DNA Research **10**, 67–77 (2003)

Complete Sequence and Analysis of the Plastid Genome of the Unicellular Red Alga *Cyanidioschyzon merolae*

Niji Ohta,[1,*] Motomichi Matsuzaki,[2,4] Osami Misumi,[3,4] Shin-ya Miyagishima,[4] Hisayoshi Nozaki,[4] Kan Tanaka,[5] Tadasu Shin-i,[6] Yuji Kohara,[6] and Tsuneyoshi Kuroiwa[7]

Department of Molecular Biology, Faculty of Science, Saitama University, 255 Shimo-Ohkubo, Sakura, Saitama, Saitama 338-8570, Japan,[1] *Department of Biomedical Chemistry, Graduate School of Medicine, University of Tokyo, 7-3-1 Hongo, Bunkyo, Tokyo 113-0033, Japan,*[2] *Bio-Oriented Technology Research Advancement Institution (BRAIN), 3-18-19 Toranomon, Minato, Tokyo 105-0001, Japan,*[3] *Department of Biological Sciences, Graduate School of Science, University of Tokyo, 7-3-1 Hongo, Bunkyo, Tokyo 113-0033, Japan,*[4] *Institute of Molecular and Cellular Biosciences, University of Tokyo, 1-1-1 Yayoi, Bunkyo, Tokyo 113-0032, Japan,*[5] *Center for Genetic Resource Information, National Institute for Genetics, 1111 Yata, Mishima, Shizuoka 411-8540, Japan,*[6] *and Department of Life Science, College of Science, Rikkyo (St. Paul's) University, 3-34-1 Nishiikebukuro, Toshima, Tokyo 171-8501, Japan*[7]

(Received 3 February 2003; revised 25 February 2003)

Abstract

The complete nucleotide sequence of the plastid genome of the unicellular primitive red alga *Cyanidioschyzon merolae* 10D (Cyanidiophyceae) was determined. The genome is a circular DNA composed of 149,987 bp with no inverted repeats. The G + C content of this plastid genome is 37.6%. The *C. merolae* plastid genome contains 243 genes, which are distributed on both strands and consist of 36 RNA genes (3 rRNAs, 31 tRNAs, tmRNA, and a ribonuclease P RNA component) and 207 protein genes, including unidentified open reading frames. The striking feature of this genome is the high degree of gene compaction; it has very short intergenic distances (approximately 40% of the protein genes were overlapped) and no genes have introns. This genome encodes several genes that are rarely found in other plastid genomes. A gene encoding a subunit of sulfate transporter (*cysW*) is the first to be identified in a plastid genome. The *cysT* and *cysW* genes are located in the *C. merolae* plastid genome in series, and they probably function together with other nuclear-encoded components of the sulfate transport system. Our phylogenetic results suggest that the Cyanidiophyceae, including *C. merolae*, are a basal clade within the red lineage plastids.

Key words: *Cyanidioschyzon merolae*; red algae; plastid; genome sequencing

1. Introduction

Plastids are unique organelles found in land plants, algae, and some protozoa. Plastids play important roles in photosynthesis and the biosynthesis of amino acids, fatty acids, vitamins, etc., in the cell. They have their own genetic systems, and their own genomes.

The origin and evolution of plastid genomes, or plastids themselves, have long been an important subject in the biological sciences. Plastids represent the endosymbiotic remnants of a free-living cyanobacterial progenitor, which lost the vast majority of its ancestral cyanobacterial genes after primary plastid endosymbiosis.[1] In order to function, plastids depend on the cell nuclei for most of their proteins and other materials. Plastid gene expression and differentiation are largely controlled by the cell nucleus, as most regulators and sigma factors are encoded in the nuclear genome.[2] How the cyanobacterial endosymbiont evolved into the plastid remains to be elucidated.

Many complete nucleotide sequences of plastid genomes have been determined and their gene contents analyzed.[3–11] The complete genome sequences of several cyanobacteria have also been determined.[12–14] This sequence information allows phylogenetic comparison and has made it possible to study the evolutionary relationships among plastids and cyanobacteria in terms of composition and structure. Furthermore, in order to under-

Communicated by Masahiro Sugiura

* To whom correspondence should be addressed. Tel. +81-48-858-3848, Fax. +81-48-858-3384, E-mail: niji@molbiol.saitama-u.ac.jp

† The nucleotide sequence data reported in this paper will appear in the DDBJ, EMBL, and GenBank nucleotide sequence databases with the accession number AB002583.

stand how this organelle has diverged since the primary endosymbiosis event, information on the nuclear genome is very important for the analysis of gene transfer between the plastid and nuclear genomes.

The red algae are thought to be one of the basal eukaryotic lineages, and may possess ancestral features of eukaryotic phototrophs.[15] The plastid genomes of the red lineage often contain genes that are involved in the biosynthesis of amino and fatty acids; however, few such genes are present in the plastid genome of the green lineage.

C. merolae is a unicellular red alga that is found in acidic hot springs,[16] and it is thought to be one of the most "primitive" eukaryotes according to many morphological characteristics.[17] The *C. merolae* cell contains one mitochondrion, one plastid with a centrally located plastid nucleoid, one Golgi body, and one microbody.[18] We have used this alga to study organelle proliferation using cytological and organelle genome analyses. The mitochondrial genome of *C. merolae* has been completely sequenced;[19] it shares many genes with higher plants, as well as *Reclinomonas americana* (Jakobid)[20] and *Acanthamoeba castellanii* (Acanthamoebidae),[21] which implies that the mitochondrial genome of *C. merolae* is very primitive. In addition, Kuroiwa's group is in the process of sequencing the entire nuclear genome of this organism. The nuclear genome of *C. merolae* is estimated at 16.4 Mbp, which is considered to be the minimum genome size in eukaryotes containing plastids. The molecular phylogeny inferred from several nuclear genes supports the basal eukaryotic position of this alga.[15] In addition, the organization of the ribosomal protein gene clusters of the plastid genomes of *C. merolae* and various plastids have been compared, and their genomic rearrangements have been discussed.[22] *C. merolae* possesses several ancestral photosynthetic eukaryote traits, and its plastid genome is a good candidate as a link between cyanobacteria and plastids.[23] Here, we report the complete nucleotide sequence of the plastid genome of *C. merolae*, with analysis of its genome structure and gene content. In addition, using a set of 8308 concatenated amino acid sequences of 41 plastid genes from various plastid lineages, we determined the phylogenetic position of the *C. merolae* plastid.

