ELSEVIER

FEMS Microbiology Letters 215 (2002) 209–219

FEMS MICROBIOLOGY Letters

www.fems-microbiology.org

Analysis of the *hli* gene family in marine and freshwater cyanobacteria

Devaki Bhaya [a],*, Alexis Dufresne [b], Daniel Vaulot [b], Arthur Grossman [a]

[a] *Department of Plant Biology, Carnegie Institution of Washington, 260 Panama Street, Stanford, CA 94305, USA*
[b] *Station Biologique, UMR 7127, CNRS, INSU et Université Pierre et Marie Curie, 29682 Roscoff Cedex, France*

Received 23 July 2002; received in revised form 15 August 2002; accepted 16 August 2002

First published online 11 September 2002

Abstract

Certain cyanobacteria thrive in natural habitats in which light intensities can reach 2000 μmol photon m^{-2} s^{-1} and nutrient levels are extremely low. Recently, a family of genes designated *hli* was demonstrated to be important for survival of cyanobacteria during exposure to high light. In this study we have identified members of the *hli* gene family in seven cyanobacterial genomes, including those of a marine cyanobacterium adapted to high-light growth in surface waters of the open ocean (*Prochlorococcus* sp. strain Med4), three marine cyanobacteria adapted to growth in moderate- or low-light (*Prochlorococcus* sp. strain MIT9313, *Prochlorococcus marinus* SS120, and *Synechococcus* WH8102), and three freshwater strains (the unicellular *Synechocystis* sp. strain PCC6803 and the filamentous species *Nostoc punctiforme* strain ATCC29133 and *Anabaena* sp. {*Nostoc*} strain PCC7120). The high-light-adapted *Prochlorococcus* Med4 has the smallest genome (1.7 Mb), yet it has more than twice as many *hli* genes as any of the other six cyanobacterial species, some of which appear to have arisen from recent duplication events. Based on cluster analysis, some groups of *hli* genes appear to be specific to either marine or freshwater cyanobacteria. This information is discussed with respect to the role of *hli* genes in the acclimation of cyanobacteria to high light, and the possible relationships among members of this diverse gene family.

1. Introduction

The major peripheral light-harvesting complex (LHC) of cyanobacteria is the water-soluble phycobilisome, which is comprised of tetrapyrrole-bound phycobiliproteins and non-pigmented linker polypeptides [1]. In vascular plants, the major LHC is composed primarily of the integral membrane Lhc polypeptides that contain three transmembrane helices and bind chlorophylls *a* and *b*. The Lhc polypeptides are encoded by a family of genes that generally contains more than 10 individuals [2–4]. However, there are more distantly related *Lhc* genes that comprise the *Lhc* extended gene family. Polypeptides encoded by this extended family include the early light-inducible proteins (ELIPs), the four transmembrane helix-containing polypeptide PsbS or PSII-S [5], and polypeptides that have one or two putative transmembrane helices [6,7]. Several genes have been identified on cyanobacterial genomes that encode single-helix members of the *Lhc* extended gene family [8,9]. These genes have been designated *hli* (high light inducible; protein designation HLIPs) [8] or *scp* (small cab-like proteins; protein designation Scps) [9]. While *hli* genes were first noted in *Synechococcus* sp. strain PCC7942 [10], they were subsequently identified in other cyanobacteria [11], red algae [12] and vascular plants [6].

The pattern of expression of *hli* genes in cyanobacteria and vascular plants is similar to that of the genes encoding ELIPs; *hli* mRNAs and encoded polypeptides accumulate under conditions that result in the absorption of excess excitation energy, including exposure to high irradiance, nitrogen starvation and low temperature [1,6,8]. *Synechocystis* sp. strain PCC6803 deleted for all four of its *hli* genes was shown to be photosensitive, and under strong illumination the cells lost all variable fluorescence and died [13]. By analogy to vascular plant ELIPs, a number of functions have been suggested for the cyanobacterial HLIPs. They may associate with pigments, perhaps transiently, serving as chlorophyll carriers [11], function in the dissipation of excess absorbed light energy within antennae complexes [6,14], or modulate the biosynthesis of chlorophyll [15]. Essentially all evidence suggests that photo-

* Corresponding author. Tel.: +1 (650) 325 1521, ext. 282; Fax: +1 (650) 325 6857.
E-mail address: devaki@andrew2.stanford.edu (D. Bhaya).

0378-1097/02/$22.00 © 2002 Federation of European Microbiological Societies. Published by Elsevier Science B.V. All rights reserved.
PII: S0378-1097(02)00934-5

synthetic organisms need HLIPs under stressful, often growth-limiting conditions.

Specific *Prochlorococcus* species thrive in high-light, nutrient-poor environments that characterize the surface waters of the open oceans, while others grow at greater depths in lower-light and higher-nutrient environments [16,17]. Although phylogenetic analyses, based on 16S rDNA and *rpoC* sequences, have established that *Prochlorococcus* is a cyanobacterial genus, it does not contain the light harvesting phycobilisomes typical of most cyanobacteria [18]. Instead, the major antennae pigment complex contains chlorophylls *a* and *b* associated with polypeptides that are similar to CP43, a polypeptide integral to the core of photosystem II that binds chlorophyll *a* [19].

An understanding of the light and nutrient habitats in which specific cyanobacterial strains thrive, and recent acquisition of complete or near complete sequence information for several cyanobacterial genomes, have made it attractive to explore both intra- and inter-species relationships among cyanobacterial *hli* genes. The genomes of seven cyanobacterial strains have been sequenced at the Kazusa DNA Research Institute, Japan; the Joint Genome Institute (JGI), USA and the Genoscope, France, and the sequences of other cyanobacterial genomes are nearly complete. Three of these cyanobacteria, the unicellular *Synechocystis* strain PCC6803 (**SC**), the filamentous *Anabaena (Nostoc)* sp. strain PCC7120 (**AN**) and *Nostoc punctiforme* strain ATCC29133 (**NT**) grow in freshwater habitats. The marine species for which complete genome information is available are represented by the high-light ecotype *Prochlorococcus marinus* strain MED4 (**PM**) and three species that grow in low/moderate light, *P. marinus* strain MIT9313 (**PL**), *P. marinus* strain SS120 (**PS**) and *Synechococcus* sp. strain WH8102 (**SN**) [17,20]. In this report we analyze the cyanobacterial *hli* gene family and discuss the differences in this gene family among the seven different cyanobacterial strains for which complete genome sequences are available.

2. Materials and methods

2.1. Genome data

Sequences of the genomes of *Synechocystis* and *Anabaena* were downloaded from the EMBL web site (http://www.ebi.ac.uk/genomes/); those of the *Prochlorococcus* strains MED4 (version of 12/19/2001), MIT9313 (version of 01/25/2002), and *Synechococcus* strain WH8102 (version of 01/26/2002) from ftp sites of the JGI (ftp://ftp.jgi-psf.org/pub/JGI_data/Microbial/prochlorococcus/final.011129; ftp://ftp.jgi-psf.org/pub/JGI_data/Microbial/prochlorococcusII/final.010823; ftp://ftp.jgi-psf.org/pub/JGI_data/Microbial/synechococcus/final.010910). Contig sequences of the genome of *N. punctiforme* (version of 01/25/2002) were downloaded from the JGI ftp site: (ftp://ftp.jgi-psf.org/pub/JGI_data/Microbial/nostoc/010409/). The genomic sequence of *Prochlorococcus* SS120 was downloaded from http://www.sb-roscoff.fr/Phyto/ProSS120. The data has been provided freely by the US DOE Joint Genome Institute for use in this publication/correspondence only.

