Volume 11 Number 17 1983 Nucleic Acids Research

# Nucleotide sequence of a *Euglena gracilis* chloroplast genome region coding for the elongation factor Tu; evidence for a spliced mRNA

Paul-Etienne Montandon and Erhard Stutz

Laboratoire de Biochimie, Université de Neuchâtel, Ch. de Chantemerle 18, CH-2000 Neuchâtel, Switzerland

Received 12 July 1983; Revised and Accepted 12 August 1983

ABSTRACT

We characterize a 1.95 kb transcription product of the Euglena gracilis chloroplast DNA fragment Eco-N + Q by S1 nuclease analysis and DNA sequencing and show that it is the product of three splicing events. Exon 1 (0.45 kb), exon 2 (0.74 kb) and 175 nucleotides of exon 3 (0.53 kb) code for the chloroplast elongation factor protein (EF-Tu). The remaining part of exon 3 and exon 4 (0.23 kb) have unidentified open reading frames. The chloroplast EF-Tu protein has 408 aminoacids and is to 70% homologous with the E. coli EF-Tu protein. The active site for aminoacyl-tRNA binding is highly conserved, while the active site for GTP/GDP binding lacks the cysteine present in the E. coli EF-Tu protein. The two introns separating exons 1, 2 and 3 are, respectively, 103 and 110 nucleotides long. The size of the third intron is not yet determined. The splicing rules for eukaryote mRNA are not followed.

## INTRODUCTION

We have recently shown (1) that the Euglena gracilis chloroplast DNA fragment Eco-N (approximately 2.9 kb) interacted with mRNA which directed the synthesis of one major stromal protein of about Mr 53 kd and two minor ones of 40 and 35 kd in a reticulocyte lysate. Passavant et al. reported (2) that a 1.3 kb segment of Eco-N hybridized with an E. coli and a Chlamydomonas chloroplast DNA probe (44) known to carry genetic information for the protein elongation factor EF-Tu. Their tentative conclusion was that Eco-N contains the single gene for the chloroplast specific EF-Tu protein. Spremulli (3) showed that chloroplast EF-Tu activity can be induced by light and this induction step is inhibited by streptomycin indicating that protein synthesis within the chloroplast is required and therefore the genetic information for chloroplast EF-Tu resides in the chloroplast genome.

Continuing our analysis on the genetic function of Eco-N we have sequenced large parts of this fragment and analysed the transcribed region

by S1 nuclease protection experiments. A preliminary account of the S1 nuclease analysis was given (4). We show in this report that Eco-N codes for a protein of 408 aminoacids which is highly homologous to the E. coli EF-Tu protein (5,6). The corresponding mRNA is not colinear with the gene which contains two small introns. Furthermore, we show that the mRNA must be part of a 1.95 kb transcript and that the transcribed region includes about 300 bases of the adjacent Eco-Q. Structural features of the chloroplast EF-Tu gene and protein are compared with the corresponding bacterial gene and protein.

## MATERIALS

Enzymes were purchased from Boehringer-Mannheim and used following instructions of the supplier. [$\alpha$-$^{32}$P]-ATP (400 Ci/mmole) was from Radiochemical Center Amersham.

## METHODS

### Cell culture and preparation of chloroplast RNA

Euglena gracilis (Z. strain, culture collection of Algae Indiana University) was grown heterotrophically in a modified Hutner's medium with vitamin B12 at 50 ng/l (7). Total RNA was extracted from purified chloroplasts (8,9) according to the procedure of Sagher et al. (10).

### Preparation of cloned DNA fragments

Chloroplast DNA fragments Eco-N and Eco-Q were coligated into pBR 322 (11). Recombinant DNA was isolated and purified as described (12). Selected DNA fragments were prepared by digestion of the recombinant DNA clones with appropriate restriction enzymes, separation of fragments on agarose gels and electrophoretic elution of the DNA from gel bands (13). Eluted DNA was extracted with phenol, afterwards with chloroform and precipitated with ethanol.

### S1 nuclease mapping

S1 nuclease resistant DNA:RNA hybrids were prepared as described (14,15). Selected DNA fragments (20 to 50 ng) were denatured in 30 µl of 80% formamide buffer at 85°C for 15 min in the presence of chloroplast RNA (20 µg) and carrier tRNA (160 µg). DNA:RNA hybridization was conducted at 46°C for 3 hours. After S1 nuclease treatment, aliquots of DNA:RNA hybrids were analysed on agarose gels under neutral and alkaline conditions (15).

Transfer of the DNA:RNA hybrids or DNA alone to nitrocellulose filters was according to Southern (16).

Detection of protected DNA:RNA hybrids and DNA fragments

To detect the protected DNA:RNA hybrids and DNA fragments transferred to nitrocellulose filters, we used as probe a nick-translated (17) HaeIII DNA fragment corresponding to probe (a) as defined in Fig. 1B. Nitrocellulose filters were incubated for 8-12 hours at 42°C in a 5x SSPE buffer (1x SSPE : 0.18 M NaCl, 0.01 M sodium phosphate, pH 7.5, 0.001 M EDTA) containing 50% (v/v) formamide, 0.1% SDS (sodium dodecylsulfate), 5x Denhard's reagent (18) and denatured Herring DNA (250 $ug \cdot ml^{-1}$). Afterwards filters were hybridized for 20 hours in the same buffer, same temperature, reducing the concentration of Denhard's reagent to 1x, in the presence of $^{32}P$-labeled DNA ($0.5 \times 10^6$ $cpm \cdot ml^{-1}$). Filters were twice washed in buffer 2xSSC (1x SSC : 0.15 M NaCl, 0.015 M sodium citrate), 0.1% SDS, for 15 min at 20°C, twice in buffer 1x SSC, 0.1% SDS for 30 min at 42°C. Autoradiography of filters was performed using Kodak X-AR-5 films.

