DNA Research **9**, 189–197 (2002)

Complete Genomic Sequence of Nitrogen-fixing Symbiotic Bacterium *Bradyrhizobium japonicum* USDA110

Takakazu Kaneko,[1] Yasukazu Nakamura,[1] Shusei Sato,[1] Kiwamu Minamisawa,[2] Toshiki Uchiumi,[3] Shigemi Sasamoto,[1] Akiko Watanabe,[1] Kumi Idesawa,[1] Mayumi Iriguchi,[1] Kumiko Kawashima,[1] Mitsuyo Kohara,[1] Midori Matsumoto,[1] Sayaka Shimpo,[1] Hisae Tsuruoka,[1] Tsuyuko Wada,[1] Manabu Yamada,[1] and Satoshi Tabata[1,*]

Kazusa DNA Research Institute, 2-6-7 Kazusa-kamatari, Kisarazu, Chiba 292-0812, Japan,[1] *Graduate School of Life Sciences, Tohoku University, Katahira, Aoba-ku, Sendai 980-8577, Japan,*[2] *and Department of Chemistry and BioScience, Faculty of Science, Kagoshima University, Kagoshima 890-0065, Japan*[3]

(Received 18 November 2002)

Abstract

The complete nucleotide sequence of the genome of a symbiotic bacterium *Bradyrhizobium japonicum* USDA110 was determined. The genome of *B. japonicum* was a single circular chromosome 9,105,828 bp in length with an average GC content of 64.1%. No plasmid was detected. The chromosome comprises 8317 potential protein-coding genes, one set of rRNA genes and 50 tRNA genes. Fifty-two percent of the potential protein genes showed sequence similarity to genes of known function and 30% to hypothetical genes. The remaining 18% had no apparent similarity to reported genes. Thirty-four percent of the *B. japonicum* genes showed significant sequence similarity to those of both *Mesorhizobium loti* and *Sinorhizobium meliloti*, while 23% were unique to this species. A presumptive symbiosis island 681 kb in length, which includes a 410-kb symbiotic region previously reported by Göttfert et al., was identified. Six hundred fifty-five putative protein-coding genes were assigned in this region, and the functions of 301 genes, including those related to symbiotic nitrogen fixation and DNA transmission, were deduced. A total of 167 genes for transposases/104 copies of insertion sequences were identified in the genome. It was remarkable that 100 out of 167 transposase genes are located in the presumptive symbiotic island. DNA segments of 4 to 97 kb inserted into tRNA genes were found at 14 locations in the genome, which generates partial duplication of the target tRNA genes. These observations suggest plasticity of the *B. japonicum* genome, which is probably due to complex genome rearrangements such as horizontal transfer and insertion of various DNA elements, and to homologous recombination.

Key words: *Bradyrhizobium japonicum* USDA110; genome sequencing; symbiosis; nodulation; nitrogen fixation

1. Introduction

Advances in DNA manipulation and sequencing technology have drastically changed the strategies for studying genetic systems of various living organisms, especially bacteria. Sequencing of the entire genomes followed by investigation of gene function using sequence information has proven to be effective; therefore, the complete genome structures of nearly 100 bacteria have been determined to date and genome sequencing of several hundred other species is in progress. Complete genomic sequences provide not only the information needed to perform functional analysis of the genes but also new insights into gene function, gene evolution, and genome evolution by comparing the gene components and gene organization in the genomes among species. This approach is called "comparative genomics."

Symbiotic nitrogen fixation is a vital component of the global nitrogen cycles and agricultural practices. The establishment of rhizobia-legume symbiosis requires the creation of developmental and functional programs. To understand the mechanisms which underlie this symbiosis, the physiology, ecology and genetics of rhizobia have been studied extensively. In addition, the recent introduction of genomic approaches has accelerated the accumulation of information on gene and genome structures as well as gene function. The complete structure of a broad host-range symbiotic plasmid (pNGR234a) of *Rhizobium* sp. NGR234 was reported in 1997.[1] Since

Communicated by Mituru Takanami

* To whom correspondence should be addressed. Tel. +81-438-52-3933, Fax. +81-438-52-3934, E-mail: tabata@kazusa.or.jp

then, the nucleotide sequences of the entire genomes of two rhizobia, *Mesorhizobium loti* and *Sinorihzobium meliloti* have been determined.[2,3] Accordingly, many comprehensive functional analyses of genes based on the genome sequences have been conducted using a microarray technology. These analyses include identification of genes that are up-regulated or down-regulated during the process of symbiotic nitrogen fixation.

Among rhizobia, *Bradyrhizobium japonicum* is the most agriculturally important species because it has the ability to form root nodules on soybeans (*Glycine max*). *B. japonicum* strain USDA110, which was originally isolated from soybean nodule in Florida, USA in 1957, has been widely used for the purpose of molecular genetics, physiology, and ecology, because of its superior characteristics regarding symbiotic nitrogen fixation.[4,5] The genome structure of *B. japonicum* USDA110 is similar to that of *M. loti* MAFF303099[2] in that many of the genes for symbiotic nitrogen fixation are clustered on the chromosome,[6,7] and a 410-kb DNA region identified as a symbiotic region is rich in these genes. A physical map of the genome and an ordered BAC (Bacterial Artificial Chromosome) library which covers most of the genome were reported.[8] To better understand the whole genetic system carried by this organism including the genes required for symbiotic nitrogen fixation, we determined the nucleotide sequence of the entire genome of *B. japonicum* USDA110, which comprises a single circular chromosome. In this paper, we describe the characteristic features of the genome of *B. japonicum* and compare them with two previously sequenced rhizobial strains, *M. loti* and *S. meliloti*.

