nodulin 20, 22, 23 and 44 genes. The 5' flanking sequence of the nodulin 23 gene was redetermined in our laboratory and contains few changes from the published sequence. The 5' flanking sequence of the nodulin 44 gene was sequenced in our laboratory (J.E. Jørgensen). TATA boxes and initiator ATGs are underlined. The cap sites are indicated by arrows. The first cap site was chosen as position +1. The cap site of the nodulin 44 gene is not known. The short boxed sequences are present in several different nodule specific plant genes.

amino acids - Cys sequences. Nevertheless it is possible to arrange the nodulin Cys pairs into structures, which are highly analogous to the 'finger regions' present in metal binding proteins (Fig.5). It is noteworthy that the conserved Leu residue is also present at the conserved position in the proposed structure. We therefore propose that the two conserved domains in the nodulins 20, 22, 23 and 44 are metal binding domains and that binding of metal ions is important for the function of these nodulins.

The nodulin 20, 22, 23 and 44 genes are evolutionary related and activated about the same time during nodule development. It is possible to align the 5' regions up to about 220 bp 5' to the cap-site, after which the sequences begin to diverge (Fig.6). Several nodulin genes including the four genes

described here are activated about the same time during nodule development. The activation of these genes coincides with a dramatic increase in the transcription of the Lb genes<sup>14</sup>. This may imply that the same mechanism is responsible for both events. It is therefore reasonable to assume that the 5' flanking regions of the nodulin and Lb genes share a common regulatory DNA sequence. The sequence 5' AAAGAT 3' is present about position -95 in these genes and is also present in the Lb genes about position -130 (unpublished observation). In nodulin 24<sup>7</sup> AAAGAT is found at position -193. The sequence 5' CTCTT 3' is present in nodulins 20, 22, 23 and 44 about position -130, while in the Lb genes it is present about position -120 and about position -80 in the inverted form. In nodulin 24 it is present at position -153 and in the inverted form at position -77. Finally both sequences are present in appropriate positions in the 5' flanking region of the Parasponia Lb gene<sup>31</sup>. The conservation of these sequences in the 5' flanking regions of the nodulin and Lb genes might suggest that they are involved in the activation of the genes. However, the sequences consist of only 5-6 bases each, and the statistical significance of the occurrence of these sequences in such genes is therefore uncertain until their presence have been established in appropriate positions in other nodulin genes.

#### ACKNOWLEDGMENTS

We would like to thank Drs. R. Goldberg and R. Fischer, UCLA, for providing the limited EcoRI and AluI/HaeIII soybean libraries. We also thank Dr. J. Key and Agrigenetics Corporation for providing the limited Sau3A soybean library. We thank Dr. K. Gausing for valuable discussions. This investigation was supported financially by the Danish State Biomolecular Engineering Programme and EEC contract BAP-0173-DK.

<sup>1</sup> Present address: Department of Biochemistry B  
Panum Institute, University of Copenhagen  
Blegdamsvej 3C  
2200 Copenhagen N, Denmark

#### REFERENCES

1. Legocki, R.P. and Verma, D.P.S. (1980) Cell 20, 153-163.
2. Bergman, H., Preddie, E., Verma, D.P.S. (1983) EMBO J. 2, 2333-2339.

