The Human 18S Ribosomal RNA Gene: Evolution and Stability

IRIS LAUDIEN GONZALEZ[1] AND ROY D. SCHMICKEL

SUMMARY

We report the 1,870-base-pair primary sequence of a human 18S rRNA gene and propose a secondary structure based on this sequence and the general mammalian structure. A basic secondary structure for the small subunit rRNA has been preserved throughout evolution by compensatory and neutral base changes in double-stranded regions. The molecule contains eight regions that can vary in structure and that comprise 432 bases, while 1,438 bases belong to regions of conserved structure among all species tested. The conserved regions show a remarkably low sequence divergence rate of 0.1% between the human and mouse genes over the approximately 80 million years since the mammalian radiation. This value may make the small subunit rDNA the most highly conserved sequence known. Sequence conservation in higher eukaryotes with multiple copies of the gene is probably achieved by the combination of strong selection and the correction of tandem genes by unequal homologous exchange.

INTRODUCTION

Ribosomes, the protein-synthesis organelles, are ancient structures that are common to all types of cells. We have completed the sequence of the human 18S rRNA gene, which permits us to place human evolution within the context of all organisms and allows genetic comparisons over enormous evolutionary distances [1–6]. Comparisons can be made at two levels: the primary nucleotide sequences and the secondary structures of the rRNA. While the primary sequences can differ greatly among prokaryotic, eukaryotic, archaebacterial,

Received July 5, 1985; revised September 5, 1985.

This work was supported by grant HD16930 from the National Institutes of Health (R. D. S.).

[1] Both authors: Department of Human Genetics, University of Pennsylvania, Philadelphia, PA 19104.

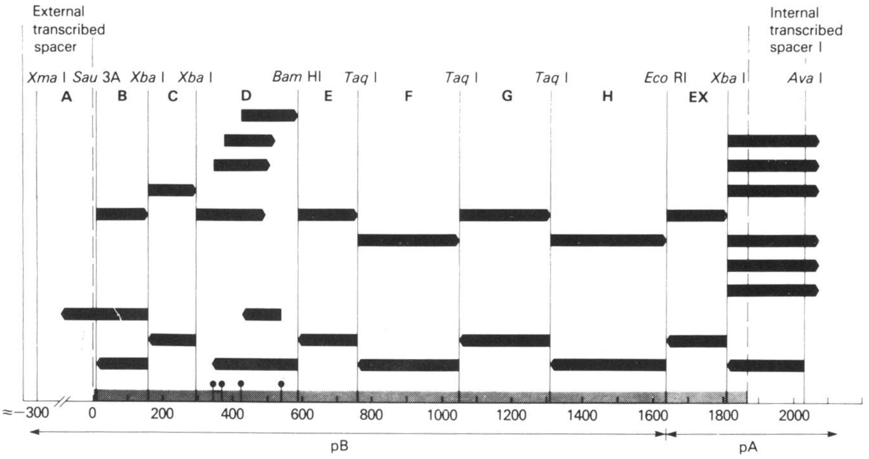

FIG. 1.—Sequencing strategy for the human 18S rRNA gene and flanking sequences

and organelle genes, the secondary structures are remarkably well preserved throughout evolution. The structurally conserved regions are separated by variable regions in which both sequence and structure have diverged among these major divisions. Several of these regions are restricted to only one group. Divergence rates between rRNA molecules are calculated separately for the "conserved" and "variable" regions [5].

MATERIALS AND METHODS

Figure 1 depicts the sequencing strategy for the human 18S rRNA gene and flanking sequences. Fragments were subcloned from the previously described rDNA-containing plasmids pB and pA4 [7]. Plasmid pB contains rDNA sequences beginning at an *Eco*RI site 5′ to the 18S gene and includes the promoter, 3.3 kilobases (kb) of external transcribed spacer and most of the 18S gene, ending at the *Eco*RI site 230 bases from the 3′ end of the gene. Plasmid pA4 starts at this latter *Eco*RI site and extends into the 28S gene. For subcloning and sequencing purposes, the gene was divided into regions based on the known restriction sites in the human [8] and rat [3, 4] 18S rRNA genes; these sites are shown by labeled vertical lines on the diagram. Section D was subdivided by *Taq*I digestion. Fragments were ligated into the appropriate M13 vectors to obtain clones in both orientations. The 5′ terminal 10 bases were sequenced in only one orientation. The sequence from the *Xba*I site near the 3′ terminus, which extends into internal transcribed spacer I, was obtained from six separate human genes; all six gave an identical sequence for the 18S terminus, but showed slight variations within the spacer. The ends of the gene were determined by analogy to those of rat [3, 4], mouse [5], rabbit [6], and *Xenopus* [9].