2. Materials and Methods

2.1. Bacterial strain and genomic libraries

Bradyrhizobium japonicum USDA110 was provided by Dr. Michael J. Sadowsky at University of Minnesota.

For DNA sequencing, five random libraries with four types of cloning vectors were constructed from the total cellular DNA of *B. japonicum* to minimize cloning bias: BRE and BRB with inserts of approximately 1.0 kb (element clones) and 2.7 kb (bridge clones), respectively, were both cloned into M13mp18; BRP with approximately 9 kb (plasmid clones) was cloned into pUC18; BRC containing 25 kb inserts (cosmid clones) was cloned in cosmid vector pKS800;[9] and BRL bearing inserts of approximately 50 kb was cloned in BAC vector pBeloBACII.

2.2. DNA sequencing

The nucleotide sequence of the entire genome of *B. japonicum* USDA110 was determined by the whole-genome shotgun strategy combined with the "bridging shotgun" method.[10]

One strand of the element clones and both strands of the clones from the other four libraries were sequenced using the Dye-terminator Cycle Sequencing kit with DNA sequencers type 377XL (Applied Biosystems, USA), and ET-Terminator kit with a RISA-384 DNA sequencer (Shimazu Corporation, Japan). The accumulated sequence files were assembled using the Phrap program (Philip Green, University of Washington, Seattle, USA). The end-sequence data from the bridge, plasmid, cosmid, and BAC clones facilitated the gap-closure process as well as accurate reconstruction of the entire genome. The final gaps in the sequences were filled by the primer walking method. A lower threshold of acceptability for the generation of consensus sequences was set at a Phred score of 20 for each base. The integrity of the reconstructed genome sequence was assessed by walking through the entire genome with the end sequences of the cosmid or BAC clones.

2.3. Gene assignment and annotation

Coding regions were assigned by a combination of computer prediction and similarity search as described previously. Briefly, prediction of protein-coding regions was carried out with the Glimmer 2.02 program.[11] Prior to prediction, the matrix was generated for the *B. japonicum* genome by training with a dataset of 1593 open reading frames (ORFs) that showed a high degree of sequence similarity to genes assigned in the genomes of *M. loti*, *S. meliloti*, and *Agrobacterium tumefaciens* at the amino acid level. All of the predicted protein-encoding regions equal to or longer than 90 bp were translated into amino acid sequences, which were then subjected to similarity search against the non-redundant protein database (nr-database) with the BLASTP program.[12] In parallel, the entire genomic sequence was compared with those in the nr-protein database using the BLASTX program to identify genes that had escaped prediction and/or those smaller than 90 bp, especially in the predicted intergenic regions. For predicted genes that did not show sequence similarity to known genes, only those equal to or longer than 150 bp were considered as candidates.

Functions of the predicted genes were assigned on the basis of the sequence similarity of their deduced products to those of genes of known function. For genes that encode proteins of 100 amino acid residues or more, a BLAST score of 10^{-20} was considered significant. A higher E-value was considered significant for genes encoding smaller proteins.

Genes for structural RNAs were assigned by similarity search against the in-house structural RNA database that had been generated based on the data in GenBank (rel. 124.0). tRNA-encoding regions were predicted by use of the tRNA scan-SE 1.21 program[13] in combination with the similarity search.

Comparison of the structures of the gene components among *B. japonicum*, *M. loti*, and *S. meloloti* was performed by taking two factors, the BLAST2 bit score and the ratio of alignment length, into consideration. A lower threshold of acceptability was set at one-fourth of the bit score reported by self-comparison of the translated amino acid sequences by the BLASTP program. Only amino acid sequences whose alignments extended over at least 0.6 times the length of the query sequence were considered similar. With lower stringency, two protein-encoding genes were considered similar if the BLAST2 score was lesser than 10^{-4}.

The GC skew analysis was performed as described by Lobry.[14]

3. Results and Discussion

3.1. Sequence determination of the entire genome

The entire genome of *B. japonicum* was sequenced according to the modified whole genome shotgun method as described in Materials and Methods. Briefly, a total of 101,181 random sequence files corresponding to approximately 6 genome-equivalents were assembled to generate draft sequences. Finishing was then carried out by visually editing the above sequences, followed by gap closing, and additional sequencing to obtain the sequence data having a Phred score of 20 or higher. The integrity of the final genome sequence was assessed by comparing the insert length of each cosmid and BAC clone with the computed distance between the end sequences of the clones. The genome of *B. japonicum* USDA110 thus deduced was a circular chromosome of 9,105,828 bp, and the average GC content was 64.1%. No plasmid was detected in the course of genome sequencing. The nucleotide position was numbered from one of the recognition sites of the restriction enzyme *Pac* I (Fig. 1 and Fig. 1 in the Supplement section) according to the physical map of the genome by Kündig et al,[6] although inconsistency was found between the published physical map and the genome sequence determined in this study with respect to another *Pac* I restriction site, possibly due to genome rearrangements.

3.2. Sequence features of the genome

The second inner circle of Fig. 1 shows the average GC percent of every 10 kb for the entire genome. Uneven distribution of GC contents was observed in several locations of the genome including a putative symbiotic region (section 3.5.2) and putative insertion elements in tRNA genes (section 3.5.4).

