Volume 8 Number 24 1980 Nucleic Acids Research

Nucleotide sequence through the 18S-28S intergene region of a vertebrate ribosomal transcription unit

Lucinda M.C.Hall and B.Edward H.Maden

Department of Biochemistry, University of Glasgow, Glasgow G12 8QQ, UK

Received 3 November 1980

ABSTRACT

We have determined the nucleotide sequence of part of a cloned ribosomal transcription unit from Xenopus laevis extending from the 3' region of the 18S gene through the 18S-28S intergene region into the start of the 28S gene. The 18S 3' region possess two tracts of high homology with the corresponding segments of other eukaryotic 18S genes (yeast and Bombyx mori) separated by a tract of low homology which in X.laevis is rich in G plus C. The first internal transcribed spacer, between the 18S and 5.8S genes, is 557 nucleotides long, very rich in G plus C (84%) and shows no sequence homology with the corresponding yeast sequence. The 5.8S rRNA sequence is revised slightly in the light of the DNA sequence. The second internal transcribed spacer, between the 5.8S and 28S genes, is 262 nucleotides long and is even richer in G plus C (88%) than the first internal spacer. 28S rRNA starts with the sequence pUCAG. This is encoded at the first of three closely linked TCAG sites in rDNA.

INTRODUCTION

In this paper we describe the nucleotide sequence of part of a Xenopus laevis ribosomal transcription unit extending from the 3' part of the 18S gene to a short distance beyond the start of the 28S gene. Our primary objective was to characterize the 18S-28S intergene region with a view to gaining insight into processing steps during ribosome maturation. In this general region of rDNA we define the tract between the 18S and 5.8S genes as the first internal transcriber spacer (ITS 1) and that between the 5.8S and 28S genes as the second internal transcribed spacer (ITS 2). At the outset of this work the lengths of these spacers were not accurately known in X.laevis, although the 5.8S gene and its immediately flanking sequences had been characterized (1). It was therefore necessary to sequence between points which could be identified with certainty as lying within the 18S and 28S genes respectively. The Eco RI site in the 18S gene provided a convenient starting point.

The beginning of the 28S was identified by experiments which will be described. Features of interest were found both in the gene and spacer parts of the sequence. We therefore provide an integrated description of this overall region of the transcription unit. We also compare our findings with those reported for yeast (2,3) and E.coli (4).

METHODS

rDNA clones

pXlrl01 was a gift from R. Reeder. It consists of a complete rDNA unit bounded by Hind III sites (figure 1), cloned in the vector pMB9. This plasmid, as pointed out previously (5), enables any sequencing objective to be approached using the rDNA of a single transcription unit. A subclone from this plasmid has been used for sequencing the 5' region of the 18S gene (5). The plasmid has also been used for transcription studies (6). To facilitate the present analysis the regions between the 18S Eco RI site and the Bam HI site in ITS 2, and between this Bam site and the Bam site in the 28S gene (figure 1), were separately subcloned into pBR322 (subclones pXlrl01L and l01M respectively). For some of the rRNA hybridization experiments two homologous subclones, derived from an independent "parent" clone, pXlrll, were used:- subclones pXlrllL and llM respectively (see figure 1 of ref.7 for derivation of these latter subclones).

Sequence analysis

This was carried out by the method of Maxam and Gilbert (8,9) with thin gels for electrophoresis (10). The sequencing strategy is outlined in figure 1. The sequences of some of the longer restriction fragments were read for up to 230 nucleotides.

rRNA hybridization and fingerprinting analysis

This was carried out as described (5).

RESULTS

The DNA sequence

The sequence is shown in figure 2. The restriction map of Boseley et al. (1), though derived from a different rDNA clone to that used here, was of assistance during the analysis. In addition the enzymes Taq 1 and Sau 3A1 were used. The latter was employed in particular to aid the analysis near the 3' end of the 18S gene, where the known rRNA sequence, $GAUCAUUA_{OH}$, (11) predicted a Sau 3A site, GATC, in the DNA. The sequence was determined throughout on both strands except for two

