Proc. Natl. Acad. Sci. USA
Vol. 80, pp. 4050–4054, July 1983
Genetics

Molecular cloning and sequence analysis of the cyanobacterial gene for the large subunit of ribulose-1,5-bisphosphate carboxylase/oxygenase

(*Anacystis nidulans*/amino acid sequence/phylogenetic tree)

KAZUO SHINOZAKI*, CHIEKO YAMADA*, NAOYUKI TAKAHATA†, AND MASAHIRO SUGIURA*

*Department of Biology, Nagoya University, Chikusa, Nagoya 464, Japan; and †National Institute of Genetics, Mishima, Shizuoka-ken 411, Japan

Communicated by Motoo Kimura, March 23, 1983

ABSTRACT **Ribulose-1,5-bisphosphate carboxylase/oxygenase consists of large subunits (LS) and small subunits. In plants, the LS is encoded in chloroplast DNA and the small subunit, in nuclear DNA. In cyanobacteria, both subunits are thought to be encoded in chromosomal DNA because of prokaryotes. The gene for the LS of ribulose-1,5-bisphosphate carboxylase/oxygenase from a cyanobacterium, *Anacystis nidulans* 6301, has been cloned in pBR322 and subjected to sequence analysis. The coding region contains 1,416 base pairs (472 codons). The deduced amino acid sequence of *A. nidulans* LS protein shows 80% homology with sequences of maize, spinach, and tobacco LS proteins; the nucleotide sequence of *A. nidulans* LS gene shows 70% homology with sequences of the plant genes. Between *A. nidulans* LS and the plant LS proteins there is exact sequence homology around the lysine residue to which the activator CO_2 binds and around the two lysine residues to which ribulose 1,5-bisphosphate binds. The amino acid sequence where the LS binds to the small subunit is also highly conserved. From comparison of the LS proteins of *A. nidulans* and the three plants, the rate of amino acid substitution is estimated to be $0.25–0.5 \times 10^{-9}$ per year per site, which is far below the median value of various types of proteins (1.2×10^{-9} for hemoglobin α). The LS protein is thus a conserved protein.**

Cyanobacteria (blue-green algae) are autotrophic prokaryotes which perform oxygenic plant-type photosynthesis. Thus, cyanobacteria are excellent organisms in which to study plant-type photosynthesis on a molecular basis. As in the case of plants and green algae, ribulose-1,5-bisphosphate carboxylase/oxygenase (RuBPCase/Oase, EC 4.1.1.39) is the key enzyme in the Calvin–Benson cycle in most cyanobacteria and is composed of eight identical large subunits (LS) of 53,000 daltons (Da) and eight identical small subunits (SS) of 12,000–14,000 Da (1). The two subunits of RuBPCase/Oase in plants and green algae are encoded separately in chloroplast and nuclear DNA. The LS protein is encoded in chloroplast DNA and is synthesized on chloroplast ribosomes (2). The SS protein is encoded in nuclear DNA and is synthesized on cytoplasmic ribosomes as a 20,000-Da precursor protein which is transported into chloroplasts and then processed to its mature size (3). Because cyanobacteria are prokaryotes and have no chloroplasts, it is of interest to determine whether their LS and SS proteins are encoded in chromosomal DNA, possibly as a single operon, or in separate genetic entities.

We found that DNA fragments containing spinach and tobacco LS genes hybridized to specific restriction fragments of total DNA isolated from several species of cyanobacteria. We have cloned and determined the sequence of the gene for LS of *Anacystis nidulans* 6301. We compared its nucleotide sequence and deduced amino acid sequence with those of plants.

MATERIALS AND METHODS

Probes and Southern Hybridization. The 0.9-MDa *Kpn* I DNA fragment containing spinach LS gene was kindly provided by R. G. Herrmann (Düsseldorf University). The 0.8-MDa *Bam*HI DNA fragment containing a part of tobacco LS gene was prepared as described (4). *A. nidulans* 6301 DNA was prepared as described (5). Southern blotting and DNA·DNA hybridization by using nick-translated LS probes were performed as described (4).

