J Mol Evol (1987) 24:236–251

Journal of
Molecular Evolution

© Springer-Verlag New York Inc. 1987

The Secondary Structure of Human 28S rRNA: The Structure and Evolution of a Mosaic rRNA Gene

Jerome L. Gorski,* Iris L. Gonzalez, and Roy D. Schmickel

Department of Human Genetics, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA

Summary. We have determined the secondary structure of the human 28S rRNA molecule based on comparative analysis of available eukaryotic cytoplasmic and prokaryotic large-rRNA gene sequences. Examination of large-rRNA sequences of both distantly and closely related species has enabled us to derive a structure that accounts both for highly conserved sequence tracts and for previously unanalyzed variable-sequence tracts that account for the evolutionary differences in size among the large rRNAs.

Human 28S rRNA is composed of two different types of sequence tracts: conserved and variable. They differ in composition, degree of conservation, and evolution. The conserved regions demonstrate a striking constancy of size and sequence. We have confirmed that the conserved regions of large-rRNA molecules are capable of forming structures that are superimposable on one another. The variable regions contain the sequences responsible for the 83% increase in size of the human large-rRNA molecule over that of *Escherichia coli.* Their locations in the gene are maintained during evolution. They are G + C rich and largely nonhomologous, contain simple repetitive sequences, appear to evolve by frequent recombinational events, and are capable of forming large, stable hairpins.

The secondary-structure model presented here is in close agreement with existing prokaryotic 23S rRNA secondary-structure models. The introduction of this model helps resolve differences between previously proposed prokaryotic and eukaryotic large-rRNA secondary-structure models.

Key words: 28S rRNA — RNA secondary structure — Evolution — Ribosome — Translation

Introduction

Determining the primary and secondary structure is an obligatory beginning in understanding the detailed molecular organization of rRNA, and in determining what specific functional roles the structural elements of the rRNA molecule serve through protein–RNA and RNA–RNA interactions. Detailed comparative and direct biochemical analyses have enabled secondary-structure models of the *Escherichia coli* 23S rRNA to be constructed (Brantant et al. 1981; Glotz et al. 1981; Noller et al. 1981). Subsequent models based on accumulated comparative and biochemical data have demonstrated the conservation of the secondary structure of this molecule during evolution, through compensatory base changes within base-paired regions (Maly and Brimacombe 1983; Noller 1984). Secondary-structural domains derived by comparative analysis have been found to correspond with physically defined ribosomal structural domains (Zimmermann 1979).

There has been less progress in determining the structure of the eukaryotic ribosome. Secondary-structure models have been determined by comparative analysis for yeast (Veldman et al. 1981), *Xenopus laevis* (Clark et al. 1984), and mouse (Michot et al. 1984) cytoplasmic large rRNAs. When compared with the prokaryotic models, each eu-

Offprint requests to: R.D. Schmickel
* *Present address:* Department of Pediatrics and Human Genetics, University of Michigan, Ann Arbor, Michigan 48105, USA

karyotic model revealed extensive conservation and a number of nonhomologous elements, or variable regions, that accounted for the increased size of these eukaryotic large rRNAs. However, discrepancies are seen within both the highly conserved and the variable structural elements when any of the eukaryotic models is compared with any other eukaryotic model.

The large number of fully sequenced cytoplasmic large-rRNA genes enables us to analyze the detailed structure of the large-rRNA gene, to study the process of the evolutionary increase in size of the large-rRNA genes, and, by comparative analysis, to derive a secondary structure of the encoded RNA. In this paper we present a model of the human 28S rRNA and compare this model with existing large-rRNA secondary-structure models.

Methods and Materials

DNA. The complete human 28S rDNA sequence and sequencing strategy have been reported (Gonzalez et al. 1985). The sequence is in total agreement with the published map of the human chromosomal 28S rRNA genes (Erickson et al. 1981; Erickson and Schmickel 1985).

Computer Analysis. The cytoplasmic large-rDNA sequences were analyzed using the Cornell sequencing programs (DeBanzie et al. 1984) to identify invariant sequence tracts of 9–28 bases in length. Identified tracts were used as anchor points for unambiguous alignment of the intervening sequences using the Align program of the NUCPROCS DNA-sequence-analysis package of PROPHET: the Smith and Waterman (1981) variation of the Sellers (1974) algorithm. The alignment was further refined by generating a secondary structure for the human 28S rRNA molecule by comparative analysis.

Potential base-paired regions were cataloged using the interactive PROPHET program (Auron et al. 1982). Evolutionarily consistent structures were chosen from among those potential base pairings by demonstrating compensatory base changes (CBCs) within the compared sequences that maintain double-stranded regions of RNA (Noller et al. 1981).

The accumulation of demonstrated CBCs supported the initial alignment and allowed further delimitation of nucleotides involved in structural features not common to all species examined (see Results). Similar methods were used to determine the secondary structures of nucleotide tracts not common to all species. In the absence of comparative data, structures were derived by computer analysis (Auron et al. 1982).

The resulting secondary-structure model was compared with existing prokaryotic (Maly and Brimacombe 1983; Noller 1984) and eukaryotic (Veldman et al. 1981; Clark et al. 1984; Michot et al. 1984) secondary-structure models of the large-rRNA molecule.

Results

Large-rDNA Sequence Alignment

The human 28S rDNA sequence was aligned with the available, fully sequenced cytoplasmic large-rDNA sequences: those from the mouse (Hassouna et al. 1984), rat (Chan et al. 1983), *X. laevis* (Ware et al. 1983), *Saccharomyces carlsbergensis* (Veldman et al. 1981), *E. coli* (Brosius et al. 1980), and *Physarum polycephalum* (Otsuka et al. 1983). The alignment (Fig. 1) defines two distinct types of sequence tracts, which differ in degree of conservation of sequence-tract length. Tracts that vary in length by at least 20 bases per 100 nucleotides compared with the *E. coli* sequence we defined as variable regions (V1–V11). Tracts that vary in length by less than 12 bases per 100 we defined as conserved regions (C1–C12). The 28S rRNA genes of all species compared can be characterized as consisting of only conserved and variable regions. Each type of sequence tract has a characteristic composition, structure and evolution.

Conserved Regions

The degree of sequence-tract length conservation among the conserved regions is striking. Sequences within these regions are highly conserved. The human sequences are 99% homologous to those of mouse and rat and 96% homologous to that of *X. laevis.* Thirty-one percent of the nucleotides within the conserved regions are invariant among the compared sequences (Fig. 1). Invariant nucleotides range from single bases to sequence tracts containing up to 14 nucleotides. Most tracts are at least four bases in length. These invariant tracts are separated by short (usually less than five bases) nonconserved sequence tracts that have remained unchanged in

