Biochem. J. (1985) 232, 725–733 (Printed in Great Britain)

Human 18 S ribosomal RNA sequence inferred from DNA sequence

Variations in 18 S sequences and secondary modification patterns between vertebrates

Fiona S. McCALLUM and B. Edward H. MADEN
Department of Biochemistry, University of Glasgow, Glasgow G12 8QQ, and
*Department of Biochemistry, University of Liverpool, P.O. Box 147, Liverpool L69 3BX, U.K.

We have determined the DNA sequences encoding 18 S ribosomal RNA in man and in the frog, *Xenopus borealis*. We have also corrected the *Xenopus laevis* 18 S sequence: an A residue follows G-684 in the sequence. These and other available data provide a number of representative examples of variation in primary structure and secondary modification of 18 S ribosomal RNA between different groups of vertebrates. First, *Xenopus laevis* and *Xenopus borealis* 18 S ribosomal genes differ from each other by only two base substitutions, and we have found no evidence of intraspecies heterogeneity within the 18 S ribosomal DNA of *Xenopus* (in contrast to the *Xenopus* transcribed spacers). Second, the human 18 S sequence differs from that of *Xenopus* by approx. 6.5%. About 4% of the differences are single base changes; the remainder comprise insertions in the human sequence and other changes affecting several nucleotides. Most of these more extensive changes are clustered in a relatively short region between nucleotides 190 and 280 in the human sequence. Third, the human 18 S sequence differs from non-primate mammalian sequences by only about 1%. Fourth, nearly all of the 47 methyl groups in mammalian 18 S ribosomal RNA can be located in the sequence. The methyl group distribution corresponds closely to that in *Xenopus*, but there are several extra methyl groups in mammalian 18 S ribosomal RNA. Finally, minor revisions are made to the estimated numbers of pseudouridines in human and *Xenopus* 18 S ribosomal RNA.

INTRODUCTION

We report here the nucleotide sequence of human 18 S ribosomal RNA (rRNA) inferred from the ribosomal DNA (rDNA) sequence. Although several 18 S ribosomal gene sequences have been published (reviewed by Nelles *et al.*, 1984), knowledge of the human sequence should be useful for a number of reasons. First, many early studies on ribosome biosynthesis in higher eukaryotes were carried out on cultured cells of human origin (HeLa cells) (Vaughan *et al.*, 1967; Jeanteur *et al.*, 1968; Warner & Soeiro, 1967; Wellauer & Dawid, 1973; Maden & Salim, 1974), but there has been no sequence data-base for the detailed interpretation of these experiments. Second, DNA sequence data are prerequisite for further experimental approaches to the biosynthesis and function of ribosomes in man, especially by the use of recombinant DNA techniques. Third, human ribosomal sequence data might be relevant to biomedical applications, for example in relation to the interaction of drugs with processes of ribosome biosynthesis or function.

In this paper we also report a correction to the *Xenopus laevis* 18 S sequence (briefly noted by Atmadja *et al.*, 1984) and a comparison with that of *X. borealis*. *Xenopus* 18 S rDNA is now particularly well characterized as a result of these and previous studies, and shows evidence of high sequence stability. We discuss other comparative 18 S sequence data in vertebrates in the light of these facts.

rRNA of higher eukaryotes contains numerous methyl groups. We summarize information on the locations of the methyl groups in 18 S rRNA in *Xenopus*, man and other mammals. This information will be relevant to gaining an understanding of the early steps in ribosome biosynthesis in the nucleolus, since most of the methyl groups are added to ribosomal precursor RNA rapidly after transcription (Maden & Salim, 1974), and methylation is functionally important in ribosome maturation (Vaughan *et al.*, 1967). We also summarize the more-limited available data on pseudouridine, since information on this class of modified nucleotides is likely to become relevant to understanding ribosome structure and assembly.

METHODS

Human 18 S rDNA

Fig. 1(*a*) shows the two cloned human rDNA fragments from which the 18 S sequence was determined. The rDNA was originally cloned as *Eco*RI fragments in bacteriophage λ vectors (Wilson *et al.*, 1978; Erickson *et al.*, 1981). The indicated fragments were recloned into the plasmid pBR322 (Erickson *et al.*, 1981; Wilson, 1982). The plasmid clones were kindly donated by Dr. G. N. Wilson. In the present work the rDNA insert from pHrB/SE was excised by restriction with *Eco*RI and *Sal*I and was purified by agarose gel electrophoresis. Fragments were excised from pHrA covering the region from the 18 S *Eco*RI site to the *Kpn*I site in ITS1. The

Abbreviations used: kb, kilobases; bp, base pairs.
* Present address of both authors.

*Sal*I–*Eco*RI fragment was cleaved with various further restriction endonucleases. The products of the various restrictions were subcloned into bacteriophage M13 vectors containing appropriate restriction site 'polylinkers' (Norrander *et al.*, 1983). Sequence analysis was carried out on single-stranded DNA templates (Fig. 1*b*) by the dideoxynucleotide terminator method (Sanger *et al.*, 1977, with subsequent modifications).

In order to sequence through the *Eco*RI site in the 18 S gene the following experiment was carried out. Human DNA, isolated from placenta, was restricted with *Pst*I and *Hin*dIII. This digestion procedure was designed to yield, among many other fragments, a *Pst*I/*Hin*dIII fragment of the 18 S gene approx. 1 kb in length, with the *Hin*dIII site near to the *Eco*RI site (Fig. 1*c*). The digest was subjected to electrophoresis on a 1% agarose gel. The 1 kb material was recovered and cloned directly into bacteriophage M13mp8, so as to place the desired rDNA insert in the correct orientation for sequencing from the *Hin*dIII site. Plaque hybridization indicated that about 1% of the recombinants contained the required rDNA insert. Three clones which gave positive signals were obtained as pure isolates. DNA from each of these clones gave the expected sequence through the *Eco*RI site.

Xenopus rDNA

The *X. borealis* rDNA clone pXbr101, which was previously used for sequencing the transcribed spacers and short regions at each end of the 18 S gene (Furlong & Maden, 1983; Furlong *et al.*, 1983), was used in this work for completing the *X. borealis* 18 S sequence analysis. This was accomplished by the method of Maxam & Gilbert (1980). Various further clones of *X. borealis* rDNA (Furlong & Maden, 1983) and *X. laevis* rDNA (Maden *et al.*, 1982; Stewart *et al.*, 1983) were used for carrying out short sequencing runs within two specific regions of the 18 S gene as summarized in the Results and discussion section.