3. Christensen, T.M.I.E. and Jochimsen, B.U. (1983) *Plant Physiol.* 72, 56-59.
4. Gebhardt, C., Oliver, J.E., Förde, B.G., Saarelainen, R., and Miflin, B.J. (1986) *EMBO J.* 5, 1429-1435.
5. Morell, M. and Copeland, L. (1985) *Plant Physiol.* 78, 149-154.
6. Thummler, F. and Verma, D.P.S. Third International Symposium on the Molecular Genetics of Plant-Microbe Interactions (1986) McGill University, Montreal, Canada, 182.
7. Verma, D.P.S., Kazazian, V., Zogbi, V., and Bal, A.K. (1978) *J.Cell.Biol.* 78, 919-936.
8. Fortin, M.G., Zelechowska, and Verma, D.P.S. (1985) *EMBO J.* 4, 3041-3046.
9. Fortin, M.G. and Verma, D.P.S. (1986) Third International Symposium on the Molecular Genetics of Plant-Microbe Interaction, McGill University, Montreal, Canada, 157.
10. Maniatis, T., Fritsch, E.F., and Sambrook, J. (1982) *Molecular Cloning. A Laboratory Manual*, Cold Spring Harbor Laboratory Press, NY.
11. Marcker, K.A., Gausing, K., Jochimsen, B., Jørgensen, P., Paludan, K., and Truelsen, E. (1981) In Banopoulos, N.J. (ed.,) *Genetic Engineering in the Plant Sciences*, Praeger Publishers 63-71.
12. Dretzen, G., Bellard, M., Sassone, Corsi, P. and Chambon, P. (1981) *Anal.Biochem.* 112, 295-298.
13. Bojsen, K., Abildsten, D., Jensen, E.Ø., and Marcker, K.A. (1984) *EMBO J.* 3, 1691-1695.
14. Marcker, A., Lund, M., Jensen, E.Ø., and Marcker, K.A. (1984) *EMBO J.* 3, 1691-1695.
15. Sanger, R., Coulson, A.R., Barrell, B.Q., Smith, A.J.H., and Roe, B.A. (1980) *J.Mol.Biol.* 143, 161-178.
16. Boel, E., Vuust, J., Norris, F., Norris, K., Wind, A., Rehfeld, J.F., and Marcker, K.A. (1983) *Proc.Natl.Acad.Sci. USA* 80, 2866-2869.
17. Maxam, A.M. and Gilbert, W. (1977) *Proc.Natl.Acad.Sci. USA* 74, 560-564.
18. Van Kammen, A. (1983) *Advances in Nitrogen Fixation Research*, Veeger, C. and Newton, W.E. Eds., Nijhoff/Junc Publ., pp.587-588.
19. Mauro, P.V., Nguyen, T., Katinakis, P., and Verma, D.P.S. (1985) *Nucl.Acids Res.* 13, 239-249.
20. Sengupta-Gopalan, C., Pitas, J.W., Thompson, D.V., and Hoffmann, L.M. (1986) *Mol.Gen.Genet.* 203, 410-420.
21. Jacobs, F.A., Purohit, F.A., Zhang, M., Fortin, M., and Verma, D.P.S. Third International Symposium on the Molecular Genetics of Plant-Microbe Interactions McGill University, Montreal, Canada, 157.
22. Von Heijne, G. (1983) *Eur.J.Biochem.* 133, 17-21.
23. Honzatko, R.B., Crawford, J.L., Monaco, H.L., Ladner, J.E., Ewards, B.F.P., Evans, D.R., Warren, S.G., Wiley, D.C., Ladner, R.C., and Lipscomb, W.N. (1982) *J.Mol.Biol.* 160, 219-263.
24. Weber, K. (1968) *Nature* 218, 1116-1119.
25. Miller, J. McLachlan, A.D., and Klug, A. (1985) *EMBO J.* 4, 1609-1614.

26. Rosenberg, U.B., Schroeder, C., Presiss, A., Kienlin, A., Côté, S., Riede, I., and Jäckle, H. (1986) *Nature* 319, 336-339.
27. Vincent, A. (1986) *Nucl.Acids Res.* 14, 4385-4391.
28. Vincent, A., Colot, H.V., and Rosbach, M. (1985) *J.Mol. Biol.* 186, 149-166.
29. Hartshorne, T.A., Blumberg, H., and Young, E.T. (1986) *Nature* 320, 283-287.
30. Berg, J.M. (1986) *Science* 232, 485-487.
31. Landsmann, J., Dennis, E.S., Higgins, T.J.V., Appleby, C.A., Kortt, A.A. and Peacock, W.J. (1986) *Nature* 324, 166-168.