DNA Research **8**, 205–213 (2001)

Complete Genomic Sequence of the Filamentous Nitrogen-fixing Cyanobacterium *Anabaena* sp. Strain PCC 7120

Takakazu Kaneko,[1] Yasukazu Nakamura,[1] C. Peter Wolk,[2] Tanya Kuritz,[2,†] Shigemi Sasamoto,[1] Akiko Watanabe,[1] Mayumi Iriguchi,[1] Atsuko Ishikawa,[1] Kumiko Kawashima,[1] Takaharu Kimura,[1] Yoshie Kishida,[1] Mitsuyo Kohara,[1] Midori Matsumoto,[1] Ai Matsuno,[1] Akiko Muraki,[1] Naomi Nakazaki,[1] Sayaka Shimpo,[1] Masako Sugimoto,[1] Masaki Takazawa,[1] Manabu Yamada,[1] Miho Yasuda,[1] and Satoshi Tabata[1,*]

Kazusa DNA Research Institute, 1532-3 Yana, Kisarazu, Chiba 292-0812, Japan[1] *and MSU-DOE Plant Research Laboratory, Michigan State University, E. Lansing, MI 48824, USA*[2]

(Received 22 September 2001)

Abstract

The nucleotide sequence of the entire genome of a filamentous cyanobacterium, *Anabaena* sp. strain PCC 7120, was determined. The genome of Anabaena consisted of a single chromosome (6,413,771 bp) and six plasmids, designated pCC7120α (408,101 bp), pCC7120β (186,614 bp), pCC7120γ (101,965 bp), pCC7120δ (55,414 bp), pCC7120ε (40,340 bp), and pCC7120ζ (5,584 bp). The chromosome bears 5368 potential protein-encoding genes, four sets of rRNA genes, 48 tRNA genes representing 42 tRNA species, and 4 genes for small structural RNAs. The predicted products of 45% of the potential protein-encoding genes showed sequence similarity to known and predicted proteins of known function, and 27% to translated products of hypothetical genes. The remaining 28% lacked significant similarity to genes for known and predicted proteins in the public DNA databases. More than 60 genes involved in various processes of heterocyst formation and nitrogen fixation were assigned to the chromosome based on their similarity to the reported genes. One hundred and ninety-five genes coding for components of two-component signal transduction systems, nearly 2.5 times as many as those in *Synechocystis* sp. PCC 6803, were identified on the chromosome. Only 37% of the Anabaena genes showed significant sequence similarity to those of Synechocystis, indicating a high degree of divergence of the gene information between the two cyanobacterial strains.

Key words: *Anabaena* sp. strain PCC 7120; genomic sequencing; heterocyst; nitrogen fixation

1. Introduction

Completion of the genomic sequence of a unicellular cyanobacterium, *Synechocystis* sp. strain PCC 6803 (hereinafter, Synechocystis),[1] in 1996 initiated a great change in the strategy of studying the function and regulation of genes in cyanobacteria. Comprehensive information on the genes in the genome has accelerated the process of identifying and characterizing the genes responsible for various biological phenomena. Such information has allowed adoption of all-encompassing approaches such as systematic gene disruption, transcriptome analysis using an array technology, and proteome analysis with 2-D gels. As a result, much knowledge has accumulated on individual genes that are involved in a variety of biological processes and is facilitating our understanding of the entire genetic system of this organism.

Anabaena sp. strain PCC 7120 (hereinafter Anabaena) is a filamentous, heterocyst-forming cyanobacterium. Heterocysts are metabolically highly active cells that have the capacity for fixation of dinitrogen, an oxygen-sensitive process, in an oxygen-containing environment. Under conditions of nitrogen deficiency, heterocysts differentiate from vegetative cells at semi-regular intervals along the filaments, generating a pattern. Facile techniques for genetic manipulation, including a highly efficient conjugation system, are available for this organism.[2] As a result, Anabaena has long been used to study the genetics and physiology of cellular differentiation, pattern formation, and nitrogen fixation. Numerous genes involved in these processes have been identified.[3] Physical maps of the chromosome and of three plasmids in Anabaena have been reported; the size of the chromo-

Communicated by Mituru Takanami

* To whom correspondence should be addressed. Tel. +81-438-52-3933, Fax. +81-438-52-3934

† Current address: Life Sciences Division, Oak Ridge National Laboratory, MS-6480, Oak Ridge, TN 37830, USA

some was estimated to be 6.42 Mb.[4,5]

Because all cyanobacteria are capable of oxygen-producing photosynthesis, each cyanobacterial species has both common and species-specific characteristics. The most effective approach to finding the genetic complement common to diverse cyanobacteria is to characterize each gene, as was done for Synechocystis, and then to search for orthologs. With this objective and the interesting biology of Anabaena in mind, we sequenced the genome of Anabaena. Here we describe the complete sequence of the chromosome and the six plasmids of Anabaena, and discuss characteristic features of its genes.

2. Materials and Methods

2.1. Bacterial strain and construction of a cloning vector

Anabaena sp. PCC 7120 was obtained from R. Haselkorn (University of Chicago).

To facilitate the construction of low-copy-number BAC vector pRL838 (GenBank AF403425), a derivative of pBAC108L,[6] an intermediate and pUC19 were fused at their unique *Bsp*LU11I sites, permitting modification of the BAC vector as a high-copy-number plasmid. The BAC vector was later freed from pUC19 by use of *Bsp*LU11I. pRL838 was originally designed to clone mapped *Anabaena* PFGE restriction fragments and to use the clones to complement mutations. To take advantage of the features bracketing the cloning region of pBAC108L, the *Sal* I and *Sst* I sites were relocated to the polylinker. The final polylinker contains unique cloning sites that are compatible with the ends generated by enzymes (see below) producing mapped sites and large restriction fragments. Inclusion of *oriT* (RK2) permits conjugal complementation. *erm* inserted downstream from *cat* provides selection in Anabaena, and homologous recombination permits replication in Anabaena.

