Biochem. J. (1986) 236, 453–460 (Printed in Great Britain)

Pea chloroplast DNA encodes homologues of *Escherichia coli* ribosomal subunit S2 and the β′-subunit of RNA polymerase

Alison L. COZENS and John E. WALKER*
MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, U.K.

The nucleotide sequence has been determined of a segment of 4680 bases of the pea chloroplast genome. It adjoins a sequence described elsewhere that encodes subunits of the F_0 membrane domain of the ATP-synthase complex. The sequence contains a potential gene encoding a protein which is strongly related to the S2 polypeptide of *Escherichia coli* ribosomes. It also encodes an incomplete protein which contains segments that are homologous to the β′-subunit of *E. coli* RNA polymerase and to yeast RNA polymerases II and III.

INTRODUCTION

The chloroplast genome is a circular double-stranded DNA molecule, on average about 150 kb in size (for reviews see Bonhert *et al.*, 1982; Whitfeld & Bottomley, 1983; Dyer, 1984, 1985). It codes for about 70 proteins. Some of these have been identified and are involved in photosynthesis. They include the large subunit of ribulose 1,5-bisphosphate carboxylase, a soluble enzyme located in the stroma, and components of the multi-subunit thylakoid membrane complexes, photosystems I and II, cytochrome *b–f* complex and ATP-synthase (reviewed by Walker & Tybulewicz, 1985). These protein components are made in the chloroplast by a transcriptional and translational machinery localized in the stroma. This apparatus is distinct from that used for production of transcripts and proteins from nuclear genes, and components of it are also encoded in the plastid DNA. For example, chloroplast ribosomal 23 S, 16 S and 5 S RNA species are transcribed from chDNA, as are all tRNAs required for chloroplast protein synthesis, although the number of different tRNAs involved has not been clearly established (review by Dyer, 1985).

Other evidence indicates that some of the ribosomal protein subunits are chloroplast gene products. In *Chlamydomonas reinhardtii* five or six out of 31–33 proteins of the large ribosomal subunit and 14 of 31 small subunit proteins are estimated to be products of chloroplast protein synthesis (Schmidt *et al.*, 1983). In spinach at least seven to ten 30 S and six to thirteen 50 S ribosomal proteins (Dorne *et al.*, 1984; Posno *et al.*, 1984) and in pea at least six 30 S and five 50 S ribosomal components are thought to be chloroplast coded (Eneas-Filho *et al.*, 1981). The presence in the chloroplast genomes of genes encoding homologues of *E. coli* ribosomal proteins has been demonstrated by hybridization and DNA sequence analysis. For example, homologues of *E. coli* ribosomal protein S19 have been found in chDNA from two species of *Nicotiniana* (Sugita & Sugiura, 1983; Zurawski *et al.*, 1984). Moreover, the gene for a homologue of the L2 subunit is found associated with it in *Nicotiniana debreyi* and spinach (Zurawski *et al.*, 1984). Also, a homologue of the *E. coli* S4 protein is encoded in maize chDNA (Subramanian *et al.*, 1983) and homologues of S12 and S7 are co-transcribed in *Euglena gracilis* with a homologue of *E. coli* elongation factor Tu (Montadon & Stutz, 1984; Passavant *et al.*, 1983).

As described in the present paper, during the analysis of a fragment of pea chDNA containing the gene for a subunit of chloroplast ATP-synthase, we have discovered that adjacent DNA sequences encode homologues of *E. coli* ribosomal protein S2 and part of the β′-subunit of RNA polymerase. This strongly suggests that these proteins are also produced from the chloroplast DNA.

MATERIALS AND METHODS

Isolation of the fragment of chDNA

A 6.2 kb *SalI-PstI* fragment derived from plasmid pPscS6, a pBR322 plasmid containing a 10.6 kb *SalI* fragment of pea chDNA (see Fig. 1; Huttly & Gray, 1984) was prepared as described previously (Cozens *et al.*, 1986).

DNA sequence analysis

Random fragments 300–800 bp long were produced from a 6.2 kb *SalI–PstI* fragment by sonication. Then they were cloned into the M13 *mp8* vector that had previously been digested with *SmaI* (Bankier & Barrell, 1983). Sequences in the resultant M13 clones were determined by the dideoxynucleotide chain termination method (Sanger *et al.*, 1977) as modified by Biggin *et al.* (1983).

Data analysis

DNA sequences were compiled in a database with the computer program DBAUTO (Staden, 1982*a*). Each nucleotide in the sequence was determined six times on average and at least once on each strand of the DNA. The DNA sequences were analysed with ANALYSEQ (Staden, 1984). A number of options in this program were used to predict protein coding regions. Because of the high A+T content of the DNA the most effective procedure was that based on the positional base

Abbreviations used: chDNA, chloroplast DNA; kb, kilobase (pair); bp, base pair.

* To whom correspondence and reprint requests should be sent.

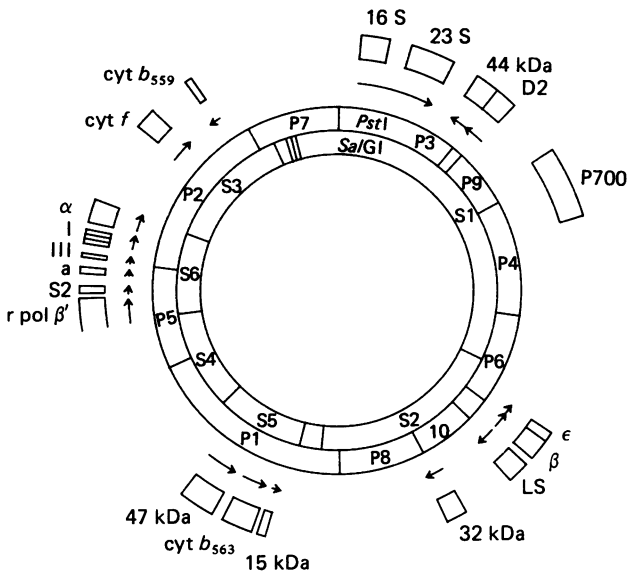


Fig. 1. Physical and genetic maps of the pea chloroplast genome

The outer arcs indicate genes that have been identified by hybridization and DNA sequencing. The arrows show the direction of transcription. Gene names (see Dyer, 1985): 16 S and 23 S, ribosomal RNAs; P700 is the P700/chlorophyll *a* apoprotein of photosystem 1; 32 kDa, 44 kDa, D2 and cyt b_{559} are components of photosystem II; cyt *f*, cyt b_{563} and 15 kDa are subunits of the cytochrome *b–f* complex; LS is the large subunit of ribulose bisphosphate carboxylase; α, I, III, a, β and ε are components of ATP synthase (Cozens *et al.*, 1986); S2 and r pol β′ are homologues of *E. coli* ribosomal subunits S2 and the β′ subunit of RNA polymerase respectively.