DNA sequencing

Subfragments of Eco-N and Eco-Q, as specified below, were cloned into the HincII site of phage M13 mp9 (19,20) and sequenced according to the dideoxy-chain termination method (21). Ligation conditions were as reported by Tait et al. (22). Transformation of E. coli JM 103 was according to Cohen et al. (23) however, competent cells were prepared in 0.02 M Tris-HCl, pH 7.2, 0.1 M $CaCl_2$ and 0.001 M NaCl. As primer we used the synthetic pentadecamer nucleotide from New England Biolabs.

In Fig. 1A we show the sites of the restriction enzymes used to generate the DNA fragments which were ligated into the phage M13 mp9 after blunt ending with the Klenow subfragment of DNA polymerase (24). 1) From Eco-N we prepared the two EcoRI-HindIII fragments of 1.3 and 1.6 kb which were digested with HinfI, HpaII or PvuII. 2) From Eco-N we prepared the HinfI fragments of 0.6 and 1.35 kb which carry the HindIII and the PvuII site, respectively. 3) Eco-Q was digested with Sau3AI or HinfI. The respective digests were directly ligated into the vector. We sequenced both strands of the entire region, although only the sequence of one strand will be displayed.

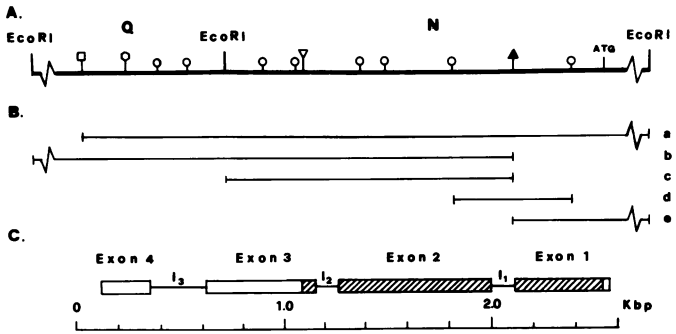



Fig. 1. Restriction sites map of DNA fragment Eco-N + Q and DNA probes used to localize the 1.95 kb transcript.
A. Restriction sites on Eco-N + Q : HaeIII, HinfI, HindIII, HpaII, Sau3AI (only this Sau3AI site is given). ATG, initiation triplet for the EF-Tu protein. B. DNA probes used for S1 nuclease mapping experiments. C. Exon-intron arrangement of the transcription unit. I, intron; boxes symbolize exon regions, hatched boxes symbolize regions coding for the structural part of the EF-Tu protein.

## RESULTS

### a) S1 nuclease analysis of Eco-N and the adjacent fragment Eco-Q

Rutti (11) showed that Eco-N interacted mainly with a mRNA of about 2 kb, and that this transcription product might also contain sequences from the adjacent fragment Eco-Q (for restriction site map see Fig. 1A). In order to size the transcribed region we used subfragments of a recombinant clone containing coligated Eco-N + Q (11) as probes in S1 nuclease experiments (Fig. 1B). Chloroplast RNA was hybridized with DNA probes (a) to (e) and the DNA:RNA hybrids were analyzed by gel electrophoresis under neutral conditions after S1 nuclease treatment. The longest major DNA:RNA hybrid of about 2 kbp was obtained with probe (a) (Fig. 2). Additional fainter bands, within a heavy background, are also discernible in this lane. [For possible origins of these bands see Discussion]. With probes (b) and (c) we obtained DNA:RNA hybrids of about 1.6 and 1.3 kbp. Considering the size and relative position of these two probes on the map given in Fig. 1B we conclude that about 0.3 kb of the 2.0 kb protected DNA region are located on Eco-Q, the rest on Eco-N and, furthermore, that the protected DNA region goes about 0.45 kb beyond the single HindIII site on Eco-N. Probe (d) yielded a DNA:RNA hybrid of about 0.6 kb the result being in line with the size of this probe and its position within Eco-N. Finally, with probe (e) we obtained a

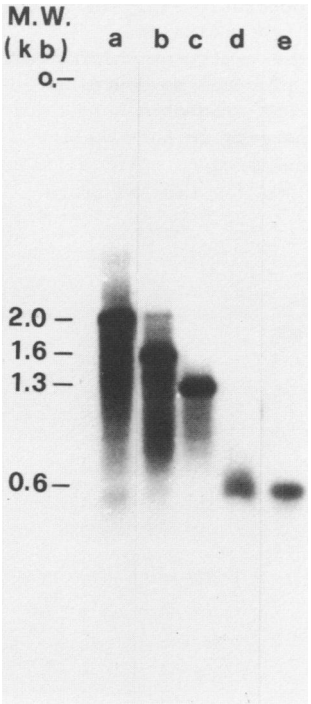



Fig. 2. Gel electrophoretic analysis of the Eco-N + Q transcript under neutral conditions. Total chloroplast RNA was hybridized to the DNA probes shown in Fig. 1B. a) HaeIII-EcoRI, 3.7 kb; b) EcoRI-HindIII, 3.4 kb; c) EcoRI-HindIII, 1.3 kb; d) HinfI, 0.6 kb; e) HindIII-EcoRI, 1.6 kb. DNA:RNA hybrids resistant to S1 nuclease were electrophoretically analysed in neutral 1.4% agarose gel. Nucleic acids were transfered to nitrocellulose paper and again hybridized to $^{32}P$ labeled probe a. O., origin of gel.

DNA:RNA hybrid which had about the same electrophoretic mobility as probe (d). In this case we rather expected a DNA:RNA hybrid of about 0.45 kb taking into account the results obtained with probes (a) to (d) [see Discussion].