RESULTS AND DISCUSSION

Human 18 S rDNA

The human 18 S rDNA sequence was determined by the strategy outlined in the preceding section. In order to ensure accuracy the sequence was covered extensively on both strands (Fig. 1*b*) and the sequencing gels were read independently by both authors at all points of difficulty. Minor difficulties were caused at a number of points by secondary structure (compression) effects or, in a few instances, artefactual band duplications. Most of these uncertainties were readily resolved by data from the complementary strand. In a few regions, and particularly between nucleotides 250 and 280, compression effects were interspersed on both strands. However, because the effects were at slightly different locations on the two strands the respective sequences could be deduced without unresolvable ambiguities. At, and immediately following, position 1776 it was not possible to establish with certainty from the dideoxynucleotide data whether there are two consecutive G residues or three. We have provisionally assigned two G residues here on the basis of comparative sequence data and secondary structure models. The 5′ and 3′ termini of the 18 S sequence were identified by their correspondence to the highly conserved terminal sequences of 18 S rRNA from human or other vertebrate sources (Eladari & Galibert, 1975; Vass & Maden, 1978; Salim & Maden, 1980). The inferred 18 S rRNA sequence is shown in Fig. 2. Aspects of the sequence will be discussed below.

Xenopus 18 S rDNA

The *X. laevis* 18 S sequence has been corrected by addition of an A residue following G-684. The presence of this A residue was first indicated as a result of encountering an *Alu*I site here, which was not predicted by the sequence. The earlier sequencing strategy had not employed *Alu*I in this region, and during sequence determination by the Maxam–Gilbert method the relevant nucleotide was masked by secondary structure on both strands. [The secondary structure on the rightwards strand had been recognized by Salim & Maden (1981), but that on the leftwards strand was inconspicuous and had been missed.] The presence of the A residue has been confirmed in several clones of *X. laevis* and *X. borealis* rDNA (all those examined) by dideoxynucleotide sequencing. We have exhaustively rechecked the *X. laevis* 18 S sequence and have found no other errors or uncertainties.

X. borealis 18 S rDNA differs at only two points from *X. laevis*. Both differences are base substitutions (Fig. 2) and both are in regions of 18 S rDNA which have been found to be variable in comparisons between more distantly related eukaryotes (Salim & Maden, 1981; Nelles *et al.*, 1984).

Previous evidence indicated that 18 S rDNA is highly homogeneous in *X. laevis* (Maden *et al.*, 1982). In the present work, sequence analysis was carried out on eight further clones of *X. laevis* rDNA and four further clones of *X. borealis* rDNA through the regions where the interspecies differences occur. This analysis did not reveal any intraspecies variation, either at the two points of interspecies difference (Fig. 2) or in the flanking sequences. From the lack of evident intraspecies heterogeneity and the very small degree of interspecies divergence it can be concluded that 18 S rDNA is extremely stable in *Xenopus*.

Two further comments may be made. First, in contrast to the 18 S gene, the transcribed spacers are highly labile in *Xenopus*. Their sequences show extensive divergence between *X. laevis* and *X. borealis* starting a few nucleotides outside the 18 S gene (Furlong & Maden, 1983; Furlong *et al.*, 1983). Moreover, there are multiple heterogeneities in the transcribed spacers of *X. laevis* (Stewart *et al.*, 1983) and *X. borealis* (B. E. H. Maden, unpublished work). Second, the 18 S gene region and transcribed spacers have now been characterized in great detail in *Xenopus*. Because all of the relevant data have been established in a single laboratory, with extensive cross-checking, particular confidence may be placed on the conclusions outlined above.

Differences between human and *Xenopus* 18 S rDNA

The human 18 S sequence is 43 nucleotides longer than that of *Xenopus*. Most of the extra nucleotides are in short blocks near the 5′ end of the sequence (Fig. 2). The largest group of extra nucleotides is between positions 240 and 280 of the human sequence. This large block of extra nucleotides lies within a region that has previously been identified as a tract of major variability between 18 S rRNA of distantly related eukaryotes (Salim & Maden, 1981; Nelles *et al.*, 1984). Outside the variable tracts in the 5′ region the majority of differences are single base

