Nucleic Acids Research, 1993, Vol. 21, No. 6 ***1449–1455***

The sequence of 28S ribosomal RNA varies within and between human cell lines

Henrik Leffers and Annette H.Andersen
Institute of Medical Biochemistry and Danish Centre for Human Genome Research, Ole Worms Allé, Building 170, University Park, Aarhus University, DK-8000 Aarhus C, Denmark

Received December 1, 1992; Revised and Accepted February 22, 1993 EMBL accession nos X69338–X69372 (incl.)

ABSTRACT

The primary structure of 28S ribosomal RNA constitutes a conserved core which is similar among most 23S-like rRNAs and expansion segments which occur at specific positions in the sequence. The expansion segments account for most of the size difference between prokaryotic (archaeal and eubacterial) and eukaryotic rRNAs and they exhibit a sequence variation which is unique among rRNAs. We have investigated the sequence variation of one of the expansion segments, V8, by sequencing a total of 111 V8 segments from 9 different human cell lines and tissues and have found 35 different variants. The variation occur mainly at two 'hot spots' which are separated by 170 nucleotides in the primary sequence but are neighbours in the secondary structure. The sequence of V8 segments varies both within and between human cell lines and tissues. The implications for the evolution of the eukaryotic 28S rRNA are discussed together with possible functions of the expansion segments. We also present a secondary structure model for the V8 segment based on comparative sequence analysis and chemical and enzymatic foot printing.

INTRODUCTION

Human 28S rRNA genes exhibit a considerable sequence variation (1, 2), most of which occurs in regions that have been expanded in eukaryotic 28S rRNAs, compared to archaeal and eubacterial 23S rRNAs. These regions have been called expansion segments (3), V (variable) regions (4) or D (divergent) regions (5). The nucleotides in the expansion segments account for most of the size difference between eukaryotic and prokaryotic 23S-like rRNAs (e.g. Human 28S rRNA: 5025 nucleotides (1) and *Escherichia coli* 23S rRNA: 2904 nucleotides (6)).

In human 28S rRNA, there are 5 expansion segments larger than 100 nucleotides and 7 smaller ones (2, 4). The two largest, V2a (872 nucleotides) located between domains I and II and V8 (701 nucleotides) located within domain IV, account for 74% of the size difference between *E.coli* 23S rRNA and Human 28S rRNA. For the nomenclature and definitions of the variable regions, we have used the system defined by Wakeman & Maden (7). There are at least three other nomenclatures, these are reviewed in (7). Vertebrate expansion segments exhibit very high G+C contents, with a value of 87% for human V8. Sequence variation is also found in the conserved parts of human 28S rRNAs where either an A or a G occurs at position 60 (8). This variation is found both in the mature 28S rRNA and in the genes (8).

The origin of the sequence variation within the expansion segments is unknown, although a possible mechanism has been suggested (9). It is also unknown whether the sequence variation is the result of characteristic differences between the rRNA gene clusters. There have been reports of differential activity of rRNA gene clusters in different human tissues (10), which could indicate that differences in the expression of rRNA genes occur between different cell lines and tissues. A well studied example of differential expression of rRNA operons is in the malaria parasite *Plasmodium berghei*, where the two rRNA operons are differentially expressed in the two developmental stages of the parasite (11, 12). If the same principle applies to human cells, we may find that different cell types express different rRNA genes with characteristic sequences in the expansion segments. Moreover, if the sequence variation is large enough, it may be possible to design oligonucleotide probes which will distinguish between the different cell types.

We decided to investigate first whether the variation observed in the expansion segments of 28S RNA genes is also present in the 28S rRNA of cytoplasmic ribosomes and, second, whether there was any evidence for cell line specificity in the variation. Previous attempts to investigate sequence variation among 28S RNA molecules have used the very crude method of S1 nuclease mapping (13). We chose to sequence cDNAs, synthesized from the 28S rRNAs, using oligonucleotide primers complementary to conserved regions of the molecule. We selected the V8 segment for study which is located within Domain 4 and is about 700 nucleotides long.

MATERIALS AND METHODS

Cells

The following cell lines and tissues were used in this study: AMA, a transformed cell line derived from amnion epithelia cells; MRC-5, normal human embryonal lung MRC-5 fibroblasts; MRC-5 V2, SV 40 transformed MRC-5 fibroblasts; MOLT 4, a transformed lymphocyte cell line; HT29 cells, an adeno carcinoma cell line; LAN 5, a neuroblastoma cell line; HFEF, human fetal ear fibroblasts (primary fibroblast culture); *antrum* and *Mucosa sigmoideum*, both human tissue samples.

Preparation of DNA, RNA and ribosomes

DNA was prepared from cell monolayers or cell pellets by suspending the cells in 20mM Tris−HCl pH 7.5; 10mM EDTA; 0.2% Triton X100, followed by protease digestion (protease K, 50 u/ml) for three hours at 40°C. CsCl (1g/ml solution) and ethidium bromide (100μg/ml) were added and the sample was centrifuged at 42,000rpm for 48 hours. The DNA band was removed and the ethidium bromide was extracted with n-butanol, before the DNA was dialysed overnight against TE buffer (20mM Tris−HCl, pH 8.0; 1 mM EDTA). After dialysis, the DNA was extracted twice with phenol and twice with chloroform, dialysed overnight against TE buffer and stored at 4°C.

Total cellular RNA was prepared by the guanidinium thiocyanate/CsCl method (14), and was stored in water at −80°C.

Ribosomes were extracted by washing monolayers with lysis buffer (10mM Tris.HCl, pH 7.5; 1mM $MgCl_2$; 60mM KCl; 6 mM β-mercaptoethanol; 0.2% Triton X100). Nuclei were removed by centrifugation (10 min. at 10,000 rpm in a Sorvall SS34 rotor) and the clear solution was loaded onto a 1.5 ml sucrose cushion (15% sucrose; 5% $(NH_4)_2SO_4$ in lysis buffer) and centrifuged at 40,000 rpm for 3.5 hours in a Beckman SW50 rotor. The ribosome pellet was washed briefly with lysis buffer and stored at −80°C. RNA was prepared by suspending the pellets in 0.3 M Na-acetate pH 5.6; 1mM EDTA followed by phenol extractions.

Cloning of rRNA genes

A genomic library of Sal I digested chromosomal DNA from Molt 4 cells was constructed in λ EMBL 4 (14). The library was screened with [^{32}P]-labelled cDNA (see below), and 10 positive clones were grown on a large scale for further analysis. Restriction fragments containing the V8 segment were excised using restriction enzymes BamHI and HincII or BamHI, HincII and PvuII, and purified by electrophoresis on a 0.8% low melting agarose gel. The fragments were cloned into M13 mp18 and mp19 vectors (15) and sequenced using the chain termination method (16). The DNA sequence was determined on wedge shaped 5% polyacrylamide gels (17).

