Volume 12 Number 8 1984 Nucleic Acids Research

The complete nucleotide sequence of mouse 28S rRNA gene. Implications for the process of size increase of the large subunit rRNA in higher eukaryotes

Nasser Hassouna, Bernard Michot and Jean-Pierre Bachellerie*

Centre de Recherche de Biochimie et de Génétique Cellulaires du CNRS, 118, route de Narbonne, 31062 Toulouse Cedex, France

Received 6 March 1984; Accepted 3 April 1984

ABSTRACT

We have determined the complete nucleotide sequence (4712 nucleotides) of the mouse 28S rRNA gene. Comparison with all other homologs indicates that the potential for major variations in size during the evolution has been restricted to a unique set of a few sites within a largely conserved secondary structure core. The D (divergent) domains, responsible for the large increase in size of the molecule from procaryotes to higher eukaryotes, represent half the mouse 28S rRNA length. They show a clear potential to form self-contained secondary structures. Their high GC content in vertebrates is correlated with the folding of very long stable stems. Their comparison with the two other vertebrates, xenopus and rat, reveals an history of repeated insertions and deletions. During the evolution of vertebrates, insertion or deletion of new sequence tracts preferentially takes place in the subareas of D domains where the more recently fixed insertions/deletions were located in the ancestor sequence. These D domains appear closely related to the transcribed spacers of rRNA precursor but a sizable fraction displays a much slower rate of sequence variation.

INTRODUCTION

A better knowledge of the eukaryotic ribosome and the processes involved in the control of its activity obviously requires detailed structural analyses of its rRNA components. The strong conservation of rRNA structure during evolution, first indicated by heterologous nucleic acid hybridizations (see (1) for review), has suggested that a common set of basic functions in all species are served by a number of homologous regions. The yeast 26S rRNA sequence (2,3) has first shown that the size differences between an eukaryotic large subunit rRNA and its prokaryotic counterpart were restricted to a few inserted domains interspersed among a set of conserved regions, as later confirmed by the *Physarum polycephalum* sequence (4). Due to the relatively fast rate of variations of these heterologous domains, little information could be gained on their potential structural organization and role in ribosome function by the sole comparison of these 2 lower eukaryotes sequences. However the present determi-

© IRL Press Limited, Oxford, England.

nation of the mouse 28S rRNA sequence, together with the very recent report of two other vertebrates sequences, *Xenopus laevis* (5) and rat (6), provides the opportunity to better analyze the process of size increase of the large rRNA during the evolution of higher eukaryotes, and its potential functional implications, through comparisons of pairs of more and more closely related species. These comparative data, extended to *E. coli* 23S rRNA (7, 8), have been analyzed in terms of potential secondary structure folding, with reference to the models previously proposed for *E. coli* (8-10) and for yeast (2). Together with the recently reported 18S rRNA (11) and 5.8S rRNA (12) sequences, the present 28S rRNA sequence now provides a complete set of the mature rRNA sequences encoded by the ribosomal transcription unit in mouse.

MATERIAL AND METHODS

Recombinant DNA :

Mouse ribosomal DNA was prepared from four recombinant plasmids constructed with two large overlapping DNA fragments (EcoRI-EcoRI : 6.7 kb, and BamHI-BamHI : 2.4 kb) which encompass the entire 28S rRNA gene and had been cloned into pBR322. Recombinant plasmid pM B2 and its subclone pMEB1 were constructed by I. Grummt (in preparation). Recombinant plasmid pMEB3, a subclone from pME6, had been previously used for sequencing the internal transcribed spacer regions of the ribosomal gene (13) and the 5'domain of 28S rRNA gene (12). Locations of these recombinants along the gene are shown in Fig. 1. Plasmid DNAs were isolated from E. coli HB101 by the clear lysate method (14) followed by CsCl-Ethidium bromide equilibrium ultracentrifugation. Supercoiled closed circular plasmid DNA was further purified by sucrose gradient ultracentrifugation.

DNA sequencing :

Restriction endonuclease analysis, purification of DNA fragments, 5'(^{32}P) end-labeling and chemical DNA sequencing were essentially carried out according to Maxam and Gilbert (15), as described previously (12).

Secondary structure analysis :

The HELCAT computer program (16) for cataloguing potentially base-paired regions was kindly provided by F. Michel. Comparative analyses of these data were performed along the lines described by Noller et al. (9).

RESULTS

1.Determination of the sequence

The sequence of mouse 28S rRNA was inferred from the sequence of the

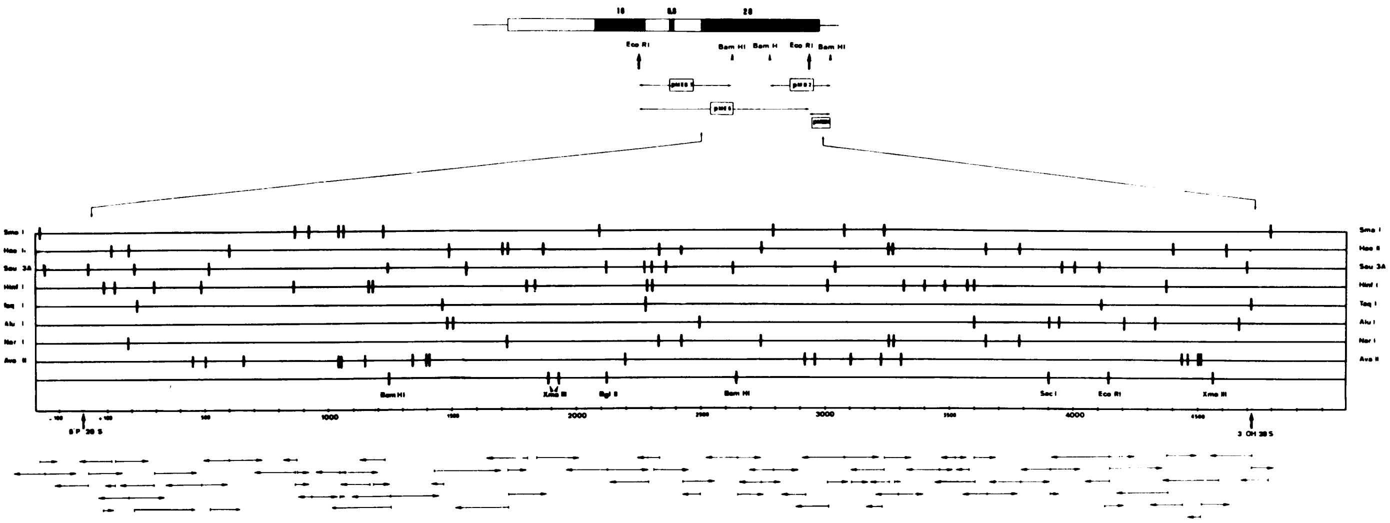


Fig. 1 : Restriction map of mouse 28S rRNA gene and sequencing strategy:

Locations of the recombinant plasmids used in this study are shown in the upper part. In the lower part the arrows are indicative of the length of sequence read, with starts of arrows corresponding to the 5'(^{32}P) labeled end. Nucleotide positions are numbered from the 5' terminus of 28S rRNA gene.