2. Materials and Methods

2.1. DNA sources

The *C. merolae* cultures given to us by Dr. G. Pint were originally mixed with *Cyanidium caldarium* Forma A (RK-1) and *Galdieria sulphuraria* (*C. caldarium* Forma B or M-8 type). *C. merolae* 10D was isolated by the single-colony isolation method on a Gellan Gum plate.[24] Cells of *C. merolae* 10D were grown in Allen's medium[25] as previously described[26] and used to isolate plastid DNA according to previously described methods.[27]

2.2. Library construction

Plastid DNA was partially digested with the restriction endonuclease *Sau*3AI and the resultant fragments were cloned in lambda DASH II (Stratagene, CA, USA). Subcloning into pBluescript II SK+ (Stratagene) was performed using *Escherichia coli* XL1-Blue (Stratagene) as the host bacterium. Exonuclease III and mung bean nuclease digestion (Stratagene) were used to create a series of overlapping deletions of the plastid insert.

2.3. DNA Sequencing

The nucleotide sequence of both strands of the plastid library was determined by the chain-termination method[28] with a Taq Dye Terminator Sequencing Kit (Applied Biosystems, CA, USA). These sequences were connected by an auto-assembler, and the resultant circular DNA sequence was refined using sequence data obtained from the *C. merolae* nuclear genome project by the whole-genome shotgun method (unpublished data). Open reading frames and transfer RNA genes were detected with the DNASIS software package (Hitachi, Japan).

2.4. Data analysis

Similarity searches of the putative open reading frames and tRNA sequences against the SwissProt and GenBank databases were performed with the program NCBI gapped BLAST[29] at the Genome Net WWW Server (http://www.genome.ad.jp/), over the Internet. Annotations of the complete plastid genomes of 11 algae, two land plants, and one protozoon were obtained from the NCBI Entrez-Genome database (http://www.ncbi.nlm.nih.gov/Entrez/), and all the protein-coding regions, except intron-coded proteins, were extracted from the data tables. For each organism, the median value of the intergenic distances, which are distances between two neighboring protein genes and take negative values when they are overlapped, was determined.

2.5. Phylogenetic analysis of cysT *and* cysW *genes*

The amino acid sequences of orthologous genes of *cysT*, *cysW*, and *modB* were extracted from the nr database using the similarity search program blastp in NCBI BLAST 2.2.2, and were aligned using CLUSTAL X[30] with the default option. After gaps in the alignment were excluded and *cysT* and *cysW* of *C. merolae* were included, a data matrix composed of 178 amino acids from 82 operational taxonomic units (OTUs) was constructed and used for the phylogenetic analysis. Neighbor-joining (NJ) trees[31] based on Kimura distances[32] were calculated

using CLUSTAL X. Bootstrap values[33] in the NJ analysis were carried out based on 1000 replications also using CLUSTAL X. The *modB* genes were designated the outgroup since they are putative paralogs of *cysT* and *cysW*. To provide a more compact figure representation, the tree was redrawn using TreeExplorer.[34]

2.6. Phylogenetic analyses of plastids based on concatenated amino acid sequences from multiple plastid genes

The data matrix of the amino acid sequences of the 41 plastid or cyanobacterial genes was the same as that used by Martin et al.,[35] except that it included the *C. merolae* sequences. The *C. merolae* sequences were aligned using the similarity search program "blastp" in BLAST 2.1 of NCBI (http://www.ncbi.nlm.nih.gov/blast/) and CLUSTAL X and then refined manually. A total of 8308 aligned amino acids from 17 OTUs was used for the phylogenetic analyses. NJ trees based on Kimura distances were calculated using CLUSTAL X; maximum parsimony (MP) trees were constructed using a heuristic search with the tree bisection-reconnection (TBR) branch-swapping algorithm, using PAUP 4.0b10;[36] and quartet puzzling-maximum likelihood (QP) analyses based on the JTT model with the discrete gamma model for site heterogeneity were carried out using TREE-PUZZLE 5.0.[37] Bootstrap values[33] in the NJ and MP analyses were based on 1000 replicates. For the QP method, quartet puzzling support (QPS) values based on 1000 puzzling steps[37] were calculated. In these phylogenetic analyses, the cyanobacterium *Synechocystis* was designated the outgroup.

3. Results and Discussion

3.1. Physical properties of the plastid genome of C. merolae

The *C. merolae* plastid genome is a circular molecule composed of 149,987 bp, and genes are distributed on both strands (Fig. 1). The genome size is within the range of those of other plastid genomes. The overall G + C content is 37.6%. This base composition is comparable with plastid genomes of land plants and is a little higher than that of algae (Table 1). There are two simple explanations for the higher G + C content. First, the higher gene density of this plastid genome causes the higher G + C content, since coding regions are usually more G + C rich than non-coding regions in plastid genomes. This explanation does not parallel the high G + C content of land plant plastid genomes, which have a lower density of genes. Additionally, the coding region of *C. merolae* has a higher G + C content than that of other algal plastid genomes. Alternatively, the high temperature of the *C. merolae* habitat might have imposed a selection pressure causing a higher G + C content to stabilize the genome; however, *C. caldarium*, which lives in similar habitats, does not have a G + C-rich genome.

Generally, plastid and cyanobacterial genomes have a pair of inverted repeats (IR) containing rRNA genes; however, the *C. merolae* plastid genome lacks one. In the Cyanidiophyceae, *C. caldarium* lacks IRs,[10] whereas *Galdieria sulphuraria*[38] contains tandem repeats with rRNA genes. Yoon et al.[39] demonstrated that *Galdieria* is positioned basal to the clade composed of *C. caldarium* and *C. merolae*. Therefore, the IR might have been lost in the common ancestor of *C. caldarium* and *C. merolae*. However, two pairs of small direct repeats in the *C. caldarium* plastid genome that contain a potential hairpin loop do not exist in the *C. merolae* plastid genome.

3.2. Genome condensation

A high degree of condensation is one of the remarkable features of this plastid genome. About 40% of the protein genes in the *C. merolae* plastid genome overlap, which is quite high compared with other plastid genomes. Plastids in the red lineage have a higher gene density than do plastids of green lineage (Table 1), but most plastids in the red lineage have few overlapping genes (Fig. 2A). The *C. merolae* plastid genome also has shorter intergenic distances than other red lineage genomes. The median intergenic distance (14 bp) is significantly smaller than that of its closest relative, *C. caldarium*, and other plastids (Table 1).

