

Complete Genome Structure of the Nitrogen-fixing Symbiotic Bacterium *Mesorhizobium loti*

Takakazu KANEKO, Yasukazu NAKAMURA, Shusei SATO, Erika ASAMIZU, Tomohiko KATO, Shigemi SASAMOTO, Akiko WATANABE, Kumi IDESAWA, Atsuko ISHIKAWA, Kumiko KAWASHIMA, Takaharu KIMURA, Yoshie KISHIDA, Chiaki KIYOKAWA, Mitsuyo KOHARA, Midori MATSUMOTO, Ai MATSUNO, Yoko MOCHIZUKI, Shinobu NAKAYAMA, Naomi NAKAZAKI, Sayaka SHIMPO, Masako SUGIMOTO, Chie TAKEUCHI, Manabu YAMADA, and Satoshi TABATA*

Kazusa DNA Research Institute, 1532-3 Yana, Kisarazu, Chiba 292-0812, Japan

(Received 22 November 2000)

Abstract

The complete nucleotide sequence of the genome of a symbiotic bacterium *Mesorhizobium loti* strain MAFF303099 was determined. The genome of *M. loti* consisted of a single chromosome (7,036,071 bp) and two plasmids, designated as pMLa (351,911 bp) and pMLb (208,315 bp). The chromosome comprises 6752 potential protein-coding genes, two sets of rRNA genes and 50 tRNA genes representing 47 tRNA species. Fifty-four percent of the potential protein genes showed sequence similarity to genes of known function, 21% to hypothetical genes, and the remaining 25% had no apparent similarity to reported genes. A 611-kb DNA segment, a highly probable candidate of a symbiotic island, was identified, and 30 genes for nitrogen fixation and 24 genes for nodulation were assigned in this region. Codon usage analysis suggested that the symbiotic island as well as the plasmids originated and were transmitted from other genetic systems. The genomes of two plasmids, pMLa and pMLb, contained 320 and 209 potential protein-coding genes, respectively, for a variety of biological functions. These include genes for the ABC-transporter system, phosphate assimilation, two-component system, DNA replication and conjugation, but only one gene for nodulation was identified.

Key words: *Mesorhizobium loti* strain MAFF303099; genome sequencing; symbiosis; nodulation; nitrogen fixation; plasmid

1. Introduction

Rhizobia, a collective name of the genera *Rhizobium*, *Sinorhizobium*, *Mesorhizobium*, and *Bradyrhizobium*, are soil and rhizosphere bacteria of agronomic importance because they form nitrogen-fixing symbioses with leguminous plants. Nodule formation and the subsequent nitrogen-fixation result from a series of interactions controlled by the exchange of molecular signals between symbiotic bacteria and host plants followed by the expression of genes from both symbiotic partners.

The genes which are involved in the process of symbiotic nitrogen-fixation are carried by both the chromosome and plasmids which are named as the symbiotic plasmid (pSym), in *Rhizobium* and *Sinorhizobium*. The nucleotide sequence of a broad host-range symbiotic plasmid (pNGR234a) of *Rhizobium* sp. NGR234 (536 kb) has been determined, and the presence of 416 protein-coding

genes including 26 genes for nodulation and 26 genes for nitrogen-fixation was reported.¹ In *S. meliloti*, physical maps of the chromosome (3.7 Mb) and two symbiotic plasmids (1.7 Mb and 1.4 Mb) have been constructed.^{2–4} In contrast, the majority of genes for nitrogen-fixing symbioses seem to be present on the chromosome in *Mesorhizobium* and *Bradyrhizobium*.^{5,6} A physical map of the chromosome of *Bradyrhizobium* (8.7 Mb) has already been generated.⁵

M. loti is a member of rhizobia which is able to form determinant-type globular nodules and to perform nitrogen-fixation on several *Lotus* species. It was reported that the symbiotic genes of *M. loti* strain ICMP3153 cluster in a 500-kb DNA region on the chromosome, which is called the “symbiosis island,” and that this element is transmissible to nonsymbiotic *Mesorhizobium* species having the ability to form nodules.⁶ To understand the genetic systems required for the entire process of symbiotic nitrogen fixation as well as those for horizontal gene transfer among natural micro-symbionts, we initiated a genome analysis of this bac-

Communicated by Mituru Takamami

* To whom correspondence should be addressed. Tel. +81-438-52-3933, Fax. +81-438-52-3934

terium. Here, we describe the complete structure of the *M. loti* genome, which consists of a single chromosome and two large plasmids, and gene complements of both the chromosome and the plasmids. This is the first paper reporting the entire genome structure of a soil bacterium that is a member of alpha-proteobacteria.