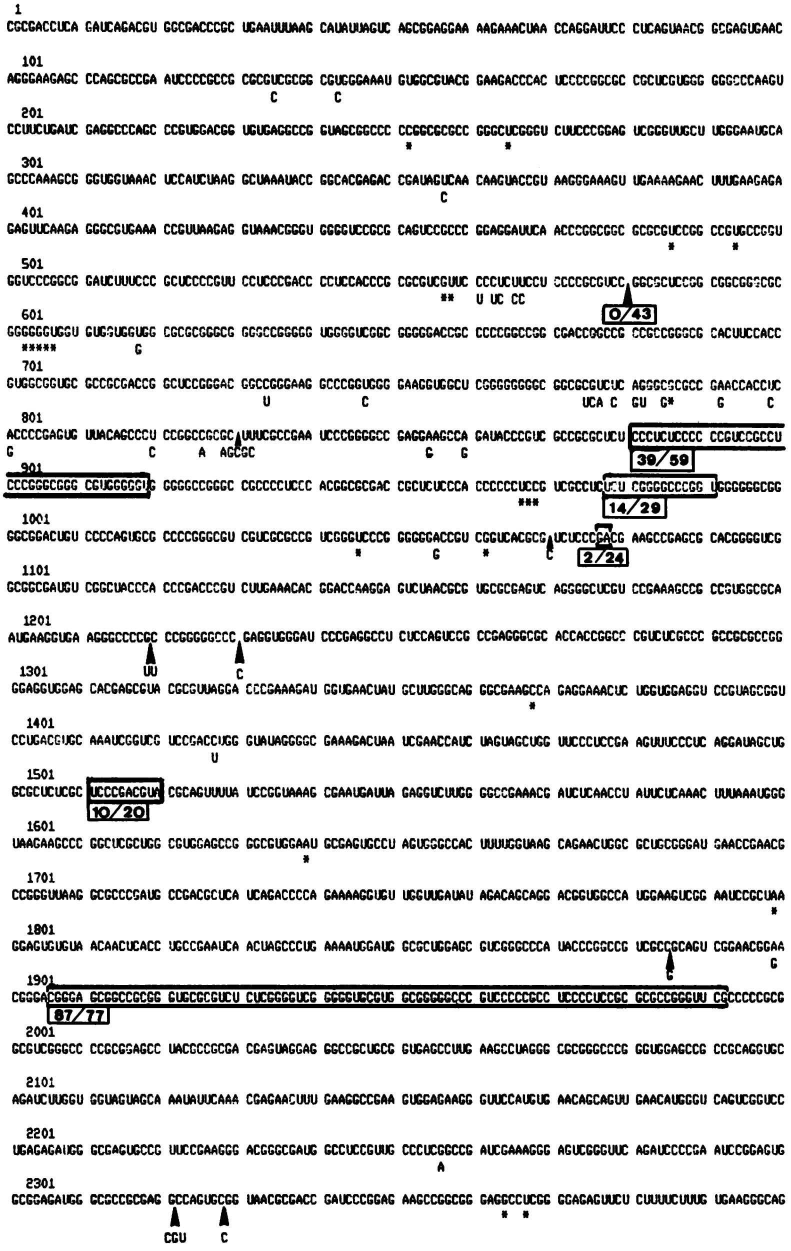


Fig. 2 : Complete primary structure of mouse 28S rRNA inferred from its gene sequence and comparison with its rat homolog.

Boxes denote sequence tracts which have extensively diverged both in sequence and size between these rodents, with the two numbers indicating

```
2401
GGCGCCCUGG AAUGGGUUCG CCCCGAGAGA GGGGCCCGUG CCUUGGAAAG CGUCGCGGUU CCGGCGGCGU CCGGUGAGCU CUCGCUGGCC CUUGAAAAUC
      U                                                       A
2501
CGGGGGAGAG GGUGUAAAUC UCGCGCCGGG CCGUACCCAU AUCCGCAGCA GGUCUCCAAG GUGAACAGCC UCUGGCAUGU UGGAACAAUG UAGGUAAGGG

2601
AAGUCGGCAA GCCGGAUCCG UAACUUCGGG AUAAGGAUUG GCUCUAAGGG CUGGGUCGGU CGGGCUGGGG CGCGAAGCGG GGCUGGGCGC GCGCCGCGGC
   *                                                                                ▲
2701                                                                                G
UGGACGAGGC GCCGCCGCCC▲UCUCCCACGU CCGGGGAGAC CCCCCGUCCU UUCCGCCCGG GCCCGCCCUC CCCUCUUCCC CGCGGGGCCC CGUCGUCCCC
                     CC                          **                                   *                 C
2801
CGCGUCGUCG CCACCUCUCU UCCCCCCUCC UUCUUCCCGU CGGGGGGCGG GUCGGGGGUC GGCGCGCGGC GCGGGCUCCG GGGCGGCGGG UCCAACCCCG
             67/81
2901
CGGGGGUUCC GGAGCGGGAG GAACCAGCGG▲UCCCCGGUGG GGCGGGGGGC CCGGACACUC GGGGGGCCGG CGGCGGCGGC GACUCUGGAC GCGAGCCGGG
  C    **            G       C    GC            **               C        ▲
3001                                                                     GA
CCCUUCCCGU GGAUCGCCUC AGCUGCGGCG GGCGUCGCGG CCGCUCCCGG GGAGCCCGGC GGGUGCCGGC GCGGGUCCCC UCCCCGCGGG GCCUCGCUCC
                   C                                                  118/93
3101
ACCCCCCCAU CGCCUCUCCC GAGGUGCGUG GCGGGGGCGG GCGGGCGUGU CCCGCGCGUG UGGGGGGAAC CUCCGCGUCG GUGUUCCCCC GCCGGGUCCG
                                                                                                          G
3201
CCCCCCGGGC CGCGGUUUUC CGCGCGGCGC CCCCGCCUCG GCCGGCGCCU AGCAGCCGAC UUAGAACUGG UGCGGACCAG GGGAAUCCGA CUGUUUAAUU
           *       C      * * *  G                                                 U
3301
AAAACAAAGC AUCGCGAAGG CCCGCGGCGG GUGUUGACGC GAUGUGAUUU CUGCCCAGUG CUCUGAAUGU CAAAGUGAAG AAAUUCAAUG AAGCGCGGGU

3401
AAACGGCGGG AGUAACUAUG ACUCUCUUAA GGUAGCCAAA UGCCUCGUCA UCUAAUUAGU GACGCGCAUG AAUGGAUGAA CGAGAUUCCC ACUGUCCCUA

3501
CCUACUAUCC AGCGAAACCA CAGCCAAGGG AACGGGCUUG GCGGAAUCAG CGGGGAAAGA AGACCCUGUU GAGCUUGACU CUAGUCUGGC ACGGUGAAGA

3601
GACAUGAGAG GUGUAGAAUA AGUGGGAGGC CCCCGGCGCC CGGCCCCGUC CUCGCGUCGG GGUCGGGGCA CGCCGGCCUC GCGGGCCGCC GGUGAAAUAC
                                              **     ▲   C    ▲       *      UC
3701                                                 U    AGGGG
CACUACUCUC AUCGUUUUUU CACUGACCCG GUGAGGCGGG GGGGCGAGCC CCGAGGGGCU CUCGCUUCUG GCGCCAAGCG UCCGUCCCGC GCGUGCGGGC
                                                                                  G A  **   *        C  *
3801
GGGCGCGACC CGCUCCGGGC ACAGUGCCAG GUGGGGAGUU UGACUGGGGC GGUACACCUG UCAAACGGUA ACGCAGGUGU CCUAAGGCGA GCUCAGGGAG

3901
GACAGAAACC UCCCGUGGAG CAGAAGGGCA AAAGCUCGCU UGAUCUUGAU UUUCAGUACG AAUACAGACC GUGAAAGCGG GGCCUCACGA UCCUUCUGAC

4001
CUUUUGGGUU UUAAGCAGGA GGUGUCAGAA AAGUUACCAC AGGGAUAACU GGCUUGUGGC GGCCAAGCGU UCAUAGCGAC GUCGCUUUUU GAUCCUUCGA

4101
UGUCGGCUCU UCCUAUCAUU GUGAAGCAGA AUUCACCAAG CGUUGGAUUG UUCACCCACU AAUAGGGAAC GUGAGCUGGG UUUAGACCGU CGUGAGACAG

4201
GUUAGUUUUA CCCUACUGAU GAUGUGUUGU UGCCAUGGUA AUCCUGCUCA GUACGAGAGG AACCGCAGGU UCAGACAUUU GGUGUAUGUG CUUGGCUGAG

4301
GAGCCAAUGG GGCCAAGCUA CCAUCUGUGG GAUUAUGACU GAACGCCUCU AAGUCAGAAU CCGCCCAAGC GGAACGAUAC GGCAGCGCCG AAGGAGCCUC

4401
GGUUGGCCCC GGAUAGCCGG GUCCCCGUCC GUCCCGCUCG GCGGGGUCCC CGCGUCGCCC CGCGGCGGCG CGGGGUCUCC CCCCGCCGGG CGUCGGGACC
                      C            ▲    UC     *           C   U ▲   C CC 64/0
4501                               C                               G
GGGGUCCGGU GCGGAGAGCC GUUCGUCUUG GGAAACGGGG UGCGGCCGGA AAGGGGGCCG CCCUCUCGCC CGUCACGUUG AACGCACGUU CGUGUGGAAC
                                                                                    ▲ *
4601                                                                                C
CUGGCGCUAA ACCAUUCGUA GACGACCUGC UUCUGGGUCG GGGUUUCGUA CGUAGCAGAG CAGCUCCCUC GCUGCGAUCU AUUGAAAGUC AGCCCUCGAC
U
4701
ACAAGGGUUU GU
```

their length in mouse and rat respectively. Outside the boxed regions, all the point differences in rat as compared to mouse are shown under the mouse sequence. Deletions in rat are denoted by a star and additions by an arrowhead.

cloned gene which appeared identical to the chromosomal genes when detailed restriction maps were determined by Southern blot hybridizations (not shown). The sequence strategy (Fig. 1) involved extensive overlaps (among others at EcoRI site, position 4128). For the few sites which were not overlapped, the absence of any short intervening oligonucleotides was directly checked through partial restriction analysis of short overlapping 5'end labelled fragments. The sequence determination on both strands was performed for about 80 % of the gene and was systematical whenever any peculiarity was found on one strand (like "silent" methylated nucleotide or band compressions due to secondary structure effects). As a result no ambiguity remains over the 4712 nucleotides of the complete sequence (Fig. 2). Partial sequence data had been reported previously by our group for the 5'terminal 585 nucleotides (12) and by others for the 3'terminal 170 nucleotides (17). In this 3'terminal segment, our present determination agrees well with those data, except for 3 changes (presence of a GC, positions 4583-4584 - presence of a A, position 4658).

