Volume 259, number 2, 273–280 FEB 08000 January 1990

The cyanelle genome of *Cyanophora paradoxa* encodes ribosomal proteins not encoded by the chloroplast genomes of higher plants

Donald A. Bryant and Veronica L. Stirewalt

Dept. of Molecular and Cell Biology, Pennsylvania State University, University Park, PA 16802, USA

Received 28 October 1989; revised version received 20 November 1989

The *rpl35*, *rpl20*, *rpl5*, *rps8*, and a portion of the *rpl6* genes of the cyanelle genome of *Cyanophora paradoxa* have been cloned, mapped and sequenced. Homologs of the *rpl35*, *rpl5*, and *rpl6* genes are not found in the chloroplasts of higher plants. The *rpl35* and *rpl20* genes most likely form a dicistronic operon which is located upstream from the *apcE-apcA-apcB* locus of the cyanelle and which is divergently transcribed from this locus. The *rpl5*, *rpl8*, and *rpl6* genes probably form a part of a larger cluster of genes encoding components of the cyanellar ribosomes. These genes are organized in a fashion similar to that observed in all procaryotes examined to date, with the exception that the *rps14* gene is not found between the *rpl5* and *rps8* coding sequences. Hypotheses concerning the origins of cyanelles and chloroplasts are discussed.

Cyanelle; Ribosomal protein; Nucleotide sequence; Chloroplast; *(Cyanophora paradoxa)*

1. INTRODUCTION

Cyanophora paradoxa is a phylogenetically ambiguous, biflagellated protozoan that belongs to a diverse group of photosynthetic organisms containing plastids known as cyanelles [1,2]. Metabolically, cyanelles appear to be the equivalent of chloroplasts [1–3]. However, the organization and composition of the photosynthetic apparatus in cyanelles most closely resembles that found in cyanobacteria and the chloroplasts of red algae [1,2,4]. The cyanelles of *C. paradoxa* possess phycobiliproteins [5,6] arranged in hemidiscoidal phycobilisomes, and the organization and structure of the thylakoid membranes is quite similar to that observed in many free-living cyanobacteria [7]. These similarities and the observation that cyanelles are enclosed in a typical lysozyme-sensitive, Gram-negative peptidoglycan layer [8] were doubtlessly responsible for the initial belief that the cyanelles of *C. paradoxa* were endosymbiotic cyanobacteria. However, the cyanelle is also genetically equivalent to the chloroplasts of higher plants. The circular cyanellar DNA of *C. paradoxa* has a G + C content of 36% [9] and is about 133 kbp in length [10–12]; this is approximately 20- to 30-fold smaller than the smallest genomes known for cyanobacteria [9] but is similar in size to those of algal and higher plant chloroplasts [13]. Although the overall distribution of genes on the cyanelle genome differs significantly from that observed in higher plant chloroplast genomes, the number and types of genes encoded is nonetheless rather similar [10–12,14]. Notable differences are the presence of genes encoding the subunits of the phycobiliproteins allophycocyanin (*apcA* and *apcB* [5,6]), phycocyanin (*cpcA* and *cpcB* [5,6]), the large core linker phycobiliprotein LCM100 (*apcE* [15]), and the small subunit of ribulose-1,5-bisphosphate carboxylase/oxygenase (*rbcS*) which is adjacent to and cotranscribed with the *rbcL* gene encoding the large subunit [16].

We are interested in the origins of chloroplasts and believe that more detailed studies of the sequences and arrangements of gene loci in the cyanelle might provide clues to the origins of these organelles. Partial nucleotide sequence analysis of the 16S rRNA of *C. paradoxa* has already shown that cyanelles, like higher plant and algal chloroplasts, are closely related phylogenetically to the cyanobacteria [17]. In the course of analyzing clones of cyanellar DNA encoding the cytochrome b-559 operon (*psbE-psbF-psbL-psbJ* [18]) and the LCM100 core linker phycobiliprotein-allophycocyanin locus (*apcE-apcA-apcB* [6,15]), two clusters of sequences encoding ribosomal proteins were detected. We report here the nucleotide sequences of the *rpl35* (*rpmI*), rpl20 (*rplT*), *rpl5* (*rplE*), and *rps8* (*rplH*) genes as well as a portion of the *rpl6* (*rplF*) genes of *C. paradoxa*. The organization of these genetic loci is similar to but distinct from that found both in *Escherichia coli* and other procaryotes and in the chloroplasts of algae and higher plants. The relationship between cyanelles and chloroplasts, and some hypotheses concerning the

Correspondence (present) address: D.A. Bryant, Institut für Molekularbiologie und Biophysik, ETH-Honggerberg-HPM, CH-8093 Zurich, Switzerland

The nucleotide sequence presented in fig.1 has been submitted to the EMBL database under the accession no. X17063. The nucleotide sequence presented in fig.3 has been submitted to the EMBL database under the accession no. X16548

Published by Elsevier Science Publishers B.V. (Biomedical Division)
00145793/90/$3.50

endosymbiotic events leading to chloroplasts and cyanelles, are discussed.