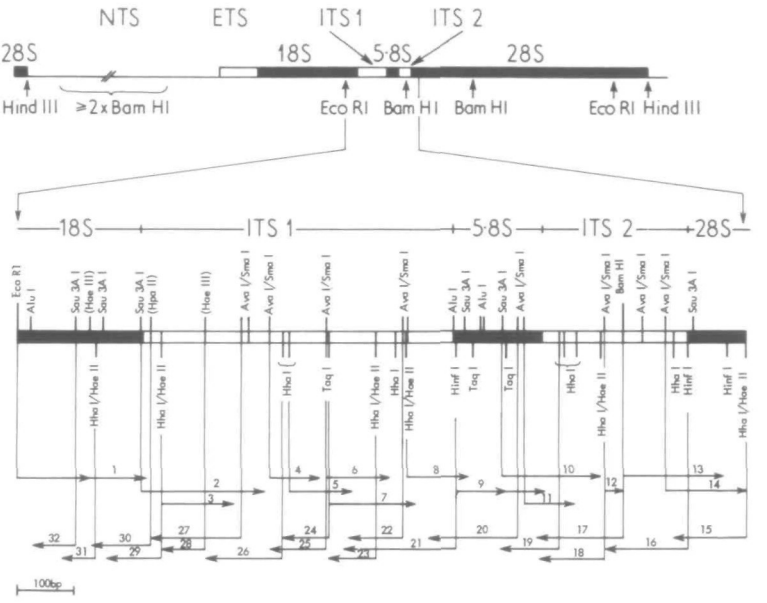


Figure 1. Upper section: unit structure of X.laevis rDNA. NTS, ETS, ITS denote non-transcribed, external transcribed and internal transcribed spacers. (Non-transcribed spacers of different repeating units differ from each other in numbers of Bam HI sites; NTS of pX1r 101 possesses two such sites). Lower section: sequencing strategy for 18S-28S intergene region. pX1r 101L (see methods section) was used to sequence up to Bam HI site in ITS 2. pX1r 101M was used for continuation into 28S gene. To sequence through the Bam site in ITS 2, a 420 bp Hinf I fragment from the parent plasmid pX1r 101 was used. Map shows all restriction sites for Alu I, Ava I, Hae II, Hha I, Hinf I, Sau 3A I and Taq I. For Hae III and Hpa II only those sites that were used are shown. (Hae III and Hpa II sites are very numerous; Ref.1). Starts of arrows indicate kinase-labelled 5' ends. Tips of arrows indicate points to which sequencing gels were read. Arrows are numbered sequentially along the two DNA strands. Preparation of restriction fragments for sequence analysis was by standard procedures. In most experiments rDNA was excised from the respective plasmid and was further restricted as necessary; the fragments were then labelled and finally restricted asymetrically for sequencing. In a few experiments strand separation was carried out (e.g. gels 3 and 26 from Hha I fragment). In one experiment, two Ava I 140 bp fragments initially comigrated but, after labelling, three of the four strands were successfully purified on a strand separation gel:- sequencing gels 6, 11 and 18. Finally, in two experiments use was made of restriction sites in both rDNA and the vector to obtain a suitable labelled fragment:- gels 12 and 27.

