Volume 304, number 2,3, 252–255 FEBS 11162 June 1992
© 1992 Federation of European Biochemical Societies 00145793/92/$5.00

The *Euglena gracilis rbcS* gene contains introns with unusual borders

Luc-Henri Tessier, Raquel L. Chan, Mario Keller, Jacques-Henry Weil and Patrice Imbault.

Institut de Biologie Moleculaire des Plantes du CNRS, Université Louis Pasteur de Strasbourg, 12 rue du Général Zimmer, 67084 Strasbourg Cedex, France

Received 27 January 1992; revised version received 29 April 1992

We have recently shown that, in *Euglena gracilis*, leader sequences are transferred by trans-splicing to the vast majority of cytoplasmic mRNAs. Trans-splicing is involved in the maturation of the *rbcS* transcript, which encodes eight small subunits of the ribulose 1,5 bisphosphate carboxylase/oxygenase. In this report, we show that the *Euglena rbcS* gene introns are different from introns found in plant *rbcS* genes. In addition these introns do not have the conserved 5′ and 3′ border sequences found in introns of eucaryotic nuclear-encoded pre-mRNAs, and they do not present any homology with self-splicing introns of groups I and II. Secondary structure analyses show that the 5′ and 3′ ends of *Euglena* introns can base-pair, suggesting that an unusual splicing mechanism exists in *Euglena*.

Euglena gracilis; rbcS gene; Intron; GT–AG rule; Splicing

1. INTRODUCTION

Euglena gracilis is a protist which has the capacity to grow on organic substrates in the dark, and also to perform photosynthesis when exposed to light. We have recently demonstrated that, in this organism, a trans-splicing mechanism transfers leader sequences to a vast majority of mRNAs [1]. This finding provides a strong argument for placing euglenoids close to trypanosomatids, and far from algae, in an evolutionary tree. However, in *Euglena* the genomic organization markedly differs from that of trypanosomes. For example, we have shown that the sequence leader (SL)-RNA genes of *Euglena* are located in repeated units which also code for the 5 S rRNA [2]: such an association is not found in trypanosomes [3] but, paradoxically, exists in several nematodes [4], which are phylogenically more distant from *Euglena*. While cis-splicing is unknown in trypanosomes [5], preliminary studies on the *rbcS* gene suggested that *Euglena* is a unique organism in which both trans-splicing and cis-splicing mechanisms co-exist [1]. Cis-splicing is presumably involved in the maturation process, since the *rbcS* gene transcript is 4.2 kb long and codes for eight small subunits of the ribulose 1,5 bisphosphate carboxylase/oxygenase [6], whereas the *rbcS* gene itself is about 15 kb in length. To gain an insight into mRNA maturation in *Euglena* we have undertaken the determination of the nucleotide sequence of the *Euglena rbcS* nuclear gene. In this paper we present the parts of this gene which code for the transit peptide (which is removed during import into the chloroplast) [7], the first small subunit, and the 3′ non-coding region: this structure corresponds to a complete plant *rbcS* gene unit. We report that the intron structure in *Euglena* in unrelated to the structure of plant *rbcS* introns, and, more generally, is different from the structure of eucaryotic pre-mRNA introns.

2. MATERIAL AND METHODS

2.1. *Genomic DNA library screening and subcloning*

Euglena DNA, partially digested with the restriction enzyme *Sau*3A, was cloned into the *Bam*HI site of Charon 40 lambda phage [8] and 3 × 10^5 recombinants were screened by plaque hybridization using ^{32}P-labeled probes specific for different regions of the *Euglena* SSU cDNA [6]. Nineteen positive clones were detected by probing with the 0.4 kb *Bgl*II restriction fragment containing the coding region for an entire mature SSU. Four of these clones (5B1, 5B2, 5E1 and 5I1) also hybridize with an oligonucleotide specific for the 5′ end of the cDNA, while three other clones hybridize with an oligonucleotide specific for the 3′ non-coding region (3C1, 3G1 and 3I1). Two phages (5E1 and 3G1) were shown, by hybridization and sequencing, to contain a common region and, therefore, to overlap. The complete *rbcS* gene distributed in the inserts of these two phages extends over 15 kb.

2.2. *Nucleotide sequencing*

Nucleotide sequencing was carried out by the dideoxy chain termination method using T7 DNA polymerase [9] and universal or specific primers on double strand DNA templates. The sequencing data were compiled and analyzed using the Wisconsin computer group program package.

