Proc. Natl. Acad. Sci. USA
Vol. 81, pp. 7946–7950, December 1984
Microbiology

Genes for the α and β subunits of phycocyanin

(cyanobacteria/*Agmenellum quadruplicatum*/photosynthesis/phycobilisome/nucleotide sequence)

R. de Lorimier*, D. A. Bryant*, R. D. Porter*, W.-Y. Liu†, E. Jay†, and S. E. Stevens, Jr.*

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

Communicated by R. H. Burris, July 27, 1984

ABSTRACT **Phycocyanin (PC) is a light-harvesting protein common to blue-green and red algae. We have isolated the genes for the two apoprotein subunits, α and β, of PC from the blue-green alga *Agmenellum quadruplicatum* PR-6. We synthesized eight sense-strand tetradecameric oligonucleotide probes that could encode a particular pentapeptide in PCα from *A. quadruplicatum*. Only one probe hybridized with total RNA from this organism. This oligonucleotide showed homology to a unique restriction fragment when used to probe Southern blots of *A. quadruplicatum* DNA. The probe-homologous 3.1-kilobase pair *Hin*dIII fragment was cloned. The nucleotide sequence of a 1.7-kilobase pair segment of this clone was determined. Two open reading frames are contained in this region, which correspond in deduced amino acid sequence to PCα and PCβ subunits. Both coding sequences are in the same orientation, separated by 105 base pairs, with the PCβ gene 5′ to the PCα gene. Each gene has a Shine–Dalgarno-type sequence near the initiation codon. Codon frequencies in the two genes may be correlated with the abundance of their products. The deduced amino acid sequences of the gene products show considerable homology with the α and β PC subunits from other species.**

In most photosynthetic organisms the major light-harvesting pigments are cyclic tetrapyrroles bound noncovalently to polypeptides intrinsic to the photosynthetic membrane. Blue-green algae (cyanobacteria) and red algae follow this pattern but, in addition, have large amounts of antenna pigments in the form of linear tetrapyrroles covalently bound to water-soluble polypeptides. These proteins, termed phycobiliproteins, occur as three major types: phycoerythrin, phycocyanin (PC), and allophycocyanin (1). Each phycobiliprotein consists of two subunits, α and β, which contain characteristic numbers and types of chromophores. Phycobiliproteins aggregate to form complexes called phycobilisomes (1). The assembly of phycobilisomes is mediated by nonpigmented linker polypeptides. These linkers also alter the spectral properties of phycobiliproteins so as to ensure an efficient transfer of absorbed light energy to the membrane (2). Phycobilisomes, in turn, are found attached to the outer surface of the thylakoid membrane, perhaps in association with photosystem II (3).

Our aim is to describe the structure of the phycobilisome and the regulation of the genes encoding its components. These studies may also shed light on the evolution of phycobiliprotein genes. As a first step, we have cloned and sequenced the genes encoding the α and β subunit apoproteins of PC from the blue-green alga *Agmenellum quadruplicatum*. A preliminary report of these results has been presented (4).

MATERIALS AND METHODS

DNA purification. *A. quadruplicatum* strain PR-6 was cultivated axenically in medium A with $NaNO_3$ (1 mg/ml) as described (5). Cells were harvested before reaching a density of 5×10^7 cells per ml, washed in 10% sucrose/50 mM Tris·HCl, pH 8.0/100 mM Na_2EDTA, and stored at −80°C. Lysis was achieved by thawing, adding egg-white lysozyme to a final concentration of 10 mg/ml, incubating at 37°C for 30 min, and adding *N*-lauroyl sarcosine (10% wt/vol stock solution) to a final concentration of 1%. An equal amount (wt/vol) of CsCl was added and DNA was purified by buoyant-density centrifugation. DNA-containing fractions were recentrifuged in the presence of ethidium bromide at 150 μg/ml. The dye was removed by *n*-butanol extraction and the DNA was dialyzed against 10 mM Tris·HCl, pH 8.0/1 mM Na_2EDTA.

RNA Purification. An exponentially growing culture of *A. quadruplicatum* was harvested, resuspended in the original volume of fresh medium A lacking nitrate, and incubated with aeration and illumination as before. The A_{620}/A_{680} ratio decreased from 0.9 to 0.3 within 24 hr, whereupon the cells were harvested, resuspended in medium A with nitrate (1 mg/ml) and incubated as before. After 8 hr, cells were harvested by centrifugation and resuspended in 10 mM Tris·HCl, pH 8.0/1 mM Na_2EDTA. Cells were lysed by passage through a French pressure cell at 20,000 psi (1 psi = 6.89 kilopascals), and RNA was isolated as described (6), except that the lysate was underlaid with saturated aqueous CsCl before centrifugation.

