The EMBO Journal vol.6 no.4 pp.1063–1071, 1987

Multiple spacer sequences in the nuclear large subunit ribosomal RNA gene of *Crithidia fasciculata*

David F.Spencer, James C.Collings, Murray N.Schnare and Michael W.Gray

Department of Biochemistry, Dalhousie University, Halifax, Nova Scotia B3H 4H7, Canada

Communicated by R.W.Davies

In *Crithidia fasciculata*, a trypanosomatid protozoan, the nuclear-encoded '28S' rRNA is multiply fragmented, comprising two large (*c* and *d*) and four small (*e*, *f*, *g* and *j*) RNA species. We have determined that the coding sequences for these RNAs (and that of the 5.8S rRNA, species *i*) are separated from one another by spacer sequences ranging in size from 31 to 416 bp. Coding and spacer sequences are presumably co-transcribed, with excision of the latter during post-transcriptional processing generating a highly fragmented large subunit (LSU) rRNA. Secondary structure modelling indicates that the *C. fasciculata* LSU rRNA complex (seven segments, including 5.8S rRNA) is held together in part by long-range intermolecular base pairing interactions that are characteristic of *intra*molecular interactions in the covalently continuous LSU (23S) rRNA of *Escherichia coli*. At least one functionally critical region (encompassing the α-sarcin cleavage site) is contained in a small RNA species (*f*) rather than in one of the two large RNAs. Within a proposed secondary structure model of *C. fasciculata* LSU rRNA, discontinuities between the different segments (created by spacer excision) map to regions that are highly variable in structure in covalently continuous LSU rRNAs. We suggest that 'rRNA genes in pieces' and discontinuous rRNAs may represent an evolutionarily ancient pattern.

Key words: Crithidia fasciculata/rRNA genes/spacers

Introduction

Eukaryotic cytoplasmic (80S) ribosomes are functionally similar, yet their constituent RNAs display considerable structural diversity, as evidenced particularly by variations in both the size (Loening, 1968; Cammarano *et al.*, 1982; Londei *et al.*, 1982) and number (Ishikawa, 1977; Gray, 1981; Cammarano *et al.*, 1982) of their constituent rRNAs. Homologous large subunit (LSU) and small subunit (SSU) rRNAs range in size between ~3.4 and 5.0 kb (25–28S) and ~1.8 and 2.3 kb (17–19S) respectively. Comparative studies of eukaryotic LSU and SSU rRNAs and their genes have revealed a striking pattern in which highly conserved stretches of primary sequence and secondary structure alternate with variable domains that differ markedly in size, base composition, and potential secondary structure (Stiegler *et al.*, 1981; Veldman *et al.*, 1981; Zwieb *et al.*, 1981; Brimacombe, 1982; Cox and Kelly, 1982; Olsen *et al.*, 1983; Clark *et al.*, 1984; Hadjiolov *et al.*, 1984; Michot *et al.*, 1984; Ellis *et al.*, 1986; Gunderson and Sogin, 1986; Schnare *et al.*, 1986a) even between closely related eukaryotes (Hassouna *et al.*, 1984). The existence of such variable regions accounts in large part for the observed differences in size among homologous SSU and LSU rRNAs.

In most eukaryotes studied to date, the LSU rRNA consists of a non-covalent complex between a low mol. wt RNA (5.8S) and a high mol. wt species (28S), whereas in *E. coli* the LSU rRNA is a single, covalently continuous molecule (23S) (Nazar, 1980, 1982; Cox and Kelly, 1981). (We use '28S rRNA' in a generic sense to designate the homologous RNA species having a sedimentation coefficient ranging from 25S to 28S in different eukaryotes). This difference is attributable to the presence of an internal transcribed spacer (ITS) separating the 5.8S and 28S coding regions in the LSU rRNA gene of almost all eukaryotes (Cox and Kelly, 1981; Nazar, 1982; but see Vossbrinck and Woese, 1986). As a result of excision of this spacer at the level of the primary transcript during post-transcriptional processing (Perry, 1976), a split LSU rRNA is produced. In some insects, additional ITSs interrupt both the 5.8S (Pavlakis *et al.*, 1979; Jordan *et al.*, 1980) and 28S (Delanversin and Jacq, 1983; Ware *et al.*, 1985; Fujiwara and Ishikawa, 1986) rRNA coding regions, so that each of these RNAs is further split. These observations indicate that the distribution of ITSs in rRNA genes is a primary determinant of the number of mature rRNA species in a given organism.

The kingdom Protista (unicellular eukaryotes) is considered to be phylogenetically the most diverse and evolutionarily the most ancient of the four eukaryotic kingdoms defined by Whittaker (1969). Its members should therefore provide valuable insights into the evolution of eukaryotic rRNA and the structural determinants of eukaryotic ribosome function. Members of the order Kinetoplastida (trypanosomatid protozoa) are particularly interesting in this regard, as recent evidence suggests that this group of organisms diverged from the main line of eukaryotes very early in evolution (Schnare *et al.*, 1986a; Sogin *et al.*, 1986). Work in our laboratory (Gray, 1979, 1981; Schnare and Gray, 1982; Schnare *et al.*, 1983) has demonstrated that one trypanosomatid, *Crithidia fasciculata*, possesses a highly unusual ribosome whose large subunit contains two high mol. wt (*c* and *d*) and four low mol. wt (*e*, *f*, *g*, and *j*) RNA species, in addition to 5S (*h*) and 5.8S (*i*) RNAs. Data from other laboratories suggest a very similar spectrum of rRNA components in *Leishmania tarentole* (Simpson and Simpson, 1978), *Trypanosoma brucei* (Cordingley and Turner, 1980) and *T. cruzi* (Hernández *et al.*, 1983), so that this pattern may be characteristic of all members of this order.

We have previously determined the complete sequences of *C. fasciculata* RNA species *e*, *f*, *g* and *j* (Schnare *et al.*, 1983); however, at the time these were published there was insufficient sequence information available from other sources to ascertain whether these novel RNAs had obvious counterparts in the LSU rRNA of other eukaryotes. To explore further the structural relationship between the multiple LSU rRNA components of *C. fasciculata* and the binary 28S:5.8S complex of conventional eukaryotic ribosomes, we have now cloned and sequenced the region encoding the nuclear LSU rRNA gene of *Crithidia*; in addition, we have directly sequenced the 5′ and 3′ ends of RNA species *c* and *d*. These data, together with our previous results

(Schnare and Gray, 1981, 1982; Schnare *et al.*, 1983), demonstrate that the nuclear LSU rRNA gene in *C. fasciculata* is discontinuous, with ITSs separating coding regions for seven distinct RNA species that together constitute the cytoplasmic LSU rRNA of this organism.

Results

Cloning and sequencing of C. fasciculata ribosomal RNA

As shown by hybridization and mapping experiments to be reported elsewhere, the nuclear rRNA genes of *C. fasciculata* (except that for 5S rRNA) are contained within a repeat 11–12 kbp in size that is separated into five fragments by digestion of genomic DNA with *Pst*I. Four of these five rDNA fragments, contiguous in the nuclear DNA, were cloned into pUC9 and the resulting recombinants (pCf1–pCf4) were recovered and characterized by restriction mapping and Southern hybridization. *Pst*I digestion conveniently fractionates *C. fasciculata* rDNA, distributing rRNA coding sequences as follows: *b* (= SSU rRNA) to insert Cf1; *i* (= 5.8S rRNA), *c* and *e* to Cf2; *d* and *f* to Cf3; *j* and *g* to Cf4.

A restriction map of that portion of the *Crithidia* rDNA repeat unit encompassing the LSU rRNA gene (Cf2, Cf3 and part of Cf4) is presented in Figure 1, which shows the relative order of the *Pst*I fragments and the rRNA coding sequences they contain. Also shown are the sequencing strategies for inserts Cf2 and Cf3 and the relevant part of Cf4. The *Pst*I junctions between Cf1 and Cf2 and between Cf2 and Cf3 were confirmed by RNA sequence information, as discussed below. The Cf3–Cf4 junction was verified by partial sequence analysis of an overlapping *Hin*dIII clone (CfH1), as indicated in Figure 1.

The sequences of Cf2, Cf3 and part of Cf4 are shown in Figure 2, which provides the complete primary structure of the LSU rRNA gene and flanking regions. Coding regions were identified by comparison with RNA sequences (dotted underline, Figure 2). The complete sequences of the *C. fasciculata* small rRNAs have been reported previously (Schnare and Gray, 1982; Schnare *et al.*, 1983), as has the 3′-terminal sequence of the SSU rRNA (Schnare and Gray, 1981); in the present study, we also determined the 5′- and 3′-terminal sequences of species *c* and *d* (data supplied to reviewers). The 3′ terminus of *d* displayed length heterogeneity, with molecules ending in either . . . CAC or . . . CACC; in contrast, the 5′ terminus of *d* and both termini of *c* were homogeneous. 5′ End-labelling of species *c* was unaffected by prior phosphatase treatment, indicating the presence of a free 5′-OH in the native molecule (data supplied to reviewers); in contrast, phosphatase treatment markedly enhanced 5′ end-labelling of wheat 26S rRNA (containing the analogous terminus) and of *Crithidia* species *d*, indicating the presence of 5′-P termini in these RNAs.

