DNA Research **10**, 137–145 (2003)

Complete Genome Structure of *Gloeobacter violaceus* PCC 7421, a Cyanobacterium that Lacks Thylakoids

Yasukazu Nakamura,[1] Takakazu Kaneko,[1] Shusei Sato,[1] Mamoru Mimuro,[2,3] Hideaki Miyashita,[2,3] Tohru Tsuchiya,[2,3] Shigemi Sasamoto,[1] Akiko Watanabe,[1] Kumiko Kawashima,[1] Yoshie Kishida,[1] Chiaki Kiyokawa,[1] Mitsuyo Kohara,[1] Midori Matsumoto,[1] Ai Matsuno,[1] Naomi Nakazaki,[1] Sayaka Shimpo,[1] Chie Takeuchi,[1] Manabu Yamada,[1] and Satoshi Tabata[1,*]

Kazusa DNA Research Institute, 2-6-7 Kazusa-kamatari, Kisarazu, Chiba 292-0818, Japan,[1] *Department of Technology and Ecology, Hall of Global Environmental Research, Kyoto University, Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501, Japan,*[2] *and Graduate School of Human and Environmental Studies, Kyoto University, Kyoto 606-8501, Japan*[3]

(Received 24 July 2003)

Abstract

The nucleotide sequence of the entire genome of a cyanobacterium *Gloeobacter violaceus* PCC 7421 was determined. The genome of *G. violaceus* was a single circular chromosome 4,659,019 bp long with an average GC content of 62%. No plasmid was detected. The chromosome comprises 4430 potential protein-encoding genes, one set of rRNA genes, 45 tRNA genes representing 44 tRNA species and genes for tmRNA, B subunit of RNase P, SRP RNA and 6Sa RNA. Forty-one percent of the potential protein-encoding genes showed sequence similarity to genes of known function, 37% to hypothetical genes, and the remaining 22% had no apparent similarity to reported genes. Comparison of the assigned gene components with those of other cyanobacteria has unveiled distinctive features of the *G. violaceus* genome. Genes for PsaI, PsaJ, PsaK, and PsaX for Photosystem I and PsbY, PsbZ and Psb27 for Photosystem II were missing, and those for PsaF, PsbO, PsbU, and PsbV were poorly conserved. *cpcG* for a rod core linker peptide for phycobilisomes and *nblA* related to the degradation of phycobilisomes were also missing. Potential signal peptides of the presumptive products of *petJ* and *petE* for soluble electron transfer catalysts were less conserved than the remaining portions. These observations may be related to the fact that photosynthesis in *G. violaceus* takes place not in thylakoid membranes but in the cytoplasmic membrane. A large number of genes for sigma factors and transcription factors in the LuxR, LysR, PadR, TetR, and MarR families could be identified, while those for major elements for circadian clock, *kaiABC* were not found. These differences may reflect the phylogenetic distance between *G. violaceus* and other cyanobacteria.

Key words: cyanobacterium; *Gloeobacter violaceus*; thylakoid membranes; genomic sequencing

1. Introduction

Gloeobacter violaceus PCC 7421 (Gloeobacter, hereafter) is a rod-shape unicellular cyanobacterium isolated from calcareous rock in Switzerland.[1] It is an obligate photoautotroph containing chlorophyll (Chl) *a*, carotenoids, and phycobiliproteins (phycoerythrin, phycocyanin and allophycocyanin) with a lower relative ratio of Chl *a* to phycobiliproteins than that of typical cyanobacteria. Gloeobacter is sensitive to strong light and its generation time is relatively long (approximately 72 hr).[1] The recent molecular phylogenetic analysis shows that the Gloeobacter linage diverged in the earliest within the radiation of cyanobacteria and chloroplasts.[2] These features strongly support the idea that this species possesses ancestral characteristics for oxygenic photosynthesis.

In fact, Gloeobacter possesses a number of unique characteristics. The cells lack thylakoid membranes, and microscopic observation strongly suggests that machinery for photosynthesis is located in the cytoplasmic membrane instead of thylakoid membranes where the machinery is found in other cyanobacteria.[1] This means that components facing the lumen in the cytoplasm in other cyanobacteria are exposed to periplasm in Gloeobacter, thus the photosynthetic electron transfer system should co-exist in the cytoplasmic membrane with a respiratory system by sharing some components. Thus, it seems likely that several processes

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

such as oxygen evolution and electron transfer mediated by cytochrome c_{553} and plastocyanin occur in the periplasm instead of the lumen. The morphology of phycobilisomes is distinct from other cyanobacteria: Phycobiliproteins form rod-shaped elements and these elements form bundle-shaped aggregates[3] which are situated vertically adjacent to the inner surface of the cytoplasmic membrane. It is also remarkable that composition of fatty acids is different: Sulfoquinovosyl diacylglycerol (SQDG), which is thought to have an important role in photosystem stabilization, is absent in Gloeobacter[4] while the content of polyunsaturated fatty acids (PUFA) is high.[1]

To reveal the genetic background responsible for these characteristics and to obtain clues to the origin and evolution of oxygenic photosynthesis, we determined the nucleotide sequence of the entire genome of Gloeobacter. Here, we describe the characteristic features of the genes and the genome of Gloeobacter, and the results of comparison with three previously sequenced Cyanobacteria: *Synechocystis* sp. PCC 6803 (Synechocystis, hereafter),[5] *Anabaena* sp. PCC 7120 (Anabaena, hereafter),[6] and *Thermosynechococcus elongatus* BP-1 (Thermosynechococcus, hereafter).[7]

2. Materials and Methods

2.1. Bacterial strain and genomic libraries

Gloeobacter violaceus PCC 7421 was obtained from the Pasteur Culture Collection, and the total cellular DNA was purified according to standard procedures.

