A GENE ORDER DATABASE OF PLASTID GENOMES

K Kurihara and T Kunisawa*

Dept of Applied Biological Sciences, Science University of Tokyo, Yamazaki 2641, Noda 278-8510, Japan
**Corresponding author.*
Email: kunisawa@rs.noda.tus.ac.jp

ABSTRACT

A gene order database of 32 completely sequenced plastid genomes was developed. The data structure is formally identical to that of the feature tables in the major GenBank/EMBL/DDBJ databases. The quality of annotations was largely improved. A normalizing gene-labeling system across the complete plastid genomes was developed so that comparative studies are made available without having to go back to sequence analysis. Many incorrect coordinates of tRNA-encoding regions found in the major databases were corrected. We attempted to distinctively label tRNA genes with the anticodon sequence CAT, which encodes either the initiator tRNA, elongator tRNA, or Ile-tRNA. The database is available at http://www.rs.noda. tus.ac.jp/~kunisawa.

Keywords: Database, Gene order, Plastid, Genome, Orthologous genes, Initiator tRNA, Elongator tRNA

1 INTRODUCTION

Nucleotide or amino acid substitutions provide measures of sequence similarity, which have been widely used to assess functional and phylogenetic relationships. The major sequence databases, GenBank, EMBL and DDBJ, have played a key role in this purpose. The advent of complete genomic sequence data has provided a new opportunity to investigate more macroscopic evolutionary events, such as duplication, inversion and transposition of parts of a genome. For example, phylogenetic trees are derived from the comparison of gene orders between different genomes (Sankoff, Leduc, Antoine, Paquin, Lang & Cedergren, 1992; Korbel, Snel, Huynen & Bork, 2002). Proteins encoded by gene pairs in a conserved order appear to interact physically (Dandekar, Snel, Huynen & Bork, 1998). The prevalence of small inversions is suggested in yeast genome evolution (Seoighe, Federspiel, Jones, Hansen, Bivolarovic, Surzycki et al., 2000). There are, however, no publicly available gene order databases. The major sequence databases can serve as surrogates for a gene order database. There are, however, a number of problems associated with the use of the major sequence databases for gene order comparison. The most serious has to do with the comparability of different database entries. Two orthologous genes may be labeled in different ways in two different database entries (sequences). This difficulty is overcome in the

COG database (Tatusov, Galperin, Natale & Koonin, 2000; Tatsuov, Natale, Garkavtsev, Tasuova, Shankavaram, Rao et al., 2001), in which a unique four-digit number is systematically assigned to each group of orthologous (plus paralogous) genes present in different genomes. Such efforts enable an easy comparison of the orders of protein-encoding genes between different genomes. However, genes coding for transfer and ribosomal RNAs are not taken into consideration in the COG database. As will be shown, errors are found at high frequencies in the major sequence databases with respect to the coordinates and/or annotations of genes specifying tRNAs. RNA-encoding genes were also neglected in a recent comparative analysis of protein sequences from 19 plastid genomes by Rivas, Lozano and Ortiz (2002). According to the GenBank/EMBL/DDBJ databases, more than 30 plastid genomes have been completely sequenced to date.

Under these circumstances, we have decided to develop a gene order database for complete plastid genomes, which can be regarded as excellent model systems in computational genomics studies because of the considerable number available and their small size. Plastid genomes are circular in shape and are of 100-200 kb in size, containing between 30 to 50 different RNA genes and 100 protein-encoding genes in land plants or 200 protein genes in red algae (Sugiura, 1995). We attempted to develop a normalizing gene-labeling system across the completely sequenced plastid genomes in a comparable way. We will show that gene order comparison presents another support both for the identification of orthologous genes with only a weak sequence similarity and for the assignment of a tRNA gene that has the anticodon sequence CAT encoding either initiator Met-tRNA, elongator Met-tRNA or Ile-tRNA. Initiator tRNA is used for the initiation of protein synthesis in all organisms including plastids, whereas elongator tRNA is used for the insertion of methionine into internal peptidic linkages (e.g., Marck & Grosjean, 2002). As well as both initiator and elongator Met-tRNAs, the same anticodon sequence CAT is shared by a peculiar Ile-tRNA species, which recognizes an isoleucine codon AUA by post-transcriptional base modification (Muramatsu, Yokoyama, Horie, Matsuda, Ueda, Yamaizumi et al., 1988). It is not always easy to distinguish among the three types of tRNA genes from sequence data alone, and their specification remains incomplete in the major databases. In the plastid genomes gene-order database, we have tried to distinctively label tRNA genes with the anticodon sequence CAT.

2 DATA

All the genomic sequence data were obtained from the GenBank/EMBL/DDBJ databases (GenBank, n.d.; EMBL, n.d.; DDBJ, n.d.). The completely sequenced plastid genomes are as follows: *Cyanophora paradoxa* (abbreviated as Cpa, GenBank/EMBL/DDBJ database accession No. U30821); *Cyanidium caldarium* (Cca, AF022186); *Cyanidioschyzon merolae* (Cme, AB002583); *Porphyra purpurea* (Ppu, U38804); *Odontella sinensis* (Osi, Z67753); *Guillardia theta* (Gth, AF041468); *Mesostigma viride* (Mvi, AF166114); *Nephroselmis olivacea* (Nol, AF137379); *Chlorella vulgaris* (Cvu, AB001684); *Chlamydomonas reinhardtii* (Cre, BK000554); *Astasia longa* (Alo, ALO294725); *Euglena gracilis* (Egr, X70810); *Chaetosphaeridium globosum* (Cgl, AF494278); *Marchantia polymorpha* (Mpo, X04465); *Anthoceros formosae* (Afo, AB086179); *Psilotum nudum* (Pnu, AP004638); *Adiantum capillus-veneris* (Aca, AY178864); *Pinus thunbergii* (Pth, D17510); *Pinus koraiensis* (Pko, AY228468); *Calycanthus fertilis* (Cfe, AJ428413); *Amborella trichopoda* (Atr, AJ506156); *Lotus japonicus* (Lja, AP002983), *Nicotiana tabacum* (Nta, Z00044); *Oenothera elata* (Oel, AJ271079); *Arabidopsis thaliana* (Ath, AP000423); *Spinacia oleracea* (Sol, AJ400848); *Epifagus virginiana* (Evi, M81884); *Oryza sativa* (Osa, X15901), *Triticum aestivum* (Tae, AB042240); *Zea mays* (Zma, X86563); *Toxoplasma gondii* (Tgo, U87145); *Eimeria tenella* (Ete, AY217738). *Medicago truncatula* chloroplast (ACO93544) is not included in the present gene order database, since no coding regions were identified in the GenBank/EMBL/DDBJ database entry.

3 ANNOTATIONS OF PROTEIN-ENCODING GENES

The first step in normalized gene-labeling across plastid genomes is the comparison of all the protein sequences from a plastid genome with all the proteins from other plastids using the FASTA computer program (Pearson & Lipman, 1988). Orthologous relationships were identified on the basis of sequence similarity. In the all-by-all FASTA analysis, we used two criteria for detecting orthologous gene pairs from different genomes; (i) a level of amino acid identity higher than 30% and (ii) a region of similarity longer than any of the halves of the either of two protein-lengths. Another gene from a third genome was included in this orthologous group if its protein sequence satisfied the homology criteria when compared to at least one member of the orthologous group. A unique label, which was taken from the GenBank/EMBL/DDBJ annotations, was assigned to the orthologous group thus identified. When orthologous genes were labeled in different ways in the major databases, we arbitrarily used one of the alternate labels (Table 1).

Table 1. Alternate gene labels.

used	synonym	used	synonym	used	synonym
carA	*trpG*	*moeB*	*chlN*	*petG*	*petE*
ccsA	*ycf5*	*nblA*	*ycf18*	*petL*	*ycf7*
cemA	*ycf10*	*ndhA*	*ndh1*	*petM*	*ycf31*
cfxQ	*cfxX/Q*	*ndhB*	*ndh2*	*petN*	*ycf6*
crtE	*preA*	*ndhC*	*ndh3*	*psbY*	*ycf32*
crtR	*desA*	*ndhD*	*ndh4*	*psbZ*	*ycf9*
cysA	*ycf85*	*ndhE*	*ndh4L*	*rpoZ*	*ycf61*
ftrC	*ftrB*	*ndhF*	*ndh5*	*tatC*	*ycf43*
hupA	*hlp*	*ndhK*	*psbG*	*thdF*	*trmE*
lysR	*ycf30*	*ntcA*	*ycf28*		
matK	*roaA*	*pdhA*	*odpA*		

Conserved open reading frames, which were not shared among at least two genomes, were represented in the form of OrfXY, where X or Y stands for an arbitrarily chosen letter of the alphabet. By contrast, non-conserved open reading frames were simply labeled “orf”. Seven paralogous gene families were found, (i) *psaA* and *psaB*, (ii) *psbA* and *psb*D, (iii) *psbE* and *psbF*, (iv) *psbL* and *psbT*, (v) *ndhA* and *ndhH*, (vi) *apcA*, *apcB*, *apcD*, *apcF*, *cpcA*, *cpcB* and *cpeB*, (vii) ycf27 and ycf29. On the basis of multiple sequence alignments by ClustalW (Thompson, Higgins, & Gibson, 1994) and gene-order comparisons between genomes we have confirmed that these paralogous genes were correctly labeled to reflect their orthologous relationships in the major databases. Using this all-by-all FASTA analysis, we were able to label more than 98% of a total of 3497 protein-encoding genes in a comparable way.

The second step is a gene order comparison for gene pairs that do not satisfy the homology criteria mentioned above. For each of these genes, their neighboring genes were examined and their gene orders were compared between genomes. We found that gene orders are well conserved for gene pairs that show sequence similarities of over 15% amino acid identity irrespective of the length of sequence similarity. A typical example where an amino acid identity is only 16% is observed in a comparison of two open reading frames labeled ORF111 from *Porphyra* (111 aa) and ycf41 from *Odontella* (113 aa) in the major databases. Although the level of sequence similarity is low, identical gene orders, ycf39-ORF111-psbI and ycf39-ycf41-psbI, are found in both genomes, suggesting an orthologous relationship between the gene pair. Based on such gene order comparisons, we were able to label about 60 protein-encoding genes, which remained unlabeled in the first step.

Using this methodology we labeled a total of 3497 individual protein-encoding genes from 32 plastid genomes in a consistent and comparable way. A substantial fraction of them, i.e. 2993 genes, were identically labeled to the major GenBank/EMBL/DDBJ databases. A complete list of differences between the present database and the major databases is given in Appendix 1. A major difference arises from the fact that most pre-existing gene-labels in the major databases are not updated when homologous relationships are found between a new sequence and pre-existing sequences. Alternate gene labels (synonyms) are not normalized in the major databases, which is another source of difference.

Plastid genomes encode many short proteins of less than 100 amino acids. This generally makes it difficult to detect orthologous relationships among them, since shorter proteins contain less information. Here, in addition to the primary sequence comparison of individual proteins, both protein length and gene order comparisons were included in orthology detection. Although we have used the identity % obtained by the FASTA alignment in the orthology criteria, this similarity measure can be replaced by the chance probability (P-value) of obtaining the FASTA sequence similarity score, which is known to be a better measure in the detection of weak similarity. Note that orthology detection based on the P-value alone becomes complex and unreliable when too many paralogs are present. The level of sequence similarity, protein length, and gene order are key elements in orthology identification

4 ANNOTATIONS OF RNA-ENCODING GENES

In the course of our initial survey of the GenBank/EMBL/DDBJ sequence entries, non-homogeneous annotations of tRNA-specifying genes were noticed. In one entry the anticodon species of a tRNA gene is listed, while in the other it is not listed. Similarly, the distinction between an initiator Met-tRNA and elongator Met-tRNA is well annotated in one entry, but is ignored in the other. Thus, we have also developed a normalizing gene-label for genes encoding tRNAs. Identification of tRNA genes was carried out with the tRNAscan computer program (Lowe & Eddy, 1997), which reports both the coordinates of a tRNA-specifying region along the genome and its corresponding anticodon species. Our search only failed to find eight tRNA-encoding regions that were annotated in the major databases, and a previously unmentioned Arg-tRNA gene was identified in the *Odontella* (Osi) plastid. The discrepancies between the GenBank/EMBL/DDBJ annotations and our search results are commented on in the present gene order database (see Section 5). Here we adopted a tRNA gene-labeling system using four letters; the first letter represents the amino acid species (in upper case) and remaining three show the anticodon sequence (in lower case), for instance, Fgaa for the Phe-tRNA with the anticodon GAA. While His-tRNAs possess an additional base at the 5' terminus, the "minus" 1 residue is often not correctly included in the major database annotations. In addition, there are a lot of mis-typing or mis-counting of base numbers for tRNA-encoding regions in the major databases. We have corrected such incorrect coordinates. Appendix 2 summarizes the corrections necessary in the major databases.

