Nucleotide sequence of the gene for the P₆₈₀ chlorophyll α apoprotein of the photosystem II reaction center from spinach

Julia Morris and Reinhold G.Herrmann*

Botanisches Institut der Universität Düsseldorf, 4000 Düsseldorf 1, FRG

Received 31 January 1984; Accepted 28 February 1984

ABSTRACT

The DNA sequence of 2210 nucleotides including the gene for the "51 kd" chlorophyll α -conjugated thylakoid membrane protein associated with the photosystem II reaction center of spinach has been determined. This protein is functionally identical with the P₆₈₀ chlorophyll α apoprotein that catalyses the primary light-induced photochemical processes of photosystem II (Camm, E.L. and Green, B.R. (1983) *Biochim. Biophys. Acta* 724 291-293). The only large open reading frame in the sequence consists of 508 triplets encoding a protein of molecular mass of 56,246 kd. The deduced amino acid sequence shows clustering of hydrophobic residues into seven core regions which probably traverse the membrane, and a large hydrophilic domain of about 200 amino acids interspersed between span VI and VII. Potential transcription promotor and terminator signals flanking the structural gene show prokaryotic-like features. Seven discrete RNA species ranging in size from 2.0 to over 5.0 kilobases display complementarity to apoprotein coding sequences implying that the region can be polycistronically transcribed. The primary transcript includes information for at least two further genes coding for subunits of the cytochrome b/f complex.

INTRODUCTION

The fine structure analysis of genes for polypeptides of the thylakoid membrane permits the dissection of the process of photosynthetic energy transduction and electron transport, and also enables a deeper understanding of the molecular basis for coordinate expression of nuclear and organellar genes during the biogenesis of this membrane. The watersplitting photosystem II in chloroplasts exists as a multisubunit structure which in spinach is composed of at least 7 nonidentical polypeptide species (1). Four of these components form a membrane spanning core that non-covalently complexes all photosynthetic and photoprotection pigments of the center; three peripherally associated proteins of M_r 34, 23 and 16 kd are located in the thylakoid lumen and appear to participate in water photolysis (2). Chlorophyll α and β -carotene are exclusively associated with two core proteins of 51 and 44 kd frequently designated CP III or CP a-1 and CP IV or CP a-2,

respectively, in their association with pigment (3, 4). Recent evidence suggests that the 51 kd protein bears both the reaction center pigment, P_{680} , that catalyzes the primary light-induced charge separation, and the pheophytin molecule which acts as an intermediate between P_{680} and the primary electron acceptor plastoquinone Q (4-7). The other core components are the herbicide-binding "32 kd" polypeptide at the reducing side of photosystem II and cytochrome b-559 (10 kd).

The gene coding for the "32 kd" herbicide-binding protein has been located on plastid chromosomes (8, 9) and sequenced (10). We have physically mapped single genes for the 51 and 44 kd proteins and cytochrome b-559 on the circular plastid chromosomes of spinach, *Oenothera* and tobacco (1, 11, 12), and shown that the three luminal proteins derive from nuclear genes (13). The gene for the P_{680} chlorophyll α apoprotein has been located close to the genes for cytochrome b6 and subunit 4 of the cytochrome complex (1, 14), some 40 kbp away from the coding sequence for the 44 kd chlorophyll α -binding protein. This analysis has also alerted us to the highly interesting possibility that posttranscriptional modification in chloroplasts could be of major importance in the biogenesis of the thylakoid membrane (1, 14).

We report here the amino acid sequence of the P_{680} chlorophyll α apo-protein as derived from nucleotide sequence analysis of the gene as well as refined information bearing on the complex transcriptional organization of this region of the plastid chromosome. Information on the genes for cytochrome b6 and subunit 4 and an evaluation of their protein sequences will be the subject of separate communications (15, 16).

MATERIALS and METHODS

Restriction endonucleases Ava I, Bam HI, Hinf I, Hind III, Hpa II, Sal I, Taq I, and Xba I were purchased from New England Biolabs (Bad Schwalbach); DNA polymerase I, its Klenow fragment, and restriction endonuclease Eco RI were from Boehringer (Mannheim). The ^{32}P α -labelled deoxyribonucleotide triphosphates dCTP and dATP (3000 Ci/mmol) were obtained from Amersham (Braunschweig).

Plasmids

The recombinant plasmids pWHsp 207 and pWHsp 208 carrying the 5.6 and 5.2 kbp Sal I primary fragments Sal I-7 and Sal I-8 from spinach (*Spinacia oleracea*) plastid DNA inserted into the Sal I site of pBR322 were selected from a shot-gun library (11, 14). Useful subclones were created by ligating Eco RI fragments from Sal I-7 into the Eco RI site of pBR325. They were designated

pWHsp 207/E1 to pWHsp 207/E4 in order of decreasing molecular mass of the inserts (1, 14).