2.Comparison of mouse 28S rRNA sequence with other homologs.

The mouse sequence has been aligned with all its available eukaryotic homologs, and with E. coli. When mouse, xenopus (5), yeast (2, 3) and physarum (4) sequences are compared all together, it is remarkable that unambiguous alignments common to the four species can be detected over a large fraction of 28S rRNA length (40 % for mouse) as shown in Fig. 3, despite the large size differences among these eukaryotic sequences (+ 39 % in mouse as compared to yeast). While very long tracks of the large rRNA molecule have been strongly conserved during evolution, the additional sequences found in higher eukaryotes are clearly clustered in a few definite areas instead of being scattered along the entire molecule. The number and the relative location of these highly divergent areas (identified as D1 to D12 and represented between brackets in Fig. 3) do not seem to depend upon the species that are considered, at least when the phylogenetic distance is high enough. Whereas only a subset of these 12 potentially variable areas may differ in size between two closely related species (such as mouse and rat, as described below), interruptions in the alignments accompanied by size variations do occur over each of these areas in the comparisons by pair between mouse, xenopus, yeast and physarum, whatever the pair of species that is considered. A similar conclusion emerges from the comparison of the four eukaryotic sequences with E. coli(7, 8). Although tracts of sequence homology (underlined by thick bars in Fig. 3) are much

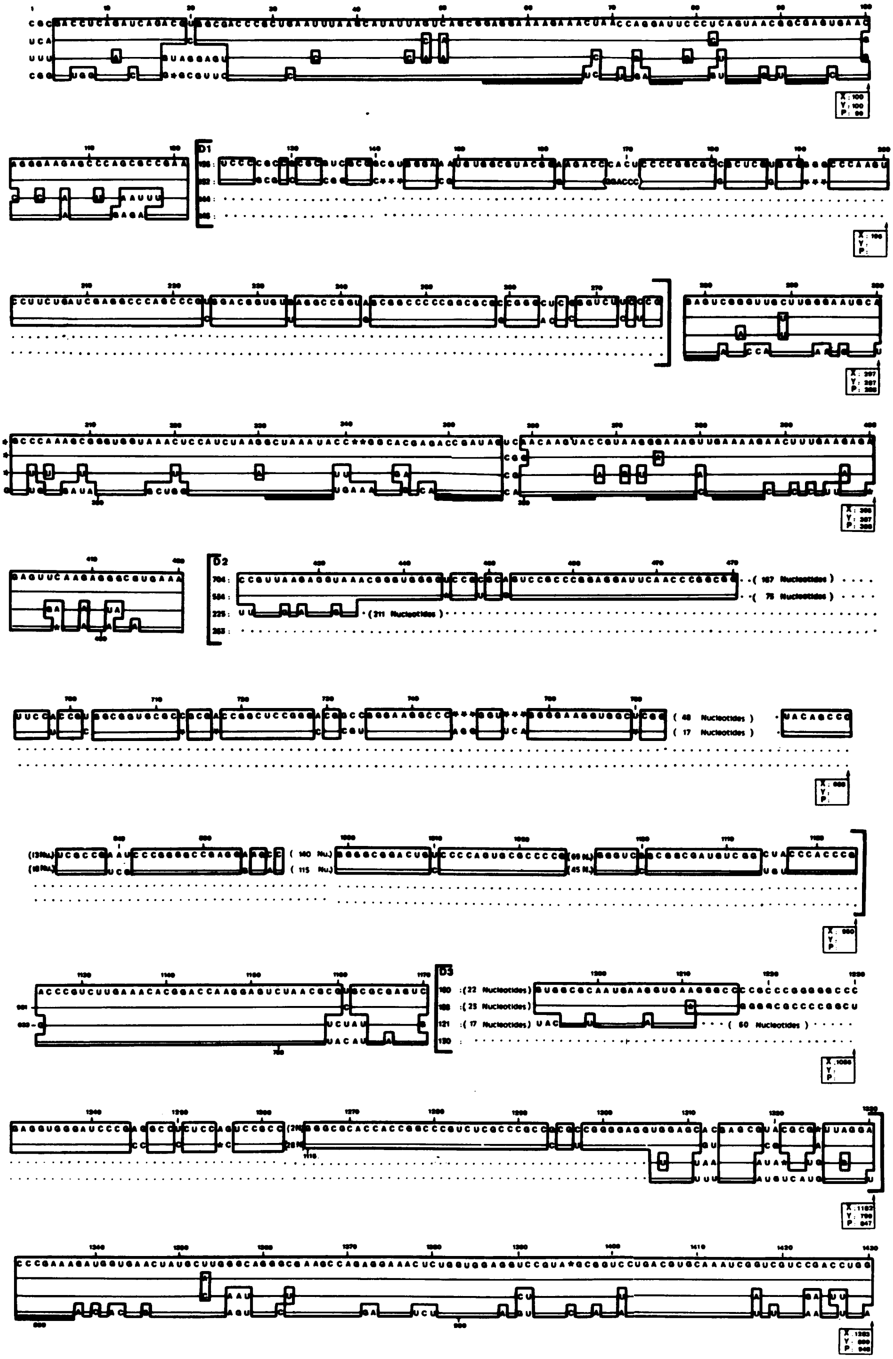
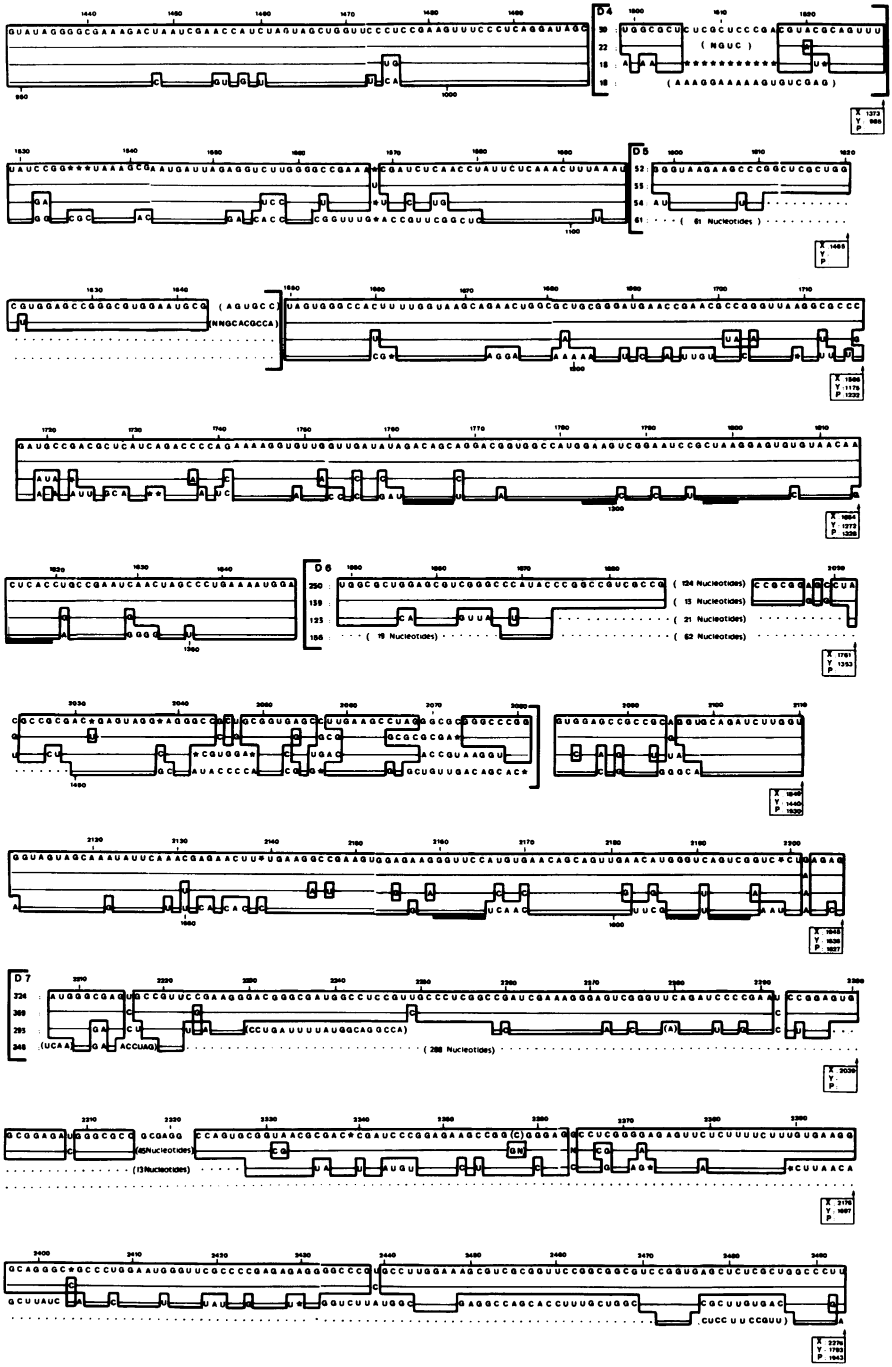
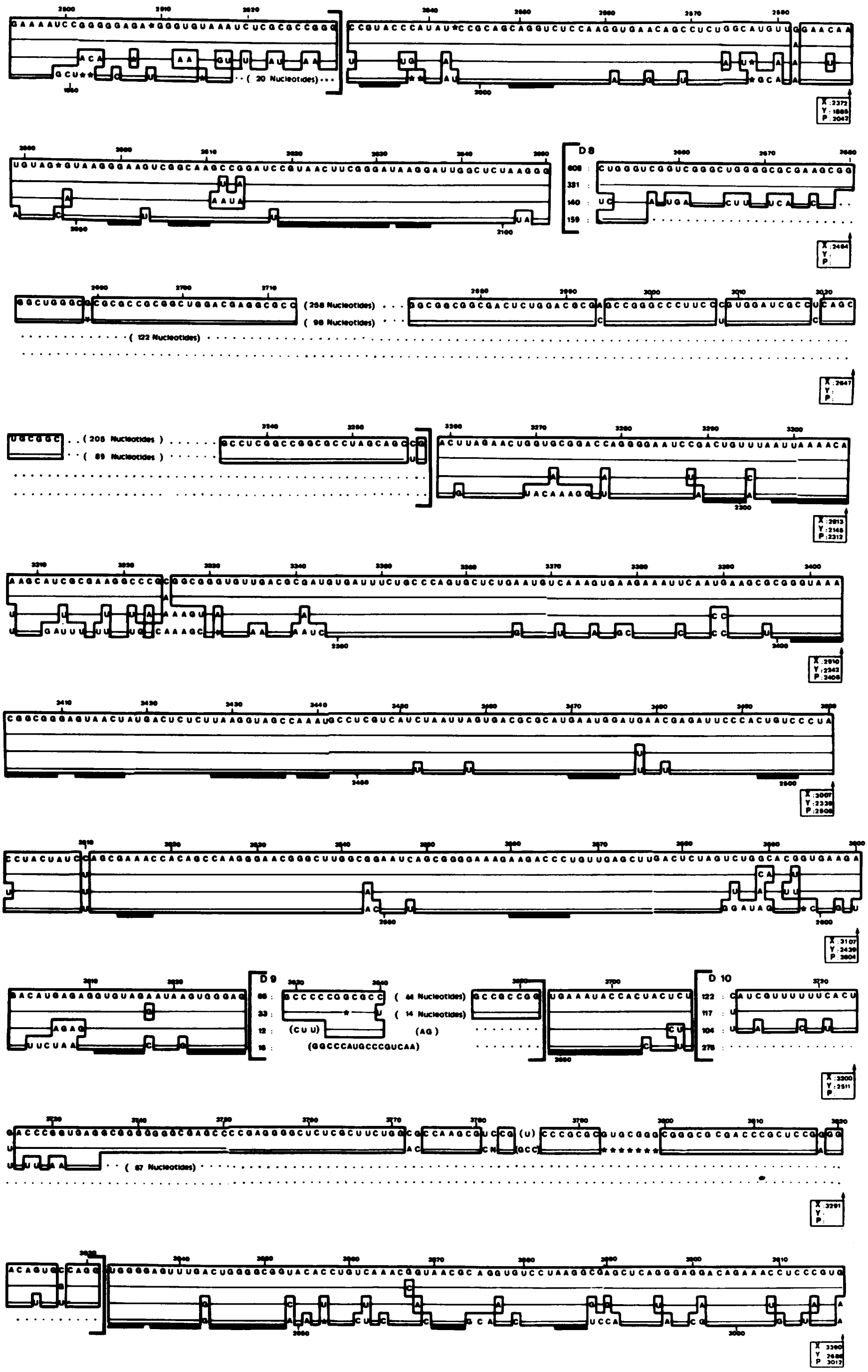