Although tRNAscan is a fine tool, every tRNA-specifying region with the anticodon sequence CAT is assigned as Met-tRNA. The tRNAscan does not distinguish between the initiator, elongator Met-tRNA and the peculiar Ile-tRNA species that recognizes an Ile codon AUA, all of which have the same anticodon sequence CAT. Thus, we have closely examined tRNA sequences identified by tRNAscan and have attempted to divide them into initiator tRNA (labeled fM), elongator tRNA (Mcat), and Ile-tRNA (Icat).

Our close examination has revealed characteristic sequences in the anticodon-loop region. Figure 1 lists nucleotide sequences that are divided into initiator tRNA (fM), together with three initiator sequences identified in the completely sequenced cyanobacteria, *Thermosynechococcus elongatus* (Tel, BA000039), *Synechocystis* sp. PCC6803 (Syn, AB001339) and *Nostoc* sp. PCC 7120 (Nos, BA000019), whose plastids are thought to share common ancestry. As shown in Figure 1, both the length and sequence of these initiator tRNA genes are well conserved across the cyanobacteria and plastids, although a one-base insertion is found at the D-loop region in an apicomplexan plastid of *Eimeria* (Ete). Most intitiator tRNA (formylated Met-tRNA, fM) sequences exhibit a uniquely conserved sequence, gCTCATAAc, where ¥

```
                         Acceptor  D              D  anticodon  anticodon    TΨC        TΨC   Acceptor
                         .******..++++........ .++++.-----.......-----.....=====.......=====******..
Thermosynechococcus      CGCGGGGTAGAGCAGTCTGGT-AGCTCGTCGGGCTCATAACCCGAAGGTCCATGGTTCAAATCCATGCCCCGCCA
Synechocystis            CGCGGGATAGAGCAGTCTGGT-AGCTCGTCGGGCTCATAACCCGAAGGTCGGTGGTTCAAATCCGCCTCCCGCCA
Nostoc                   CGCGGGATAGAGCAGCCTGGT-AGCTCGTCGGGCTCATAACCCGAAGGTCAGTGGTTCAAATCCACTTCCCGCCA

Cpa ( 62903.. 62976) -   CGCGGAGTAGAGCAGTTTGGT-AGCTCGTCGGGCTCATAACCCGAAGGTCAGTGGTTCAAATCCACTCTCCGCAA    rps4*-Ttgt*-fM-ycf36*-orf*
Cca (156623..156696) +   TGCGGAGTAGAGCAGTCTGGA-AGCTCGTCGGGCTCATAACCCGGAGGCCAATGGTTCGAATCCATTCTCCGCTA    petM -orf  -fM-orf   -psaD
Cme ( 92958.. 93031) -   CGCGGAGTAGAGCAGTCTGGT-AGCTCGTCGGGCTCATAACCCGAAGGCCAATGGTTCAAATCCATTCTCCGCTA    petM -orf  -fM-psaD  -Stga
Ppu ( 77711.. 77784) -   CGCGGGGTAGAGCAGTCTGGT-AGCTCGTCGGGCTCATAACCCGAAGGTCAATGGTTCAAATCCATTCCCCGCTA    ycf47-ycf36-fM-ycf42 -pbsA
Osi ( 62586.. 62659) +   CGCGGAGTAGAGCAGCCTGGT-AGCTCGTTGGGCTCATAACCCGAAGGTCAATGGTTCAAATCCATTCTCCGCTA    chlI -ycf47-fM-psaD  -Stga
Gth ( 89030.. 89103) +   CGCGGGGTAGAGCAGTCTGGT-AGCTCGTCGGGCTCATAACCCGAAGGCCAATGGTTCAAATCCATTCCCCGCTA    ycf47-ycf36-fM-psaD  -Stga
Mvi ( 51904.. 51977) +   CGCGGTGTAGAGCAGTCTGGT-AGCTCGTCGGGCTCATAACCCGAAGGTCGATGGTTCAAATCCACCCTCCGCAA    psaB -rps14-fM-rps16 -odpB*
Nol ( 42740.. 42813) +   TGCGGGGTAGAGCAGTCTGGT-AGCTCGCAGGGCTCATAACCCTGAAGTCGTGGGTTCAAATCCTACCTCCGCAC    rpoC2-rps2 -fM-ycf3* -petN
Cvu ( 37383.. 37456) +   AGCGGAGTAGAGCAGTCTGGT-AGCTCGTAAGGCTCATAACCTTAAGGCCGTGGGTTCGAATCCTACCTCCGCTC    orf *-Stga -fM-Ettc  -rpl20
Alo ( 37073.. 37146) -   GGCGAAGTAGAGTAAAAGGTT-AGCTCGTGGGCCTCATGACCCCAAGGTTAAAGGTTCGAATCCTTTCTTCGCCA    Rtct*-Ttgt -fM-Ggcc  -Sgct*
Egr ( 30968.. 31041) +   GGCGGAGTAGAGCAGTCAGGT-AGCTCGCAGGGCTCATAATCCTGAAGTCAGAGGTTCAAATCCTTTCTCCGCTA    Ttgt--Ggcc--fM-Sgct* -Qttg*
Cgl ( 41514.. 41587) +   CGCGGAGTAGAGCAGTCTGGT-AGCTCGCAAGGCTCATAACCTTGAAGTCATAGGTTCAAATCCTGTCTCCGCTA    petL*-rps14-fM-Ygta* -Ettc*
Mpo ( 42156.. 42229) -   CGCGGAGTAGAGCAGTCTGGT-AGCTCGCAAGGCTCATAACCTTGAGGTCATAGGTTCAAATCCTGTCTCCGCCA    psaB -rps14-fM-Ggcc* -psbZ*
Afo ( 54026.. 54099) -   CGCGGGGTAGAGCAGCCTGGT-AGCTCGCAAGGCTCATAACCTTGAGGTCACGGGTTCAAATCCTGTCTCCGCCA    psaB -rps14-fM-Ggcc* -psbZ*
Pnu ( 39655.. 39728) -   CGCGGGATAGAGCAGCTTGGT-AGCTCGTAAGGCTCATAACCTTGAGGTCACGGGTTCAAATCCCGTACCCGCAA    psaB -rps14-fM-Tggt  -psbD
Aca ( 35201.. 35274) -   CGCGGGGTGGAGCAGCTTGGT-AGCTCGCGAGGCTCATAACCTCGAGGTCACGGGTTCAAATCCCGTCTCCGCAA    psaB -rps14-fM-Tggt  -psbD
Pth ( 78886.. 78959) +   TGCGGAGTAGAGTAGTCTGGT-AGCTCGCAAGGCTCATAACCTTGAGGTCACGGGTTCAAATCCTGTCTCCGCCA    psaB -rps14-fM-Ggcc* -psbZ*
Pko ( 76762.. 76835) +   TGCGGAGTAGAGTAGTCTGGT-AGCTCGCAAGGCTCATAACCTTGAGGTCACGGGTTCAAATCCTGTCTCCGCCA    orf* -rps14-fM-Ggcc* -Stga
Cfe ( 37712.. 37785) -   CGCGGGGTAGAGCAGTTTGGT-AGCTCGCAAGGCTCATAACCTTGAGGTCACGGGTTCAAATCCTGTCTCCGCAA    psaB -rps14-fM-Ggcc* -psbZ*
Lja ( 24093.. 24166) +   CGCGGGGTAGAGCAACTTGGT-AGCTCGCAAGGCTCATAACCTTGAGGTCACGGGTTCAAATCCCGTCTCCGCAA    psaB -rps14-fM-Ggcc* -psbZ*
Nta ( 38356.. 38429) -   CGCGGGGTAGAGCAGTTTGGT-AGCTCGCAAGGCTCATAACCTTGAGGTCACGGGTTCAAATCCTGTCTCCGCAA    psaB -rps14-fM-Ggcc* -psbZ*
Oel ( 28414.. 28487) +   CGCGGGGTAGAGCAGATTGGT-AGCTCGCAAGGCTCATAACCTTGAGGTCATGGGTTCGAATCCCGTCCCCGCAC    psaB -rps14-fM-fM     -Ggcc*
Oel ( 28505.. 28578) +   CGCGGGGTAGAGCAGATTGGT-AGCTCGCAAGGCTCATAACCTTGAGGTCATGGGTTCAAATCCCGTCCCCGCAC    rps14-fM   -fM-Ggcc* -psbZ*
Ath ( 36704.. 36777) -   CGCGGGGTAGAGCAGTTTGGT-AGCTCGCAAGGCTCATAACCTTGAGGTCACGGGTTCAAATCCTGTCTCCGCAA    psaB -rps14-fM-Ggcc* -psbZ*
Sol ( 35420.. 35493) -   CGCGGGGTAGAGCAGTTTGGT-AGCTCGCAAGGCTCATAACCTTGAGGTCACGGGTTCAAATCCTGTCTCCGCAA    psaB -rps14-fM-Ggcc* -psbZ*
Evi (  5334..  5407) -   GGCGGGGTAGAGCAGTTTGGT-AGCTCGCAAGGCTCATAACCTTGAGGTCACGGGTTCAAATCCTGTCTCCGCAA    Sgga -rps14-fM-Stga  -Ettc
Osa ( 12839.. 12912) -   AGCGGAGTAGAGCAGTTTGGT-AGCTCACGAGGCTCATAACCTTGAGGTCACGGGTTCGATTCCCGTCTCCGCAC    ycf70-Gtcc -fM-Ggcc* -psbZ*
Tae ( 13003.. 13076) -   AGCGGAGTAGAGCAGTTTGGT-AGCTCACGAGGCTCATAACCTTGAGGTCACGGGTTCGATTCCCGTCTCCGCAC    Icat -Gtcc -fM-Ggcc* -psbZ*
Zma ( 13073.. 13146) -   AGCGGAGTAGAGCAGTTTGGT-AGCTCACGAGGCTCATAACCTTGAGGTCACGGGTTCGATTCCCGTCTCCGCAC    ycf70-Gtcc -fM-Ggcc* -psbZ*
Tgo (  2380..  2453) +   AGCGGGGTAGAGCAGGTTGGT-AGCTCGTCGGGCTCATGACCCGAAGGTCAGCGGTTCAAATCGGCTCCTCGTTT    Vtac*-Racg- fM-23s   -Ttgt
Tgo ( 32544.. 32617) -   AGCGGGGTAGAGCAGGTTGGT-AGCTCGTCGGGCTCATGACCCGAAGGTCAGCGGTTCAAATCGGCTCCTCGTTT    Vtac*-Racg- fM-23s   -Ttgt
Ete (  2379..  2453) +   AACGGAGTAGAGCAGTCTGGTTAGCTCATCGGGCTCATGATCCGAAGGTCAACGGTTCAATTCCGTTCTCCGTTT    Vtac*-Racg- fM-23s   -Ttgt
Ete ( 32269.. 32343) -   AACGGAGTAGAGCAGTCTGGTTAGCTCATCGGGCTCATGATCCGAAGGTCAACGGTTCAATTCCGTTCTCCGTTT    Vtac*-Racg- fM-23s   -Ttgt
                         .******..++++........ .++++.-----.......-----.....=====.......=====******..
Characteristic Feature   !         A                     gCTCATAAc                                  !
```