preference method (Staden, 1985). This exploits the preference for bases to occupy particular positions in codons, this being a function of the frequencies with which particular amino acids are found in proteins; thus it can be used to distinguish coding and non-coding regions. The calculations were weighted for a strong preference for A or T in the third codon position. The predicted protein sequences were analysed with ANALYSEP (R. Staden, unpublished work). This computer program contains a number of options for analysis of protein sequences including one used in the present study which calculates the M_r of a protein from its sequence. The protein sequences also were compared with database of the Protein Information Resource (PIR; National Biomedical Research Foundation, Georgetown University Medical Center, 3900 Reservoir Road NW, Washington, DC 20007, U.S.A.) by using the program FASTP (Lipman & Pearson, 1985). Homologous sequences were analysed further by pairwise comparison with DIAGON (Staden, 1982*b*).

RESULTS AND DISCUSSION

DNA sequence

The DNA sequence presented in Fig. 2 is 4680 bases long. It extends from the *Sal*I site and runs into the sequence encoding a membrane component of ATP-synthase which occupies most of the remainder of the *Sal*I–*Pst*I fragment (Cozens *et al.*, 1986). The A+T content of this sequence is 66.05%, corresponding to overall values of 61–67.3% for pea chDNA (Herrmann & Possingham, 1980). Open reading frames (potential genes) could be recognized in non-coding regions in both strands of the DNA. Their locations were confirmed by calculations employing the positional base preference method as explained above (see Fig. 3).

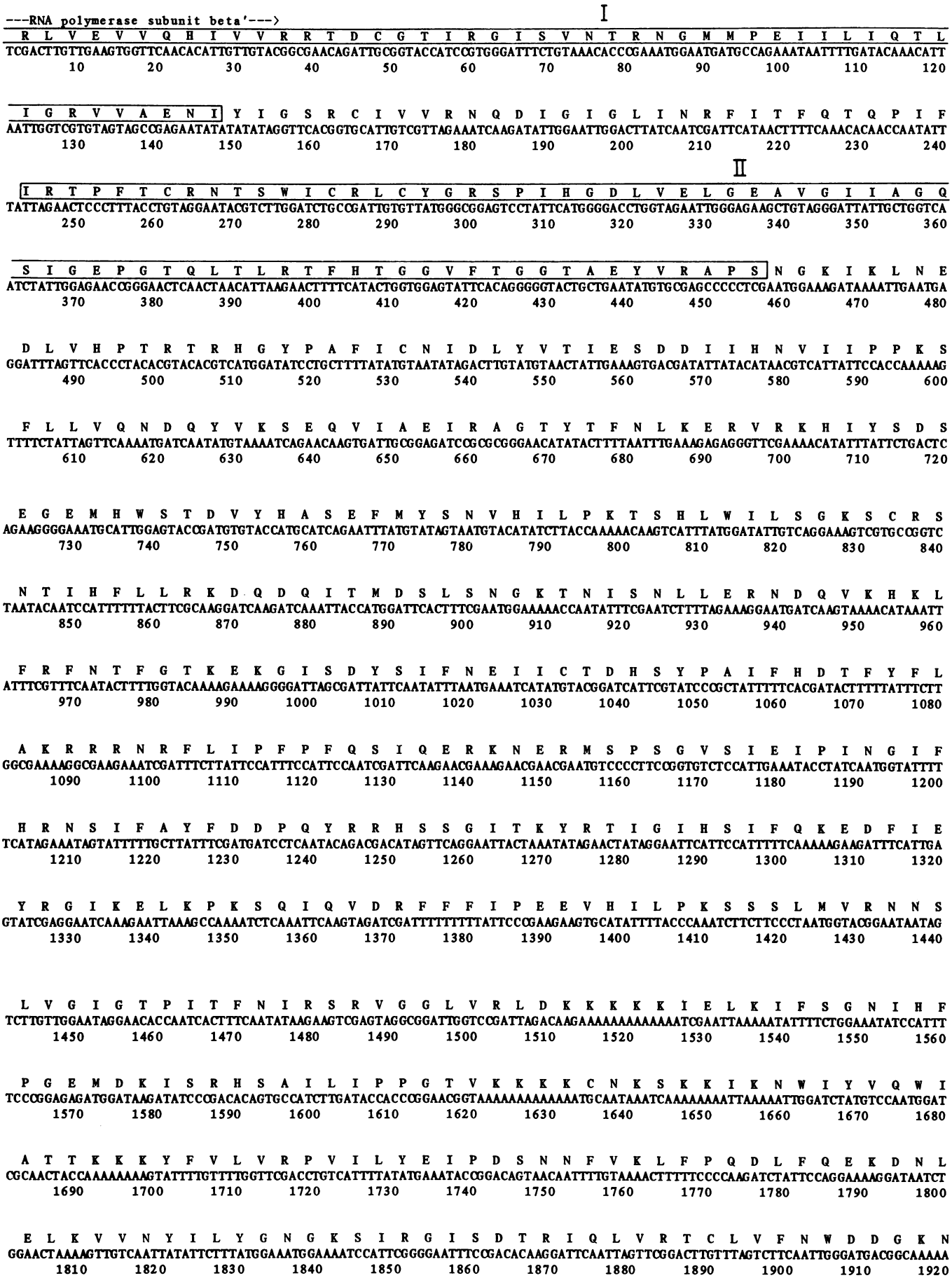
One complete open reading frame (bases 3749–4459) encodes a protein of 236 amino acids with an M_r of 26842. It is preceded by an incomplete open reading frame (bases 1–3493) and is followed by the gene coding for a protein identified as a component of chloroplast ATP-synthase (base 4669 onwards; Cozens *et al.*, 1986).