3. RESULTS AND DISCUSSION

We have determined the nucleotide sequence encoding the transit peptide, the first small subunit (SSU), and

Correspondence address: P. Imbault, Institut de Biologie Moleculaire des Plantes du CNRS, Université Louis Pasteur de Strasbourg, 12 rue du Général Zimmer, 67084 Strasbourg Cedex, France. Fax: (33) 88 61 44 42.

Published by Elsevier Science Publishers B.V.

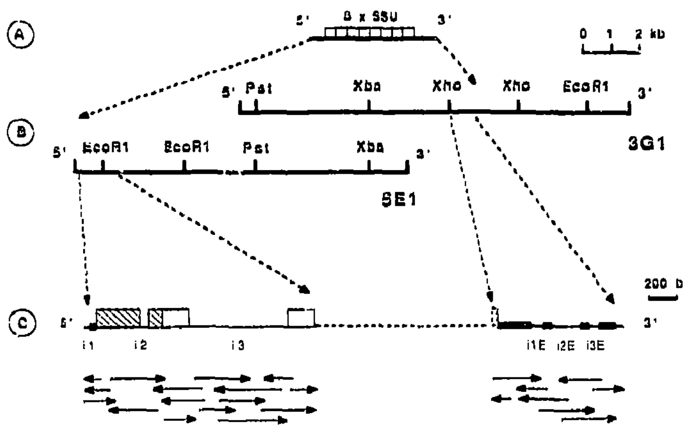

Fig. 1. (A) Map of the mRNA coding for the *Euglena* SSU [6]. The repeated sequence, encoding the eight SSUs, is indicated as 8 × SSU. (B) Maps of the inserts of the overlapping phages, 5E1 and 3G1, corresponding, respectively, to the 5′ and the 3′ parts of the mRNA. (C) Diagram of the 5′ and 3′ ends of *Euglena rbcS* gene. (Black boxes) Non-coding regions of the mRNA; (shaded boxes) regions coding for the transit peptide; (dotted boxes) regions coding for the first SSU; (dotted box surrounded by a dotted line) end of the last SSU; (solid lines) introns numbered as indicated below the line; (dotted lines) internal region of *rbcS* gene.

the 3′ non-coding region, from the inserts of the two overlapping phages, 5E1 and 3G1, which cover the complete *rbcS* gene. The structure of these different regions of *Euglena rbcS* gene are presented in Figs. 1 and 2. Three different observations lead us to the conclusion that these sequences correspond to a functional *rbcS* gene: (i) the restriction map of 5E1 is identical to that of the three other positive 5′ clones (5B1, 5B2 and 5I1); (ii) the partial sequence of the latter three clones appears to be completely homologous to the 5E1 clone; and (iii) the nucleotide sequences of the coding and non-coding regions are identical to the cDNA counterparts [6] except for four bases differences which do not alter the protein sequence.

We have previously shown that the *Euglena rbcS* gene begins with the 3′ part of an intron (i1) involved in the transfer by trans-splicing of the leader sequence to the SSU mRNA precursor [1]; this structural feature, unknown in plant *rbcS* genes [10,11], enabled us to determine the exact start position of the *Euglena rbcS* gene.

A comparison of the 5′ region of the *Euglena rbcS* gene to the corresponding region of the cDNA shows that two introns (i2 and i3) are present. The first intron, situated in the transit peptide coding sequence, extends from nucleotide 317 to 371 (53 nucleotides). The second intron extends from nucleotide 682 to 1375 (692 nucleotides) and interrupts the region encoding the mature SSU (see Fig. 2). All eucaryotic *rbcS* genes characterized so far contain introns in the coding region only. The number of introns varies from one to three, and they are always located at an invariant position [12–15] except in the green unicellular alga, *Chlamydomonas reinhardii* [16]. In higher plants the first intron is located between codons 2 and 3 of the region coding for the mature SSU, the second between codons 41 and 42, while the last intron occurs downstream within codon 59 [11]. The intron interrupting the coding region of *Euglena* mature SSU is unique, and its position is close to that of the intron located at codon 59 in the higher plant *rbcS* gene (see Fig. 2).

