

# Molecular cloning and nucleotide sequence of the $\alpha$ and $\beta$ subunits of allophycocyanin from the cyanelle genome of *Cyanophora paradoxa*

(phycobilisome/photosynthesis/rapid DNA sequencing)

DONALD A. BRYANT\*, ROBERT DE LORIMIER\*, DAVID H. LAMBERT\*, JAMES M. DUBBS\*, VERONICA L. STIREWALT\*, S. EDWARD STEVENS, JR.\*, RONALD D. PORTER\*, JOHN TAM†, AND ERNEST JAY†

\*Microbiology Program, The Pennsylvania State University, University Park, PA 16802; and †Department of Chemistry, University of New Brunswick, Fredericton, NB Canada E3B 5A3

Communicated by Jack Myers, December 27, 1984

**ABSTRACT** The genes for the  $\alpha$ - and  $\beta$ -subunit apoproteins of allophycocyanin (AP) were isolated from the cyanelle genome of *Cyanophora paradoxa* and subjected to nucleotide sequence analysis. The AP  $\beta$ -subunit apoprotein gene was localized to a 7.8-kilobase-pair *Pst* I restriction fragment from cyanelle DNA by hybridization with a tetradecameric oligonucleotide probe. Sequence analysis using that oligonucleotide and its complement as primers for the dideoxy chain-termination sequencing method confirmed the presence of both AP  $\alpha$ - and  $\beta$ -subunit genes on this restriction fragment. Additional oligonucleotide primers were synthesized as sequencing progressed and were used to determine rapidly the nucleotide sequence of a 1336-base-pair region of this cloned fragment. This strategy allowed the sequencing to be completed without a detailed restriction map and without extensive and time-consuming subcloning. The sequenced region contains two open reading frames whose deduced amino acid sequences are 81–85% homologous to cyanobacterial and red algal AP subunits whose amino acid sequences have been determined. The two open reading frames are in the same orientation and are separated by 39 base pairs. AP  $\alpha$  is 5' to AP  $\beta$  and both coding sequences are preceded by a polypurine, Shine-Dalgarno-type sequence. Sequences upstream from AP  $\alpha$  closely resemble the *Escherichia coli* consensus promoter sequences and also show considerable homology to promoter sequences for several chloroplast-encoded *psbA* genes. A 56-base-pair palindromic sequence downstream from the AP  $\beta$  gene could play a role in the termination of transcription or translation. The allophycocyanin apoprotein subunit genes are located on the large single-copy region of the cyanelle genome.

Phycobiliproteins are photosynthetic accessory pigments found in cyanobacteria and the chloroplasts of red algae and cryptomonads. Certain dinoflagellates, of which the best studied example is *Cyanophora paradoxa* (1), also contain these proteins in their chloroplast-like cyanelles. Spectroscopic properties define the three major classes of phycobiliproteins: phycoerythrins (PE,  $\lambda_{\max}$  540–570 nm); phycocyanins (PC,  $\lambda_{\max}$  610–625 nm); and allophycocyanins (AP,  $\lambda_{\max}$   $\approx$ 650 nm). Each phycobiliprotein is comprised of equimolar amounts of two nonidentical subunits,  $\alpha$  and  $\beta$ . Each subunit carries characteristic number(s) and type(s) of linear tetrapyrrole chromophores that are covalently attached to the polypeptides through one or two cysteinylthioether linkages (2, 3). Amino acid sequence analyses have been performed on proteins representing all of the spectroscopic classes (see data summarized in ref. 2). These studies indicate that the phycobiliproteins are a homologous family

descended from a single ancestral gene. Complete descriptions of the properties of the phycobiliproteins can be obtained from several recent reviews (2, 4–6). In all phycobiliprotein-containing organisms except the cryptomonads, the functional light-harvesting assembly is a supramolecular protein structure called the phycobilisome (for reviews, see refs. 2, 5–8). Phycobilisomes are largely composed of phycobiliproteins but also contain a small number of quantitatively minor, nonpigmented proteins termed “linker polypeptides.” The linker polypeptides play important roles in the assembly of the phycobilisome, aid in attachment of phycobilisomes to the thylakoid membrane, and alter the spectroscopic properties of phycobiliproteins to ensure efficient and unidirectional transfer of absorbed light energy to the reaction center complexes in the thylakoid membrane (2, 5, 9).

Because of the presence of phycobiliproteins and the presence of peptidoglycan in their cell-wall remnant, the cyanelles of *C. paradoxa* have generally been regarded as endosymbiotic, degenerated cyanobacteria (1). However, the genetic complexity of the cyanelle genome [ $\approx$ 127 kilobase pairs (kbp); ref. 10] is too low for a cyanobacterial genome (11, 12) and is clearly more similar to that observed for most higher plant and algal chloroplasts (13). By using inhibitors of protein synthesis specific for either 70S or 80S ribosomes, Grossman *et al.* (14) have concluded that the phycobiliproteins and the large linker phycobiliprotein of *C. paradoxa* are synthesized on 70S cyanelle ribosomes, while the rod linker polypeptides are apparently encoded in the nucleus and are translated on 80S cytoplasmic ribosomes. These observations are consistent with the idea that the phycobiliproteins are encoded in the cyanelle genome of *C. paradoxa*; hence, *C. paradoxa* could provide an attractive system for the isolation of genes encoding phycobiliproteins. Lemaux and Grossman (15) exploited these observations in obtaining cloned fragments from the small single-copy region of the cyanelle DNA of *C. paradoxa* that carry the coding sequence for at least the  $\beta$  subunit of PC.