*Pst*I cleavage sites occur very close to the 3′ end of the *b* coding region (positions 1–6, Figure 2) and the 5′ end of the *d* coding region (positions 3126–3131); the RNA sequences for these regions therefore provide overlaps for the Cf1–Cf2 and Cf2–Cf3 junctions. The combination of DNA and RNA sequence analysis defines the overall organization of the *Crithidia* LSU rRNA gene as one in which sequences encoding identified RNA species are separated by spacer sequences ranging in size from 31 to 416 bp.

Identification of the LSU rRNA species

From consideration of primary and potential secondary structure (to be presented in detail elsewhere) and the relative positions of their coding regions in the rDNA, species *i*, *c*, *d* and *f* were readily identified with portions of the 28S:5.8S rRNA complex of other eukaryotes, as well as parts of the 23S rRNA of *E. coli* (Table I). Species *i*, as shown previously (Schnare and Gray, 1982), is the 5.8S rRNA of *C. fasciculata*; species *c* is equivalent to the 5′ half of mouse 28S rRNA; and species *d* represents most of the 3′ half of the LSU rRNA. However, *d* lacks sequences corresponding to the 3′ terminal ~500 residues of mouse 28S rRNA, including the highly conserved (Chan *et al.*, 1983) α-sarcin cleavage site; instead, this functionally important domain (Wool, 1984) is found in species *f* in *Crithidia* (residues 4770–4783, Figure 2, are identical to the α-sarcin sequence in mouse and other eukaryotic 28S rRNAs).

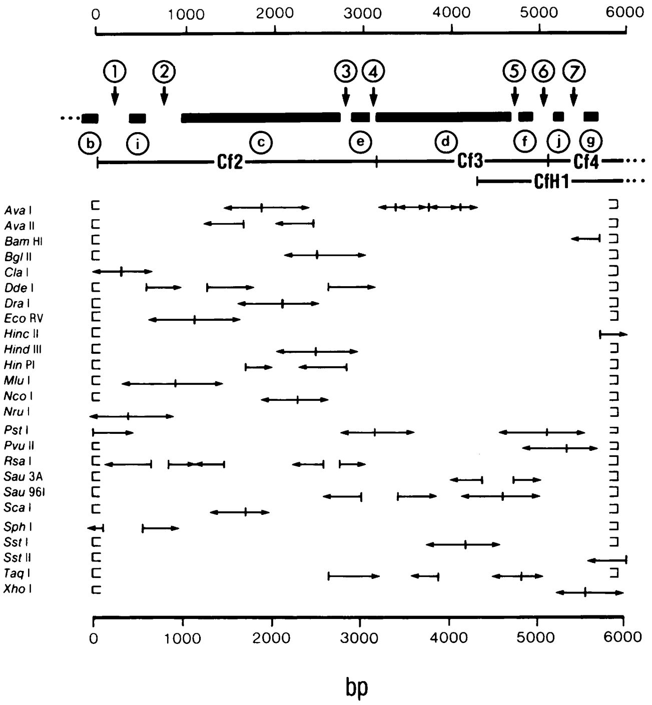
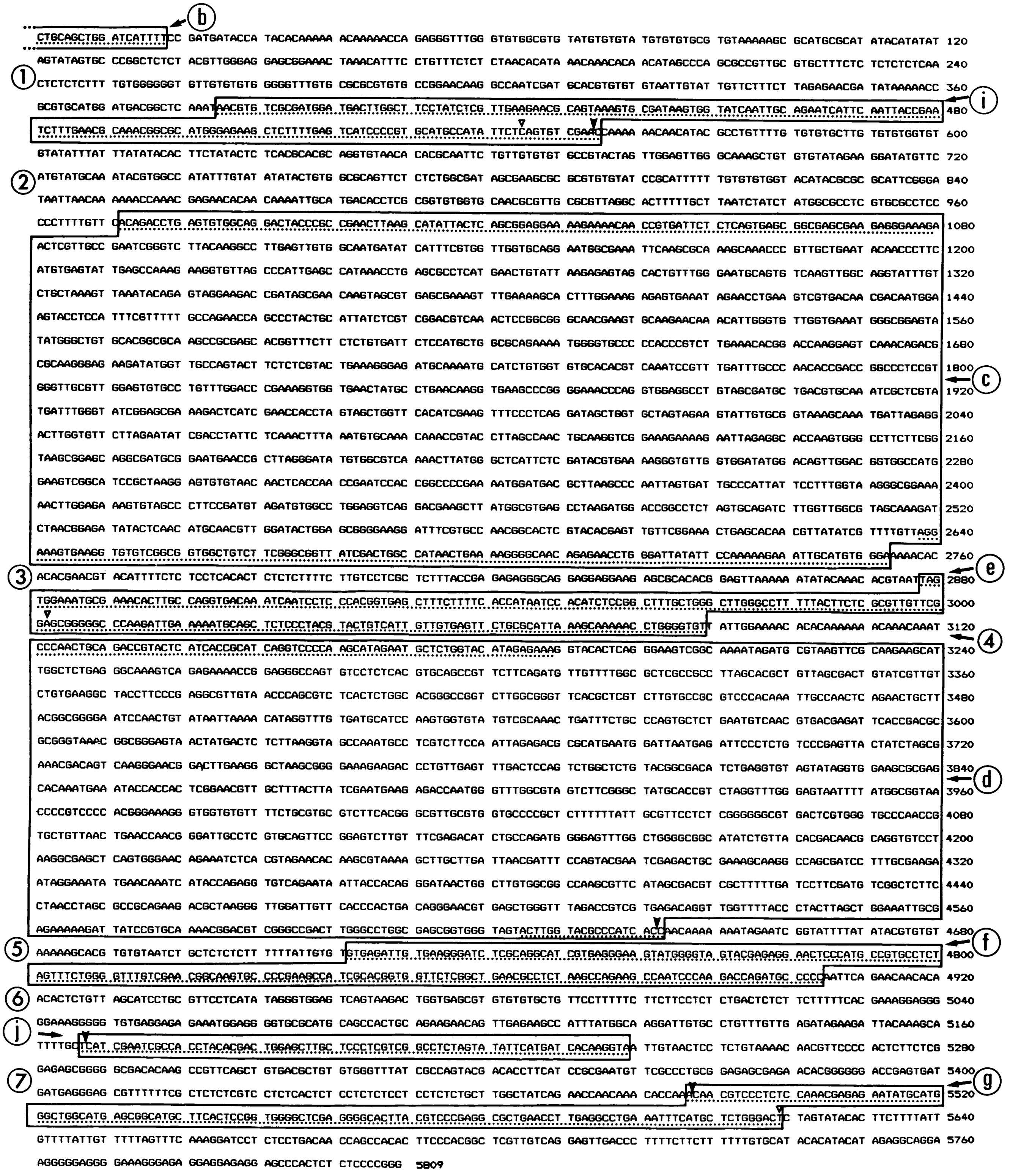


Fig. 1. Organization of the LSU rRNA gene in *C. fasciculata* nuclear DNA. Solid rectangles denote regions encoding the SSU rRNA (*b*), 5.8S rRNA (*i*), and the six segments of the discontinuous 28S rRNA (*c*, *e*, *d*, *f*, *j* and *g*). Circled numbers with vertical arrows indicate the seven internal transcribed spacers (ITS) that separate coding regions. Cloned restriction fragments used in sequencing are designated Cf2, Cf3 and Cf4 (*Pst*I) and CfH1 (*Hin*dIII). Horizontal arrows summarize the sequencing strategy, with restriction sites used listed on the left.