2.2. Sequencing strategy and data assembly

The whole-genome shotgun strategy combined with the "bridging shotgun" method[7] was adopted to determine the structure of the entire genome. Four shotgun libraries with three types of cloning vectors were generated from the total cellular DNA of Anabaena to minimize cloning bias. Libraries ANE and ANB bore inserts of approximately 1.0 kb (element clones) and 2.6 kb (bridging clones), respectively, derived from sonication and cloning in M13mp18. Library ANP (plasmid clones) bore clones of approximately 8.5 kb in pUC18, and library ANC (BAC clones) bore inserts of approximately 17 kb cloned in BAC vector pRL838.

One strand of the element clones and both strands of the clones from the other three libraries were sequenced using the Dye-terminator Cycle Sequencing kit with DNA sequencers type 377XL (Applied Biosystems, USA) according to the protocol recommended by the manufacturer. A total of 65,926 sequence files corresponding to about 5.5 genome equivalents were accumulated and assembled using the Phrap program (Philip Green, Univ. of Washington, Seattle, USA). The end-sequence data from the bridge, plasmid and BAC clones facilitated the gap-closure process as well as accurate reconstruction of the entire genome. The final gaps in the sequence were filled either by primer walking or by PCR amplification of the gap regions followed by shotgun sequencing of the products. Sequences were confirmed either by obtaining (as a minimum) sequences from both strands or sequences from the same strand using multiple clones as templates. The integrity of the reconstructed genome sequence was assessed by walking through the entire genome with the end sequences of plasmid and BAC clones.

2.3. Gene assignment and annotation

Coding regions were assigned by a combination of computer prediction and similarity search. Glimmer 2.02, a computer program based on interpolated Markov models,[8] was used to predict protein-encoding regions. Prior to prediction, a matrix was generated for the *Anabaena* genome by training with a dataset of 3426 open reading frames (ORFs) that showed a high degree of predicted amino-acid sequence similarity to known and predicted proteins of known function. All of the predicted protein-encoding regions equal to or longer than 90 bp were translated into amino-acid sequences, which were subjected to similarity search against the non-redundant protein database (nr-database) with the BLASTP program.[9] If two predicted genes overlapped on either strand, those showing similarity to known genes were preferentially taken, and the longer one was chosen unless the functionality of the shorter one was reasonably anticipated. In parallel, the entire genomic sequence was compared with those in the nr-protein database using the BLASTX program to identify genes that had escaped from prediction and/or were smaller than 90 bp, especially in the predicted intergenic regions. For predicted genes that did not show sequence similarity to known genes, only those equal to or longer than 150 bp were considered further.

Functions were assigned to the predicted genes on the basis of the similarity of their predicted products to the products of genes of known function. For genes that encode proteins of 100 amino acid residues or more, a BLAST score of 10^{-20} was considered significant. Genes with a higher E-value were taken into consideration for genes encoding smaller proteins.

Genes for structural RNAs were assigned by similarity search against the in-house structural RNA database that had been generated based on the data in GenBank (rel. 124.0). tRNA-encoding regions were predicted by use of the tRNA scan-SE 1.21 program[10] in combination with

Table 1. The size and the average GC-content of each replicon in *Anabaena* sp. PCC 7120.

	Length (bp)	Avarage GC content (%)
Chromosome	6,413,771	41.3
pCC7120α	408,101	40.5
pCC7120β	186,614	40.2
pCC7120γ	101,965	41.0
pCC7120δ	55,414	41.6
pCC7120ε	40,340	40.9
pCC7120ζ	5,584	44.2
Total	7,211,789	

the similarity search. It should be emphasized that the prediction of protein- and RNA-encoding genes in this study represent merely theoretical potential, and require experimental validation.

The complement of genes in the Synechocystis and Anabaena genomes were compared by taking three independent factors into consideration: a bit score, the BLAST2 E-value, and the ratio of an alignment length. We first used the BLAST2 algorithm[9] to identify BLASTP alignments that were determined to have an E-value not greater than 10^{-10}. A lower threshold of acceptability was set at one-third of the bit score reported by self-comparison of the translated amino acid sequences. Only amino acid sequences whose alignments extended over at least 0.6 times the length of the query sequence were considered similar. With lower stringency, two protein-encoding genes were considered similar if the BLAST2 score was lesser than 10^{-4}.

3. Results and Discussion

3.1. Sequence determination of the entire genome

The nucleotide sequence of the entire genome of Anabaena was deduced initially by assembly of a total of 65,926 sequence files, which correspond to approximately 5.5 genome equivalents, according to the modified whole genome shotgun method described in Materials and Methods. To ensure sufficient accuracy for further analysis of gene structure and function, additional sequencing was carried out to obtain the sequences from both strands or from the same strand of multiple template clones. The integrity of the final sequence was assessed by comparing the insert length of each pUC and BAC clone with the computed distance between the end sequences of the clone. The genome of Anabaena consisted of a single chromosome and six plasmids designated pCC7120α, β, γ, δ, ε, and ζ. The total size of the genome is 7,211,789 bp, and the size and the GC content of each replicon are shown in Table 1. The nucleotide position was numbered from an *Avr* II restriction site in the chromosome (Fig. 1) and pCC7120γ, *Sal* I in pCC7120α and β, and *Sna*BI in pCC7120δ, ε, and ζ (see Fig. 1 in the Supplement section).