The ribosome component

By comparison of the predicted protein sequences, with protein sequences in the Protein Information Resource database, the complete open reading frame was found to encode a protein which is closely related in sequence to that of the component S2 of the small subunit of the *E. coli* ribosome (Wittmann-Liebold & Bosserhoff, 1981; An *et al.*, 1981). Their close relationship is emphasized by pairwise comparison of their sequences with the aid of the computer program DIAGON (Fig. 4). It is supported by alignment of the two protein sequences (Fig. 5) in which the two proteins are identical in about 40% of their residues. This close relationship in primary structure and the conservation of 16 S rRNA (Schwartz & Kössel, 1980; Tohdoh & Sugiura, 1982; Graf *et al.*, 1982) and other ribosomal subunits in bacteria and chloroplasts implies that the tertiary structures of the two proteins and their arrangements in the ribosome will also be similar. *E. coli* ribosomal protein S2 is an elongated molecule (Engleman *et al.*, 1975; Tischendorf *et al.*, 1975; Georgalis *et al.*, 1981). It can be cross-linked to ribosomal proteins S1, S3, S5 and S8 (Traut *et al.*, 1980) and also to translational initiation factor IF-2 (Bollen *et al.*, 1975). Also, it can bind to 16 S rRNA (Littlechild *et al.*, 1977). Immuno-electron microscopic studies have located S2 near the cleft between head and body of the 30 S subunit (Tischendorf *et al.*, 1975; Lake, 1978). Several mutants in S2 have been isolated, some being linked to resistance to the antibiotic kasugamycin (Okuyama *et al.*, 1974; Yoshikawa *et al.*, 1975). The gene is located near minute 4 in the *E. coli* chromosome and is co-transcribed with elongation factor EF-Ts (An *et al.*, 1981).

The component of RNA polymerase

The 3.5 kb region of DNA sequence at the 5′ end of the fragment presented in Fig. 2 contains an open reading frame which starts beyond the 5′ end and terminates at base 3493. It encodes a fragment (M_r 133591) of a protein. Segments of this sequence (boxed in Fig. 2 and summarized in Fig. 6) are related particularly to the sequences of the β′-subunit of *E. coli* RNA polymerase (Squires *et al.*, 1981; Ovchinnikov *et al.*, 1982) and also more weakly to the related yeast RNA polymerases II and III (Allison *et al.*, 1985). As summarized in Figs. 1 and 6, these relationships extend over five regions, I–V, of the pea chloroplast open reading frame. Between amino acids 150 and approx. 900, no relationship is found between the putative chloroplast protein and either the RNA polymerases or any other proteins in the database. The alignment with the *E. coli* RNA polymerase β′-subunit and the pea chloroplast protein requires that a large


