Structural Analysis of Four Large Plasmids Harboring in a Unicellular Cyanobacterium, *Synechocystis* sp. PCC 6803

Takakazu Kaneko, Yasukazu Nakamura, Shigemi Sasamoto, Akiko Watanabe, Mitsuyo Kohara, Midori Matsumoto, Sayaka Shimpo, Manabu Yamada, and Satoshi Tabata*

Kazusa DNA Research Institute, 2-6-7 Kazusa-kamatari, Kisarazu, Chiba 292-0818, Japan

(Received 4 October 2003)

Abstract

The genome of the unicellular cyanobacterium *Synechocystis* sp. PCC 6803 consists of a single chromosome and several plasmids of different sizes, and the nucleotide sequences of the chromosome and three small plasmids (5.2 kb, 2.4 kb, and 2.3 kb) have already been sequenced. We newly determined the nucleotide sequences of four large plasmids, which have been identified in our laboratory (pSYSM: 120 kb, pSYSX: 106 kb, pSYSA: 103 kb, and pSYSG: 44 kb). Computer-aided analysis was performed to explore the genetic information carried by these plasmids. A total of 397 potential protein-encoding genes were predicted, but little information was obtained about the functional relationship of plasmids to host cell, as a large portion of the predicted genes (77%) were of unknown function. The occurrence of the potential genes on plasmids was divergent, and *parA* was the only gene common to all four large plasmids. The distribution data of a Cyanobacterium-specific sequence (HIP1: 5′-GCGATCGC-3′) suggested that respective plasmids could have originated from different cyanobacterial strains.

Key words: cyanobacterium; *Synechocystis* sp. PCC 6803; plasmid; genome sequencing

1. DNA Sequencing

To identify clones derived from the large plasmids for sequencing, a fraction containing plasmid DNAs was prepared from the PCC6803 cells, and subsequently resolved by gel electrophoresis. The pattern shown in Fig. 1 shows four plasmids: pSYSM (125 kb), pSYSA (110 kb), and pSYSG (45 kb), and pSYSX (106 kb) newly identified. DNA was extracted from each band and used as a probe to screen the cosmids library which had been constructed for sequencing of the PCC6803 chromosome. The selected clones
were aligned on each plasmid by a combination of hybridization and PCR using primers designed on the basis of end sequences of cosmids. The relative positions of cosmids are shown in Fig. 2. A region at coordinates 11 kb to 56 kb of pSYSX was covered by two BAC clones, pSYSX57 and pSYSX65, derived from purified pSYSX DNA. The remaining gaps for pSYSM and pSYSA were filled by PCR.

The nucleotide sequences were determined according to the shotgun method as described previously.\textsuperscript{8} Glimmer predicted a total of 550 potential protein-encoding regions in the four plasmids after training with a dataset of 102 sequences of highly probable protein-encoding genes in the plasmids. Then, sequence similarity to known genes and the relative positions were taken into account to avoid overlaps. The total number of potential protein-encoding genes starting with either an ATG, GTG, or TTG codon finally assigned to the four plasmids was 397 (Table 1). They were denoted by a serial number starting with 5 for pSYSM, 6 for pSYSX, 7 for pSYSA, and 8 for pSYSX with three letters representing the species name (s), whether the open reading frame (ORF) was longer than or shorter than 100 codons (l or s), and the transcription direction on the circular map (r or l) (Fig. 3). The average gene density varied among plasmids (one gene in every 905 bp for pSYSG and every 975 bp for pSYSA), and was higher than that of the chromosome (1 gene in every 1095 bp). We could not find any genes for tRNAs and other structural RNAs based on the similarity search and computer prediction.