Sequencing was performed according to the Sanger method [10], using two of the modifications reported by Gomer et al. [11]: the reaction was carried out with *Hin*cII buffer, at 50°C.

RESULTS

Sequence and Species Comparison

Figure 2 shows the 1,870-base sequence of the human 18S rRNA gene and partial adjacent spacer sequences. The human 18S rRNA gene sequence is

```
EXTERNAL TRANSCRIBED SPACER
-140       -130       -120       -110       -100       -90        -80
CGCTGCTCCT CCCGTCGCCG TCCGGGCCCG TCCGTCCGTC CGTCCGTCGT CCTCCTCGCT NNNNCGGGGC

-70        -60        -50        -40        -30        -20        -10
GCCGGGCCCG TCCTCACNGG CCCCCGNNNN NGTCCNGGCC CGTCGGGGCC TCGCCGCGCT CTACCTTACC


HUMAN 18S
        10         20         30         40         50         60         70
TACCTGGTTG ATCCTGCCAG TAGCATATGC TTGTCTCAAA GATTAAGCCA TGCATGTCTA AGTACGCACG

        80         90        100        110        120        130        140
GCCGGTACAG TGAAACTGCG AATGGCTCAT TAAATCAGTT ATGGTTCCTT TGGTCGCTCG CTCCTCTCCT

       150        160        170        180        190        200        210
ACTTGGATAA CTGTGGTAAT TCTAGAGCTA ATACATGCCG ACGGGCGCTG ACCCCCTTCG CGGGGGGGAT

       220        230        240        250        260        270        280
GCGTGCATTT ATCAGATCAA AACCAACCCG GTCAGCCCCT CTCCGGCCCC GGCCGGGGGG CGGGCGCCGG

       290        300        310        320        330        340        350
CGGCTTTGGT GACTCTAGAT AACCTCGGGC CGATCGCACG CCCCCCGTGG CGGCGACGAC CCATTCGAAC

       360        370        380        390        400        410        420
GTCTGCCCTA TCAACTTTCG ATGGTAGTCG CCGTGCCTAC CATGGTGACC ACGGGTGACG GGGAATCAGG

       430        440        450        460        470        480        490
GTTCGATTCC GGAGAGGGAG CCTGAGAAAC GGCTACCACA TCCAAGGAAG GCAGCAGGCG CGCAAATTAC

       500        510        520        530        540        550        560
CCACTCCCGA CCCGGGGAGG TAGTGACGAA AAATAACAAT ACAGGACTCT TTCGAGGCCC TGTAATTGGA

       570        580        590        600        610        620        630
ATGAGTCCAC TTTAAATCCT TTAACGAGGA TCCATTGGAG GGCAAGTCTG GTGCCAGCAG CCGCGGTAAT

       640        650        660        670        680        690        700
TCCAGCTCCA ATAGCGTATA TTAAAGTTGC TGCAGTTAAA AAGCTCGTAG TTGGATCTTG GGAGCGGGCG

       710        720        730        740        750        760        770
GGCGGTCCGC CGCGAGGCGA GCCACCGCCC GTCCCCGCCC CTTGCCTCTC GGCGCCCCCT CGATGCTCTT

       780        790        800        810        820        830        840
AGCTGAGTGT CCCGCGGGGC CCGAAGCGTT TACTTTGAAA AAATTAGAGT GTTCAAAGCA GGCCCGAGCC

       850        860        870        880        890        900        910
GCCTGGATAC CGCAGCTAGG AATAATGGAA TAGGACCGCG GTTCTATTTT GTTGGTTTTC GGAACTGAGG

       920        930        940        950        960        970        980
CCATGATTAA GAGGGACGGC CGGGGGCATT CGTATTGCGC CGCTAGAGGT GAAATTCCTT GGACCGGCGC

       990       1000       1010       1020       1030       1040       1050
AAGACGGACC AGAGCGAAAG CATTTGCCAA GAATGTTTTC ATTAATCAAG AACGAAAGTC GGAGGTTCGA

      1060       1070       1080       1090       1100       1110       1120
AGACGATCAG ATACCGTCGT AGTTCCGACC ATAAACGATG CCGACCGGCG ATGCGGCGGC GTTATTCCCA

      1130       1140       1150       1160       1170       1180       1190