2. MATERIALS AND METHODS

The *C. paradoxa* strain used in these studies was originally obtained from the Pasteur Culture Collection, Institut Pasteur, Paris, France and is apparently identical to the standard LB555 University of Texas (UTEX) Culture Collection strain [10,11]. The isolation of cyanellar DNA and construction of cyanellar DNA libraries of *Bam*HI and *Pst*I libraries in either pBR322 or pUC9 vectors has been described [10]. Libraries of *Eco*RI restriction fragments were produced in vector pUC19 in *E. coli* strain DH5α by similar, standard procedures [19,20]. The *Eco*RI fragment encoding *rps8* was originally cloned together with a similar-sized fragment encoding the *psbE* and *psbF* genes from an 18.0 kbp *Pst*I fragment [18]. The cloning of the 7.8 kbp *Pst*I fragment, as plasmid pCPCPst7.8, has been previously described [6]. Nucleotide sequencing of both DNA strands was performed by the synthetic 'primer walking' method as previously described [21], except that α-[^{35}S]thio-dATP and modified T7 DNA polymerase (Sequenase, US Biochemical, Cleveland, OH) were used in the chain termination method. DNA sequences were analyzed using the programs of Conrad and Mount [22] and the University of Wisconsin Genetic Computer Group (UWCGC; [23]). Previously published data were obtained from GenBank Version 60, NBRF databank Version 20, and the Swiss Protein databank Version 11.

3. RESULTS

The *rpl35* (*rpmI*) and *rpl20* (*rplT*) genes were located during sequence analysis of a 7.8 kbp *Pst*I fragment of cyanellar DNA which also encodes the *apcE-apcA-apcB* genes of *C. paradoxa* [6,10,15]. The region encoding the ribosomal proteins lies at approximately 26.8 to 27.7 on the cyanellar genome map (see [10]). Fig.1 shows the nucleotide sequence and deduced amino acids for the *rpl36* (*rpmI*) and *rpl20* (*rplT*) genes of *C. paradoxa*. The translational start codon for the *rpl35* gene lies 304 nucleotides upstream from the translational start codon for the *apcE* gene which is transcribed from the opposite DNA strand ([15]; Stirewalt, V. and Bryant, D.A., in preparation). The *rpl35* gene extends from nucleotides 81 to 278 and predicts a protein of 65

```
                                                                    50
TTAAACTTCT TTTCAGTTTT TAAAAAAAGA ATATTATAAT ATAGTTTATT AAATAATAAT TTATTTTCAC
                                                  100
AGAAAAATTC  ATG TAT AAA TTA AAA ACA CGT AAA GCA GCA GCT AAA CGT TAT AAA GCA GTA
             M   Y   K   L   K   T   R   K   A   A   A   K   R   Y   K   A   V
                                150
GGT AAT AAA AAA ATT AGT CGT CGT AAA GCG TTT CGA AGT CAC TTA TTA CAG AAA AAA AGC
 G   N   K   K   I   S   R   R   K   A   F   R   S   H   L   L   Q   K   K   S
       200                                                                 250
ACA AAT AGA AAA CGT CAA TTA TCA CAA GTA GTG ATC GCA AGC CCA GGT GAT ACT AAA AAA
 T   N   R   K   R   Q   L   S   Q   V   V   I   A   S   P   G   D   T   K   K
                                                            300
ATT TAT TTA ATG TTA CCA TAT TTA TAA ATTTTATT TTTTAAGTTT AAATAAAAAA AAAGGAATTT
 I   Y   L   M   L   P   Y   L   *
                                    350
CTAAAAA ATG ACT CGA GTA AAA CGT GGG AAT GTT GCC CGA AAA CGT CGT AAG AAA ATT TTA
         M   T   R   V   K   R   G   N   V   A   R   K   R   R   K   K   I   L
                            400
AAA TTA GCT AGT GGA TTT AGA GGT GCT CAT TCT CGT TTA TTC CGT GTT GCA AAC CAA CAA
 K   L   A   S   G   F   R   G   A   H   S   R   L   F   R   V   A   N   Q   Q
                450
GTA ATG AAA GCG TTA CGT TAT GCT TAT AAT GAT CGT AAT AAA CGT AAA CGT GAT TTT CGT
 V   M   K   A   L   R   Y   A   Y   N   D   R   N   K   R   K   R   D   F   R
500                                                                 550
GCT TTA TGG ATT GCA CGT ATT AAT GCT TCC GCA CGT TTA GAA GGT ATG ACC TAT AGT AAA
 A   L   W   I   A   R   I   N   A   S   A   R   L   E   G   M   T   Y   S   K
                                                    600
TTA ATG GGA AGT TTA AAG AAA CTT AAT ATT ATT TTA AAT AGA AAA AGT CTT TCT CAA TTA
 L   M   G   S   L   K   K   L   N   I   I   L   N   R   K   S   L   S   Q   L
                                    650
GCG ATT TAT GAT AAA GAT GCA TTT ATG GAA ATT TTA AAA ACC ATT CCT TAA TAAAGTTAATTT
 A   I   Y   D   K   D   A   F   M   E   I   L   K   T   I   P   *
                  700                                                   750
ATTAATTAAA GCTAATCAAT ATCCTATCTA GAAAAATAAA TAACAAAATA CCATAAATAA AATTTAAGGT
                                                    800
GATTTATATT AAAAAAGAAT AGTAATGAGA TTGTATACAG CCTCATTACT ATTCTTTTTT TAAAGAGTAG
                                850                                 *  L  T  P
GGGACATCCC ATTTTTTATG GTATCTAAAT TAAATAGAGT TTGTAAGATT TTTAAAGCTC GTCCTCTAAT
  S  M  G   N  K  I   T  D  L  N   F  L  T   Q  L  I   K  L  A  R   G  R  I <ORF
```

Fig.1. Nucleotide sequence and deduced amino acid sequences of the *rpl35* (*rpmI*) and *rpl20* (*rplT*) genes of the cyanelle genome of *C. paradoxa*. The *rpl35* gene begins at nucleotide 81 and ends at nucleotide 278. The *rpl20* gene extends from nucleotides 324 to 668. Polypurine sequences (Shine-Dalgarno-like, ribosome-binding sites) complementary to the 3′-terminus of the cyanelle 16S rRNA preceding each start codon are underlined. A large inverted repeat (nucleotides 761 to 810) is also underlined. A portion of an open reading frame greater than 265 codons transcribed from the opposite DNA strand is also shown (nucleotides 890 to 810).

amino acids with a molecular mass of 7607 Da and an isoelectric point of 12.08 (figs 1 and 2). A comparison of the deduced amino acid sequence of the *rpl35* gene product to the *rpmI* gene product of *E. coli* is shown in fig.2. This sequence is 38% identical and 55% similar, including conservative replacements, to the *rpmI* gene of *E. coli* [24].