A 4.6 kbp Bam HI primary fragment which represents one of two nonidentical fragments that constitute the two-molar band Bam HI-8 forms the overlap between fragments Sal I-7 and Sal I-8 (Whitfeld, personal communication; 1, 14). Attempts to purify this fragment, Bam HI-8a, from a shot-gun cloning experiment were unsuccessful; only pWHsp 408b was recovered. To obtain the recombinant DNA pWHsp 408a ultimately, a Bam HI digest of total spinach plastid DNA was fractionated by agarose gel electrophoresis, Bam HI-8 DNA electroeluted, purified and cloned in the Bam HI site of pBR325 and pBR322. The clone was recovered with a frequency of 1:20 compared to pWHsp 408b. Alternatively, the 1.6 kbp primary fragment Pvu I-10 also containing the overlap between fragments Sal I-7 and Sal I-8 (17) was isolated and ligated into the Pvu I site of pBR322 creating the recombinant clone pWHsp 510.

Plasmid DNA was prepared by the cleared lysate method (18) followed by polyethylene glycol precipitation of the nucleic acids and centrifugation to equilibrium in cesium chloride-ethidium bromide gradients.

Nucleic Acid Hybridization

Plastid RNA for Northern blots was prepared as described from etiolated spinach seedlings exposed to 18 hours light (1). The sizes of the RNA species containing the coding sequence for the P₆₈₀ chlorophyll *a* apoprotein were determined by hybridization of nick-translated DNA fragments (19) to Northern blots (20).

DNA-Sequence Analysis

DNA fragments were prepared from plasmid DNA by restriction endonuclease digestion, separated on agarose gels and electroeluted. Labelling at the 3'-ends was performed with ³²P α -labelled deoxyribonucleotides in 10 μ l of 50 mM Tris-HCl (pH 7.2), 10 mM MgSO₄, 1 mM DTT with DNA polymerase I large fragment (Klenow fragment, 2.5 units) for one hour at 20°C. After restriction endonuclease digestion or denaturation in 30 % dimethyl sulfoxide (21) the double- or single-stranded fragments, respectively, were separated on polyacrylamide gels and recovered after extraction into a salt buffer. Base-specific chemical cleavage reactions (C, C + T, A + G, G) were carried out according to Maxam and Gilbert (21). The products were fractionated on 6 % and 20 % polyacrylamide gels (0.3 mm thick) containing 7 M urea. Gels were exposed to Kodak X-Omat AR or Fuji X-ray film with Cronex Quanta III intensifying screens at -70°C. DNA and amino acid sequence analyses were performed on a Telefunken 445 computer.

Sequence Analysis of the Gene for the P₆₈₀ Chlorophyll α Apoprotein

The sequence AAGGGG located 8 to 13 base pairs proximal to the ATG start codon might function as a ribosome binding site (22) by pairing with the sequence 5'-CCUCC-3' at the 3' end of plastid 16 S rRNA during translation initiation (23, 24). At positions -187 to -182 upstream from the first methionine codon, the sequence TAGAAT which shows similarity with the -10 bacterial promotor consensus sequence TATAAT (25) can be found. Eighteen base pairs away a corresponding -35 position sequence TTGCGT (positions -211 to

Figure 1. Upper part: Partial restriction map of fragments Sal I-7 and Sal I-8 of spinach plastid DNA. Fragment number is in order of decreasing size (Sal I, Xho I, Bam HI) or in base pairs (Eco RI). Thickened line indicates the location of the P₆₈₀ chlorophyll *a* apoprotein gene. Lower part: Detailed map of the 2210 bp interval containing the P₆₈₀ chlorophyll *a* apoprotein gene and flanking sequences. The thickened arrow indicates its direction of transcription; small arrows illustrate direction and extent of individual DNA sequence analyses.

-206) also shows features in common with the *E. coli* consensus sequence TTGACA (25, 26). A thirty-nine base pair stem-loop structure (Fig. 3b) located 295 base pairs distal to the TAA stop codon has as free energy of formation $\Delta G = -41.8$ kcal calculated according to Tinoco et al. (27). Similar structures in bacterial DNA have been associated with transcription termination (26) and RNA processing (28).

Other remarkable features of the DNA sequence include (1) a seven base pair repeat TTGTTCC present three times within a span of sixty-six nucleotides between the putative promotor region, and the start of the structural gene (positions -168 to -162, -150 to -144, -105 to -102); this sequence is also present once within the structural gene (578-584); (2) a pair of nine base pair direct repeats of the sequence TATGGAAGC occurring 140 and 164 base pairs distal to the TAA stop codon (positions 1659-1667, 1693-1701), and (3) a secondary structure with a seven bp stem and a seventeen bp loop located between the putative promotor and the structural gene (Fig. 3a).