---RNA polymerase subunit beta'--->
I
II

```
 S  S  S  I  E  E  A  P  A  S  F  I  E  V  R  T  N  G  L  I  E  Y  F  L  R  I  D  L  V  K  S  N  T  S  Y  I  R  K  R  N
CAGTTCTTCGATTGAGGAGGCCCCCGCTTCCTTTATTGAAGTACGTACAAATGGTTTGATTGAGTATTTTCTAAGAATTGACTTAGTAAAATCGAATACTTCATATATTAGAAAAAGAAA
     1930      1940      1950      1960      1970      1980      1990      2000      2010      2020      2030      2040

 E  P  S  G  F  G  L  I  G  D  N  K  S  D  R  I  N  P  F  F  S  I  H  S  K  G  K  I  Q  Q  S  L  S  Q  N  H  G  T  I  R
TGAACCATCTGGTTTCGGATTGATTGGGGATAATAAATCGGATCGAATCAATCCATTTTTTCTATTCATTCCAAGGGCAAAATTCAACAATCACTTAGCCAAAATCACGGAACTATTCG
     2050      2060      2070      2080      2090      2100      2110      2120      2130      2140      2150      2160

 M  L  L  N  R  N  K  E  C  R  S  W  I  I  L  S  S  S  N  C  F  Q  M  R  P  F  N  N  E  K  S  H  N  G  I  K  K  D  P  I
TATGTTGTTGAATAGAAATAAAGAATGCCGATCTTGGATAATTTTGTCATCATCTAATTGTTTTCAAATGAGACCATTCAACAATGAAAAATCTCACAATGGGATAAAAAAAGATCCTAT
     2170      2180      2190      2200      2210      2220      2230      2240      2250      2260      2270      2280

 I  S  I  N  N  N  G  P  L  G  I  A  L  Q  V  A  N  F  Y  S  L  Y  H  L  I  T  H  N  Q  I  S  I  I  K  N  L  Q  L  D  K
AATTTCAATTAATAATAATGGCCCTTTAGGAATAGCCCTTCAAGTTGCGAATTTTTATTCACTTTATCATTTAATAACTCATAATCAGATCTCAATAATTAAAAATTTGCAACTTGACAA
     2290      2300      2310      2320      2330      2340      2350      2360      2370      2380      2390      2400

 L  T  E  I  F  Q  V  I  K  Y  Y  L  M  D  E  N  D  K  I  C  K  P  D  L  Y  S  N  I  I  L  N  P  F  H  L  N  W  F  F  L
ATTAACGGAAATTTTTCAAGTAATTAAATATTATTTAATGGACGAAAATGATAAAATTTGTAAACCCGATCTATACAGTAATATCATTTTGAATCCATTCCATTTGAATTGGTTTTTTCT
     2410      2420      2430      2440      2450      2460      2470      2480      2490      2500      2510      2520

 H  H  F  Y  C  E  K  T  F  T  R  I  S  L  G  Q  F  I  C  E  N  I  C  I  A  Q  M  K  N  R  P  H  L  K  L  K  S  G  Q  V
CCATCATTTTTATTGTGAAAAAACGTTTACAAGAATAAGTCTTGGACAATTTATTTGTGAAAATATATGTATAGCTCAAATGAAAAATAGACCACACCTAAAACTAAAATCGGGTCAAGT
     2530      2540      2550      2560      2570      2580      2590      2600      2610      2620      2630      2640

 I  I  V  Q  M  D  S  V  I  I  R  S  A  N  P  Y  L  A  T  P  G  A  T  I  H  G  H [Y  G  E  I  L  S  Q  G  D  I  L  V  T]
TATAATAGTTCAAATGGATTCTGTAATAATAAGATCGGCTAATCCTTATTTGGCAACTCCAGGAGCAACCATTCACGGCCATTATGGGGAGATCCTTTCTCAAGGAGATATTTTAGTTAC
     2650      2660      2670      2680      2690      2700      2710      2720      2730      2740      2750      2760

                                III
[F  I  Y  E  K  S  R  S  G  D  I  T  Q  G  L  P  K  V  E  Q  I  L  E  I  R  S] I  D  S  I  S  M  N  L  E  K  R  I  D  A
ATTCATATATGAAAAATCGAGATCCGGTGATATAACACAAGGTCTTCCAAAAGTGGAACAGATATTAGAAATACGTTCGATTGATTCAATATCCATGAATCTAGAAAAAAGAATTGATGC
     2770      2780      2790      2800      2810      2820      2830      2840      2850      2860      2870      2880

 W  N  E  C  I  T  K  I  I  G  I  P  W  G  F  L  I  G  A  E  L  T  I  A  Q  S  R  I  S [L  V  N  K  I  Q  K  V  Y  R  S]
TTGGAACGAGTGTATAACAAAAATTATCGGCATTCCTTGGGGATTCTTGATTGGTGCTGAGCTAACTATCGCGCAAAGTCGTATTTCTTTGGTTAATAAAATCCAAAAGGTTTATCGATC
     2890      2900      2910      2920      2930      2940      2950      2960      2970      2980      2990      3000

                      IV
 [Q  G  V  H  I  H  N  R  H  I  E  I  I  V  R  Q  I  T  S  K  V] L  V  S  E  D  G  M  S  N  I  F  L  P  G  E  L  I  G  L
CCAGGGAGTACACATCCATAATAGACATATCGAGATTATTGTGCGTCAAATAACATCCAAAGTCTTGGTTTCAGAAGATGGAATGTCTAATATATTTTTACCTGGCGAACTAATTGGATT
     3010      3020      3030      3040      3050      3060      3070      3080      3090      3100      3110      3120

 L  R  A  E  R  T  G  R  A  L  E  E  A  I  C [Y  R  A  L  L  L  G  V  T  K  T  S  L  N  T  Q  S  F  I  S  E  A  S  F  Q]
ATTGCGAGCAGAACGAACGGGGCGTGCCTTGGAAGAAGCAATTTGTTACAGAGCTTTATTATTGGGAGTAACAAAAACATCTCTGAATACTCAAAGTTTCATATCCGAAGCGAGTTTTCA
     3130      3140      3150      3160      3170      3180      3190      3200      3210      3220      3230      3240

                                                  V
 E  T  A  R  V  L  A  K  A  A  L  R  G  R  I  D  W  L  K  G  L  K  E  N  V  V  L  G  G  M  I  P  V  G  T  G  F  K  R  I
AGAAACTGCTAGAGTTTTAGCAAAAGCCGCTCTTCGGGGTCGTATTGATTGGTTGAAAGGTCTTAAAGAGAATGTTGTTTTAGGGGGAATGATACCCGTTGGTACCGGATTCAAAAGAAT
     3250      3260      3270      3280      3290      3300      3310      3320      3330      3340      3350      3360

[M  H  R  S  R] S  R  Q  H  N  K  I  T  R  K  K  K  L  F  E  V  E  I  R  N  L  L  F  H  H  R  K  L  L  D  F  A  N  F  K
AATGCACCGTTCACGGTCAAGGCAACATAACAAGATTACCCGGAAAAAAAAAATTATTCGAAGTAGAAATTAGAAATCTTTTGTTCCATCACAGAAAATTATTGGATTTTGCTAATTTCAA
     3370      3380      3390      3400      3410      3420      3430      3440      3450      3460      3470      3480

 E  F  M  *
AGAATTTATGTGATACATTACAAATATAGGATTAAATGCTTCCTAAAAGCGGGACTCATCCCCTTAAATCTTTAAATGCAGTCATTTGATTTGGTAATTAAAGTATAATAAAAAATAATA
     3490      3500      3510      3520      3530      3540      3550      3560      3570      3580      3590      3600

ATAAATAGAAATCAAAGAATGGCCTGATTCTATATTAACGGCCGATCGTCTATAAAAGAGAAGGTTCCATCGGAACAATTATTTATTGCTATTTCAGGATACTGGGTCTCTCTCTTTTTT
     3610      3620      3630      3640      3650      3660      3670      3680      3690      3700      3710      3720

                    ribosomal protein S2 --->
                    M  T  K  R  Y  W  N  I  T  F  E  E  M  M  E  A  G  V  H  F  G  H  D  T  R  K  W  N  P  R  M
TTTTTTTTTAAACAAAGTGT[GGGGA]TAAATGACTAAAAGATATTGGAACATAACTTTTGAAGAGATGATGGAAGCTGGGGTTCATTTTGGCCATGATACTAGGAAATGGAATCCTCGAAT
     3730      3740      3750      3760      3770      3780      3790      3800      3810      3820      3830      3840
```

```
  A  P  F  I  S  A  K  R  K  G  I  H  I  T  N  L  T  K  T  A  R  F  L  S  E  A  C  D  L  A  F  D  A  A  S  K  G  K  Q  F
GGCACCTTTTATATCGGCAAAACGTAAAGGTATTCATATTACAAATCTTACTAAAACTGCTCGTTTTTTATCAGAAGCTTGTGATTTGGCTTTTGATGCAGCAAGCAAAGGAAAACAATT
     3850      3860      3870      3880      3890      3900      3910      3920      3930      3940      3950      3960

  L  I  V  G  T  K  K  K  A  A  D  S  V  T  R  A  A  I  R  A  R  C  H  Y  V  N  K  K  W  L  R  G  M  L  T  N  W  Y  T  T
TTTAATTGTTGGTACAAAAAAAAAAAGCAGCCGATTCAGTAACACGGGCTGCAATAAGAGCTCGATGTCATTATGTTAATAAAAAATGGCTCCGAGGTATGTTAACGAATTGGTATACTAC
     3970      3980      3990      4000      4010      4020      4030      4040      4050      4060      4070      4080

  E  T  R  L  G  K  F  R  D  L  R  T  E  Q  K  T  G  K  L  N  S  L  P  K  R  D  A  A  M  L  K  R  Q  L  S  H  F  E  T  Y
AGAAACGAGACTTGGTAAGTTCAGGGACTTGAGAACGGAGCAAAAGACGGGGAAACTGAACTCTCTTCCAAAAAGAGATGCGGCTATGTTGAAGAGACAATTATCTCATTTCGAAACATA
     4090      4100      4110      4120      4130      4140      4150      4160      4170      4180      4190      4200

  L  G  G  I  K  Y  M  T  G  L  P  D  I  V  I  I  V  D  Q  Q  K  E  Y  T  A  L  Q  E  C  I  T  L  G  I  P  T  I  C  L  I
TCTAGGCGGTATCAAATATATGACAGGGTTACCTGATATTGTAATAATCGTTGATCAGCAAAAAGAATATACGGCTCTTCAAGAATGTATCACTTTAGGAATTCCAACGATTTGTTTAAT
     4210      4220      4230      4240      4250      4260      4270      4280      4290      4300      4310      4320

  D  T  N  C  D  P  D  L  A  D  M  S  I  P  A  N  D  D  A  I  A  S  I  R  L  I  L  N  K  L  V  F  A  I  C  E  G  R  S  S
CGATACAAATTGTGACCCAGATCTTGCAGATATGTCGATTCCGGCTAATGATGATGCTATAGCTTCCATCCGATTAATTCTTAACAAATTAGTTTTTGCAATTTGTGAGGGTCGTTCTAG
     4330      4340      4350      4360      4370      4380      4390      4400      4410      4420      4430      4440

  S  I  R  N  Y  *
CTCTATACGAAATTATTGATTAATAATAAGATAAATCCATTTTTAGATTTGGTTGGGCGGTCATAGATTTTTGGAATTGGGTATTATAGCATTACAAAATTGTGTAAAAAGAAATATTTT
     4450      4460      4470      4480      4490      4500      4510      4520      4530      4540      4550      4560

                                                                                                          ATPase a --->
                                                                                                            M  N  V  L
GTGATTAGTAGGTATTCAAAATAGAAAATCAAAGTAAAATAAGGAAATGGTTGAATCAAAATAATTCCCTTTCAAGTTATATTTTTTTATTTTAGAGGACAGGGCAATATGAATGTTCTA
     4570      4580      4590      4600      4610      4620      4630      4640      4650      4660      4670      4680
```

