DNA Research **14**, 247–256, (2007) doi:10.1093/dnares/dsm026

Complete Genomic Structure of the Bloom-forming Toxic Cyanobacterium *Microcystis aeruginosa* NIES-843

Takakazu Kaneko[1,*], Nobuyoshi Nakajima[2], Shinobu Okamoto[1], Iwane Suzuki[3], Yuuhiko Tanabe[2], Masanori Tamaoki[2], Yasukazu Nakamura[1], Fumie Kasai[2], Akiko Watanabe[1], Kumiko Kawashima[1], Yoshie Kishida[1], Akiko Ono[1], Yoshimi Shimizu[1], Chika Takahashi[1], Chiharu Minami[1], Tsunakazu Fujishiro[1], Mitsuyo Kohara[1], Midori Katoh[1], Naomi Nakazaki[1], Shinobu Nakayama[1], Manabu Yamada[1], Satoshi Tabata[1] and Makoto M. Watanabe[3]

Kazusa DNA Research Institute, 2-6-7 Kazusa-kamatari, Kisarazu, Chiba 292-0818, Japan[1]; National Institute for Environmental Studies, 16-2 Onogawa, Tsukuba, Ibaraki 305-8506, Japan[2] and Graduate School of Life and Environmental Sciences, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8572, Japan[3]

(Received 29 November 2007; accepted 2 December 2007; published online 11 January 2008)

Abstract

The nucleotide sequence of the complete genome of a cyanobacterium, *Microcystis aeruginosa* NIES-843, was determined. The genome of *M. aeruginosa* is a single, circular chromosome of 5 842 795 base pairs (bp) in length, with an average GC content of 42.3%. The chromosome comprises 6312 putative protein-encoding genes, two sets of rRNA genes, 42 tRNA genes representing 41 tRNA species, and genes for tmRNA, the B subunit of RNase P, SRP RNA, and 6Sa RNA. Forty-five percent of the putative protein-encoding sequences showed sequence similarity to genes of known function, 32% were similar to hypothetical genes, and the remaining 23% had no apparent similarity to reported genes. A total of 688 kb of the genome, equivalent to 11.8% of the entire genome, were composed of both insertion sequences and miniature inverted-repeat transposable elements. This is indicative of a plasticity of the *M. aeruginosa* genome, through a mechanism that involves homologous recombination mediated by repetitive DNA elements. In addition to known gene clusters related to the synthesis of microcystin and cyanopeptolin, novel gene clusters that may be involved in the synthesis and modification of toxic small polypeptides were identified. Compared with other cyanobacteria, a relatively small number of genes for two component systems and a large number of genes for restriction-modification systems were notable characteristics of the *M. aeruginosa* genome.

Key words: cyanobacterium; *M. aeruginosa*; microcystin; water bloom; genome sequence

1. Introduction

The planktonic, gas-vacuolated cyanobacteria, including *Anabaena*, *Aphanizomenon*, *Cylindrospermopsis*, *Microcystis*, *Nodularia*, and *Planktothrix*, cause water blooms in eutrophic lakes, ponds, and reservoirs all over the world. They also produce a diverse range of toxins, including neurotoxins [e.g. anatoxin-a, anatoxin-a(s), and saxitoxins] and hepatotoxins (e.g. microcystins, nodularins, and cylindropsermopsins), which cause a variety of human illnesses, and are responsible for deaths in native and domestic animals.[1] Among these, the unicellular, colonial cyanobacterium *Microcystis* is the most representative genus of toxic bloom-forming cyanobacteria. It produces a cyclic heptapeptide hepatotoxin, termed microcystin[1] and a depsipeptide, chymotrypsin-inhibitor called cyanopeptolin,[2] and is widely distributed geographically, from cold-temperate climates to the tropics.

Microcystis is characterized as a cyanobacterium with gas vesicles, a coccoid cell shape, a tendency to form aggregates or colonies, and an amorphous mucilage or sheath.[3] Generally, five morphospecies of *Microcystis*,

Edited by Katsumi Isono

* To whom correspondence should be addressed. Tel. +81 438-52-3935. Fax. +81 438-52-3934, E-mail: kaneko@kazusa.or.jp

M. aeruginosa, *M. ichthyoblabe*, *M. novacekii*, *M. viridis*, and *M. wesenbergii*, are recognized as the dominant species of water bloom-forming *Microcystis*.[4,5] These species were defined solely based on morphological characteristics, such as cell size, colony formation, and sheath characteristics. However, the morphology of *Microcystis* is highly variable, and the variation sometimes exceeds species criteria.[6] It has been suggested that the species definition of *Microcystis* is invalid for the following reasons: the low sequence divergence of 16S rDNA within and between morphospecies (<0.7%), the lack of correspondence between the morphospecies and the nucleotide sequences of the 16S–23S rDNA, and inability to differentiate fatty acid composition, GC content, temperature-salinity tolerance, and chemo- and photoheterotrophy among morphospecies.[7–9] In light of these considerations, and a high DNA–DNA re-association value of over 70%, which is high enough to integrate *Microcystis* into a single bacterial species,[10] Otsuka et al.[11] unified the five morphospecies into a single species under the Rules of the Bacteriological Code, and NIES-843^{T} was proposed as the type strain of *Microcystis*.

To date, the entire genome of 29 cyanobacteria with various characteristics have been sequenced (NCBI website, http://www.ncbi.nlm.nih.gov/genomes/static/eub_g.html, and CyanoBase, http://bacteria.kazusa.or.jp/cyano/). However, the genomic structures of the toxic bloom-forming cyanobacteria have not been reported, despite the level of social awareness of this species. In the present study, we describe sequencing and genomic information analysis of the entire genome of the type strain of *M. aeruginosa*, NIES-843^{T}, in order to further our understanding of the complex association between genetics, physiology, and biochemistry of this species.