Figure 1. Comparison of initiator tRNA (fM) gene sequences and gene orders. The tRNA secondary structure is indicated: Acceptor-stem, "*"; D-stem, "+"; Anticodon-stem, "-"; TYC-stem, "=". Gene orders are shwon in the direction from 5' to 3'. Genes encoded on the complementary chain are indicated "*". For more details, see text.

nucleotides located at the anticodon-loop region are shown in upper case letters, while lower case letters indicate the nucleotides at the anticodon-stem region. Exceptions are found only in euglenozoa plastids (Alo and Egr), which are colored in red in Figure 1. The other characteristic feature is that a base A is commonly found in this group of tRNA sequences at the second position of the D-stem but is not found in the other two groups, Icat and Mcat. Once tRNA-encoding genes are thus comparably labeled, we are ready to compare gene orders in the neighborhood of a tRNA gene. The present assignment of fM was examined in the light of gene order comparison. As shown at the right of Figure 1, conservation in gene order is observed in the neighborhood of the fM genes. The ribosomal protein S14 gene *rps14* and/or Ggcc are adjacent to fM in most of the green plants, suggesting a common ancestry for these fM genes. Similarly, in red algal plastids (Cca, Cme, Ppu, Osi, and Gth) fM is located adjacent to *psaD* and/or *ycf36*. In *Cyanophora* (Cpa) and euglenozoa (Alo and Egr), Tgtg is located upstream of fM. In this way, an examination of gene order conservation was helpful in the assignment of these tRNA species. It should be noted that the apicomplexan (Tgo and Ete) fM genes do not share gene orders with other plastids and that their nucleotide sequences differ considerably from others. Therefore, the present assignment of these apicomplexan tRNA sequences as fM should be viewed as a tentative one. For this reason, these apicomplexan tRNAs are labeled “fM?” in the present database. In bacteria, a mismatch (non Watson-Crick) pairing at the first position of the acceptor-stem (marked “!” at the bottom of Figure 1) constitutes an identity element of fM, which is believed to be involved in its recognition by Met-RNA transformylase (Marck, & Grosjean, 2002). An A:T pairing at the first acceptor-stem position is found in chlorophytes (Nol and Cvu) and apicomplexa (Tgo and Ete), which resembles archaeal initiators (Marck, & Grosjean, 2002). Although the *Astasia* (Alo) tRNA sequence shows a G:C pair at that position, its gene order is similar to that in another euglenozoa Egr, which shows a mismatch. Thus, the *Astasia* tRNA is likely to be an initiator fM gene, with the reservation that experimental confirmation is needed.

Figure 2 compares sequences that appear to be Ile-tRNA (Icat) with a CAT anticodon, which can recognize the Ile codon AUA by a post-transcriptional modification of the base C at the first position of the anticodon into lysidine (Muramatsu et al., 1988). All but one of these sequences exhibit a characteristic sequence, aCTCATAAt, in the anticodon-loop and -stem regions, which is uniquely found in Icat. An exception is observed in *Cynophora* (Cpa), where the first base of the characteristic sequence is replaced by G, as shown in red on the left of Figure 2. Other non-canonical bases are also observed in the apicomplxan plastids (Tgo and Ete). It is to be noted that the third position of the acceptor-stem region is commonly occupied by base A in all members of this group. This feature is not found in the other two groups of initiator and elongator tRNAs. As shown in Figure 2, some similarities in gene arrangement are found among red algal plastids (Cca, Cme, Ppu, Osi and Gth) or among green plant chloroplasts (Mvi to Zma in Figure 1). Thus, the examination of gene order further supports the present classification of most of these tRNA genes.

```
                         Acceptor  D            D  anticodon    anticodon             TΨC          TΨC  Acceptor
                         *******..+++......... .+++.-----.......-----...              ..=====......=====*******.
Thermosynechococcus      CCAGGGTTGGCCGAGCGGATG-AGGCAGCGAACTCATAATTCGCCAT------------AGGCTGGTTCGACTCCAGCACCCTGGA
Synechocystis            CCAGGGTTGGCCGAGCGGTTG-AGGCAGCGAACTCATAATTCGCCCT------------AGACAGGTTCAACTCCTGTACCCTGGA
Nostoc                   CCAGGGTTGGCCGAGCGGTTG-AGGCAGCGAACTCATAATTCGCCCA------------AGGCAGGTTCAACTCCTGCACCCTGGA

Cpa ( 85048.. 85132) +   GCATCTGTGGCCGAGCGGTTGAAGGCAGCGGGCTCATAATCCGTCATCT-GAAA-AGATATCACTGGTTCGAATCCAGTCAGATGCA   rpoC1*-rpoB* -Icat -Fgaa  -rps16
Cca (109401..109484) -   GCATCTATGGCCGAGTGGCTTAAGGCAGCGGACTCATAATCCGTCGACAT-AA--TGTCATCGCTGGTTCAAATCCGGCTAGATGCA   thdF  -chlI  -Icat -infC  -cysA
Cme ( 13223.. 13309) -   GCATCTATGGCCGAGCGGCTTAAGGCAGCGGACTCATAATCCGTGGACAAGAATTTGTCATCGCTGGTTCGAATCCAGCTGGATGCA   thdF  -chlI  -Icat -infC  -cysA
Ppu ( 34285.. 34371) -   GCATCTGTGGCCGAGGGGCCGAAGGCAGCGGACTCATAATCCGCCATTTCGAAAGAGACGTCGCTGGTTCGAATCCAGCCAGATGCA   orf*  -ycf10 -Icat -infC* -ilvH
Osi ( 96216.. 96300) +   GCATTCGTGGCCGAGTGGTTGAAGGCACCGGACTCATAATCCGTTTTCCTCT--GGAACGTCACTGGTTCGAACCCAGTCGGATGCA   orf*  -Rccg  -Icat -rpl19 -petF
Gth ( 50077.. 50162) -   GCATCTGTGGCCGAGTGGTCGAAGGCACCGGACTCATAATCCGTC-TCTTGTAAAAGACAACGCTGGTTCAAACCCAGCCGGATGCA   Rccg  -orf   -Icat -ilvH  -Ltaa
Mvi ( 21193.. 21279) +   GCATCTATTGCCGAGAGGCCGAAGGCGGCGGACTCATAATCCGTTATCTCGAAAGAGACATCGCTGGTTCGAATCCAGCTGGATGCA   Aggc* -ycf3  -Icat -rbcL* -atpB
Cvu (  8330..  8413) +   GCACCTATGGCAGAGTGGTCGATTGCACCGCACTCATAATGCGGTTTC--GAAA-GAACATCGTTGGTTCAAACCCAACTGGGTGCA   orf   -orf*  -Icat -orf*  -orf*
Alo (  8904..  8987) +   GCATTTATGGCAGAGAGGACGATAGCACGGGACTCATAATCTCGTTCC--GAAA-GGACATCGCTGGTTCAAATCCAGCTGAATGCA   rps8  -rpl36 -Icat -rps14 -rps14
Egr ( 60338.. 60421) +   GCATTTATGGCAGAGTGGACGATAGCACGGGACTCATAATCTCGCTCC--GGAA-GGACGTCGCTGGTTCGAATCCAGCTGAATGCA   rps8  -rpl36 -Icat -rps14 -rps14
Nol ( 96925.. 96997) -   GCATCCATAGCCTAGCGGTTA-AGGCAGTCGACTCATAATCGGAATA------------TCGCTGGTTCGAATCCAGCTGGATGCA   OrfAU -rbcL  -Icat -ycf62*-chlB
Nol (195929..196001) +   GCATCCATAGCCTAGCGGTTA-AGGCAGTCGACTCATAATCGGAATA------------TCGCTGGTTCGAATCCAGCTGGATGCA   OrfAU -rbcL  -Icat -ycf62*-chlB
Cgl ( 88615.. 88688) -   GCATCTATAGCCGAGTGGTTA-AGGCACCCAACTCATAATTGGAGAA-----------CTCGCAGGTTCGAATCCTGCTAGATGCA   Vgac* -OrfCL -Icat -rpl23 -rpl2
Mpo ( 80984.. 81057) -   GCATCCATGGCTGAATGGTTA-AAGCACCCAACTCATAATTGGCGAA-----------TTCACAGGTTCAATTCCTGTTGGATGCA   16s*  -Vgac* -Icat -rpl23 -rpl2
Afo (105497..105570) -   GCATCCATGGCTGAACGGTTA-AAGCACCCAACTCATAATTGGCGAA-----------TTCACAGGTTCAACTCCTGTTGGATGCA   ndhB  -ndhB  -Icat -rpl23 -rpl2
Pnu ( 84463.. 84536) -   GCATCCATGGCTGAATGGTAA-AAGCACCCGACTCATAATCGGCGAA-----------TTCGCAGGTTCAATTCCTGTTGGATGCA   Lcaa  -OrfBZ*-Icat -rpl23 -rpl2
Aca ( 82161.. 82234) -   GCATCCATGGCTGAACGGTCA-AAGCACCCAACTCATAATTGGCGAA-----------TTCACAGGTTCAACTCCTGTTGGATGCA   Racg  -Ttgt* -Icat -rpl23 -rpl2
Pth ( 65938.. 66011) -   GCATCCATGGCTGAATGGTTA-AAGCGCCCAACTCATAATTGGCGAA-----------CTCGCGGGTTCAATTCCTGCTGGATGCA   Fgaa  -OrfAN*-Icat -psbA  -rpl23
Pth (119393..119466) +   GCATCCATGGCTGAATGGTTA-AAGCGCCCAACTCATAATTGGCGAA-----------CTCGCGGGTTCAATTCCTGCTGGATGCA   Hgtg* -OrfAN*-Icat -psbA* -rpoB
Pko ( 64070.. 64143) -   GCATCCATGGCTGAATGGTTA-AAGCGCCCAACTCATAATTGGCGAA-----------CTCGCGGGTTCAATTCCTGCTGGATGCA   orf   -OrfAN*-Icat -psbA* -rpl23
Pko (116557..116630) +   GCATCCATGGCTGAATGGTTA-AAGCGCCCAACTCATAATTGGCGAA-----------CTCGCGGGTTCAATTCCTGCTGGATGCA   ftsH* -Hgtg* -Icat -psbA* -rpoB
Cfe ( 87413.. 87486) -   GCATCCATGGCTGAATGGTTA-AAGCGCCCAACTCATAATTGGCGAA-----------TTCGTAGGTTCAATTCCTGCTGGATGCA   ycf15*-ftsH* -Icat -rpl23 -rpl2
Cfe (152796..152869) +   GCATCCATGGCTGAATGGTTA-AAGCGCCCAACTCATAATTGGCGAA-----------TTCGTAGGTTCAATTCCTGCTGGATGCA   ycf15*-ftsH* -Icat -rpl23 -fM
Lja ( 84258.. 84331) -   GCATCCATGGCTGAATGGTTA-AAGCGCCCAACTCATAATTGGCGAA-----------TTCGTAGGTTCAATTCCTACTGGATGCA   Lcaa  -ftsH* -Icat -rpl23 -rpl2
Lja (148125..148198) +   GCATCCATGGCTGAATGGTTA-AAGCGCCCAACTCATAATTGGCGAA-----------TTCGTAGGTTCAATTCCTACTGGATGCA   Lcaa  -ftsH* -Icat -rpl23 -rpl2
Nta ( 88699.. 88772) -   GCATCCATGGCTGAATGGTTA-AAGCGCCCAACTCATAATTGGCGAA-----------TTCGTAGGTTCAATTCCTACTGGATGCA   ycf15*-ftsH* -Icat -rpl23 -rpl2
Nta (153854..153927) +   GCATCCATGGCTGAATGGTTA-AAGCGCCCAACTCATAATTGGCGAA-----------TTCGTAGGTTCAATTCCTACTGGATGCA   ycf15*-ftsH* -Icat -rpl23 -rpl2
Oel ( 91548.. 91621) -   GCATCCATGGCTGAATGGTTA-AAGCGCCCAACTCATAATTGGCGAA-----------TTCGTGGGTTCAATTCCTACTGGATGCA   pseud*-ftsH* -Icat -rpl23 -rpl2
Oel (161707..161780) +   GCATCCATGGCTGAATGGTTA-AAGCGCCCAACTCATAATTGGCGAA-----------TTCGTGGGTTCAATTCCTACTGGATGCA   pseud^-ftsH* -Icat -rpl23 -rpl2
Ath ( 86312.. 86385) -   GCATCCATGGCTGAATGGTTA-AAGCGCCCAACTCATAATTGGCGAA-----------TTCGTAGGTTCAATTCCTACTGGATGCA   ycf15*-ftsH* -Icat -rpl23 -rpl2
Ath (152264..152337) +   GCATCCATGGCTGAATGGTTA-AAGCGCCCAACTCATAATTGGCGAA-----------TTCGTAGGTTCAATTCCTACTGGATGCA   ycf15*-ftsH* -Icat -rpl23 -rpl2
Sol ( 84198.. 84271) -   GCATCCATGGCTGAATGGTTA-AAGCGCCCAACTCATAATTGGCGAA-----------TTCGTAGGTTCAATTCCTACTGGATGCA   orf*  -ycf2* -Icat -rpl2  -rps19
Sol (149174..149247) +   GCATCCATGGCTGAATGGTTA-AAGCGCCCAACTCATAATTGGCGAA-----------TTCGTAGGTTCAATTCCTACTGGATGCA   rps12^-ftsH* -Icat -rpl2  -OrfBU
Evi ( 21878.. 21951) -   GCATCCATGGCTTAATGGTTA-AAGCGCCCAACTCATAATTGGCGAA-----------TTTGTAGGTTCAATTCCTACTGGATGCA   Lcaa  -ftsH* -Icat -rpl23 -rpl2
Evi ( 67877.. 67950) +   GCATCCATGGCTTAATGGTTA-AAGCGCCCAACTCATAATTGGCGAA-----------TTTGTAGGTTCAATTCCTACTGGATGCA   Lcaa  -ftsH* -Icat -rpl23^-rpl2
Osa ( 83139.. 83212) -   GCATCCATGGCTGAATGGTTA-AAGCGCCCAACTCATAATTGGTAAA-----------TTTGCGGGTTCAATTCCTGCTGGATGCA   OrfAJ*-ftsH* -Icat -rpl23 -rpl2
Osa (131906..131979) +   GCATCCATGGCTGAATGGTTA-AAGCGCCCAACTCATAATTGGTAAA-----------TTTGCGGGTTCAATTCCTGCTGGATGCA   OrfAJ*-ftsH* -Icat -rpl23 -rpl2
Tae ( 82901.. 82974) -   GCATCCATGGCTGAATGGTTA-AAGCGCCCAACTCATAATTGGTAAA-----------TTTGCGGGTTCAATTCCTGCTGGATGCA   ndhB  -Lcaa  -Icat -rpl23 -rpl2
Tae (131920..131993) +   GCATCCATGGCTGAATGGTTA-AAGCGCCCAACTCATAATTGGTAAA-----------TTTGCGGGTTCAATTCCTGCTGGATGCA   ndhB  -Lcaa  -Icat -rpl23 -rpl2
Zma ( 84881.. 84954) -   GCATCCATGGCTGAATGGTTA-AAGCGCCCAACTCATAATTGGTAAA-----------TTTGCGGGTTCAATTCCTGCTGGATGCA   ftsH* -OrfAO*-Icat -rpl23 -rpl2
Zma (137783..137856) +   GCATCCATGGCTGAATGGTTA-AAGCGCCCAACTCATAATTGGTAAA-----------TTTGCGGGTTCAATTCCTGCTGGATGCA   ftsH* -OrfAO*-Icat -rpl23 -rpl2
Tgo (  6386..  6461) +   ATACTTGTGGCTGAGTGGGCAAAAGCAGTGAGCTCATAACTCATATA-----------AAACGAAAGTTCGAATCTTTTCAAGTATA   Cgca  -Ltaa #-fM   -Ygta  -Sgct
Ete (  6231..  6307) +   GTACCTGTGGCTGAGTGGTCAAAAGCGGTGGGCTCATAATCCATTTT----------TTTTCAAAAGTTCAAATCTTTTCAGGTATA   Cgca  -Ltaa  -fM   -Ygta  -Sgct
                         *******..+++......... .+++.-----.......-----...             ...=====......=====*******.
Characteristic Feature     A                            aCTCATAAt
```