2.2. Gene detection

The *hli* genes were identified by similarity searches using the four *hli* genes of SC as query sequences (gene identifiers: ssl1633{*hliC*}, ssr1789{*hliD*}, ssl2542{*hliA*}, ssr2595{*hliB*} in Cyanobase) against complete cyanobacterial genome sequences translated in their six reading frames using the tBLASTn program [21]. Because of the very small size of these genes, the tBLASTn program was operated with a default *E*-value threshold of 10 and no filter for low-complexity regions. In a second step, a multiple alignment of the *hli* genes was performed using ClustalX software [22] in order to build a profile Hidden Markov Model (profile HMM) with the hmmbuild program of the HMMER 2 package (http://hmmer.wustl.edu/) [23]. This profile HMM was calibrated with the hmmcalibrate program and used with the hmmsearch program (HMMER 2 package) to identify *hli* genes that were not detected by tBLASTn. All three of these programs were operated with the default options. Using the HMMER package five new *hli* genes were detected in PL (*hli05, 06, 07, 08, 09*) and two in SN (*hli07* and *hli08*).

2.3. Sequence clustering

Clustering of *hli* genes was achieved using the GeneRAGE algorithm (http://www.ebi.ac.uk/research/cgg/services/rage/) [24], which groups sequences based on their similarity. We chose to use BLASTp to detect similarity between Hli polypeptide sequences. The choice of a threshold value is critical to obtain the optimal ratio between specificity and sensitivity. To determine the optimal threshold, an initial all-against-all comparison of protein sequences was made using BLASTp. After analysis of these results, an *E*-value cut-off of 10^{-13} was chosen as the threshold.

2.4. Conservation of regions neighboring hli genes

To examine whether regions surrounding the *hli* genes were conserved among the different genomes, orthologous relationships between open reading frames (ORFs) flanking the *hli* genes were investigated. Initially, ORFs flanking the *hli* genes were identified using the Artemis software package (http://www.sanger.ac.uk/Software/Artemis/). Each of these flanking ORFs was then compared against the complete cyanobacterial genome sequences using tBLASTn. Whenever the *hli* genes of two different genomes were found to have similar neighboring ORFs, the neighboring ORFs were used for reciprocal comparisons using tBLASTn; ORFs that satisfied the criterion of reciprocal best hit for each other were considered to be orthologs.

3. Results and discussion

Multiple *hli* genes are present on the genomes of all seven cyanobacterial strains examined in this study (Table 1). Based on this analysis, the four marine cyanobacteria PM, PL, PS and SN have 22, 9, 13 and 8 *hli* genes, respectively, while the three freshwater cyanobacteria SC, AN and NT have 4, 8 and 9 *hli* genes, respectively. These 73 *hli* genes (Table 1) were analyzed to help understand the significance of the large number of *hli* genes on the genome (especially in PM), to determine possible relationships among the different *hli* genes and to evaluate whether the related clusters of *hli* genes represent group-specific (e.g. present only in the marine cyanobacteria or *Prochlorococcus* species etc.) and/or possible functional classes.

The genome of PM, the high-light ecotype of marine *Prochlorococcus*, encodes at least 22 *hli* genes. The number of *hli* genes on the PM genome is significantly greater than the number present on the genomes of the low-light ecotypes PL and PS (which have nine and 13 *hli* genes, respectively), the marine *Synechococcus* SN (which has eight *hli* genes) and the freshwater species SC, AN and NT (which have between four and nine *hli* genes). In SC, the *hli* gene family is required for survival in high light. A mutant of SC lacking one or two copies of the *hli* gene survives, but it is at a disadvantage relative to wild-type cells as evaluated by growth competition experiments performed in high light. If all four of the SC *hli* genes are disrupted, the cells die following exposure to high light [13]. These results suggest that HLIPs act in a cumulative manner to sustain the cells in high light, although there may also be requirements for particular gene products under specific environmental conditions. The significant increase in the number of *hli* genes on the genome of the *Prochlorococcus* high-light ecotype, in spite of the fact that this strain has the smallest genome amongst the seven genomes analyzed, may reflect a requirement for the additional gene products in coping with the persistent high-light conditions associated with the ocean surface [17]. Conversely, the smaller number of *hli* genes in PL, PS, SN and the freshwater strains, which are adapted to low/moderate-light growth conditions, is consistent with an important role for HLIPs in habitats in which the organisms are under persistent excitation pressure [25,26].

The arrangement of 22 *hli* genes in PM is shown in Fig. 1; the *hli* genes are scattered throughout the genome, with some gene clustering in particular regions. There are two instances in which two *hli* genes are contiguous (*hli*11 and *hli*12, and *hli*21 and *hli*22). In addition, two other regions of the genome contain four tandemly arranged *hli* genes (*hli*06-09 shown as A and *hli*16-19 shown as B in Fig. 1). The four tandemly arranged genes in region A are flanked by *hli*05 and *hli*10 (these genes are 4.5 kb and 3.7 kb distant from region A, respectively). Strikingly, the two clusters of tandemly arranged genes each cover 1.3 kbp and represent exact duplications; i.e. the sequence of *hli*06-09 is identical to that of *hli*16-19 at the nucleotide level. The duplicated 1.3-kbp region is comprised exclusively of four *hli* genes plus a small ORF of 59 codons (Fig. 1). This ORF has no similarity to other ORFs in the public databases, and thus may not represent a protein product. If this is the case, the duplication has resulted in the exclusive doubling of just the *hli* genes. A cursory examination of the PM genome indicates that this 1.3-kb region is the only exact duplication in excess of 1 kb in the genome. Interestingly, while two other *hli* genes (*hli*04 and *hli12*) are identical in their predicted amino acid sequences, they are not identical at the nucleotide level.

Since identification of the duplication of the *hli* gene clusters was based on genome sequence information, it was possible that the exact nucleotide match observed

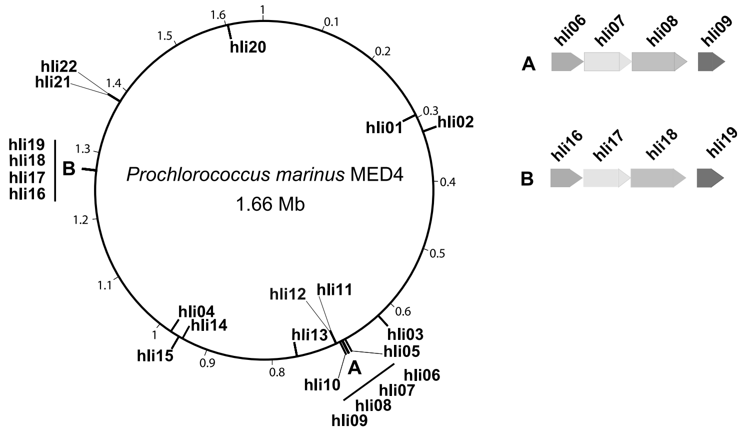

Fig. 1. Position of the 22 *hli* genes detected in the genome of *P. marinus* strain MED4. Gene groups A and B correspond to exactly duplicated regions containing the four *hli* genes (*hli*06-09 and *hli*16-19). The arrangement of the duplicated genes in groups A and B are shown on the right. Note that *hli*06, *hli*07 and *hli*08 (group A) and *hli*16, *hli*17 and *hli*18 (group B) are overlapping (see text for details).