Preparation and cloning of ribosomal cDNA

An oligonucleotide (5′-CGAATTCCCCTGGTCCGCAC-3′) complementary to 28S rRNA (positions 3585 to 3604) was synthesized. The oligonucleotide contained one mismatch (underlined) which resulted in the creation of an EcoR1 site. Annealing was performed in 20 μl TK buffer (30mM Tris−HCl, pH 7.5; 50 mM KCl) containing 30pM oligonucleotide and 9μg total RNA. The sample was heated to 95°C for 1 min., quickly transferred to 56°C and allowed to slowly cool to 40°C (20 min). 20μl extension mixture (0.6 mM dNTPs; 125 mM Tris−HCl, pH 8.4; 25mM $MgCl_2$; 5mM dithiothreitol) containing 50u AMV-reverse transcriptase (Stratagene) were added and the sample was incubated at 43°C for 40 min. RNase H digestion and second strand synthesis were performed as described in (18). The cDNA was digested with EcoRI and BamHI and cloned into MT 719 (a M13 mp19 derivative, containing a T7 RNA polymerase promotor (J.Egebjerg & HL, unpublished results)) and sequenced as described above. Another oligonucleotide (5′-GCCGCAGCTGGGGCGAT-3′), complementary to position 3239 to 3246, was used for the screening and sequencing.

Secondary structure probing and computer analysis

The conditions for the chemical and enzymatic probing of the secondary structure were as described in (19). The following reagents and enzymes were used: Dimethylsulphate (DMS) (detects unpaired A and C), Kethoxal (unpaired G), RNase T_2 (single stranded regions) and the cobra venom enzyme (CVE) (base paired segments).

Alignments, secondary structure predictions and searches in databases were made on VAX computers (Digital), using the alignment program ALMA (20) and the UWGCG program package (21), version 6.2.

RESULTS

The numbering system we have used to define the positions of the sequence variations refers to one of the genomic clones, 'Genom 6' (see below). We have used this as reference since there are several positions where all our sequences are different from the published sequence (1). The 'Genom 6' sequence was also chosen for deriving the secondary structure.

In an initial attempt to investigate the sequence variation among V8 segments, we cloned 5 complete rRNA operons (Fig. 1). The BamHI/HincII restriction fragment containing the V8 SEGMENT was isolated from each and sequenced. Three variant forms were found among the six sequences (Fig. 2). To investigate whether the genomic variation was reflected in the 28S rRNA population of the cells and whether it showed any cell line specificity, we screened 146 V8 segments from 9 different cell lines or tissues. Of these, 106 contained the complete V8 segment and 32 variant forms were found among the 106 sequences. All the variants were sequenced on both strands.

Two 'hot spots' for sequence variation were found, one at position 470 and another around position 640 (Fig. 2). Despite their separation of about 170 nucleotides in the primary structure they lie close together in the putative secondary structure (see below). The variation around position 470 is caused by a deletion of 1 to 4 C residues and/or the deletion of the sequence AC_5 (Fig. 3). The variation around position 640 is within a repeated GGC motif, and generally results from a deletion of 1 to 3 of the GGC repeats although other variations are also found in this region (Fig. 2). Variation in these two regions is found among the genomic clones and in all the cell lines and tissues we have studied.

Apart from the two 'hot spots' we have localized a number of minor variations, i.e. single nucleotide changes or insertion/deletion of one to three nucleotides (Fig. 2). The minor variations were found in combinations with variations at the two 'hot spots', thus, among clones with the A→G change at position 180, both the absence and presence of large deletions were found

Figure 1. Map of the Sal I restriction fragment containing the genes for the rRNAs. The rRNA genes are shown in thick lines and the 5 expansion segments which are larger than 100 nucleotides are indicated by open lines. The positions of the two oligonuceotides that were used in this study are shown by arrow heads.

at the two hot spots (Fig. 3). As a result of this, the number of different V8 segments we have characterized is 35 for a total of 111 sequences (including the genomic sequences). The sizes of the V8 segments vary between 694 and 712 nucleotides.

Secondary structure

A tentative secondary structure model for the V8 segment is presented in Fig. 4. Owing to the biased base composition (87% G+C), there are many base pairing possibilities within the two long helixes (II & III). The structure presented in Fig. 4 is therefore only an attempt to structure the V8 segment. It was derived from a combination of an alignment of the vertebrate sequences, the Fold program (UWGCG program package (21)) and secondary structure probing.

The secondary structure probing was made on T7 RNA polymerase transcripts of the V8 segment from clone 'genom 6'. The probing with ribonucleases and chemical modification was made using the methods described in (19). Due to the extreme base composition it was not possible to read through the whole V8 segment, but we believe that the readings between positions 650−750 are reliable (Fig. 5). In the region between 480 and 650 we were only able to detect the very strong nuclease cuts