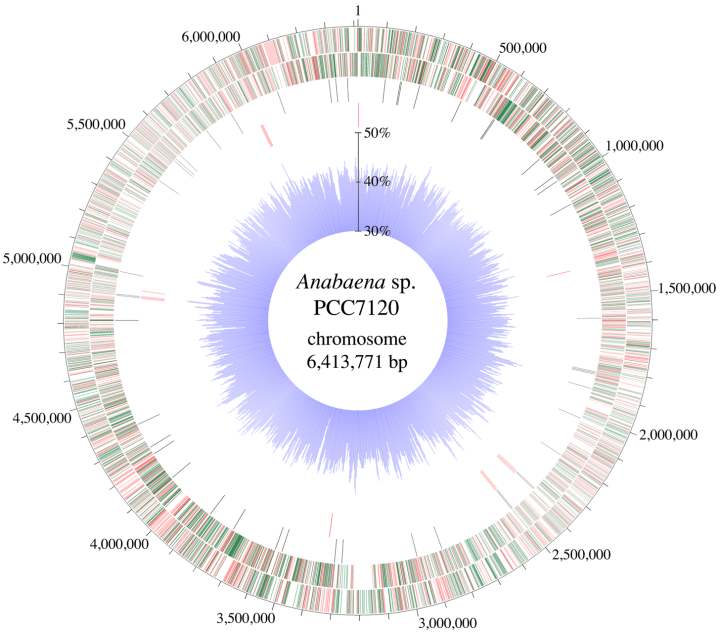

Figure 1. Circular representation of the chromosome of *Anabaena* sp. PCC 7120. The scale indicates the location in bp starting from an *Avr* II restriction site. The bars in the outermost and the second circles show the positions of the putative protein-encoding genes in the clockwise and counter-clockwise directions, respectively. Genes whose functions could be deduced by sequence similarity to genes of known functions are depicted in green, and those whose functions could not be deduced are in red. The bars in the third circle indicate the positions of predicted tRNA genes and those in the fourth circle indicate the positions of genes for structural RNAs including rRNAs and small RNAs. The innermost circle with a scale shows the average GC percent calculated by a window size of 10 kb.

3.2. Consistency with the physical map

Pulsed homogeneous orthogonal field gel electrophoresis (PHOGE)[11] analysis identified and mapped chromosomal sites in Anabaena for *Avr* II (23, plus one in pCC7120γ), *Sal* I (26, plus one in pCC7120α and two in β), *Pst* I (9), *Sph* I (3, plus one each in pCC7120α and β), and *Sst* II (5). The sizes of 7 *Aha* II, 13 *Asu* II, 1 *Fsp* I, 8 *Nco* I, 6 *Spl* I, and 5 *Stu* I fragments, plus plasmids pCC7120α, pCC7120β and pCC7120γ and three of smaller size, were measured.[4,5,12] The means and standard errors of the absolute value of the deviation of predicted lengths, positions of restriction endonuclease sites, and positions of genes (site of TLTC transposition excluded) from those determined by sequencing were 5.8±6.2 kb (n = 100), 14.1±12.2 kb (n = 66), and 17.4±12.3 kb (n = 19), respectively, illustrating the high resolution of PHOGE.[11]

Mapping was not error free; AvrU and the 767- and (obscured by plasmid pCC7120α) 412-kb *Spl* I fragments were not observed by PFGE. In double digests, *Sph* I and *Sst* II sites at 5.946–5.947 Mb were obscured by

nearby *Avr* II and *Pst* I sites, and produced unrecognized PFGE doublets with *Sal* I. We cannot account for *hetR* and *rrnC* having been mapped to the wrong restriction fragments, for the TLTC mutation having been mapped 124 kb from its correct position, and for fragments AvrR and SalU2 not having been found by sequencing.

Modified, overlapping *Hae* III sites[13–15] account for 5 *Avr* II sites and 6 *Sst* II sites missed by restriction, and for fragments NcoC1 and NcoE; an overlapping *Ava* II site[16] accounts for NcoD2; and modified *Pvu* II sites[13] overlap *Pst* I sites at nt 1,141,556 and 4,725,827 that were unobserved by restriction analysis. Twelve of the 30 *Avr* II sites are within insertion sequence IS*1594*. The site at nt 1,288,228, splitting fragment AvrI of Bancroft et al.,[4] evidently arose by transposition after 1992.

Most of the 48 genes mapped in the chromosome (*nucA* in pCC7120α:[17]) were mapped to the correct restriction fragment, or when more detailed mapping was attempted, more precisely (see above). However, Synechococcus and Synechocystis probes localized *argE*, *narA*, *narC*, *psaD1*, *psaD3*, *psaE2*, *psbD1*, and *psbD2* erroneously, indicating that mapping by heterologous hybridization entailed risk.

3.3. Assignment of protein- and RNA-encoding genes

The potential protein-encoding regions were assigned by a combination of computer prediction by the Glimmer program and similarity search, as described in Materials and Methods. Glimmer predicted a total of 6228 potential protein-encoding genes on the chromosome after training with a dataset of sequences of highly probable protein-encoding genes. By taking the sequence similarity to known genes and the relative positions into account to avoid overlaps, the total number of potential protein-encoding genes finally assigned to the chromosome was 5368 (Fig. 1). The average gene density was one gene in every 1195 bp. Six plasmids, pCC7120α, β, γ, δ, ε, and ζ, had the capacity of coding for 386, 186, 90, 66, 31, and 5 proteins, respectively, when estimated by the same procedure. The putative protein-encoding genes thus assigned to the genome starting with either an ATG, GTG, TTG, or ATT codon are denoted by serial number with three letters representing the species name (a), ORF longer than or less than 100 codons (l or s), and the reading direction on the circular map (r or l) (Fig. 1).

Four copies of rRNA gene clusters, designated as *rrnA* to *D*, were identified on the genome in the order of 16S-23S-5S at coordinates of 2,375,734–2,382,211, 2,500,525–2,505,531, 4,919,771–4,914,765, and 5,947,188–5,942,409, respectively, by sequence similarity to known bacterial rRNA genes[18–20] (Fig. 1 of the Supplement section). The 23S RNA gene in the *rrnA* cluster was disrupted by insertion of an IS element (IS*1594*). Two tRNA genes, *trnI* and *trnA*, were located between the 16S and 25S rRNA genes except for *rrnD*, where no tRNA gene was identified.