2. Materials and Methods

2.1. Bacterial strain

M. loti strain MAFF303099 was obtained from Genetic Resource Center, National Institute of Agrobiological Resources, Ministry of Agriculture, Forestry and Fisheries.

2.2. Sequencing strategy and data assembly

The whole-genome shotgun strategy combined with the "bridging shotgun" method was adopted to determine the structure of the entire genome.⁷ Four shotgun libraries with three types of cloning vectors were generated from the total cellular DNA of *M. loti* MAFF303099 to minimize the cloning bias: RLE and RLB with approximately 1.0-kb (element clones) and 2.6-kb inserts (bridge clones), respectively, both cloned into M13mp18, RLP with approximately 8.5 kb (plasmid clones) cloned into pUC18, and RLC containing 25-kb inserts (cosmid clones) cloned in cosmid vector pKS800 (K. Saeki, unpublished), a derivative of a low-copy-number plasmid pLAFR1.⁸

One strand of the element clones and both strands of the clones from three bridge libraries were sequenced using the cycle sequencing kit (Dye-terminator Cycle Sequencing kit) with DNA sequencers type 377XL (Perkin Elmer Applied Biosystems, USA) according to the protocol recommended by the manufacturer. A total of 90000 sequence files corresponding to about eight times the genome equivalent were accumulated and subjected to assembly using the Phrap program (Phil Green, Univ. Washington, Seattle, USA). The data from the bridge, plasmid and cosmid clones facilitated accurate reconstruction of the entire genome sequences as well as the gap closure process. The remaining sequence gaps were filled either by primer walking or PCR amplification of the gap regions followed by shotgun sequencing of the products. The minimum prerequisite taken for confirmation of the sequences was to obtain the sequences from both strands or to sequence the same strand using multiple clones as templates, which ensured sufficient accuracy for further analysis of gene structure. Integrity of the reconstructed genome sequence was assessed by walking of the entire genome with the end sequences of cosmid clones, each containing a DNA fragment of constant length, approximately 25 kb.

2.3. Gene assignment and annotation

Assignment of coding regions was performed by a combination of computer prediction and similarity searching. Glimmer, a computer program based on interpolated Markov models, was used for prediction of protein-coding regions.⁹ Prior to prediction, the matrix for the *M. loti* genome was generated by training with the dataset of 4218 open reading frames (ORFs) which showed a high degree of sequence similarity to the genes of known function. All the predicted protein-coding regions equal to or longer than 90 bp long were translated into amino-acid sequences, then subjected to similarity search against the non-redundant (nr) protein database with the BLASTP program.¹⁰ If two predicted genes overlapped on either strand, those showing similarity to known genes were preferentially taken, and the longer one was chosen unless the function of the shorter one was reasonably anticipated. In parallel, the entire genome sequence was compared with those in the nr-protein database using the BLASTX program to identify genes which escaped from prediction and/or those smaller than 90 bp especially in the predicted intergenic regions. For predicted genes which did not show similarity to known gene sequences, only those equal to and longer than 150 bp were taken into account.

Functional assignment of the predicted genes was performed based on the similarity to genes of known function. For genes encoding the proteins of 100 amino acid residues or longer, a BLAST score of e^{-20} was taken into account. A higher e -value was taken into consideration for genes encoding smaller proteins.

Genes for structural RNAs were assigned by similarity search against the in-house structural RNA database which had been generated based on the data in GenBank (rel. 120.0). Prediction by the tRNAscan-SE program was used for identification of tRNA-coding regions in combination with the similarity search.¹¹

3. Results and Discussion

3.1. Sequence determination of the entire genome

The nucleotide sequence of the entire genome of *M. loti* strain MAFF303099 was deduced by assembly of a total of 90,706 sequence files which correspond to approximately eight times the genome equivalent according to the modified whole genome shotgun method, as described in Materials and Methods. To ensure sufficient accuracy for further analysis of gene structure, additional sequencing was carried out to obtain the sequences from both strands or from the same strand of multiple template clones after the first assembly. Integrity of the final sequence was assessed by comparing the distance of the end sequences of each of 361 cosmid clones on the assembled sequence and its insert length for the entire genome. The length of the genome thus deduced was 7,596,297 bp.

The genome consisted of three circular molecules, a single chromosome of 7,036,071 bp and two plasmids, designated as pMLa and pMLb (351,911 bp and 208,315 bp, respectively). Average GC contents of the chromosome and pMLa and pMLb were 62.7%, 59.3% and 59.9%, respectively. The nucleotide position was numbered from a single recognition site of the restriction enzyme *Pac* I for the chromosome and pMLa, and a single *Spe* I recognition site for pMLb (see Fig. 1 in the Supplement section).