Figure 2. Comparison of Ile-tRNA gene (Icat) sequences and gene orders.

Elongator tRNA (Mcat) genes appear to possess a characteristic sequence at the anticodon-loop region, ACTCATAAG or ATTCATAAG in the non-green plastids, *Cyanophora* (Cpa) to euglenozoa (Egr), or tTTCATACg in green plant chloroplasts (Nol to Zma), as seen in Figure 3. Mismatches to the characteristic sequences are found in Cvu and Pnu, which are indicated in red in Figure 3. The Cvu tRNA gene is puzzling; a non-initiator is suggested from the G:C pair at the first position of the acceptor-stem, whereas its nearest neighbor genes, *rps14* and Ggcc, are identical to those found in the initiator tRNAs from green plant chloroplasts. Conserved orders of other tRNA genes confirm the classification as elongator tRNA.

The annotations of genes encoding ribosomal RNAs in the GenBank/EMBL/DDBJ databases were confirmed based on the multiple sequence alignment by CLUSTALW (Thompson, Higgins, & Gibson, 1994). An incorrect coordinate of *Oenothera* (Oel) 5S rRNA gene was found in this analysis. The correction is shown in Appendix 2, which lists revisions necessary to the GenBank/EMBL/DDBJ annotations of ribosomal and transfer RNA genes.

5 DATABASE STRUCTURE

The labeled arrangements of genes on the 32 plastid genomes are structured as a database, which is available online. The gene order of a plastid genome is accessible from the front page of the Web site (see Figure 4). Figure 5 illustrates our database format. The GenBank/EMBL/DDBJ database accession number is given below the species name. Columns 1-6, and 8-13 represent the coordinates of a gene. When a gene is encoded on a strand, the sequence of which is stored in the major databases, its strandness "+" is shown in column 15. Conversely, if encoded on the complementary strand, "-" is used. Pseudogenes are indicated by "&" in column 17. When tRNAscan identifies a tRNA gene that is not listed in the major databases, the symbol "%" appears in the column 17. If the tRNAscan fails to identify a tRNA gene that is annotated in the major databases, another symbol "#" is used. The gene labels developed here are shown in columns 21-30. We adopted a numbering system for exons; the second exon (counted from the 5' end) of gene *XYZ*, for instance, is labeled *XYZ*_2. Original annotations in the major databases are retained after column 41.

Amino acid or nucleotide sequence data are accessible using the link function; a user can confirm the present annotation and infer its biological function by performing the FASTA search or the CLUSTALW multiple alignment, both of which are available at various Web sites.