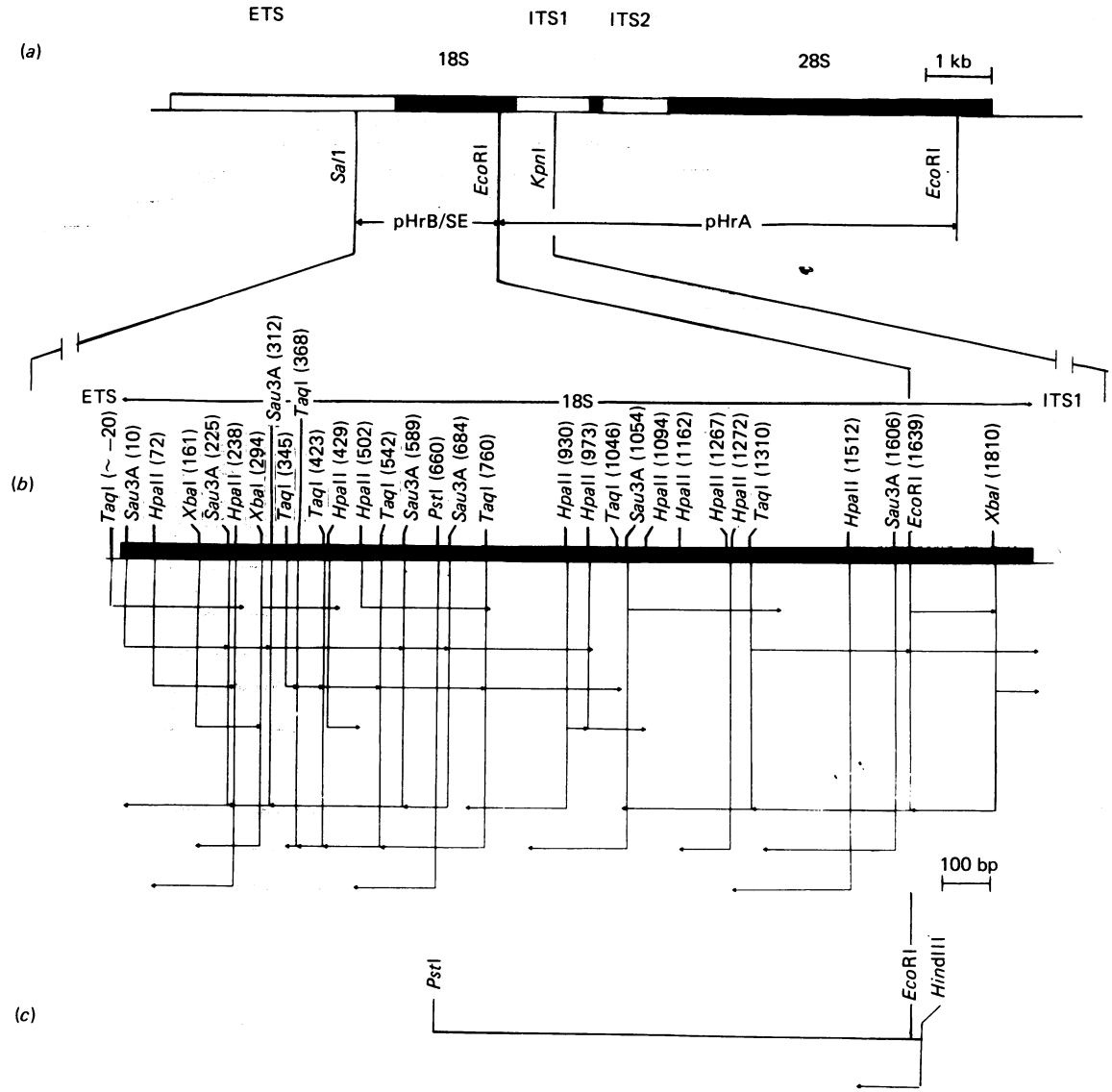

Fig. 1. Sequencing strategy for human 18 S rDNA

(*a*) The human ribosomal transcription unit showing the regions contained in the clones pHrB/SE and pHrA. (*b*) Sequencing determinations carried out on the 18 S rDNA regions of pHrB/SE and pHrA. Some M13 clones contained multiple short inserts due either to incomplete digestion of the original rDNA fragment or to ligation of multiple digestion products. Such multiple inserts were recognized in sequencing gels from the presence of sequences for the respective restriction endonuclease (e.g. *Sau*3A:GATC). (*c*) Three clones containing the indicated *Pst*I/*Hin*dIII fragment from human placental DNA were obtained and were sequenced leftwards from the *Hin*dIII site (see the Methods section). The *Hin*dIII site (AAGCTT) is at 1663 in the human sequence (Fig. 2).

substitutions, whose frequency in different parts of the sequence also follows known phylogenetic trends. Where extra material occurs in the human sequence it is not always possible to distinguish exactly which of the non-homologous nucleotides have resulted from insertions and which from substitutions. Subject to this qualification, the human sequence can be described as differing from that of *Xenopus* by about 2.3% of extra material and some 77 base changes. The latter comprise about 4.2% divergence in a common core sequence. The majority of changes, both additions and substitutions, are in the direction of higher G+C content in human than in *Xenopus* RNA. The majority of base substitutions are transitions and the majority of these affect pyrimidines in the RNA-like strand (Table 1). The excess of pyrimidine transitions over those involving purines signifies that not all of the transitions contribute to compensating base changes in helical arms of the RNA structure. In fact, only three pairs of substitutions between the *Xenopus* and human sequences generate unambiguous, compensating base changes in secondary structure models (Atmadja *et al.*, 1984; Nelles *et al.*, 1984) (pairs 321/330, 1539/1594 and 1738/1796 in the human numbering system). The great majority of the other substitutions are in single-stranded regions, including the tips and lateral bulges of several helices, especially in the model of Atmadja *et al.* (1984). This and other evidence which is relevant to distinguishing between alternative proposals

