DNA Research **18**, 393–399, (2011) doi:10.1093/dnares/dsr026
Advance Access Publication on 29 July 2011

Genomic Structure of the Cyanobacterium *Synechocystis* sp. PCC 6803 Strain GT-S

Naoyuki Tajima[1], Shusei Sato[2], Fumito Maruyama[3], Takakazu Kaneko[4], Naobumi V. Sasaki[1], Ken Kurokawa[5], Hiroyuki Ohta[6], Yu Kanesaki[7], Hirofumi Yoshikawa[7,8], Satoshi Tabata[2], Masahiko Ikeuchi[1], and Naoki Sato[1,*]

Department of Life Sciences, Graduate School of Arts and Sciences, University of Tokyo, Komaba 3-8-1, Meguro-ku, Tokyo 153-8902, Japan[1]; Kazusa DNA Research Institute, Kazusa-Kamatari 2-6-7, Kisarazu, Chiba Prefecture 292-0818, Japan[2]; Section of Bacterial Pathogenesis, Graduate School of Medical and Dental Science, Tokyo Medical and Dental University, Yushima 1-5-45, Bunkyo-ku, Tokyo 113-8510, Japan[3]; Faculty of Life Sciences, Kyoto Sangyo University, Motoyama, Kamigamo, Kita-ku, Kyoto 603-8555, Japan[4]; Department of Biological Information, Tokyo Institute of Technology, 4250-B65, Nagatsuta-cho, Midori-ku, Yokohama 226-8501, Japan[5]; Center for Biological Resources and Informatics, Tokyo Institute of Technology, 4250-B65, Nagatsuta-cho, Midori-ku, Yokohama 226-8501, Japan[6]; Genome Research Center, NODAI Research Institute, Tokyo University of Agriculture, 1-1-1 Sakuragaoka, Setagaya-ku, Tokyo 156-8502, Japan[7] and Department of Bioscience, Tokyo University of Agriculture, 1-1-1 Sakuragaoka, Setagaya-ku, Tokyo 156-8502, Japan[8]

*To whom correspondence should be addressed. Tel. +81 3-5454-6631. Fax. +81 3-5454-6998.
Email: naokisat@bio.c.u-tokyo.ac.jp

Edited by Naotake Ogasawara
(Received 8 June 2011; accepted 30 June 2011)

Abstract

***Synechocystis* sp. PCC 6803 is the most popular cyanobacterial strain, serving as a standard in the research fields of photosynthesis, stress response, metabolism and so on. A glucose-tolerant (GT) derivative of this strain was used for genome sequencing at Kazusa DNA Research Institute in 1996, which established a hallmark in the study of cyanobacteria. However, apparent differences in sequences deviating from the database have been noticed among different strain stocks. For this reason, we analysed the genomic sequence of another GT strain (GT-S) by 454 and partial Sanger sequencing. We found 22 putative single nucleotide polymorphisms (SNPs) in comparison to the published sequence of the Kazusa strain. However, Sanger sequencing of 36 direct PCR products of the Kazusa strains stored in small aliquots resulted in their identity with the GT-S sequence at 21 of the 22 sites, excluding the possibility of their being SNPs. In addition, we were able to combine five split open reading frames present in the database sequence, and to remove the C-terminus of an ORF. Aside from these, two of the Insertion Sequence elements were not present in the GT-S strain. We have thus become able to provide an accurate genomic sequence of *Synechocystis* sp. PCC 6803 for future studies on this important cyanobacterial strain.**

Key words: *Synechocystis* sp. PCC 6803; genome re-sequencing; insertion sequence; single nucleotide polymorphism; CyanoClust