A GC skew analysis was performed to locate a probable origin and terminator of DNA replication.[14] As a result, shift of GC skew was observed in two regions of the genome at coordinates 700 kb and 4890 kb, as shown in Fig. 1 (the innermost circle). In *Caulobacter crecentus*, a member of alpha-proteobacteria, the origin of DNA replication has experimentally been located between two genes, *CC0001* and *CC3763* (*hemE*), where the shift of GC skew was detected.[15,16] In *A. tumefaciens* and *S. meliloti*, the same region is anticipated to be the origin because a shift of GC skew was observed in the region spanning the *CC0001* homologue and *hemE* genes despite of no experimental evidences.[17,18] A cluster of genes containing *parA*, *parB*, *gidA*, *gidB*, *thdF*, *rho*, *aroE*, *dnaQ*, and *CC3761* homologue was conserved near this region in the above three species. Although no clear shift of GC skew was detected in *M. loti*, it is assumed that the same region functions as the replication origin, because the gene cluster containing the above gene set is conserved.[2] In *B. japonicum*, a cluster of *parA*, *parB*, *gidA*, *gidB*, *thdF*, *rho*, *CC0001* homologue, *aroE* and *dnaQ* genes (*bll0630–bll0641*) but not *hemE* (*bll2399*), was found at coordinates 675–689 kb. These observations strongly suggest that this region contains an origin of replication.

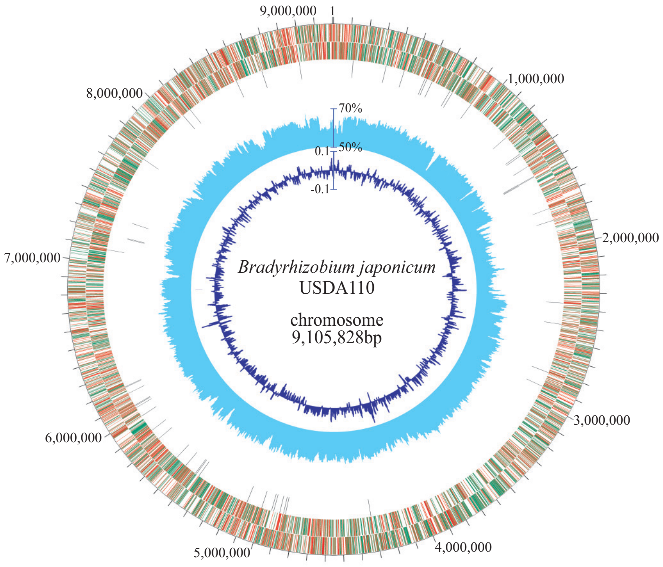

Figure 1. Circular representation of the chromosome of *Bradyrhizobium japonicum* USDA110. The scale indicates the location in bp starting from the *Pac* I recognition site. The bars in the first and second outermost circles show the positions of the putative protein-encoding genes in the clockwise and counterclockwise directions, respectively. Genes whose functions could be deduced by sequence similarity to genes of known function are depicted in green, and those whose function could not be deduced are in red. The bars in the third circle indicate the positions of predicted tRNA genes and those in the fourth circle the positions of genes for structural RNAs including rRNAs and small RNAs. The innermost and the second circles with scales show the GC skew and the average GC percent, respectively, calculated using a window-size of 10 kb.

Another shift of GC skew was detected at the diagonal position of the putative replication origin in the genome of *B. japonicum* (coordinates 4900 kb) as well

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

	Number of genes	%
Amino acid biosynthesis	173	2.1
Biosynthesis of cofactors, prosthetic groups, and carriers	179	2.2
Cell envelope	109	1.3
Cellular processes	302	3.6
Central intermediary metabolism	198	2.4
Energy metabolism	303	3.6
Fatty acid, phospholipid and sterol metabolism	212	2.5
Purines, pyrimidines, nucleosides, and nucleotides	86	1.0
Regulatory functions	567	6.8
DNA replication, recombination, and repair	96	1.2
Transcription	49	0.6
Translation	203	2.4
Transport and binding proteins	752	9.0
Other categories	1119	13.5
Subtotal of genes similar to genes of known function	**4348**	**52.3**
Similar to hypothetical protein	2506	30.1
Subtotal of genes similar to registered genes	**6854**	**82.4**
No similarity	1463	17.6
Total	**8317**	**100.0**

as those of *C. crecentus*, *A. tumefaciens* and *S. meliloti*, and it is likely that DNA replication terminates in this region. In *Escherichia coli*, a 28-mer sequence [-GGTGCGCATAATGTATATTATGTTAAAT-], designated as *dif*, is present in this region, which is required to convert a dimer chromosome to monomers after aberrant DNA duplication.[19] A similar sequence was found in the *B. japonicum* genome near the position where the shift of GC skew was detected at the coordinates 4,996,159–4,996,186, while the similar sequences were also observed in *M. loti*, *A. tumefaciens*, and *S. meliloti*.

3.3. Assignment of protein- and RNA-encoding genes

The potential protein-encoding regions were assigned using a combination of computer prediction by the Glimmer program and similarity search. Glimmer predicted a total of 13,619 potential protein-encoding genes in the genome after training with a dataset of sequences of highly probable protein-encoding genes in the genomes of three bacteria: *M. loti*, *S. meliloti*, and *A. tumefaciens*. By taking the sequence similarity to known genes and the relative positions into account to avoid overlaps as described in Materials and Methods, the total number of potential protein-encoding genes finally assigned to the genome was 8317 (Table 1 and Fig. 1 in the Supplement section). The average gene density was one gene in every 1095 bp. The putative protein-encoding genes thus assigned to the genome starting with either an ATG, GTG, TTG, or ATT codon are denoted by a serial number with three letters representing the species name (b), whether the ORF was longer than or shorter than 100 codons (l or s), and the transcription direction on the circular map (r or l) (Fig. 1). The codon usage frequency of the whole gene components in the genome is tabulated in Table 1 in the Supplement section.