A total of 48 tRNA genes representing 42 tRNA species, which are sufficient to bind all the codon species, were assigned on the chromosome by sequence similarity to known bacterial tRNA genes and computer prediction using the tRNA scan-SE program (Fig. 1; and Table 1, Table 2, Fig. 1, and Fig. 2 of the Supplement section). These genes were spread on the chromosome and were likely to be transcribed as single units, except for those in the rRNA gene clusters and 3 genes, *trnY*-GUA(1) - *trnT*-CGU - *trnG*-CCC in tandem array at coordinates 1,819,470–1,819,848 (Fig. 1). A group I intron was found in *trnL*-UAA, as was reported by Xu et al.[21] A tRNA gene cluster was found in the genome of pCC7120δ at coordinates 49,998 to 52,163 (see Fig. 1 of the Supplement section). Nineteen tRNA genes and 3 pseudogenes form a tandem array in that 2.2-kb region at distances of 3 bp to 124 bp on the same strand of DNA. None of the genes shows sequence similarity to the reported tRNA genes in any species. Whether or not these genes are functional or even transcribed remains to be clarified. Other small RNA-encoding genes showing sequence similarity to 10Sa RNA (*ssrA*),[22] 7SL RNA (*ffs*),[23] 6Sa RNA (*ssaA*),[24] and RNase P subunit B (*rnpB*)[25] were identified on the chromosome.

3.4. Functional assignment of the protein-encoding genes

Similarity search of the 5368 potential protein-encoding genes in the chromosome against the nr databases indicated that 2396 (45%) were homologues of genes of known function, 1453 (27%) showed similarity to hypothetical genes, and the remaining 1519 (28%) showed no significant similarity to any registered genes (Table 2 and Fig. 1). The six plasmid genomes contained a larger percentage of genes of unknown function, 53%, 64%, 64%, 74%, 74%, and 80% for pCC7120α, β, γ, δ, ε, and ζ, respectively.

The potential protein-encoding genes whose function could be anticipated were grouped into 14 categories with respect to different biological roles, according to the principle of Riley.[26] The numbers of genes in each category are summarized in Table 2, and the assignment of each gene is listed in CyanoBase at http://www.kazusa.or.jp/cyanobase/. On the gene map of the Supplement section (Fig. 1), the location, length and direction of these genes are indicated, with color codes corresponding to functional categories.

3.5. Characteristic features of predicted genes and the genome

3.5.1. Similarity with Synechocystis genes

A total of 3167 potential protein-encoding genes have been identified in the 3.6-Mb genome of the unicellular cyanobacterium *Synechocystis* sp. PCC 6803.[1] The

Table 2. Features of the assigned protein-coding genes and the functional classification.

	Chromosome		alpha		beta		gamma		delta		epsilon		zeta		entire genome	
		%		%		%		%		%		%		%		%
Amino acid biosynthesis	111	2.1	0	0.0	1	0.5	1	1.1	0	0.0	0	0.0	0	0.0	113	1.8
Biosynthesis of cofactors, prosthetic groups, and carriers	152	2.8	2	0.5	3	1.6	0	0.0	0	0.0	0	0.0	0	0.0	157	2.6
Cell envelope	80	1.5	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	80	1.3
Cellular processes	94	1.8	4	1.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	98	1.6
Central intermediary metabolism	70	1.3	1	0.3	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	71	1.2
Energy metabolism	98	1.8	2	0.5	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	100	1.6
Fatty acid, phospholipid and sterol metabolism	41	0.8	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	41	0.7
Photosynthesis and respiration	156	2.9	0	0.0	0	0.0	0	0.0	1	1.5	0	0.0	0	0.0	157	2.6
Purines, pyrimidines, nucleosides, and nucleotides	57	1.1	1	0.3	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	58	0.9
Regulatory functions	339	6.3	6	1.6	13	7.0	2	2.2	3	4.5	1	3.2	0	0.0	364	5.9
DNA replication, recombination, and repair	93	1.7	7	1.8	5	2.7	1	1.1	0	0.0	0	0.0	0	0.0	106	1.7
Transcription	38	0.7	1	0.3	2	1.1	0	0.0	0	0.0	0	0.0	0	0.0	41	0.7
Translation	189	3.5	4	1.0	3	1.6	3	3.3	1	1.5	2	6.5	0	0.0	202	3.3
Transport and binding proteins	294	5.5	7	1.8	9	4.8	3	3.3	0	0.0	0	0.0	0	0.0	313	5.1
Other categories	584	10.9	71	18.4	5	2.7	15	16.7	8	12.1	4	12.9	0	0.0	687	11.2
Subtotal of genes similar to genes of known function	**2396**	**44.6**	**106**	**27.5**	**41**	**22.0**	**25**	**27.8**	**13**	**19.7**	**7**	**22.6**	**0**	**0.0**	**2588**	**42.2**
Similar hypothetical proteins	1453	27.1	77	19.9	26	14.0	7	7.8	4	6.1	1	3.2	1	20.0	1569	25.6
Subtotal of genes similar to registered genes	**3849**	**71.7**	**183**	**47.4**	**67**	**36.0**	**32**	**35.6**	**17**	**25.8**	**8**	**25.8**	**1**	**20.0**	**4157**	**67.8**
No similarity	1519	28.3	203	52.6	119	64.0	58	64.4	49	74.2	23	74.2	4	80.0	1975	32.2
Total	**5368**	**100.0**	**386**	**100.0**	**186**	**100.0**	**90**	**100.0**	**66**	**100.0**	**31**	**100.0**	**5**	**100.0**	**6132**	**100.0**

Genes assigned in the chromosome and in the 6 plasmids are classified according to the similarity of their products to the products of genes of known and unknown function.

genetic complements of Synechocystis and Anabaena were compared as described in Materials and Methods. Comparison with higher stringency indicated that 2,012 *Anabaena* genes, 37% of the 5368 potential protein-encoding genes, have matched genes in the Synechocystis genome. With lower stringency, where most members of a given gene family can be grouped into a cluster, 3,600 *Anabaena* genes (67%) showed significant sequence similarity to Synechocystis genes. When reverse comparison was made, 1,322 (42%) and 2,603 (82%) of the 3167 Synechocystis genes showed sequence similarity to Anabaena genes under the defined higher and lower stringency conditions, respectively. The results indicate that significant portions of the gene components in the genomes are unique to the species.