Cloning. *A. nidulans* DNA was digested with *Eco*RI and the digest was fractionated by 0.5% agarose gel electrophoresis. The 5.6-MDa fragments were electroeluted from a gel strip and purified by DEAE-Sephacel chromatography (6). Cloning of the 5.6-MDa *Eco*RI fragments and a total *Pst* I digest of *A. nidulans* DNA was performed as described (7) but with pBR322 instead of pMB9. T4 DNA ligase was prepared as described (8).

DNA Sequence Analysis. *A. nidulans* DNA fragments for sequence analysis were prepared from plasmids pANE18 and pANP1155, which contain the 5.6-MDa *Eco*RI and the 1.5-MDa *Pst* I fragments, respectively. Base-specific chemical cleavages (G, A>C, T+C, C) were performed according to Maxam and Gilbert (9). Limited cleavage products were analyzed by electrophoresis in 12% polyacrylamide gels containing 7 M urea.

RESULTS

Location of the DNA Fragments Containing *A. nidulans* LS Gene. To determine which restriction fragments contained the LS gene, aliquots of *A. nidulans* 6301 DNA digested with several restriction endonucleases were fractionated by agarose gel electrophoresis, blotted to nitrocellulose filter sheets, and hybridized with the ^{32}P-labeled DNA fragment containing spinach or tobacco LS gene. A 5.6-MDa *Eco*RI fragment and a 1.5-MDa *Pst I* fragment were found to hybridize to the LS probe (Fig. 1). A 9.0-MDa *Eco*RI fragment of *Anavena cylindrica* M1 DNA and a 9.8-MDa *Eco*RI fragment of *Anavena variabilis* M3 DNA were also hybridized to the LS probe (data not shown).

Cloning of the LS Gene. *A. nidulans* DNA fragments, 5.6 MDa, were prepared from the *Eco*RI digest and cloned in pBR322. Clones containing the LS gene sequence were screened by hybridization of *Eco*RI digests of the recombinant DNAs with the LS probe. From 12 clones, 2 clones containing the LS gene sequence were isolated. During construction of a physical

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: RuBPCase/Oase, ribulose-1,5-bisphosphate carboxylase/oxygenase; LS, large subunit; SS, small subunit; Da, dalton(s); bp, base pair(s); RuBP, ribulose 1,5-bisphosphate.

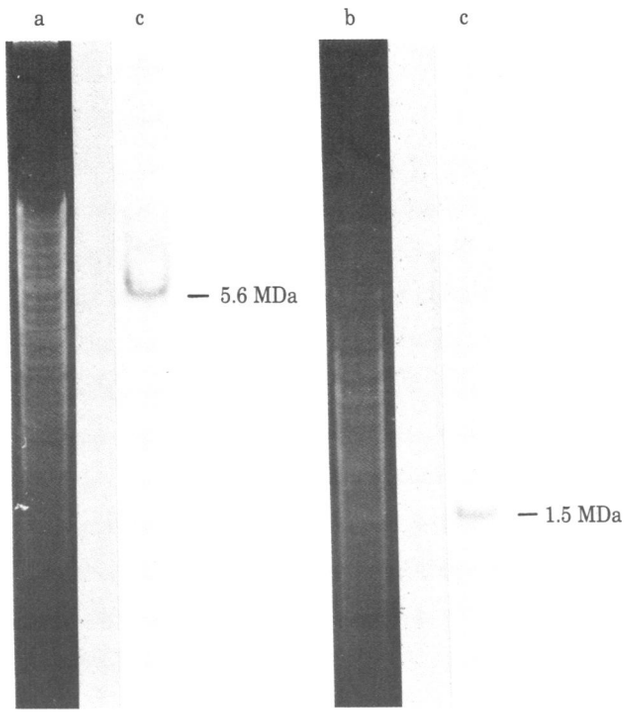

FIG. 1. Location of the DNA fragments containing *A. nidulans* LS gene. *A. nidulans* DNA was digested with *Eco*RI and *Pst* I and electrophoresed in 1% agarose gels. The DNA fragments were transferred to nitrocellulose filter sheets and hybridized with nick-translated tobacco LS DNA. Lanes: a, *Eco*RI digest; b, *Pst* I digest; c, autoradiographs of the filter hybridized with the LS probe.

