DNA RESEARCH 6, 283–290 (1999)

# Complete Structure of the Chloroplast Genome of *Arabidopsis thaliana*

Shusei SATO, Yasukazu NAKAMURA, Takakazu KANEKO, Erika ASAMIZU, and Satoshi TABATA*

*Kazusa DNA Research Institute, 1532-3 Yana, Kisarazu, Chiba 292-0812, Japan*

(Received 21 May 1999)

**Abstract**

The complete nucleotide sequence of the chloroplast genome of *Arabidopsis thaliana* has been determined. The genome as a circular DNA composed of 154,478 bp containing a pair of inverted repeats of 26,264 bp, which are separated by small and large single copy regions of 17,780 bp and 84,170 bp, respectively. A total of 87 potential protein-coding genes including 8 genes duplicated in the inverted repeat regions, 4 ribosomal RNA genes and 37 tRNA genes (30 gene species) representing 20 amino acid species were assigned to the genome on the basis of similarity to the chloroplast genes previously reported for other species. The translated amino acid sequences from respective potential protein-coding genes showed 63.9% to 100% sequence similarity to those of the corresponding genes in the chloroplast genome of *Nicotiana tabacum*, indicating the occurrence of significant diversity in the chloroplast genes between two dicot plants.

The sequence data and gene information are available on the World Wide Web database KAOS (Kazusa Arabidopsis data Opening Site) at http://www.kazusa.or.jp/arabi/.

**Key words:** *Arabidopsis thaliana*; chloroplast; genome sequencing

## 1. Introduction

The complete sequences of the chloroplast genomes were first reported for tobacco[1] and liverwort[2] in 1986. Since then, the chloroplast genome sequences of a number of land plants and algae have been determined.[3–14] The complete genome structure of a cyanobacterium *Synechocystis* sp. PCC6803, the most primitive plant-type photosynthetic organism, has also been reported.[15] The accumulation of such data has made it possible to study the evolutional relationship among the chloroplast genomes and their ancestors. One notion derived from such study is that there was a massive transfer of genes from ancestral organelles to nuclei.[16] Comparison of nuclear and chloroplast genomes at the sequence level should provide invaluable information for understanding of the origin and function of the chloroplast in cells. In this respect, *Arabidopsis thaliana*, an excellent model organism for the analysis of the complex biological processes in plants,[17] is the most appropriate material because entire genome sequencing of this plant is in progress[18,19] by international efforts in which we are involved.[20] Here we determined the complete sequence of the chloroplast genome of this plant and compared with the those of other chloroplasts reported to date. Structural similarity with the genome of a cyanobacterium *Synecohcysitis* sp. strain PCC6803 was also investigated.

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

## 2. Materials and Methods

### 2.1. DNA sources

The Mitsui P1 library of *Arabidopsis thaliana* Columbia, which has been used for sequencing of the chromosomal genome, was adopted for screening of the chloroplast DNA, as the library had been prepared from the whole cellular DNA.[21] P1 clones harboring the chloroplast genome sequences were isolated by screening the library with the following probes derived from the tobacco chloroplast:[1] pTB30 (*psaB*), pTS8 (*petB*), pPacnD (*ndhD*), and psbA-F (*psbA*), which were provided by Dr. M. Sugiura of Nagoya University.

### 2.2. DNA sequencing

The nucleotide sequence of each P1 insert was determined according to the bridging shotgun method described previously.[20] Briefly, the purified P1 DNA was subject to sonication followed by size-fractionation on agarose gel electrophoresis. Fractions of approximately 1.0 kb and 2.5 kb were respectively cloned into M13mp18 and to construct the libraries of element and bridge clones. Clones were propagated on microtiter dishes, and the supernatants were used for preparation of sequence

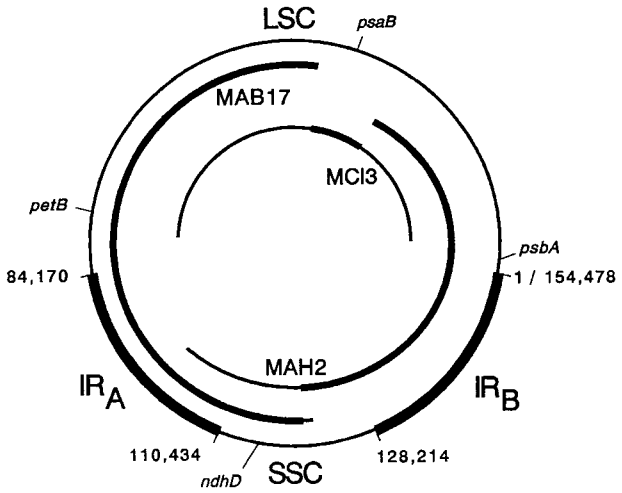


**Figure 1.** Structure of the *A. thaliana* chloroplast genome and the positions of sequenced P1 clones. The outer circle shows the overall structure of the chloroplast genome consisting of a large single-copy region (LSC), a small single-copy region (SSC) and inverted repeat regions (IRA and IRB) represented by thick lines. The positions of the genetic markers which were used for clone selection are indicated outside of the circle, and the regions covered by selected P1 clones, MAB17, MAH2, and MCI3 are indicated by inner arcs. The sequence information was obtained from the regions represented by thick lines on the clones. The initial order of the four regions deduced from the sequence data was LSC-IRA-SSC-IRB by counterclockwise as shown in this map, but the SSC sequence between IRA and IRB was inverted to conform to the indication of reported chloroplast sequences and used for the further analyses.

templates. For sequencing the element clones, single-stranded DNAs were prepared from 100 $\mu$l each of phage supernatants according to standard procedures and used directly as templates. Inserts of the bridge clones were amplified by PCR in the reaction mixture of 20 $\mu$l containing 2 $\mu$l of the phage supernatant, 50 mM KCl, 10 mM Tris-HCl (pH 9.0), 0.1% Triton X100, 1.5 mM $MgCl_2$, 50 $\mu$M each of dNTPs, 2 units of Taq polymerase (TaKaRa, Japan), and 100 nM each of the following sets of primers:

KFw (5′-GGGTTTTCCCAGTCACGAC-3′)
KRv (5′-TTATGCTTCCGGCTCGTATGTTGTG-3′)

PCR amplification was performed through 30 cycles of the temperature shift consisting of 96°C for 10 sec and 70°C for 60 sec, followed by the final extension at 70°C for 7 min in a PJ9600 thermal cycler. The products were subjected to purification by polyethylene glycol and used for the sequencing reaction.