One copy of an rRNA gene cluster was identified on the genome in the order of 16S-*trnI*-*trnA*-23S-5S at the coordinates of 1,528,226–1,533,622 (Fig. 1 in the Supplement section).

A total of 50 tRNA genes including those in the rRNA gene cluster, which represent 46 tRNA species, were assigned on the genome by sequence similarity to known bacterial tRNA genes and computer prediction using the tRNA scan-SE program (Fig. 1 and Table 1, and Table 2, Fig. 1, Fig. 2 in the Supplement section). There were three copies of *trnI*-CAU genes of completely identical sequences. Two genes, *trnN*-GUU (coordinates 6,191,247–6,191,320) and *trnF*-GAA (coordinates 6,191,318–6,191,390), were located in the same direction with a 2-bp overlap. Most of the tRNA genes were dispersed on the genome and were likely to be transcribed as single units except those in the rRNA gene cluster and *trnN*-*trnF* genes described above.

As to small RNA-encoding genes, potential genes showing sequence similarity to RNase P subunit B (*rnpB*) and tmRNAs (*ssrA*) were assigned to the genome. tmRNA is known to be involved in the degradation of aberrant proteins in the cells.[20] In *B. japonicum*, *ssrA* codes for two RNA components, tmRNA-acceptor RNA and tmRNA-coding RNA, the latter of which has been reported as unknown RNA (*sra*) involved in the development of symbiotic root nodules.[21]

3.4. Functional assignment of protein-encoding genes

Translated amino acid sequences of 8317 potential protein-encoding genes in the genome were compared with the sequences in the nr-protein databases with the BLASTP program as described in Materials and Methods. According to the results, 4348 (52%) were similar to genes of known function, 2506 (30%) showed similarity to hypothetical genes, and the remaining 1463 (18%) showed no significant similarity to any registered genes (Table 1 and Fig. 1).

The potential protein-encoding genes whose function could be anticipated were classified into 14 categories, each with different biological roles, according to the principle of Riley.[22] The numbers of genes in each category are summarized in Table 1, and the name of each gene is listed in RhizoBase at http://www.kazusa.or.jp/rhizobase/. The location, length and direction of these genes are indicated on the gene map in the Supplement section (Fig. 1), with color codes corresponding to functional categories.

3.5. Characteristic features of the predicted genes and the genome

3.5.1. Comparison of gene components among three rhizobia

To date, annotation information of the whole gene components in two rhizobia, *M. loti* and *S. meliloti*, have been published, and a total of 7281 and 6205 potential protein-encoding genes have been assigned to the respective genomes. Genes of *B. japonicum* were compared with those of *M. loti* and *S. meliloti* under the criteria described in Materials and Methods. Comparison with higher stringency indicated that 2798 *B. japonicum* genes, comprising 34% of the 8317 potential protein-encoding genes, have matched genes in both the *M. loti* and *S. meliloti* genomes, 632 of which were genes of unknown function. Seven hundred seventy-five (9%) and 584 (7%) *B. japonicum* genes were commonly found in either the *M. loti* or *S. meliloti* genome, respectively.[2,3] The remaining 4160 genes (50%) were unique to *B. japonicum.* Two hundred thirty-three of these were genes of known function, including those for NolY (*bll2016*), NoeD (*blr1633*), NodY (*blr2024*), nodulation protein (*blr2062*), NoeE (*blr2073* and *blr2074*), NolZ (*bsl2015*), and NolM (*bsr2032*). With lower stringency, where most members of a given gene family can be grouped into a cluster, 5382 *B. japonicum* genes (65%) showed significant sequence similarity to genes in both *M. loti* and *S. meliloti.* Five hundred eighty-six (7%) and 420 (5%) *B. japonicum* genes were conserved in either the *M. loti* or *S. meliloti* genome, respectively, and 1929 genes (23%) were only found in *B. japonicum.* Both results indicate that significant portions of the gene components in the genomes are unique to the species.

3.5.2. Symbiosis island

The symbiosis island was first identified in *M. loti* ICMP3153 as a transmittable DNA segment of approximately 500 kb in length containing a cluster of symbiotic genes.[23] It was shown that this region was inserted into a phe-tRNA gene in the genome with a 17-bp duplication of the 3′ terminal portion of the gene. A 611-kb DNA region with lower GC contents flanked by a complete and a partial phe-tRNA genes with a 17-bp duplication was also reported as a putative symbiosis island in *M. loti* MAFF303099.[2] In *B. japonicum*, on the other hand, a 410-kb DNA segment containing clusters of genes for symbiotic nitrogen fixation has been assigned as a symbiotic region.[7] Sequence analysis of the entire *B. japonicum* genome revealed the presence of a 680-kb DNA region of lower GC content (59.4%, see Fig. 1) at the coordinates 1681–2362 kb, in which the 410-kb symbiotic DNA region described above (coordinates 1,878,486–2,289,058) was completely included. A val-tRNA gene, instead of the phe-tRNA gene in *M. loti*, was found at one end of this region at the coordinates 2,362,342–2,362,416, but no trace of terminal duplication was detected in the other end of this region. However, a 45-bp segment of the 3′ terminal portion of the val-tRNA gene was found at the coordinates 7,939,490–7,939,534, 3528-kb apart from the intact val-tRNA gene, and a putative integrase gene (*bll7217*) was found next to this tRNA gene segment, as in *M. loti.* Considering the fact that the GC content of a region (coordinates 7934–7940 kb) adjacent to this segment were lower than the average of the entire genome (59.2% vs. 64.1%), it can be postulated that an ancient larger symbiosis island was split into two, the 680-kb DNA region (a putative symbiosis island A at the coordinates 1681–2362 kb) and the 6-kb DNA region (a putative symbiosis island B at the coordinates 7934–7940 kb), by a large-scale genome rearrangement.

