Proc. Natl. Acad. Sci. USA
Vol. 80, pp. 1835–1839, April 1983
Biochemistry

Isolation and sequence of the gene for the large subunit of ribulose-1,5-bisphosphate carboxylase from the cyanobacterium *Anabaena* 7120

(chloroplast DNA probes/sequence conservation/endosymbiont origin)

STEPHANIE E. CURTIS AND ROBERT HASELKORN

Department of Biophysics and Theoretical Biology, The University of Chicago, Chicago, Illinois 60637

Communicated by Bernard Roizman, December 27, 1982

ABSTRACT **Cloned DNA probes containing genes coding for the large subunit of ribulose-1,5-bisphosphate carboxylase (*rbc*A) of corn and of *Chlamydomonas* were used to identify, by heterologous hybridization, DNA fragments from *Anabaena* 7120 carrying the corresponding gene sequence. The same probes were used to isolate, from a recombinant λ library, a 17-kilobase-pair *Eco*RI *Anabaena* DNA fragment containing the coding sequence for the *rbc*A gene. The entire coding sequence, as well as 210 base pairs of 5′ flanking region and 210 base pairs of 3′ flanking region, was determined. Comparison of the nucleotide and amino acid sequences with those of corn, spinach, *Chlamydomonas*, and *Synechococcus rbc*A genes revealed homology of 71–77% at the nucleotide level and 80–85% at the amino acid level. Conservation of sequence is lost immediately outside the coding region on either side. Codon usage in the *Anabaena rbc*A gene is not significantly different from that in the *Anabaena* genes for nitrogenase reductase and nitrogenase β subunit.**

We are interested in understanding the mechanisms by which genes are regulated during the differentiation of heterocysts—cells specialized for aerobic nitrogen fixation—which appear at regular intervals along the filaments of certain cyanobacteria such as *Anabaena* (1). In the laboratory, the differentiation process can be manipulated by varying the concentration of combined nitrogen in the medium. Development of heterocysts is repressed in media containing nitrate, ammonia, or amino acids; differentiation and, eventually, nitrogen fixation activity are induced by withdrawal of these sources of nitrogen.

The heterocyst differs in many ways from its antecedent vegetative cell. The heterocyst has two new cell wall layers, many new enzymes of carbohydrate catabolism, and all the components of nitrogen fixation. It has lost the capacity to fix carbon dioxide and to evolve oxygen. Analysis of total cell proteins by acrylamide gel electrophoresis has shown further that many vegetative cell proteins are absent from heterocysts. In order to study the process of differentiation at the level of gene transcription, we have attempted to clone genes representative of the sets that are regulated either up or down during differentiation. The structural genes for nitrogenase and nitrogenase reductase are examples of the former class (2, 3). For the latter class, we chose the gene coding for the large subunit of *r*ibulose*b*isphosphate (RuP_2) *c*arboxylase (*rbc*A), a protein whose enzyme activity (4), antigen (5), and polypeptide band (6) disappear rapidly from developing heterocysts in cultures deprived of combined nitrogen. In this communication, we report the cloning of that gene and the determination of its complete coding sequence.

RuP_2 carboxylase has been recognized for a long time as an extremely abundant plant protein (7). It is found in the chloroplasts of higher plants, where it catalyzes the addition of CO_2 to RuP_2 in the first step of the Calvin–Bassham pathway for CO_2 fixation. The enzyme contains eight subunits of molecular weight ≈50,000 and eight of molecular weight ≈12,500. The large subunit is coded for and synthesized in the chloroplast (8). The small subunit is coded in nuclear genes (9, 10); the transcripts are spliced, translated in the cytoplasm into a precursor molecule, and transported across the chloroplast membrane (11). Transport is accompanied by removal of a signal sequence (11). The large subunit, which contains the catalytic site, is highly conserved antigenically; antibody to the tobacco protein crossreacts with algal proteins, for example (12). Small subunits from different species are much less closely related. Relatedness of the large subunit, based on biochemical parameters, physical properties, and serology, extends from plants to cyanobacteria (13); these observations have been interpreted as support for the endosymbiotic origin of chloroplasts. We therefore were interested in comparing the sequence of a cyanobacterial *rbc*A gene and its product with the chloroplast genes and their products from several plants and algae. These comparisons, which indicate common ancestry of the cyanobacterial, algal chloroplast, and plant chloroplast genes, will be presented as well.