1. Introduction

The nucleotide sequence of the genome of the cyanobacterium *Synechocystis* sp. PCC 6803 was determined by Kazusa DNA Research Institute in 1996 as the first genome of photosynthetic organism.[1] After that, this strain has been serving as a standard of cyanobacteria in various areas of research, such as photosynthesis, stress response and metabolism.[2] However, the sequenced strain (called Kazusa strain in the present study) is different from the stock in Pasteur Culture Collection (called PCC strain in the

present study). In fact, the Kazusa strain is a derivative of a 'glucose-tolerant' strain, which was obtained by J.G.K. Williams in DuPont Institute.[3] The published sequence of the Kasuza strain included some genes inactivated by a putative point mutation, a putative frame shift, or an Insertion Sequence (IS) insertion, such as a one in the *pilC* gene. The mutation within the coding sequence of the *pilC* gene was pointed out to be a possible reason for the non-motility of the Kazusa strain.[4] A 154 bp deletion was also found in the GT strain with respect to the PCC strain.[5] The location of some IS elements in the Kazusa strain is known to be different with respect to other GT and PCC strains.[6] Even within the PCC strains, different strains having different light responses have been isolated.[2] All these slightly different strains bear the common strain name PCC 6803, but we need to recognize differences in exact strains used in various studies. For this purpose, we will have to pinpoint the differences in genome sequences of various different strains.

One of the authors (N.S.) constructed 40 site-directed mutants in a previous work on comparative genomics of plants and cyanobacteria[7] using the laboratory stock of *Synechocystis* GT strain (called GT-S). We thought that this strain should be identical to the Kazusa strain, because it originated in the late 1980s from the strain owned by Dr T. Omata, which was also the source of the Kazusa strain. However, in view of the small but significant differences in genome sequence as reported earlier, it was important to establish the genetic background of our strain to assess correctly the phenotype of the above-mentioned mutants. Therefore, we attempted to analyse the genome sequence of the strain GT-S and to compare it with the reference sequence of the Kazusa strain. We found significant differences with respect to the database sequence, but we were finally convinced that the differences in the real sequences were minimal.

2. Materials and methods

2.1. Strain and genomes

Synechocystis GT-S strain was originally a gift from Dr Tatsuo Omata (Nagoya University, but he was in Riken Institute then) in the late 1980s, and then maintained in Sato laboratory as frozen glycerol stocks. In the present study, we used the stock originally frozen in the early 1990s. The cells were grown in the BG-11 medium at 32°C with aeration as described before.[8] The cells were harvested by centrifugation, and then washed twice with 4 M NaI to remove extracellular polysaccharide, and then, treated with lysozyme. DNA was released by treatment with proteinase K and sodium *N*-dodecanoylsarcosinate, extracted with phenol and chloroform and purified by CsCl ultracentrifugation.[9] As a reference, we also used an aliquot of the DNA of the original Kazusa strain, which had been stored as a stock in Kazusa DNA Research Institute.

2.2. Sequencing and data analysis

Genomic DNA was sheared by ultrasonic treatment and sequenced by a genome sequencer FLX instrument (Roche Diagnostics, Indianapolis, IN, USA) according to the manufacturer's protocol (this is usually referred to as '454 sequencing'). To find its genomic origin, namely, main genome or plasmids, each read was analysed by BLASTN[10] software version 2.2.18 using the sequences of the four plasmids as well as the main genome as targets (the accession numbers are given in Supplementary Table S5). The options were: -FF −e 0.0001 −v 2 −b 2 −m 8 −C F (no filtering, cut-off *E*-value = 0.0001, output and list sequences = 2, table-formatted output, no compositional adjustments). In the table-formatted output, only the first line corresponding to the highest identity was selected for each read, which was assigned to the genome shown therein. The authentic reads assigned for genomic DNA obtained in this way were mapped onto the reference sequence of the Kazusa strain (GenBank and RefSeq accession numbers: BA000022 and NC_000911 for the main genome) by the inGAP software version 2.3.1.[11] Unfortunately, the details of internal algorithm of the software are not clear, and there is no option related to the detection of SNPs. Therefore, all putative SNPs detected by default settings were analysed. Plasmids were also analysed by using respectively assigned reads. A list of putative SNPs was obtained as an output. Homology of affected open reading frames (ORFs) with orthologues in other cyanobacteria was analysed by the cluster data of CyanoClust database[12] prepared by the Gclust software.[13] Processing of DNA and protein sequences was performed with the SISEQ software version 1.59.[14] Sequence alignments were constructed with the Clustal X software version 2.0.9.[15] Genomic sequence was manipulated by the Artemis software version 13.0.[16]