TGACCCGCCG GGCAGCTTCC GGGAAACCAA AGTCTTTGGG TTCCGGGGGG AGTATGGTTG CAAAGCTGAA

      1200       1210       1220       1230       1240       1250       1260
ACTTAAAGGA ATTGACGGAA GGGCACCACC AGGAGTGGAG CCTGCGGCTT AATTTGACTC AACACGGGAA

      1270       1280       1290       1300       1310       1320       1330
ACCTCACCCG GCCCGGACAC GGACAGGATT GACAGATTGA TAGCTCTTTC TCGATTCCGT GGGTGGTGGT

      1340       1350       1360       1370       1380       1390       1400
GCATGGCCGT TCTTAGTTGG TGGAGCGATT TGTCTGGTTA ATTCCGATAA CGAACGAGAC TCTGGCATGC

      1410       1420       1430       1440       1450       1460       1470
TAACTAGTTA CGCGACCCCC GAGCGGTCGG CGTCCCCCAA CTTCTTAGAG GGACAAGTGG CGTTCAGCCA

      1480       1490       1500       1510       1520       1530       1540
CCCGAGATTG AGCAATAACA GGTCTGTGAT GCCCTTAGAT GTCCGGGGCT GCACGCGCGC TACACTGACT

      1550       1560       1570       1580       1590       1600       1610
GGCTCAGCGT GTGCCTACCC TACGCCGGCA GGCGCGGGTA ACCCGTTGAA CCCCATTCGT GATGGGGATC

      1620       1630       1640       1650       1660       1670       1680
GGGGATTGCA ATTATTCCCC ATGAACGAGG AATTCCCAGT AAGTGCGGGT CATAAGCTTG CGTTGATTAA

      1690       1700       1710       1720       1730       1740       1750
GTCCCTGCCC TTTGTACACA CCGCCCGTCG CTACTACCGA TTGGATGGTT TAGTGAGGCC CTCGGATCGG

      1760       1770       1780       1790       1800       1810       1820
CCCCGCCGGG GTCGGCCCAC GGCCCTGGCG GAGCGCTGAG AAGACGGTCG AACTTGACTA TCTAGAGGAA

      1830       1840       1850       1860       1870
GTAAAAGTCG TAACAAGGTT TCCGTAGGTG AACCTGCGGA AGGATCATTA


INTERNAL TRANSCRIBED SPACER
        10         20         30         40         50         60         70
ACGGAGCCCG GACGGCGGCC CGCGGCGGCG CCGCGCCGCG CTTCCCTCCG CACACCCACC CCCCCACCGC

        80         90        100        110        120        130        140
GACGGCGCGT GCGGGCGGGG CCGTGCCCGT TCGTTCGCTC GCTCGTTCGT TCGCCGCCCG GCCCGGCCGC

       150        160        170        180        190        200        210
GAGAGCCGAG AACTCGGGAG GGAGACGGGG GAGAGAGAGA GAGAGAGAGA GAGAGAGAGA GAGAGAGAGA

       220
GAAAGAAGGG CGTGT
```

FIG. 2.—Sequence of the human 18S rRNA gene and parts of the adjacent 5′ external transcribed spacer and 3′ internal transcribed spacer I. The 5′ spacer is numbered with negative numbers; the 3′ spacer starts numbering at 1.

remarkably similar to those of higher eukaryotes such as rat ([3, 4] and I. G. Wool, corrected sequence, personal communication 1985), mouse [5], and rabbit [6], and to those of lower eukaryotes. The human gene sequence was compared with the ribosomal gene sequences of 14 other organisms by using the alignments compiled by Nelles et al. [2], which include mammalian, amphibian, crustacean, fungal, archaebacterial, eubacterial, and organelle gene sequences. Thus, the comparison included organisms from all levels of evolution over a period of 3 billion years. This comparison has revealed eight variable regions [5], six of which are found in eukaryotes but not in prokaryotes. The variable regions are described below and compared among higher and lower eukaryotes, prokaryotes, and archaebacteria. The variable regions are designated V1–V8 as indicated in figure 3.