3.2. Assignment of protein-coding genes

The potential protein-coding regions were assigned by combination of computer prediction using the Glimmer program and similarity search, as described in Materials and Methods. After training with a dataset of sequences of 4218 highly probable protein-coding genes, Glimmer predicted a total of 8907 potential genes on the chromosome. By taking the sequence similarity to known genes and the relative positions into account to avoid overlaps, the total number of the potential protein-coding genes finally assigned to the chromosome was 6752. The average gene density was one gene in every 1042 bp, which is similar to those of other bacteria analyzed so far. pMLa and pMLb had the capacity to code for 320 and 209 proteins, when estimated by the same procedure. The prediction by the Glimmer program, which had been optimized for the genes on the chromosome, was less effective for the plasmid genomes: Genes which showed apparent sequence similarity to known genes often escaped prediction because of low prediction scores, suggesting that two plasmid replicons were horizontally transferred from other genetic systems during evolution.

The putative protein-coding genes thus assigned to the genome starting with either an ATG, GTG, TTG, or ATT codon are denoted by serial number with three letters representing the species name (m), ORF longer than or less than 100 codons (l or s), and the reading direction on the circular map (r or l).

It should be reminded that the genes assigned in this paper merely represent the coding potentiality of proteins and RNAs under the defined assumptions, and the real gene assignment should be validated experimentally.

3.3. Assignment of RNA-coding genes

Two copies of rRNA gene clusters were identified on the genome in the order of 16S-23S-5S at the coordinates 2,745,482–2,751,894 and 2,752,970–2,759,407 by sequence similarity to known bacterial rRNA genes^{12–14} (see Fig. 1 of the Supplement section). Two tRNA genes, *trnI* and *trnA*, were located between the 16S and 23S rRNA genes and *trnfM* downstream of the 5S rRNA gene in the respective rRNA gene clusters. The sequences of two clusters were identical except for two nucleotide residues downstream of *trnfM*. One gene for an RNA subunit of RNase P was identified.¹⁵ A total of 50 tRNA

genes representing 47 tRNA species, which are sufficient to bind all the codon species, were assigned on the chromosome by sequence similarity to known bacterial tRNA genes and computer prediction using the tRNAscan-SE program (Figs. 1 and 2, Table 1 of the Supplement section). None of the tRNA genes contained either the group I or group II intron. One notable feature of the tRNA species of *M. loti* is the presence of tRNAs which have C residue on the first position of its anticodon for every codon box corresponding to synonymous codons (see Table 1 of the Supplement section). For example, additional *trnV*-CAC, *trnK*-CUU, *trnA*-CGC, and *trnE*-CUC species are present in *M. loti*, compared with the *Escherichia coli* strain K-12 genome.¹⁶ *trnV*-CAC, *trnQ*-CUG, *trnK*-CUU, and *trnE*-CUC of *M. loti* are the species which are absent in the genome of *Synechocystis* sp. strain PCC6803.¹⁶ These tRNA species are not indispensable for translation because the codons with purine (A or G) residues in the third position can be decoded by tRNAs with a U residue at the first position of the anticodon. Therefore, it is plausible that tRNA variation is a consequence of adjustment to highly GC-pressured codons in the *M. loti* genome. No RNA coding genes were found in the plasmid genomes.

3.4. Functional assignment of protein-coding genes

Similarity search of the 6752 potential protein-coding genes in the chromosome against the nr database indicated that 3675 (54%) were homologues to genes of known function, 1423 (21%) showed similarity to hypothetical genes, and the remaining 1654 (25%) showed no significant similarity to any registered genes (Table 1). Two plasmid genomes contained a larger number of genes of unknown function, 50% and 64% for pMLa and pMLb, respectively. (Table 1).

The potential protein-coding genes whose function could be anticipated were grouped into 14 categories with respect to different biological roles, according to the principle of Riley.¹⁷ The number of genes in each category is summarized in Table 1, and the list of the name of each gene is presented in the web database, RhizoBase, at <http://www.kazusa.or.jp/rhizobase/>. On the gene map of the Supplement section (Fig. 1), the location, length and direction of these genes are indicated, with color codes corresponding to functional categories.

