Volume 9 Number 14 1981 Nucleic Acids Research

The structure of the gene for the large subunit of ribulose 1,5-bisphosphate carboxylase from spinach chloroplast DNA

Gerard Zurawski, Brigitte Perrot, Warwick Bottomley and Paul R.Whitfeld

Division of Plant Industry, CSIRO, P.O.Box 1600, Canberra City, A.C.T. 2601, Australia

Received 2 June 1981

ABSTRACT

A cloned fragment of spinach chloroplast DNA carrying the gene for the large subunit of ribulose bisphosphate (RuBP) carboxylase has been analysed by electron microscopy of R-loops, by hybridization to Northern blots of chloroplast RNA, by S1 nuclease mapping and by DNA sequencing. The transcribed region of the gene is 1690 + 3 nucleotides long and co-linear with its mRNA. It comprises a 178-179 bp 5' untranslated sequence, a 1425 bp coding region and an 85-88 bp 3' untranslated region. The deduced sequence of the 475 amino acids of the spinach large subunit protein shows 10% divergence from that of the maize large subunit protein (1). The nucleotide sequence divergence between spinach and maize over the same coding region is 16% but in the transcribed flanking regions it is 35%. Features of the spinach chloroplast gene which resemble those of bacterial genes include a 5-base Shine-Dalgarno sequence complementary to a sequence near the 3' end of chloroplast and bacterial 16S rRNA, a promoter region partially homologous to a consensus sequence of bacterial promoters, and a transcription termination region capable of forming a typical stem and loop structure.

INTRODUCTION

Ribulose 1,5-bisphosphate carboxylase, the major protein of chloroplasts, is composed of two types of subunits which together catalyse the fixation of atmospheric CO_2. One, the 55 kd large subunit, is chloroplast-DNA coded (2) and contains the catalytic site (3, 4). The other, the 12-15 kd small subunit, is nuclear coded (5) and is of unknown function. Thus, synthesis of this holoprotein depends on the expression of genes from two clearly differentiated genetic systems located in separate compartments within the one cell. Furthermore, the small subunit is synthesised on cytoplasmic ribosomes (6) and thus is the product of a typical eukaryotic system, whereas the large subunit is synthesised in the chloroplast (7) which is more analogous to a prokaryotic system. The extent to which the expression of these genes is coordinated and the nature of the mechanism whereby such coordination might be effected is far from clear. One approach to this

problem is to characterize the genes for the two protein subunits and to develop in vitro systems for their transcription and translation so that the role of possible regulatory molecules can be studied.

Synthesis of the large subunit by in vitro transcription-translation of chloroplast DNA as well as of cloned fragments of chloroplast DNA has been reported (8-12). The position of the gene on the restriction map of maize, Chlamydomonas and spinach chloroplast DNAs has been determined and clones carrying the gene from these DNAs have been constructed (8, 11, 12).

In spinach the large subunit gene lies within a 2 kb region of the 11.5 kb BamHI fragment of the chloroplast DNA (12). In this paper we describe the analysis of the 2 kb fragment by hybridization to chloroplast RNA. We also present the nucleotide sequence for 1800 bp from this region and define the coding region for the large subunit protein and the flanking transcribed regions. The nucleotide sequence and the deduced amino acid sequence for the protein are compared with the recently published corresponding sequences for the large subunit from maize chloroplasts (1).

METHODS

Plasmids were purified from cleared lysates by centrifugation in ethidium bromide CsCl gradients (13). The procedures used for the isolation of spinach chloroplast RNA, restriction enzyme digestions and agarose gel electrophoresis of DNA fragments have been described previously (14). DNA fragments were recovered from agarose gels by electroelution (15) and from acrylamide gels by diffusion into a salt buffer (16). DNA sequencing and the denaturation of fragments and separation of strands were done according to Maxam and Gilbert (16, 17). Conditions for RNA/DNA annealing and S1 nuclease digestion (18) are given in the various Figure legends.

Hybridization to Filter Bound RNA

Chloroplast RNA was electrophoresed in 1.5% agarose gels containing 5 mM methylmercuric hydroxide (19) at 5V/cm for 4-5 h and transferred to diazotized aminothiophenol paper (prepared according to the procedure developed by B. Seed, California Institute of Technology) using 0.2 M Na acetate, pH 4.0. Other conditions including those of pre- and post-hybridization washes were as described by Alwine, Kemp and Stark (20). Labeled DNA probes were hybridized to the Northern blots in 50% formamide, 0.6 M NaCl, 0.06 M Na_3 citrate, 0.2% Na dodecyl sulfate, 0.2% each of polyvinyl pyrolidone, Ficoll and bovine serum albumin and 100 µg/ml denatured sonicated calf thymus DNA, for 18 h at 45°C.

Plasmids

The preparation of pSocB149 has been described (12). The orientation of the chloroplast DNA BamHl fragment in pBR322 is such that the right-hand BamHl site (Figure 1) is close (377 bp) to the EcoRI site in the vector. EcoRI fragments of pSocB149 were subcloned into pBR325 and the desired clones selected from the tet^r, amp^r, cap^s transformants on the basis of the size of the insert. For sequencing studies pSocB149 was modified by deleting approximately 8.8 kb of the insert from the SacI site to the SalI site in pBR322 (see Figure 1). The ends of the shortened plasmid were blunt-end ligated after filling-in the SalI site and trimming the SacI site, thereby regenerating a SalI site.

Labeling of DNA

Plasmids and DNA fragments were nick-translated as described by Maniatis et al., (21). Fragments were 5' end-labeled using ^{32}P γ-ATP and T4 polynucleotide kinase after dephosphorylation with calf alkaline phosphatase (17). 3' end-labeling of restriction fragments was carried out by filling-in the single-stranded termini using ^{32}P-αdNTPs and the Klenow fragment of E. coli DNA polymerase (22). Labeled fragments were purified by gel filtration on Sephadex G75 and ethanol precipitated prior to use as hybridization probes.

R-loop Analysis

R-loop hybridization reactions (23) contained 1.5 µg/ml DNA, 17 or 34 µg/ml chloroplast RNA, 70% formamide, 0.3 M NaCl, 0.1 M PIPES, pH 8.5, 10 mM EDTA. Aliquots of 25 µl were sealed in microcapillary tubes and incubated at 54°C for 16 h. Samples were prepared for electron microscopy by dilution of 10 µl of the hybridization sample with 10 µl of spreading solution containing 70% formamide. Spreading, staining and shadowing conditions were as described by Davis, Simon and Davidson (24). Molecules were visualized and photographed with a Philips 200 electron microscope. The magnification was calibrated with a grating replica (2160 lines per mm). Length measurements were made with a Summagraphics digitizer on photographs enlarged to a print magnification of approximately 100 000 x. Measurements were converted to numbers of base pairs using ColEI plasmid (6320 bp) as a length standard.