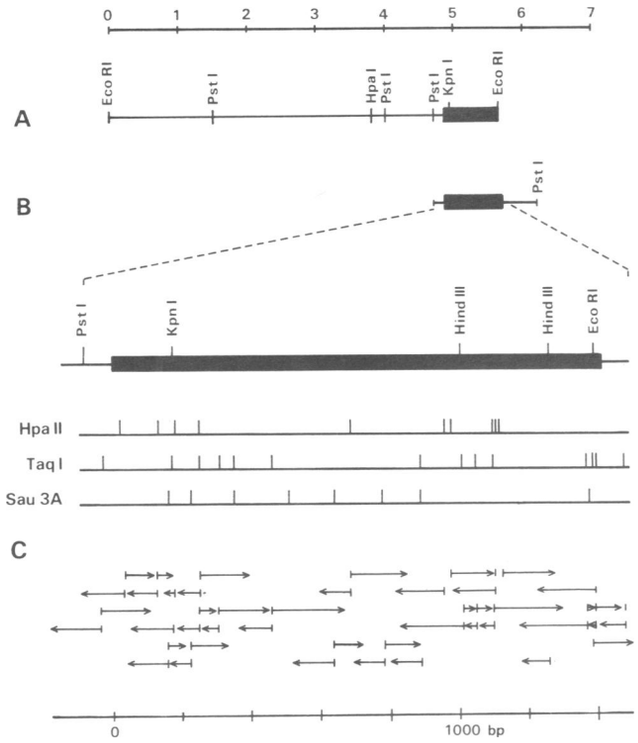

FIG. 2. Physical maps and sequence analysis strategy for *A. nidulans* LS gene. A physical map of the 5.6-MDa *Bam*HI fragment isolated from pANE18 (*A*) and a physical map of the 1.5-MDa *Pst* I fragment from pANP1155 (*B*) are shown. Thickened lines show LS coding regions. (*C*) Sequence analysis strategy in which horizontal arrows indicate the direction and extent of DNA regions analyzed.

map of the cloned 5.6-MDa *Eco*RI fragment (Fig. 2*A*), we noticed that it might not contain the entire LS gene and we attempted to clone another fragment.

A total *Pst* I digest of *A. nidulans* DNA was cloned in pBR322. Clones containing the 1.5-MDa *Pst* I fragments were screened by colony hybridization with a 0.7-MDa *Kpn* I–*Eco*RI subfragment from the cloned 5.6-MDa fragment as a probe (see Fig. 2*A*). From among 1,500 tetracycline-resistant transformants, 4 clones containing the 1.5-MDa *Pst* I fragment were isolated. A physical map of the 1.5-MDa *Pst* I fragment is shown in Fig. 2*B*.

Sequence Analysis of the LS Gene. DNA fragments containing the coding region of *A. nidulans* LS gene and its 5′ and 3′ flanking regions were analyzed according to the strategy shown in Fig. 2*C*. The nucleotide sequence of the noncoding strand (RNA-like strand) is presented in Fig. 3. The coding region of the LS gene contains 1,416 base pairs (bp). The amino acid sequence deduced from the nucleotide sequence is shown in Fig. 4. The *A. nidulans* LS protein contains 472 amino acid residues and has a molecular mass of 52,470 Da. For comparative purposes, the deduced amino acid sequences of LS proteins of tobacco (4), maize (10), and spinach (11) are also shown in Fig. 4.