The initiation codon for the *rpl20* gene occurs 45 nucleotides downstream from the termination codon of the *rpl35* gene and predicts a protein of 114 amino acids with a molecular mass of 13 354 Da and an isoelectric point of 12.24. A comparison of the sequences of several *rpl20* gene products is shown in fig.2. The *rpl20* gene product is most similar in sequence to the protein from liverwort (60% identity, 71% similarity), and is least similar (38% identity) to that of *E. gracilis*. The *C. paradoxa rpl20* gene product is equally similar to those of tobacco (52% identity) and *E. coli* (54% identity).

Both the *rpl35* and the *rpl20* genes are preceded by polypurine, Shine-Dalgarno-type sequences which are complementary to the 3′-terminus of the 16S rRNA of the cyanelle [29]. Downstream from the *rpl20* gene is a large inverted repeat which probably plays a role in transcription termination and/or mRNA stabilization. An open reading frame of greater than 265 codons, which has no significant homology to sequences currently in the various databases, terminates at this same inverted repeat from the opposite DNA strand as shown in fig.1. A putative common transcriptional terminator has also been found between the adjacent and divergently transcribed *psbDC* and *psbK* genes of the cyanelle [30].

An *Eco*RI fragment of 777 bp encoding a portion of *rpl5* (*rplE*), *rps8* (*rpsH*), and a portion of *rpl6* (*rplF*) was accidentally subcloned from an 18.0 kbp *Pst*I fragment of cyanelle DNA during the cloning and characterization of the *psbE* and *psbF* genes [18]. Subsequent mapping experiments (data not shown) indicated that this *Eco*RI fragment lies between coordinates 47 and 48 on the cyanelle genome map (see [10]). A portion of the adjacent 3.5 kbp *Eco*RI fragment (coordinates 47.7–51.2; see [10]) was also sequenced to complete the sequence of the *rpl5* gene. These sequences, and the deduced amino acid sequences, are shown in fig.3. The *rpl5* gene extends from nucleotides 41 to 586 and predicts a protein of 181 amino acids with a molecular mass of 20 482 Da and an isoelectric point of 10.49. At 11 to 16 bp in the 5′ direction from the translational start codon is a polypurine tract complementary to the 3′-terminus of the 16S rRNA of the cyanelle [29]. In fig.4, the *C. paradoxa rpl5* gene product is compared to those of *E. gracilis*, *Bacillus stearothermophilus*, *Bacillus subtilis*, *E. coli*, and *Mycoplasma capricolum*. The cyanellar protein is most similar to those of the

A

```
                  10         20         30         40         50         60
C. PARA: MYKLKTRKAA AKRYKAVGNK KISRRKAFRS HLLQKKSTNR KRQLSQVVIA SPGDTKKIYL MLPYL
E. COLI:  P I  VRG  S  F KT KG GFKHKH NLR    T  A K    H RPKAMV  K  LGLVIA C   A
```

B

```
                    10         20         30         40         50         60
C. PARADOXA: MTRVKRGNVA RKRRKKILKL ASGFRGAHSR LFRVANQQVM KALRYAYNDR NKRKRDFRAL
LIVERWORT:         Y        N  T  T   Q T  K    T    G  R  ASSHR   G    NL R
TOBACCO:        I   YI   R  T  RLF   S         T TIT  KI R  VS HR   DRK     R
E. GRACILIS:    I NNGIS K K  RKISK MK WV G  K    TG   L    RH  FY   K K NLNKT
E. COLI:     -A     VI   A H     Q  K YY  R   VY   F A I   GQ   R   RQ   Q  Q

                    70         80         90        100        110        120
C. PARADOXA: WIARINASAR LEGM---TYS KLMGSLKKLN IILNRKSLSQ LAIYDKDAFM ----EILKTI
LIVERWORT:     T VN A   DN I---S N   IEY Y KK       I A  I  L  FC S ----T I N
TOBACCO:       T    VI  ER VSY-S   R IHD Y RQ LL    I A  I  SNRNCLY MISN  I EV
E. GRACILIS:   T   GGLK YYLTINEK N IFVSF R TK TYV K L  E INVR SKS S ----HLS P
E. COLI:           A   QN I---S    FING   AS VEID  I AD I VF  V  T ----ALVEKA

                   130
C. PARADOXA: P
LIVERWORT:   ITE
TOBACCO:     DWKESTRII
E. GRACILIS: MKSTGINL
E. COLI:     KAALA
```