Function assignment of 397 potential protein-encoding genes was performed by similarity search against the nr- and Pfam databases. This search produced only 92 genes (23%) showing significant sequence similarity to genes of known function (Fig. 2 and Table 1), 142 genes (36%) to the hypothetical genes, and the remaining 163 genes (41%) were not similar to any registered genes. The potential protein-encoding genes whose function could be anticipated were classified into 8 out of 14 categories of different biological roles according to the principle of Riley.\textsuperscript{17} The number of genes in each category and their location on the maps as well as those of the chromosome are summarized in Table 1 and Figs. 2 and 3. The name of each gene is listed in CyanoBase at http://www.kazusa.or.jp/cyanobase/.

3. Remarkable Features of DNA Sequences and Predicted Genes

The sequence features of the four plasmids sequenced in this study, as well as those of three small plasmids previously reported, are listed in Table 2. The GC contents of the plasmids were lower than that of the chromosome, except for pSYSG, and showed uneven distribution along the molecules for pSYSM, pSYSX, and pSYSA, as shown in Fig. 2. This is attributable to the presence of insertion sequences (ISs), ISY\textsuperscript{203} and ISY\textsuperscript{391} (see the following section) and a homologue (sll5014) of the intron-encoded endonuclease in pSYSM. An 8-bp palindromic sequence HIP1 (5′-GCGATCGC-3′) which is commonly present in many cyanobacterial strains is found in the PCC6803 chromosome at the frequency of 1 copy in every 1131 bp. HIP1 was also present in
Figure 2. Circular representation of four large plasmids of PCC6803. The scale indicates the location in kb starting from the restriction site described in the text. The boxes in the outermost and the second circles show the positions of the potential protein-encoding genes in the clockwise and counterclockwise directions, respectively. The potential protein-encoding genes whose function could be evaluated by similarity search were classified into 8 functional categories, and are represented by different color codes. The plots in the third circle with a scale indicate the average GC percent calculated using a window-size of 1000-bp. The innermost lines show the regions covered by cosmid clones (blue), BAC clones (black) and PCR products (green).
Table 1. Features of the assigned protein-coding genes and the functional classification.

|                                 | Chromosome* | pSYSM | pSYSX | pSYSA | pSYSG |
|---------------------------------|-------------|-------|-------|-------|-------|
|                                 | %           | %     | %     | %     | %     |
| Amino acid biosynthesis         | 97          | 3.0   | 0.0   | 0.0   | 0.0   |
| Biosynthesis of cofactors, prosthetic groups, and carriers | 124         | 3.8   | 1.8   | 0.0   | 0.0   |
| Cell envelope                   | 67          | 2.1   | 0.0   | 0.0   | 0.0   |
| Cellular processes              | 78          | 2.4   | 1.8   | 0.9   | 0.0   |
| Central intermediary metabolism | 31          | 0.9   | 0.0   | 0.0   | 0.0   |
| Energy metabolism               | 93          | 2.8   | 0.0   | 0.0   | 0.0   |
| Fatty acid, phospholipid and sterol metabolism | 39         | 1.2   | 0.0   | 0.0   | 0.0   |
| Photosynthesis and respiration  | 141         | 4.3   | 0.0   | 0.0   | 0.0   |
| Purines, pyrimidines, nucleosides, and nucleotides | 41         | 1.3   | 1.8   | 0.9   | 0.0   |
| Regulatory functions            | 146         | 4.5   | 3.0   | 4.5   | 0.0   |
| DNA replication, recombination, and repair | 60         | 1.8   | 2.1   | 6.5   | 2.9   |
| Transcription                   | 30          | 0.9   | 0.0   | 0.0   | 0.0   |
| Translation                     | 168         | 5.1   | 0.0   | 0.0   | 0.0   |
| Transport and binding proteins  | 196         | 6.0   | 2.1   | 2.1   | 0.0   |
| Other categories                | 312         | 9.6   | 25.0  | 18.9  | 0.0   |
| Subtotal of genes similar to genes of known function | 1623       | 49.7  | 36.2  | 37.3  | 32.0  |
| Similar hypothetical protein    | 1133        | 34.7  | 51.9  | 38.6  | 44.1  |
| Subtotal of genes similar to registered genes | 2756       | 84.4  | 65.9  | 51.4  | 54.6  |
| No similarity                   | 508         | 15.6  | 45.1  | 60.4  | 54.5  |
| Total                           | 3264        | 100.0 | 100.0 | 100.0 | 100.0 |