In this communication we describe the molecular cloning and complete nucleotide sequences of the genes encoding the  $\alpha$  and  $\beta$  subunits of AP from the cyanelle genome of *C. paradoxa*. These sequences are discussed in comparison with our recently completed sequence analysis of the genes encoding the  $\alpha$ - and  $\beta$ -subunit apoproteins of PC from the unicellular cyanobacterium *Synechococcus* 7002 (*Agmenellum quadruplicatum* strain PR-6; ref. 16). In addition, we describe a rapid, efficient method for nucleotide sequencing using the chain-termination method on double-stranded templates and using synthetic oligonucleotides as primers.

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: AP, allophycocyanin; PC, phycocyanin; PE, phycoerythrin; kbp, kilobase pair(s); bp, base pair(s).

## MATERIALS AND METHODS

**Materials.** Restriction endonucleases, T4 DNA ligase, T4 polynucleotide kinase, and DNA polymerase I were purchased from Bethesda Research Laboratories. The Klenow fragment of DNA polymerase I was purchased from Boehringer Mannheim. [ $\gamma$ - $^{32}$ P]ATP ( $\approx 4500$  Ci/mmol; 1 Ci = 37 GBq) and [ $\alpha$ - $^{32}$ P]dATP ( $\approx 600$  Ci/mmol) were purchased from ICN. The M13 reverse sequencing primer was purchased from New England Biolabs. Technical-grade cesium chloride was used for all DNA preparations. All other chemicals were reagent grade.

**Culture Conditions, DNA Purification, and Molecular Cloning.** *C. paradoxa* (originally obtained from the Pasteur Culture Collection, Institut Pasteur, Paris) was obtained from Jessup M. Shively (Clemson University) and grown in Allen's modified medium (17) containing 10 mM Hepes buffer, pH 7.2. Cyanelles were isolated as described (18). Cyanelle DNA was extracted as described in ref. 19 using the method of Godson and Vapnek (20). Cyanelles were lysed by treatment with lysozyme (2.7 mg/ml) followed by treatment with NaDodSO<sub>4</sub> (1.33%, wt/vol). A partial *Pst* I library of the cyanelle genome was constructed in *Escherichia coli* strain RDP 145 (21) using pBR322 as the vector.

**Synthesis of Oligonucleotides.** The 16 tetradecameric oligonucleotides 3' T-A-C-G-T-(T/C)-C-T-(A/G)-C-G-(A/G/C/T)-T-A 5' (described in *Results*) were separately synthesized and purified as described (22). All other oligonucleotides were synthesized on controlled-pore glass supports using the phosphoramidite chemistry developed by Caruthers and co-workers (23, 24) in an Applied Biosystems (Foster City, CA) model 380A automated DNA synthesizer. After complete deprotection in concentrated ammonium hydroxide, the products were lyophilized twice, washed with 100% ethanol, lyophilized, suspended in 2 M ammonium acetate, and collected by centrifugation after overnight ethanol precipitation at  $-20^{\circ}\text{C}$ . The precipitates were suspended in sterile water and the absorbance at 260 nm was determined after appropriate dilution. Working stocks were approximately 15  $\mu\text{M}$  ( $A_{260\text{ nm}} \approx 2.0$ ). Oligonucleotides in the range of 14–17 bases were  $\approx 85\%$  full-length product and were sufficiently pure for use as sequencing primers without additional purification.

**Hybridization with Oligonucleotides.** Probe oligonucleotides were labeled at the 5' end with [ $\gamma$ - $^{32}$ P]ATP by T4 polynucleotide kinase. Hybridization experiments with Southern blots of DNA digested with restriction enzymes or of lysed colonies on nitrocellulose were performed as described (16, 25). Hybridization conditions for oligonucleotide probes were selected as described (25).

**Nucleotide Sequencing Studies.** Sequencing by the chain-termination (dideoxy) method (26) was carried out according to Heidecker *et al.* (27). The plasmid DNA template was linearized by digestion to completion with an appropriate restriction enzyme; the molar ratio of the primer/template was generally 10:1. Both strands of the coding sequence were determined using synthetic oligonucleotides as primers (see *Results*). Primer sequences were confirmed by overlapping sequences from the complementary strand. One portion of the sequence was completed by subcloning in pUC9 (28) and by sequencing with the M13 reverse oligonucleotide primer. Amino acid sequences and codon frequencies were deduced from nucleotide sequences by the computer program of Conrad and Mount (29).

## RESULTS AND DISCUSSION

The amino-terminal sequences of most AP  $\beta$ -subunit polypeptides begin Met-Gln-Asp-Ala-Ile (2, 30–33). If the degenerate third base of the isoleucine codon is not included,

this sequence of amino acids can be encoded by 16 sense-strand (i.e., mRNA complementary) DNA sequences: 3' T-A-C-G-T-(T/C)-C-T-(A/G)-C-G-(A/G/T/C)-T-A 5'. Each of these sequences was separately synthesized. Initially, an equimolar mixture of all 16 sequences was labeled with [ $\gamma$ - $^{32}$ P]ATP and used as a hybridization probe against a Southern blot carrying various restriction digests of the cyanelle genome of *C. paradoxa*. This mixture was also used to probe cloned cyanelle DNA fragments from a partial *Pst* I library of the cyanelle genome that had been constructed in *E. coli* using pBR322 as the cloning vector. Both types of hybridization (data not shown) indicated that a 7.8-kbp *Pst* I fragment carried a sequence complementary to the hybridization probe mixture. Subsequent hybridization experiments with various combinations of the 16 oligonucleotides and individual oligonucleotides revealed that only the sequence 3' T-A-C-G-T-T-C-T-G-C-G-T-T-A 5' hybridized to this restriction fragment.