In contrast to all *rbcS* genes sequenced to date the *Euglena rbcS* gene also contains introns in the 3′ non-coding region. The sequence of the 900 nucleotide region corresponding to the 3′ region of the cDNA was determined from the 3G1 sub-clone. This sequence reveals that the untranslated region is split by three introns located, respectively, at 244, 407 and 662 nucleotides downstream from the stop codon. These introns, i1E, i2E and i3E, are, respectively, 81, 215 and 88 nucleotides long. A comparison of the nucleotide sequences of these *Euglena* introns shows no significant homology, even in the 5′ and 3′ border regions. The 3′ border of i1 fits perfectly with the consensus sequence defined for introns of eucaryotic mRNA precursors [17]: the last two bases correspond to the invariant AG dinucleotide, and are preceeded by a stretch of pyrimidines (Fig. 2). This intron is associated with the post-transcriptionally transferred leader sequence of the SSU mRNA, which contains the invariant GT dinucleotide [1]. In contrast, all the other introns of the *rbcS* gene do not obey the so-called GT–AG rule for splice-site selection in eucaryotes. The 5′ and 3′ borders are: CT–TG for i2, AA–CC for iE1, CG–TC for iE2 and AC–CT for iE3. Intron i3 begins with the invariant GT, but it does not have a typical splice site at its 3′ end (CC instead of AG) (Fig. 2). The only common motif, present in only three cases, is a CAGPu sequence found in the vicinity of the 5′ part of the intron. In any case, alignment of the 5′ and 3′ end sequences of the introns failed to define any consensus motif. These regions have been shown to play a crucial role in eucaryotic splicing [18], and it is therefore surprising that the canonical GT–AG intron border sequence is found only in the intron involved in trans-splicing (i1) and not in the other *Euglena rbcS* introns.