For genome sequencing, four random genomic libraries with three types of cloning vectors were constructed from the total DNA to minimize cloning bias: GLE and GLB with approximately 1.0-kb (element clones) and 2.5-kb inserts (bridge clones), respectively, both cloned into M13mp18, GLP with 9.4-kb inserts on average (plasmid clones) cloned into pUC18, and GLL bearing inserts of approximately 18-kb in a BAC vector, pBeloBAC11.

2.2. DNA sequencing

The entire genome of Gloeobacter was determined by the whole-genome shotgun method in combination with the "bridging shotgun" strategy.[8]

The nucleotide sequences of one end of the element clones and both ends of the clones from the other three libraries were analyzed using the Dye-terminator Cycle Sequencing kit with DNA sequencers type 377XL (Applied Biosystems, USA). The accumulated sequences were assembled using the Phrap program (Philip Green, University of Washington, Seattle, USA). The end-sequence data from the bridge, plasmid, and BAC clones facilitated the gap-closure process as well as accurate reconstruction of the sequence of the entire genome. The final gaps in the sequences were filled by the primer walking method. The RNA sequencing method was adopted for the regions that were difficult to sequence due to high GC contents or stable secondary structures. 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 BAC clones.

2.3. Gene assignment and annotation

Protein- and RNA-encoding regions were assigned by a combination of computer prediction and similarity searches, as described previously.

Prediction of protein-encoding regions was carried out with the Glimmer 2.02 program.[9] Prior to prediction, the matrix was generated for the Gloeobacter genome by training with a dataset of 2411 open reading frames that showed a high degree of sequence similarity to genes registered in the non-redundant protein database (nr-database). 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 nr-database with the BLASTP program.[10] In parallel, all the predicted intergenic sequences were compared with those in the nr-database using the BLASTX program to identify genes that had escaped prediction. 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.

The functions of the assigned genes were deduced on the basis of the sequence similarity of their presumptive protein products to those of genes of known function and to the protein motifs in the Pfam database.[11] 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. For the motif search against the Pfam database, a score of less than 10^{-4} was considered significant.

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 using the tRNA scan-SE 1.21 program[12] in combination with similarity searches.

Comparison of the gene components between Gloeobacter and other cyanobacteria 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 less than 10^{-4}.

The GC skew analysis was performed as described by Lobry.[13]

3. Results and Discussion

3.1. Sequencing and structural features of the Gloeobacter genome

The nucleotide sequence of the entire genome of Gloeobacter was determined according to the modified whole genome shotgun method as described in Materials and Methods. A total of 38,303 random sequences corresponding to approximately 4.5 genome-equivalents were assembled to generate draft sequences. Then finishing was carried out by visually editing the above sequences and by gap closing with additional sequencing to obtain 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 BAC clone with the computed distance between the end sequences of the clones. The genome of Gloeobacter thus deduced was a circular molecule of 4,659,019 bp with an average GC content of 62%. 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 *Swa* I (Fig. 1 and Fig. 1 in the Supplement section).

The innermost circle of Fig. 1 shows uneven distribution of GC content along the genome, some of which are attributed to the presence of genes for transposases and glycosyltransferase-like proteins as is also observed in Thermosynechococcus and Anabaena.[6,7] Two regions of lower GC content at the coordinates of 189 kb–207 kb and 2894 kb–2907 kb were situated adjacent to *trnL*-UAA and *trnR*-CCU, respectively, suggesting insertion of exogenous DNA elements into the tRNA genes. This speculation is supported by the fact that a 19-bp duplication of the 3′ terminal region of *trnR*-CCU, which is likely to be transferred from the potential DNA element during the insertion process, was observed at the opposite end of the 2894 kb–2907 kb region. No such duplication was detected for the 189 kb–207 kb region, probably due to accumulation of mutations.

A GC skew analysis was performed to locate the probable origin and terminator of DNA replication, but no apparent shift was observed in any region of the genome.

An 8-bp palindromic sequence (5′-GCGATCGC-3′) named HIP1 is frequently found in the genomes of a variety of cyanobacteria.[14] HIP1 was present in the Gloeobacter genome (318 copies) but the frequency of occurrence was much lower (1 copy per 14,651 bp) than those of Thermosynechococcus (1 copy per 705 bp), Synechocystis (1 copy per 1131 bp), and Anabaena (1 copy per 1219 bp).

3.2. Assignment of RNA-encoding genes

Based on the sequence similarity to the reported structural RNAs and gene prediction by the tRNA scan-SE program, one copy of an rRNA gene cluster was assigned at the coordinates of 1,567,799–1,572,715 of the genome in the order of 16S-*trnI*-*trnA*-23S-5S in the counter-clockwise direction (Fig. 1 in the Supplement section). A total of 45 tRNA genes including those in the rRNA gene cluster representing 44 tRNA species were also identified (Fig. 1 and Table 1, Table 2, Fig. 1, and Fig. 2 in the Supplement section). Most of the tRNA genes were dispersed on the genome and were likely to be transcribed as single units, except for those in the rRNA gene cluster and *trnT*-GGU–*trnY*-GUA genes at coordinates 1,412,926–1,413,090, which were oriented in the same direction with an 11-bp gap between them.