RESULTS

Characterization of Spinach Chloroplast Large Subunit mRNA

Plasmid pSocB149, which has the 11.5 kb BamHI fragment of spinach

chloroplast DNA inserted into the BamHI site of the vector pBR322, directs the synthesis of large subunit protein in the E. coli S30 in vitro transcription-translation system (12). The susceptibility of this synthesis to prior cleavage of pSocB149 DNA by certain restriction enzymes revealed that the region coding for the large subunit protein probably lies within the 2 kb BamHI-AvaI interval (12) indicated in Figure 1. Localization of the gene entirely to this site has been confirmed by examining the R-loops formed when total chloroplast RNA is hybridized to pSocB149 linearized with SacI or SmaI. A single loop of about 1400 bp was observed (Figure 2). By measuring the distance from the loop to the ends of the linearized molecule and from a knowledge of the location of the SacI and SmaI sites in the 11.5

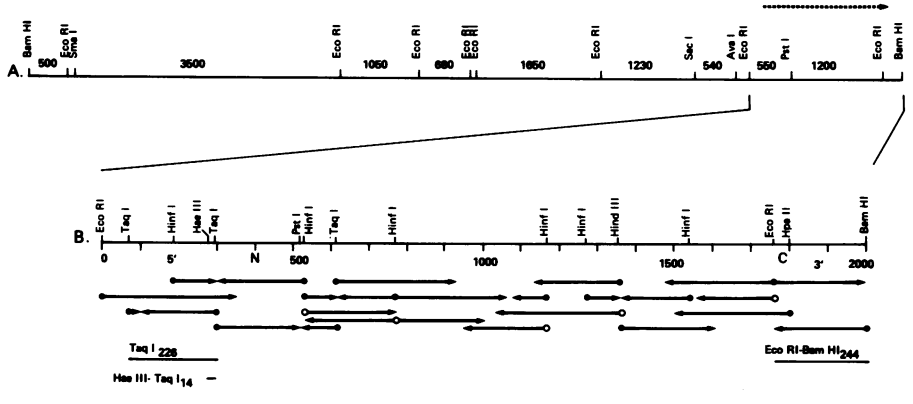


Figure 1. A Restriction Map of the 11.5 kb BamHI Fragment of Spinach Chloroplast DNA and a 2 kb Sub-fragment Containing the Large Subunit Gene. (A) The approximate locations on the 11.5 kb fragment of restriction sites for the enzymes EcoRI, SacI, SmaI and PstI are shown (distances are in base pairs) together with one relevant AvaI site. The 1750 bp and 1950 bp EcoRI fragments are those containing the PstI and SacI sites respectively. The location and direction of transcription of the large subunit gene coding region is indicated by the dashed line above the map. (B) Locations on the 2 kb EcoRI-BamHI fragment of restriction sites for the enzymes EcoRI, BamHI, HindIII, PstI, TaqI, HinfI, HpaII and one relevant HaeIII site are shown. The symbols 5', 3', N and C denote respectively the regions coding for the 5' end and the 3' end of large subunit mRNA, and the N- and C-termini of the large subunit protein. Restriction sites used for DNA sequence analysis are indicated by dots under the map; filled dots represent 3' end-labeled sites, empty dots represent 5' end-labeled sites. The arrows show the direction and extent of each sequencing run. Except for the 66 bp adjacent to the terminal EcoRI site, all 2000 bp were sequenced on both strands; sequencing was carried through all restriction sites except the terminal BamHI and the internal HindIII sites. The 3 subfragments drawn at the bottom of the diagram were used in the experiments to locate the 5' and 3' ends of the large subunit mRNA.

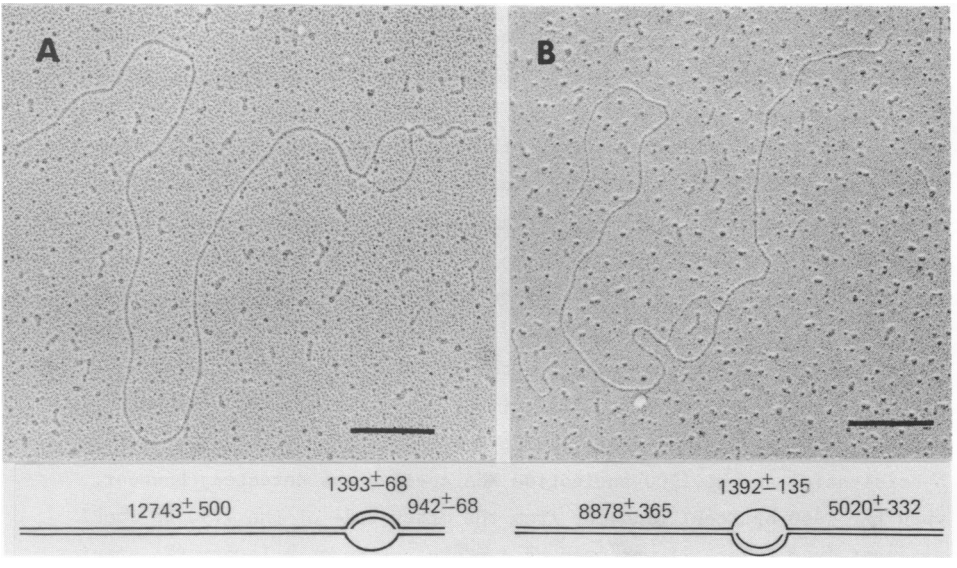


Figure 2. Electron Micrographs of R-loops Formed Between Spinach Chloroplast RNA and the Recombinant Plasmid pSocB149. The DNA was linearized by digestion with SacI (A) and SmaI (B). Line drawings show measurements in base pairs determined from 16 (A) and 22 (B) molecules. The bar is equivalent to 0.5 μm.

kb BamHI fragment (there are no sites for these two enzymes in pBR322) the R-loop was shown to coincide approximately with the 1750 bp EcoRI fragment (See Figure 1). There was no evidence of unhybridized segments in the 1400 bp R-loop, indicating that no large introns are present in the coding sequence for the large subunit protein. This result also confirmed that there is only one copy of the gene on the 11.5 kb BamHI fragment of spinach chloroplast DNA. Under the conditions used, no other R-loops were detected on pSocB149. This would suggest that transcripts of any other genes which might be present on the 11.5 kb BamHI fragment are not as abundant as those of the large subunit gene and that if R-loops for such genes are to be demonstrated then it will be necessary to use chloroplast RNA enriched for specific mRNAs.

The size of the mRNA for the large subunit as indicated by the R-loop (1400 nucleotides) is somewhat smaller than expected for a 55 kd-protein. An independent measure of its size was obtained by hybridizing the 1750 bp EcoRI fragment of pSocB149 to total chloroplast RNA which had been electrophoresed under denaturing conditions and transferred to diazotized amino-

thiophenol paper (Northern blots). The labeled probe hybridized strongly to an RNA of approximately 1700 nucleotides and faintly to smaller, yet discrete, RNAs (track 1, Figure 3). The latter are too small to code for the large subunit and presumably are breakdown products of the major 1700-nucleotide species. No RNAs larger than 1700 nucleotides hybridized to the 1750 bp EcoRI fragment and it was concluded that if the 1700-nucleotide large subunit mRNA is processed from a larger transcript then the pool size of the precursor is very small.

The same pattern of hybridization to chloroplast RNA was also obtained when the 600 bp EcoRI fragment which lies between the 1750 bp EcoRI fragment and the EcoRI site in the pBR322 vector was used as a probe (track 2, Figure 3) indicating that the transcribed region of the large subunit gene extends beyond the end of the 1750 bp EcoRI fragment (see Figure 1). No hybridization to the 1700-nucleotide RNA species was detected, however, when the 1950 bp EcoRI fragment from the other side of the 1750 bp EcoRI fragment (see Figure 1) was used as a probe, although a larger RNA species (approx. 2700 nucleotides, track 3, Figure 3) did hybridize. Compared to the large subunit mRNA, the 2700-nucleotide RNA species was present in only minor amounts.