2.3. Sequence confirmation

For each putative SNP, a genomic region of 200–300 bp was amplified (see Supplementary Table S1 for primer sequences). For each putative IS element, a genomic region of ~300 or 1500 bp was amplified (see Supplementary Table S2 for primer sequences). The amplification of a long DNA was to overcome

repeated sequences. DNA templates of both GT-S and Kazusa strains were used. The products were sequenced by conventional Sanger sequencing, using the sequencing services of MACROGEN Japan Corp. (Tokyo, Japan) or FASMAC Co. Ltd. (Atsugi, Japan).

3. Results

3.1. Identification of SNPs

We obtained 197 912 reads having an average length of 399.3 bases for the GT-S strain by 454 sequencing. Without the preliminary classification of reads, 68 single-nucleotide polymorphisms (SNPs) were obtained for the main genome, but many of them were not correct, because of the presence of highly homologous genes in plasmids. Then the reads were allocated to the main genome and the four plasmids by homology analysis as described in Section 2.2. The 173 217 reads that were classified as reads for the main genome were mapped to the reference sequence NC_000911. Using the default settings of inGAP software (see Supplementary data for the list of options), the entire genome was covered by at least one read, except four small regions (Supplementary Table S3). The analysis of such gap regions was performed separately, as described below. As a result, 31 putative SNPs were detected by the inGAP analysis. All of them were selected as highly probable SNPs for experimental validation.

Each of the putative SNPs was checked by PCR amplification and Sanger sequencing of both strands. Twenty-two SNPs (Table 1) were finally identified as the differences of the sequence of the GT-S strain with respect to the database sequence NC_000911 (identical to BA000022 with respect to the DNA sequence). To verify that these represent real differences of the two strains, we analysed, by Sanger sequencing, the DNA of the Kazusa strain, which had been stored in small aliquots. Surprisingly, all the putative SNP sites were found identical in the Kazusa strain and the GT-S strain except No. 8 (Table 1). The SNP No. 8 is the mutation

Table 1. List of putative SNPs

No.	Site	Gene	CyanoClust cluster no.	Database	GT-Kazusa	GT-S	Amino acid change	Annotation	Ref.
1	943495	*psaA*	16	G	A	A	V→I	P700 apoprotein subunit Ia	18
2	1012958	No gene	—	G	T	T	N/A	—	
3	1364187	*pyrF*	784	A	G	G	None	Orotidine 5' monophosphate decarboxylase	
4	1819782	*psbA3*	18	A	G	G	None	Photosystem II D1 protein	17
5	1819788			A	G	G	None		
6	2092571	*sll0422*	1760	A	T	T	L→ter	Asparaginase	
7	2198893	*sll0142*	15	T	C	C	None	Cation or drug efflux system protein	
8	2204584	*gspF+pilC*	917+7792	G	G	—	Frame shift	Pilin biogenesis protein	4
9	2301721	*slr0168*	6624	A	G	G	K→E	Hypothetical protein	
10	2350285.5	No gene	—	—	A	A	N/A	—	
11	2360245.5	*slr0364*	26 765+19 649	—	C	C	Frame shift	Hypothetical protein	
12	2409244	*sll0762*	2611	C	—	—	Frame shift	Hypothetical protein	
13	2419399	*ycf22*	779	T	—	—	Frame shift	Hypothetical protein	
14	2544044.5	*ssl0787*	2596	—	C	C	Frame shift	Hypothetical protein	
15	2602717	*slr0468*	31358	C	A	A	H→Q	Hypothetical protein	
16	2602734			T	A	A	I→N		
17	2748897	No gene	—	C	T	T	N/A	—	
18	3096187	*ssr1175*	796	T	C	C	I→T	Transposase	
19	3110189	No gene	—	G	A	A	N/A	—	
20	3110343	*sll0665*	1448	G	T	T	P→Q	Transposase	
21	3142651	*sps*	2831	A	G	G	None	Sucrose phosphate synthase	
22	3260096	No gene	—	C	—	—	N/A	—	