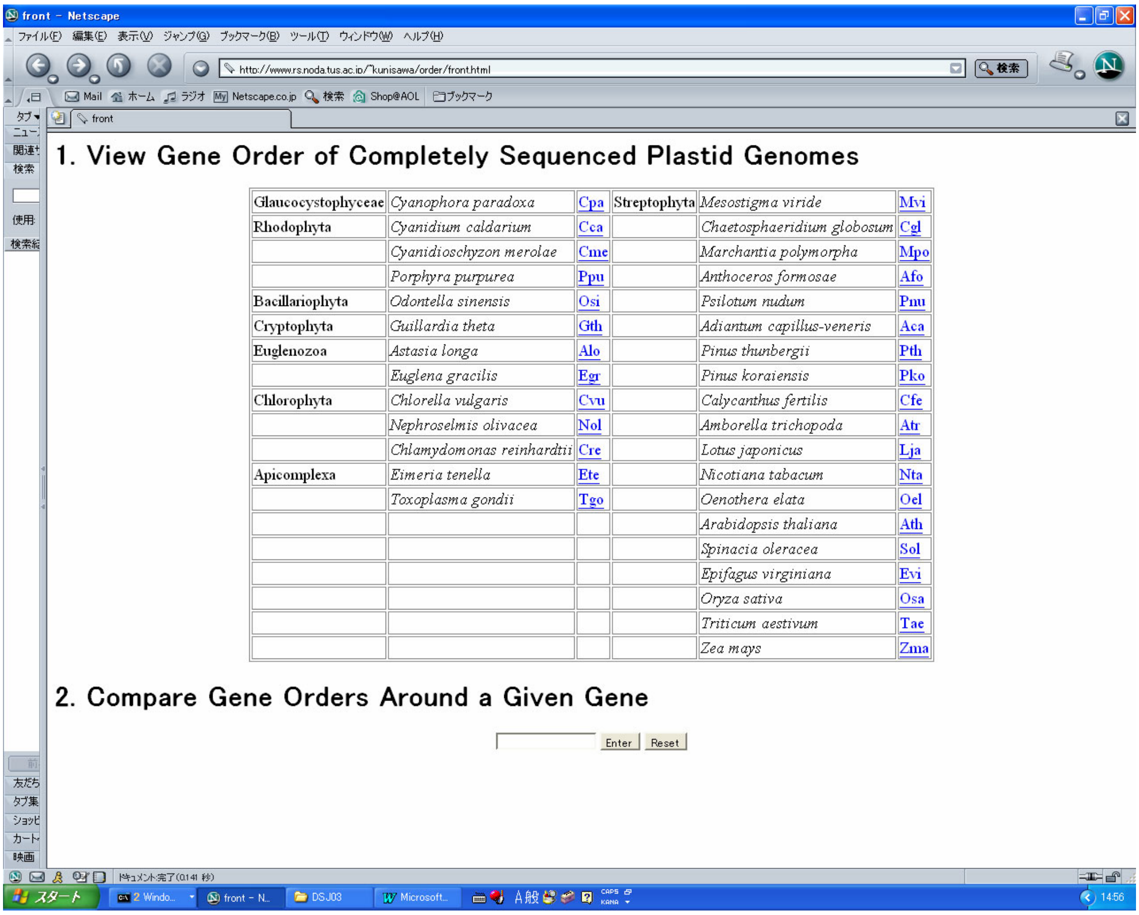
Another function of the database is to show gene orders in the neighborhood of a given gene. When a user lists one gene symbol in the front page, five genes present at both the 5' and 3' regions of the given gene are displayed, as shown in Figures 1, 2 and 3. When a gene is present in two copies on a genome, the

```
                         Acceptor  D            D   Anticodon    Anticodon TC          TC  Acceptor
                         *******..++++.........++++.----.........----.....=====.......=====*******.
Thermosynechcococcus     GGCTCAGTAGCTCAGT-GGTTAGAGCAGGGGACTCATAAGCCCAAGGTCGCAGGTTCGAATCCCGCCTGAGCCA
Synechocystis (Partial)   GCTTGGTAGCTCAGTTGGTTAGAGCAGGGGACTCATAAGCCCAAGGTCGGCGGTTCA
Nostoc                   GGCTCAGTAGCTCAGTTGGTTAGAGCACGGGACTCATAAGCCTGGGGTCGTTGGTTCAAATCCGACCTGAGCCA

Cpa ( 87985.. 88056) +   GGCTCGGTAGCTCAGTGG--TAGAGCAGGGGACTCATAAGCCCTTGGTCGTGGGTTCAAATCCCACTTGAGCCA      psaE* -ycf17 -Mcat-psbI -petL*
Cca (142204..142277) -   GGCTTAGTAGCTCAGTGGTTTAGAGCAGGGGATTCATAAGCCCAAGGTCGTAAGTTCAAGTCTTATCTAAGCCA      ycf26 -argB  -Mcat-dnaK*-rpl3
Cme (104198..104270) +   GGCTCAGTAGCTCAGAGG-TTAGAGCGGGGGACTCATAAGCCTCAGGTCGTAGGTTCAAATCTTACCTGAGCCA      apcB  -argB  -Mcat-hlp  -dnaK*
Ppu (116680..116752) -   GGCTCAGTAGCTCAGTGG-TTAGAGCAGGGGATTCATAAGCCCAAGGTCGCAGGTTCAAATCCCGCTTGAGCCA      ycf33*-argB  -Mcat-Aggc -Sgct
Osi ( 65380.. 65452) -   GGCTCTGTAGCTCAGTGG-TTAGAGCAGGGGACTCATAAGCCCAAGGTCGTAGGTTCAAATCCCACCAGAGCCA      orf   -ycf33*-Mcat-Sgct -Dgtc
Gth (119345..119417) -   GGCTTAGTAGCTCAGTGG-TTAGAGCAGGGGACTCATAAGCCCAAGGTCGCAGGTTCAAATCCCGCCTAAGCCA      ilvB* -ycf33*-Mcat-Sgct -Dgtc
Mvi ( 25330.. 25401) +   GGCTTTGTAGCTCAGCGG--TAGAGCAGGGGATTCATAAGCCCAAGGTCGCAGGTTCAAATCCCGCCAGAGCCA      atpB  -atpE  -Mcat-Gtcc*-chlI
Alo ( 60667.. 60738) -   GGTTCAATAGCTCAAAGG--TAGAGCATAGGATTCATAAGCCTCAGGTCACAAGTTCAAATCTTGTTTGAACCA      Ygta  -Hgtg  -Mcat-Wcca -Ettc
Egr (100686..100757) -   GGCTCAGTAGCTCAGAGG--TAGAGCAGGGGATTCATAAGCCCTTGGTCACAGGTTCAAATCTTGTCTGAGCCA      Ygta  -Hgtg  -Mcat-Wcca -Ettc
                         *******..++++.........++++.----.........----.....=====.......=====*******.
Characteristic Feature                                  ACTCATAAG
                                                        ATTCATAAG

                         *******..++++..... ...++++.-----.......-----.....=====.......=====*******.
Nol (  1618..  1691) +   GCCTGCTTAGCTCAGTTGGTTAGAGCGTCCGTTTCATACGCGGATTGTCACTAGTTCAAATCTAGTAGCAGGCA      Ltag  -psbM* -Mcat-ftsI -psbA
Cvu (129350..129426) +   GCCTGCTTAGCTCAGTTGGTTAGAGCATCCGTCTCATACGCGGAATGTCACTAGTTCGAATCTAGTAGCAGGCACCA   orf   -rps14 -Mcat-Ggcc -orf*
Cgl ( 47720.. 47792) +   GCCTACTTAACTCAGCGG-TTAGAGTGTCGCTTTCATACGGCGAAGGTCATTGGTTCAAATCCAATAGTAGGTA      ndhC* -Sgga  -Mcat-atpE*-atpB*
Mpo ( 53801.. 53874) +   ACCTACTTAACTCAGTGGTTTAGAGTATCGCTTTCATACGGCGAGAGTCATTGGTTCAAATCCAATAGTAGGTA      ndhC* -Vtac* -Mcat-atpE*-atpB*
Afo ( 69683.. 69755) +   ACCTACTTAACTTAGTGG-TTAGAGTATCGCTTTCATACGGCGAGAGTCATTGGTTCAAATCCAATAGTAGGTA      ndhC* -Vtac* -Mcat-atpE*-atpB*
Pnu ( 53151.. 53223) +   ACCTACTTAACTCAGTGG-TTAGAGTATCGCTTTCATAAGGCGAGAGTCATTGGTTCAAATCCAATAGTAGGTA      ndhC* -Vtac* -Mcat-atpE*-atpB*
Aca ( 49115.. 49187) +   GCCTACTTAACTCAGCGG-TGAGAGTATCGCTTTCATACGGCGAGAGTCATTGGTTCGAATCCAATAGTAGGTA      ndhC* -Vtac  -Mcat-atpE*-atpB*
Pth ( 47156.. 47228) -   ACCCACTTAACTCAGTGG-TTAGAGTATCGCTTTCATACGGCGAGAGTCATTGGTTCAAATCCAATAGTAGGTA      orf * -Vtac* -Mcat-atpE*-atpB*
Pko ( 46792.. 46864) -   ACCCACTTAACTCAGTGG-TTAGAGTATCGCTTTCATACGGCGAGAGTCATTGGTTCAAATCCAATAGTAGGTA      orf   -Vtac* -Mcat-atpE*-orf
Cfe ( 53351.. 53423) +   ACCTACTTAACTCAGTGG-TTAGAGTATTGCTTTCATACGGCGGGAGtCATTGGTTCAAATCCAATAGTAGGTA      ndhC* -Vtac* -Mcat-atpE*-atpB*
Lja (  9408..  9480) -   ACCTACTTAACTCAGCGG-TTAGAGTATCGCTTTCATACGGCGGGAGTCATTGGTTCAAATCCAATAGTAGGTA      ndhC* -Vtac* -Mcat-atpE*-atpB*
Nta ( 54589.. 54661) +   ACCTACTTAACTCAGTGG-TTAGAGTACTGCTTTCATACGGCGGGAGTCATTGGTTCAAATCCAATAGTAGGTA      ndhC* -Vtac* -Mcat-atpE*-atpB*
Oel ( 11573.. 11645) -   ACCTACTTAACTCAGTGG-TTAGAGTATTGCTTTCATACGGCAGGAGTCATTGGTTCAAATCCAATAGTAGGTA      ndhC* -Vtac* -Mcat-atpE*-atpB*
Ath ( 52056.. 52128) +   ACCTACTTAACTCAGTGG-TTAGAGTATTGCTTTCATACGGCAGGAGTCATTGGTTCAAATCCAATAGTAGGTA      ndhC* -Vtac* -Mcat-atpE*-atpB*
Sol ( 50859.. 50931) +   ACCTACTTAACTCAGCGG-TTAGAGTATTGCTTTCATACGGCGGGAGTCATTGGTTCAAATCCAATAGTAGGTA      ndhC* -Vtac* -Mcat-atpE*-atpB*
Evi (  7392..  7464) +   ACCTATTTAACTCAGTGG-TTAGAATATTGCTTTCATACGGCAGAAGTCATTGGTTCAAATCCAATAGTAGGTA      rps4* -Fgaa  -Mcat-atpB -rbcL
Osa ( 51219.. 51291) +   GCCTACTTAACTCAGTGG-TTAGAGTATTGCTTTCATACGGCGGGAGTCATTGGTTCAAATCCAATAGTAGGTA      ndhC* -Vtac* -Mcat-atpE*-atpB*
Tae ( 52034.. 52106) +   GCCTACTTAACTCAGTGG-TTAGAGTATTGCTTTCATACGGCGGGAGTCATTGGTTCAAATCCAATAGTAGGTA      ndhC* -Vtac* -Mcat-atpE*-atpB*
Zma ( 54020.. 54092) +   GCCTACTTAACTCAGTGG-TTAGAGTATTGCTTTCATACGGCGGGAGTCATTGGTTCAAATCCAATAGTAGGTA      ndhC* -Vtac* -Mcat-atpE*-atpB*
                         *******..++++..... ...++++.-----.......-----.....=====.......=====*******.
Characteristic Feature                                  tTTCATACg
```

Figure 3. Comparison of elongator tRNA (Mcat) gene sequences and gene orders.


front - Netscape
ファイル(F) 編集(E) 表示(V) ジャンプ(G) ブックマーク(B) ツール(T) ウィンドウ(W) ヘルプ(H)
http://www.rs.noda.tus.ac.jp/~kunisawa/order/front.html
検索
Mail ホーム ラジオ Netscape.co.jp 検索 Shop@AOL ブックマーク
front
ドキュメント完了(0.141 秒)
スタート
2 Windo...
front - N...
DSJ03
Microsoft...
14:56

1. View Gene Order of Completely Sequenced Plastid Genomes

Glaucocystophyceae	*Cyanophora paradoxa*	Cpa	Streptophyta	*Mesostigma viride*	Mvi
Rhodophyta	*Cyanidium caldarium*	Cca		*Chaetosphaeridium globosum*	Cgl
	Cyanidioschyzon merolae	Cme		*Marchantia polymorpha*	Mpo
	Porphyra purpurea	Ppu		*Anthoceros formosae*	Afo
Bacillariophyta	*Odontella sinensis*	Osi		*Psilotum nudum*	Pnu
Cryptophyta	*Guillardia theta*	Gth		*Adiantum capillus-veneris*	Aca
Euglenozoa	*Astasia longa*	Alo		*Pinus thunbergii*	Pth
	Euglena gracilis	Egr		*Pinus koraiensis*	Pko
Chlorophyta	*Chlorella vulgaris*	Cvu		*Calycanthus fertilis*	Cfe
	Nephroselmis olivacea	Nol		*Amborella trichopoda*	Atr
	Chlamydomonas reinhardtii	Cre		*Lotus japonicus*	Lja
Apicomplexa	*Eimeria tenella*	Ete		*Nicotiana tabacum*	Nta
	Toxoplasma gondii	Tgo		*Oenothera elata*	Oel
				Arabidopsis thaliana	Ath
				Spinacia oleracea	Sol
				Epifagus virginiana	Evi
				Oryza sativa	Osa
				Triticum aestivum	Tae
				Zea mays	Zma

2. Compare Gene Orders Around a Given Gene

Enter Reset

Figure 4. The front page.

corresponding two gene-arrangements are shown. Genes, which are marked "*", are transcribed in the right to left direction, while others are transcribed in the reverse direction. Using this function, a user can enjoy comparing gene orders between different genomes.

Figure 5. The database format. Several examples are illustrated. The top two lines indicate the column positions.

```
         1         2         3         4         5         6
123456789012345678901234567890123456789012345678901234567890123

Odontella sinensis Gene Order
(GenBank Accession No. Z67753)

   267    340 +   t Ptgg                               Sequence
   478   1545 +   p orf                  ORF355        Sequence
  2210   3694 +   r 16s                                Sequence
   ...
  9852   9923 +   t Ngtt                 9852 9922 +   Sequence
  9944  10729 +   p thiG                 thiG          Sequence
   ...
 95722  95794 + % t Rccg                               Sequence
   ...
```

6 CONCLUSIONS

We have developed a gene order database for 32 completely sequenced plastid genomes. We developed a normalizing gene-labeling system across complete genomes, by which comparative studies are made available without returning to sequence analysis. A lot of incorrect tRNA gene coordinates detected in the major databases were corrected. Incomplete annotations of tRNA genes with the anticodon sequence CAT in the major databases were improved, and their classification into initiator tRNA, elongator tRNA and Ile-tRNA genes were specified where possible. The gene order database developed here is available at http://www.rs.noda.tus.ac.jp/~kunisawa. Using this database, we are now resolving the phylogenetic relationships of plastid genomes along the lines suggested elsewhere (Kunisawa, Blanchette & Sankoff, 1997; Kunisawa, 2003). At the same time we are extending the present gene order database so that an evolutionary comparison between plastids and cyanobacteria and between plastids and host nuclear genomes will be possible.

7 REFERENCES

Dandekar, T., Snel, B., Huynen, M., & Bork, P. (1998) Conservation of gene order: a fingerprint of proteins that physically interact. *Trend Biol. Sci. 23*(9), 324-328.