There are two types of *trnL*-UAA in cyanobacteria, one with and the other without a group I intron.[15] Gloeobacter had an intron-less *trnL*-UAA. This gene was unique in that it has a shorter variable region (5 bp) than those of Thermosynechococcus, Synechocystis, and Anabaena (17 bp) and that it has a more diverse aminoacyl stem region: 5′-CCGAGCG for Gloeobacter compared to 5′-GGGGGCG for Thermosynechococcus and Anabaena, and 5′-GGGGGTG for Synechocystis. *trnfM*-CAU assigned at the coordinates of 3,218,774–3,218,507 contained an intron, as reported in other cyanobacteria.[15]

Gloeobacter had a single gene (coordinates 2,366,323–2,366,579) for tm (transfer-messenger) RNA known to be involved in degradation of aberrantly synthesized proteins. The 5′ and 3′ portions of this gene, which are separated by a 7-bp intervening sequence, had the capacity to encode an acceptor RNA and a coding RNA, respectively, and it is likely that the two RNAs generated by processing form a single functional tmRNA molecule by base-paring. Such "two-piece" type genes have been found in several *Prochlorococcus* and *Synechococcus* species, while tmRNAs in Thermosynechococcus, Synechocystis, and Anabaena are contiguously encoded.[16,17] However, sequence similarity of the tmRNA gene in Gloeobacter to that of *Prochlorococcus* was limited to its 5′ half, and the remaining half had to be assigned based on the common secondary structure. A tag peptide was predicted to be 17 amino acid residues, ATNNVVPFARARATVAA.

Potential genes for small RNAs showing sequence similarity to a B subunit of RNase P, SRP (signal recognition particle) RNA and 6Sa RNA were assigned to the genome based on sequence similarity to reported genes.

3.3. Assignment of protein-encoding genes

The potential protein-encoding regions were assigned by a combination of computer prediction by the Glimmer program and similarity search as described in Materials

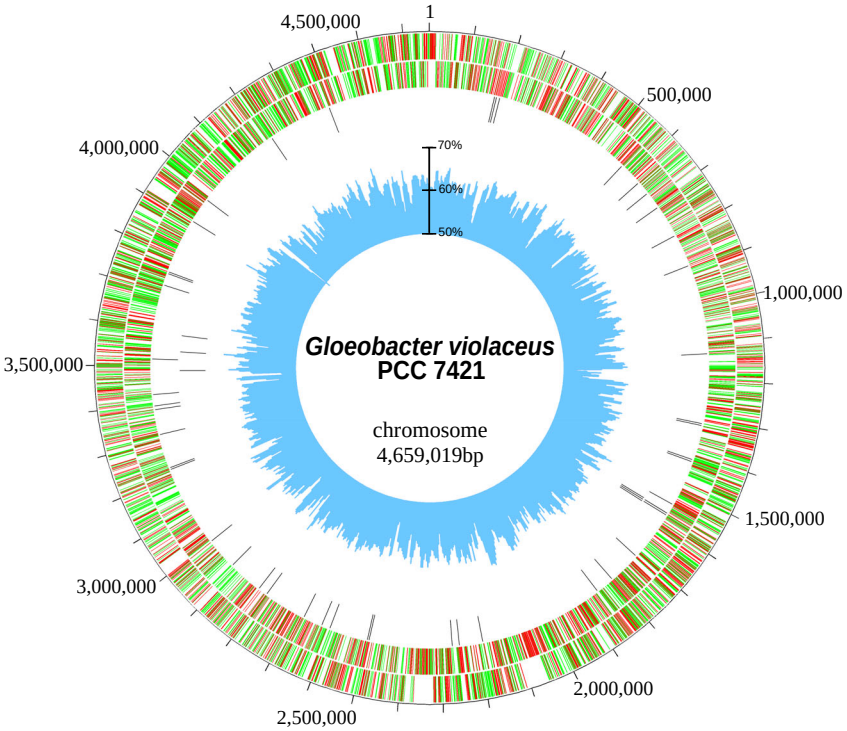

Figure 1. Circular representation of the chromosome of *Gloeobacter violaceus* sp. PCC 7421. The scale indicates the location in bp starting from the *Swa* I recognition site. The bars in the outermost and the second circles show the positions of the putative protein-encoding genes in the clockwise and counter-clockwise 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 circle with a scale shows the average GC percent calculated with a window size of 10 kb.

and Methods. Glimmer predicted a total of 4951 potential protein-encoding regions in the genome after training with a dataset of 2411 sequences of highly probable protein-encoding genes in the genome. By taking the sequence similarity to known genes and the relative positions into account to avoid overlaps, the total number of potential protein-encoding genes finally assigned to the genome was 4430 (Table 1 and Fig. 1 in the Supplement section). The average gene density was one gene every 1052 bp. The putative protein-encoding genes thus assigned starting with either an ATG, GTG, TTG, or ATT codon are denoted by a serial number with three letters representing the species name (g), 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.