The extensive overlap of genes might lead to suspicion about the pseudogenization of overlapped genes, but we consider this possibility negligible. Although the extent of overlap is quite large in comparison with other plastid genomes, most overlaps are shorter than 50 bp (equivalent to 17 amino acids) (Fig. 2A). Figure 2B shows one example of the 38-bp overlap between *rps17* and *rpl14*. The N-terminus of Rpl14 protein is highly conserved and this might weaken the C-terminal sequence of Rps17 protein. However, since the C-terminus of other algal orthologs of the Rps17 protein is not very highly conserved, the *C. merolae rps17* gene is likely to function normally. Additionally, transcripts of some plastid genes can be edited before translation. RNA editing has not yet been demonstrated in the plastid of red lineage; however, we have preliminary evidence that RNA editing occurs in this plastid genome at least at one site (unpublished data). RNA editing might recover any genomic sequence weakened by overlapping. We consider that the extent of overlap in this plastid genome does not seriously deteriorate the function of the majority of overlapping genes. Moreover, analysis of a nuclear genome shows that there are very few substitutes of overlapped nuclear genes. Therefore, it is suggested that those genes are not pseudogenes.

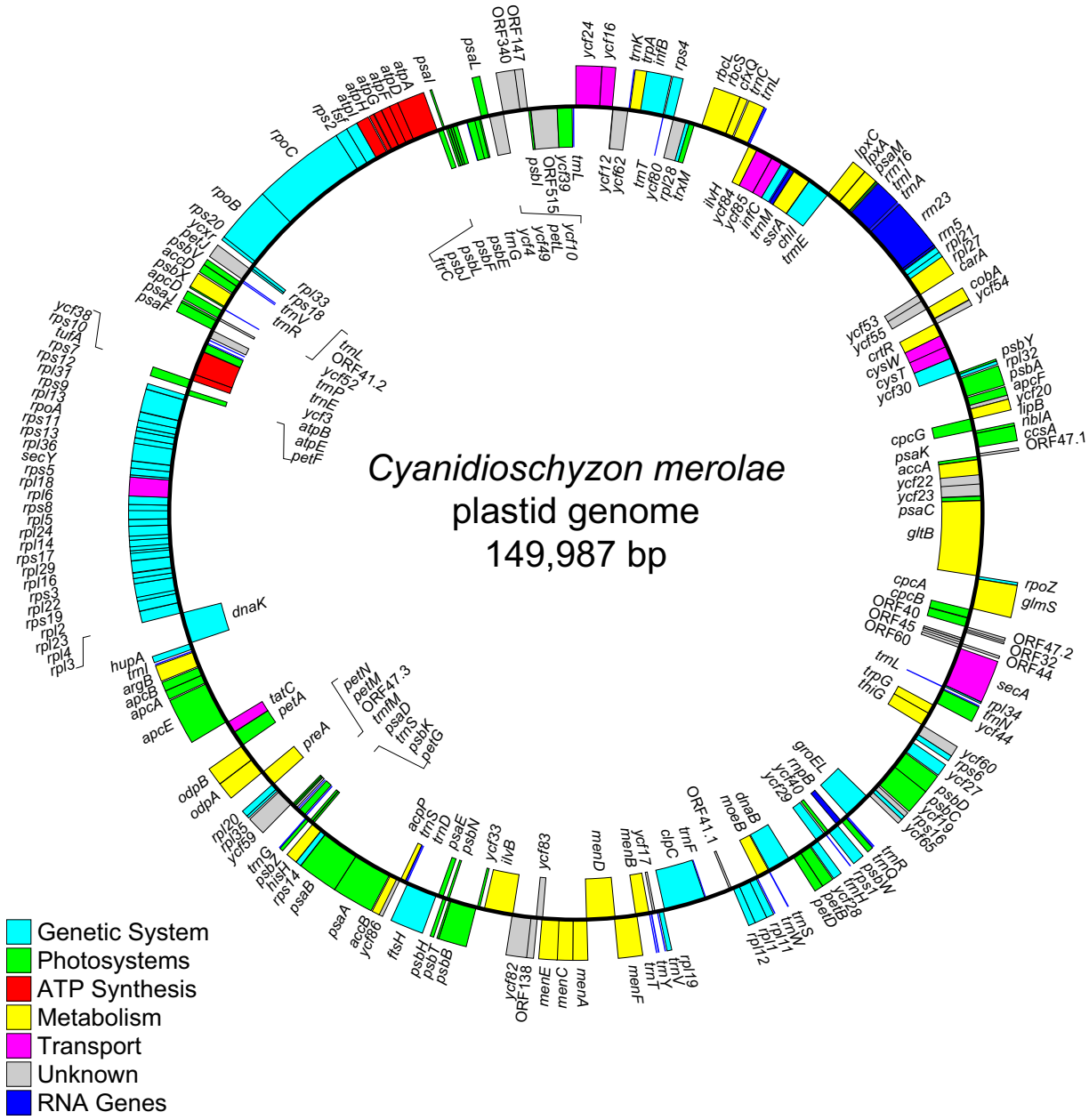

Figure 1. The plastid gene map of *C. merolae*. The outer and inner sectors represent genes on the clockwise and counterclockwise strand, respectively. Genes are color-coded by function, as shown at bottom left.

3.3. Gene content

We searched for ORFs longer than 30 codons starting with an ATG or GTG codon. The *C. merolae* plastid genome contains 243 genes, including 3 rRNAs (23S, 16S, 5S), 31 tRNAs, 1 tmRNA, 1 ribonuclease P RNA component (*rnpB*), and 207 protein genes including unidentified ORFs (Table 2). No genes containing an intron were identified.

The plastid genomes of the green lineage contain many genes involved in photosynthesis and gene expression with a small number of other functions. By contrast, plastid genomes of the red lineage contain more genes for ribosomal proteins and photosynthesis components. Moreover, they possess genes that are involved in the synthesis of amino acids, fatty acids, and pigments, among others. This is also the case for the *C. merolae* plastid genome.

The *C. merolae* plastid genome contains several genes that are rarely found in other plastid genomes, such as a sulfate-transport gene (*cysW*) (see below). The following genes are found in both *C. merolae* and *C. caldarium*, but not in *Porphyra purpurea*:[5] *crtR*, *cobA*, *glmS*, *hisH*, *lpxA*, *lpxC*, *menA-F*, *trmE*, *ycf49*, *ycf82*, *ycf83*, *ycf84*, *ycf85*. When all the plastid genes of *C. merolae*

Table 1. Number of protein-coding genes, intergenic distance, length and GC content of several plastid genomes.