3.5.2. Genes for heterocyst formation and nitrogen fixation

Genes known to be involved in the positioning of heterocyst formation, *patA* (Accession no. M87501) (*all0521*), *patS* (Accession no. AF046871) (*asl2301*), and *patN* (Accession no. AF288131) (*alr4812*); in the initiation of heterocyst differentiation, *ntcA* (Accession no. X71608) (*alr4392*), *hetR* (Accession no. M37779) (*alr2339*), *hetN* (Accession no. L22883) (*alr5358*), and *hetF* (Accession no. AF288130) (*alr3546*); and in the transition to a nondividing state, *hetC* (Accession no. U55386) (*alr2817*) and probably *hetP* (Accession no. L26915) (*alr2818*), were assigned by comparison with genes reported in Anabaena or in the related filamentous cyanobacterium, *Nostoc punctiforme*. Sixty amino acid residues of the predicted products of three genes of unknown function (*asl1930*, *alr2902*, and *alr3234*) showed similarity to an N-terminal region of the predicted product of *hetP*. Genes for the synthesis of the heterocyst envelope were identified by similarity with reported Anabaena and Nostoc genes. The gene cluster *hglE* (*alr5351*)–[2 unknown genes]–*hglD* (*alr5354*)–*hglC* (*alr5355*)–[1 unknown gene]–*hetM* (*alr5357*)–*hetN* (*alr5358*)–*hetI* (*all5359*) (Accession no. AF016890, U13677, L22883); *hglK* (Accession no. U13768) (*all0813*); and *devBCA* (Accession no. X99672) (*alr3710–alr3711–alr3712*), all of which are localized in the chromosome, play a role in the synthesis and deposition of the glycolipid layer of the heterocyst envelope. A second copy of *hglE* (*all1646*), whose predicted product showed 51% amino acid identity with that of *hglE*; three additional copies of *devBCA* (*alr3647–3649*, *alr4280–4282*, *alr4973–4975*); and a fifth copy of *devBC* (*all5347–5346*), whose predicted products showed 36% to 74% amino acid identity with the previously reported paralogs, were newly identified in this study. Chromosomal genes *hepCA* (Accession no. AF031959) (*alr2834–2835*), *hepB* (Accession no. U68035) (*alr3698*), *hepK* (Accession no. U68034) (*all4496*) and *devR* (Accession no. L44605) (*alr0442*) are involved in the synthesis of the heterocyst envelope polysaccharide (see also section 3.5.7).

Two series of chromosomal genes that are involved in nitrogen fixation, *nifVZT* (Accession no. AJ239033) (*alr1407–asr1409*) and *nifB–fdxN–nifSUHDK*–[1 unknown gene]–*nifENX*–[2 unknown genes]–*nifW–hesAB–fdxH* (Accession nos. J05111, V01482, U47055, X15553, X13522) (*all1517–1516*, *all1457–1454*, *all1440–1430*), are further subdivided by the presence of the *fdxN*- and *nifD*-localized excision elements in unrearranged

vegetative-cell chromosomes (see section 3.5.9). The predicted products of the following genes were found to show a high degree of sequence similarity (27% to 62% amino acid identity) to certain of those genes: *nifV2* (Accession no. AJ239032) (*alr2968*), 3 *nifS*-like genes (*alr2505*, *alr3088*, *alr2495*), and a *nifX*-like gene (*all2531*). Two genes that may be related to a NifJ-flavodoxin system, a *nifJ* homologue (*alr1911*) and a flavodoxin-like gene (*alr2405*), were identified in addition to the previously reported *nifJ* (Accession no. L14925) (*alr2803*).

3.5.3. Genes for two-component regulatory systems

A total of 195 genes coding for components of two-component signal transduction systems were identified on the chromosome, including 71 genes for sensory kinases, 71 for response regulators, and 53 for hybrids of a sensory kinase and a response regulator. Of these, 85 form 36 gene clusters containing 2 to 5 genes, and 110 genes are present individually. The sole 5-gene cluster, *alr3155–3159*, contains, in order, genes putatively encoding a sensory kinase, a response regulator, a phytochrome-like sensory kinase (*aphA*, *alr3157*), a response regulator, and a sensor/regulator hybrid protein. Thirteen genes for sensory kinases bear a conserved domain for a Serine/Threonine kinase. Synechocystis has no such gene,[27] suggesting that the developmental capacity of Anabaena may necessitate a more complex signal transduction system. Genes for two-component systems were also found in pCC7120α, β, δ, and ε: two genes for sensory kinases paired with response regulators, and 4 additional genes for response regulators.

3.5.4. Genes for proteins containing inteins

Inteins are defined as intervening protein sequences that are excised during a protein splicing process. In Synechocystis, 4 intein sequences have been reported: in DnaB (DNA helicase), DnaX (a subunit of DNA polymerase III), GyrB (DNA gyrase B subunit), and DnaE (the α subunit of DNA polymerase III).[28,29] There are 2 copies of *dnaB* in the genome of Anabaena, one (*all0578*) on the chromosome and the other (*all7274*) on plasmid pCC7120α. *all0578* contains a sequence corresponding to an intein of 429 amino acid residues at the position of Glycine388, whereas the copy on the plasmid has no intein. The intein in DnaE seems to be, as in Synechocystis, a split intein capable of protein trans-splicing.[30] Two split *dnaE* genes, *all3578* and *alr1054*, located more than 3 Mb apart and in reverse orientation in the chromosome, presumptively have the capacity to encode the N-terminal and C-terminal portions of the DnaE protein, respectively. One hundred and two amino acid residues at the C-terminus of the translated gene product of *all3578* and 36 amino acid residues at the N-terminus of the *alr1054* gene product show sequence similarity to the inteins of the Synechocystis *dnaE* genes, and may themselves be inteins that allow protein trans-splicing.