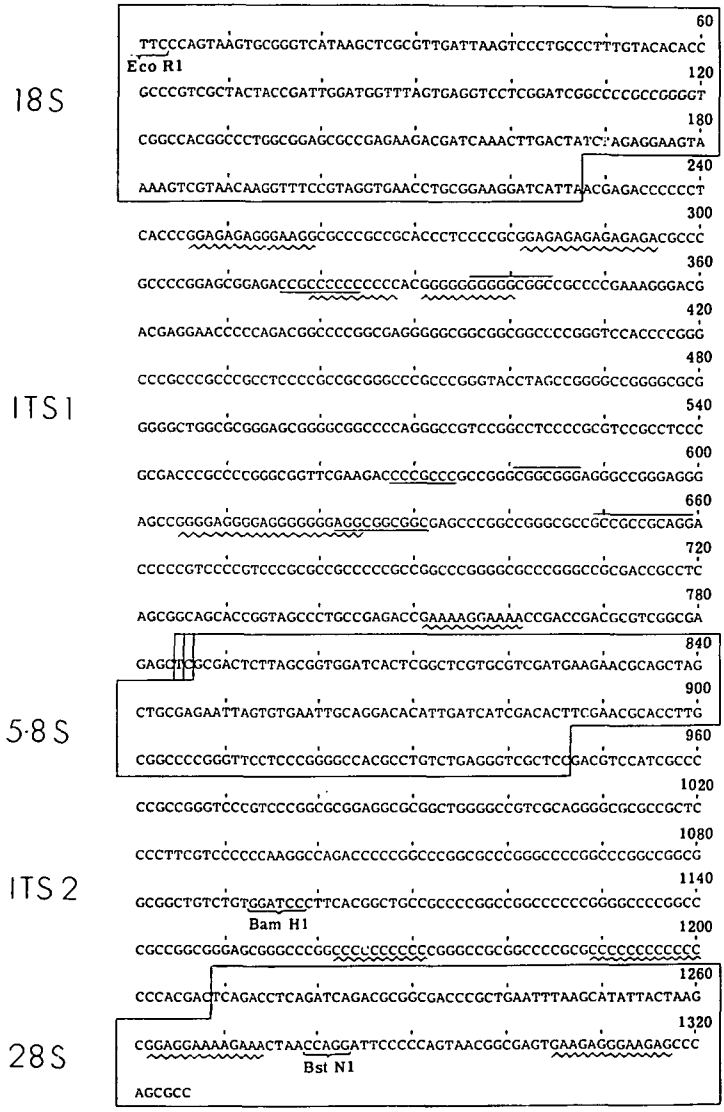


Figure 2. Sequence of X.laevis X1r101 rDNA from centre of Eco RI site in 18S gene to first Hae II site in 28S gene. Boxed regions denote 18S, 5.8S and 28S coding regions. 5.8S rRNA shows 5' terminal heterogeneity (15). Sequence shown is the "s" strand (synonymous to RNA). In ITS 1 secondary structure effects persistently interfered with gel readings in three regions on the "s" strand (horizontal line above sequence) and on the "c" strand (horizontal line below sequence). Serrated lines indicate runs of ten or more consecutive C residues or G residues or runs of more than ten A and G residues.

very short tracts in ITS 1 and a short region near the Eco RI site of the 18S gene (figure 1). Clear reading was obtained on one strand through each of these short tracts. The sequence was further corroborated by extensive overlaps along much of the length (figure 1). Pending the adoption of a definitive numbering system for the ribosomal transcription unit we have temporarily numbered nucleotides from the centre of the Eco RI site, counting the fourth nucleotide in this site as number one.

Gene-spacer boundaries

18S rRNA. A highly conserved sequence occurs at the 3' end of eukaryotic 18S rRNAs (12-14), terminating in the GAUCAUUA tract mentioned above. The corresponding DNA sequence occurs on line 4 of figure 2. The last residue in the 18S gene is A227 counting from the centre of the Eco RI site. Therefore ITS 1 commences at A228.

5.8S rRNA. The published sequence of X.laevis 5.8S rRNA (15,16) can be approximately aligned with the DNA sequence on lines 14-16 of figure 2 (see also ref.1). 5.8S rRNA shows 5' terminal heterogeneity, the longest molecules commence with pUCG (15). On this basis we place the start of the 5.8S coding region at T785. The final nucleotide in the RNA sequence (15,16) corresponds to C946 in the DNA. Therefore ITS 2 commences at G947. There are some points of discrepancy between the 5.8S rDNA sequence and the published rRNA sequence: see below.

28S rRNA. Until recently the only chemical information on the 5' end of X.laevis 28S rRNA was that the first nucleotide is pUp (17,18). The distance between the 5.8S and 28S genes was also unknown. Brand and Gerbi (19), using S1 nuclease mapping, placed the start of the 28S gene about 50 nucleotides to the right of the Bam HI site of ITS 2. On the basis of the following experiments we identify the start of the 28S gene at a point 115 nucleotides beyond the Bam site.