Data bank searches failed to reveal strong homologies between known LSU rRNA sequences and species *e*, *j* and *g*. The *e* coding sequence, located between those of *c* and *d*, maps to a region in eukaryotic 28S rRNA that displays limited primary sequence conservation but whose potential secondary structure is well preserved. In the mouse 28S rRNA secondary structure, this region encompasses helices 41–45 (Michot *et al.*, 1984); the secondary structure shown in Figure 3A can accommodate the homologous region of all published eukaryotic 28S rRNA sequences. Extensive primary sequence homology in this region is only apparent between closely related eukaryotes; however, a computer search for more limited homologies did uncover a 13-nucleotide sequence in *e* that is virtually identical with a sequence (outlined in Figure 3B) contained in helix 43 of known eukaryotic 28S rRNAs. As shown in Figure 3C, an equivalent helix and loop can be constructed with the 13-nucleotide sequence and flanking residues of *e*. Moreover, in a complete secondary structure of *e* (Figure 3C) modelled after that shown in Figure



```
CTGCAGCTGG ATCATTTTCC GATGATACCA TACACAAAAA ACAAAAACCA GAGGGTTTGG GTGTGGCGTG TATGTGTGTA TGTGTGTGCG TGTAAAAAGC GCATGCGCAT ATACATATAT 120
AGTATAGTGC CCGGCTCTCT ACGTTGGGAG GAGCGGAAAC TAAACATTTC CTGTTTCTCT CTAACACATA AACAAACACA ACATAGCCCA GCGCCGTTGC GTGCTTTCTC TCTCTCTCAA 240
CTCTCTCTTT TGTGGGGGGT GTTGTGTGTG GGGGTTTGTG CGCGCGTGTG CCGGAACAAG GCCAATCGAT GCACGTGTGT GTAATTGTAT TGTTCTTTCT TAGAGAACGA TATAAAAACC 360
GCGTGCATGG ATGACGGCTC AAATAACGTG TCGCGATGGA TGACTTGGCT TCCTATCTCG TTGAAGAACG CAGTAAAGTG CGATAAGTGG TATCAATTGC AGAATCATTC AATTACCGAA 480
TCTTTGAACG CAAACGGCGC ATGGGAGAAG CTCTTTTGAG TCATCCCCGT GCATGCCATA TTCTCAGTGT CGAACCAAAA AACAACATAC GCCTGTTTTG TGTGTGCTTG TGTGTGGTGT 600
GTATATTTAT TTATATACAC TTCTATACTC TCACGCACGC AGGTGTAACA CACGCAATTC TGTTGTGTGT GCCGTACTAG TTGGAGTTGG GCAAAGCTGT GTGTATAGAA GGATATGTTC 720
ATGTATGCAA ATACGTGGCC ATATTTGTAT ATATACTGTG GCGCAGTTCT CTCTGGCGAT AGCGAAGCGC GCGTGTGTAT CCGCATTTTT TGTGTGTGGT ACATACGCGC GCATTCGGGA 840
TAATTAACAA AAAACCAAAC GAGAACACAA CAAAATTGCA TGACACCTCG CGGTGTGGTG CAACGCGTTG CGCGTTAGGC ACTTTTTGCT TAATCTATCT ATGGCGCCTC GTGCGCCTCC 960
CCCTTTTGTT CACAGACCTG AGTGTGGCAG GACTACCCGC CGAACTTAAG CATATTACTC AGCGGAGGAA AAGAAAACAA CCGTGATTCT CTCAGTGAGC GGCGAGCGAA GAGGGAAAGA 1080
ACTCGTTGCC GAATCGGGTC TTACAAGGCC TTGAGTTGTG GCAATGATAT CATTTCGTGG TTGGTGCAGG AATGGCGAAA TTCAAGCGCA AAGCAAACCC GTTGCTGAAT ACAACCCTTC 1200
ATGTGAGTAT TGAGCCAAAG AAGGTGTTAG CCCATTGAGC CATAAACCTG AGCGCCTCAT GAACTGTATT AAGAGAGTAG CACTGTTTGG GAATGCAGTG TCAAGTTGGC AGGTATTTGT 1320
CTGCTAAAGT TAAATACAGA GTAGGAAGAC CGATAGCGAA CAAGTAGCGT GAGCGAAAGT TTGAAAAGCA CTTTGGAAAG AGAGTGAAAT AGAACCTGAA GTCGTGACAA CGACAATGGA 1440
AGTACCTCCA TTTCGTTTTT GCCAGAACCA GCCCTACTGC ATTATCTCGT CGGACGTCAA ACTCCGGCGG GCAACGAAGT GCAAGAACAA ACATTGGGTG TTGGTGAAAT GGGCGGAGTA 1560
TATGGGCTGT GCACGGCGCA AGCCGCGAGC ACGGTTTCTT CTCTGTGATT CTCCATGCTG GCGCAGAAAA TGGGGTGCCC CCACCCGTCT TGAAACACGG ACCAAGGAGT CAAACAGACG 1680
CGCAAGGGAG AAGATATGGT TGCCAGTACT TCTCTCGTAC TGAAAGGGAG ATGCAAAATG CATCTGTGGT GTGCACACGT CAAATCCGTT TGATTTGCCC AACACCGACC GGCCCTCCGT 1800
GGGTTGCGTT GGAGTGTGCC TGTTTGGACC CGAAAGGTGG TGAACTATGC CTGAACAAGG TGAAGCCCGG GGAAACCCAG GTGGAGGCCT GTAGCGATGC TGACGTGCAA ATCGCTCGTA 1920
TGATTTGGGT ATCGGAGCGA AAGACTCATC GAACCACCTA GTAGCTGGTT CACATCGAAG TTTCCCTCAG GATAGCTGGT GCTAGTAGAA GTATTGTGCG GTAAAGCAAA TGATTAGAGG 2040
ACTTGGTGTT CTTAGAATAT CGACCTATTC TCAAACTTTA AATGTGCAAA CAAACCGTAC CTTAGCCAAC TGCAAGGTCG GAAAGAAAAG AATTAGAGGC ACCAAGTGGG CCTTCTTCGG 2160
TAAGCGGAGC AGGCGATGCG GAATGAACCG CTTAGGGATA TGTGGCGTCA AAACTTATGG GCTCATTCTC GATACGTGAA AAGGGTGTTG GTGGATATGG ACAGTTGGAC GGTGGCCATG 2280
GAAGTCGGCA TCCGCTAAGG AGTGTGTAAC AACTCACCAA CCGAATCCAC CGGCCCCGAA AATGGATGAC GCTTAAGCCC AATTAGTGAT TGCCCATTAT TCCTTTGGTA AGGGCGGAAA 2400
AACTTGGAGA AAGTGTAGCC CTTCCGATGT AGATGTGGCC TGGAGGTCAG GACGAAGCTT ATGGCGTGAG CCTAAGATGG ACCGGCCTCT AGTGCAGATC TTGGTTGGCG TAGCAAAGAT 2520
CTAACGGAGA TATACTCAAC ATGCAACGTT GGATACTGGA GCGGGGAAGG ATTTCGTGCC AACGGCACTC GTACACGAGT TGTTCGGAAA CTGAGCACAA CGTTATATCG TTTTGTTAGG 2640
AAAGTGAAGG TGTGTCGGCG GTGGCTGTCT TCGGGCGGTT ATCGACTGGC CATAACTGAA AAGGGGCAAC AGAGAACCTG GGATTATATT CCAAAAAGAA ATTGCATGTG GGAAAAACAC 2760
ACACGAACGT ACATTTTCTC TCCTCACACT CTCTCTTTTC TTGTCCTCGC TCTTTACCGA GAGAGGGCAG GAGGAGGAAG AGCGCACACG GAGTTAAAAA ATATACAAAC ACGTAATTAG 2880
TGGAAATGCG AAACACTTGC CAGGTGACAA ATCAATCCTC CCACGGTGAG CTTTCTTTTC ACCATAATCC ACATCTCCGG CTTTGCTGGG CTTGGGCCTT TTTACTTCTC GCGTTGTTCG 3000
GAGCGGGGGC CCAAGATTGA AAAATGCAGC TCTCCCTACG TACTGTCATT GTTGTGAGTT CTGCGCATTA AAGCAAAAAC CTGGGGTGTT ATTGGAAAAC ACACAAAAAA ACAAACAAAT 3120
CCCAACTGCA GACCGTACTC ATCACCGCAT CAGGTCCCCA AGCATAGAAT GCTCTGGTAC ATAGAGAAAG GTACACTCAG GGAAGTCGGC AAAATAGATG CGTAAGTTCG CAAGAAGCAT 3240
TGGCTCTGAG GGCAAAGTCA GAGAAAACCG GAGGGCCAGT GTCCTCTCAC GTGCAGCCGT TCTTCAGATG TTGTTTTGGC GCTCGCCGCC TTAGCACGCT GTTAGCGACT GTATCGTTGT 3360
CTGTGAAGGC TACCTTCCCG AGGCGTTGTA ACCCAGCGTC TCACTCTGGC ACGGGCCGGT CTTGGCGGGT TCACGCTCGT CTTGTGCCGC GTCCCACAAA TTGCCAACTC AGAACTGCTT 3480
ACGGCGGGGA ATCCAACTGT ATAATTAAAA CATAGGTTTG TGATGCATCC AAGTGGTGTA TGTCGCAAAC TGATTTCTGC CCAGTGCTCT GAATGTCAAC GTGACGAGAT TCACCGACGC 3600
GCGGGTAAAC GGCGGGAGTA ACTATGACTC TCTTAAGGTA GCCAAATGCC TCGTCTTCCA ATTAGAGACG CGCATGAATG GATTAATGAG ATTCCCTCTG TCCCGAGTTA CTATCTAGCG 3720
AAACGACAGT CAAGGGAACG GACTTGAAGG GCTAAGCGGG GAAAGAAGAC CCTGTTGAGT TTGACTCCAG TCTGGCTCTG TACGGCGACA TCTGAGGTGT AGTATAGGTG GAAGCGCGAG 3840
CACAAATGAA ATACCACCAC TCGGAACGTT GCTTTACTTA TCGAATGAAG AGACCAATGG GTTTGGCGTA GTCTTCGGGC TATGCACCGT CTAGGTTTGG GAGTAATTTT ATGGCGGTAA 3960
CCCCGTCCCC ACGGGAAAGG GTGGTGTGTT TTCTGCGTGC GTCTTCACGG GCGTTGCGTG GTGCCCCGCT CTTTTTTATT GCGTTCCTCT CGGGGGGCGT GACTCGTGGG TGCCCAACCG 4080
TGCTGTTAAC TGAACCAACG GGATTGCCTC GTGCAGTTCC GGAGTCTTGT TTCGAGACAT CTGCCAGATG GGGAGTTTGG CTGGGGCGGC ATATCTGTTA CACGACAACG CAGGTGTCCT 4200
AAGGCGAGCT CAGTGGGAAC AGAAATCTCA CGTAGAACAC AAGCGTAAAA GCTTGCTTGA TTAACGATTT CCAGTACGAA TCGAGACTGC GAAAGCAAGG CCAGCGATCC TTTGCGAAGA 4320
ATAGGAAATA TGAACAAATC ATACCAGAGG TGTCAGAATA ATTACCACAG GGATAACTGG CTTGTGGCGG CCAAGCGTTC ATAGCGACGT CGCTTTTTGA TCCTTCGATG TCGGCTCTTC 4440
CTAACCTAGC GCCGCAGAAG ACGCTAAGGG TTGGATTGTT CACCCACTGA CAGGGAACGT GAGCTGGGTT TAGACCGTCG TGAGACAGGT TGGTTTTACC CTACTTAGCT GGAAATTGCG 4560
AGAAAAAGAT TATCCGTGCA AAACGGACGT CGGGCCGACT TGGGCCTGGC GAGCGGTGGG TAGTACTTGG TACGCCCATC ACCAACAAAA AAATAGAATC GGTATTTTAT ATACGTGTGT 4680
AAAAAGCACG TGTGTAATCT GCTCTCTCTT TTTTATTGTG TGTGAGATTG TGAAGGGATC TCGCAGGCAT CGTGAGGGAA GTATGGGGTA GTACGAGAGG AACTCCCATG CCGTGCCTCT 4800
AGTTTCTGGG GTTTGTCGAA CGGCAAGTGC CCCGAAGCCA TCGCACGGTG GTTCTCGGCT GAACGCCTCT AAGCCAGAAG CCAATCCCAA GACCAGATGC CCCCAATTCA GAACAACACA 4920
ACACTCTGTT AGCATCCTGC GTTCCTCATA TAGGGTGGAG TCAGTAAGAC TGGTGAGCGT GTGTGTGCTG TTCCTTTTTC TTCTTCCTCT CTGACTCTCT TCTTTTTCAC GAAAGGAGGG 5040
GGAAAGGGGG TGTGAGGAGA GAAATGGAGG GGTGCGCATG CAGCCACTGC AGAAGAACAG TTGAGAAGCC ATTTATGGCA AGGATTGTGC CTGTTTGTTG AGATAGAAGA TTACAAAGCA 5160
TTTTGCTCAT CGAATCGCCA CCTACACGAC TGGAGCTTGC TCCCTCGTCG GCCTCTAGTA TATTCATGAT CACAAGGTAA TTGTAACTCC TCTGTAAAAC AACGTTCCCC ACTCTTCTCG 5280
GAGAGCGGGG GCGACACAAG CCGTTCAGCT GTGACGCTGT GTGGGTTTAT CGCCAGTACG ACACCTTCAT CCGCGAATGT TCGCCCTGCG GAGAGCGAGA ACACGGGGGG ACCGAGTGAT 5400
GATGAGGGAG CGTTTTTTCG CTCTCTCGTC CTCTCACTCT CCTCTCTCCT CCTCTCTGCT TGGCTATCAG AACCAACAAA CACCAACCAA CGTCCCTCTC CAAACGAGAG AATATGCATG 5520
GGCTGGCATG AGCGGCATGC TTCACTCCGG TGGGGCTCGA GGGGCACTTA CGTCCCGAGG CGCTGAACCT TGAGGCCTGA AATTTCATGC TCTGGGACTC TAGTATACAC TTCTTTTATT 5640
GTTTTATTGT TTTTAGTTTC AAAGGATCCT CTCCTGACAA CCAGCCACAC TTCCCACGGC TCGTTGTCAG GAGTTGACCC TTTTCTTCTT TTTTGTGCAT ACACATACAT AGAGGCAGGA 5760
AGGGGGAGGG GAAAGGGAGA GGAGGAGAGG AGCCCACTCT CTCCCCGGG  5809
```