Table 1
hli genes detected in seven fully sequenced cyanobacterial genomes

Genome	Gene	Start	Stop	Size (aa)	GeneRAGE cluster
PM	hli01	292 721	292 575	49	6
PM	hli02	320 885	321 115	77	8
PM	hli03	634 018	634 170	51	7
PM	hli04	713 298	713 194	35	14
PM	hli05	697 543	697 695	51	12
PM	hli06	702 237	702 341	35	14
PM	hli07	702 344	702 550	69	10
PM	hli08	702 553	702 813	87	12
PM	hli09	702 844	702 969	42	15
PM	hli10	706 726	706 851	42	16
PM	hli11	713 054	713 191	46	12
PM	hli12	983 410	983 306	35	14
PM	hli13	777 735	777 469	89	17
PM	hli14	962 255	962 106	50	10
PM	hli15	968 466	968 720	85	24
PM	hli16	1 274 111	1 274 215	35	14
PM	hli17	1 274 218	1 274 427	70	10
PM	hli18	1 274 427	1 274 687	87	12
PM	hli19	1 274 718	1 274 843	42	15
PM	hli20	1 600 841	1 600 662	60	5
PM	hli21	1 388 883	1 389 014	44	12
PM	hli22	1 389 022	1 389 276	85	15
PL	hli01	413 647	413 847	67	5
PL	hli02	72 191	72 409	73	6
PL	hli03	228 052	228 201	50	7
PL	hli04	120 778	121 041	88	8
PL	hli05	746 073	745 927	49	9
PL	hli06	572 474	572 674	67	10
PL	hli07	744 236	744 114	41	11
PL	hli08	572 319	572 459	47	12
PL	hli09	572 833	572 937	35	13
PS	hli01	1 388 013	1 388 129	39	15
PS	hli02	1 489 980	1 489 831	50	7
PS	hli10	1 387 372	1 387 512	47	19
PS	hli11	1 387 549	1 387 653	35	14
PS	hli12	1 397 334	1 397 438	35	14
PS	hli13	1 387 253	1 387 369	39	20
SN	hli01	489 613	489 413	67	21
SN	hli02	1 840 235	1 840 432	66	8
SN	hli03	2 114 476	2 114 625	50	7
SN	hli04	2 299 673	2 299 924	84	2
SN	hli05	1 389 330	1 389 683	118	22
SN	hli06	1 152 224	1 152 361	46	6
SN	hli07	403 771	403 553	73	23
SN	hli08	830 619	830 443	59	5
SC	hli01	701 350	701 138	71	2
SC	hli02	982 968	983 180	71	2
SC	hli03	398 188	398 361	58	3
SC	hli04	1 141 803	1 142 015	71	4
AN	hli01	1 006 982	1 006 782	67	1
AN	hli02	607 714	607 499	72	2
AN	hli03	2 836 843	2 836 676	56	3
AN	hli04	6 277 367	6 277 543	59	4
AN	hli05	3 686 003	3 686 203	67	1
AN	hli06	3 686 251	3 686 451	67	1
AN	hli07	4 499 702	4 499 526	59	4
AN	hli08	531 645	531 526	40	4
NT[a]	hli01	81 354	81 175	60	4
NT[a]	hli02	20 142	20 354	71	2
NT[a]	hli03	181 687	181 854	56	3
NT[a]	hli04	15 331	15 122	70	2
NT[a]	hli05	77 669	77 878	70	1
NT[a]	hli06	77 923	78 123	67	1

Table 1 (*Continued*).

Genome	Gene	Start	Stop	Size (aa)	GeneRAGE cluster
NT[a]	hli07	238 527	238 324	68	1
NT[a]	hli08	95 265	95 089	59	4
NT[a]	hli09	107 200	107 024	59	4
PS	hli03	1 426 441	1 426 352	30	18
PS	hli04	118 186	118 332	49	6
PS	hli05	88 528	88 283	82	8
PS	hli06	1 297 723	1 297 986	88	17
PS	hli07	1 097 580	1 097 786	69	10
PS	hli08	1 097 458	1 097 562	35	14

Start and stop positions for each gene are specified. Abbreviations used are: PM, *P. marinus* strain MED4; PL, *P. marinus* strain MIT9313; PS, *P. marinus* strain SS120; SN, *Synechococcus* sp. strain WH8102; SC, *Synechocystis* sp. strain PCC6803; AN, *Anabaena* sp. strain PCC7120; NT, *N. punctiforme* strain ATCC29133. Marine cyanobacteria are in bold. For each gene the size of the corresponding protein and the GeneRAGE cluster number is shown.

[a] The Nostoc sequence is incomplete, so start and stop data represent information from individual contigs (hli01: contig 483; hli02: contig 397; hli03: contig 507; hli04: contig 485; hli05: contig 480; hli06: contig 480; hli07: contig 509; hli08: contig 476; hli09: contig 486).

was generated by a computational artifact that yielded improper assembly results. To establish whether or not the putative duplication was an artifact, primers within the duplicate regions paired with specific primers flanking the duplications were constructed and used for PCR. The results of these experiments demonstrated the presence of the duplicate sequences at two distinct genomic locations (data not shown, Stephanie Stillwagen, DOE JGI, Walnut Creek, CA, USA, personal communication). Identical sequences within these *hli* gene clusters may reflect a very

Table 2
Cluster analysis of the 73 *hli* genes using the GeneRAGE program (see Section 2)

Cluster	Genome	Gene
1	AN	hli01
1	AN	hli05
1	AN	hli06
1	NT	hli05
1	NT	hli07
2	SN	hli04
2	SC	hli01
2	SC	hli02
2	AN	hli02
2	NT	hli02
2	NT	hli04
3	SC	hli03
3	AN	hli03
3	NT	hli03
4	SC	hli04
4	AN	hli04
4	AN	hli07
4	AN	hli08
4	NT	hli01
4	NT	hli08
4	NT	hli09
5	PM	hli20
5	PS	hli09
5	PL	hli01
5	SN	hli08
6	PM	hli01
6	PS	hli04
6	PL	hli02
6	SN	hli06
7	PM	hli03
7	PS	hli02
7	PL	hli03
7	SN	hli03
8	PM	hli02
8	PS	hli05
8	PL	hli04
8	SN	hli02
9	PL	hli05
10	PM	hli07/17
10	PM	hli14
10	PS	hli07
10	PL	hli06
11	PL	hli07
12	PM	hli05
12	PM	hli08/18
12	PM	hli11
12	PM	hli21
12	PL	hli08
13	PL	hli09
14	PM	hli04/12
14	PM	hli06/16
14	PS	hli08/11
14	PS	hli12
15	PM	hli09/19
15	PM	hli22
15	PS	hli01
16	PM	hli10
17	PM	hli13
17	PS	hli06
18	PS	hli03
19	PS	hli10
20	PS	hli13
21	SN	hli01
22	SN	hli05
23	SN	hli07
24	PM	hli15

Genes were aligned using the BIOEDIT program. Marine cyanobacteria are in bold.

recent duplication of this locus, or the occurrence of a copy correction mechanism in the cell that maintains identity between the two sequences (although there is no experimental evidence to support such a mechanism). As more bacterial genomes are being sequenced and analyzed, it will be interesting to examine them for the presence of exact sequence duplications and gene duplications. The significance and mechanisms for creation and maintenance of these duplications is not yet understood although genome-wide studies suggest that duplicate genes are subject to specific selection and may not evolve at the same rate [27].