MATERIALS AND METHODS

Total DNA from *Anabaena* 7120 was isolated by a modification of the procedure of Bendich *et al.* (14) applied to 15-liter cultures as described (2). Cells were resuspended in 30 ml of 50 mM NaCl/50 mM Na_2EDTA/50 mM Tris·HCl, pH 8.5. Lysozyme was added to 5 mg/ml and the suspension was incubated at 37°C for 30 min. The cells were lysed by adjusting the suspension to 2.5% Triton X-100 and 1% Sarkosyl. Then, 50 g of CsCl was added and the lysate was centrifuged at 10,000 rpm in an HB-4 Sorval rotor. After removal of the floating layer of protein and lipid, the supernatant containing the nucleic acids was adjusted with CsCl to a density of 1.60 g/ml. Ethidium bromide (EtdBr, 10 mg/ml) was added to a final concentration of 300 μg/ml, and the CsCl/EtdBr mixture was centrifuged in a Beckman Ti 60 rotor at 45,000 rpm for 18–24 hr at 20°C. The DNA band was visualized with ultraviolet light, removed through the side of the tube by using a 16-gauge needle, and rebanded in a second CsCl/EtdBr gradient. EtdBr was removed by extraction with isopropanol saturated with CsCl and the remaining solution of DNA was dialyzed against several changes of TE buffer (10 mM Tris·HCl, pH 8.0/1 mM Na_2EDTA).

Phage and *Escherichia coli* strains were grown and phage and plasmid DNA were isolated as described (2, 3). Plasmid pZmB1, containing a 4.0-kilobase-pair (kbp) *Bam*HI fragment of corn

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: RuP_2, ribulose bisphosphate; EtdBr, ethidium bromide; kbp, kilobase pair(s).

chloroplast DNA, was obtained from A. A. Gatenby (15). Plasmid pCr34.1, containing a 1.0-kbp *Hin*dIII fragment of *Chlamydomonas* chloroplast DNA, was obtained from R. Hallick. This fragment is denoted R-15.2 in the publication describing the sequence of the *Chlamydomonas rbc*A gene (16). Recombinant λ phage libraries constructed by Rice *et al.* (3) were screened by plaque hybridization according to Benton and Davis (17). *Anabaena* DNA inserts were recloned from phage vectors into plasmid pBR322 and pBR325 (18). Restriction endonuclease digestions, agarose gel electrophoresis, and Southern blot hybridization with ^{32}P-labeled DNA probes were performed as described (2, 3).

DNA labeling and sequence determinations were carried out according to Maxam and Gilbert (19) with the exception of the piperidine cleavage step which was performed as described by Smith and Calvo (20). All restriction sites but one were overlapped and every sequence was determined on both strands. In order to obtain the sequences around the *Hin*dIII site, the sequence was read through the *Hin*dIII site into the vector DNA on one strand from each direction. On the other strand the sequence was obtained by reading up to the labeled end on a 20% analytical gel. Sequencing ladders were transcribed manually, entered into an Amdahl computer manually, and translated by using the program of Queen and Korn (21).

RESULTS AND DISCUSSION

We shall describe the use of clones containing *rbc*A genes from corn and *Chlamydomonas* chloroplast DNA as probes to identify *Anabaena* fragments containing homologous sequences. We were confident that heterologous probes would serve in this regard because of the extensive conservation of structure among carboxylase large subunits from many sources. In addition, using an antiserum against tobacco large subunit (kindly provided by D. Bourque), we observed crossreaction in Ouchterlony immunodiffusion tests with crude extracts of *Anabaena* and could specifically precipitate from such extracts two polypeptides whose mobility in acrylamide gels corresponded to those of the carboxylase large and small subunits (data not shown).

Two cloned *rbc*A probes were therefore hybridized to Southern blots of total *Anabaena* 7120 DNA (Fig. 1). The probe pZmB1 contains a 4.0-kbp fragment from corn chloroplast DNA consisting of the corn *rbc*A gene plus 2.5 kbp of flanking sequence (15). pZmB1 hybridized to *Anabaena Hin*dIII fragments of 10.5 and 8.5 kbp and to *Eco*RI fragments of 17.0 and 5.6 kbp. The probe pCr34.1 contains a 1.0-kbp fragment from *Chlamydomonas* chloroplast DNA containing the 3′ end of the algal *rbc*A gene (16). pCr34.1 also hybridized to the 17.0-kbp *Eco*RI fragment and the 10.5-kbp *Hin*dIII fragment.

The 5.6-kbp *Eco*RI and 8.5-kbp *Hin*dIII fragments, which hybridized to the corn probe pZmB1 but not to pCr34.1 (Fig. 1), do not contain any portion of the *Anabaena rbc*A gene. These fragments hybridized solely to DNA sequences that flank the corn *rbc*A gene in the probe pZmB1. This region has recently been shown to contain a gene for the β subunit of the chloroplast coupling factor (22). We have cloned the corresponding 5.6-kbp *Eco*RI *Anabaena* DNA fragment in recombinant λ and confirmed that it contains a gene for the β subunit of ATPase by hybridization with a probe of known sequence from spinach chloroplast DNA, kindly provided by P. Whitfeld and W. Bottomley (23). The *Eco*RI fragment containing the *Anabaena* β subunit gene is not adjacent to the fragment containing the *rbc*A gene (data not shown).