Fig.2. Comparison of the amino acid sequences of the *rpl35* (*rpmI*) (A) and *rpl20* (*rplT*) (B) genes of *C. paradoxa* with those of several other species. Hyphens indicate insertions or deletions introduced to maximize the homology. Only amino acids which differ from those of *C. paradoxa* are shown. References: *C. paradoxa rpl35* and *rpl20*, this work; *E. coli rpmI* [24]; liverwort *rpl20* [25]; tobacco *rpl20* [26]; *E. gracilis rpl20* [27]; and *E. coli rplT* [28].

```
                                                                  50
AAATTCAAAT TAATTTAAAA ATGTAAAGGA TTTACGAATT ATG GTT CAA CGG TTA AAA ACT GTA TAT
                                              M   V   Q   R   L   K   T   V   Y
                                          100
GAA CAA GAA GTC ATT AAA CAA TTA ATG ACT CGC TTC CAA TAT AAA AAT ATC CAT GAA GTA
 E   Q   E   V   I   K   Q   L   M   T   R   F   Q   Y   K   N   I   H   E   V
                      150
CCA AAA TTA AAA AAA ATA ACT GTA AAT CGT GGA TTA GGA GAA GCT GCT CAA AAT GCA AAG
 P   K   L   K   K   I   T   V   N   R   G   L   G   E   A   A   Q   N   A   K
              200
ATT TTA GAA GCT TCT GTT AAA GAA ATT ACT GAA ATT ACC GGT CAA AAA GCT ATT GTT ACT
 I   L   E   A   S   V   K   E   I   T   E   I   T   G   Q   K   A   I   V   T
250                                                               300
AGA GCC AAA AAA GCT ATT GCT GGT TTT AAA TTA AGA CAA GAT ATG CCT ATT GGT GTT ATG
 R   A   K   K   A   I   A   G   F   K   L   R   Q   D   M   P   I   G   V   M
                                                  350
GTT ACT TTA CGC GGT GAT TAT ATG TAT GCT TTC TTG GAT CGT TTA ATT AAT TTA TCT TTA
 V   T   L   R   G   D   Y   M   Y   A   F   L   D   R   L   I   N   L   S   L
                                          400
CCA AGA ATT CGA GAT TTT CGC GGT ATT ACT GCA AAA AGT TTT GAT GGT CGT GGA AAT TAT
 P   R   I   R   D   F   R   G   I   T   A   K   S   F   D   G   R   G   N   Y
                      450
AAT CTT GGT TTA AAA GAA CAA TTG ATT TTT CCA GAA GTT GAT TAT GAT AGT ATT GAA CAA
 N   L   G   L   K   E   Q   L   I   F   P   E   V   D   Y   D   S   I   E   Q
              500
ATT CGC GGT ATG GAT ATT TCA ATT GTA ACC ACT GCA AAA ACA AAT CAA GAA GGT CTT GCT
 I   R   G   M   D   I   S   I   V   T   T   A   K   T   N   Q   E   G   L   A
550                                                               600
TTA TTA AAA TCT TTA GGC ATG CCA TTT GCT GAA AGT TAA CTTCATTATTATAAGGAGGAAAA GTG
 L   L   K   S   L   G   M   P   F   A   E   S   *                         M
                                          650
GTT AAC GAT ACA ATT GCT GAT ATG CTT ACA CGT ATT CGA AAT GCA AAT TTA GCA AAA CAT
 V   N   D   T   I   A   D   M   L   T   R   I   R   N   A   N   L   A   K   H
                                      700
AAA GTT GCA CGT GTT AAA GCA ACA AAA ATA ACT CGT TGT TTA GCT AAT GTT TTA AAA GAA
 K   V   A   R   V   K   A   T   K   I   T   R   C   L   A   N   V   L   K   E
                  750
GAA GGC TTA ATT CAA AAT TTT GAA GAA ATA GAA AAT AAT CTT CAA AAT GAA TTA TTA ATA
 E   G   L   I   Q   N   F   E   E   I   E   N   N   L   Q   N   E   L   L   I
      800                                                               850
TCT CTT AAA TAT AAA GGT AAA AAA CGT CAG CCG ATA ATT ACT GCT TTA AAA AGA ATT AGT
 S   L   K   Y   K   G   K   K   R   Q   P   I   I   T   A   L   K   R   I   S
                                                  900
AAA CCA GGT TTA CGT GTT TAT GCG AAT CAT AAA GAA CTA CCT CGT GTA TTA GGA GGT TTA
 K   P   G   L   R   V   Y   A   N   H   K   E   L   P   R   V   L   G   G   L
                                          950
GGT ATT GCT ATT CTT TCA ACT TCT TCA GGA ATT ATG ACT GAT CAG ACC GCT CGT CAT AAA
 G   I   A   I   L   S   T   S   S   G   I   M   T   D   Q   T   A   R   H   K
                                      1000
GGT TGT GGT GGA GAA GTT TTA TGT TAT ATT TGG TAG ATTTATTAGGTGAATT     ATG TCT CGG
 G   C   G   G   E   V   L   C   Y   I   W   *                       M   S   R
                  1050
ATA GGT AAA CGA TTA ATT AAT ATT CCT AGC CAA GTT ACG GTT AGC ATC AAG GAT CAA GTA
 I   G   K   R   L   I   N   I   P   S   Q   V   T   V   S   I   K   D   Q   V
      1100
TTT TCT GTT AAA GGA CCA AAA GGA GAG TTA TCT AAA CAA ATT CCT TAT GGA ATT C
 F   S   V   K   G   P   K   G   E   L   S   K   Q   I   P   Y   G   I
```

Fig.3. Nucleotide sequence and deduced amino acid sequences of the rpl5 (*rplE*), *rps8* (*rpsH*), and a portion of the *rpl6* (*rplF*) genes of *C. paradoxa*. The *rpl5* gene extends from nucleotides 41 to 586, the *rps8* gene extends from nucleotides 610 to 1008, and the *rpl6* gene fragment extends from nucleotides 1025 to 1148. Polypurine tracts (Shine-Dalgarno-type, ribosome-binding sites) complementary to the 3′-terminus of the 16S rRNA of the cyanelle are underlined.

Bacillus species (59–61% identity, 77% similarity) and is slightly less homologous to those of *E. gracilis*, *E. coli* and *M. capricolum* (52–55% identity, 73–76% similarity). These values are similar to those found when comparing the proteins of the three different procaryotic genera (55–65% identity, 79–81% similarity).