Species	Lineage[a)]	Number of protein-coding genes	Median of intergenic distances (bp)	Length (bp)	GC content (%)	Accession number
Cyanidioschyzon merolae	Red	207	14	149,987	37.6	AB002583
Cyanidium caldarium	Red	200	60	164,921	32.7	NC_001840
Porphyra purpurea	Red	209	100	191,028	33.0	NC_000925
Guillardia theta	Red (2)	147	73	121,524	33.0	NC_000926
Odontella sinensis	Red (2)	140	69	119,704	31.8	NC_001713
Toxoplasma gondii	Red (2)*	26	22	34,996	21.4	NC_001799
Cyanophora paradoxa	Glaucophyte	150	112.5	135,599	30.5	NC_001675
Euglena gracilis	Green (2)	62	130	143,172	26.1	NC_001603
Astasia longa	Green (2)*	46	128	73,345	22.4	NC_002652
Nephroselmis olivacea	Green	156	250	200,799	42.1	NC_000927
Chlorella vulgaris	Green	173	243	150,613	31.6	NC_001865
Mesostigma viride	Green	105	183	118,360	30.2	NC_002186
Chaetosphaeridium globosum	Green	98	155	131,183	29.6	NC_004115
Lotus japonicus	Green	81	306	150,519	36.0	NC_002694
Arabidopsis thaliana	Green	87	243	154,478	36.3	NC_000932

a) Red lineage (red), green lineage (green), Glaucophyte, and species thought to have plastids by secondary endosymbiosis (2) are categorized. Asterisks indicate those species with non-photosynthetic plastids.

are compared with those of *C. caldarium*, eight genes occur only in *C. merolae* (*cysT*, *cysW*, *hupA*, *infB*, *petL*, *ycf22*, *ycf38*, *ycxr*), while five genes are found only in *C. caldarium* (*glnB*, *ycf26*, *ycf37*, *ycf45*, *ycf58*). *C. merolae* and *C. caldarium* share many similar genes. These common genes may have been maintained because *C. merolae* and *C. caldarium* live in an extreme environment and are closely related. Similar living conditions with strong evolutionary pressure may have led to conservation of many genes in their plastid genomes.

According to Glockner et al., *infB* is not present in the plastid genome of *C. caldarium* RK-1,[10] although we found *infB* in the plastid genome of *C. caldarium* RK-1 in a previous study.[22] These two strains of '*C. caldarium* RK-1' are probably distinctive species, as these two nucleotide sequences show only approximately 70% identity.

3.4. Genes for RNA polymerases

The *C. merolae* plastid genome contains *rpoA*, *rpoB*, *rpoC*, and *rpoZ* genes for the subunits of RNA polymerase. The single gene (*rpoC*) encoding the β' subunit of RNA polymerase, which is present in most eubacteria, is split into β' (*rpoC1*) and β'' (*rpoC2*) genes in the cyanobacteria and most plastids. However, the β' and β'' subunits are both encoded by a single *rpoC* gene in *C. merolae*, as is the case for most eubacteria, excluding cyanobacteria. There are two explanations for the presence of a single *rpoC* gene in *C. merolae*. First, it was generated by fusion of *rpoC1* and *rpoC2* specifically in the evolutionary pathway to *C. merolae*. Alternatively, the single *rpoC* gene might have been transmitted from another bacteria by horizontal transfer. When the *C. merolae rpoC* is compared with *rpoC1* and *rpoC2* in plastids and other cyanobacterial genomes and *rpoC* in bacterial genomes, its amino acid sequence is very similar to the former genes, supporting the first explanation. It is reasonable that they have been fused under the strong pressure that causes genome condensation. In fact, the length of *C. merolae rpoC* gene shortened to approximately 90% of the total length of *rpoC1* and *rpoC2* genes of other red algae.

3.5. The cysT *and* cysW *genes of* C. merolae

The plastid genome of *C. merolae* includes the *cysT* and *cysW* genes which code for components of sulfate transporter. The *cysW* gene has not previously been found in a plastid genome. Moreover, this is the first report of the existence of *cysT* in a red lineage plastid, while *cysT* genes have been identified in the plastid genome of lower green plants, such as the liverwort *Marchantia polymorpha*[4] and the green alga *Chlorella vulgaris*.[7] The products of both *cysT* and *cysW* of *C. merolae* exhibit high similarity to the corresponding proteins encoded by cyanobacteria. Furthermore, A phylogenetic tree inferred from the amino acid sequences of *cysT* and *cysW* from various bacteria and *C. merolae* was constructed