(V1) *Bp65–80:* Is present in prokaryotes and eukaryotes and largely absent in archaebacteria and organelle genes; the region is variable in length and sequence within each of these divisions. All vertebrates have identical sequences. The prokaryotes have 10 more bases than eukaryotes and consequently an enlarged hairpin.

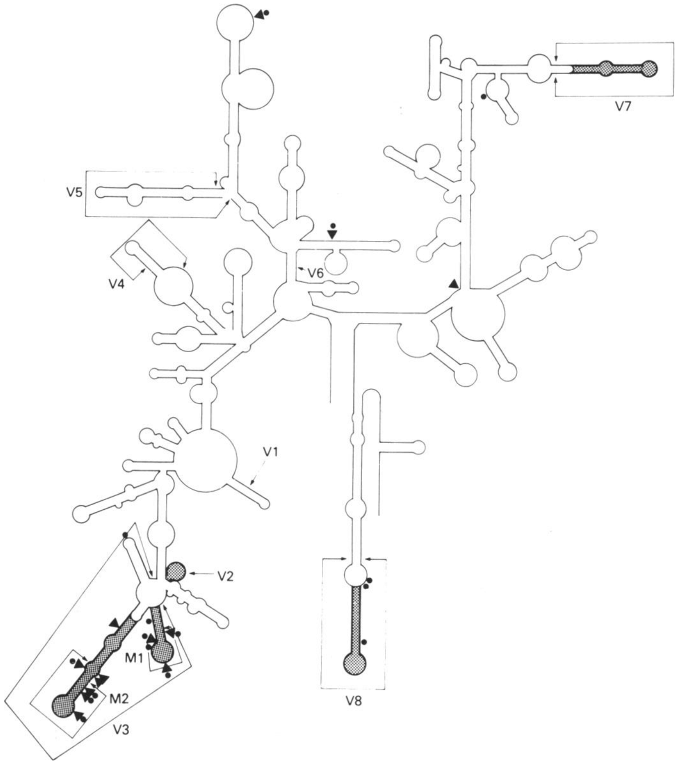

FIG. 3.—Line diagram of *E. coli* 18S rRNA secondary structure. Variable regions discussed in text are marked. Eukaryote-specific enlargements are added as *shaded regions*. *V1:* human 65–80; *V2:* 128–142; *V3:* 194–272; *V4:* 577–578; *V5:* 683–910; *V6:* 1161–1169; *V7:* 1419–1434; *V8:* 1754–1782. *Arrows* indicate differences between the human and mouse sequences. *Solid circles* are differences between human and rat sequences.

(V2) *Bp128–142:* Is found in eukaryotes but not in prokaryotes. Although it is variable among eukaryotes, the four mammalian sequences are identical and have three bases more than the *Xenopus laevis* sequence.

(V3) *Bp194–334:* This region is much larger in eukaryotes than in prokaryotes and contains two sequences that are present only in mammals, which are marked M1 (195–202) and M2 (249–272) on figure 3. M1 consists of eight bases that enlarge the double-stranded structure where it is located. M2 consists of 24 extra bases and forms the enlarged tip of another structure. The section between 320 and 334 is present only in eukaryotic and in eubacterial genes.

(V4) *Bp577–578:* A 24-base region that is present only in eubacteria.

(V5) *Bp683–910:* This region consists of two parts: the section between 683 and 850 is present only in eukaryotes; the section 851–910 differs in sequence between eukaryotes and prokaryotes but can form a similar secondary structure in both. A proposed secondary structure for the human V5 is shown in figure 4.