The complementary sequence of the hybridizing oligonucleotide, 5' A-T-G-C-A-A-G-A-C-C-G-T-A-T 3', was synthesized and both oligonucleotides were used as primers for the chain-termination (dideoxy) sequencing method using *Pst* I-digested plasmid pCpcPst7.8 as the template. The results of these sequencing experiments confirmed the presence of sequences that could encode both the  $\alpha$  and  $\beta$  subunits of AP on this restriction fragment of the cyanelle genome.

The coding region for the AP subunits of *C. paradoxa* was sequenced on both strands using chemically synthesized oligonucleotides with the strategy shown in Fig. 1. Based upon the initial sequences obtained, two new oligonucleotides were synthesized corresponding to nucleotides 991–1004 (5' G-C-A-A-T-G-C-T-T-G-C-A-G-G 3') and nucleotides 385–398 (3' G-G-A-C-C-A-C-C-A-T-T-G-C-G 5') (Fig. 2). Sequence results obtained with these oligonucleotides extended the single-strand sequence for both genes beyond the actual coding region for the  $\alpha$  and  $\beta$  subunits. Sequencing of the complementary strand for this region and appropriate overlaps were obtained by synthesizing the following oligonucleotides for use as sequencing primers: nucleotides 1181–1196 (3' G-A-C-C-A-A-A-T-C-C-A-A-T-T-A-A 5'), nucleotides 991–1004 (3' C-G-T-T-A-C-G-A-A-C-G-T-C-C 5'), and nucleotides 409–423 (5' G-A-A-A-T-G-A-C-T-G-C-T-A-C-T 3'). The complementary strand sequencing was completed by constructing a subclone in pUC9 extending from the *Bam*HI site in the insert in pCpcPst7.8 to the downstream

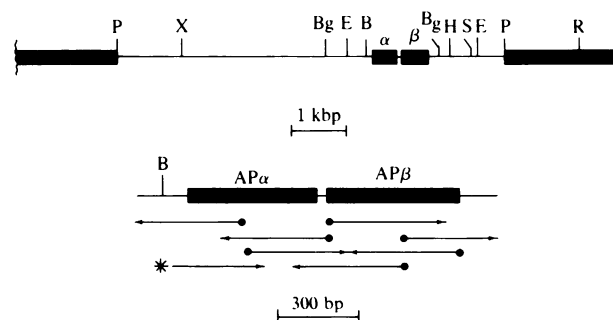


FIG. 1. Restriction map of plasmid pCpcPst7.8, which carries a 7.8-kbp *Pst* I fragment from the cyanelle genome of *C. paradoxa*, and nucleotide sequencing strategy for the AP coding region. Restriction enzymes: P, *Pst* I; X, *Xho* I; Bg, *Bgl* II; E, *Eco*RI; B, *Bam*HI; H, *Hind*III; S, *Sal* I; R, *Eco*RI site in pBR322 portion of plasmid. The arrows represent sequenced segments, with dots and the star representing 5' end points and arrowheads representing 3' end points for each strand. Dots indicate sequences obtained with synthetic oligonucleotides described in the text; the star indicates a sequence obtained using the M13 reverse sequencing primer and pCpcPst7.802 DNA as template. The thick lines indicate AP  $\alpha$ - and AP  $\beta$ -subunit coding regions.

```

TTAAGTAAGTTTTATAGTAAACAATTAATGGACTTATGAGTAAATACCTTTTACCTATAGGTATCCATTTGTCATAGAAAGTATTGACGGATCCAT 100
TCATATAATTTACTCATTATAGTATTATAGAAATGGTCCAAAGCCCAATACTTTTATCTTTTAAAAGAACCTATCTTATGAGTATCGTTACCAA 200
                                     MetSerLeuValThrLy
GTCTATTGTTAATGCTGATGCTGAAGCTCGTTACCTTAGCCAGCCGAGTTAGTGTATTAAAAGTTTGTGCGCTGGCGAAGACGTTTACGTTAT 300
sSERLeuValAsnAlaAspAlaGluAlaArgTyrLeuSerProGlyLeuLeuAspArgLeuLysSerPheAlaAlaSerGlyLeuArgArgLeuArgLeu
GCACAAATTTAAGTATAACCGTGAACGCATCGTTAGAGAAGCTGGTCAACAACCTTTTCCAAAACGCCCTGATATCGTTTCCCTGGTGGTAAACGCAT 400
ALAGLNLLeuThrAspAsnArgGluArgLeuValArgGluAlaGlyGlnGlnLeuPheGlnLysArgProAspLeuValSerProGlyLeuAsnAlaT
ATGGTGAAGAAATGACTGCTACTTGTTCAGTGACTAGATTATTATCTTCGTTTAGTAACCTTATGGTGTGTTGTTGCTGGTGTGCAACTCCAATTGAAGA 500
YRGLYGLUGLUMetThrAlaThrCysLeuArgAspLeuAspTyrTyrLeuArgLeuValThrTyrGlyValValAlaGlyAspAlaThrProLeuGluGlu
AATTGGTTAGTTGGTGTAAAGAAATGATAATCTTATAGTACTCCAGTAGCAGCTGTAGCAGAAGCGTTCGTTCCGCTAAGAGTGTAGCAACTGGT 600
ULeGlyLeuValGlyValLysGluMetTyrAsnSerLeuGlyThrProValAlaAlaValAlaGluGlyValArgSerAlaLysSerValAlaThrGly
TTACTTCCCGTGATGATGCTGCTGAAGCTGGCTTACTTCGATTACGTGATTGCTGCTTACAATAAAAATTTTGTATTTTAAACAACCTAAGGAA 700
LeuLeuSerGlyAspAspAlaAlaGluAlaGlySerTyrPheAspTyrValLeuAlaAlaLeuGlnEnd
CTGAACCTATGCAAGAGCTATTACTGCTGAATTAATGCTGCTGATGACAAGTAAATATCTGTACTGCTATCTGTAGAAAAATAAAAAGCTATTT 800
MetGlnAspAlaLeuThrAlaValLeuAsnAlaAlaAspValGlnGlyLysTyrLeuAspThrAlaSerValGluLysLeuLysSerTyrPhe
CCAACTGGTGAATTAAGAGTTCGTGACGCTGCAACTATTGCTGCTAATCTTCTGCAATCATTAAAGAGCTGTAGCTAATCCCTCTTTATTCGAT 900
eGlnThrGlyGluLeuArgValArgAlaAlaAlaThrLeuAlaAlaAsnSerSerAlaLeuLeuLysGluAlaValAlaLysSerLeuLeuTyrSerAsp
ATCACCCGCTCAGGTGGTAACTGACTACTCGCTGTTATGCTGCTGATTCGCTGACTAGATTACTATGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTG 1000
LeuThrArgProGlyGlyAsnMetTyrThrThrArgArgTyrAlaAlaCysLeuArgAspLeuAspTyrTyrValArgTyrAlaThrTyrAlaMetLeuA
CAGGTGACACATCTATTTAGATGAACGTGATTAATGGCTTAAAGAACTTATAACTCTTGGTGTACCTGAGTGTGCAACTATCCAAGCAATTC 1100
LAGLYAspThrSerLeuLeuAspGluArgValLeuAsnGlyLeuLysGluThrTyrAsnSerLeuGlyValProValGlyAlaThrLeuGlnAlaLeuGlu
AGCTGCTAAAGAAAGTAACTGCTGTTAGTAGTCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTG 1200
NAlaAlaLysGluValThrAlaGlyLeuValGlyProAspAlaGlyArgGluMetGlyLeuTyrTyrAspTyrLeuSerSerGlyLeuGlyEnd
AATCAAAACTTTTTTATAAGAATATATAAATCTACTCTCTTTTAAAACGAGGAGTAGATTTATATATATATATATATATATATATATATAAATA 1300
TATATATATATATATATATATATATATATATATATATATATATATATATATATATATATATATATATATATATATATATATATATATATATAT 1336

