# Extended Region of Nodulation Genes in *Rhizobium meliloti* 1021. II. Nucleotide Sequence, Transcription Start Sites and Protein Products

**Robert F. Fisher, Jean A. Swanson, John T. Mulligan and Sharon R. Long**

*Department of Biological Sciences, Stanford University, Stanford, California 94305*

Manuscript received April 9, 1987
Revised copy accepted July 9, 1987

## ABSTRACT

We have established the DNA sequence and analyzed the transcription and translation products of a series of putative nodulation (*nod*) genes in *Rhizobium meliloti* strain 1021. Four loci have been designated *nodF, nodE, nodG* and *nodH.* The correlation of transposon insertion positions with phenotypes and open reading frames was confirmed by sequencing the insertion junctions of the transposons. The protein products of these *nod* genes were visualized by *in vitro* expression of cloned DNA segments in a *R. meliloti* transcription-translation system. In addition, the sequence for *nodG* was substantiated by creating translational fusions in all three reading frames at several points in the sequence; the resulting fusions were expressed *in vitro* in both *E. coli* and *R. meliloti* transcription-translation systems. A DNA segment bearing several open reading frames downstream of *nodG* corresponds to the putative *nod* gene mutated in strain *nod-216.* The transcription start sites of *nodF* and *nodH* were mapped by primer extension of RNA from cells induced with the plant flavone, luteolin. Initiation of transcription occurs approximately 25 bp downstream from the conserved sequence designated the "*nod* box," suggesting that this conserved sequence acts as an upstream regulator of inducible *nod* gene expression. Its distance from the transcription start site is more suggestive of an activator binding site rather than an RNA polymerase binding site.

IN a companion study (SWANSON *et al.* 1987), we establish a physical and genetic map for the DNA between the common *nod* genes and the *nif* region. We report here the molecular analysis of a portion of this region. The determination of DNA sequence and identification of protein products for this region should provide approaches for analysis of individual gene functions and gene interactions, which will be important in the dissection of the complex host-specific nodulation of legumes by *Rhizobium.*

Analysis of function requires an understanding of gene expression. The *nodABC* genes of *Rhizobium meliloti* 1021 are not expressed in free-living cells (MULLIGAN and LONG 1985). Exposure of *R. meliloti* to plants or to plant exudates causes induction of these genes, as shown by *nodC-lacZ* fusions (MULLIGAN and LONG 1985) and by use of antibody to *nodA* protein to detect gene products (EGELHOFF and LONG 1985). The factors in alfalfa (*Medicago sativa*) exudates which cause induction are small aromatic molecules, and the most active has been identified by PETERS, FROST and LONG (1986) as the flavone luteolin (3′,4′,5,7-tetrahydroxyflavone). Studies in *R. trifolii* and *R. leguminosarum* show that, in addition to *nodABC*, other nodulation gene clusters are plant-inducible (INNES *et al.* 1985; SHEARMAN *et al.* 1986). The factors from white clover, *Trifolium repens,* responsible for induction have been isolated and identified as flavones, primarily 4′,7-dihydroxyflavone and two related molecules (REDMOND *et al.* 1986). In pea, *Pisum sativum,* a complex mixture of flavones, some of which may be glycosylated, appears to induce the *nod* genes of *R. leguminosarum* (FIRMIN *et al.* 1986); this effect is reproduced in free-living cells by the application of flavones such as apigenin and flavanones such as naringenin. The one known *nod* gene which is constitutive rather than inducible is *nodD* (MULLIGAN and LONG 1985). In *R. meliloti* and *R. leguminosarum,* induction of *nodABC* by plant exudates is in fact dependent on *nodD* expression (MULLIGAN and LONG 1985; ROSSEN *et al.* 1985).

Recent studies of *nod* gene expression in *R. leguminosarum* and *R. trifolii* show that *nodF* and other *nod* genes are also induced by treatment of cells with plant exudates or flavone inducers (SHEARMAN *et al.* 1986; REDMOND *et al.* 1986). The apparently coordinate expression of several operons of *nod* genes may be related to a highly conserved sequence which is found upstream of *nodABC, nodFE, nodH* and other sequences in *R. meliloti* (ROSTAS *et al.* 1986, DEBELLÉ and SHARMA 1986; FISHER *et al.* 1987), *R. leguminosarum* (SHEARMAN *et al.* 1986), *R. trifolii* (SCHOFIELD and WATSON 1986), and *Bradyrhizobium* sp. (*Parasponia*) (SCOTT 1986). This sequence, tentatively designated as the "*nod* box," may represent a control region. However, its relationship to *nod* gene tran-

The sequence data presented in this article have been submitted to the EMBL/GenBank Data Libraries under the accession No. Y00604.

scriptional start sites *in vivo* has not previously been examined. In this report, we determine the *in vivo* transcriptional initiation site for several *nod* genes and show that in each case the *nod* box lies 25–28 bp upstream of the start site.

## MATERIALS AND METHODS

**Strains:** Plasmids used in this study are shown in Figure 1 or described below. *R. meliloti* 1021 (EGELHOFF and LONG 1985) and RCR2011 (ROSENBERG *et al.* 1982), and *E. coli* HB101 (MANIATIS, FRITSCH and SAMBROOK 1982) and JM101 (MESSING 1983) were grown in bacteriological media as described by SWANSON *et al.* (1987).

**Materials:** Restriction enzymes were obtained from Bethesda Research Laboratories and Promega Biotec. T4 polynucleotide kinase, exonuclease III, S1 nuclease and *Pst*I linkers were also from Bethesda Research Laboratories. T4 DNA ligase and avian myeloblastosis virus reverse transcriptase were obtained from BioRad Laboratories.

**Plasmid and phage constructions:** Most of the plasmids used for DNA sequencing and protein expression are shown in Figure 1 and were constructed as follows. Subcloning was accomplished by techniques described by MANIATIS, FRITSCH and SAMBROOK (1982).

Construction of B27 and B28 depended on the intermediates described below. pRmJT16 is a 3.3-kb *Eco*RI-*Cla*I fragment of the 15.5-kb *Eco*RI fragment of pRmJT5 (SWANSON *et al.* 1987) cloned into *Eco*RI-*Cla*I-digested pBR322. The 0.2-kb *Eco*RI-*Xho*I fragment of pRmJT16 was cloned into *Eco*RI-*Sal*I-digested M13mp18 and M13mp19 to produce B27 and B28, respectively.

pRmF32 and pRmF33 were generated by cloning the 3.6-kb *Bam*HI fragment of pRmJT16 into the *Bam*HI site of pUC118; this *Bam*HI fragment contains 0.35-kb of pBR322 and 3.25 kb of *R. meliloti* DNA. These two plasmids containing the *Bam*HI fragment in opposite orientations were subsequently digested with *Pst*I and *Xba*I prior to treatment with exonuclease III to create a nested set of deletions for sequencing both strands of the entire 3.6-kb insert (HENIKOFF 1984).

pRmS15 was constructed by cloning the 2.6-kb *Sst*I-*Sph*I fragment from pRmS5 (SWANSON *et al.* 1987) into *Sst*I-*Sph*I-digested pUC18.

pRmS23 and pRmS24 were produced by cloning the 1.2-kb *Sal*I-*Sst*I fragment from pRmS15 into *Sal*I-*Sst*I-digested pUC118 and pUC119, respectively. pRmS23 was subsequently digested with *Sal*I and *Sph*I, and pRmS24 with *Bam*HI and *Sst*I, before treating with exonuclease III to generate a nested set of deletions for sequencing both strands of the insert.

Tn*5* insertions 109, 210 and 912 (see Figure 1) in pRmJT5 were subcloned into pBR322 to construct pRmJT20, pRmJT21 and pRmJT23, respectively. The 10.7-kb *Cla*I fragments (containing 5.8 kb of Tn*5* inserted into the 4.9-kb wild-type *Cla*I fragment) were inserted at the vector *Cla*I site. The 0.7-kb *Xho*I fragment from pRmJT21, containing the genome-Tn*5* junction of mutant 210, was inserted into *Sal*I-digested M13mp18 in both orientations to produce B4 and B5. The 1.7-kb *Hin*dIII fragment from pRmJT21 was cloned into *Hin*dIII-digested M13mp19 to generate B20. The 1.1-kb *Xho*I fragment from pRmJT20 was inserted into *Sal*I-cleaved M13mp19 in both orientations to make B6 and B7. B24 was constructed by inserting the 2.0-kb *Eco*RI-*Hin*dIII fragment from pRmJT23 into *Eco*RI-*Hin*dIII-digested M13mp18, and B26 was made by cloning the 1.9-kb *Xho*I-*Bgl*II fragment of pRmJT23 into *Bam*HI-*Sal*I-cleaved M13mp18.

pRmJT17 is the 5.2-kb *Cla*I fragment from pRmJT5 cloned into the *Cla*I site of pBR322. The 1.4-kb *Xho*I-*Bgl*II fragment of pRmJT17 was inserted into *Bam*HI-*Sal*I-digested M13mp18 to produce B25.

Construction of B30 involved several intermediate plasmids. pRmS505 is Tn*5* insert 505 (see Figure 1) in pRmJT5. The 11-kb *Cla*I fragment from pRmS505 (containing the 5.8-kb Tn*5* insertion inserted into the 5.2-kb *Cla*I fragment) was inserted into *Cla*I-cleaved pBR322 to generate pRmRF36. The 0.9-kb *Xho*I fragment from pRmRF36 was cloned in *Sal*I-digested M13mp19 to give rise to B30.

pRmS17 was constructed by inserting the 0.9-kb *Sst*I-*Sph*I fragment from pRmS5 in *Sst*I-*Sph*I-digested pUC18, and pRmS20 and pRmS21 were generated by cloning the 2.1-kb *Bam*HI-*Bgl*II fragment from pRmS5 into the *Bam*HI site, in both orientations, of pAD10. pRmS25 was generated by cloning the 1.3-kb *Sph*I-*Sst*I fragment from pRmS5 (SWANSON *et al.* 1987) into *Sph*I-*Sst*I-cleaved pUC19.

pRmF37, pRmF38, pRmF42 and pRmF43 are exonuclease III-digested derivatives of pRmF32. They were used as sources of *Eco*RI-*Pvu*II fragments of 1.7–2.1 kb which were cloned into *Eco*RI-*Sma*I-digested pAD10 to give rise to the expression plasmids pRmF40, pRmF41, pRmF48 and pRmF49, respectively.

pRmF51 was constructed from the 1.3-kb *Sph*I-*Sst*I fragment of pRmS25. The ends of this fragment were filled in with the Klenow fragment of DNA polymerase I. Phosphorylated *Pst*I linkers were ligated to this fragment with T4 DNA ligase and cleaved with *Pst*I, and the DNA was precipitated with ethanol. The fragment was separated from excess linkers by agarose gel electrophoresis and inserted into *Pst*I-digested pUC8 to produce pRmF51.