The arrangement of the clustered genes in the 1.3-kbp duplicate region is striking; the last nucleotide of the stop codon (TAA) for *hli06* and *hli07* is the first nucleotide of the start codon (ATG) of *hli07* and *hli08*, respectively (Fig. 1). This type of overlapping gene arrangement was demonstrated to be important for coordinating the expression of the overlapping *trpA* and *trpB* genes in the *Escherichia coli trp* operon [28]. Other examples of translational coupling have also recently been noted in *Prochlorococcus* MED4 (*phoB–PhoR*) and in the *pta–ack* bicistronic operon of *Corynebacterium glutamicum* [16,29]. Analysis of expression from the *hli* gene clusters would identify populations of polycistronic mRNAs that are transcribed from these genes, and may reveal how environmental conditions influence both the levels and distribution of distinct polycistronic transcripts.

The small size of the *hli* genes, the finding that some members of the gene family represent exact duplications, and the somewhat low degree of conservation among the different Hli proteins make classical sequence-based phylogenetic approaches difficult. In particular, bootstrap values of phylogenetic trees obtained by various methods (e.g. maximum likelihood) are very low (data not shown), raising serious concerns about the robustness of the observed relationships. To gain insights into the relationships among the 73 putative *hli* genes, we used the GeneRAGE program (Table 2). This program is a robust algorithm for quickly and accurately clustering large protein datasets into families and subfamilies [24]. Although no single program can give definitive answers regarding the relationship between genes and organisms, it does set the stage for a more complete analysis and provides information for hypothesis generation and evaluation. A number of observations resulting from these analyses are discussed below.

The 73 *hli* genes were separated into 24 clusters (these clusters contain up to seven genes, although 11 of the clusters contain a single gene representative). The clustering analysis clearly shows a strong divergence between marine and freshwater species (Fig. 2). This may indicate an early separation of the marine and freshwater cyanobacteria and the generation of divergent *hli* gene clusters within these environmentally distinct groups. However the strong divergence may also reflect very distinct evolutionary pressures that are associated with the markedly different environments in which these organisms are able to thrive.

A

FRESHWATER SPECIES

Cluster1

```
AN_hli01  ---MELY-PTDKTETA--YNGKDRNAFEFGFTPQSELWNGRLAMLGFLAYLLWDLNGYSVVRDVLHLVAYNAG 67
AN_hli05  .....TRST..LPKV.TE...V.....L..WN....I.......I..........A....L......IG.     67
AN_hli06  ....QTRPS..LPPV.PA...V.....L..W.....I.......I..........A....L......IR.     67
NT_hli06  .....TRSS..LPPV.KA...V.....L..W...A.I......AI...G......A....L.....IIS.     67
NT_hli07  MAT-.TNKVVKLSTE.KA...V....WI..WN..Q.........I..VS......A...LL......FR      68
NT_hli05  MATQ.TRSS..LPPV.PE...V.....L..W...A.I......AI..........A....L.......G.     70
```

Cluster2

```
SN_hli04  MAQTPSTDAPVIRGATVTTE-DGGRLNAFASEPRMQVVEAEQGWGFHERAEKLNGRMAMLGFIALLATEIALG-GEAFTHGLLG-LG 84
SC_hli01  -----------MTTRGFRLDQ.N-...N..I..EVY.DSSV.-A.WTKY...M...F..I..AS..IM.VVT.H.VI---.W.NS.  70
SC_hli02  -----------MTSRGFRLDQ.N-...N..I..PVY.DSSV.-A.WT.Y...M...F..I..VS...M.VIT.H.IV---.W.LS.  70
NT_hli04  -----------MTNKGF.IN.ER.Q..R..I..KIY.D.TP-RI..T.Y.......L..I...S.I.L.VFT.N.LI---.W.TSF  70
AN_hli02  -----------M.TNNAIVD..Q.LM.N..I..KVY.D.QGDRT..TPY..I....L..I...S.I.L.VFT.K.IF---...TN.Q 72
NT_hli02  -----------M.TS.AIID..Q.K..N..I..KVYID.QGDRT..TPY..M....L..I...S.I.L.VFT.H.IV---.V.AN.  71
```

Cluster3

```
SC_hli03  MSEELQPNQTPVQEDPKFGFNNYAEKLNGRAAMVGFLLILVIEYFTNQGVLAWLGLR 57
AN_hli03  ..QT-..TV..KL.E......E...R.......I....MV....A......S....K 56
NT_hli03  .TQT-..TI..KL.E......E...R.......I..A.M.....V......S....K 56
```

Cluster4

```
AN_hli08  -------------------------------MGFNHQSESWNGRLAMIGFLAAIAIEFFSGQGFLHFWNILIL 42
NT_hli01  ..........MTNASTTKVTTPVIEDRNAWRW..TP.A.I..........S..VLV.L..........G..   61
AN_hli04  ..........MTD..TTKISASVVEDRNSWRW..TP.A.I.............TL..L..........G..D  60
NT_hli09  ..........MAD..VKKTTGSVPEDPNALRW..TP...N....F......S.V.L.V.....I....G..   59
NT_hli08  ..........MTGFK..NPAPIVSEDPNAVRF..TP...N...........S..L..A.....L....G..   59
AN_hli07  ..........MSGFK...........PNAVRF..TSE.......F......SIVL..A..........G..   50
SC_hli04  MGAILCYIYLHRQPSQLVITFLTMNNENSK.F..TAFA.N..........SS.LIL.LV....V...FG..   70
```

Fig. 2. Alignment of *hli* genes in specific clusters of freshwater species (A) and marine species (B) using the GeneRAGE program. Dots mark residues that are identical to the top sequence in each cluster and dashes represent gaps. Clusters that have only a single representative are not shown.