3.5. Characteristic features of predicted genes

3.5.1. Symbiotic island

The presence of a DNA segment 500 kb long named the “symbiotic island,” where the symbiotic genes are clustered, was reported in the *M. loti* strain ICMP3153, and the nucleotide sequences of both ends of the segment have been determined.⁶ It was shown that this DNA segment is transmittable and inserted into a phe-tRNA gene with 17-bp duplication of the 3' terminal portion of

Table 1. Features of the assigned protein-coding genes and the functional classification.

	chromosome	%	non_sym	%	sym	%	pMLa	%	pMLb	%	entire genome	%
Amino acid biosynthesis	177	2.6	161	2.6	16	2.8	10	3.1	2	2.0	189	2.6
Biosynthesis of cofactors, prosthetic groups, and carriers	145	2.1	130	2.1	15	2.6	14	4.4	1	1.0	160	2.2
Cell envelope	110	1.6	107	1.7	3	0.5	3	0.9	1	1.0	114	1.6
Cellular processes	176	2.6	148	2.4	28	4.8	14	4.4	16	16.0	206	2.8
Central intermediary metabolism	120	1.8	88	1.4	32	5.5	1	0.3	0	0.0	121	1.7
Energy metabolism	326	4.8	308	5.0	18	3.1	7	2.2	7	7.0	340	4.7
Fatty acid, phospholipid and sterol metabolism	163	2.4	157	2.5	6	1.0	4	1.3	1	1.0	168	2.3
Purines, pyrimidines, nucleosides, and nucleotides	81	1.2	78	1.3	3	0.5	0	0.0	1	1.0	82	1.1
Regulatory functions	517	7.7	489	7.9	28	4.8	11	3.4	11	11.0	539	7.4
DNA replication, recombination, and repair	85	1.3	81	1.3	4	0.7	7	2.2	11	11.0	103	1.4
Transcription	54	0.8	53	0.9	1	0.2	0	0.0	0	0.0	54	0.7
Translation	190	2.8	178	2.9	12	2.1	5	1.6	1	1.0	196	2.7
Transport and binding proteins	717	10.6	686	11.1	31	5.3	41	12.8	6	6.0	764	10.5
Other categories	814	12.1	649	10.5	165	28.4	43	13.4	17	17.0	874	12.0
Subtotal of genes similar to genes of known function	3675	54.4	3313	53.7	362	62.4	160	50.0	75	75.0	3910	53.7
Similar hypothetical protein	1423	21.1	1352	21.9	71	12.2	60	18.8	38	18.2	1521	20.9
Subtotal of genes similar to registered genes	5098	75.5	4665	75.6	433	74.7	220	68.8	113	93.2	5431	74.6
No similarity	1654	24.5	1507	24.4	147	25.3	100	31.3	96	45.9	1850	25.4
Total	6752	100.0	6172	100.0	580	100.0	320	100.0	209	100.0	7281	100.0

Genes assigned in the chromosome and the two plasmids are classified according to their similarity to genes of known and unknown function. Genes on the chromosome were further divided into those inside (sym) and outside (non sym) of the symbiotic island as shown in the box.

the gene.⁶ To confirm the presence of such a symbiotic island in *M. loti* strain MAFF303099, we searched the entire chromosome and the plasmids with the 17 bp sequence of the 3' terminal portion of the phe-tRNA gene. As a result, a 610,975-bp DNA segment flanked by the 17-bp sequence on both sides was found on the chromosome at coordinates 4,644,792 to 5,255,766. One of the flanking 17-bp sequences was a part of the intact phe-tRNA gene, and a P4 integrase gene which was located near the end of the symbiotic island in ICMP3153 was also present (mll6432) in MAFF303099, in addition to the second copy (mll5763) with lower similarity outside of the opposite junction. The DNA segment contained 30 genes related to nitrogen fixation and 24 genes for nodulation. All of these characteristics strongly indicated that this 610,975-bp segment is the symbiotic island of the *M. loti* strain MAFF303099. Comparison of the structures of both ends of the islands between the two strains showed that the sequences are not conserved except for the region spanning the phe-tRNA and P4 integrase genes. Together with the fact that the island of MAFF303099 was 20% longer than that of ICMP3153, the structures of the symbiotic islands are significantly diverse between the two *M. loti* strains.

A total of 580 protein-coding genes were assigned to the symbiotic island of MAFF303099 based on computer prediction and similarity search of the sequences of known genes. The Glimmer program often failed to predict the genes of known function in the symbiotic island, suggesting exogenous origin of this DNA segment. The putative genes in the symbiotic island covered all the functional categories (Table 1) but with a different distribution compared with the genes in the remaining part of the chromosome: There was a higher percentage of genes for "Central intermediary metabolism" and "Other category" and a lower percentage of genes for "Cell envelope," "DNA replication, recombination and repair," "Transcription" and "Transport and binding proteins."

