Spinach plastid genes coding for initiation factor IF-1, ribosomal protein S11 and RNA polymerase α -subunit

Gertrud Sijben-Müller, Richard B. Hallick¹, Juliane Alt, Peter Westhoff and Reinhold G. Herrmann*

Botanisches Institut der Universität Düsseldorf, Universitätsstr. 1, 4000 Düsseldorf, FRG and

¹Department of Biochemistry, Biological Sciences West, University of Arizona, Tucson, AZ 85721, USA

Received 10 October 1985; Revised and Accepted 6 December 1985

ABSTRACT

The nucleotide sequence of 2.5 kbp from the cloned SalI fragments 8 and 11 of spinach plastid DNA has been determined. This region was found to encode three open reading frames for hydrophilic polypeptides of 77, 138, and 335 amino acids. Using the computer search algorithm of Lipman and Pearson (Science 227, 1435, 1985), these genes were identified as coding for homologues of *E. coli* initiation factor IF-1 (infA), 30S ribosomal protein S11 (rps11), and the α -subunit of DNA-dependent RNA polymerase (rpoA). The spinach plastid gene organization is infA - 381 bp spacer - rps11 - 72 bp spacer - rpoA. The genes are transcribed in vivo and appear to encode functional proteins. These findings imply that plastid chromosomes code for components of the organelle transcription apparatus.

INTRODUCTION

Plastid chromosomes from higher plants and eukaryotic alga code for a number of components of the photosynthetic membranes and the chloroplast translational apparatus. The first class includes the large subunit of the stromal enzyme ribulose biphosphate carboxylase/oxygenase (reviewed in 1), all chlorophyll *a*-conjugated apoproteins of the photosynthetic reaction centers, all cytochromes and the catalytic subunits of the thylakoid ATP synthase, a total of 15 protein species (summarized in 2). These genes have been located by hybrid selection-translation and coupled transcription-translation using appropriate recombinant DNAs and immunoprecipitation. They have been characterized by DNA sequence analysis. The second category includes the 16S, 23S, 4.5S, and 5S ribosomal RNAs of the 70S chloroplast ribosomes, a presumed full complement of the tRNAs needed for chloroplast translation, approximately one third of the 50-55 proteins of the 30S and 50S ribosome subunits, and one translation elongation factor (summarized in 3). Plastid genes for ribosomal proteins S4 (4), S7 and S12 (5), S14 (6), S19 (7, 8), L2 (8), and elongation factor EF-Tu (9-11) have been characterized. These genes have all been identified to date through nucleic acid or amino acid sequence homology of the encoded polypeptide to the corresponding *E. coli* ribosomal proteins.

We describe here potential coding loci for three hydrophilic proteins of lengths 77, 138, and 335 amino acids which were noted in the course of nucleotide sequence analysis of a 2.5 kbp region of spinach plastid DNA fragments SalI-8 and -11 adjacent to the coding locus for subunit IV (petD) of the cytochrome *b6/f* complex. The derived polypeptide sequences for these loci were compared to the sequences of known proteins in the National Biomedical Research Foundation library using the search algorithm described by Lipman and Pearson (12). In each case, highly significant homology was found for each open reading frame with a single protein. The urf77, 138, and 335 gene products have 52%, 43%, and 26% identity in amino acid sequence with initiation factor IF-1, ribosomal protein S11, and the α -subunit of DNA-dependent RNA polymerase, respectively, from *E. coli*. We conclude that these plastid genes encode the functional counterparts of the bacterial genes for components of the chloroplast transcriptional and translational apparatus.

METHODS

Recombinant DNAs and DNA Sequence Analysis

The recombinant DNAs pWHsp207, pWHsp208, pWHsp209 and pWHsp211 used in this study carry the SalI primary fragments of 6.0, 5.2, 4.0 and 0.66 kbp respectively of spinach plastid DNA in pBR322. They were selected from a shotgun library as detailed previously (13-15). The 2.3 kbp SalI-EcoRI fragment derived from the 5.2 kbp fragment SalI-8 by digestion with EcoRI was subcloned into pBR325 for nucleotide sequence analysis. Covalently closed circular plasmid DNA was prepared from cleared lysates (16) by centrifugation in ethidium bromide CsCl gradients.

DNA sequencing and separation of labelled ends by restriction cleavage or single-strand separation were done according to the chemical modification and chain cleavage procedure of Maxam and Gilbert (17). Fragments were 3' end-labelled by filling-in the single-stranded termini (18) using ^{32}P - α dNTPs (Amersham, Braunschweig) and the Klenow fragment of *E. coli* DNA polymerase (Boehringer, Mannheim). Sequence data were analysed on a Telefunken 445 computer with programs compiled by Kröger and Kröger-Block (19).

Northern Blot Analysis

The procedures used for the isolation of spinach chloroplast RNA, restriction digests, hybrid select translation, agarose gel electrophoresis of DNA fragments or RNA, transfer of RNA to nitrocellulose strips as well as pre- and posthybridization washes have been described previously (20-22). DNA fragments were recovered from agarose gels by electroelution (23) and from acryl-

amide gels by diffusion into a salt buffer (17). They were purified by Elutip-d column chromatography (Schleicher and Schüll, Dassel), and nick-translated to high specific activity as described by Rigby et al. (24). Labelled fragments were purified by gel filtration on Sephacryl S200 and ethanol precipitated prior to use as hybridization probes. Labelled DNA probes were hybridized to the Northern blots in 50% formamide, 0.75 M NaCl, 0.075 M Na₃ citrate, 0.2% Na dodecylsulfate, 0.2% each of polyvinyl pyrrolidone, Ficoll and bovine serum albumin and 100 µg/ml denatured sonicated calf thymus DNA, for 14 h at 42°C.