* The assignment of potential protein-encoding regions has been revised in 2002 (CyanoBase: http://www.kazusa.or.jp/cyanobase/Synechocystis/). Their translated sequences were worked over functional assignments and classifications by using the result subjected to similarity search against the nr-database and the examples from CyanoGenes (http://www.kazusa.or.jp/cyanobase/Synechocystis/comments/) and CYORF (http://cyano.genome.ad.jp/).

Table 2. General feature of replicons in *Synechocystis* sp. PCC 6803.

|                                  | Length (bp) | Average GC contents (%) | Average frequency of HIP1 appearance (bp) | References |
|----------------------------------|-------------|-------------------------|------------------------------------------|------------|
| Chromosome                       | 3,573,470   | 47.7                    | 1131                                     | 1)         |
| pSYSM                            | 119,895     | 43.0                    | 1518                                     | this report|
| pSYSX                            | 106,004     | 42.7                    | 1738                                     | this report|
| pSYSA                            | 103,307     | 44.5                    | 1878                                     | this report|
| pSYSG                            | 44,343      | 48.6                    | 2016                                     | this report|
| pCC5.2                           | 5,214       | 46.9                    | 579                                      | 7)         |
| pCA2.4                           | 2,378       | 42.1                    | ND                                       | 5)         |
| pCB2.4                           | 2,345       | 42.7                    | ND                                       | 6)         |

ND; not detected

plasmids, but occurred at significantly different frequencies among plasmids (Table 2). Such sequence features suggest that respective plasmids could have originated from different cyanobacterial strains.

Nine groups of Insertion Sequences (ISs), each comprising 1 to 22 members, have been reported for the PCC6803 chromosome (CyanoBase at http://www.kazusa.or.jp/cyanobase/). In PCC6803 plasmids, 15 genes for putative transposases were identified and assigned to nine ISs, which are classified into four of the nine IS groups originally found in the chromosome. Structural features and the position of each IS along with those of 80 copies of ISs in the chromosome are summarized in Table 3.16,20–22 Okamoto et al. have reported evidence of the transposition of ISY203x (Accession no. AB030081) located in pSYSX to three different positions of the chromosome by comparing the gene structures among substrains of PCC6803.22

A direct repeat of a completely identical sequence 23,639 bp in length was found at coordinates 4,092–27,731 bp and 58,810–82,449 bp in pSYSX. The average GC content of these DNA regions was lower than that of the entire pSYSX (41.2% vs. 42.7%; Fig. 2), suggesting an exogenous origin for these regions.

We compared the gene components among the seven plasmids in PCC6803 with two different stringencies described previously.23 Among the 409 potential genes (397 in four large plasmids and 12 in three small plasmids), 92 genes having homologues in different plasmids were identified with standard stringency and 122 were identified with low stringency. However, their occurrence among plasmids was diverse, and only the parA gene was common to all four large plasmids even with low stringency. A total of 48 plasmid genes showed sequence sim-
Figure 3. The gene map of four plasmids, pSYSM, pSYSX, pSYSA, and pSYSG, in PCC6803. The circular plasmids, pSYSM (119,895 bp), pSYSX (106,004 bp), pSYSA (103,307 bp), and pSYSG (44,343 bp) were opened at Mlu I, Sal I, Asc I, and Sau3AI recognition sites, respectively, and are represented by a linear map starting from this junction. Green bars show the scale in 3 kb with numerals in kb. On both sides of the scale, the potential protein-coding genes assigned on the basis of computer prediction and similarity search are shown by boxes with arrowheads indicating the reading direction. The potential genes whose function could be evaluated by similarity search were classified into 8 functional categories, and are indicated by different color codes.
**Figure 3.** Continued.