To measure the size of the protected DNA fragments we analyzed S1 nuclease resistant DNA:RNA hybrids by gel electrophoresis under alkaline conditions. As shown in Fig. 3 we obtained with probe (a) three strong bands of about 0.74, 0.53 and 0.45 kb and a faint band in the range of about 0.23 kb. The total length of protected DNA fragments, therefore, is 1.95 kb, what is close to the size of the DNA:RNA hybrid (2 kb) obtained with the same probe under neutral conditions. This strongly suggests that the mRNA is not colinear with the corresponding gene region. With probes (b) and (c) we obtained three (0.74, 0.53, 0.23 kb) and two bands (0.74, 0.45 kb), respectively. This means that a protected DNA fragment of 0.23 kb is located on Eco-Q between the HaeIII and EcoRI site while a protected DNA fragment of 0.53 kb is located à cheval of the EcoRI site, 0.45 kb being on Eco-N. Also

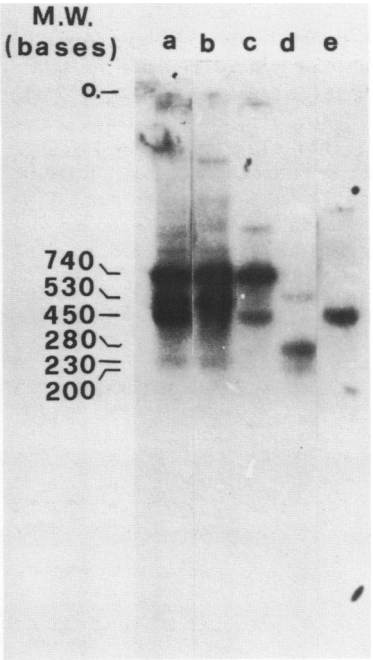



Fig. 3. Gel electrophoretic analysis of the Eco-N + Q transcript under alkaline conditions. S1 nuclease resistant chloroplast DNA:RNA hybrids were prepared as described in Fig. 2. Protected DNA fragments were analysed by electrophoresis on an alkaline 1.8% agarose gel, transfered to nitrocellulose paper and hybridized to $^{32}P$ labeled probe a. O. origin of gel.

with probe (d) we obtained two small subfragments of about 0.28 and 0.2 kb, i.e., this region must be part of two separate protected fragments. In fact it is possible to approximately place protected DNA fragments of 0.74 and 0.45 kb on the map. Probe (e) yielded a single fragment of about 0.45 kb, what agrees with all the other results and the map position of the probe.

According to the results presented in Fig. 2 and Fig. 3 the major 2 kb transcript is not colinear with the corresponding gene region, rather it is a composite of four exons separated by three introns (Fig. 1C). This was substantiated by the analysis of the DNA:RNA hybrids of probe (a) in a two-dimensional gel under neutral and alkaline conditions. We obtained the same four DNA subfragments from the 2 kb DNA:RNA hybrid under alkaline conditions (data not shown).

b) DNA sequence of the transcribed region

According to Passavant et al. (2) the smaller of the two EcoRI HindIII fragments of Eco-N (1.3 kb) codes for the C-terminal part of the chloroplast EF-Tu protein. We show 2066 nucleotides of this and the adjacent regions in Fig. 4. Only the RNA-like strand is presented along with the deduced amino-acid sequence for the chloroplast EF-Tu protein. Aligned with the chloroplast