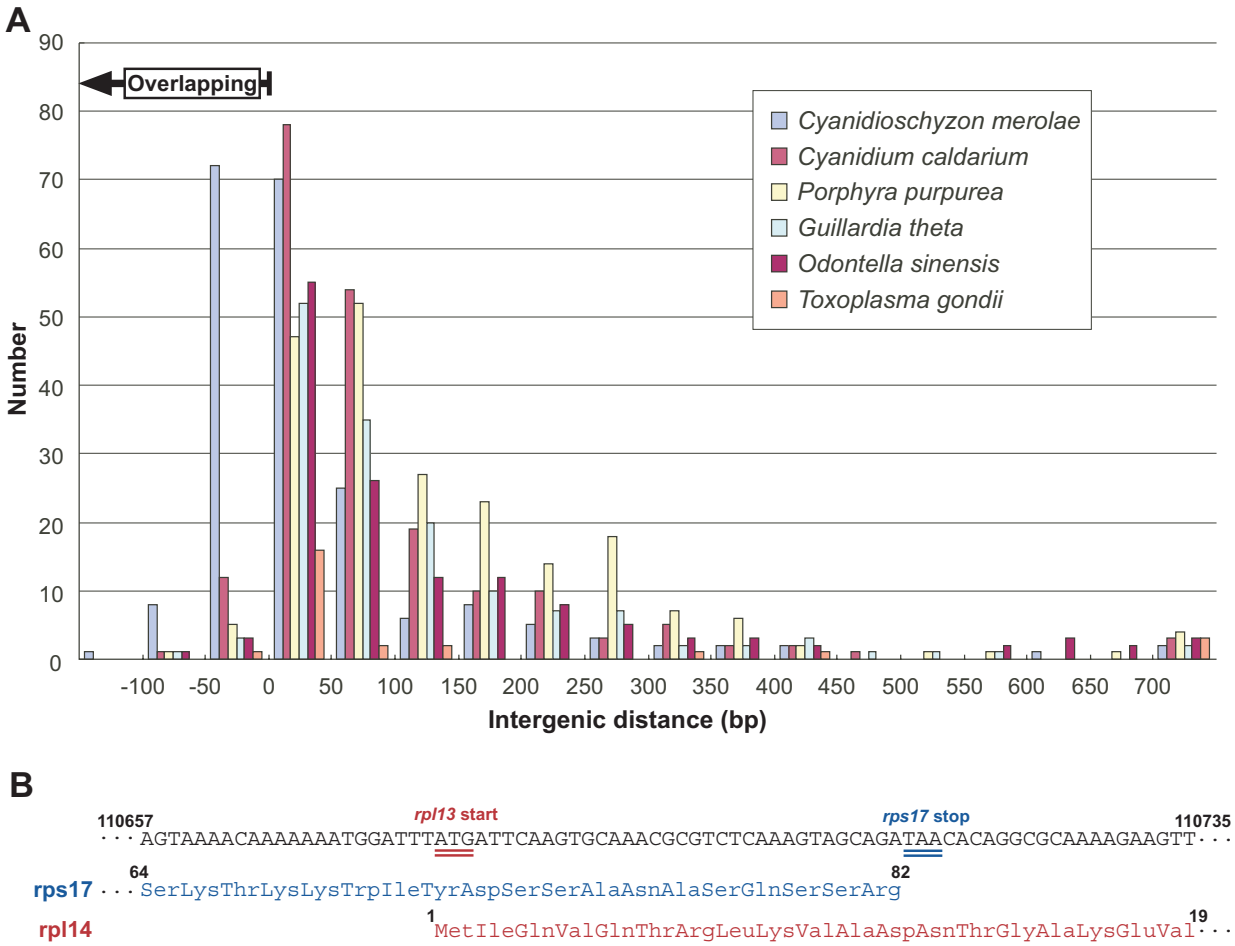

Figure 2. Overlapping genes on the *C. merolae* plastid genome. (A) Distribution of the intergenic distances in comparison with those of other species. Negative values represent overlapping. (B) An example of overlapping genes of *C. merolae*. The 3′ end of the *rps17* coding region and the 5′ end of the *rpl14* coding region share 38 bp.

by the NJ method (Fig. 3). Two major groups can be distinguished: *cysT* and *cysW* groups. The *cysW* gene of *C. merolae* was a sister to the group of cyanobacterial *cysW* genes, and the *C. merolae cysT* gene was positioned within the plastid/cyanobacterial *cysT* lineage. These results suggest that the respective genes were derived from ancestral cyanobacterium independently and were not produced by duplication.

The gene products of *cysT* and *cysW* of *C. merolae* are thought to be important for transporting a sulfate substrate into the plastid. In both enterobacteria[40] and cyanobacteria,[41] sulfate uptake is thought to require a four-component periplasmic transport system. It consists of ATP-binding protein (*cysA*) and sulfate-binding protein precursor (*sbp*), as well as the permease subunits (*cysT* and *cysW*). Recently, *cysA* and *sbp* have been identified in the nuclear genome of *C. merolae* as part of the analysis of the entire genome sequence (unpublished data). Based on the encoded genes and the degree of amino acid sequence similarity, we postulate that *C. merolae* possesses the functional components of a sulfate-transport system located in the plastid envelope. The presence of this complex on the *C. merolae* plastid envelope appears essential.

The sulfate transporter is also an interesting subject in terms of the evolution of plastids. In *M. polymorpha*, *cysA* and *cysT* (named *mbpX* and *mbpY*, respectively) were identified in the plastid genome. Although the other components of the complex have not yet been found, it was suggested that there was a functional sulfate-transport system on the analog of *C. merolae*. In contrast, neither the plastid nor nuclear genomes of higher plants, such as completely sequenced *Arabidopsis thaliana*, contain genes encoding polypeptides homologous to components of the bacterial sulfate transport complex. In higher plants, it has been suggested that other machinery has replaced the bacteria-type sulfate transport complex in the course of evolution. These genes locate the *C. merolae* plastid as an intermediate between green and red lineages, suggesting the importance of this organism in study of the plant evolution.

Table 2. Functional classification of *C. merolae* plastid genes.

classification	number	genes							
Genetic system									
Maintenance	2	*dnaB*	*hupA*						
RNA polymerase	4	*rpoA*	*rpoB*	*rpoC*	*rpoZ*				
Transcription factors	4	*ycf27*	*ycf28*	*ycf29*	*ycf30*				
Translation	4	*infB*	*infC*	*tsf*	*tufA*				
Ribosomal proteins	27	*rpl1*	*rpl2*	*rpl3*	*rpl4*	*rpl5*	*rpl6*	*rpl11*	*rpl12*
		rpl13	*rpl14*	*rpl16*	*rpl18*	*rpl19*	*rpl20*	*rpl21*	*rpl22*
		rpl23	*rpl24*	*rpl27*	*rpl28*	*rpl29*	*rpl31*	*rpl32*	*rpl33*
		rpl34	*rpl35*	*rpl36*					
	19	*rps1*	*rps2*	*rps3*	*rps4*	*rps5*	*rps6*	*rps7*	*rps8*
		rps9	*rps10*	*rps11*	*rps12*	*rps13*	*rps14*	*rps16*	*rps17*
		rps18	*rps19*	*rps20*					
tRNA maturation	1	*trmE*							
Protein quality control	4	*clpC*	*dnaK*	*ftsH*	*groEL*				
Photosystems									
Phycobilisomes	9	*apcA*	*apcB*	*apcD*	*apcE*	*apcF*	*cpcA*	*cpcB*	*cpcG*
		nblA							
Photosystem I	13	*psaA*	*psaB*	*psaC*	*psaD*	*psaE*	*psaF*	*psaI*	*psaJ*
		psaK	*psaL*	*psaM*	*ycf3*	*ycf4*			
Photosystem II	18	*psbA*	*psbB*	*psbC*	*psbD*	*psbE*	*psbF*	*psbH*	*psbI*
		psbJ	*psbK*	*psbL*	*psbN*	*psbT*	*psbV*	*psbW*	*psbX*
		psbY	*psbZ*						
Cytochrome complex	10	*petA*	*petB*	*petD*	*pet G*	*petJ*	*pet L*	*petM*	*pet N*
		ccsA	*ycf44*						
Redox system	3	*ftrC*	*petF*	*trxM*					
ATP synthesis									
ATP synthase	8	*atpA*	*atpB*	*atpD*	*atpE*	*atpF*	*atpG*	*atpH*	*atp I*
Metabolism									
Carbohydrates	5	*odpA*	*odpB*	*rbcL*	*rbcS*	*cfxQ*			
Lipids	5	*accA*	*accB*	*accD*	*acpP*	*crtR*			
Nucleotides	1	*carA*							
Amino acids	7	*argB*	*gltB*	*hisH*	*ilvB*	*ilvH*	*trpA*	*trpG*	
Complex sugars	3	*glmS*	*lpxA*	*lpxC*					
Cofactors	12	*chlI*	*cobA*	*lipB*	*menA*	*menB*	*menC*	*menD*	*menE*
		menF	*moe B*	*preA*	*thiG*				
Transport									
Transport	9	*cysT*	*cysW*	*secA*	*secY*	*tatC*	*ycf16*	*ycf24*	*ycf85*
		ycf84							
Unknown									
Conserved ORFs	25	*ycf10*	*ycf12*	*ycf17*	*ycf19*	*ycf20*	*ycf22*	*ycf23*	*ycf33*
		ycf38	*ycf39*	*ycf40*	*ycf49*	*ycf52*	*ycf53*	*ycf54*	*ycf55*
		ycf59	*ycf60*	*ycf62*	*ycf65*	*ycf80*	*ycf82*	*ycf83*	*ycf86*
		ycxr							
Unique ORFs*	14	ORF32	ORF40	ORF41.1	ORF41.2	ORF44	ORF45	ORF47.1	ORF47.2
		ORF47.3	ORF60	ORF138	ORF147	ORF340	ORF515		
RNA genes									
rRNAs	3	*rrn16*	*rrn23*	*rrn5*					
tRNAs	31	*trnA*	*trnC*	*trnD*	*trnE*	*trnF*	*trnG*	*trnG*	*trnH*
		trnI	*trnI*	*trnK*	*trnL*	*trnL*	*trnL*	*trnL*	*trnM*
		trnfM	*trnN*	*trnP*	*trnQ*	*trnR*	*trnR*	*trnS*	*trnS*
		trnS	*trnT*	*trnT*	*trnV*	*trnV*	*trnW*	*trnY*	
Miscellaneous RNAs	2	*ssrA*	*rnpB*						