In Fig. 3 we show the secondary structures proposed for introns i2, i3, i2E and i3E. In these structures base pairing brings the 5′ and 3′ ends of the introns into proximity for excision. In addition the two successive exons are also involved in base pairing, which contributes to the stability of the structure. Our data clearly indicate that the structure of *Euglena rbcS* introns is very distinctive. Preliminary results on other *Euglena* nuclear genes suggest that this structure is not unique to the *Euglena rbcS* gene (Smith, A., personnal communication and [19]). The structure of *Euglena rbcS* introns differs not only from that of higher plant and alga counterparts [15,16], but more generally from the introns of eucaryotic mRNA precursors which obey the GT–AG rule [16]. Several thousand available sequences obey the

```
                          +1
tctttcaatt tccgcccctc tgcagCACTC TTGCCGGCTC TCATTACGAT GCCATTTGAC CGTCAACCAC TTTTGTCTGG GGAGAAGGGA ATGCCAGCCA CATCTTTGTG   85
                                                     M   P  F  D   R  Q  P   L  L  S  G   E  K  G   M  P  A   T  S  L  W

GCTCGTTGGA GGTGCGGTAA TTGCAGCTGT TTGTGTCATT GTGAACACTT CCTACAATGG AACGCAGCTG TCAGTGACTG CACGTCCAAT TCAGGCAGCC GTTTCACAGG  195
  L  V  G   G  A  V   I  A  A  V   C  V  I   V  N  T   S  Y  N  G   T  Q  L   S  V  T   A  R  P  I   Q  A  A   V  S  Q V

TCTCAATGGC GCGCTTTGCA GAGTCTGGCG TTTCCCGAGG CTCTGGCAAC CGAGTCTCAC AGGCAGTTCC TCTCATGGCT GCATCTGTCG GCGCAGAGAG CGAATCTCGC  305
   S  M  A   R  F  A   E  S  G   V  S  R  G   S  G  N   R  V  S   Q  A  V  P   L  M  A   A  S  V   G  A  E  S   E  S  R

CGTTGGGTTG CGctcagatt acctcttcca gaaattttca aaatgcaaga ggtttcctga aactgAGTGC AATTCTGTTT CCCCTTTCCG GACTGTTTGC TGCCGTGGCT  415
 R  W  V   A                                                            S  A   I  L  F   P  L  S   G  L  F  A   A  V  A

CTCAAAATGG CGATGATGAA GCCTAAGGTG GCTGCCGTCC TCCCTTTTAC ATCAGAGAAG GATATGAAGG TGTGGAACCC CGTCAACAAC AAGAAGTTCG AGACCTTCTC  525
 L  K  M   A  M  M  K   P  K  V   A  A  V   L  P  F  T   S  E  K   D  M  K   V  W  N  P   V  N  N   K  K  F   E  T  F  S
                                                                     ->
CTACCTGCCC CCCCTGTCTG ACGCCCAGAT CGCCAAGCAG GTGGACATGA TCATTGCCAA GGGGCTCTCC CCCTGCCTGG AGTTCGCCGC TCCGGAGAAC AGCTTCATCG  635
  Y  L  P   P  L  S   D  A  Q  I   A  K  Q   V  D  M   I  I  A  K   G  L  S   P  C  L   E  F  A  A   P  E  N   S  F  I A

CCAATGACAA CACCGTGCGC TTCAGTGGCA CCGCTGCGGG CTACTATgtc caagcaatgc cttggtccgc atttgaaaaa gttcagtcaa aagttgccga aaccgtattg  745
   N  D  N   T  V  R   F  S  G   T  A  A  G   Y  Y

gaactcattg acgttgtcgc agaattcctc cattttttgt ggatttttta gtgaagcttt ttcactcatt tgccacatat ttccttggca cgtttttttg tgattttctt  855

ttggttttga aaattgcttc tgaaagatta aggccaaaac gttattgcct tacaattttt ttggttttgt tcagctgacc gatttttcga cgcggcttta gttgcacacc  965

acacgccaca gtctctcccc acactgcccc acagtttagt atggggttca attttgagca gaggacaaaa aaagacaggc agacattcat gatattcaca aaagaaacaa 1075

cgagggcaca cctgttgccc atttgtgcag ttttcaggac gttgttttta aaggagcaaa tcaaagggca ctgtccttac atcagaaaag catattgaga aactgcaaga 1185

agtatcttcc ttgttccaga acaaaagcac cacacaggtc cacaaccacc ccgctacaca cactctacca acaaaaccaa tgacttcaat atggaagctc cactttgcac 1295

cattcatcaa ctctgctgac ctcaattaac tctccagtag tagcaaggag tgagcgtgca ggcagtgcct gccatcccGA CAACCGGTAC TGGACCATGT GGAAGCTGCC 1405
                                                                                      D   N  R  Y   W  T  M   W  K  L  P

CATGTTCGGC TGCACGGACG CCAGCCAGGT CCTGCGCGAG ATCTCCGAGT GCCGCCGGGC CTACCCCCAG TGCTACGTCC GCCTGGCGGC CTTCGACTCC GTCAAGCAGG 1515
  M  F  G   C  T  D   A  S  Q  V   L  R  E   I  S  E   C  R  R  A   Y  P  Q   C  Y  V   R  L  A  A   F  D  S   V  K  Q V

TGCAGGTCAT CTCGTTCGTG GTGCAGCGCC CCTCCGGCAG CAGCAGCAGC AGCTGGGGCA TGGCT..................................................  1580
  Q  V  I   S  F  V   V  Q  R  P   S  G  S   S  S  S   S  W  G  M   A
                                                         <-


......CTCG AGCGGCGGCC GCTCCTGGTA AGGCGGCTTG TGTGTCTGTG TTATTTTTCC TTCGCATCGA GGTTGGGGAT CTCCAGCCTG AGTCCTGAAG CTTCTGAGTC 104E
       S   S  G  G  R   S  W  *
                        <-
CGGCTGTGGC TCCTCGCGAG CAAGCTTGCC GCGGGCACCT TGCTTTCCGA GGAAGGCTGG TTTGTGGATT TCCTCGCGAG CAAGCTTGCC GCGGGCACCT TGCTTTCCGA 214E

GGAAGGCTGG TTTGTGGATT TGTGGTGACC CGCTCCATCC ACCCCATGCT CTCTGTTTTT Taacagccat ctcttctacg gatttgattt tgttgtattc actgtggccg 324E

tggaggtcgc aaggtgaaga gatggttctc ccTTCCATTC ATTCTTGTGT GACATTTGTG CATCACAATG TGACACATAC AGAGGGGGTG GGTCTAGTGC AGCCCCCTAG 434E

ATGAacgcag gtttccatcc attttctgca atcgcgtgat gtactcatgc cagtgtacag ccttcaaact tcttgagaac gttctgagaa cctgatactc gtccggggaa 544E

cttggttgtg tgcatcgtcc cattggccaa catgtcccca caactgtcca acatgtcaat tagtgtttca cggtgttggt tgagtttctg tccgaggatg gaaatctggt 654E

tcGTTGTGAG CGGAATCAAc cTTGATGACA GTTGGGCaga gggatgctgg ttcgcctgtg gccaaaacat ggaaactgtg gaaagttgac gcattgcggt ggtggcacaa 764E

ccagcatccc acactGTTTC AGGTATCATA TTTCTGAACT CATTCATGCC CTCACCcGTG TCAAGGGACA AACAGCAACT CTTGACCACA GTTACCTCAG GCTGACCCTG 874E

TGACTACAGC TACAACTGTG GTTTGTGcac ggggtctgca ttcgtcactg accccacact gtacctgttg acctgttctg tattt                             959E
```