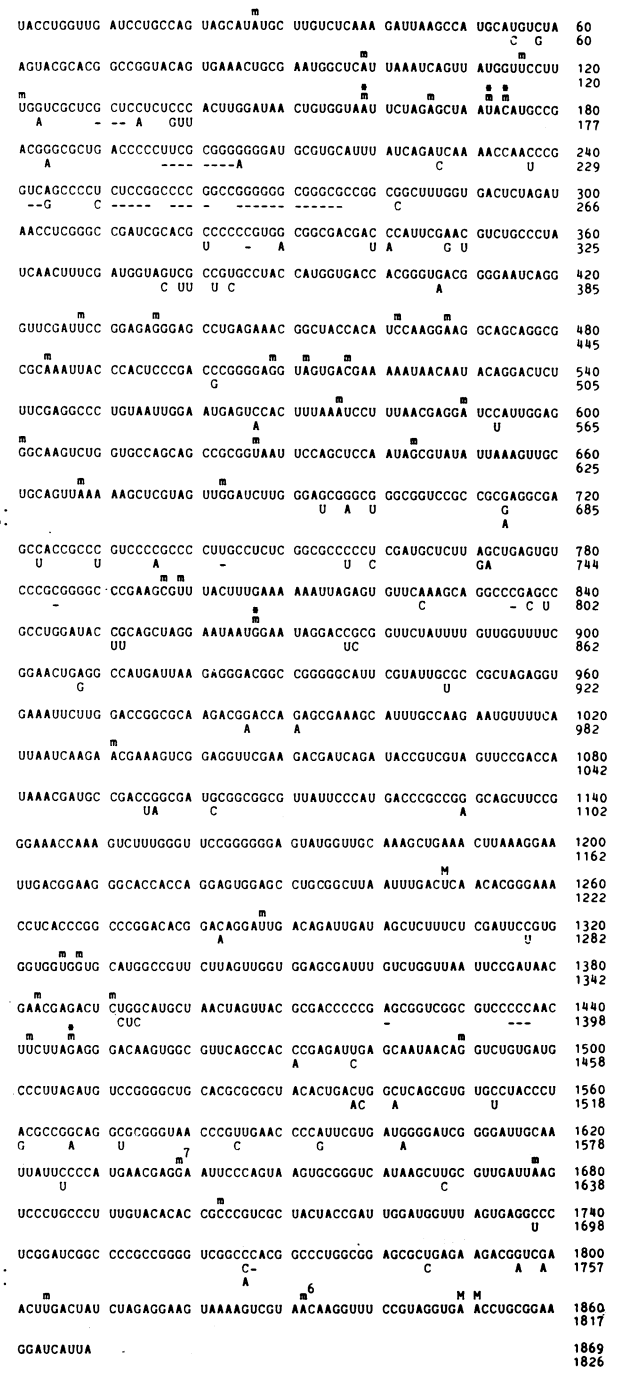

Fig. 2. Nucleotide sequence of human 18 S rRNA inferred from the DNA sequence and comparison with *Xenopus*

The first subscript line shows the positions at which the *X. laevis* sequence differs from that of human or from *X. borealis*. Where extra nucleotides occur in the human sequence dashes are shown in the aligned *X. laevis* sequence. The second subscript line shows the *X. borealis* substitutions at the two points of difference from *X. laevis* (positions 679 and 1724 in the *Xenopus* numbering system). Note that the *X. laevis* sequence has been corrected by inclusion of A-685 (equivalent to A-720 in the human sequence). A-685 is within an experimentally demonstrated *Alu*I site (AGCT) in *Xenopus*. The necessary adjustment of the *X. laevis* numbering system results in renumbering of all nucleotides downstream from A-685 (note especially those bearing methyl groups) by +1 with respect to Salim & Maden (1981), Maden (1982) and Maden *et al.* (1982).

for secondary structure in several regions of human 18 S rRNA will be discussed in detail elsewhere (F. S. McCallum & B. E. H. Maden, unpublished work). Meanwhile it will be noted that there is particularly high sequence conservation in the terminal regions and in two internal regions: 390–690 and 1140–1380 in the human numbering system.

Mammalian 18 S comparisons

Non-primate mammalian 18 S sequences have been reported from rat (Torczynski *et al.*, 1983; Chan *et al.*, 1984), mouse (Raynal *et al.*, 1984) and rabbit (Connaughton *et al.*, 1984). The rabbit sequence was determined directly from RNA; the rat and mouse sequences were from cloned rDNA.

It is clear, from comparison of these data with the human sequence and with each other, that the various mammalian 18 S sequences are closely similar. In fact the true extent of sequence conservation may be even higher than is apparent from any single pairwise comparison. Table 2 illustrates this for the human and rodent sequences, derived from rDNA. There are 15 points at which the mouse sequence (Raynal *et al.*, 1984) apparently differs from human (see the first two columns of the Table). All except three of these differences are either clustered into two subregions between nucleotides 190 and 280, where the greatest differences between human and *Xenopus* 18 S rDNA also occur (see Fig. 2), or are at isolated sites where the mouse and rat sequences are alike but differ from human (positions 140, 722 and 1095 in the human numbering system). Of the three remaining points of difference between mouse and human, there are two sites where the mouse sequence apparently lacks a nucleotide which is otherwise conserved across a broad phylogenetic range [nucleotides 286 and 1228 in the human numbering system: see footnotes (*f*) and (*k*) to Table 2 and Nelles *et al.*, 1984]. We therefore regard the status of these two apparent differences as doubtful, and we consider that there are 13 definite differences between the human and mouse 18 S sequences (indicated with a + sign in the Table). Nine of these differences are in the variable subregions between 190 and 280; elsewhere there are only four definite differences between human and mouse in the entire 18 S sequence, and at three of these positions mouse and rat are alike.

When the rat sequences are compared with each other there are a number of apparent differences between them (25 in all; Table 2). These could be interpreted as real differences between individual 18 S genes. However, we believe this to be unlikely at the majority of sites for several reasons. First, the number of apparent differences

Superscripts show the positions of methyl groups. All methylated nucleotides are common to human and *Xenopus* 18 S rRNA except those indicated with asterisks, which are unmethylated in *Xenopus*. An unqualified lower case m signifies a 2′-*O*-methyl group. Base methyl groups occur at the following positions in human and *Xenopus* (using the human numbering system): 1248, 3-(3-amino-3-carboxypropyl)-1-methylpseudouridine (amΨ, shown here as M); 1639, 7-methylguanine (m^7G); 1832, 6-methyladenine (m^6A); 1850 and 1851, 6-dimethyladenine (m^6_2A, shown here as M). Pseudouridines are not shown since only a few of these have been located (see the text and Table 4).

Table 1. Base compositions of *X. laevis* and human 18 S rDNA[a] with summary of nucleotide substitutions[b] and insertions

rDNA	No. of bases (% in parentheses) T	A	C	G	Total	G+C
X. laevis	411 (22.5%)	433 (23.7%)	466 (25.5%)	516 (28.3%)	1826	(53.8%)
X. laevis → human, lost by substitution[c]	−27	−23	−18	−9	−77	
X. laevis → human, gained by substitution[d]	11	8	33	25	77	
X. laevis → human, gained by insertion	6	1	19	17	43	
X. laevis → human, net change[e]	−10	−14	34	33	43	
Human	401 (21.5%)	419 (22.4%)	500 (26.8%)	549 (29.4%)	1869	(56.1%)

Substitutions[b]	*X. laevis* Human	No. of bases
Pyrimidine transitions	T → C	22
	C → T	11
Purine transitions	A → G	15
	G → A	6
Total transitions		54
Transversions		23

[a] The data are for the RNA-like strand of rDNA.
[b] The term 'substitution' refers to a net replacement between the aligned sequences in Fig. 2, and is used for descriptive convenience. It is not intended to imply the process(es) whereby the differences arose. At many isolated sites it is likely that the differences originated from single mutations. At other sites, especially where there are multiple differences, the actual processes of divergence cannot be reliably inferred.
[c] This line shows, with negative values, the number of positions at which the indicated base in *X. laevis* 18 S rDNA is substituted by a different base in human 18 S rDNA.
[d] This line gives the number of sites at which the indicated base in human 18 S substitutes for a different base in *X. laevis*.
[e] This line gives the algebraic sum of the numbers in the preceeding three lines.

between the two rat sequences is greater than the number of differences between the mouse and human sequences. Second, and relatedly, at nearly all points where the two rat sequences differ from each other the mouse and human sequences are in agreement, suggesting that these sites are not inherently particularly variable. Only one of the rat sequences differs from mouse and human at each of these points. Third, the rat data are in marked contrast to those discussed above pointing to the high stability of 18 S rDNA in *Xenopus*. Finally, at several points where the rat sequences differ from each other there are possible specific reasons for regarding one of the versions as doubtful, in most instances the version that differs from human and mouse (see the footnotes to Table 2). In the light of these considerations we regard only 14 of the apparent human–rat differences as definite; these are indicated with a + sign in the Table. Again, the majority of the definite differences are clustered in the variable region 190–280.