Twelve genes for the conjugal transfer proteins were identified in the symbiotic island (see RhizoBase, at <http://www.kazusa.or.jp/rhizobase/>), and nine of them, *traG-traB-trbC-trbE-trbJ-trbL-trbF-trbG-trbI*, formed a large gene cluster (mlr6395, 6397–6398, 6400–6405). There were four additional genes for conjugal transfer outside of the island, two of which, *traA* and *traD*, are adjacent to each other (mll0964 and msr0965). These genes may be involved in transmission of the symbiotic island.

Two gene clusters, each comprised of four genes for biotin synthesis (mll5828–mll5831, mll6003 and mll6005–mll6007), were assigned in the symbiotic island. Another gene cluster with the same gene set was found in plasmid pMLa. A cluster of genes for thiamine biosynthesis consisting of six genes was also identified (mll5788–mll5795), though the *thiG* gene seems to be split into two ORFs (mll5790 and msl5792) by a frameshift mutation. The

presence of these genes in the transmittable DNA segments is consistent with the observation that the different strain of *M. loti*, which is capable of nodule formation, can grow in the absence of both biotin and thiamine in the medium.⁶

Of the 580 genes assigned in the symbiotic island, 111 (19.6%) coded for transposon-related functions such as transposase, integrase, recombinase and resolvase. The high content of the genes in this category is a striking contrast to those in the genome outside of the island (30 out of 6172, 0.5%) and in the plasmids pMLa (24 out of 320, 7.5%) and pMLb (6 out of 209, 2.9%). Uneven distribution of these genes along the island apparently indicates the presence of hot spots for the insertion of transposable elements. Together with the observation that disrupted genes are present among the genes for nodulation and transposon-related function, drastic rearrangement in the symbiotic island may have taken place possibly due to insertion and transposition of transposable DNA elements.

A cluster of genes for type III secretion system were found at coordinates 5,157,467–5,168,086. Nine genes, *hrcN-y4yJ-hrcQ-hrcR-hrcS-hrcT-hrcU-y4yQ-hrcV*, formed the cluster (mlr6342–mlr6348). The genes in this system are known to be involved in induction and secretion of elicitors of the hypersensitivity response in plants.¹⁸

Of the 580 genes assigned in the symbiotic island, 250 showed a high degree of sequence similarity (equal to or higher than 30% amino acid identity) to those in the symbiotic plasmid, pNGR234a, of *Rhizobium* sp. NGR234 (536 kb).¹ This strongly suggests common ancestry of two DNA units capable of conferring nodule formation and nitrogen fixation ability. When the relative positions of each gene were compared, gene order within the gene clusters for various biological functions was often conserved. However, the relative positions of the gene clusters was rarely conserved between the two DNA units.

3.5.2. Genes related to nodulation and nitrogen fixation

Thirty-nine genes for nodulation were identified on the chromosome, and 24 of them were located in the symbiotic island (RhizoBase, at <http://www.kazusa.or.jp/rhizobase/>). Forty-six genes were assigned to the category of nitrogen fixation, of which 30 were found in the symbiotic island. Only one homologue of the *noeC* gene for nodulation was present in the plasmid genome (pMLa).

It has been reported that a NodD gene product regulates the transcription of a group of genes related to nodulation by binding to the upstream sequence named the nod-box.¹⁹ Nine genes, mlr5801, mlr5848, mlr6144, mlr6161, mlr8755, mlr6175, mlr6181, mlr6334, and mlr6386, were found to contain nod-box-like sequences

upstream. These include *nodZ-noeL-nolK* (mlr5848–5849–8749), *nodS* (mlr6161), *nolL* (mlr6181) and the two-component response regulator *y4xI* (mlr6334), whose counterparts in pNGR234a also possess the nod-box sequence.¹

A transport system of dicarboxylates such as succinate, fumarate and malate from the periplasm to the inner membrane is known to play a significant role in the energy supply during symbiosis.²⁰ Two sets of genes for the C4-dicarboxylate transport regulatory system (*dctA*, *dctB* and *dctD*) were identified inside (mll5840, mlr5841–5842) and outside (mll7237, mlr7238–7239) of the symbiotic island.

3.5.3. Plasmid genes

pMLa contained a total of 40 genes for the ABC-transporter system including 18 for ATP-binding components, 7 for substrate-binding components, and 15 for permeases, while only 5 genes for this system was identified in pMLb. pMLa also harbored two gene clusters for the assimilation of phosphate, *phnM-G-H-I-J-K* (mll9151–mll9156) and *phnM-G-H-I-J-K-L-N* (mlr9276–mlr9288).²¹ On the other hand, the genome of pMLb contained 10 genes for the two-component system, 6 for sensors and 4 for regulators, while only 2 genes for regulators were found in pMLa. A cluster of genes for cytochrome oxidase subunits 1, 2 and 3 were also identified in pMLb. The functional significance of these gene systems in the plasmid genomes is unclear.