DDBJ (n.d.) Home page of the DNA Data Bank of Japan. Available at: http://www.ddbj.nig.ac.jp.

EMBL (n.d.) Home page of the EMBL Nucleotide Sequence Database. Available at: http://www.ebi.ac.uk /embl/index.html.

GenBank (n.d.) Home page of GenBank. Available at http://www.ncbi.nlm.nih.gov/Genbank/index.html.

Korbel, J.O., Snel, B., Huynen, M.A., & Bork, P. (2002) SHOT: a web server for the construction of genome phylogenies. *Trends Genet. 18*(3), 158-162.

Kunisawa, T., Blanchette, M., & Sankoff, D. (1997) Gene Order Comparison for Phylogenetic Inference of Plastid Genomes. *Res. Commun. Biochem. Cell Mol. Biol. 1*(2), 134-142.

Kunisawa, T. (2003) Gene arrangements and branching orders of gram-positive bacteria. *J. Theor. Biol. 222*(4), 495-503.

Lowe, T.M., & Eddy, S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res. 25*(5), 955-964.

Marck, C., & Grosjean, H., (2002) tRNomics: Analysis of tRNA genes from 50 genomes of Eukarya, Archaea, and Bacteria reveals anticodon-sparing strategies and domain-specific features. *RNA 8*(10), 1189-1232.

Muramatsu, T., Yokoyama, S., Horie, N., Matsuda, A., Ueda, T., Yamaizumi, Z., Kuchino, Y., Nishimura, S., & Miyazawa, T. (1988) A novel lysine-substituted nucleoside in the first position of the anticodon of minor isoleucine tRNA from Escherichia coli. *J. Biol. Chem. 263*(19), 9261–9267.

Pearson, W.R., & Lipman, D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. U.S.A. 85*(8), 2444-2448.

Sankoff, D., Leduc, G., Antoine, N., Paquin, B., Lang, B.F., & Cedergren, R. (1992) Gene order comparisons for phylogenetic inference: evolution of the mitochondrial genome. *Proc. Natl. Acad. Sci. U.S.A. 89*(14), 6575-6579.

Rivas, J.D.L., Lozano, J.J., & Ortiz, A.R. (2002) Comparative analysis of chloroplast genomes: functional annotation, genome-based phylogeny, and deduced evolutionary patterns. *Genome Res. 12*(4), 567-583.

Seoighe, C., Federspiel, N., Jones, T., Hansen, N., Bivolarovic, V., Surzycki, R., Tamse, R., Komp, C., Huizar, L., Davis, R.W., Scherer, S., Tait, E., Shaw, D.J., Harris, D., Murphy, L., Oliver, K., Taylor, K., Rajandream, M.A., Barrell, B.G., & Wolfe, K.H. (2000) Prevalence of small inversions in yeast gene order evolution. *Proc. Natl. Acad. Sci. U.S.A. 97*(26), 14433-14437.

Sugiura, M. (1995) The chloroplast genome. *Essays Biochem. 30*, 49-57.

Tatusov, R.L., Galperin, M.Y., Natale, D.A., & Koonin, E.V. (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res. 28*(1), 33-36.

Tatusov, R.L., Natale, D.A., Garkavtsev, I.V., Tatusova, T.A., Shankavaram, U.T., Rao, B.S., Kiryutin, B., Galperin, M.Y., Fedorova, N.D., & Koonin, E.V. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res. 29*(1), 22-28.

Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res. 22*, 4673-4680.

