Volume 12 Number 24 1984 Nucleic Acids Research

Nucleotide sequence of *Rhizobium meliloti* nodulation genes

I.Török[1], E.Kondorosi[1], T.Stepkowski[2], J.Pósfai[3] and A.Kondorosi[2]

[1]Institutes of Biochemistry, [2]Genetics and [3]Biophysics, Biological Research Center of Hungarian Academy of Sciences, H-6701 Szeged, POB 521, Hungary

Received 27 September 1984; Revised and Accepted 20 November 1984

ABSTRACT

A Rhizobium meliloti DNA region, determining nodulation functions common in different Rhizobium species, has been delimited by directed Tn5 mutagenesis and its nucleotide sequence has been determined. The sequence data indicates three large open reading frames with the same polarity coding for three proteins of 196, 217 and 402 (or 426) amino acid residues, respectively. We suggest the existence of three nod genes on this region, which were designated as nodA, B and C, respectively. Comparison of the R.meliloti nodA, B, C nucleotide and amino acid sequences with those from R.leguminosarum, as reported in the accompanying paper, shows 69-72% homology, clearly demonstrating the high degree of conservation of common nod genes in these Rhizobium species.

INTRODUCTION

Rhizobium meliloti induces nitrogen-fixing root nodules on alfalfa (Medicago sativa). Genes required for nodulation (nod) and nitrogen fixation (fix), including the structural genes for the enzyme nitrogenase (nif) are carried by a very large plasmid (megaplasmid) in this bacterium (1, 2). Transfer of this sym plasmid into other Rhizobium species or into Agrobacterium tumefaciens resulted in transconjugants that formed ineffective nodules on Medicago sativa, indicating that genes coding for the early steps of nodulation and host range specificity are located on this plasmid (3, 4). Close linkage of the nod and nif genes was demonstrated in several R.meliloti strains (1, 2, 5) and physical maps of the nod-nif regions have been established (6, 7).

In R.meliloti strain 41 the essential nod genes were localized in two clusters, located about 25 and 13 kb downstream from the nifHDK operon on a 8.5 kb and on a 6.8 kb EcoRI fragment, respectively (8). Using directed Tn5 mutagenesis, a *nod* gene cluster of about 2.5-3.0 kb was found within the 8.5 kb region. In Nod$^-$ deletion mutants lacking the 8.5 kb fragment the nodulation ability on alfalfa was restored upon the introduction of *sym* plasmids of *R.leguminosarum* or *R.trifolii* (1, 8). Moreover, the *nod* region of the 8.5 kb fragment hybridized with *nod* genes from other rhizobia (9, 10). These results suggested that the 8.5 kb fragment contains *nod* genes determining functions necessary for nodulation of a wide range of legume hosts ("common" *nod* genes). The other, 6.8 kb region contains two *nod* gene regions (8), which probably determine host specificity of nodulation (*hsn* genes).

To understand in more detail the organization and regulation of nodulation genes at molecular level, we have determined the nucleotide sequence of the common *nod* gene region which was precisely delimited beforehand by directed Tn5 mutagenesis. The sequence data revealed 3 large open reading frames, all with the same polarity, in agreement with the recently demonstrated protein coding regions using *E.coli* minicells or a cell-free system (10). We suggest that this common *nod* region contains 3 genes, which were designated as *nodA*, *B* and *C*, respectively.

The sequence data were compared with those for 3 *nod* genes of *R.leguminosarum* reported in the accompanying paper (11).

MATERIALS AND METHODS

Strains and plasmids

AK631 is a compact colony morphology variant of the wild type *R.meliloti* 41. *Escherichia coli* HB101 (*pro* *leu* *thi* *lac*Y *endo*I *rec*A *hsd*S *Str*^{r}) was used for plasmid transformation and a derivative of HB101 carrying a Tn5 insertion in the chromosome at an unknown location for Tn5 mutagenesis. *E.coli* strain NM512 (*sup*^{o}), obtained from Dr. N. Murray (Edinburgh, U.K.), harboured the target plasmids for transposon mutagenesis, using bacteriophage λ::Tn5 (12) as a source of Tn5. The components of the broad host range cloning system, plasmids pRK290 and pRK2013, were used as cloning vector and for mobilization of cloned DNA (13). pPH1, a broad host range plasmid coding for gentamicin resistance (14) was used for marker-exchange of the

Tn5-carrying fragments with the wild type region. Plasmid pKSK5 (15) carries the 8.5 kb *Eco*RI *nod* fragment in pRK290. Subclones of the 8.5 kb fragment (10) were constructed in pACYC184 (16).

Enzymes and isotopes

Restriction endonucleases and other enzymes were prepared in this laboratory except *Hinf*I (kindly provided by Dr. M. Hartmenn, Jena) and *Ava*II, *Ava*I, *Aos*I, *Asu*II (from Dr. M. Szekeres, Szeged, Hungary).

Bacterial alkaline phosphatase was purchased from Worthington. γ-^{32}P ATP (~ 1000-3000 Ci/mmol) was the product of the Isotope Institute (Budapest, Hungary).

Directed Tn5 mutagenesis of the *nod* fragment

Directed Tn5 mutagenesis on pKSK5 was carried out as described earlier (8). Since the mutagenesis was not completely random and gaps remained, subclones of the 8.5 kb *Eco*RI fragment were constructed in pACYC184 and the directed Tn5 mutagenesis was done on these subclones as described (10). Tn5 insertions on the sections of the 8.5 kb fragment were precisely mapped, and recloned in pRK290.

In order to test the effect of the Tn5 insertions on the Nod phenotype, the pRK290 derivatives carrying Tn5 were introduced into AK631 using a triparental mating system (13) and the wild type region was replaced by homologous recombination (8, 17) with the homologous region carrying Tn5 at various locations and it was shown by Southern hybridization (18) that Tn5 was located exactly at the same site of the 8.5 kb fragment as in the recombinant plasmid. The phenotype of each recombinant was investigated in alfalfa plant test on 10-15 plants in separate tubes and 3-5 times repeated as described earlier (19).