Computer Analysis of DNA and Protein Sequence Data

Comparison of the derived amino acid sequences of the spinach plastid open reading frames and the 700,000 residues in the National Biomedical Research Foundation (NBRF) library was accomplished with an IBM-PC/XT or an IBM-PC/AT computer and the program FASTP (12). Search time is 8-10 min with the PC/XT and 2-3 min with the PC/AT. Evaluation of the statistical significance of a relationship was based on the criteria of Lipman and Pearson (12). These include (a) an initial and optimized "similarity score" for the aligned proteins that reflects the extent of identical residues and conservative replacements based on an amino acid replaceability matrix, and (b) a "z-value", which is equivalent to the similarity score less the mean of the random similarity scores for all of the sequences in the protein data base divided by the standard deviation of the random scores. Genuinely related sequences are viewed as those with initial similarity scores >50, that increase to between 100-300 after optimization, and those showing z-values >10.

RESULTS

DNA Sequence Analysis of SalI Fragments 8 and 11

The positions of fragments SalI-8 and SalI-11 in the spinach plastid chromosome are illustrated by Fig. 1. The sequence of the 2.5 kbp interval adjacent to petD in SalI-8 and overlapping into the 667 bp fragment SalI-11 was determined as indicated in the same figure. A computer search of the DNA sequence led to the identification of three open reading frames that begin with an initiator codon and could potentially code for polypeptides of >75 residues. All three loci have the same polarity. The nucleotide sequence of this DNA segment is shown in Fig. 2. The deduced amino acid sequences, commencing at the respective first available ATG translation start codons and ending at TAA, TAG and TGA translation stop codons, respectively, are shown below the nucleotide sequence.

Figure 1. Detail restriction map and sequence assay strategy of the SaliI-8/SaliI-11 region of the spinach plastid chromosome. The position of the genes and their direction of transcription (arrow) is indicated in the above map. Arrows commencing from solid dots in the lower map indicate extent and direction of the individual DNA sequence assays from the given restriction sites. Numbering is in base pairs. The locations of the amino termini as well as of the carboxy termini are marked NH₂ and COOH, respectively. Insert: Simplified restriction cleavage site map of the spinach plastid chromosome showing recognition sites of SaliI (25), the locations of *infA*, *rps11* and *rpoA* as well as the positions of the genes for ribosomal RNAs (*rrn*), ribosomal proteins *rps14*, *rps19*, *rpl2*, large subunit of ribulose biphosphate carboxylase/oxygenase (*rbcL*) and various thylakoid proteins (cf. 3, 6-8). The inverted repeat is marked by bold lines; the transcription polarities are indicated by arrow heads in the inner circles.

Spinach Plastid Gene for Protein Synthesis Initiation Factor IF-1 (*infA*)

The first locus begins at position 277 within fragment SaliI-11, and terminates with a TAA-terminator at position 510 (Fig. 2). If translated, the product would be a 77 residue hydrophilic polypeptide of MW 9,107. There are 19% arginine/lysine residues and 14% aspartate/glutamate residues. From a FASTP homology search with the protein sequence data base, significant homology with *E. coli* initiation factor IF-1 was identified. The alignment of the spinach sequence and *E. coli* IF-1 is shown in Fig. 3. *E. coli* IF-1 is 71 residues in length. There are 43% identical residues, and an additional 39% conservative replacements of amino acids in the spinach sequence. The two polypeptides align beginning at the amino terminus and continuing to the carboxyl end. This comparison gave initial and optimized scores of 166 and 186 with FASTP, and a z-value of 28.3. We conclude that the spinach open reading frame

codes for a protein homologous to *E. coli* IF-1, and designate the coding locus as *infA*.

Spinach Plastid Gene for Ribosomal Protein S11

The second open reading frame begins at position 891, within SallI-8, and terminates at 1307 with a TAG. The translation product would be a 138 residue hydrophilic protein of MW 14,909 with 20% arginine/lysine content. This protein was aligned with *E. coli* ribosomal protein S11 via a FASTP homology search. The result is shown in Fig. 4. There are 52% identical residues, and an additional 35% of conservative replacements. The *E. coli* S11 protein sequence is shorter at the amino terminus by 11 residues than the spinach open reading frame, but the alignment is exactly precise at the carboxyl terminus ending with 13 consecutive identical amino acids. The FASTP analysis gave initial and optimized similarity scores of 322 and 345, and a z-value of 61.4. No other significant homology with the protein data base could be detected. We conclude that the 138 codon open reading frame is the gene for chloroplast ribosomal protein S11. The gene is designated *rps11* (cf. 28).

Spinach Plastid Gene for a Polypeptide Homologous to the α -Subunit of *E. coli* DNA-Dependent RNA Polymerase

