DNA Research **9**, 123–130 (2002)

Complete Genome Structure of the Thermophilic Cyanobacterium *Thermosynechococcus elongatus* BP-1

Yasukazu Nakamura,[1] Takakazu Kaneko,[1] Shusei Sato,[1] Masahiko Ikeuchi,[2] Hiroshi Katoh,[2] Shigemi Sasamoto,[1] Akiko Watanabe,[1] Mayumi Iriguchi,[1] Kumiko Kawashima,[1] Takaharu Kimura,[1] Yoshie Kishida,[1] Chiaki Kiyokawa,[1] Mitsuyo Kohara,[1] Midori Matsumoto,[1] Ai Matsuno,[1] Naomi Nakazaki,[1] Sayaka Shimpo,[1] Masako Sugimoto,[1] Chie Takeuchi,[1] Manabu Yamada,[1] and Satoshi Tabata[1,*]

Kazusa DNA Research Institute, 2-6-7 Kazusa-kamatari, Kisarazu, Chiba 292-0812, Japan[1] *and Department of Life Sciences (Biology), The University of Tokyo, Komaba 3-8-1, Meguro, Tokyo 153-8902, Japan*[2]

(Received 21 May 2002)

Abstract

The entire genome of a thermophilic unicellular cyanobacterium, *Thermosynechococcus elongatus* BP-1, was sequenced. The genome consisted of a circular chromosome 2,593,857 bp long, and no plasmid was detected. A total of 2475 potential protein-encoding genes, one set of rRNA genes, 42 tRNA genes representing 42 tRNA species and 4 genes for small structural RNAs were assigned to the chromosome by similarity search and computer prediction. The translated products of 56% of the potential protein-encoding genes showed sequence similarity to experimentally identified and predicted proteins of known function, and the products of 34% of these genes showed sequence similarity to the translated products of hypothetical genes. The remaining 10% lacked significant similarity to genes for predicted proteins in the public DNA databases. Sixty-three percent of the *T. elongatus* genes showed significant sequence similarity to those of both *Synechocystis* sp. PCC 6803 and *Anabaena* sp. PCC 7120, while 22% of the genes were unique to this species, indicating a high degree of divergence of the gene information among cyanobacterial strains. The lack of genes for typical fatty acid desaturases and the presence of more genes for heat-shock proteins in comparison with other mesophilic cyanobacteria may be genomic features of thermophilic strains. A remarkable feature of the genome is the presence of 28 copies of group II introns, 8 of which contained a presumptive gene for maturase/reverse transcriptase. A trace of genome rearrangement mediated by the group II introns was also observed.

Key words: *Thermosynechococcus elongatus*; thermophile; genomic sequencing; cyanobacterium

1. Introduction

"Cyanobacteria" is a general term for the bacteria capable of oxgenic photosynthesis, which comprises over 1500 species with various morphologies and species-specific characteristics such as cell movement, cell differentiation, and nitrogen fixation. The most distinctive feature of the unicellular rod-shaped cyanobacterium *Thermosynechococcus elongatus* strain BP-1 is its thermophilic character.[1] It inhabits hot springs and has an optimum growth temperature of approximately 55°C. Based on the phylogenetic analysis using 16S rRNA sequences, *T. elongatus* is branched very close to the origin of cyanobacteria, while the others, including seven major lineages, are branched rather recently as "crown groups."[2] *T. elongatus* is an obligate photoautotrophic organism, pigmented with chlorophyll *a*, carotenoids and phycocyanobilin, and has long been used as a model organism for the study of photosynthesis.[3–5] Although transformation by electroporation has been attempted in thermophilic cyanobacteria, it has been successful only in *T. elongatus*. The ability of this organism to undergo natural transformation can also be used to introduce exogenous DNA into the cell (M. Ikeuchi, unpublished information). Furthermore, both photosystem I and II complexes have successfully been crystallized and subjected to X-ray 3D analysis form this organism because it provides highly stable protein complexes.[3,6]

In order to acquire knowledge on the genetic system of *T. elongatus*, which would in turn accelerate our understanding of the complex association among the genetics,

Communicated by Mituru Takanami

* To whom correspondence should be addressed. Tel. +81-438-52-3933, Fax. +81-438-52-3934, E-mail: tabata@kazusa.or.jp

physiology and biochemistry of photosynthesis, we determined the nucleotide sequence of the entire genome and analyzed the structure of all the gene components of this organism. In addition, we compared the gene structures of *T. elongatus* with those of two cyanobacterial strains, *Synechocystis* sp. PCC 6803[7] and *Anabaena* sp. PCC 7120,[8] whose complete genome structures have already been determined.

2. Materials and Methods

2.1. Bacterial strain and cloning vectors

T. elongatus strain BP-1 was obtained from a culture stock (Toray Company), which was originally isolated as a thermophilic cyanobacterium from a hot spring in Beppu, a city in the southern part of Japan.[1] It was identified as *Synechococcus elongatus* strain BP-1[4] and later given by a new genus name, *Thermosynechococcus*,[5] because it was found to be distinct from the mesophilic rod-shaped cyanobacterium, *Synechococcus*, by phylogenetic analysis of 16S rRNA sequences.[2] We established axenic culture and isolated a representative clone twice consecutively to achieve genetic uniformity.