Fig. 2. Nucleotide sequence of the 5′ and 3′ ends of *Euglena rbcS* gene. Sequences found in SSU mRNA [6] are in capital letters. Nucleotides of the 5′ end are numbered from −25 to +1581; the acceptor site of the trans-splicing is indicated by +1. Nucleotides of the 3′ end are numbered from 1E to 958E; the star corresponds to the stop codon. The upstream arrow underlines the first methionine of the mature SSU, while the downstream arrows underline the last tryptophan of the mature SSUs.

GT–AG rule, whereas only 26 described to date show a minor variation (GC) at the invariant GT [20]. Two other exceptions are known in the genes coding for the human proliferating cell nucleolar protein, P120 [21], and for the chicken cartilage matrix protein [22], in which neither GT nor AG are conserved. In the *Euglena rbcS* gene the borders are unusual and, in addition, the secondary structure of these introns is distinctive. Analysis of the nucleotide sequences of introns i2 and i3 shows that these introns do not resemble either group I or group II self-splicing introns [23] (Michel, F., personnal communication), suggesting that they belong to a novel intron class.

The presence in the *Euglena rbcS* gene of an intron with a typical AG 3′ border (intron i1), and of introns with particuliar features, indicates that two different splicing mechanisms could be involved in the maturation of the *Euglena* SSU mRNA precursor. One mechanism, carried out by a spliceosomal complex [24], would transfer the leader sequence to the 5′ end of the pre-mRNA. Such a mechanism has been described for mRNA precursor processing in Trypanosomes [25]. A second mechanism would be involved in the removal of introns located in the internal region of the *Euglena rbcS* gene. One can wonder whether these splicing events require ribonucleoprotein complexes, or whether

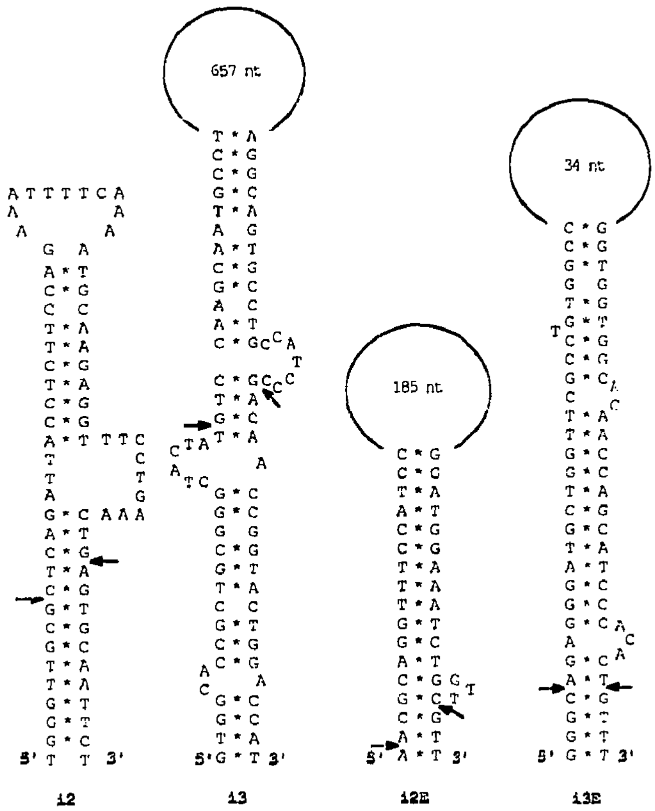

Fig. 3. Proposed secondary structures for four introns present in the *Euglena rbcS* gene. Sequences involved in base pairing are indicated by stars. Arrows indicate the splice sites.

a distinct enzymatic system, as described for the splicing of introns in nuclear tRNA precursors, could be involved [26].