Cell Line	Number Observed		Position		Position	Other Variations
V8 pub		GCGCGGCGCGCCCCCCACCCCCACCCCACG	482	GGCGGCGGC.....AGGCGGCGGAGGG	650	
Genom 6	3	GCGCGCCCCCCCCCC.ACCCCCACCCCACG	485	GGCGGCGGCGGCGGCAGGCGGCGGAGGG	659	
Genom 9	1	GCGCGCCCCCCCC...ACCCCCACCCCACG	483	GGCGGCGGCGGCGGCAGGCGGCGGAGGG	657	
Genom 5	1	GCGCGCCCCCCCC...ACCCCC.....ACG	478	GGCGGCGG.....GCGGCGGAGGG	643	
HFEF 5	1	GCGCGCCCCCCCCCC.ACCCCCACCCCACG	485	GGCGGCGGCGGCGGCAGGCGGCGGAGGG	659	
HFEF 6	1	GCGCGCCCCCCCCCC.ACCCCCACCCCACG	485	GGCGGCGGC.....AGGCGGCGGAGGG	653	
Antrum 2	1	GCGCGCCCCCCCCCC.ACCCCCACCCCACG	485	GGCGGCGGCGGCGGCAGGCGGCGGAGGG	659	A→G(180)
AMA 9	31	GCGCGCCCCCCCCC..ACCCCCACCCCACG	484	GGCGGCGGCGGCGGCAGGCGGCGGAGGG	658	
Molt4 7	1	GCGCGCCCCCCCCC..ACCCCCACCCCACG	484	GGCGGCGGCGGCGGCAGGCGGCGGAGGG	658	G→A(564)
MRC5 2	1	GCGCGCCCCCCCCC..ACCCCCACCCCACG	484	GGCGGCGG.....GCAGGCGGCGGAGGG	653	
MRC5 4	1	ACGCGCCCCCCCCC..ACCCCCACCCCACG	484	GGCGGCGGC.....AGGCGGCGGAGGG	652	G→A(457)
HT29 18	3	GCGCGCCCCCCCCC..ACCCCCACCCCACG	484	GGCGGCGGCGGCGGCAGGCGGCGGAGGG	658	A→G(180)
Antrum 7	2	GCGCGCCCCCCCCC..ACCCCCACCCCACG	484	GGCGGCGGCGG...CAGGCGGCGGAGGG	655	
AMA 1	1	GCGCGCCCCCCCC...ACCCCCACCCCACG	483	GGCGGCGGCGGCGGCAGGCGGCGGAGGG	660	+CGG(527) G→A(622)
AMA 2	1	GCGCGCCCCCCCC...ACCCCCACCCCACG	483	GGCGGCGGCGG...CAGGCGGCGGAGGG	654	
MRC5V2 2	1	GCGCGCCCCCCCC...ACCCCCACCCCACG	483	GGCGGCGGCGGCGG...GCGGCGGAGGG	655	
MRC5V2 6	8	GCGCGCCCCCCCC...ACCCCCACCCCACG	483	GGCGGCGGCGGCGGCAGGCGGCGGAGGG	657	
HFEF 3	2	GCGCGCCCCCCCC...ACCCCCACCCCACG	483	GGCGGCGGC.....AGGCGGCGGAGGG	651	
HFEF 7	2	GCGCGCCCCCCCC...ACCCCCACCCCACG	483	GGCGGCGG.....GCGGCGGAGGG	648	A→G(180)
HFEF 8	1	GCGCGCCCCCCCC...ACCCCCACCCCACG	483	GGCGGCGGCGGCGGCAGGCGGCGGAGGG	657	A→G(180)
HT29 4	2	GCGCGCCCCCCCC...ACCCCCACCCCACG	483	GGCGGCGGCGGCGG...GCGGCGGAGGG	654	A→G(180)
HT29 5	2	GCGCGCCCCCCCC...ACCCCCACCCCACG	483	GGCGGCGGCGGCGG...GCGGCGGAGGG	654	
Antrum20	2	GCGCGCCCCCCCCCC.....CACCCCACG	481	GGCGGC.....AGGCGGCGGAGGG	646	
MRC5 9	2	GCGCGCCCCCCCCCC......ACCCCACG	480	GGCGGC.....AGGCGGCGGAGGG	645	
MRC5 1	7	GCGCGCCCCCCCCC.......ACCCCACG	479	GGCGGCGG.....GCGGCGGAGGG	645	
MRC5V2 7	1	GCGCGCCCCCCCCC.......ACCCCACG	479	GGCGGC.....CAGGCGGCGGAGGG	655	
MRC5V211	1	GCGCGCCCCCCCCC.......ACCCCACG	479	GGCGGCGGCGGCGGCAGGCGGCGGAGGG	652	
HFEF 11	2	GCGCGCCCCCCCCC.......ACCCCACG	479	GGCGGCGG.....GCGGCGGAGGG	645	A→G(180)
LAN5 6	3	GCGCGCCCCCCCCC.......ACCCCACG	479	GGCGGCGG.....GCGGCGGAGGG	645	+C(601)
AMA 7	1	GCGCGCCCCCCCCC.......ACCCCACG	478	GGCGGCGGCGGCGG...GCGGCGGAGGG	652	+CGG(527) G→T(415)
AMA 8	1	GCGCGCCCCCCCCC.......ACCCCACG	478	GGCGGC.....AGGCGGCGGAGGG	642	
Molt4 6	1	GCGCGCCCCCCCCC.......ACCCCACG	478	GGCGGCGGCGG...CAGGCGGCGGAGGG	649	
MRC5 8	2	GCGCGCCCCCCCCC.......ACCCCACG	478	GGCGGCGG.....GCGGCGGAGGG	643	
LAN5 8	9	GCGCGCCCCCCCCC.......ACCCCACG	478	GGCGGCGG.....GCGGCGGAGGG	643	
AMA 3	1	GCGCGCCCCCCCC.......ACCCCACG	477	GGCGGCGGCGGCGG...GCGGCGGAGGG	651	+CGG(527) G→T(415)
MRC5 10	1	GCGCGCCCCCCCC.......ACCCCACG	477	GGCGGCGG.....GCGGCGGAGGG	642	

Figure 2. Alignment of the V8 segment sequences in the two 'hot spots' arranged according to expansion segment type. 'Cell line' refers to which cell line the V8 segment originates from and the subsequent number refers to a particular clone from the cell line. 'Numbers observed' refer to how many sequences we have of the given type. 'V8 pub' refers to the sequence published by Gonzalez *et al.*, (1985). 'Genom' refers to genomic clones. Cell lines: AMA: a transformed Human Amnion cell line; Molt 4: a transformed lymphocyte cell line; MRC5: MRC 5 fibroblasts; MRC5V2: a SV40 transformed MRC 5 cell line; HFEF: Human Fetal Ear Fibroblasts (primary culture); HT29: HT 29 cells; LAN5: LAN-5 cells; Antrum & MuSig (*Mucosa sigmoideum*): human tissue samples. The alignment was constructed and edited using the alignment editing program ALMA (20). The sequences have been submitted to the EMBL database with accession numbers from X69338 to X69372.

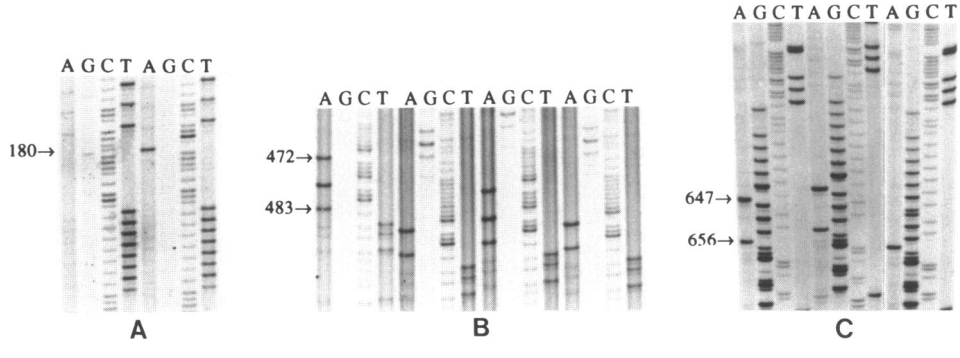


Figure 3. Examples of sequence variation within V8 segments. **A)** A→G change at position 180. **B)** Four examples of variation in the 470 region exhibiting small and large deletions. **C)** Three examples of variation in the 640 region showing variable numbers of GGC repeats and another common variation, the deletion of the sequence GAC. Numbering is according to the Genom 6 sequence.