3.5.5. Genes for proteins with WD-repeats

WD-repeat-containing proteins may be defined as those with 4 or more copies of a repeating unit carrying a motif of approximately 31 amino acids containing Trp-Asp (WD). Originally reported as regulatory proteins in eukaryotes,[31] proteins with WD-repeats have been identified in the genomes of a variety of eukaryotic species including *Arabidopsis thaliana* (59 genes), *Caenorhabditis elegans* (88 genes), and *Saccharomyces cerevisiae* (58 genes) (http://bmerc-www.bu.edu/wdrepeat/). In prokaryotes, on the other hand, only Synechocystis (5 genes)[1] and *Thermomonospora curvata* (1 gene)[32] are known to have such genes. In the Anabaena genome, 20 and 4 genes for proteins containing 4 to 15 WD-repeats were identified on the chromosome and the plasmids, respectively. The putative products of 5 of the genes were composed of the repeating unit for almost their entire lengths. Those of the remaining 19 genes containing the repeating units in their C-terminal portions bear stretches of 262 to 1128 amino acid residues at their N-termini. The N-terminal stretches of the putative products of three genes, *all0438*, *alr3119*, and *all3169*, harbored conserved regions of Serine/Threonine kinases, and those of three other genes, *alr0029*, *alr2800*, and *alr7129*, showed sequence similarity to portions of products of plant disease-resistance genes, MLA6 in barley (Accession no. AJ302292), bacterial blight-resistance protein Xa1 in rice (Accession no. AB002266), and RPM1 in *A. thaliana* (Accession no. X87851). These results may suggest that Anabaena shares a signaling pathway with eukaryotes.[33]

3.5.6. Genes for circadian rhythms

Circadian rhythms in cyanobacteria have been studied intensively in *Synechococcus* sp. PCC 7942 using luciferase as a reporter, and related genes have been isolated and characterized.[34] These include *kaiABC* as the major genetic elements of the circadian clock (Accession no. AB010691), *cikA* and *pex* as encoding components of input pathways (Accession nos. AF258464 and AB009574), proteins encoded by *rpoD2* and *cpmA* as output modifiers (Accession nos. AB006910 and AF117208), and *sasA* as an activator of *kaiBC* expression (Accession no. D14056). Presumptive counterparts of all of these genes have been identified in the Anabaena genome: *alr2884* (*kaiA*), *alr2885* and *all3328* (*kaiB*), *alr2886* (*kaiC*), *all1688* (*cikA*), *alr3979* (*pex*), *alr3800* (*rpoD2* = *sigE*), *alr3885* (*cpmA*), and *all3600* (*sasA*).

3.5.7. Clusters of genes putatively encoding glycosyltransferase-like proteins

Eighty-four genes for glycosyltransferase-like proteins have been identified on the chromosome. Sixty-three of them formed 17 clusters consisting of 2 to 17 genes organized in tandem arrays. Genes involved in polysaccharide synthesis related to heterocyst differentiation (section 3.5.2) were found in 3 gene clusters. *hepC* and *hepA* are in cluster *alr2830–alr2840* and *hepB* is in a two-gene cluster, *alr3698–3699*. Mutation of the lipopolysaccharide-biosynthetic genes *rfbP* and *rfbZ*, within the cluster *all4827–4830* leads not only to presumptive changes in the walls of vegetative cells but also to heterocysts that are defective in aerobic nitrogen fixation.[35] These results suggest that other genes in these clusters will also prove to be involved in the synthesis of polysaccharides that envelop either heterocysts or the cells that differentiate into heterocysts.

3.5.8. Transposases

A total of 145 genes were identified as presumptively encoding transposases. Eighty-six were present in the chromosome and the remaining 59 in the plasmids. Among the plasmids, pCC7120α contained transposase genes at the highest density: 44 (11.4%) of 386 genes assigned in this plasmid are transposase genes. Of the 145 such genes in the genome, 102 were located within DNA sequences that exhibit structures characteristic of IS-like elements, i.e., with inverted repeats and/or duplications at both termini. Predicted products of the transposase genes could be classified into 23 groups based on their sequence similarity to known transposases. The largest group comprised 25 members which were further divided, on the basis of the structure of the flanking regions, into a 6-member subgroup of the IS*891* family originally found in *Anabaena* sp. M-131 and PCC 7120,[36] and a 5-member subgroup identified in this study and denoted IS*An5*. The translated amino acid sequence of each member of these groups differs significantly from the others, and shows 25% to 100% identity to that of the original IS*891* transposase (Accession no. M24855).

3.5.9. Developmentally regulated genome rearrangement

Genomic rearrangements take place late during heterocyst differentiation in Anabaena. DNA elements are excised from the *fdxN*, *nifD*, and *hupL* genes by recombination, enabling these genes to be expressed.[37] The recombinase genes located in the elements are essential for excision. The *fdxN*, *nifD*, and *hupL* elements are present in vegetative cells from nt 1,716,797–1,776,224 (59,428 bp), 1,700,623–1,711,900 (11,278 bp) and 785,538–794,956 (9419 bp), respectively, of the chromosome, and encode 57 (*alr1459–all1515*), 12 (*alr1442–asr1453*), and 10 (*alr0677–all0686*) proteins including recombinases, respectively.