(V6) *Bp1161–1169:* Is highly-conserved in all eukaryotes but is absent in prokaryotes. It is part of a region for which conflicting secondary structures have been proposed.

(V7) *Bp1419–1434:* Is seen only in eukaryotes and forms the tip of a hairpin structure. Although V7 is variable among the eukaryotes, the mammals share identical sequences. The *Dictyostelium* gene possesses an extra 75-base pair (bp) sequence.

(V8) *Bp1754–1782:* Is a variable eukaryote-specific insert that shows a few differences among the mammalian sequences. This sequence also enlarges the

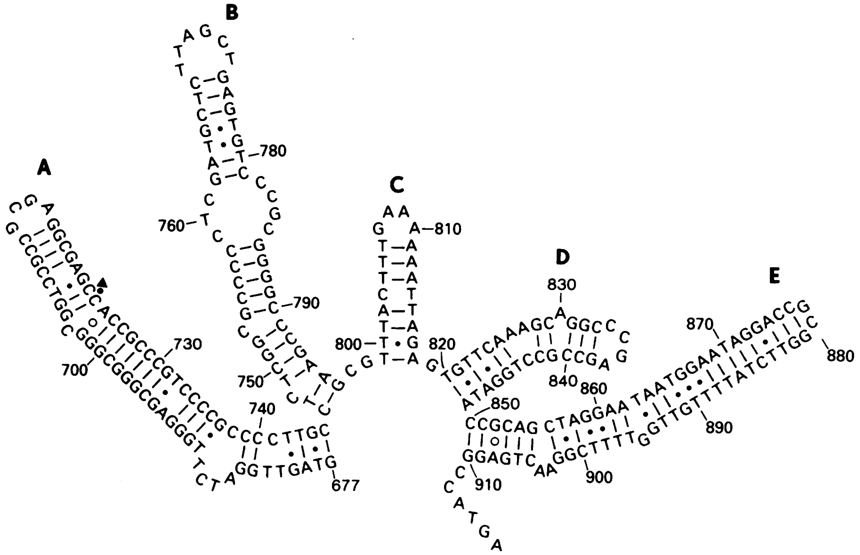

FIG. 4.—Proposed secondary structure model for region V5 between human gene bp 677–910. *Arrow* and *solid circle* indicate single difference between the human and mouse and human and rat sequences, respectively.

tip of an *E. coli* double-stranded structure. An extended sequence in the (prokaryotic-like) *Triticum aestivum* mitochondrial gene is also present here; the human mitochondrial gene [12] may have a small section that is homologous to this expansion.

Secondary Structure for Variable Region V5

The human sequence fits the secondary structure model proposed for the rat 18S gene [4]. For region V5 between human bp677 and 910, which includes a eukaryote-specific stretch, the only model available had been proposed and chemically tested for *Xenopus*, confirming the existence of one stem of the model [13]; a partial model was proposed for *Artemia* [2]. We propose a different model (fig. 4), which represents the best fit of several models constructed and tested for compensatory base changes between human, *Saccharomyces* [14], rat, mouse, rabbit, *Xenopus*, the brine shrimp *Artemia*, maize [15], rice [16], *Dictyostelium* [17], and *E. coli* [18–20] sequences. A model for secondary structure can be established that contains five helical stem or "hairpin" structures. Each stem was evaluated independently by comparison with other species. In figure 4, the stems are labeled A: 677–745, B: 746–796, C: 799–818, D: 820–849, and E: 850–910. The model was derived as follows: (1) Stem B had been demonstrated to exist in *Xenopus* by nuclease-resistance [13]; depending on digestion conditions, the last two pairs may/may not be part of the stem. The only eukaryotic sequence that gave a stem with a different shape was *Saccharomyces*. (2) Stem C consists of a short sequence that is highly conserved in all eukaryotes except *Dictyostelium;* the only base changes found involved the A at position 810 of the loop section of the *Artemia*, rice and maize genes, and elimination of the last pairing at the base of the stem in maize and rice. (3) Stem E is the equivalent of *E. coli* stem 588–651 [1]. One obtains an equivalent eukaryotic stem when one aligns the *E. coli* model loop 618–622 with human nos. 878–879 and yeast nos. 819–823; these are very similar structures, although the unpaired bases present near the top of the *E. coli* stem are missing in the human structure. Compensatory and neutral base changes are in favor of this structure. (4) Once these three stems were defined, possible structures for the two remaining stems (A and D) were constructed considering compensatory base changes. Stem A is composed of nonconserved sequences, but in this model, the secondary structures are conserved in human and other species. The *Dictyostelium* sequence is the only one that does not fit. The shape of stem D is compatible with most ribosomal sequences, although the number of base pairs is 11 in the human, 10 in *Xenopus*, *Artemia*, and *Saccharomyces*, and eight in the plants; even the *Dictyostelium* sequence fits this shape.