DISCUSSION

Primary Structure of *A. nidulans* LS. RuBPCase/Oase is one of the most extensively studied plant enzymes (1). It is now established that the enzyme catalyzes two reactions—carboxylation and oxygenation. Its carboxylase activity catalyzes the CO_2 fixation in the Calvin–Benson cycle in which one molecule of RuBP reacts with one molecule of CO_2 to yield two molecules of 3-phosphoglyceric acid. The oxygenase activity converts RuBP to one molecule of 3-phosphoglyceric acid and one molecule of phosphoglycolic acid; this reaction constitutes the first step in the photorespiratory pathway (2).

Both carboxylase and oxygenase enzyme activities reside in LS and require activation by Mg^{2+} and CO_2 before binding of the substrate RuBP (12). The CO_2 that reacts with C-2 of RuBP is different from the CO_2 molecule that activates the enzyme together with Mg^{2+}. The binding site for activator CO_2 has been determined to be the lysine at position 201 of spinach LS (13) (see Fig. 4). There is exact sequence homology around the lysine residue among *A. nidulans* and the three plant LS proteins. The affinity labeling experiments indicated that the two lysine residues at positions 175 and 334 and the cysteine residue at position 459 of spinach LS protein are involved in binding the substrate RuBP (14). Although the sequences around two of three of these amino acids are well conserved among *A. nidulans* and the plant LS proteins, the cysteine residue at position 459 in the plant LS proteins is replaced by leucine (at position 456) in *A. nidulans* LS protein. Jordan and Ogren (15) reported that the ratio of carboxylation to oxygenation catalyzed by RuBPCase/Oase is lower in cyanobacterial enzymes than in plant enzymes. This difference seems to be due to this replacement of the cysteine residue at position 459 by a leucine residue. Modifications of sequences around position 459 in plant LS genes may produce more efficient RuBPCase/Oase.

It is thought that translation of plant LS proteins is initiated at a methionine codon and that peptides from position 1 (Met) to position 14 (Lys) are cleaved off during a post-translational processing step to produce the NH_2-terminal alanine residue (11). Translation of *A. nidulans* LS protein seems also to start at a methionine codon. The sequence G-G-A-G which is complementary to the 3′ end of *A. nidulans* 16S rRNA (unpublished data) occurs 11 to 8 bp upstream from the methionine codon. The amino acid sequence upstream from the alanine residue

FIG. 3. Nucleotide sequence of *A. nidulans* LS coding region (boxed) and its flanking regions. Numbering starts at the ATG codon.

(position 12) of *A. nidulans* LS is different from the corresponding sequences of the plant LS proteins; the amino acid sequences downstream from the alanine residues resemble each other well. This suggests that *A. nidulans* LS precursor polypeptide is also processed between the lysine and alanine residues.