In summary, we infer from the combined data that the human and rodent 18 S sequences differ by just under 1%. The majority of differences are in the variable tracts between nucleotides 190–280; elsewhere there are only three or four definite differences in the entire 18 S sequence.

18 S rRNA methylation

All of the 40 methyl groups in *X. laevis* 18 S rRNA were previously located in the sequence (Salim & Maden, 1981). Most of the locations were precise; in a few instances there were short-range uncertainties of a few nucleotides.

Oligonucleotide data (Khan *et al.*, 1978) show extensive homologies between the methylation patterns of 18 S rRNA from *Xenopus*, human and other mammalian cells, with only a few differences between the *Xenopus* and mammalian methyl 'fingerprints'. The 18 S methyl 'fingerprints' from three mammalian sources were identical. From this starting point, and with additional data on the precise locations of methyl groups within some oligonucleotides (Choi & Busch, 1978; Fuke & Busch, 1979), we have been able to infer the positions of nearly all of the methyl groups in human 18 S rRNA (Fig. 2). Only two methyl groups remain to be located, both of which are on fractionally methylated nucleotides. All of the inferred locations are in agreement with those deduced independently by Connaughton *et al.* (1984) in their direct sequence analysis of rabbit 18 S rRNA, except that two of the human methylation sites do not appear in the rabbit sequence.

The methylation sites in *Xenopus* and mammalian 18 S rRNA can be categorized into three groups (Fig. 2, Table 3) according to the following homology patterns. First, the great majority of sites are common to *Xenopus* and mammals, with fully conserved primary structures. In all of the methylated oligonucleotides which are common to 18 S rRNA from HeLa cells and *Xenopus*, and which were fully analysed from both (Khan *et al.*, 1978), the methyl group was found to be in the same position in both sources. We have here assumed this to be true also in the few instances where we have relied on mammalian data from other laboratories (Choi & Busch, 1978; Fuke & Busch, 1979) on the exact positions of the methyl groups in oligonucleotides.

Table 2. Differences between sequence data for human, mouse and rat 18 S rDNA

Human[a]	Mouse[b]	Rat[c]	Rat[d]	Definite differences: Human–mouse	Definite differences: Human–rat	Definite differences: Mouse–rat	Note
G_{123}	G_{123}	$-_{123a}$	G_{123}				*e*
C_{140}	T_{140}	T_{141}	T_{140}	+	+		
$-_{196a}$	C_{197}	C_{199}	C_{197}	+	+		
G_{200}	G_{201}	C_{203}	C_{201}		+	+	
G_{203}	G_{204}	T_{206}	T_{204}		+	+	
$-_{207a}$	G_{209}	G_{211}	G_{209}	+	+		
$-_{207b}$	$-_{209a}$	G_{212}	G_{210}		+	+	
G_{208}	G_{210}	G_{213}	A_{211}				
T_{210}	T_{212}	C_{215}	C_{213}		+	+	
C_{243}	G_{245}	C_{249}	C_{246}	+		+	
C_{247}	T_{249}	C_{253}	C_{250}	+		+	
T_{250}	C_{252}	C_{256}	C_{253}	+	+		
C_{251}	T_{253}	T_{257}	T_{254}	+	+		
T_{252}	C_{254}	C_{258}	C_{255}	+	+		
G_{256}	G_{258}	$-_{261a}$	G_{259}				
C_{258}	T_{260}	T_{263}	T_{261}	+	+		
G_{270}	T_{272}	T_{275}	T_{273}	+	+		
$-_{278a}$	$-_{280a}$	$-_{283a}$	C_{282}				
$-_{280a}$	$-_{282a}$	A_{286}	$-_{284a}$				
T_{286}	$-_{287a}$	T_{292}	T_{290}				*f*
C_{321}	C_{322}	C_{327}	T_{325}				
C_{324}	C_{325}	T_{330}	$-_{327a}$				
G_{407}	G_{408}	A_{413}	G_{410}				
A_{720}	A_{721}	A_{726}	G_{723}				*g*
G_{721}	G_{722}	G_{727}	C_{724}				
C_{722}	T_{723}	T_{728}	T_{725}	+	+		
C_{725}	C_{726}	C_{731}	G_{728}				
$-_{730a}$	$-_{731a}$	$-_{736a}$	T_{734}				
$-_{736a}$	$-_{737a}$	$-_{742a}$	A_{741}				
T_{743}	T_{744}	T_{749}	$-_{747a}$				
G_{845}	G_{846}	A_{852}	G_{849}				*h*
G_{986}	G_{987}	G_{994}	A_{990}				
C_{1095}	T_{1096}	T_{1104}	T_{1099}	+	+		*j*
A_{1228}	$-_{1228a}$	$-_{1237}$	A_{1232}				*k*
A_{1295}	A_{1295}	A_{1304}	G_{1299}				
$-_{1392a}$	$-_{1392a}$	$-_{1402a}$	C_{1397}				*l*
C_{1472}	C_{1472}	C_{1484}	$-_{1476a}$				
$-_{1537a}$	$-_{1537a}$	$-_{1552a}$	A_{1542}				
C_{1542}	C1542	C_{1557}	T_{1547}				
A_{1561}	G_{1561}	A_{1577}	A_{1566}	+		+	
C_{1774}	C_{1774}	T_{1790}	T_{1779}				*m*
G_{1777}	G_{1777}	$-_{1793}$	G_{1782}				
C_{1783}	C_{1783}	C_{1799}	G_{1788}				
G_{1784}	G_{1784}	G_{1800}	C_{1789}				
Total differences				13	14	7	

[a] Nucleotides in the human 18 S rDNA column are numbered according to Fig. 2, except for numbers such as 196a (see below).

