Genomic Structure of an Economically Important Cyanobacterium, *Arthrospira* (*Spirulina*) *platensis* NIES-39

Takatomo Fujisawa [1], Rei Narikawa [2], Shinobu Okamoto [3], Shigeki Ehira [4], Hidehisa Yoshimura [2], Iwane Suzuki [5], Tatsuru Masuda [2], Mari Mochimaru [6], Shinichi Takaichi [7], Koichiro Awai [8], Mitsuo Sekine [1], Hiroshi Horikawa [1], Isao Yashiro [1], Seiha Omata [1], Hiromi Takarada [1], Yoko Katano [1], Hiroki Kosugi [1], Satoshi Tanikawa [1], Kazuko Ohmori [9], Naoki Sato [2], Masahiko Ikeuchi [2], Nobuyuki Fujita [1,*], and Masayuki Ohmori [4]

Bioresource Information Center, Department of Biotechnology, National Institute of Technology and Evaluation (NITE), 2-10-49 Nishihara, Shibuya-ku, Tokyo 151-0066, Japan[1]; Department of Life Sciences (Biology), The University of Tokyo, 3-8-1 Komaba, Meguro-ku, Tokyo 153-8902, Japan[2]; Database Center for Life Science, Research Organization of Information and Systems, 2-11-6 Yayoi, Bunkyo-ku, Tokyo 113-0032, Japan[3]; Department of Biological Sciences, Faculty of Science and Engineering, Chuo University, 1-13-27 Kasuga, Bunkyo-ku, Tokyo 112-8551, Japan[4]; Graduate School of Life and Environmental Sciences, University of Tsukuba, Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8572, Japan[5]; Natural Science Faculty, Komazawa University, 1-23-1 Komazawa, Setagaya-ku, Tokyo 154-8525, Japan[6]; Biological Laboratory, Nippon Medical School, Kosugi-cho 2, Nakahara-ku, Kawasaki 211-0063, Japan[7]; Division of Global Research Leaders, Shizuoka University, 836 Ohya, Suruga-ku, Shizuoka 422-8529, Japan[8] and Department of Life Sciences, Showa Women's University, 1-7 Taishido, Setagaya-ku, Tokyo 154-8533, Japan[9]

*To whom correspondence should be addressed. Tel. +81 3-3481-1933. Fax. +81 3-3481-8424.
E-mail: fujita-nobuyuki@nite.g.o.jp

Abstract

A filamentous non-N_2-fixing cyanobacterium, *Arthrospira* (*Spirulina*) *platensis*, is an important organism for industrial applications and as a food supply. Almost the complete genome of *A. platensis* NIES-39 was determined in this study. The genome structure of *A. platensis* is estimated to be a single, circular chromosome of 6.8 Mb, based on optical mapping. Annotation of this 6.7 Mb sequence yielded 6630 protein-coding genes as well as two sets of rRNA genes and 40 tRNA genes. Of the protein-coding genes, 78% are similar to those of other organisms; the remaining 22% are currently unknown. A total 612 kb of the genome comprise group II introns, insertion sequences and some repetitive elements. Group I introns are located in a protein-coding region. Abundant restriction-modification systems were determined. Unique features in the gene composition were noted, particularly in a large number of genes for adenylate cyclase and haemolysin-like Ca^{2+}-binding proteins and in chemotaxis proteins. Filament-specific genes were highlighted by comparative genomic analysis.

Key words: cyanobacteria; *Arthrospira*; health supplement; genome; cAMP

1. Introduction

Cyanobacteria are prokaryotes that perform oxygenic photosynthesis and constitute a large taxonomic group within the domain of eubacteria. Cyanobacteria are divided morphologically (unicellular or filamentous) or functionally (N_2-fixing and non-N_2-fixing). Filamentous species are subdivided into those with and without a heterocyst which is a differentiation from vegetative cells for fixing nitrogen.[1,2]

Arthrospira is a representative filamentous non-N_2-fixing cyanobacterium that lacks any differentiation such as for the heterocyst, akinete or hormogonium, which develops in some filamentous N_2-fixing cyanobacteria. This cyanobacterium is also well known as 'Spirulina' because of its useful property as a food. The Aztecs consumed it regularly,[3] and it is thought to be an important dietary element in tropical areas. Recently, Lake Chad has been famous for *Arthrospira* that is naturally cultivated as a food supply.[4] However, current taxonomy claims that the name 'Spirulina' for strains used as food supplements is inappropriate, and there is agreement that *Arthrospira* is a distinct genus,[5] consisting of over 30 different species including *A. platensis* and *A. maxima*. In the present study, we determined nearly the complete genome sequence of *A. platensis* NIES-39.[6]

Arthrospira platensis has physiologically particular diagnostic characteristics, and its biology is subject to a comprehensive book.[7] *Arthrospira platensis* shows vigorous gliding motility of filamentous cells (trichomes) with rotation along their long axis. Gliding is a self-propulsion across a solid or semi-solid material without the aid of any visible flagellum[8]; however, the mechanism of gliding motility is not fully understood. This organism also possesses ecologically very valuable characteristics such as alkali and salt tolerance and algal mat production on the periphery of lakes. *Arthrospira platensis* is also able to grow under high salt concentrations of ~1.5-fold higher than sea water.[9] Accordingly, it often dominates in lakes with high carbonate/bicarbonate levels and high pH levels.[10]

Gene expression during adaptation to new environmental conditions like a high salt concentration has been determined in *A. platensis*, and the importance of the cAMP signalling cascade in its regulatory mechanism in response to changes in the external environment has been highlighted.[11] Moreover, molecular analysis of adenylate cyclase genes has shown that *A. platensis* develops a number of diverse cAMP-dependent signal cascades to adapt to different severe environmental conditions.[12–14]

Arthrospira platensis has become an important industrial organic material as a health supplement, a source of beta-carotene and a natural colouring agent. It has been approved for treating symptoms of radiation sickness after the Chernobyl disaster in Russia.[15] The presence of hydrogenase in its cells also makes this cyanobacterium a useful material for clean energy production.[16]

Despite its various useful applications, very little is known about the biology, physiology and genetic system of *A. platensis*. For production of useful products, gene manipulation through genetic engineering should be considered. However, genetic transformation of *Arthrospira* has had limited success to date,[17] and thus commercial use of this organism has faced barriers due to difficulties in gene manipulation. To overcome these barriers, restriction-modification (RM) systems based on its genome sequence may prove useful.

In 1996, whole-genome sequencing was achieved for the first time in the unicellular cyanobacterium, *Synechocystis* sp. PCC 6803.[18] Since then, whole genomes of 38 strains of cyanobacteria have been sequenced (Table 1). Of these strains, four are N_2-fixing filamentous species, whereas the others are unicellular species, both N_2-fixing and non-N_2 fixing. Thus, this is the first study to sequence the genome of a filamentous non-N_2-fixing species.

This study determined the nearly complete genome sequence of *A. platensis* NIES-39 of ~6.8 Mb. Some characteristic gene sets particularly in signal transduction and gene RM systems were determined, and additional properties in the genome structure were compared with those of other cyanobacteria.

2. Materials and methods

2.1. Bacterial strain and culture conditions

Arthrospira (*Spirulina*) *platensis* strain NIES-39 was obtained from the culture collection at the National Institute for Environmental Studies, which was originally isolated from Lake Chad and maintained in the Institute of Applied Microbiology, the University of Tokyo (strain IAM M-135).[6] The cells were grown in the SOT medium[19] at 30°C under continuous illumination at 30 µmol photon $m^{-2} s^{-1}$ with aeration with 1% (v/v) CO_2.

2.2. Sequencing and assembly

Genomic DNA was isolated with Genomic-tip 500/G (Qiagen, Valencia, CA, USA) as recommended by the supplier. A DNA shotgun library with 1.5- and 5-kb inserts in pUC118 vector (Takara, Otsu, Japan) was constructed as described previously.[20] Plasmid clones were end-sequenced using dye terminator chemistry on an ABI Prism 3730 sequencer as described previously.[20] Raw sequence data corresponding to 11-fold coverage were assembled using PHRED/PHRAP/CONSED software (http://www.phrap.org).[21,22] For assembly validation, a fosmid library with 40-kb inserts in the pCC1FOS fosmid vector was constructed using the CopyControl Fosmid library production kit (Epicenter, Madison, WI, USA). Fosmid DNA was extracted from *Escherichia coli* transformants using a Montage BAC96 MiniPrep kit (Millipore, Billerica, MA, USA), and end sequencing was carried out using dye terminator chemistry on an ABI Prism 3730 as described previously.[20] Fosmid end sequences were mapped onto the

Table 1. Strains used in the comparative genomic analysis

CyanoClust grouping				Species	Abbr.	Genes	Physiological indexes		Number of Pfam domains				
Group	Filament	N$_2$-fixation	Heterocyst				Habitat	Motility	Total	HisKA	Respons_reg	PAS	GAF
I	−	−	−	*Acaryochloris marina* MBIC 11017	amr	8383	Marine	−	8273	75	161	54	74
				Gloeobacter violaceus PCC 7421	gvi	4430	Rock	−	5464	39	56	20	15
				Microcystis aeruginosa NIES 843	mar	6312	Freshwater	−	5151	20	30	1	12
				Prochlorococcus marinus SS120	pma	1883	Marine	−	1967	4	6	1	0
				Prochlorococcus marinus AS9601	pmb	1921	Marine	−	1947	4	5	1	1
				Prochlorococcus marinus MIT 9515	pmc	1906	Marine	−	1892	4	5	1	1
				Prochlorococcus marinus NATL1A	pme	2193	Marine	−	2074	5	6	1	1
				Prochlorococcus marinus MIT 9303	pmf	2997	Marine	−	2908	7	8	1	0
				Prochlorococcus marinus MIT 9301	pmg	1907	Marine	−	1928	4	6	1	1
				Prochlorococcus marinus MIT 9215	pmh	1983	Marine	−	1941	3	5	1	1
				Prochlorococcus marinus MIT 9312	pmi	1810	Marine	−	1928	5	6	1	1
				Prochlorococcus marinus MIT 9211	pmj	1855	Marine	−	1963	4	6	1	0
				Prochlorococcus marinus MED4	pmm	1717	Marine	−	1918	5	6	1	1
				Prochlorococcus marinus NATL2A	pmn	2163	Marine	−	2110	5	6	1	1
				Prochlorococcus marinus MIT 9313	pmt	2269	Marine	−	2549	6	9	1	0
				Synechococcus elongatus PCC 6301	syc	2527	Freshwater	−	3198	15	26	20	19
				Synechococcus sp. CC9605	syd	2645	Marine	−	2571	5	9	1	2
				Synechococcus sp. CC9902	sye	2307	Freshwater	−	2453	5	7	1	1
				Synechococcus elongatus PCC 7942	syf	2662	Freshwater	−	3266	16	27	20	20
				Synechococcus sp. CC9311	syg	2892	Freshwater	−	2696	12	15	1	1
				Synechococcus sp. WH 5701	syh	3346	Marine	−	nd	nd	nd	nd	nd
				Synechococcus sp. RCC307	syr	2535	Marine	−	2633	9	11	1	1
				Synechococcus sp. WH 8102	syw	2519	Marine	+	2392	5	9	1	2
				Synechococcus sp. WH 7803	syx	2533	Marine	−	2669	12	14	1	0
				Synechocystis sp. PCC 6803	syn	3569	Freshwater	+	4269	41	67	29	33
				Thermosynechococcus elongatus BP-1	tel	2476	Hot spring	−	2901	15	32	13	19
II	−	+	−	*Crocosphaera watsonii* WH 8501	cro	5958	Marine	−	nd	nd	nd	nd	nd
				Synechococcus sp. JA-3-3Ab	cya	2760	Hot spring	−	3394	18	37	13	22
				Synechococcus sp. JA-2-3B'a(2−13)	cyb	2862	Hot spring	−	3469	20	36	15	18
				Cyanothece sp. PCC 7424	cyc	5710	Freshwater	−	6407	62	120	67	47
				Cyanothece sp. PCC 7425	cyn	5327	Freshwater	−	6255	70	145	122	68
				Cyanothece sp. PCC 8801	cyp	4367	Freshwater	−	5182	38	81	32	39
				Cyanothece sp. ATCC 51142	cyt	5304	Marine	−	5490	44	81	30	33
III	−/+	−	−	*Synechococcus* sp. PCC 7002	syp	3186	Marine	−	3764	31	54	16	19
IV	+	−	−	*Arthrospira platensis* NIES 39	apl	6631	Freshwater	+	6631	70	107	76	58
V	+	+	−	*Trichodesmium erythraeum* IMS101	ter	4451	Marine	−	6159	26	47	10	24
VI	+	+	+	*Anabaena* sp. PCC 7120	ana	6130	Freshwater	−	7207	103	153	72	80
				Anabaena variabilis ATCC 29413	ava	5661	Freshwater	−	7234	102	151	73	83
				Nostoc punctiforme ATCC 29133	npu	6690	Soil	+	8751	125	194	92	121

Abbr., abbreviation; nd, not determined. Pfam domains: HisKA (PF00512), Respons_reg (PF00072), PAS (PF00989) and GAF(PF01590).

assembled sequence. Fosmid clones that link two contigs were selected and sequenced by primer walking to close any gaps. The sequencing of difficult templates was performed using a CUGA sequencing kit (Nippon Genetech, Tokyo, Japan). Contig sequences generated by Roche 454 FLX sequencer were also used to fill some gaps.