```

FIG. 2. Nucleotide sequence of the AP coding region from the cyanelle genome of *C. paradoxa*. The coding sequence for the AP  $\alpha$  subunit extends from nucleotide 184 to nucleotide 669 while the AP  $\beta$ -gene sequence extends from nucleotide 709 to nucleotide 1194. The predicted amino acid sequences of the two AP apoproteins are juxtaposed to the coding sequences. The underlined sequences upstream from AP  $\alpha$  resemble the consensus promoter sequences for *E. coli* and putative promoters for chloroplast-encoded *psbA* genes. A 56-bp palindrome, centered between nucleotides 1251 and 1252 and that might play a role in transcription termination, is also underlined.

*EcoRI* site in the insert (plasmid pCpPst7.802; see Fig. 1). The appropriate nucleotide sequence was obtained from this subclone by using the M13 reverse sequencing primer.

Two recent reviews, one describing developments for rapid DNA sequence analysis (34) and the other describing potential applications of oligonucleotides in molecular biology (35), offer little, if any, description of what we consider to be a major use of synthetic oligonucleotides—i.e., their use as specific primers for rapid, directed analysis of DNA sequences. This sequencing project was rapidly completed without a detailed restriction map of the region sequenced. The sequencing was largely, though not entirely, completed without subcloning. Since the oligonucleotide primers were chosen to provide overlapping sequences, overlapping subclones did not have to be generated nor sequenced. Finally, the method allows high-efficiency sequencing since the sequence data that can be obtained from each oligonucleotide primer are roughly constant (350–450 bp using multiple loadings on 5–6% acrylamide gels). After attempting to use random subcloning as a strategy to sequence the *lacZ* gene of *E. coli*, Kalnins et al. (36) concluded that random subcloning protocols would not be sufficient to complete most DNA sequencing studies. The results of this sequencing study suggest that a combination of directed subcloning and oligonucleotide-directed chain termination sequencing should provide a rapid approach to the problem of DNA sequence analysis.

The nucleotide sequence of the 1336-bp region of pCpPst7.8, which includes the coding sequences for the  $\alpha$ - and  $\beta$ -subunit apoproteins of *C. paradoxa* AP, is shown in Fig. 2. This region contains two long, open reading frames, which are indicated by the juxtaposed translations. The AP  $\alpha$  coding sequence extends from nucleotide 184 to nucleotide 669 while the AP  $\beta$  coding sequence extends from nucleotide 709 to nucleotide 1194. As shown in Fig. 3, each putative polypeptide is highly homologous to AP subunits from other organisms. The predicted masses of the AP subunit apoproteins are 17,277 and 17,281 daltons for AP  $\alpha$  and AP  $\beta$ , respectively.

The AP  $\alpha$ - and AP  $\beta$ -subunit genes are in the same orientation with AP  $\alpha$  located upstream from the AP  $\beta$  gene.

This organization is opposite to that found for the PC genes of *Synechococcus* 7002 (16), in which PC  $\beta$  is upstream from PC  $\alpha$ . The two coding sequences are separated by a 39 bp, A+T-rich sequence. Each coding sequence is preceded by a polypurine, Shine-Dalgarno-type sequence (5' A-A-A-G-A-A 3' for AP  $\alpha$  and 5' A-A-G-G-A-A 3' for AP  $\beta$ ) that ends 8 bp upstream from the initiator methionine codon and could play a role in translation initiation.