Functional assignment of the 4430 potential protein-encoding genes in the genome was performed by similarity search against the nr- and Pfam databases as described in Materials and Methods. A total of 1836 (41%) showed sequence similarity to genes of known function, 1635 (37%) to hypothetical genes, and the remaining 959 (22%) did not show 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 of different biological roles, according to the principle of Riley.[18] The numbers of genes in each category are summarized in Table 1, and the name of each gene is listed in CyanoBase at http://www.kazusa.or.jp/cyanobase/. 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.

3.4. Characteristic features of the predicted genes and the genome

3.4.1. Comparison of gene components among cyanobacteria

Gene components of Gloeobacter were compared with those in the chromosomes of three cyanobacteria, Thermosynechococcus (2475 genes),[7] Synechocystis (3264 genes),[5] and Anabaena (5368 genes),[6] and of Gram-negative and a Gram-positive bacteria, *Escherichia coli* (4279 genes)[19] and *Streptomyces coelicolor* (7512 genes),[20] respectively, as a control under the criteria described in Materials and Methods. Comparison with higher stringency indicated that

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

	Number of genes	%
Amino acid biosynthesis	106	2.4
Biosynthesis of cofactors, prosthetic groups, and carriers	141	3.2
Cell envelope	61	1.4
Cellular processes	86	1.9
Central intermediary metabolism	30	0.7
Energy metabolism	102	2.3
Fatty acid, phospholipid and sterol metabolism	53	1.2
Photosynthesis and respiration	158	3.6
Purines, pyrimidines, nucleosides, and nucleotides	46	1.0
Regulatory functions	195	4.4
DNA replication, recombination, and repair	68	1.5
Transcription	44	1.0
Translation	193	4.4
Transport and binding proteins	226	5.1
Other categories	327	7.4
Subtotal of genes similar to genes of known function	**1836**	**41.5**
Similar to hypothetical protein	1635	36.9
Subtotal of genes similar to registered genes	**3471**	**78.4**
No similarity	959	21.6
Total	**4430**	**100.0**

1301 Gloeobacter genes comprising 29% of the 4324 potential protein-encoding genes had matched genes in the three cyanobacterial genomes. After subtraction of genes showing sequence similarity to those of either of *E. coli* or *S. coelicolor*, 610 Gloeobacter genes could be listed as those unique to four cyanobacteria. Approximately half of them were genes of known function, which include those related to antenna components, chlorophyll synthesis, photosystems and carbon fixation.

On the other hand, even with lower stringency where most members of a given gene family can be clustered, 995 Gloeobacter genes (23%) showed no sequence similarity to genes in the five bacterial genomes used for comparison. Moreover, 684 out of 995 genes were not similar to any of the registered genes, indicating that these genes are likely to be unique to Gloeobacter.

3.4.2. Genes related to photosynthesis

Genes related to photosynthesis and their copy number in the genome of Gloeobacter as well as those of the other three cyanobacteria are listed in Table 3 in the Supplement section. There are plenty of unique characteristics in the photosynthesis of Gloeobacter as described in Introduction, and distinctive features of the genes for photosynthesis in Gloeobacter revealed in this study are as follows.

There was a set of genes for PsaA (*glr3438*), PsaB (*glr3439*), PsaC (*gsl3287*), PsaD (*glr3701*), PsaE (*gsl3408*), PsaF (*glr2732*), PsaL (*glr2236*), and PsaM (*gsl2401*), but those for PsaI, PsaJ, PsaK, and PsaX were missing. PsaF is a trans-membrane protein and its N-terminal region is believed to interact with soluble electron transfer catalysts such as plastcyanin (PetE) and cytochrome c_{553} (PetJ).[21] The presumptive product of *psaF* (*glr2732*) in Gloeobacter (181 amino acid residues) was longer than those of other cyanobacteria (approximately 164 amino acid residues), which is attributed to the longer N-terminal region of the less conserved sequence.

Genes for PsbA (*glr0779*, *glr1706*, *glr2322*, *glr2656*, *gll3144*), PsbB (*glr2999*), PsbC (*glr2324*), PsbD (*glr2323*), PsbE (*gsr0856*), PsbF (*gsr0857*), PsbH (*gsr3002*), PsbI (*gsl3634*), PsbJ (*gsr0859*), PsbK (*gsr2807*), PsbL (*gsr0858*), PsbM (*gsl2997*), PsbN (*gsl3001*), PsbO (*glr3691*), PsbT(*gsr3000*), PsbU (*gll2873*), PsbV (*gll2337*, *gll2338*), PsbX (*gsr1874*), and Psb28 (*glr1041*, *gsl0928*) were assigned to the genome, but those for PsbY, PsbZ, and Psb27 seemed to be missing. PsbO, PsbU, and PsbV, components of oxygen-evolving complex, are located on the surface of thylakoid membranes on the lumen side,[22] and the amino acid sequence of each protein is well conserved (45% to 62%) among Thermosynechococcus, Synechocystis, and Anabaena. The presumptive products of the corresponding genes in Gloeobacter, however, showed much lower sequence identity of less than 31%, correlating with the speculation that these proteins are exposed to periplasm in Gloeobacter instead of lumen, as is the case in other cyanobacteria.