The *Eco*RI fragments that showed homology to the plant probes were isolated by screening a λ Charon 4 library of cloned *Anabaena* DNA with [^{32}P]pZmB1. The clone λCh4-An600 contains the 17.0-kbp *Eco*RI fragment indicated in Fig. 1. A restriction

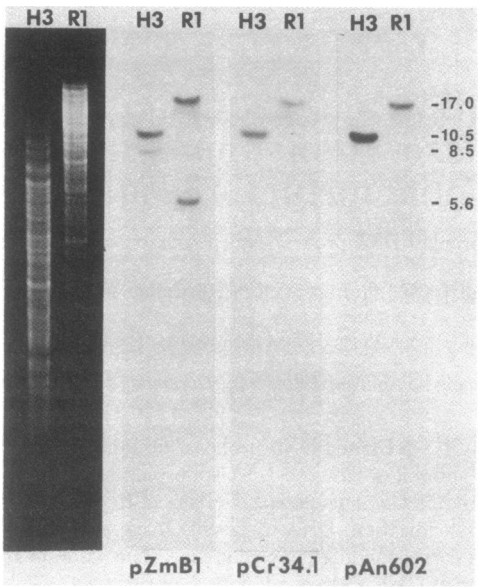

FIG. 1. Identification of the *Anabaena* gene coding for the large subunit of RuP_2 carboxylase (*rbc*A). Total *Anabaena* DNA was digested with either *Eco*RI (R1) or *Hin*dIII (H3), electrophoresed, blotted, and hybridized with the indicated probes. Sizes of the bands detected are given in kbp. The map of pAn602 is given in Fig. 2. The bands at 8.5 and 5.6 kbp detected by the corn probe but not by the *Chlamydomonas* probe contain sequences that include the gene for the β subunit of ATPase (see text).

endonuclease map of λCh4-An600 was constructed and the heterologous probes were used to determine the location and orientation of the *rbc*A gene. A 5′ probe was constructed by subcloning an internal 0.6-kbp *Pst* I fragment from the 5′ end of the corn *rbc*A gene of pZmB1. Southern blots of λCh4-An600 were hybridized with ^{32}P-labeled pZmB1, pCr34.1, or the *Pst* I fragment (data not shown). The position and length (1.5 kbp) of the hybridizing regions indicated that λCh4-An600 carries an entire *rbc*A gene and suggested that the homology is confined to the *rbc*A coding region. A subclone, pAn602, which contains most of the *rbc*A gene, was constructed by insertion of an *Eco*RI/*Hin*dIII fragment of λCh4-An600 into pBR322 (Fig. 2). When [^{32}P]pAn602 was hybridized back to *Anabaena* total DNA (Fig.

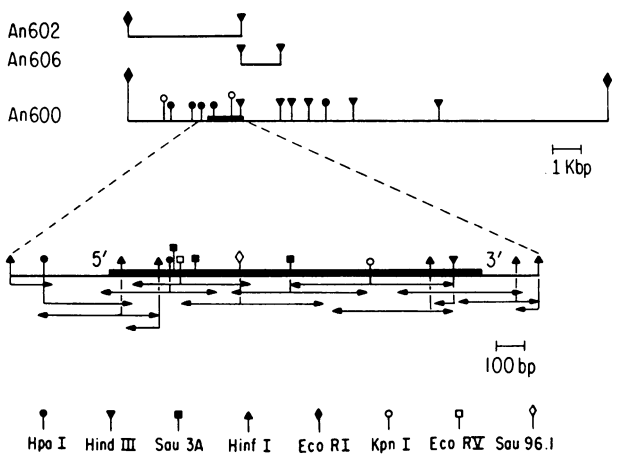

FIG. 2. Physical map of the 17-kbp *Anabaena Eco*RI DNA fragment identified in Fig. 1 and cloned in λ Charon 4A (An600). The thick bar represents the location of the *rbc*A gene, with 3′ and 5′ ends indicated. The subfragment containing *rbc*A was subjected to sequence analysis with the strategy indicated in the lower portion. To facilitate sequence analysis, subfragments of An600 were cloned in plasmids called pAn602 and pAn606, as shown.

1), it hybridized only to the 17.0-kbp *Eco*RI fragment from which it was derived and to the *Hin*dIII fragment which overlaps it in the *Anabaena* chromosome. This result, which has been reproducible with much longer exposure times for the autoradiography, indicates that *Anabaena* contains a single copy of the *rbc*A gene.

Having located and oriented the *rbc*A gene in An600, we next determined the nucleotide sequence of the gene by following the strategy outlined in the lower portion of Fig. 2. Fig. 3 displays the nucleotide sequence of the noncoding strand of the *rbc*A gene and its flanking regions. The open reading frame contains 1,428 base pairs, starting with an ATG and ending with two termination codons in tandem. Examination of the sequence flanking the open reading frame shows a possible ribosome binding sequence (24), G-G-A-A-G-G, at residues −12 to −7. The relevance of this sequence is still unknown, however, because the 3′ terminal sequence of *Anabaena* 16S rRNA has not been determined.