$\alpha$  SUBUNITS  
AP C. PARA. : M-SIVTKSIVNADAERYLSPGELDRKIFSAAGGERLRIADLTDNRERIVREAGGQLFQKRPDIVS+PGGNMAYGEE+  
AP C. CALD. : M-SIVTKSIVNADAERYLSPGELDRKIFSAAGGERLRIADLTDNRERIVREAGGQLFQKRPDIVS+PGGNMAYGEE+  
AP M. LAM. : M-SIVTKSIVNADAERYLSPGELDRKIFSAAGGERLRIADLTDNRERIVREAGGQLFQKRPDIVS+PGGNMAYGEE+  
PC SYNECHO. : KTPL EAVL SQG F NT QVLYGRLRQ ARA EA T AKADTL NG A AVVS F YTT T N FAADRG  
  
AP C. PARA. : TATCLRDLRYLRLVTVGVVAGDTPIEEIGLVGKEMVNSLGVTPVAVAEGRVSAKSAVA+-----TGLSGDDAA+  
AP C. CALD. : TATCLRDLRYLRLVTVGVVAGDTPIEEIGLVGKEMVNSLGVTPVAVAEGRVSAKSAVA+-----TGLSGDDAA+  
AP M. LAM. : TATCLRDLRYLRLVTVGVVAGDTPIEEIGLVGKEMVNSLGVTPVAVAEGRVSAKSAVA+-----TGLSGDDAA+  
PC SYNECHO. : KDK A IG M I CL GTG MD YLIA D INKTFDLS SWY EALRKHANH+ SSI AE T  
  
AP C. PARA. : EAGSYFDYVIAALQ  
AP C. CALD. : EAGSYFDYVIAALQ  
AP M. LAM. : EAGSYFDYVIAALQ  
PC SYNECHO. : TNN I A N S  
  
 $\beta$  SUBUNITS  
AP C. PARA. : MDAITAVINADVQKYLDTASVEKLSYFDTGELRVRAAATIRNNSATIKAEAVKSLY+SDITR+PGGNMAYTRR+  
AP C. CALD. : MDAITAVINADVQKYLDTASVEKLSYFDTGELRVRAAATIRNNSATIKAEAVKSLY+SDITR+PGGNMAYTRR+  
AP M. LAM. : MDAITAVINADVQKYLDTASVEKLSYFDTGELRVRAAATIRNNSATIKAEAVKSLY+SDITR+PGGNMAYTRR+  
PC SYNECHO. : A F IF R VSG AR EFISSDKL A KVVAE TK SD VSRMNA A S VTN AROLFADOPQL+IA N  
  
AP C. PARA. : YAACIRDLVYRYATYAMLGDTSLDERVNLGKETVNSLGV+PYGATIGTGAKEVIT+-----AGLV+GPDAG+R+  
AP C. CALD. : YAACIRDLVYRYATYAMLGDTSLDERVNLGKETVNSLGV+PYGATIGTGAKEVIT+-----AGLV+GPDAG+R+  
AP M. LAM. : YAACIRDLVYRYATYAMLGDTSLDERVNLGKETVNSLGV+PYGATIGTGAKEVIT+-----AGLV+GPDAG+R+  
PC SYNECHO. : M L MEIL V TFT A V ND C R VA GASVAAGR V MGKAAVAVMVD VTS DCSLLO  
  
AP C. PARA. : EMGIYDYVSSGLG  
AP C. CALD. : EMGIYDYVSSGLG  
AP M. LAM. : EMGIYDYVSSGLG  
PC SYNECHO. : TEL FETAAKAVE

FIG. 3. Comparison of AP  $\alpha$ - and AP  $\beta$ -subunit amino acid sequences. Organisms and references for sequence data: *C. para.* (*C. paradoxa*) AP  $\alpha$  and  $\beta$  sequences deduced from nucleotide sequence in this paper; *C. cald.* (*C. caldarium*) AP  $\alpha$  and AP  $\beta$  (32); *M. lam.* (*Mastigocladus laminosus* = *Fischerella* PCC 7603) AP  $\alpha$  and AP  $\beta$  (31); *Anabaena* (*Anabaena variabilis* = *Anabaena* PCC 7118) AP  $\beta$  (33). Also included are the PC  $\alpha$  and PC  $\beta$  sequences for *Synechococcus* PCC 7002 (*Synecho.*; equivalent to *Agmenellum quadruplicatum* strain PR-6; ref. 16). Blank positions indicate identity with the *C. paradoxa* sequence at that position. Stars indicate gaps inserted to maximize homology. The single-letter notation for amino acids is as follows: A, alanine; C, cysteine; D, aspartic acid; E, glutamic acid; F, phenylalanine; G, glycine; H, histidine; I, isoleucine; K, lysine; L, leucine; M, methionine; N, asparagine; P, proline; Q, glutamine; R, arginine; S, serine; T, threonine; V, valine; W, tryptophan; Y, tyrosine.

An examination of the 5' flanking sequences for the AP  $\alpha$ -subunit coding sequence reveals sequences that closely resemble the consensus *E. coli* promoter sequences (37). These sequences also exhibit remarkable similarity to the proposed promoter sequences for several chloroplast-encoded *psbA* genes (38–41). These sequences are indicated by underlining in Fig. 2. The sequence 5' T-A-T-T-G-A-C-G 3' occurs at nucleotides 86–93 and is similar to the –35 consensus sequence 5' t-g-T-T-G-A-C-A 3' for *E. coli* (37). Downstream from this sequence are several potential –10-like sequences. The sequences 5' T-A-T-A-A-T 3', located between nucleotides 104–109, and 5' T-T-A-T-A-G-T 3', located between nucleotides 118–124, most closely match the *E. coli* –10 (Pribnow box) consensus sequence 5' t-T-A-t-a-a-T 3' (37). The sequences that most closely match the *E. coli* –35 and –10 consensus sequences are separated by 25 bp, which falls outside the range observed for known *E. coli* promoters (37). Resolution of this problem will have to await mapping of the 5' end of the mRNA transcript.

