GENOME ANALYSIS OF CYANOBACTERIA

Satoshi Tabata
Kazusa DNA Research Institute, Chiba, Japan

Keywords: cyanobacteria, *Synechocystis* sp. PCC 6803, genome sequence, microarray, proteome, photosynthesis, two-component signal transduction system, plastid, gene disruption

Contents

Summary

Cyanobacteria are the only bacteria to perform oxygen-producing photosynthesis. It has generally been accepted that they are the progenitor(s) of plant plastids. In fact, cyanobacteria and plants share many similarities in both the machinery and mechanisms of photosynthesis. For this reason, cyanobacteria have long been model organisms for the study of oxygen-producing photosynthesis in higher plants. Among all phototrophic organisms, including the 1500 species of cyanobacteria, *Synechocystis* sp. PCC 6803 was the first one whose genome was fully sequenced. Genomic sequencing has revealed the structure of the genome, identified its genes, and mapped the relative location of each gene in the genome. The function of nearly half of the genes has been deduced, on the basis of sequence similarity to genes whose function is known. Genome sequencing has also allowed for implementation of systematic approaches to the study of gene function and the mechanisms of gene regulation on a genome-wide level, including systematic gene disruption, transcriptome analysis using microarray systems, and proteome analysis. Two genome databases, CyanoBase and CyanoMutants, have been established and act as a central repository for information on the gene structure and gene

function of *Synechocystis* sp. PCC 6803, respectively. As a result of genome sequencing and the establishment of these genome databases, *Synechocystis* sp. PCC 6803 is an extremely useful model for studying genetic systems of photosynthetic organisms. Work on genome sequencing is in progress to understand the genetic systems of other cyanobacterial species, including those capable of nitrogen fixation, marine inhabitants, and a thermophilic strain.

1. Introduction

Cyanobacteria, also called "blue-green algae," are one of the eleven major eubacterial phyla. Because of their varied physiological, morphological, and developmental characteristics, the 1500+ species of cyanobacteria constitute an extremely diverse group of prokaryotes. Although their phylogenetic position in the bacterial kingdom is still uncertain, cyanobacteria are believed to be genetically related to gram-positive bacteria. Cyanobacteria are capable of photosynthesis but are distinct from other photosynthetic bacteria, such as purple and green bacteria, in that they utilize H_2O as an electron donor and produce oxygen. Striking similarities in both the machinery and mechanism of photosynthesis between cyanobacteria and plants have made cyanobacteria useful model organisms for the study of oxygen-producing photosynthesis in higher plants. From an evolutionary viewpoint, it is generally accepted that ancestors of cyanobacteria acquired the ability to perform oxygen-producing photosynthesis 2.5 billion years ago and that progenitors of unicellular cyanobacteria gave rise to plant plastids through endosymbiosis.

Although cyanobacteria constitute one of the largest groups of gram-negative bacteria, only a few strains are amenable to genetic manipulation and suitable for use in physiological and genetic studies. These include the unicellular strains *Synechocystis* sp. PCC 6803, *Synechococcus* sp. PCC 7942, *Synechococcus* sp. PCC 6301, and the filamentous strain *Anabaena* sp. PCC7120. Recent progress in DNA sequencing technology has allowed for the production of large quantities of nucleotide sequence data in a short period of time. This has made possible the sequencing of complete genomes of several cyanobacterial species and has facilitated the comprehensive understanding of their genetic systems.

Synechocystis sp. PCC6803 is capable of natural transformation, which means that the cells can easily take up exogeneously added DNA. It is a photoautotroph but is capable of heterotrophic growth in the absence of light, which allows for analysis of the mechanism of oxygen-producing photosynthesis by characterization of mutants deficient in both photosystems I and II. For these reasons, this organism has been widely used for genetic and physiological studies of photosynthesis. In 1996, the sequencing of the entire genome of *Synechocystis* was completed. This was the first fully sequenced genome of a photoautotroph. Since public release of the data, a variety of genomic approaches, including systematic gene disruption, transcriptome analysis, and proteome analysis, has been performed utilizing the sequence information. In this review, characteristic features of the *Synechocystis* genome and its constitutive genes will be described on the basis of nucleotide sequence data. The current status of large-scale functional analyses of *Synechocystis* will be reviewed. The status of genome sequencing of other cyanobacterial strains will also be summarized.

2. Sequence Features of the Synechocystis sp. PCC6803 Genome

The entire genome of *Synechocystis* sp. PCC 6803 was sequenced in 1996, which makes it the fourth genome to be completely sequenced and the first among phototrophic organisms. The total length of the circular genome of *Synechocystis* is 3 573 470 bp. The ratio of nucleotides is A: 26.1%, C: 23.8%, G: 23.9%, and T: 26.2%, which results in a G+C content of 47.7 %.