2.3. Optical mapping

The sequence assemblies were constructed by optical maps (OpGen Technologies, Madison, WI, USA) and PCR.[23,24] Briefly, high molecular weight DNA was immobilized as individual molecules onto optical chips, digested with *Nco*I (New England Biolabs, Ipswich, MA, USA), fluorescently stained with YOYO-1 (Invitrogen, Carlsbad, CA, USA) and positioned onto an automated fluorescent microscope system for image capture and fragment size measurement to give high-resolution single-molecule restriction maps. Single-molecule maps were collected and then assembled to produce whole genome.

Optical maps and sequence contigs were compared as described previously.[23] Sequence FASTA files were converted to *in silico* restriction maps via MapViewer software (OpGen Technologies) for direct comparison with the optical maps. Comparisons were accomplished by aligning the sequence with the optical maps according to their restriction fragment pattern. Alignments were generated with a dynamic programming algorithm that finds the optimal location or placement of a sequence contig by global alignment of the sequence contig against the optical map. Local alignment analysis was also performed to compare segments of the sequence contigs to the optical map.

2.4. Genome analysis and annotation

Putative non-translated genes were identified using the Rfam[25] and tRNAscan-SE[26] programs and rRNA genes using the RNAmmer[27] and BLASTN[28] programs, whereas protein-coding genes were identified using the GLIMMER program by obtaining potential open reading frames >150 bp.[29,30] The genome sequence was translated into potential protein sequences in six frames and compared with sequences in the UniProt database[31] using the BLASTP program[28] for identification of additional genes not predicted by other methods and genes <150 bp, especially in the predicted intergenic regions. The start sites were manually inspected and altered based on predictions of GLIMMER and GeneHacker.[32] For functional annotation, the non-redundant UniProt database and protein signature database, InterPro,[33] were searched to assign the predicted protein sequences based on sequence similarities. Orthologous genes among cyanobacteria were manually curated using Gclust[34]

and CYORF.[35] The KEGG database was used for pathway reconstruction.[36] Signal peptides in proteins were predicted using SignalP,[37] and transmembrane helices were predicted using TMHMM.[38]

2.5. Comparative genomic analysis

We first used the hmmer (v2.3.2) program and the Pfam_ls database (release 23) to analyse the cyanobacterial genomes (Table 1).

For comparative genomic analysis to determine genes related to trichome formation and N_2-fixation, we applied the CyanoClust database.[39] We classified all putative proteins from the 39 cyanobacterial genomes into nearly orthologous protein clusters; each cluster was statistically defined to minimize its size. This is particularly useful for analysing the many regulatory proteins in cyanobacteria which have multidomains. We divided *A. platensis* and the other strains of cyanobacteria into six groups (Table 1): unicellular non-N_2-fixing cyanobacteria such as *Synechocystis* sp. PCC 6803 (group I), unicellular N_2-fixing cyanobacteria such as *Crocosphaera watsonii* (group II), *Synechococcus* sp. PCC 7002 (a unicellular cyanobacterium which forms a short filament near optimal growth temperature; D. Bryant, personal communication) (group III), filamentous non-N_2-fixing cyanobacteria (*A. platensis*, group IV), filamentous N_2-fixing non-heterocyst-forming cyanobacteria (*Trichodesmium erythraeum*, group V) and filamentous N_2-fixing heterocyst-forming cyanobacteria such as *Anabaena* sp. PCC 7120 (group VI). Gene clusters that were shared between groups were automatically extracted, and only those clusters present in all strains of the group were analysed.

3. Results and discussion

3.1. Sequencing, optical mapping and structural features of the *A. platensis* genome

The nucleotide sequence of the whole genome of *A. platensis* NIES-39 was determined using the whole-genome shotgun method. Three different libraries with 1.5-, 5- and 40-kb inserts were prepared and each end sequence was determined. A total of 92 878 random sequences corresponding to 11 genome equivalents were assembled into 18 supercontigs. Final sequence assembly was carried out by visually editing the draft sequences and by optical mapping (data not shown) to determine the relative order/orientation of each supercontig. The 18 remaining gaps were estimated to be ∼95 kb. Most, if not all, gaps were flanked by repeated clusters of tandem sequences or the phage-like sequences. Gap closing including long PCR followed by primer walking and transposon-mediated sequencing was
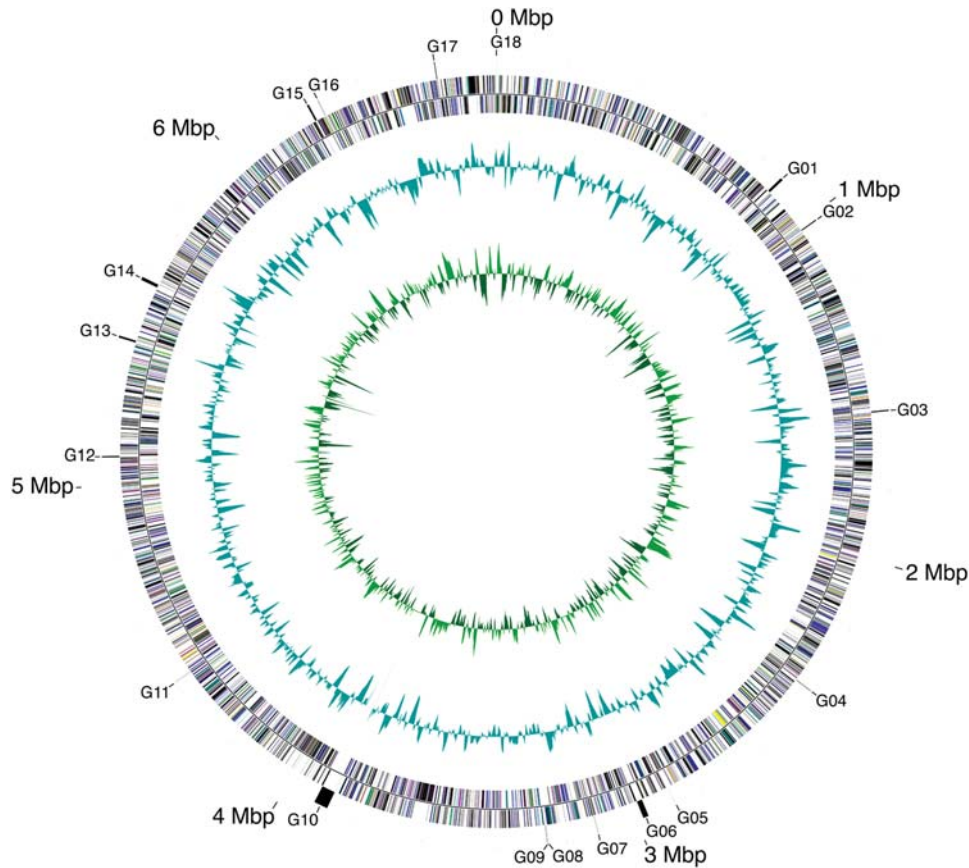
Figure 1. Schematic representation of the circular chromosome of *A. platensis*. A scale indicates the coordinates in megabase pairs. From outside to inside: circle 1, the gaps in the genome; circles 2 and 3, predicted protein-coding genes on the forward and reverse strands; circle 4, G+C content; circle 5, GC skew. Eighteen contig gaps (G01-G18) are numbered in the clockwise direction starting from the end of the longest contig. Functional categories were colour-coded according to the standard colours used by COGs. The genome sequence and annotation of *A. platensis* NIES-39 are available at GenBank/EMBL/DDBJ under accession no. AP011615.

not successful due to unusually abundant repeats. The total sequence of the 18 supercontigs has a length of 6 692 865 bp with an average G+C content of 44.3%. Taken together, the *A. platensis* genome is composed of a single, circular chromosome of 6.8 Mb (Fig. 1); no plasmid DNA sequences were detected. On GC skew analysis to locate the probable origin and terminator of DNA replication, no apparent shift of skew was detected as in other cyanobacteria.

The estimated genome comprises 6630 potential protein-coding genes (average size, 835 bp), as well as 49 RNA genes consisting of 2 sets of rRNA genes, 40 tRNA genes representing 20 tRNA species, tmRNA, the B subunit of RNase P and signal recognition particle RNA. Translated amino acid sequences were compared with sequences in the UniProt database using the BLAST program. Of the 6630 potential protein-coding genes, 5157 (78%) were orthologous or had similarity to genes of known function or hypothetical genes (*E*-value of $<10^{-3}$) and the remaining 1473 (22%) showed no significant

similarity to any registered genes. On manual curation, 2539 (38%) genes could be assigned to biological roles. Genes for general metabolism in cyanobacteria were detected with no particular difference.

3.2. Comparative genomic analysis

3.2.1. Conserved domain analysis using Pfam We obtained 2673 kinds of Pfam domains from 37 cyanobacterial genomes and 1537 kinds of Pfam domains from the *A. platensis* genome. Table 2 presents the top 50 Pfam domains in *A. platensis*, which contains highly repetitive motifs including TPR_1 (PF00515), TPR_2 (PF07719), WD40 (PF00400), HemolysinCabind (PF00353), Pentapeptide (PF00805), HNH (PF01844) and GIIM (PF08388).

