Proc. Natl. Acad. Sci. USA
Vol. 82, pp. 4157–4161, June 1985
Cell Biology

Nodulin-24 gene of soybean codes for a peptide of the peribacteroid membrane and was generated by tandem duplication of a sequence resembling an insertion element

(symbiotic nitrogen fixation/DNA sequence/hybrid-selection/*in vitro* translation and processing/immunoprecipitation)

PANAGIOTIS KATINAKIS AND DESH PAL S. VERMA*

Plant Molecular Biology Laboratory, Department of Biology, McGill University, Montreal, PQ, Canada H3A 1B1

Communicated by Oliver E. Nelson, Jr., February 25, 1985

ABSTRACT **A nodulin gene coding for a polypeptide with an apparent M_r of 24,000 (nodulin-24) was isolated from soybean (*Glycine max*). DNA sequence analysis of this gene revealed that its coding capacity is for a polypeptide of only M_r 15,100 and is interrupted by four introns. The three middle exons and their flanking segments appear to have been generated by duplications of a unit resembling an insertion sequence. This unit is bounded by a 12-base-pair inverted repeat and encompasses the 54-base-pair exon corresponding to each of three central hydrophobic domains of the protein, nodulin-24. The resulting repeated hydrophobic structure of this protein may be responsible for an apparent increase in M_r from 15,100 to 24,000. *In vitro* translation and immunological studies suggest that nodulin-24 is a precursor and is processed cotranslationally into a M_r 20,000 polypeptide. This polypeptide is a component of the membrane envelope enclosing the bacteroids (peribacteroid membrane) synthesized during symbiosis with *Rhizobium*. The low degree (<6%) of sequence divergence among the repeated units suggests that this gene has been generated recently during the evolution of symbiotic nitrogen fixation in soybean.**

Nodulins are a group of plant proteins that are induced specifically during the development of root nodules in legumes following infection with *Rhizobium*, leading to symbiotic nitrogen fixation (1). The best known of nodule-specific plant gene products are abundant proteins such as leghemoglobins and nodulin-35, nodule uricase (2, 3). In addition, a nodule-specific glutamine synthetase has been shown to exist in some species—e.g., *Phaseolus* (4). The functions of other nodulins remain unknown (see ref. 3).

We have recently identified a number of soybean nodule-specific cDNA sequences that hybrid-select mRNAs for nodulins 23, 24, 27, 44, and 100 (5, 6). Nuclear genes encoding these polypeptides have been isolated. To understand the molecular processes involved in symbiosis and to elucidate the function of nodule-specific host genes, we studied the structure and expression of the soybean nodulin gene encoding nodulin-24. Induction of this gene occurs prior to that of leghemoglobins and other nodulins and is independent of the commencement of nitrogen fixation activity in nodules (6).

During the differentiation of nodules, one of the major changes that occurs inside the infected cells is the formation of a subcellular membrane compartment in which bacteria reside (7). We have earlier demonstrated (8) that the membrane envelope enclosing the bacteroids (peribacteroid membrane) originates from the plasma membrane of the host but is modified during endosymbiosis. New polypeptides integrated into this membrane may have specific functional and/or structural roles to support the demands of the endosymbiont. Some of the nodulins may fulfill these roles. We report here that the nodulin-24 gene of soybean indeed codes for a polypeptide that appears to be part of the peribacteroid membrane and suggest a possible molecular mechanism by which this gene may have been generated.

MATERIALS AND METHODS

Plant Tissue. Soybean seeds (*Glycine max* cv. Prize) were purchased from Strayer Farm (Hudson, IA). *Glycine soja* seeds were kindly provided by Niels Nielsen (Purdue University). Plants were grown as described (9). Nodules formed as a result of inoculation with *Rhizobium japonicum* strain 61A76 were harvested 3 weeks after infection and stored under liquid nitrogen until used.

Isolation of Nucleic Acids. Poly(A)$^+$ RNA was isolated from total polysomes of 21-day nodules as described (9). Phage DNA was isolated as described (10). Plasmid DNAs were purified on CsCl/ethidium bromide gradients. Genomic DNA from *Glycine max* embryonic axes and leaves of *Glycine soja* was isolated as described (11).

Isolation of Gm N-24 from a Soybean Genomic Library. About 8×10^5 recombinant bacteriophages were screened by the method of Benton and Davis (12) from an *Alu* I-*Hae* III partial genomic library of soybean (13) by using the ^{32}P-labeled insert from pNod60, a cDNA clone for nodulin-24 (5), as a probe. The genomic clone containing nodulin-24 sequence is referred to as Gm N-24. Two full-length nodulin-24 cDNA clones, pNod18 and pNod20, were isolated by rescreening the nodule-specific cDNA library (5) with pNod60.

Southern Blotting and Hybridization. DNA digested with restriction endonucleases was electrophoresed through agarose gels and transferred (14) to GeneScreen (New England Nuclear). Pretreatment, hybridization, and washing of filters were performed as described (15). Insert DNAs used as probes were isolated from recombinant plasmids and made radioactive by nick-translation (16) to a specific activity of 0.5×10^8–1.0×10^8 cpm/μg of DNA, using [^{32}P]dCTP (specific activity 3000 Ci/mmol; 1 Ci = 37 GBq; Amersham).