One of the notable features of the *Synechocystis* genome is the presence of two types of repetitive sequences, namely, HIP1 (highly iterated palindrome) and Insertion Sequences (IS)-like elements. HIP1 is an eight-base palindromic sequence, GCGATCGC, first reported in the genomes of *Synechococcus* species and other cyanobacterial strains. The *Synechocystis* genome contains 3160 copies of HIP1, which are fairly evenly distributed and occur at an average frequency of one copy per 1131 bp. Approximately 90% of HIP1 sequences are located in potential protein-coding regions. The origin and functional significance of the HIP1 element remain to be clarified.

A total of 77 IS-like elements were identified in the *Synechocystis* genome. They are classified into nine groups (ISY100, ISY508, ISY120, ISY203, ISY352, ISY391, ISY523, ISY802, and ISY052) on the basis of sequence similarity and/or the presence of inverted-terminal repeats typical of IS-elements. Twenty-six of the IS-like elements contain open reading frames (ORFs) capable of coding for full-length transposase proteins, while the remaining ORFs are disrupted by frame-shift or deletion mutations or the insertion of other IS-like elements. It has been suggested that these elements play a significant role in the local and dynamic rearrangement of the genome structure by simple transposition and/or homologous recombination between identical IS-like sequences at different locations in the genome.

3. Assignment of RNA and Protein-Coding Genes in the Synechocystis Genome

3.1. Potential Structural RNA Genes in the Genome

Potential RNA coding genes were assigned to the genome by computer-aided analysis, which included sequence similarity searches and functional predictions. Two copies of an rRNA gene cluster, 42 tRNA genes, and a gene for an RNA subunit of RNase P were identified in the *Synechocystis* genome.

The two rRNA gene clusters each consisted of 5028 bp of identical sequence containing genes for the 16S RNA, Ile-tRNA, 23S RNA, and 5S RNA, in that order. The gene clusters were located approximately 870 kb apart in reverse orientation. Other cyanobacteria and most of the algal and plant plastids contain two copies of an rRNA gene cluster in their genomes, whereas other eubacteria whose complete genome structures have been determined contain varying numbers of copies (from one to seven) of rRNA gene clusters, suggesting a phylogenetic relationship among cyanobacteria and plant plastids.

A total of 42 tRNA genes representing 41 tRNA species were identified in the *Synechocystis* genome. Only the gene for trnI-GAU, which was located in two identical

copies of the rRNA gene cluster, was duplicated. The number of tRNA genes in the *Synechocystis* genome is limited compared to those in the *E. coli* genome (86 tRNA genes), where multiplication of genes is commonly observed. But the number of tRNA genes in the *Synechocystis* genome is sufficient for recognition of all the codons in this species. In *Synechocystis*, the tRNA genes are scattered throughout the genome, and most of the genes are found as single units. The only exceptions are the genes for trnY-GUA and trnT-GGU, which are aligned in the same direction with an 8 bp interval. In contrast, in *E. coli*, 70% of the tRNA genes form clusters consisting of between two and nine genes. It is also noteworthy that many of the tRNA genes in *Synechocystis* show a high degree of sequence similarity to those in plant plastids.

3.2. Potential Protein-Coding Genes in the Genome

Potential protein-coding regions were assigned to the genome by a combination of sequence similarity and computer predictions. As a result, a total of 3168 potential protein-coding genes were located in the *Synechocystis* genome. The gene density was one gene per 1.1 kb. The average length of the putative gene products was 326 amino acid residues. The potential protein-coding regions occupied 87.0% of the genome.

Of the 3168 potential protein-coding genes, 145 (4.6%) were identical to genes previously reported in *Synechocystis*, 935 (29.4%) were highly similar to genes whose function had been previously deduced, and 324 (10.2%) showed limited sequence similarity to known genes. Another 340 genes (10.8%) were similar to hypothetical genes of other organisms, and the remaining 1424 (45.0%) did not match any sequences in the public DNA databases at the time of publication. These results indicated that no functional information was provided for approximately 56% of the genes in the *Synechocystis* genome. The 1402 genes whose functions were deduced were classified into 15 categories according to their biological function (Table 1). Detailed information on each of the genes in the genome is provided in the *Synechocystis* genome database named CyanoBase (see Section 6 of this article).