[b] Nucleotides in mouse 18 S rDNA are numbered acording to Raynal *et al.* (1984).

[c] Nucleotides in this column are in the rat sequence of Torczynski *et al.* (1983). These author used an expanded numbering system to accommodate sequence data from other species; hence the numbers become progressively more out of register with those in the other columns towards the 3′ end of the sequence.

[d] Nucleotides in this column are in the rat sequence of Chan *et al.* (1984).

[a–d] Numbers such as 196a signify (in this example) an extra nucleotide in the rodent sequences immediately following the nucleotide corresponding to 196 in the human sequence. Thus the extra nucleotide in this example is at position 197 in the mouse sequence, 199 (the corresponding position) in the rat sequence of Torczynski *et al.* and 197 in the rat sequence of Chan *et al.*

[e] All of the eukaryotic sequences listed in Nelles *et al.* (1984) contain G or A at this position.

[f] This position is several nucleotides beyond the right hand end of the 'variable length' region, 250–280. None of the eukaryotic sequences listed in Nelles *et al.* (1984) lacks a nucleotide here, and the actual sequence is conserved from *Saccharomyces cerevisiae* to vertebrates.

[g] The rat sequence of Chan *et al.* (1984) differs from the other three sequences at several points in the region 720–743.

[h] When sequencing this region in *Xenopus* by the Maxam–Gilbert method we encountered a methylated C residue here, within an *Eco*RII site. Since methylated C reacts weakly with hydrazine it can be mistaken for T (and hence A on the other strand) if sequencing is only carried out on one strand. This appears to have been the case in this short region according to the sequencing strategy in Fig. 1 of Torczynski *et al.* (1983). (In *X. laevis* 18 S rDNA all *Eco*RII sites were confirmed by cleavage with the isoschizomer, *Bst*NI).

Table 3. Differences between methylated oligonucleotides from human and *X. laevis* 18 S rRNA

Methylated nucleotide, position in sequence		Oligonucleotide[a, b]		Note
		Designation	Sequence	
Methyl group present in human and *X. laevis* 18 S RNA: oligonucleotide difference				
Human	590	T66	(GG)^{m}AUCCAUUG	[c]
X. laevis	555	T82 *X.l.*	(GG)^{m}AUCUAUUG	
Human	1391	T45	(G)ACU^{m}CUG(GCAUG)	[d]
X. laevis	1353	T70a *X.l.*	(G)ACU^{m}CCUCCAUG	
Human	1442, 1447	T89/94	(G)UCCCCCAACUU^{m}CUUAG^{m}AG	[e]
X. laevis	1400	T89	(G)UCC - - - AACUU^{m}CUUAG(AG)	
Human	1804	T47	(GGUCG)AACU^{m}UG	
X. laevis	1761	T69 *X.l.*	(G)AUCAAACU^{m}UG	
Methyl group present in human 18 S rRNA, absent from *X. laevis*				
	159	T82 H	(G)U^{m}AAUUCUAG	[f]
	172, 174	T68	(G)CUAAUAC^{m}AU^{m}G	[f]
	867	T69 H	(G)AAUAAU^{m}GG	[f]
	1447	T94	See above	[e]
	?	T8	(G)C^{m}CCG	[g]
	?	T42	(G)C^{m}UUG	[g]

[a] The oligonucleotide designation is the number of the T_1 ribonuclease digestion product in Khan *et al.* (1978). Where necessary the letters *X.l.* and H are used to distinguish between similarly numbered oligonucleotides with different sequences in *Xenopus* and human 18 S rRNA. The sequences were deduced by a combination of evidence from rRNA and rDNA, as summarized in outline by Maden (1982). (A more detailed account is in preparation by B.E.H.M.). Superscript m denotes a 2′-*O*-methyl group. Nucleotides which precede T_1 products, or are otherwise relevant, are included in parentheses. Points of difference between human and *Xenopus* are underlined. The human oligonucleotides are identical to those from other mammalian sources (Khan *et al.*, 1978; Choi & Busch, 1978).

[b] In addition to the oligonucleotides listed as differing between human and *X. laevis* 18 S rRNA, the methyl group at position 576 in the human sequence is located in a very long T_1 ribonuclease oligonucleotide (T92 of our designation) which was fully sequenced by Fuke & Busch (1979). This oligonucleotide differs by a single base between mammals and *X. laevis* (nucleotide 567 in the human sequence; Fig. 2).

[c] The underlined C residue gives rise to a *Bam*HI site in human 18 s rDNA as well as accounting for the indicated difference between the methylated oligonucleotides between human and *X. laevis* 18 S rRNA.

[d] As a result of this short block difference between human and *X. laevis* 18 S rRNA, the 2′-*O*-methyl group is released as the alkali-stable product CmU from human 18 S RNA and CmC from *Xenopus*. This is the closest incidence yet found of a base change to a methylation site.

[e] In addition to the indicated sequence difference affecting this region, G-1447 is fractionally methylated in human 18 S rRNA. The product, AGmAG, is a characteristic feature in 'T_1 plus pancreatic' ribonuclease fingerprints of HeLa Cell 18 S rRNA and is absent from *Xenopus* (Khan & Maden, 1976).

[f] The identical sequences are present in unmethylated form in *Xenopus* (Fig. 2).

[g] These two short products are fractionally methylated in human 18 S rRNA and have not yet been located in the sequence. Note that product T8 also contains CmCCG, a conserved methylation site at human position 1703, which is also methylated in *Xenopus*. The data in this Table account for all the differences between methylated oligonucleotides of human and *Xenopus* 18 S rRNA observed by Khan *et al.* (1978).

[j] The presence of C at this point in the human sequence is confirmed by the occurrence of a *Hpa*II site (Fig. 1*b* and Fig. 2).