Fig. 1A-D. Alignment of large-rDNA sequences: H, human 28S; M, mouse 28S; X, *Xenopus laevis* 28S; Y, yeast 26S rDNA. Sequences are from the RNA-like strand in all cases. The human sequence is numbered. Shaded areas denote conserved regions. Nonshaded, bracketed areas denote variable regions (see text). Total homology to the human sequence exists unless indicated by single-base changes or the following symbols: Δ, single-base insertion relative to the human sequence; ★, absence of base relative to the human sequence; [ ], insertion of more than one base relative to human sequence. Base changes within brackets are not noted. Lower-case letters within brackets indicate sequence tracts at least 4 bases longer than the corresponding human 28S rDNA sequences as follows (letter, organism, sequence): a, mouse, CCCGGCGCGC; b, *Xenopus,* GCGCGGCGGGACCC; c, *Xenopus,* CCCCCGGGG-GCGGGGGGGGGGCGCCGGCGGGCN; d, *Xenopus,* GCCCCCCCCGACGCCTCGCGGCGGCGGGGGGGCGGGGGC; e, *Xenopus,* GCCCCCCCCGACGCCTCGCGGCGGCGGGGGGGCGGGGGC; f, mouse, GCGCCCC; g, mouse, TCCGT. Sequences at least 10 bases long and 90% homologous to the corresponding *Escherichia coli* sequences are denoted by bars below the aligned sequences. Invariant nucleotides are denoted by horizontal brackets above and below the human sequence. Base 1519 is conserved among all the species except mouse

```
1                                                                                                   100
H CGCGACCTCAGATCAGACGTGGCGACCCGCTGAATTTAAGCATATTAGTCAGCGGAGGAAAAGAAACTAACCAGGATTCCCTCAGTAACGGCGAGTGAAC
M
X TCA            C                         C A                                C                     G
Y *TT      A    GTAG AGT         C         C A A                     C     G     G   T    G          G

101                                                                                                 200
AGGGAAGAGCCCAGCGCCGAATCCCCGCCCCGCGGGGCGCGGGACATGTGGCGTACGGAAGACCCGCTCCCCGGCGCCGCTCGTGGGGGGCCCAAGTCCT
                      G    TC   Δ   A                         A
                 GCG       C          G          G            G A*    CC  G   GGC C
C  C  A   T  AATTT   AT TG TA  TTC  TGC C AG*T   AATT *** G   GGGCAA TTT   G    T  C T TCTATGTTCC TGG

201                                                                                                 300
TCTGATCGAGGCCCAGCCCGTGGACGGTGTGAGGCCGGTAGCGGCCGGCGCGCGCCCGGGTCTTCCCGGAGTCGGGTTGCTTGGGAATGCAGCCCAAAGC
                                                   [a ]  G  T
             C          T         Δ        CC  [  b   ]   C T               T
**AACAG  *CGT  TAGA G TGA AATCCC TGT  CGAG AGT CG TT TTTG   C   GAA      A   T               T T   T

301                                                                                                 400
GGGTGGTAAACTCCATCTAAGGCTAAATACCGGCACGAGACCGATAGTCAACAAGTACCGTAAGGGAAAGTTGAAAAGAACTTTGAAGAGAGAGTTCAAG
                                             CGG                *
          T         A       TT  GA           CG          A  G T       A                A        GA  A

401                                                                                                 500
AGGGCGTGAAACCGTTAAGAGGTAAACGGGTGGGGTCCGCGCAGTCCGCCCGGAGGATTCAACCCGGCGGCGGGTCCGGCCGTGTCGGCGGCCCGGCGGA
                                                                             Δ        C    []
                                   *  T    G                           *******************************
 TA      TT   G A   G  GG CA TT A  A ACAT GTGTTTT TG*********************************************

501                                                                                                 600
TCTTTCCCGCCCCCCGTTCCTCCCGACCCCTCCACCCGCCCTCCCTTCCCCCGCCGCCCCTCCTCCTCCCCGGAGGGGGCGGGCTCCGGCGGGTGCG
        T                              G GT G     [    ]    G G    GG G T    C  [    ]   GGGGT GT TG T
**********GTTAGCGC GC GG    GGG    T  G GGA  CC    G CGGC        C  [] CGGGA  ************************
*********    T  T C    TGT ***************************************************************************

601                                                                                                 700
GGGGTGGGCGGGCGGGGCCGGGGGTGGGGTCGGCGGGGGACCGTCCCCCGGACCGGCGACCGGCCGCCGCCGGGCGCATTTCCAGGCGGTGCGCCGCGAC
 T  C C                                          *      *                      C      Δ
********************** CG TC CGC G       GCGG    G  CG****NC       GC NNNN C N*TT CCT C CG CG T  GCGCG
***************************************************************** TA GG AATCT        ACT G ***********

701                                                                                                 800
CGGCTCCGGGACGGCTGGGAAGGCCCGGCGGGGAAGGTGGCTCGGGGGGCCCCGTCCGTCCGTCCGTCCTCCTCCTCCCCCGTCTCCGCCCCCCGGCCCC
                C                T                     *********************************************    GG
         C  CG            A Δ           CG CC CC    GCG GCGG ****************************************
*********************************************************************************************************

801                                                                                                 900
GCGTCCTCCCTCGGGAGGGCGCGCGGGTCGGGGCGGCGGCGGCGGCGGCGGTGGCGGCGGCGGCGGGGGCGGCGGGACCGAAACCCCCCCCGAGTGTTAC
     TC********          C AA CACCTCA***************************************************
*****************************************************************************************************
*****************************************************************************************************

901                                                                                                1000
AGCCCCCCCGGCAGCAGCACTCGCCGAATCCCGGGGCCGAGGGAGCGAGACCCGTCGCCGCGCTCTCCCCCCTCCCGGCGCCCACCCCCGCGGGAATCCC
    T  *   C  GCTT*                    A   C Δ                T TC   CCGT  GC T    G C GG****
       [ ]            TCG                   ACGCCGG C  CG  GT CTC T    GGAG    T  CG  G T CCCCCC***
*****************************************************************************************************

1001                                                                                               1100
CGCGAGGGGGGTCTCCCCCGGCGCGGCGCCGGCGTCTCCTCGTGGGGGGGCCGGGCCACCCCTCCCACGGCGCGACCGCTCTCCCACCCCTCCTCCCCGC
*************************************CGTGG G T          G                                   *     GT
********************************************             G C  GCGGGGCGGGGAA G G  AG GG C   CG T C G*******
*****************************************************************************************************

1101                                                                                               1200
GCCCCCGCCCCGGCGACGGGGGGGGTGCCGCGCGCGGGTCGGGGGGCGGGGCGGACTGTCCCCAGTGCGCCCCGGGCGGGTCGCGCCGTCGGGCCCGGGG
 T T T********************  G GC C  *                                      **  Δ           Δ  Δ
**********************************GCGC  CT  [   ]           C          TC  C  C       C  A G G  A
*******************************C AGCATCA [   ] GT  CA     TAAAT  ATAG AATGTA CTT CC    GTAA  ATTATAGCCT

1201                                                                                               1300
GAGGTTCTCTCGGGGCCACGCGCGCGTCCCCCGAAGAGGGGGACGGCGGAGCGAGCGCACGGGGTCGGCGGCGACGTCGGCTACCCACCCGACCCGTCTT
 CCG  GG********T A    TC    Δ      C***************                           T
 GCCGC GG********************************************         CGC Δ      C       T       TGT
TG    AATA********************** T CCAGCT     T A    CT C ACGTAA  [ ]  AT  [             ] TG CG

1301                                                                                               1400
GAAACACGGACCAAGGAGTCTAACACGTGCGCGAGTCGGGGGCTCGCACGAAAGCCGCCGTGGCGCAATGAAGGTGAAGGCCGGCGCGCTCGCCGGCCGA
                         G           A        TC                                     G CC    C *G GG C
                         G  C              [     ] G      *  * T            A     G      G    C**      T
                         GTC AT        GTTT  G ****GT    A  * ATAC     T     A     [    ] G GC       ****AA

1401                                                                                               1500
GGTGGGATCCCGAGGCCTCTCCAGTCCGCCGAGGGGCACCACCGGCCCGTCTCGCCCGCCGCGCCGGGGAGGTGGAGCACGAGCGCACGTGTTAGGACCC
                                    Δ                                                         T   C
          CC    C CT CC      [ d ]                         C  T        N      N  GT        G  []
   **CA AATC A   GA    **********************TGAT     TCGGAT*********    TT *     T A     AT GC       G
```