Plasmid DNA and restriction fragment preparation

This was done as described previously (20).

DNA sequence determination

Sequencing was done essentially according to the method of Maxam and Gilbert (21). The dephosphorylated DNA fragments were labelled at their 5' ends with γ-^{32}P ATP and T4 polynucleotide kinase. The labelled fragments, after a second restriction, were fractionated on acrylamide gels (5-10 %)and eluted from the gel slices by electrophoresis into dialysis bags. DNA sequencing reactions (A>C, G, C+T and C) and gel electrophoresis were carried out as described earlier (20).

Computer analysis of sequence data

Data handling and analysis of the sequence were performed by a self-prepared FORTRAN program package on a PDP/compatible minicomputer.

RESULTS AND DISCUSSION

Delimitation of *nod* genes on the 8.5 kb *Eco*RI fragment

In previous studies, based on the analysis of 17 Tn5 insertions a rough correlated physical-genetic map of the 8.5 kb *Eco*RI fragment was established (8). In order to define the *nod* region more precisely, 61 new Tn5 insertions were generated randomly by directed Tn5 mutagenesis and mapped (Fig. 1) in the same way as before (8, 10). Although Tn5 has a very low insertional specificity (22) "hot spots" on the 8.5 kb region have been observed (Fig. 1). The 8.5 kb *Eco*RI fragment of the wild--type *R.meliloti* 41 (AK631) was replaced by each mutated fragment via homologous recombination and each Tn5 insertion derivative of AK631 was tested for its symbiotic property. Figure 1 shows that Nod$^-$ mutations are located on a contiguous 2.6-3.0 kb region. The majority of the Nod$^-$ mutants were unable to induce nodule formation on *Medicago sativa* even two months after inoculation, in contrast to the wild-type AK631 which nodulated the host plant after 10 days. These mutants were unable to evoke root hair curling, as observed for other Nod$^-$ mutants, mapped in this region (8).

Two mutants mapped at the left end of the *nod* region showed a delayed Nod$^+$ phenotype : after inoculation with these strains nodules appeared with about one week delay. Tn5 insertions outside the *nod* region resulted in Nod$^+$ Fix$^+$ or Nod$^+$ Fix$^-$ phenotypes.

Sequencing of the DNA region carrying the common *nod* genes

The correlated physical-genetic map of the 8.5 kb *Eco*RI fragment (Fig. 1) shows that all Nod$^-$ mutations mapped so far, are located between the *Bam*HI site and the outermost *Sst*II site on the right. Therefore, the sequence of a 3373 bp long DNA fragment including this region was determined. In order to avoid

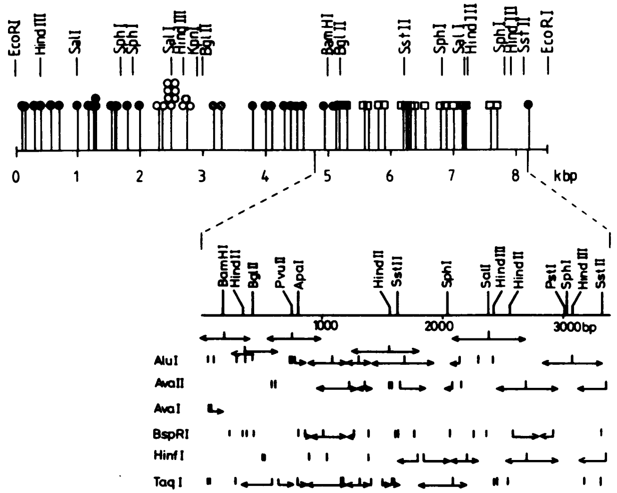

Fig. 1 Physical-genetic map of the 8.5 kb EcoRI fragment and strategy for sequencing of common nod genes.Symbiotic genes on the 8.5 kb fragment (upper part of the figure) were defined by directed Tn5 mutagenesis. Vertical lines designate locations of Tn5 insertions. Closed circles, Nod$^+$Fix$^+$; open circles, Nod$^+$Fix$^-$; hatched circles, Nod$^+$Fix$^\pm$; open squares, Nod$^-$; hatched squares, delayed Nod$^+$ (with about a week). Restriction mapping data from (8, 10). The DNA region carrying nod genes is enlarged (lower part of the figure) with a more complete restriction map and the scheme of the sequencing strategy of the common nod genes. The arrows represent the extent and direction of the determined DNA sequences.

ambiguities, almost all of the 3373 bp was sequenced in both strands (Fig. 1). The DNA sequence presented on Figure 2 starts 180 bp left of the BamHI site and extends to the second SstII site (Fig. 1).

Using a computer search, five open reading frames (i.e. nucleotide sequence between two stop codons in a given frame), larger than 500 bp appeared. Three of them are located on the same strand, in 5′ → 3′ direction, and we positioned them between the nucleotide coordinates 347-1300, 1279-1950 and 1911-3245, respectively (Fig. 2). As the co-ordinates show, the first two open reading frames in the 5′ → 3′ strand overlap. The two longest open reading frames of the complementary strand occur between the co-ordinates 3143-2465 and 471-1. This last one did not terminate in the sequenced region.