[k] Although this A residue is missing from the mouse sequence and one of the rat sequences, the A and neighbouring nucleotides are highly conserved in other eukaryotes (Nelles *et al.*, 1984).

[l] The presence of C at this point would alter a methylated oligonucleotide from (G)ACUCmUG to (G)ACUCmUCG. The version with the extra C has not been reported in oligonucleotide analyses of rat (Choi & Busch, 1978) or other mammalian 18 S rRNA (Khan *et al.*, 1978).

[m] This region (1774–1784 in the human numbering system) can present difficulties due to secondary structure on the rightwards strand. Also, in *Xenopus*, human and mouse 18 S rDNA, the sequence contains an *Eco*RII site, which was experimentally confirmed for *Xenopus* and human rDNA by cleavage with *Bst*NI.

Second, there are a few methyl groups which occur in the same position in *Xenopus* and mammals, but where base changes in the vicinities of the methyl groups give rise to differences between the respective oligonucleotides in 'fingerprints' (Table 3). It was previously anticipated that this would be the underlying basis of some of the differences between the mammalian and *Xenopus* methyl fingerprints (Khan *et al.*, 1978).

Third, there are several extra methyl groups in mammalian 18 S rRNA, marked by asterisks in Fig. 2. Interestingly, these are located in sequences which are locally conserved between *Xenopus* and mammals, but in four instances there is extra material in the mammalian sequence not very far away (the methylation sites at positions 159, 172, 174, and 1447; Fig. 2).

All of the variations in methylation patterns between *Xenopus* and mammals affect 2′-*O*-ribose methyl groups. The methylated bases, which are at positions 1248, 1639, 1832, 1850 and 1851 in the human sequence (see the legend to Fig. 2 for details), are identical between *Xenopus* and mammals (and also in most lower eukaryotes: Klootwijk & Planta, 1973; Brand *et al.*, 1978). Moreover, the sequences surrounding the base methylation sites are conserved over considerable tracts of nucleotides and across broad phylogenetic distances (Salim & Maden, 1981). It can be concluded that the base methylations, which occur late during ribosome maturation (Maden & Salim, 1974, Brand *et al.*, 1978), are even more highly conserved in their structural features and their specific, individual roles than are the ribose methylations, which occur immediately after transcription of ribosomal precursor RNA (Maden & Salim, 1974).

Pseudouridine

The numbers of pseudouridines in human, mouse and *Xenopus* 18 S rRNA were previously estimated by base composition analysis with chromatographic separation of pseudouridine from uridine (Hughes & Maden, 1978). These estimates, which were carried out before the sequences were known, involved a calculation which relied upon indirectly derived values of the 18 S chain lengths. The sequence data indicate that those chain length values were roughly 10% too high. We therefore give revised estimates of the pseudouridine contents of human and *Xenopus* 18 S rRNA (Table 4), calculated as described in the Table legend. The revised value for human 18 S rRNA is in remarkably good agreement with the value obtained for rat 18 S rRNA from oligonucleotide analyses (Choi & Busch, 1978). This finding, together with the very high sequence conservation between human and rodent 18 S rRNA and the complete conservation of methylation sites, leads to the expectation that the pseudouridines are also located at the same sites in the two sequences. *Xenopus* 18 S rRNA appears to contain several more pseudouridines than does mammalian 18 S rRNA. Again on the basis of sequence conservation and the high homology between methylation patterns it is to be expected that the majority of pseudouridines in *Xenopus* 18 S rRNA are in the same locations as in mammalian 18 S rRNA. Limited amounts of oligonucleotide data (Khan & Maden, 1976; Salim & Maden, 1980) are in agreement with this expectation.

The difficult task of locating all the pseudouridines in the overall sequence has not yet been completed for any eukaryotic 18 S rRNA, although partial data have been obtained for some vertebrate species (Choi & Busch, 1978; Salim & Maden, 1980; Connaughton *et al.*, 1984).

There are indications from base composition data that most of the pseudouridines are introduced into ribosomal precursor RNA in the nucleolus (Jeanteur *et al.*, 1968). When the locations of the many pseudouridines in mature rRNA become known, it will be possible to undertake definitive analysis of the timing of the pseudouridine modifications.

Table 4. Pseudouridine content of 18 S rRNA

Species	T residues in rDNA[a]	Ψp/(Up+Ψp) (%)[b]	Ψ residues
Human	401	9.0	36+1[c] = 37
Rat[d]			38
X. laevis	411	10.8	44+1[c] = 45

[a] The numbers are for the RNA-like strand of 18 S rDNA (Table 1). In all instances for which data are available, the site of pseudouridine in rRNA corresponds to T in the RNA-like strand of rDNA. It is therefore assumed that all pseudouridines in rRNA are encoded by T and arise by postsynthetic modification of the appropriate uridines.

[b] These percentage values are taken from Table 2 of Hughes & Maden (1978), and were the means of multiple determinations using, in separate experiments, 18 S rRNA that had been labelled *in vivo* with ^{32}P or with [^{14}C]uridine.

[c] The correction '+1' is to include the hypermodified nucleotide, 3-(3-amino-3-carboxypropyl)-1-methylpseudouridine (amΨ), which is not recovered with the bulk of the Ψp (see also Fig. 2 and Brand *et al.*, 1978).

[d] The rat data were obtained by analysis of all oligonucleotides from T_1 ribonuclease hydrolysates (Choi & Busch, 1978).

Table 5. Approximate overall rates of 18 S sequence divergence between vertebrate lineages

Divergence	Approximate time since separation of lineages (Myear)	Sequence divergence (%)	Inferred interval for 1% divergence (Myear)
Human–*X. laevis*	300	6.5[a]	45
Human–rodents	70	< 1	> 70
(*X. laevis*–*X. borealis*	10[b]	0.11	90)
Suggested average			50–70

[a] This stated divergence value between human and *Xenopus* places equal weight on substitutions and insertions, and assumes that back mutations have not had a substantial effect. These assumptions can be refined when a more comprehensive 18 S phylogeny is undertaken; the intention here is to indicate an order of magnitude for divergence rates.