Fig. 2. Complete DNA sequence of the *C. fasciculata* LSU rRNA gene. Circled lower-case letters denote coding regions, which are also outlined; circled numbers denote ITS. The dotted underline indicates sequences also determined at the RNA level. Vertical arrowheads denote additional cleavage sites that result in length heterogeneity at the 5′ termini of *j* and *g* and the 3′ termini of *i* and *d*. The DNA sequence agrees with RNA sequences we have previously published except at three positions (indicated by open triangles): (i) one C (C545) in the *i* coding sequence versus two Cs (C161 and C162) in the RNA sequence of species *i* (Schnare and Gray, 1982) (we consider the DNA sequence to be correct here); (ii) A3002 in the *e* coding sequence versus Ψ125 in the RNA sequence of species *e* (Schnare *et al.*, 1983) (the unusual chemical reactivity of residue 125 in *e*, originally interpreted as indicative of Ψ, now suggests the presence of a post-transcriptionally modified A residue at this position); and (iii) three additional 3′ terminal A residues in the RNA sequence of species *g* (Schnare *et al.*, 1983), not present at the end of the *g* coding region (these residues must therefore be added post-transcriptionally; see text).

Table I. Identities of the LSU rRNA species of *C. fasciculata*

Species	Coding region		Equivalent nucleotide positions in:	
	Nucleotides	% G + C	*E. coli* 23S rRNA	Mouse 28S rRNA
i	171	46.2	13–158	[5.8S rRNA]
c	1782	49.5	168–1415	1–2302
e	212	49.1	1420–1578	2325–2509
d	1523	52.2	1587–2625	2520–4222
f	183	57.9	2630–2788	4226–4378
j	73	50.7	–	(4379–4616)?
g	133[a]	57.9	(2810–2904)	(4617–4712)
Total	4077	51.0		

[a] The corresponding RNA species contains an additional three nucleotides not present in its coding sequence.

3A, there is quite reasonable correspondence in both primary sequence and secondary structure in those sub-regions that are the most highly conserved among eukaryotic 28S rRNAs, particularly helices 41 and 43. Comparing 99 positions that can be reliably aligned, the *Crithidia* sequence is 52–57% identical with the homologous region from yeast (*Saccharomyces cerevisiae*), rice, *Caenorhabditis elegans* or mouse 28S rRNAs. At 36 positions (36%), the same nucleotide occurs in all five sequences. Structural homology actually extends into the D7b region, with residues 3–12 in *e* having the potential to form a small helix (solid triangle, Figure 3C) that is conserved at a comparable position in eukaryotic 28S rRNAs. On the basis of these comparisons, we conclude that species *e* is the structural equivalent, albeit considerably diverged, of the region containing helices 41–44 in eukaryotic 28S rRNAs.

The position of the *g* coding region as the last in the direction of transcription suggests that it ought to be the structural equivalent of the 3′-terminal region of 28S rRNA; however, an exhaustive search failed to uncover any convincing, consistent primary sequence homologies of more than a few nucleotides between *g* and the 3′-terminal region of eukaryotic 28S rRNAs. [It should be noted that the 3′-terminal region of eukaryotic 28S rRNAs displays rather limited primary sequence conservation (and no consistent secondary structure) in broad comparisons (e.g. rat versus yeast; Chan *et al.*, 1983), with extensive homology restricted to groups of closely related organisms (e.g. mouse, human and *Xenopus*).] On the other hand, we did detect primary sequence similarity between species *g* and the 3′-terminal region of *E. coli* 23S rRNA. The alignment in Figure 4 shows a 22-nucleotide stretch (residues 98–119) in *g* that is 77% identical (17 matches out of 22) with the 3′-terminal 21 residues of *E. coli* 23S rRNA, and 50% identical (11/22) with a stretch close to the 3′ end of the mouse 28S rRNA coding sequence (positions 4683–4702). Within the 15-nucleotide stretch outlined in Figure 4, even greater similarity is evident: 87% with *E. coli* and 67% with mouse, with nine positions (60%) identical in all three sequences. On this basis, we tentatively conclude that *g* represents the 3′ end of the *Crithidia* LSU rRNA. As noted previously (see Figure 2), species *g* contains three 3′-terminal A residues that do not appear in the *g* coding sequence, and so must be added during post-transcriptional processing. Post-transcriptional 3′-oligoadenylation is also a feature of animal mitochondrial LSU rRNA (Dubin *et al.*, 1982), an observation that further supports the idea that *g* constitutes the 3′ end of the *Crithidia* LSU rRNA.

The *j* coding region lies between those of *f* and *g*, and therefore maps to a variable region (D12) in the secondary structure of eukaryotic 28S rRNA. Species *j* contains a nine-nucleotide stretch (GUCGGCCUC, residues 41–49) that is also present in the D12 region of *Xenopus laevis* 28S rRNA (residues 3878–3886) (Ware *et al.*, 1983; Clark *et al.*, 1984). However, an exhaustive computer search failed to find this sequence or a close analogue of it in the D12 region of other eukaryotic 28S rRNAs, nor did such a search reveal any other common motif(s) between *j* and the eukaryotic D12 region. At present, therefore, we cannot unambiguously equate *j* with any particular portion of eukaryotic 28S rRNA, even though it is a *bone fide* component of the large subunit of the *Crithidia* ribosome (Gray, 1981).