A total of 655 potential protein-encoding genes were tentatively assigned to putative symbiosis islands A and B. Three hundred one out of 655 genes (46%) showed sequence similarity to genes of known function. These included two clusters of genes *nifDKENX*-[8 genes]-*nifS*-[2 genes]-*nifB*-[9 genes]-*nifH*-[3 genes]-*fixBCX* (*blr1743*–*bsr1775* at the coordinates 1,907,825–1,935,084) and *nolZY*-[2 genes]-*nolA*-[1 gene]-*nodD2*-[1 gene]-*nodD1YABCSUIJ*-*nolMNO*-*nodZ*-*fixR*-*nifA*-*fixA* (*bsl2015*–*blr2038* at the coordinates 2,175,594–2,199,393) as previously reported by Kündig et al.[6] In addition, *trbLFGI* (*blr1617*-*blr1618*-*blr1619*-*blr1620*) were found at the coordinates 1,773,989–1,777,834, while *traG* was disrupted, which may impair the self-transmissibility of the symbiosis island(s) in the USDA110 strain. Four genes, *nodM* (*blr1632*), *nolK* (*bll1630*), *noeL* (*bll1631*), and *noeD* (*blr1633*), formed a gene cluster at the coordinates 1,788,234–1,794,230. Three gene clusters for unknown ABC transporters, *blr1601* (substrate-binding protein)-*blr1602* (permease)-*blr1603* (permease)-*blr1604* (ATP-binding protein), *blr1679* (permease)-*blr1680* (permease) and *blr2170* (substrate-binding protein)-*blr2171* (permease), and genes for sensor/regulator hybrid of a two-component signal transduction system, *bll2176* (sensor/regulator hybrid) and *blr2178* (sensor/regulator hybrid), were also found in the putative symbiosis islands. It is remarkable that 100 out of 167 genes for transposases were located in presumptive symbiosis island A. This accounts for 60% of the transposase genes in the genome and 15% of the putative protein-encoding genes in symbiosis island A.

3.5.3. Insertion sequences

A total of 167 genes for putative transposases were assigned to the genome of *B. japonicum.* Of these, 133 were identified as components of insertion sequences (ISs), and the remaining 34 were not due to structural rearrangements. With respect to ISs, 104 copies were classified into 20 groups, each comprising 1 to 15 members on the basis of the type and number of transposases. Structural features of each IS group are summarized in Tables 3 and 4 and Fig. 3 in the Supplement section.

Most of the members exhibited typical characteristics of ISs, such as terminal inverted repeats of 3 to 48 bp and direct repeats at the insertion sites. It has been reported that insertion of RSα occurs at a target sequence 5′-CTAG-3′ in *B. japonicum*.[24] Analysis of the entire genome sequence in this study revealed that 7 groups including RSα seemed to have preferable sequences for insertion, and it is notable that 6 of them, which comprise 26 IS members (34% of the ISs in the genome), targeted 5′-CTAG-3′. The tetramer sequence 5′-CTAG-3′ may play a significant role in the transposition of ISs and IS-mediated genome rearrangement in *B. japonicum*.

It is remarkable that 63 out of 104 ISs (60%) were located in presumptive symbiosis island A. ISs were also frequently found in a 206-kb DNA region at the coordinates 8975-0-75 kb. Of 224 putative protein-encoding genes assigned to this region, 89 (40%) did not show any sequence similarity to genes of known function, and 36 of the remaining 135 genes of known function code for transposases. This region had lower GC content (60.2%) than the rest of the genome (Fig. 1), and a gene for probable site-specific integrase/recombinase (*blr8174*) was identified near the border. These characteristic features may indicate that this region is a mobile genetic element,[25] though terminal duplication was not detected.

3.5.4. Insertion of DNA elements in tRNA genes

Duplication of portions of tRNA genes was found at 14 locations in the genome. They were separated from the respective intact genes by DNA segments of 4 to 97 kb; therefore, it is likely that these are traces of insertion of DNA elements into tRNA genes.[25] Table 5 in the Supplement section lists the species of tRNA genes, their positions, the length of duplications and DNA regions that separate the duplicated tRNA gene segments. The size of the presumptive DNA elements is diverse among the insertions, and no apparent sequence similarity was detected among them except that many of them contained putative genes for integrases. The GC contents of these DNA elements were lower than the average GC content of the genome. These features, common characteristics of mobile genetic elements, strongly suggest that insertion of the DNA elements into tRNA genes took place, though the origin of the element as well as the ability of the extant elements to transpose remain to be studied.

In and near the putative symbiosis island A, 3 copies of ile-tRNA genes (Fig. 2 in the Supplement section) and 2 copies of 49-bp 3′ terminal portions of the genes (coordinates 1,702,873–1,702,921 and 2,473,140–2,473,188, respectively) were found. Two pairs of 49-bp terminal portions were assigned as the *trnI1* and *trnI2* elements as shown in Table 5 in the Supplement section. There was another copy of the 49-bp 3′ terminal portion of the ile-tRNA gene at the coordinates 7,676,010–7,676,058, in the *trnM1* element, and the origin of this segment is unknown. This feature may indicate that complex genome rearrangements took place before and after insertion of the symbiosis island.