For generation of genomic random libraries for DNA sequencing, M13mp18, pUC18, and pBeloBACII were used as cloning vectors.

2.2. DNA sequencing

The nucleotide sequence of the entire genome of *T. elongatus* was determined by the whole-genome shotgun strategy combined with the "bridging shotgun" method.[9] Four random libraries with three types of cloning vectors were generated from the total cellular DNA of *T. elongatus* to minimize cloning bias. Libraries SEE and SEB contained inserts of approximately 1.0 kb (element clones) and 2.6 kb (bridge clones) that were derived by sonication and cloned into M13mp18. An SEP library (plasmid clones) bore approximately 5-kb fragments generated by sonication that were cloned into pUC18. An SEL library (BAC clones) contained inserts of approximately 20 kb that were cloned into the BAC vector pBeloBACII.

One strand of the element clones and both strands of the clones from the other three libraries were sequenced using the Dye-terminator Cycle Sequencing kit with DNA sequencers type 377XL (Applied Biosystems, USA) according to the protocol recommended by the manufacturer. A total of 23,278 sequence files corresponding to about 5.2 genome-equivalents were accumulated and assembled using the Phrap program (Philip Green, University of Washington, Seattle, USA). The end-sequence data from the bridge, plasmid and BAC clones facilitated the gap-closure process as well as accurate reconstruction of the entire genome. The final gaps in the sequences were filled by primer walking. A lower threshold of acceptability for the generation of consensus sequences was set at a Phred score of 20 for each base. The integrity of the reconstructed genome sequence was assessed by walking through the entire genome with the end sequences of plasmid or BAC clones.

2.3. Gene assignment and annotation

Coding regions were assigned by a combination of computer prediction and similarity search as described previously. Briefly, prediction of protein-coding regions was carried out with the Glimmer 2.02 program.[10] Prior to prediction, the matrix was generated for the *T. elongatus* genome by training with a dataset of 1802 open reading frames (ORFs) that showed a high degree of sequence similarity to experimentally identified and predicted proteins of known function at the amino acid level. All of the predicted protein-encoding regions equal to or longer than 90 bp were translated into amino acid sequences, which were then subjected to similarity search against the non-redundant protein database (nr-database) with the BLASTP program.[11] In parallel, the entire genomic sequence was compared with those in the nr-protein database using the BLASTX program to identify genes that had escaped from prediction and/or those smaller than 90 bp, especially in the predicted intergenic regions. For predicted genes that did not show sequence similarity to known genes, only those equal to or longer than 150 bp were considered as candidates.

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

Genes for structural RNAs were assigned by similarity search against the in-house structural RNA database that had been generated based on the data in GenBank (rel. 128.0). tRNA-encoding regions were predicted by use of the tRNA scan-SE 1.21 program[12] in combination with the similarity search.

Comparison of the structures of the gene components among the genomes of *T. elongatus*, Synechocystis and Anabaena was performed by taking two factors, the BLAST2 bit score and the ratio of alignment length, into consideration. A lower threshold of acceptability was set at one-fourth of the bit score reported by self-comparison of the translated amino acid sequences by the BLASTP program. Only amino acid sequences whose alignments extended over at least 0.6 times the length of the query sequence were considered similar. The GC skew analysis was performed as described by Lobry.[13]

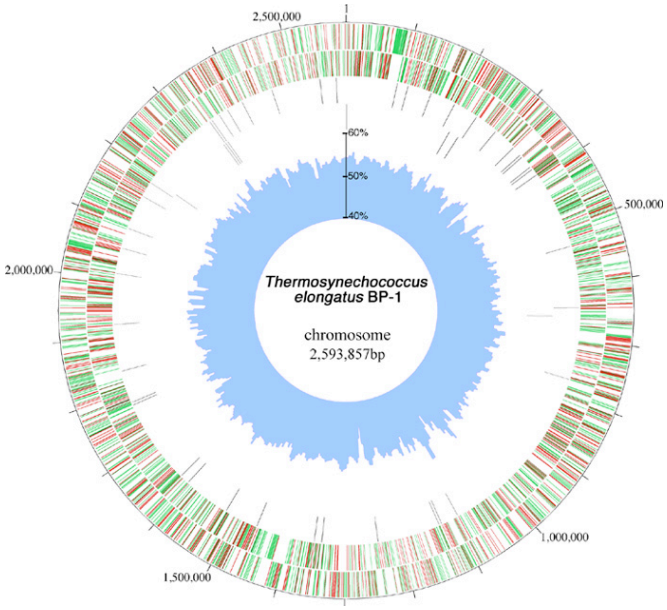

Figure 1. Circular representation of the chromosome of *Thermosynechococcus elongatus* BP-1. The scale indicates the location in bp starting from the *Swa* I recognition site. The bars in the outermost and the second circles show the positions of the putative protein-encoding genes in the clockwise and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