Figure 3.



GUAUAGGGGCGAAAGACUAAUCGAACCAUCUAGUAGCUGGUUCCCUCCGAAGUUUCCCUCAGGAUAGC
D4
30 : UGGCGCU CUCGCUCCCGA CGUACGCAGUUU
22 : (NGUC)
18 : A A A ***********U*
18 : (AAAGGAAAAAGUGUCGAG)
X 1373
Y 966
P
UAUCCGG**UAAAGCGAAUGAUUAGAGGUCUUGGGGCCGAAA*CGAUCUCAACCUAUUCUCAACUUUAAAU
D5
52 : GGGUAAGAAGCCCGGCUCGCUGG
55 :
54 : AU
61 : (61 Nucleotides)
X 1465
Y
P
CGUGGAGCCGGGCGUGGAAUGCG (AGUGCC) (NNGCACGCCA)
UAGUGGGCCACUUUUGGUAAGCAGAACUGGCGCUGCGGGAUGAACCGAACGCCGGGUUAAGGCGCCC
CG* AGGA AAAA UGU C *
X 1566
Y 1175
P 1232
GAUGCCGACGCUCAUCAGACCCCAGAAAAGGUGUUGGUUGAUAUAGACAGCAGGACGGUGGCCAUGGAAGUCGGAAUCCGCUAAGGAGUGUGUAACAA
AUA AUU GCA ** U C GAU
X 1664
Y 1272
P 1338
CUCACCUGCCGAAUCAACUAGCCCUGAAAAUGGA
GGG
D6
250 : UGGCGCUGGAGCGUCGGGCCCAUACCCGGCCGUCGCCG (124 Nucleotides) CCGCG CUA
139 : (13 Nucleotides)
123 : CA GUUA (21 Nucleotides)
186 : (19 Nucleotides) (62 Nucleotides)
X 1761
Y 1353
P
GCCGCGAC*GAGUAGG*AGGGC GCGGUGAGC UUGAAGCCUAG GCGCGGCCCGG
GCGCGCGA*
CU *CGUGGA* UGAC ACCGUAAGGU
GC AUACCCCA CG * GCUGUUGACAGCAC*
GUGGAGCCGCCG GGUGCAGAUCUUGGU
UA
GGGCA
X 1849
Y 1440
P 1530
GGUAGUAGCAAAUAUUCAAACGAGAACUU*UGAAGGCCGAAGUGGAGAAGGGUUCCAUGUGAACAGCAGUUGAACAUGGGUCAGUCGGUC*CUGAGAG
A G U CACA UCAA UUCG AAU
X 1948
Y 1636
P 1627
D7
324 : AUGGGCGAG UGCCGUUCCGAAGGGACGGGCGAUGGCCUCCGUUGCCCUCGGCCGAUCGAAAGGGAGUCGGGUUCAGAUCCCCGAAU CCGGAGUG
369 :
293 : GA C U A (CCUGAUUUUAUGGCAGGCCA) (A) C
348 : (UCAA) G A ACCUAG (288 Nucleotides)
X 2039
Y
P
GCGGAGA UGGCGCG GCGAGG CCAGUGCGGUAACGCGAC*CGAUCCCGGAGAAGCCGG(C)GGGAG GCCUCGGGGAGAGUUCUCUUUUCUUUGUGAAGG
(45Nucleotides) CG GN N CG A
(13Nucleotides) A AUGU C C AG* A *CUUAACA
X 2176
Y 1697
P
GCAGGGC*GCCCUGGAAUGGGUUCGCCCCGAGAGAGGGCCCG GCCUUGGAAAGCGUCGCGGUUCCGGCGGCGUCCGGUGAGCUCUCGCUGGCCCUU
GCUUAUC A G U UAU * GGUCUUAUGGC GAGGCCAGCACCUUUGCUGGC GCUUGUGAC
.CUCC U U CCGUU) A
X 2276
Y 1793
P 1943