With respect to the genes for phycobilisomes, *cpcG* for a rod core linker peptide and *nblA* related to degradation of phycobilisomes were absent in the Gloeobacter genome. In addition to *apcE* (*glr1245*) encoding a phycobilisome core-membrane linker polypeptide, two genes showing sequence similarity to *apcE*, *glr2806*, and *glr1262*, were found in the genome. These genes exhibited a novel structure composed of three tandem repeats having a coding capacity of N-terminal 190 amino acid residues of CpcC, a rod linker component, which is also similar to phycoerythrin-associated linker proteins, CpeC, CpeD and CpeE. *glr1262* was situated in a cluster of genes related to phycoerythrin, *pebA–pebB–glr1262–cpeC–cpeD–cpeE* (*glr1260–glr1265* at the coordinates of 1,335,838–1,344,638). There was another gene cluster related to phycoerythrin, *cpeR–cpeZ–cpeY–cpeA–cpeB–ycf58–cpeS–cpeT* (*glr1186–glr1193*) at the coordinates of 1,264,757–1,275,461 in the Gloeobacter genome.

There are two types of protochlorophyllide reductase in cyanobacteria, a light-independent enzyme inherited from photosynthetic bacteria and a light-dependent one newly acquired during its evolution. Gloeobacter

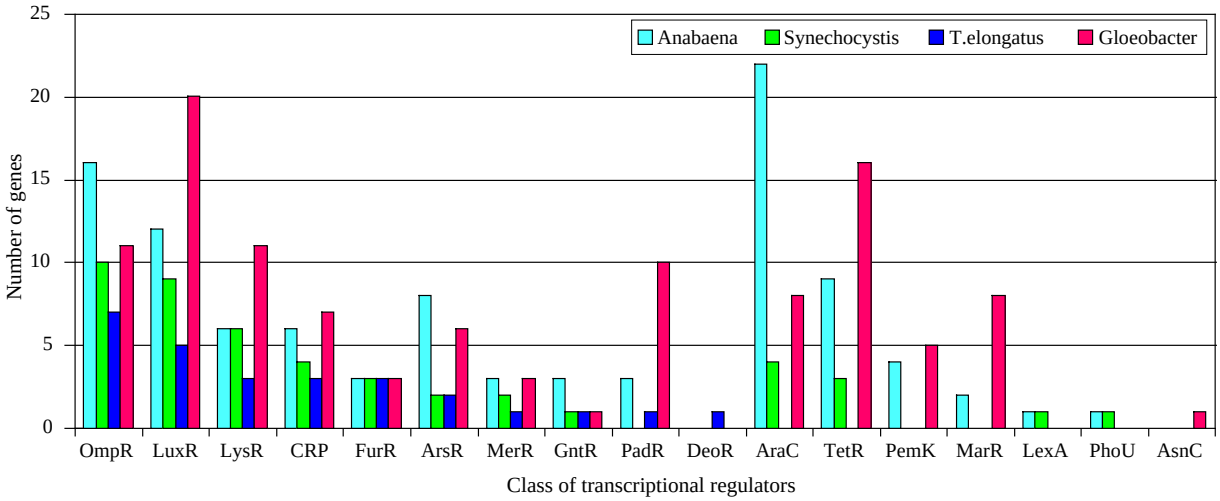

Figure 2. Genes encoding motifs of transcriptional factors in cyanobacteria. The number of the genes that are capable of coding for proteins that also contain motifs of transcriptional factors shown in the horizontal axis are plotted in the vertical axis. The color codes indicate the species of cyanobacteria as shown in the figure.

had genes for both light-independent (*chlBLN*: *glr0215*, *gll2370*, and *gll2369*, respectively) and light-dependent (*por*: *glr2486*) enzymes. This observation indicates that cyanobacteria has acquired the light-dependent enzyme in the very early phase of its evolution because Gloeobacter is believed to have diverged the earliest within the radiation of cyanobacteria.

Synechocystis have two similar genes, *pspA* for phage shock protein A and *Vipp1*.[23] *Vipp1* is known to be essential for formation of thylakoid membranes in Synechocystis and in a higher plant, *Arabidopsis thaliana*.[24] The C-terminal region of Vipp1, approximately 30 amino acid residues, is well conserved and is believed to encode a domain related to the above function, and PspA lacks this region. *glr0898* in Gloeobacter showed higher sequence similarity to that of *Vipp1* than *pspA*, but a region corresponding to the C-terminal conserved region was missing. *glr0898* may reflect the intermediate state of a gene during the evolutional process between *Vipp1* and *pspA*.

Genes for the cytochrome b_6f complex, *petA* (*glr3039*), *petB* (*gll1919*), *petC* (*glr3038*), *petD* (*gll1918*), *petG* (*gsl0511*), and *petN* (*gsl3700*), were assigned to the genome. There was another set of *petB* (*gll1870*) and *petD* (*gll1869*) with lower sequence similarity at the coordinates of 1,990,482–1,991,508. *petM* and *petL* could not be detected in this study, but this may be due to a low degree of conservation for small polypeptides. N-termini of the presumptive products of *petJ* and *petE* for soluble electron transfer catalysts were less conserved than the remaining portions, probably reflecting the difference of targeting sites, the cytoplasmic membrane in Gloeobacter and thylakoids in other cyanobacteria.