Internal transcribed spacers and post-transcriptional processing

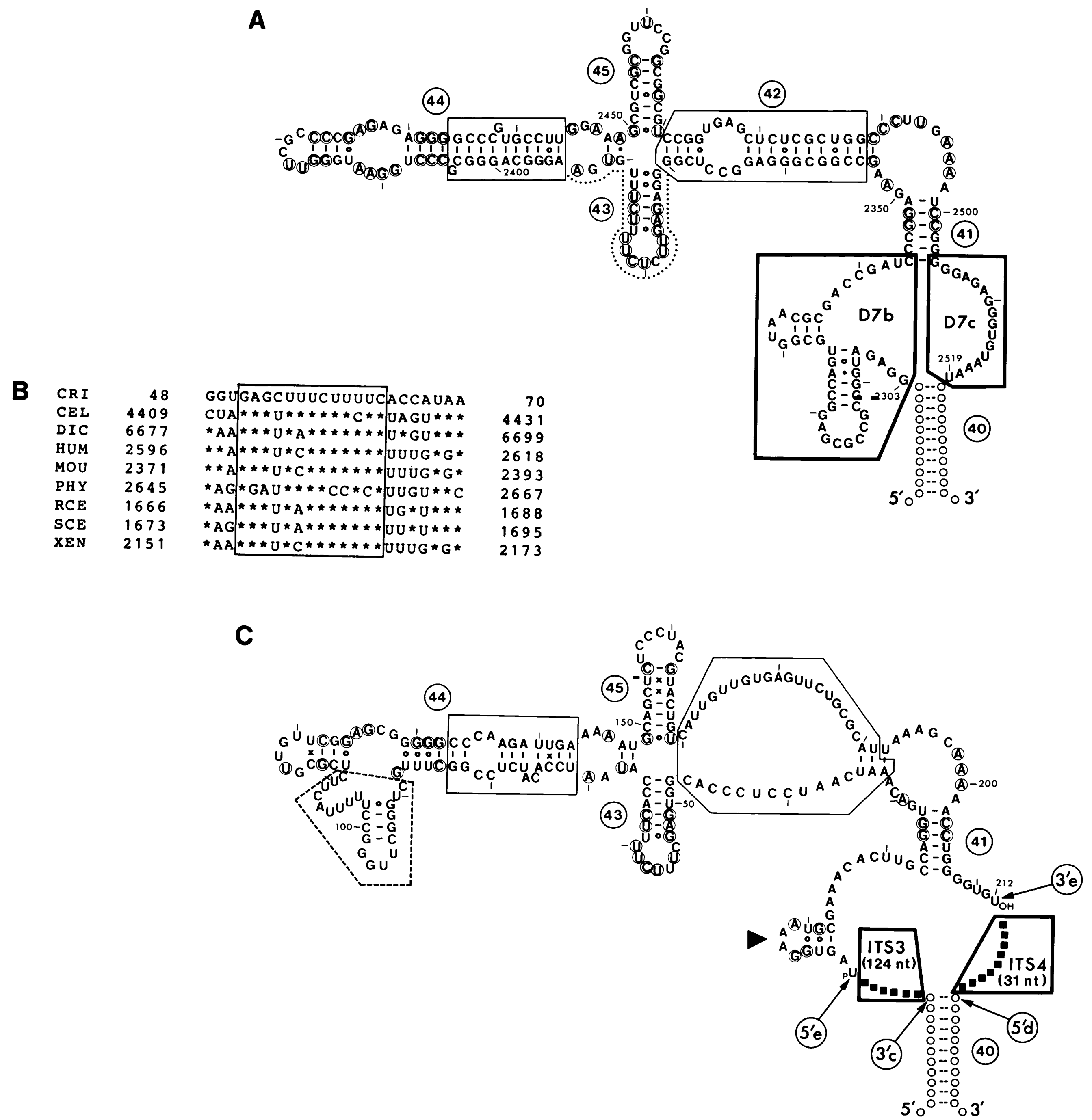
The structure of the *Crithidia* LSU rRNA gene implies that a primary transcript from this region must undergo extensive post-transcriptional processing, involving endonucleolytic removal of sequences corresponding to seven ITSs. Primary sequence and/or secondary structure signals that may specify these cleavages are of obvious interest. Certain coding/spacer boundaries (e.g. 3′-*i*/ITS2; 3′-*d*/ITS5; 3′-*f*/ITS6; ITS7/5′-*g*) are located within A + C-rich stretches; such runs also occur just upstream (ITS3/5′-*e*) or downstream (3′-*b*/ITS1; 3′-*e*/ITS4) in several other cases. The smallest ITS, ITS4, is 81% A + C. However, with one exception, the actual primary sequence at or in the vicinity of the various coding/spacer boundaries is not conserved. The exception involves the *e* coding region, a transcript of which would be flanked by imperfect direct repeats, UUA(G,U)UGGAAA; the cleavages that produce the 5′ and 3′ termini of *e* occur at precisely the same site within these repeats (see Figure 5B).

In several cases, higher-order structure appears to play a role in directing endonucleolytic cleavage of the primary transcript. As shown in Figure 5A, the 3′-*i*/ITS2 and ITS2/5′-*c* boundaries occur within short A + C-rich single-strand sequences located just at the base of a long helix; a comparable helix, which pairs the 3′ end of 5.8S rRNA with the 5′ end of 28S rRNA (helix 1 in the mouse model; Michot *et al.*, 1984), is characteristic of other eukaryotic 28S rRNAs. The 3′-*c*/ITS3 and ITS4/5′-*d* boundaries also occur within A+C-rich single-strand sequences located at the base of a helix joining the 3′ end of *c* to the 5′ end of *d* (Figure 5B). This helix corresponds to helix 40 in the mouse model (Michot *et al.*, 1984) and is conserved in both eukaryotic 28S and eubacterial 23S rRNAs.

Variable regions and spacers in eukaryotic LSU rRNA

In Figure 6, we have outlined those portions of the *E. coli* 23S rRNA secondary structure (Noller *et al.*, 1981) that are equivalent to the various *Crithidia* LSU rRNA species described here. It can be seen that the *Crithidia* RNAs account for essentially all of the *E. coli* 23S structure. Moreover, *i* and *d* each have the potential of base pairing with *c* to reconstitute helices that are formed by intramolecular pairing in *E. coli* 23S rRNA. Species *e*, *f*, *j* and *g* are equivalent to isolated domains that are not obviously tied to the rest of the secondary structure by long-range hydrogen bonding interactions.

Figure 6 also summarizes the location of the major variable regions (D1, D2, etc.) in eukaryotic 28S rRNA, designated as in Michot *et al.* (1984) As well, the arrows show the locations of discontinuities that have been found to date in eukaryotic LSU rRNAs. Each of these discontinuities is created by excision of an ITS, and in all cases, these map to variable regions in the secondary structure. In the present study, we have identified an additional variable region, designated D7c in Figure 6 (see also



CRI	48	GGU	GAGCUUUCUUUUC	ACCAUAA	70
CEL	4409	CUA	***U******C**	UAGU***	4431
DIC	6677	*AA	***U*A*******	U*GU***	6699
HUM	2596	**A	***U*C*******	UUUG*G*	2618
MOU	2371	**A	***U*C*******	UUUG*G*	2393
PHY	2645	*AG	*GAU****CC*C*	UUGU**C	2667
RCE	1666	*AA	***U*A*******	UG*U***	1688
SCE	1673	*AG	***U*A*******	UU*U***	1695
XEN	2151	*AA	***U*C*******	UUUG*G*	2173

Fig. 3. (A) Potential secondary structure of the central region (encompassing helices 41–45; circled numbers) of mouse 28S rRNA, numbered as in and slightly modified from Michot *et al.* (1984). The two thinly outlined regions display minor variation in length and potential secondary structure, and are not well conserved in sequence, among eukaryotic 28S rRNAs; heavily outlined regions (D7b and D7c) vary considerably in length and structure in 28S rRNAs. Outside of the outlined blocks, both primary sequence and potential secondary structure are well conserved, with circled residues identical in the 28S rRNAs of mouse (MOU; Hassouna *et al.*, 1984), *Caenorhabditis elegans* (CEL; Ellis *et al.*, 1986), rice (RCE; Takaiwa *et al.*, 1985) and yeast, *Saccharomyces cerevisiae* (SCE; Georgiev *et al.*, 1981). (**B**) Comparison of the highly conserved region encompassing helix 43 [dotted line, (A)]. Residues identical between *Crithidia* species *e* (CRI) and other sequences are denoted by an asterisk [the *e* sequence is numbered as in (C)]. Additional abbreviations and references: DIC, *Dictyostelium discoideum* (Ozaki *et al.*, 1984); HUM, human (Gonzalez *et al.*, 1985); PHY, *Physarum polycephalum* (Otsuka *et al.*, 1983); XEN, *Xenopus laevis* (Ware *et al.*, 1983). (**C**) Potential secondary structure of *Crithidia* species *e*, modelled after that shown in (A) (base pairs not possible in the *e* secondary structure are indicated by 'x'). Circled residues are identical in the CRI, MOU, CEL, RCE and SCE sequences. The region outlined by dashes is 20 nucleotides longer in *e* than in conventional eukaryotic 28S rRNAs [cf. (A)]. Helix 40, in which the 3' end of *c* is base paired with the 5' end of *d*, is denoted by open circles, as is the corresponding helix in (A). The heavily outlined regions indicate the positions of ITS separating the coding region of *e* from those of *c* and *d* (see Figure 2); termini resulting from spacer excision are designated by arrows (see also Figure 5B). The solid triangle denotes a conserved helix present in the D7b region of eukaryotic 28S rRNA [residues 2327–2336, (A)].