For sequencing the Tn*5* insertion sites of inserts 216, 307, 314, 316, 402, 411, 510, 614, 703, 705, 708, 805 and 913, *Bam*HI junction fragments, whose one end was in Tn*5* and contained the Kan$^R$ element, and whose other end was in the *R. meliloti* genome, were inserted into *Bam*HI-cleaved pUC118. To facilitate production of single-stranded DNA, the Kan$^R$ element was subsequently eliminated by digestion with *Hin*dIII and religation. The positions of Tn*5* inserts 304 and 906 were similarly determined by direct cloning of the appropriate *Hin*dIII-*Bam*HI fragments (with the *Hin*dIII site in Tn*5* and the *Bam*HI site in the *R. meliloti* genome) in *Bam*HI-*Hin*dIII-digested pUC118.

**DNA sequencing:** Sequencing was carried out by the dideoxy chain termination technique of SANGER, NICKLEN and COULSON (1977), in vectors M13mp18, M13mp19, pUC118 and pUC119. pUC118 and pUC119 are derivatives of pUC18 and pUC19 into which has been inserted the M13 intergenic region (J. VIEIRA, personal communication). To produce single stranded DNA, *Escherichia coli* JM101 containing pUC118 or pUC119 derivatives were infected with helper phage M13K07 in 2× YT containing 50 μg/ml ampicillin and 70 μg/ml kanamycin and shaken at 37° for 14–20 hr. Single-stranded DNA was then isolated as for M13 preparations (Amersham 1983). A series of nested deletions was created by the exonuclease III digestion procedure of HENIKOFF (1984) to sequence the segment spanned by plasmids pRmF32, pRmF33 and pRmS23, pRmS24. The sequencing strategy is shown in Figure 1. The sequence of transposon insertion positions was determined using a Tn*5*-homologous oligonucleotide as primer, as described below and by EGELHOFF *et al.* (1985). Overlapping nested deletions were organized and DNA sequence analysis conducted using SEQSORT, AA and RE programs as previously described (EGELHOFF *et al.* 1985).

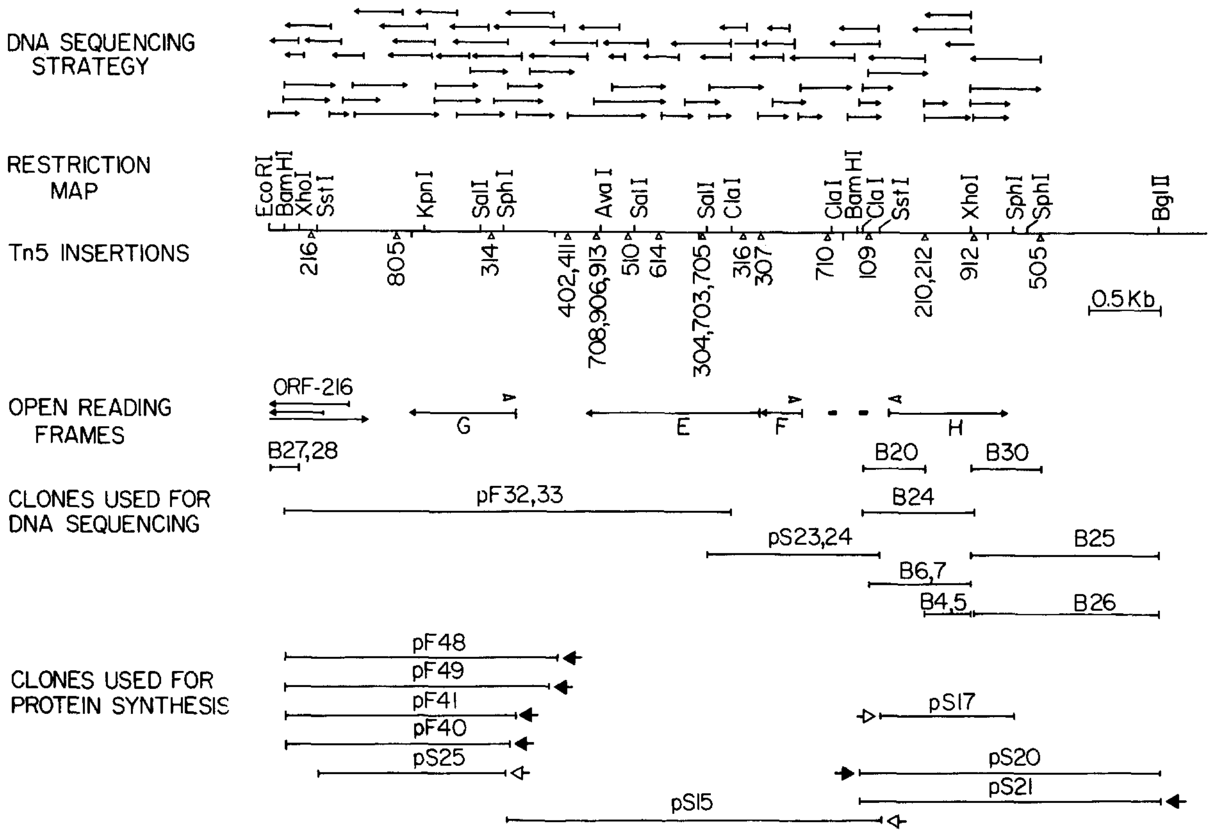


FIGURE 1.—Clones used to obtain DNA sequence, location of open reading frames, and position of transposon insertions and of synthetic oligonucleotide primers used to determine transcription initiation sites. The DNA sequencing strategy is shown in the upper part of the figure. In each case (*top of diagram*), the *tail end of the arrow* represents the end of a deletion, and the *length of the arrow* represents how much of the sequence was determined from that deletion; the end of the clone in each case would correspond to the end of the original cloned DNA segment. *Triangles* on the linear restriction map represent the insertion points for each Tn*5* mutant whose junction sequence was determined. Open reading frames which correlate with the phenotype of Tn*5* insertion mutants are designated by H, F, E, and G, as proposed by ROSTAS *et al.* (1986) and by DEBELLÉ and SHARMA (1986). Several potential protein coding sequences, present in both directions and more than one reading frame, are found in the DNA segment at the left end of the region shown here. These correlate with Tn*5* insertion 216 and appear to cross the *Eco*RI site; they are designated by ORF-216. Open arrowheads (▷) indicate position of oligonucleotides used for primer extension of RNA transcripts. Short horizontal bars (——) indicate the nod-box conserved sequence reported by ROSTAS *et al.* (1986). For improved figure clarity, the "Rm" designation (*e.g.*, p*Rm*F48) was omitted from all plasmid names. Clones used to analyze protein coding are shown in the bottom part of the figure. Expression was controlled by the *Salmonella typhimurium trp* promoter (—▶) of pAD10 (EGELHOFF and LONG 1985) or the *lac* promoter (—▷) of pUC18 and pUC19 (MESSING 1983). The direction of transcription controlled by the *trp* or *lac* promoter is shown by the orientation of the arrow.

**Protein products:** DNA segments from the *nod* gene region were cloned in expression plasmids pAD10, pUC8, pUC18, pUC19 or pUC118 so that transcription initiated either at the *Salmonella typhimurium trp* promoter (EGELHOFF and LONG 1985) (*closed arrows*, Figure 1 *bottom*) or the *E. coli lac* promoter (MESSING 1983) (*open arrows*, Figure 1 *bottom*). Plasmids (1 μg) purified by CsCl banding were incubated with a coupled transcription-translation extract from *E. coli* HB101 or *R. meliloti* RCR2011, essentially as previously described (GUNSALUS, ZURAWSKI and YANOFSKY 1979). The *R. meliloti* extract was prepared by a modification of the technique of ZUBAY *et al.* (1972) as follows: *R. meliloti* were grown in LB or M9 minimal medium to mid log phase (Klett 200, red filter) and were harvested by centrifugation. The cell pellet was weighed and resuspended in 10 mM Tris-acetate, pH 8.2, 14 mM MgOAc, 6 mM KOAc, 1 mM DTT (1 ml buffer/g cells), and broken in a French press. The *in vitro* mixture was augmented by the addition of amino acids including 15 μCi of [$^{35}$S]methionine (GUNSALUS, ZURAWSKI and YANOFSKY 1979). Reaction mixtures (25 μl) were incubated 70 min at 30° or 37°, depending on whether the source of the extract was from *R. meliloti* or *E. coli*, respectively. Protein products, processed as described previously (EGELHOFF and LONG 1985), were separated by polyacrylamide gel electrophoresis (PAGE) by the method of LAEMMLI (1970) and visualized by autoradiography. DUSHA *et al.* (1986) recently reported the independent development of a cell-free system from AK631, a derivative of Rm41.

