An intron in a ribosomal protein gene from *Tetrahymena*

Henrik Nielsen, Per Hove Andreasen[1], Hanne Dreisig[1], Karsten Kristiansen[1] and Jan Engberg

Department of Biochemistry B, Panum Institute, University of Copenhagen, Blegdamsvej 3C, DK-2200 Copenhagen N, and [1]Department of Molecular Biology, University of Odense, Campusvej 55, DK-5230 Odense M, Denmark

Communicated by R.Garrett

We have cloned and sequenced a single copy gene encoding a ribosomal protein from the ciliate *Tetrahymena thermophila*. The gene product was identified as ribosomal protein S25 by comparison of the migration in two-dimensional polyacrylamide gels of the protein synthesized by translation *in vitro* of hybrid-selected mRNA and authentic ribosomal proteins. The proteins show strong homology to ribosomal protein S12 from *Escherichia coli*. The coding region of the gene is interrupted by a 979-bp intron 68 bp downstream of the translation start. This is the first intron in a protein encoding gene of a ciliate to be described at the nucleotide sequence level. The intron obeys the GT/AG rule for splice junctions of nuclear mRNA introns from higher eukaryotes but lacks the pyrimidine stretch usually found in the immediate vicinity of the 3′ splice junction. The structure of the intron and the fact that it is found together with the well described self-splicing rRNA intron is discussed in relation to the evolution of RNA splicing.

Key words: intron/ribosomal protein/*Tetrahymena*

Introduction

Recent studies of gene structure and gene expression in ciliates have revealed several remarkable findings. To give a few examples, Cech and co-workers have demonstrated one of the most forward examples of RNA molecules harboring catalytic activity through their studies of self-splicing of the rRNA intron in *Tetrahymena* (for review, see Cech, 1985). Klobutcher *et al.* (1984) have shown that DNA splicing events occur during the maturation of a macronuclear gene in *Oxytricha* and, more recently, it has been shown that the ciliates use a variant of the universal genetic code in which the UAA and UAG triplets encode amino acids rather than translational stops. This phenomenon has been reported for the G surface antigen in *Paramecium primaurelia* (Caron and Meyer, 1985), the A immobilization antigen in *Paramecium tetraurelia* (Preer *et al.*, 1985), the α-tubulin gene in *Stylonichia lemnae* (Helftenbein, 1985) and the histone H3 (Horowitz and Gorovsky, 1985) and actin (Cupples and Pearlman, 1986) genes of *Tetrahymena thermophila*. Very few complete protein encoding genes have been described in detail from the ciliates and not one single one containing an intron. To gain a better understanding of the apparent uniqueness of the ciliates, it becomes important to characterize more protein encoding genes from this group. One very obvious source would be the ribosomal protein genes. We have initiated a study of the structure of these genes in *Tetrahymena* with special emphasis

on structures involved in their coordinative regulation (Dreisig *et al.*, 1984a, 1984b; Andreasen *et al.*, 1984). We report here the complete structure of a ribosomal protein gene from *T. thermophila* containing an intron seemingly of the type found in protein encoding genes from higher eukaryotes. Furthermore, we provide evidence that the gene product is homologous to *Escherichia coli* ribosomal protein S12.

Results

Selection of sequences coding for ribosomal proteins from an enriched cDNA library

During nutritional shift-up in *T. thermophila*, the absolute and the relative amounts of ribosomal protein mRNAs are increased compared with starved cells by a factor of 50 and 6, respectively (Andreasen *et al.*, 1984). This phenomenon was exploited to construct a cDNA library enriched in ribosomal protein gene sequences. The cDNA was synthesized using mRNA from upshifted cells as the template and cloned into the *Pst*I site of plasmid pBR322 by the G and C tailing method (Villa-Komaroff *et al.*, 1978). Individual cDNA clones containing inserts of more than some 100 bp were used to hybrid select mRNA from a mRNA preparation from upshifted cells. The selected mRNA was subsequently translated *in vitro* using a nuclease-treated rabbit reticulocyte lysate system and the resulting protein analysed by co-electrophoresis with authentic ribosomal proteins in a two-dimensional polyacrylamide gel. The evidence for the identification of one such cDNA clone, prB2, is presented in Figure 1. Migration of the protein in the gel system used in these experiments indicates that the protein is basic and therefore provides a first hand indication of its relationship to ribosomal proteins (Dreisig *et al.*, 1984a). More importantly, the protein
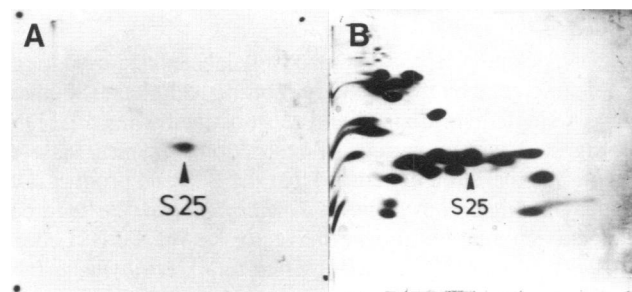
Fig. 1. Identification of the cDNA clone prB2. Linearized plasmid DNA was immobilized on a nitrocellulose filter and hybridized to poly(A)⁺ RNA from *T. thermophila*. The hybrid-selected mRNA was melted off and translated *in vitro* in a nuclease-treated rabbit reticulocyte lysate in the presence of [^{35}S]methionine. The radioactive translation product was co-electrophoresed with 50 μg of unlabelled ribosomal proteins in the pH 8.6/SDS-system (Dreisig *et al.*, 1984a). (**A**) Fluorogram showing the translation product obtained by translation *in vitro* of the mRNA hybridizing to prB2. (**B**) Coomassie-stained gel showing the proteins from the small subunit of *T. thermophila* ribosomes. The indicated ribosomal protein S25 (a comprehensive nomenclature is to be published elsewhere) co-migrates with the fluorographically detected translation product in (**A**).
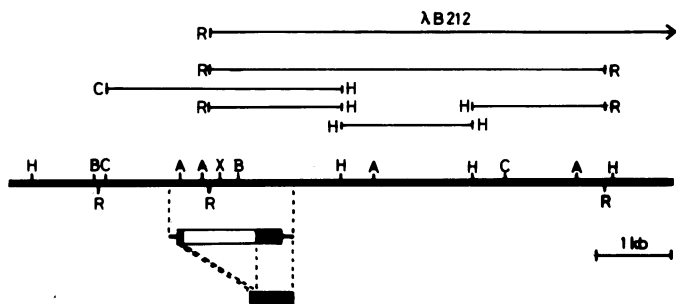
Fig. 2. Restriction enzyme map of the genomic region containing the gene coding for *T. thermophila* ribosomal protein B2 (heavy line). Above the genomic map are shown the various cloned DNAs relevant to this study. From the top is shown a λ-clone containing a 14-kb partial *Eco*RI fragment of genomic DNA. From this clone, subclones of 1.8-kb *Hin*dIII−*Eco*RI (pHRG), 1.8-kb *Hin*dIII−*Hin*dIII (pHHG2) and 1.8-kb *Hin*dIII−*Eco*RI (pHHG1) all in pUC18 were derived. The *Cla*I−*Hin*dIII fragment was cloned independently into the vector pACYC184 (pCH4) using size-selected DNA from a restriction enzyme digest of total DNA. A fragment of this was subcloned into pUC18 (pCR12 not shown). The gene map was constructed by the Southern blotting technique using total DNA and pCR12, pHRG, pHHg1 and pHHG2 as hybridization probes. Below the genomic map is shown the transcribed region including exons (filled boxes), intron (open box) and non-translated regions. The extent of the sequence found in the cDNA clone prB2 is also shown. Restriction enzyme codes are: A, *Ava*II; B, *Bgl*II; C, *Cla*I; E, *Eco*RI; H, *Hin*dIII; X, *Xba*I.