Figure 3A), that ranges in size from 13 to 21 nucleotides in different eukaryotic 28S rRNAs. *Crithidia* ITS4 maps to D7c. In a parallel study, we have examined the position of the break separating the two halves of the split 28S rRNA of an amoeboid protozoan, *Acanthamoeba castellanii* (Stevens and Pachler, 1972); this discontinuity also maps to D7c (M.N.Schnare, un-

```
MOU 28S    4683   ...UGAA-AGUCAGCCCU-CGACACAAGGGUUUGU          4712
                     ****  * ** * ***  *      * * * **
CRI g        98   ...UGAACCUUGAGGCCUGAAAUUUCAUGCUCUGGGACUaaa    136
                     ****** ****** *  **  **
ECO 23S    2884   ...UGAACCGUGAGGCUU-AACCUU                    2904
                     xxxx   x xx x x
```

Fig. 4. Comparison of the 3′ terminal sequences of mouse (MOU) 28S rRNA, *Crithidia* (CRI) species *g*, and *E. coli* (ECO) 23S rRNA. Residues identical between the CRI and either the MOU or ECO sequences are denoted by an asterisk, while residues identical in all three sequences are marked by 'x'. Lower-case letters at the end of the *g* sequence indicate the three A residues added post-transcriptionally (see text and legend to Figure 2).

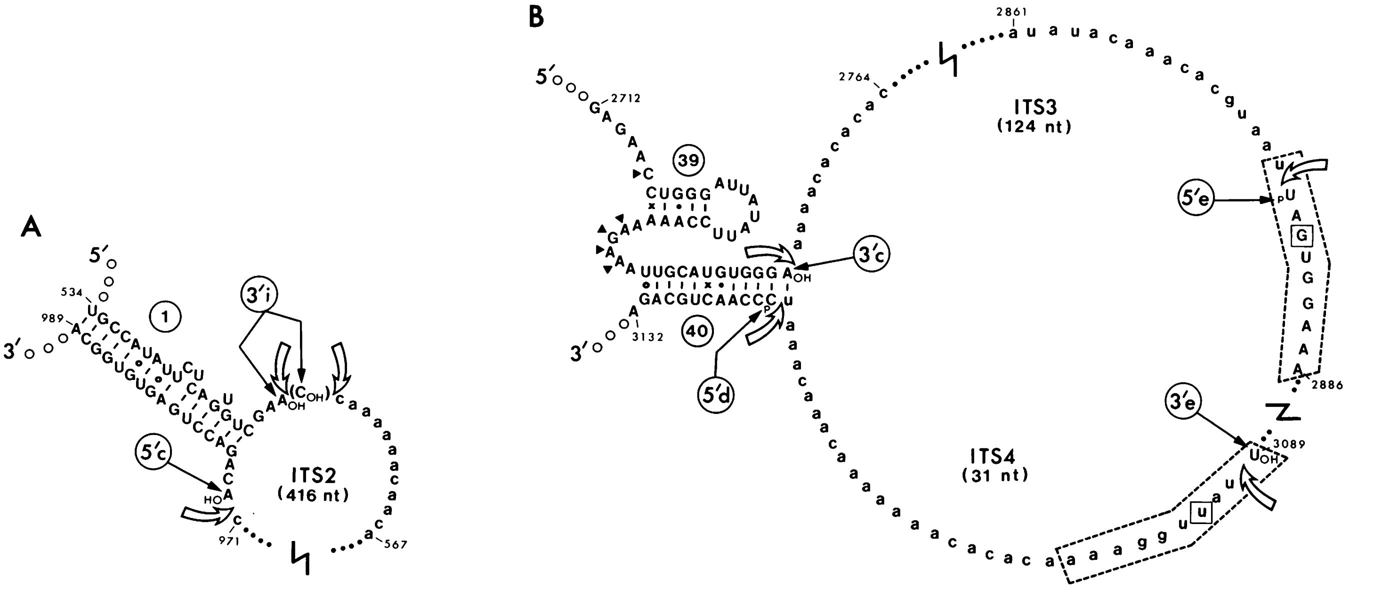


Fig. 5. Potential secondary structure in the vicinity of cleavage sites between coding and spacer sequences. Upper-case letters, coding sequences; lower-case letters, spacer sequences (residues are numbered as in Figure 2). Open curved arrows show putative sites of post-transcriptional cleavage, deduced from comparison of RNA and DNA sequences, while solid arrows designate the resulting RNA termini. Circled numbers refer to conserved helices in the secondary structure (designated as in Michot *et al.*, 1984). (**A**) Cleavage sites flanking ITS2 (note that the 3′ terminus of *i* is heterogeneous, with some molecules having an extra C). (**B**) Cleavage sites flanking ITS3 and ITS4. Solid triangles indicate additional residues in the *Crithidia* sequence, relative to other eukaryotic 28S rRNAs. Dashes enclose the 10-nucleotide imperfect direct repeats that flank the *e* coding region (the single nucleotide difference between these repeats is boxed).

published results). These results emphasize a strong correlation between ITSs in rDNA and variable regions in rRNA.

Discussion

The results reported here have a number of implications for our understanding of rRNA structure, function and evolution. In the first place, they strongly reinforce the idea that a ribosomal RNA molecule need not be a single, covalently continuous polynucleotide chain in order to function. In most eukaryotic ribosomes examined to date, a binary complex between 5.8S and 28S rRNAs is the structural equivalent of the single LSU (23S) rRNA of *E. coli*. In *C. fasciculata*, however, the LSU rRNA appears to be a complex of seven separate RNA components, at least five (and perhaps six) of which can be identified with specific regions in eukaryotic 28S and/or eubacterial 23S rRNAs. This structural equivalence implies functional equivalence, whether or not the domains in question are covalently continuous with the rest of the LSU rRNA.

Species *f*, one of the small RNAs present in the large ribosomal subunit of *Crithidia*, is particularly interesting because it contains a highly conserved sequence (the α-sarcin site) that is critical to ribosome function. Protein synthesis is inhibited when this sequence is cleaved by any of several fungal cytotoxins, including α-sarcin (Schindler and Davies, 1977; Veldman *et al.*, 1981; Endo and Wool, 1982), restrictocin and mitogillin (Fando *et al.*, 1985). Although the precise function of the α-sarcin domain remains to be elucidated, it must be of primary importance in ribosome function because hydrolysis of a single phosphodiester bond within the α-sarcin site inactivates the ribosome (Wool, 1984). Because the α-sarcin site is found exclusively in species *f* in the *Crithidia* ribosome, the domain containing it must be able to function in the form of a separate small RNA, as well as when covalently integrated into a larger molecule.

Secondary structure modelling (to be presented in detail elsewhere) indicates that the *Crithidia* LSU rRNA complex is held together in part by long-range intermolecular base pairing interactions (e.g. between *i* and *c* and between *c* and *d*) that reproduce *intra*molecular interactions that occur in the covalently continuous *E. coli* LSU (23S) rRNA (see Figure 6). The fact that species *i*, *c* and *d* are found together in the form of a stable complex after phenol extraction of *Crithidia* rRNA at low temperature (0–4°C) (Gray, 1979, 1981) supports the existence of extensive base pairing among them. Under the same extraction conditions, species *e*, *f*, *j* and *g* are not found complexed with other LSU rRNA components; indeed, in secondary structure representations of eukaryotic 28S or eubacterial 23S rRNA, the regions encompassing species *e*, *f*, *j* and *g* appear as isolated domains, not obviously base paired with the rest of the molecule. Tertiary interactions may well occur between these small RNAs and other LSU rRNA components; in this regard, it is notewor-