### *2.3. DNA sequencing and data assembly*

The sequencing reaction was performed using the cycle sequencing kits (Dye-primer Cycle Sequencing kit and Dye-terminator Cycle Sequencing kit of Perkin Elmer Applied Biosystems, USA) and reaction robots (Catalyst 800 of Applied Biosystems, USA), according to the protocol recommended by manufacturers. The DNA sequencers used were type 373XL and 377XL of Perkin Elmer Applied Biosystems. The single-pass sequence data from one end of element clones and both ends of bridge clones were accumulated and assembled using Phred-Phrap programs (Phil Green, Univ. of Washington, Seattle, USA) and the auto-assembler software of Applied Biosystems, USA.

### *2.4. Computer-assisted data analysis*

The nucleotide sequences were translated in six frames using the universal codon table, and each frame was subjected to similarity search against the non-redundant protein database, owl (release 29), using the BLASTP program.[22] Positions of each local alignment, which showed similarity with scores of 70 or more to known protein sequences, were extracted and aligned along the query sequences. If internal gaps occurred, the alignments below the score of 70 were re-searched to fill in the gaps.

Structural RNA genes were identified by similarity search against the structural RNA data set from GenBank with the BLASTN program,[22] and defined as the regions with the local alignments showing 80% or more identity to the query sequences along 50 bp or more nucleotides. For assignment of tRNA genes, the tRNAscan-SE program[23] was applied for prediction.

## 3. Results and Discussion

### *3.1. Overall structure of* A. thaliana *chloroplast genome*

The sequence of the chloroplast genome of *Arabidopsis thaliana* sp. Columbia could be constructed by assembling the sequences of three partially overlapping P1 clones. The complete genome finally deduced was 154,478 bp in size. The sequences of nucleotide positions 38,670–120,256, 120,257–154,478/1–29,018 and 29,019–38,679 were respectively obtained from clones MAB17, MAH2 and MCI3, as shown in Fig. 1. The genome consisted of a pair of inverted duplications of 26,264 bp (IRA and IRB) which are separated by long and short single copy regions of 84,170 bp (LSC) and 17,780 bp (SSC). This overall structure of the *A. thaliana* chloroplast genome is typical for land plant chloroplasts.[24,25] Although the order of the four regions originally constructed from MAB17 and MAH2 was LSC-IRA-SSC-IRB counterclockwise as shown in Fig. 1, we inverted the direction of the SSC sequence between IRA and IRB to conform to the indication of previously reported chloroplast sequences, and the sequence of the structural isoform, LSC-IRB-SSC-IRA, was used for further analyses. The overall A+T content was 63.7%, which is similar to those of tobacco (62.2%), rice (61.1%) and maize (61.5%). The A+T content of the LSC and SSC regions were 66.0% and 70.7%, respectively, whereas that of the

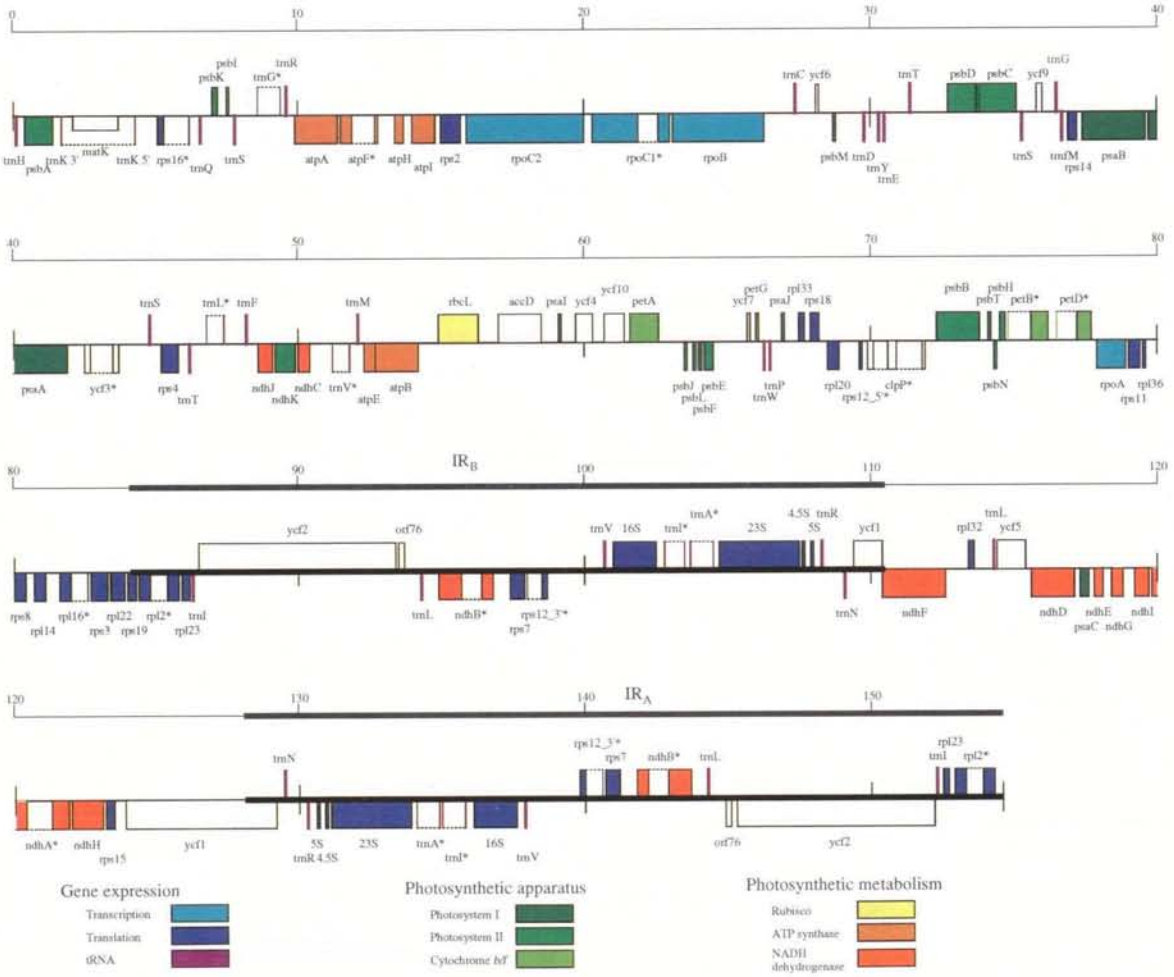


**Figure 2.** Gene organization of the *A. thaliana* chloroplast genome. The circular genome of the *A. thaliana* chloroplast was opened at the junction at IRA and LSC and is represented by a linear map starting from this junction point. The potential protein coding regions are indicated by boxes on both sides of the middle horizontal lines. The genes on the upper side are transcribed from left to right, and the lower side, from right to left. The putative genes of which the function could be deduced by similarity search are indicated by the gene names. The genes classified into 9 groups according to the biological function are shown by different color codes. The intron-containing genes are indicated by asterisks, and the position and the length of the intron is shown by the dotted horizontal line. The positions of ribosomal and tRNA genes are also shown in the map. The nucleotide sequence of the *A. thaliana* chloroplast genome appears under the accession number AP000423 in the DDBJ/GenBank/EMBL DNA databases.