Ths *rps8* gene initiates with a GTG start codon 24 bp in the 3′ direction from the termination codon of the *rpl5* gene. This start codon is preceded by a long (11 bp) polypurine tract, a portion of which could play a role in ribosome binding. In fig.3, the *rps8* gene extends from nucleotide 610 to 1008 and predicts a protein of 132

A

```
                    10         20         30         40         50         60
C. PARADOXA: MVQ-RLKTVY EQEVIKQLMT RFQYKNIHEV PKLKKITVNR GLGEAAQNAK ILEASVKEIT
E. GRACILIS:  M -   SF  LETI PK KE E G V SYR       VI    FD SC  S     VLLN LE
B. STEARO.:  MN-   EK  VK  VPA  S K N  S MQ    IE  VI M  V D V  P  A DSA E L
B. SUBTILIS: MN-   EK  NK IAPA    K N DSVMQ    IE  VI M  V D V     AIDSA E L
E. COLI:     MA-K HDY  KD  V K    E N NSVMQ   RVE   L M  V   IADK  L DNAAADLA
MYCOPLASMA:  MKS  EIK  KDQIVLE FK ELN  S MQ    IQ  VI M  I D TTDP  K D AIF LE

                    70         80         90         100        110        120
C. PARADOXA: EITGQKAIVT RAKKAIAGFK LRQDMPIGVM VTLRGDYMYA FLDRLINLSL PRIRDFRGIT
E. GRACILIS: I S   P IS K     N    KEK  V MF L   SEK  S                  Q  N
B. STEARO.:  L A  RPV       S    R   G    AK      ER  E    K  SV     V     VS
B. SUBTILIS: F A   PV       S    R  EG    AK      ER  D    K  SV     V     VS
E. COLI:     A S   PLIT K R SV     I  GY   CK      ER WE  FE   TIAV         LS
MYCOPLASMA:  KLS   P    K   SL V    EG A  AK      KK  D    K   VA    V     VS

                    130        140        150        160        170        180
C. PARADOXA: AKSFDGRGNY NLGLKEQLIF PEVDYDSIEQ IRGMDISIVT TAKTNQEGLA LLKSLGMPFAES
E. GRACILIS: KNC   S  F SF  S  SM    INF KMIK VQ LN T      E    AFF    E  I  RD-
B. STEARO.:  K A        T  I        I   KVNK V     V      N DE ARE   AL     QK-
B. SUBTILIS: K          T  I        I   KVTK V     V      N DE ARE   TQV    RQK-
E. COLI:     A          SM KR  I    I   KVDR V  L  T T     SDE  R   -AAFDF  RK-
MYCOPLASMA:  KT    F  F YT I   I      H KVIR L     T    S    K AF    QKI    EK-
```

B

```
                    10         20         30         40         50         60
C. PARADOXA:  MVNDTIADM LTRIRNANLA KHKVARVKAT KITRCLANVL KEEGLIQNFE EIENNLQNEL
LIVERWORT:      G     N  I S      G  I TVQ P   N   NI KI  FQ  F D  I DNKQ TKDI
TOBACCO:        GR    EI I S    DMD RKR V IAS  N  ENIVQI  LR  F E VR KHREKNKYF
E. COLI:     MSMQ P           GQA  NKAAVTMPSS  LKVAI        F ED - KV GDTKP
B. STEARO.:   VMT P        A     MV R EKLE P S   K EI EI   R  F RDY  Y  D K GI
MYCOPLASMA:    TT V             QR YL TVS PSS  VKLEI RI      F SD - TV GDVKKTI

                    70         80         90         100        110        120
C. PARADOXA: LISLKY-KGK KRQPI--ITA LKRISKPGLR VYANHKELPR VLGGLGIAIL STSSGIMTDQ
LIVERWORT:   ILN   -Q    KKSY--  T  R          I S    I K     M  V      R     R
TOBACCO:     VLT RH-RRN RKR YRN LN      R      I S YQRI   I   M  V      R     R
E. COLI:     ELT   FQ   AV-----VES IQ V R      I KRKDQ  K  MA L   VV    K V   R
B. STEARO.:  R F   GPNE RV-----  G               VKAH V     N           Q VL  K
MYCOPLASMA:  N E   -Q   T V---- QG   K            QAN I Q   N    S V    Q     GK

                    130
C. PARADOXA: TARHKGCGGE VLCYIW
LIVERWORT:   E  Q KI    L   V
TOBACCO:     E  LE I    I
E. COLI:     A  QA L    II  VA
B. STEARO.:  E  Q  T    IIA VI
MYCOPLASMA:  K  LANA      AF -
```

C

```
                    10         20         30         40
C. PARADOXA: MSRIGKRLIN IPSQVTVSI- KDQVFSVKGP KGELSKQIPY GI........
B. STEARO.:  -  V  KP E   AG   TV- NGNTVT         TRTFHP DM........
E. COLI:       VA APVV V AG D K - NG  ITI  K N   TRTLND AV........
MYCOPLASMA:      N  LQ   NG E K A ENNLVTIT S   T    FSP L ........
```

Fig.4. Comparison of the amino acid sequences for the *rpl5* (*rplE*) (A), *rps8* (*rpsH*) (B), and *rpl6* (*rplF*) (C) gene products of several species. Hyphens indicate insertions or deletions introduced to maximize homology. Only amino acids which differ from those of *C. paradoxa* are shown. References: *C. paradoxa*, this work; *E. gracilis* [31]; *B. stearothermophilus rpl5* [32]; *B. subtilis rpl5* [33]; *E. coli rpl5*, *rps8*, *rpl6* [34]; *Mycoplasma capricolum rpl5*, *rps8*, and *rpl6* [35]; liverwort *rps8* [25]; tobacco *rps8* [26]; *B. stearothermophilus rps8* [36]; and *B. stearothermophilus rpl6* [39].

amino acids with a molecular mass of 14 737 Da and an isoelectric point of 10.70. The *rps8* gene products of several species are compared in fig.4. The protein is significantly more similar to that of the liverwort (61% identity, 77% similarity) than any other and is surprisingly more similar to the proteins of *B. stearothermophilus* and *M. capricolum* (55–58% identity) than to those of *E. coli* and tobacco (47–48% identity). The

significance of these latter differences is not entirely clear at present, but a closer relationship to liverwort has also been suggested from the analysis of 16S rRNA sequences [17].