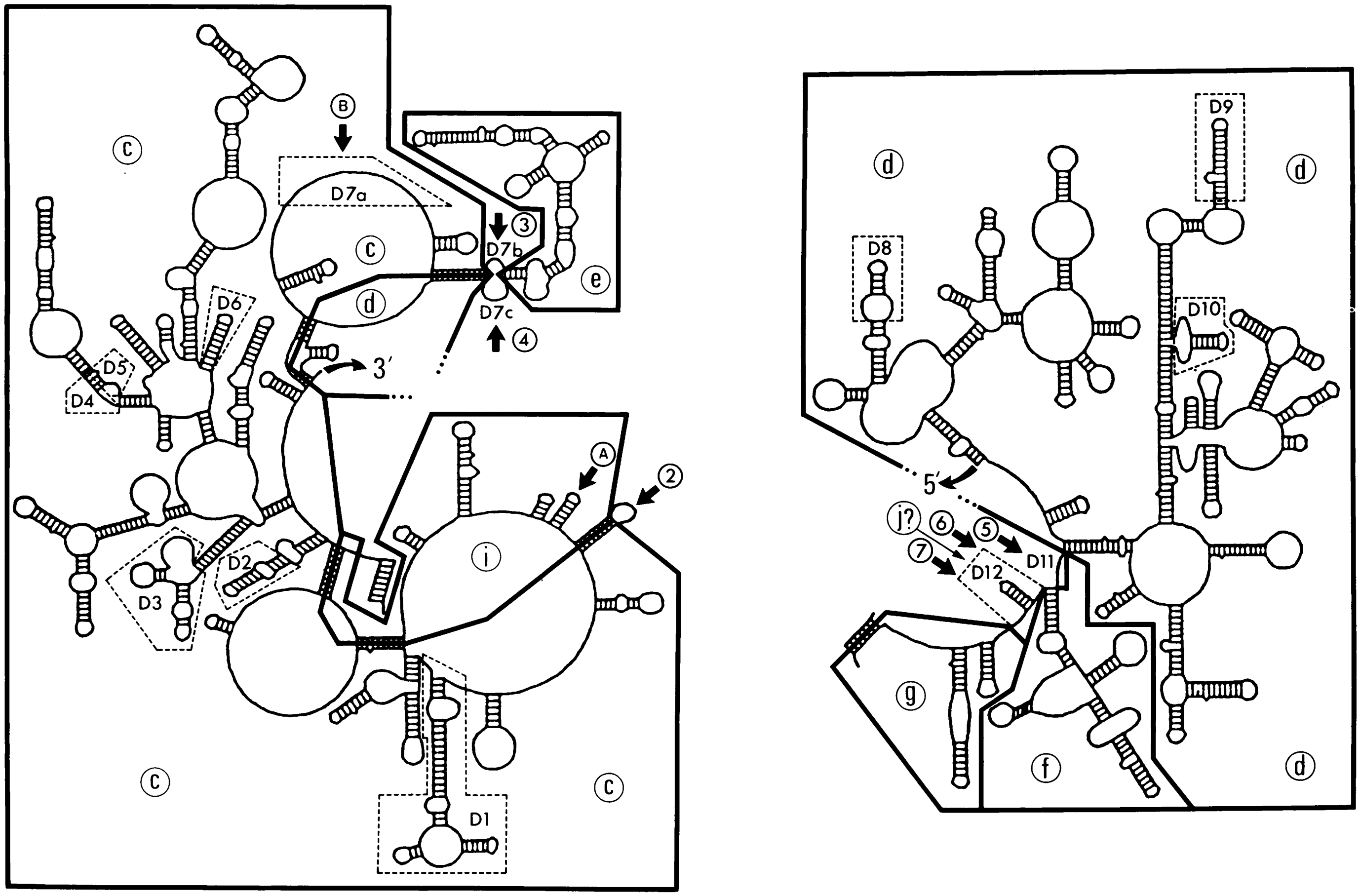


Fig. 6. Mapping of *Crithidia* coding and spacer regions to the secondary structure model of *E. coli* 23S rRNA (Noller *et al.*, 1981; see also Hogan *et al.*, 1984). Heavy lines enclose those portions of the *E. coli* secondary structure that correspond to the various *Crithidia* LSU rRNA species (indicated by circled lower-case letters). Dashed lines enclose regions that show pronounced structural differences between eukaryotic 28S and eubacterial 23S rRNAs, and also vary extensively among eukaryotic 28S rRNAs; these variable regions are designated as in Michot *et al.* (1984) (see text for a description of D7c). Circled numbers with arrows indicate the positions of ITS whose post-transcriptional excision creates discontinuities in *Crithidia* LSU rRNA. Circled upper-case letters with arrows indicate additional discontinuities created by spacer excision in 5.8S rRNA Ⓐ (Pavlakis *et al.*, 1979; Jordan *et al.*, 1980) and 28S rRNA Ⓑ (Delanversin and Jacq, 1983; Ware *et al.*, 1985; Fujiwara and Ishikawa, 1986).

thy that despite the isolation of the α-sarcin domain in secondary structure models of eukaryotic 28S rRNA, cleavage of the α-sarcin site causes a dramatic rearrangement of higher-order structure in the 60S subunit, leading to dissociation of 5.8S rRNA from its complex with 28S rRNA (Walker *et al.*, 1983). Choi (1985) has suggested that a tertiary interaction may occur between the α-sarcin site itself and 5.8S rRNA. However, any tertiary base pairing interactions between *Crithidia* species *e*, *f*, *j* or *g* and other LSU rRNA components cannot be extensive, because they do not survive low-temperature phenol extraction; if such tertiary interactions do occur in the ribosome, they may be stabilized by ribosomal proteins and/or divalent cations. Species *e* and *g*, in particular, must be only loosely associated with the other RNA and protein components in the large subunit, because they are liberated as the free RNAs when *Crithidia* ribosomes are dissociated at low magnesium (0.1 mM) concentration (Gray, 1981).

Our data clearly indicate that the discontinuous LSU rRNA in *Crithidia* is not generated by single phosphodiester bond scissions ('nicks') in an originally continuous polynucleotide chain, but by the removal of spacer sequenes. Although co-transcription of coding and spacer sequences remains to be demonstrated formally in our system, extrapolation from what is known about rDNA transcription in other eukaryotes (Long and Dawid, 1980; Sollner-Webb and Tower, 1986) makes it very likely that a single primary transcript is produced from the rDNA region in *Crithidia* nuclear DNA. Assuming this is the case, post-transcriptional processing must be a complex process, in that it involves removal of seven ITS sequences and produces mature rRNA species having chemically distinct 5′ termini (non-phosphorylated as well as phosphorylated). Detailed biochemical analysis will obviously be required to unravel the pathway of rRNA biosynthesis in *Crithidia*.

Although intact rRNA does not seem to be essential for protein synthesis (Cahn *et al.*, 1970; Kennedy *et al.*, 1981), the location of discontinuities in the primary and secondary structure is obviously critical, as evidenced by the α-sarcin effect. In this regard, we find it striking that all naturally occurring discontinuities in eukaryotic LSU rRNA are localized to variable regions. This implies that there are relaxed functional constraints in these regions, such that the basic functions of LSU rRNA are not affected by variations in size, base composition and potential secondary structure in these particular regions, or by whether these regions are covalently continuous or not.

The correlation between ITS in rDNA and variable regions in rRNA raises the question of whether there is an evolutionary connection between the two. To account for the presence of multiple ITSs in *Crithidia* rDNA, and the absence of most of them in the rDNA of other eukaryotes, one must assume either (i) that these ITSs were introduced into a continuous LSU rRNA gene after the divergence of the trypanosomatid lineage from other eukaryotes; or, (ii) that the ancestral LSU rRNA gene was, in fact, discontinuous, and that ancient ITSs have been selectively retained in the trypanosomatid lineage, but mostly 'lost' in other eukaryotes. The notion that the ancient pattern of rRNA gene organization was a modular pattern such as that seen in *Crithidia* has a number of attractive features, not the least of which is that it can readily explain the characteristic interspersed pattern of conserved and variable regions in LSU rRNA. Thus, loss of the ability to excise ITSs (a primitive ability in this model) could lead in the course of evolution to incorporation of 'degenerate spacers', as variable regions, into mature rRNA (cf. Cox and Kelly, 1982). This in turn would result in an increase in size and decrease in number of separate rRNA species, through the fusion of what were originally separate modules of rRNA structure.

Ribosomal RNA sequence data suggest that the trypanosomatids represent a particularly deep branching within the eukaryotic lineage (Schnare *et al.*, 1986a; Sogin *et al.*, 1986). However, currently available phylogenetic data are not sufficient to decide whether the common ancestor of the trypanosomatids and other eukaryotes had continuous or discontinuous rRNAs and rRNA genes. Such insights are likely to come from comparative analysis of rRNA structure and gene organization in other, early diverging eukaryotes. *Euglena gracilis* and the trypanosomatids appear to have separated from a common ancestor at the same time as, or shortly after, they separated from the main eukaryotic lineage (Sogin *et al.*, 1986); thus, characterization of the nuclear LSU rRNA gene of *Euglena* may be particularly informative. In this regard, preliminary work in our laboratory indicates that the cytoplasmic LSU rRNA of *Euglena* is also multiply fragmented (M.N.Schnare, unpublished results).

Materials and methods

Total DNA from *C. fasciculata* was prepared, digested with either *Pst*I or *Hin*dIII and ligated into pUC9, as described by Schnare *et al.* (1986a). *E. coli* JM83 cells were transformed (Hanahan, 1983) and screened by colony hybridization (Grunstein and Hogness, 1975). Colonies containing *Hin*dIII clone pCfH1 and *Pst*I clones pCf1, pCf2 and pCf3 were detected by probing with 5′ end-labelled 3 M NaCl-insoluble RNA (Schnare *et al.*, 1986a). A fourth *Pst*I clone (pCf4) was detected by probing transformed colonies with a mixture of 3′ end-labelled RNA species *g* and *j* (Schnare *et al.*, 1983). Because deletion of internal sections of its insert occurred when pCfH1 was grown in *E. coli* JM83 (rec^+), the rec^- strain DH1 (Hanahan, 1983) was used for large-scale preparation of pCfH1 DNA. Recombinant plasmids were isolated and subjected to chemical sequence analysis (Maxam and Gilbert, 1980), as described by Spencer *et al.* (1984) and Schnare *et al.* (1986b).

Crithidia RNA species *c* and *d* were purified from 3 M NaCl-insoluble RNA (Schnare *et al.*, 1986a) and were either 5′ end-labelled (Schnare *et al.*, 1985) or 3′ end-labelled (Peattie, 1979) and repurified in a 2.5% polyacrylamide gel (Schnare and Gray, 1981). The 5′- and 3′-terminal sequences were determined (Schnare *et al.*, 1983) and terminal nucleotide analysis was performed (MacKay *et al.*, 1980).

Sequence comparisons were carried out using the Beckman MicroGenie Sequence Analysis Program (Queen and Korn, 1984). Searches for strong homologies employed the 'Search' function, which will locate regions of data bank sequences longer than 40 residues that are at least 75% identical to regions of the test sequence (stretches as short as 15 residues having better homology are also detected in such a search). A search for more limited homologies used the 'Compare' function to detect stretches at least 80% identical and having at least nine correct matches.

Acknowledgements

This work was supported by a grant (A8387) from the Natural Sciences and Engineering Research Council of Canada.

References

Brimacombe,R. (1982) *Biochem. Soc. Symp.*, **47**, 49–60.

Cahn,F., Schachter,E.M. and Rich,A. (1970) *Biochim. Biophys. Acta*, **209**, 512–520.