Synthesis of Oligonucleotides. Each tetradecamer (described in *Results and Discussion*) was separately synthesized on a silica support using a rapid, solid-phase phosphite method in a custom-built, automated synthesizer (7). Products were purified by HPLC as described (7). Probe oligonucleotides were labeled at the 5′ end, using T4 polynucleotide kinase and [γ-^{32}P]ATP.

Hybridization of Oligonucleotides. RNA: Equal aliquots (10–25 μg) of RNA were mixed with 2 pmol of each labeled oligonucleotide in 50 μl of 0.3 M NaCl/0.03 M sodium citrate, pH 7.0, heated to 65°C for 20 min, and incubated at 33°C for 2 hr. RNA·oligonucleotide hybrids were detected by fractionation on nondenaturing polyacrylamide gels (linear gradient of acrylamide, 2.5–10%) according to Nobrega *et al.* (8) except that gels were kept at 10°C throughout electrophoresis. DNA: Southern blots and colony filters were probed with oligonucleotides according to the method of Hanahan and Meselson (9). The hybridization temperature was 37°C. Washes were at room temperature in 0.9 M NaCl/16 mM Na_2EDTA/0.18 M Tris·HCl, pH 8.0.

Nucleotide Sequence Determination. Restriction fragments to be sequenced were cloned in pUC8 or pUC9 (10). Se-

Abbreviation: PC, phycocyanin.

quencing by the chain-termination method (11) was carried out according to Heidecker *et al.* (12). Both strands of each fragment were sequenced by using M13 universal and reverse oligonucleotide primers. All restriction-fragment junctions were verified by obtaining overlapping sequences. Amino acid sequences and codon frequencies were deduced from nucleotide sequences by the computer program of Conrad and Mount (13).

RESULTS AND DISCUSSION

A partial amino acid sequence of PCα from *A. quadruplicatum* is known (14). A set of oligonucleotide probes for the *PCα* gene was synthesized based on amino acid residues 107–111 (Met-Asp-Glu-Tyr-Leu). This pentapeptide could be encoded by 16 combinations of the following sense-strand (i.e., mRNA-complementary) sequence: 3′ T-A-C-C-T-(A/G)-C-T-(T/C)-A-T-(A/G)-(G/A)-A 5′. The degenerate third base of the leucine codon was not included. Since successful results were obtained with the set of probes based on the CUN codons for leucine, we did not test the set based on the UU(A/G) codons of leucine.

Oligonucleotides were individually screened for complementarity to *A. quadruplicatum* RNA purified from cells that were recovering from nitrogen starvation. Our rationale was that PC mRNA might comprise a larger fraction of total RNA from these cells than from those grown steadily in sufficient nitrogen. Blue-green algae, including *A. quadruplicatum*, cease to synthesize PC when assimilable nitrogen is limiting (15, 16). Upon replenishment of this nutrient, PC reappears at a higher rate than does chlorophyll or dry weight (16). Hence, PC mRNA might be proportionally enriched.

The eight tetradecameric probes were separately hybridized in solution to total RNA from cells thus treated. The mixture was then fractionated by gel electrophoresis, and hybridized products were detected by autoradiography. Fig. 1*A* shows an autoradiogram of two lanes of such a gel. Probe 6 showed much stronger hybridization to RNA than did probe 4. The other six probes showed hybridization no stronger than that of probe 4 (results not shown). Probe 6, 3′ T-A-C-C-T-A-C-T-C-A-T-G-G-A 5′, was apparently the most complementary to a sequence represented in the total RNA and, therefore, most likely to be a PCα-specific probe. Although the molecular weight of the complementary RNA species cannot be accurately estimated from this type of gel, the major and largest hybridizing species migrated slightly more slowly than 16S rRNA.

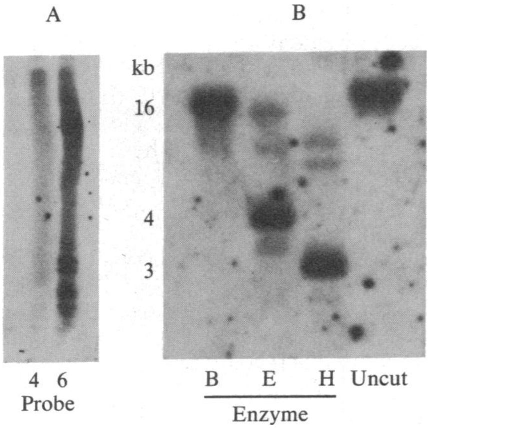

FIG. 1. (*A*) Autoradiogram of electophoretically resolved *A. quadruplicatum* PR-6 total RNA hybridized with two ^{32}P-labeled PCα probes, synthetic oligonucleotides 4 and 6. (*B*) Southern blot of *A. quadruplicatum* PR-6 DNA probed with PCα oligonucleotide probe 6. Lanes B, E, and H: *Bam*HI, *Eco*RI, and *Hin*dIII-restriction digests, respectively. Undigested (uncut) DNA is in the lane at the right. The sizes of the probe-homologous fragments, in kilobase pairs (kb), are shown on the left.