The third open reading frame is separated from *rps11* by a 72 bp spacer. It begins at position 1379 and ends at 2386 with TGA terminator. The 3'-end of this 335 codon reading frame is located 96 bp from the termination codon of the gene for subunit IV (*petD*) of the cytochrome *b/f* complex (29; Fig. 1). The initial homology search with FASTP of the protein sequence data base yielded an alignment of residues 34-178 with the corresponding region of the α -subunit of *E. coli* DNA-dependent RNA polymerase. The α -subunit sequence is based on the DNA sequence of *rpoA* (30-32). In this alignment, 34% of the 145 amino acids are identical. The initial and optimized similarity scores were 166 and 202, and the z-value was 32.2. To test for additional homology, the protein library was searched again with codons 179-335 of the spinach open reading frame that did not align in the first search. The result was an alignment of residues 255-324 with 29% identity to the corresponding carboxyl terminal region of the α -subunit polypeptide. The optimized similarity score was 105, and the z-value was 14.5. Note that both the amino- and carboxyl-halves of the spinach gene independently scored the highest alignment with the *E. coli* RNA polymerase α -subunit polypeptide. The results of the two separate alignments were combined as shown in Fig. 5. The optimized match has an 11-residue gap in the *E. coli* protein sequence near the carboxyl terminus, and two gaps of one residue each in the spinach sequence. There is no significant homology

```

          GTCGAC  TTGGTCTACG  AATCTATTCT  AATTATCAAC  GAATTCCTAG  AATTTTAGGT
          GGAATGGGGA  TTGCAATTCT  TTCTACCTCG  AGAGGTATAA  TGACAGATCG  GGAGGCTCGA
          CTAGAAGGAA  TTGGCGGAGA  AATTTTATGT  TATATATGGT  AATCCCTATA  CTATTCGAAT
          TGGATCCAAA  ATTTCTACT  TGTGAACAAA  AAAAGAGAAA  GGACAGCTTG  TCTAATATCA
          CTCCTCTATT  AGTTGATACT  TTAAGGCGGT  TTTGTCCTGA  ATG AAA GAA CAA AAA TGG
                                     Met Lys Glu Gln Lys Trp
          300
ATT CAT GAA GGT TTG ATT ACC GAA TCA CTT CCT AAT GGT ATG TTC TGG GTT CGT
Ile His Glu Gly Leu Ile Thr Glu Ser Leu Pro Asn Gly Met Phe Trp Val Arg
          50
TTA GAT AAT GAA GAT CCG ATT CTG GGT TAT GTT TCA GGA CGG ATA CGA CGT AGT
Leu Asp Asn Glu Asp Pro Ile Leu Gly Tyr Val Ser Gly Arg Ile Arg Arg Ser
          400
TCT ATA CGA ATT CTA CCG GGA GAT AGA GTT AAG ATT GAA GTA AGC CGT TAT GAT
Ser Ile Arg Ile Leu Pro Gly Asp Arg Val Lys Ile Glu Val Ser Arg Tyr Asp
          500
TCA ACT AGA GGT CGT ATA ATT TAT AGA CTG CGC AAC AAG GAT TCG AAC GAT TAA
Ser Thr Arg Gly Arg Ile Ile Tyr Arg Leu Arg Asn Lys Asp Ser Asn Asp Ter
          550
          ACAGTT  TTTCAACTTC  AACACTCCTT  TTTACAAGAA  TACAATCCGA  GATTCAAAAT
          ATTAAGAAAC  TTATTTCTCT  CCAAGAAATA  GATTTAAAT  GAAAACTAAG  GAATAATAAA
          TATGAAATA  AGAGCTTCTG  TTCGTCCAAT  TTGTGAAAAA  TGTCGACTGA  TCCGCAGACG
          GGGACGAATT  ATAGTAATTT  GTTCTAACCC  AAAACATAAA  CAACGACAAG  GATAATAATT
          CTACTATAAC  TAAAAACGAA  AAAGAATTTT  CTATAAGAA  TTGCTAAAG  ATTTGATTGA
          TGTAATAAAA  AACGGAAGGA  TTTGTTTGA  CATGAAATGG  ATATATCCAT  ATACCTTTGA
          CTCATATTTA  TGAGATGAAA  AAAT  ATG GCA AAA CCT ATA CCA AAA ATT GGT TCA
                                     Met Ala Lys Pro Ile Pro Lys Ile Gly Ser
          950
CGT AGA AAT GGA CGT ATT AGT TCG CGT AAA AGT GCA CGT AAA ATA CCA AAG GGT
Arg Arg Asn Gly Arg Ile Ser Ser Arg Lys Ser Ala Arg Lys Ile Pro Lys Gly
          1000
GTT ATT CAT GTT CAA GCA AGT TTT AAT AAT ACC ATT GTA ACT GTT ACA GAT GTA
Val Ile His Val Gln Ala Ser Phe Asn Asn Thr Ile Val Thr Val Thr Asp Val
          1050
CGA GGT CGA GTC GTT TCT TGG GCT TCT GCC GGT ACT TGT GGA TTC AGG GGT ACA
Arg Gly Arg Val Val Ser Trp Ala Ser Ala Gly Thr Cys Gly Phe Arg Gly Thr
          1100
AAA AGA GGA ACA CCA TTT GCG GCT CAA ACC GCA GCG GGA AAT GCT ATT CGT ACG
Lys Arg Gly Thr Pro Phe Ala Ala Gln Thr Ala Ala Gly Asn Ala Ile Arg Thr
          1150
GTG GTG GAA CAA GGT ATG CAA CGA GCA GAA GTC ATG ATA AAA GGT CCT AGT CTC
Val Val Glu Gln Gly Met Gln Arg Ala Glu Val Met Ile Lys Gly Pro Ser Leu
          1200
GGA AGG GAT GCA GCA TTA CGG GCT ATT CGT AGA AGC GGT ATA CTA TTA AGT TTC
Gly Arg Asp Ala Ala Leu Arg Ala Ile Arg Arg Ser Gly Ile Leu Leu Ser Phe
          1250
GTG CGA AAC GTA ACC CCT ATG CCG CAT AAT GGC TGT AGG CCT CCT AAA AAA AGA
Val Arg Asn Val Thr Pro Met Pro His Asn Gly Cys Arg Pro Pro Lys Lys Arg
          1350
CGC GTC TAG AAAAAAGAA TTGAAGAGAT TTCAAGAGAA ATAAATGATT CAAAGATATG
Arg Val Ter
          1400
ATCAATTTTG TAAAAATATTA CT  ATG GTT CGA GAG AAA ATA AGA GTA TCT ACT CAG
                                     Met Val Arg Glu Lys Ile Arg Val Ser Thr Gln
          1450
ACA CTT CAG TGG AAG TGT GTT GAA TCA AGA ACA GAT AGT AAA TGT CTT CAT TAT
Thr Leu Gln Trp Lys Cys Val Glu Ser Arg Thr Asp Ser Lys Cys Leu His Tyr