Analysis of the sequenced DNA was primarily focused on the

```
  -1     AAAATCAACATCCCCGCGCCGGAGAAGCTCATGGGGATCATCATCGAGAGGCAGCAACTCGAGCTGACGCCGGGAGC
  78    CTCCCGAGCCACGCGTTCCACGATCCTCGCAAAGAATACAAGTATCATGAAATCCGAAAGGATGATCCTGAAACGGCGA
 157    TCCGACTGGGCACGGGTTTAGTGGATCCCAGGCAATGACGGAAAGCCGAATGTGCAGTAAGGCGTCGCGCACGGCTGGG
 236    GCAAGTGCCTCGGCAGCCGGTGTCGGGATAAGTTCGCGGCCCTGCATCGAAAACAGCTCGTCGCCGAAATAGGTGCGCA
 315    GGCGCGCGATAGCCGCGCTCATGGCCGGTTGACTGAGGTTGATCCGGCGTGCGGCGGCCGTGAGCTTGCGCTCGGTCAT
 394    CAGTGCGTCGAGCGCGACGAGGAGGTTTAGATCTAGGCCCCTAAAACGCATGTGCGGCATCCATATCGCAGATGATCGT
 473    TATCCAAACAATCAATTTTACCAATCTTGCAGAGTCCTATTAGAGAACCCTGAAGTTAATGGAATCAAGGTGCGGCGCG
 552    AGAAAAGTTTCACAAGTACAGGATGGGTCCGAATTTTGAGCCGTCATCTAAGCGCTCGACCAACGGTCCAGCGCTACGG
 631    TTGGCGTCCCCGGTGTAACTTGCCGGGTACACACCACTCTCGATCGTGCTTTGAAGAAACAACACACTGGAGTTCTTAC

 710  ATG TCC TTA AAA GTG CAG TGG AAG CTA TGC TGG GAA AAT CAG CTG GAA CGT GCA GAC CAC
       M   S   L   K   V   Q   W   K   L   C   W   E   N   Q   L   E   R   A   D   H

 770  CAG GAG CTC TCA GAA TTT TTT CGA AAA TCC TAT GGG CCC ACA GGA GCG TTC CAC GCG AAA
       Q   E   L   S   E   F   F   R   K   S   Y   G   P   T   G   A   F   H   A   K

 830  CCA TTT GAG GGT GGC CGC AGT TGG GCC GGC GCG AGA CCG GAA CGC CGC GCA ATT GCT TAC
       P   F   E   G   G   R   S   W   A   G   A   R   P   E   R   R   A   I   A   Y

 890  GAC TCG GTC GGG ATA GCA AGC CAC ATG GGC GTG TTG CGC CGT TTC ATT AAG GTT GGT GAG
       D   S   V   G   I   A   S   H   M   G   V   L   R   R   F   I   K   V   G   E

 950  ACT GAT CTC CTT GTG GCT GAA CTG GGC TTA TAC GCG GTG CGG CCC GAT CTG GAG CGA ATG
       T   D   L   L   V   A   E   L   G   L   Y   A   V   R   P   D   L   E   R   M

1010  GGC ATC GCT CAC TCG GTC GGT GCT TTG ACT CCA ACT TTG CGG GAG CTT GGT GTC CCA TTC
       G   I   A   H   S   V   G   A   L   T   P   T   L   R   E   L   G   V   P   F

1070  GCC TTT GGG ACA GTT CGG CAC GCC ATG CGG AAC CAC GTT GAG AGA TAT TGC CAA AAC GGT
       A   F   G   T   V   R   H   A   M   R   N   H   V   E   R   Y   C   Q   N   G

1130  ATG GCT AGC ATT TTG ACG GGG GTT CGA GTG CGG TCG AGC ATC GCA GAG GTG AAC GCC GAT
       M   A   S   I   L   T   G   V   R   V   R   S   S   I   A   E   V   N   A   D

1190  CTC CCT TCC ACG CGC ACC GAG GAC CCA CTC GTC GTG ATA TTC CCG GTT GGA CGT CCG TTG
       L   P   S   T   R   T   E   D   P   L   V   V   I   F   P   V   G   R   P   L

1250  AAC GAA TGG CCG CCA GGT ACA TTG ATT GAA CGG AAC GGA TCG GAG CT
       N   E   W   P   P   G   T   L   I   E   R   N   G   S   E   L
                                                                  ATG AAG CAC CTC
                                                                   M   K   H   L

1309  GAT TAC ATA CAC GAG ATG CCG AGC AAC TGC GAT TAC GGG ACC GAA GAT CGT AGT ATA TAC
       D   Y   I   H   E   M   P   S   N   C   D   Y   G   T   E   D   R   S   I   Y

1369  CTG ACG TTT GAC GAC GGC CCG AAT CCA CAT TGC ACA CCG GAA ATC CTC GAT GTG CTG GCT
       L   T   F   D   D   G   P   N   P   H   C   T   P   E   I   L   D   V   L   A

1429  GAA TAC GGC GTG CCG GCG ACT TTC TTC GTC ATC GGC ACC TAT ACG AAA AGC CAG CCG GAA
       E   Y   G   V   P   A   T   F   F   V   I   G   T   Y   T   K   S   Q   P   E

1489  CTC ATT CGA CGT ATC GTC GCG GAA GGT CAC GAA GTG GCT AAC CAC ACG ATG ACC CAC CCG
       L   I   R   R   I   V   A   E   G   H   E   V   A   N   H   T   M   T   H   P

1549  GAC CTG TCA ACA TGT GGA CCT CAC GAA GTC GAA CGT GAG ATT GTC GAG GCA AGT GAG GCC
       D   L   S   T   C   G   P   H   E   V   E   R   E   I   V   E   A   S   E   A

1609  ATT ATC GCC GCT TGT CCT CAG GCC GCG GTC CGA CGC ATA CGA GCA CCT TAT GGT GTC TGG
       I   I   A   A   C   P   Q   A   A   V   R   R   I   R   A   P   Y   G   V   W

1669  AGC GAG GAA GCT CTG ACA AGA TCG GCA AGC GCT GGG CTG ACG GCA ATA CAT TGG TCG GCA
       S   E   E   A   L   T   R   S   A   S   A   G   L   T   A   I   H   W   S   A

1729  GAT CCG CGA GAT TGG TCT CGG CCA GGC GCC AAC GCG ATT GTT GAT GCA GTG CTG GAC TCG
       D   P   R   D   W   S   R   P   G   A   N   A   I   V   D   A   V   L   D   S

1789  GTT CGG CCC GGT GCA ATC GTG CTG TTG CAC GAT GGG TGC CCT CCC GAC GAA TCG GGA GCG