The direction of transcription of the large subunit mRNA was determined

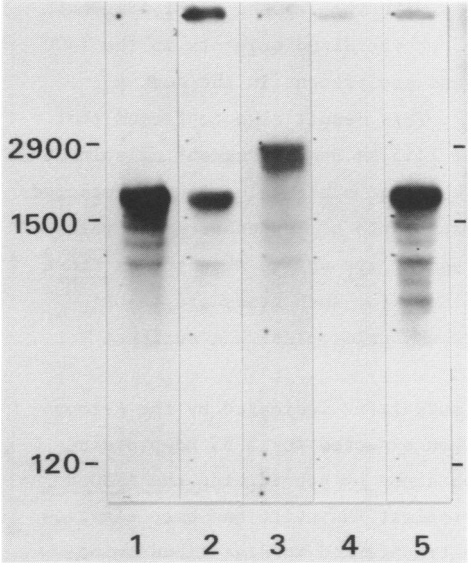


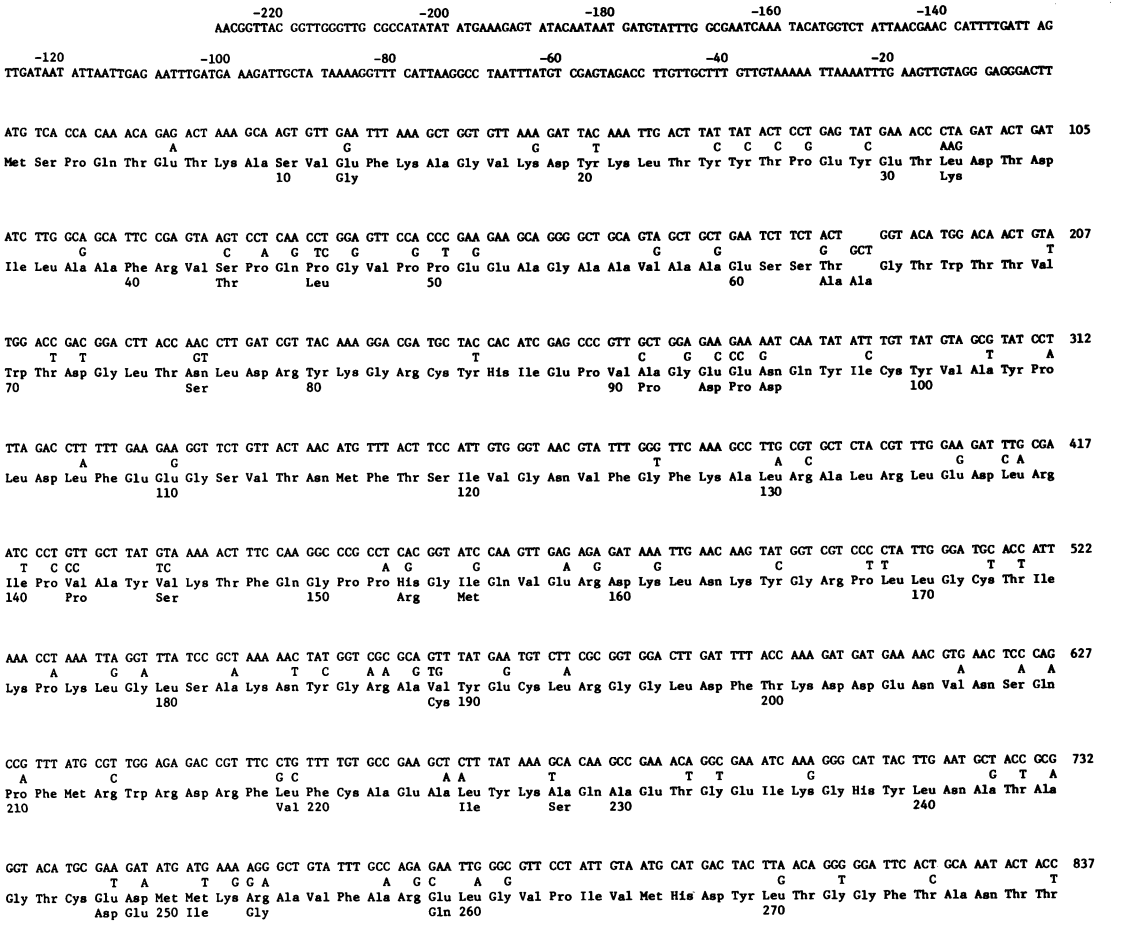
Figure 3. Determination of the Size of the mRNA for the Large Subunit Protein and the Direction of Transcription. Spinach chloroplast RNA (30 µg/slot) was electrophoresed, transferred to diazotized aminothiophenol paper, hybridized with the following ^{32}P-labeled probes and autoradiographed. (1) nick-translated 1750 bp EcoRI fragment derived from pSocB149 and recloned in pBR325 (called pSocE48); (2) nick-translated 600 bp EcoRI fragment derived from pSocB149 and recloned in pBR325; (3) nick-translated 1950 bp EcoRI fragment derived from pSocB149 and recloned in pBR325; (4) the 1200 bp EcoRI-PstI fragment of pSocE48, 3' end-labeled at the EcoRI site; (5) the 550 bp EcoRI-PstI fragment of pSocE48, 3' end-labeled at the EcoRI site. The size of the RNAs was calculated by reference to the mobility of E. coli ribosomal RNAs.

by probing Northern blots of chloroplast RNA with subfragments cut from 3' end-labeled 1750 bp EcoRI fragment with Pst 1. The 3' end-labeled strand of the 550 bp EcoRI-PstI fragment, but not the 3' end-labeled strand of the 1200 bp EcoRI-PstI fragment (see Figure 1), hybridized to the 1700-nucleotide large subunit mRNA (tracks 4 and 5, Figure 3). This result established that the direction of transcription of the gene is from the AvaI site towards the PstI site (from left to right in Figure 1), that is, towards the closer of the two inverted repeat regions on the spinach chloroplast DNA restriction map (see 12).

Sequence Analysis of the Large Subunit Gene

From the above results it seemed likely that the 1750 bp EcoRI fragment and the 250 bp EcoRI-BamHI fragment together would encompass the large subunit gene and its flanking sequences. These two fragments were therefore sequenced (16) according to the strategy diagrammed in Figure 1. The correct reading phase and the region coding for the large subunit protein were determined within this sequence by comparison of predicted translation products with known partially sequenced cyanogen bromide and tryptic peptides from barley and spinach large subunit protein (25, 26). The nucleotide sequence and the deduced amino acid sequence of the spinach large subunit protein are presented in Figure 4. The sequence is numbered from the methionine ATG triplet which precedes by 14 codons the GCT triplet that encodes the N-terminal alanine of barley (25) and wheat (Martin, personal communication) large subunit protein. For reasons discussed below, we consider translation of the large subunit of spinach starts at this methionine residue rather than at the alanine residue. The protein coding region terminates at the TAG amber codon immediately following codon number 475 for valine. Valine is known to be the C-terminal amino acid of spinach large subunit protein (3). A second stop-codon, TAA, is situated 3 codons further on.