Subcloning and DNA Sequencing. A *Taq* I fragment containing one of the repeat units (see Figs. 1 and 4) was subcloned (pR1) into pUR222 (Boehringer Mannheim). The 8-kilobase-pair (kb) *Bam*HI restriction fragment from Gm N-24 (Fig. 1) was subcloned into pBR322. The resultant recombinant plasmid, pBGm N-24, and the two full-length cDNAs (pNod18 and pNod60) were mapped with restriction endonucleases. Appropriate DNA fragments were electroeluted, subcloned in M13, mp8 or mp9 vectors (17), and

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: kb, kilobase pair(s); bp, base pair(s).

*To whom reprint requests should be addressed at: Department of Biology, McGill University, 1205 Docteur Penfield Avenue, Montreal, PQ, Canada H3A 1B1.

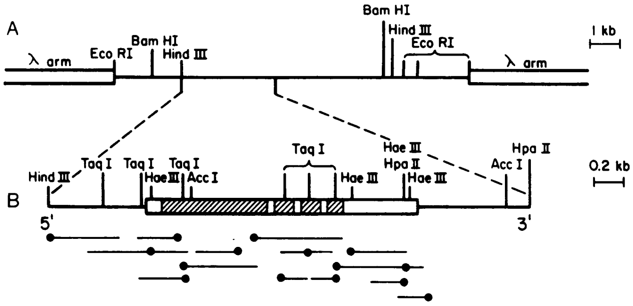

FIG. 1. Restriction map of Gm N-24 and sequencing strategy of nodulin-24 gene. (*A*) Recombinant phage (Gm N-24) DNA was mapped with *Eco*RI, *Bam*HI, and *Hin*dIII restriction enzymes. (*B*) The 8-kb *Bam*HI fragment was subcloned into pBR322, the *Hin*dIII–*Hpa* II fragment was further mapped, and appropriate fragments were electroeluted, cloned into M13 derivatives, and sequenced. A dot on one end of each fragment indicates the beginning of the sequence. Coding regions are represented by open boxes and introns are represented by hatched areas.

propagated in the host JM101, and single-stranded DNAs were purified. Sequencing reactions were performed by the dideoxy chain-termination method (18), using a 15-base-pair (bp) single-stranded primer (P-L Biochemicals). Gel electrophoresis was performed as described (19). Computer-assisted sequence analysis was accomplished with the Nuc:Aln program (20). Dot-matrix analysis was done on an IBM-PC using a program from International Biotechnologies (New Haven, CT).

Hybrid-Selection of mRNA, Translation, Processing, and Immunoprecipitation. Filters containing 50 μg of purified pNod18 were prepared and hybridized with 10 μg of nodule polysomal poly(A)$^+$ RNA as described (5). Hybrid-released mRNA was translated in a rabbit reticulocyte lysate containing [^{35}S]methionine (specific activity, 1150 Ci/mmol) in the presence or absence of microsomal membranes obtained from New England Nuclear. *In vitro* translation products were immunoprecipitated as described (5) with antiserum against peribacteroid membrane prepared by modification of the method of Verma *et al.* (8).

RESULTS

The Nodulin-24 Gene Has the Coding Capacity for a Polypeptide of Only M_r 15,100. Hybrid-released translation of nodulin-24 mRNA using a cDNA clone (pNod60) yielded a polypeptide with apparent M_r, in NaDodSO$_4$/polyacrylamide gels, of about 24,000 (5). However, sequence analysis of two full-length cDNA clones (pNod18 and pNod20) indicated that these clones only have the coding capacity for a polypeptide of 147 amino acids (M_r ≈15,100). The two cDNA sequences are almost identical throughout the coding region except for two base-pair changes and a three-nucleotide deletion in the pNod18 (see Fig. 2). pNod20 contains an extra 204 bp at the 3' untranslated region. This is consistent with the size of two nodulin-24 transcripts observed (6) and suggests the presence of two genes. Two potential poly(A) addition signals (21) are located, corresponding to 10 bp and 16 bp upstream from the site of polyadenylylation on pNod20 and pNod18, respectively (see Fig. 2). Since the genomic sequence (see below) is identical to the cDNA clone, pNod20, we believe that the total coding capacity in this gene is only for a polypeptide of M_r 15,100. The difference in apparent molecular weight and the actual size of this nodulin appears to be due to some unusual features (see below) in this polypeptide.