A **Fig. 1A.**

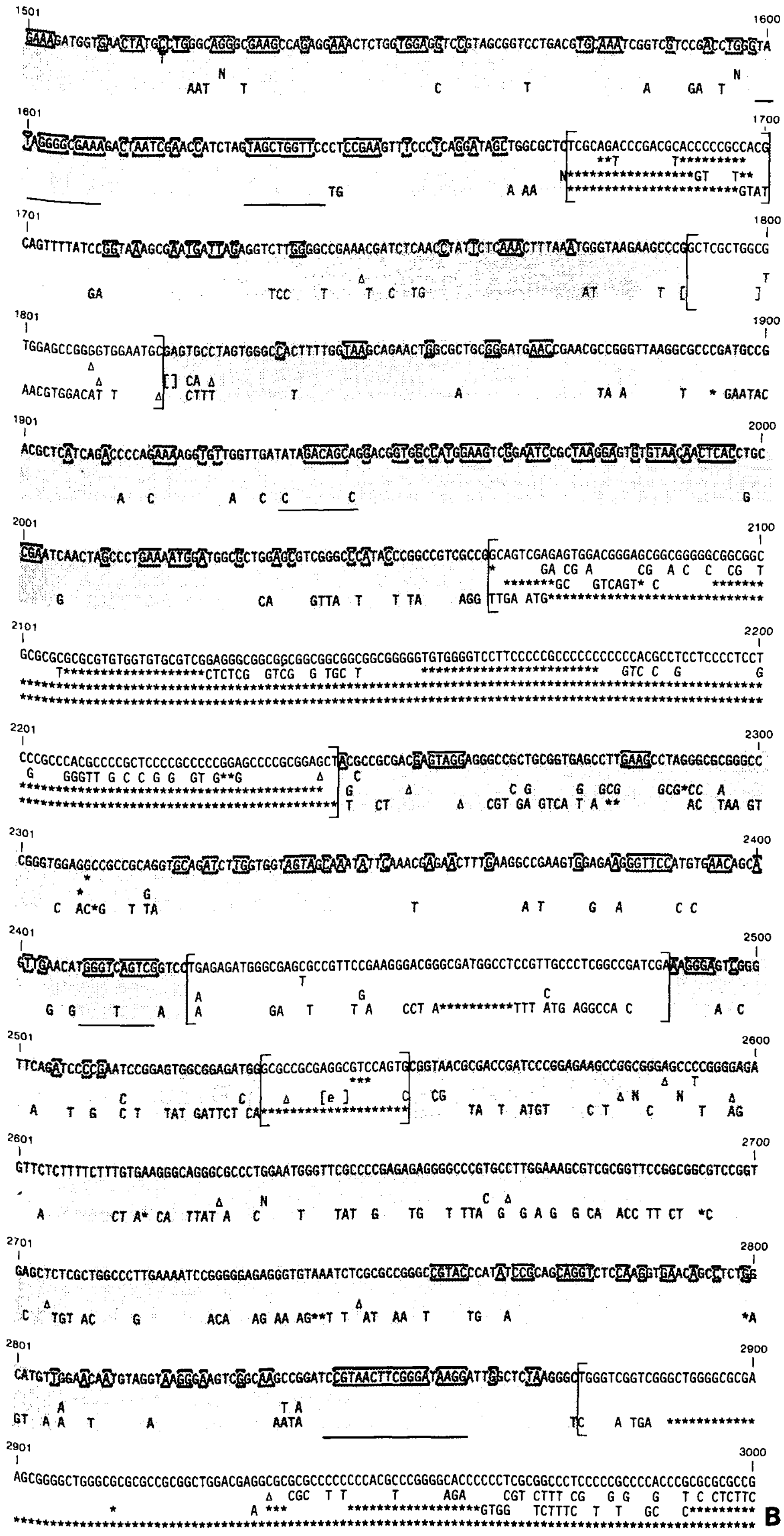

Fig. 1B.

```
3001                                                                                                3100
|                                                                                                      |
CTCGCTCCCTCCCCACCCCGCGCCCTCTCTCTCTCTCTCCCCCGCTCCCCGTCCTCCCCCCTCCCCGGGGGAGCGCCGCGTGGGGGCGCGGCGGGGGG
 C   GGGGC   GT GT C C G GTCG  G CAC      TT  C C T  T  T    GT GGGGG C    TCG GG TCG C C   GC    CTC
*********************************           *************GGC           G      ****     G   C G G *******
****************************************************************************************************

3101                                                                                                3200
|                                                                                                      |
AGAAGGGTCGGGGCGGCAGGGGCCGCGCGGCGGCCGCCGGGGCGGCCGGCGGGGGCAGGTCCCCGCGAGGGGGGCCCCGGGGACCCGGGGGGCCGGCGGC
C GG C G     T CAACCCC  G G GTT C GA  G  A GAA  A    TCC C    ***G G C       *    AC  T
*****************************************************************GC C    GGG C     ****** C    **
****************************************************************************************************

3201                                                                                                3300
|                                                                                                      |
GGCGCGGACTCTGGACGCGAGCCGGGCCCTTCCCGTGGATCGCCCCAGCTGCGGCGGGCGTCGCGGCCGCCCCCGGGGAGCCCGGCGGCGGCGCGGCGCG
  Δ  *                                      T                       T                      GT  *
   GC          C              T                         C   C *TCT  C  G GCC TCC    **********TCCT
****** C   TG TCA ACGCAGC  G G G TT     CT  TTGGTGG GCTT CT TG TA   G A TA TT************************

3301                                                                                                3400
|                                                                                                      |
CCCCCCACCCCCACCCCACGTCTCGGTCGCGCGCGCGTCCGCTGGGGGCGGGAGCGGTCGGGCGGCGGCGGTCGGCGGGCGGCGGGGCGGGGCGGTTCGT
GGT   CT    G GGGG C  G TCCAC C C CATCG CTCTCCC A  TGCGT G    G CG      G   TGT************************
  T    C  GT  **********************************G     CG   G   C N CN.CG  G     C ******       C C *****
****************************************************************************************************

3401                                                                                                3500
|                                                                                                      |
CCCCCCGCCCTACCCCCCCGGCCCCGTCCGCCCCCCGTTCCCCCCTCCTCCTCGGCGCGCGGCGGCGGCGGCGGCAGGCGGCGGAGGGGCCGCGGGCCGG
***      G GTGTGGGGG AA  TCCG   T GGT*   ***********************************************************
****************************************************************************************************
****************************************************************************************************

3501                                                                                                3600
|                                                                                                      |
TCCCCCCCGCCGGGTCCGCCCCCGGGGCCGCGGTTCCGCGCGCGCCTCGCCTCGGCCGGCGCCTAGCAGCCGACTTAGAACTGGTG]CGGACCAGGGGAAT
**                       C           TTC       [f]
*************************************************                    T                N
*****G GTG CTTGTT TAGA    CCTT GTAGGT T TT TAGAC   T G TTG TA AAT   A GAT A           ]A    A

3601                                                                                                3700
|                                                                                                      |
CCGACTGTTTAATTAAAACAAAGCATCGCGAAGGCCCGCGGCGGGTGTTGACGCGATGTGATTTCTGCCCAGTGCTCTGAATGTCAAAGTGAAGAAATTC
                                       A
 T     C         T     T      T T A AAAGT A          A

3701                                                                                                3800
|                                                                                                      |
AATGAAGCGCGGGTAAACGGCGGGAGTAACTATGACTCTCTTAAGGTAGCCAAATGCCTCGTCATCTAATTAGTGACGCGCATGAATGGATGAACGAGAT
  CC                                                                                       T

3801                                                                                                3900
|                                                                                                      |
TCCCACTGTCCCTACCTACTATCCAGCGAAACCACAGCCAAGGGAACGGGCTTGGCGGAATCAGCGGGGAAAGAAGACCCTGTTGAGCTTGACTCTAGTC
                          T
              T           T                           A                                        T

3901                                                                                                4000
|                                                                                                      |
TGGCACGGTGAAGAGACATGAGAGGTGTAGAATAAGTGGGAG[CCCCCGGCGCCCCCCCGGTGTCCCCGCGAGGGGCCCGGGGCGGGGTCCGCGGCCCTG
                                                **** ***      TCCTC CGT    Δ     CA  Δ     TC
 CA  T                       G                 *   T GT **************** CAAA      ****************
 A  TT              AGAG                      C*********************************************** T G*******

4001                                                                                                4100
|                                                                                                      |
CGGGC]CGCCGGTGAAATACCACTACTCTGATCGTTTTTTCACTGACC[CGGTGAGGCGGGGGGGCGAGCCCGAGGGGCTCTCGCTTCTGGCGCCAAGCGCC
                             C                                      Δ                        [g]
***                          T           T             NN      Δ                AC               N
*****]   A                 CT T  A   C  T   T TT[AA    CG A CT    AATT  [     ]****A   *   Δ CATT    GT

4101                                                                                                4200
|                                                                                                      |
CGCCCGGCCGGGC]GCGACCCGCTCCGGGGACAGTGCCAGGTGGGGAGTTTGACTGGGGCGGTACACCTGTCAAACGGTAACGCAGGTGTCCTAAGGCGAG
 G   TG      Δ
Δ   Δ                       A      G                                   C
ATT   ***   ]*T  T   GGTT AA   T  T              G         C    T    T    A        A         G GC

4201                                                                                                4300
|                                                                                                      |
CTCAGGGAGGACAGAAACCTCCCGTGGAGCAGAAGGGCAAAAGCTCGCTTGATCTTGATTTTCAGTACGAATACAGACCGTGAAAGCGGGGCCTCACGAT
    Δ  Δ                                                        T                  Δ   Δ
   T   A        T    A    A  A     T      C C      T            GT       A   A      T T     AT

4301                                                                                                4400
|                                                                                                      |
CCTTCTGACCTTTTGGGTTTTAAGCAGGAGGTGTCAGAAAAGTTACCACAGGGATAACTGGCTTGTGGCGGCCAAGCGTTCATAGCGACGTCGCTTTTTG
        T                                                      Δ
  TA T  C CG AA   G G  TA  C   C                               A T                   A T       C

4401                                                                                                4500
|                                                                                                      |
ATCCTTCGATGTCGGCTCTTCCTATCATTGTGAAGCAGAATTCGCCAAGCGTTGGATTGTTCACCCACTAATAGGGAACGTGAGCTGGGTTTAGACCGTC
                                        A
                     N                  A
 T                         ACC          GT
```

C

Fig. 1C.

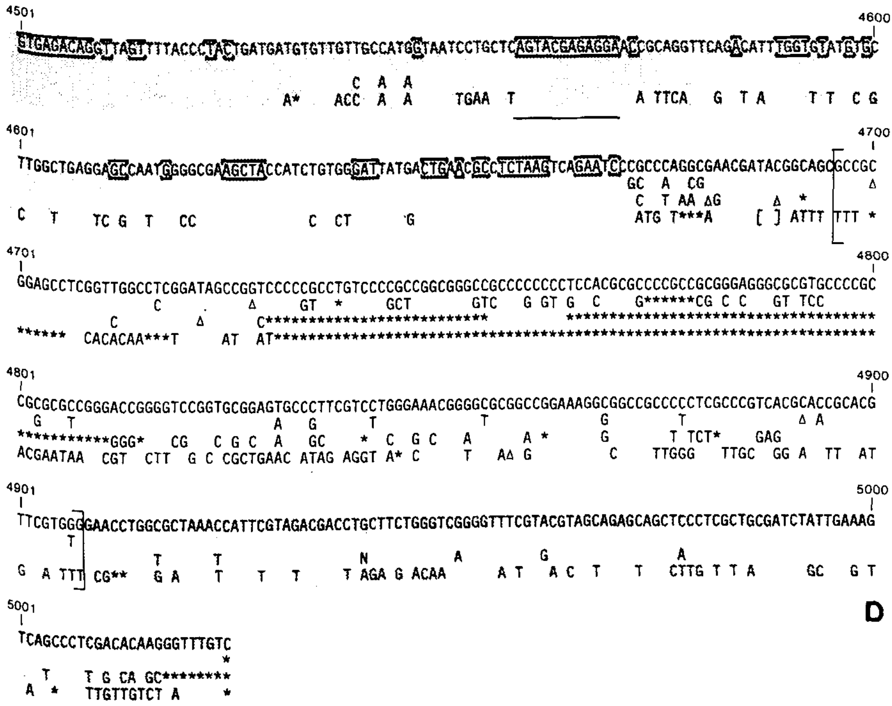

Fig. 1D.

length during evolution. A total of 319 invariant tracts comprising 800 bases are distributed within 11 of the 12 conserved regions.

Secondary-Structure Model of Human 28S rRNA

The proposed secondary structure of the human 28S and 5.8S rRNAs is shown in Fig. 2. This bimolecular structure is divided into seven domains defined by long-range RNA base pairing. Six domains are largely superimposable on those of the Noller (1984) *E. coli* 23S rRNA secondary structure. No sequence tracts or structures homologous to domain 2, composed entirely of variable nucleotides, exist within *E. coli* 23S rRNA. Our model consists of 144 helixes. The conserved regions have 94 helixes; the variable regions, 50.

Conserved Structural Elements

Using the criterion that the existence of a helix can be considered confirmed if two CBCs have been demonstrated in at least two different homologous sequence tracts (Woese et al. 1983), we confirmed the existence of 69 of the 94 helixes making up the 28S rRNA conserved regions (Tables 1–7). Thirteen unconfirmed helixes are supported by a single CBC; their existence is considered probable. Four of the 13 helixes (26, 38, 45, and 50) are supported either by CBCs in mitochondrial large rRNAs or by the analysis of chemically modified *E. coli* 23S rRNA (see notes to Tables 1–7). Comparative support is unavailable for seven helixes (2, 13, 18, 48, 57, 62, and 93). Five helixes (4–8) are formed from 5.8S rRNA. Support for their existence has been presented previously (see Vaughn et al. 1984).

Several helixes exhibit noncanonical compensatory base pairing (pairing other than Watson–Crick, such as A–G). The existence of A–G base pairing is supported by comparative analysis, RNA chemical modification (Noller 1984), and magnetic resonance spectroscopy (Kan et al. 1983). The occurrences and locations of these base pairs are noted in Tables 1–7.