On the basis of the nucleotide sequence data, spinach large subunit protein contains 461 amino acid residues (MW 51200), or 475 residues (MW 52760) if the N-terminal leader sequence from methionine to alanine is included. The predicted protein sequence immediately distal to the N-terminal alanine matches, except for 2 amino acid substitutions, the 46 known N-terminal residues of the barley large subunit protein (25). Also 22 of the 24 known N-terminal residues from the wheat large subunit protein (Martin, personal communication) match the predicted spinach large subunit N-terminal sequence. Of the 475 predicted residues, 216 are confirmed by


-220 -200 -180 -160 -140
AACGGTTAC GGTTGGGTTG CGCCATATAT ATGAAAGAGT ATACAATAAT GATGTATTTG GCGAATCAAA TACATGGTCT ATTAACGAAC CATTTTGATT AG
-120 -100 -80 -60 -40 -20
TTGATAAT ATTAATTGAG AATTTGATGA AAGATTGCTA TAAAAGGTTT CATTAAGGCC TAATTTATGT CGAGTAGACC TTGTTGCTTT GTTGTAAAAA TTAAAATTTG AAGTTGTAGG GAGGGACTT
ATG TCA CCA CAA ACA GAG ACT AAA GCA AGT GTT GAA TTT AAA GCT GGT GTT AAA GAT TAC AAA TTG ACT TAT TAT ACT CCT GAG TAT GAA ACC CTA GAT ACT GAT 105
Met Ser Pro Gln Thr Glu Thr Lys Ala Ser Val Glu Phe Lys Ala Gly Val Lys Asp Tyr Lys Leu Thr Tyr Tyr Thr Pro Glu Tyr Glu Thr Leu Asp Thr Asp
10 Gly 20 30 Lys
ATC TTG GCA GCA TTC CGA GTA AGT CCT CAA CCT GGA GTT CCA CCC GAA GAA GCA GGG GCT GCA GTA GCT GCT GAA TCT TCT ACT GGT ACA TGG ACA ACT GTA 207
Ile Leu Ala Ala Phe Arg Val Ser Pro Gln Pro Gly Val Pro Pro Glu Glu Ala Gly Ala Ala Val Ala Ala Glu Ser Ser Thr Gly Thr Trp Thr Thr Val
40 Thr Leu 50 60 Ala Ala
TGG ACC GAC GGA CTT ACC AAC CTT GAT CGT TAC AAA GGA CGA TGC TAC CAC ATC GAG CCC GTT GCT GGA GAA GAA AAT CAA TAT ATT TGT TAT GTA GCG TAT CCT 312
Trp Thr Asp Gly Leu Thr Asn Leu Asp Arg Tyr Lys Gly Arg Cys Tyr His Ile Glu Pro Val Ala Gly Glu Glu Asn Gln Tyr Ile Cys Tyr Val Ala Tyr Pro
70 Ser 80 90 Pro Asp Pro Asp 100
TTA GAC CTT TTT GAA GAA GGT TCT GTT ACT AAC ATG TTT ACT TCC ATT GTG GGT AAC GTA TTT GGG TTC AAA GCC TTG CGT GCT CTA CGT TTG GAA GAT TTG CGA 417
Leu Asp Leu Phe Glu Glu Gly Ser Val Thr Asn Met Phe Thr Ser Ile Val Gly Asn Val Phe Gly Phe Lys Ala Leu Arg Ala Leu Arg Leu Glu Asp Leu Arg
110 120 130
ATC CCT GTT GCT TAT GTA AAA ACT TTC CAA GGC CCG CCT CAC GGT ATC CAA GTT GAG AGA GAT AAA TTG AAC AAG TAT GGT CGT CCC CTA TTG GGA TGC ACC ATT 522
Ile Pro Val Ala Tyr Val Lys Thr Phe Gln Gly Pro Pro His Gly Ile Gln Val Glu Arg Asp Lys Leu Asn Lys Tyr Gly Arg Pro Leu Leu Gly Cys Thr Ile
140 Pro Ser 150 Arg Met 160 170
AAA CCT AAA TTA GGT TTA TCC GCT AAA AAC TAT GGT CGC GCA GTT TAT GAA TGT CTT CGC GGT GGA CTT GAT TTT ACC AAA GAT GAT GAA AAC GTG AAC TCC CAG 627
Lys Pro Lys Leu Gly Leu Ser Ala Lys Asn Tyr Gly Arg Ala Val Tyr Glu Cys Leu Arg Gly Gly Leu Asp Phe Thr Lys Asp Asp Glu Asn Val Asn Ser Gln
180 Cys 190 200
CCG TTT ATG CGT TGG AGA GAC CGT TTC CTG TTT TGT GCC GAA GCT CTT TAT AAA GCA CAA GCC GAA ACA GGC GAA ATC AAA GGG CAT TAC TTG AAT GCT ACC GCG 732
Pro Phe Met Arg Trp Arg Asp Arg Phe Leu Phe Cys Ala Glu Ala Leu Tyr Lys Ala Gln Ala Glu Thr Gly Glu Ile Lys Gly His Tyr Leu Asn Ala Thr Ala
210 Val 220 Ile Ser 230 240
GGT ACA TGC GAA GAT ATG ATG AAA AGG GCT GTA TTT GCC AGA GAA TTG GGC GTT CCT ATT GTA ATG CAT GAC TAC TTA ACA GGG GGA TTC ACT GCA AAT ACT ACC 837
Gly Thr Cys Glu Asp Met Met Lys Arg Ala Val Phe Ala Arg Glu Leu Gly Val Pro Ile Val Met His Asp Tyr Leu Thr Gly Gly Phe Thr Ala Asn Thr Thr
Asp Glu 250 Ile Gly Gln 260 270

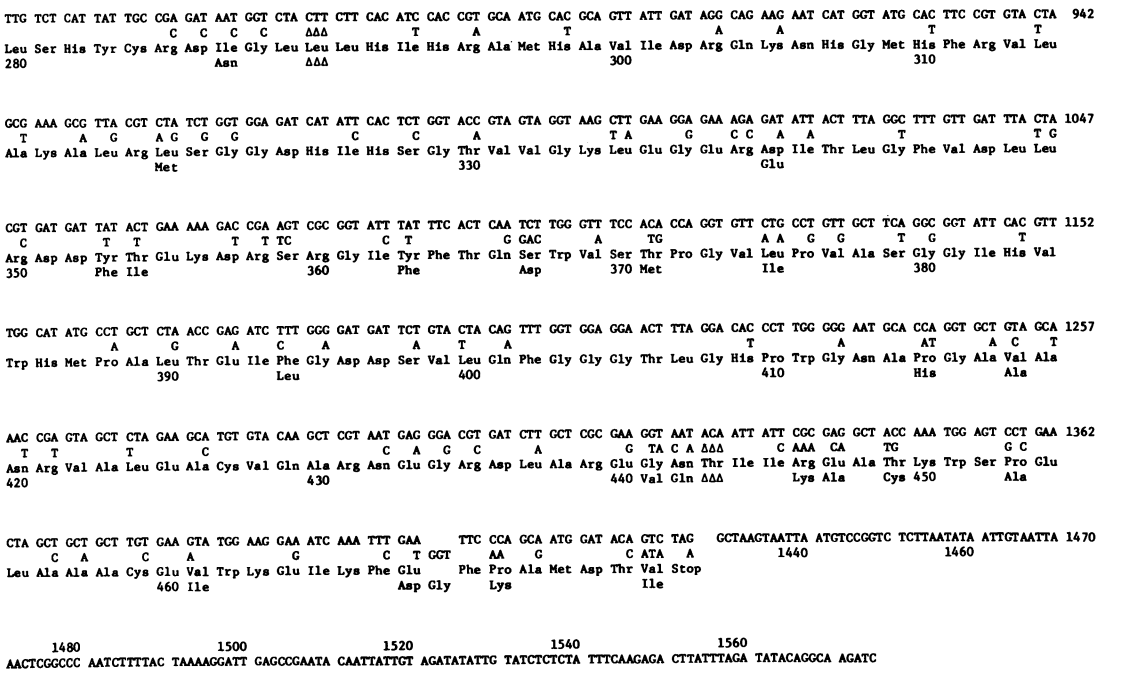


Figure 4. The Nucleotide Sequence of the Spinach Chloroplast Large Subunit Gene and Flanking Regions. The sequence of the non-transcribed strand of the gene is arranged in codons and the corresponding amino acids are indicated. Numbering starts at the ATG triplet at which translation is probably initiated. Nucleotide differences between the spinach large subunit gene and the maize large subunit gene (1) and any resulting amino acid changes are shown below the spinach sequence. Regions for which protein sequence data from spinach are available are: amino acids 165-177, 320-339, 451-463 (3 tryptic peptides located near the catalytic site), 251-259 and 298-305 (25, 27).

