Volume 14 Number 7 1986 Nucleic Acids Research

DNA sequence of *Rhizobium trifolii* nodulation genes reveals a reiterated and potentially regulatory sequence preceding *nodABC* and *nodFE*

Peter R.Schofield[1] and John M.Watson*

CSIRO, Division of Plant Industry, GPO Box 1600, Canberra, ACT 2601, Australia

Received 13 February 1986; Accepted 13 March 1986

ABSTRACT

The Rhizobium trifolii nod genes required for host-specific nodulation of clovers are located on 14 kb of Sym (symbiotic) plasmid DNA. Analysis of the nucleotide sequence of a 3.7 kb portion of this region has revealed open reading frames corresponding to the nodABCDEF genes. A DNA sequencing technique, using primer extension from within Tn 5, has been used to determine the precise locations of Tn5 mutations within the nod genes and the phenotypes of the corresponding mutants correlate with their mapped locations. The predicted nodA and nodB genes overlap by four nucleotides and the nod F and nodE genes overlap by a single nucleotide, suggesting that translational coupling may ensure the synthesis of equimolar amounts of these gene products. The nodABC and nodFE genes constitute separate transcriptional units and each is preceded by a conserved 76-bp sequence which may be involved in the regulation of expression of these genes.

INTRODUCTION

The soil bacterium Rhizobium is capable of infecting leguminous plants and inducing a highly-differentiated structure, the root nodule, within which atmospheric nitrogen is reduced to ammonia by the bacteria. The symbiotic interaction is characterized by the high degree of specificity observed between temperate legumes and their corresponding fast-growing species of Rhizobium microsymbiont.

Bacterial genes required for nodulation (nod) and nitrogen fixation (nif or fix) are located on large Sym (symbiotic) plasmids in the fast-growing Rhizobium species. In R. trifolii and R. leguminosarum the host-specific nodulation genes are located on 14 kb and 10 kb of Sym plasmid DNA, respectively (1, 2), indicating that there are relatively few bacterial genes involved in nodule induction. In R. meliloti, two separate clusters of nod genes have been identified (3, 4).

We have determined the nucleotide sequence of a 3.7 kb portion of the R. trifolii strain ANU843 nodulation gene region. This region has been shown to contain parts of three nodulation gene transcriptional units by:

© IRL Press Limited, Oxford, England.

i) directed Tn5 mutagenesis (5); ii) directed MudI1734 mutagenesis (6) and iii) analysis of functions encoded on cloned restriction fragments (7).

The availability of a range of Tn 5-induced, nodulation-deficient R. trifolii mutants enabled us to utilize the Tn5 insertions as sites from which to directly access nod gene sequences. Previously, Tn 5 insertion sites have been determined by sequencing from the HpaI site, located 185 bp from the end of Tn5 (8, 9). To facilitate sequencing into flanking DNA, we synthesized an oligonucleotide complementary to the ends of Tn5, for use in chain-termination sequencing.

The nodulation genes, identified as open reading frames in the nucleotide sequence, were found to correlate with the locations of Tn 5 insertions known to result in either the loss of plant root hair curling ability or altered plant host range.

Analysis of the DNA sequence revealed the presence of two copies of a conserved 76-bp sequence. Hybridization analysis also indicates the presence of a third copy of this conserved sequence in the 14 kb nod gene region. The two copies, identified within the sequenced region are located at the 5' ends of two of the nod gene transcriptional units suggesting that these sequences may be involved in regulating the coordinate expression of these genes.

MATERIALS AND METHODS

Strains and plasmids

The 14 kb fragment of R. trifolii strain ANU843 nod DNA (pRt587), identified previously (9), was used as a source of DNA for sequence analysis. The nod::Tn 5 mutations used in this study were described by Djordjevic et al., (5): Strains referred to as A22 through O7 are strain ANU845 (pRt032::Tn5) insertions. Strain numbers 245, 246, 249, 252, 258, 274, 276, 277, 297 and 851 are derivatives of strain ANU843 containing Tn5 insertions in the Sym plasmid. M13 sequencing vectors mp18 and mp19 were used in E. coli strain JM107 (10).

DNA sequence determination and analysis

Standard molecular biology techniques for electrophoresis, DNA preparation and hybridization were carried out essentially as described previously (11).

The 3.7 kb EcoRI - BglII restriction fragment was sequenced by the chain-termination method (12), using M13 sequencing vectors. The DNA sequence was determined on both strands using overlapping Sau3A and TaqI

fragments in conjunction with sequence obtained in both directions from the EcoRI, PstI, BamHI, SmaI, SphI, ClaI and BglII sites indicated in Figure 5. The DNA sequence was compiled and analysed using SEQ and ANALYSEQ (13) programs.

Oligonucleotide synthesis

Both the universal M13 sequencing primer (3'-TGACCGGCAGCAAAATG-5') and the Tn5 sequencing primer (see below) (3'-TTCATCGCAGGACTTGC-5') were synthesized in our laboratory by the phosphoramidite method (14)

Tn5-mediated sequence determination

To undertake chain-termination sequencing from Tn5 into flanking Rhizobium DNA, a 17-mer oligonucleotide complementary to the sequence from nucleotides 16 to 32 in the arms of Tn5, was synthesized. This sequence was selected as it possessed least homology to other sequences in the arms of Tn5 (15). DNA fragments extending from HindIII sites within Tn5 to EcoRI, BamHI or BglII sites in flanking Rhizobium DNA were cloned into mp18 and detected by using γ-[^{32}P]ATP-labelled Tn5 primer. Tn5 primer-extended, chain-termination sequencing from a given insertion site allows DNA sequence to be determined on different strands in each direction such that the terminal 10 bp of Tn5 can be seen (see Fig. 2). In each instance, the two sequences can be aligned with respect to each other by virtue of the 9-bp duplication of host DNA generated by Tn5 insertion (15).