2. Materials and methods

2.1. Genome sequencing

Microcystis aeruginosa NIES-843 (*M. aeruginosa*) was obtained from Microbial Culture Collection of National Institute for Environmental Studies (MCC-NIES) (http://www.nies.go.jp/biology/mcc/home.htm). Total cellular DNA was purified according to standard procedures, and two genomic libraries for sequencing were constructed, using two types of cloning vectors: MA, with inserts of ∼2.6 kb cloned into pUC118 (Takara Bio. Inc., Japan), and MAB, with inserts of ∼34 kb cloned into a BAC vector pCC1BAC (Epicentre Bio., USA).

Genome sequencing was performed using the whole-genome shotgun method in combination with BAC end sequencing. The nucleotide sequences of both ends of the clones from the MA and the MAB libraries were analyzed using a Dye-terminator Cycle Sequencing kit and the DNA sequencing systems MegaBase4000 (GE Healthcare, USA), type 3730XL (Applied Biosystems, USA), and DeNOVA-5000HT (Shimadzu Co., Japan). The accumulated sequences were assembled using the Phrap program (http://phrap.org/). The end-sequence data from the BAC clones facilitated the gap-closure process, and provided the scaffolding for reconstruction of the sequence of the entire genome. The final gaps in the sequence were filled either by primer walking or by sequencing PCR products that encompassed the gaps. The integrity of the reconstructed genome sequence was assessed by walking through the genome with the end sequences of the BAC clones.

2.2. Gene assignment, annotation, and information analyses

RNA- and protein-encoding regions were assigned by a combination of computer prediction and similarity searches, as described previously.[12]

Genes for structural RNAs were identified by similarity searches against an in-house structural RNA database that had been constructed based on the data in CyanoBase (http://bacteria.kazusa.or.jp/cyano/). tRNA- and rRNA-encoding regions were predicted by use of the tRNA scan-SE 1.23 program[13] and the RNAmmer ver.1.1d program,[14] respectively, in combination with similarity searches.

The prediction of protein-encoding regions was carried out with the Glimmer 3.02 prediction program.[15] Prior to prediction, a matrix was generated for the *M. aeruginosa* genome by training with a data set of 3406 open reading frames that showed a high degree of sequence similarity to genes registered in the translated EMBL protein database Rel. 37.2 (TrEMBL). All of the predicted protein-encoding regions equal to or longer than 150 bp were translated into amino acid sequences, which were then subjected to similarity searches against the TrEMBL database using the BLASTP program.[16] In parallel, all the predicted intergenic sequences were compared with sequences in the TrEMBL database using the BLASTX program, to identify genes that were not detected by the prediction process. For predicted genes that did not show sequence similarity to known genes, only those equal to or longer than 150 bp were considered candidates.

Functions of the assigned genes were deduced based on the sequence similarity of their translated protein products to those of genes of known function and to the protein motifs in the InterPro database (ver. 16.0).[17] A BLAST score of 10^{-5} was considered significant.

Assignment of Clusters of Orthologous Groups of proteins (COGs) of predicted gene products was carried out by the BLASTP analysis against the COG reference data set[18] (http://www.ncbi.nlm.nih.gov/COG/). A BLAST E-value of less than $E = 10^{-10}$ was considered significant. After filtering, COG assignments of the

putative gene products were generated according to COG identification, taking the best-hit pair in the reference data set.

Multicopy DNA elements of longer than 500 bp having a capacity to encode a putative transposase were identified as insertion sequences (ISs) using the BLAST2 program, then classified by RECON1.05[19] and IS finder (www-is.biotoul.fr).

Multiple copy elements of less than 600 bp long flanked by inverted repeats were identified as miniature inverted-repeat transposable elements (MITEs) using the BLAST2 and the RECON programs.

3. Results and discussion

3.1. Sequencing and structural features of the M. aeruginosa genome

The nucleotide sequence of the entire genome of *M. aeruginosa* was determined using the modified whole genome shotgun method, as described in Section 2. A total of 55 246 random sequences corresponding to ∼6.5 genome equivalents were assembled to generate draft sequences. Finishing was carried out by visually editing the draft sequences and by additional sequencing to close the gaps. The integrity of 98.6% of the final genome sequence was assessed by comparing the insert length of anchored BAC clones with the computed distance between the end sequences of the clones. Sixteen remaining gaps were closed by sequencing of the ends of the plasmid clones or of PCR products. The genome of *M. aeruginosa* was a circular molecule of 5 842 795 bp with an average GC content of 42.3%. No plasmid was detected during the course of this study. Nucleotide positions were assigned based on the predicted translational initiation site of the homolog of *sll0611* (solanesyl diphosphate synthase gene in *Synechocystis* sp. PCC 6803)[20] (Fig. 1 and Supplementary Fig. 1). The innermost circle of Fig. 1 shows the distribution of GC content in the genome. There was no characteristic pattern, according to GC skew analysis. An 8 bp palindromic sequence (5′GCGATCGC3′) termed HIP1 is frequently found in the genomes of a variety of cyanobacteria.[21] HIP1 was present in the *M. aeruginosa* genome (1821 copies), and the frequency of occurrence (1 copy/3209 bp) was 2.6–2.8-fold lower than *Synechocystis* sp. PCC 6803 (1 copy/1131 bp), and *Anabaena* sp. PCC 7120 (1 copy/1219 bp).

3.2. Structural features of M. aeruginosa genes

3.2.1. RNA-encoding genes Two copies of an rRNA gene cluster were assigned at coordinates 1 885 814–1 890 709 and 3 593 859–3 598 753 of the genome, in the sequence of 16S-*trnI*-23S-5S, and in the opposite direction to what is shown in Supplementary Fig. 1. A total of 42 tRNA genes, including those in the rRNA gene clusters, were identified, representing 41 tRNA species (Supplementary Figs 1 and 2, Supplementary Table 1). Most of the tRNA genes were dispersed throughout the genome and are likely to be transcribed as single units, with the exception of those in the rRNA gene clusters and *trnT*-GGU-*trnY*-GUA (Supplementary Fig. 2). No introns were found in any tRNA gene species. *M. aeruginosa* had a single gene for transfer-messenger RNA (tmRNA), which is known to be involved in the degradation of aberrantly synthesized proteins. Putative genes for small RNAs that showed sequence similarity to the B subunit of RNase P, signal recognition particle (SRP) RNA, and 6Sa RNA were assigned based on sequence similarity to reported genes.