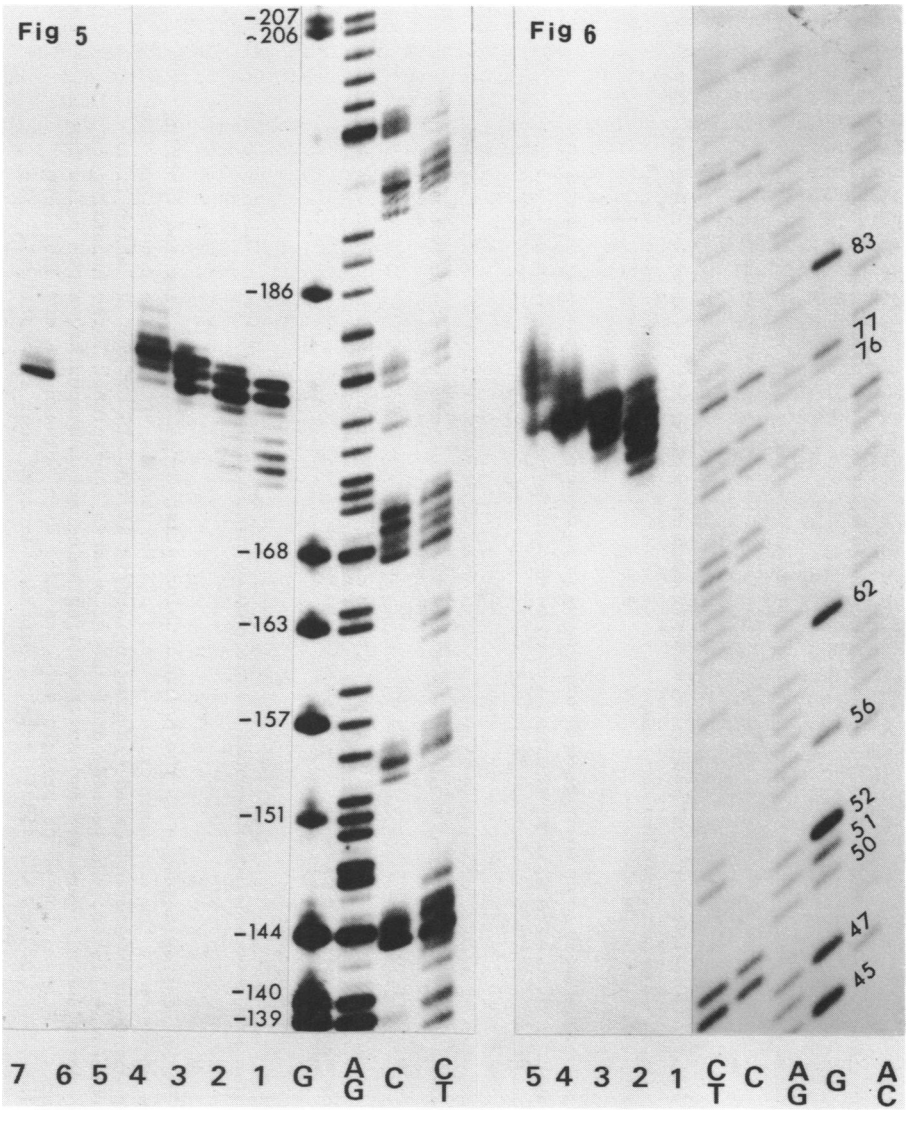


Figure 5. Determination of the 5' end of Spinach Large Subunit mRNA by S1 Nuclease Mapping and cDNA Synthesis. ^{32}P 5' end-labeled TaqI-226S fragment (15 ng) and spinach chloroplast total RNA (48 µg) in 10 µl of 10 mM Tris-HCl (pH 7.4), 10 mM $MgCl_2$, 50 mM NaCl were denatured at 90°C for 2 min and annealed by cooling slowly to room temperature over 30 min. 20 µl of 45 mM Na Acetate (pH 4.6), 75 mM NaCl, 1.5 mM $ZnSO_4$, 7.5% glycerol containing nuclease S1 (Miles) were added and the reaction was incubated for 15 min at 37°C. The reaction was stopped by the addition of 70 µl of TE (10 mM Tris-HCl, pH 8.4, 1 mM EDTA) and 100 µl of TE-saturated phenol. The reaction

was re-extracted with TE-saturated phenol, twice ethanol precipitated, ethanol washed, dried and taken up in 4 µl formamide-dyes (28). The samples were electrophoresed on a 8% acrylamide-7M urea thin gel (28). Tracks C + T, C, A + G, G are Maxam and Gilbert sequencing reactions of ^{32}P 5' end-labeled TaqI-226S DNA. Tracks 1-4 are reactions with respectively 1000, 500, 200, 100 units of S1 nuclease. Track 5 is a reaction with E. coli ribosomal RNA used in place of spinach RNA (1000 units of S1 were added). In tracks 6 and 7 the HaeIII-TaqI-14 bp fragment (0.5 ng ^{32}P 5' end-labeled at the TaqI site) was annealed with 6 µg of E. coli ribosomal RNA (Track 6) or 12 µg of spinach chloroplast RNA (Track 7) as described above. 40 µl of 62.5 mM Tris HCl (pH 8.3), 175 mM KCl, 12.5 mM $MgCl_2$, 37.5 mM β-mercapto-ethanol, 625 µM deoxyribonucleotides with 10 units of reverse transcriptase were added and the reaction incubated for 1 h at 42°C. The reaction was stopped and samples were prepared for gel electrophoresis as described above. Because this gel shows the sequence of the complementary DNA strand, the numbers against the G residues correspond to C residues in the sequences shown in Figures 4 and 7. The nucleotides corresponding to the 5' end of spinach large subunit mRNA are indicated in Figure 7.

Figure 6. Determination of the 3' end of Spinach Large Subunit mRNA by S1 Nuclease Mapping. EcoRI-BamHI-244 bp fragment (60 ng), ^{32}P 3' end-labeled at the EcoRI site, was annealed to 12 µg of spinach chloroplast total RNA as described in Figure 5. Tracks 2-5 show the products of treatment with respectively 1000, 500, 200 and 100 units of S1 nuclease separated on a 8% acrylamide-7M urea thin gel. Track 1 shows the product of treatment with 1000 units of S1 of a reaction with E. coli ribosomal RNA replacing the spinach RNA. Conditions for S1 treatment and sample preparation are as described in Figure 5. Tracks A + C, G, A + G, C, C + T are the products of Maxam and Gilbert sequencing reactions of the EcoRI-BamHI-244 bp fragment ^{32}P 3' end-labeled at the EcoRI site. Because this gel shows the sequence of the complementary DNA strand, the numbers against the G residues correspond to C residues in the sequences shown in Figures 4 and 8. The nucleotides corresponding to the ends of the major S1-resistant products are indicated in Figure 8.

protein sequence analysis of peptides from barley and spinach large subunit protein (Figure 4; Ref. 25, 26).