Positive correlation was detected between the number of genes and number of Pfam domains of the 38 cyanobacterial species ($r = 0.9X$, $P < 0.05$), including *A. platensis* (Supplementary Fig. S1). Eighteen

Table 2. Top 50 of Pfam domains of *A. platensis*

domains	apl	amr	ana	ava	cya	cyb	cyc	cyn	cyp	cyt	gvi	mar	npu	pma	pmb	pmc	pme	pmf	pmg	pmh	pmi	pmj	pmm	pmn	pmt	syc	syd	sye	syf	syg	syn	syp	syr	syw	syx	tel	ter
TPR_1	415	300	164	158	58	73	202	98	144	75	97	197	300	8	12	7	50	207	13	13	12	24	8	86	86	34	9	11	35	12	55	62	8	18	10	40	628
TPR_2	398	308	166	163	76	90	199	115	147	72	123	202	298	7	16	7	49	210	13	13	10	20	9	87	76	42	11	16	43	12	53	63	13	21	17	48	593
WD40	220	398	230	282	4	4	197	72	65	128	153	97	320	2	2	2	3	2	2	2	2	2	3	0	0	0	0	0	0	0	45	11	0	1	0	10	174
HemolysinCabind	191	172	72	60	0	0	53	3	10	27	10	52	123	0	6	0	3	17	0	0	0	0	0	1	3	7	0	5	7	10	14	0	27	11	5	0	156
Pentapeptide	160	252	98	89	40	44	111	106	101	111	68	50	123	8	6	7	8	13	6	6	6	8	7	8	15	20	11	11	20	12	48	62	13	13	13	41	107
HNH*	111	9	15	11	4	3	7	4	7	14	4	11	7	2	3	3	2	2	3	3	3	2	4	2	2	1	2	3	1	2	11	4	2	2	2	7	8
Response_reg	107	161	153	151	37	36	120	145	81	81	56	30	194	6	5	5	6	8	6	5	6	6	6	6	9	26	9	7	27	15	67	54	11	9	14	32	47
GIIM*	103	6	5	2	0	0	5	0	4	6	1	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	12
HATPase_c	82	95	139	135	26	27	76	102	47	52	49	23	168	6	5	5	6	9	6	5	6	6	6	8	17	7	7	18	13	48	36	10	7	11	21	38
PAS_3	77	47	75	80	13	17	51	156	30	31	17	1	88	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	20	0	34	18	0	0	0	13	7
PAS	76	54	72	73	13	15	67	122	32	30	20	1	92	1	1	1	1	1	1	1	1	1	1	1	1	20	1	1	20	1	29	16	1	1	1	13	10
PAS_4	76	54	73	81	19	24	57	128	28	32	30	3	85	1	0	0	1	1	0	0	0	0	0	1	1	19	1	1	19	1	24	12	1	1	1	14	5
DUF820	70	26	56	48	15	21	67	53	85	54	93	89	63	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	2	0	35	26	0	0	0	1	7
HisKA	70	75	103	102	18	20	62	70	38	44	39	20	125	4	4	4	5	7	4	3	5	4	5	5	6	15	5	5	16	12	41	31	9	5	12	15	26
ABC_tran	65	79	106	105	55	63	78	88	65	73	68	66	106	31	32	28	34	43	30	31	32	29	31	34	44	50	40	38	51	38	60	60	36	47	42	49	56
GAF	58	74	80	83	22	18	47	68	39	33	15	12	121	0	1	1	1	0	1	1	1	0	1	1	0	19	2	1	20	1	33	19	1	2	0	19	24
Pkinase	46	58	49	56	9	10	28	30	16	20	15	17	56	0	0	0	0	2	0	0	0	0	0	1	5	0	1	5	1	7	8	1	0	1	0	11	42
SBBP*	46	0	0	0	0	1	0	7	0	0	9	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	13
HEAT_PBS	41	9	55	38	15	14	20	19	20	16	25	27	28	1	1	0	5	4	1	1	1	2	0	4	4	14	5	7	14	5	13	14	8	3	6	9	33
PPC	41	3	17	12	1	0	2	0	1	8	9	6	5	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	1	2	0	7	0	0	0	75
Pkinase_Tyr	40	54	46	52	8	9	26	23	16	20	12	16	51	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	5	0	6	8	0	0	10	38
Transposase_35	40	2	25	13	47	65	46	4	30	7	2	92	2	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1	2	0	0	0	32	4
Transposase_14*	39	3	0	0	16	0	5	1	1	0	3	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	26	0	0	0	0	0	15
Glycos_transf_1	38	48	52	45	7	8	34	56	33	36	29	28	57	10	8	12	12	12	6	8	9	7	8	9	10	13	12	19	13	11	26	19	15	14	17	12	30
Glycos_transf_2	38	26	48	45	8	8	42	48	35	33	24	34	38	9	10	13	4	14	6	7	6	6	11	4	8	11	12	11	12	9	24	11	12	11	10	20	38
Methyltransf_11	38	74	42	41	27	26	35	54	33	36	44	32	54	10	14	12	12	19	10	12	11	10	15	11	15	23	14	12	23	15	25	19	17	18	19	11	51
RVT_1*	37	9	8	3	0	0	6	0	4	7	1	5	2	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	2	1	0	0	5	11

Table 2. Continued

domains	apl	amr	ana	ava	cya	cyb	cyc	cyn	cyp	cyt	gvi	mar	npu	pma	pmb	pmc	pme	pmf	pmg	pmh	pmi	pmj	pmm	pmn	pmt	syc	syd	sye	syf	syg	syn	syp	syr	syw	syx	tel	ter
GGDEF	33	54	14	14	3	4	34	19	23	31	1	2	21	0	0	0	0	0	0	0	0	0	0	0	0	17	0	0	17	4	23	13	0	0	0	8	4
Methyltransf_12	33	66	34	32	23	22	28	46	28	34	39	28	46	8	11	10	12	19	9	9	7	9	12	10	13	20	14	13	20	14	24	17	15	16	17	10	44
BPD_transp_1	27	24	34	39	28	34	25	30	23	23	19	19	31	8	5	8	8	14	6	5	7	9	6	8	13	22	13	12	23	11	21	22	13	13	15	16	19
Epimerase	27	54	37	42	16	14	30	44	31	30	35	27	56	16	18	20	18	22	16	17	14	12	18	16	17	16	18	20	16	19	25	17	21	19	20	13	27
Transposase_2	27	2	24	12	42	62	42	4	26	5	2	84	2	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1	2	0	0	0	32	6
PIN	26	10	17	9	1	2	10	18	10	14	22	31	14	1	1	1	1	1	1	1	1	1	1	1	1	4	1	1	5	3	13	22	1	5	2	1	1
AAA	25	34	34	37	23	18	28	25	19	19	25	23	37	13	13	13	15	16	13	13	12	12	14	15	15	16	18	14	16	16	18	17	14	16	15	14	37
DAO	25	32	29	37	20	22	31	28	22	25	28	20	34	16	16	13	15	23	12	11	14	14	15	15	21	23	25	21	24	27	25	22	21	22	21	13	24
MMR_HSR1	25	35	32	28	20	21	24	30	24	28	18	24	31	16	16	16	16	17	17	17	17	16	16	16	17	21	18	18	21	19	23	19	17	15	17	24	27
PG_binding_1	24	1	29	28	0	0	3	3	1	2	1	4	26	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1	7	0	0	0	1	11
adh_short	23	47	38	41	14	12	28	39	23	28	43	25	79	10	9	8	9	16	7	9	11	10	7	8	16	12	13	16	12	20	17	12	15	17	16	11	16
CBS	23	24	21	21	8	6	18	16	11	13	9	10	19	3	3	3	3	5	4	4	3	4	3	3	5	11	6	6	11	5	8	14	5	6	5	7	13
SLH	23	24	29	32	10	9	9	11	11	13	11	8	40	2	0	0	0	4	0	0	0	1	0	0	4	7	6	4	7	6	9	11	2	4	4	8	10
AAA_5	22	24	20	23	12	11	13	19	16	13	19	19	30	10	9	10	10	13	9	9	9	12	11	10	13	13	15	14	13	13	17	12	15	14	14	12	29
Guanylate_cyc*	22	7	6	5	1	2	2	5	2	3	0	1	8	0	0	0	0	2	0	0	0	0	0	0	2	2	1	0	2	1	3	1	0	0	1	2	13
Radical_SAM	22	24	25	26	13	15	31	31	23	20	24	21	25	11	10	10	8	14	12	11	10	12	10	8	13	17	16	13	17	14	22	19	13	15	15	16	20
SMC_N	20	13	20	22	11	11	20	24	18	16	18	18	22	10	9	7	12	8	7	6	7	8	6	11	11	8	11	8	9	6	12	6	8	10	9	9	13
FHA	19	38	13	13	10	9	15	23	13	15	6	7	13	0	0	0	0	2	0	0	0	0	0	0	2	0	1	2	1	11	9	1	0	1	20	14	
Abhydrolase_1	18	24	24	27	10	11	24	26	15	20	23	14	29	4	5	6	5	4	5	5	5	4	6	5	4	13	6	5	13	5	15	14	6	3	6	11	18
GTP_EFTU	18	18	19	19	15	14	15	14	18	17	12	13	18	9	10	11	11	12	10	9	10	10	10	11	12	12	13	11	12	11	17	16	11	11	11	14	17
Hydrolase	18	21	25	33	9	9	19	22	13	23	14	15	26	4	2	4	5	6	2	4	2	5	3	5	7	18	7	7	18	6	15	16	9	8	8	10	12
NAD_binding_4	18	25	23	24	8	6	17	21	18	17	16	18	29	8	5	8	7	11	7	8	7	6	9	7	10	11	8	9	11	8	16	7	8	9	10	11	12
3Beta_HSD	17	25	17	19	5	5	16	21	13	14	19	11	22	6	6	3	7	7	6	7	5	4	8	8	7	8	7	8	8	5	11	7	8	6	7	8	10
Aminotran_1_2	17	12	17	20	12	13	14	13	15	15	17	15	22	7	8	5	7	10	8	6	5	6	5	8	9	11	9	7	13	9	13	11	12	7	7	11	16
CbiA	17	29	24	27	13	12	19	25	22	19	19	20	29	7	8	8	9	6	8	8	8	8	8	8	6	11	8	8	12	9	19	19	9	6	8	11	18

See Table 1 for abbreviations

*Highest domain numbers in *A. platensis*

Table 3. Summary of comparative genomic analysis using CyanoClust

	Physiological profiling				
	Common	*A. platensis*-specific	Heterocyst-specific	N_2-fixing-specific	Filament-specific
Group I	O	X	X	X	X
Group II	O	X	X	O	X
Group III	O	X	X	X	X/O
Group IV	O	O	X	X	O
Group V	O	X	X	O	O
Group VI	O	X	O	O	O
No. of clusters	694	1066	223	8	29/7
No. of genes in *A. platensis*	938	2056	0	0	31/7

For cyanobacteria grouping, see Table 1.

domains (DUF1064, BOF, Cytochrom_C_2, DUF1624, DUF1667, DUF2234, DUF2273, DUF268, DUF310, DUF579, Hp0062, JmjC, LAB_N, PhoU_div, PMSR, PPTA, SoxG and YibE_F) are unique to the *A. platensis* genome. The abundance of signalling domains, such as GAF (guanylate cyclase/adenylate cyclase/FhlA), PAS (Per/Arnt/Sim), HisKA and Respons_reg, depends on the habitat of the cyanobacteria; i.e. soil and freshwater cyanobacteria possess larger numbers of signalling domains than marine cyanobacteria (Supplementary Fig. S2). *A. platensis*, which lives in high salt lakes, was thus categorized as a soil/freshwater cyanobacterium.

3.2.2. Physiological profiling using CyanoClust
Arthrospira platensis is the first strain of filamentous non-N_2-fixing cyanobacteria to have its genome sequenced. Whole-genome sequencing allows for comparative genomic analysis especially in regard to trichome formation and N_2-fixation. Using the CyanoClust database, we extracted 694 gene clusters (938 genes in *A. platensis*) that were common to all six cyanobacteria groups (Table 3 and Supplementary Table S1). These clusters may be closely related to housekeeping, photosynthetic or cyanobacteria-specific genes. We also extracted 1066 gene clusters (2056 genes in *A. platensis*) that were specific to *A. platensis* (Table 3 and Supplementary Table S2); 71 (94 genes in *A. platensis*) common to only groups IV (*A. platensis*) and V (*T. erythraeum*) (Supplementary Table S3); and 223 common to the heterocyst-forming cyanobacteria. The latter clusters included known heterocyst-related genes such as *patN* and *hetP* (Table 3 and Supplementary Table S4). In addition, eight clusters were specifically conserved among N_2-fixing cyanobacteria (groups II, V and VI) (Table 3 and Supplementary Table S5); most of these encode *nif*-related genes.