Variable Regions

Eleven variable regions account for the 83% increase in size of human large rRNA over that of *E. coli.* The human variable regions consist primarily of stretches of short, G + C-rich (82.4% G + C compared with 55.8% in conserved regions) repetitive sequences most clearly demonstrated in the largest variable regions: V2, V5, V8, and V11. In humans, the most common repeated sequences are poly(GGC), poly(CT), poly(CCT), poly(G), and poly(C). Within the V8 region, a 257-base, 87% G + C tract (see Fig. 1, bases 2940–3190) exhibits first a 81% pyrimidine sequence tract of 126 bases followed by a 66% purine tract of 125 bases. The presence of long purine- and pyrimidine-rich stretches is expected to favor the formation of long, stable hairpins in the RNA encoded by these re-

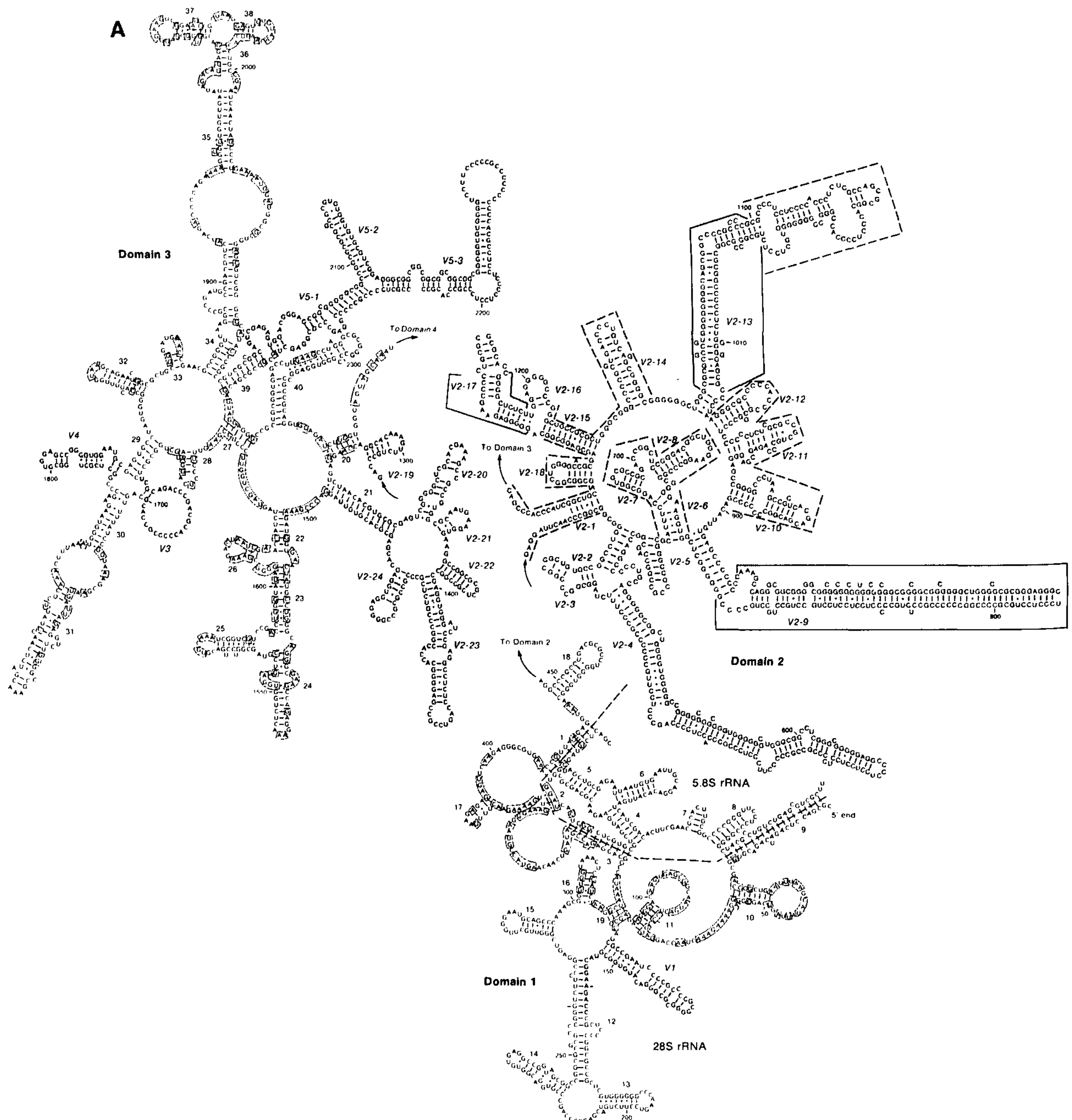

Fig. 2A, B. Domains 1–7 of the human 28S rRNA molecule. Helixes are consecutively labeled and correspond to those listed in Tables 1–7. Helixes composed of variable sequence tracts are denoted as belonging to a variable region (e.g., V2 is the second variable region) and by relative position. Variable sequence tracts are denoted by shading. Invariant nucleotides are boxed. Within V2 and V8, sequences conserved among the vertebrates are denoted by dashed lines, sequences specific to mammals are enclosed in solid lines

gions—a structural feature unique to the variable regions (Fig. 2). Such structures have been observed by electron microscopic studies of mammalian large rRNA (Schibler et al. 1977).

Our alignment identifies conserved collinear sequence tracts within the variable regions (Fig. 1). These conserved tracts differ from those in the conserved regions being (1) less numerous, (2) conserved over shorter evolutionary distances, and (3) interrupted by sequences of low homology that vary in length during evolution. Only one conserved sequence tract, V2-19, a 31-nucleotide stretch capable of forming a conserved structural element (Fig. 2) is totally conserved in all eukaryotes examined. All other conserved tracts are vertebrate specific (Figs. 1 and 2).

Secondary Structure of the Variable Regions

The variable-region sequence tracts are capable of forming secondary structures that do not disrupt the superimposable secondary structures of the con-

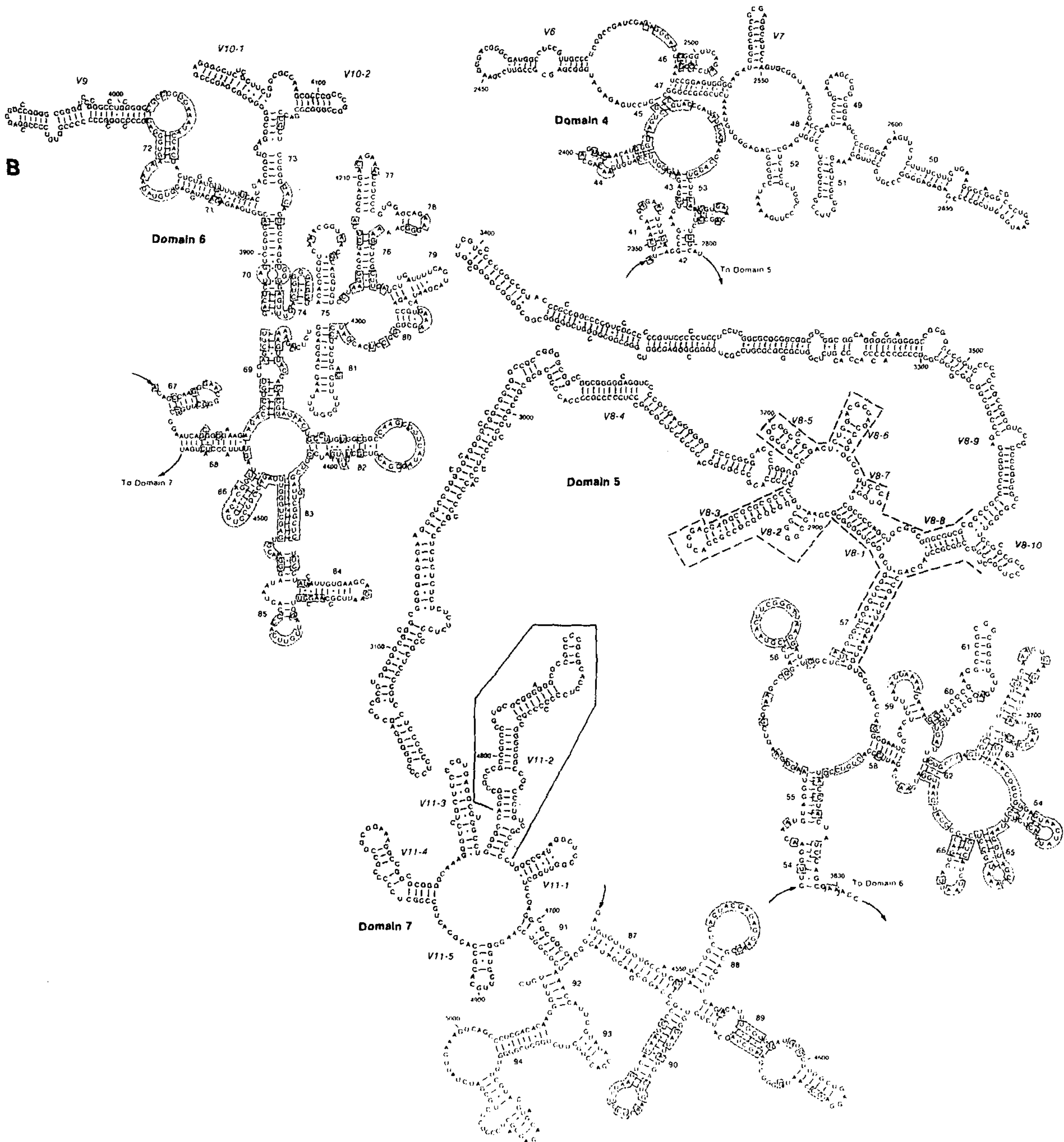

Fig. 2. Continued

served regions. For the compared sequences the secondary structures of the variable regions are similar but not necessarily superimposable. This is best demonstrated in domain 2, which has a secondary-structure core conserved in all vertebrates. Mammal- or human-specific variable sequence tracts are capable of either extending core helixes (V2-13, V2-17) or forming mammal- or human-specific structural elements (V2-9) (see Fig. 2).

Our alignment provides comparative support for the existence of helixes that make up more than 60% of the human variable sequence tracts. Helixes V1, V2-1, V2-20, and V2-22 are confirmed by the demonstration of at least two CBCs in each (Fig. 3). The insertion and/or deletion of complementary sequence tracts define helixes V7, V9, and V10-2 (see Fig. 4). Complementary sequence tracts unique to human or mammalian rDNA support the existence of helixes unique to humans (V2-9, V2-13, and V2-17) and of those unique to but not necessarily conserved among mammals (V2-2, V2-5, V5, V8-4, and V8-9). Twenty-one helixes contain sequence tracts that are conserved in all vertebrate variable regions; these tracts represent conserved core structures within the variable regions of these species (Figs. 1 and 2).