3. Results and Discussion

3.1. Sequence determination of the entire genome

The entire genome of *T. elongatus* was sequenced according to the modified whole genome shotgun method as described in Materials and Methods. Initially, a total of 23,278 random sequence files, which correspond to approximately 5.2 genome-equivalents, were assembled to generate low-accuracy draft sequences. Finishing was carried out by visually editing the above draft sequences, followed by gap closing and additional sequencing to obtain the sequence data having a Phred score of 20 or higher. The integrity of the final genome sequence was assessed by comparing the insert length of each pUC and BAC clone with the computed distance between the end sequences of the clones. The genome of *T. elongatus* thus deduced was a circular chromosome of 2,593,857 bp, and the average GC content was 53.9%. No plasmid was detected in the course of genome sequencing. The nucleotide position was numbered from a recognition site of the restriction enzyme *Swa* I (Fig. 1 and Fig. 1 in the Supplement section).

3.2. Sequence features of the genome

The inner circle of Fig. 1 shows the average GC percent of every 10 kb for the entire genome. No obvious uneven distribution was observed except a deep spike at the coordinates 1,240,000–1,250,000 where the gene for a glycosyl transferase-like protein is located. Similar spikes were also found in the genome of *Anabaena* at the region where the genes for the glycosyl transferase-like proteins were clustered. GC skew analysis was performed to locate a probable origin and terminator of DNA replication,[13] but no apparent shift of skew was detected (data not shown).

HIP1 is an 8-base palindromic sequence, GCGATCGC, first reported in the genomes of *Synechococcus* species and other cyanobacterial strains.[14] By surveying the entire genome sequence of *T. elongatus*, 3681 copies of HIP1 were detected. The average frequency of occurrence was 1 copy/705 bp, which is much higher than those in Synechocystis (1 copy/1131 bp) and Anabaena (1 copy/1219 bp). HIP1 was found more frequently in protein-encoding regions (1 copy/672 bp) than in RNA-encoding regions (1 copy/4328 bp) and intergenic regions (1 copy/1209 bp). Moreover, only one copy of HIP1 was identified in the 110,339-bp regions corresponding to mobile introns and insertion sequences (IS), which were likely horizontally transferred into the genome during the course of evolution, suggesting that HIP1 is characteristic of cyanobacterial genomes.

3.3. Assignment of protein- and RNA-encoding genes

The potential protein-encoding regions were assigned using a combination of computer prediction by the Glimmer program and similarity search. Glimmer predicted a total of 2743 potential protein-encoding genes in the genome after training with a dataset of sequences of highly probable protein-encoding genes. By taking the sequence similarity to known genes and the relative positions into account to avoid overlaps, as described in Materials and Methods, the total number of potential protein-encoding genes finally assigned to the genome was 2475 (Fig. 1 and Fig. 1 in the Supplement section). The average gene density was one gene in every 1048 bp. The putative protein-encoding genes thus assigned to the genome starting with either an ATG, GTG, TTG, or ATT codon are denoted by a serial number with three letters representing the species name (t), whether the ORF was longer than or shorter than 100 codons (l or s), and the transcription direction on the circular map (r or l) (Fig. 1). The codon usage frequency of the whole gene components in the genome is tabulated in Table 1 in the Supplement section.

One copy of an rRNA gene cluster was identified on the genome in the order of 16S-23S-5S at the coordinates of 2,330,963–2,336,733 (Fig. 1 in the Supplement section). The 23S RNA gene in the cluster contained a group I intron (TelI2) capable of encoding a homing endonuclease

Table 1. Features of the assigned protein-encoding genes and their functional classification.

		%
Amino acid biosynthesis	99	4.0
Biosynthesis of cofactors, prosthetic groups, and carriers	120	4.8
Cell envelope	56	2.3
Cellular processes	78	3.2
Central intermediary metabolism	23	0.9
Energy metabolism	72	2.9
Fatty acid, phospholipid and sterol metabolism	29	1.2
Photosynthesis and respiration	124	5.0
Purines, pyrimidines, nucleosides, and nucleotides	41	1.7
Regulatory functions	87	3.5
DNA replication, recombination, and repair	62	2.5
Transcription	29	1.2
Translation	151	6.1
Transport and binding proteins	153	6.2
Other categories	266	10.7
Subtotal of genes similar to genes of known function	**1390**	**56.2**
Similar to hypothetical protein	837	33.8
Subtotal of genes similar to registered genes	**2227**	**90.0**
No similarity	248	10.0
Total	**2475**	**100.0**

with a single-LAGLIDADG motif.[15] This type of group I intron has been reported in the genome of a chlamydia *Simkania negevensis* ZT,[16] that is phylogenetically close to cyanobacteria, and in the chloroplast and mitochondrial genomes of green algae. They are classified into four subfamilies based on the type of endonucleases and on the position of insertion. The group I intron in *T. elongatus* belongs to the I-*Cre* I subfamily.[15] The length of the intron was 745 bp, which is 143 bp shorter than that of the typical member of this family, but the position of insertion was conserved, that is between nt 2565 and 2566 from the 5′ end of the gene. The translated amino acid sequence of the encoded endonuclease was 31% identical with a I-*Cre* I protein.[17]