IR-regions is 57.7% due to the presence of an rRNA gene cluster.

The shifts of the border positions between the two inverted repeat regions (IRA and IRB) and two single copy regions (LSC and SSC) have been observed among various chloroplast species.[26–29] To evaluate the difference of the IR lengths in the chloroplast genomes between *A. thaliana* (26,264 bp) and tobacco (25,339 bp), the exact IR border positions were compared with respect to the adjacent genes between two species. Whereas a very small shift (2 bp) was observed for the junction of IRA and LSC, larger shifts were present at other three junctions. The same tendency was seen in the positions of the IR border between rice and maize chloroplast genome.[5] In *A. thaliana*, the junction between LSC and IRB is located within the *rps19* gene, and the junction between IRB and SSC is within the *ndhF* gene. In tobacco, these two genes are located in the single copy regions.

### 3.2. *Structural features of the putative protein-coding genes*

The potential protein-coding regions were deduced as described in Materials and Methods, and the positions of a total of 87 genes including 79 unique gene species and 8 duplicated genes in the inverted repeat regions were localized on the map (Fig. 2). The predicted amino acid sequences of the *A. thaliana* chloroplast genes were then compared to those in the completely sequenced plastid and cyanobacterial genomes (Table 1). All of them showed the highest identity to those of tobacco, although

**Table 1**. List of the potential protein-coding genes assigned in the *A. thaliana* chloroplast genome and identity with the orthologous genes. The translated amino acid sequences of 79 assigned genes were compared with those of the corresponding genes in the genomes of *Nicotiana tabacum*, *Zea mays*, *Oryza sativa*, *Marchantia polymorpha*, *Pinus thunbergii*, *Epifagus virginiana*, *Euglena gracilis*, *Cyanophora paradoxa*, *Odontella sinensis*, *Porphyra purpurea*, and *Synechocystis* sp. PCC6803.

*Gene expression*

***Transcription***

| | *Nicotiana tabacum* | *Zea mays* | *Oryza sativa* | *Marchantia polymorpha* | *Pinus thunbergii* | *Epifagus virginiana* | *Euglena gracilis* | *Cyanophora paradoxa* | *Odontella sinensis* | *Porphyra purpurea* | *Synechocystis* sp. PCC6803 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *rpoA* | 79.60% | 68.00% | 67.10% | 53.30% | 62.50% | | | 38.50% | 36.50% | 37.50% | 38.10% | sll1818 |
| *rpoB* | 92.00% | 79.10% | 79.60% | 68.60% | 71.40% | | 46.80% | 45.70% | 47.40% | 48.10% | 47.20% | sll1787 |
| *rpoC1* | 91.10% | 79.10% | 79.30% | 67.70% | 67.20% | | 43.50% | 51.50% | 44.80% | 52.90% | 52.20% | slr1265 |
| *rpoC2* | 74.60% | 69.50% | 61.20% | 44.90% | 66.70% | | 29.00% | 41.50% | 36.00% | 36.90% | 39.80% | sll1789 |

***Translation***

| | *Nicotiana tabacum* | *Zea mays* | *Oryza sativa* | *Marchantia polymorpha* | *Pinus thunbergii* | *Epifagus virginiana* | *Euglena gracilis* | *Cyanophora paradoxa* | *Odontella sinensis* | *Porphyra purpurea* | *Synechocystis* sp. PCC6803 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *rpl2* | 96.90% | 68.00% | 87.40% | 63.20% | 66.90% | 87.70% | 51.10% | 52.60% | 56.40% | 46.90% | 42.60% | sll1802 |
| *rpl14* | 90.20% | 79.70% | 83.70% | 80.30% | 78.70% | | 59.00% | 69.70% | 61.50% | 67.20% | 66.40% | sll1806 |
| *rpl16* | 90.30% | 85.60% | 85.60% | 79.90% | 78.40% | 79.90% | 66.90% | 70.90% | 64.40% | 68.90% | 70.40% | sll1805 |
| *rpl20* | 82.90% | 68.10% | 68.10% | 59.10% | 66.10% | 72.60% | 33.90% | 48.70% | 50.00% | 45.30% | 51.40% | sll0767 |
| *rpl22* | 72.30% | 57.70% | 55.90% | 56.40% | 63.60% | | 46.00% | 51.80% | 50.50% | 52.70% | 51.40% | sll1803 |
| *rpl23* | 95.70% | 87.80% | 85.90% | 58.20% | 59.30% | 71.40% | 35.30% | | 47.60% | 37.60% | 38.00% | sll1801 |
| *rpl32* | 84.60% | 64.90% | 64.90% | 66.00% | 68.10% | | 45.70% | | 52.20% | 54.30% | 44.70% | ssr1736 |
| *rpl33* | 86.40% | 74.20% | 74.20% | 71.20% | 71.20% | 78.80% | | 58.50% | 56.90% | 50.80% | 57.10% | ssr1398 |
| *rpl36* | 100.00% | 91.90% | 91.90% | 86.50% | 75.70% | 91.90% | 64.90% | 78.40% | 64.90% | 70.30% | 76.30% | ssr0006 |
| *rps2* | 89.40% | 77.00% | 77.90% | 72.30% | 71.70% | 80.70% | 40.40% | 48.90% | 46.60% | 50.90% | 50.00% | sll1250 |
| *rps3* | 87.60% | 67.70% | 63.20% | 62.40% | 62.70% | 71.80% | 34.30% | 42.90% | 41.50% | 42.40% | 42.90% | sll1804 |
| *rps4* | 89.10% | 79.10% | 80.10% | 73.10% | 63.70% | 70.90% | 48.50% | 58.20% | 51.00% | 57.70% | 57.60% | slr0469 |
| *rps7* | 97.40% | 84.00% | 83.30% | 76.80% | 83.90% | 92.90% | 40.60% | 53.50% | 43.50% | 58.70% | 52.30% | sll1597 |
| *rps8* | 84.30% | 75.70% | 75.70% | 60.40% | 67.90% | 77.60% | 39.90% | 47.80% | 44.00% | 45.50% | 52.60% | sll1809 |
| *rps11* | 88.40% | 65.00% | 65.70% | 74.40% | 78.30% | 78.30% | 42.50% | 55.00% | 50.40% | 51.90% | 54.30% | sll1817 |
| *rps12* | 95.10% | 85.40% | 84.60% | 89.40% | 87.00% | 92.20% | 69.10% | 80.50% | 76.40% | 76.40% | 80.50% | sll1096 |
| *rps14* | 90.20% | 79.70% | 83.70% | 80.30% | 78.70% | 44.60% | 59.00% | 69.70% | 61.50% | 67.20% | 52.00% | slr0628 |
| *rps15* | 86.40% | 81.10% | 74.40% | 57.00% | 59.10% | | | | | | 35.80% | sll1784 |
| *rps16* | 78.50% | 75.90% | 58.60% | | | | | 48.10% | 40.50% | 52.60% | 49.40% | ssr0482 |
| *rps18* | 86.10% | 67.30% | 67.30% | 74.00% | 77.10% | 78.90% | 51.90% | 55.20% | 51.50% | 52.20% | 46.30% | ssr1399 |
| *rps19* | 88.00% | 64.90% | 64.90% | 78.30% | 71.70% | 75.80% | 52.70% | 66.30% | 58.70% | 60.90% | 63.00% | ssl3432 |