Category	**Gene number**
Amino acid biosynthesis	84
Biosynthesis of cofactors, prosthetic groups, and carriers	108
Cell envelope	64
Cellular processes	62
Central intermediary metabolism	31
Energy metabolism	86
Fatty acid, phospholipid, and sterol metabolism	35
Photosynthesis and respiration	131
Nucleic acid metabolism	38
General regulatory functions	147
DNA replication, recombination, and repair	49
Transcription	24
Translation	144
Transport and binding proteins	158
Other categories	255

Function unknown	1751
Total	3167

Table 1. Gene category list of the genome of *Synechocystis* sp. PCC6803

4. Characteristic Features of *Synechocystis* Genes

4.1. Genes for Photosynthesis

A total of 128 genes involved in the various processes of photosynthesis have been identified on the basis of sequence similarity to known photosynthetic genes. The relative locations of these genes in the genome are shown in Figure 1. The following genes are present in multiple copies in the *Synechocystis* genome: *cpcC* (two copies), *cpcG* (two copies), *ctaC* (two copies), *ctaD* (two copies), *ctaE* (two copies), *ndhD* (six copies), *ndhF* (three copies), *petC* (three copies), *petF* (four copies), *psaK* (two copies), *psbA* (two copies), *psbC* (*isiA*) (two copies), and *psbD* (two copies). Unlike in higher plants, the genes coding for the small subunits of photosystems I and II (*psaG*, *psaH*, *psaN*, *psbP*, *psbQ*, *psbR*, *psbS*, *psbTn*, and *psbW*) were not identified in the *Synechocystis* genome.

-
-
-

Bibliography

Bryant D.A. (ed.) (1994). *The Molecular Biology of Cyanobacteria*. Kluwer Academic Publishers: The Netherlands, 881 pp. [This book summarizes knowledge obtained by molecular genetics of cyanobacteria.]

Hughes J., Lamparter T., Mittmann F., Hartmann E., Gärtner W., Wilde A., and Börner T. (1997). A prokaryotic phytochrome. *Nature* **386**, 663. [This paper describes the features of a phytochrome gene found in the genome of cyanobacteria.]

Kaneko T., Sato S., Kotani H., Tanaka A., Asamizu E., Nakamura Y., Miyajima N., Hirosawa M., Sugiura M., Sasamoto S., Kimura T., Hosouchi T., Matsuno A., Muraki A., Nakazaki N., Naruo K., Okumura S., Shimpo S., Takeuchi C., Wada T., Watanabe A., Yamada M., Yasuda M., and Tabata S. (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 Research* **3**, 109–136. [This is the first report of the genome structure of a cyanobacterium *Synechocystis* sp. PCC6803.]

Mizuno T., Kaneko T., and Tabata S. (1996). Compilation of all genes encoding bacterial two-component signal transducers in the genome of the cyanobacterium, *Synechocystis* sp. strain PCC 6803. *DNA Research* **3**, 407–414. [This paper summarize the genes for two-component signal transduction system in the genome of *Synechocystis* sp. PCC6803.]

Nakamura Y., Kaneko T., Hirosawa M., Miyajima N., and Tabata S. (1998). CyanoBase, a WWW database containing the complete nucleotide sequence of the genome of *Synechocystis* sp. strain PCC6803 *Nucleic Acids Research* **26**, 63–67. [This paper presents the genome database for *Synechocystis* sp. PCC6803.]

Nakamura Y., Kaneko T., Miyajima N., and Tabata S. (1999). Extension of CyanoBase. CyanoMutants:repository of mutant information on *Synechocystis* sp. Strain PCC 6803 *Nucleic Acids Resesrch* **27**, 66–68. [This paper presents the mutant database for *Synechocystis* sp. PCC6803.]

Sazuka T., Yamaguchi M., and Ohara O. (1999). Cyano2Dbase updated: Linkage of 234 protein spots to corresponding genes through N-terminal microsequencing. *Electrophoresis* **20**, 2160–2171. [This paper presents the proteome database for *Synechocystis* sp. PCC6803.]

Xiong J., Fischer W.M., Inoue K., Nakahara M., and Bauer C.E. (2000). Molecular evidence for the early evolution of photosynthesis. *Science* **289**, 1724–1730. [This paper discusses the origin and the evolution of photosynthesis by comparison of structures of photosynthetic genes.]

Biographical Sketch

Satoshi Tabata is project leader at the Kazusa DNA Research Institute. He studied as an undergraduate in the Department of Biology, School of Science, Kobe University (1973–1977) and then at the Institute for Chemical Research, Kyoto University, receiving his Ph.D. in 1983. The author was a Postdoctoral Research Associate at the University of California, San Diego (1983–1984), an Assistant Professor at Kyoto University (1985–1987), an Assistant Professor and then Associate Professor at Nagoya University (1988–1993), and then became a Senior Researcher at the Kazusa DNA Research Institute in 1994. His major working field is plant molecular genetics.