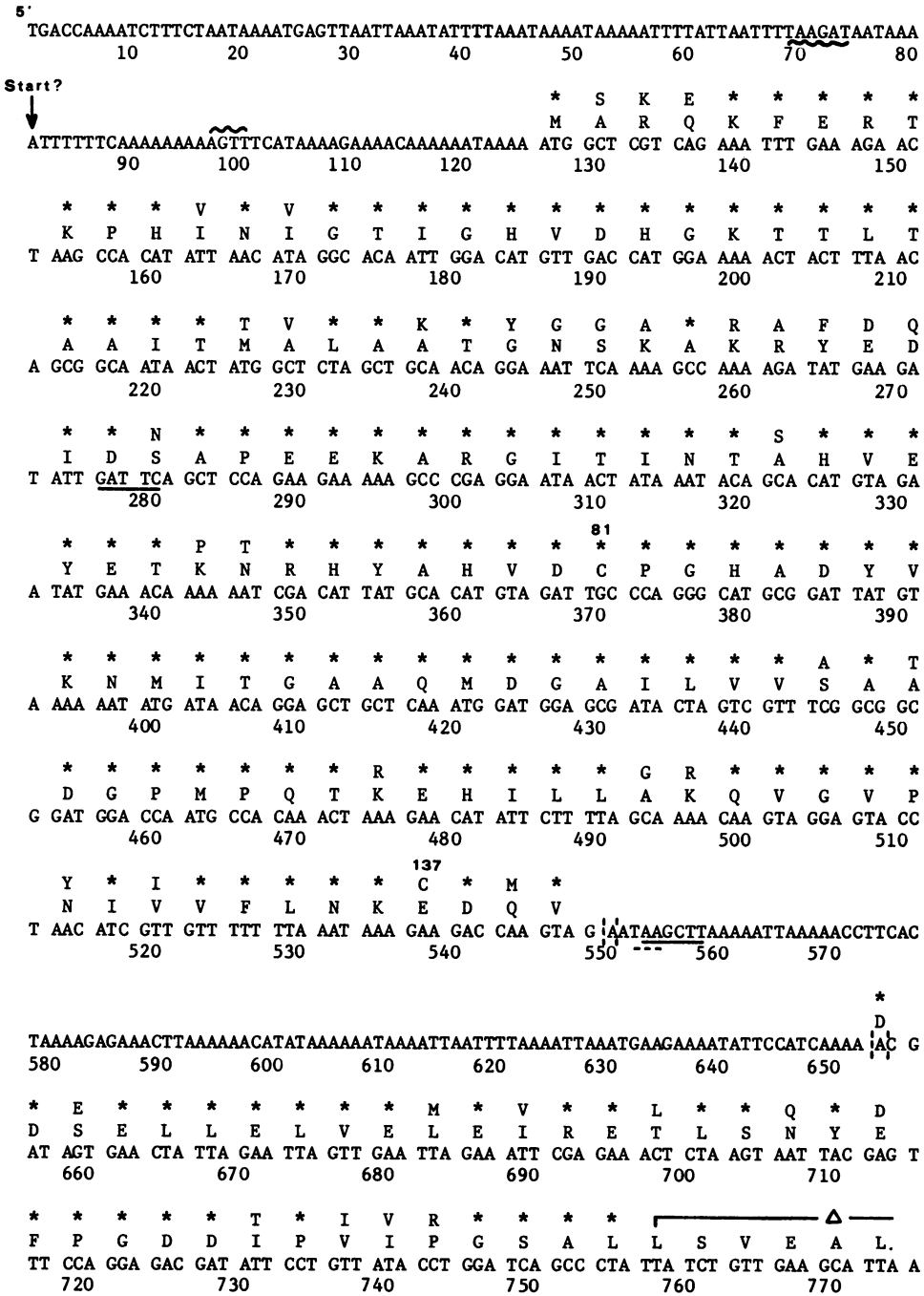
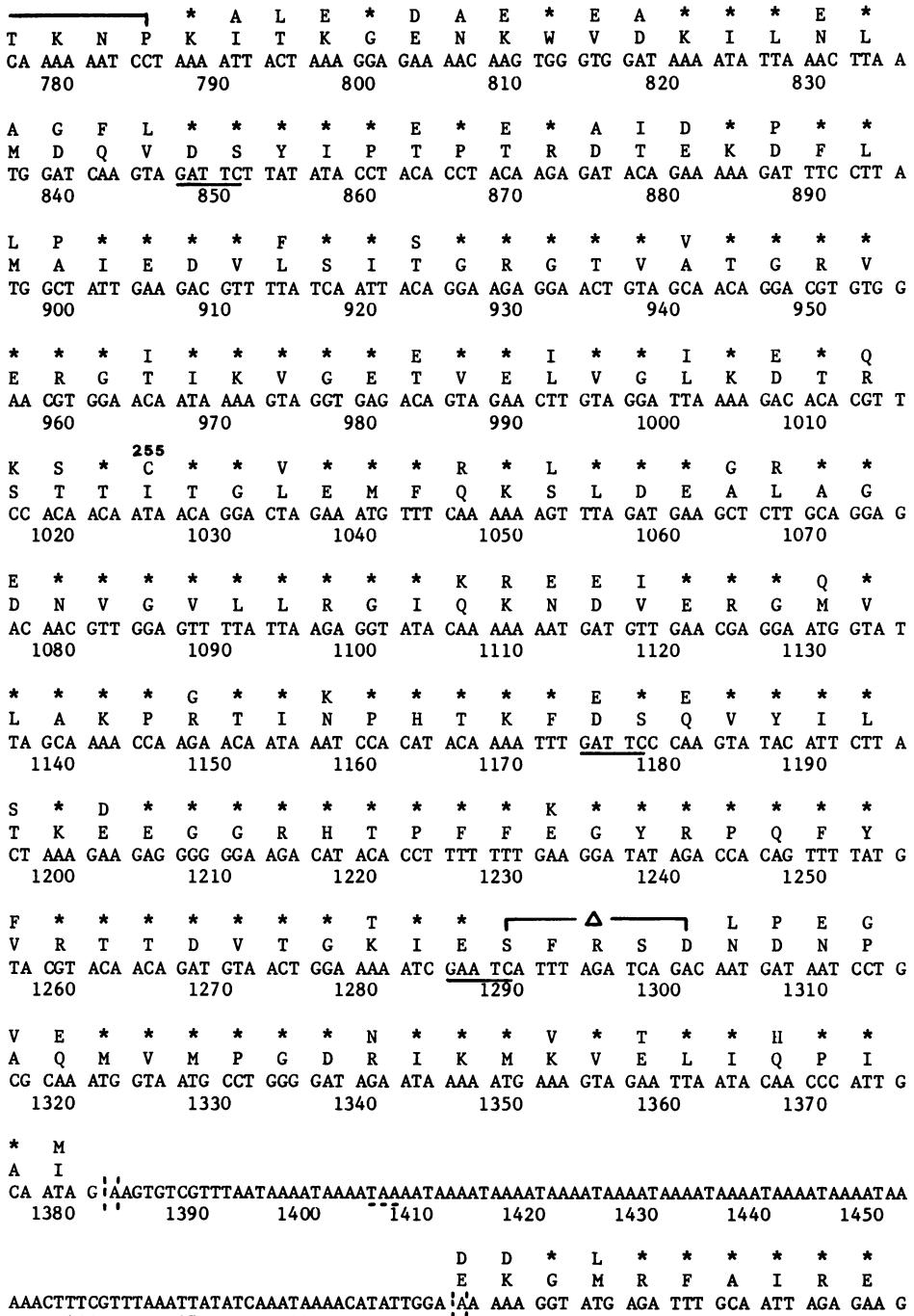

5'
TGACCAAAATCTTTCTAATAAAATGAGTTAATTAAATATTTTAAATAAAATAAAAATTTTATTAATTTTAAGATAATAAA
10 20 30 40 50 60 70 80
Start?
* S K E * * * * *
M A R Q K F E R T
ATTTTTTCAAAAAAAAAGTTTCATAAAAGAAAACAAAAAATAAAA ATG GCT CGT CAG AAA TTT GAA AGA AC
90 100 110 120 130 140 150
* * * V * V * * * * * * * * * * * * * *
K P H I N I G T I G H V D H G K T T L T
T AAG CCA CAT ATT AAC ATA GGC ACA ATT GGA CAT GTT GAC CAT GGA AAA ACT ACT TTA AC
160 170 180 190 200 210
* * * * T V * * K * Y G G A * R A F D Q
A A I T M A L A A T G N S K A K R Y E D
A GCG GCA ATA ACT ATG GCT CTA GCT GCA ACA GGA AAT TCA AAA GCC AAA AGA TAT GAA GA
220 230 240 250 260 270
* * N * * * * * * * * * * * * * S * * *
I D S A P E E K A R G I T I N T A H V E
T ATT GAT TCA GCT CCA GAA GAA AAA GCC CGA GGA ATA ACT ATA AAT ACA GCA CAT GTA GA
280 290 300 310 320 330
81
* * * P T * * * * * * * * * * * * * * *
Y E T K N R H Y A H V D C P G H A D Y V
A TAT GAA ACA AAA AAT CGA CAT TAT GCA CAT GTA GAT TGC CCA GGG CAT GCG GAT TAT GT
340 350 360 370 380 390
* * * * * * * * * * * * * * * * * A * T
K N M I T G A A Q M D G A I L V V S A A
A AAA AAT ATG ATA ACA GGA GCT GCT CAA ATG GAT GGA GCG ATA CTA GTC GTT TCG GCG GC
400 410 420 430 440 450
* * * * * * * R * * * * * G R * * * * *
D G P M P Q T K E H I L L A K Q V G V P
G GAT GGA CCA ATG CCA CAA ACT AAA GAA CAT ATT CTT TTA GCA AAA CAA GTA GGA GTA CC
460 470 480 490 500 510
137
Y * I * * * * * C * M *
N I V V F L N K E D Q V
T AAC ATC GTT GTT TTT TTA AAT AAA GAA GAC CAA GTA G AATAAGCTTAAAAATTAAAAACCTTCAC
520 530 540 550 560 570
*
D
TAAAAGAGAAACTTAAAAAACATATAAAAAATAAAATTAATTTTAAAATTAAATGAAGAAAATATTCCATCAAAA AC G
580 590 600 610 620 630 640 650
* E * * * * * * * M * V * * L * * Q * D
D S E L L E L V E L E I R E T L S N Y E
AT AGT GAA CTA TTA GAA TTA GTT GAA TTA GAA ATT CGA GAA ACT CTA AGT AAT TAC GAG T
660 670 680 690 700 710
* * * * * T * I V R * * * * Δ
F P G D D I P V I P G S A L L S V E A L.
TT CCA GGA GAC GAT ATT CCT GTT ATA CCT GGA TCA GCC CTA TTA TCT GTT GAA GCA TTA A
720 730 740 750 760 770