Link (39) has noted the presence of a eukaryotic "TATA-like" sequence motif located between the –35-like and –10-like elements of most chloroplast protein genes, but not for chloroplast tRNA and rRNA genes. The consensus sequence of this element for *psbA* genes from the chloroplasts of several higher plants is 5' (C/T)-A-T-A-T-A-A 3'; this closely resembles the consensus promoter sequence for genes transcribed by RNA polymerase II. Strikingly, this sequence element, 5' C-A-T-A-T-A-A 3' (nucleotides 102–108; underlined in Fig. 2), also occurs between the –35-like sequence and one of the –10-like sequences located upstream from the AP  $\alpha$ -subunit gene. Link (39) has shown in an *in vitro* transcription system that, in the absence of the –35-like sequence, a region containing this TATA-like element still allows a basal level of specific transcription of the mustard *psbA* gene. He has suggested that the light-regulated expression of the *psbA* gene might involve a "promoter-switch" involving these different elements. Whether the homologous sequences observed upstream from AP  $\alpha$  could play a similar role in *C. paradoxa* is unknown.

Downstream from the AP  $\beta$ -subunit gene is a 56-bp palindrome centered between nucleotides 1251 and 1252 (see Fig. 2). The free energy of formation for a hairpin structure from this sequence would be about –36.6 kcal/mol (1 cal = 4.18 J) (42). A hairpin structure formed as a result of this sequence could play a role in either transcription or translation termination. Similar palindromic sequences have been noted downstream from several chloroplast genes (38) and are a common feature upstream from transcription termination sites in prokaryotes (37).

The restriction map of the 7.8-kbp *Pst* I fragment that carries the AP genes of *C. paradoxa* (Fig. 1) has been compared to the restriction map of the cyanelle genome of *C. paradoxa* reported by Kuntz *et al.* (43). The restriction pattern of the 7.8-kbp fragment matches exactly the pattern observed at approximately three o'clock in the restriction map published by Kuntz *et al.* (43), a result that we have confirmed through mapping experiments with our isolate of *C. paradoxa* (unpublished observations). This places the AP genes in the large single-copy region of the cyanelle genome and approximately 30 kbp distant from the PC  $\beta$ -subunit gene, which has been mapped near the center of the small single-copy region of the cyanelle genome (15).

The deduced amino acid sequences of the *C. paradoxa* AP  $\alpha$ - and AP  $\beta$ -subunit polypeptides have been compared in Fig. 3 with the complete amino acid sequences for the  $\alpha$  and  $\beta$  subunits of AP from the cyanobacterium *Mastigocladus laminosus* (31), to the  $\alpha$  and  $\beta$  subunits of AP from the red alga *Cyanidium caldarium* (32), to the  $\beta$  subunit of AP from the cyanobacterium *Anabaena variabilis* (33), and to the  $\alpha$  and  $\beta$  subunits of PC from the cyanobacterium *Synechococcus* 7002

(16). The sequences have been aligned for maximal homology based on comparisons of both the amino acid sequences of the apoproteins and nucleotide sequences of the genes. The pairs of AP  $\alpha$ -subunit sequences have an average amino acid homology of  $81 \pm 3\%$  while the pairs of AP  $\beta$  subunits have an average amino acid homology of  $85 \pm 0.6\%$ . When aligned for maximal homology, the two AP apoprotein subunits from *C. paradoxa* have amino acid sequences that are 40% homologous to one another.

The predicted amino acid sequence of the AP  $\beta$  subunit is highly homologous at both its NH<sub>2</sub> and COOH termini and is identical in length (161 amino acids) to other AP  $\beta$  subunits whose amino acid sequences have been determined (2, 30–33). This suggests that proteolytic processing of the AP  $\beta$  subunit does not occur *in vivo*. However, the NH<sub>2</sub> termini of most AP  $\alpha$  subunits for which amino acid sequences have been determined begin with the sequence Ser-Ile-Val-Thr-Lys ... (2, 30–32), while the NH<sub>2</sub> terminus of the *C. paradoxa* AP  $\alpha$  subunit is predicted to be Met-Ser-Ile-Val-Thr-Lys .... Whether the methionine is proteolytically removed from the *C. paradoxa* AP  $\alpha$  polypeptide is not known, since amino acid sequencing of this AP subunit has not been performed.

We have compared the coding sequences for the AP genes from *C. paradoxa* and the PC genes from *Synechococcus* 7002 (16). The AP  $\alpha$ /AP  $\beta$  coding sequences represent the most homologous pairing while the PC  $\alpha$ /AP  $\beta$  represent the least. The corrected percent divergences of replacement and silent substitutions were calculated for the best aligned sequences of these two pairings by the method of Perler *et al.* (44). The AP  $\alpha$ /AP  $\beta$  pairing gives corrected divergence values of 78% and 58% for silent and replacement mutations, respectively. These values are roughly comparable with those derived from comparison of the  $\alpha$ - and  $\beta$ -globin gene coding sequences of higher eukaryotes (45). The PC  $\alpha$ /AP  $\beta$  pairing gives values of 139% and 84% for silent and replacement mutations, respectively. Although the higher value for replacement mutations is probably an accurate reflection of the divergence between AP and PC, the high value for silent mutations in this latter comparison may also be partially a reflection of the large differences in G+C content (12, 46) and codon usage for the two organisms. The currently available nucleotide sequence data for biliprotein genes are much too limited to allow definitive statements concerning the evolution of this gene family.

A question of interest to those who have studied phycobiliprotein primary structures and the evolution of this protein family concerns the origin of the second chromophore-binding site on the PC  $\beta$  subunit. The single chromophore-binding sites of the AP  $\alpha$ , AP  $\beta$ , and PC  $\alpha$  subunits are clearly homologous to one of the chromophore-binding sites of the PC  $\beta$  subunit (see data compiled in ref. 2). It has been suggested that an insertion event centered about the COOH-terminal chromophore-binding cysteine added about 10 amino acids to the PC  $\beta$  ancestral gene and thereby generated a second chromophore-binding site (31, 32).