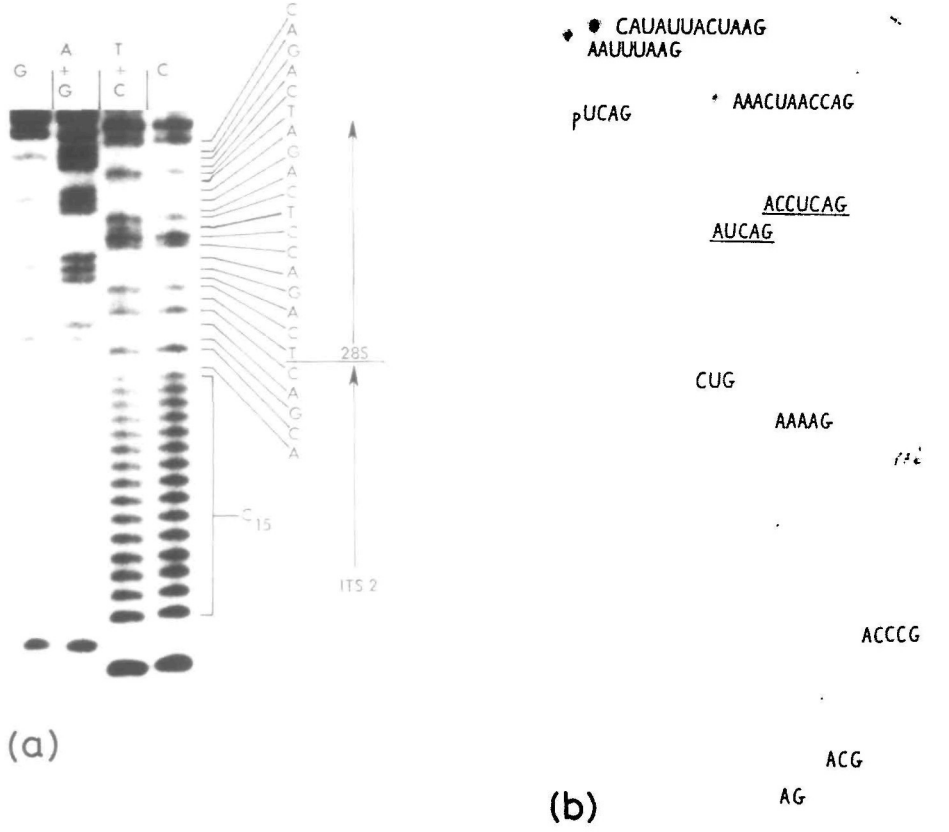
(i) The 5' terminal oligonucleotide of 28S rRNA was identified by fingerprinting analysis. 28S rRNA was hybridized to pX1r11M (which contain a homologous rDNA segment to pX1r101M, used for sequencing). The hybridized RNA was eluted and a T1 ribonuclease fingerprint was prepared. General procedures for these steps have been described (5). All uridine-containing products were screened for pUp by alkaline hydrolysis. One such product was found, and analysis with pancreatic ribonuclease established the sequence as pUCAG.

(ii) The above finding indicates that the 28S gene starts with TCAG. Sequencing was therefore carried out rightwards from the 5.8S gene in search of TCAG sites. Three closely linked sites were found, the first of which is just after a conspicuous run of C residues (figures 2, 3a).

(iii) To identify which, if any, of these three sites was the start of the 28S gene we hybridized 28S rRNA to an rDNA restriction fragment bounded on the right by the Bst NI site in figure 2. If the 28S gene starts at one of the three TCAG sites the hybridized rRNA should yield a very simple fingerprint containing pUCAG and 10-12 other spots. In particular, if the gene starts at the first TCAG site the RNA fingerprint will contain ACCUCAG and AUCAG. If it starts at the second site the fingerprint will contain AUCAG but not ACCUCAG. If it starts at the third site neither of these products will be present. The only other products expected in the "one uridylate" region of the fingerprint are CUG, at the bottom of this region and AAACUAACCAG, at the top and merging with the two uridylate regions. AUCAG and ACCUCAG, if present, will migrate to intermediate positions in the second dimension of the fingerprint. (pUCAG runs in the two uridylate region by virtue of its extra phosphate). Figure 3b shows the fingerprint that was obtained. Rather a small quantity of material was recovered from the hybrid and it was necessary to rely upon electrophoretic mobility data for identification of the products. This was possible using the principles outlined above in conjunction with the DNA sequence and with mobility data already obtained in (i) on reference products such as AAAAG, CUG and pUCAG. The mobilities of all products are in accord with the predictions of the DNA sequence. In particular AUCAG and ACCUCAG, which could not be confused with any other predicted products, are both present in good yield establishing that the 28S coding sequence starts at the first of the three TCAG sites.