For comparative purposes, the nucleotide sequence of the gene for maize large subunit protein (1) is also shown in Figure 4. The similarity of the two sequences provides additional confirmation of the proposed structure of the spinach large subunit protein. The sequence data also verifies the earlier tentative conclusion that there are no introns in the large subunit gene.

The 5' End of the Large Subunit mRNA

The region coding for the 5' end of large subunit mRNA was located using the S1 nuclease mapping procedure of Berk and Sharp (18). The TaqI-226 bp fragment, lying between base pairs 286 and 60 proximal to the N-terminal methionine codon (Figure 1), was labeled at the 5' ends and strand separated. Approximately 120 residues of the slow strand (TaqI-226s) were

protected from S1 nuclease digestion by prior annealing of the DNA to spinach chloroplast RNA (Figure 5). DNA sequence analysis confirmed that TaqI-226s corresponded to the transcribed strand of the large subunit gene (Figure 5). The precise length of the S1-spared TaqI- 226s fragment was determined by electrophoresis on a DNA sequencing gel using a DNA sequence ladder generated from TaqI-226s as a marker (Figure 5). Based on this analysis, the region coding for the 5' end of large subunit mRNA is 178-179 nucleotides upstream from the ATG triplet at which translation is presumed to initiate (Figure 7).

This result was confirmed by reverse transcriptase extension of a DNA fragment annealed to spinach chloroplast RNA. A 14 bp HaeIII-TaqI fragment (Figure 1; - 71 to -58, Figure 7) primed cDNA synthesis on the chloroplast RNA template to a length of 119 and 120 residues (Figure 5). Thus this method also shows the length of the 5' untranslated region of spinach large subunit mRNA to be 178-179 nucleotides (58 + 119 or 120).

The 3' End of the Large Subunit mRNA

The 244 bp BamHI-EcoRI fragment (Figure 1) which overlaps the region coding for the C-terminal amino acid of large subunit protein, was 3' end-labeled at the EcoRI site, denatured, annealed to spinach chloroplast RNA, and treated with S1 nuclease. The products, which were analysed on a DNA sequencing gel using as a marker a sequencing ladder generated from the same labeled restriction fragment (Figure 6), indicated that the region coding for the 3' end of the large subunit mRNA was between nucleotides 1510 and 1513 (Figure 4). Thus the 3' untranslated region of large subunit mRNA is 85-88 residues, giving a total length for the mRNA of 1688-1692 residues, a value agreeing well with the size determined above by hybridization to Northern blots.

DISCUSSION

Large Subunit Protein

Although the determined N-terminus of large subunit protein of barley (25) and wheat (Martin, personal communication) is alanine, there is strong evidence to suggest that translation of the spinach large subunit protein initiates 14 codons prior to the alanine codon. There is a methionine codon at this point in both the spinach and maize genes (Figure 4) and protein synthesis is usually presumed to start with formylmethionine. Two nucleotide changes occur in the 13 codons between the methionine and alanine codons of spinach and maize; one results in an amino acid substitution and

the other in a neutral change (Figure 4). This modest degree (5%) of divergence is closer to that found in the protein coding portion of the genes (16% divergence) and far less than the extensive divergence (35%) found in the untranslated parts of the genes (see below). The sequence 5'GGAGG3' occurs at positions -10 to -6 upstream from the methionine codon (Figure 4) of both the spinach and maize large subunit genes. This sequence is complementary to part of the sequence 3'AUUCCUCCA5' which is found at the 3' end of maize chloroplast 16S RNA (29). Such complementarity occurs frequently at a comparable position in the ribosome binding site of prokaryotic mRNAs and is thought to play a crucial role in the initiation of protein synthesis (30, 31). No sequence showing such extensive complementarity to the 3' end of 16S chloroplast RNA occurs immediately prior to the N-terminal alanine residue (Figure 4). Since no relevant protein sequence data is available for spinach or maize large subunit proteins it is possible that the N-terminus of these proteins is methionine while alanine is found at the N-terminus of barley and wheat. However, Langridge (32) has found that the spinach large subunit protein synthesized in an E. coli cell-free system is 1000-2000 daltons larger than the purified spinach protein. Furthermore, treatment of this larger protein with soluble extracts from chloroplasts converts it to the same size as the purified large subunit protein. On this basis we think that translation of the spinach and maize large subunit proteins is initiated at the methionine codon, but that a post-translational processing event cleaves the protein adjacent to the alanine residue.

The small subunit of ribulose bisphosphate carboxylase is synthesized in a precursor form, the function of which may be ensure that the cytoplasmically synthesized protein is transported into the chloroplast (33-35). It is unlikely that the role of a precursor form of the large subunit protein would be concerned with the transport of the protein across a membrane, but other possible functions, including that of maintaining the subunit in an appropriate conformation until it is assembled into the holoprotein, can be suggested (see Ref. 32). Comparison of the leader sequence of the large subunit from spinach with that of the small subunit from pea (36) does not reveal any homology.

Comparison of the deduced sequence for the spinach protein with the published sequences determined for parts of the barley protein and with the complete deduced sequence for the maize protein shows the protein to be strongly conserved. Of the 475 amino acid residues in the spinach large

subunit 49 are changed (10% divergence) in maize (Figure 4). Of these changes, 45 are amino acid substitutions and 4 are deletions/ additions. These changes are not randomly distributed throughout the molecule. In particular, 32 of the 33 amino acids comprising two of the three tryptic peptides known to be located at or near the catalytic site of the spinach large subunit protein (27) are conserved in spinach and maize (Figure 4). The C-terminal 35 residues, which encompass the third tryptic peptide known to be located at or near the catalytic site, are, however, particularly variable with 12 changes (34% divergence). As sequence information from other species becomes available a more complete picture of the regions of the molecule that cannot tolerate changes will emerge. One expected feature of the changes observed between the spinach and maize large subunit genes is that, while the amino acid sequences show only 10% divergence, the nucleotide sequences have diverged by 16% (Figure 4). This discrepancy results from the large number of neutral (usually in the third codon position) nucleotide changes.

Analysis of the codons used in translating the spinach large subunit protein shows that only 6 of the possible 61 are not used at all. Three of these correspond to the 3 not used in the maize gene : ACG (thr), CGG (arg) and AGC (ser). In addition, CTC (leu), ATA (ile) and TCG (ser) do not appear in the large subunit gene from spinach. As might be expected for chloroplast DNA with an overall composition of 38% G+C (37) there is a marked bias towards the use of codons having A or T in the third position (71%). Within the protein coding region the base composition is 44% G+C. Twenty-seven tRNAs have been identified in spinach chloroplasts and a further 8 possible tRNAs have been detected (38). This number is sufficient to translate all 55 codons used, allowing for G-U base-pairing invoked in the wobble hypothesis (39).