studied in this case co-migrated with the small subunit protein S25. In general, translation *in vitro* of *Tetrahymena* mRNA gives rise to small, apparently truncated protein products, probably because of differences in codon usage between the rabbit and *Tetrahymena*, in particular the unusual use of UAA and UAG as amino acid coding triplets rather than translational stops in *Tetrahymena*. A similar blocking of translation *in vitro* in the rabbit reticulocyte lysate has been reported for mRNA from *Paramecium* (Meyer *et al.*, 1984). We have found that when the reticulocyte lysate is supplemented with a subcellular fraction prepared from *Tetrahymena* cells a much higher fraction of the *in vitro* synthesized proteins co-migrate with the native proteins (P.H.Andreasen *et al.*, in preparation). However, in the case of prB2, the unusual use of stop codons is not expected to influence the identification of the protein product because the gene contains no UAA or UAG codons (see later).

Cloning of the single copy B2 gene

The cDNA clone prB2 was radioactively labeled and used to map the genomic gene by the Southern blotting technique (Southern, 1975). A single hybridizing band was obtained when total DNA was digested with a variety of restriction enzymes (data not shown). We therefore concluded that the B2 gene product is encoded by a single copy gene in *T. thermophila*. We then constructed a genomic library, in the vector λgtWESλ5B (Leder *et al.*, 1977), of partially *Eco*RI-digested total *Tetrahymena* DNA size selected to 10−14 kb on a sucrose gradient. Using prB2 as a probe, 400 000 recombinant plaques were screened which corresponds to 3−4 complete libraries. Twelve hybridization-positive plaques were obtained, plaque purified through a second plating and characterized by restriction enzyme mapping. Surprisingly, none of the clones contained sequences corresponding to the 5′ end of the cDNA. We therefore designed a strategy to clone a specific restriction enzyme fragment containing this DNA. Total DNA was digested with *Hin*dIII and fractionated on a sucrose gradient. The fraction containing the gene, as deduced from the above-mentioned Southern blot, was further digested with *Cla*I and cloned into the *Hin*dIII and *Cla*I sites of

pACYC184. The resulting clones were screened by colony hybridization (Grunstein and Hogness, 1975) using the 3′ end of the gene as probe (pHRG, Figure 2). In this way, a restriction enzyme fragment containing the entire gene was obtained. With these clones at hand, extensive mapping of the gene was carried out by Southern blotting of total DNA, supporting our previous conclusion that only a single B2 gene is present in the *T. thermophila* genome (Figure 2, experimental data not shown). This low copy number of a gene coding for a ribosomal protein is similar to the situation in yeast in which most, if not all, ribosomal protein genes are found in one or two copies per haploid genome equivalent (Leer *et al.*, 1985). However, the presence of only one copy of the B2 gene is not universal among the *Tetrahymenas*. In a screening of six different species, cases of two and perhaps three copies of the B2 gene were found (unpublished results).
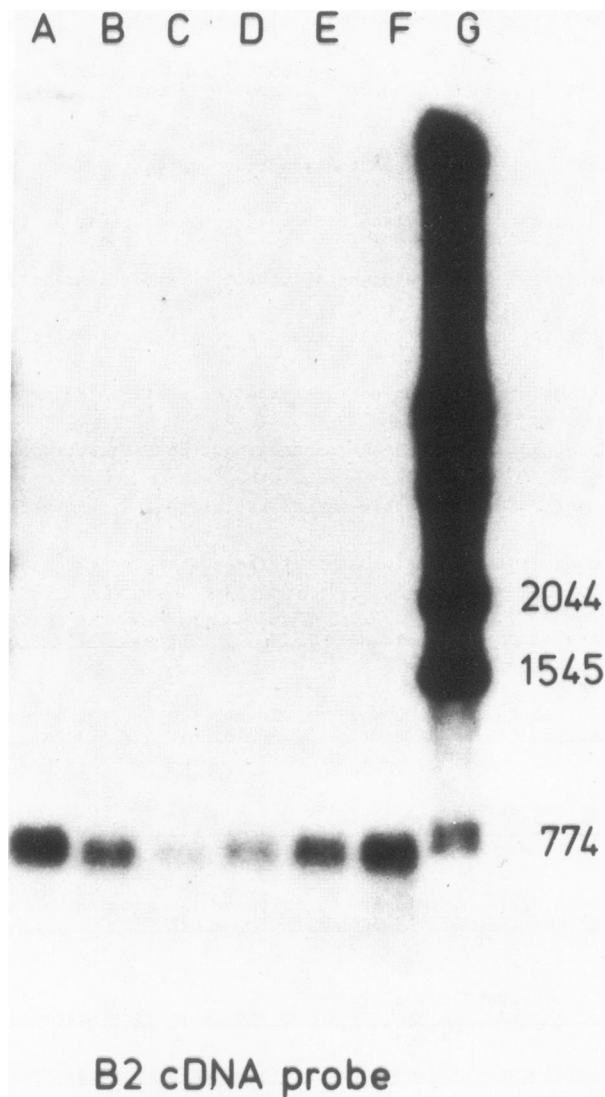
Sequence analysis

The DNA sequence of the cDNA and the genomic copy of the gene is given in Figure 3. From a comparison of these two sequences it is evident that the gene is composed of two exons containing 68 bp and 358 bp, respectively, of coding region and split in an arginine codon by a 979-bp intron. The AT content of coding regions is 56% compared with >80% for the immediately flanking spacers and 76% for the intron. These high numbers reflect a very high overall AT content of the *T. thermophila* genome of 75% (Allen and Gibson, 1973).