**Transcript mapping:** RNA was isolated from induced *R. meliloti* 1021 grown in M9 minimal liquid medium (with addition of 15 μM luteolin for 3 hr) by the technique of C. YANOFSKY (personal communication) as previously described by FISHER *et al.* (1987). Briefly, cells were harvested on ice, resuspended in a lysis buffer, and extracted directly with phenol, after which the nucleic acids were subjected to several cycles of DNAase treatment followed by phenol extraction. Synthetic end-labeled oligonucleotides complementary to the coding regions for *nodG*, *nodF* and *nodH* (see Figures 1 and 2) were incubated with the RNA preparations, and avian myeloblastosis virus reverse transcriptase was used to extend each primer to the 5′ end of the

```
A
BamHI             15                  30                  45                  60
GGG ATC CAG CCT GTG AAT ATG AGA ACG CCG AGC CGC AGA GGG GAT GAG CGC AAT GTA GCA

                  75                  90                 105                 120
GGG AGA AAC ATA CTC GCG CCG GCA GTC GAT TTG ATT GTT GTG ATC TAC GGC AGA GAT GTG

                 135                 150                 165                 180
TCG TTC TTG GTG CAC TGG ATA GCC GTC CAC TCG TCA CAG ACA TCC ATT TCA CGG ATG GCC

                 195                 210           #710  225                 240
GAC ATC CAA ACA ATC CAT TTT ACC AAT CCC ACT GAT ATG TAG CAC AAG CTG CCC ACG ATA
                                                                        mRNA▶
                 255                 270                 285                 300
GGG AGG CCA ATG ATG TTC TTC GTC ATC GGA GGC TTC TGC ACA GAC CAG CCG GAT GTC CGG

                 315                 330                 345                 360
CTC TGC GGG CAT TAG GCT TAG CCA GTC GCG CAC GCC TGA TGA TAA TTT TCG TAT CGG GCC

                 375                 390                 405                 420
GCC TCA GGA ATT TGA GCC GCC GTG CGT CGA ACA CAA GCT AAA GGC AAC AGA ATG GTA GAT
                                                                    MET Val Asp

                 435                 450                 465                 480
CAA CTC GAA AGC GAA ATC ATT GGC ATC ATC AAG AAC CGT GTC GAA TCG GAG GGC GGC GAT
Gln Leu Glu Ser Glu Ile Ile Gly Ile Ile Lys Asn Arg Val Glu Ser Glu Gly Gly Asp

                 495                 510                 525                 540
GGA GAG ACC GCG TTA ATA GTC GGC GAT TTA ACG GCT GCC ACT GAA TTG ACC GCG CTT GGT
Gly Glu Thr Ala Leu Ile Val Gly Asp Leu Thr Ala Ala Thr Glu Leu Thr Ala Leu Gly

                 555                 570                 585                 600
GTC GAT TCT CTC GGA TTG GCA GAC ATC ATC TGG GAC GTG GAA CAG GCC TAC GGT ATC AGG
Val Asp Ser Leu Gly Leu Ala Asp Ile Ile Trp Asp Val Glu Gln Ala Tyr Gly Ile Arg

                 615                 630                 645                 660
ATC GAG ATG AAC ACG GCC GAG GCG TGG TCG GAT CTC CAG AAC GTC GGC GAC ATA GTG GGA
Ile Glu MET Asn Thr Ala Glu Ala Trp Ser Asp Leu Gln Asn Val Gly Asp Ile Val Gly

                 675#307             690                 705                 720
GCC ATC CGA GGC TTG CTC ACT AAG GGG GCT TGA ATG GAC AGG CGC GTT GTC ATC ACC GGA
Ala Ile Arg Gly Leu Leu Thr Lys Gly Ala  .  MET Asp Arg Arg Val Val Ile Thr Gly

                 735                 750                 765                 780
ATG GGC GGC CTA TGC GGA CTG GGC ACC GAC ACC ACC TCC ATC TGG AAA TGG ATG CGC GAA
MET Gly Gly Leu Cys Gly Leu Gly Thr Asp Thr Thr Ser Ile Trp Lys Trp MET Arg Glu

                 795                #316                 825                 840
GGC CGC TCC GCC ATC GGG CCG CTT CTC AAT ACA GAG CTT CAC GGC CTG AAG GGC ATA GTG
Gly Arg Ser Ala Ile Gly Pro Leu Leu Asn Thr Glu Leu His Gly Leu Lys Gly Ile Val

                 855                 870                 885                 900
GGC GCT GAG GTC AAG GCG CTG CCT GAC CAC AAC ATC GAC CGC AAG CAG CTC GTA TCG ATG
Gly Ala Glu Val Lys Ala Leu Pro Asp His Asn Ile Asp Arg Lys Gln Leu Val Ser MET

                 915                 930                 945                 960
GAT CGC ATT AGC GTG CTT GCC GTG ATT GCA GCG CAC GAA GCC ATG CGC CAG GCC GGG CTT
Asp Arg Ile Ser Val Leu Ala Val Ile Ala Ala His Glu Ala MET Arg Gln Ala Gly Leu

                 975                 990                1005                1020
TCC TGC AAT GAA GGA AAT GCC CTT CGG TTC GGC GCG ACC GTG GGC GTC GGC TTG GGA GGA
Ser Cys Asn Glu Gly Asn Ala Leu Arg Phe Gly Ala Thr Val Gly Val Gly Leu Gly Gly

                1035                1050                1065                1080
TGG GAC GCT ACC GAA AAA GCA TAC CGT ACC CTC CTT GTC GAC GGG GGG ACC CGT ACT GAA
Trp Asp Ala Thr Glu Lys Ala Tyr Arg Thr Leu Leu Val Asp Gly Gly Thr Arg Thr Glu

           #304 1095                1110                1125                1140
ATC TTC ACT GGT GTA AAG GCT ATG CCG AGT GCC GCC GCC TGC CAG GTC AGC ATG AGC CTC
Ile Phe Thr Gly Val Lys Ala MET Pro Ser Ala Ala Ala Cys Gln Val Ser MET Ser Leu

                1155                1170                1185                1200
GGC CTG CGG GGC CCG GTC TTC GGC GTC ACC TCC GCC TGT TCC TCG GCC AAC CAT GCG ATC
Gly Leu Arg Gly Pro Val Phe Gly Val Thr Ser Ala Cys Ser Ser Ala Asn His Ala Ile

                1215                1230                1245                1260
GCT TCG GCG GTA GAC CAG ATC AAG TGC GGC CGG GCC GAC GTC ATG CTC GCG GGG GGC AGC
Ala Ser Ala Val Asp Gln Ile Lys Cys Gly Arg Ala Asp Val MET Leu Ala Gly Gly Ser

                1275                1290                1305                1320
GAC GCG CCA CTA GTC TGG ATT GTG CTG AAG GCA TGG GAA GCT ATG CGC GCA CTC GCT CCG
Asp Ala Pro Leu Val Trp Ile Val Leu Lys Ala Trp Glu Ala MET Arg Ala Leu Ala Pro

                1335                1350                1365                1380
GAT ACT TGC CGA CCC TTC TCC GCC GGC AGG AAA GGC GTG GTA CTG GGC GAG GGT GCA GGC
Asp Thr Cys Arg Pro Phe Ser Ala Gly Arg Lys Gly Val Val Leu Gly Glu Gly Ala Gly

                1395#614            1410                1425                1440
ATG GCC GTG CTG GAA AGC TAT GAA CAT GCC ACC GCT CGC GGT GCA ACA ATA CTC GCG GAG
MET Ala Val Leu Glu Ser Tyr Glu His Ala Thr Ala Arg Gly Ala Thr Ile Leu Ala Glu

                1455                1470                1485                1500
GTC GCC GGC GTC GGC CTT TCC GCC GAT GCG TTC CAT ATC ACA GCG CCG GCT GTC CAT GGG
Val Ala Gly Val Gly Leu Ser Ala Asp Ala Phe His Ile Thr Ala Pro Ala Val His Gly

                1515                1530                1545                1560
CCG GAG TCG GCG ATG CGC GCT TGC CTT GCC GAT GCA GGA CTC AAT GCC GAG GAC GTC GAC
Pro Glu Ser Ala MET Arg Ala Cys Leu Ala Asp Ala Gly Leu Asn Ala Glu Asp Val Asp

                1575                1590            #510                    1620
TAC CTC AAC GCG CAC GGC ACC GGC ACC AAG GCC AAC GAT CAA AAC GAA ACT ACG GCT ATC
Tyr Leu Asn Ala His Gly Thr Gly Thr Lys Ala Asn Asp Gln Asn Glu Thr Thr Ala Ile

                1635                1650                1665                1680
AAG CGC GTC TTC GGA GAC CAT GCT TAT TCG ATG TCC ATA TCT TCC ACC AAG TCC ACC CAC
Lys Arg Val Phe Gly Asp His Ala Tyr Ser MET Ser Ile Ser Ser Thr Lys Ser Thr His

                1695                1710                1725                1740
GCG CAC TGT ATC GGC GCA GCG AGT GCG CTT GAA ATG ATC GCC TGT GTG ATG GCG ATC CAA
Ala His Cys Ile Gly Ala Ala Ser Ala Leu Glu MET Ile Ala Cys Val MET Ala Ile Gln

                1755                1770                1785                1800
GAA GGA GTC GTG CCG CCG ACC GCC AAC TAT CGT GAG CCA GAT CCC GAT TGC GAT CTA GAC
Glu Gly Val Val Pro Pro Thr Ala Asn Tyr Arg Glu Pro Asp Pro Asp Cys Asp Leu Asp

                1815        #913    1830                1845                1860
GTG ACG CCA AAC GTG CCG CGT GAG CGT AAG GTG CGC GTT GCC ATG AGC AAC GCC TTC GCC
Val Thr Pro Asn Val Pro Arg Glu Arg Lys Val Arg Val Ala MET Ser Asn Ala Phe Ala

                1875                1890                1905                1920
ATG GGT GGC ACG AAC GCA GTT CTG GCA TTC AAG CAG GTA TGA GCC TGT CAG TTG CTT CCT
MET Gly Gly Thr Asn Ala Val Leu Ala Phe Lys Gln Val  .

                1935                1950                1965                1980
CGA TGA TCA CCA CTT GCA GGG CGG CAT GAG ACG CCC CCT GTG CCG TTG CAT CAA TGA AGT

                1995                2010                2025#411            2040
CCG TCA AGC GGG ATC GGG CAT GGC TCT CGT CAG GAG AAG GAG CGA CTG GTC TGG GCA GCG

                2055                2070                2085                2100
TTT GAC CCT AGG CCA GCG CTT CGA GAT CGC GTG ATT GGC CGG CGT TCA TGT AAG GCA GTT

                2115                2130                2145                2160
TTT CTG GTG GCG CGA TGA AGT AAT GCT GGT TGT TCT TGC GGC ATT GCC GGC GTG TAC CGG

                2175                2190                2205                2220
CAT CAG CGT GAA TGC ATC GCA GCA AGC AAG TCG GAG CTT GCA AGG TGC TGT CTG ACC TAG

                2235                2250                2265                2280
CGC CGT GTA TCA GCC GCG AAG CCG GGC TCG CCG CTG CGC AAT ATT AAG GCG GAG CTG CGG

                2295                2310                2325                2340
CGC AAG GCG ACG ATC AGC CGC GGC AAG GAA GCC ACG ATG AAC ACG AAG ACG CTG GCG CGG

                2355                2370                2385                2400
TGA CGT TAA AAG ACA CCA TCA CCA GCC TAC ACG ATA TGA GAA GAG GTT AAG ACA ATG TTC
                                                                        MET Phe

                2415                2430                2445                2460
GAA TTG ACC GGG CGC AAG GCG CTC GTC ACG GGC GCA TCA GGA GCC ATA GGA GGG GCT ATC
Glu Leu Thr Gly Arg Lys Ala Leu Val Thr Gly Ala Ser Gly Ala Ile Gly Gly Ala Ile

                2475                2490                2505                2520
GCC CGC GTG CTG CAT GCT CAG GGC GCT ATC GTC GGA CTG CAC GGC ACC CAA ATT GAA AAA
Ala Arg Val Leu His Ala Gln Gly Ala Ile Val Gly Leu His Gly Thr Gln Ile Glu Lys

                2535                2550#314            2565                2580
CTG GAG ACA CTG GCA ACT GAG CTT GGA GAC CGG GTC AAG CTG TTC CCG GCT AAT CTG GCC
Leu Glu Thr Leu Ala Thr Glu Leu Gly Asp Arg Val Lys Leu Phe Pro Ala Asn Leu Ala

                2595                2610                2625                2640
AAT CGA GAC GAA GTC AAG GCG CTT GGT CAG AGA GCG GAA GCC GAT CTT GAA GGC GTC GAC
Asn Arg Asp Glu Val Lys Ala Leu Gly Gln Arg Ala Glu Ala Asp Leu Glu Gly Val Asp

                2655                2670                2685                2700
ATC CTG GTC AAC AAT GCT GGC ATC ACC AAG GAT GGA TTG TTC TTG CAC ATG GCA GAC CCC
Ile Leu Val Asn Asn Ala Gly Ile Thr Lys Asp Gly Leu Phe Leu His MET Ala Asp Pro

                2715                2730                2745                2760
GAC TGG GAC ATT GTG CTG GAG GTC AAC CTC ACC GCC ATG TTC CGA CTC ACC CGC GAG ATC
Asp Trp Asp Ile Val Leu Glu Val Asn Leu Thr Ala MET Phe Arg Leu Thr Arg Glu Ile

                2775                2790                2805                2820
ACC CAG CAG ATG ATA CGC CGT CGA AAT GGC CGC ATC ATC AAT GTC ACT TCG GTC GCC GGC
Thr Gln Gln MET Ile Arg Arg Arg Asn Gly Arg Ile Ile Asn Val Thr Ser Val Ala Gly

                2835                2850                2865                2880
GCC ATC GGC AAT CCA GGC CAG ACC AAT TAC TGC GCC TCC AAG GCC GGT ATG ATC GGC TTT
Ala Ile Gly Asn Pro Gly Gln Thr Asn Tyr Cys Ala Ser Lys Ala Gly MET Ile Gly Phe

                2895                2910                2925                2940
TCC AAG TCG CTG GCG CAG GAG ATC GCT ACG CGA AAC ATC ACT GTC AAC TGC GTC GCC CCG
Ser Lys Ser Leu Ala Gln Glu Ile Ala Thr Arg Asn Ile Thr Val Asn Cys Val Ala Pro

                2955                2970                2985                3000
GGC TTC ATC GAA TCG GCA ATG ACC GAT AAG CTC AAT CAC AAA CAG AAG GAG AAA ATC ATG
Gly Phe Ile Glu Ser Ala MET Thr Asp Lys Leu Asn His Lys Gln Lys Glu Lys Ile MET

                3015                3030                3045                3060
GTG GCG ATC CCG ATC CAC CGC ATG GGC ACC GGT ACC GAA GTC GCG TCC GCC GTT GCG TAT
Val Ala Ile Pro Ile His Arg MET Gly Thr Gly Thr Glu Val Ala Ser Ala Val Ala Tyr

                3075                3090                3105                3120
CTC GCT TCC GAT CAC GCC GCC TAT GTC ACC GGA CAG ACC ATT CAC GTG AAC GGC GGT ATG
Leu Ala Ser Asp His Ala Ala Tyr Val Thr Gly Gln Thr Ile His Val Asn Gly Gly MET

                3135                3150                3165                3180
GCA ATG ATT TGA AGG CGG TCG GGC CTA CGG ATG AGT GGG CTT GCA TTT GCA TAC GCC AGC
Ala MET Ile  .

                3195                3210#805            3225                3240
CTA TCA GCG CAA TGA TGA TAA CGG CAT AAA GGC CAT TGC ACT TTC CGA AAG CTG AGG AAG

                3255                3270                3285                3300
CAA GCC ATT ATG GAT AGT GCA CCT GTC AGC AAT ACT GAA CGG TCT CAA CGG AAT AGC CTG

                3315                3330                3345                3360
CGA TTG AGC GCT CCG GTC CCA GCA GCA ATA GCT CGG CCC CAT ATG AAG ACG CTG TCT CGC

                3375                3390                3405                3420
TCG GCG CCG GCG CAT CAG CGC GGA ACG TCA GAT AGC GCA AAC GCT TTA GTG CGG CGT TGC

                3435                3450                3465                3480
TTA GCG CCA TTA CGT CGC GCC ACC GTC TTG CCG CGG TGA TCC CAC GCA TTG GGA TGC CTT

                3495                3510                3525                3540
GAG CGA GCT GAG CTG CCG AGG CGT AAC CCG GAT AGG TTT CCT GAA CAT AGA ACA AGG CCA

                3555                3570                3585                3600
CAA ATG TCT CTT CCC CAT CTT CGG CGG CTT GAA GCC GAA GCG ATC CAT GTC ATT CGA GAA

                3615                3630                3645                3660
GTT GTT GCG ACA TTC TCC AAT CCG GTC GTG CTT TAC TCG ATC GGC AAA GAC TCC TCG GTA

                3675                3690                3705                3720
CTG CTG CAC CTG GCG ATG AAG GCG TTC TAC CCC GCC AAG CCG CCA TTT CCA TTC CTG CAT

                3735                3750                3765                3780
GTA GAT ACC AAA TGG AAG TTC CGG GAG ATG ATC GAG TTT CGC GAC CGG ATG GCG CGA GAG

                3795                #216                3825                3840
CTC GGC TTC GAT CTC CTC GTC CAC GTC AAT CAG GAC GGG GTC GAG CAG GGC ATC GGG CCA

                3855                3870                3885                3900
TTC ACG CAC GGT TCC AAC GTG CAC ACC CAT GTC ATG AAG ACG ATG GGG CTC CGG CAG GCG

                3915                3930                3945                3960
CTC GAG AAA TAC GGT TTC GAC GCG GCG CTC GCA GGC GCG CGG CGC GAC GAG GAG AAG TCG

                3975                3990                4005                4020
CGC GCC AAG GAA CGC ATC TTC TCG ATT CGC AGC GCC CAG CAC GGC TGG GAT CCG CAG CGC

                4035                4050                4065                4080
CAG CGG CCC GAG ATG TGG AAG ACT TAC AAT ACG CGG GTC GGA CAA GGC GAG ACG ATG CGA

                4095                4110
GTC TTC CCG CTT TCC AAC TGG ACC GAA TTC
```