Transcriptional Organization of the P₆₈₀ Apoprotein Gene

Hybrid selection experiments employing immobilized cloned Eco RI sub-fragments from fragment Sal I-7 and from adjacent regions of fragment Sal I-8 had disclosed that a coherent 7 kbp region was capable of selecting RNA species whose translation products reacted immunologically with antisera to the 51 kd protein (1). Northern blot analysis of spinach ptRNA probed with nick-translated fragments Sal I-8 or Sal I-7 revealed multiple banded patterns with discrete species ranging from 2.0 to more than 5.0 kb (Fig. 4). Hybridization of nitrocellulose strips from the same gel with LSU- or 32 kd-gene probes gave the known single reacting RNA species of 1.7 and 1.25 kb (Bisanz unpublished data, 10, 29), suggesting that the variable sizes of the RNA species transcribed from the Sal I-8/Sal I-7 segment result from posttranscriptional processing or variable read-through of transcriptional termination signals rather than nonspecific secondary degradation.

In order to more exactly define the relationship among these giant RNA species four nick-translated DNA fragments were hybridized to Northern strips of spinach plastid RNA from plant material which was actively synthesizing thylakoid membranes and hence would be expected to possess a relatively high concentration of RNA species for thylakoid membrane proteins (1). Fragment A contains 160 bp at the amino terminus, fragment B 622 bp at the carboxyl terminus plus 340 bp of 3' flanking sequence of the gene for the P₆₈₀ chlorophyll α apoprotein, fragment C, the primary fragment Xho I-17, 600 bp of the cytochrome b6 structural gene and fragment D 400 bp of the subunit 4 gene.

-350-300
 CTGTATAGATTAGAAAGGTTTAGTTAAAGATTTAAGTTAAGAAGACAGA
-250
 ATAAATAAAGGAAAATTCCTACCAACATAGCCTTCCAATTATGAATAAGTATGAAAGCATTCTACTCATAGACGAT
-200
 GCTATCAACTCCGATTTGGGATTTCCTACTTATOGAGTATAGAAATAGATTTCCTTCCTTCCTTCTACAAACATA
-150
 ATTGTTCATTATTACCAACAGATAGAACAAACATTAAACCTTGTTCCTGAGATAATCCATTGAACAAAAGGATC
-100
-50
 CATTGCATAGTCATAGTCTTTTCCAATGCAATAAAGTTACATAGTGTTCATTTTCTTTTGATAAAGGGGTATTTTCC

130

ATG GGT TTG CCT TGG TAT CGT GTT CAT ACT GTT GTA TTG AAT GAT CCC GGC CGG TTG
 MET Gly Leu Pro Trp Tyr Arg Val His Thr Val Leu Asn Asp Pro Gly Arg Leu
 6090

ATT TCT GTC CAT ATC ATG CAT ACA GCT CTA GTT GCT GGT TGG GCC GGT TCG ATG GCT
 Ile Ser Val His Ile Met His Thr Ala Leu Val Ala Gly Trp Ala Gly Ser Met Ala
 120150

CTA TAT GAA TTA GCT GTT TTT GAT CCC TCT GAC CCC GTT CTT GAT CCA ATG TGG AGA
 Leu Tyr Glu Leu Ala Val Phe Asp Pro Ser Asp Pro Val Leu Asp Pro Met Trp Arg
 180210

CAG GGT ATG TTC GTT ATA CCT TTC ATG ACT CGT TTA GGA ATA ACC AAT TCG TGG GGC
 Gln Gly Met Phe Val Ile Pro Phe Met Thr Arg Leu Gly Ile Thr Asn Ser Trp Gly
 240270

GGT TGG AGT AIT ACA GGG GGT ACT ATA ACG GAT CCT AGC AIT TGG AGT TAC GAA GGT
 Gly Trp Ser Ile Thr Gly Gly Thr Ile Thr Asp Pro Ser Ile Trp Ser Tyr Glu Gly
 300330

GTG GOC GGA GOG CAT AIT ATG TTT TCG GGC TTG TGC TTT TTG GOC GCT AIT TGG CAT
 Val Ala Gly Ala His Ile Met Phe Ser Gly Leu Cys Phe Leu Ala Ala Ile Trp His
 360390

TGG GTA TAT TGG GAT TTA GAA ATA TTT TCG TAG GAA CGT ACA GGA AAA CCT TCT TTG
 Trp Val Tyr Trp Asp Leu Glu Ile Phe Ser Asp Glu Arg Thr Gly Lys Pro Ser Leu
 420450

GAT TTG OCT AAG AIT TTT GGA AIT CAT TTA TTT CTT TOC GGG GTG GCT TGT TTT GGT
 Asp Leu Pro Lys Ile Phe Gly Ile His Leu Phe Leu Ser Gly Val Ala Cys Phe Gly
 480510

TTT GGT GCA TTT CAT GTA ACA GGC TTG TAC GGG CCC GGA ATA TGG GTA TOC GAT CCT
 Phe Gly Ala Phe His Val Thr Gly Leu Tyr Gly Pro Gly Ile Trp Val Ser Asp Pro
 540570