```

```

1500
GGA CGC TTT ATT CTC TCT CCA CTT ATG AAG GGT CAA GCT GAT ACA ATA GGC ATT
Gly Arg Phe Ile Leu Ser Pro Leu Met Lys Gly Gln Ala Asp Thr Ile Gly Ile

1550
GCG ATG CGA AGA GCG TTA CTT GGA GAA ATA GAA GGA ACA TGT ATC ACA CGT GCA
Ala Met Arg Arg Ala Leu Leu Gly Glu Ile Glu Gly Thr Cys Ile Thr Arg Ala

1600
AAA TCT GAA AAA ATA CCA CAC GAA TAC TCT ACC ATA TTA GGT ATT CAA GAA TCA
Lys Ser Glu Lys Ile Pro His Glu Tyr Ser Thr Ile Leu Gly Ile Gln Glu Ser

1650
GTA CAC GAA ATT TTA ATG AAT TTG AAA GAA ATA GTA TTG AGA AGT AAT CTA TAT
Val His Glu Ile Leu Met Asn Leu Lys Glu Ile Val Leu Arg Ser Asn Leu Tyr

1700
GGA GCT TGT GAG GCG TCT ATT TGT GTT AGG GGC CCC AGA GGT GTA ACT GCT CAA
Gly Ala Cys Glu Ala Ser Ile Cys Val Arg Gly Pro Arg Gly Val Thr Ala Gln

1750
GAT ATC ATC TTG CCA CCT TAT GTA GAA ATA GTT GAC AAT ACG CAG CAT ATC GCC
Asp Ile Ile Leu Pro Pro Tyr Val Glu Ile Val Asp Asn Thr Gln His Ile Ala

1800
AGC TTG ACG GAA CCA ATT GAT TTG TGT ATT GGA TTA CAA CTC GAG AGA AAT CGG
Ser Leu Thr Glu Pro Ile Asp Leu Cys Ile Gly Leu Gln Leu Glu Arg Asn Arg

1850
GGA TAT CAT ATA AAA GCG CCA AAT AAC TTT CAA GAC GGA AGT TTT CCT ATA GAT
Gly Tyr His Ile Lys Ala Pro Asn Asn Phe Gln Asp Gly Ser Phe Pro Ile Asp
900
GCT CTA TTC ATG CCT GTT CGG AAC GTG AAT CAT AGT ATT CAT TCT TAT GGA AAT
Ala Leu Phe Met Pro Val Arg Asn Val Asn His Ser Ile His Ser Tyr Gly Asn
1950
GGG AAT GAA AAA CAA GAG ATA CTT TTT CTC GAA ATA TGG ACA AAT GGT AGT TTA
Gly Asn Glu Lys Gln Glu Ile Leu Phe Leu Glu Ile Trp Thr Asn Gly Ser Leu
2000
ACT CCG AAA GAA GCA CTT TAT GAA GCC TCT CGG AAT TTA ATT GAT TTA TTA ATT
Thr Pro Lys Glu Ala Leu Tyr Glu Ala Ser Arg Asn Leu Ile Asp Leu Leu Ile
2100
CCT TTC CTA CAT GCG GAA GAA AAC GTA AAT TTA GAG GAC AAT CAA CAC AAA GTT
Pro Phe Leu His Ala Glu Glu Asn Val Asn Leu Glu Asp Asn Gln His Lys Val
2150
TCT TTA CCC CTT TTT ACC TTT CAT AAT AGA TTG GCT GAA ATA AGG AAA AAC AAA
Ser Leu Pro Leu Phe Thr Phe His Asn Arg Leu Ala Glu Ile Arg Lys Asn Lys
2200
AAA AAA ATA GCA TTG AAA TTC ATT TTT ATT GAC CAA TTA GAA TTG CCA CCT AGG
Lys Lys Ile Ala Leu Lys Phe Ile Phe Ile Asp Gln Leu Glu Leu Pro Pro Arg
2250
ATC TAC AAT TGC CTA AAA AAA TCC AAT ATA CAC ACA TTA TTG GAC CTT TTG AAT
Ile Tyr Asn Cys Leu Lys Lys Ser Asn Ile His Thr Leu Leu Asp Leu Leu Asn
2300
AAC AGT CAA GAA GAT CTT ATT AAA ATG AAG CAT TTT CGC ATA GAA GAC GTA AAA
Asn Ser Gln Glu Asp Leu Ile Lys Met Lys His Phe Arg Ile Glu Asp Val Lys
2350
CAA ATA TTT GGC ACT CTA GAA AAG CAT TTC GTA ATT GAT TTA AAA AAT AAA AGA
Gln Ile Phe Gly Thr Leu Glu Lys His Phe Val Ile Asp Leu Lys Asn Lys Arg
2400
TGA AAAA AAGT AATTGA ATAGATGTAT CTAGGGAAAA TTCACGTTGA AGTGACTATT
Ter
2450
CCCTAGATAC ACACGCCGTG TTATTTTCATA ATTAAATCAA TTTAAAAAAG GCCTAAAGTT
2500

```

Figure 2. DNA sequence of 2.5 kbp of the *Sall* fragments 8 and 11 of spinach plastid DNA, and the translation of the loci designated *infA*, *rps11*, and *rpoA*. These genes are believed to encode, respectively, homologues of *E. coli* initiation factor IF-1, ribosomal protein S11, and the α -subunit of DNA-dependent RNA polymerase. Sequences capable of forming secondary structures are underlined or overlined. Putative ribosome-binding sites are boxed; for C-terminal boxed sequences see Fig. 7.

```

      10      20      30      40      50      60
MKEQKWIHEGLITESLPNGMFWVRDNEPILGYVSGRIRRSSIRILPGDRVKIEVSRVD
:..  .: . :.....: :... . .....: .....:
AKEDNIEMQGTVLETLPNTMFRVELENGHVTAHISGKMRKNYIRILTGDKVTVELTPYD

      70
STRGRIIYRLRNKDSND
.:.....:
LSKGRIVFRSR

```

Figure 3. Amino acid sequence alignment between urf77 encoded in SalI fragment 11 of spinach plastid DNA (top) and protein synthesis initiation factor IF-1 of *E. coli* (bottom). The *E. coli* IF-1 protein was sequenced by Pon et al. (26). The alignment was generated with the homology search algorithm described by Lipman and Pearson (12). The two polypeptides have 42.9% identity (:) in amino acid sequence. An additional 39% of the amino acids can be aligned (.) via an amino acid replaceability matrix (see text).

between the two polypeptides for the first 30, or last 10-20 amino acids. The overall homology has 26% identical residues over a 301 amino acid overlap. Within this overlap an additional 38% of the codons align via the homology replaceability matrix used by FASTP.