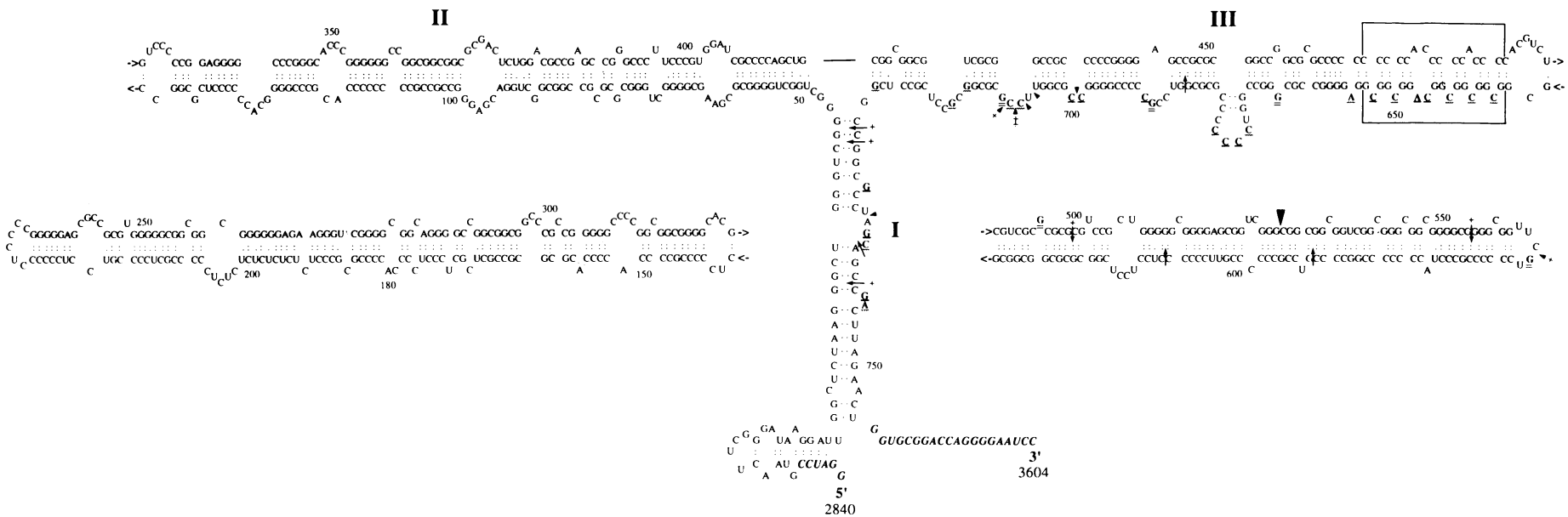


Figure 4. Secondary structure of the V8 segment from clone Genom 6. The 5′-BamHI site and the 3′-recognition sequence for the oligonucleotide primer are in italics. Nucleotides accessible to chemical modification by DMS (unpaired A & C) and kethoxal (unpaired G) are underlined; weak reactivity—dotted line, medium reactivity—single line and strong reactivity—double underlining. RNase T2 cuts (single strand-specific) are shown by filled arrow heads and nuclease CVE cuts (double strand-specific) are shown by arrows (++ strong reactivity; + medium; − weak). The position of the two hot spots is boxed and the possible AMA specific insertion is shown by a large arrow head. Numbering is according to the Genom 6 sequence.

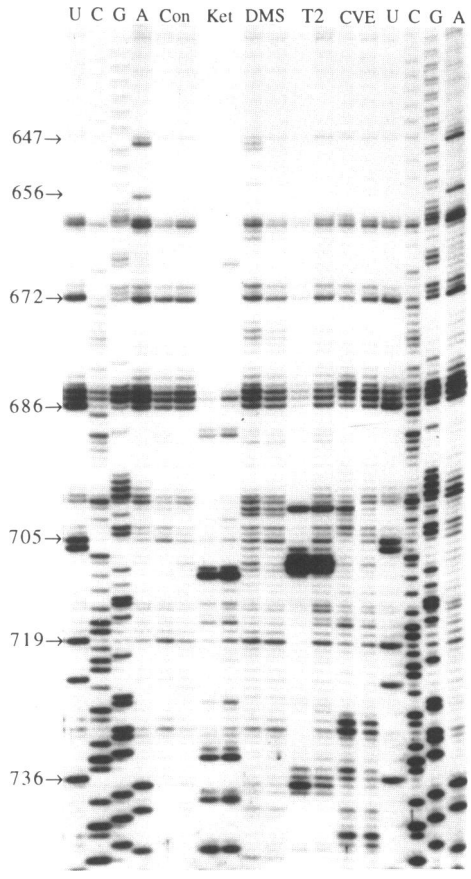


Figure 5. Enzymatic and chemical probing of the V8 segment. Two independent samples of each chemical modification and enzymatic digestions were run. Sequencing reactions were performed directly on the RNA. U, C, G, A, sequencing reactions. Con, control RNA; Ket, kethoxal-treated RNA; DMS, RNA reacted with DMS; T2 and CVE, RNA digested with ribonuclease T2 and cobra venom RNase, respectively. Numbering is according to the Genom 6 sequence.

and strong sites for chemical modifications. The accessible sites are shown on the secondary structure model (Fig. 4). Most of the V8 segment is quite inaccessible to the various probes, indicating that it folds into a compact secondary structure. However, the results of this study support the structure presented in Figure 4.

DISCUSSION

Our data establish that the sequence variations among 28S rRNA genes are also present within the 28S rRNAs of cytoplasmic ribosomes. The variations in the V8 segment are not random but are confined to a few regions which, with a single exception (A→G-180), are all located within helix III. There are two hot spots for variation, one around position 470 and one around position 640. In general, there seems to be a correlation between the sizes of the two deletions, clones with a large deletion at position 470 will often also have a large deletion around position 640. However, we have detected clones with almost every combination of small and large deletions (Fig. 2).

There are reports of other variations within the V8 segment (2). These variations occur at the two hot spots and within helix III where a deletion of two residues (CT) in a $(CT)_9$ motif has been observed (2), a variation that we have not detected. The

Table 1. Comparison of the sizes of the largest expansion segments.