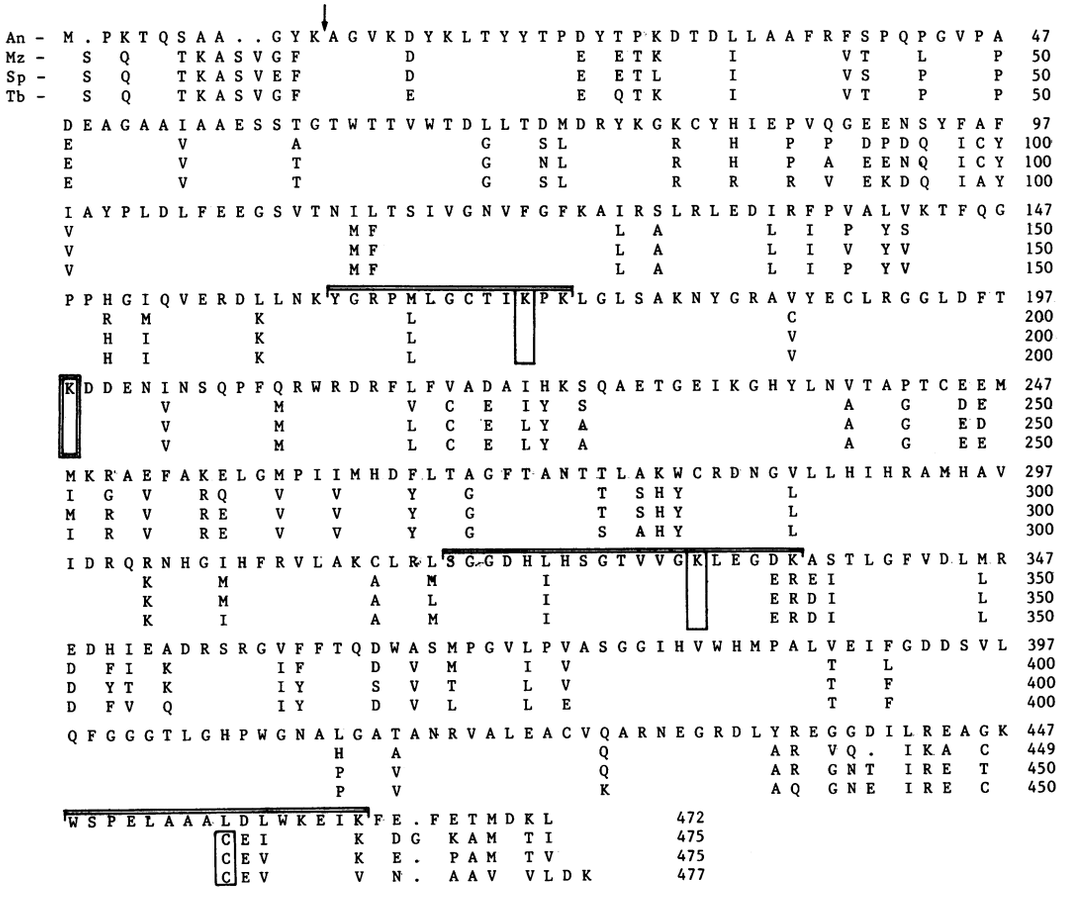

FIG. 4. Deduced amino acid sequences of *A. nidulans* LS (An), maize (Mz), spinach (Sp), and tobacco (Tb) LS proteins. Only amino acid residues different from *A. nidulans* LS (top row) are shown in the other sequences. Dots indicate deleted amino acids. Lysine residues (K) boxed with double lines bind to the activator CO_2. Double lines indicate regions containing the active sites. Lysine (K) and cysteine (C) residues to which ribulose 1,5-bisphosphate (RuBP) binds are boxed. An arrow indicates a possible processing site.

Table 1. Comparison of amino acid and nucleotide sequences among the LS genes of *A. nidulans*, tobacco, spinach, and maize

	Amino acids	Nucleotide			
		Total	1st	2nd	3rd
		Relative to *A. nidulans*			
Spinach	81.1	69.7	83.1	89.8	36.2
	(0.21 ± 0.04)	(0.39 ± 0.04)	(0.19 ± 0.04)	(0.11 ± 0.03)	(1.43 ± 0.30)
Tobacco	79.7	69.8	82.2	90.3	36.6
	(0.23 ± 0.04)	(0.39 ± 0.04)	(0.20 ± 0.05)	(0.10 ± 0.03)	(1.40 ± 0.29)
Maize	78.6	68.8	81.6	90.0	34.7
	(0.24 ± 0.05)	(0.40 ± 0.04)	(0.21 ± 0.05)	(0.11 ± 0.03)	(1.53 ± 0.34)
		Relative to tobacco			
Spinach	92.8	90.5	94.9	96.0	81.0
	(0.07 ± 0.03)	(0.10 ± 0.01)	(0.05 ± 0.02)	(0.04 ± 0.02)	(0.22 ± 0.05)
Maize	90.3	85.0	92.4	95.4	67.1
	(0.10 ± 0.03)	(0.17 ± 0.01)	(0.08 ± 0.03)	(0.05 ± 0.02)	(0.43 ± 0.08)

Numbers indicate percentage of homologous sequences relative to *A. nidulans* LS and to tobacco LS. The number of amino acid sites compared is about 470 and therefore that of total nucleotide sites is about 1,410. The evolutionary distance (K) and its standard error (σ_K) are given in parentheses.