Examination of the sequence in the 3′ direction has shown a possible prokaryotic terminator, an inverted repeat followed by a stretch of T residues (25) at positions 1,450–1,508. Because the *Eco*RI–*Hin*dIII fragment of pAn602 provides an active promoter, in *E. coli*, for expression of β-galactosidase or tetracycline resistance in suitable fusion vectors (C. Richaud, personal communication), we expect the *Anabaena rbc*A promoter to resemble a typical prokaryotic one. The *rbc*A 5′ flanking region contains several candidates for such promoter sequences. The nucleotide sequences of *rbc*A genes from another prokaryote and three eukaryotes are available. The *Anabaena rbc*A coding sequence shows 71–77% identity with the other *rbc*A genes (Ta-

```
-210
AGCGTTAACT  ATGCACTCCT  AGATGAGCAA  GACACTGGTG  AAGAGGATTA  CCACTAAAGC  TAAGTGTTAG
-140
TTGCAGAAAG  GTCGCTGACC  TCTACCAAAA  GATTATTCCT  GTTTTTCTCG  TGCTGATAGG  GAGGTAGGGC
-70
AATTGTGAGA  GGAAATTGTA  CCAAAACGTG  ATTAGATAAG  TAAAAAGAGT  GACATCTTGG  AAGGATAGAT

                                                                              72
ATG TCT TAC GCT CAA ACG AAG ACT CAG ACA AAA TCT GGG TAT AAA GCC GGG GTT CAA GAT TAC AGA CTA ACT
Met Ser Tyr Ala Gln Thr Lys Thr Gln Thr Lys Ser Gly Tyr Lys Ala Gly Val Gln Asp Tyr Arg Leu Thr
                                                                             144
TAT TAC ACA CCT GAT TAC ACA CCT AAA GAT ACA GAT ATT CTG GCG GCA TTC CGT GTT ACA CCC CAG CCC GGA
Tyr Tyr Thr Pro Asp Tyr Thr Pro Lys Asp Thr Asp Ile Leu Ala Ala Phe Arg Val Thr Pro Gln Pro Gly
                                                                             216
GTT CCC TTT GAG GAA GCG GCT GCG GCA GTA GCG GCT GAG TCT TCT ACT GGT ACT TGG ACG ACC GTA TGG ACA
Val Pro Phe Glu Glu Ala Ala Ala Ala Val Ala Ala Glu Ser Ser Thr Gly Thr Trp Thr Thr Val Trp Thr
                                                                             288
GAC CTG TTA ACC GAT CTA GAT CGT TAC AAA GGT CGT TGC TAC GAT ATC GAA CCA GGT CCC GGC GAA GAC AAC
Asp Leu Leu Thr Asp Leu Asp Arg Tyr Lys Gly Arg Cys Tyr Asp Ile Glu Pro Val Pro Gly Glu Asp Asn
                                                                             360
CAA TCC ATT GCC TAC ATC GCT TAT CCT TTG GAT CTG TTT GAA GAA GGC TCC ATC ACC AAC GTT TTG ACC TAC
Gln Phe Ile Ala Tyr Ile Ala Tyr Pro Leu Asp Leu Phe Glu Glu Gly Ser Ile Thr Asn Val Leu Thr Ser
                                                                             432
ATT GTA GGT AAC GTA TTT GTT TTT AAA GCA TTA CGC GCA TTG CGT TTG GAA GAC ATT CGC TTT CCT GTT GCT
Ile Val Gly Asn Val Phe Gly Phe Lys Ala Leu Arg Ala Leu Arg Leu Glu Asp Ile Arg Phe Pro Val Ala
                                                                             504
TAC ATC AAG ACC TTC CAA GGC CCT CCT CAC GGT ATC CAA GTT GAG CGT GAC AAA TAA AAC AAA TAT GGC CGT
Tyr Ile Lys Thr Phe Gln Gly Pro Pro His Gly Ile Gln Val Glu Arg Asp Lys Leu Asn Lys Tyr Gly Arg
                                                                             576
CCT CTG TTG GGT TGT ACC ATC AAA CCA AAA TTA GGT CTG TCT GCT AAG AAC TAC GGA CGC GCT GTA TAC GAG
Pro Leu Leu Gly Cys Thr Ile Lys Pro Lys Leu Gly Leu Ser Ala Lys Asn Tyr Gly Arg Ala Val Tyr Glu
                                                                             648
TGT TTG CGC GGT GGT TTG GAC TTC ACC AAA GAC GAC GAA AAC ATT AAC TCC GCA CCA TTC CAA AGA TGG CGC
Cys Leu Arg Gly Gly Leu Asp Phe Thr Lys Asp Asp Glu Asn Ile Asn Ser Ala Pro Phe Gln Arg Trp Arg