Organism	Expansion Segment: V2a	V8	V11	Total	Reference
Animals					
Human	872	700	228	1801	(1)
Rat	783	594	229	1601	(27)
Mouse	696	612	225	1533	(5)
X.laevis	526	335	153	1014	(28)
D.melanogaster	350	218	177	745	(29)
H. homus	379	84	128	591	(30)
C.elegans	241	165	138	544	(31)
T. thermophila	(32)	245	126	108	478
T. brucei	242	238	a	–	(33)
C. faciculata	212	218	a	–	(34)
Plants					
P. micans	299	142	108	549	(35)
C. lemon	235	153	125	513	(36)
A. thaliana	233	152	127	512	(37)
S. alba	233	152	127	512	(38)
Tomato	232	153	121	506	(39)
Rice	221	151	125	497	(40)
Fungi					
D.discoideum	392	86	b	–	(41
P. polycephalum	254	163	180	597	(42)
M. racemosus	318	124	126	568	(43)
S.cerevisia	216	155	137	508	(44)

a: This expansion segment cannot be precisely located in this organism.
b: Not sequenced.

total number of variant V8 segments reported so far (this study included) is as high as 39, a surprisingly high number when compared to the reported number of genes (300–400; 22, 23). Moreover, it is unlikely that we have detected all the variants since we still detect almost as many new variants per 10 sequences now as we did at the start of this study.

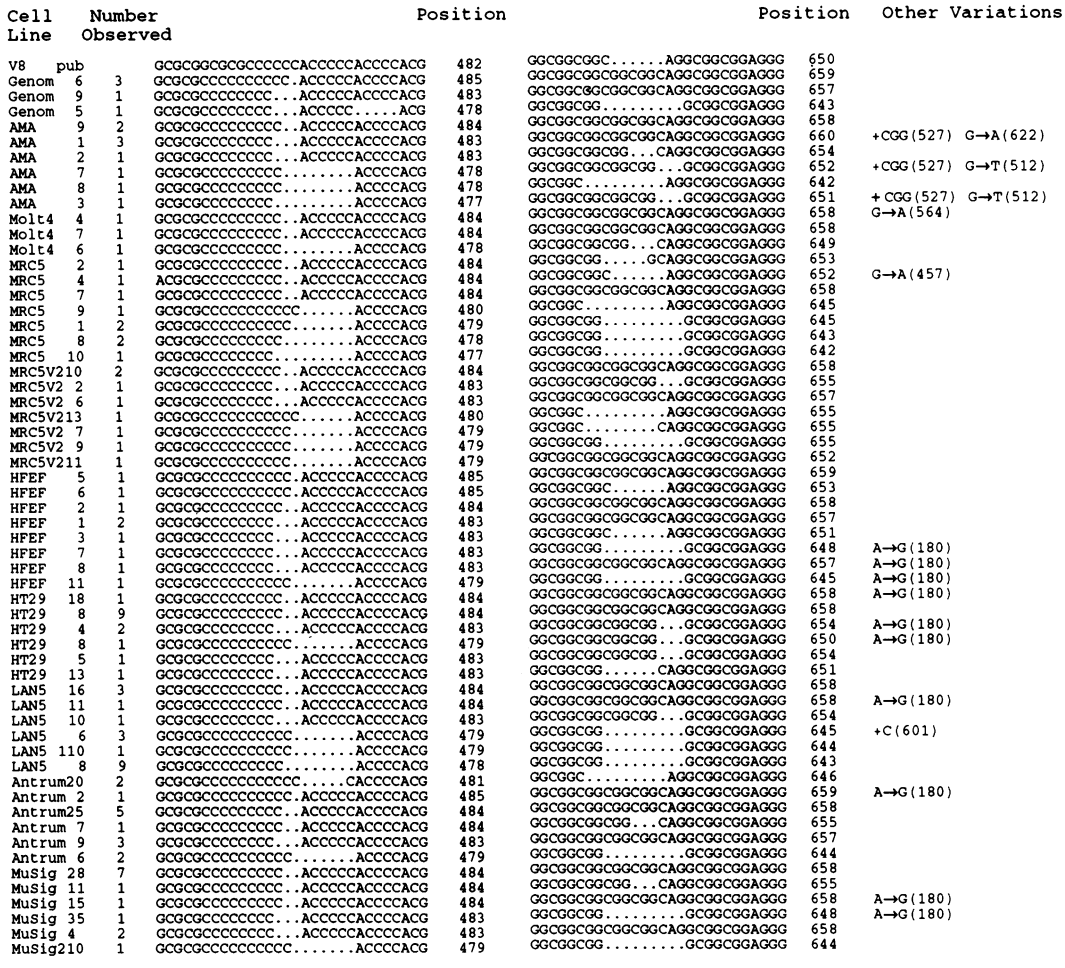
Cell line specific variation

We investigated whether there was any cell line-specificity in the variation by sequencing V8 segments from different cell lines and tissues. Although the number of sequenced V8 segments is inadequate to make any definite conclusions, we have found some potential candidates for cell line specific variation. Among these is an insertion of three nucleotides (CGG) at position 527 in three out of nine V8 segments from AMA cells (Fig. 4). This insertion has not been detected in V8 segments from other cell lines. Within the two hot spots, we do not find any obvious correlation between cell line and the size and/or combination of deletions (Fig. 6). However, there seems to be a difference between the proportion of V8 segments with large and small deletions. Thus, of 18 sequenced V8 segments from LAN-5 cells, 13 have a large deletion in both hot spots whereas in HT29 cells, 10 out of 15 sequences have only a small or no deletion around position 470 and no deletions around position 640 (Fig. 6).

This suggests that rRNA genes are expressed differentially in different cell lines which would agree with earlier results which show that different rRNA gene clusters are active in different human tissues (10). The mechanism for such differential activity could involve an inactivation of some of the rRNA gene clusters by changes in the chromatin structure. This is supported by the known presence of an active and inactive chromatin structure in rRNA genes which are maintained throughout the cell cycle (24).