GAAAAUCCGGGGGAGA*GGGUGUAAAUCUCGCGCCGGG
(20 Nucleotides)
CCGUACCCAUAU*CCGCAGCAGGUCUCCAAGGUGAACAGCCUCUGGCAUGUU
GGAACAA
X: 2372 Y: 1885 P: 2042
UGUAG*GUAAGGGAAGUCGGCAAGCCGGAUCCGUAACUUCGGGAUAAGGAUUGGCUCUAAGGG
D 8
608 : 331 : 140 : 159 :
CUGGGUCGGUCGGGCUGGGGCGCGAAGCGG
X: 2484 Y: P:
GGCUGGG
CGCGCCGCGGCUGGACGAGGCGCC
(258 Nucleotides)
(98 Nucleotides)
(122 Nucleotides)
GGCGGCGGCGACUCUGGACGCG
GCCGGGCCCUUCC
UGGAUCGCG
CAGC
X: 2647 Y: P:
UGCGGC
(205 Nucleotides)
(89 Nucleotides)
GCCUCGGCCGGCGCCUAGCAGG
ACUUAGAACUGGUGCGGACCAGGGGAAUCCGACUGUUUAAUUAAAACA
X: 2813 Y: 2148 P: 2312
AAGCAUCGCGAAGGCCCG
GGCGGGUGUUGACGCGAUGUGAUUUCUGCCCAGUGCUCUGAAUGUCAAAGUGAAGAAAUUCAAUGAAGCGCGGGUAAA
X: 2910 Y: 2242 P: 2408
CGGCGGGAGUAACUAUGACUCUCUUAAGGUAGCCAAAUGCCUCGUCAUCUAAUUAGUGACGCGCAUGAAUGGAUGAACGAGAUUCCCACUGUCCCUA
X: 3007 Y: 2338 P: 2505
CCUACUAUC
AGCGAAACCACAGCCAAGGGAACGGGCUUGGCGGAAUCAGCGGGGAAAGAAGACCCUGUUGAGCUUGACUCUAGUCUGGCACGGUGAAGA
X: 3107 Y: 2439 P: 2604
GACAUGAGAGGUGUAGAAUAAGUGGGAG
D 9
66 : 33 : 12 : 16 :
GCCCCCGGCGCC
(44 Nucleotides)
(14 Nucleotides)
(CUU)
(AG)
(GGCCCAUGCCCGUCAA)
GCCGCCG
UGAAAUACCACUACUCU
D 10
122 : 117 : 104 : 276 :
AUCGUUUUUUCACU
X: 3200 Y: 2511 P:
ACCCGGUGAGGCGGGGGGGCGAGCCCCGAGGGGCUCUCGCUUCUGG
CAAGCG
CG
(U)
CCCGCG
UGCGGG
CGGCGCGACCCGCUCCG
(67 Nucleotides)
X: 3291 Y: P:
ACAGU
CAGG
UGGGGAGUUUGACUGGGGCGGUACACCUGUCAAACGGUAACGCAGGUGUCCUAAGGCGAGCUCAGGGAGGACAGAAACCUCCCGUG
X: 3390 Y: 2686 P: 3012

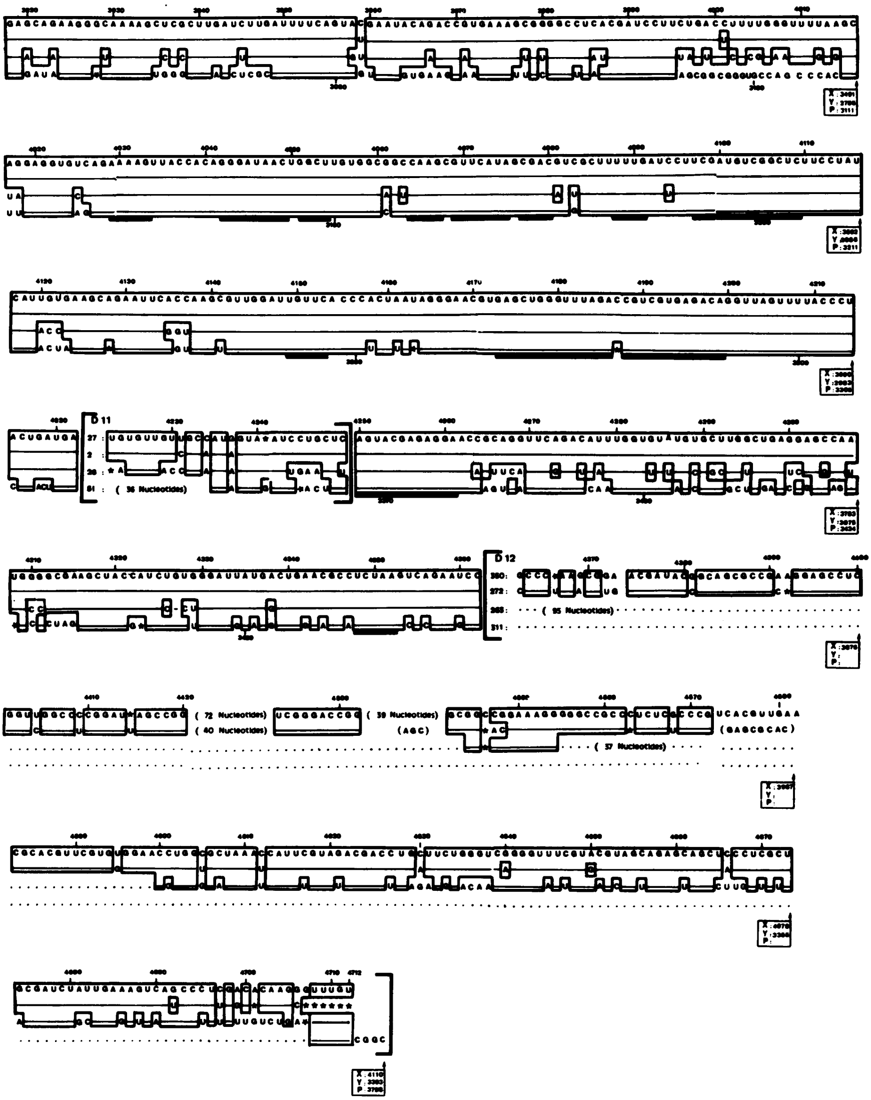


Fig. 3 : Comparison of mouse 28S rRNA sequence-top line- with the other eukaryotic homologs : amphibian Xenopus laevis (5) - 2nd line-yeast, Saccharomyces carlsbergensis (4) -3rd line- and slime mold Physarum polycephalum (4) - bottom line.

Whenever the four sequences can be unambiguously aligned, the conserved nucleotides are boxed (horizontal lines indicate identity with the mouse sequence). Sequence tracts (> 4 nucleotides) which are common to these eukaryotes and to E. coli (7, 8) are denoted by a thick bar under the boxes. Whenever the alignment between the four sequences is not possible due to extensive divergence plus size differences, the sequence is shown between large square brackets. For these 12 less conserved areas (denoted D1 to D12 from their relative location from 5' end) which are responsible for the large size variations of eukaryotic large rRNAs, the respective size in each species is indicated by a number on the left-hand side. Within these areas, significant homologies restricted to yeast and both vertebrates or to both vertebrates only are also indicated by boxes while tracts where no residual homology can be detected between any pair of species are usually denoted by a dotted line with the sole indication of their nucleotide number.

shorter in that case and could be poorly significant on the sole basis of sequence comparison, the compared analysis of secondary structure models (8-10 and our accompanying paper), in which they map at identical positions, definitely establishes they are remnants of the common ancestor sequence, thus allowing unambiguous alignments to be made. Such alignments with E. coli 23S rRNA are again interrupted over 12 locations by divergent tracts the length of which has varied between E. coli and these eukaryotes. It is important to note that these variable regions have precisely the same relative location along the molecule as revealed by the sole comparison of eukaryotic sequences. It therefore appears clearly that the potential for expansion or reduction in size of the large rRNA during evolution is restricted to a unique set of a few sites within a largely conserved structural core.

3. Common structural core and domains of variable size :

We have constructed a secondary structure model for mouse 28S rRNA (see accompanying paper) with reference to the folding patterns previously described for E. coli (8-10) and yeast (2) and to the folding potentials of the other eukaryotic sequences aligned as in Fig. 3. The boundaries of the areas where size variations have taken place between pro-and eukaryotes can be appreciated with a much better accuracy when comparisons of secondary structure models are taken into account than by the sole sequence alignment : within the areas of interrupted sequence alignments a number of conserved secondary structure features can nonetheless be identified in all species which improves accordingly the mapping of the size-variable segments. Results of this refined mapping are summarized in Table 1. It is remarkable that outside these size-variable areas, the four eukaryotes and E. coli share an almost identical secondary structure, the validity of which is supported by a number of compensatory changes distributed over the majority of the proposed duplexes (see accompanying paper). This common structure core represents 85 % of the length of E. coli 23S rRNA.

The location of these size-variable areas (see Table 1 for coordinates) within the conserved secondary structure core is depicted in Fig. 4 using a representation of the E. coli 23S rRNA folding model (9). It is

Table 1 : Sites of major size variations in large rRNA during evolution.