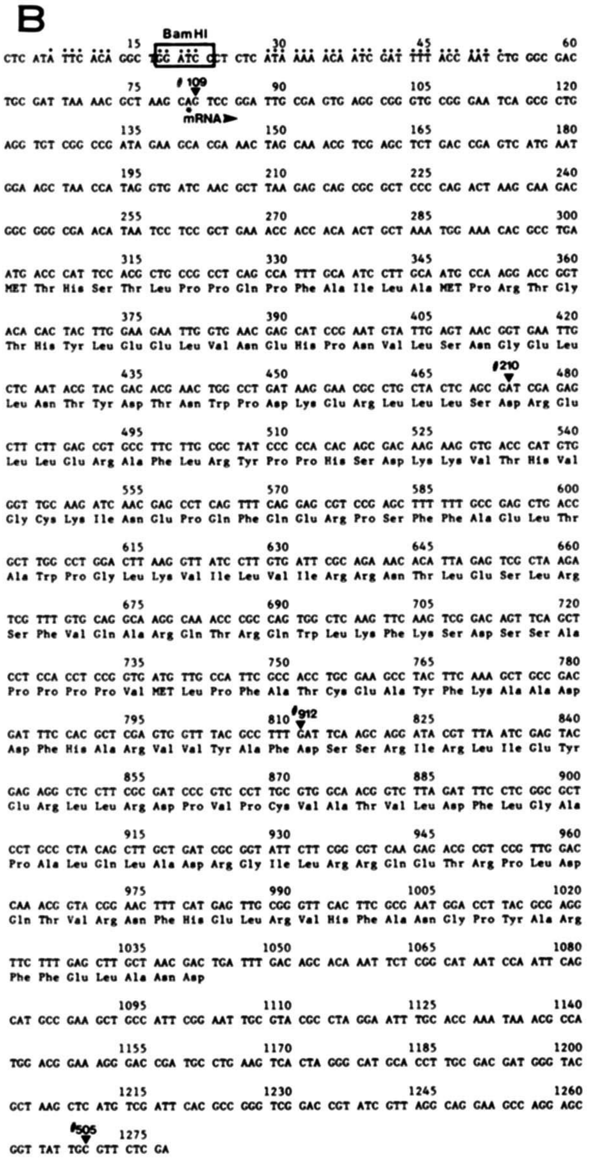


FIGURE 2.—DNA sequence of putative *nod* genes *F*, *E*, *G*, and *H*. In (A) the nucleotide sequence beginning upstream of *nodF* and proceeding beyond *nodG* to the *Eco*RI site is presented (refer to Figure 1 for map). In (B), the sequence reads in the opposite direction (*i.e.*, toward *nifHDK*), and begins at a position corresponding to nucleotide 23 in (A). The *Bam*HI site which is present in both (A) and (B) is boxed to facilitate alignment of the two sequences. The transcription start sites deduced in Figure 5 are indicated as asterisks at nucleotide 234–237 in (A) (*nodF* transcript) and 80 in (B) (*nodH* transcript). Positions of transposon Tn*5* insertions whose junctions have been sequenced are indicated by the insertion number, with a filled triangle pointing to the left-most base in the 9 base pair repeat created by the insertion. The consensus *nod* box sequences are indicated by dots (·). An inverted repeat, capable of forming a hairpin with 10 of 14 matches (ΔG = −16.2 kcal/mol), is shown in (A) as diverging arrows spanning nucleotides 287–318. Several strains had Tn*5* insertions at identical nucleotides; only one of each of these is shown. The groups are as follows: (304, 703, 705); (210, 212); (708, 906, 913); (402, 411). Strains with different first numerals are products of separate mutagenesis experiments, and thus are independent insertions at the identical nucleotide.