The region upstream of the predicted initiation codon is strikingly asymmetric with respect to the base composition of the two DNA strands. The RNA-like strand is extremely purine rich. This is especially pronounced for the transcribed part of the 5′ flanking region (for mapping of the initiation site, see later) in which 56 out of 72 nucleotides are purines (77%), 49 (68%) being As. This A-richness of the immediate 5′ region is also observed in the actin gene (39 out of 56 nucleotides) and the histone H4 gene (32 out of 45 nucleotides). The significance of this asymmetry is unknown but it certainly has implications for the secondary structure of the 5′ end of the mRNA. Hairpins are very unlikely to be formed and stacking interactions between neighbouring purines expected to dominate. We have not been able to identify any larger sequence element common to the 5′ ends of the known *Tetrahymena* protein encoding genes.

The presence of a poly(A) tail on the cloned cDNA was used to define the sense strand and to locate the precise site of the polyadenylation which is at a position 139 nucleotides downstream of the UGA stop codon. This distance seems to be very well conserved among ribosomal protein genes from *Tetrahymena* (unpublished results). A repeat sequence, $(TPyAA)_6$, is found in the B2 gene around the site of polyadenylation. The pyrimidine is a T in the first two repeats and a C in the last four repeats. Only three of the repeats are transcribed so that polyadenylation in this case takes place at the site where a T is substituted by a C in the repeat. We have not found this sequence motif in other *Tetrahymena* genes.

The sequence surrounding the AUG start codon conforms to the consensus sequence derived from translation initiation in higher eukaryotes, CC(A/G)CCAUG(G) (Kozak, 1984), in the sense that the most conserved nucleotide, the purine at −3 (most often an A) and the G in position +4 is found in the B2 gene of *Tetrahymena*. It is noteworthy that all protein encoding genes reported from *Tetrahymena* thus far have a triple A stretch preceding the AUG triplet. For a discussion of other aspects of the coding sequence, see later.

```
AGATAAATTAGAATGAAAAGAAGAAATAGTATAAAGAAAAGAAATATATAAAGGGGAAATTTAATGAAGTCTAATTAAATAAATGTATGTGTATATGGTTTGAATATGCATTTGTATTT
                            -150                                                                    -100
```

↓

```
                                                                                      Met GlyⱯVal Gly LysↃPro Arg G
AAGATACTTAAAAGCGGTGTGGAAGATTAAAATTAAAAGAAAACAAAATAATTAAAGTAAAAAATATAAAAAAGAAAATAAAAATCAAA ATG GGT GTT GGT AAA CCT AGA G
              -50                                                                             1
```

```
ly Ile Arg Ala Gly Arg Lys Leu Ala Arg His Arg Lys Asp Gln Arg
GT ATT AGA GCC GGT AGA AAA TTG GCT AGA CAC AGA AAG GAC CAA AG gtgaaatattttttaataacaatcaatctacagcaagcatgaatctgaggagtgtt
                                                                                 100
```

```
gcatcaaagctatagcaatccttctctttattttcaaaacgctctcaaatgaatgatgtatcaattacaaatccattccattacattccccaattttgacaaacatttaaaaattatt
                                    200
```

```
tttattcattctcatcaattttgttaaatgcatctaattaatgtcttcaaccaacaattctactaagatataatcgtttctctgcgcttagtttgctttatccttcttacggacctatt
                               300
```

```
gaaaatattttgtatatagctcaaaaagccaaataatctttaataaataaaagcagacagattttgttttatcttttgagttagaattctaaagtttctagctaactttcaatcaacaa
        400
```

```
agctgttgcaaagatttaagaaatagaataaaaagcaaacccctaaattagattatttgaaataattaggtttttatctagaatgattgattgaaattctcagattgaatattaatcga
      500                                                                                               600
```

```
tcaaattacaaaaattatgaactatatataaaatagaagttaaatagaggagactatgtataacaaaagcagcaataaataaataaatgaacaaatggtttgataaagcatttagtaaa
                                    700
```

```
attaaagagcattaacgctttataacgaatttgaaaaatgaatgtttaaaaatgaataagcaaataggtttgttgtttgtaatttccaaataaattttacttgaaatttttgtcttttg
                                800
```

```
aaagaaagatctcaactaactttaacaagcatataaactaataaagattttttgcttaatttatttcgaggatggatggataataaaatatcatcttaaagataagcggatatgaaaaca
                                 900
```

```
                                                                                               Trp Ala Asp Asn Asp Phe Asn
taaacaaattaataaaaaaataaatgattgcttgccatta┌ttctcttcctct┐aaaataaatttattaaattcaattttaaataaaaag A TGG GCT GAT AAC GAT TTC AAC
                                         1000
```

```
Lys Arg LeuⱯLeu Gly Ser Arg Trp ArgↃAsn Pro Phe Met Gly Ala Ser His Ala LysↃGlyⱯLeu Val Thr Glu Lys Ile Gly Ile Glu Ser
AAG AGA CTT TTG GGC TCC AGA TGG AGA AAT CCT TTT ATG GGT GCT TCT CAC GCT AAG GGT TTA GTC ACT GAA AAG ATC GGT ATT GAA TCT
                   1100
```

```
Lys Gln Pro Asn Ser AlaⱯVal Arg LysↃCysⱯVal ArgↃVal Leu Leu Arg Lys Asn Ser Lys LysↃIle Ala Ala PheⱯVal Pro Met Asp Gly
AAG CAA CCT AAC TCT GCC GTC AGA AAG TGT GTT AGA GTT TTG CTC AGA AAG AAC TCC AAG AAG ATC GCT GCT TTC GTT CCT ATG GAT GGT
                   1200
```

```
CysⱯLeu Asn Phe Leu Ala Glu Asn Asp Glu Val Leu Val Ala Gly Leu Gly ArgↃGln Gly His Ala Val Gly Asp Ile Pro Gly Val Arg
TGC CTT AAC TTC TTA GCT GAA AAC GAC GAA GTC TTG GTT GCT GGT CTT GGT AGA CAA GGT CAC GCC GTC GGT GAT ATT CCG GGT GTT AGA
                   1300
```

```
Phe LysↃValⱯVal CysⱯVal LysↃGly Ile Ser Leu Leu Ala Leu Phe Lys Gly Lys Lys Glu Lys Arg  *
TTC AAA GTT GTT TGT GTC AAG GGT ATT TCT CTC TTG GCT CTT TTC AAG GGT AAG AAG GAA AAG CGT TGA GCTGCATATTAATAATAATACTTCCAA
                   1400
```

```
TAAATTAAATAAAAAGAGTATTTGCCTTAATATATATTATTAACTGTAAATGAAAATATATAGTAATTATTTAAATTTACTCTCTCAAACATCTTTTTAAATTAATTAATCAATCAATC
                                1500
```

```
AATCAATTCATAGTTAAATCCTTCTCATTTTTAAATACTTTTCATATATCTCAATACATACATTTGTACCATTTATATTTTTTATAGTTTAACAACAATAAAAACAATAAAAAAGAATA
                   1600
```

