Mol Gen Genet (1984) 194:432–443

MGG

Sequence and evolution of the regions between the *rrn* operons in the chloroplast genome of *Euglena gracilis bacillaris*

M. Raafat El-Gewely[1], Robert B. Helling[1,2], and Joost G. Th. Dibbits[1]

[1] Division of Biological Sciences, University of Michigan, Ann Arbor, Michigan 48109, USA
[2] Institut de Biologie Moléculaire et Cellulaire, 15 rue Descartes, F-67084 Strasbourg, France

Summary. The rRNA genes are arranged in three sequential operons preceded by a fourth partial operon. Part or all of a 1462 nucleotide sequence extending from within the 3′-end of the 23S rRNA gene, across the 5S rRNA gene and a presumptive transcription terminator, to within the first structural gene (for 16S rRNA) of the *rrn* operon was determined for each region between operons. Homologies of the 3′-end of the 23S rRNA gene with the 4.5S rRNA genes of higher plant chloroplasts, and of the 5S rRNA gene with other 5S rRNA genes were examined. The region preceding the 16S rRNA gene, which is expected to contain sites for initiation and regulation of *rrn* transcription, includes a 305 base-pair sequence with substantial homology with structural genes elsewhere in the chloroplast genome. The homologies suggest that this portion of the leader evolved from copies of parts of the structural genes which had been inserted before the 16S rRNA genes. Thus the chloroplast *rrn* leader may provide a unique opportunity to study how a regulatory sequence evolved from well-defined structural genes.

Introduction

The ribosomal RNA (rRNA) genes are of considerable importance for at least two reasons. First, they are central to metabolism and are among the most important genes in any organism. Except at slow growth rates the rate of synthesis of rRNA in *Escherichia coli* is closely coupled to ribosome formation (Gausing 1977). Under such conditions ribosomes are believed to operate at maximal efficiency. The rates of rRNA and ribosome production increase exponentially with growth rate, rather than linearly like most mRNAs, in accord with the central importance of protein synthesis in adjusting to the environment (Maaløe and Kjeldgaard 1966; Ingraham et al. 1983). The importance of the rRNA genes in higher organisms is exemplified by the large number of copies usually present and their selective amplification in certain active tissues. A second reason that the rRNA genes are of importance is that they and the tRNA genes are the most ancient we can recognize, reflecting their essential role in all organisms. Information about these genes may help us to understand some of the earliest steps in the evolution of life.

The chloroplast DNA of *Euglena gracilis* var. *bacillaris* (or B strain) includes three complete sets of rRNA genes (Helling et al. 1979; El-Gewely et al. 1981) preceded by a fourth partial rRNA gene set (Koller, Delius and Helling, unpublished; Fig. 1). Each gene-set (*rrn* operon) encodes (in order of transcription) 16S rRNA, $tRNA^{Ile}$, $tRNA^{Ala}$, 23S rRNA and 5S rRNA (El-Gewely et al. 1981; Hallick 1983). The repeated sets differ in restriction sites and length as the result of differences in the regions separating the operons (Helling et al. 1979; El-Gewely et al. 1981). These regions between successive operons must contain the sites where transcription is initiated and terminated. The same regions have been shown to hybridize with ^{125}I-labelled chloroplast 4S RNA (El-Gewely et al. 1981), with chloroplast RNA labelled using nucleotidyl transferase (using strain Z; Orozco et al. 1980), and with purified tRNAs (both strains B and Z; Keller et al. 1982; Kuntz et al. 1983; Kuntz and Helling, unpublished) suggesting that one or more tRNA genes might be present. We should like to understand the regulation and evolution of the *rrn* operons and so have sequenced these regions between operons.

Materials and methods

Strains and DNA. Sequences were determined from cloned chloroplast DNA segments *Bam*E, *Bam*F, *Eco*M and *Eco*R in plasmids pRBHO26, pRBHO22, pMIL19 and pMIL12 respectively (El-Gewely et al. 1981), present in *E. coli* K12. Plasmid DNA preparations, endonuclease-cleavage, agarose gel electrophoresis and fragment purification were as described (El-Gewely et al. 1982; El-Gewely and Helling 1980). When the sizes of fragments to be purified were too close to be separated cleanly by zonal centrifugation, agarose gel electrophoresis and electroelution into 3MM paper were used (Girvitz et al. 1980).

End-labelling experiments and sequencing reactions. Sequences were determined after labelling with ^{32}P the 3′- or 5′-ends generated by endonuclease *Bam*HI or *Xho*I, or the 5′-ends generated by *Bst*EII or *Eco*RI digestion of the cloned chloroplast DNA (Fig. 1), asymmetrically cleaving the labelled fragment, and separating the resulting labelled pieces by sucrose gradient sedimentation or agarose gel electrophoresis as appropriate. The 3′-labellings of *Bam*HI-generated ends and *Xho*I-generated ends were made using *E. coli* DNA polymerase Klenow fragment at 0° C (Maxam and Gilbert 1980). The 5′-labellings were made using polynucleotide kinase following calf intestine alkaline phospha-

Offprint requests to: R. B. Helling

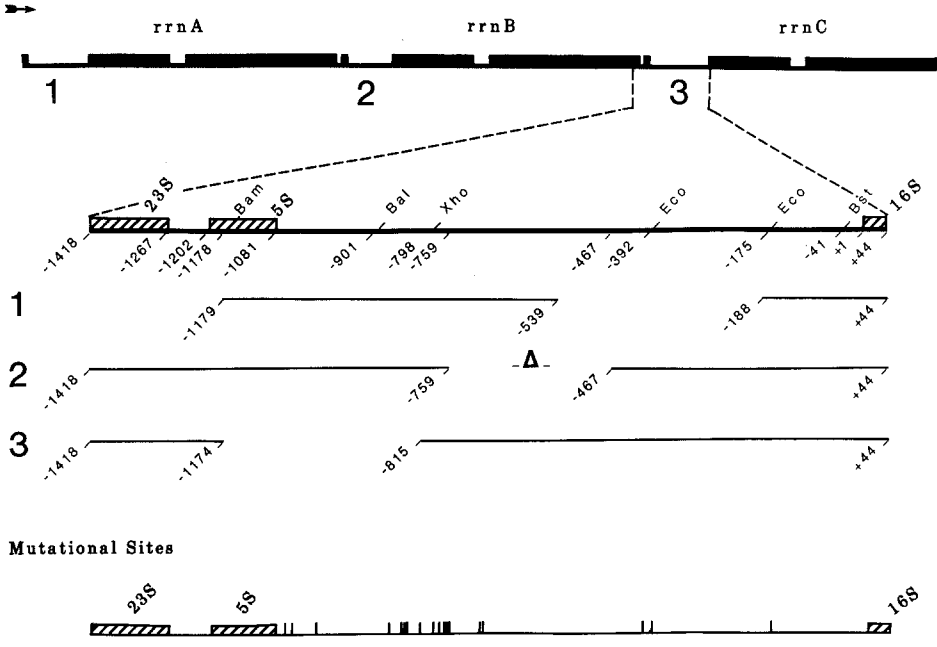

Fig. 1. Organization of the *rrn* operons, regions sequenced and locations of base-pair differences among the repeats. Top: Overall map of the repeated rRNA genes from the 5S rRNA gene of the partial gene-set (left) across the three *rrn* operons. The order of transcription of each operon is from left to right. The sequences of each repeat which were actually determined are shown by the horizontal lines. The locations of the nucleotide differences between repeats (Table 1) are shown below. No *Eco*RI sites are present in region 1, sites at −392 and −175 are present in region 2, and only the site at −392 is present in region 3. The other endonuclease cleavage sites are present in each region