                                                                             720
GAT CGC TTC TTG TTT GTA GCT GAT GCC ATC ACC AAA GCA CAA GCA GAA ACA GGC GAA ATC AAA GGT CAC TAC
Asp Arg Phe Leu Phe Val Ala Asp Ala Ile Thr Lys Ala Gln Ala Glu Thr Gly Glu Ile Lys Gly His Tyr
                                                                             792
CTA AAC GTG ACC GCT CCT ACC TGT GAA GAA ATG CTA AAA CGG GCT GAG TAC GCT AAA GAA CTC AAA CAG CCC
Leu Asn Val Thr Ala Pro Thr Cys Glu Glu Met Leu Lys Arg Ala Glu Tyr Ala Lys Glu Leu Lys Gln Pro
                                                                             864
ATC ATC ATG CAC GAC TAC CTG ACC GCA GGT TTC ACA GCT AAC AAC ACC TTG GCT CGT TGG TGT CGT GAC AAC
Ile Ile Met His Asp Tyr Leu Thr Ala Gly Phe Thr Ala Asn Thr Thr Leu Ala Arg Trp Cys Arg Asp Asn
                                                                             936
GGT CTT CTA CTG CAC ATC CAC CGC GCG ATG CAC GCA GTA ATC GAC CGT CAA AAG AAC CAC GGT ATC CAC TTC
Gly Leu Leu Leu His Ile His Arg Ala Met His Ala Val Ile Asp Arg Gln Lys Asn His Gly Ile His Phe
                                                                            1008
CGT GTA TTG GCT AAA GCC CTA CGT CTA TCT GGT GGT GAC CAC ATC CAC ACC GGT ACC GTA GTA GGT AAA TTG
Arg Val Leu Ala Lys Ala Leu Arg Leu Ser Gly Gly Asp His Ile His Thr Gly Thr Val Val Gly Lys Leu
                                                                            1080
GAA GGT GAA CGC GGT ATC ACA ATG GGC TTC GTT GAC CTA CTA CGT GAA AAC TAC GTT GAG CAA GAC AAG TCT
Glu Gly Glu Arg Gly Ile Thr Met Gly Phe Val Asp Leu Leu Arg Glu Asn Tyr Val Glu Gln Asp Lys Ser
                                                                            1152
CGC GGT ATT TAC TTT ACC CAA GAC TGG GCT TCT CTA CCT GGT GTA ATG GCA GTT GCT TCC GGT GGT ATC CAC
Arg Gly Ile Tyr Phe Thr Gln Asp Trp Ala Ser Leu Pro Gly Val Met Ala Val Ala Ser Gly Gly Ile His
                                                                            1224
GTA TGG CAT ATG CCA GCG TTG GTA GAA ATC TTC GGT GAT GAC TTC GTA CTA CAA TTC GGT GGT GGT ACA CTC
Val Trp His Met Pro Ala Leu Val Glu Ile Phe Gly Asp Asp Ser Val Leu Gln Phe Gly Gly Gly Thr Leu
                                                                            1296
GGA CAC CCT TGG GGT AAC GCT CGT GGT GCA ACC GCT AAC CGT GTA GCT TTG GAA GCT TGC GTC CAA GCA CGT
Gly His Pro Trp Gly Asn Ala Arg Gly Ala Thr Ala Asn Arg Val Ala Leu Glu Ala Cys Val Gln Ala Arg
                                                                            1368
AAC GAA GGT CGT AAC TTG GCT CGT GAA GGT AAC GAC GTT ATC CGT GAA GCT GCT AAG TGG TCT CCT GAA TTG
Asn Glu Gly Arg Asn Leu Ala Arg Glu Gly Asn Asp Val Ile Arg Glu Ala Ala Lys Trp Ser Pro Glu Leu
                                                                            1434
GCT GTC GCT TGC GAA CTG TGG AAA GAA ATC AAG TTC GAG TTT GAG GCA ATG GAT ACC GTC TGA TAA
Ala Val Ala Cys Glu Leu Trp Lys Glu Ile Lys Phe Glu Phe Glu Ala Met Asp Thr Val End End
                                                                            1504
AGGCTGAAGA  TGCAAGGATG  AAGGTTAAAT  GGAAAATACC  TTTCATCCTT  ACATTCTTTA  AACTTCATCC
                                                                            1574
TTTTAAAGGG  CTGGGTCAAG  CATGAATCTC  AAGCAAATAG  CGAAAGATAC  AGCCAAAACG  CTCCAAAGCT
                                                                            1644
ACCTGACTTA  TCAGGCGCTA  ATGACTGTGT  TGGCACAGCT  AGGCGAAATG  AATCCGCCGT  TAGCATTATG
```