Fig. 2. Nucleotide sequence of pea chDNA from a *SalI* site to the beginning of the gene for the ATP-synthase a subunit and sequences of encoded proteins

Transcription and translation are from left to right. Proposed ribosome binding sites are boxed. The boxed protein sequences are homologous to the β'-subunit of *E. coli* RNA polymerase.

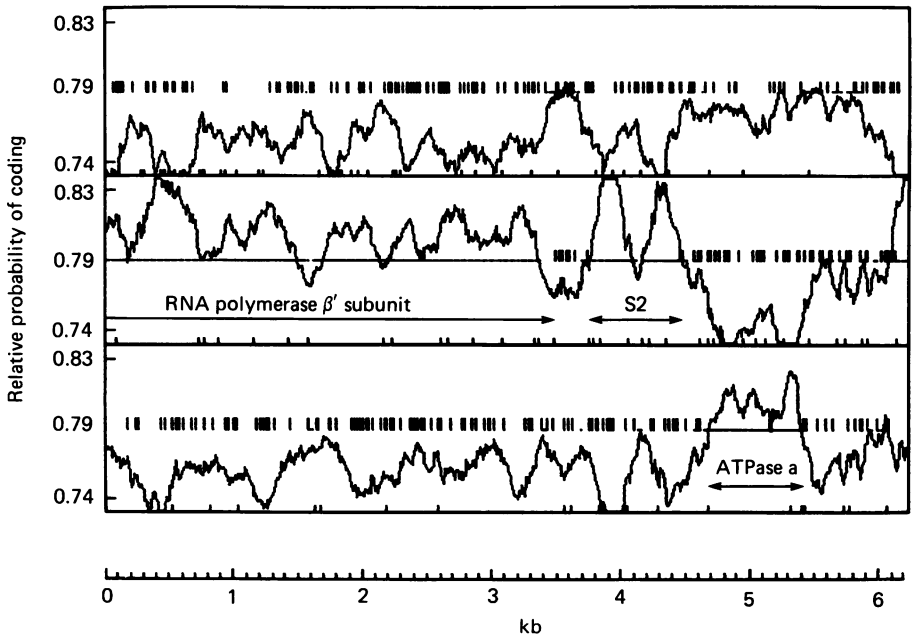


Fig. 3. Gene predictions in the DNA sequenced

The positional base preference method (Staden, 1985) was employed. The three boxes represent the translational reading frames of the DNA. The scale is in kb. The top of the scale on the ordinate represents the level that would be reached in a gene encoding a protein of average amino acid composition. Points under the midline are likely to be non-coding. Vertical bars on the abscissae represent possible start codons (ATG), those on the midlines indicate potential stop codons. The positions of the potential genes for the β'-subunit of RNA polymerase and ribosomal protein S2 are shown. The ATPase a subunit has been described elsewhere (Cozens *et al.*, 1986).