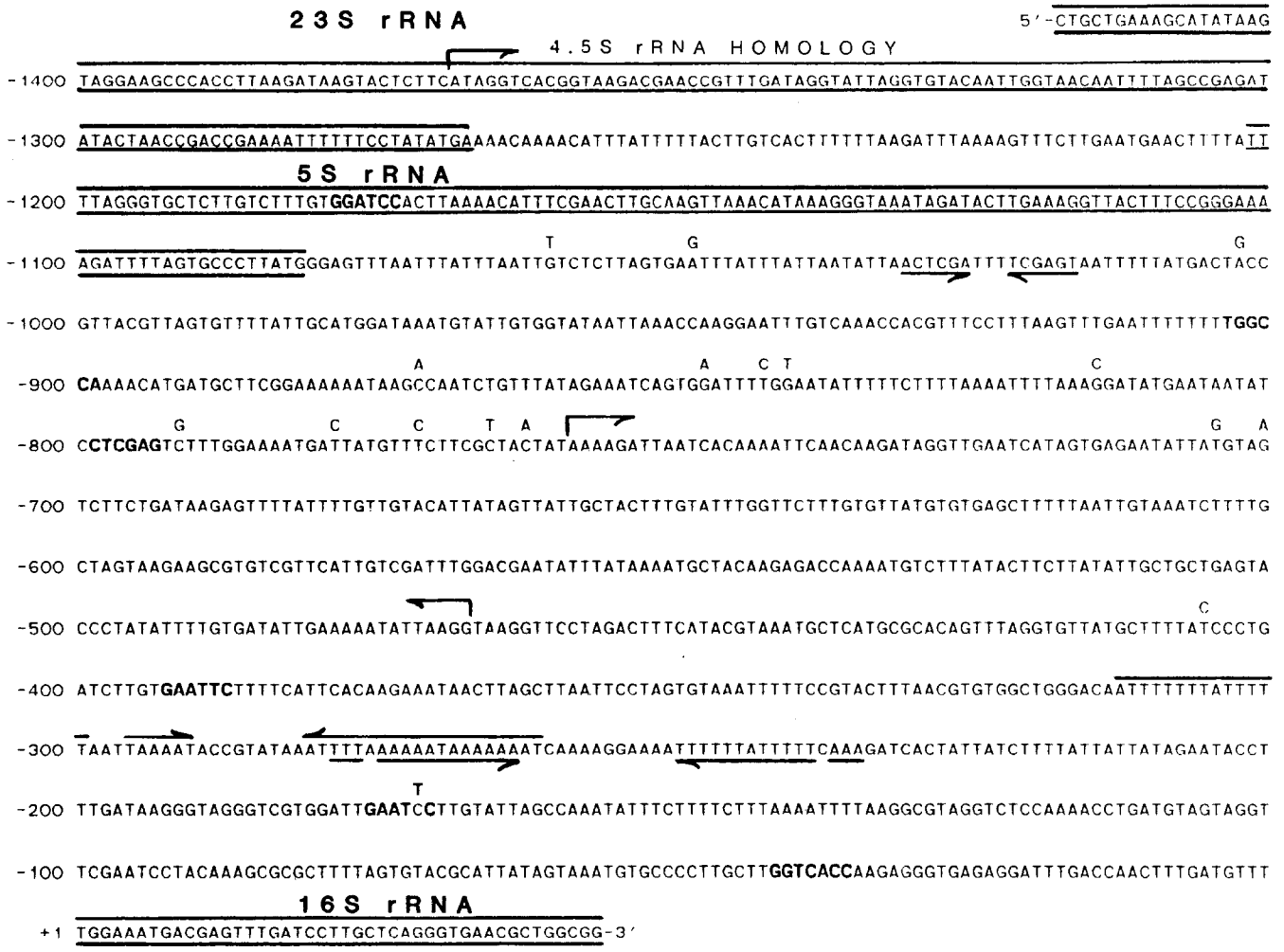

Fig. 2. The sequence of the RNA-like DNA strand of the chloroplast *rrn* operons from the 3′-end of the 23S rRNA gene to the 5′-end of the 16S rRNA gene. The ends of the 5S rRNA and 23S rRNA have not been determined but are predicted to correspond approximately to the positions shown. Single nucleotide differences among the repeated sequences are shown (see Table 1). The sequence deleted from region 2 (Fig. 1) relative to regions 1 and 3 extends from −759 to −467 as denoted by arrows. Arrows at approximately −1025 underline an inverted repeat sequence which may be part of a transcription-terminator. Lines with arrows in the region −313 to −235 designate alternative inverted repeat sequences. Specific endonuclease cleavage sites are shown in bolder type (see Fig. 1)

tase treatment (Maxam and Gilbert 1980), or using the exchange method (Berkner and Folk 1980). Some restriction enzymes are shipped in a high phosphate buffer which inhibits polynucleotide kinase and alkaline phosphatase (Chaconas and DeSande 1980). Thus after cleavage by endonucleases, care was taken to remove inhibitory salts by ethanol precipitation before labelling at the 5′-ends. For the exchange reaction 0.1 mM spermidine was included in the kinase reaction buffer. When 5′-labelled DNA was digested with endonucleases, 30 mM phosphate was added to prevent the removal of the ^{32}P by phosphatase possibly contaminating the commercial restriction enzyme preparations.

Cleavage reactions and gel electrophoresis were carried out according to Maxam and Gilbert (1980) except for the cytidine reaction. For this reaction the procedure of Rubin and Schmid (1980) was followed.

Results

The overall sequence

A 1462 nucleotide sequence extending from −151 in the 23S rRNA gene to +44 in the 16S rRNA gene is presented in Fig. 2. The entire sequence of this region between *rrnA* and *rrnB* (region 2) and most of the regions 1 and 3 sequences were determined, as shown in Fig. 1. The complete 16S rRNA gene sequence from *E. gracilis* strain Z has been published (Graf et al. 1982). The segment of the 16S rRNA shown in Fig. 2 is identical with the corresponding sequence of the Z strain. The sequences coding for all the rRNAs can be recognized by their homology with the corresponding genes from other genomes.

The ends of the 5S and 23S rRNAs have not been determined and so could vary slightly from the positions denoted in Fig. 2 (see later). With this qualification, the two genes are separated by 65 nucleotide pairs, and 1081 nucleotide pairs separate the 5S and 16S rRNA genes. This latter region has a G+C frequency of 29%. The G+C content of total chloroplast DNA is approximately 28% (Schmitt et al. 1981).

Sequence differences among the repeated gene-sets

The region between operons *rrnA* and *rrnB* (region 2, Fig. 1) was known to lack several hundred base-pairs present in the other leader regions (Helling et al. 1979; El-Gewely et al. 1981; Koller, Delius and Helling, in preparation). The complete sequence of this region was determined and found to lack a continuous stretch of 292 base pairs present in the other repeats (Figs. 1, 2). Deletions are believed to occur between short sequence repeats usually (Albertini et al. 1982). However no strong similarity of the sequences at the break points of the deletion in region 2 is apparent (Fig. 2).

Seventeen single base-pair substitutions were found among the regions sequenced and an additional one is inferred because region-1 lacks an *Eco*RI cleavage sequence (Helling et al. 1979) present at −393 to −388 in regions 2 and 3 (Table 1, Fig. 1). The overall frequency of mutant bases among the regions sequenced is less than 0.6%. Ten of the substitutions correspond to transition mutations; the other seven, to transversions (Table 1). One substitution (at −172 in region 2) probably resulted from a transition converting the sequence GAATCC to GAATTC and thereby creating an *Eco*RI cleavage site (Helling et al. 1979). As a result of this second *Eco*RI site (Fig. 1), region 2 includes a 217 nucleotide-pair *Eco*RI segment separating *Eco*S and *Eco*O (Helling et al. 1979; El-Gewely et al. 1981).

Table 1. Nucleotide differences among the repeats

Position	Nucleotide at position in region[a]		
	1	2	3
−1061	G	T	
−1049	A	G	
−1003	A	G	
−872	C	A	
−848	G	A	
−843	T	C	
−841	G	T	
−815	G	C	C
−792	C	G	G
−779	T	C	C
−772	T	C	C
−766	C	T	C
−763	C	A	A
−705	T		G
−701	G		A
−406		T	C
−393 to −388	Not GAATTC[b]	GAATTC	GAATTC
−172	C	T	C

[a] Interoperon regions as designated in Fig. 1

[b] Regions 2 and 3 have an *Eco*RI cleavage site (GAATTC) at −393 to −388. Region 1 has not been sequenced at this position but is not cleaved by *Eco*RI endonuclease

The 3′-end of the 23S rRNA gene

The 5′- and 3′-ends of rRNA genes are generally complementary, and it is believed that the corresponding regions of the transcripts pair. The paired structures are probably required for processing the transcript to form the mature RNAs (Gegenheimer and Apirion 1981). Therefore we looked for the 3′-end of the 23S rRNA gene in the overall sequence by looking for a segment complementary to the 5′-end (Graf et al. 1980; Orozco et al. 1980). Eleven of 13 nucleotides in the sequence −1276 to −1264 (Fig. 2) are complementary to a sequence across the 5′-end of the gene (12 of 13 if a G-T pair is allowed). The presumptive secondary structure of the transcript is shown in Fig. 3. The paired sequences at the ends of the 23S rRNA are very similar to those at the ends of the $tRNA^{Ala}$ so the same processing activity may cleave free both $tRNA^{Ala}$ and 23S rRNA as shown. The extensive pairing of the region between the 16S and 23S rRNAs together with the highly folded structures of mature RNAs suggests that very little of the transcript remains as a single-stranded random coil.

Higher plant chloroplasts contain a 4.5S rRNA whose corresponding DNA is separated by a spacer segment of about 100 nucleotide pairs from the 23S rRNA gene (Takaiwa and Sugiura 1982). The 4.5S rRNA is homologous to the 3′-end of prokaryotic 23S rRNA. (Prokaryotes lack the 4.5S rRNA). The 5′-end of the 23S rRNA gene of the chloroplasts is complementary to the 3′-end of the 4.5S rRNA gene rather than to the 3′-end of the 23S rRNA gene Takaiwa and Sugiura 1982). Therefore the 4.5S RNA