Nine conjugation-related genes formed a cluster, *trbB-C-D-E-J-L-F-G-I*, in pMLb (mll9603–mll9612). Ten genes for the conjugation transfer proteins were identified in pMLa (mlr9249–mlr9251, mlr9253, mlr9255–mlr9256), two of which, mlr9249 and mlr9258, correspond to *trbC* and *trbG*, respectively. However, the remaining eight genes showed significant sequence similarity to those on plasmid pXF51 of *Xylella fastidiosa*,²² a Gram-negative plant pathogenic bacterium, suggesting that the two plasmids have a different origin.

Three genes for DNA replication, *repA*, *repB*, and *repC*, were identified in both pMLa (mll9353–9351) and pMLb (mll9654–9652), as was in pNGR234a (*y4cK*, *y4cJ* and *y4cI*).¹ The consensus DNA sequence of OriV (approximately 154 bp) was observed in the intergenic regions between *repB* and *repC* genes in pMLa and pMLb at coordinates 308,883–309,034 and 123,540–123,690, respectively.²³

3.5.4. Comparison of codon usages among the genomes and symbiotic island

Codon usage frequency of the genes in the chromosome and the two plasmids are tabulated in Table 2 of the Supplement section. Those inside and outside of the symbiotic island were calculated separately. It is remarkable that the GC contents of both coding regions (63.7%)

and the third-position of the codons (80.2%) are the highest in the chromosome outside of the symbiotic island. Codon usage of the genes in the symbiotic island and in the plasmids are similar to each other: GC contents of the coding regions are slightly lower than that of the symbiotic island, and GC contents of the third-position are nearly 10% lower than that of the symbiotic island.

3.5.5. Other features

In addition to the various characteristics of genes in the *M. loti* genome described above, other remarkable features noted so far are as follows:

1. Two genes for the sigma-54 factor (*rpoN*), which is involved in transcriptional control of major nitrogen-fixation genes such as *nifHDK*,²⁴ were present (mll3196 and mll5872).
2. A putative gene product of a large subunit of a nitrate reductase gene (mlr2864) contained an insertion of approximately 90 amino acid residues. This may be a putative intein,²⁵ although the conserved residues for inteins were not apparent.
3. Five pairs of chaperonin *groES-groEL* genes were present in the genome: One (msl5812–mll5810) inside and three (mll8202–8201, mlr2393–2394, and mll2233–2232) outside of the symbiotic island, and one (msr9341–mlr9342) on plasmid pMLa.
4. Four clusters of genes for flagellar structure, assembly and motility were identified in the 47-kb region at the approximate coordinates 2,336,711–2,382,749. The first and the second clusters which contain five (mll2899–mll2905) and 20 (mlr2907–mlr2933) genes, respectively, were conserved in *S. meliloti* and *Agrobacterium tumefaciens* to various extents.^{26,27} The remaining two clusters comprising seven genes (mlr2937–msr2943) and two genes (mlr2957 and mlr2958) were unique to *M. loti*. Two additional gene clusters for the same function were found at the approximate coordinates 4,486,611–4,493,619 (7 genes, msr5593–mlr5602) and 5,286,190–5,293,683 (6 genes, mll6475–mlr6488), both of which showed sequence similarity to those of *Caulobacter crescentus*.²⁸
5. Succinoglycan is an acidic exopolysaccharide which is known to play a significant role in nodule formation.²⁹ The presence of a large cluster of genes for a family of glycosyl transferases, *exoP-N-O-M-A-L-K-H-I-T-W-V-U-X-Y-F-Q-Z-B*, required for the synthesis of succinoglycan have been reported in *S. meliloti*.³⁰ In *M. loti*, *exoP-N-O-M-A-L-T* [5 genes]-*K* [1 genes]-*U* [7 genes]-*X-Y-F-Q* formed a cluster (mlr5249–mlr5276) at coordinates of 4,179,643–4,209,505, while *exoI* (mll0560, mll8119,

and *mlr0479*), *exoZ* (*mlr6758* and *mlr8032*) and *exoB* (*mlr7878*) were separately present in the genome. There were no *exoH*, *exoW* and *exoV* genes on the chromosome or in the plasmids, whereas three genes (*mlr5265*, *mlr5266* and *mlr5268*) homologous to those for sugar modification were found between *exoT* and *exoK*. It is therefore plausible that these three genes in *M. loti* simply substitute for *exoH*, *exoW* and *exoV* in *S. meliloti*, or produce different types of succinoglycan.