*Photosynthesis*

***Photosynthetic apparatus***

| | *Nicotiana tabacum* | *Zea mays* | *Oryza sativa* | *Marchantia polymorpha* | *Pinus thunbergii* | *Epifagus virginiana* | *Euglena gracilis* | *Cyanophora paradoxa* | *Odontella sinensis* | *Porphyra purpurea* | *Synechocystis* sp. PCC6803 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *petA* | 90.3% | 87.5% | 87.5% | 78.8% | 81.2% | | | 62.1% | 50.3% | 60.4% | 54.8% | sll1317 |
| *petB* | 98.6% | 97.2% | 99.1% | 94.9% | 94.0% | | 87.9% | 92.1% | 84.2% | 89.3% | 83.2% | slr0342 |
| *petD* | 99.3% | 98.8% | 98.1% | 94.4% | 93.1% | | | 83.8% | 75.5% | 78.1% | 75.6% | slr0343 |
| *petG* | 97.3% | 97.3% | 97.3% | 83.8% | 83.8% | | 64.7% | 73.0% | 59.5% | 64.9% | 70.3% | ssr0010 |
| *psaA* | 98.0% | 96.3% | 95.5% | 92.8% | 91.6% | | 79.2% | 82.2% | 79.8% | 82.3% | 80.5% | slr1834 |
| *psaB* | 97.7% | 96.2% | 96.7% | 92.5% | 92.4% | | 82.4% | 81.7% | 78.2% | 83.0% | 79.8% | slr1835 |
| *psaC* | 100.0% | 93.8% | 95.1% | 91.4% | 96.3% | | 91.4% | 88.9% | 86.6% | 90.1% | 90.1% | ssl0563 |
| *psaI* | 96.8% | 86.1% | 86.1% | 71.0% | 61.3% | | | 59.4% | 60.0% | 63.3% | 50.0% | ssr0004 |
| *psaJ* | 95.5% | 92.9% | 88.6% | 81.0% | 68.2% | | 62.2% | 60.5% | 61.0% | 51.4% | 47.5% | ssr0008 |
| *psbA* | 99.7% | 98.3% | 98.9% | 97.2% | 96.3% | | 86.6% | 90.5% | 89.9% | 89.9% | 85.1% | sll1867, slr1181, slr1311 |
| *psbB* | 98.6% | 96.3% | 95.5% | 90.9% | 90.7% | | 73.8% | 79.4% | 79.1% | 78.7% | 76.1% | slr0906 |
| *psbC* | 98.5% | 95.1% | 96.4% | 95.6% | 95.3% | | 77.7% | 85.0% | 80.1% | 83.1% | 81.0% | sll0851 |
| *psbD* | 98.9% | 94.9% | 98.0% | 96.6% | 96.6% | | 87.0% | 87.0% | 87.8% | 87.0% | 85.0% | sll0849, slr0927 |
| *psbE* | 100.0% | 97.6% | 97.6% | 88.0% | 94.0% | | 71.6% | 76.8% | 68.3% | 68.3% | 69.1% | ssr3451 |
| *psbF* | 97.4% | 97.4% | 97.4% | 94.9% | 94.9% | | 83.3% | 70.6% | 79.4% | 76.5% | 82.4% | ssr4006 |
| *psbH* | 93.2% | 87.1% | 90.4% | 68.5% | 80.8% | | 60.3% | 66.7% | 63.2% | 64.9% | 68.3% | ssl2598 |
| *psbI* | 100.0% | 100.0% | 97.2% | 94.4% | 88.9% | | 71.9% | 80.6% | 77.8% | 75.0% | 72.2% | ssr1001 |
| *psbJ* | 97.5% | 92.5% | 90.0% | 90.0% | 87.5% | | 61.5% | 70.0% | 63.9% | 66.7% | 67.6% | ssr1008 |
| *psbK* | 82.0% | 72.1% | 68.9% | 58.2% | 57.1% | | 50.0% | 66.7% | 69.0% | 70.5% | 69.8% | ssr1005 |
| *psbL* | 97.3% | 97.4% | 97.4% | 94.7% | 86.8% | | 78.9% | 76.3% | 86.8% | 76.3% | 75.0% | ssr0007 |
| *psbM* | 100.0% | 97.1% | 100.0% | 87.9% | 72.7% | | | 71.0% | | | 55.9% | ssr1003 |
| *psbN* | 100.0% | 97.7% | 97.7% | 86.0% | 86.0% | | 44.7% | 62.8% | 60.5% | 65.1% | 48.8% | ssr0009 |
| *psbT* | 93.9% | 93.9% | 93.9% | 87.5% | 93.8% | | 80.0% | 76.7% | 71.9% | 76.7% | 51.7% | ssr0001 |