3.5.10. Plasmid genes

More than 100 genes were cloned and sequenced in Anabaena prior to our study, three of which we can now assign to plasmids. Adenine-specific DNA methyltransferase (Accession no. AF220506), *all7280*, is located on pCC7120α. ζ-Carotene desaturase (Accession no. D26095), *all7255*, which converts ζ-carotene to lycopene,[38] is located on pCC7120α; another copy of this gene was identified on the chromosome (*all2382*). Three of the 11 putative genes for σ factors involved in the initiation of transcription were assigned to the plasmids. These include previously reported *sigB* (Accession no. M95760) (*all7615*), and 2 newly identified genes, *sigB3* (*all7608*) and *sigB4* (*all7179*).

Plasmid-localized genes have the following features that merit comment.

1. Genes encoding the sugar non-specific nuclease, *nucA* (Accession no. X64706), *all7362*, and its corresponding inhibitor, *nuiA* (Accession no. X77568), *alr7361*, were shown to be present on pCC7120α.[17,39] We have found two additional pairs of paralogous genes in the plasmids. One pair, *alr7261–all7262*, is also on pCC7120α. A second pair, *alr8011–all8013*, whose putative *nucA* gene is split into two ORFs (*alr8011* and *asr8012*) by a stop codon, is on pCC7120γ.

2. Genes for site-specific recombinases were assigned in pCC7120α (*alr7076* and *alr7203*), β (*alr7511*), γ (*alr8001*), δ (*all8545*), and ε (*all9019*). The sequences of these genes showed amino acid identities of 24% to 96%.

3. Homologues of *parAB* genes, whose products would be expected to be involved in the partitioning of plasmids during cell duplication,[40] were presumptively identified in plasmids pCC7120α (*alr7082* and *alr7083*), pCC7120β (*alr7581* and *alr7582*), pCC7120γ (*alr8006* and *alr8007*), and pCC7120ε (*alr9026* and *alr9027*). Only a presumptive *parA* gene (*alr8562*) was found on pCC7120δ.

4. pCC7120β contains a cluster of genes presumptively encoding proteins related to DNA replication: DNA polymerase III β subunit (*alr7569*), DNA polymerase III γ and τ subunits (*alr7570*), DNA polymerase III δ' subunit (*alr7575*), and single-strand DNA binding protein (*alr7579*).

5. Six genes, *all7592*, *all7610*, *all7618*, *alr7622*, *all7631*, and *alr7635* that are presumptively involved in cation transport are on pCC7120β.

6. pCC7120γ bears a cluster of 3 genes (*all8088*, *all8089*, and *all8090*) that presumptively encode an ABC phosphonate transporter.

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, divided into 26 entries. The accession numbers are as follows: AP003581 (nucleotide positions 1–348,050), AP003582 (348,001–690,650), AP003583 (690,601–1,030,250), AP003584 (1,030,251–1,378,550), AP003585 (1,378,501–1,720,550), AP003586 (1,720,501–2,069,550), AP003587 (2,069,501–2,413,050), AP003588 (2,413,001–2,747,520), AP003589 (2,747,471–3,089,350), AP003590 (3,089,301–3,422,800), AP003591 (3,422,751–3,770,150), AP003592 (3,770,101–4,118,350), AP003593 (4,118,301–4,451,850), AP003594 (4,451,801–4,795,050), AP003595 (4,795,001–5,142,550), AP003596 (5,142,501–5,491,050), AP003597 (5,491,001–5,833,850), AP003598 (5,833,801–6,176,600), and AP003599 (6,176,551–6,413,771) for the chromosome; AP003600 (1–341,950) and AP003601 (341,901–408,101) for pCC7120α; AP003602 (1–186,614) for pCC7120β; AP003603 (1–101,965) for pCC7120γ; AP003604 (1–55,414) for pCC7120δ; AP003605 (1–40,340) for pCC7120ε; AP003606 (1–5,584) for pCC7120ζ.

Acknowledgements: This work was supported by the Kazusa DNA Research Institute Foundation and by the U.S. Department of Energy under grant DOE-FG02-91ER20021.