The translational start codon of the *rpl6* gene occurs 17 bp downstream from the termination codon of the *rps8* gene. At present only the first 41 codons of the gene have been sequenced, since the gene is interrupted by an *Eco*RI site employed in the cloning. Nonetheless, the portion of the gene sequenced is clearly identifiable. The amino terminal sequences of several *rpl6* gene products are compared in fig.4. The cyanellar *rpl6* gene product exhibits significant homology (37–46%) to those of several procaryotes.

4. DISCUSSION

The *rpmI* and *rplT* genes of *E. coli* are arranged in a larger gene cluster, 5′-*thrS-infC-rpmI-rplT-pheS-pheT*-3′ which produces multiple, overlapping transcripts and whose regulation is complex [38]. Nonetheless, it is probable that *rpmI* and *rplT* are usually co-transcribed, and that many transcripts also include *infC*, which encodes initiation factor III. In *C. paradoxa*, *rpl35* and *rpl20* are probably transcribed as a dicistronic mRNA since no potential transcription terminator occurs between the two genes but a large inverted repeat occurs downstream from the translation termination codon of *rpl20*. The *rpl35* gene has not been reported to occur in higher plant chloroplast genomes [25,26,39], and the *rpl20* gene is possibly transcribed as a monocistronic transcript in these organisms.

In *E. coli*, the *rpl5*, *rps8*, and *rpl6* genes are part of a large cluster of genes referred to as the *spc* (spectinomycin) operon [34]. This locus encodes ten ribosomal proteins: 5′-*rpl14-rpl24-rpl5-rps14-rps8-rpl6-rpl18-rps5-rpl30-rpl15-secY*-'X'-3′. The *spc* operon in turn in flanked upstream by the S10 operon [40], encoding eleven ribosomal proteins, and downstream by the α operon, encoding the α subunit of RNA polymerase and four additional ribosomal proteins [41]. Evrard et al. [42] have recently located the genes for several ribosomal proteins of the S10 operon on the cyanelle genome. These also include some genes which are nuclear-encoded in higher plants. These genes occur approximately 2.5 kbp upstream from *rpl5* near map coordinates 50.5 to 52 of the cyanelle genome. Considering the map distance between these sequences and their orientation relative to those reported here, it is possible that most of the intervening genes of the S10 and *spc* operons are encoded in the cyanelle in contrast to the situation in higher plant chloroplasts.

Of the three genes in the second locus described in this report, only *rps8* is encoded in the chloroplasts of higher plants [25,26,39] with *rpl5* and *rpl6* homologs presumably being encoded in the nucleus. The chloroplast genome of liverwort contains a much abridged version of these three ribosomal protein operons encoding a total of nine ribosomal proteins [25,43], and similar abridged versions occur in the tobacco [26] and rice [39] chloroplast genomes. In each of the cases, the *rps8* gene is flanked upstream by *rpl14* and downstream by *infA*. Interestingly, the chloroplast genome of *E. gracilis* has very recently been reported to contain an operon consisting of the five genes 5′-*rpl16-rpl14-rpl5-rps8-rpl36*-3′ [31]. The gene order of this locus differs from that in higher plant chloroplasts only by the presence of the *rpl5* gene. It is important to note that the *rps14* gene is chloroplast encoded in higher plants and that this gene occurs downstream from and is co-transcribed with the *psaAB* genes [42,45]. In *E. gracilis*, *rps14* is chloroplast encoded, and the gene is found downstream from *rpl36* [31]. Significantly, in *C. paradoxa* the *rps14* gene is 'missing' from this locus, as it is in higher plants and *E. gracilis*. In contrast, the gene order 5′-*rpl5-rps14-rps8-rpl6*-3′ has been shown to be conserved in diverse procaryotes including *M. capricolum* [35], *B. subtilis* [33], and even in the archaebacterium *Methanococcus vannieli* [46]. The location of the *rps14* gene in *C. paradoxa* is not yet known, but the gene has not been found in a 1.5-kbp region immediately downstream from *psaB* (Stirewalt, V. and Bryant, D.A., unpublished results).

There are essentially two ways to view the cyanelle: (1) as a relatively recent reinvention of the type of endosymbiotic event which led to chloroplasts and in which the cyanelle is convergently evolving towards the chloroplast; or (2) as a distinctive modern-day descendent from an ancestral organism that possibly evolved into organisms harboring chloroplasts on the one hand and cyanelles on the other. Although there are conflicting views concerning the origins of higher plant chloroplasts [17,47,48], recent evidence suggests that view 2 may be correct and that the phycobiliprotein-containing cyanelles are in fact more closely related to the chloroplasts of higher plants than are the chloroplasts of *E. gracilis* and *Chlamydomonas reinhardtii* [17,42]. This presents an apparent paradox, since the green algae share with the higher plants and certain procaryotes such as *Prochloron* sp. and *Prochlorothrix hollandica* the property of synthesizing chlorophyllous antennae proteins which bind chlorophylls *a*, *b*, and xanthophylls [15]. This apparent paradox may be quite simply explained by the following hypothesis: ancestral cyanobacteria which participated in the endosymbiotic event leading to chloroplasts and cyanelles had the biochemical potential to produce antennae consisting of both phycobiliproteins and chlorophylls *a* and *b* (or related chlorophylls) although both systems may not have been expressed simultaneously. It is known that some present-day cyanobacteria switch from a predominantly phycobiliprotein antenna to a predominantly chloro-

phyllous antenna under conditions of nutrient deprivation [49]. Modern procaryotic and eucaryotic algal descendents would have retained one or the other antennae system, but no cyanobacteria have yet been described which exhibit both potentials. This scenario nicely explains why all oxygen-evolving procaryotes and plastids form a single cluster when analyzed on the basis of 16S rRNA similarities [18,47].