Unresolved Structures

Conflicting secondary structure models have been proposed and "proven" [21, 22] for the *Xenopus* and yeast section corresponding to human nos. 1141–1164. Unfortunately, the human sequence provides no clarifying base changes.

The short segment from 570 to 594 includes the eubacteria-specific V4. This stretch can form two possible hairpins, neither of which has received support from compensatory base changes.

DISCUSSION

The availability of the 18S rRNA sequences from human and other species permits us to compare ribosomal structure over vast evolutionary periods. At the primary sequence level, one finds considerable divergence between the major kingdoms of organisms. However, sequence is remarkably similar within groupings of more recent emergence, such as the vertebrates: the human and *Xenopus* 18S genes have an overall divergence rate of only 2.5%, and humans and rodents have overall divergence rates of 0.45% and 0.37% (see table 1). Closer examination reveals that the differences among these genes are concentrated in the "variable regions" described above. For example, there are 17 differences between the human sequence and the rat sequence [4]; 12 of these are clustered in the 432 bases of the variable regions, giving an eight-fold higher divergence rate than in the conserved regions (table 1). Most of the nucleotide base differences among the mammalian rRNA genes are found in regions V3 and V8. Region V3 also contains two segments that are present only in the mammalian genes (M1 and M2 in fig. 3). These eight regions are the small subunit equivalent of the "variable" or "joining" regions that are also seen in the large subunit rRNA gene [23, 24]. The 18S gene variable regions seem to obey certain size contraints and do not have the great sequence variability that is found in the variable regions of the 28S subunit, where only the primate sequences were close enough to allow comparison [23]. The divergence rates between the conserved regions of the human and rodent 18S genes are less than 1/3 those found for comparable regions of the large subunit gene [23]. Comparison between the small and large subunit gene sequences indicates that the 18S gene is more stable, although both genes contribute to the same functional structure.

When secondary structure is considered, a remarkable conservation is observed among all organisms [1]. The secondary structure has changed very little over 3 billion years, pointing to the importance of the structure for rRNA function.

The great stability of the small subunit rRNA (secondary structure conservation) may be due to two factors: (1) *Selection*—The small subunit rRNA may

TABLE 1

Divergence between the Human and Other Vertebrate 18S Genes

	Overall		In variable regions		In conserved regions		
	(1,870 bases)		(432 bases)		(1,438 bases)		
	No. differences	% divergence	No. differences	% divergence	No. differences	% divergence	Variable / Conserved
Rat*	17	0.45	12	1.4	5	0.17	8
Mouse	14	0.37	11	1.3	3	0.1	13
Xenopus	99	2.64	55	6.4	44	1.5	4.3

Note: Divergence rate was calculated as no. base changes divided by the combined target region.

* I. G. Wool, revised sequence, personal communication, 1985.

be closely related to the original translation organelle and functional requirements would constitute the major constraints on the secondary structure of the molecule. In an RNA molecule, primary sequence and secondary structure are much more closely coordinated than mRNA sequence and protein structure [1]. (2) *Unequal homologous exchange*—Uniformity is promoted in higher organisms that have tandem copies of the genes on separate chromosomes by the mechanism of unequal homologous exchange. Gene conversion results from the mechanism of homologous exchange. Subsequent unequal homologous exchange permits amplification of the correct sequence [25]. Neither of these two factors alone would explain the extreme constancy of these genes over time. By itself, the rigid selection imposed on an essential organelle is not fully operant here because of the redundancy of the gene. There are over 300 copies of the gene in each nucleus, and a single variant among them makes no phenotypic difference. As a single factor, the presence of tandem copies leads to rapid divergence with the help of unequal homologous exchange. However, the two factors acting together promote the extreme level of sequence conservation. It is likely that the ability to correct (or diverge) by unequal homologous exchange plus the selective pressure on blocks of genes can lead to the most effective genetic conservation.