The sequences as well as the gene information shown in this paper are available in the Web database, Rhizobase, at <http://www.kazusa.or.jp/rhizobase/>. The sequence data analyzed in this study have been registered in the DDBJ/GenBank/EMBL databases by dividing them into 24 entries. The accession numbers are as follows: AP002994 (nucleotide positions 1–347,660), AP002995 (347,655–694,550), AP002996 (694,545–1,044,163), AP002997 (1,044,158–1,373,866), AP002998 (1,373,861–1,721,610), AP002999 (1,721,605–2,067,898), AP003000 (2,067,893–2,415,969), AP003001 (2,415,964–2,761,746), AP003002 (2,761,741–3,111,238), AP003003 (3,111,233–3,460,348), AP003004 (3,460,343–3,798,921), AP003005 (3,798,916–4,131,550), AP003006 (4,131,545–4,473,431), AP003007 (4,473,426–4,821,836), AP003008 (4,821,831–5,168,650), AP003009 (5,168,645–5,508,325), AP003010 (5,508,320–5,849,176), AP003011 (5,849,171–6,195,680), AP003012 (6,195,675–6,542,221), AP003013 (6,542,216–6,890,165), and AP003014 (6,890,160–7,036,071) for the chromosome; AP003015 (1–307,548) and AP003016 (307,543–351,911) for pMLa; AP003017 (1–208,315) for pMLb.

Acknowledgements: We thank Drs. M. Kawaguchi, K. Minamizawa and T. Uchiumi for their valuable discussions. We also thank Dr. K. Saeki for providing the pKS800 cosmid vector. This work was supported by the Kazusa DNA Research Institute Foundation.

References

- Freiberg, C., Fellay, R., Bairoch, A., Broughton, W. J., Rosenthal, A., and Perret, X. 1997, Molecular basis of symbiosis between *Rhizobium* and legumes, *Nature*, **387**, 394–401.
- Capela, D., Barloy-Hubler, F., Gatiús, M. T., Gouzy, J., and Galibert, F. 1999, A high-density physical map of *Sinorhizobium meliloti* 1021 chromosome derived from bacterial artificial chromosome library, *Proc. Natl. Acad. Sci. USA*, **96**, 9357–9362.
- Barloy-Hubler, F., Capela, D., Barnett, M. J. et al. 2000, High-resolution physical map of the *Sinorhizobium meliloti* 1021 pSyma megaplasmid, *J. Bacteriol.*, **182**, 1185–1189.
- Barloy-Hubler, F., Capela, D., Batut, J., and Galibert, F. 2000, High-resolution physical map of the pSymb megaplasmid and comparison of the three replicons of *Sinorhizobium meliloti* strain 1021, *Curr. Microbiol.*, **41**, 109–113.
- Kundig, C., Hennecke, H. and Gottfert, M. 1993, Correlated physical and genetic map of the *Bradyrhizobium japonicum* 110 genome, *J. Bacteriol.*, **175**, 613–622.
- Sullivan, J. T. and Ronson, C. W. 1998, Evolution of rhizobia by acquisition of a 500-kb symbiosis island that integrates into a phe-tRNA gene, *Proc. Natl. Acad. Sci. USA*, **95**, 5154–5149.
- Kaneko, T., Tanaka, A., Sato, S. et al. 1995, Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. I. Sequence features in the 1 Mb region from map positions 64% to 92% of the genome., *DNA Res.*, **2**, 153–166.
- Ditta, G., Stanfield, S., Corbin, D., and Helinski, D. R. 1980, Broad host range DNA cloning system for gram-negative bacteria: construction of a gene bank of *Rhizobium meliloti*, *Proc. Natl. Acad. Sci. USA*, **77**, 7347–7351.
- Delcher, A. L., Harmon, D., Kasif, S., White, O., and Salzberg, S. L. 1999, Improved microbial gene identification with GLIMMER, *Nucleic Acids Res.*, **27**, 4636–4641.
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. 1990, Basic local alignment search tool, *J. Mol. Biol.*, **215**, 403–410.
- Lowe, T. M. and Eddy, S. R. 1997, tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence, *Nucleic Acids Res.*, **25**, 955–964.
- Wang, E. T., van Berkum, P., Sui, X. H., Beyene, D., Chen, W. X., and Martinez-Romero, E. 1999, Diversity of rhizobia associated with *Amorpha fruticosa* isolated from Chinese soils and description of *Mesorhizobium amorphae* sp. nov., *Int. J. Syst. Bacteriol.*, **49**, 51–65.
- Minnick, M. F., Mitchell, S. J., McAllister, S. J., and Battisti, J. M. 1995, Nucleotide sequence analysis of the 23S ribosomal RNA-encoding gene of *Bartonella bacilliformis*, *Gene*, **162**, 75–79.
- Bulygina, E. S., Galchenko, V. F., Govorukhina, N. I. et al. 1990, Taxonomic studies on methylotrophic bacteria by 5S ribosomal RNA sequencing, *J. Gen. Microbiol.*, **136**, 441–446.
- Brown, J. W., Haas, E. S., James, B. D., Hunt, D. A., Liu, J. S., and Pace, N. R. 1991, Phylogenetic analysis and evolution of RNase P RNA in proteobacteria, *J. Bacteriol.*, **173**, 3855–3863.
- Nakamura, Y. and Tabata, S. 1999, Codon-anticodon assignment and detection of codon usage trends in seven microbial genomes, *Microb. Comp. Genomics*, **2**, 299–312.
- Riley, M. 1993, Functions of the gene products of *Escherichia coli*, *Microbiol. Rev.*, **57**, 862–952.
- Hueck, C. J. 1998, Type III protein secretion systems in bacterial pathogens of animals and plants, *Microbiol. Mol. Biol. Rev.*, **62**, 379–433.
- Rostas, K., Kondorosi, E., Horvath, B., Simoncsits, A., and Kondorosi, A. 1986, Conservation of extended promoter regions of nodulation genes in *Rhizobium*, *Proc. Natl. Acad. Sci. USA*, **83**, 1757–1761.
- Yurgel, S., Mortimer, M. W., Rogers, K. N., and Kahn, M. L. 2000, New substrates for the dicarboxylate transport system of *Sinorhizobium meliloti*, *J. Bacteriol.*, **182**,