3.5.5. Genes related to symbiotic nitrogen fixation

Genes related to various steps of the symbiotic nitrogen fixation process have been identified in the genome of *B. japonicum*, as listed in the gene category list of RhizoBase at http://www.kazusa.or.jp/rhizobase/. Remarkable features noted for the *B. japonicum* genome so far are itemized below.

1. Genes showing sequence similarity to that of *virA* (*blr2697*) and *virG* (*blr2694*) were identified. Homologues of *virD* and *virE* were not detected.

2. There was a large gene cluster (*blr2358–bll2381*) at the coordinates 2,561,178–2,595,273 containing *exoP* (succinoglycan biosynthesis transport protein: *bll2362*), *exoZ* (probable exopolysaccharide production protein: *blr2369*),[26] 6 genes similar to those for glycosyl transferase, which is likely to be involved in exopolysaccharide synthesis.

3. Two adjacent genes, *blr1993* and *blr1994*, located in the presumptive symbiosis island were assigned as putative polygalacturonase and pectinesterase, respectively. Outside of the symbiosis island, *bll7427*, *blr3367*, and *bll2241* showed significant sequence similarity to those for ligninase (*ligE*), cellulase, and endo-1,4-beta-xylase, respectively. These genes may take part in the degradation of plant cell walls during the infection process.[27]

4. The sequence of the putative gene product of *blr0241* is 80% identical with that of ACC deaminase of *M. loti* (*mlr5932*). This gene is likely to be involved in the reduction of ethylene concentration during the nodule formation process.[28]

5. *nolK*/*noeL*/*nodM*/*noeD* (*bll1630-bll1631-blr1632-blr1633*) formed a gene cluster in the presumptive symbiosis island at the coordinates 1,788,234–1,794,230. A homologue of *nodQ* (*bll1475*) was also found but was 178 kb distant from *nolK* outside of the symbiosis island.

6. Of the four copies of *dnaJ* homologues (*bll0184*, *blr0680*, *blr2626*, *blr4653*), *blr0680* showed the highest degree of sequence similarity to that of *dnaJ* in *Rhizobium leguminosarum*, which was reported to be

involved in the regulation of nodule formation and nitrogen fixation.[29]

7. *blr1755* showed significant sequence similarity to *iscN* of *R. etli*, which plays a critical role in nitrogen fixation.[30] It formed a cluster with the *fixU* and *nifS* homologues (*bsr1757* and *blr1756*, respectively), as in *R. etli.*

8. Two adjacent genes, *nfeD* (*bll4952*) and a homologue of a gene for stomatin-like protein (*bll4951*), both of which are assumed to be involved in nodulation competitiveness,[31,32] were identified at the coordinates 5,494,753–5,496,917.

9. The putative products of *fixK2/fixLJ/fixNOPQ/fixGHIS* genes (*bll2757-bll2759-bll2760-blr2763-blr2764-bsr2765-blr2766-blr2767-blr2768-blr2769-bsr2770*) are the oxygen-sensing cascade and high-affinity terminal oxidase that are required for microaerobic respiration and nitrogen fixation in nodules.[33] This gene cluster was located at a position 674 kb from main symbiosis island A, similar to its location on the *M. loti* chromosome.[2] This system may have originated from ancient non-symbiotic bradyrhizobia.

3.5.6. Other features

In addition to the various characteristics of genes in the *B. japonicum* genome described above, other remarkable features are itemized below.

1. All the genes for the key carbon fixation enzymes seem to be present. These include genes related to the Calvin-Benson-Bassham (CBB) cycle (*cbbR-cbbFPTALSX*) which formed a cluster (*bll2580–blr2587*) at the coordinates 2,822,117–2,831,210, a PEP carboxylase gene (*blr2955*), and a carbonic anhydrase gene (*bll2065*). *bll1137*, *bll4865*, and *bll4863* also showed sequence similarity to the gene for carbonic anhydrase. A gene showing sequence similarity to that for ribulose-phosphate 3-epimerase (*blr2588*) was located downstream of *blr2587*, as in *S. meliloti*,[34] suggesting that this gene may be a part of the gene cluster for the CBB cycle. These fixation systems of carbon dioxide are probably required for chemoautotrophic growth of *B. japonicum* USDA110 in the presence of hydrogen gas as an energy source.[4]

2. Ni-Mo hydrogenase mediates uptake of dihydrogen via nitrogenase, and increases N_2 fixation efficiency. Two sets of a hydrogenase gene cluster were found, one inside and another outside of the symbiotic region (coordinates 1,888,916–1,902,575 and 7,624,531–7,649,197, respectively). Interestingly, a complete set of *hupNCUVSLCDFGHIJK*, *hypABFCD'E*, and *hoxXA* were located outside of the symbiosis island, and were highly homologous to the uptake hydrogenase genes of *B. japonicum* strain USDA122DES.[35] On the other hand, some variant genes for uptake hydrogenase (*hupSLC*, *hypAFDE*) were found inside the symbiosis region.[7]

3. It was shown that *B. japonicum* had a NapAB-type periplasmic dissimilatory nitrate reductase, as does *S. meliloti*.[36] *napEDABC* formed a cluster, while *norCB*, *nirK*, and *nosZ* were dispersed throughout the genome.[37] The putative products of these genes are likely to be involved in denitrification from nitrate to N_2.