Features of the sequence

Table 1 summarizes the base compositions of different regions of the sequence. The ITS regions are very rich in G plus C whereas the neighbouring ribosomal sequences are only slightly so. In both of the spacers there are quite long runs of C residues and in ITS1 there are also polypurine tracts. ITS 1 is particularly deficient in T residues. Detailed features of the sequence, together with available comparative data, may be relevant to understanding eukaryotic ribosome structure and processing, and are now discussed.



pUCAG/ACCUCAG/AUCAG/ACG/CG/G/CG/ACCCG/CUG/AAUUUAAG/

CAUAUUACUAAG/CG/G/AG/G/AAAAG/AAACUAACCAG/

Figure 3. Identification of start of 28S coding sequence. (a) shows detail of sequencing gel reading rightwards from last Sma I site in ITS 2 (gel 14 of figure 1). Three closely linked TCAG sites occur after the conspicuous run of C residues (T is rather faint and the third TCAG was read more easily from lanes with longer separation: not shown. The sequence was confirmed on the other strand). Below the figure is shown the predicted RNA sequence up to the point encoded by the Bst NI site in figure 2 if the 28S gene starts at the first TCAG site. T1 ribonuclease cleavage points are indicated. (b) shows a T1 ribonuclease fingerprint of the region of 28S rRNA that hybridizes to rDNA to the left of the Bst NI site. G and CG have run off the end of the fingerprint and there is a small amount of incompletely digested material at the origin of the second dimension. Otherwise the mobilities of all the products match the predictions from the DNA sequence.

	Length	T	A	G	C	G + C
18S (3' region)	227	50	51	66	60	
		22%	22.5%	29%	26.5%	55.5%
ITS 1	557	19	69	230	239	
		3.5%	12.5%	41%	43%	84%
5.8S	≤162	35	30	48	49	
		21.5%	18.5%	30%	30%	60%
ITS 2	262	18	13	92	139	
		7%	5%	35%	53%	88%
28S (5' region)	118	16	39	32	31	
		13.5%	33%	27%	26.5%	53.5%

Table 1. Base composition of gene and spacer segments through the 18S - 28S intergene region. (5.8S data calculated for maximum sequence length).

18S 3' region. Because the last 20 nucleotides of 18S rRNA are almost invariant between eukaryotes (12-14) it was of interest to extend interspecies comparisons across a larger part of the 18S gene. Figure 4 compares the 18S rDNA sequences of X.laevis, silkworm (20) and yeast (2) from the Eco RI site through to the start of ITS 1. Large

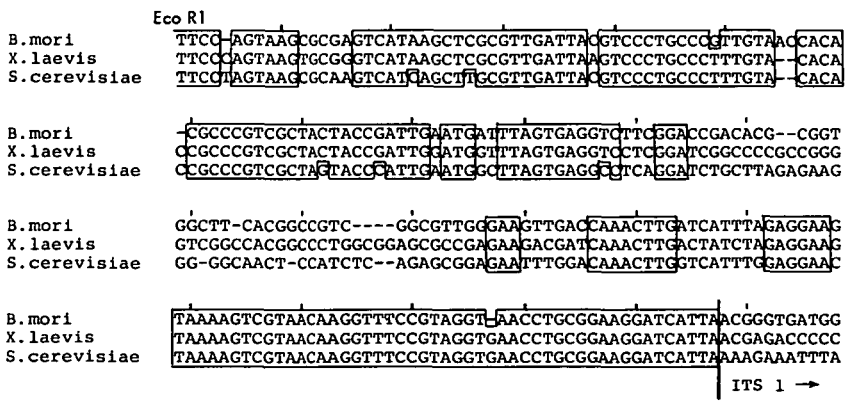


Figure 4. Comparison of 3' region of 18S rDNA from Bombyx mori (20), Xenopus laevis, and Saccharomyces cerevisiae (2). Tracts of three or more nucleotides showing complete homology between the three species are boxed. Prime symbols denote every tenth nucleotide of the Xenopus sequence from the centre of the Eco RI site, as in figure 2.

parts of the gene sequences are identical in the three species. However, in the region between nucleotides 100-170 (numbered from the Eco RI site) homology is much lower than elsewhere. Brimacombe (21) has proposed a hairpin secondary structure for part of this region in yeast. The Xenopus and silkworm sequences can be fitted to a similar secondary structure, with nucleotides 123-126 forming the terminal loop in Xenopus. Evidently this is a region where ribosome function is not critically dependent upon a unique primary structure, but conservation of secondary structure may well be important.