3.1. Fresh water species

Clusters 1–4 contain all 22 *hli* genes of the freshwater species, AN, NT and SC (Fig. 2A and Table 2). Of these, clusters 2, 3 and 4 contain at least one representative from each species; while cluster 1 contains three genes each from NT and AN (AN_*hli 01*, AN_*hli05* AN_*hli06* and NT_*hli05*, NT_*hli06*, NT_*hli07*). Cluster 2 contains two representatives from NT and SC each (NT_ *hli02* and NT_*hli04*, SC_*hli01* and SC_*hli02*) and one from AN (*hli02*) and SN *(hli04*), cluster 3 contains a single representative from each of the freshwater species (SC_*hli03*, AN_*hli03*, NT_*hli03*) and cluster 4 contains seven genes, with three each from NT and AN and one from SC (*hli*04). Based on nearest neighbor analysis, AN_*hli02* and NT_*hli*02 (both in cluster 2), AN_*hli04* and NT_*hli01* (both in cluster 4), and AN_*hli03* and NT_*hli03* (both in cluster 3) all share neighboring genes, as do NT_*hli04* and SC_*hli02* (both in cluster 2) (Fig. 3). These results suggest the following features of the *hli* gene families:

1. There may be three basic *hli* gene 'forms' in the freshwater species represented by the sequences in clusters 2–4. Furthermore, strong sequence similarity among pairs of polypeptides representative of the different freshwater species, encoded by genes within these groups, combined with nearest neighbor analyses with respect to these genes, maybe suggestive of orthologous relationships among specific members of these gene clusters. Whether the genes in clusters 2, 3, and 4 have distinct functions is still unclear. Recent evidence suggests that a severe phenotype is associated only with a mutant in which all four *hli* genes are inactivated;

B

MARINE SPECIES

Cluster 5

```
PL_hli01  ---------MASESPLDSNTSAEPVS--SEELNAWRRGFTPQAEIWNGRMAMAGLIIGISVLLLLRLVMPADCRAWLN 67
SN_hli08  ..........-.Q-.EK-.GGVAEPVG.D.....K............L..I..SA.LA.V..V.VF-AGN       59
PS_hli09  .MNSQSTNKEKK..---------TQ.VEKS.....K..............SI...L.LI..I.INKFYG         60
PM_hli20  .........KK..KINLK--ETKKVVDKQ...L.K..............TI.IG.ILI.IA.ISKF-SSI       60
```

Cluster 6

```
PM_hli01  -------------------------MNEDN-QPRFGFVNFAETWNGRMAMMGILIGLGTELITGQSILRQIGIG 48
PL_hli02  MRIYCHQDGQAISMLCYIDEILESP.A-.KP...............L....VI...S...L......S.M.L. 73
PS_hli04  .........................SPEDIE..Y....Y..I....L..L..V...S...L...G..G...F. 49
SN_hli06  .........................S-...-A.............L....FV.......L...G..S...L. 46
```

Cluster 7

```
PM_hli03  MIKPDIVPKRKLPRYGFHFYNEKLNGRMAMIGFIALILTELFLKHGLLLW 50
PS_hli02  ..D.K.I.E....S....NHT.N....W........VIV.FK.G..I.IR 50
PL_hli03  .LE.T.I.Q.RK......SH..........L.....MVV.AT.G....I. 50
SN_hli03  .LE.TDI.Q.R...F...GHT......A..L.....LAV.IK.G....I. 50
```

Cluster 8

```
PL_hli04  MTSPKQNLPGDQDLPSEQAVFEGSEQGSESSEVQPPINSATTGDPPTFGWSAYAERVNGRFAMIGLAAVLLIEVVSRDTFVHWAGLVS 88
SN_hli02  ....SEPPATASVPET----------------------....S.V.A....G..........V.FT.I.V..AI.G...L.....LP 66
PS_hli05  ...-NNPELSKVESSKSESQENND.TNDVQMTP-----....P.I.S....G............FI.I....TI.KSG.L......P 82
PM_hli02  ...--NQEQNN.EAMELEKTNSEEIKIE.Q--------.IEIE.RYE....N.S.IT......L.FL.II...LI.QKS.LN...IF  77
```

Cluster 10

```
PM_hli07/17  MSNSSYT--TTESGGRQNMFPSETRPYIDESVSYDGYPQNAEKVNGRWAMIGFVALLGAYVTTGQIIPGIF 69
PM_hli14     .A..---QV.........................S..K............................... 68
PL_hli06     .TS.--.NVI..D......YA..P.MQ..PE--.TAFSKE..LA...G.....LSAVV..LF....L.... 67
PS_hli07     .TS.AQAQI.....N......V.AQ.QLV.N--.S..IED...A.........I......L.S........ 69
```

Cluster 12

```
PM_hli05     ----------------------------------MNSKKVKVLETKTVEKEKVVAEKLNGRFAMIGFIAAIGAYLTTGQIIPGFV 51
PL_hli08     ..................................----MK.TPK.NR..NQ.LT..RV..MA..M..W..V...........V. 47
PM_hli11     ..................................-----M.NN.P.L.....I...........M..V.LV.............I 46

                                                                                        6
```

Cluster 14

```
PM_hli04/12  MTPEAERFNGWAAMLGFVAAVGAYVTTGQIIPGWF 35
PM_hli06/16  ...D......L.......................F. 35
PS_hli08/11  ......K.............F...A........I. 35
PS_hli12     ...Q..K.......I...C...S.A........I. 35
```

Cluster 15

```
PM_hli09/19  -----------------------------------------MENSKPNYWQNAERTNGRMAMMGFFALVVNYGLFGWIIPGIF 42
PS_hli01     ..........................................N.N---..TI.......L..I.L...II...F........Y 39
PM_hli22     MSPLTGFIIVVIAITLQFTLYTIKRLQEPLDPNLFDSQKSPK.N.R.KSF.K...I....L..V.LL.......F.......FI 84
```

Cluster 17

```
PM_hli13  MKEEKPPL-KNSDNSPTENLKEETNNTSSDNEYSKWVDNQGDEVKDVFGFNSSAELVNGRAAMIGFLMLLLTELVFKGRPVTSSIFGIN 88
PS_hli06  .RM.DNLNQ..EEDRFD...IGSRKEITGTSD-A.....NDN..TQ.....EN.....S.......I..I....I.N.K...L...... 88
```

Fig. 2 (*Continued*).

mutants lacking one or two *hli* genes do not exhibit this phenotype which may be indicative of a redundancy or overlapping gene functions [13].

2. AN and NT have multiple copies of closely related *hli* genes within clusters 1 and 4 (and cluster 2 for NT). However, in SC there appears to have been only one recent duplication (*hli01* and *hli02* for which the encoded amino acid sequences are 87% identical; 94% similar). Clusters 4 and 1 have three representatives each from AN and NT, but one or none from SC, respectively. The significance of this is unclear since all three species grow in relatively low-light environments. However, both NT and AN are multi-cellular and developmentally complex (both species can differentiate nitrogen-fixing heterocysts which contain a highly modified photosynthetic apparatus that does not evolve oxygen). It would be particularly interesting to follow expression patterns of the *hli* genes under different growth conditions (e.g. low-nitrogen, high-light) as well as to examine regions upstream of the *hli* coding regions to identify conserved sequence elements.
3. Only one marine cyanobacterial HLIP sequence, SN_*hli04*, clusters with the sequences from the freshwater organisms (within cluster 2). The amino acid similarity between SN_*hli04* and the most closely related gene in cluster 2, NT_*hli02*, is 43%. The significance of this clustering is not apparent at this time.
4. A comparison of sequences in NT, AN and SC demonstrates that there is little conservation of genes that flank the *hli* genes between the filamentous and unicellular cyanobacteria (Fig. 3). There is evidence that there have been rearrangements of genomes in some freshwater cyanobacteria, including SC and there is also evidence for recent transposition events in SC [30,31]. This makes it less surprising that neighborhood conservancy is not maintained between NT/AN and SC.