Secondary structure

Determination of the secondary structure of the V8 segment is difficult since conventional phylogenetic comparisons cannot be



Cell Line	Number Observed	Position		Position		Other Variations
V8 pub		GCGCGGCGCGCCCCCCACCCCCACCCCACG	482	GGCGGCGGC.....AGGCGGCGGAGGG	650	
Genom 6	3	GCGCGCCCCCCCCCC.ACCCCCACCCCACG	485	GGCGGCGGCGGCGGCAGGCGGCGGAGGG	659	
Genom 9	1	GCGCGCCCCCCCCC...ACCCCCACCCCACG	483	GGCGGCGGCGGCGGCAGGCGGCGGAGGG	657	
Genom 5	1	GCGCGCCCCCCCCC...ACCCCC.....ACG	478	GGCGGCGG.....GCGGCGGAGGG	643	
AMA 9	2	GCGCGCCCCCCCCC..ACCCCCACCCCACG	484	GGCGGCGGCGGCGGCAGGCGGCGGAGGG	658	
AMA 1	3	GCGCGCCCCCCCCC...ACCCCCACCCCACG	483	GGCGGCGGCGGCGGCAGGCGGCGGAGGG	660	+CGG(527) G→A(622)
AMA 2	1	GCGCGCCCCCCCCC...ACCCCCACCCCACG	483	GGCGGCGGCGG...CAGGCGGCGGAGGG	654	
AMA 7	1	GCGCGCCCCCCCCC.....ACCCCACG	478	GGCGGCGGCGGCGG...GCGGCGGAGGG	652	+CGG(527) G→T(512)
AMA 8	1	GCGCGCCCCCCCCC.....ACCCCACG	478	GGCGGC.....AGGCGGCGGAGGG	642	
AMA 3	1	GCGCGCCCCCCCC.....ACCCCACG	477	GGCGGCGGCGGCGG...GCGGCGGAGGG	651	+CGG(527) G→T(512)
Molt4 4	1	GCGCGCCCCCCCCC..ACCCCCACCCCACG	484	GGCGGCGGCGGCGGCAGGCGGCGGAGGG	658	G→A(564)
Molt4 7	1	GCGCGCCCCCCCCC..ACCCCCACCCCACG	484	GGCGGCGGCGGCGGCAGGCGGCGGAGGG	658	
Molt4 6	1	GCGCGCCCCCCCCC.....ACCCCACG	478	GGCGGCGGCGG...CAGGCGGCGGAGGG	649	
MRC5 2	1	GCGCGCCCCCCCCC..ACCCCCACCCCACG	484	GGCGGCGG.....GCAGGCGGCGGAGGG	653	
MRC5 4	1	ACGCGCCCCCCCCC..ACCCCCACCCCACG	484	GGCGGCGGC.....AGGCGGCGGAGGG	652	G→A(457)
MRC5 7	1	GCGCGCCCCCCCCC..ACCCCCACCCCACG	484	GGCGGCGGCGGCGGCAGGCGGCGGAGGG	658	
MRC5 9	1	GCGCGCCCCCCCCCCC.....ACCCCACG	480	GGCGGC.....AGGCGGCGGAGGG	645	
MRC5 1	2	GCGCGCCCCCCCCCC.....ACCCCACG	479	GGCGGCGG.....GCGGCGGAGGG	645	
MRC5 8	2	GCGCGCCCCCCCCC.....ACCCCACG	478	GGCGGCGG.....GCGGCGGAGGG	643	
MRC5 10	1	GCGCGCCCCCCCC.....ACCCCACG	477	GGCGGCGG.....GCGGCGGAGGG	642	
MRC5V210	2	GCGCGCCCCCCCC..ACCCCCACCCCACG	484	GGCGGCGGCGGCGGCAGGCGGCGGAGGG	658	
MRC5V2 2	1	GCGCGCCCCCCCC...ACCCCCACCCCACG	483	GGCGGCGGCGGCGG...GCGGCGGAGGG	655	
MRC5V2 6	1	GCGCGCCCCCCCC...ACCCCCACCCCACG	483	GGCGGCGGCGGCGGCAGGCGGCGGAGGG	657	
MRC5V213	1	GCGCGCCCCCCCCCCC.....ACCCCACG	480	GGCGGC.....AGGCGGCGGAGGG	655	
MRC5V2 7	1	GCGCGCCCCCCCCC.....ACCCCACG	479	GGCGGC.....CAGGCGGCGGAGGG	655	
MRC5V2 9	1	GCGCGCCCCCCCCCC.....ACCCCACG	479	GGCGGCGG.....GCGGCGGAGGG	655	
MRC5V211	1	GCGCGCCCCCCCCC.....ACCCCACG	479	GGCGGCGGCGGCGGCAGGCGGCGGAGGG	652	
HFEF 5	1	GCGCGCCCCCCCCCC.ACCCCCACCCCACG	485	GGCGGCGGCGGCGGCAGGCGGCGGAGGG	659	
HFEF 6	1	GCGCGCCCCCCCCC.ACCCCCACCCCACG	485	GGCGGCGGC.....AGGCGGCGGAGGG	653	
HFEF 2	1	GCGCGCCCCCCCCC..ACCCCCACCCCACG	484	GGCGGCGGCGGCGGCAGGCGGCGGAGGG	658	
HFEF 1	2	GCGCGCCCCCCCC...ACCCCCACCCCACG	483	GGCGGCGGCGGCGGCAGGCGGCGGAGGG	657	
HFEF 3	1	GCGCGCCCCCCCC...ACCCCCACCCCACG	483	GGCGGCGGC.....AGGCGGCGGAGGG	651	
HFEF 7	1	GCGCGCCCCCCCC...ACCCCCACCCCACG	483	GGCGGCGG.....GCGGCGGAGGG	648	A→G(180)
HFEF 8	1	GCGCGCCCCCCCC...ACCCCCACCCCACG	483	GGCGGCGGCGGCGGCAGGCGGCGGAGGG	657	A→G(180)
HFEF 11	1	GCGCGCCCCCCCCCC.....ACCCCACG	479	GGCGGCGG.....GCGGCGGAGGG	645	A→G(180)
HT29 18	1	GCGCGCCCCCCCC..ACCCCCACCCCACG	484	GGCGGCGGCGGCGGCAGGCGGCGGAGGG	658	A→G(180)
HT29 8	9	GCGCGCCCCCCCC..ACCCCCACCCCACG	484	GGCGGCGGCGGCGGCAGGCGGCGGAGGG	658	
HT29 4	2	GCGCGCCCCCCCC...ACCCCCACCCCACG	483	GGCGGCGGCGGCGG...GCGGCGGAGGG	654	A→G(180)
HT29 8	1	GCGCGCCCCCCCCCC.....ACCCCACG	479	GGCGGCGGCGGCGG...GCGGCGGAGGG	650	A→G(180)
HT29 5	1	GCGCGCCCCCCCC...ACCCCCACCCCACG	483	GGCGGCGGCGGCGG...GCGGCGGAGGG	654	
HT29 13	1	GCGCGCCCCCCCC...ACCCCCACCCCACG	483	GGCGGCGG.....CAGGCGGCGGAGGG	651	
LAN5 16	3	GCGCGCCCCCCCCC..ACCCCCACCCCACG	484	GGCGGCGGCGGCGGCAGGCGGCGGAGGG	658	
LAN5 11	1	GCGCGCCCCCCCCC..ACCCCCACCCCACG	484	GGCGGCGGCGGCGGCAGGCGGCGGAGGG	658	A→G(180)
LAN5 10	1	GCGCGCCCCCCCC...ACCCCCACCCCACG	483	GGCGGCGGCGGCGG...GCGGCGGAGGG	654	
LAN5 6	3	GCGCGCCCCCCCCC.....ACCCCACG	479	GGCGGCGG.....GCGGCGGAGGG	645	+C(601)
LAN5 110	1	GCGCGCCCCCCCCCC.....ACCCCACG	479	GGCGGCGG.....GCGGCGGAGGG	644	
LAN5 8	9	GCGCGCCCCCCCC.....ACCCCACG	478	GGCGGCGG.....GCGGCGGAGGG	643	
Antrum20	2	GCGCGCCCCCCCCCCC.....CACCCCACG	481	GGCGGC.....AGGCGGCGGAGGG	646	
Antrum 2	1	GCGCGCCCCCCCCC.ACCCCCACCCCACG	485	GGCGGCGGCGGCGGCAGGCGGCGGAGGG	659	A→G(180)
Antrum25	5	GCGCGCCCCCCCCC..ACCCCCACCCCACG	484	GGCGGCGGCGGCGGCAGGCGGCGGAGGG	658	
Antrum 7	1	GCGCGCCCCCCCC..ACCCCCACCCCACG	484	GGCGGCGGCGG...CAGGCGGCGGAGGG	655	
Antrum 9	3	GCGCGCCCCCCCC..ACCCCCACCCCACG	483	GGCGGCGGCGGCGGCAGGCGGCGGAGGG	657	
Antrum 6	2	GCGCGCCCCCCCCC.....ACCCCACG	479	GGCGGCGG.....GCGGCGGAGGG	644	
MuSig 28	7	GCGCGCCCCCCCC..ACCCCCACCCCACG	484	GGCGGCGGCGGCGGCAGGCGGCGGAGGG	658	
MuSig 11	1	GCGCGCCCCCCCC..ACCCCCACCCCACG	484	GGCGGCGGCGG...CAGGCGGCGGAGGG	655	
MuSig 15	1	GCGCGCCCCCCCC..ACCCCCACCCCACG	484	GGCGGCGGCGGCGGCAGGCGGCGGAGGG	658	A→G(180)
MuSig 35	1	GCGCGCCCCCCCC...ACCCCCACCCCACG	483	GGCGGCGG.....GCGGCGGAGGG	648	A→G(180)
MuSig 4	2	GCGCGCCCCCCCC...ACCCCCACCCCACG	483	GGCGGCGGCGGCGGCAGGCGGCGGAGGG	658	
MuSig210	1	GCGCGCCCCCCCCCC.....ACCCCACG	479	GGCGGCGG.....GCGGCGGAGGG	644	