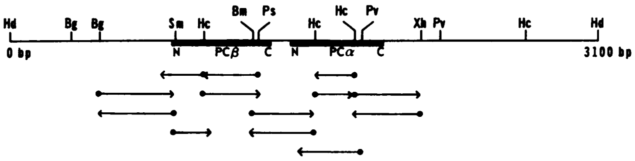

FIG. 2. Restriction map of *A. quadruplicatum* PR-6 *Hin*dIII fragment from pAQPR1, and nucleotide-sequencing strategy for PC coding region. Restriction enzymes: Bg, *Bgl* II; Bm, *Bam*HI; Hc, *Hin*cII; Hd, *Hin*dIII; Ps, *Pst* I; Pv, *Pvu* II; Sm, *Sma* I; Xh, Xho I. The arrows represent sequenced segments, with dots at 5′ endpoints and arrowheads at 3′ endpoints for each strand. The thick lines show PCα and PCβ coding regions, with NH_2 and COOH termini (N and C, respectively) indicated. bp, Base pairs.

Oligonucleotide 6 was used to probe a Southern blot of *A. quadruplicatum* total DNA. Single *Bam*HI, *Eco*RI, and *Hin*dIII fragments showed strong hybridization (Fig. 1*B*), suggesting that a unique sequence is homologous to this probe. Also shown in Fig. 1*B* is a lane of undigested DNA, wherein the probe hybridized at a location that corresponds to the limiting mobility of linear DNA. This is evidence that the observed hybridization was not due to small plasmids found in this species (17), the mobilities of which do not coincide with high molecular weight linear DNA.

To clone the probe-homologous sequence, *A. quadruplicatum* DNA was digested with *Hin*dIII and fractionated by agarose gel electrophoresis. A fraction in the range 2.3–4.0 kilobase pairs was eluted from the gel and ligated to *Hin*dIII-digested pBR325 (18). Ligation products were used to transform *Escherichia coli* strain RDP145 (19). Chloramphenicol-resistant transformants (4×10^3) were screened for homology to probe 6 by colony hybridization. Six positive clones were detected and analyzed by restriction endonuclease digestion. All six clones contained the same *A. quadruplicatum* DNA fragment, and both orientations of this insert were found. A restriction map of the insert from one clone, pAQPR1, is shown in Fig. 2. Southern blots of pAQPR1 restriction digests were hybridized with probe 6. The results (not shown) fixed the homology to the smallest *Hin*cII fragment (214 base pairs) of the insert. A partial nucleotide sequence of this region was obtained by using oligonucleotide 6 as a primer for the chain-termination sequencing method. An amino acid sequence derived from these data corresponded to a PCα segment NH_2-terminal to the pentapeptide on which the probe sequence was based.

Maxicell experiments (20) showed that the *Hin*dIII fragment insert of pAQPR1 directed the synthesis of PCα and PCβ apoproteins in *E. coli* (unpublished results). Similar studies on deletions of this clone showed that the PCβ coding sequence was located upstream from that for PCα. To complete the characterization of these two coding regions, a segment of the pAQPR1 insert from the inward *Bgl* II site to the *Xho* I site was sequenced. Fig. 2 shows the sequencing strategy and Fig. 3 shows the nucleotide sequence.