3.2.2. Protein-encoding genes The potential protein-encoding regions were assigned using a combination of computer prediction by the Glimmer program and similarity search, as described in Section 2. By taking into account sequence similarity to known genes and relative position, to avoid overlaps, the total number of putative protein-encoding genes assigned to the genome was 6312. The average gene density was one gene per 925 bp. The putative protein-encoding genes that started with ATG, GTG, TTG, or ATT were denoted by a serial number with the prefix 'MAE', representing the species name *M. aeruginosa* (Supplementary Fig. 1). The codon usage frequency of all the gene components in the genome is listed in Supplementary Table 2. It should be noted that the putative genes assignments in this paper represent coding potential based on a defined set of assumptions.

Functional assignment of the 6312 putative protein-encoding genes was performed by similarity searches against the InterPro and TrEMBL databases, as described in Section 2. The number of genes with sequence similarity to genes of known function was 2588 (41%), 2304 (36%) showed sequence similarity to hypothetical genes, and the remaining 1433 (23%) did not show significant similarity to any registered genes.

COG assignment of the translated gene products was carried out by BLASTP search against the COG reference data set. A total of 3304 putative gene products encoded by *M. aeruginosa* were assigned to 1373 COG identifications in 20 COG categories (Supplementary Fig. 3). In comparison, 2812 of 3314 genes assigned in *Synechocystis* sp. PCC 6803 were classified into 20 COG categories using the same parameters. As shown in Supplementary Fig. 3, there was marked over-representation in the 'Replication, recombination, and repair' category in *M. aeruginosa*, due to the presence of a large number of transposase genes, as described in Section 3.3.1.

3.2.3. Functional domains A search of the genome sequences of *M. aeruginosa* and 27 additional cyanobacteria against the InterPro database (release v12.0) using

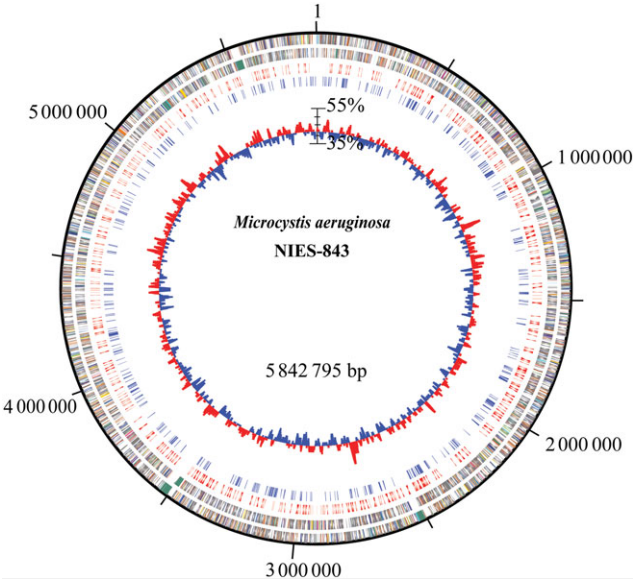

Figure 1. Schematic representation of the circular chromosome of *M. aeruginosa*. The scale indicates location in bp, starting with the initiation codon (ATG) of MAE00010, an ortholog of slr0611 in *Synechocystis* sp. PCC 6803. The bars in the outermost and the second circles show the positions of the putative protein-encoding genes in the clockwise and counter-clockwise directions, respectively. The putative genes are represented by 20 color codes, based on COG assignments (Supplementary Fig. 3). The bars in the third and fourth circles indicate the positions of MITEs (in red) and ISs (in blue). The innermost circles with scales show the percent average GC content calculated using a window size of 10 kb. The sequences as well as the gene information shown in this paper are available in CyanoBase (http://bacteria.kazusa.or.jp/cyano/). The sequence data analyzed in this study has been registered in DDBJ/GenBank/EMBL under accession number AP009552.

the InterProScan (v3.2) program detected 2017 Pfam domains (Supplementary Table 3). A total of 1384 domains were found in the *M. aeruginosa* genome, 777 of which were common to all cyanobacterial species examined. Nineteen domains were unique to *M. aeruginosa*. They consisted of the following IDs: DUF204, PP2C, 2_5_RNA_ligase, CHAP, Chlorophyllase, DUF1080, DUF1702, Eco57I, EcoRI, Gal-bind_lectin, Glyco_transf_22, Lipoxygenase, NodZ, PEPCK_ATP, PhnA_Zn_Ribbon, Ricin_B_lectin, Thymidylat_synt, Transposase_31, and tRNA-synt_1f, including two domains of restriction enzymes (CyanoBase, http://bacteria.kazusa.or.jp/cyano/). The top 30 Pfam domains in *M. aeruginosa* are listed in Supplementary Table 4. Note that *M. aeruginosa* harbors a smaller number of domains related to signal transduction systems (HisKA, Response_reg, GAF, PAS, and HAMP) compared with other freshwater cyanobacteria (Supplementary Table 3, Supplementary Fig. 4). This may be due to the thick gelatinous sheath material that covers *M. aeruginosa* cells and protects them from severe environmental changes.

3.3. Characteristic features of the genes and the genome

3.3.1. Mobile DNA elements ISs are small mobile DNA elements capable of transposition mediated by a self-encoded transposase, and can be classified into various families. ISs have been reported in diverse genera of cyanobacteria, including *Synechocystis* sp. PCC 6803 (73 copies in five families),[22] *Anabaena* sp. PCC 7120 (65 copies in seven families),[12] *Thermosynechococcus elongatus* BP-1 (59 copies in four families),[23] and *Gloeobacter violaceus* PCC 7421 (22 copies in four families)[24] (Fig. 2). A total of 452 copies of ISs were assigned in the *M. aeruginosa* genome that fit the parameters described in Section 2. They appeared to be distributed rather evenly throughout the genome (Fig. 1). The putative ISs could be classified into 35 groups of 13 families on the basis of similarity and the type of transposase (www-is.biotoul.fr) (Fig. 2, Table 1, and Supplementary Table 5). Four types of ISs have been reported in *M. aeruginosa*, PCC 7806.[25] By combining and re-analyzing all the ISs identified in the two strains, we were able to classify the ISs in *Microcystis* into 37 groups. They were designated as ISMae1 to ISMae37, four of which (ISMae1 to ISMae4) have previously been characterized in *M. aeruginosa* PCC 7806 (Table 1). ISMae1 and ISMae4 were common to two *Microcystis* species (10 copies and 33 copies, respectively, in NIES-843), whereas ISMae3 was absent from, and only truncated fragments of ISMae2 were found in the genome of NIES-843.