TAT GGA CTA ACT GGA AAA GTA CAA CCT GTA TGT TCA GCA TGG GGC GTG GAA GGT TTT
 Tyr Gly Leu Thr Gly Lys Val Gln Pro Val Cys Ser Ala Trp Gly Val Glu Gly Phe
 600630

GAT OCT TTT GTT CCA AGA GGA ATA GOC TCT CAT CAT AIT GCA GCA GGA ACA TTA GGT
 Asp Pro Phe Val Pro Arg Gly Ile Ala Ser His His Ile Ala Ala Gly Thr Leu Gly
 660690

ATA TTA GCA GGT CTA TTC CAT CTT AGT GTT OGT TCG OCT CAA CGT TTA TAC AAA GGA
 Ile Leu Ala Gly Leu Phe His Leu Ser Val Arg Ser Pro Gln Arg Leu Tyr Lys Gly
 720750

TTA CGT ATG GGT AAC AIT GAA ACC GTT CTT TCT AGT AGT ATC GCT GCT GTC TTT TTT
 Leu Arg Met Gly Asn Ile Glu Thr Val Leu Ser Ser Ser Ile Ala Ala Val Phe Phe
 780810

GCA GCT TTT GTT GTT GCG GGA ACT ATG TGG TAT GGT TCC GCA ACT ACC CCG ATC GAA
 Ala Ala Phe Val Val Ala Gly Thr Met Trp Tyr Gly Ser Ala Thr Thr Pro Ile Glu
 840870

TTA TTT GGC CCC ACT CGT TAT CAA TGG GAT CAG GGA TAC TTT CAG CAA GAA ATA TAC
 Leu Phe Gly Pro Thr Arg Tyr Gln Trp Asp Gln Gly Tyr Phe Gln Gln Glu Ile Tyr

870 900
 CGA CGA GTA AGT GCT GGG CTA GCC GAA AAT CAA AGT TTC TCA GAA GCT TGG TCA AAA
 Arg Arg Val Ser Ala Gly Leu Ala Glu Asn Gln Ser Phe Ser Glu Ala Trp Ser Lys
 930 960
 ATT CCC GAG AAA TTG GCT TTT TAT GAT TAC ATT GGA AAT AAT CCG GCG AAA GGG GGA
 Ile Pro Glu Lys Leu Ala Phe Tyr Asp Tyr Ile Gly Asn Asn Pro Ala Lys Gly Gly
 990 1020
 TTA TTT AGA GCA GGC TCA ATG GAC AAC GGC GAT GGA ATA GCT GTT GGA TGG TTA GGA
 Leu Phe Arg Ala Gly Ser Met Asp Asn Gly Asp Gly Ile Ala Val Gly Trp Leu Gly
 1050 1080
 CAC OCT ATT TTT AGA GAT AAA GAA GGA CGT GAA CTC TTT GTA CGA CGT ATG CCT ACT
 His Pro Ile Phe Arg Asp Lys Glu Gly Arg Glu Leu Phe Val Arg Arg Met Pro Thr
 1110 1140
 TTT TTT GAA ACC TTT CCG GTC GTT TTG ATA GAT GGG GAC GGA ATT GTT CGA GCT GAT
 Phe Phe Glu Thr Phe Pro Val Val Leu Ile Asp Gly Asp Gly Ile Val Arg Ala Asp
 1170
 GTT OCT TTT AGA AGA GCA GAA TCT AAG TAT AGC GTT GAA CAA GTA GGT GTA ACT GTT
 Val Pro Phe Arg Arg Ala Glu Ser Lys Tyr Ser Val Glu Gln Val Gly Val Thr Val
 1200 1230
 GAG TTT TAC GGC GGC GAA CTA AAC GAA GTT AGT TAT AGC GAT CCT GCT ACT GTC AAA
 Glu Phe Tyr Gly Gly Glu Leu Asn Glu Val Ser Tyr Ser Asp Pro Ala Thr Val Lys
 1260 1290
 AAA TAT GCT AGA CGT GCT CAA TTA GGT GAA ATT TTC GAA TTA GAC CGT GCT ACT TTG
 Lys Tyr Ala Arg Arg Ala Gln Leu Gly Glu Ile Phe Glu Leu Asp Arg Ala Ser Leu
 1320 1350
 AAA TCA GAT GGT GTT TTT CGC AGT AGT CCG AGG GGT TGG TTT ACT TTT GGA CAT GCT
 Lys Ser Asp Gly Val Phe Arg Ser Ser Pro Arg Gly Trp Phe Ser Phe Gly His Ala
 1380 1410
 TTC TTT GCC CTA CTC TTC TTC GGA CAC ATT TGG CAT GGG TCT AGA ACC TTG TTC
 Ser Phe Ala Leu Leu Phe Phe Phe Gly His Ile Trp His Gly Ser Arg Ser Leu Phe
 1440 1470
 AGA GAT GTT TTT GCT GGG ATT GAC CCC GAT TTG GAT GTT CAA GTA GAA TTT GGC GCA
 Arg Asp Val Phe Ala Gly Ile Asp Pro Asp Leu Asp Val Gln Val Glu Phe Gly Ala
 1500 1530
 TTC CAA AAA ATT GGA GAT CCA ACT ACA AGA AGA CAA GGA GTT TAA TACAACATTGCTTTA
 Phe Gln Lys Ile Gly Asp Pro Ser Ser Arg Arg Gln Gly Val -c-
 1580
 TGATTTTTTGGCTCTATTTTTTTATGGTTTGCACATAGGATACTAGAGCAATCTTGATTAAATCACCGGCTTTTC
 1630 1680
 TTTTTTTTTTTTTCATTATTTGGGAAAAATAGTCCCAAGTAAACAGGTATGGAAGCTATAATTGTAAACACCAATCGA
 1730
 ATCTATGGAAGCATTGGTTTATACATTCTTATTAGTCTCTACTTAGGGATTATTTTTTTTGGCATCTTTTTTCGG
 1780 1830
 GAACTCTGAAATTTCAACTGAAAAATGAAATGATTTTTTCATTATTTCAAATTGACGTAATGAGCCTTCCAATAT
 1860
TGGAAGGCTCATTGCTCGAC