Appendix 1. Differences between the gene order database (left) and GenBank/EMBL/DDBJ (right) annotations.

```
>Cpa
  5727   7598 +   p dnaK       dnaK-A
  8038   9663 -   p groEL      groEL-A
  9727  10038 -   p groES      groES-A
 10419  11003 +   p clpP       clpPI
 23257  23358 -   p petM       petX
 38535  38915 +   p rpl12      rpl7
 41932  42933 +   p ycf48      orf333
 44000  44599 -   p clpP       clpP2
 46922  47893 -   p crtE       preA
 56389  56697 -   p ycf49      orf102
 57632  57829 -   p psbZ       ycf9
 65817  66308 -   p ycf51      orf163
 66487  66576 +   p petN       ycf6
101705 102271 -   p bioY       orf188
106514 107098 -   p clpP       clpP1
107479 107790 +   p groES      groES-B
107854 109479 +   p groEL      groEL-B
109919 111790 -   p dnaK       dnaK-B
117520 118092 +   p carA       trpG
118714 119682 -   p ccsA       ycf5
124256 124990 -   p cysA       orf244
130030 131004 -   p lysR       ycf30
131084 131200 -   p psbY       ycf32
133050 134432 +   p moeB       chlN

>Cca
 26413  27249 -   p cemA       ycf10
 36169  37032 -   p crtR       desA
 37126  38079 -   p lysR       ycf30
 38214  38327 +   p psbY       ycf32
 42478  42642 +   p nblA       ycf18
 42694  43575 +   p ccsA       ycf5
 51101  51319 +   p rpoZ       ycf61
 59064  59651 -   p carA       trpG
 90499  91686 -   p moeB       chlN
 95562  96191 +   p ycf29      ompR
105097 106002 +   p cfxQ       cfxX
108065 108799 -   p cysA       ycf85
149345 150064 +   p tatC       ycf43
152129 153145 -   p pdhA       odpA
153181 154152 +   p crtE       preA
156102 156197 +   p petN       ycf6
157586 157771 -   p psbZ       ycf9

>Cme
  6763   7071 -   p trxA       trxM
 12050  12703 -   p cysA       ycf85
 14638  15978 -   p thdF       trmE
 28154  29065 -   p lysR       ycf30
 49265  49828 -   p carA       trpG
 59840  60328 +   p ntcA       ycf28
 95041  95988 -   p crtE       preA
 95999  96955 +   p pdhA       odpA
127076 127831 +   p ycf63      ycxr
144872 145702 -   p cemA       ycf10

>Ppu
  1095   2402 +   p moeB       chlN
  2454   3929 -   p ycf46      ORF491
  5020   5544 -   p ycf52      ORF174
```

```
 12526  13242 -   p ycf53      ORF238
 13279  14241 -   p ycf55      ORF320
 14421  14747 +   p ycf54      ORF108
 14758  15711 -   p lysR       ycf30
 15930  16040 +   p psbY       ycf32
 24455  25807 +   p ycf80      ORF450
 27357  27452 -   p petL       ycf7
 31543  32448 +   p cfxQ       ORF301
 34445  35281 -   p cemA       ycf10
 37652  38101 +   p ycf58      ORF149
 38383  38718 -   p ycf41      ORF111
 70041  70805 +   p tatC       ORF254
 73102  74136 -   p pdhA       odpA
 74450  75421 +   p crtE       preA
 76976  77575 -   p ycf42      ORF199
 78397  78612 -   p ycf47      ORF71
 78688  78786 -   p petM       ycf31
 78840  78929 -   p petN       ycf6
 79260  80309 +   p ycf59      ORF349
 83167  84864 -   p ycf45
 85043  85240 +   p ycf86
108332 108520 -   p psbZ       ycf9
114350 116236 +   p ftsH       ycf25
126158 127306 -   p moeB       ORF382
127331 127558 -   p psaC       ORF75
131048 131761 -   p ntcA       ycf28
132611 132826 +   p ycf40      ORF71
142122 142421 -   p ycf65      ORF99
143163 143759 -   p upp        ORF198
153023 153634 -   p ycf60      ORF203
154703 155272 +   p carA       trpG
155281 156594 -   p ycf44      ORF437
156863 157585 -   p dsbD       ORF240
158849 159025 +   p nblA       ycf18
159774 160268 +   p cpcA       cpeA
160407 161366 +   p ccsA       ycf5
169721 169948 +   p rpoZ       ORF75
177513 178496 +   p ycf62      ORF327
179531 179887 -   p ftrC       ftrB

>Osi
   478   1545 +   p OrfBY      ORF355
  7318   7428 +   p psbY       ycf32
  8139   9077 -   p ccsA       ycf5
 45509  46438 -   p lysR       ycf30
 46632  46796 -   p rpl32      rpl32'
 46931  47041 -   p psbY       ycf32
 52814  53881 -   p OrfBY      ORF355
 54453  54695 -   p acpP       acp
 58339  60273 +   p ftsH       ycf25
 78255  78344 +   p petN       ycf6
 78384  78512 +   p petM       ycf31
 82800  82895 +   p petL       ycf7
 86186  86371 -   p psbZ       ycf9
 94783  95082 +   p ycf66      ORF99
 98256  99047 -   p tatC       ycf43

>Gth
 12381  12680 -   p ycf65      orf99
 23432  24364 -   p lysR       ycf30
 24489  24602 +   p psbY       ycf32
 32662  32892 -   p rpoZ       orf75
 33657  34562 -   p ccsA       ycf5
 45154  46002 -   p ycf80      orf282
 55486  55794 -   p ftrC       ftrB
 57914  58009 -   p petL       ycf7
 58663  59499 -   p cemA       ycf10
 84369  85241 +   p tatC       ycf43
 88009  88098 +   p petN       ycf6
 88132  88230 +   p petM       ycf31
 90036  90224 -   p psbZ       ycf9
 96179  96424 +   p acpP       acpA
 96530  96811 +   p hupA       hlp
117067 118962 +   p ftsH       ycf25

>Mvi
 41989  42078 +   p petN       ycf6
 62835  63506 +   p bioY       orf223
 65929  66117 -   p psbZ       ycf9
 66600  69272 +   p ftsH       ycf2
 97970  99310 +   p moeB       chlN
102606 102833 -   p rpoZ       ycf61

>Nol
 43898  43990 +   p petN       ycf6
 65537  65725 -   p psbZ       ycf9
 69697  80925 +   p ftsH       ycf2
 92744  93454 -   p OrfAT      orf236
 99084  99818 -   p OrfAU      orf244
100202 100513 -   p OrfAV      orf103a
100510 100923 -   p OrfAW      orf103b
101052 101516 -   p OrfAX      orf154
102356 104140 -   p OrfAY      orf594
104394 105563 -   p OrfAZ      orf389
105762 106607 -   p OrfBA      orf281
106961 107290 -   p OrfBB      orf109
107401 107622 -   p OrfBC      orf73
107792 108331 -   p OrfBD      orf179
108331 109068 -   p OrfBE      orf245
109711 110043 +   p OrfBF      orf110
110120 110530 +   p OrfBG      orf136
110799 111110 -   p OrfBH      orf103b
111722 112489 -   p OrfBI      orf255
114476 114667 -   p OrfBJ      orf63a
116083 116313 -   p OrfBK      orf76
116661 117002 -   p OrfBL      orf113
117723 118094 -   p OrfBM      orf123
118659 119447 -   p OrfBN      orf262
125851 127146 -   p moeB       chlN
130930 131121 +   p OrfBO      orf63b
134353 134565 +   p OrfBP      orf70
158361 158573 -   p OrfBP      orf70
161805 161996 -   p OrfBO      orf63b
165780 167075 +   p moeB       chlN
173479 174267 +   p OrfBN      orf262
174832 175203 +   p OrfBM      orf123
175924 176265 +   p OrfBL      orf113
176613 176843 +   p OrfBK      orf76
178259 178450 +   p OrfBJ      orf63a
180437 181204 +   p OrfBI      orf255
181816 182127 +   p OrfBH      orf103b
182396 182806 -   p OrfBG      orf136
182883 183215 -   p OrfBF      orf110
183858 184595 +   p OrfBE      orf245
184595 185134 +   p OrfBD      orf179
185304 185525 +   p OrfBC      orf73
```

```
185636 185965 +   p OrfBB      orf109
186319 187164 +   p OrfBA      orf281
187363 188532 +   p OrfAZ      orf389
188786 190570 +   p OrfAY      orf594
191410 191874 +   p OrfAX      orf154
192003 192416 +   p OrfAW      orf137b
192413 192724 +   p OrfAV      orf103a
193108 193842 +   p OrfAU      orf244
199472 200182 +   p OrfAT      orf236

>Cvu
 24561  24785 +   p OrfAQ      ORF74
 45442  45591 +   p OrfAQ      ORF49b
 50297  50485 -   p psbZ       ycf9
 61443  62678 -   p accD       AccD
 64391  64516 +   p OrfAQ      ORF41b
 74243  74368 -   p OrfAR      ORF41c
 88702  90216 -   p ycf62      ORF504
 92172  92300 -   p OrfAQ      ORF42c
 93076  93264 +   p OrfAS      ORF62
 99596  99745 -   p OrfAR      ORF49c
102214 103521 -   p moeB       chlN
106516 107463 +   p ccsA       ycf5
110072 112531 +   p ycf1       ORF819
130569 130736 -   p OrfAS      ORF55c
135963 136121 -   p OrfAR      ORF52d
144485 144649 +   p OrfAQ      ORF54d

>Alo
 11316  11804 +   p OrfCM      ORF162
 14857  15720 +   p OrfAP      ORF287
 30165  30956 +   p OrfAP      ORF263
 32305  33066 +   p OrfAP      ORF253
 33196  33831 +   p OrfAP      ORF211
 33982  34485 +   p OrfAP      ORF167
 39421  41097 -   p OrfCM      ORF558
 61721  62035 +   p OrfBQ      ORF104
 62152  63432 +   p OrfBQ      ORF426

>Egr
 18724  20100 +   p ycf13      mat1
 72138  72335 -   p psbZ       ycf9

>Cgl
 89877  90179 -   p OrfCL      orf100
 99526 100902 +   p moeB       chlN
118964 120340 -   p moeB       chlN
129687 129989 +   p OrfCL      orf100

>Mpo
  4001   4105 +   p psbM       ORF34
 22162  22263 -   p ycf12      ORF33
 22516  22614 +   p psaM       ORF32
 22997  23107 -   p psbI       ORF36a
 23438  23605 -   p psbK       ORF55
 24053  25594 +   p chlB       ORF513
 26976  28088 +   p matK       ORF370i
 29909  36319 +   p ftsH       ORF2136
 37012  38124 +   p cysA       mbpX
 41647  41835 +   p psbZ       ORF62
 51233  51742 -   p ndhJ       ORF169
 51793  52524 -   p ndhK       psbG
 52515  52877 -   p ndhC       ndh3
 58065  59015 +   p accD       ORF316
 59525  60079 +   p ycf4       ORF184
 60151  61455 +   p cemA         ORF434
 62794  62916 -   p psbJ         ORF40
 63036  63152 -   p psbL         ORF38
 64152  64247 +   p petL         ORF31
 64370  64483 +   p petG         ORF37
 65027  65155 +   p psaJ         ORF42b
 70669  70776 +   p psbT         ORF35
 70863  70994 -   p psbN         ORF43
 71092  71316 +   p psbH         ORF74
 75300  75413 -   p rpl36        secX
 91101  93179 -   p ndhF         ndh5
 93886  94095 +   p rpl32        ORF69
 94183  95049 +   p cysT         ORF288
 95482  96444 +   p ccsA         ORF320
 96665  98164 -   p ndhD         ndh4
 98289  98534 -   p psaC         frxA
 98757  99059 -   p ndhE         ndh4L
 99113  99688 -   p ndhG         ORF191
 99779 100330 -   p ndhI         frxB
102202 103380 -   p ndhH         ORF392
103873 105267 -   p ycf1         ORF464
105329 108535 -   p ycf1         ORF1068
110104 110973 -   p chlL         frxC

>Afo
 33690  40868 +   p ftsH         ycf2
 77577  79103 +   p cemA         ycf10
126778 127644 +   p cysT         ORF288
137666 139087 -   p ycf1         ORF473
139569 142664 -   p ycf1         ORF1031
142902 144317 -   p moeB         chlN

>Pnu
 60532  61881 +   p cemA         ycf10
 84908  85159 +   p OrfBZ        orf83
 88517  88867 -   p OrfCA        orf116
 98449  98718 -   p OrfCB        orf89
 98787  99053 -   p OrfCC        orf88
 98809  99165 +   p OrfCD        orf119
 99050  99289 -   p OrfCE        orf79
 99132  99347 +   p OrfCF        orf71
124100 124315 -   p OrfCF        orf71
124158 124397 +   p OrfCE        orf79
124282 124638 -   p OrfCD        orf119
124394 124660 +   p OrfCC        orf88
124729 124998 +   p OrfCB        orf89
134580 134930 +   p OrfCA        orf116
138288 138539 -   p OrfBZ        orf83

>Aca
 56428  57822 +   p cemA         ycf10
 98198 104512 +   p ftsH         ycf2
124805 126181 -   p moeB         chlN
128339 134653 -   p ftsH         ycf2

>Pth
  7983   8129 +   p OrfCG        ORF48a
  8594   8695 +   p ycf12        ORF33
 26778  26867 +   p petN         ORF29
 27451  27672 +   p OrfCQ        ORF73a
 28111  28374 -   p OrfCR        ORF87
 30226  30366 -   p OrfCT        ORF46b
 30742  30867 +   p OrfCU        rps12
 30988  31119 +   p OrfCV        rps12
 31594  31803 +   p OrfCW        rps12
```

```
 33851  34039 -   p petL         ORF62b
 38271  39056 -   p cemA         ORF261
 39194  39364 -   p OrfCX        ORF56a
 39724  40278 -   p ycf4         ORF184
 48310  48480 +   p OrfCY        rps12
 48477  48677 -   p OrfCZ        ORF66
 50267  50431 -   p OrfCH        ORF54a
 50602  50739 +   p OrfDF        rps12
 51051  51128 -   p ycf12        ORF25
 51599  51745 -   p OrfCG        ORF48b
 64251  64442 +   p ycf72        rps12
 66046  66180 +   p OrfAN        rps12
 71552  71875 -   p OrfAB        ORF107
 71742  71954 +   p OrfAA        rps12
 79389  79577 -   p psbZ         ORF62
 83970  84164 -   p OrfCH        ORF64b
 86346  86573 +   p ycf68        ORF75a
 86897  87019 -   p OrfDA        ORF40e
 92296  92511 -   p OrfDB        ORF71
 93946  95349 +   p moeB         chlN
 95542 100812 +   p ycf1         rps12
104925 105887 -   p ccsA         ORF320
108685 108906 -   p OrfDC        ORF73b
112617 118781 -   p ftsH         ORF2054
119224 119358 -   p OrfAN        ORF44b

>Pko
 12583  12765 -   p atpF         ORF60a
 27076  27231 +   p OrfCQ        ORF51a
 27769  28086 -   p OrfCR        ORF105
 28783  28977 -   p OrfCS        ORF64a
 30271  30420 -   p OrfCT        ORF49b
 30796  30921 +   p OrfCU        ORF41a
 31042  31173 +   p OrfCV        ORF43b
 31615  31824 +   p OrfCW        ORF69a
 39163  39306 -   p OrfCX        ORF47c
 47944  48147 +   p OrfCY        ORF67b
 48115  48315 -   p OrfCZ        ORF66
 48970  49110 -   p OrfBX        ORF46b
 49570  49767 -   p OrfCG        ORF62a
 49579  49809 +   p OrfDF        ORF76b
 53626  54327 +   p petB         ORF233
 64178  64504 +   p OrfAN        ORF107
 64970  65218 +   p ndhK         ORF82
 69170  69304 +   p ycf3         ORF44e
 69635  69847 +   p OrfAA        rps12
 82100  82297 -   p OrfCH        ORF65
 84263  84490 +   p ycf68        ORF75
 84734  84940 -   p OrfDA        ORF68b
 90107  90418 -   p OrfDB        ORF103
 91835  93247 +   p moeB         chlN
 93548  93829 +   p ycf1         ORF93
 95141  96868 +   p ycf1         ORF575
105236 105505 -   p OrfDC        ORF89b
112157 113782 -   p ftsH         ORF541
115043 116008 -   p ftsH         ORF321

>Cfe
 29424  29513 +   p petN         ycf6
 36975  37163 +   p psbZ         ycf9
 50242  50988 -   p ndhK         psbG
 87550  94413 +   p ftsH         ycf2
 94535  94768 +   p ycf15        ycf2
 99347  99574 -   p OrfDD        rps12
115263 116234 +   p ccsA         ycf5
```

```
140708 140935 +   p OrfDD      rps12
145869 152732 -   p ftsH       ycf2

>Lja
 11104  11796 +   p ndhK       psbG
 25158  25346 -   p psbZ       ycf9
 32379  32468 -   p petN       ycf6
 59718  60407 +   p cemA       ycf10
 63079  63174 +   p petL       ycf7
 84420  91316 +   p ftsH       ycf2
110740 111711 +   p ccsA       ycf5
141140 148036 -   p ftsH       ycf2

>Nta
  7835   8020 +   p psbK       ORF98
 29535  29624 +   p petN       ycf6
 37594  37782 +   p psbZ       ycf9
 46248  46472 -   p OrfAC      rps12
 62638  63192 +   p ycf4       ORF184
 63415  64104 +   p cemA       ORF229
 88885  95727 +   p ftsH       ycf2
 96060  96407 -   p OrfAD      ORF115
 96119  96397 +   p OrfAE      ORF92
 96556  96795 +   p OrfAF      ORF79
101951 102346 -   p OrfAG      rps12
102102 102314 +   p OrfAH      ORF70B
110597 110824 -   p OrfAI      rps12
111029 112081 +   p ycf1       ORF350
115061 115228 +   p rpl32      rp132
116344 117285 +   p ccsA       ycf5
131802 132029 +   p OrfAI      ORF75
140280 140675 +   p OrfAG      ORF131
140312 140524 -   p OrfAH      rps12
145831 146070 -   p OrfAF      rps12
146219 146566 +   p OrfAD      ORF115
146229 146507 -   p OrfAE      rps12
146545 146808 -   p ycf15      ycf15’
146899 153741 -   p ftsH       ycf2’

>Oel
 29173  29361 -   p psbZ       ycf9
 66473  67162 +   p cemA       ycf10
 91742  98584 +   p ftsH       ycf2
 99214  99429 -   p OrfAD      rps12
 99767  99937 +   p OrfAF      ORF56
104576 104869 -   p OrfBR      rps12
104953 105264 +   p OrfBS      ORF103
106252 106431 +   p OrfBT      rrn16
107541 107789 -   p OrfBU      rps12
107627 107890 -   p OrfBU      rps12
108551 108727 +   p OrfBV      ORF58
108729 109088 +   p ycf68      ORF119
110571 110735 -   p OrfBW      rps12
118885 119844 +   p ccsA       ycf5
142593 142757 +   p OrfBW      ORF54
144240 144599 -   p ycf68      ORF119
144601 144777 -   p OrfBV      ORF58
145438 145701 +   p OrfBU      ORF87
145539 145787 +   p OrfBU      ORF82b
146897 147076 -   p OrfBT      rrn16
148064 148375 -   p OrfBS      rps12
148459 148752 +   p OrfBR      ORF97
153391 153561 -   p OrfAF      ORF56
153899 154114 +   p OrfAD      ORF71
154744 161586 -   p ftsH       ycf2
```

```
>Ath
 28089  28178 +   p petN       ycf6
 35751  35939 +   p psbZ       ycf9
 49257  49934 -   p ndhK       psbG
 60741  61430 +   p cemA       ycf10/cemA
 65712  65807 +   p petL       ORF31
 86474  93358 +   p ftsH       ycf2
 93495  93728 +   p ycf15      orf77
114461 115447 +   p ccsA       ycf5
123884 129244 -   p ycf1       rps7
140704 141171 +   p rps7       orf77
144921 145154 -   p ycf15      rpl23
145291 152175 -   p ftsH       rpl23

>Sol
  1783   3300 -   p matK       maturase
 27285  27374 +   p petN       ycf6
 29409  29660 -   p OrfCS      rps12
 34644  34832 +   p psbZ       ycf9
 42595  42663 -   p OrfAA      ycf3
 42608  42703 +   p OrfAB      ORF31
 43188  43271 -   p OrfAC      rps12
 91797  91970 +   p OrfAF      ORF57
 96758  96949 -   p OrfCI      ORF63
 97056  97199 +   p OrfAH      ORF47
 97339  97503 -   p OrfAG      ORF54
100243 100461 +   p OrfCJ      ORF72
106348 107793 +   p ycf1       ORF482
112317 113288 +   p ccsA       ycf5
132984 133202 -   p OrfCJ      ORF72
135942 136106 +   p OrfAG      ORF54
136246 136389 -   p OrfAH      ORF47
136496 136687 +   p OrfCI      ORF63
141475 141648 -   p OrfAF      rps12
142690 149085 -   p ftsH       ycf2

>Evi
 22045  28695 +   p ftsH       ORF2216
 42887  48103 -   p ycf1       ORF1738
 61133  67783 -   p ftsH       ORF2216

>Osa
  1668   3296 -   p matK       ORF542
 11937  12125 +   p psbZ       ORF62
 14077  14346 -   p ycf70      ORF91
 17556  17645 -   p petN       ORF29
 47992  48471 -   p ndhJ       ORF159
 48569  49309 -   p ndhK       psbG
 56553  56873 +   p accD       ORF106
 57222  57332 +   p psaI       ORF36
 57702  58259 +   p ycf4       ORF185
 58677  59369 +   p cemA       ORF230
 61565  61687 -   p psbJ       ORF40
 63531  63626 +   p petL       ORF31
 63799  63912 +   p petG       petE
 64622  64756 +   p psaJ       ORF44
 67638  68288 -   p clpP       rps12
 70490  70597 +   p psbT       ORF35
 80915  81163 -   p OrfBX      rps12
 81286  81699 +   p ycf72      ORF137
 83534  83620 +   p ftsH       ORF28
 83997  84746 +   p OrfAJ      ORF249
 90227  90442 -   p OrfAG      ORF72
 90501  90659 -   p OrfAK      ORF85
 93241  93642 +   p ycf68      ORF133
 99016  99087 +   p OrfAL      ORF23
100206 100397 +   p OrfAM      ORF63
101229 101399 +   p ndhH       ORF56
104352 104543 +   p rpl32      ORF63
105236 106201 +   p ccsA       ORF321
110000 110536 -   p ndhI       ORF178
114721 114912 -   p OrfAM      ORF63
116031 116102 -   p OrfAL      ORF23
121476 121877 -   p ycf68      ORF133
124360 124617 +   p OrfAK      ORF85
124676 124891 +   p OrfAG      ORF72
130372 131121 -   p OrfAJ      ORF249
131498 131584 -   p ftsH       ORF28
133419 133832 -   p ycf72      ORF137
133955 134203 +   p OrfBX      ORF82

>Tae
 12018  12206 +   p psbZ       ycf9
 17643  17732 -   p petN       ycf6
 65073  65273 +   p rpl33      psl33
105310 106278 +   p ccsA       ycf5
126718 127188 +   p rps7       rps 7

>Zma
  1674   3308 -   p matK       matk
 12017  12205 +   p psbZ       ORF62
 14498  14707 -   p ycf70      ORF69
 19081  19170 -   p petN       ORF29
 59666  60223 +   p ycf4       ORF185
 65352  65447 +   p petL       ORF31
 65611  65724 +   p petG       petE
 72401  72502 +   p psbT       rps12
 82689  82913 -   p OrfBX      ORF75
 83036  83449 +   p ycf72      rps12
 85019  85159 +   p OrfAO      rps12
 85141  85245 +   p ftsH       rps12
 85276  86001 +   p ftsH       rps12
 86288  86707 +   p ftsH       rps12
 87515  87814 +   p ycf15      rps12
 87875  88396 +   p OrfAJ      rps12
 94364  94621 -   p OrfAK      ORF85
 97093  97497 +   p ycf68      ORF133
 98712  98861 +   p OrfCK      ORF49
102866 102937 +   p OrfAL      rps12
104074 104265 +   p OrfAM      rps12
108995 109960 +   p ccsA       rps12
118472 118663 -   p OrfAM      ORF63
119800 119871 -   p OrfAL      ORF23
123876 124025 -   p OrfCK      ORF49
125240 125644 -   p ycf68      ORF133
128116 128373 +   p OrfAK      rps12
128423 128599 +   p OrfAG      rps12
134341 134862 -   p OrfAJ      ORF173
134923 135222 -   p ycf15      ORF99
136030 136449 -   p ftsH       ORF139
136736 137461 -   p ftsH       ORF241
137492 137596 -   p ftsH       ORF34
137578 137718 -   p OrfAO      ORF46
139288 139701 -   p ycf72      ORF137
139824 140048 +   p OrfBX      ORF75

>Tgo
 15007  15138 +   p OrfCO      ORF B
 16035  16352 +   p OrfCN      ORF E
 16395  18692 +   p clpC       clp
 18806  19015 +   p OrfCP      ORF C

>Ete
 14656  14793 +   p OrfCO      ORF-B
 15734  16060 +   p OrfCN      ORF-E
 16108  18333 +   p clpC       CLP
 18466  18732 +   p OrfCP      ORF-C
 27990  29426 -   p ycf24      ORF-G
```