MITEs are a subset of non-autonomous mobile DNA elements that do not encode a transposase. They can be identified as multiply copy elements of less than 600 bp long flanked by inverted repeats.[26] We searched for MITEs that were present at more than 20 copies in the *M. aeruginosa* genome, and identified 517 copies that were 150–435 bp in length, including partial segments.

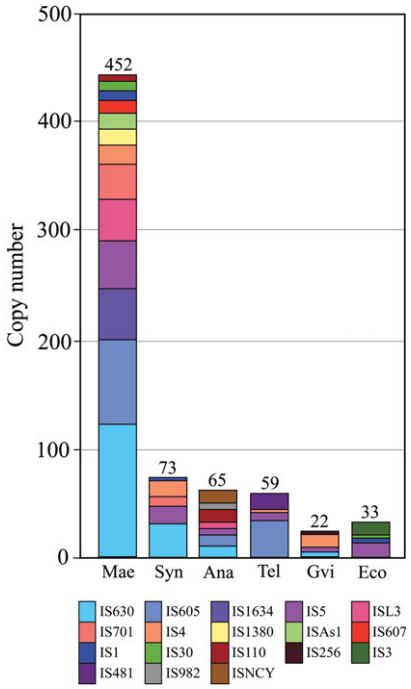

Figure 2. Distribution of IS families in cyanobacteria and *Escherichia coli*. Bars indicate the distribution of known IS families, which are represented by 18 color codes. Mae, *Microcystis aeruginosa* NIES-843; Syn, *Synechocystis* sp. PCC 6803; Ana, *Anabaena* sp. PCC 7120; Tel, *Themosynechococcus elongates* BP-1; Gvi, *Gloeobacter violaceus* PCC 7421; Eco, *Escherichia coli* K-12.

These elements were classified into eight groups on the basis of similarity (MaeMITEa–MaeMITEh), as shown in Fig. 1, Table 1, and Supplementary Table 6.

One of the most distinctive features of the *M. aeruginosa* genome was its extremely high content of mobile DNA elements. A total of 688 kb of the genome (11.8% of the entire genome) were occupied by ISs (583 kb, 10.0% of the genome) and MITEs (105 kb, 1.8% of the genome). Traces of mutations and/or genome rearrangements caused by transposition of these elements have been detected in our preliminary analysis. These observations suggested that there is a high degree of genome plasticity in *M. aeruginosa* due to the large number of mobile repetitive elements. However, the implication as to why the genome of this cyanobacterium is so rich in various mobile elements and how and to what extent they have contributed to the phylogenetic establishment as well as to the genomic characteristic of *M. aeruginosa* is at the moment not clear. Further analysis of these mobile elements with respect to these points remains to be performed.

3.3.2. Genes for synthesis of toxic and bioactive peptides Three non-ribosomal peptide synthetase (NRPS) gene clusters were found in the genome of *M. aeruginosa*: the microsystin synthase gene cluster (*mcyA-J*) at coordinates 3 486 436–3 541 027, the cyanopeptolin synthase gene cluster (*mcnA-C* and *mcnE-G*) at coordinates 5 526 971–5 557 378, and a novel unknown NRPS gene cluster at coordinates 5 202 708–5 219 745.

The microsystin synthase (*mcy*) gene cluster consist of two subclusters that are transcribed in opposite

Table 1. Summary of ISs and MITEs in the *M. aeruginosa* genome

Name	Length (bp)	Family	Copy number
ISMae1	955	IS1	10
ISMae4	1223	IS5	33
ISMae5	1694	IS4	19
ISMae6	1686	IS5	8
ISMae7	519	IS5	3
ISMae8	1093	IS30	8
ISMae9	1370	IS110	7
ISMae10	1330	IS605	49
ISMae11	1473	IS605	9
ISMae12	1835	IS605	6
ISMae13	1295	IS605	5
ISMae14	1366	IS605	4
ISMae15	1431	IS605	3
ISMae16	1379	IS605	2
ISMae17	1510	IS607	9
ISMae18	1240	IS607	2
ISMae19	1410	IS607	2
ISMae20	1897	IS607	4
ISMae21	956	IS630	22
ISMae22	1144	IS630	21
ISMae23	1237	IS630	15
ISMae24	1089	IS630	14
ISMae25	1262	IS630	13
ISMae26	1041	IS630	12
ISMae27	1148	IS630	9
ISMae28	1599	IS630	8
ISMae29	1224	IS630	5
ISMae30	1991	IS1380	16
ISMae31	2168	IS1634	41
ISMae32	896	IS1634	4
ISMae33	653	IS1634	3
ISMae34	1368	IS701	32
ISMae35	1421	ISAs1	16
ISMae36	1424	ISL3	32
ISMae37	921	ISL3	6
MaeMITEa	219	—	240
MaeMITEb	187	—	132
MaeMITEc	182	—	44
MaeMITEd	233	—	30
MaeMITEe	212	—	27
MaeMITEf	429	—	24
MaeMITEg	176	—	21
MaeMITEh	179	—	20

directions: one encodes NRPSs with accessory domains (*mcyA*, *mcyB*, and *mcyC*)[27] and the other encodes PKS (*mcyD*), mixed PKS-NRPS (*mcyE* and *mcyG*) and a racemase (*mcyF*).[28] Two additional genes (*mycI* and *mycJ*) that participate in microcystin synthesis and a gene possibly involved in microcystin transportation (*mcyH*) are located downstream of the *mcyDEFG* subcluster.[29–30] Overall, the primary structure of the *mcy* gene cluster in *M. aeruginosa* was consistent with those reported previously. An intriguing hypothesis is that the ISs and MITEs that were scattered around the *mcy* gene cluster are involved in the transposition of the cluster between individuals, thus contributing to the instability (presence or absence)[31] and genetic diversity of *mcy* genes in natural populations of *M. aeruginosa*.[32,33]

It has been suggested that a gene for thioesterase, *mcyT*, located within the *mcy* gene cluster in another microcystin producing cyanobacterium, *Planktothrix agardhii*, is responsible for the induction of cyanotoxin production.[34,35] In *M. aeruginosa*, a putative gene for the thioesterase superfamiliy protein was found outside of the *mcy* gene cluster (MAE08510, at coordinates 747 362–747 778).