GT-Kazusa and GT-S are *Synechocystis* sp. PCC 6803 strain GT in Kazusa DNA Research Institute and Sato Laboratory. 'Site' and 'Database' refers to the sequences in BA000022 or NC_000911. Insertion site numbers represent the last position of insertion site + 0.5. N/A indicates that the amino acid change is not applicable because SNP site is not in an ORF.

within the *pilC* gene, which had been reported earlier.[4] The two putative SNPs in the *psbA3* coding region were identical to the corresponding sites of the *psbA2* gene. Since the correct *psbA3* sequence had been published before the genome sequence,[17] these putative SNPs are probably sequencing artefacts in NC_000911. A putative SNP site in the *psaA* gene also matches the previously published sequence.[18] In other cases, we have no clear explanation, and might be sequencing errors and/or mutations in cosmid clones used in the original sequencing.

Unfortunately, the mapping of reads on to the reference genome was not perfect using the obtained reads. In 14 short regions, no reads or at most two reads were mapped (Supplementary Tables S3 and S4). These regions were amplified by PCR for both GT-S and Kazusa strains (results not shown). Conventional sequencing of the PCR products confirmed that there is no sequence difference in 11 of these regions with respect to the database sequence. The remaining three regions having two reads were close to one another and located within a 3 kb region. Clean PCR amplification of this 3 kb region was not successful because of repeated sequences. However, the presence of two reads led us to tentatively conclude that there is no sequence difference in these regions.

3.2. Analysis of plasmids

Plasmids were also analysed by inGAP mapping. There were no putative SNPs in pSYSM and pSYSG (Supplementary Table S5). In pSYSA, four sites were reported as putative SNPs, but all of them represent sites having only two reads and one of the reads matched database sequence. Therefore, these were not considered as SNPs in pSYSA. In pSYSX, four sites within or near *ssr6089* gene were detected as putative SNPs. Analysis using the CyanoClust database indicated that this plasmid contains 30 kb homologous regions, *ssr6002–slr6038* and *ssr6062–slr6094*. The *ssr6089* gene has a nearly identical homologue *ssr6030*. However, the sequence corresponding to the four putative SNPs were identical in the two genes in the database sequence NC_005232. Therefore, the SNP calling was not due to mixing of reads for homologous genes. The SNPs could possibly represent mutations in the strain GT-S, but final validation is hampered by high similarity of the long homologous regions.

3.3. Alteration of ORFs due to frame shift

There are five cases in which a single gene is split into a pair of genes as a result of frame shift. Figure 1A shows the site of putative SNP 12, namely the *sll0762–sll0763* region. There is an extraneous C in the database sequence, and accordingly, the removal of this C results in fusion of the two ORFs. This new ORF encoding a hypothetical protein has well-conserved orthologues in other cyanobacteria (*Anabaena*, *Cyanothece*, *Arthrospira* etc.) as shown by the alignment of the cluster 2611 of the CyanoClust (Fig. 1B).

To correct the database sequence to obtain the GT-S genome sequence, we should combine (i) *slr0162* (*gspF*) and *slr0163* (*pilC*), (ii) *slr0364* and *slr0366*, (iii) *sll0762* and *sll0763* (this is described above), (iv) *sll0751* (*ycf22*) and *sll0752*, and (v) *ssl0787* and *ssl0788* (Supplementary Figs S1 and S2). In addition, the extended C-terminus of Sll0422 protein should be removed after correction for the nucleotide change (Supplementary Fig. S1). All these changes except (i) also apply to the real sequence of Kazusa strain.

3.4. Large indels

We also checked large indels (insertion/deletions). The exact sites of insertion of various IS elements have already been analysed.[6] Among them, ISY*203*b insertion between *slr1862* and *slr1863* and ISY*203*g insertion between *sll1473* and *sll1475* were found in the Kazusa strain but not in the GT-S strain. ISY*203*e insertion between *ssl2982* and *slr1636* was detected in both Kazusa and GT-S strains (Table 2) but not in another GT strain in Ikeuchi laboratory. It has also been known that a 154 bp element upstream of the *slr2031* gene is deleted in the GT strains.[5] This deletion was shared by all GT strains analysed in the present study.