Fig. 3. The complete DNA sequence of the genomic inserts from plasmids pCR12 and pHRG (using pCH4 for overlap) and the insert from the cDNA clone prB2 was determined by the Maxam—Gilbert method. Both strands were sequenced. The gene and part of the flanking sequences is displayed. The coding regions are shown in upper case letters with amino acids in accordance with the universal genetic code. The intron sequence is in lower case letters. The sequence also found in the cDNA is underlined. The following features are indicated in the figure. (1) Major site of transcription initiation as evidenced by primer extension analysis. Boxed sequence: pyrimidine stretch possibly involved in splicing of the intron. Predicted cleavage sites in the protein sequence for peptidases thermolysin (Y) and trypsin (1). These sites were used for the identification of the B2 gene product through a comparison of the fragment pattern generated by digestion of the native protein as identified on two-dimensional gels (Figure 1) with that expected from the gene sequence (to be published elsewhere).

The most significant feature of the B2 gene is the presence of an intron in the 5' end of the coding region. The splice junctions could not be determined unequivocally from a comparison of the cDNA and genomic sequences because the sequence AAAG occurs at both splice junctions. However, only when positioned as shown in Figure 3, the intron obeys the GT/AG rule of eukaryotic introns. The B2 intron conforms very well to the extended consensus sequence at the 5' splice site (A/C)AG⌷GT (A/G)AGT (Mount, 1982) but poorly to the consensus at the 3' splice site (C/T)ₙN(C/T)AG⌷G, *op.cit.*). Most interestingly, a stretch of As is found at the 3' end of the intron instead of the pyrimidine stretch found in all introns except those of type I (Michel and Dujon, 1983). Perhaps the 13 pyrimidines found 35 nucleotides upstream of the 3' splice site is a substitute for this pyrimidine stretch. Several putative branch points (Keller,

1984) can be found in the intron, but none close to the 3' splice junction. The intron does not contain the sequence elements characteristic of type I or type II introns (Michel and Dujon, 1983). The largest open reading frame within the intron could, allowing for translation of UAA and UAG triplets, encode a protein of 87 amino acids. However, this reading frame is not characteristically biased in its codon usage.

Transcription analysis

We have previously shown that the relative as well as the absolute amount of mRNA coding for ribosomal proteins is regulated through a starvation—refeeding cycle (Andreasen *et al.*, 1984). This is also the case for the transcript homologous to the prB2 probe as can be seen from the Northern blot in Figure 4. In this case RNA from an equal number of cells from various

Fig. 4. Northern blot analysis of prB2-specific transcripts. Total RNA was extracted from 40×10^5 cells at various time points during a starvation−refeeding cycle, separated on a 1.4% agarose gel containing formaldehyde and transferred to a nitrocellulose filter. The blot was probed with the cDNA probe prB2 nick-translated to a specific activity of 2×10^7 d.p.m./µg. **Lanes A−F**: RNA isolated from exponentially growing cells (**A**), cells starved for 2 h (**B**) and 24 h (**C**), and cells 0.5 h (**D**), 1.0 h (**E**) and 2.0 h (**E**) after refeeding. **Lane G**: DNA size markers.

time points throughout the starvation−refeeding cycle were applied to the gel. A 40-fold decrease in B2 transcript was observed during the starvation period and a subsequent 40-fold increase relative to this level upon refeeding. A similar result was obtained when equal amounts of RNA were applied to the gel (data not shown). Thus the transcription pattern of the B2 gene is in accordance with it being a ribosomal protein gene.

The Northern blot analysis showed evidence of a single B2 gene transcript with a size of ~700 nucleotides. In order to map the 5′ end of this transcript we carried out both primer extension and nuclease protection experiments. For the primer extension experiment a synthetic oligonucleotide complementary to a part of the 5′ exon was radioactively labeled, annealed to poly(A)$^+$ RNA isolated from upshifted cells and extended by reverse transcriptase in the presence of unlabeled dNTPs. The extension product was analyzed on 8% denaturing polyacrylamide
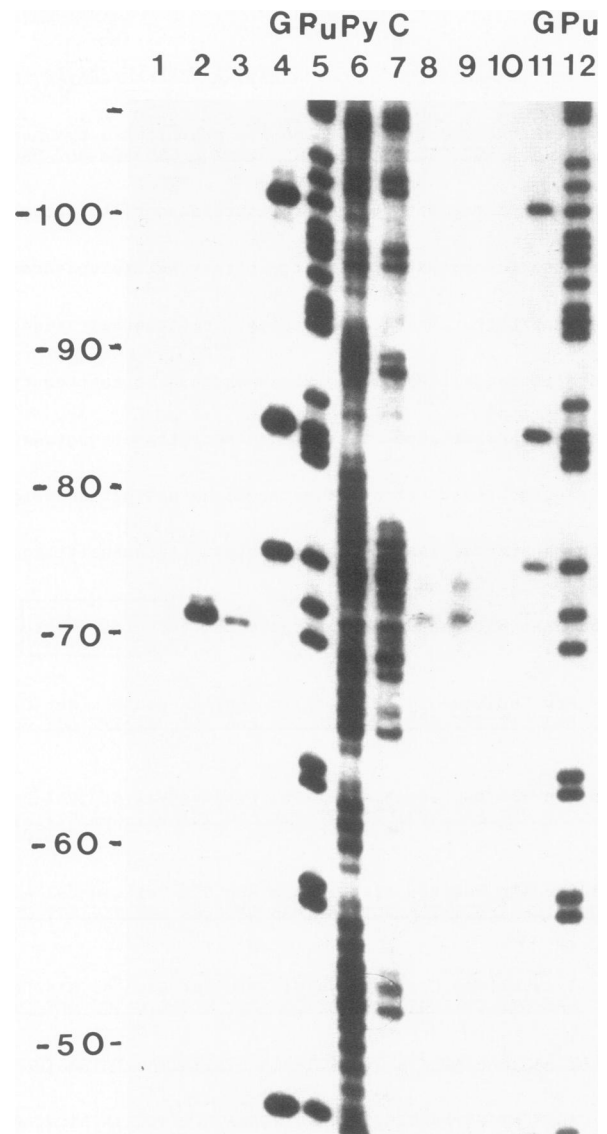
Fig. 5. Mapping of the transcription initiation site by primer extension and nuclease protection analysis. A synthetic 18-mer, 5′-GGCTCTAATACC-TCTAGG-3′, complementary to position 16−33 of the coding region of the B2 gene (Figure 2) was labeled at the 5′ end using T4 polynucleotide kinase and [γ-^{32}P]ATP. The labelled oligonucleotide was annealed to poly(A)$^+$ RNA isolated from upshifted cells and extended by AMV reverse transcriptase in the presence of unlabeled dNTPs. The extension products were analysed on an 8% denaturing polyacrylamide gel in parallel with the samples from the nuclease protection experiment and Maxam−Gilbert sequence ladder of the 5′ end-labeled restriction fragment used as probe in the protection experiment. For this experiment a 5′ end-labeled HpaII fragment was hybridized to poly(A)$^+$ RNA or total cytoplasmic RNA and subsequently treated with Mung bean nuclease (cf. Materials and methods). **Lane 1**: primer extension with no primer added (control for self-priming). **Lane 2**: primer extension with annealing at 65°C for 10 min and incubation with reverse transcriptase for 15 min at 37°C and 80 min at 43°C (see Materials and methods for details). **Lane 3**: similar experiment with annealing at 65°C for 15 min and 37°C for 15 min followed by incubation with reverse transcriptase for 15 min at 37°C and 45 min at 43°C. **Lanes 4−7, 11** and **12**: Maxam−Gilbert sequence ladder. Resistant fragments from nuclease protection experiments using poly(A)$^+$ RNA (**lane 8**), total cytoplasmic RNA (**lane 9**) or no RNA (**lane 10**). Note that the primer was synthesized one nucleotide shorter at its 5′ end than the sequenced strand. For this reason, the length of the primer extension products can be read directly from the sequencing ladder without considering the one nucleotide faster migration of Maxam−Gilbert fragments relative to their sequences. In contrast, the size of the nuclease resistant fragments should be corrected with −1 nucleotide.