Acknowledgements: We wish to thank Mrs. Hoeft for the synthesis of the oligonucleotide. R.L.C. was supported by a grant from CONICET-Argentina.

REFERENCES

[1] Tessier, L.-H., Keller, M., Chan, R.L., Fournier, R., Weil, J.-H. and Imbault, P. (1991) EMBO J. 10, 556–601.

[2] Keller, M., Tessier, L.-H., Chan, R.L., Weil, J.-H. and Imbault, P. (submitted).

[3] De Lange, T., Liu, A.Y.C., van der Ploeg, L.H.T., Borst, P., Tromp, M.C. and van Boom, J.H. (1983) Cell 34, 891–900.

[4] Krause, M. and Hirsh, D. (1987) Cell 49, 753–761.

[5] Agabian, N. (1990) Cell 61, 1157–1160.

[6] Chan, R.L., Keller, M., Canaday, J., Weil, J.-H. and Imbault, P. (1990) EMBO J. 9, 333–338.

[7] Smith, S.M. and Ellis, R.J. (1979) Nature 227, 680–685.

[8] Dunn, I.S. and Blattner, F.R. (1987) Nucleic Acids Res. 15, 2677–2698.

[9] Sanger, F., Nicklen, S. and Coulson, A.R. (1977) Proc. Natl. Acad. Sci. USA 74, 5463–5467.

[10] Dean, C., Pichersky, E. and Dunsmuir, P. (1989) Annu. Rev. Plant. Physiol. 40, 415–439.

[11] Manzara, T. and Gruissem, W. (1988) Photosynth. Res. 16, 117–138.

[12] Lebrun, M., Waksman, G. and Freyssinet, G. (1987) Nucleic Acids Res. 15, 4360–4364.

[13] Wimpee, C.F., Stiekema, W.J. and Tobin, E.M. (1983) in: Plant Molecular Biology, pp. 391–401, Liss, New York.

[14] Grandbastien, M.A., Berry-Lowe, S., Shirley, B.W. and Meagher, R.B. (1986) Plant Mol. Biol. 7, 451–465.

[15] Sugita, M., Manzara, T., Pichersky, F., Cashmore, A. and Gruissem, W. (1987) Mol. Gen. Genet. 209, 247–256.

[16] Goldschmidt-Clermont, M. and Rahire, M. (1987) J. Mol. Biol. 191, 421–432.

[17] Chambon, P. and Breathnach, R. (1981) Annu. Rev. Biochem. 50, 349–383.

[18] Wieringa, B., Meyer, F., Reiser, J. and Weissmann, C. (1983) Nature 301, 38–43.

[19] Muchhal, U.S. and Schwartzbach, S.D. (1991) 3rd Int. Congress Soc. Plant Mol. Biol., Tucson, 6–10 October, 1991, Poster Abstract 232.

[20] Jackson, I.J. (1991) Nucleic Acids Res. 19, 3795–3798.

[21] Larson, R.G., Henning, D., Haidar, M.A., Jhiang, S., Lin, W.L., Zhang, W.N. and Bush, H. (1990) Cancer Commun. 2, 63–71.

[22] Kiss, I., Deak, F., Holloway, R.G., Delius, H., Mebus, K.A., Frimburger, E., Argroves, W.S., Tsonis, P.A., Winterbotton, N. and Goetinck, P.F. (1989) J. Biol. Chem. 264, 8126–8134.

[23] Michel, F. and Dujon, B. (1983) EMBO J. 2, 33–38.

[24] Luhrman, R., Kastner, B. and Bach, M. (1990) Biochim. Biophys. Acta 1087, 265–292.

[25] van der Ploeg, L.H.T., Liu, A.Y., Michels, P.A.M., de Lange, T., Borst, P., Majumder, H.K., Weber, J., Veeneman, G.H. and van Boom, J. (1982) Nucleic Acids Res. 10, 3591–3604.

[26] A.R. Krainer and T. Maniatis (1988) in: Transcription and Splicing (B.D. Hames and D.M. Glover, Eds.), pp. 131–136, IRL Press.