Location			Size of the equivalent tract in					
Identification of the divergent domain in eukaryotes	Boundaries in Mouse	E. coli	E. coli	Physarum	Yeast	Xenopus	Mouse	Rat
D1	122-277	264-374	111	186	144	152	156	154
D2	436-1124	425-577	53	246	216	499	689	776
D3	1166-1315	602-655	54	119	111	175	150	153
D4	1507-1525	845-849	5	9	7	12	19	29
D5	1606-1635	927-932	6	52	34	30	30	30
D6	1879-2032	1164-1185	22	63	27	44	154	145
D7 a	2207-2265	1359-1377	19	80	49	59	59	59
b	2302-2342	1416-1419	4	30	22	83	41	41
D8	2648-3259	1713-1745	33	155	153	334	611	594
D9	3629-3686	2127-2161	35	12	8	29	60	63
D10	3727-3819	2200-2223	24	260	75	83	93	89
D11	4221-4225	2626-2629	4	27	2	5	5	5
D12	4379-4619	2789-2812	24	215	154	170	241	179
	Total size		394	1454	1002	1675	2308	2317
	(Fraction of rRNA length)		13,5%	38,4%	30,4%	40,7%	48,9%	49,1%

noteworthy that none of them has been proposed to be involved in base-paired interactions with either adjacent regions of the conserved core or any distal segment in E. coli (8-10). Their constituting independent domains for secondary structure folding is also indicated by examination of all the eukaryotic sequences, as shown below. Moreover, the mouse sequence data confirm major trends in the evolution of these areas, which were previously apparent from the examination of xenopus (5) and rat (6) sequences, i.e. a large size increase from lower to higher eukaryotes with a very low content in A (about 5 %) and a very high GC content in vertebrates (80-85 %, with for most areas a roughly similar content in G and C). It must be stressed that very similar trends are also apparent for the internal transcribed spacers of the ribosomal gene during the evolution of higher eukaryotes when comparing yeast (20-22), xenopus (18), rat (19) and mouse (13). As summarized in Table 1, expansion of 28S rRNA in higher eukaryotes is most dramatic in two domains, termed D2 and D8 (total size in mouse 1301, as compared to 369 in yeast and in 86 E. coli). This is also apparent in the comprehensive representation of the local expansions within 28S rRNA during the evolution of eukaryotes (Fig. 5).

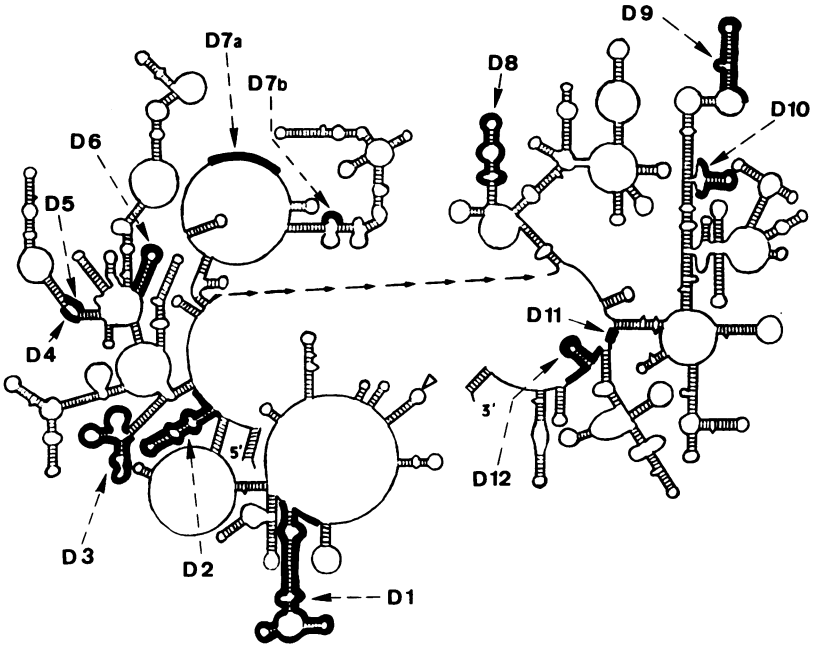


Fig. 4 : Sites of major size variation during evolution mapped within the E. coli 23S rRNA secondary structure.

The 23S rRNA secondary structure is represented as proposed by Noller et al. (9). The segments of variable size during evolution are depicted by thick lines (coordinates of their boundaries are shown in Table 1). The denomination of each corresponding divergent area in the eukaryotic alignments shown in Fig. 3 is also indicated. The linkage between the 5' and the 3' halves of the molecule (displayed on the left and the right half of the page respectively) is denoted by a string of arrows. The location of the eukaryote-specific interruption of the large rRNA (between the 3' end of 5.8S rRNA and the 5' end of 28S rRNA) is shown by an open triangle. Areas of potentially variable size within eukaryotic 5.8S rRNA or its prokaryotic equivalent have not been considered here.

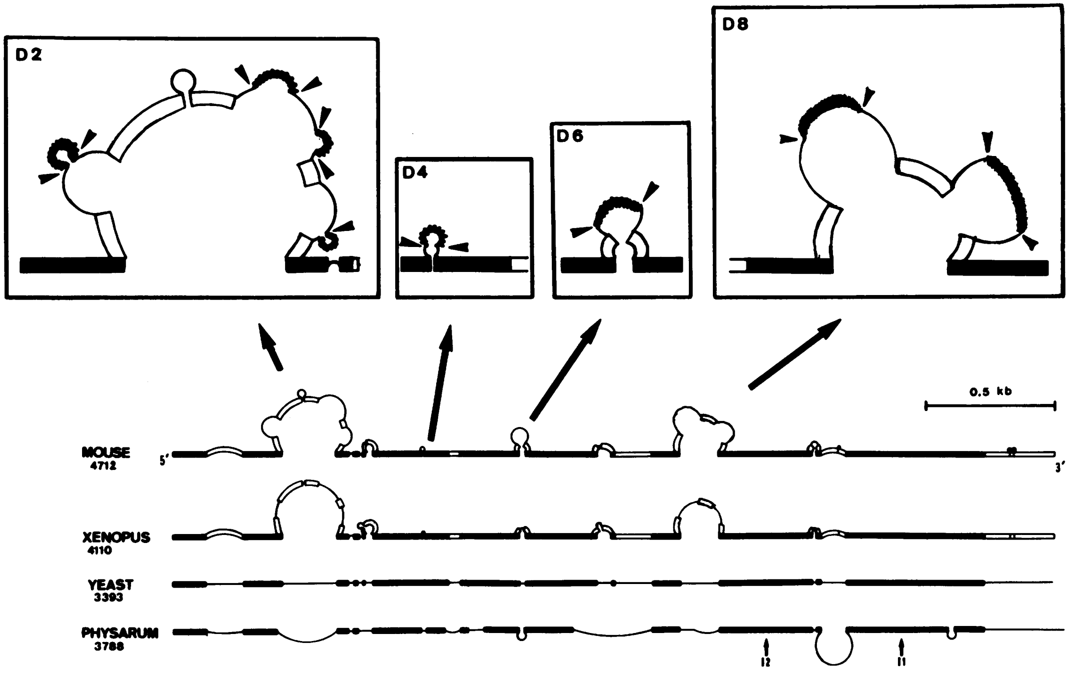


Fig. 5 : Hot-spots for the enlargement of 28S rRNA from lower to higher eukaryotes.

All these molecules have been aligned by reference to yeast 26S rRNA, the shortest eukaryotic representative which is accordingly shown as an horizontal line in its entirety. Black boxes denote regions conserved between yeast and the vertebrates (more than 80 % homology) - with most of them also present in Physarum (locations of Physarum introns are depicted by arrows). Regions where size variations are located are represented by circular lines, with their lengths proportional to their size. Open boxes denote regions which have largely diverged between yeast and vertebrates but which are highly homologous among vertebrates; they have been used for mapping areas of major enlargement between xenopus and mouse (denoted by secondary "bubbles"). Insets denote, for the rat molecule, the location of the tracts which are highly divergent and differ in size between mouse and rat : they are depicted as thick wavy lines (boxed sequence tracts in Fig. 2).

4. The process of size increase in higher eukaryotes

New information on this problem can be gained by comparing a pair of moderately distant species (mouse/xenopus) and a pair of closely related species (mouse/rat), due to the presence among the vertebrates, of a number of conserved tracts, within these globally rapidly evolving areas. As schematized in Fig. 5, the size increase among vertebrate 28S rRNAs is not uniforly distributed over the entire length of each of the size-variable "D" domains : it is instead circumscribed over a few subareas. It is remarkable that the newly fixed insertions/deletions (identified by the mouse-/rat comparison) are all precisely located within the sequence tracts which had been modified the more recently during the vertebrate evolution (identified by the mouse/xenopus comparison).

a) Mouse/Xenopus : within the 12 divergent "D" domains, conserved tracts between xenopus and mouse (> 10 nucleotides with at least 70 % homology) amount to 1353 nucleotides (corresponding global homology : 92.8 %). Over D2 domain, length differences between mouse and xenopus can be unambiguously ascribed to 4 small subareas, which are depicted as "secondary" bubbles in Fig. 5. Similarly two such subareas can be identified within D8 domain. A refined mapping can also be carried out for the other D areas (as schematized in Fig. 5).

b) Mouse/rat : Although the sequence conservation between the two rodents is very high (see Fig. 2), it is drastically interrupted (over a few discrete areas. Nine segments can be detected (boxed tracts in Fig. 2) which have largely varied in sequence and size between both rodents. It is remarkable that all these variable segments, which amount to 401 nucleotides in mouse, can be precisely mapped within the same subareas of the "D" domains (defined as in Fig. 3) where length differences can be detected between mouse and xenopus, as depicted in Fig. 5 (insets). Six out of these nine segments are located within D2 (four) and D8 (two) domains thus confirming these two areas as the major potential sites for size expansion in higher eukaryotes.

There is not a unique trend for the size variation of these nine sequence tracts between both rodents (some are larger in mouse, others are larger in rat) and the total size of the molecule is nearly identical in both species (4712 vs. 4718). These tracts have about the same markedly unbalanced base content as the entire "D" domains of the 3 vertebrates (very low in A, about 80 % in G + C) with roughly similar numbers of G and C within each segment).