3.5. Finally validated differences of the two strains

All previous description was based on the comparison using the database sequence as the sole reference. Given that there are a number of changes that have to be made for the database sequence, we summarize our results as the differences between the real sequences of GT-S and GT-Kazusa. The two sequences are essentially identical except a single frame-shift mutation in the *pilC* gene and two more insertions of ISY*203* in GT-Kazusa with respect to GT-S.

4. Discussion

The present study revealed that a significant number of differences are present in the database sequence and the genome sequences of laboratory strains of the same 'species' *Synechocystis* sp. PCC 6803. The detailed analysis using the genomic DNA of both Kazusa and GT-S strains indicated that the detected 21 putative SNPs were, in fact, differences in the database sequence, but not real differences in

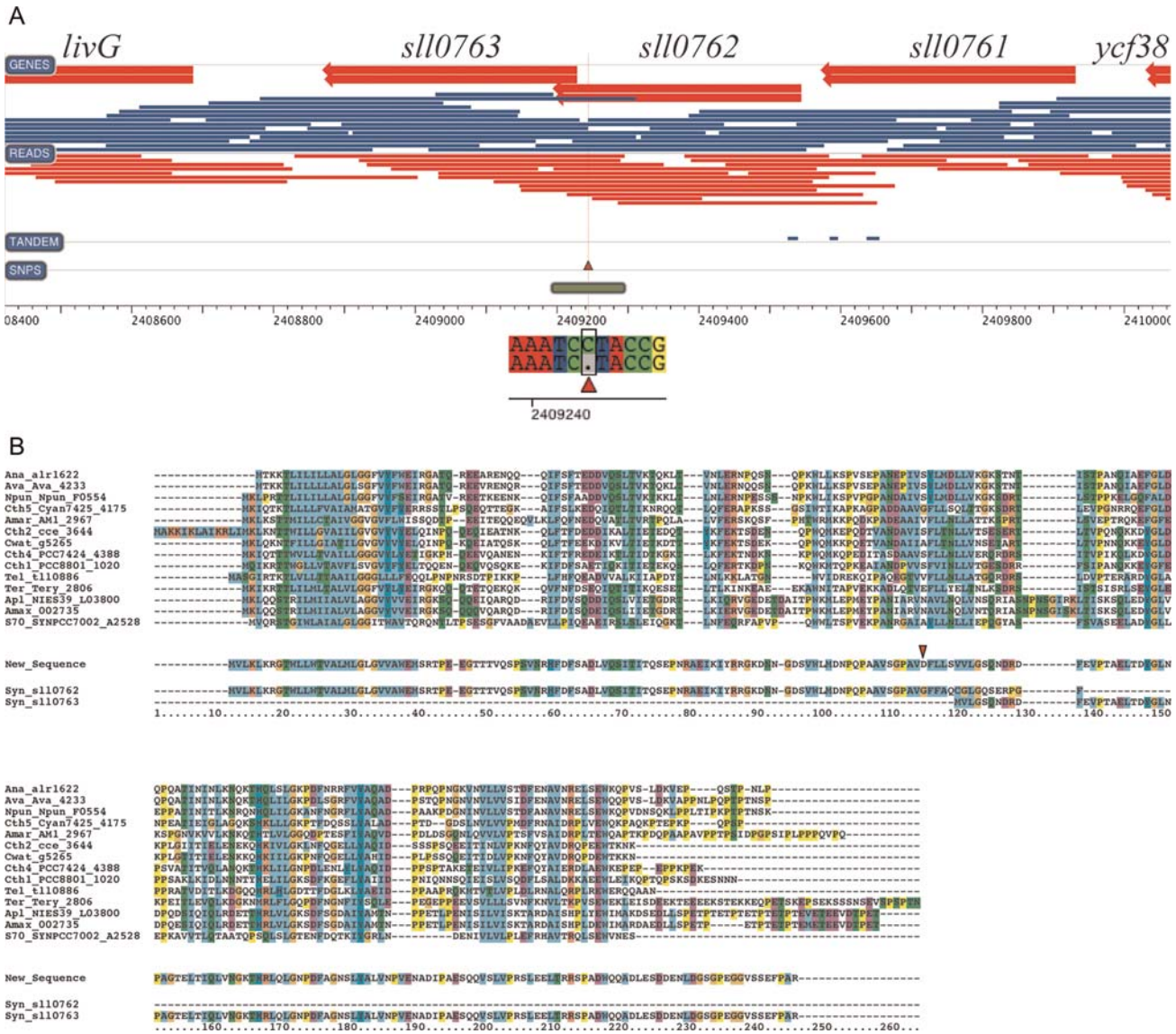

Figure 1. Correction of ORFs due to a frame shift in the *sll0762–sll0763* retion. (A) Output of an SNP site in the reference sequence of the Kazusa strain by the inGAP software. The upper DNA sequence indicates the reference sequence of the Kazusa strain (GenBank and RefSeq accession numbers: BA000022 and NC_000911), and the lower DNA sequence indicates the sequence of the GT-S strain. Each arrow represents a gene. Each arrowhead indicates an SNP site. (B) New alignment with a corrected sequence. Homology of affected ORFs with corresponding sequences in other cyanobacteria was analysed by the CyanoClust database version 4, and the cluster 2611 was found. Sequences were retrieved and a new alignment was obtained by the Clustal X software. 'New_Sequence' indicates the corrected sequence. Arrowhead indicates the nucleotide variations detected as putative SNP site.

Table 2. List of ISY*203*s detected in GT strains

IS name	Transposase gene	Database	GT-Kazusa	GT-S
ISY*203*b	*sll1780*	Yes	Yes	No
ISY*203*e	*slr1635*	Yes	Yes	Yes
ISY*203*g	*sll1474*	Yes	Yes	No