A total of 42 tRNA genes representing 42 tRNA species were assigned on the genome by sequence similarity to known bacterial tRNA genes and computer prediction using the tRNA scan-SE program (Fig. 1 and Table 1, Table 2, Fig. 1, and Fig. 2 in the Supplement section). These genes were dispersed on the genome and were likely to be transcribed as single units, except for those in the rRNA gene cluster, *trnA*-UGC and *trnI*-GAU. *trnL*-UAA contained a group I intron (TelI1) at the coordinates nt 2,355,846–2,356,084, as reported in several cyanobacterial strains.[18,19] The intron was 239 bp in length, and the sequence was 65.4% identical with that of Anabaena. Insertion of a group II intron (TelI4h) into *trnI*-CAU was observed as the first example in eubacteria (see section 3.5.5). It is reasonable to speculate that excision of this intron actually takes place in the cell because this is the only *trnI*-CAU gene in the genome.

As to small RNA-encoding genes, potential genes showing sequence similarity to SRP RNA,[20] tm RNA,[21] 6Sa RNA (*ssaA*),[22] and RNase P subunit B (*rnpB*)[23] were assigned to the genome.

3.4. Functional assignment of protein-encoding genes

Translated amino acid sequences of 2475 potential protein-encoding genes in the genome were compared with the sequences in the nr-protein databases with the BLAST program as described in Materials and Methods. According to the results, 1390 (56%) were homologous to genes of known function, 837 (34%) showed similarity to hypothetical genes, and the remaining 248 (10%) showed no significant similarity to any registered genes (Table 1 and Fig. 1).

The potential protein-encoding genes whose function could be anticipated were classified into 14 categories, each with different biological roles, according to the principle of Riley.[24] The numbers of genes in each category are summarized in Table 1, and the name of each gene is listed in CyanoBase at http://www.kazusa.or.jp/cyanobase/. On the gene map in the Supplement section (Fig. 1), the location, length and direction of these genes are indicated, with color codes corresponding to functional categories.

3.5. Characteristic features of the predicted genes and the genome

3.5.1. Comparison of gene constituents among three cyanobacteria

To date, annotation information of all the gene components in two cyanobacteria, *Synechocystis* sp. PCC 6803 and *Anabaena* sp. PCC 7120, have been published, and a total of 3167 (later revised to 3264) and 5368 potential protein-encoding genes have been assigned to the respective chromosomes [CyanoBase at http://www.kazusa.or.jp/cyanobase/]. *T. elongatus* genes were compared with those of Synechocystis and Anabaena under the criteria described in Materials and Methods. The result indicated that 1569 *T. elongatus* genes, comprising 63% of the 2475 potential protein-encoding genes, have matched genes in both the Synechocystis and Anabaena genomes, 522 of which were genes of unknown function. One hundred sixteen (5%) and 254 (10%) *T. elongatus* genes were commonly found in either the Synechocystis or Anabaena genome, respectively. The remaining 536 genes (22%) were unique to *T. elongatus*. One hundred seventy of these were genes of known functions, including 26 genes for transporters,

13 genes for the two-component system and Ser/Thr kinases, and 46 genes which are probably exogenous origin, such as those for the restriction-modification system, transposases and reverse transcriptases.

3.5.2. Genes related to photosynthesis

Genes related to photosynthesis and their copy number in the genomes of three cyanobacteria are listed in Table 3 in the Supplement section.

The complete sets of PSI and PSII genes including *psaX* were identified in the genome. Two copies of *psbV* that codes for cytochrome *c*550 were tandemly duplicated only in *T. elongatus*.[5] There were three copies of *psbA* that encodes reaction center D1 complex of photosystem II, two of which were tandemly arranged. Single genes were found for subunits of cytochrome b6/f complex (*petA–petD*, *petG*, *petM*, and *petN*). In contrast with multiple *petC* genes in Synechocystis and Anabaena, *T. elongatus* contained a single *petC* gene for Rieske-type FeS center protein for subunits of cytochrome b6/f complex, in contrast with multiple genes in other two cyanobacteria. Moreover, a universal gene for the small membrane-spanning component (*petL*) is missing. It is noteworthy that *petE* for plastocyanin, a mobile electron carrier, is missing in *T. elongatus*, while genes for alternative carriers, cytochrome *b*6 (*tll1283/petJ*) and cytochrome M (*tll2429*), were found as usual. A complete set of genes for CO_2 fixation enzymes were found, although that for RubisCO activase was absent, as in Synechocystis. As to phycobilisome components, *T. elongatus* contained complete sets of *cpcA–cpcG* (for phycocyanin) and *apcA–apcF* genes (allophycocyanin). In addition to four genes for typical single membrane-spanning proteins of CAB/ELIP/HLIP superfamily (*tsl2208*, *tsr1755*, *tsr1916*, and *tsr0446*), which may be involved in tolerance against high light, a novel subfamily of the CAB/ELIP/HLIP superfamily (*tsl0063*) was identified.