Structure of the Nodulin-24 Gene. Hybridization of a genomic clone (Gm N-24) containing a 12-kb insert, with a full-length nodulin-24 cDNA (pNod20), showed (Fig. 1*A*) that the coding region of this gene is present on the *Bam*HI fragment that was subcloned into pBR322. A detailed organizational map of the region containing the coding sequence (2.4 kb) is shown in Fig. 1*B*, including the strategy for sequencing. The nucleotide sequences are depicted in Fig. 2. Comparison of the sequence of the genomic clone with that of the cDNA (pNod20), using the Nuc:Aln program of Wilbur and Lipman (20), indicated that the largest possible open reading frame encodes 147 amino acids and showed that this nodulin gene contains four introns. Three of the introns are bounded by the consensus sequence 5' G-T.../G-A 3' found in most functional eukaryotic genes (22). However, the donor site of the fourth intron contains the sequence 5' G-C.../G-A 3'. This type of 5' splice junction has been observed in a number of eukaryotic genes (23, 24) and seems to be utilized efficiently *in vivo* (25). There are present on the 5' and 3' flanking regions the consensus sequence's "TATA" box and poly(A) addition signal of functional eukaryotic genes. In the absence of the protein sequence, we have assigned, among the two possible initiator methionine codons, the first as an initiator.

S1 nuclease mapping using a *Hin*dIII/*Hae* III fragment (see Fig. 1*B*) and poly(A)$^+$ RNA from 3-week nodules (data not shown) revealed two potential transcription start sites, one of which (upstream and marked as base 1 in Fig. 2) is more pronounced. The second transcript may be due to another nodulin-24 gene or may represent dual promoters. The latter is consistent with two potential TATA boxes in this gene. Detailed examination of the nucleotide sequence of this gene revealed several interesting features: (*i*) it contains three almost identical exons (exons 2, 3, and 4), (*ii*) the intervening sequences flanking these exons are conserved, and (*iii*) introns 2 and 3 are almost identical. Dot-matrix analysis (26) of Gm N-24 and pNod20 revealed (Fig. 3 *A* and *B*) the existence of a number of direct repeats in both the cDNA and genomic sequences. Three of the repeats in Gm N-24 are 180–190 nucleotides long and are arranged in tandem. Each of the three repeating units in genomic DNA consists of an almost identical exon (found as a tandem repeat in the cDNA) plus 5' and 3' flanking intron sequences (see Fig. 4).

Possible Origin of the Repeat Units in the Nodulin-24 Gene. Comparison of the sequence of the three repeats (R1, R2, and R3) shows high homology (R1/R2, 98%; R1/R3, 96%; R2/R3, 94%). Since the exon sequences encompassed in each repeat are almost identical, the observed differences are primarily due to the intron sequences. Analysis of the repeat sequence R1 (Fig. 4*A*) suggested that it has features of an insertion element (29). A 12-bp inverted sequence permitting a hairpin structure exists on each 5' and 3' end of R1 and R3 (short arrow) but is not duplicated or may have been eliminated in R2. This 12-bp sequence in R3 is flanked at the 5' end by an imperfect direct repeat (Fig. 4 *A* and *B*, underlined by arrows) present at the 3' end of R1 (position 1265, Fig. 2). Thus, the entire structure (R1, R2, and R3) also has features of an insertion element (Fig. 4*B*).