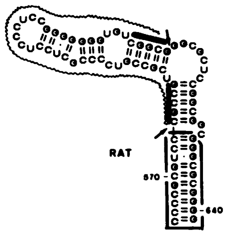


Fig. 6 : Secondary structure in rat and mouse 28S rRNA in the vicinity of the rat-specific insert.

The 43 nucleotide long insert in rat is denoted by a wavy line. The corresponding site in mouse is shown by 2 arrows. Within the rat insert, the distal regions (overlined by a thick bar) represent an inverted repeat. Except for the insert both sequences are identical in this area. The helical stem common to both rodents is boxed.

c) Insertions/Deletions : A 64 nucleotide long tract in mouse (positions 4466-4529) seems to correspond to an exact insert in the rat sequence. However it is not clear from the rat paper (6) if this location, which corresponds exactly to an AvaI site, has been overlapped in the sequence determination. On the other hand, a 43 nucleotide long segment in rat constitutes a perfect insert into the mouse sequence (positions 580-581). The absence of this segment in mouse (definitely established by sequence overlaps) corresponds to the amputation of the tip of a very long helical stem (only partially displayed in Fig. 6) involving more than 200 nucleotides (acc. paper). The inverted repeat at both ends of this rat insert could obviously have direct implications on the mechanism generating this insertion (or deletion). Insertions identified in Zea mays chloroplast 23S rRNA

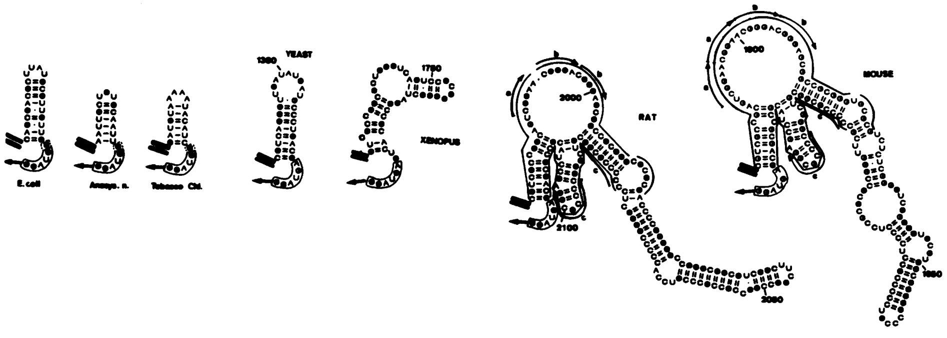


Fig. 7 : Secondary structure of the size-variable "D6" area during evolution.

Boundaries of this domain are precisely defined by a duplex conserved in all pro and eukaryotes on the 5' side (denoted by 2 thick bars) and by an invariant oligonucleotide (boxed) in equivalent location on the 3' side. For mouse and rat, arrows delineate 3 pairs of directs repeats, denoted "a", "b" and "c", present in both species (however one copy of "a" is missing in rat). Overlined sequences are identical in both rodents. Anacystis nidulans and tobacco chloroplast 23S rRNA sequences are taken from (23).

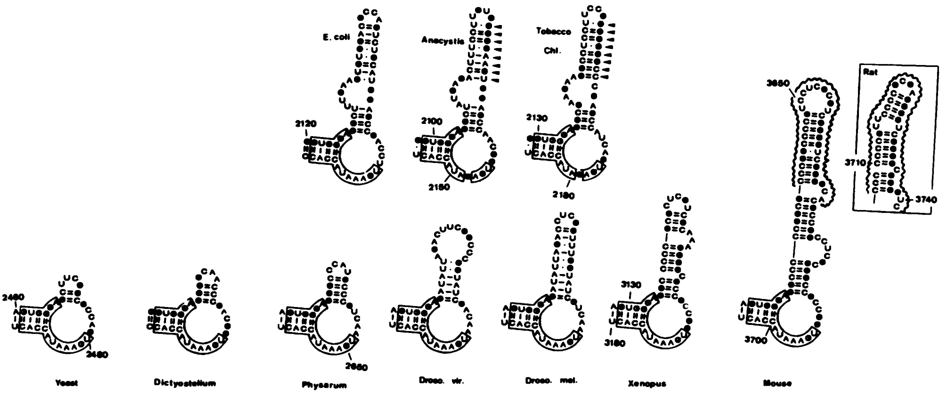


Fig. 8 : Folding of the size-variable "D9" area during evolution.
Boundaries of this domain are defined by the boxed structures common to all species (with compensatory base changes in the distal part of the stem. For rat, differences with mouse are restricted to the terminal part of the variable duplex (wavy line) which is represented in an inset. Partial sequence date available for Dictyostelium discoïdeum (24) and for Drosophila melanogaster (25) and Drosophila virilis (26) in this area have also been taken into account. Folding of the homologous domain in prokaryotic (or prokaryote-related) sequences is also shown. The secondary structure proposed by Branlant et al. (8) for E. coli is perfectly confirmed by a series of compensatory base-changes (denoted by arrow-heads) in Anacystis nidulans and tobacco chloroplast (23).

(27) as compared to E. coli have been previously shown to contain terminal inverted repeats.

d) Size increase and secondary structure folding : Correlated with the markedly unbalanced base content of these regions, the frequent occurrence of inverted or direct repeats (see Fig. 7) may be operative in maintaining their high potential for variation among higher eukaryotes, particularly through DNA strand slippages during replication (28). More generally, the reformation of exceptionnally stable giant intra-DNA strand helices, which could easily occur within the replication fork for most of the variable areas of the 28S rRNA gene, can provide a basis for their continued sequence instability. A systematical examination of the folding potential of all the eukaryotic "D" domains confirms that the areas of divergence between rat and mouse are preferentially located within the terminal (loop-proximal) part of long helical stems. This is shown in Fig. 7 and 8 for two domains of moderate length for which unequivocal folding patterns are more easily derived. A most telling example of a giant helix is shown for "D8" domain (fig. 9), which has been dramatically expanded in higher eukaryotes

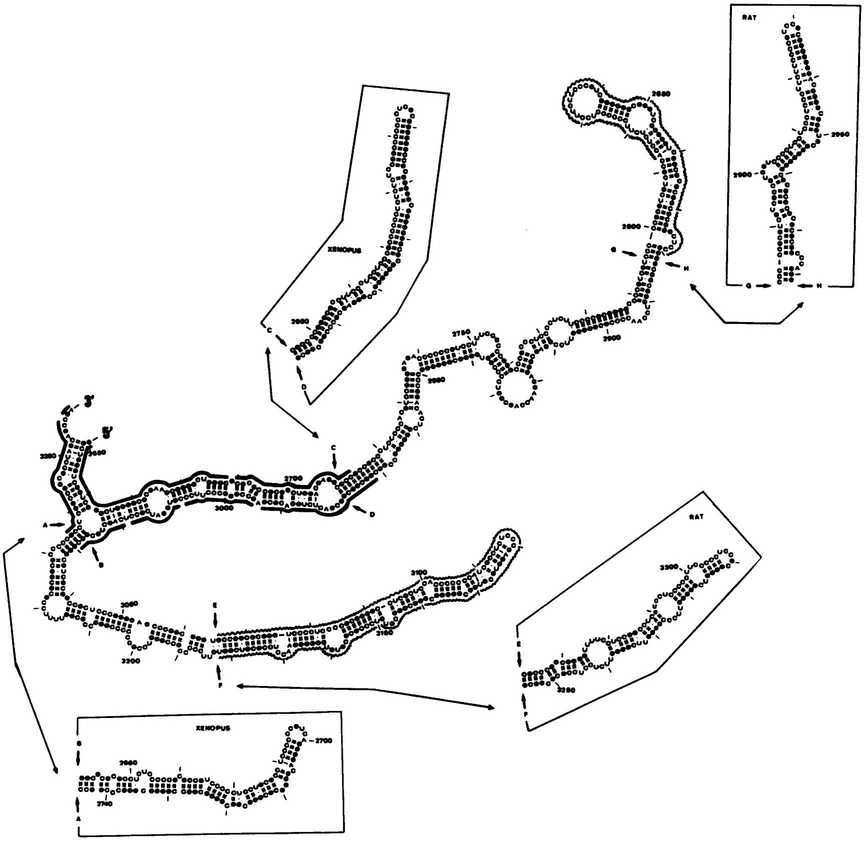


Fig. 9 : Size expansion and secondary structure of the "D8" domain in vertebrates.

The mouse sequence is folded in its entirety , with the areas of extensive sequence conservation between mouse (or rat) and xenopus denoted by a thick overline. The folding of subareas of xenopus D8 domain which are highly divergent from mouse is represented in insets (lettered arrows delineate the junction with the structure common to mouse). The wavy lines denote areas where sequence and size differences between mouse and rat are restricted, with the corresponding region in rat shown in insets. Regions of the rat 28S rRNA which are not represented can be folded like the mouse sequence.