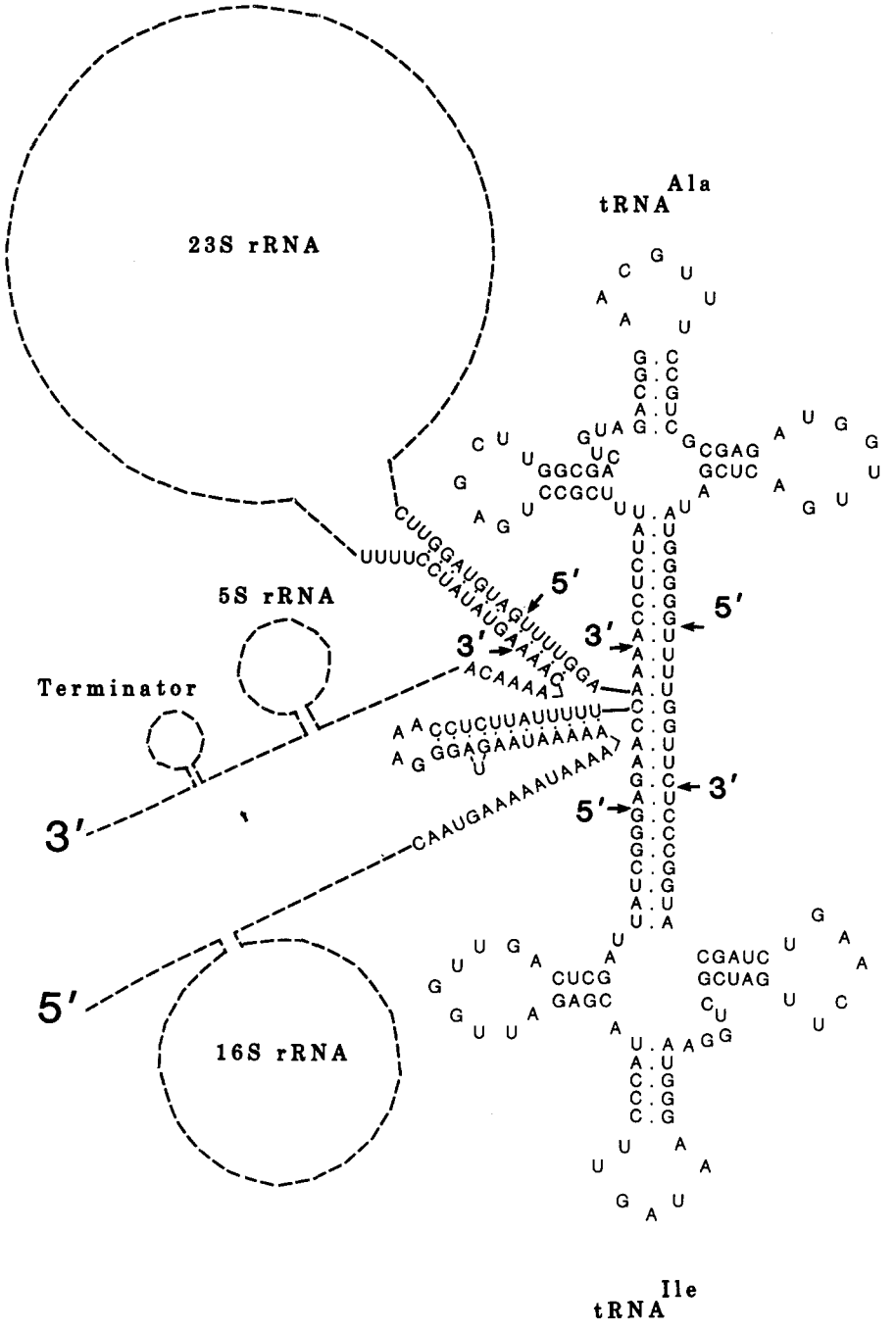

Fig. 3. Possible secondary structure of part of the chloroplast *rrn* transcript as predicted from the corresponding gene sequence. The sequences are depicted without base modifications (after Graf et al. 1980). The sequence between the 16S rRNA genes was determined in strain Z (Graf et al. 1980; Orozco et al. 1980) and the region across the 3′-end of the 23S rRNA gene was sequenced in strain B (this paper). The ends of the 23S rRNA have not been determined but are predicted to correspond approximately to the positions shown

gene is believed to have evolved from the 3′-end of an ancestral 23S rRNA gene and to be the structural and functional equivalent of the 3′-end of large subunit rRNAs.

Direct sequence comparison of the 3′-end of the 23S rRNA gene of *E. gracilis* with the corresponding sequences of *E. coli* and higher plant chloroplasts reveals that −1369 to −1268 is 50% homologous with the corresponding rRNA sequence of *E. coli* and about 56% homologous with the 4.5S rRNA sequences of higher plant chloroplasts (Fig. 4). The *E. gracilis* sequence to the 5′-side of −1369 is homologous to the 3′-end of the higher plant chloroplast 23S genes but not to the spacer sequence between the 23S and 4.5S rRNA genes. The 4.5S rRNAs from chloroplasts of higher plants (except wheat and maize) contain a nine base-pair insertion relative to other equivalent RNAs (Fig. 4). *E. gracilis* 23S rRNA lacks eight of the nine base-pairs. It is not possible to determine from these data whether an ancestor of *Euglena* chloroplasts had the full insertion. The *E. gracilis* sequence can be folded into a possible secondary structure (Fig. 5) very similar to the model structure proposed for 4.5S rRNA and for the 3′-end of 23S rRNA by Matchatt et al. (1981).

The 5S rRNA gene

Each of the four 5S rRNA genes contains a *Bam*HI cleavage site (Gray and Hallick 1979; El-Gewely et al. 1981), allowing the rapid location and identification of the sequence by comparison with published 5S rRNA sequences. The complete sequence of the 5S rRNA gene of *rrnA*, and partial

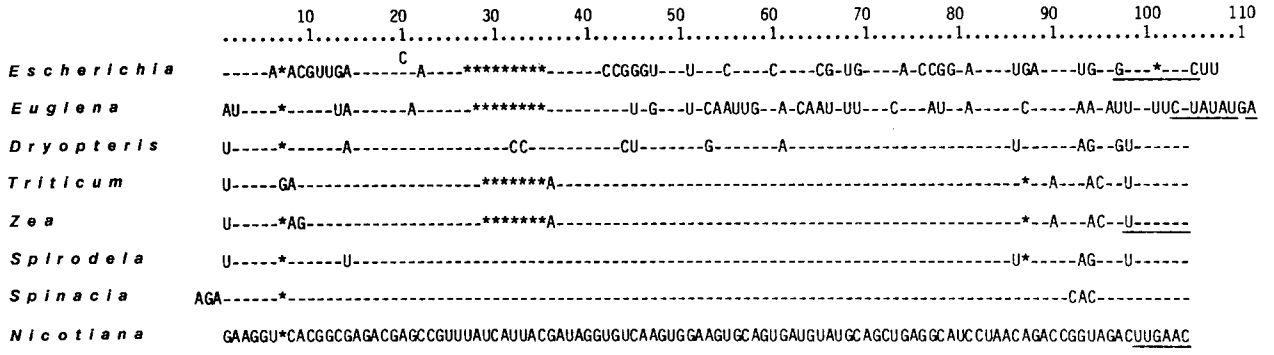

Fig. 4. Homology of the 3′-end of the *E. gracilis* 23S rRNA with 4.5S rRNAs from higher plant chloroplasts. All sequences are compared with the 4.5S rRNA sequence of *Nicotiana* (Takaiwa and Sugiura 1982). Bars indicate identical nucleotides and asterisks, deletions. Underlined bases are complementary to the 5′-end of the large rRNA. Other references are *Escherichia* (Brosius et al. 1980), *Dryopteris* (Takaiwa et al. 1982), *Triticum* (Wildeman and Nazar 1980), *Zea* (Edwards et al. 1981), *Spirodela* (Keus et al. 1983), and *Spinacia* (Kumagai et al. 1982)

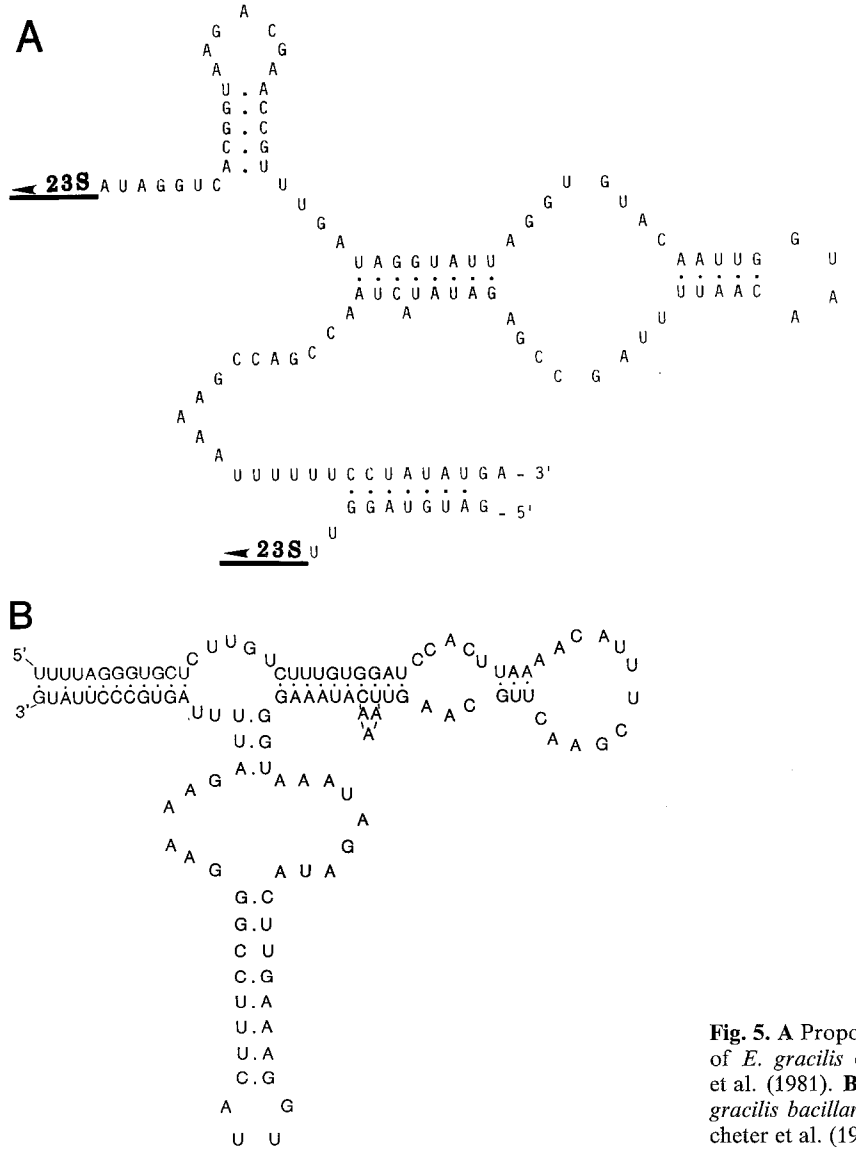

Fig. 5. A Proposed secondary structure of the 3′-end of 23S rRNA of *E. gracilis* chloroplasts according to the model of Matchatt et al. (1981). **B** Proposed secondary structure of 5S rRNA of *E. gracilis bacillaris* chloroplasts according to the model of De Wacheter et al. (1982)