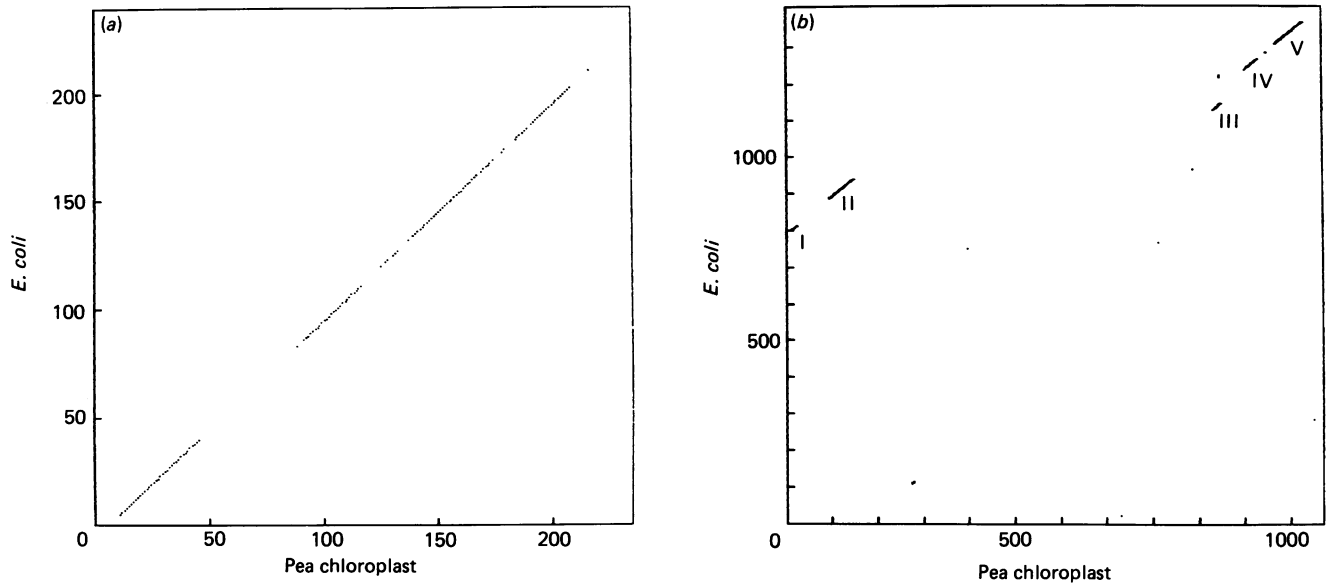


Fig. 4. Comparisons of the encoded proteins with the aid of the computer program DIAGON

(*a*) *E. coli* ribosomal subunit S2 (ordinate) versus the protein predicted by the complete open reading frame in the pea chDNA (abscissa; bases 3749–4459 in Fig. 1); (*b*) the β'-subunit of *E. coli* RNA polymerase (ordinate) versus the protein predicted by the incomplete open reading frame (abscissa; bases 1–3493). The sequences of homologous segments I–V are shown in Fig. 6. The numbers on the axes correspond to the lengths of the amino acid chains. The calculations were made with a window of 25 amino acids and a score of 290.

```
                1                                                   50
Pea chloroplast MTKRYWNITF EEMMEAGVHF GHQTRKWNPR MAPFISAKRK GIHIINLTKT
E.coli                ATVSM RDMLKAGVHF GHQTRYWNPK MKPFIFGARN KVHIINLEKT

               51                                                  100
Pea chloroplast ARFLSEACDL AFDAASKGKQ FLIVGTKKKA ADSVTRAAIR ARCHYVNKKW
E.coli          VPMFNEALAE LNKIASRKGK ILFVGTKRAA SEAVKDAALS CDQFFVNHRW

              101                                                  150
Pea chloroplast LRGMLTNWYT TETRLGKFRD LRTEQKTGKL NSLPKRDAAM LKRQLSHFET
E.coli          LGGMLTNWKT VRQSIKRLKD LETQSQDGTF DKLTKKEALM RTRELEKLEN

              151                                                  200
Pea chloroplast YLGGIKYMTG LPDIVIIVDQ QKEYTALQEC ITLGIPTICL IDTNCDPDLA
E.coli          SLGGIKDMGG LPDALFVIDA DHEHIAIKEA NNLGIPVFAI VDTNSDPDGV

              201                                              245
Pea chloroplast DMSIPANDDA IASIRLILNK LVFAICEGRS SSIRNY
E.coli          DFVIPGNDDA IRAVTLYLGA VAATVREGRS QDLASQAEES FVEAE
```

Fig. 5. Alignment of the sequences of the protein predicted from the pea chloroplast complete open reading frame with that of the *E. coli* ribosomal subunit S2

Identities are boxed.

insertion of about 250 amino acids be present in the chloroplast protein. This has two possible implications. Firstly, that the pea chloroplast enzyme differs substantially in structure from the bacterial enzyme, or secondly, that the pea chloroplast gene contains an intron of approx. 750 bp. In the absence of further experimental evidence it is not possible, at present, to know which interpretation is correct.

Chloroplasts have been shown to contain two, and possibly three, different RNA polymerase activities. In *E. gracilis* two types of activity have been isolated. One polymerase is tightly bound to chDNA and is selective for transcription of rRNA genes. It contains three major subunits of M_r 116000–118000, 83000–88000 and 24000–26000 (Narita *et al.*, 1985). The other is in the soluble fraction and is selective for transcription of tRNAs. It has not been determined whether chloroplast mRNA is transcribed by the soluble extract (Greenberg *et al.*, 1984). Higher plant chloroplast RNA polymerases that transcribe mRNA are composed of multiple subunits with a broad M_r range (Kidd & Bogorad, 1980). The pea enzyme may contain polypeptides of M_r 180000, 140000, 110000, 95000, 65000, 47000 and 27000 (Tewari & Goel, 1983), and it is conceivable that the fragment sequenced in the present work corresponds to one of these subunits.

