Sequence and organization of the human mitochondrial genome

Article · April 1981

DOI: 10.1038/290457a0

14 authors, including:

Jacques Drouin
Institut de recherches cliniques de Montréal
233 PUBLICATIONS **21,800** CITATIONS
SEE PROFILE

Ian C Eperon
University of Leicester
86 PUBLICATIONS **14,404** CITATIONS
SEE PROFILE

Peter H. Schreier
University of Cologne
67 PUBLICATIONS **13,177** CITATIONS
SEE PROFILE

the loops will continuously change, and the structure of the ring, although periodic, will be highly complex and variable.

This theory cannot account for the third, diffuse ring. Perhaps this ring consists of small (micrometre-sized) particles which are generated in the main ring by erosion due to sputtering and meteoroid impacts and then removed by electromagnetic forces.

This article was inspired by comments from P. Goldreich and S. Tremaine during the Voyage 1 Saturn encounter. I thank C. D. Murray, R. Farouki and T. Gold for useful contributions. The data are from early Voyager Bulletin Mission Status Reports and I thank the Voyager Imaging Team (leader B. A. Smith) for their generosity in releasing this information before their own publication. This research was supported by NSF grant AST 79164-74-A01.

Received 16 December 1980; accepted 18 February 1981.

1. Goldreich, P. & Tremaine, S. *Nature* **277**, 97-99 (1979).
2. *Voyager Bull., Mission Status Rep.* No. 56 (31 October 1980).
3. Smith, B. A. *et al. Science* (in the press).
4. Lin, D. N. C. & Papaloizou, J. *Mon. Not. R. astr. Soc.* **186**, 799-812 (1979).
5. Julian, W. H. & Toomre, A. *Astrophys. J.* **146**, 810-830 (1966).
6. Greenberg, R. *Astr. J.* **78**, 338-346 (1973).
7. Dermott, S. F., Farouki, R. & Murray, C. D. (in preparation).

Sequence and organization of the human mitochondrial genome

S. Anderson, A. T. Bankier, B. G. Barrell, M. H. L. de Bruijn, A. R. Coulson, J. Drouin*, I. C. Eperon, D. P. Nierlich*, B. A. Roe*, F. Sanger, P. H. Schreier*, A. J. H. Smith, R. Staden & I. G. Young*

MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK

The complete sequence of the 16,569-base pair human mitochondrial genome is presented. The genes for the 12S and 16S rRNAs, 22 tRNAs, cytochrome c oxidase subunits I, II and III, ATPase subunit 6, cytochrome b and eight other predicted protein coding genes have been located. The sequence shows extreme economy in that the genes have none or only a few noncoding bases between them, and in many cases the termination codons are not coded in the DNA but are created post-transcriptionally by polyadenylation of the mRNAs.

MITOCHONDRIA have a separate autonomously replicating DNA genome (for a review see ref. 1). Although most of the proteins present in the mitochondrion are encoded by nuclear DNA, a few are coded for by mitochondrial (mt) DNA and are synthesized by the separate mt translation system. The rRNAs and tRNAs of this system are mitochondrially encoded and it is likely that the other components, although nuclear coded, are distinct from those in the cytoplasmic translation system. The enzymes specific for the unique mitochondrial transcription and replication processes are probably also nuclear coded. We have recently shown that mammalian mitochondria have a different genetic code[2,3] as have other mitochondria systems that have been studied[4-7], and a different decoding mechanism[7-9].

Replication of mammalian mtDNA proceeds by initiation of heavy (H) strand synthesis at a specific origin[10,11] resulting in the formation of a displacement (D) loop with a newly synthesized H strand of ~680 bases known as 7S DNA. In mouse cells a large proportion of mtDNA exists in the D-loop form but only a few of these nascent H strands grow to full length. Initiation of light (L) strand synthesis is at a specific origin[12] and does not occur until this region has been exposed by H-strand synthesis.

Transcription of mammalian mtDNA is also unique in that both strands seem to be completely transcribed from promoters situated in the D-loop region[13,14]. These primary transcripts are then processed to give the 12S and 16S rRNAs, tRNAs and a number of presumptive mRNAs, which are not capped but are polyadenylated[15]. The small size of the ribosomal RNAs, the chloramphenicol sensitivity of mitochondrial protein synthesis and the use of formyl-methionine in initiation of translation have led to the suggestion that the mitochondrial genetic system arose by endosymbiosis of primitive bacteria within a host cell. The sequence of the rRNA genes, however, shows only low homology with present-day prokaryotic rRNAs[16].