Of the 29 gene clusters (31 genes in *A. platensis*, Table 4 and Supplementary Table S6) common to

only filamentous cyanobacteria (groups IV, V and VI), *fraC* and *fraG* (*sepJ*) are already known to have an association with trichome formation.[40,41] Notably, seven gene clusters (seven genes in *A. platensis*, Table 4 and Supplementary Table S7) were found to be common to the filamentous groups and *Synechococcus* sp. PCC 7002, which may be a primitive filamentous species (group III) and may also be related to trichome formation. Moreover, several genes (*patU*, *hetR* and *hetF*)[42–44] that are required for heterocyst maturation and N_2 fixation are conserved in *A. platensis*, although no nitrogenase genes were detected; this suggests that heterocyst formation could be developmentally coupled with trichome formation. Several contiguous genes (bolded in Table 4) are syntenically conserved among the five trichome-forming cyanobacteria. Particularly, NIES39_A00790 is located just downstream of *fraC*, implying that it is another candidate required for trichome development.

3.3. Mobile DNA elements
Group II introns include ribozymes and retroelements in the genomes of organelles, *Eubacteria* and *Archaea*. In bacteria, group II introns primarily act as retroelements consisting of six RNA structural domains (DI−DVI) and an intron-encoded protein, whereas in mitochondria and plastids, they frequently lack the intron-encoded protein and are mobile. At least 150 group II introns were identified in the *A. platensis* genome; 71 of these encode reverse transcriptase/maturase, whereas the other 79 do not. Additionally, 88 sequences were assigned to the group II catalytic intron (RF00029), DV and DVI. This number was much higher than that in previously reported abundant species (27 copies of group II introns in *T. elongatus*[45] and 28 copies of RF00029 in *T. erythraeum*). These results suggested that

Table 4. Gene clusters conserved only in filamentous cyanobacteria

Cluster No.	Gene ID	Annotation
Specific for groups IV, V and VI		
4981	**NIES39_A00790**	Hypothetical protein
5158	**NIES39_A00800**	Filament integrity protein (*fraC*)
4909	NIES39_A01310	NUDIX hydrolase
3571	NIES39_A03140	Hypothetical protein
4548	NIES39_A04850	Hypothetical protein
3571	NIES39_C00130	Hypothetical protein
4852	NIES39_C00870	Hypothetical protein
2616	NIES39_D00070	Hypothetical protein
4736	NIES39_D00920	Hypothetical protein
4710	NIES39_E01660	Nuclease (SNase-like)
4588	NIES39_E02590	Serine/threonine protein kinase
4667	NIES39_F00600	Probable glycosyl transferase
4881	**NIES39_K02770**	Hypothetical protein
4277	**NIES39_K02780**	Hypothetical protein
4548	NIES39_L02410	Hypothetical protein
4747	NIES39_L03440	Hypothetical protein
2616	NIES39_L06030	Hypothetical protein
5149	NIES39_L06180	Hypothetical protein
5020	NIES39_M00320	Hypothetical protein
4731	NIES39_M02510	DUF6 transmembrane protein (*sepJ, fraG*)
5143	**NIES39_N00400**	Hypothetical protein (*patU*)
2616	**NIES39_N00410**	Hypothetical protein
5032	NIES39_N00980	Hypothetical protein
4826	NIES39_O03720	Hypothetical protein
5058	**NIES39_O03830**	hypothetical protein
5102	**NIES39_O03850**	Hypothetical protein
4991	NIES39_O04240	Hypothetical protein
4590	NIES39_O06790	Alpha/beta hydrolase fold
4996	**NIES39_Q00570**	Hypothetical protein
5045	**NIES39_Q00580**	Hypothetical protein
4607	NIES39_R00660	Hypothetical protein
Specific for groups III, IV, V and VI		
4298	NIES39_C02810	Hypothetical protein
4333	NIES39_C03480	Heterocyst differentiation protein (*hetR*)
4276	NIES39_D04070	Hypothetical protein
4505	NIES39_J00800	Hypothetical protein
3819	NIES39_J02400	Heterocyst differentiation protein (*hetF*)
4311	NIES39_O06320	Hypothetical protein
4255	NIES39_Q02800	Hypothetical protein

Bolded genes are contiguous.

group II introns have been self-propagating extensively in *A. platensis* cells for a long time.

Group I introns include ribozymes that catalyse their own RNA splicing to produce ligated exons (mature RNA) and the excised intron. We detected two group I introns in one gene comprising three ORFs (NIES39_L05970, NIES39_L05980 and NIES39_L05990) which encode class I ribonucleotide reductases (RNRs; Fig. 2). Notably, a stop codon was found just after the insertion position in each intron, indicative of translational regulation. It is extremely rare for a group I intron to be inserted in protein-coding genes in bacteria, although one report has detected this insertion in class II RNR of the cyanobacterium *N. punctiforme*.[46] At least three different classes of RNRs that differ in primary sequence, substrate and cofactor requirements have been described in cyanobacteria. Since class I RNRs have no homology with class II RNRs, the essential DNA biosynthesis step from RNA could be an evolutionary hot spot for targeting these selfish genes. Cyanobacterial genome comparison revealed that classes I and II are distributed complementally in cyanobacteria, except for *Gloeobacter violaceus* in which no known RNR genes have been detected (Supplementary Table S8), whereas class III anaerobic RNRs are also present in some species. Many class I and II RNR genes are interrupted by introns and/or inteins which are spliced out for sequencing of RNA or protein. Inteins in *A. platensis* were also found in DnaB (NIES39_M02320), DnaX (NIES39_D01990) and thymidylate synthase (NIES39_Q01490) genes.

Novel phage-like sequences spanning from 12.5 to 25.5 kb were found in at least 18 loci in the *A. platensis* genome. Although there are a number of variations in deletion, insertion and rearrangement in each sequence, they consist of phage infection-related genes (e.g. NIES39_F00750) and small conserved putative genes, generating a direct repeat in some case. Such phage-like sequences contribute to at least 295 kb in the whole genome. As for transposases of the insertion sequence, 139 genes were found, which is comparable to other cyanobacteria.[18] A total 612 kb of the genome are composed of group II introns, phage-like sequences, insertion sequences and some repetitive elements, which include clustered regulatory interspaced short palindromic repeat (CRISPR) sequences (see Section 3.5.2).

3.4. Signal transduction proteins
3.4.1. cAMP and c-di-GMP signal transduction
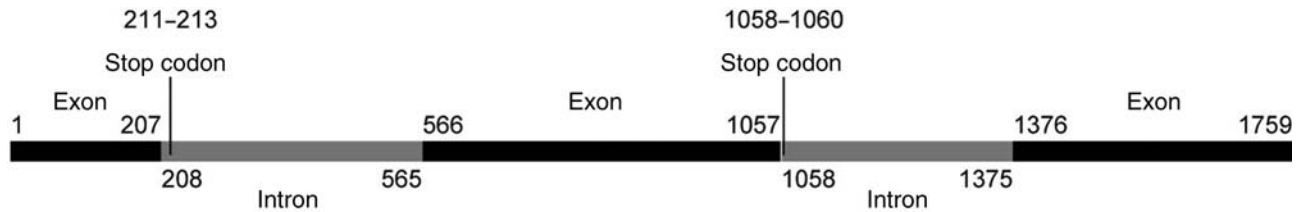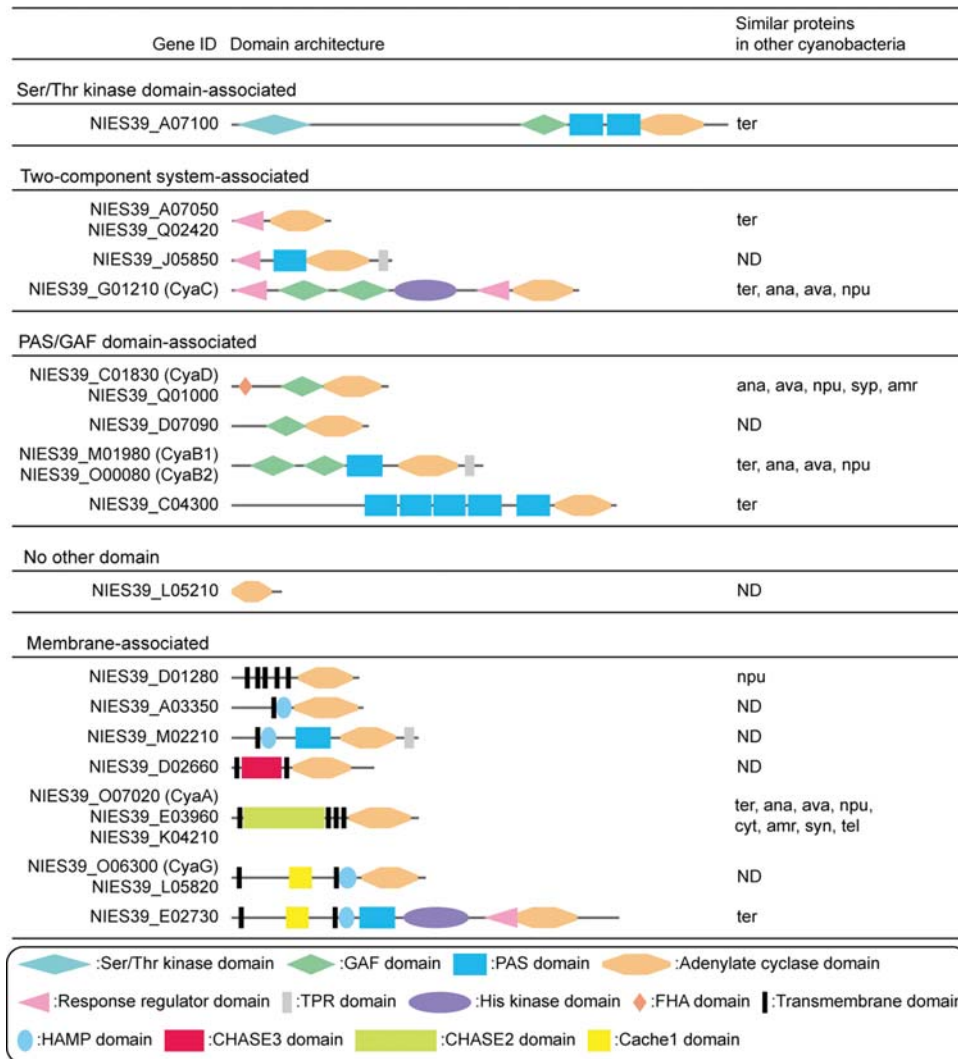cAMP is an important signalling molecule in

Figure 2. Gene structure of the ribonucleotide reductase (RNR) gene with two group I intron insertions.

Figure 3. Domain architecture of adenylate cyclases in *A. platensis*. See Table 1 for abbreviations.

cyanobacteria.[47] *A. platensis* genome encodes 22 adenylate cyclases that produce cAMP from ATP (Fig. 3). This number is quite prominent among cyanobacterial genomes, which usually encode less than 10 adenylate cyclases. Even *T. erythraeum*, the most closely related cyanobacterium to *A. platensis*, contains only 13 adenylate cyclase genes. In addition to the genes *cyaC*,[13,14] *cyaG*,[13,14] *cyaA*, *cyaB1*, *cyaB2* and *cyaD* homologues (Fig. 3),[48] many novel genes were detected. Another unique characteristic of

A. platensis is its 10 membrane-associated adenylate cyclases, compared with most cyanobacteria which have only one or two. Thus, the cAMP signal transduction response to external stimuli may be highly developed in *A. platensis*. As for the domain structure of adenylate cyclase, NIES39_A07100 bears a unique domain architecture: a Ser/Thr protein kinase domain in the N-terminus and GAF, PAS and adenylate cyclase domains in the C-terminus (Fig. 3). This domain architecture is quite similar to that of the

HstK family,[49] a hybrid of Ser/Thr protein kinase and histidine kinase, except that the C-terminal histidine kinase domain is replaced with the adenylate cyclase domain.

The *A. platensis* genome also contains 14 genes encoding cNMP-binding domains. Four of them [NIES39_C02730 (NtcA), NIES39_D00910 (CRP), NIES39_M01440 and NIES39_O00090] encode the helix-turn-helix DNA-binding motif of the CRP family, whereas the remaining encode unique domain architecture proteins. NIES39_A00950 bears a unique domain architecture, namely, an ATP:ADP antiporter-like domain in the N-terminus and a cNMP-binding domain in the C-terminus. The C-terminal cNMP-binding domain contains residues critical for binding cAMP. A previous report showed that *A. platensis* trichomes rapidly aggregate to form a mat in response to externally added cAMP.[50] This aggregation is accompanied with an increase in the intracellular ATP level and a decrease in the intracellular ADP level,[12] suggesting that NIES39_A00950 may contribute to this cAMP-dependent aggregation.