It has been reported that the Gloeobacter cells do not contain SQDG, which is one of the major fatty acid components of thylakoid membranes.[4] In accordance with this, neither *sqdB* nor *sqdX* required for biosynthesis of SQDG[25,26] was found in the Gloeobacter genome. SQDG plays a significant role in accommodation of photosynthetic apparatus on thylakoid membranes together with phosphatidylglycerol. Gloeobacter, in the absence of both thylakoids and SQDG, should have an alternative mechanism to secure photosynthesis on the cytophasmic membrane.

3.4.3. Genes related to signal transduction

The translated amino acid sequences of all the assigned genes in the Gloeobacter genome as well as three cyanobacterial genomes were subjected to search against the Pfam database to find DNA-binding motifs specific to transcription factors. As a result, 110 out of 4430 presumptive gene products in Gloeobacter were assigned as putative transcription factors, and were classified into 14 families, as shown in Fig. 2. It is notable that Gloeobacter contained a relatively large number of transcription factors in the LuxR, LysR, PadR, TetR, and MarR families, and that 14 out of 20 members of the LuxR family were hybrids with response regulators in the two-component signal transduction system.

A total of 76 genes were assigned to be members of the two-component system, including 27 genes for sensor histidine kinases, 12 for hybrids of sensor histidine kinases and response regulators, and 37 for response regulators. Twenty-seven out of 37 genes for response regulators contained motifs of transcription factors.

There were 15 genes for serine/threonine protein kinases, but none of them were similar to those of other three cyanobacteria except for portions of the protein kinase domains. The presumptive products of *gll0585*,

glr1096, and *glr4072* contained 5 to 7 repeats of a TPR domain known to be related to protein-protein interactions. Among 7 genes for protein phosphatase, *gll0589* showed a low level of sequence similarity to *pppA* for serine/threonine protein phosphatase. The presumptive products of *gll2757*, *gll1405*, and *gll0158* contained a conserved domain of protein phosphatase 2C.

Gloeobacter had a larger number of genes (14 genes) for sigma factors than Thermosynechococcus (7 genes), Synechocystis (9 genes) and Anabaena (12 genes). These include one gene (*glr2572*) for SigA, 5 genes (*gll1334*, *gll0203*, *gll3008*, *gll3762*, *gll4359*) for group 2 sigma, 6 genes (*gll1739*, *gll2708*, *glr1158*, *glr1764*, *glr3584*, *glr4357*) for ECF (extracytoplasmic function)-type sigma and 2 genes (*gll0669*, *glr2668*) for group 3 sigma. *gll0669* was similar to genes for ECF-type sigma factors in *Streptomyces* and alpha-proteobacteria, and genes surrounding *gll0669* showed sequence similarity to those of proteobacteria, suggesting an exogenous origin of this region.

3.4.4. Genes related to circadian rhythm

Circadian rhythms have intensively been studied in cyanobacteria, and genes involved in various processes of circadian timing and regulation have been identified in many species of cyanobacteria.[27] These genes include *kaiABC* as the major genetic elements of the circadian clock, *sasA*, *cikA*, *ldpA*, and *pex* as input modifiers, and *rpoD2* and *cpmA* as output modifiers. Even after intensive search, we could not detect *kaiABC* in the Gloeobacter genome. It is therefore likely that Gloeobacter does not have a genetic controlling system for circadian rhythms and that cyanobacteria have acquired this system after divergence of the Gloeobacter lineage. Alternatively, Gloeobacter might have lost such genes. With respect to genes for input modifiers, only *gll3769* could be identified as *ldpA*, and none of the remaining genes showed significant sequence similarity to *sasA*, *cikA*, and *pex*. On the contrary, two genes, *gll3762* and *gll1462*, were assigned as *rpoD2* and *cpmA*, respectively, as output modifiers.

3.4.5. Genes for WD repeat proteins

A total of 14 genes for proteins containing WD-repeats were identified in the Gloeobacter genome. The presumptive products of 13 genes contained 6 to 15 WD-repeats at the C-terminal portions with unique N-terminal regions of 260–1090 amino acid residues, while that of *glr0535* was composed of 8 WD-repeats for the entire length. Eleven genes encoding 14 or 15 WD-repeats showed sequence similarity to WD-repeat protein genes in Anabaena and could be classified into two groups. *gll0725*, *gll3661*, *glr2244*, and *glr1630* were capable of coding for polypeptides with unique N-terminal portions of 1050–1090 amino acid residues and were similar to *all0283*, *all0284*, and *all2124* in Anabaena. *gll2655*, *gll2888*, *gll4351*, *gll4356*, *glr1175*, *glr1965*, and *glr2821* formed another group encoding unique N-terminal regions of 469–581 amino acid residues showing sequence similarity to *alr7129*, *alr2800*, and *alr0029*. These genes seem to be shared only by Gloeobacter and Anbaena, but their biological role remains to be studied.

A Hat protein containing 11 WD-repeats at the C-terminus is located in thylakoid membranes, and is known to be involved in high-affinity transport of inorganic carbon.[28] Genes with conserved N-terminal unique regions of approximately 500 amino acid residues of the Hat protein are found in Thermosynechococcus, Synechocystis and Anabaena but not in Gloeobacter, suggesting the presence of an alternative system for the transport of inorganic carbon in the cytoplasmic membrane.