Binding of a 4′-phosphopantethein (4-PPT) to a peptidyl carrier protein is required for microcystin biosynthesis. This binding reaction is catalyzed by the 4-PPT transferase (4-PPTase).[36] A putative 4-PPTase gene (MAE07060) was found in the genome of *M. aeruginosa*. It is highly likely that this gene is involved in microcystin biosynthesis, since this was the only candidate gene for 4-PPT identified in the genome. Of note, the gene showed sequence similarity to that of the 4-PPTase that is involved in nodularin synthesis in *Nodularia spumigena*,[37] and was located distant from the *mcy* gene cluster.

Cyanopeptolin is an another well-known bioactive peptide,[2] and a gene cluster for the synthesis of this peptide contained presumptive genes for four NRPSs (McnE, McnC, McnB, and McnA), a protein with unknown function, a transposase, McnF (ABC transporter-like protein), and McnG (unknown). *Microcystis* cf. *wesenbergii* NIVA-CYA 172/5 has a *mcnD* gene encoding a halogenase that is required for halogenation of cyanopeptolins.[38] However, we did not identify a *mcnD* homolog in the genome of strain NIES-843, which indicates that this strain may produce a non-halogenated variant of cyanopeptolin. As was the case with the *mcy* gene cluster, several genes for transposases were located at the both ends of the cyanopeptolin gene cluster, which may also account for the instability and possible lateral transposition of this gene cluster.

We also identified a series of putative genes for NRPSs that was superficially similar to the cyanopeptolin gene cluster, and a putative polyketide synthase gene cluster (PKS) of unknown function (coordinates 2 508 556–2 513 289). The presence of these gene clusters suggested that additional small polypeptides are produced in *M. aeruginosa*, although the production of polypetides other than microcystins has not been reported in this strain.

3.3.3. Gas vesicle development The gas vesicles allow the *Microcystis* cells to float at the surface of various water environments, position them under the favorable light and oxygen conditions for growth, leading to bloom formation.[3] Thus, the gas vesicles are of taxonomical and ecological importance.

M. aeruginosa NIES-843 harbored the same set of *gvp* genes (*gvpAI*, *AII*, *AIII*, *C*, *N*, *J*, *X*, *K*, *F*, *G*, *V*, and *W* at the coordinates 3 399 358–3 407 222) as those in *M. aeruginosa* PCC 7806 and 9354.[39] The overall physical organization of the *gvp* region in the genome of NIES-843 is also equivalent to those of PCC 7806 and 9354. The major constituent of the gas vesicles is a small hydrophobic protein, GvpA, encoded by *gvp AI–AIII*, which are arranged along ribs that form the cylinder and cones.[40] The amino acid sequence of GvpA is highly conserved and identical even between strains that produce gas vesicles of different width.[41] The amino acid sequence of GvpA in NIES-843 was identical to those of PCC 7806 and BC 8401, except that replacement one amino acid occurred for GvpAI in PCC 7806. Another major component is GvpC, encoded by *gvpC*, a hydrophilic protein of ∼160 amino acid residues, which attaches to the outer surface of the gas vesicle[42] and might affect the gas vesicle width.[43] Four consecutive 33 amino acid residue repeat (33RR), which is highly conserved in some genera of cyanobacteria, were identified in the inferred amino acid sequences of GvpC in NIES-843, as is in PCC 7806.[39] The amino acid sequence of NIES-843 GvpC showed 93% and 95% identity with those of PCC 7806 and BC 8401, respectively, except one missing 33RR in BC 8401.

3.3.4. Genes for two-component regulatory systems Cyanobacteria generally harbor a substantial number of genes that encode elements of two-component regulatory systems, the most simple of which is represented by histidine kinase and a response regulator.[44] For example, 88 genes related to two-component systems have been identified in the genome of *Synechocystis* sp. PCC 6803 (3.95 Mb for the chromosome and the plasmids), and 185 genes in the genome of *Anabaena* sp. PCC 7120 (7.21 Mb).[12,20,45] In contrast, 22 genes for histidine kinases and 23 genes for response regulators were identified in the 5.83 Mb genome of *M. aeruginosa* (Supplementary Table 7 and Supplementary Fig. 5). Five genes for histidine kinases, MAE03210 (*hik34*), MAE14410 (*hik2*), MAE36080 (*hik33*, *dspA*, *nblS*, *dfr*), MAE52650 (*sphS*), MAE60820 (*sasA*), are conserved in all the cyanobacterial genomes sequenced to date, including that of *M. aeruginosa*, which suggests that they have an essential

role in cyanobacteria.[46] In addition to orthologs of previously reported genes for histidine kinases in cyanobacteria, *M. aeruginosa* was unique in that it harbored two genes for histidine kinases (MAE46010 and MAE48940), and two genes for hybrid histidine kinases (MAE21690 and MAE37480).

Five sets of genes for histidine kinases and response regulators were located adjacent to each other, and the remaining genes were scattered throughout the chromosome. Cognate pairs could be deduced for only 11 sets of histidine kinases and response regulators, based on previous studies of orthologous two-component systems in *Synechocystis* sp. PCC 6803 and *Synechococcus elongatus* PCC 6301,[47] as well as the gene organization in *M. aeruginosa* (Supplementary Fig. 5). Pairwise relationships between the remaining components have yet to be clarified.