***Photosynthetic metabolism***

| | *Nicotiana tabacum* | *Zea mays* | *Oryza sativa* | *Marchantia polymorpha* | *Pinus thunbergii* | *Epifagus virginiana* | *Euglena gracilis* | *Cyanophora paradoxa* | *Odontella sinensis* | *Porphyra purpurea* | *Synechocystis* sp. PCC6803 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *atpA* | 94.3% | 86.4% | 86.8% | 88.1% | 87.0% | | 77.4% | 77.5% | 72.1% | 76.1% | 72.7% | sll1326 |
| *atpB* | 92.8% | 91.0% | 91.8% | 89.7% | 88.7% | | 83.2% | 80.8% | 80.7% | 82.4% | 80.2% | slr1329 |
| *atpE* | 87.1% | 74.8% | 72.5% | 65.1% | 70.2% | | 40.3% | 35.7% | 36.9% | 47.7% | 42.0% | slr1330 |
| *atpF* | 88.0% | 72.6% | 74.3% | 50.3% | 63.6% | | 27.0% | 30.4% | 24.2% | 24.0% | 25.2% | sll1324 |
| *atpH* | 98.8% | 97.5% | 97.5% | 97.5% | 96.3% | | 82.7% | 90.1% | 87.5% | 85.2% | 81.3% | ssr2615 |
| *atpI* | 94.0% | 90.8% | 91.2% | 85.2% | 85.6% | | 71.4% | | 67.7% | 70.2% | 67.5% | sll1322 |
| *ndhA* | 81.5% | 80.4% | 80.2% | 70.1% | | | | | | | 60.5% | sll0519 |
| *ndhB* | 96.6% | 95.1% | 94.8% | 70.6% | | | | | | | 53.8% | sll0223 |
| *ndhC* | 90.0% | 85.0% | 85.8% | 71.7% | | | | | | | 65.0% | slr1279 |
| *ndhD* | 86.3% | 78.6% | 79.3% | 69.5% | | | | | | | 52.4% | sll0027, sll1733, slr0331, slr1291 |
| *ndhE* | 91.1% | 75.2% | 76.2% | 74.7% | | | | | | | 58.0% | sll0522 |
| *ndhF* | 76.1% | 69.3% | 67.3% | 54.7% | | | | | | | 57.5% | sll0226, sll1732, slr0644 |
| *ndhG* | 77.8% | 75.0% | 76.7% | 56.5% | | | | | | | 41.8% | sll0521 |
| *ndhH* | 91.6% | 85.8% | 85.5% | 82.3% | | | | | | | 69.3% | slr0261 |
| *ndhI* | 94.6% | 83.0% | 82.4% | 78.1% | | | | | | | 67.5% | sll0520 |
| *ndhJ* | 90.4% | 82.1% | 82.1% | 74.7% | | | | | | | 53.8% | slr1281 |
| *ndhK* | 89.8% | 83.7% | 84.0% | 71.6% | | | | | | | 67.7% | slr1280 |
| *rbcL* | 94.1% | 92.6% | 93.7% | 92.6% | 93.1% | | 85.9% | 84.0% | 58.7% | 57.2% | 81.0% | slr0009 |

***Others***

| | *Nicotiana tabacum* | *Zea mays* | *Oryza sativa* | *Marchantia polymorpha* | *Pinus thunbergii* | *Epifagus virginiana* | *Euglena gracilis* | *Cyanophora paradoxa* | *Odontella sinensis* | *Porphyra purpurea* | *Synechocystis* sp. PCC6803 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *accD* | 65.7% | | 47.3% | 70.2% | 62.8% | 57.2% | | | | 52.6% | 57.7% | sll0336 |
| *clpP* | 84.7% | 68.0% | 66.5% | 74.0% | 61.3% | 84.2% | | 38.6% | | | 49.7% | sll0534, slr0144, slr0165, slr0542 |
| *matK* | 63.9% | 53.2% | 52.5% | 33.9% | 43.2% | 47.6% | | | | | | |
| *ycf1* | 81.1% | | | 31.0% | 39.4% | 33.9% | | | | | | |
| *ycf2* | 88.7% | 52.7% | | 19.2% | 26.4% | 58.7% | | | | | | |
| *ycf3* | 100.0% | 93.7% | 92.1% | 87.2% | 86.4% | | | 74.2% | 57.4% | 66.9% | 63.5% | slr0823 |
| *ycf4* | 88.0% | 80.5% | 80.0% | 64.3% | 71.2% | | 35.8% | 41.0% | 38.3% | 47.5% | 45.9% | sll0226 |
| *ycf5* | 65.9% | 65.0% | 65.3% | 53.1% | 52.6% | | | 41.4% | 41.6% | 38.2% | 40.8% | sll1513 |
| *ycf6* | 100.0% | 96.6% | 100.0% | 86.2% | 89.7% | | | 79.3% | 69.0% | 69.0% | 55.2% | ssr0004 |
| *ycf7* | 93.5% | 83.9% | 83.9% | 67.7% | 69.0% | | | | 35.5% | 38.7% | | |
| *ycf9* | 95.2% | 83.9% | 90.3% | 83.9% | 75.8% | | 48.3% | 41.9% | 31.4% | 47.5% | 39.3% | sll1281 |
| *ycf10* | 79.5% | 62.0% | 59.0% | 49.4% | 56.4% | | | | | 30.4% | 29.5% | sll1485, slr1596 |
| *orf76* | 76.0% | 53.4% | | | | 30.6% | | | | | | |