The spinach plastid gene encodes a polypeptide of MW 38,441. The peptide has 25% polar, ionizable residues, including 43 arginine/lysine and 40 aspar-

```

      10      20      30      40      50      60
MAKPIPKIGSRNRGRISRSKARK-IPKGVHIVQASFNNTIVTVDVRGRVSVWASAGTCG
.....: .....: .....: .....:
AKAPIRARKVRKQVSDGVAHIHASFNNITIVTITDRQGNALGWATAGGSG

      70      80      90     100     110     120
FRGTKRGTPFAAQTAAGNAIRTVEQGMQRAEVMIKGPSLGRDAALRAIRRSIGILLSFVR
:.....: .....: : : :... :.....: .....: .: . .
FRGSRKSTPFAAQVAAERCADAVKEYGIKNLEVMVKGPGPGRESTIRALNAAGFRITNIT

      130
NVTPMPHNGCRPPKKRRV
.:.....:
DVTPIPHNGCRPPKKRRV

```

Figure 4. Amino acid sequence alignment between ribosomal protein S11 encoded in SalI fragment 8 of spinach plastid DNA (top) and ribosomal protein S11 of *E. coli* (bottom). The *E. coli* S11 protein was sequenced by Kamp and Wittmann-Liebold (27). The alignment was generated with the homology search algorithm described by Lipman and Pearson (12). The two polypeptides have 52% identity (:) in amino acid sequence over a 127 residue overlap. An additional 35% of the amino acids can be aligned (.) via an amino acid replaceability matrix.


```

      10      20      30      40      50      60
MVREKIRVSTQTLQWKCVESRTDSKCLHYGRFILSPLMKGQADTIGIAMRRALLGEIEGT
      .. :.: :.: :.: :.: :
      MQGSVTEFLKPRLVDIEQVSSTHAKVTLEPLERGFGHTLGNALRRILLSSMPGC

      70      80      90      100     110     120
CITRAKSEKIPHQYSTILGIQESVHEILMNLKEIVLRSNLYGTCEASICVRGPRGVTAQD
      :. :. :. :.: :.: :.: :.: :.: :.: :.: :.: :.: :.: :.: :.: :.:
AVTEVEIDGVLHEYSTKEGVQEDILEILLNLKGLAVRVQKGDEVILTlnKSGIGPVTAAD

      130     140     150     160     170
IILPPYVEIVKNTQHIASLT-EPIDLCIGLQLERNRGY-HIKAPNNFQDGSFPIDALFM-
      :. :.: :.: :.: :.: :.: :.: :.: :.: :.: :.: :.: :.: :.: :.: :.:
ITHDGDVEIVKPKQHVICHLTDENASISMRIKVQRGRGVVPASTRIHSEEDERPIGRLDVD

      180     190     200     210     220     230
-PVRNVNHSIHSYGNNGNEKQEILFLEIWTNGSLTPKEALYEASRNLIIDLLIPFLHAEENV
      :. :. :. :. :. :. :. :. :. :. :. :. :. :. :. :. :. :. :. :. :.
ACYSPVERIAYNVEAARVEQRTDLDKLVIEMETNGTIDPEEAIRRAATILAEQLEAFVDL

      240     250     260     270     280     290
NLEDNQHKVSLPLFTFHNRLAEIRKNKKKIALKFIFIDQLELPPRIYNCLKKSNIHTLLD
      :. :.: :.: :.: :.: :.: :.: :.: :.: :.: :.: :.: :.: :.: :.: :.:
R-DVRQPEVKEEKPEFDPIL-----LRPVDDLELTVRSANCLKAEAIHYIGD

      300     310     320     330
LLNNSQEDLIKMKHFRIEKVKQIFGTLEKHFVIDLKNKR
      :.: :.: :. :.: :.: :.: :.: :.: :.: :.: :.: :.: :.: :.: :.:
LVQRTEVELLKTPLNGKKSLSLTKDVLASRGLSLGMRLNWPASIAD

```

Figure 5. Amino acid sequence alignment between urf335 encoded in SalI fragment 8 of spinach plastid DNA (top) and the α -subunit polypeptide of *E. coli* DNA-dependent RNA polymerase (bottom). The alignment was generated with the homology search algorithm described by Lipman and Pearson (12). The two polypeptides show a 25.9% identity over a 301 residue overlap.

tate/glutamate. This is comparable to the MW 36,470 *E. coli* α -subunit polypeptide with 29% polar, charged residues.

Expression of the Genes

In order to determine if the *infA*, *rps11*, and *rpoA* loci are expressed in vivo, initial hybrid select translation and Northern blot analysis have been performed. The results of RNA blot hybridization with DNA probes from different sites within the DNA region display more than half a dozen RNA species, ranging from 6.0 kb to less than 1.0 kb. Some of these transcripts span the SalI-8/-11 as well as the SalI-11/-9 junctions. From this preliminary data, it appears that *infA*, *rps11*, and *rpoA* are all transcribed in vivo. The complex transcription pattern is reminiscent to that of the adjacent gene cluster *psbB* (51 kd chlorophyll α apoprotein) - *petB* (cytochrome *b6*) - *petD* (subunit

Figure 6. Translation of hybrid selected RNA species (22) in a rabbit reticulocyte assay. Spinach chloroplast RNA was hybridized to immobilized SalI fragments of spinach plastid DNA. The selected and released RNA fractions were translated in the presence of ^{35}S -methionine and the products analysed by polyacrylamide gel electrophoresis (dodecylsulfate-containing 10-15% gel) and fluorography. Cell-free products synthesized from RNA selected by (lane 1) fragment SalI-9; (2) SalI-11; (3) SalI-8; and (4) SalI-7. The map positions of these fragments are outlined in Fig. 1. Lanes 5 and 6 are shorter exposures of the patterns shown in lanes 3 and 4 (7 vs. 2 days). Potential candidates for the RNA polymerase α -subunit, ribosomal protein S11 and initiation factor IF-1 are marked by arrows. The identity of most other polypeptide species is known, including the 51 kd chlorophyll *a* apoprotein of photosystem II (51 kd), its immunologically reacting truncated products (single dots), cytochrome *b*₆ (cyt *b*₆), subunit IV (SU4) of the cytochrome *b/f* complex and its truncated product (double dots; cf. also 15). Molecular weight scale in kd on the left.