[b] This estimate of the time since separation of *X. laevis* and *X. borealis* derives from serum albumin data (Bisbee *et al.*, 1977). The value may be less accurate than the preceding ones, which are from palaeontological estimates. However, it may be noted that during the same period the transcribed spacers of *X. laevis* and *X. borealis* have diverged to the extent that there is little remaining homology (Furlong & Maden, 1983; Furlong *et al.*, 1983) indicating a divergence rate at least 100 times more rapid in the transcribed spacers than in the 18 S gene.

Concluding comments

The work described in this paper has established the DNA sequence encoding human 18 S rRNA, has located nearly all of the methyl groups in the inferred rRNA sequence and has given a refined estimate of the number of pseudouridine residues in 18 S rRNA. The comparative data generated from this work reinforce earlier evidence (summarized in Nelles *et al.*, 1984) and the 18 S sequence is characterized by high evolutionary stability. The vertebrate data are consistent with an overall rate of sequence divergence of roughly 1% per 50–70 million years (Table 5). Moreover, the rate of change is non-uniform along the sequence: some regions are practically constant between *Xenopus* and man. As previously noted (Salim & Maden, 1981) most of the methylation sites are concentrated within highly conserved regions. However, an intriguing complication is raised by the finding that some locally conserved sequences are methylated in mammals but not in *Xenopus* (Fig. 2, Table 3). It is becoming apparent that secondary modification is inter-related in a complex manner with primary structure, conformation and ribosome assembly. These topics will be discussed further in the light of secondary structure models (Atmadja *et al.*, 1984; Nelles *et al.*, 1984) in a subsequent report (F. S. McCallum & B. E. H. Maden, unpublished work).

Note added in proof (received 24 September 1985)

The results of sequencing by the Maxam–Gilbert method support our assignment of two rather than three G residues at and immediately following position 1776 in the human sequence.

We thank Dr. G. N. Wilson for the clones pHrB/SE and pHrA. This work was supported by the Medical Research Council.

REFERENCES

Atmadja, J., Brimacombe, R. & Maden, B. E. H. (1984) Nucleic Acids Res. **12**, 2649–2667
Bisbee, C. A., Baker, M. A., Wilson, A. C., Hadzi-Azimi, I. & Fischberg, M. (1977) Science **195**, 785–787
Brand, R. C., Klootwijk, J., Planta, R. J. & Maden, B. E. H. (1978) Biochem. J. **169**, 71–77
Chan, Y. L., Gutell, R., Noller, H. F. & Wool, I. G. (1984) J. Biol. Chem. **259**, 224–230
Choi, Y. C. & Busch, H. (1978) Biochemistry **17**, 2551–2560
Connaughton, J. F., Rairkar, A., Lockard, R. E. & Kumar, A. (1984) Nucleic Acids Res. **12**, 4731–4745
Eladari, M. E. & Galibert, F. (1975) Eur. J. Biochem. **55**, 247–255
Erickson, J. M., Rushford, C. L., Dorney, D. J., Wilson, G. N. & Schmickel, R. D. (1981) Gene **16**, 1–9
Fuke, M. & Busch, H. (1979) Nucleic Acids Res. **7**, 1131–1135
Furlong, J. C. & Maden, B. E. H. (1983) EMBO J. **2**, 443–448
Furlong, J. C., Forbes, J., Robertson, M. & Maden, B. E. H. (1983) Nucleic Acids Res. **11**, 8183–8196
Hughes, D. G. & Maden, B. E. H. (1978) Biochem. J. **171**, 781–786
Jeanteur, P., Amaldi, F. & Attardi, G. (1968) J. Mol. Biol. **33**, 757–775
Khan, M. S. N. & Maden, B. E. H. (1976) J. Mol. Biol. **101**, 235–254
Khan, M. S. N., Salim, M. & Maden, B. E. H. (1978) Biochem. J. **169**, 531–542
Klootwijk, J. & Planta, R. J. (1973) Eur. J. Biochim. **39**, 325–333
Maden, B. E. H. (1982) in The Cell Nucleus (Busch, H. & Rothblum, L., eds), vol. 10, pp. 319–351, Academic Press, New York
Maden, B. E. H. & Salim, M. (1974) J. Mol. Biol. **88**, 133–164
Maden, B. E. H., Forbes, J. M., Stewart, M. A. & Eason, R. (1982) EMBO J. **1**, 597–601
Maxam, A. M. & Gilbert, W. (1980) Methods Enzymol. **65**, 499–560
Nelles, L., Fang, B.-L., Volckaert, G., Vandenberghe, A. & De Wachter, R. (1984) Nucleic Acids Res. **12**, 8749–8768
Norrander, J., Kempe, T. & Messing, J. (1983) Gene **26**, 101–106
Raynal, F., Michot, B. & Bachellerie, J.-P. (1984) FEBS Lett. **167**, 263–268
Salim, M. & Maden, B. E. H. (1980) Nucleic Acids Res. **8**, 2871–2884
Salim, M. & Maden, B. E. H. (1981) Nature (London) **291**, 205–208
Sanger, F., Nicklen, S. & Coulson, A. R. (1977) Proc. Natl. Acad. Sci. U.S.A. **74**, 5463–5467
Stewart, M. A., Hall, L. M. C. & Maden, B. E. H. (1983) Nucleic Acids Res. **11**, 629–646
Torcynski, R., Bollon, A. P. & Fuke, M. (1983) Nucleic Acids Res. **11**, 4879–4890
Vass, J. K. & Maden, B. E. H. (1978) Eur. J. Biochem. **85**, 241–247
Vaughan, M. H., Soeiro, R., Warner, J. R. & Darnell, J. E. (1967) Proc. Natl. Acad. Sci. U.S.A. **58**, 1527–1534
Warner, J. R. & Soeiro, R. (1967) Proc. Natl. Acad. Sci. U.S.A. **58**, 1984–1990
Wellauer, P. K. & Dawid, I. B. (1973) Proc. Natl. Acad. Sci. U.S.A. **70**, 2827–2831
Wilson, G. N., Hollar, B. A., Waterson, J. R. & Schmickel, R. D. (1978) Proc. Natl. Acad. Sci. U.S.A. **75**, 5367–5371
Wilson, G. N. (1982) in The Cell Nucleus (Busch, H. & Rothblum, L., eds.), vol. 10, pp. 287–318, Academic Press, New York

Received 19 July 1985; accepted 16 August 1985