Only two long open reading frames are contained in this sequence. These are shown in Fig. 3, juxtaposed with the deduced amino acid sequences. The coding sequence for PCβ is nucleotides 382–900, and that for PCα is nucleotides 1006–1494. Each putative polypeptide is homologous to PC subunits from other organisms (Fig. 4). The predicted molecular masses of the PC subunits are 17,620 and 18,335 daltons for PCα and PCβ, respectively. These are reasonably close to the molecular masses measured by Gardner *et al.* (26) for PCα (16.0 kDa) and PCβ (18.5 kDa) from this organism. The sequence of the 15 NH_2-terminal amino acid residues of *A. quadruplicatum* PCβ has been determined (26) and exactly matches that deduced from the nucleotide sequence. The

```
AGATCTTTTTACAAGATGTAATGTTTAAATGCCGGCAGACGTTGTATAACATTTACCTAAGATTAAGAGTCACTCGCAGTACTCCTTAGAAACCCCATAG   100

GTTCCAAGGAACTAGCATGAACTTTATCTGGCAACTTTAAGAATCTGAGAAATTCAATGAATGTAAAGTTTCTTAAATGCCAAGGTGAAAAACAAGCAAA   200

AATAGCTGACACTCTTAATTGGCTTTGGGGATTAAGTTTCCAACTCGAAAACAAAACCTTTTATCGACTCTAGGATTTTGTTTTCAGCAAGAGAGCCCCT   300

CAGCACTTGCTTCACTCTTGTTAGTAAGCAAACCGCACAAAATAAATCCCACTCATCAAAATATAAGTAGGAGATAAAAACATGTTTGATATTTTTACCC   400
                                                                                 MetPheAspIlePheThrA

GGGTTGTTTCTCAGGCTGATGCCCGAGGTGAGTTCATTTCTAGCGACAAGCTCGAAGCTCTCAAAAAAGTTGTTGCCGAAGGTACCAAGCGTTCTGATGC   500
rgValValSerGlnAlaAspAlaArgGlyGluPheIleSerSerAspLysLeuGluAlaLeuLysLysValValAlaGluGlyThrLysArgSerAspAl

CGTAAGCCGCATGACCAACAATGCGTCTTCCATCGTTACTAACGCTGCTCGTCAACTCTTCGCTGACCAGCCCCAACTCATCGCTCCCGGTGGAAATGCT   600
aValSerArgMetThrAsnAsnAlaSerSerIleValThrAsnAlaAlaArgGlnLeuPheAlaAspGlnProGlnLeuIleAlaProGlyGlyAsnAla

TACACCAACCGTCGCATGGCTGCTTGTCTTCGCGACATGGAAATCATCCTCCGTTATGTAACCTACGCAACCTTCACTGGTGATGCGTCTGTACTCAACG   700
TyrThrAsnArgArgMetAlaAlaCysLeuArgAspMetGluIleIleLeuArgTyrValThrTyrAlaThrPheThrGlyAspAlaSerValLeuAsnA

ACCGCTGCCTCAATGGCCTCCGTGAAACCTACGTTGCGCTTGGTGTTCCCGGTGCTTCCGTTGCTGCTGGTGTACGTGCAATGGGTAAAGCTGCTGTAGC   800
spArgCysLeuAsnGlyLeuArgGluThrTyrValAlaLeuGlyValProGlyAlaSerValAlaAlaGlyValArgAlaMetGlyLysAlaAlaValAl

GATTGTTATGGATCCCGCTGGTGTAACTTCCGGTGACTGCAGCTCTCTCCAACAGGAAATCGAACTCTACTTCGAAACTGCTGCAAAAGCTGTTGAATAA   900
aIleValMetAspProAlaGlyValThrSerGlyAspCysSerSerLeuGlnGlnGluIleGluLeuTyrPheGluThrAlaAlaLysAlaValGluEnd

TCTTTTTTAATTCAACTCTGACATTTTTCGTTTTAAGTCTTACCGATACCGTAAGACGCTCTTTTAAAGAAAATTATTGATAATCCATAGGGAGATAATC  1000

TGACAATGAAAACCCCTCTTACCGAAGCAGTAGCACTCGCTGATTCTCAAGGCCGTTTCCTCAGCAACACTGAGCTCCAGTACCTCTATGGTCGTCTTCG  1100
     MetLysThrProLeuThrGluAlaValAlaLeuAlaAspSerGlnGlyArgPheLeuSerAsnThrGluLeuGlnTyrLeuTyrGlyArgLeuAr

TCAAGGTGCTTTCGCCCTTGAAGCGGCTCAAACGTTGACTGCAAAAGCTGACACCCTCGTTAATGGTGCTGCTCAAGCGGTTTACAGCAAGTTCCCCTAC  1200
gGlnGlyAlaPheAlaLeuGluAlaAlaGlnThrLeuThrAlaLysAlaAspThrLeuValAsnGlyAlaAlaGlnAlaValTyrSerLysPheProTyr

ACCACCAGCACTCCTGGCAACAACTTCGCTGCTGACCAGCGCGGTAAAGACAAGTGTGCTCGTGACATCGGTTACTACCTCCGCATGGTTACCTACTGCC  1300
ThrThrSerThrProGlyAsnAsnPheAlaAlaAspGlnArgGlyLysAspLysCysAlaArgAspIleGlyTyrTyrLeuArgMetValThrTyrCysL

TAGTTGCTGGTGGTACTGGCCCCATGGATGAGTACCTCATCGCTGGTGTTGACGAAATCAACCGTACTTTCGATCTTTCTCCCAGCTGGTATGTTGAAGC  1400
euValAlaGlyGlyThrGlyProMetAspGluTyrLeuIleAlaGlyValAspGluIleAsnArgThrPheAspLeuSerProSerTrpTyrValGluAl

TCTCAAGCACATCAAAGCAAACCATGGTTTGACTGGCGATGCTGCTACTGAAACTAACAACTACATCGACTACGCAATTAACGCCCTCAGCTAATTTTGC  1500
aLeuLysHisIleLysAlaAsnHisGlyLeuThrGlyAspAlaAlaThrGluThrAsnAsnTyrIleAspTyrAlaIleAsnAlaLeuSerEnd

TTAGTCTAGGCCCGGATGGGTAAGTGGTTTTCAGCTTAAGTGTTGGGTTCTACTTACTTCTCCGGGTCTTGCTCTATCTAAAAACATTGGTTTAACAAGG  1600

AGTATTAGGCAAATGCCAGTTACTGTCGTGCCTCTCGCTTGGGAACCGCTGCGTTTGACCAATCACCCGTCGAACTGCGCGCTAACTATTCTCGA
```