The unusual intron–exon arrangement in the nodulin-24 gene raised the question whether this structure (R1, R2, and R3) is actually in the soybean genome or was created during subcloning. *Hae* III restriction enzyme cuts outside of the three tandemly repeated units. If this repeat structure is present in the genome, then the *Hae* III fragment should appear as a single hybridizing band of 1.3 kb in the genomic DNA. As shown in Fig. 5, using a subcloned repeat (pR1), only one band of the expected size was observed. The notable insertion element features of this structure suggested to us that it might also be present in other locations in the soybean genome. Fig. 5 (lane 3) shows that sequences related to this repeat unit (R1) are present in a few copies in *Glycine max* genome. A search of the genome of *Glycine soja*, the closest ancestral relative of modern soybean (*Glycine max*), revealed (Fig. 5, lane 4) some common and some unique location(s) of

```
-60       -50       -40       -30       -20       -10       1
CTACTCCAACTCCTTTATATAGAGTATATATTCCCACAAATTTTCTCATCTTTTGTTACTAAACA

  10       20        30        40        50          60
           ▼
AACTCGATCTGTTGTAATTTATTTAGTACGTATTGAAAA ATG GGT TCA AAG ATG GCT ATA
                                        Met Gly Ser Lys Met Ala Ile

  70           80           90             100          110
CTG ATC CTA GGC CTG TTG GCC ATG CTC CTT TTG ATC ACC TCA GAA GTG GCA
Leu Ile Leu Gly Leu Leu Ala Met Leu Leu Leu Ile Thr Ser Glu Val Ala

  120         130          140       150       160       170
GCC AGG AAT TTA AAA GAG G/GCAAGTTAATTATAATGTTATATATCATCTTACCTTATATGG
Ala Arg Asn Leu Lys Glu

 180       190       200       210       220       230       240
TTCTCATTTCAAATAAGAGAATATTAATAAAGGCTTTAAAAACACTAGCTAATTTTTAAAAAAAAGTA

  250       260       270       280       290       300       310
CTATACTTCTAAAATATTTTTGTTGTAACCATTTTATAATTTTTTATCGACTTAAATATTTTCTCGCC

    320       330       340       350       360       370       380
CCTGCAATTATGTGTTTTTGTATACTTTTTATCCTTGCACTTTTCCTAATAATCCTTGTAAAATTCTC

     390       400       410       420       430       440
TTTTTTATGGTTTTGGACTTATAATTTTATTTGTTTAGTTCCTGTAACATTTTTTATTTTGTCGTTGC

450       460       470       480       490       500       510
AATATTTGAATAATTTGCTTTAGTTTCAGTCTCTATTATATTTTTTTTATAGAGGACTAATTCAGAAT

 520       530       540       550       560       570       580
AAAGAAAAATATTACAGGTCCTAAGAAATTAATAACAAACAAAACCGACAAAAATAAAGGTTTCATAA

  590       600       610       620       630       640       650
AGAGCATTTCAAGAAAAAAAAAAATCAGCAATAAATAAAAAAAAGTCCTAATAACAAGAGTAGTATTT

   660       670       680       690       700       710       720
AAACCCATTTTTTTGGTGAATTTAAAGGGATCTTAATTATACGCCTAATATATATAATGGGTGTGCAA

     730       740       750       760       770       780
AATTCATTTTTCTATAGTAAATGGTTCACGTGTTAAGGGTTAATGTGTTCCAG/CA GGT GAG GCT
                                                        Ala Gly Glu Ala

   790         800           810          820           830
GTT CAA GAG ACA AAT GAA GTG GCT GAT GCC AAA TTA GTT GCT G/GTGGTGTTTT
Val Gln Glu Thr Asn Glu Val Ala Asp Ala Lys Leu Val Ala

840       850       860       870       880       890       900
CTTTCTCTTCCCTCTCCTCCCTCAAAATCATATACACTCTAATTAATGGGTGTGCAAAATTAATTTTT

 910       920       930       940       950       960
CGATATATATTAATGTTTGACGTGTTAATATTTAATTAATGTGTTCCAG/CA GGT GAG GCT GTT
                                                    Ala Gly Glu Ala Val

970          980          990          1000           1010       1020
CAA GAG ACA AAT GAA GTG GCT GAT ACC AAA TTA GTT GGT G/GTGGTGTTTTCTTT
Gln Glu Thr Asn Glu Val Ala Asp Thr Lys Leu Val Gly

     1030      1040      1050      1060      1070      1080
CTCTTCCTCTCCTCCCTCACAAATCATATACACCTAATTAATGGTGTGCAAAATTAATTTTTCGATAT

        1100      1110      1120      1130          1140          1150
ATAGTAAATGTTTGACGTGTTAATATTTAATTAATGTGTTCCAG/ CA GGT GAG GCT GTT CAA
                                                Ala Gly Glu Ala Val Gln

        1160         1170          1180           1190      1200
GAG ACA AAT GAA GTG GCT GAT ACC AAA TTA GTT GGT G/GTGGTGTTTTCTTTCTCT
Glu Thr Asn Glu Val Ala Asp Thr Lys Leu Val Gly

  1210      1220      1230      1240      1250      1260      1270
TCCTCTCCTCCCTCAAAAATCATATACACCTAATTAATGGTGTGCAAAATTAATTTTTCGATATATAG

    1280      1290      1300      1310        1320          1330
TAAATGTTTGACGTGTTAATACTTAATTAATGTGTTTGCAG/CA GGT GGG GTT GTT AAA CAG
                                             Ala Gly Gly Val Val Lys Gln

    1340         1350         1360         1370         1380
AGA AAT AAA GTG GGT TAT GGC AAA TTA GTT GGT GTT GGT GGT TAT GAT TAT
Arg Asn Lys Val Gly Tyr Gly Lys Leu Val Gly Val Gly Gly Tyr Asp Tyr

  1390           1400          1410         1420          1430
GGG AAT TGG AAT GGT GGC CAA CGC TCT CCA TAT GGA ACG GGA GCT ATT TGC
Gly Asn Trp Asn Gly Gly Gln Arg Ser Pro Tyr Gly Thr Gly Ala Ile Cys

 1440         1450          1460          1470          1480
ATG AGA GGC TGT TGT TTT CCA TCC TCG TTG GGA GGT TCG GTA AGT TGC TGC
Met Arg Gly Cys Cys Phe Pro Ser Ser Leu Gly Gly Ser Val Ser Cys Cys

1490         1500           1510       1520       1530       1540
CCG CAT GAA TGG CAG TAA CCGCCCTACATGGCATGGCTGCTGCTTATTTCATGAATAAAATC
Pro His Glu Trp Gln  *

          1560       1570      1580      1590      1600      1610
TCTTGTTGTGAGAGGCTTGTCTTTTCTCAAGTATAAATAATGTAGAATAATACGTACGTCACTGTGCC

.620       1630      1640      1650      1660      1670      1680
TGTGTGACAAGTTTAAGAGCAAATTGAAATAATCTTCTGTGCAAATATTTTGTTTTAAAATCGCGACT
                                          ●
  1690      1700      1710      1720      1730      1740      1750
TTTGCTTCTATTATTGTTATTATTATTATTATAAAACATGCAGTTCCTTATGTAACTGGGGTTAAGGG

      1830      1840      1850      1860      1870      1880
TTTGATTTATCAATACGTAGGTAATAAGAATTATATACAACAACGACTTCAGCGCATGAAGAAGTA
                                            ●
```