gels using a sequence ladder as size marker (Figure 5, lanes 1—7). A single prominent band was observed (Figure 5, lanes 2 and 3) corresponding to a transcript starting at position T_{-70} relative to the translation start codon (cf. Figure 3). A few weaker bands were observed at A_{-150} and A_{-152} but these are believed to represent extension products from irrelevant mRNAs, as no corresponding bands were detected in the nuclease protection experiment. The probe used for the nuclease protection experiment was a restriction fragment 5' end-labeled at the *Hpa*II site at position +32 and extending to a *Dra*I site 500 bp upstream. Following hybridization to preparations of poly(A)$^+$ or total cytoplasmic RNA and digestion with Mung bean nuclease, the protected fragments were size fractionated on a sequencing gel (Figure 5, lanes 8—10). Two closely spaced, prominent bands were detected (Figure 5, lanes 8 and 9) the smaller of which corresponded exactly in size (after proper mobility corrections, see legend to Figure 5) to the extension product of the primer extension experiment. The results from the two independent mapping procedures support each other and suggest the transcription start site to be at a position corresponding to position T_{-70}. Transcription from this position would give rise to a spliced mRNA product of close to 700 nucleotides in size which is in very good agreement with the Northern blotting data. It has not been excluded that larger but unstable B2 transcripts exist since the techniques used will only detect stable intermediates.

The relationship between the TATA-box and the transcription site in *Tetrahymena* is far from settled. In the B2 gene, no obvious TATA-box is found at the eukaryotic consensus distance of 30—40 bp from T_{-70}. The situation is similarly unclear in the case of the actin gene (Cupples and Pearlman, 1986). Considering the fact that this distance can be experimentally manipulated without effect on the initiation site in an organism that otherwise conform to the consensus (yeast, Chen and Struhl, 1985) it is premature to make a conclusion on this issue.

The B2 gene product

The B2 gene can be translated into a 142-amino acid protein of mol. wt 16 000 containing 23% basic (lysine + arginine) residues. The reading frame contains no UAA or UAG triplets believed to encode glutamine and glutamic acid or glutamine, respectively, in the ciliates (Caron and Meyer, 1985; Preer et al., 1985; Helftenbein, 1985; Horowitz and Gorovsky, 1985). The codon usage, tabulated in Figure 6 is biased in the same way as reported for the histone H3-I, H3-II (Horowitz and Gorovsky, 1985) and H4 (Bannon et al., 1984) and the actin gene (Cupples and Pearlman, 1986) from T. *thermophila*. Apparently there is a preference for a pyrimidine in the third position of a codon in cases where the amino acid and the first two positions are given. The use of the rare arginine codon CGU as the last codon in the B2 reading frame is noteworthy. This is the first reported use of this particular codon from >2000 ciliate codons sequenced thus far (126 of these are arginine codons, 122 being AGA). Only 35 of the possible 64 codons are used in the B2 reading frame. This restricted codon usage was also reported in the case of the other *Tetrahymena* protein encoding genes and has led to the speculation that some codons have been completely lost in the ciliates (Syvanen, 1986).

As shown in Figure 1, we have identified the B2 gene product as ribosomal protein S25. In order to identify this protein in terms of its *E. coli* equivalent, we made a computer comparison of the sequence of the *T. thermophila* S25 ribosomal protein, as translated from the gene sequence, and all *E. coli* ribosomal proteins. The computer search clearly identified the small subunit

1	UUU	*Phe*	4	UCU	*Ser*	0	UAU	*Tyr*	2	UGU	*Cys*
5	UUC	*Phe*	2	UCC	*Ser*	0	UAC	*Tyr*	1	UGC	*Cys*
2	UUA	*Leu*	0	UCA	*Ser*	0	UAA	*Ter*	1	UGA	*Ter*
5	UUG	*Leu*	0	UCG	*Ser*	0	UAG	*Ter*	2	UGG	*Trp*
4	CUU	*Leu*	4	CCU	*Pro*	0	CAU	*His*	1	CGU	*Arg*
2	CUC	*Leu*	0	CCC	*Pro*	3	CAC	*His*	0	CGC	*Arg*
0	CUA	*Leu*	0	CCA	*Pro*	3	CAA	*Gln*	0	CGA	*Arg*
0	CUG	*Leu*	1	CCG	*Pro*	0	CAG	*Gln*	0	CGG	*Arg*
4	AUU	*Ile*	1	ACU	*Thr*	1	AAU	*Asn*	0	AGU	*Ser*
2	AUC	*Ile*	0	ACC	*Thr*	6	AAC	*Asn*	0	AGC	*Ser*
0	AUA	*Ile*	0	ACA	*Thr*	3	AAA	*Lys*	14	AGA	*Arg*
3	AUG	*Met*	0	ACG	*Thr*	14	AAG	*Lys*	0	AGG	*Arg*
8	GUU	*Val*	9	GCU	*Ala*	4	GAU	*Asp*	15	GGU	*Gly*
5	GUC	*Val*	3	GCC	*Ala*	2	GAC	*Asp*	1	GGC	*Gly*
0	GUA	*Val*	0	GCA	*Ala*	5	GAA	*Glu*	0	GGA	*Gly*
0	GUG	*Val*	0	GCG	*Ala*	0	GAG	*Glu*	0	GGG	*Gly*