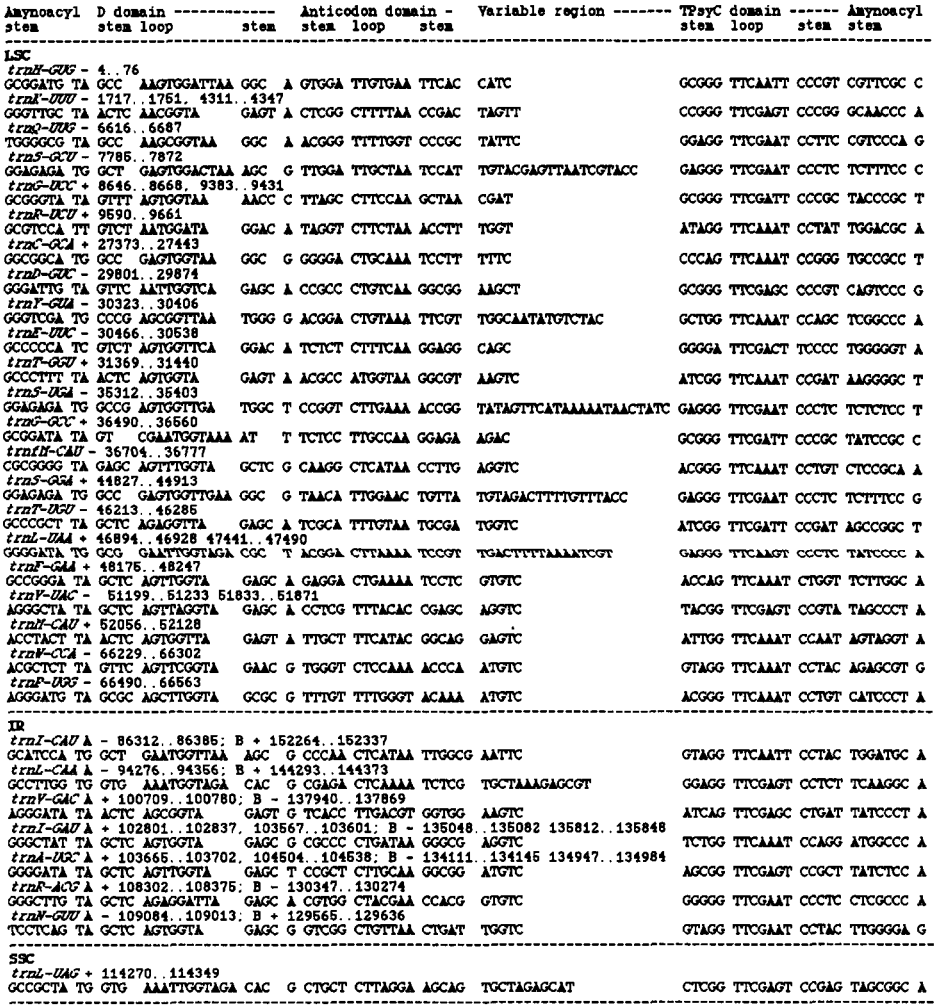


**Figure 3**. Structure of the tRNA genes in the *A. thaliana* chloroplast genome. The nucleotide sequences, nucleotide positions in the genome and the structural domains for the 37 tRNA genes are tabulated.

**Table 2**. The codon-anticodon recognition pattern and codon usage for the *A. thaliana* chloroplast genome. Numerals indicate the frequency of usage of each codon in 22,978 codons in 79 species of potential protein coding genes.

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| UUU | F | 979 | *trnF*-GAA | UCU | S | 495 | *trnS*-GGA | UAU | Y | 704 | *trnY*-GUA |
| UUC | F | 415 | | UCC | S | 248 | | UAC | Y | 148 | |
| UUA | L | 872 | *trnL*-UAA | UCA | S | 336 | *trnS*-UGA | UAA | - | 49 | |
| UUG | L | 440 | *trnL*-CAA | UCG | S | 163 | | UAG | - | 19 | |
| CUU | L | 504 | *trnL*-UAG | CCU | P | 373 | *trnP*-UGG | CAU | H | 394 | *trnH*-GUG |
| CUC | L | 154 | | CCC | P | 171 | | CAC | H | 128 | |
| CUA | L | 320 | | CCA | P | 259 | | CAA | Q | 652 | *trnQ*-UUG |
| CUG | L | 139 | | CCG | P | 117 | | CAG | Q | 170 | |
| AUU | I | 1024 | *trnI*-GAU | ACU | T | 481 | *trnT*-GGU | AAU | N | 852 | *trnN*-GUU |
| AUC | I | 341 | | ACC | T | 216 | | AAC | N | 257 | |
| AUA | I | 632 | *trnI*-CAU | ACA | T | 375 | *trnT*-UGU | AAA | K | 1016 | *trnK*-UUU |
| AUG | M | 519 | *trnM*-CAU *trnfM*-CAU | ACG | T | 115 | | AAG | K | 270 | |
| GUU | V | 483 | *trnV*-GAC | GCU | A | 594 | *trnA*-UGC | GAU | D | 704 | *trnD*-GUC |
| GUC | V | 152 | | GCC | A | 186 | | GAC | D | 166 | |
| GUA | V | 455 | *trnV*-UAC | GCA | A | 342 | | GAA | E | 944 | *trnE*-UUC |
| GUG | V | 176 | | GCG | A | 132 | | GAG | E | 270 | |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| UGU | C | 203 | *trnC*-GCA | CGU | R | 299 | *trnR*-ACG | AGU | S | 359 | *trnS*-CCU |
| UGC | C | 68 | | CGC | R | 103 | | AGC | S | 95 | |
| UGA | - | 12 | | CGA | R | 313 | | AGA | R | 380 | *trnR*-UCU |
| UGG | W | 400 | *trnW*-CCA | CGG | R | 100 | | AGG | R | 127 | |
| | | | | | | | | GGU | G | 534 | *trnG*-GCC |
| | | | | | | | | GGC | G | 149 | |
| | | | | | | | | GGA | G | 637 | *trnG*-UCC |
| | | | | | | | | GGG | G | 248 | |