5' Untranslated Region of the Large Subunit Gene

Transcription of the mRNA for the spinach large subunit gene is initiated 178-179 nucleotides upstream from the ATG codon at which translation of the large subunit protein presumably starts (Figure 7). The size of the 5' untranslated region of the mRNA is surprising, particularly since transcription of the maize large subunit gene initiates only 61 $\pm$ 2 nucleotides prior to the methionine start codon (1). The lack of 117 of these nucleotides from the maize large subunit mRNA indicates that at least this part of the spinach leader region has little effect on the translatability of the mRNA. It is possible that sequences in this region are involved in the regulation

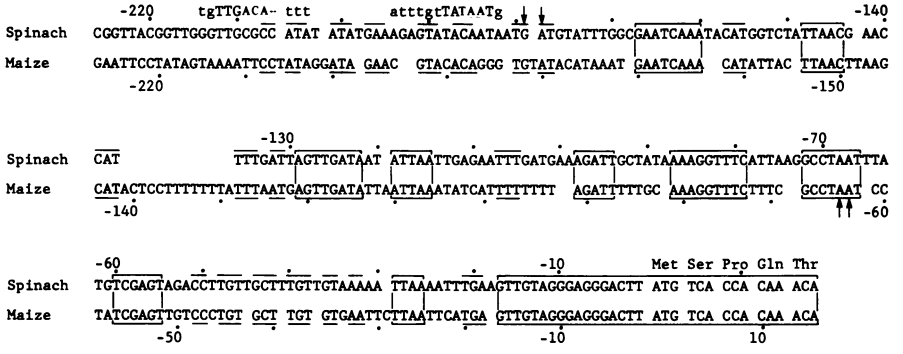


Figure 7. Comparison between Sequences 5' to the Coding Region of the Large Subunit Genes of Spinach and Maize (1) Chloroplasts DNAs. The numbering of the spinach sequence (top) and the maize sequence (bottom) is relative to the ATG triplet at which translation of the large subunit protein is initiated. The sequences are aligned to give maximal homology with boxes indicating regions of extensive homology and lines indicating regions of lesser homology. The arrows at -63 and -64 of the maize sequence and -178 and -179 of the spinach sequence indicate the respective large subunit mRNA transcription start sites. The most common E. coli promoter sequence deduced by Siebenlist (41) is indicated between nucleotides -214 and -182 of the spinach sequence. Higher case letters in this common sequence indicate that a base appears more frequently in that position of promoters than bases indicated by lower case letters (see Ref. 41).

of transcription and it is interesting to note that in maize, in contrast to spinach, the large subunit mRNA is differentially expressed in mesophyll and bundle sheath cells (40).

Comparison of the 20 nucleotides preceeding the ATG initiation codons of the spinach and maize gene reveals about 95% homology (Figure 7), a reflection of the importance of this region as a probable ribosome binding site (see above). Further upstream the homology is markedly reduced and, if the comparison is made over the 227 nucleotides in spinach and maize which have so far been sequenced in this region only 65% of them match. The comparable figure for homology over the first 227 nucleotides of the coding region of the maize and spinach genes is 85%.

In Figure 7 the untranslated sequences preceding the coding regions of the two genes have been aligned to maximize the extent of the homology and reveal stretches of perfect matching up to 9 nucleotides in length. The changes that have occurred through evolution of this region include extensive (up to 13 nucleotides) deletions (or additions) as well as base substitutions. Because one consequence of these changes has been the relocation of the

promoter for the large subunit gene, the comparison shown in Figure 7 can be viewed as part of a mutational analysis of the respective promoter regions.

Does the Promoter for the Large Subunit Gene have Features in Common with Bacterial Promoters?

Nucleotide sequence studies of the 16S rRNA gene from maize chloroplast DNA have indicated a probable prokaryote origin for at least that part of the DNA (42). The retention of homology between the 3' end of 16S rRNA and the ribosome binding site of the maize and spinach large subunit mRNA is a further indication of the prokaryote-like nature of genes on chloroplast DNA. The question arises then as to whether the promoters for chloroplast DNA genes have features in common with bacterial promoters.

Extensive studies of promoters from the bacterium Escherichia coli show that two regions immediately preceding transcription initiation sites are generally conserved (41). The sequence TATAAT is often found about 5 nucleotides prior to the transcription start site. The other region to show considerable conservation occurs approximately 35 nucleotides upstream from the same site. It should however be emphasized that examples of promoters with nucleotide changes in either of these conserved regions exist.

A comparison between the spinach large subunit promoter region and the E. coli consensus promoter sequence reveals homology with both conserved regions (Figure 7). In this context it may be significant that substantial amounts of large subunit protein are synthesized in an E. coli cell-free transcription-translation system directed by the EcoRI 1750 bp fragment (unpublished observations). This result shows that E. coli RNA polymerase must be capable of initiating transcription somewhere in the 350 bp region prior to the large subunit gene translation start codon. It will be of interest to determine whether the promoter utilized by E. coli RNA polymerase corresponds to the promoter used in vivo by spinach chloroplast RNA polymerase.

The maize large subunit gene promoter, however, shows no obvious homology to E. coli promoters (1). Sequence analysis of further chloroplast DNA promoter regions will help to elucidate their evolutionary origin.

3' Untranslated Region of the Large Subunit Gene

The spinach large subunit mRNA has 82-85 untranslated nucleotides at the 3' end (Figure 8). The region immediately preceding the 3' end of the mRNA is capable of forming a stem and loop secondary structure (Figure 9).

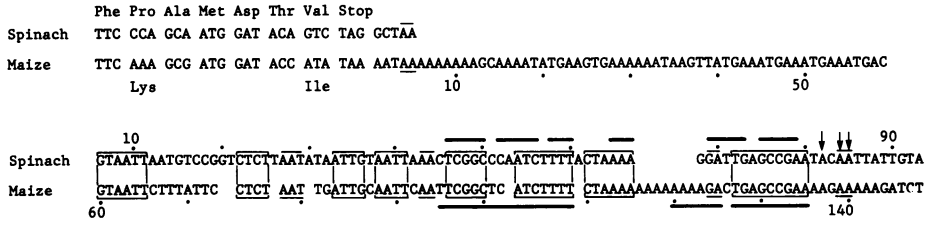


Figure 8. Comparison between Sequences 3' to the Coding Region of the Large Subunit Genes of Spinach and Maize (1) Chloroplast DNAs. The spinach sequence (top) and the maize sequence (bottom) are numbered relative to the translation stop codons. The sequences are aligned to show maximal homology with boxes and lines indicating the regions of homology as for Figure 7. The arrows at nucleotides 82, 84, 85 of the spinach sequence indicate the positions at which transcription of the spinach large subunit mRNA terminates. The thick bars indicate those nucleotides which base-pair to form the structures shown in Figure 9.

Analogous secondary structures occur at prokaryotic transcription termination sites and have been shown by mutational analysis to be required for efficient transcription termination (43). Such structures have not been implicated in transcription termination of most eukaryote genes, although stem and loop structures exist at the 3' ends of some genes transcribed by RNA polymerase III (44, 45).

Comparison of the nucleotide sequences of spinach and maize chloroplast DNAs immediately following the large subunit gene translation stop codon reveals little homology (Figure 4). However, if the spinach sequence is displaced downstream by 54 nucleotides the next 80 nucleotides can be aligned with the maize sequence with extensive homology (Figure 8). Of these nucleotides, 41 include the region capable of forming the stem and loop structure and the site of transcription termination for the spinach mRNA (Figure 9). Thus the maize chloroplast DNA 3' to the large subunit gene coding region is also capable of forming a stem and loop structure (Figure 9). In pea chloroplast DNA this same 41-nucleotide region is completely homologous to the spinach sequence while the adjacent regions show extensive divergence (unpublished observations). If it is assumed that these stem and loop structures do have a role in transcription termination then the variation in the sequence of the stem and in the size of the loop between spinach and maize suggests that these aspects of the structure can tolerate this variation. Extensive studies of stem and loop structures which function as transcription terminators in E. coli show that neither