Fig. 4.



```
*   *   *   *   *   *   *   *   *   *   A   K   V   L   G  TER
G   G   R   T   V   G   A   G   V   V   L   S   I   I   Q  TER
GT GGT AGA ACT GTG GGT GCC GGC GTT GTT TTA TCA ATA ATA CAA TAATTAATTTATGAATCTAAG
       1530        1540        1550        1560        1570        1580

AGATATAAATAATATGACACTTTCAAAAAATGAAAATATAAAAGCAAAACAAAAACAAATAAACCTCCCAAAAATATTGC
1590       1600       1610       1620       1630       1640       1650       1660

GTCAAGAAATAAAAGAAAATAATAAAATAATAAAATGGTTTTATAATATTGTAATGTTATTAGGAGGAATCGGATTTTTA
1670       1680       1690       1700       1710       1720       1730       1740

ATAGTAGGAATATCAAGTTATATAGGAAATAATCTAATATATTTTTTAGACGCAAGCGAAATAATATTTTTTCCTCAAGG
1750       1760       1770       1780       1790       1800       1810       1820

AATAACAATGTGTTTTTATGGTACTTGCGGAATTTTATTTAGTATAAATCAAATAAGTATAATTTTAAACGGAGTTGGCG
1830       1840       1850       1860       1870       1880       1890       1900

AAGGTTATAACGAATTCAACAAAGAGTTAAATTTAATGACAATTTACCGAAAGGGTAAACAAGGAAAAAATTCTGATATA
1910       1920       1930       1940       1950       1960       1970       1980

AATATAACGTATTCACTTAAAGACATAGTAAAAGAATGAGATAACTGCAATTAATATACGTCTCAATAAAATATATAAGT
1990       2000       2010       2020       2030       2040       2050       2060
```

Fig. 4. Nucleotide sequence of the EF-Tu gene and adjacent regions. Only the RNA like strand is shown along with the deduced aminoacid sequence of the EF-Tu chloroplast protein. ATG, initiation triplet (position 126); TAA, termination triplet (position 1566). The aminoacid sequence of the E. coli EF-Tu protein is aligned with the chloroplast EF-Tu protein. Aminoacid identity is marked by an asterisk, - Δ - indicates that the EF-Tu protein of E. coli lacks the corresponding aminoacid sequence found in the chloroplast counterpart. The position of all three cystein residues in the E. coli EF-Tu protein are marked. Vertical dashed lines mark suggested splicing sites (exon - intron boundaries); horizontal dashed underlines mark the first stop codon in introns; a horizontal dashed overline marks the start of an URF (unidentified reading frame) in exon 3; a wave underline marks a possible RNA polymerase binding site (position 70); a wave overline marks a possible ribosome binding site (position 97); an arrow marks a possible transcription initiation site (position 81); EcoRI, HindIII, HinfI and HpaII restriction sites are underlined.

EF-Tu aminoacid sequence is the aminoacid sequence of the E. coli EF-Tu protein (5,6). The sequencing results can be summarized as follows : 1) The sequencing data allow to exactly locate the structural part of the chloroplast EF-Tu gene. The start codon (nucleotide position 126 to 128) is within the large EcoRI-HindIII DNA fragment (1.6 kb) which codes for 140 out of the 408 aminoacids, i.e., about 34% of structural gene are on the large EcoRI-HindIII fragment leaving only 66%, on the small EcoR-HindIII fragment.
2) An abrupt change in nucleotide sequence occurs at position 551 immediately before the HindIII site. This represents the end of the first exon, which is most likely spliced to the second exon which starts at position 653 as

will be discussed. Thereby an intron of 103 nucleotides is cleaved out. The second exon reaches from position 653 to position 1381 coding for the next 243 aminoacids of the EF-Tu protein. 3) A second intron of 110 nucleotides with a very high AT content leads to the third exon which starts at position 1491 and codes for the last 25 aminoacids of the EF-Tu protein (C-terminal). 4) According to the S1 nuclease analysis exon 3 does not stop at or close to the end of the EF-Tu gene, but rather extends into the adjacent fragment Eco-Q having a total size of about 0.53 kb. A computer analysis of the corresponding DNA sequence reveals that a major unidentified open reading frame (URF) starts at position 1576 and ends at position 2023. Exon 3 could code for a polypeptide of about 149 aminoacids (aminoacid sequence not displayed). 5) The 2 kb transcript contains a forth exon of about 0.23 kb which must be placed somewhere between exon 3 and the $tRNA^{Gln}$ gene very recently mapped and sequenced on Eco-Q (25). We have also sequenced this DNA segment and found e.g. a major URF open for 101 aminoacids (not shown on Fig. 4). This hypothetical URF is separated by an intron of 110 positions from exon 3. In Fig. 1C the precise position of exons 1, 2 and 3 are shown along with a tentative position of exon 4.