3.5.3. Genes likely to be involved in thermophilic character

Genes for three types of fatty acid desaturases (*desA*, *desB*, and *desD*) are missing in contrast with mesophilic Synechocystis,[25] although three related copies in the fourth type were found (*tlr1653/desC*, *tll1719/desC*, and *tlr2380/desC*). This agrees with the absence of highly unsaturated fatty acids in lipids, which are popular in many thermophiles. On the other hand, more genes were found for heat-shock proteins than in Synechocystis. Namely, three additional *dnaJ*-like genes (*tll0881*, *tll1433*, and *tlr0324*) were identified in addition to the four sets of *dnaK/dnaJ* and three common *dnaJ*-like genes (*tlr0758*, *tlr0132*, and *tll0182*). Genes for HtpG, HspA, Hsp33, and GroEL1/EL2/ES and a complete set of genes for ClpB (*tlr1389*, *tll2453*) and ClpC (*tll0307*), ClpX (*tlr0510*) and ClpP (*tlr0509*, *tlr1071*, *tll1759*) are present, as in Synechocystis.

3.5.4. Genes for signal transduction

For the two-component signal transduction system, 17 and 27 potential genes for His kinases and response regulators, respectively, were identified in the *T. elongatus* genome. They include highly conserved genes such as those for a drug sensor (*tlr0437/sll0698* in Synechocystis), a phosphate sensor (*tll0925/phoR*), a KaiC-interacting protein (*tlr0029/sasA*), a phytochrome-like circadian input kinase (*tll0899/cikA*) and three motility-related *cheA*-like proteins (*tlr0349/pilL*, *tll0568/pixL*, and *tll1021/sll1296* in Synechocystis). A total of 11 genes for Ser/Thr protein kinases were found in *T. elongatus*. Some of them are conserved in the Synechocystis genome (*tlr0445/spkA*, *tlr1326/spkB*, *tlr2432/spkF*, and *tlr2304/spkG*). Seven genes for presumptive protein phosphatases including GlnB phosphatase (*tlr2243*) and two genes for adenylate cyclase (*tll2280*, *tll2410*) were detected.

Twenty-seven genes were assigned as those coding for transcription factors. They are categorized into the LuxR family (4 genes), OmpR family (7 genes), LysR family (3 genes: *rbcR*, *ntcB*, and *ndhR*), CRP family (3 genes including *ntcA*), ArsR family (2 genes) and FUR family (3 genes), while the others are single genes such as those for a heat-shock gene repressor HrcA (*tll0761*), a DeoR-type repressor (*tlr0491*), a GntR-type repressor (*tll2117*) and a MerR-type regulator (*tll1888*). No genes belonging to LexA- or AraC-families were detected. A complete set of genes for sigma factors for RNA polymerase (*sigA–D*, *sigF*, *sigG1/G2*), except a *sigE* homologue, were identified in the genome.

T. elongatus has five genes that presumptively code for bacteriophytochrome, including those for a *pixJ1* homolog (*tll0569*) required for positive phototaxis and *cikA* (*tll0899*), while genes for popular cyanobacterial phytochromes (*cph1* and *cph2*), which are assumed to be an ancestor of plant phytochrome, were not detected. In addition, two genes homologous to those for flavin-binding cryptochrome-like proteins (*tll0552* and *tll0425*) and one for flavin-binding photoreceptor, phototropin (*tll1282*) were found.

3.5.5. Genes with group II introns

Group II introns code for the self-splicing ribozyme mainly found in organelle genomes in fungi and plants and in some eubacteria as well.[26,27] It is reported that some group II introns function as retro mobile genetic elements with the help of the maturase/reverse transcriptase activity encoded within the introns. There were 28 copies of group II introns in the *T. elongatus* genome (Table 4 and Fig. 3 in the Supplement section). Twenty-five of them exhibited a conserved secondary

structure (domains I–VI) and had conserved bases in domain V, a probable reaction center of ribozyme, while the remaining 3 (TelI3g, TelI3m, and TelI3n) had deletions in domains I to IV. These introns were classified into two types with respect to their lengths: TelI3a–3t had a consensus length of 844 bp, and TelI4a–4h, which contain a DNA segment encoding a presumptive maturase/reverse transcriptase in domain IV, had a consensus length of 2384 bp.

Sequence identity among the members of each type of the group II introns was quite high: 87.2–100% for the type TelI3 and 85.3–100% for the type TelI4 introns at the nucleotide level. The nucleotide sequences of 7 members of TelI3 (TelI3a, TelI3b, TelI3c, TelI3d, TelI3e, TelI3i, and TelI3o), 4 other members of TelI3 (TelI3h, TelI3f, TelI3j, and TelI3s,) and 2 members of TelI4 (TelI4b and TelI4c) were completely identical. The fact that the structures of the group II introns are highly conserved indicates that retrotransposition of the group II introns occurred in this organism recently in the evolution.