IV; 2, 29, 33) which is transcribed into a 6 kbp polycistronic RNA, and subsequently extensively modified by nucleolytic cleavage and intercistronic splicing (2, 29 and unpublished results).

The results of hybrid select translation using the fragments SalI-7, -8, -11 or -9 are shown in Fig. 6. For each fragment characteristic translation patterns are obtained. Included are translation products matching the mole-

Figure 7. Nucleotide sequence of the intercistronic region between the genes for RNA polymerase α -subunit (*rpoA*) and subunit IV of the cytochrome *b/f* complex (*petD*; 29). The sequence includes the last two codons for each of these genes and possible transcription terminator structures. The stem-and-loop structure of only one strand is shown. Arrow in the upper sequence marks a *petD* mRNA 3' terminus determined by S1 protection analysis. Numbering of nucleotides corresponds to that in Fig. 2. Note the base correction at position 2418 (cf. 29). Corresponding hexanucleotide motifs flanking the dyad structure on both strands are boxed (cf. text).

cular weights expected for IF-1, ribosomal protein S11, and the RNA polymerase α -subunit. These proteins therefore are potential candidates for the chloroplast polypeptides, but their actual identification must await further characterization.

The putative *infA* - *rps11* - *rpoA* operon and the *psbB* - *petB* - *petD* operon map on opposite strands and are transcribed towards each other (Fig. 1). Their RNA and protein patterns are different. A secondary structure (nucleotide positions 2406-2451) appears to function as termination signal for the polycistronic *psbB* - *petB* - *petD* transcription unit. Based on S1 nuclease mapping, the 3'-end of the transcript occurs at nucleotide 2403 (unpublished result). The same secondary structure feature may also serve as the transcription terminator for *rpoA* (Fig. 7). It is worth noting in this context that the complementary strands share two hexanucleotide motifs at corresponding positions, preceding this dyad structure (Fig. 7).

DISCUSSION

Significance of the *infA*, *rps11* and *rpoA* Identifications

In order to compare the possible significance of the identification of spinach plastid *infA*, *rps11*, and *rpoA* genes with plastid genes previously

TABLE 1. Similarity scores, homology, and z-values for alignment of chloroplast proteins and their *E. coli* homologues. Proteins were aligned with the FASTP algorithm over the number of amino acids in the "overlap" column, with the identical residues listed as "%Homology". "Initial" and "Optimal" similarity scores, and "z-value" (12) are described in the text. All of the alignments are for plastid DNA-coded polypeptides derived from DNA sequence data, except rpl12 which is from a protein sequence of a nuclear gene product.

Gene	Source	Initial	Optimal	%Homology	Overlap	z-Value	Ref.
infA	Spinach	166	186	42.9	70	28.3	-
rpoA	Spinach	166	207	32.7	150	33.2	-
rps11	Spinach	322	345	52.0	127	61.4	-
rpl2	Spinach <i>N. debneyi</i>	545	552	49.8	70	96.8	(8)
rpl12	Spinach	157	246	48.8	121	39.4	(34)
rps4	Maize	118	322	36.1	205	52.5	(4)
rps7	<i>Euglena</i>	139	320	40.5	153	54.8	(5)
rps12	<i>Euglena</i>	442	465	69.7	122	85.3	(5)
rps14	<i>Marchantia</i>	135	206	45.9	98	32.7	(6)
rps19	Tobacco Spinach <i>N. debneyi</i>	267	275	51.1	90	48.4	(7, 8)
tufA	<i>Euglena</i>	673	685	70.1	403	125	(11)

identified as homologues for bacterial ribosomal protein genes, the previously published sequences for chloroplast ribosomal proteins were each analysed against the NBRF Library via the FASTP homology search. The results are summarized in Table 1. In each case there was a single highly significant match with the previously identified *E. coli* ribosomal protein. The optimized similarity scores ranged from 206-685, with z-values in the range 32.7 to 125. Amino acid sequence identity varied from 36.1% to 70.1%. For each of these gene products, no other alternative protein in the NBRF Library gave an optimized similarity score >100. Using this group of plastid genes as a basis of comparison, the identification of rps11 would fall near the middle of the comparison group, while both infA and rpoA are at the low end, yet clearly in the "highly significant" group as defined by Lipman and Pearson (12). Identification of infA, rps11, and rpoA are also appropriate in the sense of the biological context for these genes. The gene products of these loci would be logical components of the unique chloroplast translational and transcriptional apparatus.

Significance of the Proteins

In *E. coli*, S11 has been designated "fidelity protein" since its absence causes decreased fidelity of translation (35, 36). No specific function has as yet been assigned to IF-1 which is the smallest of the translational initiation factors in *E. coli* (26, 37). It appears to associate with the 30S ribosomal subunit and the 70S ribosome, seems to stimulate binding and activities of IF-2 and IF-3, and act in recycling of IF-2 at the 70S level.