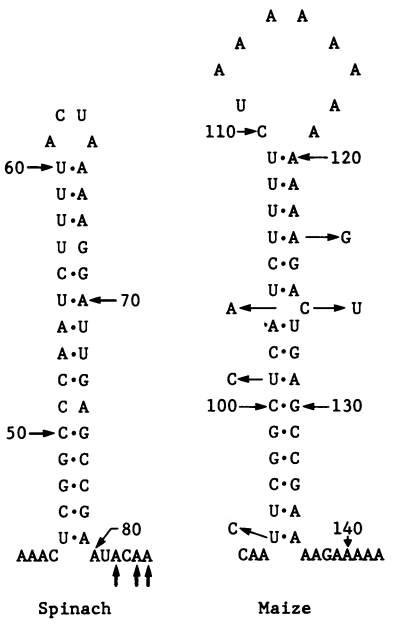


Figure 9. The Proposed Stem and Loop Secondary Structure at the 3' end of Spinach Large Subunit mRNA and the Analogous Structure in Maize. The heavy arrows indicate the positions at which transcription of the spinach large subunit gene terminates. The nucleotide numbers correspond to those given in Figure 8. The nucleotide changes required to convert the stem region of the proposed maize secondary structure to the stem region of the spinach secondary structure are indicated. The estimated free energies (ΔG) of formation of the stem and loop structures calculated according to Tinoco et al. (46) and Borer et al. (47) are -16 kcal for spinach and -20 kcal for maize.

the sequence of the stem nor the size of the loop are as important as is the ability to form a stable stem region (42). Mutants relieving transcription termination almost always destabilize the stem regions (45). However, confirmation that the stem and loop structures which can be formed at the 3' ends of the pea and maize large subunit genes are in fact present at the 3' end of the respective mRNAs, and therefore are probably involved in transcription termination, must await the determination of the 3' end of the pea and maize large subunit mRNA.

ACKNOWLEDGEMENT

We thank K. Ferguson, S. Craig and C. Miller for advice and assistance with the electron microscopy. One author (G.Z.) is a Queen Elizabeth II Fellow.

REFERENCES

1. McIntosh, L., Poulsen, C. and Bogorad, L. (1980) Nature 288, 556-560.
2. Chan, P.H. and Wildman, S.G. (1972) Biochim. Biophys. Acta 277, 677-680.
3. Sugiyama, T. and Akazawa, T. (1970) Biochemistry 9, 4499-4504.
4. Nishimura, M. and Akazawa, T. (1973) Biochem. Biophys. Res. Commun. 54, 842-848.
5. Kawashima, N. and Wildman, S.G. (1972) Biochim. Biophys. Acta 262, 42-49.

6. Gray, J.C. and Kekwick, R.G.O. (1973) FEBS Lett. 38, 67-69.
7. Blair, G.E. and Ellis, R.J. (1973) Biochim. Biophys. Acta 319, 223-234.
8. Coen, D.M., Bedbrook, J.R., Bogorad, L. and Rich, A. (1977) Proc. Nat. Acad. Sci. USA 74, 5487-5491.
9. Bedbrook, J.R., Coen, D.M., Beaton, A.R., Bogorad, L. and Rich, A. (1979) J. Biol. Chem. 254, 905-910.
10. Bottomley, W. and Whitfeld, P.R. (1979) Eur. J. Biochem. 93, 31-39.
11. Malnoe, P., Rochaix, J.-D., Chua, N.H. and Spahr, P.-F. (1979) J. Mol. Biol. 133, 417-434.
12. Whitfeld, P.R. and Bottomley, W. (1980) Biochem. Internat. 1, 172-178.
13. Katz, L., Kingsbury, D.T. and Helinski, D.R. (1973) J. Bacteriol. 114, 577-591.
14. Whitfeld, P.R., Herrmann, R.G. and Bottomley, W. (1978) Nucleic Acids Res. 5, 1741-1751.
15. Wienand, J., Schwartz, Z. and Feix, G. (1979) FEBS Lett. 98, 319-323.
16. Maxam, A.M. and Gilbert, W. (1977) Proc. Nat. Acad. Sci. USA 74, 560-564.
17. 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.
18. Berk, A.J. and Sharp, P.A. (1977) Cell 12, 721-732.
19. Bailey, J.M. and Davidson, N. (1976) Anal. Biochem. 70, 75-85.
20. Alwine, J.L., Kemp, D.J. and Stark, G.R. (1977) Proc. Nat. Acad. Sci. USA 74, 5350-5354.
21. Maniatis, T., Kee, S.G., Efstratiadis, A. and Kafatos, F.C. (1976) Cell 8, 163-182.
22. Wu, R. (1970) J. Mol. Biol. 51, 501-521.
23. Thomas, M., White, R.L. and Davis, R.W. (1976) Proc. Nat. Acad. Sci. USA 73, 2294-2298.
24. Davis, R.W., Simon, M. and Davidson, N. (1971) in Methods in Enzymology, Grossman, L. and Moldave, K., Eds., Vol. 21, pp 413-428. Academic Press, New York.
25. Poulsen, C., Martin, B. and Svendsen, I. (1979) Carlsberg Res. Commun. 44, 191-199.
26. Stringer, C.D. and Hartman, F.C. (1978) Biochem. Biophys. Res. Commun. 80, 1043-1048.
27. Hartman, F.C., Norton, I.L., Stringer, C.D. and Schloss, J.V. (1978) in Photosynthetic Carbon Assimilation, Siegelman, H.W. and Hind, G., Eds., pp 245-269. Plenum Press, New York.
28. Sanger, F. and Coulson, A.R. (1975) J. Mol. Biol. 94, 441-448.
29. Schwarz, Z. and Kössel, H. (1979) Nature 279, 520-522.
30. Shine, J. and Dalgarno, L. (1974) Proc. Nat. Acad. Sci. USA 71, 1342-1346.
31. Scherer, G.F.E., Walkinshaw, M.D., Arnott, S. and Morré, D.J. (1980) Nucleic Acids Res. 8, 3895-3907.
32. Langridge, P. (1981) FEBS Lett. 123, 85-89.
33. Dobberstein, B., Blobel, G. and Chua, N.-H. (1977) Proc. Nat. Acad. Sci. USA 74, 1082-1085.
34. Chua, N.-H. and Schmidt, G.W. (1978) Proc. Nat. Acad. Sci. USA 72, 6110-6114.
35. Highfield, P.E. and Ellis, R.J. (1978) Nature 271, 420-424.
36. Bedbrook, J.R., Smith, S.M. and Ellis, R.J. (1980) Nature 287, 692-697.
37. Whitfeld, P.R. and Spencer, D. (1968) Biochim. Biophys. Acta 157, 333-343.
38. Driesel, A.J., Crouse, E.J., Gordon, K., Bohnert, H.J., Herrmann, R.G., Steinmetz, A., Mubumbila, M., Keller, M., Burkard, J. and Weil, J.H. (1979) Gene 6, 285-306.
39. Crick, F.H. (1966) J. Mol. Biol. 19, 548-555.

40. Link, G., Coen, D.M. and Bogorad, L. (1978) Cell 15, 725-731.
41. Siebenlist, U. (1979) Nucleic Acids Res. 6, 1895-1907.
42. Schwarz, Z. and Kössel, H. (1980) Nature 283, 739-742.
43. Rosenberg, M. and Court, D. (1979) Annu. Rev. Genet. 13, 319-353.
44. Korn, L.J. and Brown, D.D. (1978) Cell 15, 1145-1156.
45. Gerlach, W.L. and Dyer, T.A. (1980) Nucleic Acids Res. 8, 4851-4865.
46. Zurawski, G. and Yanofsky, C. (1980) J. Mol. Biol. 142, 123-129.
47. 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.
48. Borer, P.N., Dengler, B., Tinoco, I. and Uhlenbeck, O. (1974) J. Mol. Biol. 86, 843-853.