3.3.5. Genes for the regulation of phosphate uptake A two-component system, Hik7 (SphS) and Rre29 (SphR), regulates the expression of genes that encode proteins for phosphate acquisition, i.e. the ABC-type phosphate transporter, the periplasmic alkaline phosphatase, and the extracellular nuclease, in *Synechocystis* sp. PCC 6803.[47] SphR has been shown to bind to the pho box, PyTTAAPyPy(T/A), which is located upstream of the genes it regulates.[47] *M. aeruginosa* contained several genes that presumably encode proteins involved in phosphate acquisition: three operons that encoded subunits for ABC-type phosphate transporters (MAE18310–MAE18280, MAE18380–MAE18340, and MAE09280–MAE09250), three monocistronic genes for phosphate-binding periplasmic proteins (MAE18390, MAE32380, and MAE38290), and two genes for alkaline phosphatases (MAE50240 and MAE16640). The consensus core sequence of the pho box motif was present upstream of the MAE18310 and MAE18380 operons (Supplementary Fig. 6), which suggested that these operons are regulated by homologs of Hik7 (SphS, a product of MAE52650) and Rre29 (SphR, a product of MAE52640) in *M. aeruginosa*, under phosphate-deprivation conditions. MAE52630, a homolog of *sphU*, which encodes a regulator of the SphS–SphR two-component system, was located downstream of the SphS–SphR genes,[48] as observed in *Anabaena* sp. PCC 7120.[12]

3.3.6. Transcription factors In addition to genes for the response regulators, 43 genes encoded putative transcriptional factors in the *M. aeruginosa* genome (Supplementary Table 8). This number is comparable to other cyanobacteria. Some of them were common to other cyanobacteria and their biological functions have been investigated. These include genes for NtcA (a global regulator for nitrogen assimilation),[49] NtcB (a regulator of nitrate assimilation),[50] RbcR (a regulator of genes for Rubisco),[51] NdhR (a regulator of genes for subunits of NAD dehydrogenase),[52] SyCrp1 (a cAMP receptor protein for cell motility),[53] HrcA (an inhibitor of genes for GroESL),[54] and ZiaR (a regulator of the zinc efflux system).[55] However, the functions of most of the putative transcriptional factors, including 10 genes that were unique to *M. aeruginosa*, have yet to be clarified.

3.3.7. Serine/threonine protein kinases and phosphates The genome of *M. aeruginosa* contained 24 genes for putative serine/threonine protein kinases, including multidomain kinases, and two genes for serine/threonine protein phosphatases (Supplementary Table 9). Although some of the genes for serine/threonine protein kinases were homologous to those characterized in other cyanobacteria, such as *Synechocystis* sp. PCC 6803 and *Anabaena* sp. PCC 7120, the function of the most of these genes in *M. aeruginosa* is unknown. Two genes encoding a presumptive serine/threonine protein phosphatase showed sequence similarity to the gene for PphA in *Synechocystis* sp. PCC 6803 and *Anabaena* sp. PCC 7120. PphA is involved in the dephosphorylation of P_{II}, which regulates nitrogen assimilation.[56] *M. aeruginosa* contained two copies of putative *glnB* genes, which encode P_{II} proteins (MAE59130 and MAE57460), whereas other cyanobacteria contain a single copy of the *glnB* gene, with the exception of *Gloeobacter violaceus* PCC 7421.

3.3.8. Restriction-modification system We searched genes for restriction-modification enzymes (RM) in the *M. aeruginosa* genome using a BLAST search against an RM gene set in the Restriction Enzyme database (REBASE).[57] We identified 62 putative RM genes, including four Type I RMs and 58 Type II RMs. Five solitary restriction enzymes and 27 solitary methyltransferases were assigned (Supplementary Table 10). In addition, we found four potential RM-related loci (MAE05060–MAE05130, MAE10650–MAE10670, MAE60340–MAE60360, and MAE29970–MAE30000) that were disrupted by the insertion of ISs. The recognition sequences of *Nsp*HI (RCATGY) (1 in every 254 kb), *Ava*III (ATGCAT) (1 in every 216 kb), and *Nsp*V (TTCGAA) (1 in every 209 kb) were present at an extremely low frequency in the *M. aeruginosa* genome (Supplementary Table 10).

3.3.9. Genes related to photosynthesis Genes related to photosynthesis are listed in Supplementary Table 11. Complete sets of genes for both photosystem I (PSI) and photosystem II (PSII) were present in the *M. aeruginosa* genome. There were several distinctive features of these genes in *M. aeruginosa*, as follows: (i) a C-terminal partial segment of *psbA* (MAE10410) was found, in addition to five copies of a putative *psbA* gene, which encodes the reaction center D1 complex of PSII; (ii)

a putative *psbN* gene, which encodes a PSII small subunit, was duplicated (MAE36550 and MAE36570) only in *M. aeruginosa*; (iii) seven copies of a putative *ndhD* gene, which encodes NADH dehydrogenase subunit 4, were assigned—the *ndhD5* gene was present in triplicate (MAE23750, MAE23770, and MAE23790), whereas *ndhD1*, *ndhD2*, *ndhD3*, and *ndhD4* were present in a single copy; (iv) in addition to complete sets of putative *cpcA*–*cpcG* (phycocyanin) and *apcA*–*apcF* (allophycocyanin), another gene cluster (MAE51670–MAE51680) that showed sequence similarity to *cpcA*–*cpcB* was present.

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

Funding: This work was supported by the National Institute for Environmental Studies Foundation and the Kazusa DNA Research Institute Foundation.