FIG. 3. Nucleotide sequence of the *rbc*A gene from *Anabaena* 7120. Numbering of residues begins with the first A of the methionine codon. A possible ribosome binding site at positions −7 to −12 is underlined.

Table 1. Sequence homologies between *rbc*A genes of *Anabaena* and other organisms

	% homology with other genes			
	Synechococcus	*Chlamydomonas*	Spinach	Corn
Nucleotides	77	76	74	71
Amino acids	84	84	85	80

Data for these comparisons are from the following references: *Synechococcus*, B. Reichelt and S. Delaney, personal communication; *Chlamydomonas*, ref. 16; spinach, ref. 26; and corn, refs. 27 and 28.

ble 1). It is somewhat surprising that the *Anabaena rbc*A gene is about as homologous to the gene from *Synechococcus*, a unicellular cyanobacterium, as it is to the gene from *Chlamydomonas*, a green alga, or to the spinach gene. The spinach and corn gene sequences are 84% homologous (26). Conservation of nucleotide sequence ends immediately outside the coding region in both directions.

The amino acid sequence of the *Anabaena* RuP_2 carboxylase large subunit, determined by translating the gene sequence, is compared with the sequence of large subunits from other species in Fig. 4. The degree of homology is astonishing: there are few positions at which amino acids are not identical or related by single base substitutions. Most differences at the nucleotide level are found in the third position of codons, resulting in no amino acid change and a high percentage of homology at the amino acid level (Table 1). There are several exceptions to this generalization. One is the corn sequence between amino acid residues 442 and 448 where there are numerous two-base changes and an apparent deletion of one residue. A second exception to rigid conservation occurs at the beginning of the gene. The *Anabaena* protein is one amino acid longer than the protein from chloroplasts and five longer than that from *Synechococcus*. Amino acid sequence data from barley (32) indicate that the mature barley protein commences with an alanine residue (position 16 in Fig. 4). It is interesting to note that it is at the lysine residue preceding alanine-16 that the strong homology between the five *rbc*A sequences of Fig. 4 begins. Langridge (33) reported that the spinach protein synthesized in *E. coli* cell-free extracts is 1–2 kilodaltons larger than the mature protein but could be reduced to the size of the latter by treatment with a soluble chloroplast extract. As suggested previously (26), it is possible that the amino-terminal peptides of *rbc*A proteins are discarded, and thus variations at this end are tolerated. Proof of this hypothesis awaits the determination of *rbc*A amino termini from sources other than barley. In contrast, although the carboxy termini of the large subunit sequences presented in Fig. 4 vary somewhat in amino acid composition, they are invariant in length.

The conservation of amino acid sequence around the cysteine and lysine residues that are specifically affinity labeled by substrate analogs (29) and the lysine residue at the site of carbamate formation in enzyme activated by CO_2 (30) (see Fig. 4) has already been noted (15, 26, 31). These residues begin to define the active site of the enzyme. Residues 171–188 and 191–206 are fully conserved in all five proteins; from residue 147 to 212 there are only nine single amino acid replacements in all. A cysteine residue near the carboxyl terminus of the proteins is photoaffinity labeled in the spinach enzyme, but that residue is not present in the *Synechococcus* sequence and so it may not be an essential component of the active site.

Codon utilization in the *Anabaena rbc*A gene (Table 2) displays many asymmetries. This distribution of preferred codons is also seen in the two other *Anabaena* genes for which sequence data are available, nitrogenase reductase (*nif*H) (34) and the β subunit of nitrogenase (*nif*K) (35). The codons UCG, CCG, and CGA are not used in the three *Anabaena* genes whose sequences are known (34, 35). However, these absences do not permit reduction in the number of tRNA species needed for translation because all three can be translated by using tRNA