Comparison of *A. nidulans* LS Gene with Plant LS Genes. The primary structure of the LS proteins is biochemically conservative because both active sites of RuBPCase/Oase are located on it. In contrast, the SS proteins are considered most likely to have changed more freely in the course of evolution because of fewer functional and structural constraints. To study the evolution of *A. nidulans* LS gene in detail, we compared its nucleotide sequence and the deduced amino acid sequence with those of tobacco (4), maize (10), and spinach (11). Because it is generally held that the rate of mutant substitutions in the functional genes differs depending on the position in the codon, nucleotide sequences were compared not only as a whole but also at each position of codons in the LS gene. The results are given in Table 1 in terms of the observed fraction of the site(s) for which two sequences are identical.

To estimate the evolutionary distance (K)—i.e., the total number of nucleotide or amino acid substitutions per site that have occurred in the course of evolution since the divergence of two species—we used the simple formula proposed by Jukes and Cantor (16) and Kimura and Ohta (17). Although the formula usually underestimates the distance unless some underlying assumptions are warranted, it is valid as the first approximation (18). If we let L and n be the number of possible states per site ($L = 4$ for nucleotide site and $L = 20$ or is assumed to be indefinitely large for amino acid site) and the number of sites compared, respectively, the following formula can be used to estimate K in terms of both nucleotide and amino acid substitutions

$$K = -\frac{L-1}{L}\ln\left(\frac{Ls-1}{L-1}\right). \quad [1]$$

Also, the standard error (σ_K) of K may be computed by

$$\sigma_K = \frac{L-1}{Ls-1}\sqrt{\frac{s(1-s)}{n}}. \quad [2]$$

The rate of substitution per site per year can then be obtained by dividing K by $2T$, T being the time since the divergence of two species. In the following, we assumed that T for monocotyledons (maize) and dicotyledons (spinach and tobacco) is $1–2 \times 10^8$ years (19, 20) but we did not assume any value for T for cyanobacteria and higher plants other than spinach and tobacco. Instead, assuming the rough constancy of molecular evolutionary rate among different lineages (21), we shall estimate these divergence times.

From Table 1 we calculate that for LS genes the rate of amino acid substitution is $0.25–0.5 \times 10^{-9}$ per year per site or 0.25–0.5 pauling, one pauling being one amino acid substitution per 10^9 years (22). This value is similar to that for cytochrome *c* (0.3 pauling) and insulin (0.4 pauling) (23). Among various proteins the median value is represented by hemoglobin α (1.2 pauling) so that the LS proteins have been changing at a rate far below the median and the value is really conservative.

In Fig. 5, we give the phylogenetic trees constructed by the evolutionary distances in terms of amino acid substitution and nucleotide substitution for the entire sequences of LS genes. Not only the topology but also the relative length of each corresponding branch in both trees are similar. Using formula 1 for amino acid substitution, we estimated the divergence time of *A. nidulans* and higher plants as $2.3–4.6 \times 10^8$ years, which is 2.3 times earlier than that of tobacco and maize. Thus, we suggest that the common ancestor leading to cyanobacteria and chloroplasts in higher plants existed several hundred million