       V   R   P   G   A   I   V   L   L   H   D   G   C   P   P   D   E   S   G   A

1849  CTT ACG GGT CTG CGT GAC CAA ACG CTT ATG GCG CTT TCC CGT ATC GTC CCG GCG CTG CAT
       L   T   G   L   R   D   Q   T   L   M   A   L   S   R   I   V   P   A   L   H

1909  GAG CGT GGT TTT GCA ATT CGC CCA CTT CCT CCG CAT CAC TGAACAGACGAGAACCC ATG TAC
       E   R   G   F   A   I   R   P   L   P   P   H   H                   M   Y

1971  CTG CTT GAC ACA ACC AGC ACC GCC GCT ATC TCA ATC TAC GCG CTG CTC TTG ACC GCC TAC
       L   L   D   T   T   S   T   A   A   I   S   I   Y   A   L   L   L   T   A   Y
```

```
2031  AGG AGC ATG CAA GTC CTA TAT GCT CGG CCG ATA GAC GGT CCA GCA GTG TCG GCA GAA CCG
      R   S   M   Q   V   L   Y   A   R   P   I   D   G   P   A   V   S   A   E   P

2091  GTC GAG ACC CGC CCT CTG CCA GCC GTG GAT GTT ATC GTC CCC AGC TTC AAT GAG GAC CCA
      V   E   T   R   P   L   P   A   V   D   V   I   V   P   S   F   N   E   D   P

2151  GGC ATC CTC TCG GCG TGC CTC GCG TCC ATT GCA GAC CAG GAT TAT CCT GGA GAA TTG CGA
      G   I   L   S   A   C   L   A   S   I   A   D   Q   D   Y   P   G   E   L   R

2211  GTC TAT GTC GTT GAT GAT GGT TCT CGG AAC CGC GAG GCC ATT GTG CGT GTA CGC GCC TTC
      V   Y   V   V   D   D   G   S   R   N   R   E   A   I   V   R   V   R   A   F

2271  TAT TCG CGC GAT CCG AGG TTC AGC TTC ATT CTG CTC CCA GAG AAC GTC GGA AAG CGG AAA
      Y   S   R   D   P   R   F   S   F   I   L   L   P   E   N   V   G   K   R   K

2331  GCG CAG ATT GCC GCG ATA GGC CAA TCC TCT GGG GAT TTG GTG CTG AAT GTC GAC TCG GAC
      A   Q   I   A   A   I   G   Q   S   S   G   D   L   V   L   N   V   D   S   D

2391  AGC ACG ATC GCT TTC GAT GTG GTC TCC AAG CTT GCC TCG AAG ATG CGA GAT CCA GAG GTC
      S   T   I   A   F   D   V   V   S   K   L   A   S   K   M   R   D   P   E   V

2451  GGT GCG GTT GTG GGT CAA CTC ACG GCT AGC AAT TCG GGT GAC ACT TGG CTG ACT AAA TTG
      G   A   V   V   G   Q   L   T   A   S   N   S   G   D   T   W   L   T   K   L

2511  ATC GAC ATG GAG TAT TGG CTT GCC TGT AAC GAA GAA CGC GCG GCA CAG TCT CGC TTC GGT
      I   D   M   E   Y   W   L   A   C   N   E   E   R   A   A   Q   S   R   F   G

2571  GCT GTT ATG TGT TGC TGC GGC CCT TGT GCT ATG TAC CGT CGG TCG GCG CTC GCT TCG CTG
      A   V   M   C   C   C   G   P   C   A   M   Y   R   R   S   A   L   A   S   L

2631  CTA GAC CAG TAC GAA ACG CAA CTG TTT CGC GGT AAG CCA AGC GAC TTC GGT GAG GAC CGC
      L   D   Q   Y   E   T   Q   L   F   R   G   K   P   S   D   F   G   E   D   R

2691  CAT CTG ACG ATT CTC ATG TTG AAG GCA GGC TTT CGA ACT GAG TAC GTT CCA GAC GCC ATA
      H   L   T   I   L   M   L   K   A   G   F   R   T   E   Y   V   P   D   A   I

2751  GTG GCA ACC GTC GTC CCG GAT ACG CTG AAA CCA TAT CTG CGC CAA CAA CTG CGT TGG GCA
      V   A   T   V   V   P   D   T   L   K   P   Y   L   R   Q   Q   L   R   W   A

2811  CGC AGC ACG TTC CGT GAC ACG TTT CTA GCG CTC CCT CTG TTG CGC GGC CTC AGC CCT TTT
      R   S   T   F   R   D   T   F   L   A   L   P   L   L   R   G   L   S   P   F

2871  CTC GCA TTT GAC GCG GTC GGA CAG AAT ATC GGG CAA CTG TTG CTC GCC CTG TCG GTG GTG
      L   A   F   D   A   V   G   Q   N   I   G   Q   L   L   L   A   L   S   V   V

2931  ACG GGT CTT GCG CAT CTC ATA ATG ACC GCC ACA GTG CCA TGG TGG ACA ATT TTG ATT ATT
      T   G   L   A   H   L   I   M   T   A   T   V   P   W   W   T   I   L   I   I

2991  GCG TGC ATG ACC ATT ATA CGC TGC AGC GTC GTA GCA TTG CAT GCT CGC CAA CTT AGA TTT
      A   C   M   T   I   I   R   C   S   V   V   A   L   H   A   R   Q   L   R   F

3051  CTT GGC TTC GTT CTG CAC ACA CCC ATC AAC CTC TTT CTC ATA CTT CCG CTG AAA GCT TAT
      L   G   F   V   L   H   T   P   I   N   L   F   L   I   L   P   L   K   A   Y

3111  GCG TTG TGT ACA TTG TCC AAT AGC GAC TGG CTG TCA CGC TAC TCC GCG CCA GAA GTA CCA
      A   L   C   T   L   S   N   S   D   W   L   S   R   Y   S   A   P   E   V   P

3171  GTC AGC GGG GGA AAG CAG ACC CCA ATT CAA ACC TCC GGT CGA GTG ACA CCT GAC TGC ACT
      V   S   G   G   K   Q   T   P   I   Q   T   S   G   R   V   T   P   D   C   T

3231  TGC AGC GGC GAG TGACAGTAGCATGACTGGAAACGGGCGAGTTTTGAGACAGGAAGCGGAAAATCAATTGTCAGA
      C   S   G   E

3306  TCGTGAGATGGCCCAAGATCTCCGCGGTGGCTTGAGCCGAGTCCGTTCGAATGGAAGGACCAAACAGT
```