*) The numbers indexing ORFs refer to the number of amino acid residues in the deduced polypeptide.

3.6. Phylogenetic analyses of plastids based on concatenated amino acid sequences from multiple plastid genes

Figure 4 shows the phylogenetic tree constructed using the concatenated amino acid sequences of the 41 plastid genes. The phylogenetic relationships resolved in this study were essentially consistent with those found

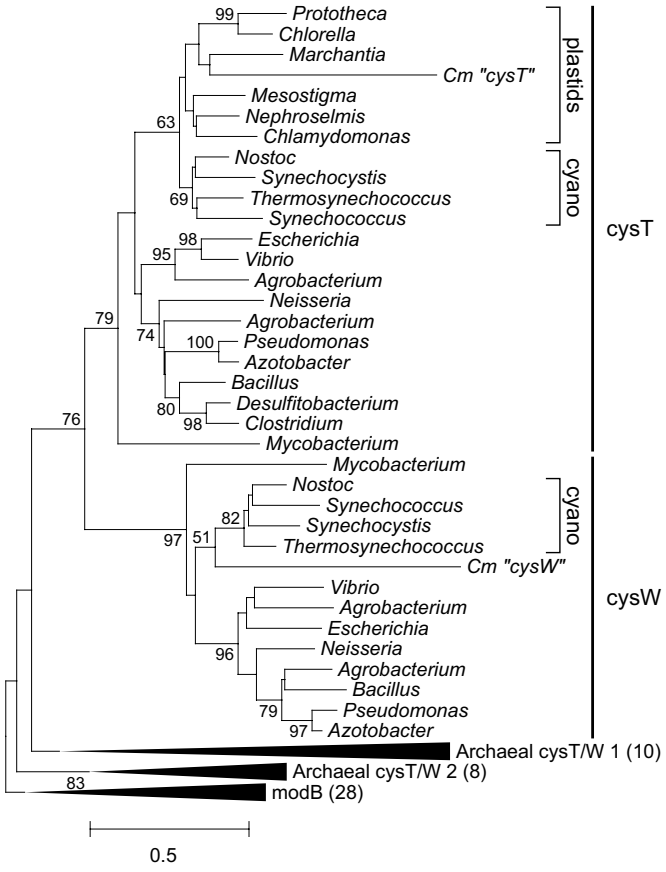

Figure 3. Phylogenetic relationships of *cysT* and *cysW* genes from various plastids and prokaryotes, with *mod*B genes designated as the outgroup. The tree was constructed based on 178 amino acid sequences by the neighbor-joining (NJ) method[31] using on Kimura distances.[32] Branch lengths are proportional to Kimura distances, which are indicated by the scale bar below the tree. Numbers at branches represent the bootstrap values (50% or more) based on 1000 replications. "Cm" indicates genes from *C. merolae*. Triangles represent compressed branches, and the number of OTUs compressed is shown in the parenthesis.

in Martin's analysis,[35] except for relationships regarding *C. merolae*. The plastids principally belong to the green lineage, red lineage, and glaucocystophyte *Cyanophora*. The red lineage is subdivided into two sister clades, one containing two cyanidiophycean algae (*Cyanidium* and *Cyanidioschyzon*) and the other composed of two secondary plastid-containing algae (the diatom *Odontella* and the cryptophyte *Guillardia*) and the rhodophycean alga *Porphyra*. Although high bootstrap/QPS values (99–100%) in the NJ and QP analyses supported the robustness of these two clades, the MP method weakly resolved the latter clade (with 55% bootstrap values). Within the latter clade, *Porphyra* and *Guillardia* were resolved as sister OTUs with 79% and 100% bootstrap/QPS values in the MP and NJ/QP calculations, respectively.

Some previous plastid phylogenies showed that the secondary plastids of Heterokontophyta, such as *Odontella*, were phylogenetically related to the Cyanidiales or to Cyanidiophyceae.[39,42] However, Yoon et al. demonstrated that the secondary plastids from the cryptophytes, haptophytes, and heterokonts constitute a monophyletic group that is sister to the rhodophycean lineage (excluding the Cyanidiales),[43] based on a large number of OTUs and the large amount of sequence data (total of 5827 nucleotides). Our phylogenetic results are consistent with those of Yoon et al.,[43] in that Cyanidiophyceae is a basal clade within the red lineage of plastids. However, our study suggests non-monophyly of the secondary plastids, as in Martin et al.[35] This discrepancy may be due to the small numbers of OTUs analyzed in this study (see Zwickl et al.;[44] Yoon et al.[43]) or it may result from differences in the sequence data used between Yoon et al.[43] (nucleotide sequences) and Martin et al.[35]/this study (amino acid sequences). The addition of OTUs from a wide range of plastids in the red lineage to the present multiple plastid-gene phylogeny or other studies using data independent of plastid gene sequences may resolve this problem.

4. Conclusions and Perspectives

The complete plastid genome sequence of *C. merolae* provides us with valuable information for understanding the processes of plastid evolution and the phylogenetic relationships among photosynthetic organisms. The 150-kb *C. merolae* plastid genome contains 243 genes. When compared with other plastid genomes, the higher gene density of *C. merolae* is characteristic. It is suggested that the *C. merolae* plastid genome was involved in degenerate evolution after the plastid endosymbiosis. The *cysT* and *cysW* genes have been identified in a red lineage plastid genome for the first time. This is valuable information, and will help determine the evolutionary relationships of three lineages: the cyanobacteria, the red lineage, and the green lineage. In order to understand the evolution and phylogenetic relationships of the plastids produced by the endosymbiosis of an ancestral cyanobacterium, comprehensive nuclear genome analysis is very important, as shown for the sulfate transporter. The entire gene composition of *C. merolae* will soon be reported, as the sequence of nuclear genome of *C. merolae* was recently completed. This will provide important information about the formation and evolution of the chloroplast.