Figure 6. Alignment of the sequences of the two 'hot spots' for variation arranged according to cell type. Number observed refer to how many V8 segments of that type we have sequenced from a given cell line or tissue sample. Cell lines are as in the legend to Fig. 2.

applied because the V8 segment sequences from different organisms are highly conserved in some regions and very divergent in others. Thus, possible compensating base changes cannot be precisely located. However, vertebrate V8 segments can all be folded into a T-like structure, with a short helix I and two long helixes (II & III) (Fig. 4) (3, 4, 25). None of the secondary structure models are identical, although the overall topography is similar among the models and roughly similar structures have been found in invertebrates (26). The T-like structure is supported by electron microscopy data, where evidence for two long symmetrical helices at the position where the V8 segment is located was presented (7). Moreover, our secondary structure probing results are in very good agreement with the secondary structure presented in Fig. 4.

Phylogenetic comparisons

Apparently, there is a correlation between the complexity of the organism and the size of the expansion segments (Table 1). The largest expansion segments are found among mammals with a continuous decrease in size when moving 'down' in evolution (Table 1). An alignment of human and other vertebrate V8 segments shows that all are highly conserved in helix I and in the 'inner' part of helix II. The region towards the apex loop of helix II is missing in *Xenopus* and, in the same region, there are small deletions in the rodent rRNAs. The most variable region is helix III. Most of this helix is absent in *Xenopus*, whereas the rodent rRNAs are almost identical to the human rRNA except in the neighbourhood of the two hot spots, where the similarity disappears. Most of the region from the hot spots toward the apex loop is missing in both mouse and rat.

Origin and function of the expansion segments

The function of the expansion segments remains unknown. They could function at three levels: 1) at the DNA level as hot spots for recombination, which would tend to maintain a homogeneous rRNA gene population; 2) at the transcription level where they could function as transcriptional enhancers or anti terminators, or 3) at the RNA level where some of the eukaryote-specific features of the ribosome could arise either through the expansion segments or through proteins which bind to them. The variation found within the expansion segments could indicate that they are hot spots for recombination, since slippage of a few nucleotides in the recombination process would lead to the accumulation of variants. However, this is contradicted by the conservation in the secondary structure of the RNA, a conservation shown both in this study and in an earlier study of expansion segments from insects where a few compensating base changes were found (26). A possible mechanism for generating the variation and at the same

time maintaining the secondary structure has been proposed by Hancock and Dover (9). They suggest a mechanism called 'compensatory slippage' in which replicational slippage at one position will be compensated by slippage at another position, resulting in the maintenance of the overall secondary structure. However, this mechanism is similar to the mechanism operating on the conserved core of the rRNA molecule, where the secondary structure is conserved despite very large variations in the primary sequence. This indicate that the expansion segments should be regarded as integral parts of the rRNA molecules and suggest that they function at the RNA level. Thus their function is likely to be related to the function of eukaryotic ribosomes.

Finally, the sequence variation in the 28S rRNA population must originate from variation among the rRNA genes and can thus be utilized as a tool in the investigation of the organization of rRNA operon clusters, particularly in view of efforts to physically map the human genome. In addition, mapping and sequencing of rRNA operon clusters may answer some important questions concerning the evolution of rRNA genes: In particular, do some rRNA operons get amplified in some cell types? This could explain the apparent differences in the 28S rRNA populations among different cell lines.