the two genomes. The final balance sheet indicates that we found a single nucleotide change and two IS insertions between the Kazusa strain and the GT-S strain. The time of separation of the two strains may be estimated as the mid-1980s according to the opinions of concerned people, which are now quite obscure. The time interval until the DNA isolation for sequencing may be roughly estimated as about 10 years for the Kazusa strain. The GT-S strain was stocked in the early 1990s, and re-plated in 2010 for the present analysis. The effective time interval from the separation of the two sub-strains was also about 10 years. The results suggest that nucleotide change (mutation) could be kept to a minimum (only one, in this case) if due attention is paid for maintenance of strains, but IS mobilization may be more frequent (two events). The rapid mobilization of IS could be limited to the particular element

ISY*203*, but we do not know the actual trigger of activation of this IS element. We, therefore, should be careful about IS activation in the maintenance of laboratory stocks. We will need a convenient way of detecting a mobilized ISY*203* to be sure about our research using the GT strain.

The nucleotide changes as a result of re-sequencing caused significant effects on gene annotation. As mentioned, five genes had been thought split into two by a single nucleotide difference before this analysis. The length of another gene was also changed. The IS element inserted in the *sll1474* (*ccaS*) gene is known to inactivate it.[6] Altogether, the nucleotide changes (whether sequencing errors or real mutations) have an important impact on molecular biological researches using cyanobacteria or other bacteria. A single run of new generation sequencing with some additional PCR experiments can establish identity of the organism that is being used in the laboratory. This will become a standard of molecular genetics in microbiology.

The genomic database is very important in not only experimental studies but also computational analysis. The use of correct sequence is a prerequisite for detailed comparative genomics research. The 21 sites per 3.6 Mb genome are significantly large number for present-day level of genome analysis. The correction of the standard sequence will be especially useful in *Synechocystis* sp. PCC 6803, which is a standard cyanobacterial strain in various areas of research such as photosynthesis and stress response among others. We hope our data deposited as a new separate entry will be useful for all those who are using this cyanobacterium in various researches.