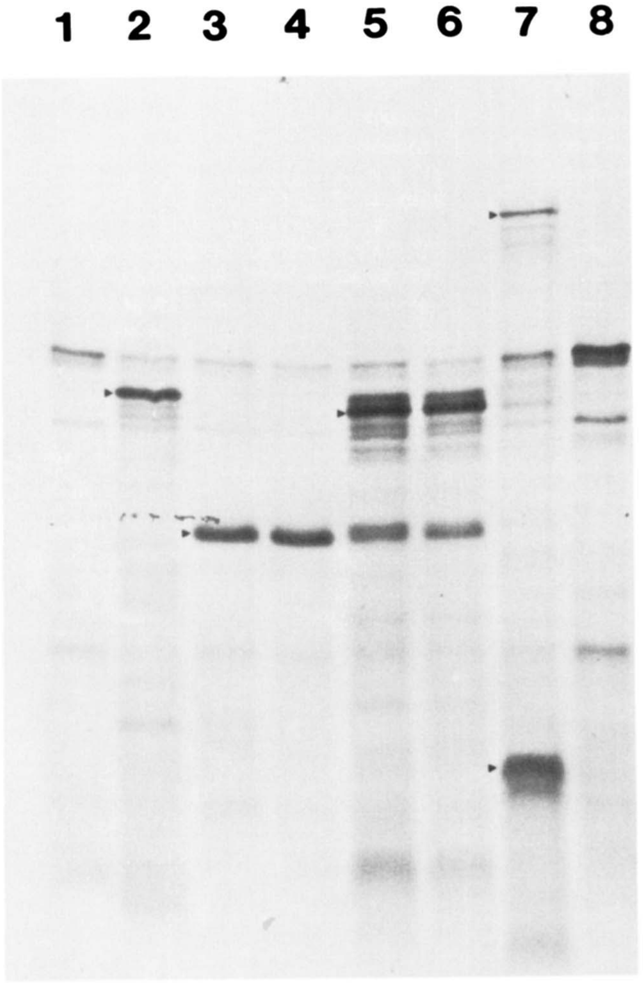


FIGURE 3.—*In vitro* expression of *nod* protein products. Coupled transcription-translation was conducted with an *R. meliloti* extract and analyzed on SDS-polyacrylamide gels as described in Materials and Methods. Plasmids directing *in vitro* protein synthesis are as follows: Lane 1: vector pAD10 (control for lanes 2–6). Lane 2: plasmid pRmS20; Lane 3: plasmid pRmF41; Lane 4: plasmid pRmF40; Lane 5: plasmid pRmF48; Lane 6: plasmid pRmF49; Lane 7: plasmid pRmS15; Lane 8: vector pUC19 (control for lane 7).

corresponding transcript (WILLIAMS and MASON 1985). In addition, the appropriate cloned single stranded DNAs were annealed with the same unlabeled oligonucleotide primers, and sequencing ladders were generated by dideoxy chain termination reactions from these primers. The sequencing ladders and RNA-complementary primer extension products were electrophoresed in parallel on sequencing gels to establish the position of the transcript initiation sites.

## RESULTS

**DNA sequence of the extended *nod* gene region:** In the *R. meliloti* genome, mutations in a region mapping between *nodDABC* and *nifHDK* cause severely delayed and Nod⁻ phenotypic changes (SWANSON *et al.* 1987). We determined the nucleotide sequence of this DNA segment and also located the precise transposon insertion sites of all the Tn*5* mutants which mapped in this region (Figure 1). We were thereby able to correlate Nod$^-$ phenotype directly with the position of a given Tn*5* insertion. The DNA sequence, shown in Figure 2, A and B, was analyzed for open reading frames (ORFs) and other features to better understand the molecular and genetic organization of this region.

**ORFs defining *nodF*, *nodE* and *nodG*.** In Figure 2A, three ORFs are presented, starting at nucleotides 412, 694 and 2395. The first two of these lie within a segment in which Tn*5* insertions cause significant delays and reductions in nodulation of alfalfa (SWANSON *et al.* 1987). These two ORFs have been designated *nodF* and *nodE*, according to the convention of SHEARMAN *et al.* (1986) for *R. leguminosarum*, DEBELLE and SHARMA (1986) and ROSTAS *et al.* (1986) for *R. meliloti*, and SCHOFIELD and WATSON (1986) for *R. trifolii*; they are equivalent to the *R. meliloti* strain 41 genes designated *hsnA* and *hsnB* by HORVATH *et al.* (1986). *nodF* specifies a protein of 93 amino acids ($M_r$ 9,760) and *nodE* encodes one of 402 amino acids ($M_r$ 41,779). Downstream of *nodE* is a DNA segment approximately 500 bp long, in which several short ORFs initiating with Met were found, but none were larger than about 40 amino acids. This is followed (at nucleotide 2395 of Figure 2A) by an ORF which we designated *nodG* (245 amino acids, $M_r$ 26,058) after DEBELLÉ and SHARMA (1986) [also called *hsnC* (HORVATH *et al.* 1986)].

Tn*5* insertions generated in the accompanying study (SWANSON 1987) (Figure 1) were located within the sequence shown in Figure 2A. Several points are noteworthy. Transposons 307, 316, 304, 703, 705, 614, 510, 708, 906 and 913 had previously been shown to cause marked decreases and delays in nodulation. The insertion point for transposon 307 lies within the ORF for *nodF*, and the others lie in the ORF for *nodE*. By contrast, strains 402 and 411, which display no altered symbiotic phenotype, have Tn*5* inserted about 120 bp downstream from the end of *nodE*; this provides a bracket for the Nod$^-$ phenotype, and is consistent with the ORFs determined by DNA sequence analysis.

An almost normal nodulation phenotype is seen with mutant 314 (nucleotide 2553, Figure 2A), which was the only Tn*5* insertion found in the large ORF of *nodG*. A transposon insertion *nif*-distal to 314, 805, also shows a slight delay in nodulation, although its position does not coincide with that of a significant ORF.

Transposon insertion 216 was found to have a severely altered Nod$^-$ phenotype, resulting in a pronounced delay in nodule formation (SWANSON *et al.* 1987). Interestingly, several large ORFs were found when the sequence of the DNA flanking insert 216 was determined. Two of these read in the same direction as *nodF*, *E*, and *G*, and begin with Met residues at nucleotides 3544 and 3732 (Figure 2A). The two lie in different reading frames, and each is continuous through the *Eco*RI site. Protein product analysis (see below) and preliminary sequence analysis downstream of the *Eco*RI site is consistent with the site at nucleotide 3544 being functional in translation initiation. In the opposite orientation, an ORF extends from the *Eco*RI site through nucleotide 3430 (Figure 2A). Which, if any, of these ORFs constitutes a gene is currently under analysis by additional DNA sequence determination of this region, and by transcript, protein, and complementation analyses.

**ORF defining *nodH*:** Figure 2B shows the DNA sequence of a large ORF reading divergently from the ORFs shown in Figure 2A. This ORF has been designated *nodH* by DEBELLÉ and SHARMA (1986) and ROSTAS *et al.* (1986), and *hsnD* by HORVATH *et al.* (1986).

Three Tn*5* insertions, two of which are probably siblings, were found to lie in the ORF of *nodH*; all of these, 210, 212, and 912, caused very marked reductions in nodule number and a long delay in the appearance of the few nodules which did form (SWANSON *et al.* 1987). The phenotypes of these transposon insertions thus correlate perfectly with their position within *nodH*. A downstream transposon, 505, which exhibits no altered phenotype, lies outside the ORF for *nodH*.

**Protein products:** Several features of the nucleotide sequence were confirmed by analysis of protein products encoded by specific segments of the *nod* gene region. In the first set of analyses, *nod* gene segments were cloned into the expression vector pAD10 (EGELHOFF and LONG 1985), in which transcription is driven by the *trp* promoter of *S. typhimurium*. We had previously shown (FISHER *et al.* 1987) that the almost identical *trp* promoter of *E. coli* is recognized and efficiently utilized by *R. meliloti* RNA polymerase. Coupled transcription-translation was conducted *in vitro* with an *R. meliloti* extract to express radiolabeled polypeptides for analysis by PAGE and autoradiography. Clones pRmS20 and pRmS21 (see Figure 1) were used to analyze the *nodH* gene segment. An *in vitro* translation product of apparent molecular weight 29,000 was produced by expression of clone pRmS20 (Figure 3, lane 2, *arrow*), corresponding to the ORF for *nodH* and in excellent agreement with the predicted size of 28,552. By contrast, clone pRmS21, in which transcription proceeds in the opposite direction, produced no insert-specific translation products (data not shown).

Clones pRmF48, pRmF40, pRmF41 and pRmF40 were used to analyze the segment of DNA encoding *nodG* and *nod-216* (Figure 1). A protein product of approximately 28,000 is generated only by plasmids pRmF48 and pRmF49 (Figure 3, lanes 5 and 6). This size agrees well with that predicted by the DNA sequence of this region for *nodG* (26,058). The inserts in these two plasmids contain DNA which lies directly upstream of the putative *nodG* translation start site. Plasmids pRmF40 and pRmF41, whose promoter-proximal insert ends begin downstream of the *nodG* translation start site (see Figure 1), do not synthesize the *nodG* product (Figure 3, lanes 3 and 4). A smaller protein product of approximately 20,000 is produced by pRmF40, pRmF41, pRmF48 and pRmF49 (Figure 3, lanes 3–6). The size of this product corresponds well with that of an insert-vector fusion polypeptide which is specified by the DNA sequence of the *nod-216* region, and is consistent with translation initiation occurring at nucleotide 3544 of Figure 2A for ORF-216.