Appendix 2. Differences between the gene order database (left) and GenBank/EMBL/DDBJ (right) annotations.

```
>Aca
  6164   6235 -   t Qttg     trnG, tRNA-Gln
 62123  62196 -   t Ptgg     62124  62197 -
 97858  97934 +   t Hgtg     97859  97932 +
105171 105242 -   t Ngtt     105171 105232 -
127609 127680 +   t Ngtt     127619 127681 +
134917 134993 -   t Hgtg     134919 134992 -

>Afo
 33134  33210 +   t Hgtg     33134  33208 +

>Alo
 60748  60822 -   t Hgtg     60749  60822 -

>Ath
     2     76 -   t Hgtg     4  76 -
 35312  35404 +   t Stga     35312  35403 +

>Cca
 96237  96312 -   t Hgtg     96238  96311 -
109401 109484 -   t Icat     Mcat
156623 156696 +   t fM       Mcat

>Cfe
     2     78 -   t Hgtg     1  78 -
  7251   7322 -   t Qttg     7250  7322 -

>Cgl
 43670  43745 +   t Hgtg     43670  43744 +

>Cme
 19230  19302 +   t Atgc     19231  19302 +
 59178  59253 +   t Hgtg     59178  59251 +
 70621  70692 +   t Tggt     70621  70691 +
126900 126972 -   t Rtct     126901 126972 -

>Cpa
 85048  85132 +   t Icat     Mcat

>Cvu
130470 130556 -   t Sgga     +

>Egr
 30968  31041 +   t fM       Mcat
 60996  61067 +   t Cgca     60996  61056 +
133369 133484 -   r 5s       pseudogene

>Ete
  2379   2453 +   t fM?      Mcat
  5909   5983 +   t Hgtg     5910    5983 +
  6231   6307 +   t Icat?    Mcat
 32269  32343 -   t fM?      Mcat

>Evi
 70022     69 -   t Hgtg     70023  69 -

>Gth
  9223   9296 -   t Hgtg     9223    9295 -
 50077  50162 -   t Icat     Mcat
 89030  89103 +   t fM       Mcat
```

```
>Lja
    46    122 -   t Hgtg     48    121 -

>Mpo
 29595  29671 +   t Hgtg     29595  29669 +
 42156  42229 -   t fM       Met

>Mvi
 58526  58599 -   t Hgtg     58527  58599 -
 71518  71591 +   t Ptgg     71508  71591 +

>Nol
 26457  26530 -   t Hgtg     26458  26529 -
 85192  85273 -   t Ygta     85192  85274 -

>Nta
     4     80 -   t Hgtg     6     80 -
 47119  47205 +   t Sgga     47119  47197 +

>Oel
     7     82 -   t Hgtg     8     82 -
 28414  28487 +   t fM       Met
 28505  28578 +   t fM       Met
 38059  38129 -   t Cgca     38049  38129 -
 72024  72097 -   t Wcca     72025  72097 -
138601 138672 +   t Ngtt     138601 138682 +
139542 139662 -   r 5s       139543 139662 -

>Osa
  1373   1407 -   t Kttt_1   1363   1397 -
  3895   3931 -   t Kttt_2   3902   3938 -
  6616   6687 -   t Qttg     6615   6687 -
 13003  13050 -   t Gtcc_1   13010  13050 -
 64229  64302 -   t Ptgg     64229  64303 -
 81050  81126 +   t Hgtg     81050  81124 +
133992 134068 -   t Hgtg     133991 134068 -


>Osi
  9852   9923 +   t Ngtt     9852   9922 +
 34019  34090 +   t Qttg     34020  34089 +
 95722  95794 + % t Rccg     not listed
 96216  96300 +   t Icat     96190  96274 +

>Pko
  1341   1375 -   t Kttt_2   1376   3863 -
  3864   3900 -   t Kttt_1   3864   3898 -
  8639   8658 +   t Gtcc_1   8739   8761 +
  9420   9470 +   t Gtcc_2   9423   9469 +
 28496  28569 -   t Dgtc     28497  28569 -
 29427  29553 + # t Gtcc     5' fragment
 46792  46864 -   t Mcat     46792  46918 -
 49033  49109 +   t Hgtg     49033  49107 +
 68475  68561 -   t Sgga     68475  68561 +
 76959  77029 -   t Ggcc     76960  77209 -
 77775  77863 +   t Stga     77775  77862 +
102643 102716 +   t Pggg     102643 102716 -
116167 116243 -   t Hgtg     116169 116243 -
```

```
>Pnu
  7055   7131 -   t Hgtg     7057   7131 +
138107 138187 +   t Lcaa     138107 138188 +

>Ppu
  8465   8546 -   t Lgag     8465   8536 -
 26578  26648 -   t Gtcc     26585  26648 -
 32547  32617 +   t Cgca     32547  32618 +
 34285  34371 -   t Icat     Mcat
116603 116674 -   t Aggc     116604 116674 -
132450 132524 -   t Hgtg     132451 132523 -
138880 138953 -   t Racg     138890 138953 -

>Pth
 29392  29518 + # t Gtcc     5' fragment
 47156  47228 -   t Mcat     47156  47288 -
 70440  70526 -   t Sgga     70440  70526 +
118990 119066 -   t Hgtg     118992 119066 -

>Sol
    75 150725 -   t Hgtg     1      74 -
  8887   8934 +   t Gtcc_2   8887   8944 +
  9042   9113 +   t Rtct     9052   9113 +
 29731  29803 -   t Ettc     29741  29803 -
 30262  30333 +   t Tggt     30262  30343 +
 35159  35229 +   t Ggcc     35159  35228 +
 44024  44110 +   t Sgga     44023  44110 +
149174 149247 +   t Icat     pseudogene

>Tae
  1385   1419 -   t Kttt_1   1385   1409 -
  3907   3943 -   t Kttt_2   3814   3943 -
  6687   6758 -   t Qttg     6686   6803 -
 64063  64136 -   t Ptgg     64063  64137 -
 80812  80888 +   t Hgtg     80812  80886 +
116344 116417 -   t Racg     116344 116419 -
134006 134082 -   t Hgtg     134005 134082 -

>Tgo
  2380   2453 +   t fM?      Mcat
  5935   6008 +   t Hgtg     5936   6008 +
  6386   6461 +   t Icat?    Mcat
 32544  32617 -   t fM?      Mcat

>Zma
 82800  82876 +   t Hgtg     82800  82874 +
139861 139937 +   t Hgtg    139863 139937 -
```