ACKNOWLEDGEMENTS

We thank Jan Christiansen for providing some of the RNA samples and Roger A.Garrett, Jan Christiansen, Niels Larsen, Julio E.Celis and Jan Egebjerg for stimulating discussions and RAG for help with the manuscript. J.Egebjerg participated in the probing of the secondary structure. The work has been supported by the Danish Cancer Foundation and the Danish Centre for Human Genome Research.

REFERENCES

1. Gonzalez, I. L., Gorski, J. L., Campen, T. J., Dorney, D.J., Erickson, J.M., Sylvester, J.E. & Schmickel, R.D. (1985). Proc. Nat. Acad. Sci. 82, 7666–7670.
2. Maden, B. E. H., Dent, C. L., Farrell, T. E., Garde, J., McCallum, F. S. & Wakeman, J. A. (1987). Biochem. J. 246, 519–527.
3. Clark, C. G., Tague, B. W., Ware, V. C. & Gerbi, S. A. (1984). Nucl. Acids Res. 12, 6197–6220.
4. Gorski, J. L., Gonzalez, I. L. & Schmickel, R. D. (1987). J. mol. Evol. 24, 236–251.
5. Hassouna, N., Michot, B. & Bachellerie, J. P. (1984). Nucl. Acids Res. 12, 3563–3583.
6. Brosius, J., Dull, T. J. & Noller, H. F. (1980). Proc. Nat. Acad. Sci., U.S.A. 77, 201–204.
7. Wakeman, J. A. & Maden, E. H. (1989). Biochem. J. 258, 49–56.
8. Qu, L.-H., Nicoloso, M. & Bachellerie, J.-P. (1991) Nucl. Acids Res. 19, 1015–1019.
9. Hancock, J. M. & Dover, G. A. (1990). Nucl. Acids Res. 18, 5949–5954.
10. de Capoa, A., Marlekaj, P., Baldini, A., Rocchi, M. and Archidiacono, N. (1985). Hum. Genet. 69, 212–217.
11. Gunderson, J. H., Sogin, M. L., Wollett, G., Hollingdale, M., de la Cruz, V. F., Waters, A. P. and McCutchan, T. F. (1987). Science 238, 933–937.
12. Waters, A. P., Syin, C. and McCutchan, T. F. (1989). Nature 342, 438–440.
13. Gonzalez, I. L., Sylvester, J. E. & Schmickel, R. D. (1988). Nucl. Acids Res. 16, 10213–10224.
14. Sambrook, J., Fritsch, E. F., and Maniatis, T. (1989) Molecular cloning, a laboratory manual, second edition. Cold Spring Harbour Laboratory Press.
15. Norrander, J., Kempe, T. & Messing, J. (1983). Gene, 26, 101–114.
16. Sanger, F., Nicklen, S. & Coulsen, A. R. (1977). Proc. Natl. Acad. Sci., U.S.A. 74, 5463–5467.
17. Ansorge, W. & Labeit, S. (1984). J. Biochem. Biophys. Methods, 9, 33–47.
18. Gubler. U. (1988). Nucleic Acids Res. 16, 2726.
19. Egebjerg, J., Leffers, H., Christensen, A., Andersen, H. D. & Garrett, R. A. (1987). J. Mol. Biol. 196, 125–136.
20. Thirup, S. and Larsen N. E. (1990). Proteins: Structure, function and genetics 7, 291–295.
21. Devereux, J., Haeberli, P. & Smities, O.(1984). Nucl. Acids Res. 12. 387–395.
22. Bross, K. & Krone, W. (1972). Humangenetik 14, 137–145.
23. Schmickel, R. D. (1973). Pediat. Res. 7, 5–13.
24. Conconi, A., Widmer, R. M., Koller, T. & Sogo, J. M. (1989). Cell 57, 753–761.
25. Michot, B., Hassouna, N. & Bachellerie, J.-P. (1984). Nucl. Acids Res. 12, 4259–4279.
26. Linares, A. R., Hancock, J. M. & Dover, G. A. (1991). J. Mol. Biol. 219. 381–390.
27. Chan, Y. L., Olvera, J. & Wool, I. G. (1984). Nucl. Acids. Res. 11, 7819–7831.
28. Ware, V. C., Tague, B. W., Clark, C. G., Gourse, R. L., Brand, R. C. & Gerbi, S.A. (1983). Nucl. Acids. Res. 11, 7795–7817.
29. Tautz D., Hancock J.M., Webb D.A., Tautz C., Dover G.A. (1988). Mol. Biol. Evol. 5, 366–376.
30. Degnan, B. M., Yan, J., Hawkins, C. J. & Lavin, M. F. (1990). Nucl. Acids Res. 18, 7063–7070.
31. Ellis, R. E., Sulston, J.E. & Coulson, A. R. (1986). Nucl. Acids Res. 14, 2345–2364.
32. Engberg, J. & Nielsen, H. (1990). Nucl. Acids Res. 18, 6915–6919.
33. Sloof, p., Van den Burg, J., Voogd, A., Benne, R., Agostinelli, M., Borst, P., Gutell, R. & Noller, H. (1985). Nucl. Acids. Res. 13, 4171–4190.
34. Spencer, D. F., Collings, J. C., Schnare, M. N. & Gray, M. W. (1987). EMBO J. 6, 1063–1071.
35. Lenaers, G. Maroteaux, L., Michot, B. & Herzog, M. (1989). J. Mol. Evol. 29, 40–51.
36. Koloscha, V. O. & Fodor, I. I. (1986). EMBL database accession number: X05910.
37. Unfried, I. & Gruendler, P. (1990). Nucl. Acids Res. 18, 4011.
38. Rathgeber, J. & Capesius, I. (1990). Nucl. Acids Res. 18, 1288.
39. Kiss, T., Kiss, M. & Solymosy, F. (1989). Nucl. Acids Res. 17, 796.
40. Takaiwa, F. Oono, K., Iida, Y. & Sugiura, M. (1985). Gene 37, 255–259.
41. Ozaki, T., Hoshikawa, Y., Iida, Y. & Iwabuchi, M. (1984). Nucl. Acids Res. 12, 4171–4184.
42. Otsuka, T., Nomiyama, H., Yoshida, H., Kukita, T., Kuhara, S. & Sakaki, Y. (1983). Proc. Natl. Acad. Sci. USA. 80, 3163–3167.
43. Ji, G. E. & Orlowski, M. (1990). Curr. Genet. 17, 499–506.
44. Georgiev, O. I., Nikolaev, H., Hadjiolov, A. A., Skryabin, K. G., Zakharyev, V. M. & Bayev, A. A. (1981). Nucl. Acids Res. 9, 6953–6958.