- 4216–4221.
21. Chen, C. M., Ye, Q. Z., Zhu, Z. M., Wanner, B. L., and Walsh, C. T. 1990, Molecular biology of carbon-phosphorus bond cleavage. Cloning and sequencing of the *phn* (*psiD*) genes involved in alkylphosphonate uptake and C-P lyase activity in *Escherichia coli* B, *J. Biol. Chem.*, **265**, 4461–4671.
 22. Simpson, A. J. G., Reinach, F. C., Arruda, P. et al. 2000, The genome sequence of the plant pathogen *Xylella fastidiosa*, *Nature*, **406**, 151–157.
 23. Chain, P. S., Hernandez-Lucas, I., Golding, B., and Finan, T. M. 2000, oriT-directed cloning of defined large regions from bacterial genomes: identification of the *Sinorhizobium meliloti* pExo megaplasmid replicator region, *J. Bacteriol.*, **182**, 5486–5494.
 24. Kullik, I., Fritsche, S., Knobel, H., Sanjuan, J., Hennecke, H., and Fischer, H. M. 1991, *Bradyrhizobium japonicum* has two differentially regulated, functional homologs of the sigma 54 gene (*rpoN*), *J. Bacteriol.*, **173**, 1125–1138.
 25. Perler, F. B. 2000, InBase, the Intein Database, *Nucleic Acids Res.*, **28**, 344–345.
 26. Sourjik, V., Sterr, W., Platzer, J., Bos, I., Haslbeck, M., and Schmitt, R. 1998, Mapping of 41 chemotaxis, flagellar and motility genes to a single region of the *Sinorhizobium meliloti* chromosome, *Gene*, **223**, 283–290.
 27. Deakin, W. J., Parker, V. E., Wright, E. L., Ashcroft, K. J., Loake, G. J., and Shaw, C. H. 1999, *Agrobacterium tumefaciens* possesses a fourth flagelin gene located in a large gene cluster concerned with flagellar structure, assembly and motility, *Microbiology*, **145**, 1397–1407.
 28. Skerker, J. M. and Shapiro, L. 2000, Identification and cell cycle control of a novel pilus system in *Caulobacter crescentus*, *EMBO J.*, **19**, 3223–3234.
 29. Reuber, T. L. and Walker, G. C. 1993, Biosynthesis of succinoglycan, a symbiotically important exopolysaccharide of *Rhizobium meliloti*, *Cell*, **74**, 269–280.
 30. Glucksmann, M. A., Reuber, T. L., and Walker, G. C. 1993, Family of glycosyl transferases needed for the synthesis of succinoglycan by *Rhizobium meliloti*, *J. Bacteriol.*, **175**, 7033–7044.