Fig. 2 Nucleotide sequence of the R.meliloti common nod genes (nodA, B and C) and the deduced amino acid sequences (single letter code).
Sequences conserved in R.leguminosarum are underlined (for R.leguminosarum nodA, B, C sequences see accompanying paper) (11).

left to right directed strand containing the three open reading frames, since these were compatible with previous results where three proteins could be mapped from this DNA strand in E.coli minicells and in E.coli coupled transcription/translation system with molecular weights of 23, 28.5 and 44 kd, respectively (10). In these experiments the proteins were synthesized only when this region was placed behind a strong E.coli promoter. Recently a coupled transcription/translation system was established using R.meliloti cell-free extract by Dusha et al. (unpublished). In this system the number and size of proteins expressed from this region were the same as in the E.coli system and again no expression was found from their own promoter(s), unless the promoter was supplied by the vector.

Translation initiation codons and evaluation of the protein coding regions

In the first open reading frame (between co-ordinates 347-1300) there are two ATG triplets in phase. The first one is located at position 443 and the second at 710. The size and position of the 23 kd polypeptide is consistent with the ATG at position 710 being the translational initiation codon of the open reading frame. The calculated molecular weight of this polypeptide is 21840 d consisting of 196 amino acid residues, as deduced from DNA sequence. This value, differs only slightly from the molecular weight of the polypeptide experimentally observed in E.coli minicells (23 kd).

A 32.3 kd protein could also be encoded in this open reading frame (from position 443 to 1300), which was sometimes detected in E.coli minicells (33 kd). Only the second ATG (position 710) is preceded by a Shine-Dalgarno (23) sequence (GGAG, at position 699). This sequence probably serves as a ribosomal binding site also in R.meliloti although direct experimental proof is still lacking.

The second open reading frame (between co-ordinates 1279 - - 1950), starts with an ATG at position 1297. It is particularly interesting, that this ATG is overlapped by the termination codon (TGA, position 1298) of the preceding open reading frame so that the ATGA sequence contains both the termination and initiation codons in separate reading frames. This overlapping termination-initiation codon sequence ATGA occurs frequently in bacteriophage λ (24) and it was found also in some overlapping bacterial genes (25, 26, 27).

The molecular weight of the polypeptide is 23756 d, consisting of 217 amino acid residues as calculated from the sequence, which show a discrepancy with the 28.5 kd polypeptide detected in this region by Schmidt et al. (1984). We do not know the reason for this difference. A Shine-Dalgarno sequence (GGAG, at position 1291) was also detected upstream of the ATG at position 1297.

The third and largest open reading frame (between co--ordinates 1911-3245) can encode a polypeptide of 44125 d consisting of 402 amino acid residues (as calculated from DNA sequence) and starts from an ATG at position 2036. This ATG is preceded by another ATG in phase (position 1965), which corresponds to a polypeptide of molecular weight of 46759 d, consisting of 426 amino acid residues. In E.coli minicells a 44 kd polypeptide was synthesized from this region (10). Both initiation codons are preceded by Shine-Dalgarno-like sequences (ACGAG at position 1955 and GGAG at position 2031). Therefore it is still not clear where the translation of the nodC gene product starts.

The existence of 3 large open reading frames, together with the 3 polypeptide coding region determined previously (10) suggests that the common nod region consists of 3 genes, which we designate nodA (with molecular weight of 21840 d), nodB (23756 d) and nodC (44125 d or nodC' 46759 d), respectively. Since nodA,B and C may form one transcriptional unit and Tn5 causes strong polar mutations, further genetic complementation analysis is required to support the existence of three nodulation genes.

The putative nodA gene product contains 2, the putative nodB product has 5 and both putative nodC products contain 11 cysteine residues. Interestingly, in the middle region of the nodC product, 4 cysteines are located very close to each other (4 out of 6 consecutive amino acid residues are cysteines), which is very likely to be important in determining either the structure or the function of this protein.