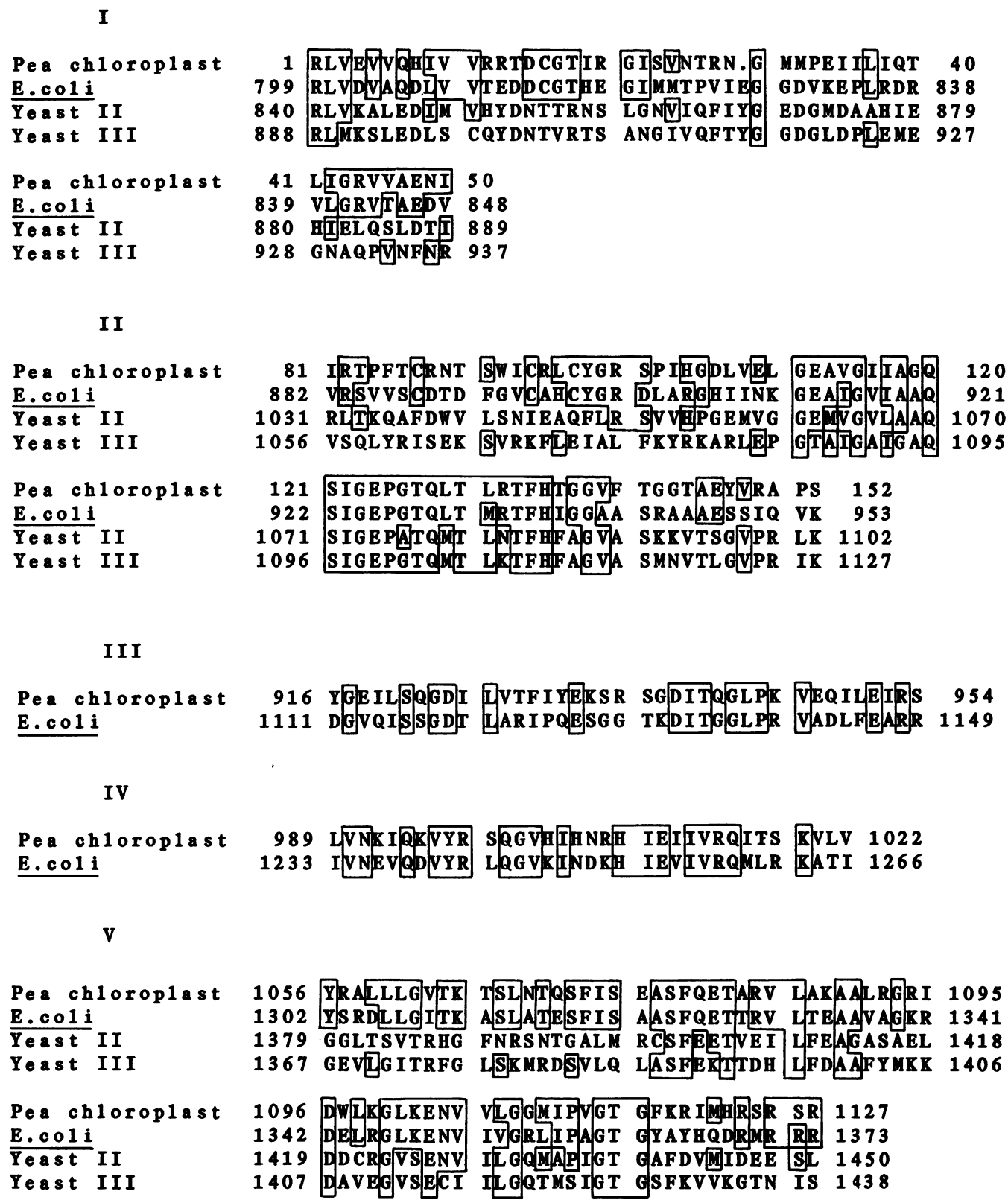


Fig. 6. Alignment of homologous segments of the protein predicted from the incomplete open reading frame in the pea chloroplast genome, the β′-subunit of *E. coli* RNA polymerase and yeast RNA polymerases II and III

Parts I–V corresponds to homologies found in Fig. 4(*b*). See also Fig. 2. Identities between the pea chloroplast protein and *E. coli* RNA polymerase and/or yeast RNA polymerases II and III are boxed.

Gene arrangements

The region of the pea chloroplast genome discussed in this paper has an arrangement of genes that is probably as follows: RNA polymerase β'-subunit:ribosomal subunit S2:ATPase *a*:ATPase III (see Fig. 1). Transcription of this region has not been studied closely, but it is known that a transcript extends at least from within S2, through ATPase *a* and ATPase III (Cozens *et al.*, 1986). Moreover, the lack of extensive non-coding sequences around the S2 gene makes it somewhat unlikely that major promoters will lie there, and therefore it is possible that all of these genes are co-transcribed. Co-transcription of genes with unrelated function is known in *E. coli*. For example, the α-subunit of RNA polymerase is co-transcribed with several ribosomal proteins (Post *et al.*, 1980). It is therefore of some interest to find that the chloroplast β'-subunit of RNA polymerase is associated with ribosome subunit S2, although the *E. coli* homologue is associated with a translational factor, EF-Ts (An *et al.*, 1981).

We are grateful to Dr J. C. Gray for supplying plasmid pPscS6. A.L.C. is supported by an MRC Research Studentship.