We further hypothesize that the ancestral endosymbiont might have been similar to a modern-day *Nostoc* species, since these cyanobacteria are known to enter into a wide range of symbiotic associations with diverse hosts [50]. In support of this proposal, the 16S rRNA sequences of Giovannoni et al. [17] indicate that cyanelles, and by extension, chloroplasts, are more closely related to several filamentous cyanobacterial species than to unicellular species which they examined. Moreover, the cyanellar *atpI* locus appears to be more similar in organization and sequence to that of *Anabaena* (*Nostoc*) sp. PCC7120 than to that of *Synechococcus* sp. PCC6301 ([51-53], Bryant, D.A., Annarella, M.B. and Stirewalt, V.L., in preparation).

Acknowledgements: The authors gratefully acknowledge the assistance of Franz Suter in learning the University of Wisconsin data analysis programs, Iwan Bissig for his assistance with manuscript preparation, and Dr Herbert Zuber and the Institut für Molekularbiologie and Biophysik for their hospitality during the preparation of this manuscript. This work was supported by NSF Grant DMB-8504294, NIH Grant GM-31625, and Pennsulvania State University Experiment Station Project 2874 to D.A.B.

REFERENCES

[1] Trench, R.K. (1982) in: Progress in Phycological Research, vol. 1 (Round, F.E. and Chapman, D.J. eds) pp. 257–288, Elsevier, Amsterdam.
[2] Wasmann, C.C., Loffelhardt, W. and Bohnert, H.J. (1987) in: The Cyanobacteria (Fay, P. and Van Baalen, C. eds) pp. 303–324, Elsevier, Amsterdam.
[3] Jaynes, J.M. and Vernon, L.P. (1982) Trends Biochem. Sci. 7, 22–24.
[4] Bryant, D.A. (1987) in: Photosynthetic Picoplankton (Platt, T. and Li, W.K.W. eds) pp. 423–500, Canad. Bull. Aquatic Sci, vol. 214, Dept. Fisheries Oceans, Ottawa, Canada.
[5] Lemaux, P.G. and Grossman, A. (1985) EMBO J. 4, 1911–1919.
[6] Bryant, D.A., De Lorimier, R., Lambert, D.H., Dubbs, J.M., Stirewalt, V.L., Stevens, S.E. jr., Porter, R.D., Tam, J. and Jay, E. (1985) Proc. Natl. Acad. Sci. USA 82, 3242–3246.
[7] Giddings, T.J. jr., Wassman, C. and Staehelin, L.A. (1983) Plant Physiol. 71, 409–419.
[8] Aitken, A. and Stanier, R.Y. (1979) J. Gen. Microbiol. 112, 219–223.
[9] Herdman, M. and Stanier, R.Y. (1977) FEMS Microbiol. Lett. 1, 7–12.
[10] Lambert, D.H., Bryant, D.A., Stirewalt, V.L., Dubbs, J.M., Stevens, S.E. jr. and Porter, R.D. (1985) J. Bacteriol. 153, 659–664.
[11] Bohnert, H.J., Michalowski, C., Bevacqua, S., Mucke, H. and Loffelhardt, W. (1985) Mol. Gen. Genet. 201, 565–574.
[12] Ko, K., Jaynes, J.M. and Strauss, N.A. (1985) Plant Sci. 42, 115–123.
[13] Palmer, J.D. (1985) Annu. Rev. Genet. 19, 325–354.
[14] Kuntz, M., Crouse, E.J., Mubumbila, M., Burkard, G., Weil, J.-H., Bohnert, H.J., Mucke, H. and Loffelhardt, W. (1984) Mol. Gen. Genet. 194, 508–512.
[15] Bryant, D.A. (1988) in: Light-Energy Transduction in Photosynthesis: Higher Plants and Bacterial Models (Stevens, S.E. jr. and Bryant, D.A. eds) pp. 69–90, Am. Soc. Plant Physiol., Rockville.
[16] Starnes, S.M., Lambert, D.H., Maxwell, E.S., Stevens, S.E. jr., Porter, R.D. and Shively, J.M. (1985) FEMS Microbiol. Lett. 28, 165–169.
[17] Giovannoni, S.J., Turner, S., Olsen, G.J., Barns, S., Lane, D.J. and Pace, N.R. (1988) J. Bacteriol. 170, 3584–3592.
[18] Cantrell, A. and Bryant, D.A. (1987) Photosyn. Res. 16, 65–81.
[19] Maniatis, T., Fritsch, E.F. and Sambrook, J. (1982) Molecular Cloning: A Laboratory Manual, Cold Spring Harbor Laboratory, Cold Spring Harbor.
[20] Ausubel, F.M., Brent, R., Kingston, R.E., Moore, D.D., Seidman, J.G., Smith, J.A. and Struhl, K. (1987) Current Protocols in Molecular Biology, Wiley, New York.
[21] Cantrell, A. and Bryant, D.A. (1987) Plant Mol. Biol. 9, 453–468.
[22] Conrad, B. and Mount, D.W. (1982) Nucleic Acids Res. 10, 31–38.
[23] Dereveux, J., Haeberli, P. and Smithies, O. (1984) Nucleic Acids Res. 12, 387–395.
[24] Wada, A. and Sako, T. (1987) J. Biochem. 101, 817–820.
[25] Ohyama, K., Fukuzawa, H., Kohchi, T., Shirai, H., Sano, T., Sano, S., Shirai, H., Umesono, K., Shiki, Y., Takeuchi, M., Chang, Z., Aota, S.-I., Inokuchi, H. and Ozeki, H. (1986) Nature 322, 572–574.
[26] Shinozaki, K., Ohme, M., Tanaka, M., Wakasugi, T., Hayashida, N., Matsubayashi, T., Zaita, N., Chunwongse, J., Obokata, J., Yamaguchi-Shinozaki, K., Ohto, C., Torazawa, K., Meng, B.Y., Kusuda, J., Takaiwa, F., Kato, A., Tohdoh, N., Shimada, H and Suguira, M. (1986) EMBO J. 5, 2043–2049.
[27] Manzara, T. and Hallick, R. (1987) Nucleic Acids Res. 15, 3927.
[28] Wittmann-Liebold, B. and Seib, C. (1979) FEBS Lett. 103, 61–65.
[29] Janssen, I., Mucke, H., Loffelhardt, W. and Bohnert, H.J. (1987) Plant Mol. Biol. 9, 479–484.
[30] Stirewalt, V.L. and Bryant, D.A. (1989) Nucleic Acids Res., in press.
[31] Christopher, D.A. and Hallick, R.B. (1989) Nucleic Acids Res. 17, 7591–7608.
[32] Kimura, J. and Kimura, M. (1987) FEBS Lett. 210, 85–90.
[33] Henkin, T.M., Moon, S.H., Mattheakis, L.C. and Nomura, M. (1989) Nucleic Acids Res. 17, 7469–7486.
[34] Cerretti, D.P., Dean, D., Davis, G.R., Bedwell, D.M. and Nomura, M. (1983) Nucleic Acids Res. 11, 2599–2616.
[35] Ohkubo, S., Muto, A., Kawauchi, Y., Yamao, F. and Osawa, S. (1987) Mol. Gen. Genet. 210, 314–322.
[36] Kimura, J. and Kimura, M. (1987) J. Biol. Chem. 262, 12150–12157.
[37] Kimura, M., Rawlings, N. and Appelt, K. (1981) FEBS Lett. 136, 58–64.
[38] Grunberg-Manago, M. (1987) in: *Escherichia coli* and *Salmonella typhimurium:* Cellular and Molecular Biology, vol. 2 (Neidhardt, F.C. et al. ed.) pp. 1386–1409, Am. Soc. Microbiol., Washington.
[39] Hiratsuka, J., Shimada, H., Whittier, R., Ishibashi, T., Sakamoto, M., Mori, M., Kondo, C., Honji, Y., Sun, S.-R., Meng, B.-Y., Li, Y.-Q., Kanno, A., Nishizawa, Y., Hirai, A., Shinozaki, K. and Suguira, M. (1989) Mol. Gen. Genet. 217, 185–194.
[40] Zurawski, G. and Zurawski, S.M. (1985) Nucleic Acids Res. 13, 4521–4526.