3.2. Marine species

Clusters 5–8 all contain a single representative from each of the four marine species analyzed (Fig. 2B). Furthermore, these genes also all share flanking gene neighbors (Fig. 3). This may indicate that these four gene clusters are essential for all marine species (somewhat similar to the case in freshwater species where there are three clusters that have at least one representative from each of the species). Although the phylogenetic relationships among the gene groupings are difficult to evaluate, based on sequence similarity, the marine gene clusters 6–8 are possibly most closely related to the freshwater gene clusters 2 and 3. It is unclear if the apparent similarities between these groups reflect the specific functions of the proteins, but it would be interesting to explore this possibility by examining the influence of diverse environmental conditions on the expression of genes within these clusters.

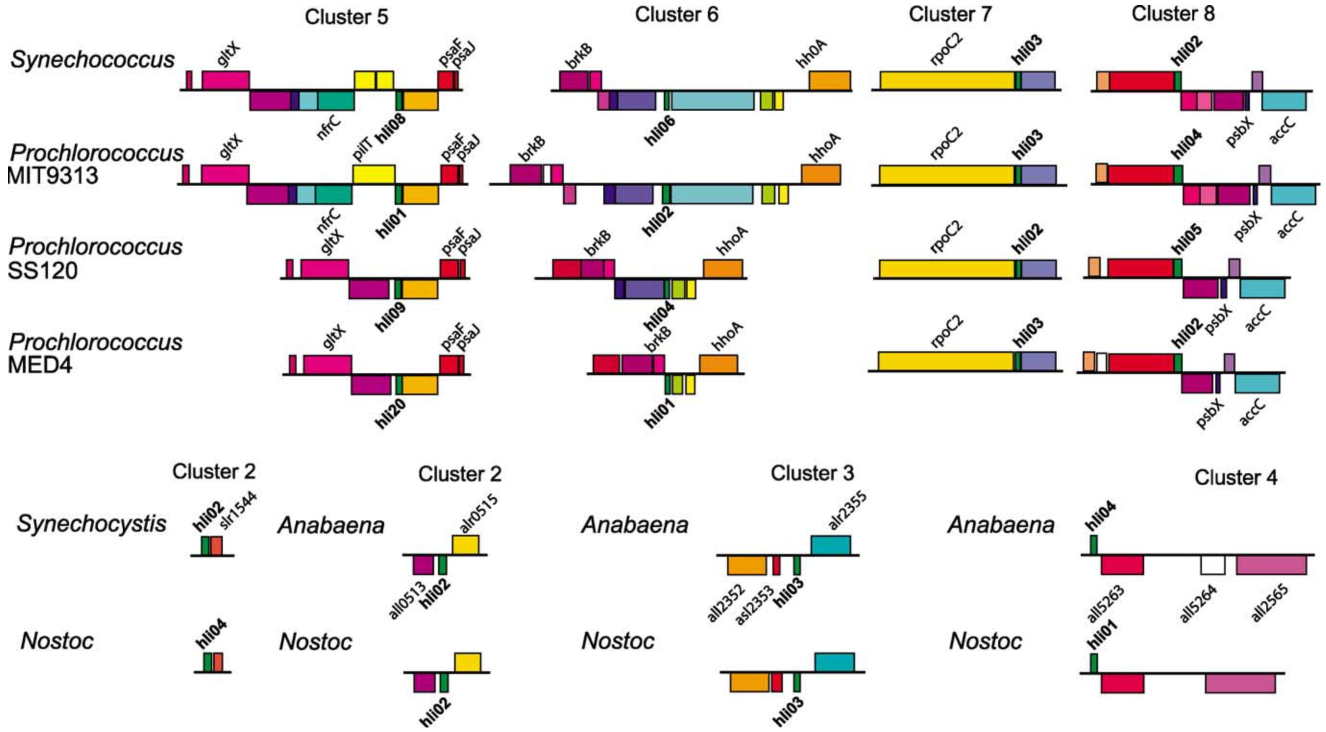

Fig. 3. Conserved genes in the neighborhood of *hli* genes arranged according to the GeneRAGE clusters. ORFs that are painted with the same color are considered orthologs. For clusters 5–8 only genes for which the assignment of a gene name is unambiguous have been labeled (e.g. *rpoC2*). For clusters 2–4 the Cyanobase Gene Identifier for flanking genes of *Synechocystis* or *Anabaena* have been labeled for identification of flanking genes.

Five clusters (10, 12, 14, 15 and 17) contain multiple genes from the marine species but not all of the species are represented in these clusters (Fig. 2B). Furthermore, 11 clusters (9, 11, 13, 16 and 18–24) each contain only one representative gene. This indicates that there are several marine *hli* genes (three from PL {*hli05, hli07, hli09*}; two from PM {*hli10* and *hli15*}; three from PS {*hli3, hli10* and *hli13*}; three from SN {*hli01, hli05* and *hli07*}) that have diverged to the point of not being grouped together using the GeneRAGE program. Clusters with multiple family members from one species may represent an evolutionary trend toward duplication of specific genes; however until more experimental evidence is available, it is not possible to draw any direct conclusions based simply on sequence similarities.

In Fig. 4, 23 out of the 44 *hli* genes within the *Prochlorococcus* species (PL, PM and PS) are aligned using ClustalX. It is quite striking that within this group the C-terminus of the HLIPs maintain a strongly conserved motif TGQIIPGF/IF. This motif is not conserved in clusters 5, 6, 7 and 8 (which contain one representative from each of the marine cyanobacterial species) or in clusters 17 and 18. It is also notably missing in all of the freshwater strains. The fact that this motif is only present in a subset of the *Prochlorococcus* HLIPs may be indicative of specialized function.

3.3. Concluding remarks

We have attempted an analysis of the large *hli* gene family that is ubiquitous in all cyanobacterial species examined so far (as well as in other groups). This study was motivated by the initial observation that there was an apparent over-representation of *hli* genes in PM, the high-light adapted marine strain. The presence of a very large *hli* gene family in PM is consistent with recent results showing that a mutant of SC lacking all four copies of the *hli* gene was unable to survive in high light [13], raising the obvious question of whether the number of *hli* genes in an organism could be correlated with adaptation/acclimation to the light environment.

To attempt to answer this question, we took a bioinformatics approach to analyze the 73 *hli* genes identified in seven recently sequenced cyanobacterial strains. There are a number of problems associated with the use of small genes to construct a phylogeny. The construction of cyanobacterial phylogenies, is particularly problematic since there are issues related to the ancient history of this group; over the course of evolution cyanobacteria may have experienced both lateral gene transfer and the formation of gene mosaics [32,33]. Furthermore, it is even difficult to determine which bacteria are most closely related to cyanobacteria; some sister groups are considered to be the *Deinococcales* and spirochetes and more recent analyses suggests a relationship with low GC Gram-positive bacteria such as *Halobacterium* and *Aquifex aeolicus* [34–36]. To avoid these potential pitfalls we used a clustering analysis method (GeneRAGE) that does not make any assumptions about phylogenetic relatedness between genes.