FIG. 3. Nucleotide sequence of PC coding region. The coding sequence for PCβ is nucleotides 382–900, and that for PCα is nucleotides 1006–1494.

partial amino acid sequence of PR-6 PCα derived from peptide analysis (14) corresponds in general to that derived from the nucleotide sequence, but there are discrepancies. The deduced amino acid sequence presented here shows more homology with PCα subunits from other organisms than does the reported partial sequence (14). The similarities between nucleotide- and peptide-derived amino acid sequences identify these two coding regions as PC subunit genes.

The nucleotide sequence of the PCα gene from bases 1324 to 1337, as numbered in Fig. 3, is exactly complementary to probe 6. Thus, the use of RNA hybridization to screen candidate oligonucleotides identified the best-matched probe.

Plasmid pAQPR1 was used to probe Southern blots of *A. quadruplicatum* total DNA and proved homologous to unique restriction fragments (data not shown). Therefore, the PCα and PCβ genes are apparently present as unique sequences. The possibility remains that these genes occur in more than one copy per genome; for example, within large and highly homologous repeats or on a large multicopy plasmid that is linearized by the DNA purification procedure.

The NH_2- and COOH-terminal residues of the deduced PCα amino acid sequence are the same as those found by peptide analysis (14, 26). Therefore, this subunit apparently does not undergo proteolytic post-translational processing. The same inference can be drawn for the NH_2-terminus of PCβ (26). Data on the COOH terminus are not available, but the COOH-terminal sequences of PCβ from other organisms match that of the deduced sequence for *A. quadruplicatum* PCβ, rendering processing unlikely. These conclusions are in agreement with studies on translation of *Cyanidium caldarium* RNA *in vitro*, which showed that post-translational cleavage at the NH_2-termini of PC subunits does not occur in that organism (27).

The PCα and PCβ genes are in the same orientation, with the PCβ gene located upstream of the PCα gene. Maxicell experiments on pAQPR1 deletions show that, in *E. coli*, the PC subunit genes are cotranscribed from a promoter between the inward *Bgl* II site and the 5′ end of the PCβ coding sequence (unpublished results). Comparisons of the promoter sequences for other cyanobacterial genes with the sequence 5′ to the coding sequence of the PCβ gene revealed no extensive homology. The two PC coding sequences are separated by an apparently noncoding segment of 105 base pairs. Both coding regions are preceded by Shine–Dalgarno-type sequences, which may be involved in the translation of PC mRNA (28). Beginning 13 bases 5′ to the PCβ initiation codon is the sequence A-G-G-A-G, and 15 bases 5′ to the PCα initiation codon is G-G-A-G. These two sequences are complementary to a segment at the 3′ end of the only reported blue-green algal 16S rRNA: 5′ ... C-C-U-C-C-U-U-U3′ (29, 30).

Codon frequences in the PCα and β genes are similar (Table 1). Qualitatively, the preferred synonymous codons of *E. coli* genes (31, 32) generally match those of the PC genes, but there are exceptions. Most notably, the preferred leucine (CUG) and proline (CCG) codons of *E. coli* are never used in *A. quadruplicatum* PC genes. Thus, the compositions of isoaccepting tRNA species may be rather different in the two organisms. Codon optimization has been demonstrated for abundantly expressed genes in *E. coli* (31, 32) and is expected for PC genes since PC subunits are among the most abundant proteins in blue-green algae. One criterion of codon optimization is the choice of uracil or cytosine as the 3′ base in codons of certain amino acids (31, 32). This criterion is not so dependent on iso-tRNA composition since N_1N_2U and N_1N_2C are usually decoded by the same tRNA. Gouy

```
                          20                  40                  60                  80   *
(A) Ag) MKTPLTEAVALADSQGRFLSNTELQYLYGRLRQGAFALEAAQTLTAKADTLVNGAAQAVYSKFPYTTSTPGNNFAADQRGKDKCARDIGY
    Sy) S         A         S    VAF  F  A SG A  KA ANN  S  D   N                  STPE  A
    Ma) V   I D I A  T           AVN  YQRA AS    RA   N QR ID       Q    LIQ S P Y   A   S       H
    Cy)     I   I A  N           AVN  YQRA AS    RS  SN ER I             TSQ P PQY SSAV  A

               100                 120                 140                 160
    Ag) YLRMVTYCLVAGGTGPMDEYLIAGVDEINRTFDLSPSWYVEALKHIKANHGLTGDAATETNNYIDYAINALS
    Sy)    I   A        I    L  L    TK   A         Y       S  SRD A S    IL
    Ma)    II  S        L       LN   DA E      I    Y       S Q  N A T    V
    Cy)        C  V             LE    T            NY       S Q  N A T

                          20                  40                  60                  80 *
(B) Ag) MFDIFTRVVSQADARGEFISSDKLEALKKVVAEGTKRSDAVSRMTNNASSIVTNAARQLFADQPQLIAPGGNAYTNRRMAACLRDMEIIL
    Sy) T  A  K  A        L DAQ D SLRL    N  I T N I G      A    A   E  S            -
    Ma) AY V  K      S    L NEQ D  AN  K  N  L V N I S   T       A  EE         S TR GT
    Cy)  L A AK  AQ       L NTQ D  S M S  N  L V N I S   A       A  SE     Q   T

               100                 120                 140            *    160
    Ag) RYVTYATFTGDASVLNDRCLNGLRETYVALGVPGASVAAGVRAMGKAAVAIVMDPAGVTSGDCSSLQQEIELYFETAAKAVE
    Sy)       V      I D    D      L       L  E   K KD      S RN I Q    A-IS LGS  DK  A  A
    Ma)   I   ILA                  Q   T  S   V IQK KE  IN AN  N I K    A IS VAS  DR  A  A
    Cy)    S  IIA  S I D           Q          V IEK KDS I  AN  S I T    A MA VGT  DR  T  Q
```