Figure 2. DNA nucleotide sequence of the non-transcribed strand of the
 P₆₈₀ chlorophyll *a* apoprotein gene and flanking regions from *S. oleracea*.
 The corresponding amino acids are indicated. Numbering of nucleotides
 starts 350 bp proximal to the first methionine codon of open reading frame.
 An arrow marks the putative transcription start site. The -10 and -35
 regions of the putative promoter are boxed and nucleotides which can base-
 pair to form stem-loop structures (Fig. 3) are underlined.

Figure 3. Proposed stem-loop secondary structures located in the (a) 5' and (b) 3' flanking sequences to the P₆₈₀ chlorophyll *a* apoprotein gene. Numbering of nucleotides follows Figure 2. Free energy for formation, ΔG , was calculated according to Tinoco et al. (27).

Figure 4. (a) Diagram showing location of DNA fragments A, B, C and D in relation to the genes for the P₆₈₀ chlorophyll *a* apoprotein, cytochrome b6 and subunit 4 of the cytochrome complex. (b) Hybridization between nick-translated fragments A - D and Northern blots of spinach plastid RNA. (A) Fragment Bam HI-Hpa II, 160 bp; (B) fragment Sal I-Hind III, 955 bp; (C) fragment Xho I-17, 205 bp; (D) fragment Hpa II-Hpa II, 300 bp. Fragment A was hybridized to a Northern blot from a different gel.

The location of these fragments is depicted in Fig. 4.

All four fragments hybridize to RNA species which are larger than 4 kb in size. Strikingly, all four fragments also hybridize strongly to a component of approximately 2.5 kb although fragments A and D are more than 4000 bp apart. In addition, fragments A and B hybridize with components of 2000 and 2100 bp, fragment C and D with components of 1500 and 1300 bp, and fragment C alone with a species around 900 bp in size.

DISCUSSION

The nucleotide sequence of 2210 base pairs of spinach plastid DNA revealed one large non-interrupted open reading frame of 1524 nucleotides. Its deduced amino acid sequence codes for a 508 amino acid residue protein of molecular mass 56,246. This figure deviates significantly from the apparent molecular weight of 51 kd for the chlorophyll *a*-conjugated protein associated with photosystem II estimated according to electrophoretic mobility in SDS-containing polyacrylamide gels. The protein does not appear to be synthesized as a precursor after in vitro translation using either an RNA-programmed rabbit reticulocyte system or a DNA-programmed *E. coli* assay; the largest component reacting immunologically with antisera elicited against the 51 kd protein possesses the same electrophoretic mobility in various gel systems as the authentic protein obtained from spinach thylakoid membranes (1). The discrepancy might be explainable if translation were to begin at another methionine codon within the open reading frame. If the fifth methionine residue (amino acid sixty, Fig. 2), which is preceded by a Shine-Dalgarno type ribosomal binding sequence (GGAG), were the first translational codon, the protein would have an Mr of 49,865. However, the DNA sequence of the corresponding region of the P₆₈₀ chlorophyll *a* apoprotein gene from *Eucenothera* plastome I shows 95 % conservation of the sixty amino acids between the two methionine residues although the nucleotide sequences differ by 8 % (Offermann, personal communication; 30). Nine of fourteen nucleotide differences between these two sequences are silent. This high degree of amino acid homology indicates that the full 508 amino acids are present in the primary translation product. A similar inconsistency between the size of a protein based on its electrophoretic mobility and that calculated from its deduced amino acid sequence has been noted for other hydrophobic proteins, e.g. the precursor of the herbicide-binding "32 kd" thylakoid membrane protein (10), the P₇₀₀ chlorophyll *a* apoprotein of photosystem I (Kirsch, personal communication) and beef heart mitochondrial ATP synthase subunits (31).