There are 33 GGDEF and 15 EAL domains, respectively, involved in the synthesis and hydrolysis of c-di-GMP,[51] a widely distributed second messenger for biofilm formation in bacteria. One GGDEF (NIES39_C01090) and three GGDEF/EAL (NIES39_C03530, NIES39_L00190 and NIES39_Q02220) proteins contain putative flavin-binding PAS domains, suggestive of light- or redox-responsive c-di-GMP signalling for movement, aggregation or biofilm formation.

3.4.2. Two-component signal transduction systems

Two-component signal transduction systems canonically consist of two proteins, a histidine kinase and a response regulator.[52] There are 84 putative genes for histidine kinases in the *A. platensis* genome. Among them, 33 encode hybrid histidine kinases containing not only a transmitter domain but also single or multiple receiver domain(s) or histidine phospho-transfer domain(s), which accept the phosphate group via intramolecular and/or intermolecular phospho-transfer. There are also 65 putative genes for response regulators.

Orthologous histidine kinases of *Synechocystis*, Hik2 (slr1147), Hik7 (SphS, sll0337), Hik8 (sasA, sll0750), Hik33 (sll0698) and Hik34 (slr1285), are conserved in almost all cyanobacterial genomes. The *A. platensis* genome contains orthologues of Hik2 (NIES39_J01640), Hik8 (NIES39_A03740), Hik33 (NIES39_M02760) and Hik34 (NIES39_J00420), but not SphS, a kinase that induces proteins to acquire inorganic phosphate (Pi) under Pi-deficient conditions. Although some marine cyanobacteria have been found to lack apparent orthologues of SphS, *A. platensis* is the first example among freshwater cyanobacteria.

Accordingly, it does not contain a gene for SphU which regulates SphS.[53] However, the response regulator SphR and its target genes *phoA* (alkaline phosphatase) and *pts* (high affinity Pi transporter) are present in the *A. platensis* genome, suggesting a different regulation system in this cyanobacterium.

Some histidine kinases in *A. platensis* contain multiple domains of PAS/PAC and/or GAF. These structures are also prevalent in other histidine kinases from filamentous cyanobacteria, such as *Anabaena* sp. PCC 7120[54] and *N. punctiforme*.[55] Like in other cyanobacterial genomes, most genes for histidine kinases and response regulators in the *A. platensis* genome are not collocated to each other on the chromosomes, except for the 18 cases (Supplementary Table S9). Several genes encode fragments of histidine kinases in the *A. platensis* genome suggestive of pseudogenes (NIES39_A02200−NIES39_A02210, NIES39_D02110−NIES39_D02130−NIES39_D02160, NIES39_L02740−NIES39_L02750, NIES39_L05670−NIES39_L05680−NIES39_L05690).

3.4.3. Transcription factors

Cyanobacteria have multiple σ^{70}-type sigma factors of RNA polymerase and no σ^{54}-type sigma factor. In the *A. platensis* genome, there are seven σ^{70}-type sigma factors. NIES39_L06110 encodes a group 1 sigma factor, which is also called the principal sigma factor and is essential for cell viability. In addition, three group 2 sigma factors (NIES39_A03770, NIES39_A08720 and NIES39_C01040), which show a high degree of sequence similarity with the principal sigma factor but are non-essential for cell growth, and three group 3 sigma factors (NIES39_C05220, NIES39_D05870 and NIES39_O06410), which are divergent from group 1 and 2 sigma factors in amino acid sequence, are identified. The B-, C- and D-type group 2 sigma factors and the F- and G-type group 3 sigma factors are conserved among non-marine cyanobacteria,[56,57] all of which are present in *A. platensis*. An additional group 3 sigma factor which does not form a clade with previously analysed sigma factors was also determined.

In the *A. platensis* genome, there are 66 putative genes for transcriptional regulators. Compared with the genome size, this number is quite small for a non-marine cyanobacterium, being more comparable to marine cyanobacteria.[58] For example, the freshwater cyanobacterium *Anabaena* sp. PCC 7120, having a comparable genome size to *A. platensis*, contains more than 100 transcriptional regulators.[54] Some well-conserved transcriptional regulators among cyanobacteria were found in *A. platensis*, including genes for NtcA (NIES39_C02730),[59] NtcB (NIES39_C02400),[60] NdhR (NIES39_K02860),[61] HrcA (NIES39_O04340)[62] and CRP (NIES39_D00910).[63] Two adjacent genes (NIES39_D06090 and NIES39_D06100)

encode RbcR paralogues.[64] Although *A. platensis* does not form a heterocyst, a gene for HetR (NIES39_C03480) has recently been found.[65] *Arthrospira platensis* contains 20 response regulators that have a so-called receiver domain at the N-terminus and the helix-turn-helix type DNA-binding domain at the C-terminus. Of these regulators, 16 are classified in the OmpR family whereas the remaining belong to the LuxR family. Orthologous proteins of NrrA (NIES39_A06330),[66] ManR (NIES39_G00800),[67,68] RpaA (NIES39_H00910),[69] SphR (NIES39_J04020),[70] RpaB (NIES39_K03840)[69] and NblR (NIES39_O05300)[69] were also identified.

3.4.4. PAS/GAF domains

Most non-marine cyanobacteria contain a large number of PAS and GAF domains.[54] The PAS and GAF domains are important signalling molecules that serve as specific sensors for light, redox, oxygen and many other signals or a structural element for dimerization.[71,72] We detected 131 PAS domains using a custom HMM profile of *Anabaena* and *Nostoc* PAS domains.[73] This number is almost comparable to those of other freshwater and soil cyanobacteria such as *Anabaena* and *Nostoc*. Six PAS domains encoded by NIES39_C01090, NIES39_C03530, NIES39_M00920, NIES39_L00190, NIES39_M02160 and NIES39_Q02220 are predicted to bind a flavin; the first three retain the conserved Cys residue that is essential for the photocycle of the LOV-type photoreceptors. One PAS domain encoded by NIES39_A03380 is predicted to bind a heme.

The *A. platensis* genome contains 58 GAF domains. Four GAF domains encoded by NIES39_M01980 (CyaB1) and NIES39_O00080 (CyaB2) are predicted to bind cAMP,[74] whereas 18 are categorized into the phytochrome-type or cyanobacteriochrome-type GAF domain subfamilies that bind linear tetrapyrrole and exhibit reversible photoconversion.[75] We found no Cph1 and AphA orthologues, which are cyanobacterial phytochromes, although bacteriophytochrome-type AphB (NIES39_A02210) and AphC (NIES39_C03450) orthologues are detected.[76] In addition, there are two green/red reversible AnPixJ-type (NIES39_J03990_GAF2 and NIES39_C00690_GAF1)[77] and six blue/green reversible TePixJ-type (NIES39_K04670_GAF1, NIES39_E01230_GAF1, NIES39_L03820_GAF1, NIES39_E04120_GAF3, NIES39_D04330_GAF1 and NIES39_M02490_GAF1) GAF domains.[75,78]

3.4.5. Chemotaxis regulators

In the *A. platensis* genome, there are eight methyl-accepting chemotaxis proteins (MCPs), which may serve as sensors for cell motility. In contrast with photoreceptor MCPs (e.g. PixJ) in many other cyanobacteria,[79] any of *Arthrospira* MCPs does not harbour any photoreceptor domain. Three of these MCP genes are clustered with other chemotaxis-regulating *che* genes such as *cheY*, *cheW* and *cheA* (NIES39_A07840−NIES39_A07910, NIES39_E01010−NIES39_E01070 and NIES39_H00230− NIES39_H00290 gene clusters). Two *che* clusters (NIES39_A07840−NIES39_A07910 and NIES39_H00230−NIES39_H00290) lack the *patA* genes which are present in the *che* clusters of other cyanobacteria. These *che* clusters contain additional *che* genes (*cheB*, *cheC* and *cheR* in the NIES39_A07840−NIES39_A07910 gene cluster and *cheB* and *cheR* in the NIES39_H00230−NIES39_H00290 gene cluster), which have not been detected in *che* clusters of other cyanobacteria such as *Synechocystis*, whose motility and phototaxis regulation is well understood. CheR and CheB are known to methylate and demethylate the MCP, respectively, and thereby function as a kind of molecular memory for flagellar regulation in some proteobacteria such as *E. coli*.[80] CheC is a phosphatase for the CheY response regulator.[81] As MCPs of these clusters have no photosensory domains, these genes may function in chemotaxis. Molecular memory and a rapid turnover system of phosphorylation can facilitate highly sensitive chemotactic responses.

3.4.6. Ser/Thr protein kinases and protein phosphatases

There are 43 genes that encode Ser/Thr protein kinases in the *A. platensis* genome. As in other filamentous cyanobacteria such as *Anabaena* and *Nostoc*,[54] HstK family (two), WD40-containing (nine) and TPR-containing (five) Ser/Thr kinases were detected.

Fourteen genes encoding phospho-Ser, -Thr and -Tyr phosphatases were also identified in the *A. platensis* genome. Among them, nine belong to the PP2C family including NIES39_O07130 and NIES39_R00570 that have response regulator domains. Other members of this family include NIES39_A03470, NIES39_C00770, NIES39_J01920, NIES39_J03440, NIES39_K03740, NIES39_M01820 and NIES39_M02650 (PphA) which dephosphorylates the nitrogen regulator PII protein.[82] In addition, two phospho-Tyr (NIES39_B01000 and NIES39_N00600) and three PP1/2A/2B-type (NIES39_D01240, NIES39_G01060 and NIES39_G01070) phosphatases were identified.

3.5. Genome protection

3.5.1. RM system

We detected three sets of type I RM systems (*hsdMSR*) in the *A. platensis* NIES-39 genome (Supplementary Table S10 and Supplementary Fig. S3) which are shared between a closely related strain reported by a Chinese group.[83] Among them, two RM-specific genes (*hsdS*, NIES39_A06660 and NIES39_C00340) are
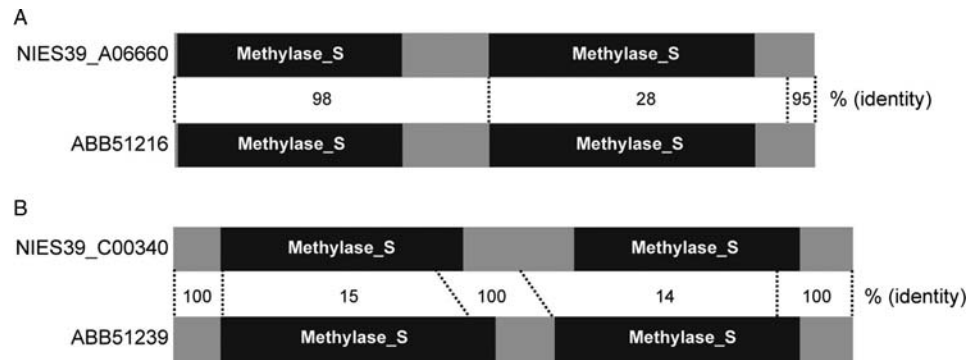
Figure 4. Domain architecture of HsdS proteins. The target recognition domain roughly corresponds to Methylase_S (PF01420).

mosaically conserved between the two strains (Fig. 4), whereas the full-length methylase and restriction genes *hsdM* and *hsdR* are highly conserved (>98%). HsdS protein is composed of two target recognition domains and the N-terminal, central and C-terminal regions. NIES39_A06660 is highly conserved with ABB51216 throughout the entire protein except for the second target recognition domain (Fig. 4A). NIES39_C00340 is highly conserved with ABB51239 at the N-terminal, central and C-terminal regions but not at the target recognition domains (Fig. 4B). These mosaically conserved HsdS proteins may be generated by one or two homologous recombinations through a mechanism similar to domain swapping of *hsdS* genes in other bacteria.[84] This enables *A. platensis* to acquire the restriction system against a wider range of exogenous DNA elements, interfering with genetic transformation of *A. platensis*. On the other hand, another *hsdS*-like reading frame, which is interrupted by an inframe stop codon, is totally conserved between these two strains.