sequences of the genes of *rrnB* and of the fourth (partial) gene-set were determined (Fig. 1). No difference among the homologous sequences was observed. The deduced 5S rRNA from this DNA segment (Fig. 2) has the marker sequences conserved in most 5S rRNAs, namely YRC_{35}, CRV_{48}, $AARC_{74}$, RRV_{95}, RAR_{121} and YGY_{140}, when appropriate position alignment is used (Fig. 6). The corresponding secondary structure of *E. gracilis* chloroplastic

```
                                ....1....1....1....1....1....1....1....1....1....1....1....1.
                                          10        20        30        40        50
                                Conserved Squences:                   YRC         CRU

1-  E. gracilis bacillaris (chloroplast)        .UUUUAGGGUGCUC.UUGUC..UUUGUGGAUCCACUUAAAA....CAUUUCGAAC.

2-  Dryopteris acuminata (chloroplast;fern)     .UAUUCUGGUGUCC.CAGGC..GUAGAGGAACCACACCGAUC...CAUCUCGAAC.

3-  Spinacia oleracea (chloroplast;spinach)     .UAUUCUGGUGUCC.UAGGC..GUAGAGGAACCACACCAAUC...CAUCCCGAAC.

4-  Anacystis nidulans (cyanobacterium)         ...UCCUGGUGUCU.AUGGC..GGUAUGGAACCACUCUGACC..CCAUCCCGAAC.

5-  Escherichia coli (eubacterium)             ..UGCCUGGCGGCC.GUAGC..GCGGUGGUCCCAC.CUGACC..CCAUGCCGAAC.

6-  Triticum aestivum (mitochondrion)          .AAACCGGGC.ACU.ACGGUGAGACGUGAAAACAC.CCGAUC..CCAUUCCGACC.

7-  E. gracilis Z (cytosol)                    ..GGCGU.ACGGCC.AUACU..ACCGGGAAUACAC.CUGAAC..CCGU.UCGAUU.

8-  Spinacia oleracea (cytosol)                ...GGGU.GCGAUC.AUACC..AGCACUAAUGCAC.CGGAUC..CCAU.CAGAAC.

9-  Homo sapiens (cytosol;human)               ...GUCU.ACGGCC.AUACC..ACCCUGAACGCGC.CCGAUC..UCGU.CUGAUC.


      ...1....1....1....1....1....1....1....1....1....1....1....1....1....1....1....1....1....1....1
         60        70        80        90       100       110       120       130       140       150
                    AARC                RRU                 RAR                 YGY

1-  UUGC.....AAGUUAAAC.AUAAAG.GGU..AAAUAGAU.ACUUGAA.AGGUUACUUUCCGGGAAAAG....AUUU..UAGUGCCCUUAUG...

2-  UUGG.....UGGUGAAACUCUGCCGCGGU..AAC.CAAU.ACUCGG..GGGGGGCCCUGCGGAAAAAU....AGCU..CGAUGCCAGGAUA...

3-  UUGG.....UGGUUAAACUCUACUGCGGU..GAC..GAU.ACUGUAG.GGGAGGUCCUGCGGAAAAAU....AGCU..CGACGCCAGGAUG...

4-  UCAG.....UUGUGAAACAUACCUGCGGC..AAC..GAU.AGCUCCC.GGGUAGCCGGUCGCUAAAAU....AGCU..CGACGCCAGGUC....

5-  UCAG.....AAGUGAAACGCCGUAGCGCC..GAU..GGU.AGUGUGG.GGUCU.CCCCAUGCGAGAGU....AGGG..AACUGCCAGGCAU...

6-  UCGAUAUAUAUGUGGAAUCGUCUUGCGCCA.UAU...GU.ACUG....AAAUUGUU...CGGGAGACA....UGGU..CAAAGCCCGGAAA...

7-  UCAG.....AAGUUAAGCCUGGUCAGGCCC.AGUU.AGU.ACUGAGGUGGGCGACCACUUGGGAACACU...GGGU...GCUGU.ACGCUU...

8-  UCCG.....CAGUUAAGCGUGCUUGGGCGA.GAGU.AGU.ACUAGGAUGGGUGACCUCCUGGGAAGUCC...UCGU...GUUGC.ACCCCU...

9-  UCGG.....AAGCUAAGCAGGGUCGGGCCU.GGUU.AGU.ACUUGGAUGGGAGACCGCCUGGGAAUACC...GGGU...GCUGU.AGGCUU...
```

Fig. 6. The proposed primary sequence of 5S rRNA of *E. gracilis bacillaris* and the sequences of eight other 5S rRNA species aligned as described in Erdmann et al. (1983). In the conserved marker sequences (shown above), Y designates pyrimidine, and R, purine

Table 2. A complete homology matrix involving comparisons of the 5S rRNA sequences from *E. gracilis* and eight other species. The numbers indicate the number of identical nucleotides between the compared species using the alignment shown in Fig. 6. The sequences of the other eight species are compiled in Erdmann et al. (1983)

	1	2	3	4	5	6	7	8	9
1. *E. gracilis bacillaris* (chloroplast)	(121)	70	72	59	54	52	48	53	50
2. *Dryopteris acuminata* (chloroplast; fern)	70	(122)	107	81	66	61	56	53	59
3. *Spinacia oleracea* (chloroplast; spinach)	72	107	(122)	81	66	62	50	53	56
4. *Anacystis nidulans* (blue-green alga)	59	81	81	(120)	71	62	54	56	53
5. *Escherichia coli*	54	66	66	71	(120)	63	70	62	68
6. *Triticum aestivum* (mitochondrion)	52	61	62	62	63	(122)	51	51	51
7. *Euglena gracilis Z* (cytosol)	48	56	50	54	70	51	(121)	78	90
8. *Spinacia oleracea* (cytosol; spinach)	53	53	53	56	62	51	78	(120)	81
9. *Homo sapiens* (cytosol; human)	50	59	56	53	69	51	90	81	(120)

5S rRNA is shown in Fig. 5. Although the exact end points of the 5S rRNA were not determined, this would not alter the major conclusions presented here.

A large number of 5S rRNA sequences have been determined (Erdmann et al. 1983). In Fig. 6 we have compared the 5S rRNA from *E. gracilis* chloroplasts with eight other 5S rRNAs of diverse origins. The *Euglena* chloroplastic 5S rRNA appears to be more similar to the 5S rRNAs from plant chloroplasts than to the other rRNAs listed. Direct homology comparisons between the 5S rRNAs from *E. gracilis bacillaris* and the above eight species confirm this (Table 2). Several important points can be deduced from these comparisons.

1) The least homology among all comparisons is that between the *Euglena* 5S rRNAs from the chloroplast and from the cytosol. This finding is consistent with the endosymbiotic hypothesis for the origin of chloroplasts.

2) The 5S rRNA of cyanobacteria represented by *A. nidulans* is closer to chloroplast 5S rRNA in general than *E. coli* 5S rRNA is to chloroplast 5S rRNA. The opposite is true for cytosolic 5S rRNAs.

3) *Euglena* chloroplast 5S rRNA shows the strongest homology to 5S rRNAs of other chloroplasts. However the other chloroplast 5S rRNAs are much closer to each other than to the *Euglena* 5S rRNA (see also Fig. 7).

4) The highest homology to the human 5S rRNA is

```
1-Spinacia oleracea      UAUUCUGGUGUCCUAGGCGUAGAGGAACCACACCAAUCCAUCCCGAACUUGGUGGUUAAACUCUACUGCGGUGACGAUA
2-Lemna minor            ----------------------------------------------------------------------------------
3-Phaseolus vulgaris     ----------------------------------------------------------------d-----------------
4-Nicotiana tabacum      ----------------------------------------------------------------------------------
                                    C
5-Vicia faba             ----------CU------------------A------------------------------A-----------------
6-Dryopteris acuminata   ----------C-------------------G-------U-------------G-------G-C-----A--C-AU
                                                                                                    UA
7-E.gracilis bacillaris  -U--AG----CU--U-U-U-U-U---U----UUA--Ad---UU--------CAA-------dA--AA-d---A-A----
```

```
1-   CUGUAGGGGAGGUCCUGCGGAAAAAUAGCUCGACGCCAGGAUG
2-   ----------------------------------------A--
3-   --------------------------------G------A--
4-   ----------------------------------------
5-   --------------------------------G------A--
6-   ACUCG----G--C-------------------U--------A
7-   --UG-AA--UUACUU-C---G----G-UU-UAGU---CUU---
```

Fig. 7. Sequence comparison of chloroplast 5S rRNAs of higher plants and *E. gracilis bacillaris*. All sequences are compared with the sequence of 5S rRNA of *Spinacia* chloroplasts. Bars denote identical nucleotides and the letter "d" represents single nucleotide deletions. The region corresponding to the Bam sequence in *E. gracilis* DNA is underlined. The *Euglena* sequence has a UA after A_{90} (Fig. 6) which is not present in the other chloroplast 5S rRNA genes. Higher plant sequences are from Erdmann et al. (1983)