Table 3. Structural features of insertion sequences in the PCC6803 plasmids.

| ISY100 (ISS1987)* | ISY508 | ISY120 | ISY203 (ISAS)* | ISY352 | ISY391 | ISY523 (ISS5)* | ISY532 |
|------------------|--------|--------|----------------|--------|--------|----------------|--------|
| size (bp)        | 947    | 968    | 802            | 1174   | 1410   | 1379           | 871    |
| inverted repeat (bp) | 24     | 24     | 22             | 36     | 26     | 35             | 17     |
| direct repeat     | TA     | TA     | 9 bp           | IS4    | TA     | TA             | 3 bp   |
| family            | IS630/TC1 | IS630/TC1 | IS4            | IS4    | IS4    | IS4            | IS4    |
| number of full-length copies ** | 3,026(20) | 0(2) | 0(3) | 3,026(9) | 0(5) | 0(2) | 1,400(12) |
| number of partial copies ** | 0(2) | 0(1) | 0(8) | 0(0) | 0(3) | 2,026(1) | 0(1) |

* Reference 20

** A total number of IS in the plasmids are shown. The alphabetical characters indicate plasmids in which ISs are found: A:pSYSA, G:pSYSG, M:pSYSM, X:pSYSX. Numerals in parentheses show ISs that are identified in the PCC6803 chromosome.

Flanking by gene(s) for cation transporters slr0798 in the chromosome and slr6042/slr6043 in pSYSX. These findings suggest that both gene sets have a similar function for transduction of cation signals.

The WD-repeat is a protein motif prevalent in the genomes of cyanobacteria, and five genes for WD-repeat proteins have been reported in the PCC6803 chromosome. In this study, slr8038 in pSYSG was identified as a gene presumptively coding for a WD-repeat protein. The putative product of slr8038 contained 14 repeat units at the C-terminal portion. The N-terminal region of 600 amino acid residues long was unique in PCC6803, but seemed to share the sequence with the putative WD-repeat proteins in Anabaena sp. PCC 7120 (Alr7129, Alr2800, and Alr0029) and Gloeobacter violaceus PCC 7421 (Gll2655, Gll2888, Gll4351, Gll4356, Gll1175, Gll1965, and Gll2821). The function of these proteins remains to be studied.

Other notable features of the DNA sequences and the predicted genes are as follows.

1. The sequences and the order of five to seven protein-encoding genes were significantly conserved between pSYSM and pSYSX in the regions at the approximate coordinates 95–102 kb and 31–44 kb, 84–91 kb and 58–66 kb, and 67–76 kb and 75–89 kb, respectively. This implies either an evolutionary relationship between these two plasmids or the occurrence of a segmental transfer between the two plasmids.

2. Genes for glycosyl transferase families (sll5043, sll5044, sll5048, sll5050, slr5054, slr5055, slr5056,
and slr5077) formed a cluster at coordinates 40–58 kb in pSYSM. Homologues of polysaccharide transporter (slr5049 and slr5052) were also found within this cluster, suggesting that the genes in this cluster are involved in the biosynthesis and transport of unknown exopolysaccharides.

3. ndhK2 (formerly psbG2, slr8031) in pSYSG, which presumptively codes for a subunit of NADH dehydrogenase, is a cryptic gene.24 It has been reported that mutation of ndhK (slr1280) in the chromosome was suppressed by ndhK2 activated by a DNA rearrangement,24 indicating that the two genes are functionally complementary. However, the biological role of ndhK2 remains to be elucidated.

4. Four genes, slr5035, slr5086, slr6056, and slr8026, presumably encoding transcriptional regulators, were assigned to pSYSM, pSYSX, and pSYSG, but a gene for sigma factors, which is found in a plasmid of *Anabaena* sp. *PCC 7120,*8 was not identified.