NIES-39 harbours eight clusters of the type II RM systems, four of which are identical to those of the strain of the Chinese group,[83] although three are frame shifted. The other four clusters are novel in NIES-39. On the basis of this information of the RM systems, genetic transformation in *A. platensis* requires further improvement.

3.5.2. CRISPR system
CRISPRs are widespread in the genomes of many bacteria and almost all archaea, and they confer resistance to phages together with a group of CRISPR-associated (Cas) proteins.[85–87] The *A. platensis* genome contains three regions of CRISPRs, which are located in close proximity to Cas proteins. CRISPR1 (at coordinates 2897603−2897927 and 2903170−2904079) is organized as 17 almost identical sequences of 35 nt interspaced by non-identical sequences of 34−43 nt. CRISPR2 (at coordinates 3587555−3589797) and CRISPR3 (at coordinates 5189874−5191453) are organized as 29 and 23 identical sequences of 36 and 35 nt interspaced by non-identical sequences of 37−49 and 32−41 nt, respectively. We also found 18 Cas genes. One of the three Cas1 proteins encoded by NIES39_F00150 is a fusion protein, having a C-terminal Cas1 domain and a reverse transcriptase domain similar to that of group II introns.

3.6. Membrane transporters
The *A. platensis* genome contains ∼180 genes encoding putative membrane transporters. NapA-type Na^+/H^+ antiporters (Nha3 in *S. elongatus* PCC 7942) are known to be involved in salt tolerance at alkaline pH in some cyanobacteria.[88–90] *Arthrospira platensis* possesses seven putative Na^+/H^+ antiporters, one of which (NIES39_C00590) is an orthologue of NapA and Nha3.

Accumulation of bicarbonate in the cytoplasm is essential for photosynthesis under high pH conditions.[91] *Arthrospira platensis* possesses two sets of genes for CO_2-uptake modulation, NDH-1 (NdhF3-D3-CupA-CupS for high affinity and NdhF4-D4-CupB for low affinity), which converts CO_2 to bicarbonate in the cytoplasm (Supplementary Table S11). Genes *sbtA* and *bicA* encode the sodium-dependent bicarbonate transporter in *A. platensis*. Two closely related copies of *bicA* (NIES39_B00700, NIES39_B00710), which are tandemly arranged on the genome, were also found. There is only one operon for the ABC-type transporter for either nitrate (NrtA-B-C-D) or bicarbonate (CmpA-B-C-D).

3.7. Photosynthesis-related genes
3.7.1. Photosystem components
Almost all photosynthesis genes were detected in *A. platensis* (Supplementary Table S11). For photosystem II, the reaction centre D1 protein is encoded by at least three identical copies and one divergent copy of *psbA*. Additional variant genes include cytochrome c550-like (NIES39_J05860) and psb28-2 (NIES39_A01290). For photosystem I, all known genes including *psaX* were detected. Genes for

cytochrome b_6/f, ATP synthase, and NDH were detected as in other cyanobacteria. We detected two copies of cytochrome c_6, but no typical electron carrier proteins in the thylakoid lumen such as plastocyanin and cytochrome c_M.

3.7.2. Genes for tetrapyrrole biosynthesis

One set of single-copy genes involved in tetrapyrrole biosynthesis was identified in the *A. platensis* genome. Almost all genes involved in porphyrin biosynthesis were assigned. For siroheme and heme biosynthesis, uroporphyrinogen-III C-methyltransferase (NIES39_H00180) and two types of ferrochelatase (HemH: NIES39_K00630 and NIES39_M01510) were also identified. The latter encodes the so-called ferrochelatase-like protein possessing characteristic poly-His sequence at the C-terminus, which are also found in some cyanobacteria. For heme metabolism and subsequent phycobilin biosynthesis, heme oxygenase (HO: NIES39_Q00930), heme *o* synthase (NIES39_O03120), protoheme IX farnesyltransferase (NIES39_E02030) and biliverdin reductase (BvdR: NIES39_B00110) were assigned. For chlorophyll *a* biosynthesis, all genes were identified as single-copy genes. As in other cyanobacteria, isoforms catalyzing the same reaction, i.e. aerobic (HemF: NIES39_D01090) and anaerobic (HemN: NIES39_E03410) coproporphyrinogen III oxidase and light-dependent (Por: NIES39_R00680) and dark-operative (ChlL, ChlN and ChlB: NIES39_L05610, NIES39_L05630 and NIES39_L01300, respectively) protochlorophyllide oxidoreductase, were identified, whereas only the aerobic form of Mg-protoporphyrin IX monomethylester cyclase (AcsF/CrdI/CHL27: NIES39_B00770) was found. These isoforms were assumed to function in the adaption to different oxygen environments.

3.7.3. Genes for biosynthesis of carotenoid, lipid and others

All known cyanobacterial genes involved in carotenoid biosynthesis were found except for beta-carotene ketolase (CrtW or CrtO) in the *A. platensis* genome. However, 3-hydroxyechinenone, which is produced by ketolase, is reported to bind to an orange carotenoid protein (NIES39_N00720) to regulate energy dissipation from phycobilisome to photosystem II.[92,93] Therefore, a yet unknown type of ketolase may be present in the genome.

All known cyanobacterial genes involved in biosynthesis of glycerolipids, fatty acids, lipoic acid, vitamin E, lipopolysaccharides and polyhydroxyalkanoates were assigned. A gene for ω-3 fatty acid desaturase (*desB*) was not found, which is consistent with the biochemical analysis of *A. platensis*.[94]

The biosynthesis genes for cyanotoxins such as non-ribosomal peptide toxins (microcystins), ribosomal depsipeptide toxins (microviridins), alkaloid toxins (anatoxins and saxitoxins) and urea-derived toxins (cylindrospermopsin) were not present in the genome, which concurs with the long-time use of this organism as a food.

3.7.4. Carboxysome, reactive oxygen species protection and gas vesicles

A. platensis has eight genes for β-type carboxysome, which converts the accumulated bicarbonate ion to CO_2 for ribulose 1,5-bisphosphate carboxylase. These genes are split into two operons (*ccmK1-K2* and *ccmK3-K4-L-M-N-O*). Only one gene (*ccmM*) encodes carbonic anhydrase for conversion of bicarbonate to CO_2.

We also detected some enzymes which protect against reactive oxygen species, except for catalase. However, the *A. platensis* genome harbours many genes for thioredoxin peroxidase, peroxiredoxin and other putative peroxidases including five genes for bacterioferritin co-migratory proteins (NIES39_A03490, NIES39_E01510, NIES39_E02230, NIES39_L02380, NIES39_O06760).

Genus *Arthrospira* possesses gas vesicles for buoyancy of its trichomes, which is in contrast with no gas vesicles in the morphologically similar genus *Spirulina*.[95] Consequently, *A. platensis* has six gas vesicle genes (*gvpA-C-N-J, gvpV* and *gvpW*); GvpA is a structural protein to form the skeleton of the gas vesicle.

3.8. Cell surface proteins and motility

3.8.1. Cell surface and extracellular proteins

Concerning the cell surface structure and extracellular proteins, 42 genes were found to encode putative extracellular proteins that contain 191 haemolysin-like Ca^{2+}-binding domains in the *A. platensis* genome (Table 2). This number is much higher than that in many other cyanobacteria. Fifteen of these genes consist of merely haemolysin-like Ca^{2+}-binding domains, whereas 14 possess an additional SBBP-like domain, which is often found in S-layer-related proteins (i.e. NIES39_B00690, NIES39_C00400, NIES39_D06770, NIES39_D06910, NIES39_D06930, NIES39_D06980, NIES39_D07000, NIES39_D07020, NIES39_Q00770, NIES39_Q00840, NIES39_Q02140, NIES39_Q02160, NIES39_Q02170 and NIES39_Q02190).[96] NIES39_D06770 also possesses the CalX-beta domain which is present in the cytoplasmic domains of CalX Na^+/Ca^{2+} exchangers to expel calcium from cells,[97] and NIES39_Q02140, NIES39_Q02160, NIES39_Q02170 and NIES39_Q02190 possess both the PPC domain found at the C-terminus of secreted bacterial peptidases and the CalX-beta domain. Three other genes (NIES39_C03760, NIES39_F00640 and NIES39_J05690) also possess the PPC domain. Other genes possess domains for cadherin (NIES39_C05180),
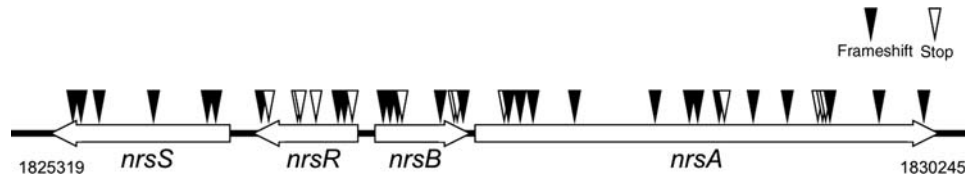
Figure 5. Selective mutations in the *nrsS/R/B/A* region.

SCP (NIES39_D00540, NIES39_D06960 and NIES39_L04630), Lipase-GDSL (NIES39_E00490), pro-isomerase (NIES39_L02390) and Chlam-PMP (NIES39_L00320 and NIES39_M01680); NIES39_L00320 also possesses the CalX-beta domain. Although pathogenic haemolysins are well known in Enterobacteria,[98,99] it is interesting that non-toxic *A. platensis* has a large number of genes that contain the haemolysin-like Ca^{2+}-binding domain.

The *A. platensis* genome has no genes coding siderophore to scavenge iron from the environment.

3.8.2. Gliding motility The molecular mechanism for gliding motility has been poorly understood in cyanobacteria. The outer surface of gliding cyanobacteria in Oscillatoriaceae consists of a parallel, helically arranged protein array (S-layer).[100,101] The S-layer protein, oscillin, required for gliding motility in *Phormidium*[102] is conserved in *A. platensis* (NIES39_A01430, 42% identity). NIES39_A01430 consists of 19 haemolysin-like Ca^{2+}-binding domains. This protein is also homologous to SwmA (27% identity) for swimming motility in unicellular *Synechococcus* sp. WH 8102.[103,104]

Gliding motility of cyanobacteria is reported to be driven by a secretion of mucilage from the junctional pore complex (JPC).[105] The JPC consists of a channel through the outer membrane and the peptidoglycan layer. Although the subunits of JPC have not been identified, comparative genomics analysis may reveal their genes.

3.8.3. Type IV pilus-related proteins Type IV pili are involved in twitching motility in many bacteria[79,106–108] and have recently been reported to function in gliding motility in *N. punctiforme*[109] and twitching motility in unicellular *Synechocystis*.[110] *A. platensis* exhibits vigorous gliding motility and harbours type IV pilus-related genes (Supplementary Table S11), but shows no twitching motility, suggesting that this organism utilizes type IV pili as machinery for gliding motility like *N. punctiforme*.[109] Typical *pil* genes for pili biogenesis are the *pilMNOQ* operon, *pilB1T1C* operon, *pilD* (prepilin leader peptidase) and three prepilin subunit genes (NIES39_C03030, NIES39_C00840 and NIES39_

C00850). Additionally, *A. platensis* harbours the rather unusual *pilT2*-like (NIES39_A08210) and *pilC*-like (NIES39_A03250) genes. These copies may be involved in novel regulation of the type IV pili or type II secretion system, which has not yet been studied in cyanobacteria.