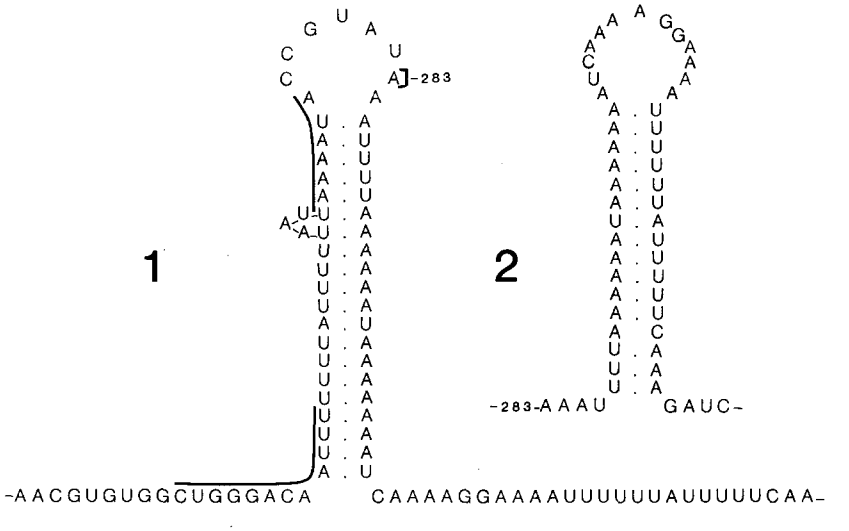

Fig. 8. Stem-loop formation expected in a transcript of the *rrn* leader. The stem-loop 1 is calculated to be stable with respect to the unpaired single strand by 14.5 kcal mol^{-1}, and the alternative stem loop 2, by 7.8 kcal mol^{-1}, using the values compiled by Salser (1977). The lined portions of stem-loop 1 correspond to the −35 and −10 regions of a consensus prokaryotic promoter and the approximate position of the first nucleotide in a resulting transcript

that of the *Euglena* cytosol, suggesting a closer homology of the nuclear genes of *Euglena* to animals than to plants (Delihas et al. 1981). Similar conclusions were reached from $tRNA^{Phe}$ sequence comparisons (Chang et al. 1981).

Possible transcription signals

Examination of the segment following the 5S rRNA gene reveals a sequence similar to that of a transcription terminator in prokaryotes. A transcript of this region is expected to form a stem-loop structure (stem-segments underlined beginning at −1031, Fig. 2) followed by a uridylate-rich sequence, typical of rho-independent prokaryotic transcription terminators (Holmes et al. 1983). Therefore we expect the 3′-end of the *rrn* transcripts to be at about −1015 to −1010.

Two sets of inverted repeat segments of length greater than ten base-pairs are found between −313 and −235 (Fig. 2). A transcript across either set would be expected to form a stem-loop structure through self-pairing (Fig. 8). Except for a single "C" in structure 2, both stems contain the bases "A" and "U" only. The two stem-loops are alternative structures because each incorporates a common segment (Fig. 2), so both could not be present in a single transcript. A sequence similar to a prokaryotic promoter (Pribnow 1979; Hawley and McClure 1983) is present in this region (−10 box from −297 to −290; Fig. 8). Transcripts initiated from this possible promoter (i.e., transcripts beginning at about −283) would lack the sequence from −313 to −291 necessary for forming stem 1 and so would have the potential for forming the second stem loop only (Figs. 2, 8).

Sequence homologies

A short portion of the *rrn* leader sequence of *E. gracilis* strain Z has been reported, and homology of that segment with the sequence extending from the 3′-end of the 16S

Table 3. Pairwise sequence comparisons among homologous regions

Sequences compared	Number of nucleotides	K (Difference per site)
E. gracilis leader / *E. gracilis* structural regions		
Homology 1	172	0.31
Homology 2	133	0.25
Total structural	195	0.27
Total spacers	110	0.31
Overall total	305	0.28
E. gracilis leader / *E. coli* structural genes	195	0.42
E. gracilis structural genes / *E. coli* structural genes	198	0.23
E. gracilis leader / *E. gracilis* ancestral	195	0.24
E. gracilis structural genes / E. gracilis ancestral	195	0.04
E. gracilis ancestral / *E. coli* structural genes	195	0.18

Comparisons were made as described (Mijata et al. 1982), without correction for multiple mutations at single sites. Segments Homology 1 and Homology 2 are defined in the legend to Fig. 9. The last three K values were calculated assuming that the *Euglena* structural and leader sequences diverged from a common ancestral sequence subsequent to the divergence of the lines giving rise to *Euglena* and *E. coli*. *E. gracilis* ancestral refers to the closest common ancestor of the leader and structural sequences

rRNA gene across the $tRNA^{Ile}$ gene was suggested (Orozco et al. 1980; Graf et al. 1980; Mijata et al. 1982). Comparison of the *bacillaris* leader with the latter sequence reveals that the leader segment from −353 to −182 shows 69% identity with a sequence extending from the 3′-terminal 42 nucleotides of the 16S rRNA gene through the first 51 nucleotides of the $tRNA^{Ile}$ gene (Table 3). A "dot plot" shows the homology clearly (Fig. 9A). Discontinuities in the plot at two positions demonstrate that the leader differs by deletions of 3 and 8 base-pairs from the 16S rRNA-spacer segment (arrows 1, 2). Single base-pair differences are also present (Fig. 10).

This region of the leader was known to hybridize with chloroplast $tRNA^{Ile}$ and $tRNA^{Trp}$ (Keller et al. 1980; Kuntz et al. 1982; Kuntz and Helling, unpublished). When we compared the leader with the $tRNA^{Trp}$ gene (Hollingsworth and Hallick 1982), we observed substantial sequence identity (Figs. 9B, 10). The segment −181 to −49 is 75% identical with a sequence extending from within the $tRNA^{Met}$ gene through the $tRNA^{Trp}$ gene into the $tRNA^{Glu}$ gene (Table 3). The location of the junction in the leader separating the regions of two ancestral homologies corresponds to the middles of two putative ancestral tRNA genes (Figs. 9, 10). Comparison of the tRNAs reveals that they share common sequences in the pseudouridine loop, the extraloop, and the anticodon loop (Fig. 11). This suggests that after a copy of one ancestral sequence was incorporated in the leader, a copy of the other sequence was incorporated by recombination at the region of near sequence identity. The region spanning the junction would correspond to a hybrid gene

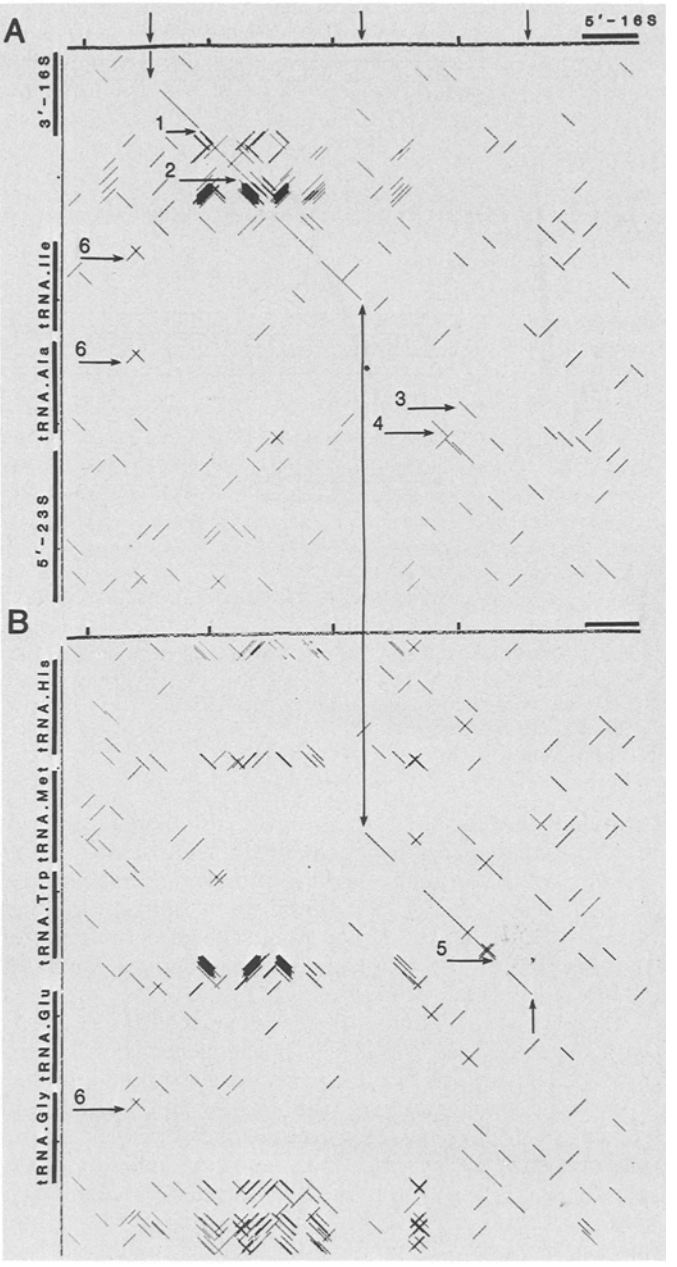

Fig. 9A, B. Homology between the *E. gracilis bacillaris* chloroplast *rrn* leader sequence and other chloroplast DNA sequences. **A** and **B**, double dot-homology matrices with the leader sequence from −418 to +44 positioned across the top. The other sequences (from strain Z) are positioned on the vertical axis. The order of transcription is left to right and top to bottom. Similarity strings are plotted for 10-base pair sequences in which at least 8 base pairs are identical (window size 10, stringency 8), using a computer program designed by J. Adams (unpublished). The original plots distinguished strings of stringencies 8, 9 and 10 by color, and the strings representing homology stood out more obviously from the background. Strings with a negative slope reflect direct comparisons between the sequences. Strings with a positive slope reflect similarity between the sequence on the vertical axis and the opposite strand of the leader (inverted sequence). Vertical arrows reflect limits of homology specified in Fig. 10. **A** Vertical axis includes a 446 nucleotide sequence encoding the 3′-end of 16S rRNA-$tRNA^{Ile}$-$tRNA^{Ala}$-5′-end of 23S rRNA (Graf et al. 1980; Orozco et al. 1980). **B** Vertical axis includes a 490 nucleotide sequence encoding $tRNA^{Met-2}$-$tRNA^{Trp}$-$tRNA^{Glu}$ (Hollingsworth and Hallick 1982). The gene cluster begins with a gene for $tRNA^{Tyr}$ which is not shown. Arrow 5 designates a single base-pair addition to the leader relative to the sequence on the vertical axis .