This method may be used to determine the nucleotide sequence of regions of DNA that have been genetically probed by Tn5 mutagenesis or for rapidly determining the genetic location of random Tn5 mutants of other bacterial species if homologous genes have already been characterized.

RESULTS AND DISCUSSION

DNA sequence of the R. trifolii nod genes

The DNA sequence of the EcoRI-BglII nod gene fragment was determined as outlined above and the 3668-bp nucleotide sequence is presented in Figure 1. Several open reading frames were identified, the predicted amino acid sequences of which are homologous to those of nod genes previously characterized in other Rhizobium species (16, 17, 4, 18). The six genes corresponding to these reading frames are designated nodA, B, C, D, E and F. Since we have been unable to detect the corresponding R. trifolii nod gene products (7), we utilized DNA sequence analysis programs that predict probable coding sequences on the basis of unequal

codon usage. In all cases, both Staden's (13) and Fickett's (19) programmes indicate that the open reading frames translated in Figure 1 are within coding regions of the DNA.

The predicted R. trifolii nod gene products (nodA, 21.6 kD; nodB,

```
GAATTCGAATCTCGCGTCGCCCGCATAGGCAAGCTGTTCGGCCACCACGGCGTCGCGATTTCTGGAACCGTCGTCAGCCACATAAACTCGCAACATCCCCGGATAACTCCGTGGTTGGCA   120
CTTAAGCTTAGAGCGCAGCGGGCGTATCCGTTCGACAAGCCGGTGGTGCCGCAGCGCTAAAGACCTTGGCAGCAGTCGGTGTATTTGAGCGTTGTAGGGGCCTATTGAGGCACCAACCGT
 F E F R A D G A Y A L Q E A V V A D R N R S G D D A V Y V R L M G P Y S R P Q C
     * *   *                         * *   * * * *   *       *   *   *

AGGGATGCCAGGCAATCGGACAGGACGCGCGGGACTCTCGTTGAACGAAGGCACGATGATGTCCACGCTTGGCAGAGAGCTCTCTTCGACGGTGTTTTCCGCTGTTTCCGACTTGAGAGT   240
TCCCTACGGTCCGTTAGCCTGTCCTGCGCGCCCTGAGAGCAACTTGCTTCCGTGCTACTACAGGTGCGAACCGTCTCTCGAGAGAAGCTGCCACAAAAGGCGACAAAGGCTGAACTCTCA
 P I G P L R V P R A P S E N F S P V I I D V S P L S S E E V T N E A T E S K L T
                     *   * * *   * * *   * *   *

TCGGAAAGCGTGAAGGACCTGTACACTCCTATAGGAGGAGGAAAGCAGTCCATAAAGCGAGACGGCGGCGATGCTGGTTGTTTCAAGCAAGGTCATGGTTTCTCGTCGATTTAGTTTTAC   360
AGCCTTTCGCACTTCCTGGACATGTGAGGATATCCTCCTCCTTTCGTCAGGTATTTCGCTCTGCCGCCGCTACGACCAACAAAGTTCGTTCCAGTACCAAAGAGCAGCTAAATCAAAATG
 R F A H L V Q V S R Y S S S L L G Y L S V A A I S T T E L L T M
 *   *     * *       *       *     *   *   *     * * *   * *   *

TGGGGAAGTGAACGAAGGGAAAATCCCACGGCTATGCAGTGCCGGGATAATCGCCAAGAGCGCCGAAAGTGTTTGATCACGCAGTCCGGTGAGCTTGCAGTTCCCGATCTCATCAGGTGGG   480
ACCCCTTCACTTGCTTCCCTTTTAGGTGCCGATACGTCACGGCCCTATTAGCGGTTCTCGCGGCTTTCACAAACTAGTGCGTCAGGCCACTCGAACGTCAAGGGCTAGAGTAGTCCACCC
Q P L S R L S F G R S H L A P I I A L L A S L T Q D R L G T L K C N G I E D P P
  * *         * * *     * * *         * *     * * * *                   *   * *

CAACCGTCGTGCAAAAGCACGATTGCACCTGGCCGAGCAGCAGCGAGTACCGCGTCAACGATAGAATCGACTCCGGGACGGGACCAGTCTCTCGGGTCTATCGACCAGTGAACAGGTCGA   600
GTTGGCAGCACGTTTTCGTGCTAACGTGGACCGGCTCGTCGTCGCTCATGGCGCAGTTGCTATCTTAGCTGAGGCCCTGCCCTGGTCAGAGAGCCCAGATAGCTGGTCACTTGTCCAGCT
C G D H L L V I A G P R A A A L V A D V I S D V G P R S W D R P D I S W H V P R
  * * * * * * * * * * *       * *     * *       * *     * * * * *   * * *

AGTCCGGCGTTCATCGATCTTGAAAGGACATCCGCGTTCCAAGCGCCGTATCGTCGTATATGTTGAACCGCAACTTGGGGACACGCAGAGACGATGGCTTCGTTTGTCTCCTTTATCTCT   720
TCAGGCCGCAAGTAGCTAGAACTTTCCTGTAGGCGCAAGGTTCGCGGCATAGCAGCATATACAACTTGGCGTTGAACCCCTGTGCGTCTCTGCTACCGAAGCAAACAGAGGAAATAGAGA
L G A N M S R S L V D A N W A G Y R R I H Q V A V Q P C A S V I A E N T E K I E
* * *     *     *         *   * *   *             * * * *     * *             *

                                          ▼F3                                                              ▼C2
CGTTCGATCGCCTCCGAGTCACAGGCCGTAAGATCGGGGTGGGTCATCGTGTGATTGGCGACCAAGTGACCTTCTGCTGCAACACGGCGGACAAGATCGGGGTGGTCTTTGACATGCCCA   840