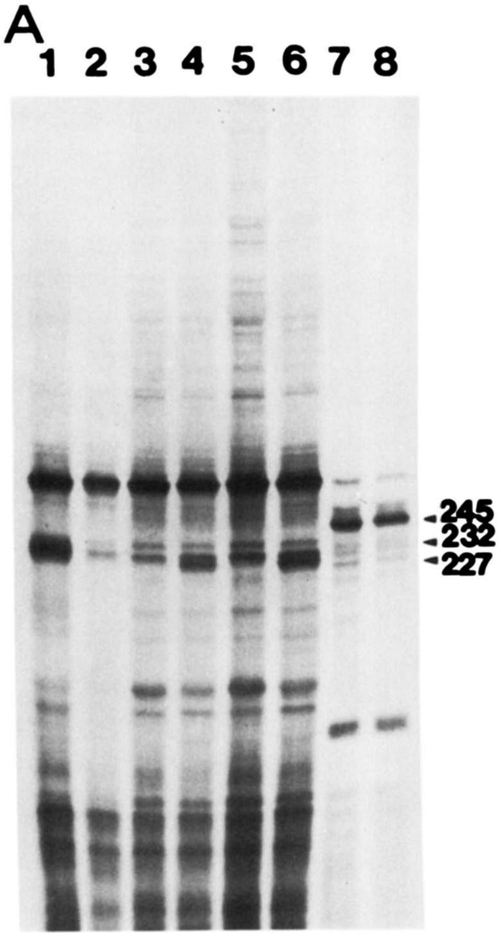
Plasmid pRmS15, which includes the DNA segment spanning *nodF* and *nodE*, produces proteins of approximately 42,500 and 13,000 (Figure 3, lane 7, *arrows*), which correspond well with the predicted sizes of 41,779 and 9,760 for the *nodE* and *nodF* ORFs, respectively (Figure 2A). Expression of pRmS15 also gives rise to a polypeptide migrating at about 8,800. This presumably results from the fusion of the *nodG* sequence to an in-phase ORF lying downstream of the *Sph*I site in the pUC19 polylinker. Such a fusion would result in the synthesis of a hybrid polypeptide of 7,978. It is not known whether this product arises *in vitro* from the same transcript as the *nodF* and *nodE* proteins.

Analysis of the DNA sequence of *nodG* was complicated in two positions by regions of band compression on the sequencing gels. We therefore decided to confirm the choice and expected size of the putative ORF we had deduced for *nodG* by independent methods. The full length (245 amino acid residue) *nodG* gene product is expressed by pRmF48 and pRmF49 (Figure 4A, lanes 7 and 8; 4B, *top line*). We made use of fusion plasmids which permit testing of multiple open reading frames, and created fusions of more than one segment of the *nodG* sequence, to confirm that the indicated ORF was correct. We cloned the *nodG* segment into the *Pst*I site of pUC8, pUC9 and pUC118, which fuses the *nodG* coding sequence to the vector *lacZ* in three different reading frames, and analyzed protein expression directed by these constructs. pRmF51 used a *Pst*I linker to fuse the *nodG* coding sequence to *lacZ* in pUC8; the sequence of this fusion is shown in the bottom line of Figure 4B. When this plasmid was used to direct protein synthesis in a coupled transcription-translation system, it generated the predicted 232 amino acid residue polypeptide shown in Figure 4A, lane 1). When this same *Pst*I-linkered fragment was inserted into pUC9 and pUC118, creating fusions in the other two reading frames, and used to direct protein synthesis, no large insert-specific proteins were produced (data not shown). In addition, pRmS25 fused *lacZ* in frame to the *nodG* coding sequence at the *Sph*I site (Figure 4B, *middle line*), resulting in a 227 amino acid fusion protein (Figure 4A, lanes 4 and 6). Thus these two independent sets of fusions confirmed that the predicted *nodG* ORF shown in Figure 2A was correct. This analysis also demonstrates that insertion 314 interrupts the *nodG* ORF; thus the lack of altered phenotype of this insertion mutant is in contrast to the *nodG*::Tn*5* insertion mutants reported by HORVATH *et al.* (1986).

Because insert-specific expression of pRmS25 and pRmF51 was controlled by the *E. coli* wild-type *lac* promoter, expression *in vitro* by *E. coli* extracts was dramatically enhanced upon addition of exogenous cyclic AMP (cAMP) to the template-extract mixture (Figure 4A, lanes 1 *vs.* 2; lanes 4 and 6 *vs.* 3 and 5). Expression of the full-length *nodG* protein from the *trp* promoter on pRmF48 and pRmF49 is shown in Figure 4A, lanes 7 and 8. In other experiments we were able to show that cAMP was unable to stimulate expression from the *lac* promoter in *R. meliloti* extracts (data not shown).

**Transcription initiation sites:** To determine the transcription start sites of the *nod* gene proteins, we isolated RNA from *R. meliloti* grown under free-living conditions and under conditions which induce *nod* gene expression. Synthetic oligonucleotides complementary to the coding regions of *nodG*, *nodF* and *nodH* (Figure 1) were used to carry out both primer extension reactions with the RNA, and a series of DNA sequencing reactions with single strand derivatives of pRmF32 (for *nodG*), pRmS23 (for *nodF*) and B20 (for *nodH*). Electrophoresis on sequencing gels revealed a single defined start site for the *nodH* induced transcript (Figure 5, *left panel*) and distinct *nodF* induced transcription start sites at a few adjacent nucleotides (Figure 5, *right panel*). Adjacent lanes show extension on uninduced transcripts. The transcription start sites are indicated in the DNA sequence shown in Figure 2. The primer extension products for the *nodF* induced transcript also included a prominent lower molecular weight band, which is consistent with a transcript 5′ end at nucleotide 284 (Figure 2A). This corresponds to the beginning of an inverted repeat capable of generating a stable RNA hairpin ($\Delta G = -16.2$ kcal/mol). While it is possible that this base represents an alternative *in vivo* transcription initiation site, it is also possible that the primary transcript is degraded or processed, and that the hairpin secondary structure serves to stabilize the RNA from further degradation (BELASCO *et al.* 1985).

Primer extension attempts using a *nodG* primer yielded no defined RNA-complementary product (data not shown). This may indicate the absence of *nodG*-homologous RNA in the *R. meliloti* cells, or it



B

```
                                                                             Sph I
ATGTTCGAATTGACCGGGCGCAAGGCGCTCGTCACGGGCGCATCAGGAGCCATAGGAGGGGCTATCGCCCGCGTGCTGCATGCTCAGGGC...
MetPheGluLeuThrGlyArgLysAlaLeyValThrGlyAlaSerGlyAlaIleGlyGlyAlaIleAlaArgValLeuHisAlaGlnGly...

                                                                             Sph I
                                                      ATGACCATGATTACGCCAAGCTTGCATGCTCAGGGC...
                                                         ThrMetIleThrProSerLeuHisAlaGlnGly...

                                                                           Pst I
                                       ATGACCATGATTACGAATTCCCGGGGATCCGTCGACCTGCAGGCTCAGGGC...
                                          ThrMetIleThrAsnSerArgGlySerValAspLeuGlnAlaGlnGly...
```

FIGURE 4.—Use of protein fusions to confirm *nodG* open reading frame. (A) Coupled transcription-translation was conducted with an *E. coli* extract as described by EGELHOFF and LONG (1985). Lanes 1 and 2: plasmid pRmF51 in the presence and absence of 0.05 mM cAMP, respectively. Lanes 3–6, plasmid pRmS25 under the following conditions: lane 3, in the absence of cAMP and IPTG; lane 4, in the presence of 0.5 mM cAMP; lane 5, in the presence of 1 mM IPTG; lane 6, in the presence of 0.5 mM cAMP and 1 mM IPTG. Lanes 7 and 8: plasmids pRmF48 and pRmF49, respectively, in the absence of cAMP. The sizes (in amino acid residues) of the native and fusion proteins are indicated at the right. (B) DNA sequence and deduced amino acid sequence of native *nodG* (*top line*) present in pRmF48 and pRmF49; of the pRmS25 fusion (*middle line*); and of the pRmF51 fusion (*bottom line*). Details of construction of pRmS25 and pRmF51 are presented in Materials and Methods. Relevant restriction sites at which fusions were generated are underlined and overlined. *lacZ* sequences are italicized. Common *lacZ* sequences between the two fusions are underlined.

may indicate that the RNA starts further upstream than the end of the DNA template (pRmF32) used to generate the comparison sequencing ladder. In the latter situation, this would be consistent with *nodG* being part of the *nodFE* operon.

## DISCUSSION

The regulation, expression and function of genes involved in nodulation are not yet understood mechanistically. Analysis of open reading frames and transcripts provides an initial set of clues to the function and regulation of these genes. The four ORFs described here agree exactly with those reported recently by DEBELLÉ and SHARMA (1986) for the sibling strain *R. meliloti* 2011. Another recent DNA sequence, determined by HORVATH *et al.* (1986) in *R. meliloti* strain 41, largely agrees with that shown here for *nodH* and *nodE*; they designate these genes as *hsnD* and *hsnB*, respectively. In contrast, substantial differences, including alternative choice of reading frames in portions of the sequences, exist between our *nodG* and *nodF* sequence and the (*nodG*) *hsnC* and (*nodF*) *hsnA* sequence of HORVATH *et al.* (1986). We used repeated sequencing of *nodG* and translational fusions to confirm the *nodG* sequence presented here for strain 1021. A single bp difference in the *nodF* (*hsnA*) sequences of strains 1021 and 41 results in a frameshift which completely accounts for the differences in the carboxyl third of the *nodF* proteins.