c) The chloroplast EF-Tu protein

The aminoacid sequence of the chloroplast EF-Tu protein, as deduced from the DNA sequencing data, shows a remarkably high similarity with the E. coli EF-Tu protein (about 70%). In Fig. 4 the two sequences are aligned from the start to the stop codon. In the N-terminal part and up to about aminoacid position 140 the sequence homology between the two proteins is in the range of 90%. This includes a long conservative stretch of 32 aminoacids in the vicinity of cystein 81 which is part of the active site involved in the aminoacyl-tRNA binding (rev. 26). The chloroplast EF-Tu protein has cystein 81 but does not have cystein 137 which is part of the GTP/GDP binding site. The middle part of the chloroplast EF-Tu is less homologous in its sequence with the bacterial counterpart (50 to 70%). The sequential homology is twice interrupted by insertions of ten and five aminoacids, as symbolized in Fig. 4 by a deletion in the E. coli protein. A third cystein in the E. coli protein (pos. 255) is replaced by isoleucine in the chloroplast EF-Tu protein. No particular function is attributed to this cystein. The C-terminal part contains a stretch of 16 aminoacids homologous with the E. coli counterpart and terminates in phase with the E. coli protein, the last five aminoacids,

TABLE 1. A comparison of the codon usage in the *Euglena gracilis* chloroplast and *E. coli* tuf A genes

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Phe | UUU | 10 | (1) | Ser | UCU | 2 | (7) | Tyr | UAU | 7 | (2) | Cys | UGU | 0 | (1) |
| | UUC | 1 | (13) | | UCC | 2 | (3) | | UAC | 2 | (8) | | UGC | 1 | (2) |
| Leu | UUA | 18 | (0) | | UCA | 7 | (0) | Ter | UAA | 1 | (1) | Ter | UGA | 0 | (0) |
| | UUG | 0 | (0) | | UCG | 1 | (0) | | UAG | 0 | (0) | Trp | UGG | 1 | (1) |
| Leu | CUU | 5 | (1) | Pro | CCU | 9 | (0) | His | CAU | 10 | (1) | Arg | CGU | 5 | (21) |
| | CUC | 0 | (1) | | CCC | 1 | (0) | | CAC | 0 | (10) | | CGC | 0 | (2) |
| | CUA | 6 | (0) | | CCA | 9 | (1) | Gln | CAA | 11 | (0) | | CGA | 4 | (0) |
| | CUG | 0 | (26) | | CCG | 0 | (19) | | CAC | 2 | (8) | | CGG | 0 | (0) |
| Ile | AUU | 12 | (3) | Thr | ACG | 12 | (13) | Asn | AAU | 11 | (0) | Ser | AGU | 3 | (0) |
| | AUC | 2 | (26) | | ACC | 0 | (16) | | AAC | 6 | (7) | | AGC | 0 | (0) |
| | AUA | 18 | (0) | | ACA | 23 | (1) | Lys | AAA | 27 | (18) | Arg | AGA | 13 | (0) |
| Met | AUG | 13 | (10) | | ACG | 0 | (0) | | AAG | 2 | (5) | | AGG | 0 | (0) |
| Val | GUU | 13 | (24) | Ala | GCU | 8 | (13) | Asp | GAU | 19 | (4) | Gly | GGU | 6 | (19) |
| | GUC | 1 | (0) | | GCC | 4 | (1) | | GAC | 7 | (21) | | GGC | 2 | (21) |
| | GUA | 17 | (10) | | GCA | 11 | (5) | Glu | GAA | 27 | (30) | | GGA | 23 | (0) |
| | GUG | 3 | (3) | | GCG | 16 | (8) | | GAG | 3 | (0) | | GGG | 3 | (1) |

Those codons for which the corresponding tRNA have been sequenced (references in the text) are underlined. The *E. coli* values are from (27).

however, being different in the two EF-Tu proteins.

Table 1 shows the codon usages for translating the chloroplast and bacterial EF-Tu protein. Euglena chloroplasts use codons rich in A and/or U, while E. coli prefers in the tuf A and tuf B gene codons rich in G and/or C. The difference in codon usage is manifested in a considerable difference in the nucleotide sequence of the chloroplast and the bacterial (27,28) genes, with sequence homologies in the range of 50 to 60%. Several Euglena chloroplast tRNA genes have been sequenced (25, 29-32), and we underline in Table 1 the corresponding 16 codons. We see that 14 of the 16 codons were used, several at very high frequency. Only $tRNA^{His}_{CAC}$ and $tRNA^{Ser}_{AGC}$ were not used for the synthesis of the EF-Tu protein.

d) Introns and splicing sites

Higher plant chloroplast protein genes sofar analyzed seem not to contain introns (e.g. 33-36). For Euglena gracilis this seems not to be the case : Stiegler et al. (37) have already suggested, based on hybridisation experiments, that the chloroplast rbc L* gene contains several introns. Our data allow for the first time to exactly locate two introns within a chloroplast protein gene. Guided by the aminoacid sequence of the bacterial protein, we could define the two most likely splicing frames for introns 1 and 2 as indicated in Fig. 4. The first splicing event maintains the reading frame, cutting out an intron which starts with a nonsense codon and is extremely rich in AT. The second splicing event allows to keep in phase with the aminoacid positions of the E. coli protein and also removes a AT rich DNA stretch containing stop codons and eleven tandemly repeated pentanucleotides. We have placed the third intron at position 2020 considering the S1 nuclease result and the trinucleotide motive [GAA] of the first and second splicing event. This third splicing event also eliminates an AT rich segment which starts with a nonsense codon. Assuming that the URFs of exons 3 and 4 are spliced and translated there would result a single protein of about Mr 26 000.