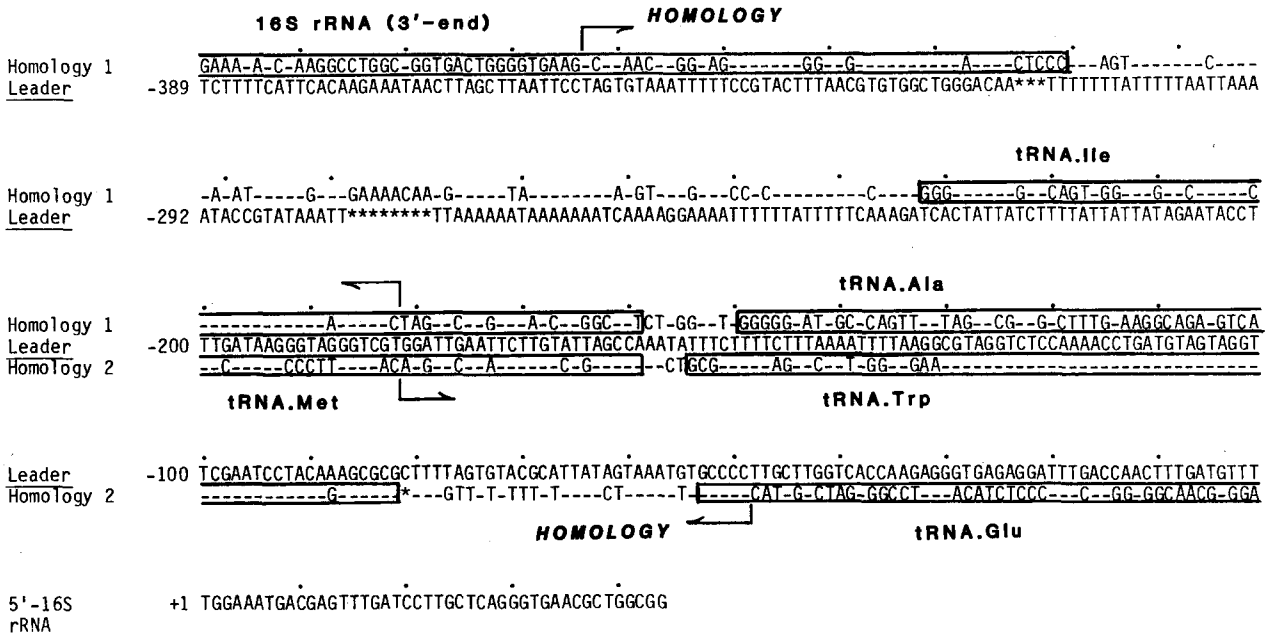

Fig. 10. Direct comparisons of the leader and structural sequences. Homology 1 includes DNA of strain Z encoding the 3′-end of the 16S rRNA-tRNA^{Ile}-tRNA^{Ala} (partial). The 3′-terminal sequence of the 16S rRNA of strain B was determined by Steege et al. (1982). Homology 2 includes DNA of strain Z encoding tRNA $^{Met-2}$ (partial)-tRNA^{Trp}-tRNA^{Glu} (partial). Sequences are of the RNA-like DNA strand and are oriented 5′ to 3′. Dashes indicate a nucleotide identical with that in the leader. Asterisks indicate the nucleotide is not present (deletion)

or pseudogene for a tRNA. However, although a transcript of the leader in this region might fold into a tRNA-like configuration, the amino acid acceptor stem is not expected to pair properly and so a transcript is unlikely to form a real tRNA (Fig. 11 A); nor would the acceptor stem for a transcript of the following segment derived from the tRNA^{Trp} gene pair, properly (Fig. 11 B).

A sequence at about −100 in the leader shows similarity with the 16S-tRNA^{Ile}-tRNA^{Ala}-23S sequence (Fig. 9 A, arrow 3). This reflects homology between the pseudouridine and extraloop regions of the tRNA^{Ala}, tRNA^{Trp} and pseudo-tRNA^{Trp} genes. However the overall sequence comparisons show clearly that the immediate ancestor of the pseudo-tRNA^{Trp} gene is the tRNA^{Trp} gene (82% homology) and not the tRNA^{Ala} gene (50% homology). The strings marked by arrows at 6 (Fig. 9) reflect nearly identical sequences in the leader and in the coding regions for the dihydrouridine stem-loops of tRNA for Ala, Ile and Gly.

Discussion

The extra 16S rRNA gene of strain Z together with its flanking regions was sequenced by Roux et al. (1983). As they noted, the 216-nucleotide sequence preceding the extra gene is 98% identical with our B strain sequence just before the 16S rRNA genes. The region following the extra gene of strain Z has been shown by heteroduplex analysis to contain a segment similar to part of the region between the rRNA gene sets (Koller and Delius 1982). Comparison of the latter sequences reveals 79% homology between a 268-nucleotide stretch beginning 16 nucleotides after the extra 16S rRNA gene and the B strain sequence −1051 to −772 (Fig. 12). The homologous segments are identical for the first 49 nucleotides, a sequence containing the presumptive *rrn* operon transcription-terminator. When considered together with the 98% sequence identity of the extra 16S rRNA gene and the 16S rRNA genes of the complete operons, and the conservation of the 15 nucleotides following the extra gene, the results suggest that the extra gene may be functional. Roux et al. (1983) were unable to find a transcription product from this gene. However *E. gracilis* can grow photosynthetically and non-photosynthetically and it is possible that the extra gene is functional in some growth conditions but not in others, or that a transcript is rapidly broken down.

Our sequencing has revealed more ancient rearrangements as well. Comparison of the leader sequences with the homologous structural regions from strain Z (Figs. 8, 9) shows that the two have diverged approximately 27% (Table 3). By using *E. coli* sequences as a reference, most of this divergence can be seen to have taken place in the leader (0.24 versus 0.04 in the structural genes). Miyata et al. (1982) made a similar conclusion based on knowledge of part of the leader sequence from strain Z. They explained the greater divergence in the leader as the expected result of accumulation of mutations in non-functional DNA free from selective constraints.

The B strain was isolated in North America, probably in New Jersey, by N.A. Coria (possibly this strain was referred to in Am. J. Hygiene 21, 111–120 (1935), L. Provasoli, personal communication; Provasoli et al. 1948). The Z strain (Pringsheim 25) was "axenized" by E.G. Pringsheim from a strain of H. de Saedeleer, and was probably isolated in Germany (Pringsheim and Pringsheim 1952; Hutner et al. 1956). The two strains are expected to have diverged for more than 150×10^6 years, since the two continental plates were last in contact (Bambach et al. 1981). (This is true, of course, only if the standard strains are the descendents of those original isolates. Strain mix-ups have probably occurred at least one (Helling et al. 1979)). Assuming at least one generation per week, this corresponds to about 10^{10} generations for each strain. At a rate of about