FIG. 2. Nucleotide sequence of soybean nodulin-24 gene. Putative transcriptional control regions [TATA box(es), poly(A) addition signals] are underlined. The deduced amino acid sequence corresponding to that of cDNA (pNod20) is shown beneath the nucleotide sequence and the three almost identical exons are underlined. The potential poly(A) addition signals $^{A}_{T}$AATAA on the two cDNA clones (pNod18 and pNod20) are present, corresponding to positions 1646 and 1844, and 3′ ends of the respective transcripts are marked by dots beneath the bases. Nod18 had 3 bp deleted at position 777 and 2 bp altered at positions 826 (thymine) and 1009 (cytosine), resulting in codon change (valine and proline, respectively). The nucleotide sequence is numbered from the major transcription start site (determined by S1 nuclease mapping). A minor promoter (denoted by an arrowhead) is also shown. *, Termination codon.

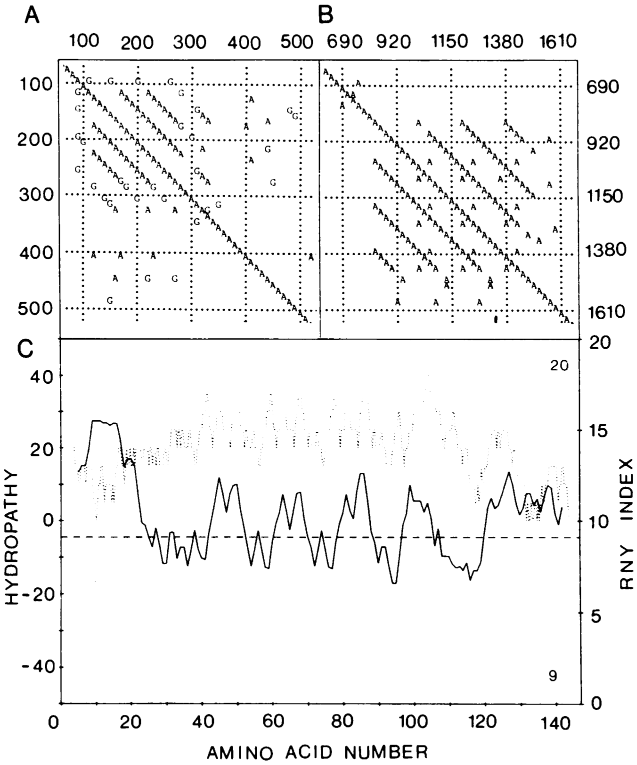

FIG. 3. Dot-matrix sequence analysis (*A* and *B*) and hydropathy and RNY (purine, N, pyrimidine) analysis of the sequence encoded by pNod20 (*C*). Comparison of part of the nucleotide sequence (representing coding region) of cDNA (pNod20) with itself (*A*) and the region of the nodulin-24 gene containing three repeats (*B*). The window size in *A* is 9 nucleotides and in *B* it is 19 nucleotides. Symbol "A" represents 90% and "G" represents 60% sequence homology. Hydropathy analysis (27) was used to predict hydrophobic (positive values) and hydrophilic stretches (negative values) at a span set of 9 (solid line) (*C*). The score of RNY nucleotide (dotted line) was determined in a span of 20 nucleotides, using a computer program based on Shepherd's method (28).

this sequence. It should be noted that the extra hybridization bands in lanes 3 and 4 are due to intron sequences of the R1 unit since hybridization with the cDNA (pNod20), which contains the three tandemly repeating exons, showed only one fragment (Fig. 5, lane 1). These results suggest that sequences related to the intron region of the repeat (R1) are also present elsewhere in the genome. Furthermore, the size of the *Hae* III fragment (lane 2) differs between *Glycine max* and *Glycine soja* (data not shown).

The Nodulin-24 Gene Codes for a Membrane Protein. Analysis of the derived amino acid sequence of the putative nodulin-24 protein has revealed features that could help in assigning a role and location to this nodulin gene product. The hydropathy (27) plot indicates (Fig. 3*C*) that nodulin-24 may be a transmembrane protein. The 54-bp exons correspond to each of the 18 amino acid hydrophobic domains. The calculated molecular weight of nodulin-24 is in contrast with the apparent molecular weight as measured on $NaDodSO_4$/polyacrylamide gels (5). This discrepancy could be explained by anomalous binding of $NaDodSO_4$ to the repeated hydrophobic domains, giving rise to an uneven charge distribution. Similar aberrant migrations on $NaDodSO_4$/polyacrylamide gels have been reported for a number of proteins containing repeated regions (30–32).