The amino acid sequences of nod gene products were analysed for regions of hydrophobicity (Fig. 3) by computer. The degree of hydrophobicity is calculated as the relative hydrophobicities

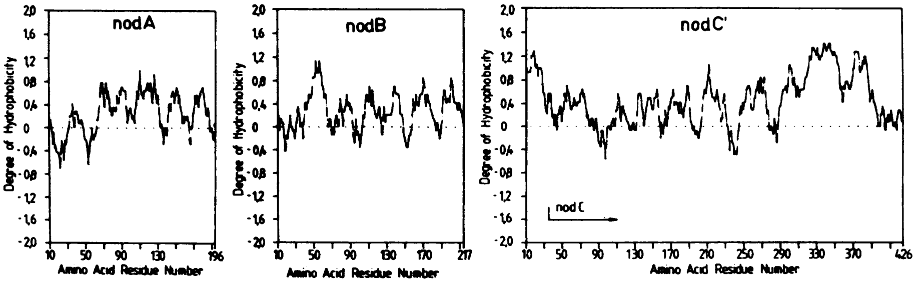

Fig. 3 Relative hydrophobicities of the putative nod gene products. The amino acid sequence of the protein (Fig. 2) was analysed for hydrophobic areas (a moving average of 14 amino acids residues) by a program prepared at our Institute. Higher values indicate greater hydrophobicity. nodC' corresponds to the 46759 d protein, nodC is the 44125 d protein.

of amino acids (28) with a moving average over 14 amino acid residues.

Although all proteins contain hydrophobic regions, it is apparent that the nodC gene product is highly hydrophobic. The most hydrophobic part of the nodC protein is the carboxy-terminal region between amino acid residue positions 306-348 with a hydrophobicity index of 2.22 according to (29) and between positions 362-380 with a hydrophobicity index of 2.74. Such arrangement of hydrophobic regions was found also in E.coli outer membrane proteins (30). The N-terminus of the putative 46759 d molecular weight polypeptide is also highly hydrophobic; this region is missing from the putative 44125 d molecular weight protein product of the nodC gene. The high hydrophobicity of the N-terminal region is characteristic for the signal peptide of proteins transported from the inner part of the cell (31); however, their hydrophobic region is preceded by positively charged amino acid(s), usually by lysine. Moreover, it is not clear whether the larger protein is the in vivo nodC gene product, which questions the significance of this finding. Nevertheless, on the basis of high hydrophobicity of the nodC gene product one may speculate that it interacts with the membrane. The other putative nod gene products have no hydrophobic N-terminal leader peptide regions and there is only one longer hydrophobic stretch of amino acid residues in nodA between position

Table 1 Codon utilization of the nodA, B, C and nifH genes in R.meliloti

		nodA	nodB	nodC	nodC'	nifH			nodA	nodB	nodC	nodC'	nifH			nodA	nodB	nodC	nodC'	nifH			nodA	nodB	nodC	nodC'	nifH
Phe	UUU	4	2	7	7	2	Ser	UCU	0	1	3	3	1	Tyr	UAU	2	2	7	7	3	Cys	UGU	0	2	4	4	0
	UUC	4	2	9	9	3		UCC	3	1	6	6	7		UAC	2	4	4	7	7		UGC	2	3	7	7	5
Leu	UUA	2	0	0	0	0		UCA	1	1	1	2	1	Stop	UAA	0	0	0	0	0	Stop	UGA	1	1	1	1	1
	UUG	6	1	10	11	1		UCG	4	4	9	9	2		UAG	0	0	0	0	0	Trp	UGG	4	3	6	6	0
	CUU	2	4	6	7	5	Pro	CCU	1	5	6	6	2	His	CAU	0	4	3	3	3	Arg	CGU	3	6	4	4	1
	CUC	4	3	13	14	13		CCC	2	2	2	2	1		CAC	6	8	1	1	5		CGC	5	2	14	14	7
	CUA	2	0	3	3	1		CCA	5	3	12	12	1	Gln	CAA	1	1	9	9	4		CGA	3	4	4	4	1
	CUG	3	9	16	18	10		CCG	4	9	5	5	5		CAG	3	2	6	6	5		CGG	6	2	4	4	2
Ile	AUU	4	5	10	10	3	Thr	ACU	3	1	4	4	1	Asn	AAU	1	1	5	5	4	Ser	AGU	1	2	0	0	0
	AUC	2	6	6	8	17		ACC	1	3	6	9	6		AAC	5	3	4	4	9		AGC	3	4	11	13	1
	AUA	2	4	6	6	1		ACA	3	3	5	6	3	Lys	AAA	3	1	4	4	2	Arg	AGA	2	1	1	1	2
Met	AUG	5	4	8	9	10		ACG	2	6	8	8	2		AAG	2	1	6	6	15		AGG	0	0	1	2	0
Val	GUU	5	2	6	6	2	Ala	GCU	5	5	8	9	4	Asp	GAU	3	8	9	9	5	Gly	GGU	6	5	10	10	7
	GUC	4	7	14	14	8		GCC	4	4	9	11	14		GAC	3	6	15	16	11		GGC	5	4	7	7	14
	GUA	0	0	3	3	4		GCA	4	8	9	9	7	Glu	GAA	7	10	6	6	10		GGA	3	2	4	4	2
	GUG	7	5	11	11	7		GCG	4	7	13	14	9		GAG	9	6	9	9	13		GGG	4	3	3	3	6

100-117 with hydrophobicity index of 1.44, and such a region was not found in the putative nodB gene product.

Codon usage

The codon usage (Table 1) of the nodA, B and C genes as well as of another sequenced gene of R.meliloti, nifH (20)is not random. The most significant asymmetry in codon usage was found for codons AGG (arg) and UUA (leu). In nodA, B and nifH AGG is not used and in nodC only in 9% of cases. UUA is not used in nodB, C and nifH and also only infrequently used in nodA (10%).

Another group of codons, such as UCU (ser),UCA (ser),AGU (ser) and CUA (leu), is not used in one or two of these genes (UCU in nodA, UGA in nodC and nifH) or used seldomly.

The third type of the codon is represented by those that are not utilized in one gene but normally used in an other one. Codon GUA (val) is normally used in nifH, and less frequently in nodC but not at all in nodA and nodB. The non-usage of codon CAU (his) seems to be specific for nodA, and UGU (cys) for nodA and nifH, since these codons are used quite frequently in the other genes.