4. Homologues of *trbBCDEJLFGI* (*bll8288-bll8287-bsl8286-bll8285-bll8284-bll8282-bll8281-bll8280-bll8279*), *traG* (*bll0043*), and *traF* (*bll0047*), which were essential for conjugative transfer of Ti-plasmid in *A. tumefaciens*,[38] were assigned to the putative mobile genetic element at the coordinates 8975-0-75 kb (see section 3.5.3). Taking into consideration the fact that the homologues are found on either the symbiosis island or pMLb in *M. loti*, these genes may play a role in the transfer of this genetic element. Inconsistency between the previously reported physical map of the genome and the sequence determined in this study was attributed to this region, suggesting that this element may be self-transmissibile.[6]

5. A significant number of genes for flagella formation and mobility, including those for chemotaxis, *cheAWY* (*blr2192-blr2193-blr2194*), were found in *B. japonicum*.[39]

6. A homologue of a *tfdA*-like gene (ketoglutarate-dependent 2,4-dichlorophenoxyacetate dioxygenase), which is involved in the degradation of 2,4-D (2,4-dichlorophenol) and phenoxyacetate, was found (*bll1493*).[40]

7. *B. japonicum* USDA110 simultaneously possesses type III and IV secretion systems. Gene homologues of *ctpACDFGHI*, *pilA* for the type IV secretion system (*bsl7141-bll1440-bll1439-bll1437-bll1436-bll1435-bll1434*, and *bsl6587*) were found mainly outside of symbiosis islands, whereas those for *rhcC$_1$C$_2$JNQRSTUV* for the type III secretion system (*bll1811-bll1842-blr1813-blr1816-blr1818-blr1819-bsr1820-blr1821-blr1822-bll1800*) were located on symbiosis island A. Interestingly, genes for the type III and type IV secretion systems exist within symbiosis islands of *M. loti* strains MAFF303099 and R7A, respectively.[41] These secretion systems may be involved in fine turning of host-specific nodulation.

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 appears in DDBJ/GenBank/EMBL under accession number BA000040.

Acknowledgements: This work was supported by the Kazusa DNA Research Institute Foundation. Thanks are also due to Shimadzu Corporation and Wako Pure Chemical Industries, Ltd. for technical assistance.

References

1. 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.
2. Kaneko, T., Nakamura, Y., Sato, S. et al. 2000, Complete genome structure of the nitrogen-fixing symbiotic bacterium *Mesorhizobium loti*, *DNA Res.*, **7**, 331–338.
3. Galibert, F., Finan, T. M., Long, S. R. et al. 2001, The composite genome of the legume symbiont *Sinorhizobium meliloti*, *Science*, **293**, 668–672.
4. Minamisawa, K. and Mitsui, H. 2000, In: Bollag, J. and Stotzky, G. (eds) Soil Biochemistry, Vol. 10. Marcel Dekker, Inc., New York, pp. 349–377.
5. Mathis, J. N., McMillin, D. E., Champion, R. A., and Hunt, P. G. 1997, Genetic variation in two cultures of *Bradyrhizobium japonicum* USDA110 differing in their ability to impact drought tolerance to soybean, *Curr. Microbiol.*, **35**, 363–366.
6. Kündig, C., Hennecke, H., and Göttfert, M. 1993, Correlated physical and genetic map of the *Bradyrhizobium japonicum* 110 genome, *J. Bacteriol.*, **175**, 613–622.
7. Göttfert, M., Röthlisberger, S., Kündig, C., Beck, C., Marty, R., and Hennecke, H. 2001, Potential symbiosis-specific genes uncovered by sequencing a 410-kilobase DNA region of the *Bradyrhizobium japonicum* chromosome, *J. Bacteriol.*, **183**, 1405–1412.
8. Tomkins, J. P., Wood, T. C., Stacey, M. G. et al. 2001, A marker-dense physical map of the *Bradyrhizobium japonicum* genome, *Genome Res.*, **11**, 1434–1440.
9. Hattori, Y., Omori, H., Hanyu, M. et al. 2002, Ordered cosmid library of the *Mesorhizobium loti* MAFF303099 genome for systematic gene disruption and complementation analysis, *Plant Cell Physiol.*, **43**, 1542–1557.
10. 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.
11. 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.
12. Altschul, S. F., Madden, T. L., Schaffer, A. A. et al. 1997, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.*, **25**, 3389–3402.
13. 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.
14. Lobry, J. R. 1996, Asymmetric substitution patterns in the two DNA strands of bacteria, *Mol. Biol. Evol.*, **13**, 660–665.
15. Brassinga, A. K., Siam, R., and Marczynski, G. T. 2001, Conserved gene cluster at replication origins of the alpha-proteobacteria *Caulobacter crescentus* and *Rickettsia prowazekii*, *J. Bacteriol.*, **183**, 1824–1829.
16. Nierman, W. C., Feldblyum, T. V., Laub, M. T. et al. 2001, Complete genome sequence of *Caulobacter crescentus*, *Proc. Natl. Acad. Sci. USA*, **98**, 4136–4141.
17. Capela, D., Barloy-Hubler, F., Gouzy, J. et al. 2001, Analysis of the chromosome sequence of the legume symbiont *Sinorhizobium meliloti* strain 1021, *Proc. Natl. Acad. Sci. USA*, **98**, 9877–9882.
18. Goodner, B., Hinkle, G., Gattung, S. et al. 2001, Genome sequence of the plant pathogen and biotechnology agent *Agrobacterium tumefaciens* C58, *Science*, **194**, 2323–2328.
19. Blakely, G. W., Davidson, A. O., and Sherratt, D. J. 2000, Sequential strand exchange by XerC and XerD during site-specific recombination at *dif*, *J. Biol. Chem.*, **275**, 9930–9936.
20. Keiler, K. C., Shapiro, L., and Williams, K. P. 2000, tmRNAs that encode proteolysis-inducing tags are found in all known bacterial genomes: A two-piece tmRNA functions in *Caulobacter*, *Proc. Natl. Acad. Sci. USA*, **97**, 7778–7883.
21. Ebeling, S., Kündig, C., and Hennecke, H. 1991, Discovery of a rhizobial RNA that is essential for symbiotic root nodule development, *J. Bacteriol.*, **173**, 6373–6382.
22. Riley, M. 1993, Functions of the gene products of *Escherichia coli*, *Microbiol. Rev.*, **57**, 862–952.
23. 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.
24. Kaluza, K., Hahn, M., and Hennecke, H. 1985, Repeated sequences similar to insertion elements clustered around the *nif* region of the *Rhizobium japonicum* genome, *J. Bacteriol.*, **162**, 535–542.
25. Hayashi, T., Makino, K., Ohnishi, M. et al. 2001, Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12, *DNA Res.*, **8**, 11–22.
26. Becker, B. U., Kosch, K., Parniske, M., and Muller, P. 1998, Exopolysaccharide (EPS) synthesis in *Bradyrhizobium japonicum*: sequence, operon structure and mutational analysis of an *exo* gene cluster, *Mol. Gen. Genet.*, **259**, 161–171.
27. Mateos, P. F., Baker, D. L., Petersen, M. et al. 2001, Erosion of root epidermal cell walls by *Rhizobium* polysaccharide-degrading enzymes as related to primary host infection in the *Rhizobium*-legume symbiosis, *Can. J. Microbiol.*, **47**, 475–487.
28. Penmetsa, R. V. and Cook, D. R. 1997, A legume ethylene-insensitive mutant hyperinfected by its Rhizobial symbiont, *Science*, **275**, 527–530.
29. Labidi, M., Laberge, S., Vezina, L. P., and Antoun, H. 2000, The *dnaJ* (hsp40) locus in *Rhizobium leguminosarum* bv. *phaseoli* is required for the establishment of an effective symbiosis with *Phaseolus vulgaris*, *Mol. Plant Microbe Interact.*, **13**, 1271–1274.
30. Dombrecht, B., Tesfay, M. Z., Verreth, C. et al. 2002,