ITS 1. The first and last parts of this sequence contain some A-rich purine tracts; the central region is very deficient in both A and T. Much of this central region can be fitted to a series of hairpin structures (figure 5). Parts, at least, of these structures seem likely to exist in ribosomal precursor RNA since strong secondary structure effects were encountered when sequencing the corresponding

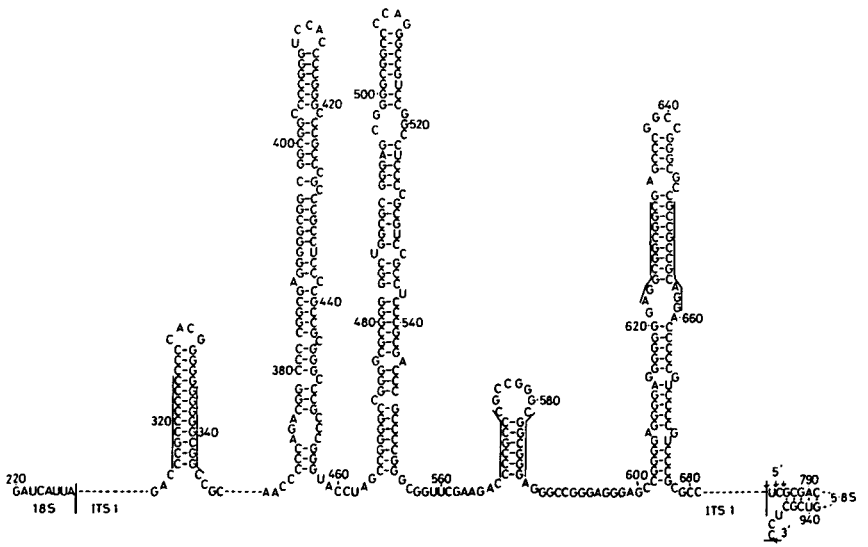


Figure 5. Possible structural features in ITS 1 showing hairpins in central region and single stranded processing points at each end of 5.8S gene (see also Boseley et al., (1)). The hairpins would be very stable because of the large number of GC pairs. Vertical lines indicate secondary structure interactions inferred from DNA sequencing gels. A-rich regions at the start and end of ITS 1 do not give rise to any obvious secondary structures.

parts of rDNA (figures 2 and 5). The first and last parts of ITS 1, with their imbalance between A and T, would not give rise to internal helical features.

Boseley et al sequenced the last 80 nucleotides before the start of the 5.8S gene (1) using a different clone of X.laevis rDNA. The last 50 nucleotides before the 5.8S gene are identical in the two clones but there are several differences in the preceding nucleotides: -

```
                    * **          *   *      * *
This study      :   CCCGGGCCGCGACCGCCTCAG-CGGCAGCACCGGTAG
Boseley et al   :   GCGCGGCCGCGACCGC-TCAGACGGCAGCCCGGGTAG
```

ITS 1 has also been sequenced in yeast (2,3). The yeast spacer is considerably shorter than the X.laevis spacer (~ 360 nucleotides compared with 557 nucleotides), is generally AT-rich and diverges from the Xenopus sequence almost immediately after the 18S gene - spacer boundary (figure 4). The Xenopus and yeast sequences differ from each other on both sides of the boundary between ITS 1 and the 5.8S gene (2,3,22).