On the basis of the available nucleotide sequence data we propose a more detailed scheme that requires several events for which the temporal sequence cannot presently be determined. The 3' termini of the *Synechococcus* 7002 PC genes and the *C. paradoxa* AP genes have been aligned to allow maximal homology (Fig. 4). Inspection of the PC  $\beta$ -subunit gene reveals two regions that are not homologous to sequences occurring in the other genes. These regions flank the second chromophore-binding site. We propose that the PC  $\beta$  gene increased in length via unequal sister chromatid exchange (47) or replication fork slippage ("looping out"; ref. 48). These events may possibly have involved the three 9-bp direct repeats that occur in the PC  $\beta$  gene (Fig. 4). Either type of event could have effectively duplicated a region coding for 9–10 amino acids in the region NH<sub>2</sub> terminal to the chro-

```

AP α: CCA GTA GCA GCT GTA GCA GAA GGC GTT CGT TCC GCT ANG AGT GTA GCA --- --- --- --- --- ACT GGT
AP β: CCT GTA GGT GCA ACT ATC CAA GCA ATT CAA GCT GCT AAA GAA GTA ACT --- --- --- --- --- GCT GGT
PC α: CCC AGC TGG TAT GTT GAA GCT CTC AAG CAC ATC AAA GCA AAC CAT --- --- --- --- --- GGT
PC β: CCC GGT GCT TCC GTT GCT GCT GGT GTA CGT GCA ATG GGT AAA GCT GCT GTA GCG ATT GTT ATG GAT CCC GCT GCT

AP α: TTA CTT TCC GGT GA T GAT GCT GCT --- --- --- GAA GCT GGC TCT TAC TTC GAT TAC GTG ATT GCT GCT TTA CAA TAA
AP β: TTA GTA --- GGT CC T GAT GCT GGT --- --A GA- GAA ATG GGT ATT TAC TAC GAC TAC ATT TCT TCT GGT TTA GGT TAA
PC α: TTG ACT --- GGC GA T GCT GCT ACT --- --- --- GAA ACT AAC AAC TAC ATC GAC TAC GCA ATT AAC GCC CTC AGC TAA
PC β: GTA ACT TCC GGT GACT GCA GCT CT CTC CAA CAG GAA ATC GAA CTC TAC TTC GAA ACT GCT GCA AAA GCT GTT GAA TAA

```

FIG. 4. Comparison of the 3' termini of the coding sequences for *C. paradoxa* AP genes and the *Synechococcus* 7002 PC genes (16). Dashes indicate gaps inserted to maximize homology. The three 9-bp direct repeats that may have played a role in PC  $\beta$ -subunit elongation (see *Results and Discussion*) are underlined. The proposed insertion and deletion events that could have shifted the reading frame to create a codon (TGC) for a chromophore-binding cysteine in the PC  $\beta$  gene are indicated by the symbols  $\uparrow$  and  $\downarrow$ , respectively.

mophore-binding site near the COOH terminus of the PC  $\beta$  polypeptide. In support of the duplication hypothesis for the increased length of the PC  $\beta$  gene, we have previously noted that there is a significant degree of nucleotide homology between the two segments bounded by these three direct repeats (16).

A secondary event appears to have been an "C" insertion mutation (see Fig. 4) that retained an aspartic acid residue (GAT  $\rightarrow$  GAC) but shifted the reading frame so as to produce a cysteine codon (TGC) where alanine (GCN) had previously occurred. An apparent deletion 8 bp downstream from the "C" insertion returns the reading frame to that originally used. The origin of the 6–9 bp insertion/deletion that occurs 3' to these events (COOH terminal to the second chromophore-binding cysteine) is uncertain. Events occurring in this region may have been required to optimize the newly acquired chromophore-binding site.

With the inclusion of previous work (16), we have now cloned and sequenced the genes coding for both subunits of examples of two of the three major spectroscopic classes of phycobiliproteins. We hope to obtain a better understanding of the evolution of the phycobiliprotein gene family as additional nucleotide sequence data become available. The cloned AP and PC genes will provide hybridization probes for studying the expression of these genes, which is known to vary in response to light intensity, light wavelength (complementary chromatic adaptation), and nutrient availability.

This work was supported by U.S. Public Health Service Grant GM 31625 and U.S. Department of Agriculture Grant 82-CRRC-1-1080.