To test whether nodulin-24 is a nodule-specific membrane peptide, the hybrid-selected mRNA was translated *in vitro* by using rabbit reticulocyte lysate with and without microsomal membranes (33). Data presented in Fig. 6 show that nodulin-24 is cotranslationally processed into a polypeptide of appar-

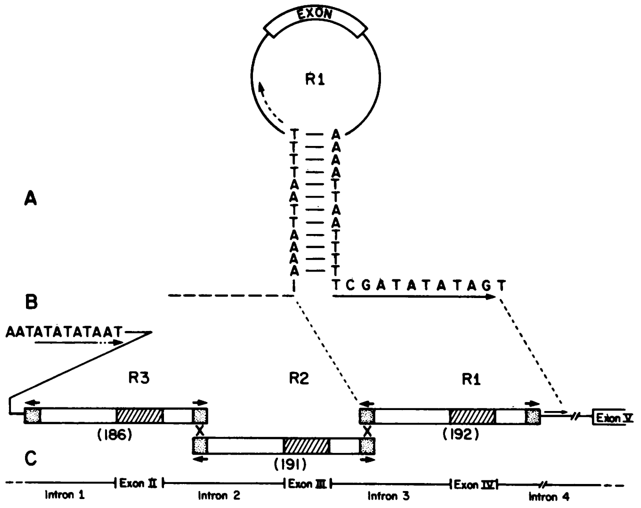

FIG. 4. Schematic organization of the repeats in nodulin-24 gene. (*A*) Repeat (R1) as defined in the text is shown as a stem–loop structure; the observed and the putative (lost during a possible unequal crossing-over event) direct repeat is denoted by the arrow and dashes, respectively. (*B*) Diagrammatic representation of the possible steps involved in the generation of the tandem array of these repeat(s). Open boxes indicate introns and hatched boxes indicate exons. Stippled boxes denote the 12-bp hairpin structure shown in *A* and thick arrows over them show orientation. The diagonal crosses demark the location of recombinational crossing-over events required to restore the continuity of the three repeats as shown in *C*. The imperfect direct repeats described in the text are marked by thin arrows in *A* and *B*.

ent M_r 20,000. Both precursor and the product (Fig. 6, lanes 5 and 6) are immunoreactive with antibodies against membrane envelope enclosing the bacteroids (peribacteroid membrane). This, along with the hydropathy analysis, suggests that this nodulin is a component of the peribacteroid membrane. Furthermore, a nodule-specific protein crossreactive with the antibody against a chemically synthesized peptide, representing the hydrophobic repeated domain of nodulin-24, has been observed in the peribacteroid membrane fraction and appears to be modified post-translationally (unpublished data).

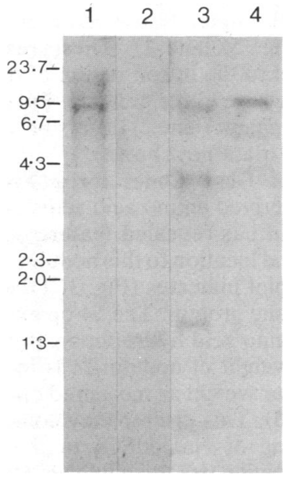

FIG. 5. Southern blot analysis of soybean genomic DNA with pNod20 and pR1 as probes. DNAs (10 μg each) isolated from *Glycine max* (lanes 1–3) and *Glycine soja* (lane 4) were digested with *Eco*RI (lanes 1, 3, and 4) and *Eco*RI + *Hae* III (lane 2), electrophoresed in 1% (wt/vol) agarose gels, transferred to GeneScreen, and hybridized with ^{32}P-labeled pNod20 (lane 1) or pR1 inserts (lanes 2–4). Bacteriophage λ *Hin*dIII and pBR329 *Hin*fI fragments were used as molecular weight markers. Sizes are shown in kb.

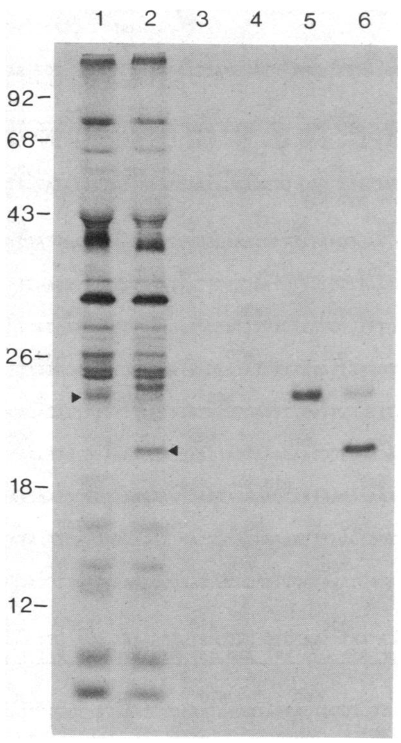

FIG. 6. Processing and immunoprecipitation of *in vitro* translation product of nodulin-24 mRNA. RNAs were translated in a rabbit reticulocyte lysate containing 100 μCi each of [^{35}S]methionine in the presence or the absence of microsomal membranes. Translation products were allowed to react with antiserum prepared against peribacteroid membrane (unpublished data), precipitated with Sepharose-protein A, and electrophoresed in a NaDodSO$_4$/15% polyacrylamide gel. Shown are translation products of nodule poly(A)$^+$ RNA in the presence (lane 2) or the absence (lane 1) of microsomal membranes (note the positions of nodulin-24 and its processed product, arrowheads). Lanes 3 and 4 are the product of RNA hybrid-selected by pBR322 (controls without and with added membranes). The translation products, using hybrid-selected nodulin-24 mRNA, are shown in lanes 5 and 6 without and with added membranes, respectively. Protein markers are shown as $M_r \times 10^{-3}$.