Table 1. Domain 1 (613 nucleotides, 20 stems)

Stem no.[a]	Human bases paired in stem	Number of CBCs[b]	Models in agreement[c]					Notes
			American *E. coli*	German *E. coli*	Yeast	*Xenopus*	Mouse	
1	4–14[d]/411–421	10 irr	(+)	(+)	+	+	+	
2	16–18[d]/370–372	0 irr, ilp, sl	(+)	(+)	(+)	(+)	X	e
3	22–32[d]/333–344	9 irr	(+)	(+)	(+)	+	+	
4	33–40[d]/91–98[d]		X	+	+	+	+	
5	45–50[d]/55–60[d]		X	+	+	+	+	
6	65–72[d]/83–90[d]		X	+	+	+	+	
7	107–110[d]/113–116[d]		(+)	(+)	(+)	+	X	
8	118–124[d]/130–136[d]		(+)	(+)	(+)	+	+	
9	137–157[d]/2–22	7Y, sl	(+)	(+)	+	+	+	f
10	26–33/49–57	3 irr	(+)	+	+	+	(+)	
11	77–82/101–106	1	+	+	+	+	+	
V1	115–131/136–153	8Y, 1M, ilp, sl	–	–	+	+	+	
12	156–179/246–268	15YMX, sl, irr, ilp	(+)	(+)	(+)	(+)	+	
13	183–190/198–206		(+)	(+)	(+)	X	X	g
14	218–228/234–244	1X	(+)	(+)	(+)	X	X	
15	274–280/289–295	4 ilp	+	+	+	+	+	
16	301–306/313–318	2	+	+	+	+	+	
17	381–384/389–392	2	+	+	X	+	X	
18	427–434/445–452		(+)	(+)	(+)	+	X	h
19	108–111/321–324	1	+	+	X	+	X	

[a] "V" before number denotes a stem composed of variable sequence tracts

[b] Compensatory base changes (CBCs) were determined as described in Methods from human, yeast, *Physarum*, and *E. coli* alignments unless denoted as follows: M, mouse; X, *Xenopus laevis*; and Y, yeast-specific CBCs compared with the human sequence. sl, 1- to 2-base slippage of aligned nucleotides; equivalent base pairing of aligned sequences is possible by the displacement of one strand by a nucleotide. irr, stem length varies among compared sequences by 1–2 base pairs. ilp, placement or size of internal loops varies among sequences compared; noncanonical base pairs within stems, such as A–G base pairs, are noted

[c] +, identical base-paired helix. (+), nearly identical stem with modifications. X, different secondary structure proposed. –, not applicable; equivalent bases do not exist

[d] 5.8S rRNA, 1–159 bases. Otherwise 28S rRNA, 1–454 bases

[e] This stem is anomalous. Our structure agrees with the American model. A bulge is present in the German model. All sequences can form a similar stem with slippage

[f] This stem could vary in size at the expense of stem 8. Its existence is confirmed by 7 CBCs between the yeast and human sequences. A homologous stem exists in both of the *E. coli* models. The nucleotides between the 3′ 5.8S terminus and the 5′ 28S origin are processed out of the initial transcript to generate the free rRNA ends of this stem

[g] These stems are confirmed in *E. coli*. All species are capable of forming three stems similar in structure to those proposed for *E. coli*. However, precise alignments are not possible since small species-specific insertions and deletions are present. Stem 12 is confirmed proximally with 15 CBCs. Our models for stems 13 and 14 are not supported by CBCs but are more similar to the *E. coli* model than to the remaining eukaryotic models

[h] This stem is confirmed in the *E. coli* models (Maly and Brimacombe 1983). A slippage in alignment is necessary to form a similar stem in eukaryotes. These sequences are highly conserved and no eukaryotic CBCs are available to support this stem. The mouse model forms a mouse-specific structure with these nucleotides

Comparisons with Other Large-rRNA Secondary-Structure Models

The human 28S rRNA secondary structure is compared to two *E. coli* 23S rRNA models [the American (Noller 1984) and German (Maly and Brimacombe 1983) models] and eukaryotic cytoplasmic large-rRNA models for yeast (Veldman et al. 1981), *X. laevis* (Clark et al. 1984), and mouse (Michot et al. 1984) in Tables 1–7. With the exception of the structural features of the variable regions, the domains of any one species are superimposable upon those of another. Our model disagrees with the American, German, yeast, *Xenopus*, and mouse models in only 11, 7, 21, 13, and 14 stems, respectively. Sixty-three CBCs support the existence of these contested helixes (see Fig. 3 and the notes to Tables 1–7). Helix 20, which defines domain 3 and is supported by 4 CBCs, was not proposed in the yeast or *Xenopus* models.

Most of the differences between the human and *E. coli* models (9 of 11 discrepancies in the American model and 5 of 7 in the German) are restricted to three specific regions of the secondary structure: the 5.8S rRNA helixes, domain 4, and the last three helixes of the large-rRNA molecule. Helixes 48–52 of domain 4 (see Fig. 2) represent the most uncertain region of the *E. coli* models (Maly and Brimacombe

Table 2 Domain 2 (bases 445–1293; 839 bases, 18 stems)

Stem no.	Human bases paired in stem	Number of CBCs	Models in agreement				
			American *E. coli*	German *E. coli*	Yeast	*Xenopus*	Mouse
V2-1	464–469/1275–1280	3Y irr	–	–	X	X	X
V2-2	473–479/648–654		–	–	–	–	X
V2-3	480–486/490–499		–	–	–	–	X
V2-4	503–569/575–640		–	–	–	–	X
V2-5	656–664/669–675		–	–	–	–	X
V2-6	677–683/732–739		–	–	–	X	X
V2-7	686–689/694–697		–	–	–	X	X
V2-8	706–712/722–728		–	–	–	X	X
V2-9	740–810/813–895		–	–	–	–	X
V2-10	902–914/920–937		–	–	–	X	X
V2-11	941–955/960–971		–	–	–	X	X
V2-12	973–980/989–995		–	–	–	X	X
V2-13	996–1060/1066–1139		–	–	–	–	X
V2-14	1147–1158/1163–1174		–	–	–	X	X
V2-15	1181–1191/1244–1254		–	–	–	X	X
V2-16	1194–1195/1203–1204		–	–	–	–	X
V2-17	1205–1220/1205–1242		–	–	–	–	X
V2-18	1256–1261/1269–1273		–	–	–	X	X

Symbols and abbreviations as in Table 1

1983). Our model is in general agreement with the *E. coli* models and, as in *E. coli*, the helixes in our model are capable of forming alternative stem loop structures (Maly and Brimacombe 1983). We propose an additional stem in this region, helix 52, which is supported by 3 CBCs (Fig. 3). A helix analogous to our helix 50 is present in both prokaryotic models but precise sequence alignments and confirming CBCs are unavailable. RNA–RNA cross-linking studies support this helix's existence in *E. coli* (Stiege et al. 1982, 1983).

There is extensive evidence that eukaryotic 5.8S rRNA is homologous to the first 160 nucleotides of *E. coli* 23S rRNA and that it interacts through base pairing with the 5′ end of the eukaryotic large rRNA (reviewed in Walker and Pace 1983). Vaughn et al. (1984) have used comparative approaches similar to ours to yield a universal secondary-structure model for the 5.8S rRNA molecule and its interactions with the 28S rRNA. The Vaughn model is in excellent agreement with ours and with the German *E. coli* model.

At the 3′ end of the human 28S rRNA molecule, the existence of helix 94 is confirmed by 5 eukaryotic CBCs. No interaction similar to the base pairing of the 5′ and 3′ ends of the *E. coli* 23S rRNA molecule has been convincingly demonstrated in eukaryotes (Vaughn et al. 1984). The discrepancies between the eukaryotic and prokaryotic models within domain 7 are probably related to fundamental differences in structure. No invariant sequences are present in this region (Fig. 1).

The only other differences between our model and the prokaryotic models are in helixes 26, 44, and 45. Each of these helixes is supported by a single CBC and agrees with either the German (helix 26) or the American (helixes 44 and 45) model.

Discussion

Conserved Structural Features

We have demonstrated that for the species examined the rRNA secondary structures of the conserved sequence tracts are superimposable and in close agreement. We have identified 800 invariant nucleotides within the cytoplasmic large-rRNA sequence and placed them in a secondary-structural context. The conserved distribution of invariant nucleotides separated by specific lengths of nonconserved nucleotides represents a consensus sequence for the conserved regions. Many invariant nucleotide tracts include bulged nucleotides within helixes, possibly a common feature of RNA–protein contact sites (Peattie et al. 1981). The existence of a conserved superimposable structural core within the 28S rRNA molecule is supported by experiments that demonstrate conserved heterologous rRNA–protein interactions (Gourse et al. 1981; El-Baradi et al. 1985).

The conservation of rRNA sequences and secondary structures implies that they have been selected for during evolution and possess functional

Table 3. Domain 3 (bases 1291–2333; 1043 bases, 32 stems)

Stem no.	Human bases paired in stem	Number of CBCs	Models in agreement: American *E. coli*	German *E. coli*	Yeast	*Xenopus*	Mouse	Notes
V2-19	1294–1297/1306–1309		–	–	X	+	X	
20	1310–1316/2327–2333	4	+	(+)	X	X	+	
21	1319–1332/1483–1496	4YX	+	+	+	+	+	
V2-20	1336–1347/1354–1363	2Y	+	+	+	(+)	+	
V2-21	1364–1366/1374–1376		+	(+)	X	X	X	
V2-22	1380–1386/1391–1397	3M	–	–	X	X	+	
V2-23	1398–1421/1431–1454		–	–	X	X	(+)	
V2-24	1457–1462/1472–1477		–	–	X	X	(+)	
22	1504–1510/1622–1628	6 irr	+	+	+	+	+	
23	1514–1524/1593–1603	6	+	+	+	+	+	
24	1527–1540/1545–1559	7	+	+	+	+	+	
25	1563–1572/1581–1588	7 ilp	+	+	+	+	+	
26	1607–1609/1618–1620	1	X	+	+	+	X	a
27	1640–1645/2259–2264	4	+	+	+	+	+	
28	1651–1654/1659–1663	4 ilp	+	+	+	+	(+)	b
29	1667–1673/1820–1826	7	+	+	(+)	X	+	c
V3	1676–1677/1700–1701		–	–	–	–	+	
30	1703–1712/1775–1786	6 irr (A–G)	+	+	X	X	+	
31	1723–1743/1747–1765	11 irr, ilp, sl	+	+	X	+	+	
V4	1790–1799/1805–1819		–	–	X	(+)	(+)	
32	1833–1842/1848–1857	9 irr	+	+	+	+	+	
33	1862–1864/1871–1873	1	+	+	(+)	+	+	d
34	1877–1906/2034–2055	12 ilp, irr	+	+	+	+	+	
35	1922–1935/2004–2016	9 irr (3 A–G)	+	+	+	+	(+)	
36	1943–1947/1996–2000	3	+	+	+	+	+	
37	1949–1956/1966–1973	7 (2 A–G)	+	+	+	+	+	e
38	1979–1982/1991–1994	1 (A–G)	X	+	(+)	X	X	f
39	2056–2065/2246–2254	5Y irr, ilp	+	+	+	+	+	
V5-1	2075–2076/2221–2244		–	–	–	–	X	
V5-2	2097–2110/2115–2127		–	–	–	–	X	
V5-3	2128–2162/2181–2219		–	–	–	–	X	
40	2268–2291/2299–2319	11 ilp (A–G)	+	+	(+)	(+)	(+)	g