Fig. 6. Codon usage table. The exon sequences from Figure 2 was translated using the universal genetic code. A total of 143 codons including the UGA-stop codon is tabulated. The codon usage is characteristic of *Tetrahymena* as is the restricted use of codons, in this case only 35 out of the 64 codons (see text for discussion).
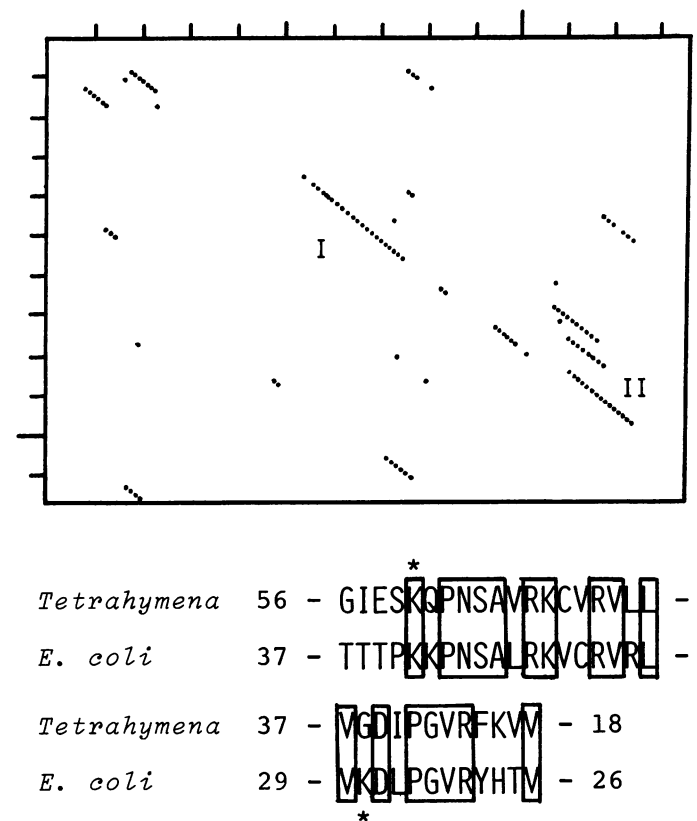

Tetrahymena	56 —	GIES K PNSA RK CVRV
E. coli	37 —	TTTP KK PNSA RK VCRVR
Tetrahymena	37 —	MGDIPGVRFK VM — 18
E. coli	29 —	MKDLPGVRYHTM — 26

Fig. 7. Comparison of ribosomal proteins S25 from *T. thermophila* and S12 for *E. coli*. The top figure shows a dot-matrix comparison of the proteins over their entire lengths. Horizontal axis: 142-amino acid (including initiating methionine) *T. thermophila* S25 protein deduced from the DNA sequence (Figure 2). Vertical axis: 123-amino acid (excluding initiating methionine) *E. coli* S12 protein (Post and Nomura, 1980). The regions containing amino acids 42 and 87 in the *E. coli* sequence (I and II) are aligned in the bottom figure. These amino acids (indicated by asterisks) are changed in most streptomycin-resistant mutants of *E. coli* (Funatsu and Wittman, 1972). Identical amino acids in the two sequences are boxed. Most of the non-identical pairs in the two regions are neutral substitutions. The numbers refer to the length of the sequences not displayed.

2715

E. coli protein S12 as being the most homologous to the *Tetrahymena* S25 protein. In Figure 7 is shown a dot-matrix comparison of the two proteins over their entire lengths. Most interestingly, two of the regions of best homology encompass the amino acids 42 and 87 in the *E. coli* protein, known to be the sites of change in most streptomycin-resistant mutants (Funatsu and Wittmann, 1972). Several amino acids in these regions are identical in the two sequences (Figure 7, lower part).

Discussion

Until recently, the presence of introns in protein encoding genes in the ciliates was an open question. Then Martindale *et al.* (1986) described a conjugation-specific gene containing numerous introns based on comparison of restriction enzyme maps of a cDNA and a genomic clone. Our finding of an intron in the gene coding for ribosomal protein S25 in *Tetrahymena* demonstrates for the first time the sequence of a ciliate mRNA intron and shows that the GT/AG consensus rule for splice junctions extends to this group, one of the earliest groups to diverge from the eukaryotic mainstream during evolution. The fact that the *Tetrahymena* intron conforms to the eukaryotic consensus rule does not necessarily mean that the splicing mechanism is the same. In this respect it is interesting to note the lack of the pyrimidine stretch at the 3' splice junction characteristic of all introns other than those of type I (and some introns in tRNAs). It is not known whether a conspicuous pyrimidine stretch located 35 bp upstream of the splice junction should be regarded as a substitute for the pyrimidine stretch found at the splice junction in other introns. Sequences completely homologous to the branch sites involved in lariat formation (Keller, 1984) in higher organisms and yeast was not found close to the 3' end of the intron. On the other hand, the consensus sequence for branch sites in higher eukaryotes is very degenerate and the premesis for searching for sites of similar function in *Tetrahymena* unclear. Given the classification of the *Tetrahymena* intron based on the GT/AG rule, on the one hand, and its divergence from other introns obeying the same rule on the other hand, it is possible that further work on this intron will help the recent speculations on the evolutionary relatedness of the various forms of splicing (Tabak and Grivell, 1986; Cech, 1986). One way to approach this question could be to study the sequence of the same intron in the closely related *T. pigmentosa* in which two genes coding for ribosomal protein S25 have been found (unpublished results).

One other aspect of the presence of a nuclear mRNA-type intron in *Tetrahymena* is that the nucleus of *Tetrahymena* is the first genetic compartment in which this type of intron and a type I intron, namely the rRNA intron, are found to co-exist. Thus these two major types of introns are not mutually exclusive. Although the rRNA intron is capable of acting as an enzyme *in vitro* using other polyribonucleotides as substrates (Zaug and Cech, 1986), it is not likely to play a role in the splicing of mRNA *in vivo*. First, strains lacking the rRNA intron are known (Nielsen *et al.*, 1985), and second, the outspliced intron is metabolically unstable *in vivo* (Brehm and Cech, 1983). Finally it should be noted that the small nuclear RNAs, some of which are implicated in mRNA splicing in higher eukaryotes, are well described in *Tetrahymena* (Pedersen *et al.*, 1985). We are in the process of characterizing several of these at the sequence level in order to obtain a tool for studying this aspect of RNA splicing in *Tetrahymena*.