3.9. Microevolution

Arthrospira platensis NIES-39 harbours many pseudogenes, whose coding regions fragmentally correspond to known genes or ORFs (Supplementary Table S12). Many of them are not just accidental mutants resulting from single mutations but are generated by at least two events, frameshift and/or nonsense mutations. An extreme case is the operon of *nrsS/R/B/A*, which is inactivated by at least 27 frameshift mutations and 13 nonsense mutations (Fig. 5). NrsS/R involve the two-component signal transduction system for nickel sensing and transcriptional regulation, whereas NrsB/A form a cation-efflux pump to remove heavy metals.[111] The extensive degradation of its operon strongly suggests that *A. platensis* NIES-39 is in the process of microevolution to adapt to its specific location in Lake Chad.

Rapid evolution is also noted in the *hsdS* gene which determines the RM specificity (see Section 3.5.1). Since many strains of *A. platensis* and *A. maxima* are the target of genome projects and economical applications, the microevolution of their genomes and diversity in phenotype are critical themes for post-genome studies.

References

1. Desikachary, T.V. 1973, In: Carr, N.G. and Whitton, B.A. (eds.), *Botanical Monographs*, Blackwell Scientific

Publications: Oxford, London, Edinburgh, Melbourne, pp. 473−81.

2. Doolittle, W.F. 1982, In: Carr, N.G. and Whitton, B.A. (eds.), *The Biology of Cyanobacteria, Botanical Monographs*, University of California Press: Berkeley and Los Angels, pp. 307−31.

3. Van Tuerenhout, D. 2005, *The Aztecs: New Perspectives*, ABC-CLIO.

4. Belay, A. 2007, In: Gershwin, M.E. and Belay, A. (eds.), *Spirulina in Human Nutrition and Health*, CRC press: Florida, pp. 1−26.

5. Castenholz, R.W., Rippka, R., Herdman, M. and Wilmotte, A. 2007, In: Boone, D.R., Castenholz, R.W. and Garrity, G.M. (eds.), *Bergey's Manual of Systematic Bacteriology*, 2nd edition. Springer: Berlin, pp. 542−3.

6. Kasai, F., Kawachi, M., Erata, M., et al. 2009, NIES-collection list of strains 8th edition, *Jpn. J. Phycol.*, **57**, 162.

7. Vonshak, A. 1997, *Spirulina platensis (Arthrospira): Physiology, Cell-Biology, and Biotechnology*, Taylor & Francis Group: London.

8. Hoiczyk, E. 2000, Gliding motility in cyanobacteria: observations and possible explanations, *Arch. Microbiol.*, **174**, 11−7.

9. Zeng, M.T. and Vonshak, A. 1998, Adaptation of *Spirulina platensis* to salinity stress, *Comp. Biochem. Physiol. A*, **120**, 113−8.

10. Kebede, E. 1997, Response of *Spirulina platensis* (*Arthrospira fusiformis*) from Chitu, Ethiopia, to salinity stress from sodium salts, *J. Appl. Phycol.*, **9**, 551−8.

11. Ohmori, K., Ehira, S., Kimura, S. and Ohmori, M. 2009, Changes in the amount of cellular trehalose, the activity of maltooligosyl trehalose hydrolase, and the expression of its gene in response to salt stress in the cyanobacterium *Spirulina platensis*, *Microb. Environ.*, **24**, 52−6.

12. Ohmori, K. and Ohmori, M. 2002, cAMP stimulates Na^+-dependent ATP formation in the alkalophylic cyanobacterium *Spirulina platensis*, *Microb. Environ.*, **17**, 144−7.

13. Yashiro, K., Sakamoto, T. and Ohmori, M. 1996, Molecular characterization of an adenylate cyclase gene of the cyanobacterium *Spirulina platensis*, *Plant Mol. Biol.*, **31**, 175−81.

14. Kasahara, M., Unno, T., Yashiro, K. and Ohmori, M. 2001, CyaG, a novel cyanobacterial adenylyl cyclase and a possible ancestor of mammalian guanylyl cyclases, *J. Biol. Chem.*, **276**, 10564−9.

15. Loseva, L.P. and Dardynskaya, I.V. 1993, *Spirulina* natural sorbent of radionucleides, 6th International Congress of Applied Algology, Czech Republic.

16. Amao, Y. and Nakamura, N. 2006, Biohydrogen production with the light-harvesting function of grana from *Spirulina* and colloidal platinum, *Int. J. Hydro. Energy*, **29**, 39−42.

17. Kawata, Y., Yano, S., Kojima, H. and Toyomizu, M. 2004, Transformation of *Spirulina platensis* strain C1 (*Arthrospira* sp. PCC 9438) with Tn5 transposase−transposon DNA-cation liposome complex, *Mar. Biotechnol.*, **6**, 355−63.

18. 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.

19. Ogawa, T. and Terui, G. 1970, Studies on the growth of *Spirulina platensis*, *J. Ferment. Technol.*, **48**, 361−7.

20. Sekine, M., Tanikawa, S., Omata, S., et al. 2006, Sequence analysis of three plasmids harboured in *Rhodococcus erythropolis* strain PR4, *Environ. Microbiol.*, **8**, 334−46.

21. Ewing, B. and Green, P. 1998, Base-calling of automated sequencer traces using phred. II. Error probabilities, *Genome Res.*, **8**, 186−94.

22. Ewing, B., Hillier, L., Wendl, M.C. and Green, P. 1998, Base-calling of automated sequencer traces using phred. I. Accuracy assessment, *Genome Res.*, **8**, 175−85.

23. Chen, Q., Savarino, S. and Venkatesan, M. 2006, Subtractive hybridization and optical mapping of the enterotoxigenic *Escherichia coli* H10407 chromosome: isolation of unique sequences and demonstration of significant similarity to the chromosome of *E. coli* K-12, *Microbiology*, **152**, 1041−54.

24. Reslewic, S., Zhou, S., Place, M., et al. 2005, Whole-genome shotgun optical mapping of *Rhodospirillum rubrum*, *Appl. Environ. Microbiol.*, **71**, 5511−22.

25. Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A. and Eddy, S. 2003, Rfam: an RNA family database, *Nucleic Acids Res.*, **31**, 439−41.

26. 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**, 955−64.

27. Lagesen, K., Hallin, P., Rodland, E., Staerfeldt, H.-H., Rognes, T. and Ussery, D. 2007, RNAmmer: consistent and rapid annotation of ribosomal RNA genes, *Nucleic Acids Res.*, **35**, 3100−8.

28. 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**, 3389−402.

29. Delcher, A.L., Harmon, D., Kasif, S., White, O. and Salzberg, S.L. 1999, Improved microbial gene identification with GLIMMER, *Nucleic Acids Res.*, **27**, 4636−41.

30. Salzberg, S.L., Delcher, A.L., Kasif, S. and White, O. 1998, Microbial gene identification using interpolated Markov models, *Nucleic Acids Res.*, **26**, 544−8.

31. The UniProt, C 2009, The Universal Protein Resource (UniProt) 2009, *Nucleic Acids Res.*, **37**, D169−74.

32. Yada, T. and Hirosawa, M. 1996, Detection of short protein coding regions within the cyanobacterium genome: application of the hidden Markov model, *DNA Res.*, **3**, 355−61.

33. Hunter, S., Apweiler, R., Attwood, T.K., et al. 2009, InterPro: the integrative protein signature database, *Nucleic Acids Res.*, **37**, D211−5.

34. Sato, N. 2009, Gclust: trans-kingdom classification of proteins using automatic individual threshold setting, *Bioinformatics*, **25**, 599−605.

35. Furumichi, M., Sato, Y., Omata, T., Ikeuchi, M. and Kanehisa, M. 2002, CYORF: community annotation of cyanobacteria genes, *Genome Inform.*, **13**, 402−3.

36. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. and Hattori, M. 2004, The KEGG resource for deciphering the genome, *Nucleic Acids Res.*, **32**, D277−80.

37. Bendtsen, J.D., Nielsen, H., von Heijne, G. and Brunak, S. 2004, Improved prediction of signal peptides: SignalP 3.0, *J. Mol. Biol.*, **340**, 783−95.

38. Krogh, A., Larsson, B.R., von Heijne, G. and Sonnhammer, E. 2001, Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes, *J. Mol. Biol.*, **305**, 567−80.

39. Sasaki, V.N. and Sato, N. 2010, CyanoClust: comparative genome resources of cyanobacteria and plastids, *Database*, doi: 10.1093/database/BAP025.

40. Bauer, C.C., Buikema, W.J., Black, K. and Haselkorn, R. 1995, A short-filament mutant of *Anabaena* sp. strain PCC 7120 that fragments in nitrogen-deficient medium, *J. Bacteriol.*, **177**, 1520−6.

41. Mullineaux, C.W., Mariscal, V., Nenninger, A., et al. 2008, Mechanism of intercellular molecular exchange in heterocyst-forming cyanobacteria, *EMBO J.*, **27**, 1299−308.

42. Wong, F.C. and Meeks, J.C. 2001, The *hetF* gene product is essential to heterocyst differentiation and affects HetR function in the cyanobacterium *Nostoc punctiforme*, *J. Bacteriol.*, **183**, 2654−61.

43. Buikema, W.J. and Haselkorn, R. 1991, Characterization of a gene controlling heterocyst differentiation in the cyanobacterium *Anabaena* 7120, *Genes Dev.*, **5**, 321−30.

44. Zhang, W., Du, Y., Khudyakov, I., et al. 2007, A gene cluster that regulates both heterocyst differentiation and pattern formation in *Anabaena* sp. strain PCC 7120, *Mol. Microbiol.*, **66**, 1429−43.

45. Nakamura, Y., Kaneko, T., Sato, S., et al. 2002, Complete genome structure of the thermophilic cyanobacterium *Thermosynechococcus elongatus* BP-1, *DNA Res.*, **9**, 123−30.

46. Meng, Q., Zhang, Y. and Liu, X.Q. 2007, Rare group I intron with insertion sequence element in a bacterial ribonucleotide reductase gene, *J. Bacteriol.*, **189**, 2150−4.

47. Ohmori, M. and Okamoto, S. 2004, Photoresponsive cAMP signal transduction in cyanobacteria, *Photochem. Photobiol. Sci.*, **3**, 503−11.

48. Katayama, M. and Ohmori, M. 1997, Isolation and characterization of multiple adenylate cyclase genes from the cyanobacterium *Anabaena* sp. strain PCC 7120, *J. Bacteriol.*, **179**, 3588−93.

49. Phalip, V., Li, J.H. and Zhang, C.C. 2001, HstK, a cyanobacterial protein with both a serine/threonine kinase domain and a histidine kinase domain: implication for the mechanism of signal transduction, *Biochem. J.*, **360**, 639−44.

50. Ohmori, K., Hirose, M. and Ohmori, M. 1993, An increase in the intracellular concentration of cAMP triggers formation of an algal mat by the cyanobacterium *Spirulina platensis*, *Plant Cell Physiol.*, **34**, 169−71.

51. Romling, U. and Simm, R. 2009, Prevailing concepts of c-di-GMP signaling, *Contrib. Microbiol.*, **16**, 161−81.

52. Stock, J.B., Ninfa, A.J. and Stock, A.M. 1989, Protein phosphorylation and regulation of adaptive responses in bacteria, *Microbiol. Rev.*, **53**, 450−90.

53. 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.

54. Ohmori, M., Ikeuchi, M., Sato, N., et al. 2001, Characterization of genes encoding multi-domain proteins in the genome of the filamentous nitrogen-fixing cyanobacterium *Anabaena* sp. strain PCC 7120, *DNA Res.*, **8**, 271−84.

55. Ashby, M.K. and Houmard, J. 2006, Cyanobacterial two-component proteins: structure, diversity, distribution, and evolution, *Microbiol. Mol. Biol. Rev.*, **70**, 472−509.

56. Yoshimura, H., Okamoto, S., Tsumuraya, Y. and Ohmori, M. 2007, Group 3 sigma factor gene, *sigJ*, a key regulator of desiccation tolerance, regulates the synthesis of extracellular polysaccharide in cyanobacterium *Anabaena* sp. strain PCC 7120, *DNA Res.*, **14**, 13−24.