Symbols and abbreviations as in Table 1

[a] This stem is preferred. A single CBC and complementary constant nucleotides exist. All mitochondria share a deletion of this stem (Maly and Brimacombe 1983). The German model considers this stem confirmed

[b] The various models differ with the respect to the length of the proximal portion of this stem

[c] *X. laevis* sequence contains an insertion and does not allow the CBCs that all other species demonstrate

[d] Models differ in the length of the stem

[e] Two A–G base pairs exist as CBCs to conserve this stem

[f] This stem exists only in the German model but is confirmed by CBCs in *Saccharomyces cervisiae* mitochondrion (Maly and Brimacombe 1983) and by a single A–G CBC between *E. coli* and the eukaryotes

[g] Various models exhibit slippage in base pairing and variation on where internal loops occur. Existence of this stem in prokaryotes is supported by RNA crosslinking data (Stiege et al. 1982)

significance. Support for the notion that translation may be a function of interacting RNA molecules (Noller 1980; Woese 1980) has grown since RNA was shown to have enzymelike properties (Zaug and Cech 1986). Prokaryotic large rRNA has been implicated to be functionally significant in the assembly of the 50S subunit, in ribosomal protein binding, in translational fidelity, in tRNA binding, and in the peptidyl transferase and elongation-GTPase activities (Noller 1984). By analogy, the nucleotides involved in these functions within the human 28S rRNA molecule can be identified.

Comparative analyses with more species may reduce the number of nucleotides considered invariant and will provide a firmer statistical basis for their identification.

Variable Regions

The limited number of sites for the variable regions suggests that a limited number of positions in which size variation is compatible with translational function exist within the structural core of the large-rRNA molecule. The presence of a mammal-specific

Table 4. Domain 4 (bases 2334–2803; 470 bases, 15 stems)

Stem no.	Human bases paired in stem	Number of CBCs	Models in agreement					Notes
			American *E. coli*	German *E. coli*	Yeast	*Xenopus*	Mouse	
41	2348–2353/2361–2366	5	+	+	+	+	+	
42	2368–2371/2798–2801	5	+	+	X	+	+	
43	2375–2380/2778–2783	3 irr	+	+	X	+	+	
44	2386–2392/2406–2412	8	+	X	X	(+)	(+)	a
45	2418–2420/2756–2758	1	+	X	X	X	X	b
V6	2432–2445/2460–2476		–	–	X	X	X	
46	2496–2500/2509–2513	4 irr (A–G)	+	+	X	+	+	
47	2515–2527/2744–2755	10 ilp, irr	+	+	X	+	+	
V7	2531–2538/2543–2550		–	–	X	+	+	
48	2564–2567/2697–2700		+	X	X	X	+	
49	2569–2573/2584–2588	1	X	+	X	X	X	
50	2592–2632/2639–2671		+	+	X	X	+	
51	2675–2681/2689–2695	6Y	X	+	X	X	X	
52	2704–2709/2726–2731	3Y	X	X	X	X	X	
53	2784–2787/2793–2796	4	+	+	+	+	+	

Symbols and abbreviations as in Table 1

[a] The German and yeast models exhibit relative slippage. Our favored stem structure exhibits 8 CBCs and conserves the complementary base pairing of the constant nucleotides in the proximal portion. An AA dinucleotide bulge appears to be a constant feature

[b] A single CBC supports this stem, present only in the American model. We consider this a preferred structure; it is composed primarily of complementary constant nucleotides. Both this and stem 44 are strongly supported by physical analysis of an *E. coli* r-protein–RNA complex of this region (Vester and Garrett 1984)

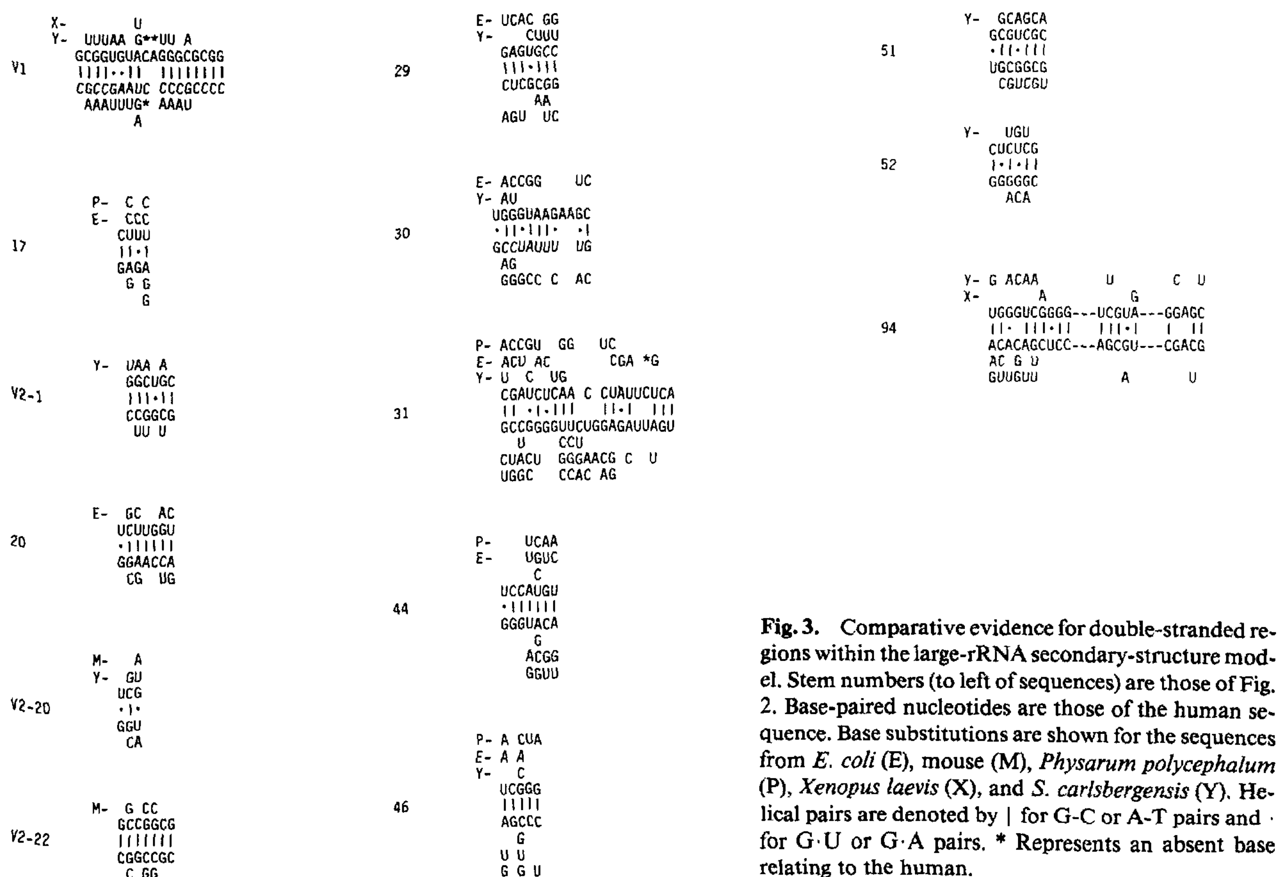

Fig. 3. Comparative evidence for double-stranded regions within the large-rRNA secondary-structure model. Stem numbers (to left of sequences) are those of Fig. 2. Base-paired nucleotides are those of the human sequence. Base substitutions are shown for the sequences from *E. coli* (E), mouse (M), *Physarum polycephalum* (P), *Xenopus laevis* (X), and *S. carlsbergensis* (Y). Helical pairs are denoted by | for G-C or A-T pairs and · for G·U or G·A pairs. * Represents an absent base relating to the human.

variable region (V5) suggests that our list of variable regions may not include all possible such regions in other species. Evidence that variable regions exist within the mature large-rRNA molecules includes (1) the agreement in nucleotide composition and size between the large-rRNA molecules and what is predicted from their DNA sequences, (2) the detection by electron microscopy of the predicted struc-

Table 5. Domain 5 (bases 2804–3833; 1030 bases, 19 stems)

Stem no.	Human bases paired in stem	Number of CBCs	Models in agreement: American *E. coli*	German *E. coli*	Yeast	*Xenopus*	Mouse	Notes
54	2804–2809/3822–3827	3	+	(+)	(+)	+	(+)	a
55	2815–2821/3813–3819	11 irr	+	+	+	(+)	+	b
56	2840–2844/2860–2864	2	+	+	+	+	+	
57	2869–2885/3569–3584		(+)	(+)	X	(+)	(+)	c
V8-1	2888–2897/3241–3250		–	–	X	X	X	
V8-2	2901–2903/2907–2909		–	–	X	X	X	
V8-3	2910–2923/2930–2943		–	–	X	X	X	
V8-4	2944–3062/3069–3191		–	–	–	–	+	
V8-5	3193–3197/3202–3207		–	–	X	X	X	
V8-6	3210–3214/3222–3226		–	–	X	X	X	
V8-7	3230–3232/3237–3239		–	–	X	X	X	
V8-8	3256–3263/3557–3564		–	–	X	X	X	
V8-9	3265–3395/3400–3535		–	–	–	–	+	
V8-10	3536–3542/3548–3555		–	–	–	–	X	
58	3595–3601/3798–3804	5 irr	+	+	+	+	+	
59	3608–3610/3620–3622	1 sl	(+)	+	(+)	(+)	+	d
60	3623–3630/3652–3659	6	+	+	+	+	+	
61	3634–3638/3642–3646	4	+	+	(+)	+	+	
62	3664–3667/3787–3790		(+)	(+)	X	(+)	(+)	e
63	3673–3685/3694–3715	4 ilp, irr	(+)	(+)	+	+	(+)	f
64	3722–3727/3735–3740	2	+	+	+	+	+	
65	3744–3748/3755–3759	sl	+	+	+	+	+	
66	3760–3765/3773–3778	2 irr	+	+	+	+	+	