5. Databases

The genome sequence of the strain GT-S was deposited in the DDBJ/GenBank/EMBL database under the accession number AP012205.

Acknowledgements: The authors thank people in Ikeuchi and Sato laboratories for discussion.

Supplementary data: Supplementary data are available at www.dnaresearch.oxfordjournals.org.

Funding

This work was supported in part by the Global Center of Excellence (GCOE) Program 'From the Earth to "Earths"' from the MEXT, Japan.

References

1. 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**, 109–36.
2. Ikeuchi, M. and Tabata, S. 2001, *Synechocystis* sp. PCC 6803—a useful tool in the study of the genetics of cyanobacteria, *Photosynth. Res.*, **70**, 73–83.
3. Williams, J.G.K. 1988, Construction of specific mutations in photosystem II photosynthetic reaction center by genetic engineering methods in *Synechocystis* 6803, *Methods Enzymol.*, **167**, 766–78.
4. Bhaya, D., Bianco, N.R., Bryant, D. and Grossman, A. 2000, Type IV pilus biogenesis and motility in the cyanobacterium *Synechocystis* sp. PCC6803, *Mol. Microbiol.*, **37**, 941–51.
5. Katoh, A., Sonoda, M. and Ogawa, T. 1995, A possible role of 154-base pair nucleotides located upstream of ORF440 on CO_2 transport of *Synechocystis* PCC6803. In: *Photosynthesis: from Light to Biosphere* Mathis, P. (ed.), *vol. 3*, Springer: Dordrecht, 481–4.
6. 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**, 265–73.
7. Ishikawa, M., Fujiwara, M., Sonoike, K. and Sato, N. 2009, Orthogenomics of photosynthetic organisms: bioinformatic and experimental analysis of chloroplast proteins of endosymbiont origin in *Arabidopsis* and their counterparts in *Synechocystis*, *Plant Cell Physiol.*, **50**, 773–88.
8. Sato, N. 1994, Effect of exogenous glucose on the accumulation of monoglucosyl diacylglycerol in the cyanobacterium *Synechocystis* PCC 6803, *Plant Physiol. Biochem.*, **32**, 121–6.
9. Porter, R.D. 1988, DNA transformation, *Methods Enzymol.*, **167**, 703–12.
10. Altschul, S.F., Madden, T.L., Schäffer, A.A., et al. 1997, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.*, **25**, 3389–402.
11. Qi, J., Zhao, F., Buboltz, A. and Schuster, S.C. 2010, inGAP: an integrated next-generation genome analysis pipeline, *Bioinformatics*, **26**, 127–9.
12. Sasaki, N.V. and Sato, N. 2010, CyanoClust: comparative genome resources of cyanobacteria and plastids, *Database*, **2010**, bap025.
13. Sato, N. 2009, Gclust: *trans*-kingdom classification of proteins using automatic individual threshold setting, *Bioinformatics*, **25**, 599–605.
14. Sato, N. 2000, SISEQ: manipulation of multiple sequence and large database files for common platforms, *Bioinformatics*, **16**, 180–1.
15. Thompson, J.D., Higgins, D.G. and Gibson, T.J. 1994, CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence

weighting, position-specific gap penalties and weight matrix choice, Nucleic Acids Res., **22**, 4673–80.
16. Rutherford, K., Parkhill, J., Crook, J., et al. 2000, Artemis: sequence visualization and annotation, *Bioinformatics*, **16**, 944–5.
17. Metz, J., Nixon, P. and Diner, B. 1990, Nucleotide sequence of the *psbA3* gene from the cyanobacterium *Synechocystis* PCC 6803, *Nucleic Acids Res.*, **18**, 6715.
18. Smart, L.B. and McIntosh, L. 1991, Expression of photosynthesis genes in the cyanobacterium *Synechocystis* sp. PCC 6803: *psaA–psaB* and *psbA* transcripts accumulate in dark-grown cells, *Plant Mol. Biol.*, **17**, 959–71.