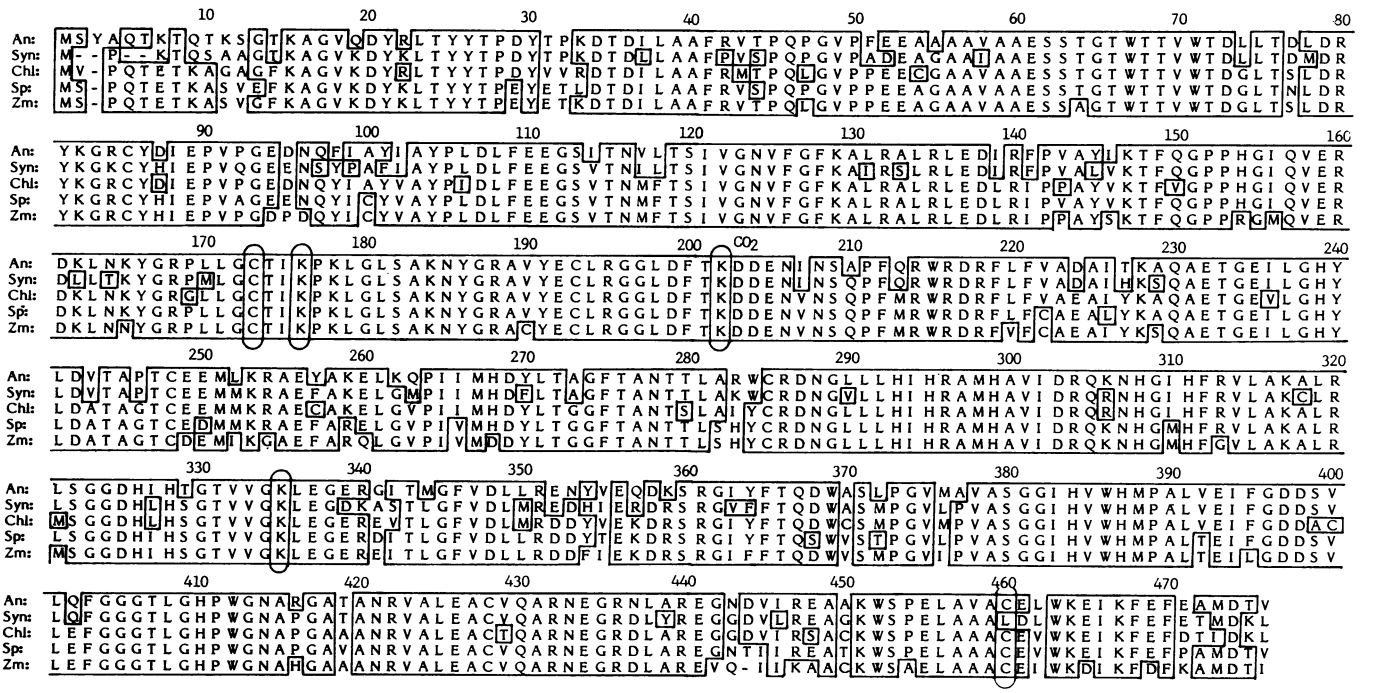

FIG. 4. Comparison of amino acid sequences of the large subunit of RuP_2 carboxylase from *Anabaena* (An), *Synechococcus* (Syn), *Chlamydomonas* chloroplasts (Chl), spinach chloroplasts (Sp), and corn chloroplasts (Zm). Residue numbering refers to the *Anabaena* sequence. Data for sources other than *Anabaena* were obtained from references mentioned in the legend to Table 1. Boxes surround positions at which at least three of the proteins have the same amino acid. Circled residues have been suggested as participants in the catalytic site (29–31). The lysine residue labeled CO_2 is the site of carbamate formation in enzyme activated by CO_2 (30).

Table 2. Codon utilization in the *Anabaena rbcA* gene

UUU	Phe	8	UCU	Ser	9	UAU	Tyr	4	UGU	Cys	4
UUC		12	UCC		4	UAC		15	UGC		3
UUA	Leu	4	UCA		1	UAA	ochre	1	UGA	opal	1
UUG		15	UCG		0	UAG	amber	0	UGG	Trp	9
CUU	Leu	1	CCU	Pro	11	CAU	His	1	CGU	Arg	18
CUC		2	CCC		5	CAC		12	CGC		9
CUA		11	CCA		4	CAA	Gln	12	CGA		0
CUG		8	CCG		0	CAG		3	CGG		1
AUU	Ile	6	ACU	Thr	4	AAU	Asn	0	AGU	Ser	0
AUC		19	ACC		18	AAC		17	AGC		0
AUA		0	ACA		10	AAA	Lys	18	AGA	Arg	2
AUG	Met	8	ACG		2	AAG		7	AGG		0
GUU	Val	11	GCU	Ala	25	GAU	Asp	12	GGU	Gly	31
GUC		3	GCC		4	GAC		16	GGC		6
GUA		15	GCA		13	GAA	Glu	23	GGA		3
GUG		1	GCG		6	GAG		8	GGG		2

needed for other codons. In the cases of UCG and CGA, the sequences of the corresponding seryl$_2$ tRNA and arginyl tRNA from *E. coli* have been determined and each has been shown to recognize three codons *in vitro*. The arginyl tRNA anticodon sequence is ICG; the seryl$_2$ tRNA anticodon sequence is 5-carboxymethoxy-UGA (36). The *nif* genes are expressed only in the absence of a combined nitrogen source, whereas *rbcA* is expressed only when such a source is present. Under aerobic conditions, *nif* gene expression requires, in addition, heterocyst differentiation. It might have been expected, therefore, that the distribution of codons utilized in the two classes of genes would differ, allowing a further level of differential expression beyond control of transcription. The distributions do not differ; therefore, the postulated translational control can be ruled out.

Note Added in Proof. The nucleotide sequence of the RuP_2 carboxylase large subunit gene from tobacco chloroplasts was determined recently by Shinozaki and Sugiura (37). The sequence is similar to the sequences reported here.

We are grateful to B. Reichelt and S. Delaney for exchange of sequence data prior to publication. We thank D. Bourque, R. Hallick, A. Gatenby, P. Whitfeld, and W. Bottomley for gifts of antisera and probes, V. DeRose for excellent technical assistance, P. Lammers and S. Robinson for helpful advice, and C. Sherman for preparing Figs. 3 and 4. This work was supported by Grant GM 21823 from the National Institutes of Health and Grant 5901-0410 from the U.S. Department of Agriculture, Science and Education Administration, through the Competitive Research Grants Office. S.E.C. was a Postdoctoral Fellow of the National Institute of General Medical Sciences.