Symbols and abbreviations as in Table 1

[a] The models differ in the length of the distal portion of this stem. The distal nucleotides are constant. No CBCs exist to support the extension of this stem

[b] *Xenopus* model exhibits slippage and altered base pairing compared with the other models

[c] This stem is variable and forms the base of the V8 stems. Psoralen crosslinking supports the existence of this stem in prokaryotes (Turner and Noller 1983). This stem is partially deleted in other organisms (Glotz et al. 1981). All sequences are capable of forming a stem with slippages in alignment. No definitive CBCs are demonstrated. These sequences are highly conserved among eukaryotes

[d] This stem is in all models, but only a single CBC exists

[e] This stem is confirmed by CBCs if mismatched bases or internal loops are allowed (Maly and Brimacombe 1983). Alternatively, this stem can be composed largely of constant nucleotides, as in our model

[f] The distal portion of this stem is variable and cannot be described in terms of CBCs. This stem is smaller in eukaryotes. Similar structures can be constructed for all species

tures within the variable regions (Schibler et al. 1977), (3) the protection from S1 nuclease of rRNA–rDNA hybrids (Ware et al. 1983), and (4) the direct sequencing of the variable rRNA regions (Hassouna et al. 1984). This implies that the products of these genes are mosaic RNA molecules.

The nature of the processes that have generated sequence variability within these regions in vertebrates can be inferred from the sequence alignments. The presence of homologous collinear tracts within these regions indicates that these sequences share common ancestral origins. The nucleotides of those tracts are capable of forming eukaryote- or vertebrate-specific core structural elements within the variable regions, which implies that these structures have been selected for during evolution and may have functional significance. Differences in the spacing between these homologous tracts indicate a history of repeated duplications and/or deletions in the intervening divergent regions. The overall lack of homology in the variable regions implies that multiple changes have occurred. These regions consist primarily of repetitive simple sequences, which, in model systems, have been shown to arise by unequal homologous recombinational events in DNA that are not maintained by selection (Smith 1976). Similar results have been obtained in comparative studies of ribosomal external transcribed (Furlong et al. 1983), internal transcribed (Hall and Maden 1980; Subrahmanyam et al. 1982; Furlong and Maden 1983; Michot et al. 1983), and nontranscribed spacers (Erickson and Schmickel 1985).

The lack of sequence conservation in large tracts of the variable regions raises the possibility that the nucleotides of those regions serve no function. Alternatively, the variable sequences may serve species-specific functions. Eukaryotic rDNA transcription is highly specialized and is known to be dependent on species-specific rDNA sequences and ancillary factors (Grummt et al. 1982). The human variable regions share sequence homology with the herpes simplex type I virus (Jones et al. 1985). Ribosomal

Table 6. Domain 6 (bases 3834–4529; 696 bases, 23 stems)

Stem no.	Human bases paired in stem	Number of CBCs	Models in agreement: American *E. coli*	German *E. coli*	Yeast	*Xenopus*	Mouse	Notes
67	3837–3842/3851–3856	3 irr	+	+	+	+	+	
68	3860–3874/4515–4529	9 irr, ilp	+	+	+	+	(+)	
69	3878–3890/4340–4353	5	+	+	+	+	(+)	
70	3891–3906/4133–4149	6 irr	+	+	+	+	+	
71	3908–3924/4025–4042	12 ilp, irr	+	+	+	+	+	
72	3935–3939/4019–4023	1	+	+	+	+	+	
V9	3942–3971/3977–4006	1Y	(+)	(+)	–	(+)	(+)	
73	4046–4058/4119–4128	6Y ilp	(+)	+	+	+	+	
V10-1	4059–4070/4074–4087		–	–	X	+	X	
V10-2	4096–4102/4109–4115	3M	–	–	X	+	(+)	
74	4152–4155/4159–4162	2 (A–G)	+	+	+	+	+	
75	4163–4170/4182–4189	5 irr	+	+	+	+	+	
76	4194–4202/4242–4250	6	+	+	+	+	+	
77	4205–4210/4218–4223	7	+	+	+	+	+	
78	4229–4230/4237–4238	1Y	(+)	+	X	X	X	a
79	4253–4261/4268–4277	10 irr, ilp	+	+	+	+	+	b
80	4278–4281/4286–4289	4 irr	+	+	+	+	+	
81	4301–4312/4322–4331	13 ilp (3 A–G)	+	+	+	+	+	
82	4360–4372/4390–4404	6 irr, ilp	+	+	+	(+)	(+)	c
83	4408–4423/4473–4491	3	+	+	+	+	+	
84	4424–4434/4441–4452	5 ilp	+	+	(+)	+	+	
85	4453–4455/4465–4467	2	+	+	+	+	+	
86	4494–4500/4505–4512	1	+	+	+	+	+	

Symbols and abbreviations as in Table 1

[a] Both prokaryotic models form a stem from these nucleotides. Base pairing in eukaryotes is limited and definitive CBCs are lacking

[b] This stem reveals eukaryotic/prokaryotic specificity and strong intrakingdom conservation. Similar structures can be formed in the two kingdoms if irregularities in alignment are allowed

[c] Models vary in extent of distal base pairing of this stem

Table 7. Domain 7 (bases 4530–5025; 496 bases, 12 stems)

Stem no.	Human bases paired in stem	Number of CBCs	Models in agreement: American *E. coli*	German *E. coli*	Yeast	*Xenopus*	Mouse	Notes
87	4532–4547/4675–4690	5MXY, irr, ilp (A–G)	+	+	+	+	+	
88	4549–4556/4574–4580	14 ilp	+	+	+	+	+	
89	4581–4606/4611–4636	7 ilp	+	(+)	+	+	+	
90	4637–4649/4661–4673	6 ilp (A–G)	+	+	+	+	+	
91	4693–4702/4912–4920	5 irr	+	+	+	+	+	
V11-1	4707–4713/4722–4728		–	–	X	X	X	
V11-2	4731–4772/4776–4818		–	–	–	–	X	
V11-3	4819–4830/4835–4846		–	–	X	X	X	
V11-4	4852–4863/4872–4885		–	–	X	X	X	
V11-5	4894–4897/4904–4907	1M, 2Y, sl	–	–	X	X	X	
92	4921–4925/5018–5022	1Y	X	X	X	–	(+)	
93	4930–4933/4939–4942		X	X	X	(+)	(+)	a
94	4946–4968/4972–5015	5XY irr	X	X	+	(+)	(+)	b

Symbols and abbreviations as in Table 1

[a] This stem is speculative, without supported CBCs. All eukaryotic sequences can form a similar structure

[b] The proximal portion of this stem is confirmed with 5 supporting CBCs. Alternative base pairing is possible distally

interactions with viruses or antibiotics could provide powerful selective forces and generate rapid evolutionary changes. Strong complementarity (47 of 58 bases) of V2 and V8 with the ferritin heavy-chain mRNA, which is preferentially translated in vitro, has been found (Jain et al. 1985). This may be an example of a species-specific interaction that regulates gene expression. Such interactions could

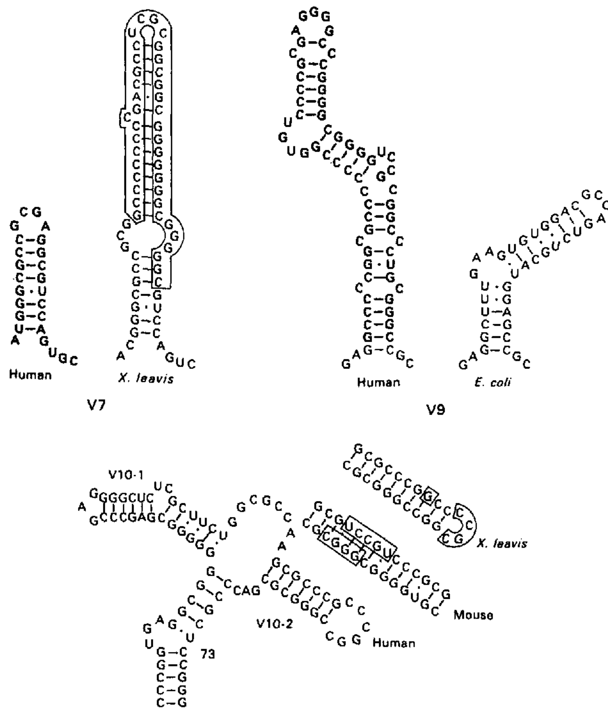

Fig. 4. Comparative analysis of proposed secondary-structural features supported by sequence-tract duplications or deletions. Portions of the human 28S rRNA molecule are shown. Shaded areas denote those sequence tracts not present (deleted) in the yeast 26S rRNA gene. Boxed nucleotides denote sequence tracts inserted relative to the human 28S rRNA gene sequence for the sequence of a given species

also provide selective pressure for the production of variability within these regions.

Higher-Order rRNA Structures

Rapid progress is being made in understanding the topology of rRNA within the ribosomal subunits (reviewed in Brimacombe and Stiege 1985). The structural constancy of the ribosomal subunits demonstrated by electron microscopy (Henderson et al. 1984) and the strong conservation of rRNA secondary-structural elements support the contention that eukaryotic and prokaryotic rRNAs share a common structural organization. The presence of the variable structures within the large rRNAs of higher eukaryotes may aid in the understanding of the topology of the rRNA in that the presence of those structures may limit what inter- and intramolecular interactions are permitted.