The *Rhizobium etli* gene *iscN* is highly expressed in bacteroids and required for nitrogen fixation, *Mol. Genet. Genomics*, **267**, 820–828.

31. You, Z., Gao, X., Ho, M. M., and Borthakur, D. 1998, A stomatin-like protein encoded by the *slp* gene of *Rhizobium etli* is required for nodulation competitiveness on the common bean, *Microbiology*, **144**, 2619–2627.
32. Garcia-Rodriguez, F. M. and Toro, N. 2000, *Sinorhizobium meliloti nfe* (nodulation formation efficiency) genes exhibit temporal and spatial expression patterns similar to those of genes involved in symbiotic nitrogen fixation, *Mol. Plant Microbe Interact.*, **13**, 583–591.
33. Nellen-Anthamatten, D., Rossi, P., Preisig, O. et al. 1998, *Bradyrhizobium japonicum* FixK2, a crucial distributor in the FixLJ-dependent regulatory cascade for control of genes inducible by low oxygen levels, *J. Bacteriol.*, **180**, 5251–5255.
34. Finan, T. M., Weidner, S., Wong, K. et al. 2001, The complete sequence of the 1,683-kb pSymB megaplasmid from the N_2-fixing endosymbiont *Sinorhizobium meliloti*, *Proc. Natl. Acad. Sci. USA*, **98**, 9889–9894.
35. Maier, R. J. and Triplett, E. W. 1996, Toward more productive, efficient, and competitive nitrogen-fixing symbiotic bacteria, *Crit. Rev. Plant Sci.*, **15**, 191–234.
36. Barnett, M. J., Fisher, R. F., Jones, T. et al. 2001, Nucleotide sequence and predicted functions of the entire *Sinorhizobium meliloti* pSymA megaplasmid, *Proc. Natl. Acad. Sci. USA*, **98**, 9883–9888.
37. Mesa, S., Göttfert, M., and Bedmar, E. J. 2001, The *nir*, *nor*, and *nos* denitrification genes are dispersed over the *Bradyrhizobium japonicum* chromosome, *Arch. Microbiol.*, **176**, 136–142.
38. Li, P. L., Everhart, D. M., and Farrand, S. K. 1998, Genetic and sequence analysis of the pTiC58 *trb* locus, encoding a mating-pair formation system related to members of the type IV secretion family, *J. Bacteriol.*, **180**, 6164–6172.
39. Barbour, W. M., Hattermann, D. R., and Stacey, G. 1991, Chemotaxis of *Bradyrhizobium japonicum* to soybean exudates, *Appl. Environ. Microbiol.*, **57**, 2635–2639.
40. Itoh, K., Kanda, R., Sumita, Y. et al. 2002, *tfdA*-like genes in 2,4-dichlorophenoxyacetic acid-degrading bacteria belonging to the *Bradyrhizobium-Agromonas-Nitrobacter-Afipia* cluster in alpha-Proteobacteria, *Appl. Environ. Microbiol.*, **68**, 3449–3454.
41. Sullivan, J. T., Trzebiatowski, J. R., Cruickshank, R. W. et al. 2002, Comparative sequence analysis of the symbiosis island of *Mesorhizobium loti* strain R7A, *J. Bacteriol.*, **184**, 3086–3095.