	T	C	A	G	
T	Phe 31	Ser 7	Tyr 10	Cys 2	T
	Phe 10	Ser 4	Tyr 7	Cys 1	C
	Leu 13	Ser 5	Stop 1	Stop -	A
	Leu 13	Ser 5	Stop -	Trp 17	G
C	Leu 4	Pro 13	His 12	Arg 11	T
	Leu 2	Pro 7	His 2	Arg 1	C
	Leu 6	Pro 3	Gln 10	Arg 4	A
	Leu 1	Pro 4	Gln 3	Arg 1	G
A	Ile 17	Thr 13	Asn 5	Ser 10	T
	Ile 3	Thr 5	Asn 3	Ser 3	C
	Ile 10	Thr 6	Lys 11	Arg 11	A
	Met 11	Thr 1	Lys 2	Arg 1	G
G	Val 22	Ala 20	Asp 21	Gly 18	T
	Val 4	Ala 6	Asp 5	Gly 11	C
	Val 11	Ala 10	Glu 20	Gly 23	A
	Val 3	Ala 3	Glu 2	Gly 8	G

Figure 5. Tabulation according to codon specificity of the 508 amino acids of the P₆₈₀ chlorophyll *a* apoprotein as deduced from the DNA sequence.

The codon usage for the 508 amino acid residues of the "51 kd" protein has been tabulated in Figure 5. Although all possible codon species are present for each amino acid, a marked preference (71%) is shown for codons having A or T in the wobble position. Such biased codon usage has been noted for other plastid genes, e.g. genes for the large subunit of ribulose 1,5-bisphosphate carboxylase in corn (32) and spinach (29), genes for the herbicide-binding "32 kd" thylakoid membrane protein in spinach and tobacco (10), as well as spinach and pea genes for ATP synthase subunits (33-35).

The amino acid composition of the P₆₈₀ chlorophyll *a* apoprotein displays hydrophobic character (48%) expected for an intrinsic membrane polypeptide. The hydropathy profile calculated according to Kyte and Doolittle (36) exhibits domain structure in which the hydrophobic residues are clustered into seven discrete core regions probably traversing the membrane in the form of α -helices (Fig. 6). A large hydrophilic domain of nearly 200 amino acid residues, that is more than one third of the protein, is interspersed between spanning segments VI and VII. It is tempting to speculate that this segment, which contains four pairs of arginine residues, associates with the water-splitting polypeptides. Core segments VI and VII and a small hydrophobic region at amino acid residues 360-370 may serve to anchor this section within the lipid matrix of the membrane. Spanning segments III and IV, V and VII each contain a pair of histidine residues separated by 13 or 14 amino acids.

Figure 6. Hydropathy profile of the 508 residue amino acid sequence in which amino acid residue number is plotted against the average hydrophobicity of 11 surrounding residues, according to the method of Kyte and Doolittle (36). The average of the first eleven residues is positioned at the first amino acid, etc. Positions of membrane spanning regions are marked I-VII.

In addition, a cysteine residue is located between the histidines of segments III and IV. This situation is reminiscent of hydrophobic segments of the chloroplast cytochrome b6 which also contain pairs of histidine residues separated by 13 and 14 amino acids that are believed to coordinate heme groups (16). The carefully spaced histidine or cysteine residues in the P₆₈₀ chlorophyll *a* apoprotein might serve to mediate the noncovalent binding of trapping center chlorophyll *a* and pheophytin. A comparison between the amino acid sequence of this protein from other plant species and between other chlorophyll *a* binding proteins may clarify this point. The membrane-internal, nine amino acid sequence -his-leu-phe-leu-ser-gly-val-ala-cys-, starting at residue 142, is nearly identical to an eight residue sequence -his-phe-leu-leu-gly-val-ala-cys- from amino acid 118 in the herbicide-binding "32 kd" protein of spinach, tobacco (10), soybean (37) and *Amaranthus* (38; amino acids 82-90 in this instance). Such conserved sequences are potential candidates for plastoquinone binding sites. A more detailed description of the three-dimensional structure of the P₆₈₀ chlorophyll *a* apoprotein will be the subject of a forthcoming paper (in collaboration with Dr. Argos, Purdue University).

Prokaryotic-like promotor sequences have been identified in the 5' flanking sequences of several plastid structural genes including the gene for the large subunit of ribulose 1,5-bisphosphate carboxylase in spinach (29) and tobacco (39), the overlapping genes for the beta and epsilon subunits of chloroplast ATP synthase (40, 41), and the gene for the herbicide-binding

"32 kd" thylakoid membrane protein in spinach and tobacco (10). A possible bacterial-like (26, 42) promotor sequence TAGAAT and its associated -35 region sequence TTGCGT, separated by 18 bp, are located 182 and 206 nucleotides, respectively, proximal to the first methionine of the open reading frame. At the 3' end, a thirty-nine base pair stem-loop structure is located 295 base pairs distal to the TAA stop codon (Fig. 3). Similar stem-loop structures can be identified in the 3' untranslated regions of several plastid genes (10, 15, 29, 32, 37, 39, 40). In *E. coli* such stem-loop structures have been identified as transcription termination sites (26) and in *E. coli* T7 phage as RNAase III processing sites (43).