DISCUSSION

The Nodulin-24 Gene Encodes a Polypeptide of the Peribacteroid Membrane. We have demonstrated that one of the nodulin genes of soybean codes for a polypeptide of M_r 15,100 that migrates in NaDodSO$_4$/polyacrylamide gels with an apparent M_r of 24,000. This polypeptide is cotranslationally processed into a product of apparent M_r 20,000. The features that are responsible for an apparent increase in its molecular weight reside in the processed part of the molecule, and the three repeated hydrophobic domains may be responsible for it. This peptide appears to be an integral part of the peribacteroid membrane synthesized during endosymbiosis. This membrane compartment is essential for effective symbiotic nitrogen fixation (7, 34). Since this nodulin was initially identified by using antisera against total soluble proteins (5), it suggests that some of this protein may be present in the cytoplasm (free or in the form of vesicles) to account for antigenic response in this fraction.

Generation of the Nodulin-24 Gene. Nucleotide sequence analysis of the nodulin-24 gene suggested a novel aspect of gene generation in eukaryotic cells, possibly via duplication of an inserted sequence (such as R1, in Fig. 4) containing an amino acid domain. In the case of nodulin-24, this domain confers a hydrophobic character that may have provided selective advantage for this product to be a part of the peribacteroid membrane. The RNY index (Fig. 3*C*), an indicator of evolutionary conservation of coding sequence (28), suggests that this domain has an ancient origin but may have recently moved into its present location to generate the nodulin-24 gene.

The data presented here suggest that the primordial repeat unit, the intron region of which is also present in other locations of the genome (both in *Glycine max* and *Glycine soja*, see Fig. 5), acquired the exon domain before moving into the nodulin-24 loci. Intron sequences surrounding the exon have features that are reminiscent of an insertion event (i.e., 3-bp inverted repeat followed by a 4-bp direct repeat). It could be argued that the duplication of the postulated insertion element "R1" occurred before or after insertion in this gene. As shown in Fig. 4*B*, the entire array of R1, R2, and R3 can be folded to construct a formal stem-loop structure that is flanked by an imperfect direct repeat, suggesting that duplication occurred before insertion. However, the presence of inverted repeats on each end of R1 and R3 are indicative of duplication at the target site following insertion (35). In either event, the observed final structure would be the same. We cannot explain the presence of a direct repeat (similar to the one on the 3′ end of R1) inside the "loop" structure (Fig. 4*A*, curved-dashed arrow) with this model.

Duplication of exons has been proposed as a mechanism for the generation of human preproglucagon gene (36). Evidence for internal duplication has also been found for a number of eukaryotic genes (37, 38). Recently, in-frame insertions have been shown to generate strain-specific protein size polymorphism (39). However, in the case of nodulin-24, this duplication involves both intron and exon regions.

The intron/exon of the nodulin-24 gene complies with the general concept that the exons mark the boundaries of structural or functional domains in the encoded protein and facilitate the evolutionary shuffling of such domains (40). The gene structure of globin (41), bovine rhodopsin (42), and β-crystallin (43) also support this concept.

Evolutionary Implications. One additional implication of our findings concerns the origin of the postulated insertion sequence(s) in this nodulin gene. The fact that very little divergence (<6%) exists between the R1, R2, and R3 tandemly repeated units, which also constitute two introns, implies that this gene was generated very recently in evolution. In comparison to the rate of divergence in leghemoglobin genes (44), the two conserved introns (introns 2 and 3) in nodulin-24 genes have their origin <10 million years ago. However, based upon the RNY values, the exons encompassed by these repeats appear to be ancient. It can be postulated that these rearrangements were the result of environmentally induced mobilization of an insertion sequence (45, 46), which could have occurred during the evolution of the symbiotic state, possibly as a result of an early pathogenic relationship between these two organisms.

We thank Greg Brown, Eric Olson, and Carmen Sapienza for their critical reading of this manuscript, Marc Fortin for providing antibodies prepared against purified peribacteroid membranes of soybean nodules, Rose-Marie Riha and Alda Pombo-gentile for their excellent technical help, Joel Delisle for his computer analysis, and Yvette Mark for her secretarial assistance in preparing this manuscript. This work was supported by an operational and a strategic research grant from the Natural Sciences and Engineering Research Council of Canada and the Quebec Ministry of Education.