GCAAGCTAGCGGAGGCTCAGTGTCCGGCATTCTAGCCCCACCCAGTAGCACACTAACCGCTGGTTCACTGGAAGACGACGTTGTGCCGCCTGTTCTAGCCCCACCAGAAACTGTACGGGT
R E I A E S D C A T L D P H T M T H N A V L H G E A A V R R V L D P H D K V H G
                    * * * * * * *   * * *   * * * *     * *   *   *

                                                                                                  ▼249
CCGAGGACGAAAAAGGTGGCTGGTACGCGGTGTTCGGCGAGCACATCCAGGATCTGGCCAGTGCAGTGCGGATTGGGACCGTCGTCGAACGTCAAGTAGATGTTACGATCGTCGCTTCCT   960
GGCTCCTGCTTTTTCCACCGACCATGCGCCACAAGCCGCTCGTGTAGGTCCTAGACCGGTCACGTCACGCCTAACCCTGGCAGCAGCTTGCAGTTCATCTACAATGCTAGCAGCGAAGGA
G L V F F T A P V R H E A L V D L I Q G T C H P N P G D D F T L Y I N R D D S G
*     * * * * * *         * * * *       * *   * * * * * * * * * *     *

GCCTCATCGAACGGCACTTCGCTTATATAAGCGCGCCGCTTCATAGCTCGCATCCGTTTCGTTCGATCAATGATCCAGAAGGCCACTCGCTCATTGGTCGTCCTATGGGAAATACCAGGA   1080
CGGAGTAGCTTGCCGTGAAGCGAATATATTCGCGCGGCGAAGTATCGAGCGTAGGCAAAGCAAGCTAGTTACTAGGTCTTCCGGTGAGCGAGTAACCAGCAGGATACCCTTTATGGTCCT
                             L E C G N R E I L S G S P W E S M P R G I P F V L V
A E D F P V E S I Y A R R K M                        *   *     * *     *       *
                            *

CGAGCACGTCTTCGGTGCGCGTGGAGGGCATATCCGGCAGCGCATCTGGAAGGGTCGAGCGTACACGAACCCCCGTGACGATATTGGAGATACCGTCTCGGCAGAATCTCTCCACGTGGT   1200
GCTCGTGCAGAAGCCACGCGCACCTCCCGTATAGGCCGTCGCGTAGACCTTCCCAGCTCGCATGTGCTTGGGGGCACTGCTATAACCTCTATGGCAGAGCCGTCTTAGAGAGGTGCACCA
 L V D E T R T S P M D P L A D P L T S R V R V G T V I N S I G D R C F R E V H N
 *   * *   *                         * * * *   *             *               *

TCCGCATAGCATGTCGGACGGTACCGAACGCGAACGGCACCGCCAGCTCCTGCAAAGCCGGAGCCAAAGCGCGGATCGAGTGAGGGATGCCAAGCCCCTCCAGATCCGGCCGCACGCCGT   1320
AGGCGTATCGTACAGCCTGCCATGGCTTGCGCTTGCCGTGGCGGTCGAGGACGTTTCGGCCTCGGTTTCGCGCCTAGCTCACTCCCTACGGTTCGGGGAGGTCTAGGCCGGCGTGCGGCA
 R M A H R V T G F A F P V A L E Q L A P A L A R I S H P I G L G E L D P R V G Y
     *   * * * * *   * * *   *     *   *           * *   * *     * * * * * *   *

ACAAGCCAAGTTCGGCGACCAGCAAATCGGTATCACCAACTTTTATGAATCGACGCAGCAACCCCATGTGACTTGCTATGCCTTTTGAATCGTAAGCAATTGCCCGGCGTTCCGGCCTCG   1440
TGTTCGGTTCAAGCCGCTGGTCGTTTAGCCATAGTGGTTGAAAATACTTAGCTGCGTCGTTGGGGTACACTGAACGATACGGAAAACTTAGCATTCGTTAACGGGCCGCAAGGCCGGAGC
 L G L E A V L L D T D G V K I F R R L L G M H S A I G K S D Y A I A R R E P R A
 * * * * * * * * *         * * * * * *   *   *       *     * *   * * *   * * * *

▼252                ▼H9
CACCCGCCCAACTGCGGCCAGTCTCGAAAGGTTTGGCATTGAATGCGCCAGTCGGACCATAGGTTGCTCGGAAAAATTCAGACAGTTCCGCATGATCCGACGGTTCGAGATCATTTTCCC   1560
GTGGGCGGGTTGACGCCGGTCAGAGCTTTCCAAACCGTAACTTACGCGGTCAGCCTGGTATCCAACGAGCCTTTTTAAGTCTGTCAAGGCGTACTAGGCTGCCAAGCTCTAGTAAAAGGG
 G A W S R G T E F P K A N F A G T P G Y T A R F F E S L E A H D S P E L D N E W
 *   * * *     * * *   *       *

                                                                                 ▼246
AGGTTATTTTCCACCGCACTCCAGCAGACATGCAAAGCTCCAATTCGTTCCTCAGCCAACAAGCCCAAAGCAGCAGAGATTGGTTCAGCGCACACGCTGATAGGACAGATGGGACCGCTT   1680
TCCAATAAAAGGTGGCGTGAGGTCGTCTGTACGTTTCGAGGTTAAGCAAGGAGTCGGTTGTTCGGGTTTCGTCGTCTCTAACCAAGTCGCGTGTGCGACTATCCTGTCTACCCTGGCGAA
 T I K W R V G A S M

GTCGTCTATTTGCGACAGATACGATCCGCCTCCGCTCGCTAGTTTGCTTTGGCGCCTTGATAAGTTCCGGCCTTTCTAATAAGCACTCCGAAAGATTGGTAAAATTGATTGTTTGGATCG   1800
CAGCAGATAAACGCTGTCTATGCTAGGCGGAGGCGAGCGATCAAACGAAACCGCGGAACTATTCAAGGCCGGAAAGATTATTCGTGAGGCTTTCTAACCATTTTAACTAACAAACCTAGC
                 * * *   * * * * * * * * * * * * * * * * *   * * * * * *   * * *
       ▼K7       M R F K G L D L N L L V A L D A L M T E R K L T A A A R S I N L
CAATCATCTACAGCGTGGATCGAGAATGCGTTTTAAGGGCCTGGATCTAAATCTTCTCGTGGCACTCGATGCTTTGATGACCGAGCGGAAGCTCACCGCGGCAGCGCGTAGCATCAACCT   1920
GTTAGTAGATGTCGCACCTAGCTCTTACGCAAAATTCCCGGACCTAGATTTAGAAGAGCACCGTGAGCTACGAAACTACTGGCTCGCCTTCGAGTGGCGCCGTCGCGCATCGTAGTTGGA
```