## DISCUSSION

### 1. Transcription product analysis

The S1 nuclease analysis yields a single transcript of 1.95 kb which is the product of three splicing events. About 64% of its length code for the EF-Tu protein and the rest may or may not be translated. The question arises

whether the remaining part of Eco-N is transcribed under the prevailing experimental condition (light grown, heterotrophic cells). The autoradiograph of Fig. 2 and especially lanes a and b show additional bands, e.g., in the range of 1.8 and 1.6 kb within a strong background. These bands could be other mRNAs not related to the 1.95 kb transcript since the Eco-N contains additional 1.4 kb of unassigned sequences upstream of the EF-Tu gene. The results presented here, however, strongly suggest that the fainter bands of lower mobility in lanes a and b are DNA:RNA hybrids of the 1.95 kb transcript which are shortened by partial S1 nuclease attack of the RNA linking the DNA fragments. This assumption is corroborated by the absence of similar bands in lane (e), where, e.g., additional DNA:RNA hybrids should become apparent if the 1.4 kb region of Eco-N is transcribed (see Fig. 1).

With probe (e) we obtain under neutral and alkaline conditions a DNA:RNA hybrid of 0.6 kb and a DNA segment of 0.45 kb, respectively. The sequencing results show that there are 428 positions from the HindIII site to the AUG triplet of the tuf A* gene. This value is in line with the "alkaline" S1 nuclease results, however it does not exactly match the "neutral" S1 nuclease results. This size discrepancy could be due to an irregular gel electrophoretic mobility of the DNA:RNA hybrid, or to the presence of a short undetected exon in the leader part.

2. Regulatory sites and splicing events

We have sequenced the region upstream to the EF-Tu gene as shown in Fig. 4. This region contains at position 69 a 5'-TAAGAT- motive which may qualify as "Pribnow" box (38). Transcription might start at position 81 (arrow) where an A is followed by a cluster of Ts. At position 97 we recognize the tetranucleotide 5'-AGTT- which is complementary to the 3' end of the chloroplast 16S rRNA gene 3'-ccTCAAc- (12) and may function as ribosome recognition site (39). It is noteworthy that the underlined A deviates from the bacterial Shine-Dalgarno sequence which has a C at this position.

Both, the S1 nuclease analysis and the DNA sequencing data clearly show that the chloroplast EF-Tu gene contains two introns. Evidence for a third intron within the 1.95 kb transcript is based solely on the S1 nuclease analysis. Within the EF-Tu gene, it is possible to exactly locate the splicing sites if we accept the aminoacid sequence and therefore the reading frame of the bacterial gene as reference points. For both splicing events the splicing sites are almost identical, however, the splicing rules for

eukaryotic mRNAs (40) are not followed. Rather we see that the motive ($I_1$) 5'-GA:AU-AA:CG- for the Euglena chloroplast RNA resembles e.g. the splicing motives of maize chloroplast $tRNA^{Ile}$ (41) which reads 5'-UA:AU-CA:CG-.

According to the S1 nuclease results the stable 1.95 transcript ends about 300 nucleotides beyond the EcoRI (N/Q) site. In this region we detected several runs of "T" which may qualify as transcription stop signals (data not shown). We may add that the genes for $tRNA^{Gln}$ and $tRNA^{Ser}$ (25) are on the same strand with the same polarity as the tuf*A gene.

3. The chloroplast EF-Tu protein

The aminoacid sequence of the chloroplast EF-Tu protein shows a high degree of homology with the bacterial counterpart, leaving no doubt that the chloroplast genome codes for this important catalytic protein as suggested by Spremulli (3) and by Hallick's group (2). The EF-Tu protein contains a total of 408 aminoacids and has calculated molecular weight of 45011. To the best of our knowledge the protein has not yet been purified in preparative amounts although attempts to do so are made (personal note of Prof. Ciferri). Therefore a comparison with the native protein is not yet possible. According to the deduced aminoacid sequence the chloroplast EF-Tu protein is highly conserved especially in the N-terminal part, including the cystein 81 which is part of the active site involved in the aminoacyl tRNA binding step (26). On the other hand cystein 137 is replaced in the chloroplast protein by aspartic acid, i.e., the function of the SH-group in this GTP/GDP binding site can be taken over by other aminoacids. Beck and Spremulli (42) observed that the Euglena cytoplasmic EF-1 protein is quite resistant towards the SH-binding reagent N-ethylmaleimide, suggesting that the -SH groups are not always essential for EF-Tu activity.

How does the size of the chloroplast EF-Tu as deduced from the sequencing data correlate with the proteins synthesized in a rabbit reticulocyte lysate under the direction of Eco-N selected mRNA ? In our previous study we postulated that Eco-N codes for a major stromal protein of Mr 53 000 which in its papain and chymotrypsin digest pattern was identical to a stromal protein of the same size synthesized in isolated chloroplasts (1). In those experiments two additional minor proteins of Mr 35 000 and 40 000 were discernable which we suggested to be premature translation termination products of the major protein. Assuming that the most abundant mRNA, i.e. the 1.95 kb transcript was preferentially selected and translated in the previous experiment and

considering the data given here we postulate that the protein with the apparent Mr 53 000 is related to the EF-Tu protein. Very recently Miller et al. (43) showed using the same approach as ours (1) that the reticulocyte lysate synthesizes a major protein of Mr close to 45 000 which they thought to be the EF-Tu protein. These results would be in perfect agreement with the calculated size of our sequencing results.

ACKNOWLEDGMENTS

We are grateful to Drs S.T. Surzycki (Bloomington) and J.D. Rochaix (Geneva) for providing bacterial and Chlamydomonas chloroplast clones containing the EF-Tu gene, which were used in preliminary experiments. We are grateful to B. Rutti for providing the Euglena DNA clones containing Eco-N + Q and to Dr. O. Hagenbüchle (Lausanne) for introducing us to the dideoxy sequencing method. We thank C. Aegerter for technical help and Ch. Bachmann for secretarial help. The computer programs were obtained from R. Staden (Cambridge). This research is supported by Fonds national suisse de la Recherche scientifique (3.183.82 to E.S.).

*A terminology equivalent to that used for specifying E. coli genes is applied. rbc L, gene for the large subunit of ribulosebisphosphatecarboxylase; tuf A, gene for the elongation factor Tu. [Nomenclature according to rules discussed in Porto Portese, Nato-FEBS Meeting, August 1982].