Most of the group II introns were found either in ISs, the group II introns, or intergenic regions with two exceptions: *trnI*-CAU (as described in Section 3.3) and *tll2478*. It is likely that *tll2478* was structurally and functionally disrupted by insertion, because TelI3j was located in a coding region of *tll2478* in the reverse orientation with respect to transcription; therefore, splicing of the intron from the transcript is not expected to occur. Insertion of a group II intron into domain IV of other group II introns was often observed: TelI3i in TelI3h, TelI3a in TelI4a, TelI3b in TelI4b, TelI3c in TelI4c, TelI3d in TelI4d, and TelI3e in TelI4e. Insertion of ISEL2f into TelI4g was also found. Because domain IV of the group II introns is known to be variable, it is speculated that such insertions do not impair ribozyme activity and mobility.

There was evidence of genome rearrangement mediated by the group II introns. Traces of separated halves of a gene, homologous with *sll0683* in Synechocystis, were found adjacent to TelI4f and TelI4g, which are approximately 400 kb apart in the genome. One possible explanation for this arrangement would be that insertion of a TelI4 intron into an original *sll0683* homologue took place, which was then subjected to recombination with another copy of TelI4 in the reverse orientation far apart in the genome, followed by additional conversion of the segment containing either copy of the separated half of the gene.

3.5.6. Insertion sequences

A total of 70 insertion sequences (ISs), 52 of which were likely to retain intact structures and 18 disrupted, were identified in the *T. elongatus* genome. These ISs could be classified into five groups on the basis of similarity, types of transposases, and lengths of inverted repeats generated by insertion.[28] Structural features of each IS group are summarized in Table 5, Table 6, and Fig. 4 in the Supplement section.

It is notable that ISEL3 had a capacity coding for two polypeptides showing sequence similarity to the IS200 family and IS605-TnpB family transposases in the reverse orientation. One remarkable structural feature of ISEL4 is that the ORF for the transposase does not close within the element. For ten members of ISEL4, the ORF ends at the stop codon present in the flanking sequences duplicated during insertion. For the remaining four members, the termination of translation is likely to occur outside of the element. Extreme examples are ISEL4m and ISEL4p, in which the ORFs are fused with the downstream potential protein-encoding genes.

Insertion of the group II intron, TelI3, into three members of ISEL1 was observed (Fig. 4 in the Supplement section). For ISEL1u and ISEL1v, most of the ORFs encoding the putative transposases are replaced by TelI3p and TelI3r, respectively, leaving only the terminal 119 bp 5′- and 163 bp 3′-regions of the ISs intact, while TelI3k is simply inserted between nt 119 and 120 of ISEL1j. These ISs may still be active in transposition when the *trans*-acting transposase is present.

3.6. Other features of the genes and the genome of T. elongatus

Structural features of the genes and the genome of *T. elongatus* that merit comment are the following.

1. The presence of inteins has been reported in two cyanobacteria, Synechocystis (DnaE, DnaB, GyrB, and DnaX) and Anabaena (DnaE and DnaB).[8] In *T. elongatus*, only one intein was identified in DnaE. The intein in DnaE seemed to be a split intein, as in Synechocystis and Anabaena, capable of protein *trans*-splicing. Two split *dnaE* genes, *tll2056* and *tll2069*, located 10.2 kb apart in the chromosome, and presumably had the capacity to code for the N-terminal and C-terminal portions of DnaE, respectively.

2. In the *T. elongatus* genome, two genes for proteins containing WD repeats were identified. The putative product of *tlr0489* contained seven WD repeats, composed of the repeating unit for their entire length. The putative product of *tlr1498* containing the 11 repeating units in the C-terminal portion bears a stretch of 635 amino acid residues at the N-terminus. This N-terminal portion showed a high degree of sequence similarity to those of the translated amino acid sequences of three genes: *slr0143* in Synechocystis, and *alr2791* and *all0759* in Anabaena.[8] The presumptive gene products of these genes contain WD-repeats at the C-termini, suggesting that this gene is conserved among cyanobacteria

though its biological role remains to be clarified.

3. Genes for circadian rhythms in cyanobacteria have intensively been studied in *Synechococcus* sp. PCC 7942.[29] These genes include *kaiABC* as the major genetic elements of the circadian clock, *sasA* as an activator of *kaiBC* expression, *cikA* and *pex* as encoding components of input pathways, *rpoD2* and *cpmA* as output modifiers. Presumptive counterparts of all of these genes have been identified in the *T. elongatus* genome: *tlr0481* (*kaiA*), *tlr0482* (*kaiB*), *tlr0483* (*kaiC*), *tlr0029* (*sasA*), *tll0899* (*cikA*), *tlr1955* (*pex*), *tll0831* and *tlr0264* (*rpoD2*/*sigE*), and *tll1189* (*cpmA*).