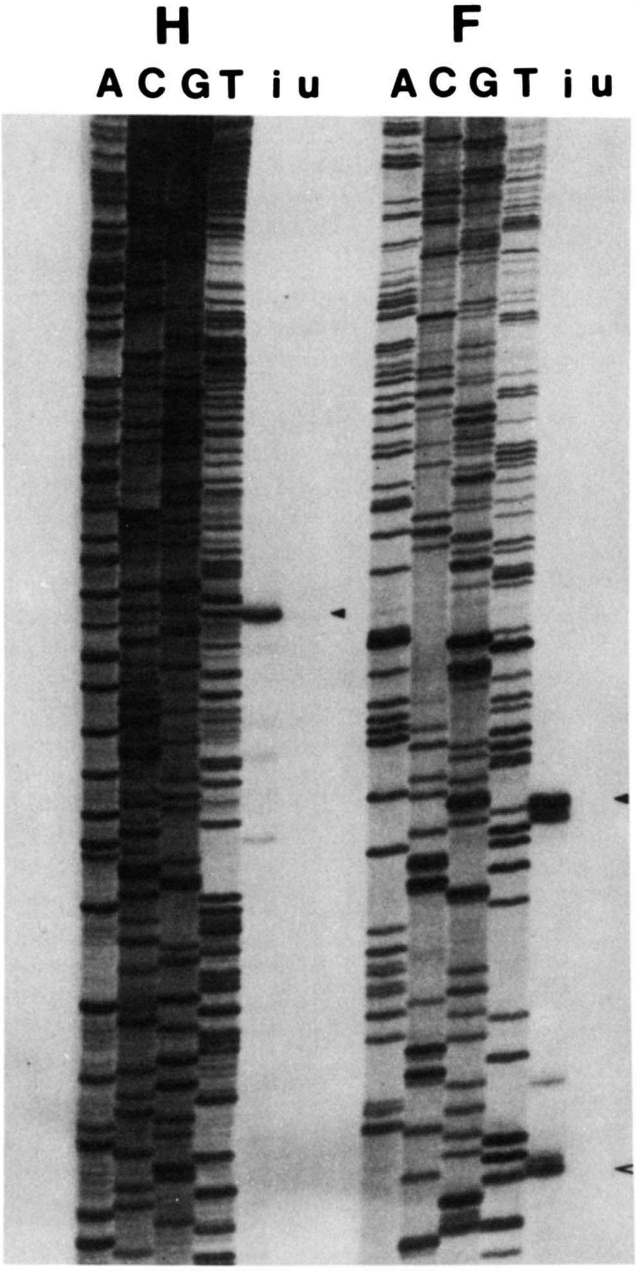


FIGURE 5.—Primer extension to determine transcription start sites for *nodF* and *nodH*. DNA primers complementary to 15-nucleotide segments within the structural genes of *nodF* and *nodH* were used to direct DNA synthesis complementary to transcripts isolated from luteolin-induced (i) and uninduced (u) *R. meliloti*. Products were separated by electrophoresis on sequencing gels adjacent to dideoxy-termination sequencing ladders generated using the same oligomers as primers on appropriate single stranded DNA templates. (*Left*) A single major transcript start site for *nodH* mRNA is indicated (*arrow*). Minor bands seen at lower molecular weights are sometimes more prominent. (*Right*) Four potential start sites are seen for the *nodF* transcript (*arrow*) (see summary on sequence, Figure 2A). Prominent lower molecular weight bands (*open arrow*) correspond in position to an inverted repeat in the sequence of the *nodF* leader, which could cause formation of an RNA hairpin secondary structure (Figure 2A).

DNA sequence determinations have been made for *nod* genes in *R. leguminosarum* (SHEARMAN *et al.* 1986) and *R. trifolii* (SCHOFIELD and WATSON 1986), to which *nodF* and *nodE* from *R. meliloti* show substantial homology. SHEARMAN *et al.* (1986) have pointed out homology between the deduced amino acid sequence of *R. leguminosarum nodF* and of acyl carrier protein from *E. coli* and barley; homologous sequences are largely conserved in *R. meliloti nodF* as well. While one potential function for the *nodF* gene product might be in lipid synthesis or modification, it has been recently reported that *E. coli* acyl carrier protein functions in the synthesis of an extracellular $\beta$-1,2-glucan (THERISOD, WEISSBORN and KENNEDY 1986). Since this molecule is present in *Rhizobium* and *Agrobacterium* (PUVANESARAJAH *et al.* 1985), both of which stimulate abnormal plant growth, a role for a specialized acyl carrier protein in glucan biosynthesis should be investigated. No DNA sequence homologies to other known genes are obvious for *nodE*; a hydropathy analysis (KYTE and DOOLITTLE 1982) of the predicted amino acid sequence indicates it to be largely hydrophobic (grease index = +0.10). Thus *nodE* joins the ranks of other *nod* gene proteins likely to be membrane localized (JACOBS, EGELHOFF and LONG 1985; JOHN *et al.* 1985; EVANS and DOWNIE 1986). The *nodG* amino acid sequence was shown by DEBELLÉ and SHARMA (1986) to have homology to that of ribitol dehydrogenase of *Klebsiella pneumoniae*; its hydropathy also reveals substantial hydrophobic character (grease index = +0.09). Using the FASTP protein comparison program (LIPMAN and PEARSON 1985), we found significant *nodG* amino acid sequence homologies to alcohol dehydrogenase and glucose dehydrogenase as well. The *nodH* protein coding region has a very unusual feature, in that the polypeptide has a high proline content (21 out of 245 residues). This may give rise to a protein with an unusual tertiary structure; however, the polypeptide expressed *in vitro* from the *nodH* clone migrates with the expected mobility in SDS-polyacrylamide gels (Figure 3).

An *in vitro* transcription-translation expression system from *R. meliloti* was useful in identifying and defining protein products of the *nod* genes. In these experiments, two exogenous promoters were used to direct expression of the *nod* genes *in vitro*. We had previously shown that *R. meliloti* RNA polymerase could efficiently initiate transcription from an enteric *trp* promoter (FISHER *et al.* 1987). In addition, we also utilized the *E. coli lac* promoter to direct *nod* gene expression in the presence of *R. meliloti* extracts. However, we found that addition of cAMP to the extract, which greatly enhances *lac* promoter function in *E. coli* (ZUBAY, SCHWARTZ and BECKWITH 1970), has no detectable effect on *in vitro* use of the *lac* promoter by the *R. meliloti* extract (data not shown). It is possible that the catabolite activation protein (cAMP receptor protein) which complexes with cAMP, and binds near and enhances function of the *lac* promoter in *E. coli*, does not exist in *Rhizobium* or does not bind cAMP or the *E. coli* target DNA sequence.

The *Rhizobium* extract directs the synthesis of *nodF* and *nodE* gene products and also what is probably a *nodG* fusion protein from pRmS15 (Figure 3, lane 7). Since in pRmS15 the *nodF* translation start site is over 600 bp downstream from the vector *lac* transcription start site, it is possible that the transcript which directs synthesis of these polypeptides originates not from the *lac* promoter, but from a sequence within the *nod* gene clone. The *nodF* regions studied in *R. leguminosarum* and *R. trifolii* are transcribed only in flavone-induced cells (SHEARMAN *et al.* 1986; REDMOND *et al.* 1986), and it would be unexpected and interesting to observe transcription arising from a *nodF* promoter in an *in vitro* extract isolated from noninduced cells. Towards this end, we are determining the transcript start sites for these *in vitro* products. This constitutes a first step toward analyzing the factors involved in inducible *nod* promoter function. The expression of a possible *nodG* fusion protein from plasmid pRmS15 suggests either that it is expressed from its own promoter, or that the *nodF-nodE in vitro* transcript may read through to *nodG*. Although it has been proposed by ROSTAS *et al.* (1986) that *nodG* (*hsnC*) must have its own promoter due to the Nod$^+$ phenotype of transposon insertions between *nodE* (*hsnB*) and *nodG* (*hsnC*), this can only be confirmed by transcription analysis. Our studies of RNA from this region failed to detect a transcription start site between *nodE* and *nodG*, and left open the possibility that transcription initiates much further upstream.

ROSTAS *et al.* (1986) cloned and sequenced six segments of the *R. meliloti* strain 41 genome which displayed considerable homology in a 50-bp region. Three of these segments lay upstream of *R. meliloti nod* genes *nodA*, *nodF* (*hsnA*) and *nodH* (*hsnD*). SCHOFIELD and WATSON (1986), SCOTT (1986), and SHEARMAN *et al.* (1986), studying *R. trifolii*, *Bradyrhizobium* sp. (Parasponia), and *R. leguminosarum*, respectively, also observed this highly conserved sequence upstream of *nodA* and *nodF* in these species. This sequence ("*nod* box") has been postulated to regulate co-ordinately *nod* gene function (ROSTAS *et al.* 1986), but until now this function remained speculation since no transcription initiation sites had been determined for any inducible *nod* genes. In this study, our primer extension mapping demonstrated that the transcript start sites for *nodH* and *nodF* lie downstream of each *nod* box by 28- and 26-bp, respectively (Figure 2, A and B). In work to be published elsewhere, the *nodA* transcript start site has also been mapped at a similar distance from its *nod* box. The features of these three promoters and their behavior when regulated by an additional locus, *syrM*, are discussed in a separate study (J. T. MULLIGAN and S. R. LONG, unpublished data). The position of the transcript start sites in relation to the *nod* box is consistent with the idea that the *nod* box functions as an upstream regulatory sequence. However, the *nod* box is centered further upstream than the usual consensus sequence for a prokaryotic RNA polymerase binding site (MCCLURE 1985; REZNIKOFF *et al.* 1985). The transcription start sites demonstrated here for *nodF* and *nodH* appear to rule out the involvement of the sequences homologous to the *nif* promoter in regulation of expression (ROSTAS *et al.* 1986), since these sequences lie within the transcribed leader of at least one of the genes. Thus, in *Rhizobium*, the nature of a regulatory element analogous to the *E. coli* −10 consensus sequence remains to be determined for *nod* gene promoters.

The phenotype for one transposon insertion mutant, 710, which lies between the *nodF nod*-box and transcription start site for *nodF*, shows a significant reduction in the number of nodules formed compared to wild-type (SWANSON *et al.* 1987). However, another mutant, 109, which lies one base inside the transcript leader for *nodH*, has a wild-type nodulation phenotype. Transposon insertions mapped by HORVATH *et al.* (1986) to positions just upstream of *nodH* and *nodF* translation start sites, and thus likely to lie in the transcript leader region, also have completely Nod$^+$ phenotypes. This is likely due to the documented nonpolarity of Tn*5* (CORBIN, BARRAN and DITTA 1983; HORVATH *et al.* 1986, MULLIGAN and LONG 1985). This observation reinforces the importance of conducting detailed RNA and protein analyses to accompany genetic studies of regulation.