5. ParA and ParB are involved in partitioning of plasmids during cell duplication.25 Either one or two copies of a *parA* homologue were identified in each of four large plasmids but not in three small plasmids: slr5066 and slr5105 in pSYSM, slr6036 and slr6093 in pSYSX, slr7044 in pSYSA and slr8015 in pSYSG, although the degree of similarity varied between 23.1% and 99.5% at the amino acid level. *parB* was found only in pSYSG (slr8016).

In this study, we sequenced four large plasmids in *PCC6803* to investigate characteristic features common to plasmids, including the previously sequenced small plasmids, and to find clues to understand the functional role of the plasmids in the cell. As a consequence, we could not observe any obvious common features among plasmids except that a significant proportion of the presumptive protein-encoding genes were those of unknown function (61% to 81%). Further detailed analysis of the sequence data with the aid of computers may provide a clearer picture of plasmids in cyanobacteria. With respect to the gene components in plasmids, some genes, such as those for the two-component system, the cation transporters, the WD-repeat protein, the glycosyl transferase families, and *ndhK*, are likely to benefit the host cells. This should be experimentally examined by using the natural transformation property in this organism.

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 numbers are as follows: AP004310 for pSYSM, AP004311 for pSYSA, AP004312 for pSYSG, and AP006585 for pSYSX.

Acknowledgements: This work was supported by the Kazusa DNA Research Institute Foundation.

References

1. Kaneko, T., Sato, S., Kotani, H. et al. 1996, Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain *PCC 6803*. II. Sequence determination of the entire genome and assignment of potential protein-coding regions, *DNA Res.*, 3, 109–136.
2. Imsche, M. and Tabata, S. 2001, *Synechocystis* sp. *PCC 6803* — a useful tool in the study of the genetics of cyanobacteria, *Photosynthesis Res.*, 70, 73–83.
3. Houmard, J. and Tandeau de Marsac, N. 1988, In: Packer, L. and Glazer, A. N. (eds) Methods Enzymol., Vol. 167. Academic Press, San Diego, pp. 808–847.
4. Van der Plas, J., Oosterhoff-Teerstra, R., Borrias, M., and Weisbeek, P. 1992, Identification of replication and stability functions in the complete nucleotide sequence of plasmid pUH24 from the cyanobacterium *Synechococcus* sp. *PCC 7942*, *Mol. Microbiol.*, 6, 653–664.
5. Yang, X. and McFadden, B. A. 1993, A small plasmid, pCA2.4, from the cyanobacterium *Synechocystis* sp. strain *PCC 6803* encodes a *rep* protein and replicates by a rolling circle mechanism, *J. Bacteriol.*, 175, 3981–3991.
6. Yang, X. and McFadden, B. A. 1994. The complete DNA sequence and replication analysis of the plasmid pCB2.4 from the cyanobacterium *Synechocystis* sp. *PCC 6803*, *Plasmid*, 31, 131–137.
7. Xu, W. and McFadden, B. A. 1997. Sequence analysis of plasmid pCC5.2 from cyanobacterium *Synechocystis* sp. *PCC 6803* that replicates by a rolling circle mechanism, *Plasmid*, 37, 95–104.
8. Kaneko, T., Nakamura, Y., Wolk, C. P. et al. 2001, Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium *Anabaena* sp. strain *PCC 7120*, *DNA Res.*, 8, 205–213.
9. Matveyev, A. V., Young, K. T., Meng, A., and Elhai, J. 2001, DNA methyltransferases of the cyanobacterium *Anabaena* sp. *PCC 7120*, *Nucleic Acids Res.*, 29, 1491–1506.
10. Linden, H., Misawa, N., Saito, T., and Sandmann, G. 1994, A novel carotenoid biosynthesis gene coding for zeta-carotene desaturase: functional expression, sequence and phylogenetic origin, *Plant Mol. Biol.*, 24, 369–379.
11. Khudyakov, I. Y. and Golden, J. W. 2001, Identification and inactivation of three group 2 sigma factor genes in *Anabaena* sp. strain *PCC 7120*, *J. Bacteriol.*, 183, 6667–6675.
12. Castets, A.-M., Houmard, J., and Tandeau de Marsac, N. 1986, Is cell motility a plasmid-encoded function in the cyanobacterium *Synechocystis* 6803?, *FEBS Microbiol. Lett.*, 37, 277–281.
13. Kotani, H., Kaneko, T., Matsubayashi, T., Sato, S., Sugiiura, M., and Tabata, S. 1994, A physical map of the genome of a unicellular cyanobacterium *Synechocystis* sp. strain *PCC 6803*, *DNA Res.*, 1, 303–307.
14. Kotani, H., Tanaka, A., Kaneko, T., Sato, S., Sugiiura, M., and Tabata, S. 1995, Assignment of 82 known genes and gene clusters on the genome of the unicellular cyanobacterium *Synechocystis* sp. strain *PCC 6803*, *DNA Res.*, 2, 133–142.
15. Casse, F., Boucher, C., Julliot, J. S., Michel, M., and Denarie, J. 1979, Identification and characterization of large plasmid in *Rhizobium meliloti* using agarose elec-
16. Kaneko, T., Tanaka, A., Sato, S. et al. 1995, Sequence analysis of the genome of the unicellular cyanobacterium Synechocystis sp. strain PCC 6803. I. Sequence features in the 1 Mb region from map positions 64% to 92% of the genome, *DNA Res.*, **2**, 153–166.