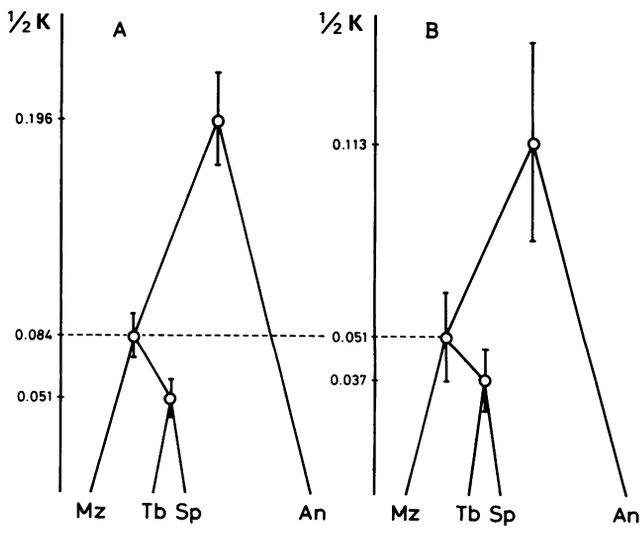

FIG. 5. Phylogenetic trees for *A. nidulans* (An), maize (Mz), tobacco (Tb), and spinach (Sp), constructed on the basis of the nucleotide substitution (*A*) and the amino acid substitution (*B*) in their LS genes. Each branch of the trees represents evolutionary divergence time.

Table 2. Codon usage of the LS genes of *A. nidulans* (A), tobacco (Tb), spinach (Sp), and maize (Mz)

	2nd position																			
	U					C					A					G				
1st		An	Tb	Sp	Mz		An	Tb	Sp	Mz		An	Tb	Sp	Mz		An	Tb	Sp	Mz	3rd
U	Phe	5	12	12	12	Ser	2	7	9	6	Tyr	2	10	14	9	Cys	2	5	5	6	U
	Phe	19	9	8	9	Ser	5	2	4	4	Tyr	12	8	5	8	Cys	5	4	4	5	C
	Leu	0	9	8	12	Ser	0	3	2	5	Term	1	1	0	1	Term	0	0	0	0	A
	Leu	10	10	10	11	Ser	10	0	0	1	Term	0	0	1	0	Trp	9	8	8	8	G
C	Leu	1	10	10	5	Pro	5	11	12	6	His	1	9	6	11	Arg	13	11	12	9	U
	Leu	10	0	0	1	Pro	7	2	3	2	His	15	5	9	4	Arg	11	5	5	10	C
	Leu	0	6	11	8	Pro	1	5	5	9	Gln	8	9	8	10	Arg	0	6	6	4	A
	Leu	22	6	2	1	Pro	9	3	2	5	Gln	4	4	3	3	Arg	3	1	0	0	G
A	Ile	1	9	11	11	Thr	1	16	15	17	Asn	1	9	8	6	Ser	0	2	5	3	U
	Ile	21	10	8	9	Thr	24	7	10	7	Asn	13	6	8	7	Ser	2	3	0	0	C
	Ile	0	2	0	4	Thr	0	5	9	6	Lys	17	21	19	18	Arg	0	7	4	3	A
	Met	12	8	10	12	Thr	4	1	0	0	Lys	9	4	4	8	Arg	0	0	2	2	G
G	Val	4	16	15	12	Ala	19	22	21	17	Asp	6	23	21	23	Gly	23	23	21	23	U
	Val	13	1	1	1	Ala	9	6	4	5	Asp	25	4	5	6	Gly	17	2	5	1	C
	Val	0	17	16	14	Ala	7	13	14	17	Glu	24	24	26	20	Gly	0	13	13	14	A
	Val	10	2	2	3	Ala	10	4	4	5	Glu	5	9	7	8	Gly	4	8	6	9	G

Term, termination.

years ago. Because the progenitor of eukaryotes had already appeared on Earth by this epoch, this suggestion is compatible with the endosymbiotic theory of chloroplasts in plants.