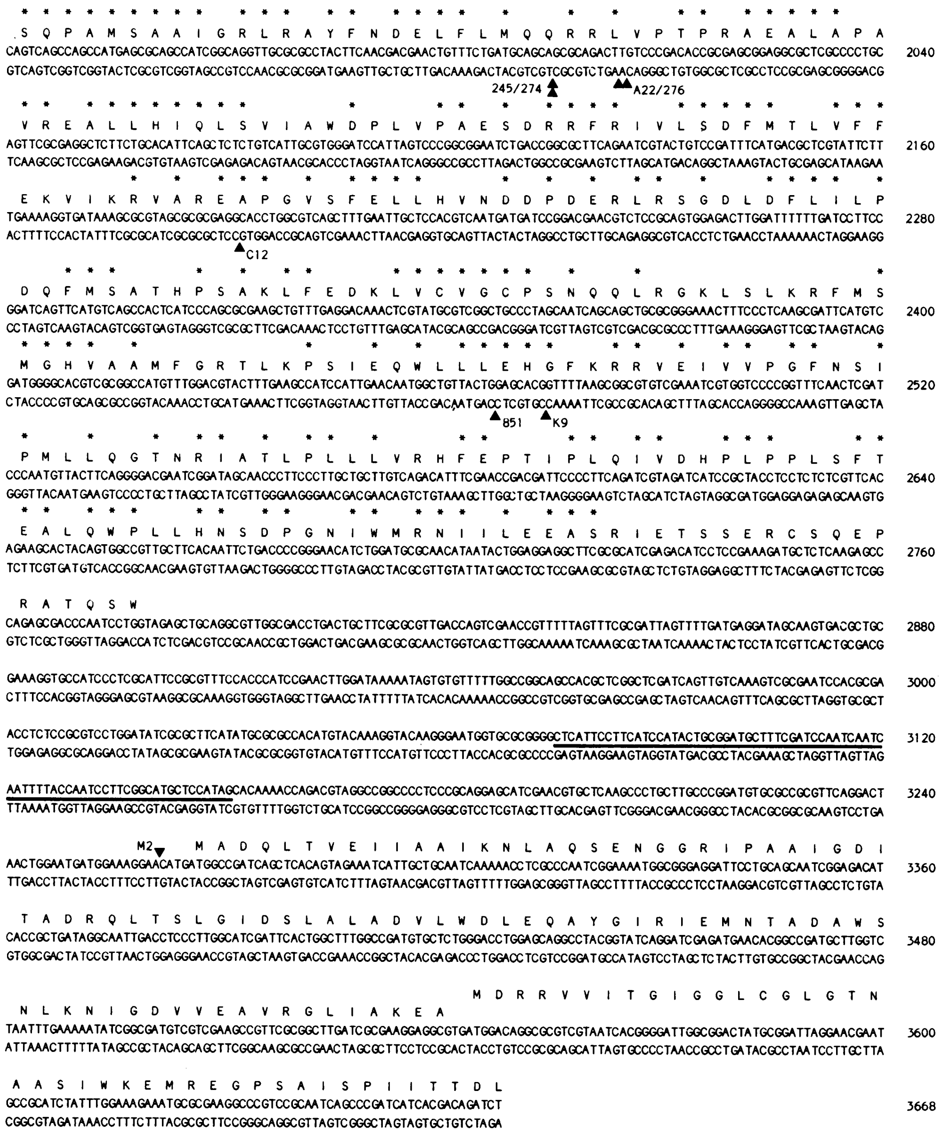

Figure 1. Nucleotide sequence of the 3668-bp EcoRI- BglII nod gene fragment of R. trifolii strain ANU843. The predicted amino acid sequences of the nod gene products are indicated (see Fig. 5 for summary). Asterisks indicate the invariant amino acids in the homologous gene products of R. trifolii, R. leguminosarum, R. meliloti and Parasponia Rhizobium sp. Arrowheads indicate the Tn5 insertion points which were determined by Tn5 primer-directed, chain-termination sequencing. The 75/76-bp conserved sequences, flanking the nodD gene, are indicated by solid lines drawn between the two strands.