We are grateful to M. ORBACH for gifts of pUC118, pUC119, and M13K07, and for his helpful advice on their use to generate single stranded DNA for sequencing; T. EGELHOFF, M. YELTON and P. GOLLNICK for materials and suggestions on coupled transcription/translation experiments; S. KUSTU and J. KEENER for suggestions on S-30 protocols; and C. YANOFSKY for advice on mRNA isolation. The expert technical assistance of M. WORTHINGTON is greatly appreciated. We also thank F. DEBELLÉ for communicating unpublished DNA sequence data, and gratefully acknowledge the exchange of unpublished information from J. DENARIE, A. KONDOROSI and his colleagues, J. A. DOWNIE, A. W. B. JOHNSTON, M. A. DJORDJEVIC, K. SCOTT, J. WATSON, and B. G. ROLFE. We are indebted to N. FEDERSPIEL, H. BRIERLEY, M. YELTON and K. PETERS for comments on the manuscript, and A. BLOOM for her skill in its preparation. This research was supported by contract DE-AS03-82-ER12084 from the Department of Energy Division of Biological Energy Research.

## LITERATURE CITED

Amersham Corporation, 1983 M13 Cloning and Sequencing Handbook. Amersham Corp., Arlington Heights, Ill.

BELASCO, J. G., J. T. BEATTY, C. W. ADAMS, A. VON GABAIN and S. N. COHEN, 1985 Differential expression of photosynthesis genes in *R. capsulata* results from segmental differences in stability within the polycistronic *rxcA* transcript. Cell **40:** 171–181.

CORBIN, D., L. BARRAN and G. DITTA. 1983 Organization and expression of *Rhizobium meliloti* nitrogen fixation genes. Proc. Natl. Acad. Sci. USA **80:** 3005–3009.

DEBELLÉ, F. and S. B. SHARMA, 1986 Nucleotide sequence of *Rhizobium meliloti* RCR2011 genes involved in host specificity of nodulation. Nucleic Acids Res. **14:** 7453–7472.

DUSHA, I., J. SCHRODER, P. PUTNOKY, Z. BANFALVI and A. KONDOROSI, 1986 A cell-free system from *Rhizobium meliloti* to study the specific expression of nodulation genes. Eur. J. Biochem. **160:** 69–75.

EGELHOFF, T. T. and S. R. LONG, 1985 *Rhizobium meliloti* nodulation genes: identification of *nodDABC* gene products, purification of *nodA* protein, and expression of *nodA* in *Rhizobium meliloti*. J. Bacteriol. **164:** 591–599.

EGELHOFF, T. T., R. F. FISHER, T. W. JACOBS, J. T. MULLIGAN and S. R. LONG, 1985 Nucleotide sequence of *Rhizobium meliloti* 1021 nodulation genes: *nodD* is read divergently from *nodABC*. DNA **4:** 241–248.

EVANS, I. J. and J. A. DOWNIE, 1986 The *nodI* gene product of *Rhizobium leguminosarum* is closely related to ATP-binding bacterial transport proteins; nucleotide sequence analysis of the *nodI* and *nodJ* genes. Gene **43:** 95–101.

FIRMIN, J. L., K. E. WILSON, L. ROSSEN and A. W. B. JOHNSTON, 1986 Flavonoid activation of nodulation genes in *Rhizobium* reversed by other compounds present in plants. Nature **324:** 90–92.

FISHER, R. F., H. L. BRIERLEY, J. T. MULLIGAN and S. R. LONG, 1987 Transcription of *Rhizobium meliloti* nodulation genes: identification of a *nodD* transcription initiation site *in vitro* and *in vivo*. J. Biol. Chem. **262:** 6849–6855.

GUNSALUS, R. P., G. ZURAWSKI and C. YANOFSKY, 1979 Structural and functional analysis of cloned deoxyribonucleic acid containing the *trpR-thr* region of the *Escherichia coli* chromosome. J. Bacteriol. **140:** 106–113.

HENIKOFF, S., 1984 Unidirectional digestion with exonuclease III creates targeted breakpoints for DNA sequencing. Gene **28:** 351–359.

HORVATH, B., E. KONDOROSI, M. JOHN, J. SCHMIDT, I. TÖRÖK, Z. GYÖRGYPAL, I. BARABAS, U. WIENEKE, J. SCHELL and A. KONDOROSI, 1986 Organization, structure and symbiotic function of *Rhizobium meliloti* nodulation genes determining host specificity for alfalfa. Cell **46:** 335–343.

INNES, R. W., P. L. KUEMPEL, J. PLAZINSKI, H. CANTER-CREMERS, B. G. ROLFE and M. A. DJORDJEVIC, 1985 Plant factors induce expression of nodulation and host-range genes in *Rhizobium trifolii*. Mol. Gen. Genet. **201:** 426–432.

JACOBS, T. W., T. T. EGELHOFF and S. R. LONG, 1985 Physical and genetic map of a *Rhizobium meliloti* nodulation gene region and nucleotide sequence of *nodC*. J. Bacteriol. **162:** 469–476.

JOHN, M., J. SCHMIDT, U. WIENEKE, E. KONDOROSI, A. KONDOROSI and J. SCHELL. 1985 Expression of the nodulation gene *nodC* of *Rhizobium meliloti* in *Escherichia coli*: role of the *nodC* gene product in nodulation. EMBO J. **4:** 2425–2430.

KYTE, J. and R. F. DOOLITTLE, 1982 A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. **157:** 105–132.

LAEMMLI, U. K., 1970 Cleavage of structural proteins during the assembly of the head of bacteriophage T4. Nature **227:** 680–685.

LIPMAN, D. J. and W. R. PEARSON, 1985 Rapid and sensitive protein similarity searches. Science **227:** 1435–1441.

MANIATIS, T., E. F. FRITSCH and J. SAMBROOK, 1982 *Molecular Cloning*. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.

MCCLURE, W. R., 1985 Mechanism and control of transcription initiation in prokaryotes. Annu. Rev. Biochem. **54:** 171–204.

MESSING, J., 1983 New M13 vectors for cloning. Methods Enzymol. **101:** 20–78.

MULLIGAN, J. T. and S. R. LONG, 1985 Induction of *Rhizobium meliloti nodC* expression by plant exudate requires *nodD*. Proc. Natl. Acad. Sci. USA **82:** 6609–6613.

PETERS, N. K., J. W. FROST and S. R. LONG, 1986 A plant flavone, luteolin, induces expression of *Rhizobium meliloti* nodulation genes. Science **233:** 977–980.

PUVANESARAJAH, V., F. M. SCHELL, G. STACEY, C. J. DOUGLAS and E. W. NESTER, 1985 Role for 2-linked-β-D-glucan in the virulence of *Agrobacterium tumefaciens*. J. Bacteriol. **164:** 102–106.

REDMOND, J. W., M. BATLEY, M. A. DJORDJEVIC, R. W. INNES, P. L. KUEMPEL and B. G. ROLFE, 1986 Flavones induce expression of nodulation genes in *Rhizobium*. Nature **323:** 632–634.

REZNIKOFF, W. S., D. A. SIEGELE, D. W. COWING and C. A. GROSS, 1985 The regulation of transcription initiation in bacteria. Annu. Rev. Genet. **19:** 355–387.

ROSENBERG, C., F. CASSE-DELBART, I. DUSHA, M. DAVID and C. BOUCHER, 1982 Megaplasmids in the plant-associated bacteria *Rhizobium meliloti* and *Pseudomonas solanacearum*. J. Bacteriol. **150:** 402–406.

ROSSEN, L., C. A. SHEARMAN, A. W. B. JOHNSTON and J. A. DOWNIE, 1985 The *nodD* gene of *Rhizobium leguminosarum* is autoregulatory and in the presence of plant exudate induces the *nodA, B, C* genes. EMBO J. **4:** 3369–3373.

ROSTAS, K., E. KONDOROSI, B. HORVATH, A. SIMONCSITS and A. KONDOROSI, 1986 Conservation of extended promoter regions of nodulation genes in *Rhizobium*. Proc. Natl. Acad. Sci. USA **83:** 1757–1761.

SANGER, F., S. NICKLEN and A. R. COULSON, 1977 DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA **74:** 5463–5467.

SCHOFIELD, P. R. and J. M. WATSON, 1986 DNA sequence of *Rhizobium trifolii* nodulation genes reveals a reiterated and potentially regulatory sequence preceding *nodABC* and *nodFE*. Nucleic Acids Res. **14:** 2891–2903.

SCOTT, K. F., 1986 Conserved nodulation genes from the nonlegume symbiont *Bradyrhizobium* sp. (*Parasponia*). Nucleic Acids Res. **14:** 2905–2919.

SHEARMAN, C. A., L. ROSSEN, A. W. B. JOHNSTON and J. A. DOWNIE, 1986 The *Rhizobium leguminosarum* nodulation gene *nodF* encodes a polypeptide similar to acyl-carrier protein and is regulated by *nodD* plus a factor in pea root exudate. EMBO J. **5:** 647–652.

SWANSON, J. A., J. K. TU, J. OGAWA, R. SANGA, R. F. FISHER and S. R. LONG, 1987 Extended region of nodulation genes in *Rhizobium meliloti* 1021. I. Phenotypes of Tn*5* insertion mutants. Genetics **117:** 181–189.

THERISOD, H., A. C. WEISSBORN and E. P. KENNEDY, 1986 An essential function for acyl carrier protein in the biosynthesis of membrane-derived oligosaccharides of *E. coli*. Proc. Natl. Acad. Sci. USA **83:** 7236–7240.

WILLIAMS, J. G. and P. J. MASON, 1985 Hybridization in the analysis of RNA. pp. 139–160. In: *Nucleic Acid Hybridization: A Practical Approach*, Edited by B. D. HAMES and S. J. HIGGINS. IRL Press, Washington, D.C.

ZUBAY, G., D. SCHWARTZ and J. BECKWITH, 1970 The mechanism of action of catabolite sensitive genes. Cold Spring Harbor Symp. Quant. Biol. **35:** 433–435.

ZUBAY, G., D. E. MORSE, W. J. SCHRENK and J. H. MILLER, 1972 Detection and isolation of the repressor protein for the tryptophan operon of *Escherichia coli*. Proc. Natl. Acad. Sci. USA **69:** 1100–1103.

Communicating editor: J. ROTH