REFERENCES

Allison, L. A., Moyle, M., Shales, M. & Ingles, C. J. (1985) Cell **42**, 599–610
An, G., Bendiak, D. S., Mamelak, L. A. & Friesen, J. D. (1981) Nucleic Acids Res. **9**, 4163–4171
Bankier, A. T. & Barrell, B. G. (1983) in Techniques in Nucleic Acid Biochemistry (Flavell, R. A., ed.), B508/1–B508/31, Elsevier Scientific Publishers, Ireland
Biggin, M. D., Gibson, T. J. & Hong, G. F. (1983) Proc. Natl. Acad. Sci. U.S.A. **80**, 3963–3965
Bohnert, H. J., Crouse, E. J. & Schmitt, J. M. (1982) Encycl. Plant Physiol. New Ser. **14B**, 475–530
Bollen, A., Heimark, R. L., Cozzone, A., Traut, R. R., Hershey, J. W. & Kahan, L. (1975) J. Biol. Chem. **250**, 4310–4314
Cozens, A. L., Walker, J. E., Phillips, A. L., Huttly, A. K. & Gray, J. C. (1986) EMBO J. **5**, 217–222
Dorne, A. M., Lescure, A. M. & Mache, R. (1984) Plant Mol. Biol. **3**, 83–90
Dyer, T. A. (1984) in Chloroplast Biogenesis (Baker, N. R. & Barker, J., eds.), pp. 23–69, Elsevier Science Publishers, Amsterdam
Dyer, T. A. (1985) Oxford Surveys Plant Mol. Cell Biol. **2**, 147–177
Eneas-Filho, I., Hartley, M. R. & Mache, R. (1981) Mol. Gen. Genet. **184**, 484–488
Engleman, D. M., Moore, P. B. & Schoenborn, B. P. (1975) Proc. Natl. Acad. Sci. U.S.A. **72**, 3888–3892
Georgalis, Y., Giri, L. & Littlechild, J. A. (1981) Biochemistry **20**, 1061–1064
Graf, L., Roux, E. & Stutz, E. (1982) Nucleic Acids Res. **10**, 6369–6381
Gray, J. C., Phillips, A. L., Howe, C. J., Willey, D. L., Huttly, A. K., Doherty, A., Bowman, C. M. & Dyer, T. A. (1984) in Biosynthesis of the Photosynthetic Apparatus (Thorber, J. P. *et al.*, eds.), pp. 295–307, A. R. Liss, New York
Greenberg, B. M., Narita, J. O., De Luca-Flaherty, C., Graissem, W., Rushlow, K. A. & Hallick, R. B. (1984) J. Biol. Chem. **259**, 14880–14887
Herrman, R. G. & Possingham, I. V. (1980) in Chloroplasts (Reinert, J., ed.), pp. 45–96, Springer-Verlag, Berlin
Huttly, A. K. & Gray, J. C. (1984) Mol. Gen. Genet. **194**, 402–409
Kidd, G. H. & Bogorad, L. (1980) Biochim. Biophys. Acta **609**, 14–30
Lake, J. A. (1978) in Advanced Techniques in Biological Electron Microscopy II (Koehler, J., ed.), pp. 173–211, Springer-Verlag, Berlin, Heidelberg, New York
Lerbs, S., Briat, J.-F. & Mache, R. (1983) Plant Mol. Biol. **2**, 67–74
Lipman, D. J. & Pearson, W. R. (1985) Science **227**, 1435–1441
Littlechild, J. A., Dijk, J. & Garrett, R. A. (1977) FEBS Lett. **74**, 292–300
Montadon, P. E. & Stutz, E. (1984) Nucleic Acids Res. **12**, 2851–2859
Narita, J. O., Rushlow, K. E. & Hallick, R. B. (1985) J. Biol. Chem. **260**, 11194–11199
Okuyama, A., Yoshikawa, M. & Tanaka, N. (1974) Biochem. Biophys. Res. Commun. **60**, 1163–1169
Ovchinnikov, Yu, A., Monastyrskaya, G. S., Gubanov, V. V., Guryev, S. O., Salomatina, I. S., Shuvaeva, T. M., Lipkin, V. M. & Sverdlov, E. D. (1982) Nucleic Acids Res. **10**, 4035–4044
Passavant, C. W., Stiegler, G. L. & Hallick, R. B. (1983) J. Biol. Chem. **258**, 693–695
Phillips, A. L. & Gray, J. C. (1984) Mol. Gen. Genet. **194**, 477–484
Posno, M., van Noort, M., Debise, R. & Groot, S. P. (1984) Curr. Genet. **8**, 147–154
Post, L. E., Arfsten, A. E., Davis, G. R. & Nomura, M. (1980) J. Biol. Chem. **255**, 4653–4659
Sanger, F., Nicklen, S. & Coulson, A. R. (1977) Proc. Natl. Acad. Sci. U.S.A. **74**, 5463–5467
Schmidt, R. J., Richardson, C. B., Gillham, N. W. & Boynton, J. W. (1983) J. Cell Biol. **96**, 1451–1463
Schwartz, Z. & Kössel, H. (1980) Nature (London) **283**, 739–742
Shine, J. & Dalgarno, L. (1974) Proc. Natl. Acad. Sci. U.S.A. **71**, 1342–1346
Squires, C., Krainer, A., Barry, G., Shen, W. F. & Squires, C. L. (1981) Nucleic Acids Res. **9**, 6827–6840
Staden, R. (1982*a*) Nucleic Acids Res. **10**, 4731–4751
Staden, R. (1982*b*) Nucleic Acids Res. **10**, 2951–2961
Staden, R. (1984) Nucleic Acids Res. **12**, 521–538
Staden, R. (1985) Genet. Eng. **7**, 67–114
Subramanian, A. R., Steinmetz, A. & Bogorad, L. (1983) Nucleic Acids Res. **11**, 5277–5286
Sugita, M. & Sugiura, M. (1983) Nucleic Acids Res. **11**, 1913–1918
Tewari, K. K. & Goel, A. (1983) Biochemistry **22**, 2142–2148
Tischendorf, G. W., Zeichardt, H. & Stöffler, G. (1975) Proc. Natl. Acad. Sci. U.S.A. **72**, 4820–4824
Tohdoh, N. & Sugiura, M. (1982) Gene **17**, 213–218
Traut, R. R., Lambert, J. M., Boileau, G. & Kenny, J. W. (1980) in Ribosomes (Chambliss, G. *et al.*, eds.), pp. 89–110, University Park Press, Baltimore
Walker, J. E. & Tybulewicz, V. L. J. (1985) in The Molecular Biology of the Photosynthetic Apparatus (Arntzen, C., Bogorad, L., Bonitz, S. & Steinback, K., eds.), pp. 141–153, Cold Spring Harbor
Whitfeld, P. R. & Bottomley, W. (1983) Annu. Rev. Plant Physiol. **34**, 279–310
Wittmann-Liebold, B. & Bosserhoff, A. (1981) FEBS Lett. **129**, 10–16
Yoshikawa, M., Okuyama, A. & Tanaka, N. (1975) J. Bacteriol. **122**, 796–797
Zurawaski, G., Bottomley, W. & Whitfeld, P. R. (1984) Nucleic Acids Res. **12**, 6547–6558

Received 9 December 1985/13 January 1986; accepted 22 January 1986