1. Haselkorn, R. (1978) *Annu. Rev. Plant Physiol.* **29**, 319–344.
2. Mazur, B. J., Rice, D. & Haselkorn, R. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 186–190.
3. Rice, D., Mazur, B. J. & Haselkorn, R. (1982) *J. Biol. Chem.* **257**, 13157–13163.
4. Adams, D. G. & Carr, N. G. (1981) *Crit. Rev. Microbiol.* **9**, 35–100.
5. Codd, G. A. & Stewart, W. D. P. (1977) *FEMS Microbiol. Lett.* **2**, 247–249.
6. Fleming, H. & Haselkorn, R. (1974) *Cell* **3**, 159–170.
7. Singer, S. J., Eggman, L., Campbell, J. M. & Wildman, S. G. (1952) *J. Biol. Chem.* **197**, 233–239.
8. Chan, P. H. & Wildman, S. G. (1972) *Biochim. Biophys. Acta* **277**, 677–680.
9. Kawashima, N. & Wildman, S. G. (1972) *Biochim. Biophys. Acta* **262**, 32–39.
10. Bedbrook, J. R., Smith, S. M. & Ellis, R. J. (1980) *Nature (London)* **287**, 692–697.
11. Chua, N.-H. & Schmidt, G. W. (1978) *Proc. Natl. Acad. Sci. USA* **75**, 6110–6114.
12. Dorner, R. W., Kalmi, A. & Wildman, S. G. (1958) *Biochim. Biophys. Acta* **29**, 240–245.
13. Kieras, F. J. & Haselkorn, R. (1968) *Plant Physiol.* **43**, 1264–1270.
14. Bendich, A. J., Anderson, R. S. & Ward, B. L. (1980) in *Genome Organization and Expression in Plants*, ed. Leaver, C. J. (Plenum, New York), pp. 31–35.
15. Gatenby, A. A., Castleton, J. A. & Saul, M. W. (1981) *Nature (London)* **291**, 117–121.
16. Dron, M., Rahine, M. & Rochaix, J.-D. (1982) *J. Mol. Biol.* **162**, 775–793.
17. Benton, W. & Davis, R. (1977) *Science* **196**, 180–181.
18. Bolivar, F. & Backman, K. (1979) *Methods Enzymol.* **68**, 245–267.
19. Maxam, A. M. & Gilbert, W. (1980) *Methods Enzymol.* **65**, 499–560.
20. Smith, D. R. & Calvo, J. M. (1980) *Nucleic Acids Res.* **8**, 2255–2275.
21. Queen, C. L. & Korn, L. J. (1980) *Methods Enzymol.* **65**, 595–609.
22. Krebbers, E. T., Larrinua, I. M., McIntosh, L. & Bogorad, L. (1982) *Nucleic Acids Res.* **10**, 495–505.
23. Zurawski, G., Bottomley, W. & Whitfeld, P. R. (1982) *Proc. Natl. Acad. Sci. USA* **79**, 6260–6264.
24. Shine, J. & Dalgarno, L. (1974) *Proc. Natl. Acad. Sci. USA* **71**, 1342–1346.
25. Rosenberg, M. & Court, D. (1979) *Annu. Rev. Genet.* **13**, 319–353.
26. Zurawski, G., Perrot, B., Bottomley, W. & Whitfeld, P. (1981) *Nucleic Acids Res.* **9**, 3251–3270.
27. McIntosh, L., Poulsen, C. & Bogorad, L. (1980) *Nature (London)* **288**, 556–560.
28. Poulsen, C. (1981) *Carlsberg Res. Commun.* **46**, 259–278.
29. Hartman, F. C., Norton, I. L., Stringer, C. D. & Schloss, J. V. (1978) in *Photosynthetic Carbon Assimilation*, eds. Siegelman, H. W. & Hind, G. (Plenum, New York), pp. 245–269.
30. Lorimer, G. H. (1981) *Biochemistry* **20**, 1236–1240.
31. Lorimer, G. H. (1981) *Annu. Rev. Plant Physiol.* **32**, 349–383.
32. Poulsen, C., Martin, B. & Svendsen, I. (1979) *Carlsberg Res. Commun.* **44**, 191–199.
33. Langridge, P. (1981) *FEBS Lett.* **123**, 85–89.
34. Mevarech, M., Rice, D. & Haselkorn, R. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 6476–6480.
35. Mazur, B. J. & Chui, C.-F. (1982) *Proc. Natl. Acad. Sci. USA* **79**, 6782–6786.
36. RajBhandary, U. L. (1976) in *Cell Biology*, eds. Altman, P. L. & Katz, D. D. (Federation of American Societies for Experimental Biology, Bethesda, MD), pp. 306–308.
37. Shinozaki, K. & Sugiura, M. (1982) *Gene* **20**, 91–102.