FIG. 4. Comparison of PCα (*A*) and PCβ (*B*) amino acid sequences. Organisms and references for sequence data: Ag (*A. quadruplicatum* PR-6), PCα and β sequences deduced from nucleotide sequence in this paper; Sy (*Synechococcus* 6301), PCα (21) and PCβ (22); Ma (*Mastigocladus laminosus*), PCα and PCβ (23); Cy (*Cyanidium caldarium*), PCα (24) and PCβ (25). Blank positions indicate identity with the *A. quadruplicatum* sequence at that position. Dashes indicate gaps inserted to maximize homology. An asterisk above a C indicates a phycocyanobilin-binding cysteine. 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; and Y, tyrosine.

and Gautier have defined a parameter, called the P2 index, for evaluating optimization based on the choice of 3′ pyrimidine (32). The calculated values of this index for the PCα and PCβ genes are 0.79 and 0.70, respectively (a value of 1.0 represents maximum optimization). These values are in the range of those found for highly expressed *E. coli* genes, such as that for ribosomal protein S10 (P2 index = 0.75), and suggest that codon frequencies in PC genes reflect optimization.

Two other potential coding sequences have been identified in the nucleotide sequence shown in Fig. 3. The first is nucleotides 157–273, and the second begins at nucleotide 1613 and extends beyond the sequenced region. The length of the hypothetical polypeptide encoded by the former open reading frame is 38 residues, and that of the latter is at least 28 residues. Both coding sequences have high proportions of codons that are rare in PCα and PCβ genes (Table 1). For the putative 38-residue polypeptide, no Shine–Dalgarno-type sequence longer than a trinucleotide (G-A-G) is discernible at an appropriate distance upstream of the initiation codon of its coding sequence. The coding sequence beginning at nucleotide 1613 is preceded by A-A-G-G-A-G, which could constitute a 16S rRNA binding sequence. The significance of these two open reading frames is unknown.

The deduced amino acid sequences of PC subunits from PR-6 are shown in Fig. 4 in comparison with the other three complete sets of PC sequences: those of *Synechococcus* 6301 (21, 22), *Mastigocladus laminosus* (23), and *C. caldarium* (24, 25). Overall, there is considerable homology between the four sequences of each subunit. The average homologies between all pair-wise combinations are 72.5 ± 5.4% for α subunits and 71.5 ± 4.2% for β subunits. This comparison shows no statistically significant difference in amino acid sequence conservation between subunit types.

Amino acid sequence analyses of phycobiliproteins belonging to all spectroscopic classes indicate that the subunits of these proteins are closely related descendents of a single ancestral gene (25–29, 33, 34). A comparison of the deduced amino acid sequences of the α and β subunits of *A. quadruplicatum* PC reveals ≈27% homology, although allowance for the numerous conservative substitutions would increase this value considerably. The overall nucleotide homology, when the genes are aligned for maximal homology, is ≈46%. Although the amino acid sequence homology is most notable around the common chromophore-binding cysteine (residue 84 in PCα and residue 82 in PCβ), the nucleotide homology is

Table 1. Codon frequencies in *A. quadruplicatum* PR-6 PCα and PCβ coding sequences