Materials and methods

Cells

T. thermophila B1868 (mating type VII) was grown in complex proteose peptone/yeast extract/glucose medium. Cells were starved by washing into 10 mM Tris−HCl, pH 7.5 and refed by the addition of complex medium as previously described (Dreisig *et al.*, 1984a).

Isolation of mRNA

Total cellular RNA was isolated by the guanidinium thiocyanate method (Chirgwin *et al.*, 1979). The RNA was passed twice over an oligo(dT)-cellulose column (Collaborative Research) following the manufacturer's recommendations. Approximately 2.5% of the total RNA was isolated in this way as poly(A)$^+$ RNA.

Construction of cDNA library

The cDNA library was constructed from poly(A)$^+$ RNA upshifted cells using conventional methods (Maniatis *et al.*, 1982). The second strand synthesis was self-primed and the double-stranded cDNA blunt-ended wtih S_1 nuclease before cloning into the *Pst*I site of pBR322 by the homopolymeric tailing method.

In vitro translation of hybrid-selected mRNA

The B2 cDNA clone was used to hybrid select complementary mRNA from poly(A)$^+$ RNA as described by Maniatis *et al.* (1982), except that the hybridization buffer was 50% formamide, 20 mM Pipes−NaOH, pH 6.4, 0.2% SDS, 0.6 M NaCl, 100 μg/ml bovine tRNA (Sigma), 100 μg/ml polyadenylic acid (Sigma), 1 U/μl RNasin (Promega) and 1 mM dithioerythritol (DTE) and the hybridization reaction was begun at 70°C, slowly adjusted to 42°C and continued at 42°C for at least 12 h. The hybrid-selected mRNA was subsequently translated in a nuclease-treated rabbit reticulocyte lysate system (Pelham *et al.*, 1976) according to the manufacturer's (Amersham) recommendations with the following exceptions: lysate concentration was reduced to 40%, Spermidine−HCl, pH 7.0 was added to 0.5 mM, a mixture of 19 unlabeled amino acids (not including methionine) was added to a final concentration of 50 μM each and the translations were carried out in the presence of 1.5 μCi/μl of [^{35}S]methionine (>800 Ci/mmol, NEN). The translation products were analysed by two-dimensional gel electrophoresis in the pH 8.6/SDS system (Dreisig *et al.*, 1984a) followed by fluorography as previously described (Andreasen *et al.*, 1984).

Hybridization methods

The gene map was constructed by Southern blotting (Southern, 1975) and conventional hybridization techniques (Maniatis *et al.*, 1982). The transcription mapping was likewise carried out by the Northern blotting technique. In this case, the various RNAs were separated on formaldehyde/agarose gels (Maniatis *et al.*, 1982). In both cases we used nitrocellulose membranes (Schleicher and Schuell) and nick-translated DNA hybridization probes. The RNA prepared for the Northern blots was total RNA prepared as previously described from cells at various stages during a starvation−refeeding cycle (Andreasen *et al.*, 1984).

Cloning and sequencing

A genomic library of total *Tetrahymena* DNA was constructed in the vector λgtWESλB. High mol. wt DNA was extracted from a proteinase K-digested cell lysate and partially digested with *Eco*RI. The digest was fractionated on a sucrose gradient and the DNA in the fractions containing fragments > 10 kb was ligated to the isolated vector arms of the phage λgtWES (Leder *et al.*, 1977). The ligated DNA was *in vitro* packaged using a packaging extract prepared according to Hohn and Murray (1977). The recombinant phages were plated on *E. coli* LE 392. It should be noted that a genomic library prepared in this way contains DNA originating from both nuclei in *Tetrahymena*, the macronucleus (45n) and the micronucleus (2n). The λ-library was screened for B2 sequences by the plaque screening procedure (Benton and Davis, 1977) using a nick-translated cDNA clone (prB2) as hybridization probe. Since none of the λ clones contained the entire gene, a plasmid cloning strategy was designed aimed at cloning a specific restriction fragment containing the entire gene from total DNA. Such a strategy is feasible because of the low complexity of the *Tetrahymena* genome. The resulting transformants were screened by the colony hybridization method (Grunstein and Hogness, 1975) using radioactively labeled pHRG (Figure 2) as hybridization probe. The clones obtained in this way contained the entire B2 gene and included a substantial overlap with the original λ-clones.

Radioactively end-labelled restriction fragments of the genomic copy of the B2 gene and the cDNA clone were DNA sequenced by the chemical sequencing method (Maxam and Gilbert, 1977). Overlapping fragments of both strands were sequenced in all cases.

Computer analysis

DNA sequence analysis was carried out using the ANALYSEQ programs obtained from Dr R.Staden (Medical Research Council, Cambridge) and the NAQ and PSQ program from National Biomedical Research Foundation (NBRF),

Georgetown. Comparisons of the sequence of the B2 gene product and ribosomal protein sequences were performed by using the SEARCH, ALIGN and DOT-MATRIX programs from NBRF. The scoring values were obtained from a standard mutational data matrix (NBRF) and the *E. coli* ribosomal protein sequences from the NBRF protein database kindly provided by the DNA sequence library at the European Molecular Biology Laboratory, Heidelberg.

Primer extension analysis

A synthetic 18-mer complementary to the 5' end of the gene (Figures 2 and 5) was radioactively labeled at the 5' end using T4 polynucleotide kinase and $[\gamma\text{-}^{32}P]$-ATP (Maniatis *et al.*, 1982). 5 pmol of labeled primer were mixed with 2 μg of poly(A)$^+$ RNA and made 10 mM Tris−HCl pH 8.0, 1 mM EDTA in a volume of 5 μl. The mixture was heated for 2 min at 90°C for denaturation, 3 μl of 0.5 M KCl was added and the reaction transferred to 65°C for 10 min for annealing. The volume was adjusted to 40 μl by the addition of a buffer consisting of 50 mM Tris−HCl, pH 8.3, 10 mM MgCl$_2$, 30 mM KCl, 1 mM dithiothreitol (DTT), 1 mM of each of the dNTPs and 0.5 U/μl of RNasin. 30 U of AMV reverse transcriptase was added and the reaction allowed to proceed for 15 min at 37°C and 80 min at 43°C. The reaction was then stopped by the addition of 25 μl of 50 mM EDTA and the RNA removed by digestion with 10 μg of RNAse for 30 min at 37°C. After phenol extraction and alcohol precipitation, the primer extension product was resuspended in a loading buffer containing 80% formamide and analysed on an 8% denaturing polyacrylamide gel.