References

1. Codd, G. A., Lindsay, J., Young, F. M., Morrison, L. F. and Metcalf, J. S. 2005, In: Huisman, J., Matthijs, H. C. P. and Visser, P. M.(eds.), *Harmful Cyanobacteria*, Netherlands: Springer, pp. 1–23.
2. Bister, B., Simone Keller, S., Baumann, H. I., et al.. 2004, Cyanopeptolin 963A, a chymotrypsin inhibitor of *Microcystis* PCC 7806, *J. Nat. Prod.*, **67**, 1755–1757.
3. Holt, J. G., Krieg, N. R., Sneath, P. H. A., Staley, J. T. and Williams, S. T. 1994, In: Hensyl, W. R.(ed.), *Bergey's Manual of Determinative Bacteriology*, 9th Ed., Baltimore: Williams & Wilkins, pp. 377–425.
4. Komárek, J. 1991, A review of water-bloom forming *Microcystis* species, with regard to populations from Japan, *Arch. Hydrobiol. Suppl. Algol. Stud.*, **64**, 115–127.
5. Watanabe, M. 1996, In: Watanabe, M. F., Harada, K., Carmichael, W. W. and Fujiki, H.(eds.), *Toxic Microcystis*, Boca Raton: CRC Press, pp. 13–34.
6. Otsuka, S., Suda, S., Li, R., Matsumoto, S. and Watanabe, M. M. 2000, Morphological variability of colonies of *Microcystis* morphospecies in culture, *J. Gen. Appl. Microbiol.*, **46**, 39–50.
7. Otsuka, S., Suda, S., Li, R., Watanabe, M., Oyaizu, H., Matsumoto, S. and Watanabe, M. M. 1998, 16S rDNA sequences and phylogenetic analyses of *Microcystis* strains with and without phycoerythrin, *FEMS Microbiol. Lett.*, **164**, 119–124.
8. Otsuka, S., Suda, S., Li, R., et al. 1999, Phylogenetic relationships between toxic and non-toxic strains of the genus *Microcystis* based on 16S to 23S internal transcribed spacer sequences, *FEMS Microbiol. Lett.*, **172**, 15–21.
9. Otsuka, S., Suda, S., Li, R., et al. 1999, Characterization of morphospecies and strains of the genus *Microcystis* (Cyanobacteria) for a reconsideration of speceis classification, *Phycol. Res.*, **47**, 189–197.
10. Wayne, L. G., Brenner, D. J., Colwell, R. R., et al. 1987, International Committee on Systematic Bacteriology. Report of the ad hoc committee on reconciliation of approaches to bacterial systematics, *Int. J. Syst. Bacteriol.*, **37**, 463–464.
11. Otsuka, S., Suda, S., Shibata, S., Oyaizu, H., Matsumoto, S. and Watanabe, M. M., 2001, A proposal for the unification of five species of the cyanobacterial genus *Microcystis* Kützing ex Lemmermann 1907 under the Rules of the Bacteriological Code, *Int. J. Syst. Evol. Microbiol.* **51**, 873–879.
12. 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.
13. Lowe, T. M. and Eddy, S. R. 1997, tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence, *Nucleic Acids Res.*, **25**, 955–964.
14. Lagesen, K., Hallin, P., Rødland, E. A., Staerfeldt, H. H., Rognes, T. and Ussery, D. W. 2007, RNAmmer: consistent and rapid annotation of ribosomal RNA genes, *Nucleic Acids Res.*, **35**, 3100–3108.
15. 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.
16. 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.
17. Mulder, N. J., Apweiler, R., Attwood, T. K., et al. 2007, New developments in the InterPro database, *Nucleic Acids Res.*, **35**, D224–D228.
18. Tatusov, R. L., Fedorova, N. D., Jackson, J. D., et al. 2003, The COG database: an updated version includes eukaryotes, *BMC Bioinformatics*, **4**, 41.
19. Bao, Z. and Eddy, S. R. 2002, Automated de novo identification of repeat sequence families in sequenced genomes, *Genome Res.*, **12**, 1269–1276.
20. 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.
21. 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* PCC 6301 involving a highly iterated palindrome (HIP1), *Mol. Microbiol.*, **7**, 189–195.
22. Kaneko, T., Nakamura, Y., Sasamoto, S., et al. 2003, Structural analysis of four large plasmids harboring in a unicellular cyanobacterium, *Synechocystis* sp. PCC 6803*, DNA Res.*, **10**, 221–228.
23. 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.
24. Nakamura, 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.
25. Mlouka, A., Comte, K. and Tandeau de Marsac, N. 2004, Mobile DNA elements in the gas vesicle gene cluster of the planktonic cyanobacteria *Microcystis aeruginosa*, *FEMS Microbiol. Lett.*, **237**, 27–34.