One of the most obvious results from the analyses presented above suggests that there is a significant distinction between *hli* genes in the marine and freshwater strains. Since there has not yet been an extensive analysis of genes across various cyanobacterial species, it is useful to compare this data in the context of some recent molecular phylogenetic studies of various cyanobacterial species using 16S rRNA sequence data [37,38]. Honda et al. used 16S rRNA sequences from a variety of 44 different freshwater and marine strains to determine evolutionary lineages within the cyanobacteria. Based on maximum likelihood and neighborhood joining trees generated with

```
PM_hli04/12  ---------------------------------------------------MTPEAERFNGWAAMLGFVAAVGAYVTTGQIIPGWF
PM_hli05     ...................................MNSKKVKVLETKTVEKEKVV..KL..RF..I..I..I...L........FV
PM_hli06/16  ......................................................D......L......................F.
PM_hli07/17  .................MSNSSYTTTESGGRQNMFPSETRPYIDESVSYDGYPQN..KV..RW..I....LL............I.
PM_hli08/18  MSPLAVFLILIVSLTALLVASLTKQFQEENLIYSNKNQMTNSNTKTKTIEKEKVV..TL..RF..I.LI......L........FV
PM_hli09/19  ............................................MENSKPNYWQN...T..RM..M..F.L.VN.GLF.W....I.
PM_hli10     ............................................MEFVKKF..EK..K...K.....MF.LI...YF....V..I.
PM_hli11     ........................................MKNNEPKLVEKEKIV..KL..RF..M....L....L........FI
PM_hli14     ..................MANSQVTTESGGRQNMFPSETRPYIDESVSYDSYPKN..KV..RW..I....LL............I.
PM_hli15     ..MIEKKGDNIRSENFYPDSNYYLDQDNTPEETTLPEDQIFNTKKFEWPNSYWFI...T..RL..I..M.VIIN.TLF.W.AYPIL
PM_hli21     ..........................................MAKIKSVEKEKIV..KL..RF..I..I......L........FV
PM_hli22     ..MSPLTGFIIVVIAITLQFTLYTIKRLQEPLDPNLFDSQKSPKMNNRKKSFWKN..IT..RL..V.LL.L.VN.GFF.W....FI
PS_hli01     ...............................................MNNNYWTI...T..RL..I.LF.LIIN.GFF.W....IY
PS_hli07     .................MTSSAQAQITTESGNRQNMFPVEAQPQLVENYSGYIED..KA..RW..I..I.LL...L.S......I.
PS_hli08/11  .........................................................K.............F...A........I.
PS_hli10     .......................................MKTSTSSTKVETSKVL..KI..R..LI.VI.LL...SA.......YL
PS_hli12     ......................................................Q..K.......I...C...S.A........I.
PS_hli13     ...............................................MDQAN.SVLDIAF.RP..I..ILLL.T.LV.......T.
PL_hli05     .....................................MPISDFLKETINDC..NK..SL..RI..V.ML.LMVT.LA..D....V.
PL_hli06     ...................MTSSTNVITEDGGRQNMYASEPRMQIDPEYTAFSK...LA..RG..I..LS..V..LF....L..I.
PL_hli07     .............................................MTIADF.SNK..TW..RV.....LV.I.T..V..E....I.
PL_hli08     .......................................MKKTPKTNRVENQKLT...V..M...M..W......L........VV
PL_hli09     ....................................................NEN..LQ..RW..I..IG.LAS.AA.......L.
```

Fig. 4. Comparison of all the *hli* sequences of *Prochlorococcus* strains that share the conserved C-terminus motif (TGQIIPGI/FF). Dots mark residues that are identical to the top sequence and dashes represent gaps.

these data, they suggested that there were at least seven different evolutionary lineages within the cyanobacteria. These trees also indicate that unicellular and filamentous species may be closely arranged on the same branch of the tree, but that freshwater and marine strains are often well separated. This is consistent with our results where, for example, unicellular SC *hli* genes are much more closely related (based on clustering) to the *hli* genes from freshwater, filamentous species than to the *hli* genes from unicellular marine strains. The generation of phylogenies based on 16S rRNA sequences was often taken as the benchmark for phylogenetic analyses. However with the explosion of information associated with the generation of complete genome sequences from a variety of prokaryotes, a number of individual genes within these genomes can be used to help establish phylogenetic relationships among the cyanobacteria. Differences in phylogenetic relationships that are obtained when different genes are used to evaluate such relationships suggest that the generation of a single, consistent evolutionary tree may not be easy, especially since multiple pressures imposed by specific environments may differentially influence the apparent rates at which specific genes evolve. In the case of the *hli* genes, the evolutionary pressure for this gene family to evolve and adapt to high-light conditions may create a phylogeny that is not necessarily consistent with a 16S rRNA tree. This raises the interesting possibility of analyzing and classifying a variety of molecular markers potentially indicative of specific environmental pressures (high-light or nutrient-stress). Analyses which focus on the proliferation or drastic reduction of genes in a particular adapted ecotype or species (for instance, the proliferation of *hli* genes in a high-light-adapted strain versus in low-light-adapted ecotypes or the steep reduction in two-component regulatory systems in marine species relative to freshwater species may allow us to gain an insight into environmental selection pressures, and the extent to which such pressures has shaped the individual cyanobacterial species. Since we now have a large data base of information from a range of different cyanobacteria that have adapted to very different ecological niches, this orientation represents an attractive approach for future work.

Acknowledgements

We thank F. Partensky the coordinator of the *Prochlorococcus* SS120 genome sequencing project and the Genoscope, France (M. Salanoubat) for providing us with access to the genome sequence prior to publication. This study is partly supported by the MARGENES program (EU contract xxx). A.D. is supported by a doctoral fellowship from Région Bretagne. The cooperation between Stanford and Roscoff is supported by an NSF–CNRS bilateral grant.

References

[1] Grossman, A.R., Bhaya, D. and He, Q. (2001) Tracking the light environment by cyanobacteria and the dynamic nature of light harvesting. J. Biol. Chem. 276, 11449–11452.

[2] Green, B.R., Durnford, D.G. and Jones, R.L. (1996) The chlorophyll-carotenoid proteins of oxygenic photosynthesis. Annu. Rev. Plant Physiol. Plant Mol. Biol. 47, 685–714.

[3] Sandona, D., Croce, R., Pagano, A., Crimi, M. and Bassi, R. (1998) Higher plants light harvesting proteins. Structure and function as revealed by mutation analysis of either protein or chromophore moieties. Biochim. Biophys. Acta 1365, 207–214.

[4] Jansson, S. (1994) The light-harvesting chlorophyll *a*/*b*-binding proteins. Biochim. Biophys. Acta 1184, 1–19.

[5] Kim, S., Sandusky, P., Bowlby, N.R., Aebersold, R., Green, B.R., Vlahakis, S., Yocum, C.F. and Pichersky, E. (1992) Characterization of a spinach psbS cDNA encoding the 22 kDa protein of photosystem II. Fed. Eur. Biochem. Soc. Lett. 314, 67–71.

[6] Jansson, S., Andersson, J., Jung-Kim, S. and Jackowski, G. (2000) An *Arabidopsis thaliana* protein homologous to cyanobacterial high-light-inducible proteins. Plant Mol. Biol. 42, 345–351.