Acknowledgements: This work was supported by Grants-in-Aid from New Energy and Industrial Technology Development Organization (No. 01B66006c to N. O.), from the Ministry of Education, Culture, Sports, Science, and Technology of Japan (No. 1320611 to T. K.), and from the Promotion of Basic Research Activities for Innovative Biosciences (ProBRAIN) to T. K.

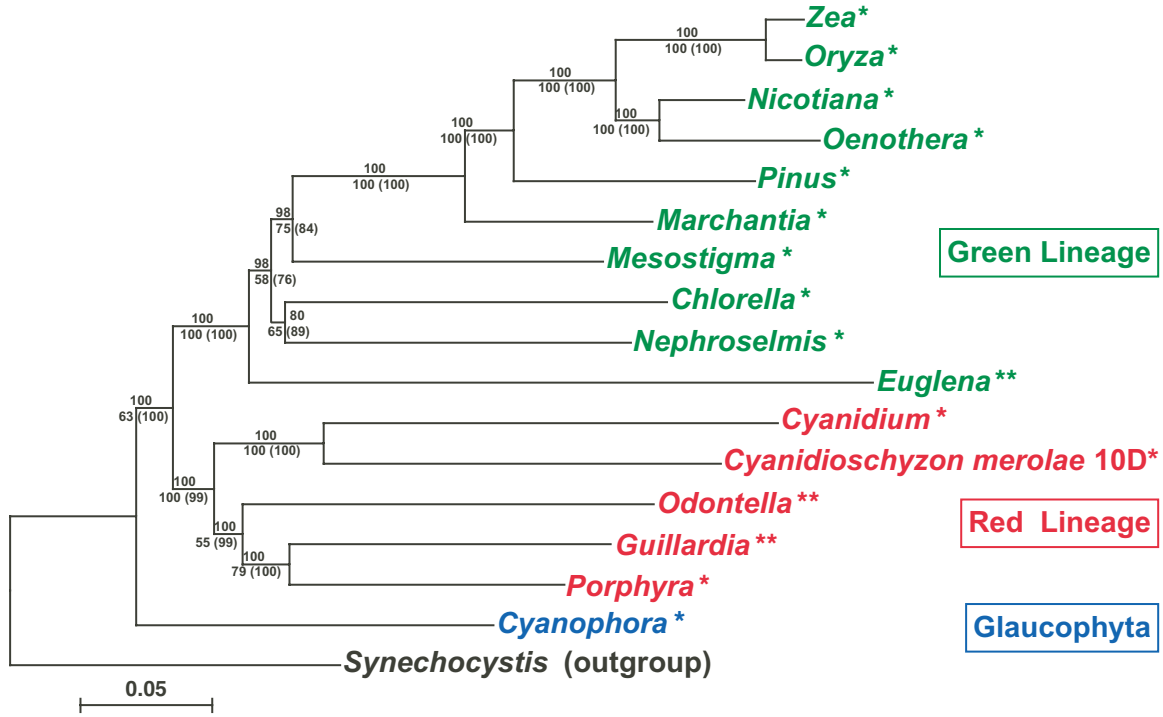

Figure 4. Neighbor-joining (NJ) trees[31] based on Kimura distances[32] using 8308 amino acid sequences from the 41 plastid genes[35] of 17 OTUs representing a wide-range of photosynthetic eukaryotic taxa. Branch lengths are proportional on Kimura distances, which are indicated by the scale bar below the tree. Numbers above branches represent the bootstrap values (50% or more) based on 1000 replications of the NJ method (based on the Kimura distances). Numbers without and with parentheses below branches are bootstrap values (50% or more) based on 1000 replications of the full heuristic MP analysis (with simple addition sequence) and QPS values (50% or more) by quartet puzzling-maximum likelihood calculation using TREE-PUZZLE 5.0[37] based on JTT model with the discrete gamma model for the site-heterogeneity. Single and double asterisks indicate the primary and secondary plastids, respectively.

References

1. Taylor, F. J. R. 1974, Implications and extensions of the serial endosymbiosis theory of the origin of eukaryotes, *Taxon*, **23**, 299–258.
2. Tanaka, K., Oikawa, K., Ohta, N., Kuroiwa, H., Kuroiwa, T., and Takahashi, H. 1996, Nuclear encoding of a chloroplast RNA polymerase sigma subunit in a red alga, *Science*, **272**, 1932–1935.
3. Shinozaki, K., Ohme, M., Tanaka, M. et al. 1986, The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression, *EMBO J.*, **5**, 2043–2049.
4. Ohyama, K., Fukazawa, H., Kohchi, T. et al. 1986, Chloroplast gene organization deduced from complete sequence of liverwort *Marchantia polymorpha* chloroplast DNA, *Nature*, **322**, 572–574.
5. Reith, M. E. and Munholland, J. 1995, Complete nucleotide sequence of the *Porphyra purpurea* chloroplast genome, *Plant Mol. Biol. Rep.*, **13**, 333–335.
6. Hiratsuka, J., Shimada, H., Whittier, R. et al. 1989, The complete sequence of the rice (*Oryza sativa*) chloroplast genome: intermolecular recombination between distinct tRNA genes accounts for a major plastid DNA inversion during the evolution of the cereals, *Mol. Gen. Genet.*, **217**, 185–194.
7. Wakasugi, T., Nagai, T., Kapoor, M. et al. 1997, Complete nucleotide sequence of the chloroplast genome from the green alga *Chlorella vulgaris*: the existence of genes possibly involved in chloroplast division, *Proc. Natl. Acad. Sci. USA*, **94**, 5967–5972.
8. Sato, S., Nakamura, Y., Kaneko, T., Asamizu, E., and Tabata, S. 1999, Complete structure of the chloroplast genome of *Arabidopsis thaliana*, *DNA Res.*, **6**, 283–290.
9. Douglas, S. E. and Penny, S. L. 1999, The plastid genome of the cryptophyte alga, *Guillardia theta*: complete sequence and conserved synteny groups confirm its common ancestry with red algae, *J. Mol. Evol.*, **48**, 236–244.
10. Glöckner, G., Rosenthal, A., and Valentin, K. 2000, The structure and gene repertoire of an ancient red algal plastid genome, *J. Mol. Evol.*, **51**, 382–390.
11. Kowallik, K. V., Stöbe, B., Schaffran, I., and Freier, U. 1995, The chloroplast genome of a chlorophyll a + c-containing alga, *Odontella sinensis*, *Plant Mol. Biol. Rep.*, **13**, 336–342.
12. Kaneko, T., Sato, S., Kotani, H. et al. 1996, Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determi-