5.8S rDNA. The DNA sequence found here is identical to that reported by Boseley et al., for their rDNA clone (1). This fact, and additional considerations outlined below, lead us to propose revisions to the published 5.8S rRNA sequence (15,16). One of the revisions is of interest in the present context since it relates to the 5' processing point of the molecule. Several vertebrate 5.8S rRNAs possess terminal heterogeneity, which must be introduced during processing. On the basis of detailed sequence data from rat (23) and comparative fingerprinting data from other species (16,24) the vertebrate 5.8S starting sequence was believed to be p(C)GAC... However, Ford and Mathieson reported that some 5.8S rRNA molecules in Xenopus start with pUCG (15). This sequence can only be accommodated in the rDNA sequence by proposing that the longest 5.8S rRNA molecules start with pUCGCGAC... Insertion of the extra GC doublet into the 5' end of the sequence is numerically balanced by lack of an expected GC at a point some 50 nucleotides into the sequence. Finally there is a difference between one version of the RNA sequence (15) and the DNA sequence in a short region centred on nucleotide 110 of the RNA sequence. The differences (and implied revisions) are summarized below:-

```
                                        50          110
Ref. 15                  p(UC)GAC...   AGCGCU...   GCACUUCCG...
Ref. 16                  p(UC)GAC...   AGCGCU...   GCACCUUGCG...
Inferred from DNA        p(UC)GCGAC...  AGCU  ...   GCACCUUGCG...
```

The revision at the 5' end bears upon processing as follows. Boseley et al., (1), before Ford and Mathieson's work was published, assumed that 5.8S rRNA starts with p(C)GAC... They pointed out that according to the secondary structure model for isolated 5.8S rRNA(23,16) the 5' processing point would be within a double stranded region of the precursor molecule, whereas the 3' processing point would be within a single stranded region (model A of ref. 1, see also fig. 5). However, the revised sequence places the 5' processing point 2-3 nucleotides to the left of the site previously envisaged. The new site is outside the region of interaction with the 3' penultimate part of 5.8S rRNA (23,16), and therefore, like the 3' processing site, is a single stranded region (figure 5).

ITS 2. The most striking aspect of ITS 2 is the very high content of C residues (53%). This spacer differs from ITS 1 in having no A-rich purine tracts and the proportions of A and T are roughly equal. Much of the sequence could fold into hairpin secondary structures but the two long tracts of C residues near the start of the 28S gene are likely to remain single stranded. Again some nucleotide differences are seen when the sequence from this study is compared to that obtained by Boseley et al., (1) for the first 30 nucleotides of ITS 2 in a different rDNA plasmid. No data are yet available to permit a comparison with other eukaryotes.

28S rDNA. Within 20 nucleotides of the start of the 28S gene the sequence TCAGA occurs three times. Similar repetition is not seen at the 5' ends of either the 18S (5) or 5.8S genes. Also of interest in the short region sequenced are two A-rich purine tracts similar to those seen in ITS 1. In this case the rest of the sequence does not have an exceptionally high proportion of G plus C but there are again many less T residues than A residues.

DISCUSSION

The experiments described in this paper establish the nucleotide sequence through the 18S-28S intergene region of a cloned ribosomal transcription unit of X.laevis and define the exact locations of the gene-spacer boundaries. The two internal transcribed spacers are extremely rich in G plus C and contain tracts of ten or more consecutive G or consecutive C residues. These characteristics distinguish the spacers from the gene regions; the transitions between spacer and gene

type sequence occur quite abruptly. When the sequences of Xenopus and yeast are compared the transition from largely conserved gene sequence to highly variant spacer sequence also occurs abruptly at the gene-spacer boundaries. It is interesting that in the 3' region of the 18S gene the part which shows least homology between Xenopus and yeast is, in Xenopus, also very rich in G plus C.

The high degree of evolutionary variability between transcribed spacer sequences (2,3) makes it difficult to generalise on the possible functional roles of different parts of these sequences. It is likely that secondary structure, as well as sequence, plays an important role in the processing of precursor rRNA. From the sequence data so far available (this work, 5, and unpublished work from this laboratory) it appears that extensive base pairing would not occur between the spacers flanking the mature rRNAs. Rather the gene-spacer boundaries are in regions which would probably remain single stranded in the precursor molecule. This contrasts with the situation in E. coli in which the spacers on either side of both 16S and 23S rRNA can form a long base pairing stem (4,25). The stem contains sites for ribonuclease III, one of the key enzymes in prokaryotic rRNA processing. However this implies simultaneous cleavage on both sides of the rRNA whereas in Xenopus the spacers are excised sequentially (26). Furthermore there is apparently no transcribed spacer beyond the 3' end of the 28S gene (27). Veldman et al. also note that in yeast complementarity is not seen between the sequences flanking various rRNA molecules (3).