The genes for the P₆₈₀ chlorophyll *a* apoprotein, cytochrome b6 and subunit 4 of the cytochrome b/f complex map close together (Fig. 4) and are encoded by the same DNA strand. When RNA selected by hybridization to specific, cloned DNA fragments containing sequences from either one of these structural genes is translated in a heterologous cell-free system, all three proteins are recovered in each instance although in differing amounts (1, 14). This finding establishes that this DNA segment of the spinach plastid chromosome is polycistronically transcribed. Northern blot analysis shows, in fact, one large RNA species, over 5.3 kb in length, that could carry information for all three proteins (Fig. 4). Nucleotide sequence analysis displays three possible transcriptional termination signals, one distal to the P₆₈₀ chlorophyll *a* apoprotein gene (Fig. 3), one between the genes for cytochrome b6 and subunit 4 and one in the 3' flanking region to the latter gene (15). Hence, the longest transcript beginning at the putative promotor of the photosystem II gene probably arises by read-through at two stem-loop structures. Smaller RNA species, 2.0 kb for the P₆₈₀ chlorophyll *a* apoprotein, around 1.5 kb for cytochrome b6 plus subunit 4 and 0.9 kb for cytochrome b6 alone, could arise through either alternate transcriptional activity or processing of the larger transcripts. Separate promoters have not been identified for the genes for the cytochrome complex. However, stem-loop structures may also act as processing sites in a similar manner as those identified in *E. coli* phages (43). All RNA species which are smaller than 5 kb but contain sequences complementary to all three genes could arise by RNA processing *plus* (intercistronic) splicing. An exact size coincidence between RNA species from different segments of this DNA region is unlikely but has not been entirely ruled out.

To date, an eukaryotic-like posttranscriptional maturation has not been described for plastid mRNA. Studies on transcription of the rRNA operon in isolated chloroplasts (44) and the existence of split rRNA (45) and tRNA

genes (e.g., 46) have shown that plastids harbour enzymatic equipment for posttranscriptional endonucleolytic modification and splicing for these RNA classes. It will be interesting to unravel the details of the expression of this DNA segment and to correlate the outlined events with the biogenesis of the photosystem II reaction center and the cytochrome b/f complex.

ACKNOWLEDGEMENTS

We thank Ms. Ursula Borucki for help in the preparation of the manuscript, Ms. Barbara Schiller for assistance with photographic work. This study was supported by a grant from the Stiftung Volkswagenwerk.

*To whom correspondence should be addressed

REFERENCES

1. Westhoff, P., Alt, J. and Herrmann, R.G. (1983) *EMBO J.* 2, 2229-2237.
2. Akerlund, H.E., Jansson, C. and Andersson, B. (1982) *Biochim. Biophys. Acta* 681, 1-10.
3. Delepelaire, P. and Chua, N.-H. (1979) *Proc. Natl. Acad. Sci. USA* 76, 111-115.
4. Camm, E.L. and Green, B.R. (1983) *Biochim. Biophys. Acta* 724, 291-293
5. Nakatani, H.Y., Ke, B., Dolan, E. and Arntzen, C.T. (1983) *Biochim. Biophys. Acta*, in press.
6. Klimov, V.V., Ke, B. and Dolan, E. (1980) *FEBS Lett.* 118, 123-126.
7. Breton, J. (1982) *FEBS Lett.* 147, 16-21.
8. Bedbrook, J.R., Link, G., Coen, D.M., Bogorad, L. and Rich, A. (1978) *Proc. Natl. Acad. Sci. USA* 75, 3060-3064.
9. Driesel, A.J., Speirs, J. and Bohnert, H.J. (1980) *Biochim. Biophys. Acta* 610, 297-310.
10. Zurawski, G., Bohnert, H.J., Whitfeld, P.R. and Bottomley, W. (1982) *Proc. Natl. Acad. Sci. USA* 79, 7699-7703.
11. Herrmann, R.G., Westhoff, P., Alt, J., Winter, P., Tittgen, J., Bisanz, C., Sears, B.B., Nelson, N., Hurt, E., Hauska, G., Viebrock, A. and Sebald, W. (1983) in *Structure and Function of Plant Genomes*, Cifferi, O. and Dure, L., Eds., pp. 143-154. Academic Press, New York.
12. Westhoff, P., Widger, W.R., Alt, J. and Herrmann, R.G., in preparation.
13. Westhoff, P., Bartlett, S.G., Jansson, C., Klein-Hitpass, L., Larsson, C., Berzborn, R. and Herrmann, R.G. (1983) *Sixth Internatl. Congress on Photosynthesis, Abstracts Vol. II*, p. 330, Brussels.
14. Alt, J., Westhoff, P., Sears, B.B., Nelson, N., Hurt, E., Hauska, G. and Herrmann, R.G. (1983) *EMBO J.* 2, 979-986.
15. Heinemeyer, W., Alt, J. and Herrmann, R.G. (1984) *FEBS Lett.*, submitted.
16. Widger, W.R., Cramer, W.A., Herrmann, R.G. and Trebst, A. (1984) *Proc. Natl. Acad. Sci. USA*, in press.
17. Herrmann, R.G. and Whitfeld, P.R. (1982) in *Methods in Chloroplast Molecular Biology*, Edelman, M., Hallick, R. and Chua, N.-H., Eds., pp. 451-468, Elsevier/North-Holland, Amsterdam.
18. Katz, L., Kingsbury, D.T. and Helinski, D.R. (1983) *J. Bacteriol.* 114, 577-591.