57. Osanai, T., Ikeuchi, M. and Tanaka, K. 2008, Group 2 sigma factors in cyanobacteria, *Physiol. Plant.*, **133**, 490−506.

58. Wu, J., Zhao, F., Wang, S., et al. 2007, cTFbase: a database for comparative genomics of transcription factors in cyanobacteria, *BMC Genomics*, **8**, 104.

59. 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**, 643−7.

60. 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**, 5840−7.

61. McGinn, P., Price, D., Maleszka, R. and Badger, M. 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**, 218−29.

62. Nakamoto, H., Suzuki, M. and Kojima, K. 2003, Targeted inactivation of the *hrcA* repressor gene in cyanobacteria, *FEBS Lett.*, **549**, 57−62.

63. 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**, 6241−5.

64. 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**, 576−83.

65. Zhang, J.Y., Chen, W.L. and Zhang, C.C. 2009, *hetR* and *patS*, two genes necessary for heterocyst pattern formation, are widespread in filamentous nonheterocyst-forming cyanobacteria, *Microbiology*, **155**, 1418−26.

66. Ehira, S. and Ohmori, M. 2006, NrrA, a nitrogen-responsive response regulator facilitates heterocyst development in the cyanobacterium *Anabaena* sp. strain PCC 7120, *Mol. Microbiol.*, **59**, 1692−703.

67. Ogawa, T., Bao, D.H., Katoh, H., Shibata, M., Pakrasi, H.B. and Bhattacharyya-Pakrasi, M. 2002, A two-component signal transduction pathway regulates manganese homeostasis in *Synechocystis* 6803, a photosynthetic organism, *J. Biol. Chem.*, **277**, 28981−6.

68. Yamaguchi, K., Suzuki, I., Yamamoto, H., et al. 2002, A two-component Mn^{2+}-sensing system negatively regulates expression of the *mntCAB* operon in *Synechocystis*, *Plant Cell*, **14**, 2901−13.

69. Ashby, M.K. and Mullineaux, C.W. 1999, Cyanobacterial *ycf27* gene products regulate energy transfer from phycobilisomes to photosystems I and II, *FEMS Microbiol. Lett.*, **181**, 253−60.

70. Aiba, H. and Mizuno, T. 1994, A novel gene whose expression is regulated by the response-regulator, SphR, in response to phosphate limitation in *Synechococcus* species PCC 7942, *Mol. Microbiol.*, **13**, 25−34.

71. Ponting, C.P. and Aravind, L. 1997, PAS: a multifunctional domain family comes to light, *Curr. Biol.*, **7**, R674−7.

72. Aravind, L. and Ponting, C.P. 1997, The GAF domain: an evolutionary link between diverse phototransducing proteins, *Trends Biochem. Sci.*, **22**, 458−9.

73. Narikawa, R., Okamoto, S., Ikeuchi, M. and Ohmori, M. 2004, Molecular evolution of PAS domain-containing proteins of filamentous cyanobacteria through domain shuffling and domain duplication, *DNA Res.*, **11**, 69−81.

74. Kanacher, T., Schultz, A., Linder, J.U. and Schultz, J.E. 2002, A GAF-domain-regulated adenylyl cyclase from *Anabaena* is a self-activating cAMP switch, *EMBO J.*, **21**, 3672−80.

75. Ikeuchi, M. and Ishizuka, T. 2008, Cyanobacteriochromes: a new superfamily of tetrapyrrole-binding photoreceptors in cyanobacteria, *Photochem. Photobiol. Sci.*, **7**, 1159−67.

76. Okamoto, S., Kasahara, M., Kamiya, A., Nakahira, Y. and Ohmori, M. 2004, A phytochrome-like protein AphC triggers the cAMP signaling induced by far-red light in the cyanobacterium *Anabaena* sp. strain PCC7120, *Photochem. Photobiol.*, **80**, 429−33.

77. Narikawa, R., Fukushima, Y., Ishizuka, T., Itoh, S. and Ikeuchi, M. 2008, A novel photoactive GAF domain of cyanobacteriochrome AnPixJ that shows reversible green/red photoconversion, *J. Mol. Biol.*, **380**, 844−55.

78. Rockwell, N.C., Njuguna, S.L., Roberts, L., et al. 2008, A second conserved GAF domain cysteine is required for

the blue/green photoreversibility of cyanobacteriochrome Tlr0924 from *Thermosynechococcus elongatus*, *Biochemistry*, **47**, 7304−16.

79. Yoshihara, S., Suzuki, F., Fujita, H., Geng, X.X. and Ikeuchi, M. 2000, Novel putative photoreceptor and regulatory genes required for the positive phototactic movement of the unicellular motile cyanobacterium *Synechocystis* sp. PCC 6803, *Plant Cell Physiol.*, **41**, 1299−304.

80. Baker, M.D., Wolanin, P.M. and Stock, J.B. 2006, Signal transduction in bacterial chemotaxis, *Bioessays*, **28**, 9−22.

81. Rao, C.V., Glekas, G.D. and Ordal, G.W. 2008, The three adaptation systems of *Bacillus subtilis* chemotaxis, *Trends Microbiol.*, **16**, 480−7.

82. Irmler, A. and Forchhammer, K. 2001, A PP2C-type phosphatase dephosphorylates the PII signaling protein in the cyanobacterium *Synechocystis* PCC 6803, *Proc. Natl Acad. Sci. USA*, **98**, 12978−83.

83. Zhao, F., Zhang, X., Liang, C., Wu, J., Bao, Q. and Qin, S. 2006, Genome-wide analysis of restriction-modification system in unicellular and filamentous cyanobacteria, *Physiol. Genomics*, **24**, 181−90.

84. O'Sullivan, D., Twomey, D.P., Coffey, A., Hill, C., Fitzgerald, G.F. and Ross, R.P. 2000, Novel type I restriction specificities through domain shuffling of HsdS subunits in *Lactococcus lactis*, *Mol. Microbiol.*, **36**, 866−75.

85. Haft, D.H., Selengut, J., Mongodin, E.F. and Nelson, K.E. 2005, A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes, *PLoS Comput. Biol.*, **1**, e60.

86. Barrangou, R., Fremaux, C., Deveau, H., et al. 2007, CRISPR provides acquired resistance against viruses in prokaryotes, *Science*, **315**, 1709−12.

87. Sorek, R., Kunin, V. and Hugenholtz, P. 2008, CRISPR—a widespread system that provides acquired resistance against phages in bacteria and archaea, *Nat. Rev. Microbiol.*, **6**, 181−6.

88. Wang, H.L., Postier, B.L. and Burnap, R.L. 2002, Polymerase chain reaction-based mutageneses identify key transporters belonging to multigene families involved in Na^+ and pH homeostasis of *Synechocystis* sp. PCC 6803, *Mol. Microbiol.*, **44**, 1493−506.

89. Laloknam, S., Tanaka, K., Buaboocha, T., et al. 2006, Halotolerant cyanobacterium *Aphanothece halophytica* contains a betaine transporter active at alkaline pH and high salinity, *Appl. Environ. Microbiol.*, **72**, 6018−26.

90. Billini, M., Stamatakis, K. and Sophianopoulou, V. 2008, Two members of a network of putative Na+/H+ antiporters are involved in salt and pH tolerance of the freshwater cyanobacterium *Synechococcus elongatus*, *J. Bacteriol.*, **190**, 6318−29.

91. Kaplan, A., Hagemann, M., Bauwe, H., Kahlon, S. and Ogawa, T. 2008, In: Herrero, A. and Flores, E. (eds.), *The Cyanobacteria: Molecular Biology, Genomics and Evolution*, Caister: UK, pp. 305−34.

92. Wilson, A., Ajlani, G., Verbavatz, J.M., Vass, I., Kerfeld, C.A. and Kirilovsky, D. 2006, A soluble carotenoid protein involved in phycobilisome-related

energy dissipation in cyanobacteria, *Plant Cell*, **18**, 992−1007.

93. Kerfeld, C.A., Sawaya, M.R., Brahmandam, V., et al. 2003, The crystal structure of a cyanobacterial water-soluble carotenoid binding protein, *Structure*, **11**, 55−65.

94. Murata, N., Wada, H. and Gombos, Z. 1992, Modes of fatty-acid desaturation in cyanobacteria, *Plant Cell Physiol.*, **33**, 933−41.

95. Guglielmi, G., Rippka, R. and Tandeau De Marsac, N. 1993, Main properties that justify the different taxonomic position of *Spirulina* spp. and *Arthrospira* spp. among cyanobacteria, *Bull. Inst. Oceanogr.*, **12**, 13−23.

96. Smith, D.R., Doucette-Stamm, L.A., Deloughery, C., et al. 1997, Complete genome sequence of *Methanobacterium thermoautotrophicum* deltaH: functional analysis and comparative genomics, *J. Bacteriol.*, **179**, 7135−55.

97. Schwarz, E.M. and Benzer, S. 1997, Calx, a Na-Ca exchanger gene of *Drosophila melanogaster*, *Proc. Natl Acad. Sci. USA*, **94**, 10249−54.

98. Wiles, T., Kulesus, R. and Mulvey, M. 2008, Origins and virulence mechanisms of uropathogenic *Escherichia coli*, *Exp. Mol. Pathol.*, **85**, 11−9.

99. Hahn, H. and Specht, U. 2003, Secretory delivery of recombinant proteins in attenuated *Salmonella* strains: potential and limitations of Type I protein transporters, *FEMS Immunol. Med. Microbiol.*, **37**, 87−98.

100. Hoiczyk, E. and Baumeister, W. 1995, Envelope structure of four gliding filamentous cyanobacteria, *J. Bacteriol.*, **177**, 2387−95.

101. Read, N., Connell, S. and Adams, D. 2007, Nanoscale visualization of a fibrillar array in the cell wall of filamentous cyanobacteria and its implications for gliding motility, *J. Bacteriol.*, **189**, 7361−6.

102. Hoiczyk, E. and Baumeister, W. 1997, Oscillin, an extracellular, Ca^{2+}-binding glycoprotein essential for the gliding motility of cyanobacteria, *Mol. Microbiol.*, **26**, 699−708.

103. McCarren, J. and Brahamsha, B. 2007, SwmB, a 1.12-megadalton protein that is required for nonflagellar swimming motility in *Synechococcus*, *J. Bacteriol.*, **189**, 1158−62.

104. McCarren, J., Heuser, J., Roth, R., Yamada, N., Martone, M. and Brahamsha, B. 2005, Inactivation of *swmA* results in the loss of an outer cell layer in a swimming *Synechococcus* strain, *J. Bacteriol.*, **187**, 224−30.

105. Hoiczyk, E. 1998, Structural and biochemical analysis of the sheath of *Phormidium uncinatum*, *J. Bacteriol.*, **180**, 3923−32.

106. 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.

107. Terauchi, K. and Ohmori, M. 2004, Blue light stimulates cyanobacterial motility via a cAMP signal transduction system, *Mol. Microbiol.*, **52**, 303−9.

108. Okamoto, S. and Ohmori, M. 2002, The cyanobacterial PilT protein responsible for cell motility and transformation hydrolyzes ATP, *Plant Cell Physiol.*, **43**, 1127−36.

109. Duggan, P.S., Gottardello, P. and Adams, D.G. 2007, Molecular analysis of genes in *Nostoc punctiforme* involved in pilus biogenesis and plant infection, *J. Bacteriol.*, **189**, 4547−51.

110. Yoshihara, S. and Ikeuchi, M. 2004, Phototactic motility in the unicellular cyanobacterium *Synechocystis* sp. PCC 6803, *Photochem. Photobiol. Sci.*, **3**, 512−8.

111. Lopez-Maury, L., Garcia-Dominguez, M., Florencio, F.J. and Reyes, J.C. 2002, A two-component signal transduction system involved in nickel sensing in the cyanobacterium *Synechocystis* sp. PCC 6803, *Mol. Microbiol.*, **43**, 247−56.