This study illustrates why ribosomal genes are found in tandem arrangements. The redundancy frees them from strong selective pressures. If they had separated, they would probably have diverged and become inactive. Redundant genes must remain tandem to preserve their identity. The effectiveness of this arrangement is inferred by the extraordinary conservation of the primary sequence. By comparison with organisms from distant kingdoms, it is evident that the primary sequence is not essential for function. Nevertheless, the primary sequence of the vertebrates has been maintained.

Finally, this study emphasizes the important feature that has been found with every study—the secondary structure of the 18S rRNA molecule is of utmost importance for its function and, although very complex, it has remained unchanged for eons.

REFERENCES

1. Noller HI: *Ann Rev Biochem* 53:119–162, 1984
2. Nelles L, Fang B-L, Volckaert G, Vandenberghe A, De Wachter R: *Nucleic Acids Res* 12:8749–8768, 1984
3. Torczynski R, Bollon AP, Fuke M: *Nucleic Acids Res* 11:4879–4890, 1983
4. Chan Y-L, Gutell R, Noller HF, Wool IG: *J Biol Chem* 259:224–230, 1984
5. Raynal F, Michot B, Bachellerie J-P: *FEBS Lett* 167:263–268, 1984
6. Connaughton JF, Rairkar A, Lockard RE, Kumar A: *Nucleic Acids Res* 12:4731–4745, 1984
7. Erickson JM, Rushford CL, Dorney DJ, Wilson GN, Schmickel RD: *Gene* 16:1–9, 1981
8. Wilson GN, Szura LL, Rushford C, Jackson D, Erickson J: *Am J Hum Genet* 34:32–49, 1982
9. Salim M, Maden BEH: *Nature* 291:205–208, 1981
10. Sanger F, Nickeln S, Coulson AR: *Proc Natl Acad Sci USA* 74:5463–5467, 1977
11. Gomer R, Datta S, Firtel R: *BRL Focus* 7(1):6–7, 1985

12. Eperon IC, Anderson S, Nierlich DP: *Nature* 286:460–467, 1980
13. Atmadja J, Brimacombe R, Maden BEH: *Nucleic Acids Res* 12:2649–2667, 1984
14. Rubtsov PM, Musakhanov MM, Zakharyev VM, Krayer AS, Skryabin KG, Bayer AA: *Nucleic Acids Res* 8:5779–5794, 1980
15. Messing J, Carlson J, Hagen G, Rubenstein I, Oleson A: *DNA* 3:31–40, 1984
16. Takaiwa F, Oono K, Sugiura M: *Nucleic Acids Res* 12:5441–5448, 1984
17. Ohlsen GJ, McCarroll R, Sogin ML: *Nucleic Acids Res* 11:8037–8049, 1983
18. Brosius J, Palmer ML, Kennedy JP, Noller HF: *Proc Natl Acad Sci USA* 75:4801–4805, 1978
19. Carbon P, Ehresmann C, Ehresmann B, Ebel J-P: *Eur J Biochem* 100:399–410, 1979
20. Woese CR, Gutell R, Gupta R, Noller HF: *Microbiol Rev* 47:621–669, 1983
21. Mankin AS, Kopylov AM, Bogdanov AA: *FEBS Lett* 134:11–14, 1981
22. Hogan JJ, Gutell RR, Noller HF: *Biochemistry* 23:3322–3330, 1984
23. Gonzalez IL, Gorski JL, Campen TJ, et al.: *Proc Natl Acad Sci USA* 82:7666–7670, 1985
24. Gorski JL, Gonzalez IL, Schmickel RD: Submitted for publication
25. Arnheim N, Krystal M, Schmickel R, Wilson G, Ryder O, Zimmer E: *Proc Natl Acad Sci USA* 77:7323–7327, 1980

MEETING: *International Congress of Human Genetics*, September 22–26, 1986, West Berlin.