Acknowledgments. We are deeply grateful to Suzanne Leahy for her assistance and expertise in preparing the illustrations. We thank Wayne P. Rindone for his assistance in using the PROPHET programs and Lori Cole for preparing the manuscript. We also thank Drs. Donna George and Paula Henthorn for their enthusiasm and criticism. This work was supported in part by National Institutes of Health grant HD 16930 to R.D.S. I.L.G. and J.L.G. were supported by National Institutes of Health post-doctoral training grant GM 07511. J.L.G. is supported by National Institutes of Health Physician Scientist Award K-11-HD 00610.

References

Auron PE, Rindone WP, Vary CPH, Celentano JJ, Vournakis JN (1982) Computer-aided prediction of RNA secondary structures. Nucleic Acids Res 10:403–419

Branlant C, Krol A, Machatt A, Pouyet J, Ebel JP, Edwards K, Kossel H (1981) Primary and secondary structures of *Escherichia coli* MRE 600 23S ribosomal RNA. Comparison with models of secondary structure for maize chloroplast 23S rRNA and for large portions of mouse and human 16S mitochondrial rRNAs. Nucleic Acids Res 9:4303–4324

Brimacombe R, Stiege W (1985) Structure and function of ribosomal RNA. Biochem J 229:1–17

Brosius J, Dull TJ, Noller HF (1980) Complete nucleotide sequence of a 23S ribosomal RNA gene from *Escherichia coli*. Proc Natl Acad Sci USA 77:201–204

Chan YL, Olvera J, Wool IG (1983) The structure of rat 28S ribosomal ribonucleic acid inferred from the sequence of nucleotides in a gene. Nucleic Acids Res 11:7819–7831

Clark CG, Tague BW, Ware VC, Gerbi SA (1984) *Xenopus laevis* 28S ribosomal RNA: a secondary structure model and its evolutionary and functional implications. Nucleic Acids Res 12:6197–6220

DeBanzie JS, Steeg EW, Lis JT (1984) Update for users of the Cornell sequence analysis package. Nucleic Acids Res 12:619–625

El-Baradi TTAL, Raue HA, DeRegt VCHF, Verbree EC, Planta RJ (1985) Yeast ribosomal protein L25 binds to an evolutionary conserved site on yeast 26S and *E. coli* 23S rRNA. EMBO J 4:2101–2107

Erickson JM, Schmickel RD (1985) A molecular basis for discrete size variation in human ribosomal DNA. Am J Hum Genet 37:311–325

Erickson JM, Rushford CL, Dorney DJ, Wilson GN, Schmickel RD (1981) Structure and variation of human ribosomal DNA: molecular analysis of cloned fragments. Gene 16:1–9

Furlong JC, Maden BEH (1983) Patterns of major divergence between the internal transcribed spacers of ribosomal DNA in *Xenopus borealis* and *Xenopus laevis*, and of minimal divergence within ribosomal coding regions. EMBO J 2:443–448

Furlong JC, Forbes J, Robertson M, Maden BEH (1983) The external transcribed spacer and preceding region of *Xenopus borealis* rDNA: a comparison with the corresponding region of *Xenopus laevis* rDNA. Nucleic Acids Res 11:8183–8196

Glotz C, Zwieb C, Brimacombe R, Edwards K, Kessel H (1981) Secondary structure of the large subunit ribosomal RNA from *Escherichia coli*, *Zea mays* chloroplast, and human and mouse mitochondrial ribosomes. Nucleic Acids Res 9:3287–3306

Gonzalez IL, Gorski JL, Campen TJ, Dorney DJ, Erickson JM, Sylvester JE, Schmickel RD (1985) Variation among human 28S ribosomal RNA genes. Proc Natl Acad Sci USA 82:7666–7670

Gourse RL, Thurlow DL, Gerbi SA, Zimmermann RA (1981) Specific binding of a prokaryotic ribosomal protein to a eukaryotic ribosomal RNA: implications for evolution and autoregulation. Proc Natl Acad Sci USA 78:2722–2726

Grummt I, Roth E, Paule MR (1982) Ribosomal RNA transcription in vitro is species specific. Nature 296:173–174

Hall LMC, Maden BE (1980) Nucleotide sequence through the 18S–28S intergene region of a vertebrate ribosomal transcription unit. Nucleic Acids Res 8:5993–6005

Hassouna N, Michot B, Bachellerie JP (1984) The complete nucleotide sequence of mouse 28S rRNA gene. Implications

for the process of size increase of the large subunit rRNA in higher eukaryotes. Nucleic Acids Res 12:3563–3583

Henderson E, Oakes M, Clark MW, Lake JA, Matheson AT, Zillig W (1984) A new ribosome structure. Science 225:510–512

Jain SK, Crampton J, Gonzalez IL, Schmickel RD, Drysdale JW (1985) Complementarity between ferritin H mRNA and 28S ribosomal RNA. Biochem Biophys Res Commun 131:863–868

Jones TR, Parks CL, Spector DJ, Hyman RW (1985) Hybridization of herpes simplex virus DNA and human ribosomal DNA and RNA. Virology 144:384–397

Kan LS, Chandrasegaran S, Pulford SM, Miller PS (1983) Detection of a guanine–adenine base pair in a decadeoxyribonucleotide by proton magnetic resonance spectroscopy. Proc Natl Acad Sci USA 80:4263–4265

Maly P, Brimacombe R (1983) Refined secondary structure models for the 16S and 23S ribosomal RNA of *Escherichia coli*. Nucleic Acids Res 11:7263–7286

Michot B, Bachellerie JP, Raynal F (1983) Structure of mouse rRNA precursors. Complete sequence and potential folding of the spacer regions between 18S and 28S rRNA. Nucleic Acids Res 11:3375–3396

Michot B, Hassouna N, Bachellerie JP (1984) Secondary structure of mouse 28S rRNA and general model for the folding of the large rRNA in eukaryotes. Nucleic Acids Res 12:4259–4279

Noller HF (1980) Structure and topography of ribosomal RNA. In: Chamblis G, Craven GR, Davies J, Davis K, Kahan L, Nomura M (eds) Ribosomes: structure, function and genetics. University Park Press, Baltimore, pp 3–22

Noller HF (1984) Structure of ribosomal RNA. Annu Rev Biochem 53:119–162

Noller HF, Kop JA, Wheaton V, Brosius J, Gutell RR, Kopylov AM, Dohme F, Herr W (1981) Secondary structure model for 23S ribosomal RNA. Nucleic Acids Res 9:6167–6189

Otsuka T, Nomiyama H, Yoshida H, Kukita T, Kuhara S, Sakaki Y (1983) Complete nucleotide sequence of the 26S rRNA gene of *Physarum polycephalum*: its significance in gene evolution. Proc Natl Acad Sci USA 80:3163–3167

Peattie DA, Douthwaite S, Garrett RA, Noller HF (1981) A "bulged" double helix in a RNA–protein contact site. Proc Natl Acad Sci USA 78:7331–7335

Schibler U, Wyler T, Hagenbuchle O (1977) Changes in size and secondary structure of the ribosomal transcription unit during vertebrate evolution. J Mol Biol 94:503–517

Sellers PH (1974) Theory and computation of evolutionary distances. SIAM J Appl Math 26:787–793

Smith GP (1976) Evolution of repeated DNA sequences by unequal crossover. Science 191:528–535

Smith TF, Waterman MS (1981) Identification of common molecular sequences. J Mol Biol 147:195–197

Stiege W, Zwieb C, Brimacombe R (1982) Precise localization of three intra-RNA crosslinks in 23S RNA, and one in 5S RNA, induced by treatment of *Escherichia coli* 50S ribosomal subunits with bis-(2-chloroethyl)-methylamine. Nucleic Acids Res 10:7211–7229

Stiege W, Glotz C, Brimacombe R (1983) Localization of a series of intra-RNA crosslinks in the secondary and tertiary structure of 23S RNA, induced by ultraviolet irradiation of *Escherichia coli* 50S ribosomal subunits. Nucleic Acids Res 11:1687–1706

Subrahmanyam CS, Cassidy B, Busch H, Rothblum LI (1982) Nucleotide sequence of the region between the 18S rRNA sequence and the 28S rRNA sequence of rat ribosomal DNA. Nucleic Acids Res 10:3667–3680

Turner S, Noller HF (1983) Identification of sites of 4′-(hydroxymethyl)-4,5′,8-trimethylpsoralen cross-linking in *Escherichia coli* 23S ribosomal ribonucleic acid. Biochemistry 22: 4159–4164

Vaughn JC, Sperbeck SJ, Ramsey WJ, Lawrence CB (1984) A universal model for the secondary structure of 5.8S ribosomal RNA molecules, their contact sites with 28S ribosomal RNAs, and their prokaryotic equivalent. Nucleic Acids Res 12:7479–7502

Veldman GM, Klootwijk J, deRegt VCHF, Planta RJ, Branlant C, Krol A, Ebel JP (1981) The primary and secondary structure of yeast 26S rRNA. Nucleic Acids Res 9:6935–6952

Vester B, Garrett RA (1984) Structure of a protein L 23 RNA complex located at the A-site domain of the ribosomal peptidyl transferase center. J Mol Biol 179:431–452

Walker TA, Pace NR (1983) 5.8S ribosomal RNA. Cell 33: 320–322

Ware VC, Tague BW, Clark CG, Gourse RL, Brand RC, Gerbi SA (1983) Sequence analysis of 28S ribosomal DNA from the amphibian *Xenopus laevis*. Nucleic Acids Res 11:7795–7817

Woese CR (1980) Just so stories and Rube Goldberg machines: speculations on the origin of the protein synthetic machinery. In: Chamblis G, Craven GR, Davies J, Davis K, Kahan L, Nomura M (eds) Ribosomes: structure, function and genetics. University Park Press, Baltimore, pp 357–373

Woese CR, Gutell R, Gupta R, Noller HF (1983) Detailed analysis of the higher-order structure of 16S-like ribosomal ribonucleic acids. Microbiol Rev 47:621–669

Zaug AJ, Cech TR (1986) The intervening sequence of *Tetrahymena* is an enzyme. Science 231:470–476

Zimmermann RA (1979) Interactions among protein and RNA components of the ribosome. In: Chamblis G, Craven GR, Davies J, Davis K, Kahan L, Nomura M (eds) Ribosomes: structure, function and genetics. University Park Press, Baltimore, pp 135–169

Received November 25, 1985/Revised May 28, 1986