3.4.6. Insertion sequences

Seventy-four genes for putative transposases were assigned to the Gloeobacter genome, of which 52 were identified as components of insertion sequences (ISs). A total of 47 copies of ISs, 32 of which are likely to retain intact structure, were classified into 14 groups mainly on the basis of the type of the transposases and the length of inverted repeats. Structural features of each IS group are summarized in Tables 4 and 5 and in Fig. 3 in the Supplement section.

Traces of large-scale genome rearrangements by an IS-mediated homologous recombination could be detected. ISGL5 generates 8-bp direct repeats on both sides of the IS during insertion, as shown in Table 5 in the Supplement section. However, the 8-bp sequence on one side of ISGL5b, GATTTACC, was found not on the other side of this IS but on the proximal side of ISGL5d located 827 kb apart in the opposite direction. This observation strongly suggests that inversion of the 827-kb segment of the genome took place by homologous recombination between two ISGL5 elements. Evidence of another genome rearrangement could be found in the structure of ISGL2d and ISGL2e. ISGL2d (179 bp) and ISGL2e (260 bp) are located 432 kb apart in the opposite direction, but are likely to be 5′ and 3′ portions, respectively, of a single ISGL2 (963 bp) because the same 6-bp sequence, CCTTAG, was present next to the inverted repeat sequence of each IS. Together with the fact that ISGL12b, a segment of ISGL12, was found adjacent to ISGL2d, it is probable that insertion of ISGL12 into ISGL2 followed by recombination with another ISGL12 in the genome occurred during the evolution of this species.

We could detect very recent insertion of an IS during the course of genome sequencing. After assembly of the random sequences generated from multiple Gloeobacter cells, two sequences for a single gene encoding an HlyD family secretion protein (*gll2454*), one with and the other

without ISGL4, were obtained. The inserted IS had an identical 1160-bp sequence with those of ISGL4c, ISGL4d and ISGL4g, and a 10-bp repeat, GGTAAAAGC, was observed on both sides of the IS. These observations strongly suggest that insertion of ISGL4 occurred during the propagation of the Gloeobacter cells.

3.4.7. Inteins

Genes encoding proteins with inteins have been reported in Synechocystis (*dnaE*, *dnaB*, *gyrB*, and *dnaX*), Anabaena (*dnaE* and *dnaB*) and Thermosynechococcus (*dnaE*).[29] Gloeobacter had two genes encoding an intein, *gll0034* (*dnaB*) and *gll3966* for ribonucleotide reductase subunit alpha. No intein was identified in DnaE (Glr3934), GyrB (Gll2506), or DnaX (Glr1609).

The inteins in DnaB in Synechocystis and Anabaena share similar characteristics: They are 428 and 429 amino acid residues in size, and located at the coordinates of 390–817 and 381–809 amino acid residues from the N-termini, respectively. The DnaB intein in Gloeobacter, on the other hand, was 258 amino acid residues long and was situated 222–479 amino acid residues from the N-terminus, suggesting that it has a different origin than the other two cyanobacteria.

A gene encoding ribonucleotide reductase subunit alpha had not been identified in cyanobacteria, and *gll3966* was the first example. The presumptive product of *gll3966* had two inteins at the coordinates of 194–606 and 756–1122 amino acid residue from the N-terminus. The first intein showed sequence similarity to an intein Pab RIR1-2 in ribonucleoside diphosphate reductase in *Pyrococcus abyssi*, and the second intein was similar to an intein Dra-RF78101 RIR1 found in ribonucleoside diphosphate reductase alpha subunit in *Deinococcus radiodurans* R1, ATCC13939/RF78101.

3.4.8. Group II intron

Group II introns are the self-splicing ribozyme. Some of the group II introns are known to transpose with the aid of self-encoded maturase.[30,31] One copy of the group II intron (GlvI1) could be identified at the coordinates of 168,850–171,364 in the Gloeobacter genome (Fig. 1 in the Supplement section). Domain IV of GlvI1 had a coding capacity of a maturase (*gll0177*), which was similar to maturase genes in Cal.x1 of *Calothrix* and TelI4e (*tlr1161*), TelI4c (*tlr0522*) and TelI4b (*tlr0308*) of Thermosynechococcus. GlvI1 was located in the intergenic region between *glr0176* and *glr0178*, therefore, it is not known whether this intron is functional or not.

The sequences as well as the gene information shown in this paper are available in the Web database, CyanoBase, at http://www.kazusa.or.jp/cyanobase/. The sequence data analyzed in this study have been registered in DDBJ/GenBank/EMBL (the accession number: BA000045).