23.5 kD; nodD, 36.0 kD; nodF, 9.8 kD and partial sequence of nodC and nodE) were compared to those of R. leguminosarum (16, 20), R. meliloti (17, 4, 21) and Parasponia Rhizobium (22). Invariant amino acids are indicated by asterisks in Figure 1. Comparisons of Rhizobium nod gene products reveal 70-80% amino acid homology between pairs of species. With 50-60% of the amino acids being invariant, between any pair of species, there remains sufficient homology between at least the nodA, B, C and D gene products to allow functional interspecies complementation (23, 3, 24, 25).

For the nodA gene, our data are consistent with that obtained from R. meliloti (17, 21). In both instances, the predicted nodA gene product is 196 amino acids long, whereas the proposed R. leguminosarum nodA gene product differs at both the amino- and carboxy-terminal ends, but these differences could be accounted for by single nucleotide insertions or deletions. In both R. leguminosarum and R. meliloti, several potential initiation codons are present at the beginning of the nodC coding sequences. Comparison with the R. trifolii nodC sequence, which has only one initiation codon, would suggest that the most likely translation start site for nodC in the other two species is the more upstream initiation codon.

Location of Tn5 insertions in nod genes

Fifteen independent Tn5 insertion points were determined (Fig. 2) and some insertional specificity was noted (see Fig. 1), as found in E. coli (26). The locations of these Tn 5 insertions, and their corresponding mutant phenotypes (5), correlate with the predicted nod gene coding regions. Insertions in nodA (246, 252, H9 and M7), nodB (249, M16, C21 and F3) and nodC (C8, L11, C10, L21, A29, L9, 277, K13, O7 and J7) result in a root hair curling deficient (Hac^-) and Nod^-phenotype. The four nodB mutants, however, induce some root hair distortions (5), which suggest that, not only are these mutants located in a second nodulation gene, but that nodC may have an independent promoter. The proximity of nodB and C (22-bp intergenic space) and the lack of any promoter-like sequences, do not support this conclusion. Mutants of nodD (245, 274, A22, 276, C12, 851 and K9) are all Nod^- and Hac^-. Two Tn5 insertions (246 and K7), located in the 232-bp intergenic region between nodD and nodA, are Nod^+. Both of these Nod^+ mutants, however, are located very close to the point of translational initiation (246 is 51 bp upstream of nodA; K7 is 18 bp upstream of nodD). Whilst the Tn 5 insertions have disrupted either leader or promoter sequences, the observed nodulation may have resulted from weak expression

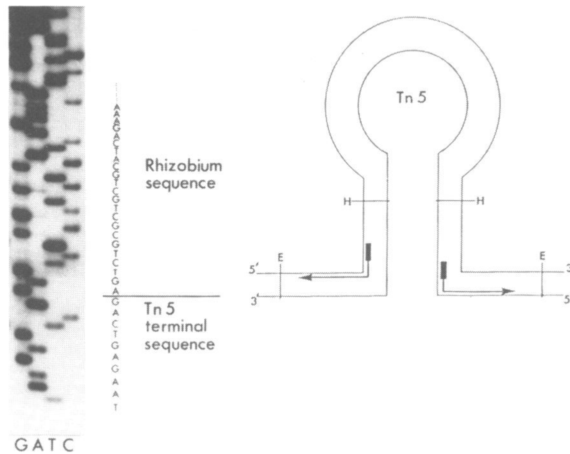

Figure 2. Tn 5 primer-directed sequencing strategy. DNA fragments, extending from the HindIII (H) sites within Tn 5 to EcoRI(E) (BamHI or BglII) sites in flanking R. trifolii DNA, were cloned into M13mp18. The synthetic 17-mer Tn 5 primer (see Materials and Methods) was then used to direct chain-termination sequencing reactions using single-stranded recombinant M13 DNA as templates.

of the nod genes from a promoter within the arms of Tn5. These mutants show slightly reduced and delayed nodulation (5) further suggesting that the Tn 5 insertions may have affected the transcription of these genes. In E. coli, low-level constitutive expression of the lac (27) and dnaG (28) genes has been observed downstream of Tn5 insertions. Weak transcription from a promoter within the arms of Tn5, is only detectable when the Tn5 insertion is located close to a ribosome binding site, thus preventing rho-dependent termination and allowing the expression of downstream genes (29).

Host range mutants (M2, M23, C7, 297, H7, K11 and 258) nodulate subterranean clover, peas and beans and (with the exception of M2) are unable to nodulate white clover (5). Mutants M2 (see below) and M23 appear to map within the nodF gene whilst the others (C7, 297, H7, K11 and 258) are thought to be located within the nodE gene. The amino-terminal portions of the nodE gene of R. trifolii and R. leguminosarum are homologous (20) and the nodE gene of R. leguminosarum encodes a 48 kD protein (18).