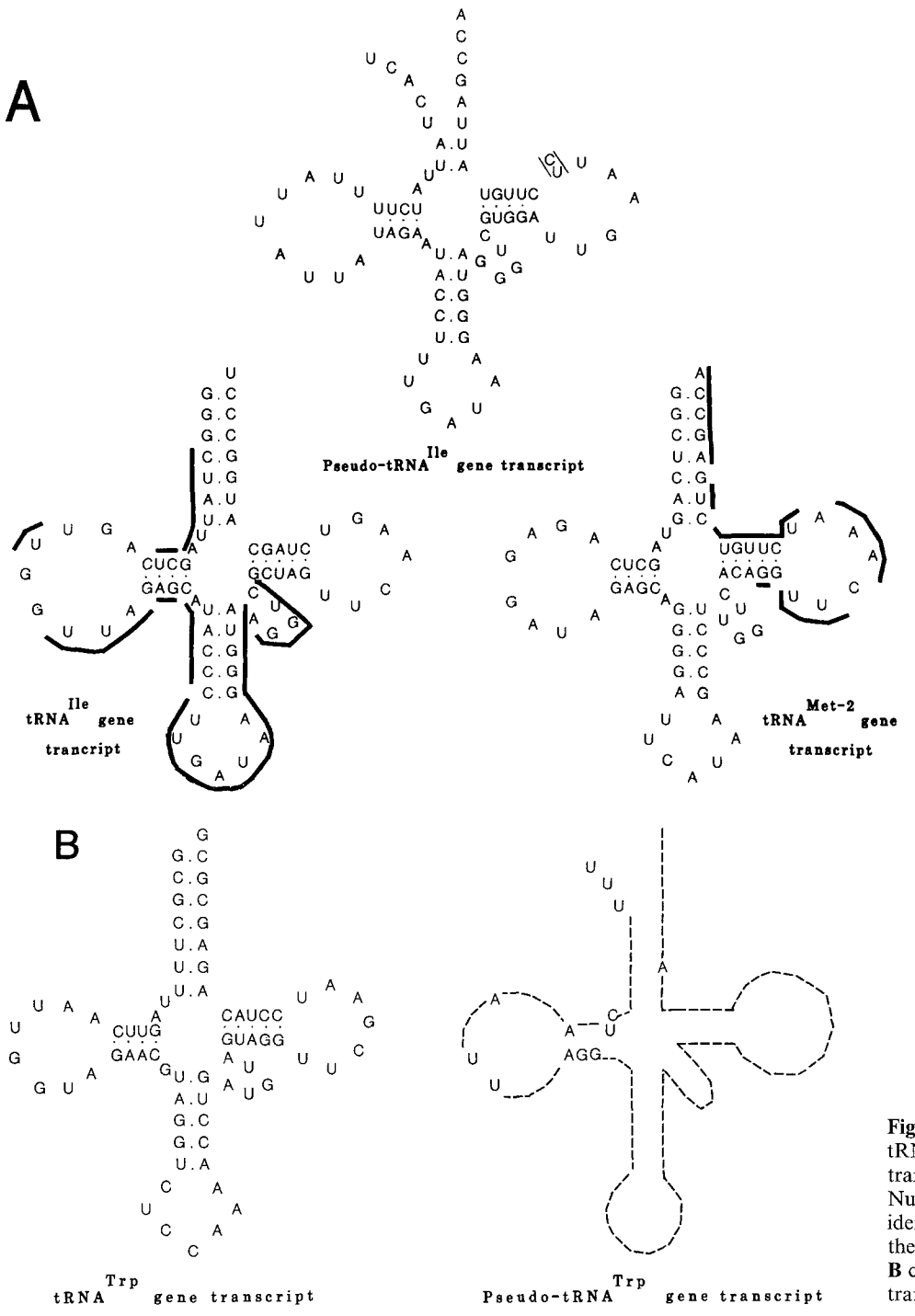

Fig. 11 A. Secondary structures of $tRNA^{Ile}$ and $tRNA^{Met-2}$ and of a transcript of the pseudo-$tRNA^{Ile}$ DNA. Nucleotides in the tRNAs which are identical in the corresponding region of the pseudo-tRNA are lined.
B comparison of $tRNA^{Trp}$ and a transcript of the pseudo-$tRNA^{Trp}$ DNA

10^{-10}–10^{-11} mutations per base-pair per generation (Drake 1969), we would expect a substantial number of differences between the *rrn* leaders of the two strains if the hypothesis of Miyata et al. (1982) were correct. In fact the 305 base-pair segment of the leader considered in Table 3 differs from the equivalent sequence of the Z strain (R. Hallick, personal communication) by only 4 base-pairs (divergence of 0.013). This suggests that most of the evolution in the leader took place prior to separation of the B and Z strains, and that the hypothesis of Miyata et al. (1982) cannot be correct. We present an alternative explanation for the greater divergence of the leader than of the structural genes.

One possible pathway for this evolution is presented in Fig. 13. In an ancestral organism a segment was removed or was copied from the middle of an *rrn* operon and inserted just before the 16S rRNA gene. Presumably the insertion resulted from interaction between short segments of similar sequence in the two DNA segments (Edlund and Normark 1981; Adams 1982). If the insertion was between the gene and its promoter, transcription may have been still possible across the inserted segment. Selection for improved control of transcription would result in the rapid accumulation of nucleotide substitutions. This contrasts with insertion of DNA into a structural gene. In such a case the function of the gene product would generally be destroyed, and res-

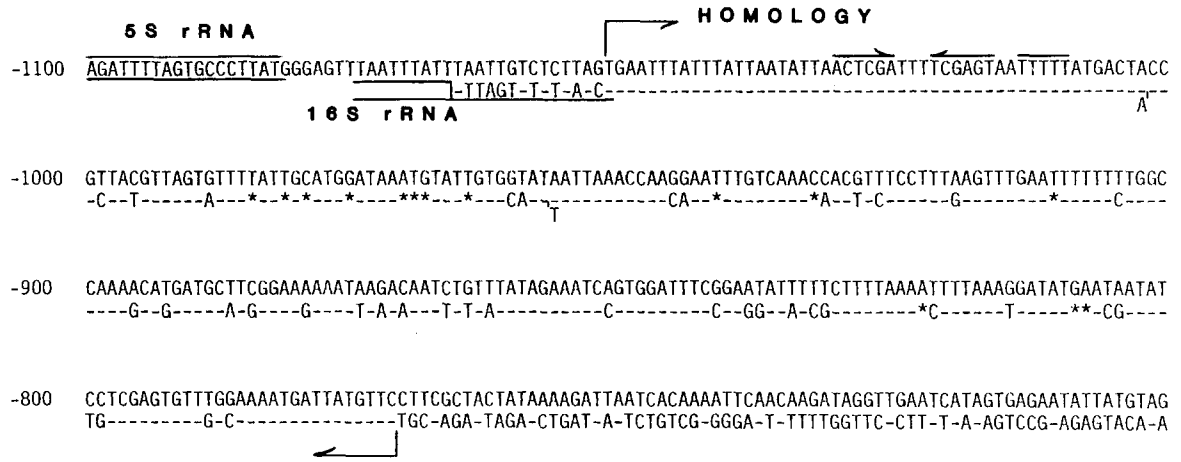

Fig. 12. Homology of the sequence following the extra 16S rRNA gene of strain Z and a sequence between the complete *rrn* operons of the B strain. The continuous sequence is from Fig. 2. The sequence compared with it is from Roux et al. (1983). Bars indicate identical nucleotides; asterisks, deletions. A presumptive transcription terminator sequence is designated by overlining at −1031 to −1010

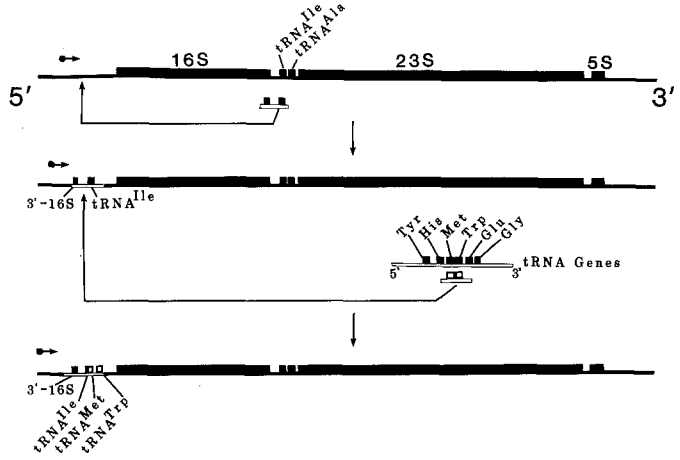

Fig. 13. Possible pathway by which the leader of *E. gracilis* chloroplast *rrn* operons evolved

toration of function could not be achieved by a few simple mutations.

Subsequently a copy of part of the operon containing $tRNA^{Met}$, $tRNA^{Trp}$ and $tRNA^{Glu}$ was inserted in the leader by recombination in the region of $tRNA^{Ile}$-$tRNA^{Met}$ homology. Again selection would be expected to result in accumulation of mutations improving transcription-control across the inserted sequence. (This assumes that at least part of the transcription initiation and/or control originates in the region upstream from −48). A segment of the leader which is likely to have derived from the $tRNA^{Trp}$ gene shows sequence similarity with a segment at the 5′-end of the 23S rRNA gene (Fig. 9A, arrow 4). We suspect that the similar regions of the 23S rRNA and $tRNA^{Trp}$ genes and their derivatives may have been instrumental in the pairing of leader and structural genes at the time of incorporation of the second block of genes into the leader.

During its evolution the *E. gracilis* chloroplast genome has undergone a series of rearrangements in the *rrn* region. We now understand the structural details of some of these rearrangements at the level of the nucleotide sequences involved. Knowledge of the *rrn* leader sequence provides a unique opportunity to study the consequences of the conversion of structural gene sequences with well-known functions to a sequence with a regulatory role.

Acknowledgements. We thank P. Oudet (Strasbourg) for his help with the initial computer analysis, J. Adams for developing the program used for the final analysis and for his interest and help, R. Arking and E. Lau for their contributions, L. Provasoli and J. Schiff for help in ascertaining the origins of the strains. E. Crouse for criticism, E. Stutz and R. Hallick for unpublished results, D. Bay for his photography, and M. Madouse for typing. M.R. El-G. was on leave from Alexandria University, and J.D. was on leave from the Dept. of Genetics, Agricultural University, Wageningen, the Netherlands. RBH was a Fellow of the European Molecular Biology Organization and thanks J. Weil for his hospitality. Supported in part by N.I.H. Grant GM25565.