E. coli DNA-dependent RNA polymerase has a subunit structure of $\beta'\beta\alpha_2\sigma$ (38). The subunit molecular weights based on electrophoretic mobility are 165,000, 155,000, 39,000, and 95,000. The bacterial subunit genes are designated rpoC (β' -subunit), rpoB (β -subunit), rpoA (α -subunit), and rpoD (σ -subunit). Watson and Surzycki (39) have reported that both plastid and nuclear DNA from *Chlamydomonas reinhardtii* contain sequences homologous with rpoB and rpoC but this conclusion is based on Southern hybridizations at low stringency. More recently, a *Euglena* plastid gene homologous to rpoC (similarity score 321) has been identified and partially sequenced (Hallick and Radebough, unpublished observations), a finding that provides independent evidence for homologues of *E. coli* RNA polymerase subunit genes being encoded in plastid DNA.

Two different types of RNA polymerase activities appear to be present in spinach chloroplasts. One is a soluble enzyme active in tRNA and mRNA transcription (40), and the other is associated with a DNA-protein complex and is preferentially active in rRNA synthesis (41). The latter enzyme has major subunits of 69, 60, 55, 34 and 15 kd (42). The subunit structure of the spinach chloroplast soluble polymerase activity is not known, but a possibly analogous enzyme from maize has subunits of 180, 160, 66, 51, 43, 41, 28, and 15 kd (43). A pea chloroplast RNA polymerase has been described with subunits of 180, 140, 110, 95, 65, 47, and 27 kd (44). It is not yet possible to relate the putative spinach plastid rpoA gene product to any of these chloroplast RNA polymerase subunits. The maize and spinach enzymes each have a subunit that is in approximately the 38-39 kd size range expected for the spinach plastid rpoA gene product, but the identification of the role, if any, of the rpoA polypeptide in chloroplast transcription must await further study.

Significance of rps11 and rpoA as Plastid Genes

In *E. coli*, genes for ribosomal protein S11 and the α -subunit of DNA-dependent RNA polymerase are both part of the α -operon. The α -operon is a unit of gene regulation transcribed in the order rpsM (ribosomal protein S13) - rpsK (ribosomal protein S11) - rpsD (ribosomal protein S4) - rpoA (α -sub-

unit) - rplQ (ribosomal protein L17; 31, 45). The occurrence of closely spaced rps11 - ropA genes in the spinach plastid chromosome is reminiscent of a partial bacterial α -operon. The presence of rps11 adjacent to this 335 residue open reading frame is consistent with its identification as rpoA. Translation of the 4 ribosomal protein genes in *E. coli* is regulated by the levels of the S4 protein (30). Ribosomal protein S4 has been shown to be plastid coded in maize, but the gene is not flanked by other α -operon genes (4).

There are other examples of plastid genes being encoded in operons that are organized as in *E. coli*, but with some genes being deleted. The *E. coli* strep operon contains the genes rpsL (S12) - rpsG (S7) - fusA (EF-G) - tufA (EF-Tu) (46), while in *Euglena* plastids, the organization is rps12 - rps7 - tufA (5) and in tobacco plastids the organization is rps12 - rps7 (Hildebrand, Bourque, and Hallick, unpublished observation). The *E. coli* S10 operon encodes in order of transcription the genes for ribosomal proteins S10, L3, L4, L23, L2, S19, L22, S3, L16, L29, and S17, but in spinach, tobacco, soybean, and *N. debneyi* plastids, only the genes for proteins S19 and L2 appear to be linked (8). Finally, the *E. coli* atp operon bears the genes for the eight ATP synthase subunit species in the order subunits a - c (proteolipid) - b - atpD (δ) - atpA (α) - atpC (γ) - atpB (β) - atpE (ϵ , 47, 48). The chloroplast ATP synthase contains 9 subunit species (cf. 49). In spinach, six of them are plastid encoded in form of two operons, including atpI (homologous to *E. coli* subunit a; Hennig and Herrmann, unpublished data) - atpH (proteolipid) - atpF (homologous to subunit b) - atpA - and, 40 kbp away from this cluster, atpB - atpE (cf. 20, 49, 50). The positional and sequence conservation of all these genes is one of the most striking general feature, in that they demonstrate common phylogenetic roots between plastids and eubacteria.

ACKNOWLEDGEMENTS

We thank Prof. Dr. R. Mache (Grenoble) and Dr. A. R. Subramanian (Berlin) for their help in the initial identification of rps11, Ms. Ursula Borucki for the preparation of the manuscript. This study was supported by grants from the Stiftung Volkswagenwerk and the Deutsche Forschungsgemeinschaft (He 693). RBH was supported by grants GM35625 and GM35665 from the National Institute of Health.