It is difficult to explain the significance of non-randomness in codon usage but there are several examples where codon AGG (arg) is not used, for instance in several outer membrane proteins of E.coli (32).

Direction of transcription of nod genes

The direction of the 3 large open reading frames indicates that the nodA, B and C genes have the same polarity. The lack of longer sequences between the structural genes suggests that the 3 genes form one transcriptional unit. Moreover, no characteristic transcriptional terminator sequences were found in the analysed region. When the common nod region was placed behind a strong E.coli promoter, all the three polypeptides were produced (10).

It is likely that transcription of these nod genes starts from the 5′ flanking region of the nodA gene. No large open reading frame with the same polarity was found upstream from the nodA gene in the sequenced region. In the opposite direction (from right to left), however, a large open reading frame not terminating in the sequenced region was detected which may correspond to a gene transcribed from the other DNA strand.

At the 5′ flanking region of the nodA gene two Tn5 insertion mutants with delayed nodulation phenotype were mapped (Fig. 1). It is possible that these insertions affect the expression of nodA gene (and probably nodB and C genes as well), since Tn5 mutations further upstream from the nodA gene within the sequenced region had no detectable effect on nodulation (Fig. 1).

Unfortunately, we have not found the appropriate conditions where mRNA from the common nod region is synthesized at detectable levels. Lack of detectable expression of the common nod genes was found in studies with E.coli minicells and in a coupled transcription/translation system (10). Thus, the promoter region(s) of the nodA, B, C is still unknown.

Comparison of sequence data for the nodA, B, C genes of R.meliloti with that of R.leguminosarum

The nucleotide and amino acid sequence of nodA, B, C genes and of their gene products from R.leguminosarum are reported in this issue (11). Comparison of these data with those of R.meliloti (Fig. 2; conserved bases or amino acids are underlined) showed that the organization of the 3 genes is fairly similar and the nucleotide sequencesof nodA, B and C genes share 72%, 69% and 71.4% homology, respectively (Fig. 4). The determined nod gene sequences from the two Rhizobium species differ in deletions or insertions at six locations. In nodA a possible single nucleotide insertion or deletion (around position 820 in the R.meliloti sequence) resulted in a frame-

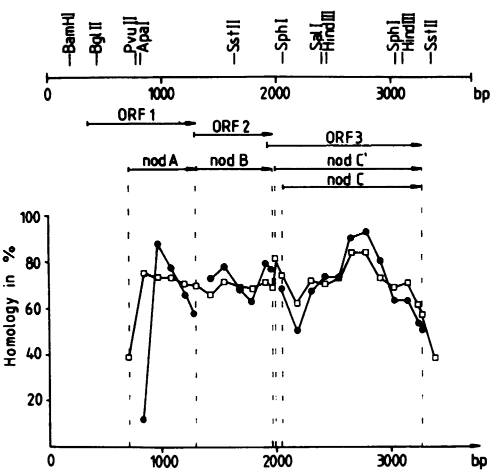

Fig. 4 Nucleotide and amino acid sequence homology between the common nod region of R.meliloti and R.leguminosarum. The homology was calculated as a percentage of the common nucleotides or amino acid residues found in R.leguminosarum in comparison to those in R.meliloti. Each point represents the homology of 120 nucleotides or 40 amino acid residues with the exception of the end of each gene where the homology percentage of the remaining nucleotides or amino acids was calculated. The physical map of R.meliloti common nod region and the open reading frames (ORF1, 2 and 3) corresponding to nodA, B and C (or C') genes are also shown. Squares: nucleotide sequence homology, circles: amino acid sequence homology. The first and last squares represent the nucleotide homology of the 5' and 3' flanking sequences of the common nod region.

shift mutation. Therefore the N-terminal regions of the putative nodA gene products differ. A frameshift mutation was found also at the C-terminal region of the same gene around position 1265. In the R.meliloti nodB sequence a six-base-pair deletion not altering the reading frame was identified between positions 1846 and 1847. Two three-base-pair insertions not influencing the reading frame were noticed in the nodC sequence of R.meliloti. Moreover a deletion between positions 3156 and 3157 generates a 19 bp long frame shift which is compensated by an insertion around position 3176.

The nucleotide sequences between the nod structural genes and a 38 bp region upstream from nodA (68%) are also highly conserved. The DNA region downstream from the nodC, however, shares much less homology.

The nodA, B and C gene products of R.meliloti and R.leguminosarum show 58%, 69% and 70% conservation of amino acid residues, respectively.
The middle regions of the three polypeptide sequences are highly conserved and the C-terminal regions are more diverged (Fig. 4). This type of conservation is fairly common as observed also for the nitrogenase reductase protein from R.meliloti and from other nitrogen-fixing organisms (20). The highest amino acid conservation was found in the nodC gene product where the homology for 78 amino acid residues was 95% (between amino acid residue positions 182 and 260). Since this region contains the highly clustered cysteine residues it is very likely that this region is an important part of the protein.

The fairly high sequence conservation for the nodA, B and C genes provides a convincing confirmation of our earlier suggestion, based on interspecies complementation and DNA hybridization experiments (1, 8, 10), that these nod genes are indeed common in different Rhizobium species. The nucleotide sequence of the nodC gene of R.meliloti strain 1021 was also determined recently and showed complete homology to the nodC gene of R.meliloti 41, except two base pair changes (S. Long and T. Jacobs, personal communication).

One must note that these 3 genes are probably not the only ones which are involved in the determination of common nodulation functions. Preliminary data indicate that a DNA region left to the nodA, B, C genes influences nodulation.
Although no Tn5 insertion mutations with clear Nod$^-$ phenotype have been localized on the left EcoRI-BamHI segment of the 8.5 kb fragment (Fig. 1), Nod$^-$ deletion mutants lacking the 8.5 kb fragment but containing the cloned common nod genes on a 3.6 kb AvaI-EcoRI fragment nodulate alfalfa only with one-two weeks delay.
Finally, the determined nod sequence data will be certainly helpful for the elucidation of the biochemical process and genetic control of nodule initiation and development which are still completely obscure.