4. A gene for sulfide quinone reductase (*tll0288*) was found. This gene may function to support photosynthetic growth without photosystem II by utilization of sulfide, which is a common substrate in hot springs.[30]

5. No catalase gene was identified in the *T. elongatus* genome, whereas two superoxide dismutase genes (*tlr0036*/*sodM* and *tll1519*/*sodF*) were present. Hydrogen peroxide radicals, produced by superoxide dismutase, might be quenched by some peroxidases such as thioredoxin peroxidase (*tll1454*).

6. Regarding the genes for transformation by exogenously added DNA, those for genetic recombination (*recA*, *recF*, *recG*, *recJ*, and *recQ*), type I and III, but not type II, restriction-modification systems were identified.[31] *pilB* (*tll0122*), *pilM* (*tlr2341*), *pilN* (*tlr2342*), *pilO* (*tlr2343*), *pilQ* (*tlr2344*), *comA* (*tll2339*), *comE* (*tll1702*), and *comM* (*tlr0594*), which are likely to be involved in the ability for natural transformation and phototactic motility,[32] were also found.

7. Some cyanobacteria are known to produce biopolymers such as cellulose and poly-hydroxybutyrate.[33,34] There are two genes (*tlr1795* and *tll0007*) showing a high degree of sequence similarity to those for cellulose synthase in *T. elongatus*, while genes known to be involved in biosynthesis of poly-hydroxybutyrate were not detected.

8. There is a "cold-spot" for insertion of the ISs and the mobile introns at the approximate coordinates of 1720 kb–2020 kb of the genome (Table 4 and Table 6 in the Supplement section). No IS and the mobile intron was found in this 300 kb region, even though these elements as a whole occupied approximately 4.3% of the entire genome with an average density of 1 element in every 28 kb. Furthermore, this region overlaps with a "hot-spot" of genes conserved among three cyanobacterial species at the approximate coordinates of 1553 kb–1972 kb. In this region, 74% of the *T. elongatus* genes are common with those in both Synechocystis and Anabaena, which is considerably higher than that for the entire genome (63%). The biological significance of this region needs to be studied further in the future.

The sequences as well as the gene information shown in this paper are available in the Web database, CyanoBase, at http://www.kazusa.or.jp/cyanobase/. The sequence data analyzed in this study have been registered in DDBJ/GenBank/EMBL under accession number BA000039.

Acknowledgements: This work was supported by the Kazusa DNA Research Institute Foundation.

References

1. Yamaoka, T., Satoh, K., and Katoh, S. 1978, Photosynthetic activities of a thermophilic blue-green alga, *Plant Cell Physiol.*, **19**, 943–954.
2. Honda, D., Yokota, A., and Sugiyama, J. 1999, Detection of seven major evolutionary lineages in cyanobacteria based on the 16S rRNA gene sequence analysis with new sequences of five marine *Synechococcus* strains, *J. Mol. Evol.*, **48**, 723–739.
3. Zouni, A., Witt, H. T., Kern, J. et al. 2001, Crystal structure of photosystem II from *Synechococcus elongatus* at 3.8 A resolution, *Nature*, **409**, 739–743.
4. Sonoike, K. and Katoh, S. 1989, Simple estimation of the differential absorption coefficient of P-700 in detergent-treated preparations, *Biochim. Biophys. Acta*, **976**, 210–213.
5. Katoh, H., Itoh, S., Shen, J. R., and Ikeuchi, M. 2001, Functional analysis of *psbV* and a novel c-type cytochrome gene *psbV2* of the thermophilic cyanobacterium *Thermosynechococcus elongatus* strain BP-1, *Plant Cell Physiol.*, **42**, 599–607.
6. Jordan, P., Fromme, P., Witt, H. T., Klukas, O., Saenger, W., and Krauss, N. 2001, Three-dimensional structure of cyanobacterial photosystem I at 2.5 A resolution, *Nature*, **411**, 909–917.
7. Kaneko, T., Sato, S., Kotani, H. et al. 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**, 109–136.
8. Kaneko, T., Nakamura, Y., Wolk, C. P. et al. 2001, Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium *Anabaena* sp. strain PCC 7120, *DNA Res.*, **8**, 205–213.
9. Kaneko, T., Tanaka, A., Sato, S. et al. 1995, Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC 6803. I. Sequence features in the 1 Mb region from map positions 64% to 92% of the genome, *DNA Res.*, **2**, 153–166.
10. Delcher, A. L., Harmon, D., Kasif, S., White, O., and Salzberg, S. L. 1999, Improved microbial gene identification with GLIMMER, *Nucleic Acids Res.*, **27**, 4636–4641.
11. Altschul, S. F., Madden, T. L., Schaffer, A. A. et al. 1997, Gapped BLAST and PSI-BLAST: a new generation of