Mutant M2 has a unique phenotype, being able to nodulate white and subterranean clover, as well as peas and beans (5). In this mutant, Tn5 is located between the proposed ribosome binding site (30) and the initiation codon of the nodF gene (Fig. 3). This insertion results in a duplication of

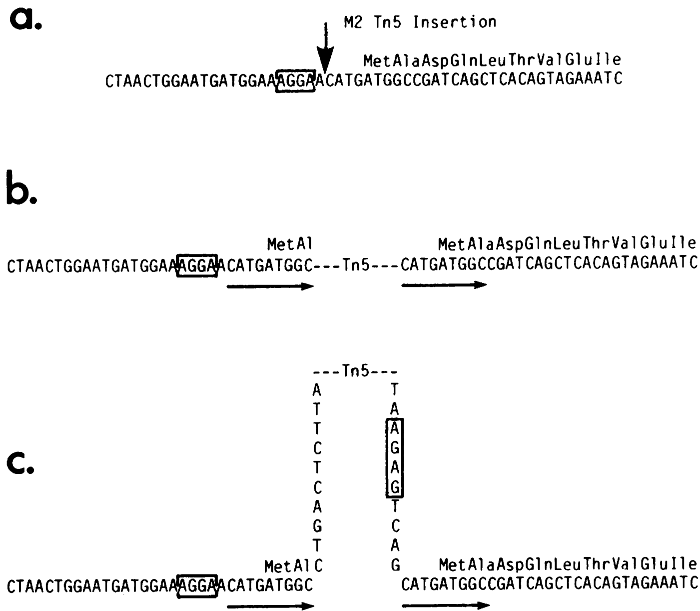

Figure 3. Nature of the Tn5-induced mutation in host-range mutant M2. (a) DNA sequence analysis (see Fig. 1) predicts an open reading frame (nodF) preceded by a ribosome-binding site (boxed) in the wild-type sequence. (b) Insertion of Tn5 results in the direct duplication of 9bp of host DNA (underscored arrows) (15). (c) The presence of a potential ribosome-binding site (boxed) in the arm of Tn5 may allow translation of the putative Tn5-promoted nodF transcript. The unique phenotype of mutant M2 (see text) may be a consequence of differential levels of transcription or differential translation due to differing intervals between the ribosome-binding site and initiation codon in the wild-type (5bp) and mutant M2 (8bp) sequences.

9 bp of host DNA which regenerates the nodF gene initiation codon. A potential ribosome binding site (AGAG) is located in the terminal nucleotides of Tn5 (Fig. 3), and therefore weak transcription from within the arms of Tn5 (see above) could allow translation of the nodF gene product. The ability of mutant M2 to nodulate white clover, unlike the other host range mutants, may be a consequence of the unique location of Tn5 in this derivative. For example, the extended host range of this mutant may result from plant-independent, Tn5 - promoted transcription of the nodFE genes.

R. trifolii host range genes have previously been identified by their ability to confer altered host range on other Rhizobium species (5,7). However, two putative host-range genes, nodF and nodE appear to be conserved between R. trifolii and R. leguminosarum (20), suggesting that

host range may be determined, not by widely-differing gene products, but by differences in the active sites of the nod gene products. The nodFE gene products may target or modify a bacterial signal for the appropriate host plant nodulation response. An incorrectly processed signal may elicit nodule initiation on legumes other than the normal host. The unique host range of mutant M2 may result, not from an incorrectly-processed signal, but from abnormal (i.e. Tn5-promoted) transcription of the nodFE genes.

Operon structure of nod genes

Previously, the nodA and nodB genes have been shown to overlap in R. meliloti by the sequence ATGA (17). We observe an identical overlap in the nodA and nodB genes of R. trifolii, at position 1001-1004. Additionally, we observe an overlap between the nodF translation stop and the nodE translation initiation codons TGATG at position 3542-3546. The presence of these overlapping sequences may indicate that translational coupling (31) occurs.

The presence of overlapping nodA-nodB and nodF-nodE genes, as well as the proximity of nodB and nodC, suggest that these R. trifolii nod genes are present in two separate operons, namely: nodABC and nodFE. The nodD gene coding sequence is read divergently from nodABC indicating that the nodD gene constitutes a separate operon. The presence of a terminator structure (32) (two inverted repeats located between positions 2815 and 2833, forming an 8 out of 9-bp stem and a 2-bp loop, followed by a T-rich region) at the 3' end of the nodD gene suggests that the nodFE operon is transcribed separately from the nodD gene. A second terminator - like structure is located within the coding region of nodD forming an 8-bp stem between positions 2244 and 2260 followed by a T-rich sequence (8 of 11 bases). The significance of this sequence is not clear, but it may be that nodD expression is controlled by this sequence via a mechanism such as rho-dependent termination (33).

Examination of the sequences located 5' of each of the nod gene operons does not reveal homology with either the E. coli (32) or nif (34) consensus promoter sequences.

The R. trifolii nod gene region contains two reiterated sequences, the first located between nucleotide 1756 - 1830 and the second between nucleotides 3076 - 3151. These two sequences are inverted with respect to each other and are 78% homologous. The sequences are located 235 bp upstream of the nodABC genes and 114 bp upstream of the nodFE genes. In R. meliloti there is an analogous sequence, located 192 bp upstream of the

Rhizobium trifolii	nodABC	CGCATTC-TCGATCCACGCTGTAGATGATTGCGATCCAAACAATCAATTTTACCAATCTTTCGGAGTGCTTATTAG
Rhizobium trifolii	nodFE	CTCATTCCTTCATCCATACTGCGGATGCTTTCGATCCAATCAATCAATTTTACCAATCCTTCGGCATGCTCCATAG
Rhizobium meliloti	nodABC	GCATGTGCGGCATCCATATCGCAGATGATCGTTATCCAAACAATCAATTTTACCAATCTTGCAGACT-CCTATTAG
Consensus		-----T-----ATCCA----G--GATG-T----ATCCAA-CAATCAATTTTACCAATCTT-C-G--T-C----TAG