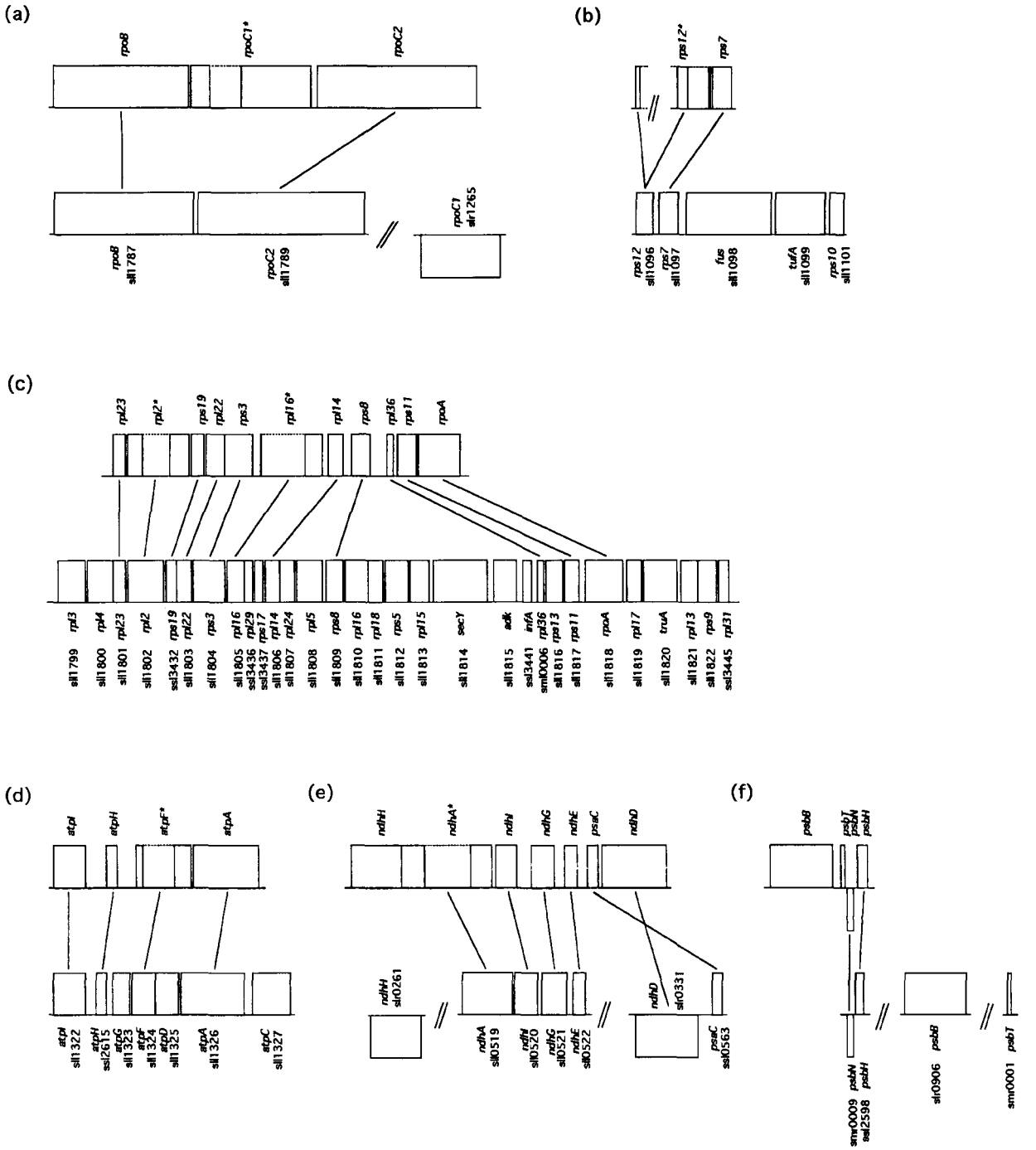


**Figure 4**. Structural comparison of gene clusters between the genomes of the *A. thaliana* chloroplast and the cyanobacterium *Synechocystis* sp. PCC6803. The clusters of the functionally related genes in *A. thaliana* chloroplast genome (*upper*) were compared with the corresponding regions in the cyanobacterial genome (*lower*). Genes are represented by open boxes on upper (transcribed from left to right) or lower (transcribed from right to left) side of the middle horizontal lines for each genome. The gene clusters compared are: (a) *rpoB-rpoC1-rpoC2*, (b) *rps12-rps7*, (c) *rpl23-rpl2-rps19-rpl22-rps3-rpl16-rpl14-rps8-rpl36-rps11-rpoA*, (d) *atpI-atpH-atpF-atpA*, (e) *ndhH-ndhA-ndhI-ndhG-ndhE-psaC-ndhD*, and (f) *psbB-psbT-psbN-psbH*.

the value varied from 64% to 100% depending on the gene species, indicating that significant diversity was generated between the two dicot plants. To obtain information on the relationship between gene function and divergence, comparison was made by dividing the identified genes into the following three functional categories: genes related to gene expression, photosynthetic apparatus and photosynthetic metabolism. The genes classified into gene expression varied from 72% to 100% (average identity: 87.9%), those classified into photosynthetic

metabolism from 76% to 100% (average identity: 89.7%), and those classified into photosynthetic apparatus from 82.0% to 100% (average identity: 97.0%). It is apparent that those of the former two gene categories are more divergent. Sequence conservation was also observed with the genes in the cyanobacterial genome (25.2% to 90.1%: average identity 59.6%) with the highest values in the category of photosynthetic apparatus (47.5% to 90.1%: average identity 70.4%).

### *3.3. Structural features of putative RNA-coding genes*

The *A. thaliana* chloroplast genome contained two copies of ribosomal RNA gene clusters (16S - 23S - 4.5S - 5S) in the two inverted repeat regions (Fig. 2). Each cluster was intervened by two tRNA genes, *trnI* and *trnA*, in the 16S and 23S spacer. The sequence identities with those of tabacco chloroplast were 98–100% in the coding regions and 92% on avarage in the spacer regions.

A total of 37 tRNA genes (30 gene species) representing 20 amino acid species were identified in the genome by similarity search and computer prediction. The nucleotide sequences of the gene products and the positions are summarized in Fig. 3. Six of the 30 tRNA gene species, *trn*K-UUU, *trn*G-UCC, *trn*L-UAA, *trn*V-UAC, *trn*I-GAU, and *trn*A-UGC, contained intervened sequences of 513–5,520 bp long at either the anticodon stem, the anticodon loop, or D-stem.

On the basis of the structural information derived from the entire protein and tRNA gene constituents of the genome, the frequency of codon usage and the recognition patterns of the codons and the corresponding anticodons were deduced (Table 2). No significant codon usage bias was observed. Thirty tRNA species can sufficiently recognize all the codons used in the genome except for *trn*R-ACG, where only one species of *trn*R was found for four kinds of the codons with different third letters. One possible explanation is that only the first two letters of the codons are recognized by tRNA. Alternatively, it could be that the corresponding tRNAs are supplied from the nuclear genome.

### *3.4. Genome structure in comparison with cyanobacterium*

Seventy-four out of 79 genes in the *A. thaliana* chloroplast genome were commonly found in the *Synechocystis* genome (Table 2). The structures of 13 clusters consisting of two or more adjoining genes with related functions in the chloroplast genome were compared with those of the corresponding regions in the cyanobacterial genome. The number and relative positions of the genes were conserved in 7 small clusters: *rpl33-rps18*, *atpB-atpE*, *ndhC-psbG-ndhJ*, *psaA-psaB*, *psbD-psbC*, *psbE-psbF-psbL-psbJ*, and *petB-petD*. Deletion, addition, or rearrangement of the genes in either genome were observed in 6 clusters: *rpoB-rpoC1-rpoC2*, *rps12-rps7*, *rpl23-rpl2-rps19-rpl22-rps3-rpl16-rpl14-rps8-rpl36-rps11-rpoA*, *atpI-atpH-atpF-atpA*, *ndhH-ndhA-ndhI-ndhG-ndhE-psaC-ndhD*, *psbB-psbT-psbN-psbH*, as shown in Fig. 4. Limited conservation of gene organization among the genomes of liverwort chloroplast, *Escherichia coli*, and *Synechococcus* has also been noted.[30] These observations would not only provide evidence supporting the endosymbiotic theory in which ancestral photosynthetic prokaryotes of cyanobacteria are the origin of plant chloroplasts, but also suggest that gene shuffling took place during the establishment of the cyanobacterium species.