REFERENCES

1. Rutti, B., Keller, M., Ortiz, W. and Stutz, E. (1981) FEBS Lett. 134, 15-19.
2. Passavant, C.W., Stiegler, G.L. and Hallick, R.B. (1983) J. Biol. Chem. 258, 693-695.
3. Spremulli, L.L. (1982) Arch. Biochem. Biophys. 214, 734-741.
4. Montandon, P.E., Rutti, B. and Stutz, E. (1983) Experientia 39, 669.
5. Jones, D.M., Petersen, T.E., Nielsen, K.M., Magnusson, S., Sottrup-Jensen, L., Gansing, K. and Clark, B.F.C. (1980) Eur. J. Biochem. 108, 507-526.
6. Arai, K., Clark, B.F.C., Duffy, L., Jones, M.D., Kaziro, Y., Laursen, R.A., L'Italien, J., Miller, D.L., Nagarkatti, S., Nakamura, S., Nielsen, K.M., Petersen, T.E., Takahashi, K. and Wade, M. (1980) Proc. Natl. Acad. Sci. USA 77, 1326-1330.
7. Vasconcelos, A.C., Pollak, M., Mendiola, L.R., Hoffmann, H.P., Brown, D.H., Price, C.A. (1971) Plant Physiol. 47, 217-221.
8. Ortiz, W., Reardon, E.M. and Price, C.A. (1980) Plant Physiol. 66, 291-294.
9. Ortiz, W. and Stutz, E. (1980) FEBS Lett. 116, 298-302.
10. Sagher, D., Grosfeld, H. and Edelman, M. (1976) Proc. Natl. Acad. Sci. USA 73, 722-726.
11. Rutti, B. (1982) Ph.D. thesis, University of Neuchâtel, Switzerland.
12. Graf, L., Roux, E., Stutz, E. and Kössel, H. (1982) Nucleic Acids Res. 10, 6369-6381.
13. Wienand, U., Schwarz, Z. and Feix, G. (1979) FEBS Lett. 98, 319-323.
14. Berk, A.J., Sharp, P.A. (1977) Cell 12, 721-732.

15. Favaloro, J., Treisman, R., Kamen, R. (1980) Methods in Enzymol. 65, (Grossman, L., Moldave, K. eds, Academic Press, New York, pp. 718-749.
16. Southern, E.M. (1975) J. Mol. Biol. 98, 503-517.
17. Rigby, P.W.J., Dieckman, H., Rhodes, C. and Berg, P. (1977) J. Mol. Biol. 113, 237-251.
18. Denhard, D. (1966) Biochem. Biophys. Res. Commun. 23, 641-646.
19. Messing, J., Crea, R., Seeberg, P.H. (1980) Nucleic Acids Res. 9, 309-321.
20. Messing, J. and Vieira, J. (1982) Gene 19, 269-276.
21. Sanger, F., Coulson, A.R., Barrel, B.G., Smith, A.J.H. and Roe, B.A. (1980) J. Mol. Biol. 143, 161-178.
22. Tait, R.C., Rodriguez, R.L. and West, R.W. (1980) J. Biol. Chem. 255, 813-815.
23. Cohen, S.W., Chang, A.C.Y. and Hsu, L. (1972) Proc. Natl. Acad. Sci. USA 69, 2110-2114.
24. Wartell, R.M. and Reznikoff, W.S. (1980) Gene 9, 307-310.
25. Karabin, G.D. and Hallick, R.B. (1983) J. Biol. Chem. 258, 5512-5518.
26. Miller, D.L. and Weissbach, H. (1977) in Molecular Mechanisms of Protein Synthesis, Weissbach, H. and Pestka, S. eds, pp. 323-373, Academic Press, New York.
27. Yokota, T., Sugisaki, H., Takanami, M. and Kaziro, Y. (1980) Gene 12, 25-31.
28. An, G. and Friesen, J.D. (1980) Gene 12, 33-39.
29. Graf, L., Kössel, H. and Stutz, E. (1980) Nature 286, 908-910.
30. Orozco, E.M., Rushlow, K.E., Dodd, J.R. and Hallick, R.B. (1980) J. Biol. Chem. 255, 10997-11003.
31. Orozco, E.M. and Hallick, R.B. (1982) J. Biol. Chem. 257, 3265-3275.
32. Hollingsworth, M.J. and Hallick, R.B. (1982) J. Biol. Chem. 257, 12795-12799.
33. McIntosh, L., Paulsen, C. and Bogorad, L. (1980) Nature 288, 556-560.
34. Zurawski, G., Perrot, B., Bottomley, W. and Whitfeld, P.R. (1981) Nucleic Acids Res. 9, 3251-3270.
35. Zurawski, G., Bohnert, H.J., Whitfeld, P.R. and Bottomley, W. (1982) Proc. Natl. Acad. Sci. USA 79, 7699-7703.
36. Howe, C.J., Auffret, A.D., Doherty, A., Bowman, C.M., Dyer, T.A. and Gray, J.C. (1982) Proc. Natl. Acad. Sci. USA 79, 6903-6907.
37. Stiegler, G.L., Mattews, H.M., Bingham, S.E. and Hallick, R.B. (1982) Nucleic Acids Res. 10, 3427-3444.
38. Pribnow, D. (1975) Proc. Natl. Acad. Sci. USA 72, 784-788.
39. Shine, J. and Dalgarno, L. (1974) Proc. Natl. Acad. Sci. USA 71, 1342-1346.
40. Sharp, P.A. (1981) Cell 23, 643-646.
41. Koch, W., Edwards, K. and Kössel, H. (1981) Cell 25, 203-213.
42. Beck, C.M. and Spremulli, L.L. (1982) Arch. Biochem. Biophys. 215, 414-424.
43. Miller, E., Jurgenson, J.E., Reardon, E.M. and Price, C.A. (1983) J. Biol. Chem. in the press.
44. Watson, J.C. and Surzycky, S.J. (1982) Proc. Natl. Acad. Sci. USA 79, 2264-2267.