Figure 4. Comparison of the conserved 75/76-bp nucleotide sequences preceding the nodABC and nodFE genes in R. trifolii and R. meliloti. The consensus sequence indicates the nucleotides conserved in the three copies, including a highly-conserved (96%) 27-bp core sequence (underlined).

nodABC genes (21). A comparison of the R. meliloti sequence with those of R. trifolii reveals a total of 59% homology between all copies (Fig. 4). In a core sequence of 27 bp, the homology is 96%. Additionally, this conserved sequence is found preceding the nodKABC genes of a slow-growing Rhizobium sp. isolated from the non-legume Parasponia (22).

To determine if additional copies of this sequence occur in the 14 kb nod region, the 557-bp PstI fragment (positions 2785-3342) was hybridized to a Southern blot of various restriction digests of pRt587 DNA. Two additional fragments hybridized with this probe. The first carries the reiterated sequence preceding the nodABC genes whilst the second carries a region that includes other nod genes (6) (Region IV in Figure 5).

Regulation of nod gene expression

To examine the possibility of this reiterated sequence being involved in the regulation of nod gene expression, the locations of the operons defined by DNA sequence analysis (Fig. 5) were correlated with transcriptional and regulatory information derived from R. trifolii (6). Using Mud lac fusions, Innes et al. (6) identified four R. trifolii transcriptional units. One of these, corresponding to the nodD gene, is constitutively expressed. The other three are activated in the presence of clover root exudate. These plant-activatable transcriptional fusions correlate with the locations of the reiterated nod-gene consensus sequences on the 5' sides of the plant-inducible nodABC and nodFE genes, as well as in the vicinity of the Region IV gene(s).

In R. meliloti and R. leguminosarum, the nodD gene has been shown to be a positively-acting autoregulatory gene whose product, in the presence of plant root exudate, activates expression of the nodABC and nodFE operons (35,36). The promoters of many positively-regulated genes are characterized by the presence of conserved nucleotide sequences located upstream of the normal consensus promoter sequence (37). These observations suggest that the nod gene-specific consensus sequences may be involved in the

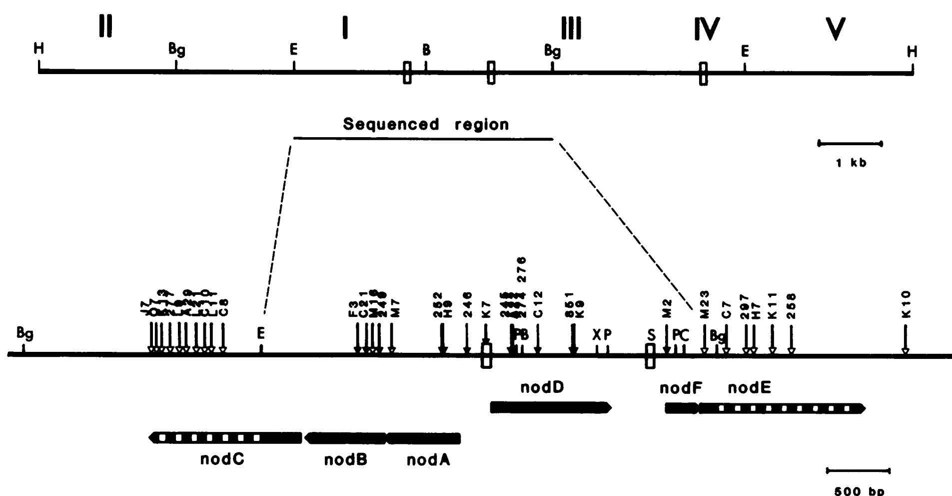

Figure 5. The 14kb HindIII nod gene region of R. trifolii strain ANU843 cloned in pRt587(9). Regions I-V carry genes involved in clover host-specific nodulation (5,7). The lower part of the figure indicates the locations of Tn5 insertions (arrows) and the extent of the predicted nodA, B, D and F coding regions (solid arrows) as well as the expected lengths of the nodC and E genes (broken arrows). Open boxes indicate the locations of the 75/76-bp conserved sequences. Restriction sites are indicated as : B(BamHI); Bg(BglII); C(ClaI); E(EcoRI); H(HindIII); P(PstI); S(SphI) and X(XmaI).

plant-activated, coordinate expression of Rhizobium nod genes.

The identification of a nod gene-specific, reiterated sequence preceding the putative nodABC and nodFE operons, and its correlation with plant root exudate-activated transcriptional units, suggests that Rhizobium nodulation genes are under complex regulatory controls that involve not only a bacterially-encoded activator, but also a signal molecule from the plant partner in the symbiosis. The data presented in this paper should facilitate definition of the promoters of each of the nod gene transcriptional units and, by comparison of various Rhizobium-legume symbioses, enable elucidation of the roles played by individual bacterial nod genes in the complex process of host-specific nodulation.