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

References

1. Rippka, R., Waterbury, J., and Cohen-Bazire, G. 1974, A cyanobacterium which lacks thylakoids, *Arch. Microbiol.*, **100**, 419–436.
2. Honda, D., Yokota, A., and Sugiyama, J. 1999, Detection of seven major evolutionary lineages in cyanobacteria based on the 16S rRNA gene sequence analysis with new sequences of five marine *Synechococcus* strains, *J. Mol. Evol.*, **48**, 723–739.
3. Guglielmi, G., Cohen-Bazire, G., and Bryant, D. A. 1981, The structure of *Gloeobacter violaceus* and its phycobilisomes, *Arch. Microbiol.*, **129**, 181–189.
4. Selstam, E. and Campbell, D. 1996, Membrane lipid composition of the unusual cyanobacterium *Gloeobacter violaceus* sp. PCC7421, which lacks sulfoquinovosyl diacylglycerol, *Arch. Microbiol.*, **166**, 132–135.
5. Kaneko, T., Sato, S., Kotani, H. et al. 1996, Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions, *DNA Res.*, **3**, 109–136.
6. Kaneko, T., Nakamura, Y., Wolk, C. P. et al. 2001, Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium *Anabaena* sp. strain PCC7120, *DNA Res.*, **8**, 205–213.
7. Nakamura, Y., Kaneko, T., Sato, S. et al. 2002, Complete genome structure of the thermophilic cyanobacterium *Thermosynechococcus elongatus* BP-1, *DNA Res.*, **9**, 123–130.
8. 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.
9. 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.
10. 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.
11. Bateman, A., Birney, E., Cerruti, L. et al. 2002, The Pfam protein families database, *Nucleic Acids Res.*, **30**, 276–280.
12. 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.
13. Lobry, J. R. 1996, Asymmetric substitution patterns in the two DNA strands of bacteria, *Mol. Biol. Evol.*, **13**, 660–665.
14. Gupta, A., Morby, A. P., Turner, J. S., Whitton, B. A., and Robinson, N. J. 1993, Deletion within the metallothionein locus of cadmium-tolerant *Synechococcus*

PCC6301 involving a highly iterated palindrome (HIP1), *Mol. Microbiol.*, **7**, 189–195.
15. Paquin, B., Kathe, S. D., Nierzwicki-Bauer, S. A., and Shub, D. A. 1997, Origin and evolution of group I introns in cyanobacterial tRNA genes, *J. Bacteriol.*, **179**, 6798–6806.
16. Williams, K. P. 2002, Descent of a split RNA, *Nucleic Acids Res.*, **30**, 2025–2030.
17. Gaudin, C., Zhou, X., Williams, K. P., and Felden, B. 2002, Two-piece tmRNA in cyanobacteria and its structural analysis, *Nucleic Acids Res.*, **30**, 2018–2024.
18. Riley, M. 1993, Functions of the gene products of *Escherichia coli*, *Microbiol. Rev.*, **57**, 862–952.
19. Blattner, F. R., Plunkett, G., 3rd, Bloch, C. A. et al. 1997, The complete genome sequence of *Escherichia coli* K-12, *Science*, **277**, 1453–1474.
20. Bentley, S. D., Chater, K. F., Cerdeno-Tarraga, A. M. et al. 2002, Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2), *Nature*, **417**, 141–147.
21. Wynn, R. M., Omaha, J., and Malkin, R. 1989, Structural and functional properties of the cyanobacterial photosystem I complex, *Biochemistry*, **28**, 5554–5560.
22. Shen, J. R., Qian, M., Inoue, Y., and Burnap, R. L. 1998, Functional characterization of *Synechocystis* sp. PCC6803 delta *psbU* and delta *psbV* mutants reveals important roles of cytochrome c-550 in cyanobacterial oxygen evolution, *Biochemistry*, **37**, 1551–1558.
23. Westphal, S., Heins, L., Soll, J., and Vothknecht, U. C. 2001, Vipp1 deletion mutant of *Synechocystis*: a connection between bacterial phage shock and thylakoid biogenesis?, *Proc. Natl. Acad. Sci. USA*, **98**, 4243–4248.
24. Kroll, D., Meierhoff, K., Bechtold, N. et al. 2001, VIPP1, a nuclear gene of *Arabidopsis thaliana* essential for thylakoid membrane formation, *Proc. Natl. Acad. Sci. USA*, **98**, 4238–4242.
25. Guler, S., Seeliger, A., Hartel, H., Renger, G., and Benning, C. 1996, A null mutant of *Synechococcus* sp. PCC7942 deficient in the sulfolipid sulfoquinovosyl diacylglycerol, *J. Biol. Chem.*, **271**, 7501–7507.
26. Guler, S., Essigmann, B., and Benning, C. 2000, A cyanobacterial gene, *sqdX*, required for biosynthesis of the sulfolipid sulfoquinovosyldiacylglycerol, *J. Bacteriol.*, **182**, 543–545.
27. Iwasaki, H. and Kondo, T. 2000, The current state and problems of circadian clock studies in cyanobacteria, *Plant Cell Physiol.*, **41**, 1013–1020.
28. Hisbergues, M., Gaitatzes, C. G., Joset, F., Bedu, S., and Smith, T. F. 2001, A noncanonical WD-repeat protein from the cyanobacterium *Synechocystis* PCC6803: structural and functional study, *Protein Sci.*, **10**, 293–300.
29. Perler, F. B. 2002, InBase: the Intein Database, *Nucleic Acids Res.*, **30**, 383–384.
30. Martinez-Abarca, F. and Toro, N. 2000, Group II introns in the bacterial world, *Mol. Microbiol.*, **38**, 917–926.
31. Dai, L. and Zimmerly, S. 2002, Compilation and analysis of group II intron insertions in bacterial genomes: evidence for retroelement behavior, *Nucleic Acids Res.*, **30**, 1091–1102.