17. Riley, M. 1993, Functions of the gene products of *Escherichia coli*, *Microbiol. Rev.*, **57**, 862–952.

18. Robinson, N. J., Robinson, P. J., Gupta, A., Bleasby, A. J., Whitton, B. A., and Morby, A. P. 1995, Singular over-representation of an octameric palindrome, HIP1, in DNA from many cyanobacteria, *Nucleic Acids Res.*, **23**, 729–735.

19. Kaneko, T. and Tabata, S. 1997, Complete genome structure of the unicellular cyanobacterium Synechocystis sp. PCC 6803, *Plant Cell Physiol.*, **38**, 1171–1176.

20. Cassier-Chauvat, C., Poncelet, M., and Chauvat, F. 1997, Three insertion sequences from the cyanobacterium Synechocystis PCC 6803 support the occurrence of horizontal DNA transfer among bacteria, *Gene*, **195**, 257–266.

21. Urasaki, A., Sekine, Y., and Ohtsubo, E. 2002, Transposition of cyanobacterium insertion element ISY100 in *Escherichia coli*, *J. Bacteriol.*, **184**, 5104–5112.

22. Okamoto, S., Ikeuchi, M., and Ohmori, M. 1999, Experimental analysis of recently transposed insertion sequences in the cyanobacterium Synechocystis sp. PCC 6803, *DNA Res.*, **6**, 265–273.

23. Nakaamura, Y., Kaneko, T., Sato, S. et al. 2003, Complete genome structure of Gloeobacter violaceus PCC 7421, a cyanobacterium that lacks thylakoids, *DNA Res.*, **10**, 137–145.

24. Steinmuller, K., Ellersiek, U., and Bogorad, L. 1991, Deletion of the *psbG1* gene of the cyanobacterium Synechocystis sp. PCC 6803 leads to the activation of the cryptic *psbG2* gene, *Mol. Gen. Genet.*, **226**, 107–112.

25. Motallebi-Veshareh, M., Rouch, D. A., and Thomas, C. M. 1990, A family of ATPases involved in active partitioning of diverse bacterial plasmids, *Mol. Microbiol.*, **4**, 1455–1463.