ACKNOWLEDGEMENTS

We thank Drs. L.R. Rossen, A.W.B. Johnston, A. Downie and S. Long for exchanging informations prior to publication and for critical discussions; Drs. C. Kari and G.B. Kiss for criticism of the manuscript and Anton Józsefné for skilled technical assistance. One of us (T.S.) was a fellow of the International Training Course at the Biological Research Center, Szeged.

*To whom correspondence should be addressed

REFERENCES

1. Banfalvi, Z., Sakanyan, V., Koncz, C., Kiss, A., Dusha, I. and Kondorosi, A. (1981) Mol. Gen. Genet. 184, 318-325.
2. Rosenberg, C., Boistard, P., Denarie, J. and Casse-Delbart, F. (1981) Mol. Gen. Genet. 184, 326-333.
3. Kondorosi, A., Kondorosi, E., Pankhurst, C.E., Broughton, W.J. and Banfalvi, Z. (1982) Mol. Gen. Genet. 188, 433-439.
4. Wong, C.H., Pankhurst, C.E., Kondorosi, A. and Broughton, W.J. (1983) J. Cell Biol. 97, 787-794.
5. Long, S.R., Buikema, W.J. and Ausubel, F.M. (1982) Nature 298, 485-488.
6. Buikema, W.J., Long, S.R., Brown, S.E., van den Bos, R.C., Earl, C.D. and Ausubel, F.M. (1983) J. Mol. App. Genet. 2, 249-260.
7. Kondorosi, A., Kondorosi, E., Banfalvi, Z., Broughton, W.J., Pankhurst, C.E., Randhawa, G.S., Wong, C.H. and Schell, J. (1983) in Molecular Genetics of the Bacteria-Plant Interaction, Pühler, A. Ed., pp. 55-63. Springer-Verlag, Berlin--Heidelberg.
8. Kondorosi, E., Banfalvi, Z. and Kondorosi, A. (1984) Mol. Gen. Genet. 193, 445-452.
9. Kondorosi, A., Kondorosi, E., Banfalvi, Z., Dusha, I., Putnoky, P., Toth, J. and Bachem, C. (1984) in Proceedings of the XV. Int. Congress of Genetics, New Delhi (in press)
10. Schmidt, J., John, M., Kondorosi, E., Kondorosi, A., Wieneke, U., Schröder, G., Schröder, J. and Schell, J. (1984) EMBO J. 3, 1705-1711.
11. Rossen, L.R., Johnston, A.W.B. and Downie, J.A. (1984) Nucl. Acids. Res. (this issue).
12. Berg, D.E., Davies, J., Allet, B. and Rochaix, J. (1975) Proc. Natl. Acad. Sci. USA 72, 3628-3632.
13. Ditta, G., Stanfield, S., Corbin, D. and Helinski, D.R. (1980) Proc. Natl. Acad. Sci. USA 77, 7347-7351.
14. Beringer, J.E., Beynon, J., Buchanan-Wollaston, A.V. and Johnston, A.W.B. (1978) Nature 276, 633-634.
15. Kondorosi, E., Banfalvi, Z., Slaska-Kiss, K. and Kondorosi, A. (1983) in UCLA Symposia on Molecular and Cellular Biology, Plant Molecular Biology, Goldberg, R. Ed., pp. 259--275, A.R. Liss, Inc., New York.
16. Chang, A.C.Y. and Cohen, S.N. (1978) J. Bacteriol. 134, 1141-1156.

17. Ruvkun, G.B. and Ausubel, F.M. (1981) Nature, 289, 85-88.
18. Southern, E.M. (1975) J. Mol. Biol. 98, 503-517.
19. Kondorosi, A., Svab, Z., Kiss, G.B. and Dixon, R.A. (1977) Mol. Gen. Genet. 151, 221-226.
20. Török, I. and Kondorosi, A. (1981) Nucl. Acids Res. 21, 5711-5723.
21. Maxam, A. and Gilbert, W. (1977) Proc. Natl. Acad. Sci. USA 74, 560-564.
22. de Bruijn, F.J. and Lupski, J.R. (1984) Gene, 27, 131-149.
23. Shine, J. and Dalgarno, L. (1974) Proc. Natl. Acad. Sci. USA 71, 1342-1346.
24. Sanger, F., Coulson, A.R., Hong, G.F., Hill, D.F. and Petersen, G.B. (1982) J. Mol. Biol. 162, 729-773.
25. Yamamoto, T., Tamura, T., Yokota, T. and Takano, T. (1982) Mol. Gen. Genet. 188, 356-359.
26. Normark, S., Bergström, S., Edlung, T., Grundström, T., Jaurin, B., Lindberg, F.P. and Olsson, O. (1983) Ann. Rev. Genet. 17, 499-525.
27. Cole, S.T., Grundström, T., Jaurin, B., Robinson, J.J. and Weiner, J.H. (1982) Eur. J. Biochem. 126, 211-216.
28. Taylor, W.R. and Thornton, J.M. (1984) J. Mol. Biol. 173, 487-514.
29. Segrest, J.P. and Feldmann, R.J. (1974) J. Mol. Biol. 87, 853-858.
30. Overbeeke, N., Bergmans, H., van Mansfeld, F. and Lugtenberg, B. (1983) J. Mol. Biol. 163, 513-532.
31. Watson, M.E.E. (1984) Nucl. Acids Res. 12, 5145-5164.
32. Hackett, J. and Reeves, P. (1983) Nucl. Acids Res. 11, 6487-6495.