1. Legocki, R. P. & Verma, D. P. S. (1980) *Cell* **20**, 153–163.
2. Bergmann, H., Preddie, E. & Verma, D. P. S. (1983) *EMBO J.* **2**, 2333–2339.
3. Verma, D. P. S. & Nadler, K. (1984) in *Genes Involved in Microbe-Plant Interactions*, eds. Verma, D. P. S. & Hohn, T. (Springer, New York), pp. 57–93.
4. Cullimore, J. V., Lara, M., Lea, P. J. & Miflin, B. J. (1983) *Planta* **157**, 245–253.
5. Fuller, F., Kunstner, P. W., Nguyen, T. & Verma, D. P. S. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 2594–2598.
6. Fuller, F. & Verma, D. P. S. (1984) *Plant Mol. Biol.* **3**, 21–28.
7. Verma, D. P. S. (1982) in *The Molecular Biology of Plant Development*, eds. Smith, H. & Grierson, D. (Blackwell, Oxford), pp. 437–466.
8. Verma, D. P. S., Kazazian, V., Zogbi, V. & Bal, A. K. (1978) *J. Cell Biol.* **78**, 919–936.
9. Verma, D. P. S., Nash, D. T. & Schulman, H. M. (1974) *Nature (London)* **251**, 74–77.
10. Maniatis, T., Hardison, R. C., Lacy, E., Lauez, J., O'Connel, C., Quon, D., Sim, D. K. & Efstratiadis, A. (1978) *Cell* **15**, 687–701.
11. Varsanyi-Breiner, A., Gusella, J. F., Keys, C., Housman, D. E., Sullivan, D., Brisson, N. & Verma, D. P. S. (1979) *Gene* **7**, 317–334.
12. Benton, W. D. & Davis, R. W. (1977) *Science* **196**, 180–182.
13. Fisher, R. L. & Goldberg, R. B. (1982) *Cell* **29**, 651–660.
14. Southern, E. M. (1975) *J. Mol. Biol.* **98**, 503–517.
15. Wahl, G. M., Stern, M. & Stark, G. R. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 3683–3687.
16. Rigby, P. W. J., Dieckmann, M., Rhodes, C. & Berg, P. (1977) *J. Mol. Biol.* **113**, 237–251.
17. Messing, J. & Vieira, J. (1982) *Gene* **19**, 269–276.
18. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 5463–5467.
19. Biggin, M. D., Gibson, T. J. & Hong, G. F. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 3963–3965.
20. Wilbur, W. J. & Lipman, D. J. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 726–730.
21. Nevins, J. R. (1983) *Annu. Rev. Biochem.* **52**, 441–446.
22. Mount, S. M. (1982) *Nucleic Acids Res.* **10**, 459–472.
23. Erbil, C. & Niessing, J. (1983) *EMBO J.* **2**, 1339–1343.
24. King, R. C. & Piatigorsky, J. (1983) *Cell* **32**, 707–712.
25. Fisher, D. H., Dodgson, J. B., Hughes, S. & Engel, J. D. (1984) *Proc. Natl. Acad. Sci. USA* **81**, 2733–2737.
26. Maizel, J. V. & Lenk, R. P. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 7665–7669.
27. Kyte, K. & Doolittle, R. F. (1982) *J. Mol. Biol.* **157**, 105–132.
28. Shepherd, J. C. W. (1981) *J. Mol. Evol.* **17**, 94–102.
29. Iida, S., Meyer, J. & Arber, W. (1983) in *Mobile Genetics Elements*, ed. Shapiro, J. A. (Academic, New York), pp. 159–221.
30. Saito, H., Kranz, D. M., Takagaki, Y., Hayday, A. C., Eisen, H. N. & Tonegawa, S. (1984) *Nature (London)* **309**, 757–762.
31. Ozaki, L. S., Svec, P., Nussenzweig, R. S., Nussenzweig, V. & Godson, G. N. (1983) *Cell* **34**, 815–822.
32. Mostov, K. E., Friedlander, M. & Blobel, G. (1984) *Nature (London)* **308**, 37–43.
33. Jackson, R. C. & Blobel, G. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 5598–5602.
34. Verma, D. P. S. & Long, S. (1983) *Int. Rev. Cytol. Suppl.* **14**, 211–245.
35. Kleckner, N. (1981) *Annu. Rev. Genet.* **15**, 341–404.
36. Bell, G. I., Sanchez-Pescador, R., Laybourn, P. J. & Najarian, R. C. (1983) *Nature (London)* **304**, 368–370.
37. Eifertman, A. E., Young, P. R., Scot, R. W. & Tilghman, S. M. (1981) *Nature (London)* **294**, 713–718.
38. Blake, C. (1983) *Trends Biochem. Sci.* **8**, 11–13.
39. Hudspeth, M. E. S., Vincent, R. D., Perlman, P. S., Shumard, D. S., Treisman, L. O. & Grossman, L. J. (1984) *Proc. Natl. Acad. Sci. USA* **81**, 3148–3152.
40. Gilbert, W. (1978) *Nature (London)* **271**, 501.
41. Go, M. (1981) *Nature (London)* **291**, 90–95.
42. Nathans, J. & Hogness, D. H. (1983) *Cell* **34**, 807–814.
43. Inara, G., Piatigorsky, J., Norman, B., Slingsby, C. & Blundell, T. (1983) *Nature (London)* **302**, 310–315.
44. Brown, G. G., Lee, J. S., Brisson, N. & Verma, D. P. S. (1984) *J. Mol. Evol.* **21**, 19–32.
45. Bukhari, A. I., Shapiro, J. A. & Adhya, S. L. (1977) *DNA Insertion Elements, Plasmids and Episomes* (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY).
46. Roeder, G. S. & Fink, G. R. (1980) *Cell* **21**, 239–249.