*To whom correspondence should be addressed

REFERENCES

1. Bottomley, W. (1980) in Results and Problems in Cell Differentiation, Reinert, J., Ed., Vol 10, pp. 179-199. Springer Verl., Berlin, Heidelberg, New York.
2. Herrmann, R.G., Westhoff, P., Alt, J., Tittgen, J. and Nelson, N. (1985) in Molecular Form and Function of the Plant Genome, van Vloten-Doting, L., Groot, G. and Hall, T.C., Eds., pp. 233-256. Plenum Publ. Corp., New York, London.
3. Crouse, E.J., Schmitt, J.M. and Bohnert, H.J. (1985) Plant Mol. Biol. Reporter 3, 43-89.
4. Subramanian, A.R., Steinmetz, A. and Bogorad, L. (1983) Nucl. Acids Res. 11, 5277-5286.
5. Montandon, P.E. and Stutz, E. (1984) Nucl. Acids Res. 12, 2851-2859.
6. Umesono, K., Inokuchi, H., Ohya, K. and Ozeki, H. (1984) Nucl. Acids Res. 12, 9551-9565.
7. Sugita, M. and Sugiura, M. (1983) Nucl. Acids Res. 11, 1913-1918.
8. Zurawski, G., Bottomley, W. and Whitfield, P.R. (1984) Nucl. Acids Res. 12, 6547-6558.
9. Watson, J.C. and Surzycki, S.C. (1982) Proc. Natl. Acad. Sci. USA 79, 2264-2267.
10. Passavant, C.W., Stiegler, G.L. and Hallick, R.B. (1982) J. Biol. Chem. 258, 693-695.
11. Montandon, P.E. and Stutz, E. (1983) Nucl. Acids Res. 11, 5877-5892.
12. Lipman, D.J. and Pearson, W.R. (1985) Science 227, 1435-1441.
13. Herrmann, R.G., Seyer, P., Schedel, R., Gordon, K., Bisanz, C., Winter, P., Hildebrandt, J.W., Wlaschek, M., Alt, J., Driesel, A.J. and Sears, B.B. (1980) in Biological Chemistry of Organelle Formation, Bücher, Th., Sebald, W. and Weiss, H., Eds., pp. 97-112. Springer Verl., Berlin, Heidelberg, New York.
14. Herrmann, R.G., Westhoff, P., Alt, J., Winter, P., Tittgen, J., Bisanz, C., Sears, B.B., Nelson, N., Hurt, E., Hauska, G., Viebrock, A., Sebald, W. (1983) in Structure and Function of Plant Genomes, Cifferi, O. and Dure III, L., Eds., pp. 143-154. Plenum Publ. Corp., New York, London.
15. Alt, J., Westhoff, P., Sears, B.B., Nelson, N., Hurt, E., Hauska, G. and Herrmann, R.G. (1983) EMBO J. 2, 979-986.
16. Clewell, D.B. (1972) J. Bact. 110, 667-676.
17. Maxam, A.M. and Gilbert, W. (1980) Methods Enzymol. 65, 499-560.
18. Wu, R. (1970) J. Mol. Biol. 51, 501-521.
19. Kröger, M. and Kröger-Block, A. (1984) Nucl. Acids Res. 12, 113-120.
20. Westhoff, P., Nelson, N., Bünemann, H. and Herrmann, R.G. (1981) Curr. Genet. 4, 109-120.
21. Westhoff, P., Alt, J., Nelson, N., Bottomley, W., Bünemann, H. and Herrmann, R.G. (1983) Plant Mol. Biol. 2, 95-107.
22. Bünemann, H., Westhoff, P. and Herrmann, R.G. (1982) Nucl. Acids Res. 10, 7163-7180.
23. Wienand, U., Schwarz, Z. and Feix, G. (1979) FEBS Lett. 98, 319-323.
24. Rigby, P.W.J., Dieckmann, M., Rhodes, C. and Berg, P. (1977) J. Mol. Biol. 113, 237-251.
25. Herrmann, R.G., Whitfield, P.R. and Bottomley, W. (1980) Gene 8, 179-191.
26. Pon, C.L., Wittmann-Liebold, B. and Gualerzi, C.O. (1979) FEBS Lett. 101, 157-160.
27. Kamp, B. and Wittmann-Liebold, B. (1980) FEBS Lett. 121, 117-122.
28. Hallick, R.B. and Bottomley, W. (1983) Plant Mol. Biol. Reporter 1, 38-43.
29. Heinemeyer, W., Alt, J. and Herrmann, R.G. (1984) Curr. Genet. 8, 543-549.
30. Post, L.E., Arfsten, A.E., Davis, G.R. and Nomura, M. (1980) J. Biol. Chem. 255, 4653-4659.

31. Meek, D.W. and Hayward, R.S. (1984) Nucl. Acids Res. 12, 5813-5821.
32. Bedwell, D., Davis, G., Gosink, M., Post, L., Nomura, M., Kestler, H., Zengel, J.M. and Lindahl, L. (1985) Nucl. Acids Res. 13, 3891-3903.
33. Morris, J. and Herrmann, R.G. (1984) Nucl. Acids Res. 12, 2837-2850.
34. Bartsch, M., Kimura, M. and Subramanian, A.R. (1982) Proc. Natl. Acad. Sci. USA 79, 6871-6875.
35. Nomura, M., Mizushima, S., Ozaki, M., Traub, P. and Lowry, C.W. (1969) Cold Spring Harbor Symp. Quant. Biol. 34, 49-61.
36. Pongs, O., Stöffler, G. and Wittmann, H.G. (1971) Eur. J. Biochem. 23, 7-11.
37. Thach, R.E., Hershey, J.W.B., Kolakofsky, D., Dewey, K.F. and Remold-O'Donnel, E. (1969) Cold Spring Harbor Symp. Quant. Biol. 34, 277-284.
38. Burgess, R.R. (1976) in RNA Polymerase, Losick, R. and Chamberlin, M.J., Eds., pp. 69-100. Cold Spring Harbor Lab., New York.
39. Watson, J.C. and Surzycki, S.J. (1983) Curr. Genet. 7, 201-210.
40. Grussem, W., Greenberg, B.M., Zurawski, G., Prescott, D.M. and Hallick, R.B. (1983) Cell 35, 815-828.
41. Briat, J.-F., Laulhere, J.P. and Mache, R. (1979) Eur. J. Biochem. 98, 285-292.
42. Briat, J.-F. and Mache, R. (1980) Eur. J. Biochem. 111, 503-509.
43. Kidd, G.H. and Bogorad, L. (1980) Biochim. Biophys. Acta 609, 14-30.
44. Tewari, K.K. and Goel, A. (1983) Biochemistry 22, 2142-2148.
45. Post, L.E. and Nomura, M. (1979) J. Biol. Chem. 254, 10604-10606.
46. Nomura, M., Morgan, E.A. and Jaskunas, S.R. (1977) Ann. Rev. Genet. 11, 297-347.
47. Nielsen, J., Hansen, F.G., Hoppe, J., Friedl, P. and von Meyenburg, K. (1981) Mol. Gen. Genet. 184, 33-39.
48. Saraste, M., Gay, N.J., Eberle, A., Runswick, M.J. and Walker, J.E. (1981) Nucl. Acids Res. 9, 5287-5296.
49. Westhoff, P., Alt, J., Nelson, N. and Herrmann, R.G. (1985) Mol. Gen. Genet. 199, 290-299.
50. Zurawski, G., Bottomley, W. and Whitfeld, P.R. (1982) Proc. Natl. Acad. Sci. USA 79, 6260-6264.