**Acknowledgments:** We thank S. Sasamoto, K. Kawashima, S. Shinpo, A. Matsuno, A. Watanabe, M. Yasuda, M. Yatabe, and M. Yamada for excellent technical assistance. We are grateful to Mitsui Plant Biotechnology Research Institute, the Research Institute of Innovative Research for the Earth, the Arabidopsis Biological Resource Center at Ohio State University, and Dr. M. Sugiura for providing the DNA libraries and the DNA markers. This work was supported by the Kazusa DNA Research Institute Foundation.

## References

1. 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.
2. 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.
3. Hiratsuka, J., Shimada, H., Whittier, R. et al. 1898, 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.
4. Wolfe, K. H., Morden, C. W., and Palmer, J. D. 1992, Function and evolution of a minimal plastid genome from a nonphotosynthetic parasitic plant, *Proc. Natl. Acad. Sci. USA*, **89**, 10648–10652.
5. Maier, R. M., Neckermann, K., Igloi, G. L. et al. 1995, Complete sequence of the maize chloroplast genome: Gene content, hotspots of divergence and fine tuning of genetic information by transcript editing, *J. Mol. Biol.*, **251**, 614–628.
6. Wakasugi, T., Tsudzuki, J., Ito, S. et al. 1994, Loss of all ndh genes as determined by sequencing the entire chloroplast genome of the black *pine Pinus thunbergii*, *Proc. Natl. Acad. Sci. USA*, **91**, 9794–9798.
7. Wakasugi, T., Nishikawa, A., Yamada, K. et al. 1998, Complete nucleotide sequence of the plastid genome from a fern, *Psilotum nudum*, *Endocytobiosis Cell Res.*, **13** (Suppl.), 147.
8. 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.
9. Hallick, R. B., Hong, L., Drager, R. G. et al. 1993, Complete sequence of *Euglena gracilis* chloroplast DNA, *Nucleic Acid Res.*, **21**, 3537–3544.
10. Reith, M. and Munholland, J. 1995, Complete nucleotide sequence of the *Prophyra purpurea* chloroplast genome, *Plant. Mol. Biol. Rep.*, **13**, 333–335.
11. Kowallik, K. V., Stoebe, B., Schaffran, I. et al. 1995, The chloroplast genome of a chlorophyll a+c-containing alga, *Odontella sinensis*, *Plant. Mol. Biol. Rep.* **13**, 336–342.
12. Stirewalt, V. L., Michalowski, C. B., Loffelhardt, W. et al. 1995, Nucleotide sequence of the cyanelle genome from *Cyanophora paradoxa*, *Plant. Mol. Biol. Rep.*, **13**, 327–9332.
13. Ohta, N., Sato, N., and Kuroiwa, T. 1998, Analysis of the plastid genome of protoflorideophyceous algae *Cyanidioschzon merolae*, *Plant. Cell Physiol.*, **39** (Suppl.), s54.
14. 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.
15. Kaneko, T., Sato, S., Kotani, H. et al. 1996, Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions, *DNA Res.*, **3**, 109–139.
16. Martin, W., Stobe, B., Goremykin, V. et al. 1998, Gene transfer to the nucleus and the evolution of chloroplasts, *Nature*, **393**, 162–165.
17. Meyerowitz, E. M. and Somervill, C. R. (eds) 1994, *Arabidopsis*, Cold Spring Harbor Laboratory Press.
18. Kaiser, J. 1996, First global sequencing effort begins, *Science*, **274**, 30.
19. Meinke, D. W., Cherry, J. M., Dean, C. et al. 1998, *Arabidopsis thaliana*: A model plant for genome analysis, *Science*, **282**, 662–682.
20. Sato, S., Kotani, H., Nakamura, Y. et al. 1997, Structural analysis of *Arabidopsis thaliana* chromosome 5. I. Sequence features of the 1.6 Mb regions covered by twenty physically assigned P1 clones, *DNA Res.*, **4**, 215–230.
21. Liu, Y.-G., Mitsukawa, N., Vazquez-Tello, A., and Whittier, R. F. 1995, Generation of a high-quality P1 library of *Arabidopsis* suitable for chromosome walking, *Plant J.*, **7**, 351–358.
22. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. 1990, Basic local alignment search tool, *J. Mol. Biol.*, **215**, 403–410.
23. Lowe, T. M. and Eddy, S. R. 1997, tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence, *Nucl. Acids Res.*, **25**, 955–964.
24. Sugiura, M. 1992, The chloroplast genome, *Plant. Mol. Biol.*, **19**, 149–168.
25. Sugiura, M., Hirose, T., and Sugita, M. 1998, Evolution and mechanism of translation in chloroplasts, *Annu. Rev. Genet.*, **32**, 437–459.
26. Sugita, M., Kato, A., Shimada, H. et al. 1984, Sequence analysis of the junctions between a large inverted repeat and single-copy regions in tobacco chloroplast DNA, *Mol. Gen. Genet.*, **194**, 200–205.
27. Moon, E. and Wu., R. 1988, Organization and nucleotide sequence of genes at both junctions between the two inverted repeats and the large single-copy region in the rice chloroplast genome, *Gene*, **70**, 1–12.
28. Prombona, A. and Subramanian, A. R. 1989, A new rearrangement of angiosperm chloroplast DNA in Rye (*Secale cereale*) involving translocation and duplication of the ribosomal rpS15 gene, *J. Biol. Chem.*, **264**, 19060–19065.
29. Maier, R. M., Dory, I., Igloi, G. et al. 1990, The *ndhH* genes of gramminean plastomes are linked with the junctions between small single copy and inverted repeat regions, *Curr. Genet.*, **18**, 245–250.
30. Ozeki, H., Ohyama, K., Fukuzawa, H. et al. 1987, Genetic system of chloroplasts, *Cold Spring Harbor Symp. Quant. Biol.*, **52**, 791–804.