Cammarano,P., Londei,P., Mazzei,F. and Biagini,R. (1982) *Comp. Biochem. Physiol.*, **73B**, 435–449.

Chan,Y.-L., Endo,Y. and Wool,I.G. (1983) *J. Biol. Chem.*, **258**, 12768–12770.

Choi,Y.C. (1985) *J. Biol. Chem.*, **260**, 12769–12772.

Clark,C.G., Tague,B.W., Ware,V.C. and Gerbi,S.A. (1984) *Nucleic Acids Res.*, **12**, 6197–6220.

Cordingley,J.S. and Turner,M.J. (1980) *Mol. Biochem. Parasitol.*, **1**, 91–96.

Cox,R.A. and Kelly,J.M. (1981) *FEBS Lett.*, **130**, 1–6.

Cox,R.A. and Kelly,J.M. (1982) *Biochem. Soc. Symp.*, **47**, 11–48.

Delanversin,G. and Jacq,B. (1983) *C.R. Acad. Sci. Ser. III*, **296**, 1041–1044.

Dubin,D.T., HsuChen,C.C., Timko,K.D., Azzolina,T.M., Prince,D.L. and Ranzini,J.L. (1982) In Slonimski,P., Borst,P. and Attardi,G. (eds), *Mitochondrial Genes*. Cold Spring Harbor Laboratory, New York, pp. 89–98.

Ellis,R.E., Sulston,J.E. and Coulson,A.R. (1986) *Nucleic Acids Res.*, **14**, 2345–2364.

Endo,Y. and Wool,I.G. (1982) *J. Biol. Chem.*, **257**, 9054–9060.

Fando,J.L., Alaba,I., Escarmis,C., Fernandez-Luna,J.L., Mendez,E. and Salinas,M. (1985) *Eur. J. Biochem.*, **149**, 29–34.

Fujiwara,H. and Ishikawa,H. (1986) *Nucleic Acids Res.*, **14**, 6393–6401.

Georgiev,O.I., Nikolaev,N., Hadjiolov,A.A., Skryabin,K.G., Zakharyev,V. and Bayev,A.A. (1981) *Nucleic Acids Res.*, **9**, 6953–6958.

Gonzalez,I.L., Gorski,J.L., Campen,T.J., Dorney,D.J., Erickson,J.M., Sylvester,R.E. and Schmickel,R.D. (1985) *Proc. Natl. Acad. Sci. USA*, **82**, 7666–7670.

Gray,M.W. (1979) *Can. J. Biochem.*, **57**, 914–926.

Gray,M.W. (1981) *Mol. Cell Biol.*, **1**, 347–357.

Grunstein,M. and Hogness,D.S. (1975) *Proc. Natl. Acad. Sci. USA*, **72**, 3961–3965.

Gunderson,J.H. and Sogin,M.L. (1986) *Gene*, **44**, 63–70.

Gutell,R.R., Weiser,B., Woese,C.R. and Noller,H.F. (1985) *Prog. Nucleic Acids Res. Mol. Biol.*, **32**, 155–216.

Hadjiolov,A.A., Georgiev,O.I., Nosikov,V.V. and Yavachev,L.P. (1984) *Nucleic Acids Res.*, **12**, 3677–3693.

Hanahan,D. (1983) *J. Mol. Biol.*, **166**, 557–580.

Hassouna,N., Michot,B. and Bachellerie,J.-P. (1984) *Nucleic Acids Res.*, **12**, 3563–3583.

Hernández,R., Nava,G. and Castañeda,M. (1983) *Mol. Biochem. Parasitol.*, **8**, 297–304.

Hogan,J.J., Gutell,R.R. and Noller,H.F. (1984) *Biochemistry*, **23**, 3330–3335.

Ishikawa,H. (1977) *Comp. Biochem. Physiol.*, **58B**, 1–7.

Jordan,B.R., Latil-Damotte,M. and Jourdan,R. (1980) *Nucleic Acids Res.*, **8**, 3565–3573.

Kennedy,T.D., Hanley-Bowdoin,L.K. and Lane,B.G. (1981) *J. Biol. Chem.*, **256**, 5802–5809.

Loening,U.E. (1968) *J. Mol. Biol.*, **38**, 355–365.

Londei,P., Cammarano,P., Mazzei,F. and Romeo,A. (1982) *Comp. Biochem. Physiol.*, **73B**, 423–434.

Long,E.O. and Dawid,I.B. (1980) *Annu. Rev. Biochem.*, **49**, 727–764.

MacKay,R.M., Spencer,D.F., Doolittle,W.F. and Gray,M.W. (1980) *Eur. J. Biochem.*, **112**, 561–576.

Maxam,A.M. and Gilbert,W. (1980) *Methods Enzymol.*, **65**, 499–560.

Michot,B., Hassouna,N. and Bachellerie,J.-P. (1984) *Nucleic Acids Res.*, **12**, 4259–4279.

Nazar,R.N. (1980) *FEBS Lett.*, **119**, 212–214.

Nazar,R.N. (1982) *FEBS Lett.*, **143**, 161–162.

Noller,H.F., Kop,J., Wheaton,V., Brosius,J., Gutell,R.R., Kopylov,A.M., Dohme,F., Herr,W., Stahl,D.A., Gupta,R. and Woese,C.R. (1981) *Nucleic Acids Res.*, **9**, 6167–6189.

Olsen,G.J., McCarroll,R. and Sogin,M.L. (1983) *Nucleic Acids Res.*, **11**, 8037–8049.

Otsuka,T., Nomiyama,H., Yoshida,H., Kukita,T., Kuhara,S. and Sakaki,Y. (1983) *Proc. Natl. Acad. Sci. USA*, **80**, 3163–3167.

Ozaki,T., Hoshikawa,Y., Iida,Y. and Iwabuchi,M. (1984) *Nucleic Acids Res.*, **12**, 4171–4184.

Pavlakis,G.N., Jordan,B.R., Wurst,R.M. and Vournakis,J.N. (1979) *Nucleic Acids Res.*, **7**, 2213–2238.

Peattie,D.A. (1979) *Proc. Natl. Acad. Sci. USA*, **76**, 1760–1764.

Perry,R.P. (1976) *Annu. Rev. Biochem.*, **45**, 605–629.

Queen,C. and Korn,L.J. (1984) *Nucleic Acids Res.*, **12**, 581–599.

Schindler,D.G. and Davies,J.E. (1977) *Nucleic Acids Res.*, **4**, 1097–1110.

Schnare,M.N. and Gray,M.W. (1981) *FEBS Lett.*, **128**, 298–304.

Schnare,M.N. and Gray,M.W. (1982) *Nucleic Acids Res.*, **10**, 2085–2092.

Schnare,M.N., Spencer,D.F. and Gray,M.W. (1983) *Can. J. Biochem. Cell. Biol.*, **61**, 38–45.

Schnare,M.N., Heinonen,T.Y.K., Young,P.G. and Gray,M.W. (1985) *Curr. Genet.*, **9**, 389–393.

Schnare,M.N., Collings,J.C. and Gray,M.W. (1986a) *Curr. Genet.*, **10**, 405–410.

Schnare,M.N., Heinonen,T.Y.K., Young,P.G. and Gray,M.W. (1986b) *J. Biol. Chem.*, **261**, 5187–5193.

Simpson,L. and Simpson,A.M. (1978) *Cell*, **14**, 169–178.

Sogin,M.L., Elwood,H.J. and Gunderson,J.H. (1986) *Proc. Natl. Acad. Sci. USA*, **83**, 1383–1387.

Sollner-Webb,B. and Tower,J. (1986) *Annu. Rev. Biochem.*, **55**, 801–830.

Spencer,D.F., Schnare,M.N. and Gray,M.W. (1984) *Proc. Natl. Acad. Sci. USA*, **81**, 493–497.

Stevens,A.R. and Pachler,P.F. (1972) *J. Mol. Biol.*, **66**, 225–237.

Stiegler,P., Carbon,P., Ebel,J,-P. and Ehresmann,C. (1981) *Eur. J. Biochem.*, **120**, 487–495.

Takaiwa,F., Oono,K., Iida,Y. and Sugiura,M. (1985) *Gene*, **37**, 255–259.

Veldman,G.M., Klootwijk,J., de Regt,V.C.H.F., Planta,R.J., Branlant,C., Krol,A. and Ebel,J.-P. (1981) *Nucleic Acids Res.*, **9**, 6935–6952.

Vossbrinck,C.R. and Woese,C.R. (1986) *Nature*, **320**, 287–288.

Walker,T.A., Endo,Y., Wheat,W.H., Wool,I.G. and Pace,N.R. (1983) *J. Biol. Chem.*, **258**, 333–338.

Ware,V.C., Tague,B.W., Clark,C.G., Gourse,R.L., Brand,R.C. and Gerbi,S.A. (1983) *Nucleic Acids Res.*, **11**, 7795–7817.

Ware,V.C., Renkawitz,R. and Gerbi,S.A. (1985) *Nucleic Acids Res.*, **13**, 3581–3597.

Whittaker,R.H. (1969) *Science*, **163**, 150–160.

Wool,I.G. (1984) *Trends Biochem. Sci.*, **9**, 14–17.

Zwieb,C., Glotz,C. and Brimacombe,R. (1981) *Nucleic Acids Res.*, **9**, 3621–3640.

Received on January 12, 1987

Note added in proof

These sequence data have been submitted to the EMBL/GenBank Data Libraries under the accession number Y00055.