19. Maniatis, T., Kee, S.G., Efstratiadis, A. and Kafatos, F.C. (1976) *Cell* 8, 163-182.
20. Thomas, P.S. (1980) *Proc. Natl. Acad. Sci. USA* 77, 5201.
21. Maxam, A.M. and Gilbert, W. (1980) in *Methods in Enzymology*, Grossman, L. and Moldave, K., Eds., Vol. 65, pp. 499-560, Academic Press, New York.
22. Shine, J. and Dalgarno, L. (1974) *Proc. Natl. Acad. Sci. USA* 71, 1342-1346.
23. Schwarz, Z. and Kössel, H. (1980) *Nature* 283, 739-742.
24. Tohdoh, N. and Sugiura, M. (1982) *Gene* 17, 213-218.
25. Pribnow, D. (1975) *J. Mol. Biol.* 99, 419-443.
26. Rosenberg, M. and Court, D. (1979) *Ann. Rev. Genet.* 13, 319-353.
27. Tinoco, I., Borer, P.N., Dengler, B., Levine, M.D., Uhlenbeck, O.C., Crothers, D.M. and Gralla, J. (1973) *Nature (New Biology)* 246, 40-41.
28. Altman, S., Guervier-Takada, C., Frankfort, H.M. and Robertson, H.D. (1982) in *Nucleases*, Linn, S.M. and Roberts, J.R., Eds., pp. 243-274, Cold Spring Harbor, New York.
29. Zurawski, G., Perrot, B., Bottomley, W. and Whitfeld, P.R. (1981) *Nucleic Acids Res.* 9, 3251-3270.
30. Morris, J., Offermann, K. and Herrmann, R.G. (1983) *Sixth Internatl. Congress on Photosynthesis, Abstracts Vol. 2*, 321.
31. Darley-Usmar, M.V. and Fuller, S.D. (1981) *FEBS Lett.* 135, 164-166.
32. McIntosh, L., Poulsen, C. and Bogorad, L. (1980) *Nature* 288, 556-560.
33. Zurawski, G., Bottomley, W. and Whitfeld, P.R. (1982) *Proc. Natl. Acad. Sci. USA* 79, 6260-6264.
34. Howe, C.J., Auffret, A.D., Doherty, A., Bowman, C.M., Dyer, T.A. and Gray, J.C. (1982) *Proc. Natl. Acad. Sci. USA* 79, 6903-6907.
35. Alt, J., Winter, P., Sebald, W., Moser, J.G., Schedel, R., Westhoff, P. and Herrmann, R.G. (1983) *Curr. Gen.* 7, 129-138.
36. Kyte, J. and Doolittle, R.F. (1982) *J. Mol. Biol.* 157, 105-132.
37. Spielmann, A. and Stutz, E. (1983) *Nucleic Acids Res.* 11, 7157-7167.
38. Hirschberg, J. and McIntosh, L. (1983) *Science* 222, 1346-1348.
39. Shinozaki, K. and Sugiura, M. (1982) *Gene* 20, 91-102.
40. Krebbers, E.T., Larrinua, I.M., McIntosh, L. and Bogorad, L. (1982) *Nucleic Acids Res.* 10, 4985-5002.
41. Shinozaki, K. and Sugiura, M. (1982) *Nucleic Acids Res.* 10, 4923-4934.
42. Hawley, D.K. and McClure, W.R. (1983) *Nucleic Acids Res.* 11, 2237-2255.
43. Rosenberg, M. and Kramer, R.A. (1977) *Proc. Natl. Acad. Sci. USA* 74, 984-988.
44. Bohnert, H.J., Driesel, A.J. and Herrmann, R.G. (1976) in *Genetics and Biogenesis of Chloroplasts and Mitochondria*, Bücher, T., Neupert, W., Sebald, W. and Werner, S., Eds., pp. 629-644, North-Holland Publishing Company, Oxford.
45. Rochaix, J.D. and Malnoe, P. (1978) *Cell* 15, 661-670.
46. Takaiwa, F. and Sugiura, M. (1982) *Nucleic Acid Res.* 10, 2665-2676.