protein database search programs, *Nucleic Acids Res.*, **25**, 3389–3402.
12. Lowe, T. M. and Eddy, S. R. 1997, tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence, *Nucleic Acids Res.*, **25**, 955–964.
13. Lobry, J. R. 1996, Asymmetric substitution patterns in the two DNA strands of bacteria, *Mol. Biol. Evol.*, **13**, 660–665.
14. Gupta, A., Morby, A. P., Turner, J. S., Whitton, B. A., and Robinson, N. J. 1993, Deletion within the metallothionein locus of cadmium-tolerant *Synechococcus* PCC 6301 involving a highly iterated palindrome (HIP1), *Mol. Microbiol.*, **7**, 189–195.
15. Lucas, P., Otis, C., Mercier, J. P., Turmel, M., and Lemieux, C. 2001, Rapid evolution of the DNA-binding site in LAGLIDADG homing endonucleases, *Nucleic Acids Res.*, **29**, 960–969.
16. Everett, K. D., Kahane, S., Bush, R. M., and Friedman, M. G. 1999, An unspliced group I intron in 23S rRNA links Chlamydiales, chloroplasts, and mitochondria, *J. Bacteriol.*, **181**, 4734–4740.
17. Chevalier, B. S., Monnat, R. J. Jr., and Stoddard, B. L. 2001, The homing endonuclease I-*Cre* I uses three metals, one of which is shared between the two active sites, *Nat. Struct. Biol.*, **8**, 312–316.
18. Paquin, B., Kathe, S. D., Nierzwicki-Bauer, S. A., and Shub, D. A. 1997, Origin and evolution of group I introns in cyanobacterial tRNA genes, *J. Bacteriol.*, **179**, 6798–6806.
19. Rudi, K. and Jakobsen, K. S. 1999, Complex evolutionary patterns of tRNA Leu(UAA) group I introns in the cyanobacterial radiation, *J. Bacteriol.*, **181**, 3445–3451.
20. Gorodkin, J., Knudsen, B., Zwieb, C., and Samuelsson, T. 2001, SRPDB (Signal Recognition Particle Database), *Nucleic Acids Res.*, **29**, 169–170.
21. Watanabe, T., Sugita, M., and Sugiura, M. 1998, Identification of 10Sa RNA (tmRNA) homologues from the cyanobacterium *Synechococcus* sp. strain PCC 6301 and related organisms, *Biochim. Biophys. Acta*, **1396**, 97–104.
22. Watanabe, T., Sugiura, M., and Sugita, M. 1997, A novel small stable RNA, 6Sa RNA, from the cyanobacterium *Synechococcus* sp. strain PCC 6301, *FEBS Lett.*, **416**, 302–306.
23. Vioque, A. 1992, Analysis of the gene encoding the RNA subunit of ribonuclease P from cyanobacteria, *Nucleic Acids Res.*, **20**, 6331–6337.
24. Riley, M. 1993, Functions of the gene products of *Escherichia coli*, *Microbiol. Rev.*, **57**, 862–952.
25. Sakamoto, T. and Murata, N. 2002, Regulation of the desaturation of fatty acids and its role in tolerance to cold and salt stress, *Curr. Opin. Microbiol.*, **5**, 208–210.
26. Martinez-Abarca, F. and Toro, N. 2000, Group II introns in the bacterial world, *Mol. Microbiol.*, **38**, 917–926.
27. Dai, L. and Zimmerly, S. 2002, Compilation and analysis of group II intron insertions in bacterial genomes: evidence for retroelement behavior, *Nucleic Acids Res.*, **30**, 1091–1102.
28. Ohtsubo, E. and Sekine, Y. 1996, In: Saedler, H., and Giel, A. (eds) Current topics in microbiology and immunology, Springer, pp. 1–26.
29. Iwasaki, H. and Kondo, T. 2000, The current state and problems of circadian clock studies in cyanobacteria, *Plant Cell Physiol.*, **41**, 1013–1020.
30. Mullineaux, C. W. 2001, How do cyanobacteria sense and respond to light?, *Mol. Microbiol.*, **41**, 965–971.
31. Matveyev, A. V., Young, K. T., Meng, A., and Elhai, J. 2001, DNA methyltransferases of the cyanobacterium *Anabaena* PCC 7120, *Nucleic Acids Res.*, **29**, 1491–1506.
32. Yoshihara, S., Geng, X., and Ikeuchi, M. 2002, *pilG* gene cluster and split *pilL* genes involved in pilus biogenesis, motility and genetic transformation in the cyanobacterium *Synechocystis* sp. PCC 6803, *Plant Cell Physiol.*, **43**, 513–521.
33. Hai, T., Hein, S., and Steinbuchel, A. 2001, Multiple evidence for widespread and general occurrence of type-III PHA synthases in cyanobacteria and molecular characterization of the PHA synthases from two thermophilic cyanobacteria: *Chlorogloeopsis fritschii* PCC 6912 and *Synechococcus* sp. strain MA19, *Microbiology*, **147**, 3047–3060.
34. Nobles, D. R., Romanovicz, D. K., and Brown, R. M., Jr. 2001, Cellulose in cyanobacteria. Origin of vascular plant cellulose synthase?, *Plant Physiol.*, **127**, 529–542.