Mung bean nuclease protection experiments

The probe in these experiments was a 531-bp *Dra*I−*Hpa*II restriction fragment extending from position −500 to +32 according to the numbering of Figure 3. The probe had been selectively labeled at the *Hpa*II site with $[\gamma\text{-}^{32}P]$ATP and polynucleotide kinase. 0.1 μg samples of the probe were alcohol precipitated together with 3 μg of poly(A)$^+$ RNA or 100 μg of total cytoplasmic RNA from upshifted cells followed by resuspension in 60 μl of 70% formamide, 1 mM EDTA, 0.1 M Pipes buffer pH 7.0 and 0.1 M NaCl. After denaturation at 75°C for 10 min the samples were cooled gradually to 52°C over a period of 5 h. Hybridization was continued at 52°C for an additional 15 h and then terminated by addition of 500 μl ice cold S1 buffer (Maniatis *et al.*, 1982) and transfer to a dry ice cooling bath. 30 U of Mung bean nuclease (New England Biolabs) were added during thawing of the samples and incubation was continued at 30°C for 1 h. The nuclease treatment was terminated by addition of 2 vol alcohol and the precipitated samples were analysed on sequencing gels after resuspension and heating in 80% formamide loading buffer.

Isolation of ribosomal proteins

Ribosomes and ribosomal subunits were isolated as previously (Dreisig *et al.*, 1984a). Ribosomal proteins were extracted with MgCl$_2$/acetic acid and precipitated with acetone (Kristiansen and Kruger, 1978). The ribosomal proteins were fractionated by two-dimensional polyacrylamide gel electrophoresis in the pH 8.6/SDS system (Dreisig *et al.*, 1984a).

Acknowledgements

References

Allen,S. and Gibson,I. (1973) In Elliott,A.M. (ed.), *Biology of Tetrahymena*. Dowden, Hutchison and Ross, Stroudsberg, Pennsylvania, p. 360.
Andreasen,P.H., Dreisig,H. and Kristiansen,K. (1984) *Eur. J. Biochem.*, **140**, 485−492.
Bannon,G.A., Bowen,J.K., Yao,M.-C. and Gorovsky,M.A. (1984) *Nucleic Acids Res.*, **12**, 1961−1975.
Benton,W.D. and Davis,R.W. (1977) *Science*, **196**, 180−182.
Brehm,S.L. and Cech,T.R. (1983) *Biochemistry*, **22**, 2390−2397.
Caron,F. and Meyer,E. (1985) *Nature*, **314**, 185−188.
Cech,T.R. (1985) *Int. Rev. Cytol.*, **93**, 3−22.
Cech,T.R. (1986) *Cell*, **44**, 207−210.
Chen,W. and Struhl,K. (1985) *EMBO J.*, **4**, 3273−3280.
Chirgwin,J.M., Przybyla,A.E., MacDonald,R.J. and Rutter,W.J. (1979) *Biochemistry*, **18**, 5294−5299.
Cupples,C.G. and Pearlman,R.F. (1986) *Proc. Natl. Acad. Sci. USA*, in press.
Dreisig,H., Andreasen,P.H. and Kristiansen,K. (1984a) *Eur. J. Biochem.*, **140**, 469−475.
Dreisig,H., Andreasen,P.H. and Kristiansen,K. (1984b) *Eur. J. Biochem.*, **140**, 477−483.
Funatsu,G. and Wittmann,H.G. (1972) *J. Mol. Biol.*, **68**, 547−550.
Grunstein,M. and Hogness,D. (1975) *Proc. Natl. Acad. Sci. USA*, **72**, 3961−3965.
Helftenbein,E. (1985) *Nucleic Acids Res.*, **13**, 415−433.
Hohn,B. and Murray,K. (1977) *Proc. Natl. Acad. Sci. USA*, **74**, 3259−3263.
Horowitz,S. and Gorovsky,M.A. (1985) *Proc. Natl. Acad. Sci. USA*, **82**, 2452−2455.
Keller,W. (1984) *Cell*, **39**, 423−425.
Klobutcher,L.A., Jahn,C.L. and Prescott,D.M. (1984) *Cell*, **36**, 1045−1055.
Kozak,M. (1984) *Nucleic Acids Res.*, **12**, 857−872.
Kristiansen,K. and Kruger,A. (1978) *Biochim. Biophys. Acta*, **521**, 435−451.
Leder,P., Tiermeier,D. and Enquist,L. (1977) *Science*, **196**, 175−177.
Leer,R.J., Van Ramsdonk-Duin,M.M.C., Mager,W.H. and Planta,R.J. (1985) *Curr. Genet.*, **9**, 273−277.
Maniatis,T., Fritsch,E.F. and Sambrook,J. (1982) *Molecular Cloning, A Laboratory Manual*. Cold Spring Harbor Laboratory Press, NY.
Martindale,D.W., Martindale,H.M. and Bruns,P.J. (1986) *Nucleic Acids Res.*, **14**, 1341−1354.
Maxam,A. and Gilbert,W. (1977) *Methods Enzymol.*, **65**, 499−560.
Meyer,E., Caron,F. and Guiard,B. (1984) *Biochimie*, **66**, 403−412.
Michel,F. and Dujon,B. (1983) *EMBO J.*, **2**, 33−38.
Mount,S.M. (1982) *Nucleic Acids Res.*, **10**, 459−472.
Nielsen,H., Simon,E.M. and Engberg,J. (1985) *J. Protozool.*, **32**, 480−485.
Pedersen,N., Hellung-Larsen,P. and Engberg,J. (1985) *Nucleic Acids Res.*, **11**, 4203−4223.
Pelham,H.R.B. and Jackson,R.J. (1976) *Eur. J. Biochem.*, **67**, 247−256.
Post,L.E. and Nomura,M. (1980) *J. Biol. Chem.*, **255**, 4660−4666.
Preer,J.R.,Jr, Preer,L.B., Rudman,B.M. and Barnett,A.J. (1985) *Nature*, **314**, 188−190.
Southern,E.M. (1975) *J. Mol. Biol.*, **98**, 503−517.
Syvanen,M. (1986) *Trends Genet. Sci.*, **2**, 63−66.
Tabak,H.F. and Grivell,L.A. (1986) *Trends Genet. Sci.*, **2**, 51−55.
Villa-Komaroff,L., Efstratiadis,A., Broome,S., Lomedico,P., Tizard,R., Naker, S.P., Chick,W.L. and Gilbert,W. (1978) *Proc. Natl. Acad. Sci. USA*, **75**, 3727−3731.
Zaug,A.J. and Cech,T.R. (1986) *Science*, **231**, 470−475.