[7] Heddad, M. and Adamska, I. (2000) Light stress regulated two helix proteins in Arabidopsis thaliana related to the chlorphyll *a*/*b*-binding gene family. Proc. Natl. Acad. Sci. USA 97, 3741–3746.

[8] Dolganov, N.A.M., Bhaya, D. and Grossman, A.R. (1995) Cyanobacterial protein with similarity to the chlorophyll *a*/*b*-binding proteins of higher plants: evolution and regulation. Proc. Natl. Acad. Sci. USA 92, 636–640.

[9] Funk, C. and Vermaas, W. (1999) A cyanobacterial gene family coding for single-helix proteins resembling part of the light-harvesting proteins from higher plants. Biochemistry 38, 9397–9404.

[10] Dolganov, N. and Grossman, A.R. (1999) A polypeptide with similarity to phycocyanin a subunit phycocyanobilin lyase involved in degradation of phycobilisomes. J. Bacteriol. 181, 610–617.

[11] Kaneko, T., Sato, S., Kotani, H., Tanaka, A., Asamizu, E., Nakamura, T., Miyajima, N., Hirosawa, M., Sugiura, M., Sasamoto, S., Kimura, T., Hosouchi, T., Matsuno, A., Muraki, A., Nakazaki, N., Naruo, K., Okumura, S., Shimpo, S., Takeuchi, C., Wada, T., Watanabe, A., Yamada, M., Yasuda, M. and Tabata, S. (1996) Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC 6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. DNA Res. 3 (Suppl.), 185–209.

[12] Reith, M.E. and Munholland, J. (1995) Complete nucleotide sequence of the *Porphyra purpurea* chloroplast genome. Plant Mol. Biol. 13, 333–335.

[13] He, Q., Dolganov, N., Bjorkman, O. and Grossman, A.R. (2001) The high light-inducible polypeptides in *Synechocystis* PCC6803. Expression and function in high light. J. Biol. Chem. 276, 306–314.

[14] Montane, M.H. and Kloppstech, K. (2000) The family of light-harvesting-related proteins (LHCs, ELIPs, HLIPs): was the harvesting of light their primary function? Gene 258, 1–8.

[15] Xu, H., Vavilin, D., Funk, C. and Vermaas, W. (2002) Small cab-like proteins regulating tetrapyrrole biosynthesis in the cyanobacterium *Synechocystis* sp. PCC6803. Plant Mol. Biol. 49, 149–160.

[16] Scanlan, D.J. and West, N.J. (2002) Molecular ecology of the marine cyanobacterial genera *Prochlorococcus* and *Synechococcus*. FEMS Microbiol. Ecol. 40, 1–12.

[17] Partensky, F., Hess, W.R. and Vaulot, D. (1999) Prochlorococcus, a marine photosynthetic prokaryote of global significance. Microbiol. Mol. Biol. Rev. 63, 106–127.

[18] Urbach, E., Robertson, D.L. and Chisholm, S.W. (1992) Multiple evolutionary origins of prochlorophytes within the cyanobacterial radiation. Nature 355, 267–270.

[19] Ting, C.S., Rocap, G., King, J. and Chisholm, S.W. (2002) Cyano-

bacterial photosynthesis in the oceans: the origins and significance of divergent light-harvesting strategies. Trends Microbiol. 10, 134–142.

[20] Moore, L.R., Goericke, R. and Chisholm, S.W. (1995) Comparative physiology of *Synechococcus* and *Prochlorococcus*: Influence of light and temperature on growth, pigments, fluorescence and absorptive properties. Mar. Ecol. Prog. Ser 116, 259–275.

[21] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) 'Gapped BLAST and PSI-BLAST:' a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

[22] Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F. and Higgins, D.G. (1997) The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. 24, 4876–4882.

[23] Eddy, S.R. (1998) Profile hidden Markov models. Bioinformatics 14, 755–763.

[24] Enright, A.J. and Ouzounis, C.A. (2000) GeneRAGE: a robust algorithm for sequence clustering and domain detection. Bioinformatics 16, 451–457.

[25] Wyman, M., Fay, P. (1987) Acclimation to the natural light climate. In: The Cyanobacteria (Fay, P., Baalen, C.V., Eds.), pp. 347–376. Elsevier, Amsterdam.

[26] Fogg, G.E. (1987) Marine planktonic cynaobacteria. In: The Cyanobacteria (Fay, P., Baalen, C.V., Eds.), pp. 393–414. Elsevier, Amsterdam.

[27] Wagner, A. (2002) Selection and gene duplication: a view from the genome. Genome Biol. 3, 1012–1013.

[28] Das, A. and Yanofsky, C. (1989) Restoration of a translation stop-start overlap reinstates translational coupling in a mutant trpB′-trpA gene pair of the *Escherichia coli* tryptophan operon. Nucleic Acids Res. 17, 9333–9340.

[29] Reinscheid, D.J., Schnicke, S., Rittmann, D., Zahnow, U., Sahm, H. and Eikmanns, B.J. (1999) Cloning, sequence analysis, expression and inactivation of the *Corynebacterium glutamicum* pta-ack operon encoding phosphotransacetylase and acetate kinase. Microbiology 145, 503–513.

[30] Kaneko, T. and Tabata, S. (1997) Complete genome structure of the unicellular cyanobacterium *Synechocystis* sp. PCC6803. Plant Cell Physiol. 38, 1171–1176.

[31] Okamoto, S., Ikeuchi, M. and Ohmori, M. (1999) Experimental analysis of recently transposed insertion sequences in the cyanobacterium *Synechocystis* sp. PCC 6803. DNA Res. 6, 265–273.

[32] Lawrence, J.G. and Ochman, H. (2002) Reconciling the many faces of lateral gene transfer. Trends Microbiol. 10, 1–4.

[33] Juhala, R.J., Ford, M.E., Duda, R.L., Youlton, A., Hatfull, G.F. and Hendrix, R.W. (2000) Genomic sequences of bacteriophages HK97 and HK022: pervasive genetic mosaicism in the lamboid bacteriophages. J. Mol. Evol. 299, 27–51.

[34] Zhaxybayeva, O. and Gogarten, J.P. (2002) Bootstrap, Bayesian probability and maximum likelihood mapping: exploring new tools for comparative genome analyses. BMC Genomics 3, 4.

[35] Gupta, R.S. (1998) Protein phylogenies and signature sequences: A reappraisal of evolutionary relationships among archaebacteria, eubacteria, and eukaryotes. Microbiol. Mol. Biol. Rev. 62, 1435–1491.

[36] Gupta, R.S. and Golding, G.B. (1996) The origin of the eukaryotic cell. Trends Biochem. Sci. 21, 166–171.

[37] Honda, D., Yokota, A. and Sugiyama, J. (1999) Detection of seven major evolutionary lineages in cyanobacteria based on 16S rRNA gene analysis with new sequences of five marine *Synechococcus* strains. J. Mol. Evol. 48, 723–739.

[38] Ishida, T., Watanabe, M.M., Sugiyama, J. and Yokota, A. (2001) Evidence for polyphyletic origin of the members of the orders of *Oscillatoriales* and *Pleurocapsales* as determined by 16S rDNA analysis. FEMS Micobiol. Lett. 201, 79–82.