nation of the entire genome and assignment of potential protein-coding regions, *DNA Res.*, **3**, 109–136.
13. 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.
14. Nakamura, Y., Kaneko, T., Sato, S. et al. 2002, Complete genome structure of the thermophilic cyanobacterium *Thermosynechococcus elongatus* BP-1, *DNA Res.*, **9**, 123–130.
15. Nozaki, H., Matsuzaki, M., Takahara, M. et al. 2003, The phylogenetic position of red algae revealed by multiple nuclear genes from mitochondria-containing eukaryotes and an alternative hypothesis on the origin of plastids, *J. Mol. Evol.*, in press.
16. De Luca, P., Taddei, R., and Varano, L. 1978, '*Cyanidioschyzon merolae*': a new alga of thermal acidic environments, *Webbia*, **33**, 37–44.
17. Kuroiwa, T. 1998, The primitive red algae *Cyanidium caldarium* and *Cyanidioschyzon merolae* as model system for investigating the dividing apparatus of mitochondria and plastid, *Bioessays*, **20**, 344–354.
18. Kuroiwa, T., Kuroiwa, H., Sakai, A., Takahashi, H., Toda, K., and Itoh, R. 1998, The division apparatus of plastids and mitochondria, *Int. Rev. Cytol.*, **181**, 1–41.
19. Ohta, N., Sato, N., and Kuroiwa, T. 1998, Structure and organization of the mitochondrial genome of the unicellular red alga *Cyanidioschyzon merolae* deduced from the complete nucleotide sequence, *Nucleic Acids Res.*, **26**, 5190–5298.
20. Lang, B. F., Burger, G., O'Kelly, C. J. et al. 1997, An ancestral mitochondrial DNA resembling a eubacterial genome in miniature, *Nature*, **387**, 493–497.
21. Burger, G., Plante, I., Lonergan, K. M., and Gray, M. W. 1995, The mitochondrial DNA of the amoeboid protozoon, *Acanthamoeba castellanii*: complete sequence, gene content and genome organization, *J. Mol. Biol.*, **245**, 522–537.
22. Ohta, N., Sato, N., Ueda, K., and Kuroiwa, T. 1997, Analysis of a plastid gene cluster reveals a close relationship between *Cyanidioschyzon* and *Cyanidium*, *J. Plant Res.*, **110**, 235–245.
23. Ohta, N., Sato, N., and Kuroiwa, T. 1999, The organellar genomes of *Cyanidioschyzon merolae*. In: Seckbach, J. (ed) Enigmatic Microorganisms and Life in Extreme Environments, Kluwer Academic Publishers, Dordrecht, pp. 139–149.
24. Toda, K., Takahashi, H., Itoh, R., and Kuroiwa, T. 1995, DNA contents of cell nuclei in two Cyanidiophyceae: *Cyanidioschyzon merolae* and *Cyanidium caldarium* Forma A, *Cytologia*, **60**, 183–188.
25. Allen, M. B. 1959, Studies with *Cyanidium caldarium*, an anomalously pigment chlorophyta, *Arch. Mikrobiol.*, **32**, 270–277.
26. Suzuki, K., Ohta, N., and Kuroiwa, T. 1992, Isolation of the cell-nuclear, mitochondrial, and chloroplast DNA from the ultra-small eukaryote *Cyanidioschyzon merolae*, *Protoplasma*, **171**, 80–84.
27. Ohta, N., Nagashima, H., Kawano, S., and Kuroiwa, T. 1992, Isolation of the chroloplasts DNA and the sequences of *trnK* gene of *Cyanidium caldarium strain* RK-1, *Plant Cell Physiol.*, **33**, 657–661.
28. Sambrook, J., Fritsch, E. F., and Maniatis, T. 1989, Molecular cloning: a Laboratory Manual, 2nd Ed, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.
29. Altschul, S. F., Madden, T. L., Schäffer, A. A. et al. 1997, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.*, **25**, 3389–3402.
30. Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F., and Higgins, D. G. 1997, The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools, *Nucleic Acids Res.*, **25**, 4876–4882.
31. Saitou, N. and Nei, M. 1987, The neighbor-joining method: a new method for reconstructing phylogenetic trees, *Mol. Biol. Evol.*, **4**, 406–425.
32. Kimura, M. 1983, The Neutral Theory of Molecular Evolution, Cambridge University Press, Cambridge, England.
33. Felsenstein, J. 1985, Confidence limits on phylogenies: an approach using bootstrap, *Evolution*, **38**, 16–24.
34. Tamura, K. 2001, TreeExplorer, http://evolgen.biol.metro-u.ac.jp/TE/TE_man.html
35. Martin, W., Rujan, T., Richly, E. et al. 2002, Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus, *Proc. Natl. Acad. Sci. USA*, **99**, 12246–12251.
36. Swofford, D. L. 2002, PAUP* 4.0: Phylogenetic Analysis Using Parsimony, version 4.0b10. Computer program distributed by Sinauer Associates, Inc., Fitchburg.
37. Strimmer, K. and von Haeseler, A. 1996, Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies, *Mol. Biol. Evol.*, **13**, 964–969.
38. Maid, U. and Zetsche, K. 1992, A 16 kb small single-copy region separates the plastid DNA inverted repeat of the unicellular red alga *Cyanidium caldarium*: physical mapping of the IR-flanking regions and nucleotide sequences of the *psbD-psbC*, *rps16*, 5S rRNA and *rpl21* genes, *Plant Mol. Biol.*, **19**, 1001–1010.
39. Yoon, H. S., Hackett, J. D., and Bhattacharya, D. 2002, A single origin of the peridinin- and fucoxanthin-containing plastids in dinoflagellates through tertiary endosymbiosis, *Proc. Natl. Acad. Sci. USA*, **99**, 11724–11729.
40. Ames, G. F. L. 1986, Bacterial periplasmic transport systems: structure, mechanism, and evolution, *Annu. Rev. Biochem.*, **55**, 397–425.
41. Laudenbach, D. E. and Grossman, A. R. 1991, Characterization and mutagenesis of sulfur-regulated genes in a cyanobacterium: evidence for function in sulfate transport, *J. Bacteriol.*, **173**, 2739–2750.
42. Oliveira, M. C. and Bhattacharya, D. 2000, Phylogeny of the Bangiophycidae (Rhodophyta) and the secondary endosymbiotic origin of algal plastids, *Am. J. Bot.*, **87**, 482–492.
43. Yoon, H. S., Hackett, J. D., Pinto, G., and Bhattacharya, D. 2002, The single, ancient origin of chromist plastids,

Proc. Natl. Acad. Sci. USA, **99**, 15507–15512.
44. Zwickl, D. J. and Hillis, D. M. 2002, Increased taxon sampling greatly reduces phylogenetic error, *Syst. Biol.*, **51**, 588–598.