References

Adams J (1982) Gene duplication and the birthday problem. Nature 296:176–177

Albertini AM, Hofer M, Calos MP, Miller JH (1982) On the formation of spontaneous deletions: the importance of short sequence homologies in the generation of large deletions. Cell 29:319–328

Bambach RK, Scotese CR, Ziegler AM (1981) Before Pangea: the geographies of the paleozoic world. Am Sci 68:26–38

Berkner KL, Folk WR (1980) Polynucleotide kinase exchange as an assay for class II restriction endonucleases. In: Grossman L, Moldave K (eds) Methods in Enzymology 651 pp. 28–36. Academic Press, New York

Brosius J, Dull TJ, Noller HF (1980) Complete nucleotide sequence of a 23S ribosomal RNA gene from *Escherichia coli*. Proc Natl Acad Sci USA 77:201–204

Chang SH, Hecker LI, Brum CK, Schnabel JJ, Hackman JE, Silberklang M, RajBhandary UL, Barnett WE (1981) The nucleotide sequence of *Euglena* cytoplasmic phenylalanine transfer RNA. Evidence for possible classification among the animal rather than the plant kingdom. Nucl Acids Res 9:3199–3204

Delihas N, Anderson J, Andresini W, Kaufman L, Lyman H (1981) The 5S ribosomal RNA of *Euglena gracilis* cytoplasmic ribosomes is closely homologous to the 5S RNA of the trypanosomatid protozoa. Nucl Acids Res 9:6627–6633

DeWachter R, Chen MW, Vandenberghe A (1982) Conservation of secondary structure in 5S ribosomal RNA: A uniform model for eukaryotic, eubacterial, archaebacterial and organelle sequences is energetically favorable. Biochemie 64:311–329

Drake JW (1969) Comparative spontaneous mutation rates. Nature 221:1128–1132

Edlund T, Normark S (1981) Recombination between short DNA homologies causes tandem duplication. Nature 292:269–271

Edwards K, Bedbrook J, Dyer T, Kössel H (1981) 4.5S rRNA from *Zea mays* chloroplasts shows structural homology with the 3′-end of prokaryotic 23S rRNA. Biochem Int 2:533–538

El-Gewely MR, Helling RB (1980) Preparative separation of DNA-ethidium bromide complexes by zonal density gradient centrifugation. Anal Biochem 102:423–428

El-Gewely MR, Lomax MI, Lau ET, Helling RB, Farmerie W, Barnett WE (1981) A map of specific cleavage sites and tRNA genes in the chloroplast genome of *Euglena gracilis bacillaris*. Mol Gen Genet 181:296–305

El-Gewely MR, Helling RB, Farmerie W, Barnett WE (1982) Location of a phenylalanine tRNA gene on the physical map of the *Euglena gracilis* chloroplast genome. Gene 17:337–339

Erdman VA, Huysmans E, Vandenberghe A, DeWachter R (1983) Collection of published 5S and 5.8S ribosomal RNA sequences. Nucl Acids Res 11:r105–r133

Gausing K (1977) Regulation of ribosome production in *Escherichia coli*: synthesis and stability of ribosomal RNA and of ribosomal protein messenger RNA at different growth rates. J Mol Biol 115:335–354

Gegenheimer P, Apirion D (1981) Processing of prokaryotic ribonucleic acid. Microbiol Rev 45:502–541

Girvitz S, Bacchetti S, Rainbow AJ, Graham FL (1980) A rapid and efficient procedure for the purification of DNA from agarose gels. Anal Biochem 106:492–496

Graf L, Kossel H, Stutz E (1980) Sequencing of 16S-23S spacer in a ribosomal RNA operon of *Euglena gracilis* reveals two tRNA genes. Nature 286:908–910

Graf L, Roux E, Stutz E, Kössel H (1982) Nucleotide sequence of a *Euglena gracilis* chloroplast gene coding for the 16S rRNA: homologies to *E. coli* and *Zea mays* chloroplast 16S rRNA. Nucl Acids Res 10:6369–6381

Gray PW, Hallick RB (1979) Isolation of *Euglena gracilis* chloroplast 5S ribosomal RNA and mapping the 5S rRNA gene on chloroplast DNA. Biochemistry 18:1820–1825

Hallick RB (1983) Chloroplast DNA. In: Buetow D (ed) The Biology of *Euglena*, vol 4. Academic Press, New York, in press

Hawley DK, McClure WR (1983) Compilation and analysis of *Escherichia coli* promoter DNA sequences. Nucl Acids Res 11:2237–2255

Helling RB, El-Gewely MR, Lomax MI, Baumgartner JE, Schwarzbach SD, Barnett WE (1979) Organization of the chloroplast ribosomal RNA genes of *Euglena gracilis bacillaris*. Mol Gen Genet 174:1–10

Hollingsworth MJ and Hallick RB (1982) *Euglena gracilis* chloroplast transfer RNA transcription units. Nucleotide sequence analysis of a $tRNA^{Tyr}$-$tRNA^{His}$-$tRNA^{Met}$-$tRNA^{Trp}$-$tRNA^{Glu}$-$tRNA^{Gly}$ gene cluster. J Biol Chem 257:12795–12799

Holmes WM, Platt T, Rosenberg M (1983) Termination of transcription in *E. coli*. Cell 32:1029–1032

Hutner SH, Bach MK, Ross GIM (1956) A sugar-containing basal medium for vitamin B_{12}-assay with *Euglena*; application to body fluids. J Protozool 3:101–112

Ingraham JL, Maaløe O, Neidhardt FC (1983) Growth of the bacterial cell. Sinauer, Sunderland, Massachusetts, USA

Keller M, Burkard G, Bohnert HJ, Mubumbila M, Gordon K, Steinmetz A, Heiser D, Crouse EJ, Weil JH (1980) Transfer RNA genes associated with the 16S and 23S rRNA genes of *Euglena* chloroplast DNA. Biochem Biophys Res Commun 95:47–54

Keus RJA, Roovers DJ, Dekker AF, Groot GSP (1983) The nucleotide sequence of the 4.5S and 5S rRNA genes and flanking regions from *Spirodela oligorhiza* chloroplasts. Nucl Acids Res 11:3405–3410

Koller B, Delius H (1982) Parts of the sequence between the complete rRNA operons are repeated on either side of the extra 16S rRNA gene in chloroplast DNA of *Euglena gracilis* strain Z. FEBS Lett 140:198–202

Kumagai I, Pieler T, Subramanian AR, Erdmann VA (1982) Nucleotide sequence and secondary structure analysis of spinach chloroplast 4.5S RNA. J Biol Chem 257:12924–12928

Kuntz M, Keller M, Crouse EJ, Burkard G, Weil JH (1982) Fractionation and identification of *Euglena gracilis* cytoplasmic and chloroplastic tRNAs and mapping of tRNA genes on chloroplast DNA. Curr Genet 6:63–69

Maaløe O, Kjeldgaard NO (1966) Control of macromolecular synthesis. W.A. Benjamin, New York

Machatt MA, Ebel J-P, Branlant C (1981) The 3′-terminal region of bacterial 23S ribosomal RNA: structure and homology with the 3′-terminal region of eukaryotic 28S rRNA and with chloroplast 4.5S rRNA. Nucl Acids Res 9:1533–1549

Maxam AM, Gilbert W (1980) Sequencing end-labeled DNA with base-specific chemical cleavages. In: Grossman L, Moldave K (eds) Methods in enzymology, vol 651. Academic Press, New York, pp 499–560

Miyata T, Kikuno R, Ohshima Y (1982) A pseudogene cluster in the leader region of the *Euglena* chloroplast 16S–23S rRNA genes. Nucl Acids Res 10:1771–1780

Orozco EM, Rushlow KE, Dodd JR, Hallick RB (1980) *Euglena gracilis* chloroplast ribosomal RNA transcription units. II. Nucleotide sequence homology between the 16S–23S ribosomal RNA spacer and the 16S ribosomal RNA leader regions. J Biol Chem 255:10997–11003

Pribnow D (1979) Genetic control signals in DNA. In: Goldberger RF (ed) Biological regulation and development I, Plenum Press, New York, pp 219–227

Pringsheim EG, Pringsheim O (1952) Experimental elimination of chromatophores and eye-spot in *Euglena gracilis*. New Phytol 51:65–76

Provasoli L, Hutner SH, Schatz A (1948) Streptomycin-induced chlorophyll-less races of *Euglena*. Proc Soc Exptl Biol Med 69:279–282

Roux E, Graf L, Stutz E (1983) Nucleotide sequence of 'truncated tRNA operon' of the *Euglena gracilis* chloroplast operon. Nucl Acids Res 11:1957–1969

Rubin CM, Schmid CW (1980) Pyrimidine-specific chemical reactions useful for DNA sequencing. Nucl Acids Res 8:4613–4619

Salser W (1977) Globin mRNA sequences: Analysis of base pairing and evolutionary implications. Cold Spring Harbor Symp Quant Biol 42:985–1002

Schmitt JM, Bohnert H-J, Gordon KHJ, Herrmann R, Bernardi G, Crouse EJ (1981) Compositional heterogeneity of the chloroplast DNAs from *Euglena gracilis* and *Spinacia oleracea*. Eur J Biochem 117:375–382

Steege DA, Graves MC, Spremulli LL (1982) *Euglena gracilis* chloroplast small subunit rRNA sequence and base pairing potential of the 3′ terminus, cleavage by colicin E3. J Biol Chem 257:10430–10439

Takaiwa F, Kusuda M, Sugiura M (1982) The nucleotide sequence of chloroplast 4.5S rRNA from a fern, *Dryopteris acuminate*. Nucl Acids Res 10:2257–2260

Takaiwa F, Sugiura M (1982) The complete nucleotide sequence of a 23S rRNA gene from tobacco chloroplasts. Eur J Biochem 124:13–19

Wildeman AG, Nazar RW (1980) Nucleotide sequence of wheat chloroplastid 4.5S ribonucleic acid. J Biol Chem 255: 11896–11900

Communicated by G.A. O'Donovan

Received August 1, 1983

Note added in proof

Karabin et al. (J Biol Chem 258:14790) show the 5S rRNA corresponds to nucleotides −1,198 to −1,084, and suggest different ends and pairing of the 23S rRNA gene.