26. Bureau, T. E., Ronald, P. C. and Wessler, S. R., 1996, A computer-based systematic survey reveals the predominance of small inverted-repeat elements in wild-type rice genes. *Proc. Natl. Acad. Sci. USA*, **93**, 8524–8529.
27. Nishizawa, T., Asayama, M., Fujii, K., Harada, K. and Shirai, M., 1999, Genetic analysis of the peptide synthetase genes for a cyclic heptapeptide microcystin in *Microcystis* spp., *J. Biochem.*, **126**, 520–529.
28. Dittmann, E. and Börner, T. 2005, Genetic contributions to the risk assessment of microcystin in the environment, *Toxicol. Appl. Pharmacol.*, **203**, 192–200.
29. Nishizawa, T., Ueda, A., Asayama, M., et al. 2000, Polyketide synthase gene coupled to the peptide synthetase module involved in the biosynthesis of the cyclic heptapeptide microcystin, *J. Biochem.*, **127**, 779–789.
30. Tillett, D., Dittmann, E., Erhard, M., von Döhren, H., Börner, T. and Neilan, B. A. 2000, Structural organization of microcystin biosynthesis in *Microcystis aeruginosa* PCC 7806: an integrated peptide-polyketide synthetase system, *Chem. Biol.*, **7**, 753–764.
31. Tanabe, Y., Kasai, F. and Watanabe, M. M., 2007, Multilocus sequence typing (MLST) reveals high genetic diversity and clonal population structure of the toxic cyanobacterium *Microcystis aeruginosa*, *Microbiology*, **153**, 3695–3703.
32. Mikalsen, B., Boison, G., Skulberg, O. M., et al., 2003, Natural variation in the microcystin synthetase operon mcyABC and impact on microcystin production in *Microcystis* strains, *J. Bacteriol.*, **185**, 2774–2785.
33. Tanabe, Y., Kaya, K. and Watanabe, M. M. 2004, Evidence for recombination in the microcystin synthetase (*mcy*) genes of toxic cyanobacteria *Microcystis* spp., *J. Mol. Evol.*, **58**, 633–641.
34. Schneider, A. and Marahiel, M. A. 1998, Genetic evidence for a role of thioesterase domains, integrated in or associated with peptide synthetases, in non-ribosomal peptide biosynthesis in *Bacillus subtilis*, *Arch. Microbiol.*, **169**, 404–410.
35. Christiansen, G., Fastner, J., Erhard, M., Börner, T. and Dittmann, E. 2003, Microcystin Biosynthesis in Planktothrix: Genes, Evolution, and Manipulation, *J. Bacteriol.*, **185**, 564–572.
36. Mootz, H. D. and Marahiel, M. A. 1997, Biosynthetic systems for nonribosomal peptide antibiotic assembly, *Curr. Opin. Chem. Biol.*, **1**, 543–551.
37. Copp, J. N. and Neilan, B. A. 2006, The phosphopantetheinyl transferase superfamily: phylogenetic analysis and functional implications in cyanobacteria, *Appl. Environ. Microbiol.*, **72**, 2298–2305.
38. Tooming-Klunderud, A., Rohrlack, T., Shalchian-Tabrizi, K., Kristensen, T. and Jakobsen, K. S. 2007, Structural analysis of a non-ribosomal halogenated cyclic peptide and its putative operon from *Microcystis*: implications for evolution of cyanopeptolins, *Microbiology*, **153**, 1382–1393.
39. Mlouka, A., Comte, K., Castets, A-M., Bouchier, C. and Tandeau de Marsac, N. 2004, The gas vescle gene cluster from *Microcystis aeruginosa* and DNA rearrangements that led to loss of cell buoyancy, *J. Bacteriol.*, **186**, 2355–2365.
40. Walsby, A. E. 1994, Gas vesicles, *Microbiol. Rev.*, **58**, 94–144.
41. Beard, S. J., Handley, B. A., Hayes, P. K. and Walsby, A. E. 1999, The diversity of gas vesicle genes in *Planktothrix rubescens* from Lake Zürich, *Microbiology*, **145**, 2757–2768.
42. Walsby, A. E. and Hayes, P. K. 1988, The minor cyanobacterial gas vescle protein, GVPc, is attached to the outer surface of the gas vesicle, *J. Gen. Microbiol.*, **134**, 2647–2657.
43. Dunton, P. G. and Walsby, A. E. 2005, The diameter and critical collapse pressure of gas vesicles in *Microcystis* are correlated with GvpCs of different length, *FEMS Microbiol. Lett.*, **247**, 37–43.
44. Minezaki, Y., Homma, K. and Nishikawa, K. 2005, Genome-wide survey of transcription factors in prokaryotes reveals many bacteria-specific families not found in archaea, *DNA Res.*, **12**, 269–280.
45. 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 Res.*, **3**, 407–414.
46. Ashby, M. K., Houmard, J. and Mullineaux, C. W. 2002, The *ycf27* genes from cyanobacteria and eukaryotic algae: distribution and implications for chloroplast evolution, *FEMS Microbiol. Lett.*, **214**, 25–30.
47. Suzuki, S., Ferjani, A., Suzuki, I. and Murata, N. 2004, The SphS–SphR two component system is the exclusive sensor for the induction of gene expression in response to phosphate limitation in *Synechocystis*, *J. Biol. Chem.*, **279**, 13234–13240.
48. Juntarajumnong, W., Hirani, T. A., Simpson, J. M., Incharoensakdi, A. and Eaton-Rye, J. J. 2007, Phosphate sensing in *Synechocystis* sp. PCC 6803: SphU and the SphS–SphR two-component regulatory system, *Arch. Microbiol.*, **188**, 389–402.
49. Vega-Palas, M. A., Madueno, F., Herrero, A. and Flores, E. 1990, Identification and cloning of a regulatory gene for nitrogen assimilation in the cyanobacterium *Synechococcus* sp. strain PCC 7942, *J. Bacteriol.*, **172**, 643–647.
50. Aichi, M., Takatani, N. and Omata, T. 2001, Role of NtcB in activation of nitrate assimilation genes in the cyanobacterium *Synechocystis* sp. strain PCC 6803, *J. Bacteriol.*, **183**, 5840–5847.
51. Maier, U. G., Fraunholz, M., Zauner, S., Penny, S. and Douglas, S. 2000, A nucleomorph-encoded CbbX and the phylogeny of RuBisCo regulators, *Mol. Biol. Evol.*, **17**, 576–583.
52. McGinn, P. J., Price, G. D., Maleszka, R. and Badger, M. R. 2003, Inorganic carbon limitation and light control the expression of transcripts related to the CO_2-concentrating mechanism in the cyanobacterium *Synechocystis* sp. strain PCC6803, *Plant Physiol.*, **132**, 218–229.
53. Yoshimura, H., Hisabori, T., Yanagisawa, S. and Ohmori, M. 2000, Identification and characterization of a novel cAMP receptor protein in the cyanobacterium *Synechocystis* sp. PCC 6803, *J. Biol. Chem.*, **275**, 6241–6245.
54. Nakamoto, H., Suzuki, M. and Kojima, K. 2003, Targeted inactivation of the *hrcA* repressor gene in cyanobacteria, *FEBS Lett.*, **549**, 57–62.

55. Thelwell, C., Robinson, N. J. and Turner-Cavet, J. S. 1998, An SmtB-like repressor from Synechocystis PCC 6803 regulates a zinc exporter, *Proc. Natl. Acad. Sci. USA*, **95**, 10728–10733.
56. Kloft, N., Rasch, G. and Forchhammer, K. 2005, Protein phosphatase PphA from *Synechocystis* sp. PCC 6803: the physiological framework of PII-P dephosphorylation, *Microbiology*, **151**, 1275–1283.
57. Roberts, R. J., Vincze, T., Posfai, J. and Macelis, D. 2007, REBASE—enzymes and genes for DNA restriction and modification, *Nucleic Acids Res.* **35**, D269–D270.