Amino Acid	Codon	Frequency PCα	Frequency PCβ	Amino Acid	Codon	Frequency PCα	Frequency PCβ
Arg	CGA	0	1	Gly	GGA	0	1
	CGC	2	4		GGC	4	1
	CGG	0	1		GGG	0	0
	CGU	5	6		GGU	9	10
	AGA	0	0	Val	GUA	1	6
	AGG	0	0		GUC	0	0
Leu	CUA	1	0		GUG	0	0
	CUC	9	10		GUU	6	10
	CUG	0	0	Lys	AAA	4	4
	CUU	4	2		AAG	3	2
	UUA	0	0	Asn	AAC	8	4
	UUG	2	0		AAU	1	3
Ser	UCA	0	0	Gln	CAA	4	3
	UCC	0	3		CAG	2	3
	UCG	0	0	His	CAC	1	0
	UCU	2	6		CAU	1	0
	AGC	5	3	Glu	GAA	5	8
	AGU	0	0		GAG	2	1
Thr	ACA	0	0	Asp	GAC	6	5
	ACC	6	7		GAU	4	5
	ACG	1	0	Tyr	UAC	9	4
	ACU	8	4		UAU	2	1
Pro	CCA	0	0	Cys	UGC	1	2
	CCC	3	4		UGU	1	1
	CCG	0	0	Phe	UUC	5	4
	CCU	2	0		UUU	0	2
Ala	GCA	5	3	Ile	AUA	0	0
	GCC	2	3		AUC	5	5
	GCG	2	4		AUU	1	3
	GCU	14	17	Met	AUG	3	6
				Trp	UGG	1	0

essentially constant over the entire lengths of the two coding sequences.

The deduced amino acid sequences of the *A. quadruplicatum* PC subunits were also compared to the α and β subunits of *M. laminosus* allophycocyanin and *C. caldarium* allophycocyanin (33, 34). When aligned to allow maximal homology, the PCβ subunit shows the same degree of homology ($\approx$36%) to all allophycocyanin subunits. The PCα subunit is less homologous to the allophycocyanin subunits from these organisms; PCα exhibits slightly higher homology to the allophycocyanin α subunits (31%) than to the β subunit (28%). These homologies are similar to pair-wise comparisons of other phycocyanin and allophycocyanin sequences (33, 34).

PCβ subunits for which amino acid sequences have been determined are slightly longer (170–172 amino acids) than the corresponding PCα subunits (162 amino acids; see Fig. 4). Additionally, the PCβ subunits have an additional chromophore-binding site located at cysteine-153, near the COOH terminus of this subunit. It has been proposed that an insertion event, whereby a short stretch of 10 or 11 amino acids was introduced into the ancestral PCβ gene product near its COOH terminus, could have generated this second chromophore-binding site (33, 34). An analysis of the nucleotide and amino acid sequences suggests that an imperfect recombination event, which effectively duplicated a short stretch of 8 or 9 amino acids, may have been responsible for lengthening members of the PCβ gene family. The nucleotide sequence 5'GGT-GTA-CGT-GCA-ATG-GGT-AAA-GCT3', encoding amino acids 129–137 of PCβ, is very similar to the immediately adjacent 5'GCT-GTA-***-GCG-ATT-GTT-ATG-GAT3', encoding amino acids 138–144. A significant level of nucleotide homology (46%, equivalent to that observed throughout the remainder of the coding sequences) is evident between the regions encoding amino acids 141–162 in PCα and amino acids 157–172 in PCβ. This optimal alignment does require that a certain number of insertions/deletions have occurred, since the PCβ sequence is four amino acids longer than the corresponding PCα sequence.

We have observed a bewildering array of structural features in comparing the coding sequences for the two genes, including direct repeats, inverted repeats, and palindromes. Of particular interest are two direct repeats of nine nucleotides, 5'G-C-T-G-G-T-G-T-A3', in the vicinity (nucleotides 766–774 and 817–825) of the apparently duplicated sequence discussed above. Interestingly, a 38-nucleotide imperfect palindrome in the PCα coding sequence, centered between nucleotides 1406 and 1444, is also centered on the homologous region affected by the proposed duplication. Whether these structural elements played any role in the events leading to the lengthening of the members of the PCβ gene family cannot be determined with the limited sequence data available.

With clones of the PCα and PCβ genes, we can begin to address a number of questions regarding phycobiliproteins. Phycobiliproteins such as phycoerythrin and allophycocyanin may be cloned using PC genes or fragments of these genes as probes, since these proteins share certain amino acid sequences. By analyzing nucleotide sequences and arrangements, we may gain insights as to the evolution of the phycobiliprotein gene family. Expression of phycobiliproteins is known to be regulated by such factors as light wavelength, light intensity, and nutrient limitation. Cloned genes for these proteins will facilitate the study of their expression. Furthermore, the capacity of some blue-green algae, including *A. quadruplicatum* PR-6, for genetic transformation may allow phycobiliproteins to be specifically altered *in vivo* by directed mutagenesis.