Table 2 shows the codon usage of the LS genes of *A. nidulans* and plants. Codon usage in *A. nidulans* LS gene is different from usages in the plant LS genes. For example, GUC and GUG codons are frequently used among valine codons in *A. nidulans* LS, whereas GUU and GUA codons are frequently used in the plant LS genes. ACC is the most frequent codon among threonine codons in *A. nidulans* LS gene but ACU is the most frequent codon in the plant LS genes. Significant differences of the codon usage are also observed in Phe, Leu, Ile, Ser, Tyr, His, Asn, Asp, and Gly codons. These differences are due to synonymous changes at the third position of codons. In eukaryotic genes, codons involving CG in this order (i.e., codons CGN and NCG) and codon NTC are almost deficient (24). However, in the LS genes of *A. nidulans* and the plant chloroplasts these codons are used with normal frequency, which supports the prokaryotic feature of chloroplast genes.

We thank Prof. R. G. Herrmann for providing us with the spinach LS DNA fragment and Dr. Noboru Tomioka, Dr. Ken-ichi Aoki, and Dr. Kazuo Asama for helpful discussion. This work was supported in part by Grants-In-Aid from the Ministry of Education, Science and Culture.

1. Jensen, R. H. & Bahr, J. T. (1977) *Annu. Rev. Plant Physiol.* **28**, 329–400.
2. Poulsen, C. (1981) *Carlsberg Res. Commun.* **46**, 259–278.
3. Ellis, R. J. (1981) *Annu. Rev. Plant Physiol.* **32**, 111–137.
4. Shinozaki, K. & Sugiura, M. (1982) *Gene* **20**, 91–102.
5. Tomioka, N., Shinozaki, K. & Sugiura, M. (1982) *Mol. Gen. Genet.* **148**, 359–363.
6. Tohdoh, N. & Sugiura, M. (1982) *Gene* **17**, 213–218.
7. Sugiura, M. & Kusuda, J. (1979) *Mol. Gen. Genet.* **172**, 137–141.
8. Sugiura, M. (1980) *Anal. Biochem.* **108**, 227–229.
9. Maxam, A. M. & Gilbert, W. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 560–564.
10. McIntosh, L., Poulsen, C. & Bogorad, L. (1980) *Nature (London)* **288**, 556–560.
11. Zurawski, G., Perrot, B., Bottomley, W. & Whitfeld, P. (1981) *Nucleic Acids Res.* **9**, 3235–3270.
12. Badger, M. R. & Lorimer, G. H. (1976) *Arch. Biochem. Biophys.* **175**, 723–729.
13. Lorimer, G. H. (1981) *Biochemistry* **20**, 1236–1240.
14. Schloss, J. V., Stringer, C. D. & Hartman, F. C. (1978) *J. Biol. Chem.* **253**, 5707–5711.
15. Jordan, D. B. & Ogren, W. L. (1981) *Nature (London)* **291**, 513–515.
16. Jukes, T. H. & Cantor, C. R. (1969) in *Mammalian Protein Metabolism*, ed. Munro, H. N. (Academic, New York), pp. 21–132.
17. Kimura, M. & Ohta, T. (1972) *J. Mol. Evol.* **2**, 87–90.
18. Takahata, N. & Kimura, M. (1981) *Genetics* **98**, 641–657.
19. Muller, J. (1981) *Bot. Rev.* **47**, 1–142.
20. Asama, K. (1982) *Bull. Natl. Sci. Mus. Ser. C* **8**, 43–58.
21. Kimura, M. (1968) *Nature (London)* **217**, 624–626.
22. Kimura, M. (1969) *Proc. Natl. Acad. Sci. USA* **63**, 1181–1188.
23. Dayhoff, M. O. (1972) *Atlas of Protein Sequence and Structure* (Natl. Biomed. Res. Found., Washington, DC), Vol. 5.
24. Kafatos, F., Efstratiadis, A., Forget, B. G. & Weissman, S. M. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 5618–5622.