- Trench, R. (1982) in *Progress in Phycological Research*, eds. Round, F. & Chapman, D. J. (Elsevier, Amsterdam), Vol. 1, pp. 257–288.
- Glazer, A. N. (1984) *Biochim. Biophys. Acta* **768**, 29–51.
- Lundell, D. J., Glazer, A. N., DeLange, R. J. & Brown, D. M. (1984) *J. Biol. Chem.* **259**, 5472–5480.
- Scheer, H. (1981) *Angew. Chem. Int. Ed. Engl.* **20**, 241–261.
- Glazer, A. N. (1982) *Annu. Rev. Microbiol.* **36**, 173–198.
- Cohen-Bazire, G. & Bryant, D. A. (1983) in *The Biology of the Cyanobacteria*, eds. Carr, N. G. & Whitton, B. A. (Blackwell, Boston), pp. 143–190.
- Gantt, E. (1980) *Int. Rev. Cytol.* **66**, 45–80.
- Gantt, E. (1981) *Annu. Rev. Plant Physiol.* **32**, 327–347.
- Redlinger, T. & Gantt, E. (1982) *Plant Physiol.* **79**, 5542–5546.
- Bohnert, H. J., Crouse, E. J., Pouyet, J., Mucke, H. & Löffelhardt, W. (1982) *Eur. J. Biochem.* **126**, 381–388.
- Herdman, M., Janvier, M., Rippka, R. & Stanier, R. Y. (1979) *J. Gen. Microbiol.* **111**, 73–85.
- Herdman, M. & Stanier, R. Y. (1977) *FEMS Microbiol. Lett.* **1**, 7–11.
- Edelman, M. (1981) in *The Biochemistry of Plants*, ed. Marcus, A. (Academic, New York), Vol. 6, pp. 249–301.
- Grossman, A., Talbott, L. & Egelhoff, T. (1983) *Carnegie Inst. Washington Yearb.* **82**, 112–116.
- Lemaux, P. G. & Grossman, A. (1984) *Proc. Natl. Acad. Sci. USA* **81**, 4100–4104.
- de Lorimier, R., Bryant, D. A., Porter, R. D., Liu, W.-Y., Jay, E. & Stevens, S. E., Jr. (1984) *Proc. Natl. Acad. Sci. USA* **81**, 7946–7950.
- Allen, M. M. (1968) *J. Phycol.* **4**, 1–4.
- Heinhorst, S. & Shively, J. M. (1983) *Nature (London)* **304**, 373–374.
- Maniatis, T., Fritsch, E. F. & Sambrook, J. (1982) *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY), p. 92.
- Godson, G. N. & Vapnek, D. (1973) *Biochim. Biophys. Acta* **299**, 516–520.
- Buzby, J. S., Porter, R. D. & Stevens, S. E., Jr. (1983) *J. Bacteriol.* **154**, 1446–1450.
- Jay, E., Macknight, D., Lutze-Wallace, C., Harrison, D., Wishart, P., Liu, W.-Y., Asundi, V., Pomeroy-Cloney, L., Rommens, J., Eglington, L., Pawlak, J. & Jay, F. (1984) *J. Biol. Chem.* **259**, 6311–6317.
- Beaucage, S. L. & Caruthers, M. H. (1981) *Tetrahedron Lett.* **22**, 1859–1862.
- Matteucci, M. D. & Caruthers, M. H. (1981) *J. Am. Chem. Soc.* **103**, 3185–3191.
- Hanahan, D. & Meselson, M. (1983) *Methods Enzymol.* **100**, 333–342.
- Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 5463–5467.
- Heidecker, G., Messing, J. & Gronenborn, B. (1980) *Gene* **10**, 69–73.
- Vieira, J. & Messing, J. (1982) *Gene* **19**, 259–268.
- Conrad, B. & Mount, D. W. (1982) *Nucleic Acids Res.* **10**, 31–38.
- Glazer, A. N., Apell, G. S., Hixson, C. S., Bryant, D. A., Rimon, S. & Brown, D. M. (1976) *Proc. Natl. Acad. Sci. USA* **73**, 428–431.
- Sidler, W., Gysi, J., Isker, E. & Zuber, H. (1981) *Hoppe-Seyler's Z. Physiol. Chem.* **362**, 611–628.
- Offner, G. D. & Troxler, R. F. (1983) *J. Biol. Chem.* **258**, 9931–9940.
- DeLange, R. J., Williams, L.-C. & Glazer, A. N. (1981) *J. Biol. Chem.* **256**, 9558–9566.
- Deininger, P. L. (1983) *Anal. Biochem.* **135**, 247–263.
- Itakura, K., Rossi, J. J. & Wallace, R. B. (1984) *Annu. Rev. Biochem.* **53**, 323–356.
- Kalnins, A., Otto, K., Rütger, U. & Müller-Hill, B. (1983) *EMBO J.* **2**, 593–597.
- Rosenberg, M. & Court, D. (1979) *Annu. Rev. Genet.* **13**, 319–353.
- Whitfield, P. R. & Bottomley, W. (1983) *Annu. Rev. Plant Physiol.* **34**, 279–310.
- Link, G. (1984) *EMBO J.* **3**, 1697–1704.
- Spielmann, A. & Stutz, E. (1983) *Nucleic Acids Res.* **11**, 7157–7167.
- Sugita, M. & Sugiura, M. (1984) *Mol. Gen. Genet.* **195**, 308–313.
- Tinoco, J., Borer, P. N., Dengler, B., Levine, M. D., Uhlenbeck, O. C., Crothers, D. M. & Gralla, J. (1973) *Nature (London) New Biol.* **246**, 40–41.
- Kuntz, M., Crouse, E. J., Mubumbila, M., Burkard, G., Weil, J.-H., Bohnert, H. J., Mucke, H. & Löffelhardt, W. (1984) *Mol. Gen. Genet.* **194**, 508–512.
- Perler, F., Efstratiadis, A., Lomedico, P., Gilbert, W., Kolodner, R. & Dodgson, J. (1980) *Cell* **20**, 555–566.
- Efstratiadis, A., Posakony, J. W., Maniatis, T., Lawn, R. M., O'Connell, C., Spritz, R. A., DeRiel, J. K., Forget, B. G., Weissman, S. M., Slightom, J. L., Blechl, A. E., Smithies, O., Baralle, F. E., Shoulders, C. C. & Proudfoot, N. J. (1980) *Cell* **21**, 653–668.
- Herdman, M., Janvier, M., Waterbury, J. B., Rippka, R., Stanier, R. Y. & Mandel, M. (1979) *J. Gen. Microbiol.* **111**, 63–71.
- Anderson, R. P. & Roth, J. R. (1977) *Annu. Rev. Microbiol.* **31**, 473–505.
- Streisinger, G., Okada, Y., Emrich, J., Newton, J., Tsugita, A., Terzaghi, E. & Inouye, M. (1966) *Cold Spring Harbor Symp. Quant. Biol.* **31**, 77–84.