[41] Bedwell, D., Davis, G., Gosink, M., Post, L., Nomura, M., Kestler, H., Zengel, J.M. and Lindahl, L. (1985) Nucleic Acids Res. 13, 3891-3903.
[42] Evrard, J.L., Kuntz, M. and Weil, J.H. (1989) J. Mol. Evol., in press.
[43] Fukuzawa, H., Kohchi, T., Sano, T., Shirai, H., Umesono, K., Inokuchi, H., Ozeki, H. and Ohyama, K. (1988) J. Mol. Biol. 203, 333-351.
[44] Meng, B.Y., Tanaka, M., Wakasugi, T., Ohme, M., Shinozaki, K. and Sugiura, M. (1988) Curr. Genet. 14, 395-400.
[45] Kirsch, W., Seyer, P. and Herrman, R.G. (1986) Curr. Genet. 10, 843-855.

Note added in proof: Additional sequence characterization of the region upstream from *rpl5* suggests that the gene order in this locus include additional ribosomal protein genes and is *5' rps3-rpl16-rps17-rpl14-rpl5-rps8-rpl6 3'*.

[46] Auer, J., Spicker, G. and Bock, A. (1989) J. Mol. Biol. 209, 21-36.
[47] Turner, S., Burger-Wiersma, T., Giovannoni, S.J., Mur, L.R. and Pace, N.R. (1989) Nature 337, 380-382.
[48] Morden, C.W. and Golden, S.S. (1989) Nature 337, 382-385.
[49] Riethman, H., Bullerjahn, G., Reddy, K.J. and Sherman, L.A. (1988) Photosyn. Res. 18, 133-162.
[50] Whitton, B.A. (1972) in: The Biology of the Blue-Green Algae (Carr, N.G. and Whitton, B.A. eds) pp. 415-434, Blackwell Scientific, Oxford.
[51] McCarn, D.F., Whitaker, R.A., Alam, J., Brba, J.M. and Curtis, S.E. (1988) J. Bacteriol. 170, 3448-3458.
[52] Cozens, A.L. and Walker, J.E. (1987) J. Mol. Biol. 194, 359-383.
[53] Curtis, S.E. (1988) in: Light-Energy Transduction in Photosynthesis: Higher Plants and Bacterial Models (Stevens, S.E. jr. and Bryant, D.A. eds) pp. 305-319, Am. Soc. Plant Physiol., Rockville.