The skilled assistance of James Dubbs and Veronica Stirewalt is gratefully acknowledged. This work was supported by United States Department of Agriculture Grant 82-CRCR-1-1080 and United States Public Health Service Grants GM31625 and GM28609.

1. Glazer, A. N. (1982) *Annu. Rev. Microbiol.* **36,** 173–198.
2. Yu, M. H. & Glazer, A. N. (1982) *J. Biol. Chem.* **257,** 3429–3433.
3. Giddings, T. H., Jr., Wasmann, C. & Staehelin, C. A. (1983) *Plant Physiol.* **71,** 409–419.
4. de Lorimier, R., Bryant, D. A., Porter, R. D., Jay, E. & Stevens, S. E., Jr. (1984) in *Abstracts of the Annual Meeting of the American Society for Microbiology*, ed. Neidhardt, F. C. (Am. Soc. Microbiol., Washington, DC), p. 96.
5. Stevens, S. E., Jr., & Porter, R. D. (1980) *Proc. Natl. Acad. Sci. USA* **77,** 6052–6056.
6. Maniatis, T., Fritsch, E. F. & Sambrook, J. (1982) in *Molecular Cloning* (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY), p. 196.
7. 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.
8. Nobrega, F. G., Dieckmann, C. L. & Zagoloff, A. (1983) *Anal. Biochem.* **131,** 141–145.
9. Hanahan, D. & Meselson, M. (1983) in *Methods In Enzymology*, eds. Wu, R., Grossman, L. & Moldave, K. (Academic, New York), Vol. 100, pp. 333–342.
10. Vieira, J. & Messing, J. (1982) *Gene* **19,** 259–268.
11. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 5463–5467.
12. Heidecker, G., Messing, J. & Gronenborn, B. (1980) *Gene* **10,** 69–73.
13. Conrad, B. & Mount, D. W. (1982) *Nucleic Acids Res.* **10,** 31–38.
14. Gardner, E. E. (1980) Dissertation (Univ. of Texas, Austin, TX).
15. Lau, R. H., MacKenzie, M. M. & Doolittle, W. F. (1977) *J. Bacteriol.* **132,** 771–778.
16. Paone, D. A. M. & Stevens, S. E., Jr. (1981) *Plant Physiol.* **67,** 1097–1100.
17. Roberts, T. M. & Koths, K. E. (1976) *Cell* **9,** 551–557.
18. Bolivar, F. (1978) *Gene* **4,** 121–136.
19. Buzby, J. S., Porter, R. D. & Stevens, S. E., Jr. (1983) *J. Bacteriol.* **154,** 1446–1450.
20. Sancar, A., Wharton, R. P., Seltzer, S., Kacinski, B. M., Clarke, N. D. & Rupp, W. D. (1981) *J. Mol. Biol.* **148,** 45–62.
21. Freidenreich, P., Apell, G. S. & Glazer, A. N. (1978) *J. Biol. Chem.* **253,** 212–219.
22. Walsh, R. G., Wingfleld, P., Glazer, A. N. & DeLange, R. J. (1980) *Fed. Proc. Fed. Am. Soc. Exp. Biol.* **39,** 2060 (abstr.).
23. Frank, G., Sidler, W., Widmer, H. & Zuber, H. (1978) *Hoppe-Seyler's Z. Physiol. Chem.* **359,** 1491–1507.
24. Offner, G. D., Brown-Mason, A. S., Ehrhardt, M. M. & Troxler, R. F. (1981) *J. Biol. Chem.* **256,** 12167–12175.
25. Troxler, R. F., Ehrhardt, M. M., Brown-Mason, A. S. & Offner, G. D. (1981) *J. Biol. Chem.* **256,** 12176–12184.
26. Gardner, E. E., Stevens, S. E., Jr., & Fox, J. L. (1980) *Biochim. Biophys. Acta* **624,** 187–195.
27. Belford, H. S., Offner, G. D. & Troxler, R. F. (1983) *J. Biol. Chem.* **258,** 4503–4510.
28. Shine, J. & Dalgarno, L. (1974) *Proc. Natl. Acad. Sci. USA* **71,** 1342–1346.
29. Tomioka, N. & Sugiura, M. (1983) *Molec. Gen. Genet.* **191,** 46–50.
30. Williamson, S. E. & Doolittle, W. F. (1983) *Nucleic Acids Res.* **11,** 225–235.
31. Ikemura, T. (1981) *J. Mol. Biol.* **151,** 389–409.
32. Gouy, M. & Gautier, C. (1982) *Nucleic Acids Res.* **10,** 7055–7074.
33. Sidler, W., Gysi, J., Isker, E. & Zuber, H. (1981) *Hoppe-Seyler's Z. Physiol. Chem.* **362,** 611–628.
34. Offner, G. D. & Troxler, R. F. (1983) *J. Biol. Chem.* **258,** 9931–9940.