A final point concerns the presence of genes coding for tRNAs in ribosomal transcription units of prokaryotes and mitochondria (28,29). We are unable to find any sequences consistent with a tRNA gene (with or without an intervening sequence) in either ITS 1 or ITS 2 using the criteria of conserved nucleotides (especially GTΨC) combined with cloverleaf secondary structure.

ACKNOWLEDGEMENTS

We thank John Forbes and Mary Robertson for technical assistance. This work was supported by a grant from the Medical Research Council.

REFERENCES

1. Boseley, P.G., Tuyns, A. and Birnstiel, M.L. (1978) Nucleic Acids Research 5, 1121-1137.
2. Skryabin, K.G., Krayev, A.S., Rubtsov, P.M. and Bayev, A.A.(1979) Dokl, Akad, Nauk. USSR. 247, 761-765.

3. Veldman, G.M., Brand, R.C., Klootwijk, J. and Planta, R.J.(1980) Nucleic Acids Research 8, 2907-2920.
4. Young, R.A. and Steitz, J.A. (1978). Proc. Nat. Acad. Sci. U.S.A. 75, 3593-3597.
5. Salim, M. and Maden, B.E.H. (1980). Nucleic Acids Research 8, 2871-2884.
6. Trendelenburg, M.F. and Gurdon, J.B. (1978). Nature 276, 292-294.
7. Maden, B.E.H. (1980). Nature, in the press.
8. Maxam, A.M. and Gilbert, W. (1977). Proc. Nat. Acad. Sci. U.S.A. 74, 560-564.
9. Maxam, A.M. and Gilbert, W. (1980). Methods in Enzymology 65, 499-560.
10. Sanger, F. and Coulson, A.R. (1978). FEBS Letters 87, 107-110.
11. Vass, J.K. and Maden, B.E.H. (1978). Eur. J. Biochem. 85, 241-247.
12. Hagenbuchle, O., Santer, M., Steitz, J.A. and Mans, R.J. (1978). Cell 13, 551-563.
13. De Jonge, P., Klootwijk, J. and Planta, R.J. (1977). Nucleic Acids Research 4, 3655-3663.
14. Alberty, H., Raba, M. and Gross, H.J. (1978). Nucleic Acids Research 5, 425-434.
15. Ford, P.J. and Mathieson, T., (1978) Eur. J. Biochem. 87, 199-214.
16. Khan, M.S.N. and Maden, B.E.H. (1977). Nucleic Acids Research 4, 2495-2505.
17. Slack, J.M.W. and Loening, U.E. (1974). Eur.J. Biochem. 43, 59-67.
18. Khan, M.S.N. and Maden, B.E.H. (1976). J. Mol.Biol. 101, 235-254.
19. Brand, R.C. and Gerbi, S.A. (1979). Nucleic Acids Research 7, 1497-1511.
20. Samols, D.R., Hagenbuchle, O. and Gage, L.P. (1979). Nucleic Acids Research 7, 1109-1119.
21. Brimacombe, R. (1980). Biochemistry International 1, 162-171.
22. Rubin, G.M. (1973). J. Biol. Chem. 248, 3860-3875.
23. Nazar, R.N., Sitz, T.O. and Busch, H. (1975). J. Biol. Chem. 250, 8591-8597.
24. Hampe, A., Elardi, M.E. and Galibert, F. (1976). Biochimie 58, 943-951.
25. Bram, R.J., Young, R.A. and Steitz, J.A.(1980). Cell 19, 393-401.
26. Wellauer, P.K. and Dawid, I.B. (1974). J. Mol. Biol. 89, 379-395.
27. Sollner-Webb, B., and Reeder, R.H. (1979). Cell 18, 485-499.
28. Young, R.A., Macklis, R. and Steitz, J.A. (1979). J. Biol. Chem. 254, 3264-3271.
29. Eperon, I.C., Anderson, S. and Neirlich, D.P. (1980). Nature 286, 460-467.