(see Fig. 5). While the folding of such a long domain (about 0.6 kb) would appear difficult to predict on the sole basis of primary sequence, this task is facilitated by the unbalanced base content and the presence of simple sequence tracts. We have derived a Y-shaped structure, with a short 13 bp stalk and two very long arms of unequal lengths (the larger one, on the 5' side, including about 360 nucleotides). Such a folding pattern is not only highly preferred on a thermodynamical basis, it is also favoured by direct secondary structure mapping carried out by E.M. observation of mature rat 28S rRNA (29). The characteristic double hairpin loop detected in that work (see Fig. 1 in (29), note that the assignment of 5', 3' polarity was incorrect) precisely corresponds, both in size and location, to the long arms of the Y-shaped structure (the short stalk proposed in Fig. 9 is likely to be denatured in the conditions used for the E.M. observation). Comparison of the 3 vertebrates in this "D8" domain allows additional correlations to be made between secondary structure folding and phylogenetic status. Folding patterns (Fig. 9) are closely analogous except for length differences in the giant stems. It is remarkable that a long, stalk-proximal portion of one of the giant stems is conserved in the 3 vertebrates while the entire stems are conserved between the rodents but their terminal tips. The preferential addition of new sequence tracts in the areas where the former enlargement had already taken place during the evolution of higher eukaryotes, together with the secondary structure arrangement of the large tracts of remnant sequences, makes the expansion pattern in this D domain clearly reminiscent of a continued "growing tip" process.

5. Spacer-like domains in mature 28S rRNA.

By their high potential to form self-contained very stable stem structures and by their history of repeated insertion and deletion events, the so-called "D" domains of 28S rRNA gene in higher eukaryotes are again closely related to the transcribed spacers of the ribosomal transcription unit (18, 13, 30, 31). Although the presence of very short introns cannot definitely be ruled out so far, all the experimental evidences suggest that most (if not all) the transcripts of "D" domains are present in mature 28S rRNA of higher eukaryotes : very similar size and base content of sequenced genes and mature rRNAs, detection of the characteristic GC-rich giant stems (29) in mature 28S rRNA as mentioned above, protection from S1 nuclease of rRNA-DNA hybrids (5). A more direct evidence has been obtained recently for D1 domain, in a variety of eukaryotes, through rRNA sequencing, using reverse transcriptase (L.H. Qu and J.P. Bachellerie, in preparation). These experiments moreover confirm the extremely high sequence homogeneity of the ribosomal gene family (about 200 repeats) in mouse. However, contrarily to what is found for the internal transcribed spacer regions (13) relatively large subareas of the "D" domains of 28S rRNA are conserved between distant vertebrates such as mouse and xenopus (Fig. 3). Within the "D" domains, the slower rate of variation of these subareas is clearly confirmed by the mouse-rat comparison : their overall degree of divergence is 0.60 % instead of 7.7 % for the remaining parts of the D domains (even without taking into account the 9 segments which have varied extensively between both rodents), while a value of 0.27 % was obtained for the entire common core (Table 2). This relatively slow rate of variation and the presence of closely related secondary structure features (as exemplified in Fig. 9) raise the possibility of their being involved in functions shared by moderately distant eukaryotes. More should be learned on this point by identifying the molecular interactions (RNA-RNA or RNA-proteins) in which these definite domains may be involved in higher eukaryotes, either during the ribosome cycle in the cytoplasm or even during its assembly and transport from nucleolar sites.

ACKNOWLEDGMENTS

We are grateful to I. Grummt for providing us with pMB2 and pMEB1 recombinants, to S. Gerbi for communicating the *X. laevis* 28S rRNA sequence before publication, and to F. Michel for providing computer programs and for his generous assistance in their adaptation. We thank M. Parès for typing the manuscript. The technical assistance of J. Feliu and N. Joseph was appreciated. We thank Pr. J.P. Zalta for his continued support. This work was supported by a grant from C.N.R.S. ("ATP Organisation du Génome Eucaryote"), by an ADRC contract (6107), by a fellowship from a Ligue Nationale Française Contre le Cancer to B.M. and a scholarship (Ministère de la Coopération) to N.H.

*To whom correspondence should be sent

REFERENCES

1. Gerbi, S.A., Gourse, R.L. and Graham, C.G. (1982) in The Cell Nucleus, Busch, H. and Rothblum, L.I. Eds., Vol. X, pp. 351-386, Academic Press, New York.
2. 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.
3. Georgiev, O.I., Nikolaev, N., Hadjiolov, A.A., Skryabin, K.G., Zakharyev, V.M. and Bayev, A.A. (1981) Nucleic Acids Res. 9, 6953-6958.
4. Otsuka, T., Nomiyama, H., Yoshida, H., Kukita, T., Kuhara, S. and Sakaki, Y. (1983) Proc. Natl. Acad. Sci. USA 80, 3163-3167.
5. Ware, V.C., Tague, B.W., Clark, C.G., Gourse, R.L., Brand, R.C. and Gerbi, S. (1983) Nucleic Acids Res. 11, 7795-7817.
6. Chan, Y.L., Olvera, J. and Wool, I.G. (1983) Nucleic Acids Res. 11, 7819-7831.
7. Brosius, J., Dull, T.J. and Noller, H.F. (1980) Proc. Natl. Acad. Sci. USA 77, 201-204.
8. Branlant, C., Krol, A., Machatt, M.A., Pouyet, J., Ebel, J.P., Edwards, K. and Kossel, H. (1981) Nucleic Acids Res. 9, 4303-4324.
9. Noller, H.F., Kop, J., Wheaton, V., Brosius, J., Gutell, R.D., Kopylov,

A., Dohme, F., Herr, W., Stahl, D.A., Gupta, R. and Woese, C.R. (1981) Nucleic Acids Res. 9, 6167-6189.

10. Glotz, C., Zwieb, C., Brimacombe, R., Edwards, K. and Kössel, H. (1981) Nucleic Acids Res. 9, 3287-3306.
11. Raynal, F., Michot, B. and Bachellerie, J.P. (1984) FEBS Lett. 167, in press
12. Michot, B., Bachellerie, J.P. and Raynal, F. (1982) Nucleic Acids Res. 10, 5273-5283.
13. Michot, B., Bachellerie, J.P. and Raynal, F. (1983) Nucleic Acids Res. 11, 3375-3391.
14. Goebel, W. and Bonewald, R. (1975) J. Bacteriol. 123, 658-665.
15. Maxam, A.M. and Gilbert, W. (1980) Methods in Enzymol. 65, 499-560.
16. Michel, F., Jacquier, A. and Dujon, B. (1982) Biochimie 64, 867-881.
17. Kominami, R., Mishima, Y., Urano, Y., Sakai, M. and Muramatsu, M. (1982) Nucleic Acids Res. 10, 1963-1979.
18. Hall, L.M.C. and Maden, B.E.H. (1980) Nucleic Acids Res. 8, 5993-6005.
19. Subrahmanyam, C.S., Cassidy, B., Busch, H. and Rothblum, L.I. (1982) Nucleic Acids Res. 10, 3667-3680.
20. Skryabin, K.G., Kraev, A.S., Rubstov, P.M. and Baev, A. (1979) Dokl. Akad. Nauk. SSR 247, 761-765.
21. Veldman, G.M., Brand, R.C., Klootwijk, J. and Planta, R.J. (1980) Nucleic Acids Res. 8, 2907-2920.
22. Veldman, G.M., Klootwijk, J., Van Heerikhuizen, H. and Planta, R.J. (1981) Nucleic Acids Res. 9, 4847-4862.
23. Kumano, M., Tomioka, N. and Sugiura, M. (1983) Gene 24, 219-225.
24. Gourse, R.L., Thurlow, D.L., Gerbi, S.A. and Zimmermann, R.A. (1981) Proc. Natl. Acad. Sci. USA 78, 2722-2726.
25. Roiha, H. and Glover, D.M. (1981) Nucleic Acids Res. 9, 5521-5532.
26. Rae, P.M.M., Kohorn, B.D. and Wade, R.P. (1980) Nucleic Acids Res. 8, 3491-3504.
27. Edwards, K. and Kössel, H. (1981) Nucleic Acids Res.9, 2853-2869.
28. Efstratiadis, A., Posakony, J.W., Maniatis, T., Lawn, R.M., O'Connell, C., Sprintz, R.A., DeRiel, J.K., Forget, B.G., Weissman, S.M., Slightom, J.L., Blechl, A.E., Smithies, O., Baralle, F.E., Shoulders, C.C. and Proudfoot, N.J. (1980) Cell 21, 653-668.
29. Schibler, U., Wyler, T. and Hagenbüchle, O. (1975) J. Mol. Biol. 94, 503-517.
30. Furlong, J.C. and Maden, B.E.H. (1983) The EMBO Journal 2, 443-448.
31. Furlong, J.C., Forbes, J., Robertson, M. and Maden, B.E.H. (1983) Nucleic Acids Res. 11, 8183-8196.