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**, 109–136.
2. Wolk, C. P., Ernst, A., and Elhai, J. 1994, In: Bryant, D. A. (ed) The molecular biology of cyanobacteria, Kluwer Academic Publishers, Dordrecht, pp. 769–823.
3. Wolk, C. P. 2000, In: Brun, Y. V. & Shimkets, L. J. (eds) Prokaryotic development, American Society for Microbiology, Washington DC, pp. 83–104.
4. Bancroft, I., Wolk, C. P., and Oren, E. V. 1989, Physical and genetic maps of the genome of the heterocyst-forming cyanobacterium *Anabaena* sp. strain PCC 7120., *J. Bacteriol.*, **171**, 5940–5948.
5. Kuritz, T., Ernst, A., Black, T. A., and Wolk, C. P. 1993, High-resolution mapping of genetic loci of *Anabaena* PCC 7120 required for photosynthesis and nitrogen fixation, *Mol. Microbiol.*, **8**, 101–110.
6. Shizuya, H., Birren, B., Kim, U. J. et al. 1992, Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector., *Proc. Natl. Acad. Sci. USA*, **89**, 8794–8797.
7. Kaneko, T., Tanaka, A., Sato, S. et al. 1995, Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. I. Sequence features in the 1 Mb region from map positions 64% to 92% of the genome., *DNA Res.*, **2**, 153–166.
8. 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.
9. 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, *Nucl. Acids Res.*, **25**, 3389–3402.
10. 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.
11. Bancroft, I. and Wolk, C. P. 1988, Pulsed homogeneous orthogonal field gel electrophoresis (PHOGE), *Nucleic Acids Res.*, **16**, 7405–7418.
12. Simon, R. D. 1978, Survey of extrachromosomal DNA found in the filamentous cyanobacteria, *J. Bacteriol.*, **136**, 414–418.
13. Lambert, G. R. and Carr, N. G. 1984, Resistance of DNA from filamentous and unicellular cyanobacteria to restriction endonuclease cleavage, *Biochim. Biophys. Acta*, **781**, 45–55.
14. Padhy, R. N., Hottat, F. G., Coene, M. M., and Hoet, P. P. 1988, Restriction analysis and quantitative estimation of methylated bases of filamentous and unicellular cyanobacterial DNAs, *J. Bacteriol.*, **170**, 1934–1939.
15. Matveyev, A. V., Young, K. T., Meng, A., and Elhai, J. 2001, DNA methyltransferases of the cyanobacterium *Anabaena* PCC 7120, *Nucl. Acids Res.*, **29**, 1491–1506.
16. Duyvesteyn, M. G. C., Korsuize, J., de Waard, A., Vonshak, A., and Wolk, C. P. 1983, Sequence-specific endonucleases in strains of *Anabaena* and *Nostoc*, *Arch. Microbiol.*, **134**, 276–281.
17. Muro-Pastor, A. M., Kuritz, T., Flores, E., Herrero, A., and Wolk, C. P. 1994, Transfer of a genetic marker from a megaplasmid of *Anabaena* sp. strain PCC 7120 to a megaplasmid of a different *Anabaena* strain, *J. Bacteriol.*, **176**, 1093–1098.
18. Ligon, P. J., Meyer, K. G., Martin, J. A., and Curtis, S. E. 1991, Nucleotide sequence of a 16S rRNA gene from *Anabaena* sp. strain PCC 7120, *Nucleic Acids Res.*, **19**, 4553.
19. Kumano, M., Tomioka, N., and Sugiura M. 1983, The complete nucleotide sequence of a 23S rRNA gene from a blue-green alga, *Anacystis nidulans*, *Gene*, **24**, 219–225.
20. Corry, M. J., Payne, P. I., and Dyer, T. A. 1974, The nucleotide sequence of 5S rRNA from the blue-green alga *Anacystis nidulans*, *FEBS Lett.*, **46**, 63–66.
21. Xu, M. Q., Kathe, S. D., Goodrich-Blair, H., Nierzwicki-Bauer, S. A., and Shub, D. A. 1990, Bacterial origin of a chloroplast intron: conserved self-splicing group I introns in cyanobacteria, *Science*, **250**, 1566–1570.
22. Watanabe, T., Sugita, M., and Sugiura, M. 1998, Identification of 10Sa RNA (tmRNA) homologues from the cyanobacterium *Synechococcus* sp. strain PCC6301 and related organisms, *Biochim. Biophys. Acta*, **1396**, 97–104.

23. Gorodkin, J., Knudsen, B., Zwieb, C., and Samuelsson, T. 2001, SRPDB (Signal Recognition Particle Database), *Nucleic Acids Res.*, **29**, 169–170.
24. Watanabe, T., Sugiura, M., and Sugita, M. 1997, A novel small stable RNA, 6Sa RNA, from the cyanobacterium *Synechococcus* sp. strain PCC6301, *FEBS Lett.*, **416**, 302–306.
25. Vioque, A. 1992, Analysis of the gene encoding the RNA subunit of ribonuclease P from cyanobacteria, *Nucleic Acids Res.*, **20**, 6331–6337.
26. Riley, M. 1993, Functions of the gene products of *Escherichia coli*, *Microbiol. Rev.*, **57**, 862–952.
27. 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.
28. Pietrokovski, S. 1996, A new intein in cyanobacteria and its significance for the spread of inteins, *Trends Genet.*, **12**, 287–288.
29. Gorbalenya, A. E. 1998, Non-canonical inteins, *Nucleic Acids Res.*, **26**, 1741–1748.
30. Wu, H., Hu, Z., and Liu, X. Q. 1998, Protein trans-splicing by a split intein encoded in a split DnaE gene of *Synechocystis* sp. PCC6803, *Proc. Natl. Acad. Sci. USA*, **95**, 9226–9231.
31. Neer, E. J., Schmidt, C. J., Nambudripad, R., and Smith, T. F. 1994, The ancient regulatory-protein family of WD-repeat proteins, *Nature*, **371**, 297–300.
32. Janda, L., Tichy, P., Spizek, J., and Petricek, M. 1996, A deduced *Thermomonospora curvata* protein containing serine/threonine protein kinase and WD-repeat domains, *J. Bacteriol.*, **178**, 1487–1489.
33. van der Biezen, E. A. and Jones, J. D. 1998, The NB-ARC domain: a novel signalling motif shared by plant resistance gene products and regulators of cell death in animals, *Curr. Biol.*, **8**, 226–227.
34. Iwasaki, H. and Kondo, T. 2000, The current state and problems of circadian clock studies in cyanobacteria, *Plant Cell Physiol.*, **41**, 1013–1020.
35. Xu, X., Khudyakov, I., and Wolk, C. P. 1997, Lipopolysaccharide dependence of cyanophage sensitivity and aerobic nitrogen fixation in *Anabaena* sp. strain PCC 7120, *J. Bacteriol.*, **179**, 2884–2891.
36. Bancroft, I. and Wolk, C. P. 1989, Characterization of an insertion sequence (IS891) of novel structure from the cyanobacterium *Anabaena* sp. strain M-131, *J. Bacteriol.*, **171**, 5949–5954.
37. Golden, J. 1998, In: de Bruijn, F. J., Lupski, R. & Weinstock, G. M. (eds) Bacterial genomes physical structure and analysis, Chapman and Hall, New York, pp. 162–173.
38. 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.
39. Muro-Pastor, A. M., Herrero, A., and Flores, E. 1997, The *nuiA* gene from *Anabaena* sp. encoding an inhibitor of the NucA sugar-non-specific nuclease, *J. Mol. Biol.*, **268**, 589–598.
40. 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.