ACKNOWLEDGEMENTS

We thank Christine Hayes for assistance with DNA sequencing, Judy Tellam for oligonucleotide synthesis, Sharon Long and colleagues , Kieran Scott, Allan Downie, Lone Rossen and Claire Shearman for sharing

R. meliloti, Parasponia Rhizobium and R. leguminosarum nod gene sequences prior to publication, Michael Djordjevic and Barry Rolfe for some Tn5 mutant DNAs, Lina Onorato for preparation of the manuscript and John Shine for his continued support of this work. We also thank Siiri Iismaa and Allan Downie for their constructive comments during the preparation of this manuscript. This work has been supported in part by an Agrigenetics Sponsored Research Program and by the Wool Research Trust Fund on the recommendation of the Australian Wool Corporation.

*To whom correspondence should be addressed

[1]Present address: Department of Molecular Biology, Genentech, Inc., 460 Point San Bruno Boulevard, South San Francisco, CA 94080, USA

REFERENCES

1. Schofield, P.R., Ridge, R.W., Rolfe, B.G., Shine, J. and Watson, J. (1984) Plant Mol. Biol. 3, 3-11.
2. Downie, J.A., Hombrecher, G., Ma, Q-S., Knight, C.D., Wells, B. and Johnson, A.W.B. (1983) Mol. Gen. Genet. 190, 359-365.
3. Kondorosi, E., Banfalvi, Z. and Kondorosi, A. (1984) Mol. Gen. Genet. 193, 445-452.
4. Jacobs, T.W., Egelhoff, T.T. and Long, S.R. (1985) J. Bacteriol. 162, 469-476.
5. Djordjevic, M.A., Schofield, P.R. and Rolfe, B.G. (1985) Mol. Gen. Genet. 200, 463-471.
6. Innes, R.W., Kuempel, P.L., Plazinski, J., Canter-Cremers, H., Rolfe, B.G. and Djordjevic, M.A. (submitted for publication).
7. Schofield, P.R., Ph.D. Thesis, (1984) Australian National University, Canberra.
8. Bossi, L. and Ciampi, M.S. (1981) Mol. Gen. Genet. 183, 406-408.
9. Schofield, P.R., Djordjevic, M.A., Rolfe, B.G., Shine, J. and Watson, J. (1983) Mol. Gen. Genet. 192, 459-465.
10. Yanish-Peron, C., Vieira, J., Messing, J. (1985) Gene 33, 103-119.
11. Maniatis, T., Fritsch, E.F. and Sambrook, J. (1982) Molecular Cloning: A Laboratory Manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.
12. Sanger, F., Nicklen, S. and Coulsen, A.R. (1977) Proc. Natl. Acad. Sci. USA 74, 5463-5476.
13. Staden, R. (1984) Nucleic Acids Res. 12, 551-567.
14. Beaucage, S.L. and Caruthers, M.H. (1981) Tetrahedron Lett. 22, 1859-1862.
15. Auerswald, E.A., Ludwig, G. and Schaller, H. (1981) Cold Spring Harbor Symp. Quant. Biol. 45, 107-113.
16. Rossen, L., Johnston, A.W.B. and Downie, J.A. (1984) Nucleic Acids Res. 12, 9497-9508.
17. Tǫrǫk, I., Kondorosi, E., Stepkowski, T., Pǫsfai, J. and Kondorosi, A. (1984) Nucleic Acids Res. 12, 9509-9524.
18. Downie, J.A., Knight, C.D., Johnston, A.W.B. and Rossen, L. (1985) Mol. Gen. Genet. 198, 255-262.
19. Fickett, J.W. (1982) Nucleic Acids Res. 10, 5303-5318.
20. Downie, J.A. and Shearman, C. pers. comm.

21. Egelhoff, T.T., Fisher, R.F., Jacobs, T.W., Mulligan, J.T. and Long, S.R. (1985) DNA 4, 241-248.
22. Scott, K.F., this volume.
23. Banfalvi, Z., Sakanyan, V., Koncz, C., Kiss, A., Dusha, I. and Kondorosi, A. (1981) Mol. Gen. Genet. 184, 318-325.
24. Djordjevic, M.A., Schofield, P.R., Ridge, R.W., Morrison, N.A., Bassam, B.J., Plazinski, J., Watson, J.M. and Rolfe, B.G. (1985) Plant Mol. Biol. 4, 147-160.
25. Fisher, R.F., Tu, J.K. and Long, S.R. (1985) Appl. Env. Microbiol. 49, 1432-1435.
26. Berg, D.E., Schmandt, M.A. and Lowe, J.B. (1983) Genetics 105, 813-828.
27. Berg, D.E., Weiss, A. and Crossland, L. (1980) J. Bacterial. 142, 439-446.
28. Lupski, J.R., Smiley, B.L. Blattner, F.R. and Godson, G.N. (1982) Mol. Gen. Genet. 185, 120-128.
29. Schmid, M.B. and Roth, J.R. (1983) J. Bacteriol. 153, 1114-1119.
30. Shine, J. and Dalgarno, L. (1975) Nature 254, 34-38.
31. Normark, S., Bergstrøm, T., Jaurin, B., Lindberg, F.P. and Olsson, O. (1983) Ann. Rev. Genet. 17, 499-525.
32. Rosenberg, M. and Court, D. (1979) Ann. Rev. Genet. 13, 319-353.
33. Holmes, W.M., Platt, T. and Rosenberg, M. (1983) Cell 32, 1029-1032.
34. Ausubel, F.M. (1984) Cell 37, 5-6.
35. Mulligan, J.T. and Long, S.R. Proc. Natl. Acad. Sci. USA 82, 6609-6613.
36. Rossen, L., Shearman, C.A., Johnston, A.W.B. and Downie, J.A. EMBO J. (in press).
37. Raibaud, O. and Schwartz, M. (1984) Ann. Rev. Genet. 18, 173-206.