To understand this genetic system more fully we have determined the complete DNA sequence of both the human and the bovine mitochondrial genomes. We report here the sequence of human mtDNA and an interpretation of this sequence in terms of its coding capacity, gene arrangement and gene expression. The bovine sequence will be published elsewhere.

In the accompanying articles Ojala *et al.*[17] and Montoya *et al.*[18] report the sequence analysis of the stable polyadenylated RNAs. The comparison of the DNA and RNA sequences reveals the unique nature of the precise transcript processing in mammalian mitochondria.

Organization of the genome

The DNA sequence is shown in Fig. 1 along with the predicted coding sequences. This is presented schematically in Fig. 2. The identification of the genes and their precise location in the DNA sequence are based on several unique features, the evidence for which is discussed below and in the accompanying articles[17,18]. The sequence shows extreme economy of organization in that there are none or very few noncoding bases between adjacent genes except for the D-loop region where we have been unable to assign any coding function. The sequence shown is that of the L strand which we define as the main coding strand containing the sense sequence of the rRNAs and most of the tRNAs and mRNAs. Thus, these RNAs are transcribed from (and therefore hybridize to) the H strand.

Origins of DNA replication

The 5′ end of the major 7S DNA species found in the D-loop of HeLa cell mtDNA[10] corresponds to nucleotide 191 of the DNA sequence. Other human 7S DNA species having different 5′ ends have also been reported[11,19,20], but the precise location of these termini in the sequence presented here is not known. The size of the 7S DNA (≤680 nucleotides) and mapping data[10,19,20]

*Present addresses: Department of Biochemistry and Biophysics, School of Medicine, University of California, San Francisco, California 94143, USA (J.D.); Department of Microbiology, College of Letters and Science, University of California, Los Angeles, California 90024, USA (D.P.N.); Chemistry Department, Kent State University, Kent, Ohio 44242, USA (B.A.R.); Institut für Genetik der Universität, 5 Koln 41, Weyertal 121, FRG (P.H.S.); Department of Biochemistry, John Curtin School of Medical Research, Australian National University, PO Box 334, Canberra City, ACT 2601, Australia (I.G.Y.).

0028-0836/81/150457—09$01.00

Table 1 Human mitochondrial genetic code

Phe	UUU	77	Ser	UCU	32	Tyr	UAU	46	Cys	UGU	5
	UUC	141		UCC	99		UAC	89		UGC	17
Leu	UUA	73		UCA	83	Ter	UAA	–	Trp	UGA	93
	UUG	16		UCG	7		UAG	–		UGG	11
Leu	CUU	65	Pro	CCU	41	His	CAU	18	Arg	CGU	7
	CUC	167		CCC	119		CAC	79		CGC	25
	CUA	276		CCA	52	Gln	CAA	81		CGA	29
	CUG	45		CCG	7		CAG	9		CGG	2
Ile	AUU	125	Thr	ACU	51	Asn	AAU	33	Ser	AGU	14
	AUC	196		ACC	155		AAC	131		AGC	39
Met	AUA	167		ACA	133	Lys	AAA	85	Ter	AGA	–
	AUG	40		ACG	10		AAG	10		AGG	–
Val	GUU	30	Ala	GCU	43	Asp	GAU	15	Gly	GGU	24
	GUC	49		GCC	124		GAC	51		GGC	88
	GUA	70		GCA	80	Glu	GAA	64		GGA	67
	GUG	18		GCG	8		GAG	24		GGG	34

The genetic code of human mtDNA differs from the universal code in that UGA codes for tryptophan and not termination, AUA codes for methionine not isoleucine, and AGA and AGG are termination rather than arginine codons. In addition, AUA and possibly AUU are initiation codons as well as AUG. Boxes of four codons are each read by a single tRNA with U in the first position of the anticodon; boxes of two codons are read by tRNAs with G:U wobble anticodons. The number of methionyl tRNAs and their codon response are unclear and discussed in the text. Also shown are the total number of each amino acid codon found in the genes and predicted genes shown in Figs 1 and 2. Codons ending in A and C predominate (36.37 and 40.93%, respectively) and those ending in U and G are used less frequently (16.33 and 6.36%, respectively).

indicate that the 7S DNA, and presumably the D-loop, do not extend into the $tRNA^{Pro}$ gene. Although H-strand synthesis during replication initiates in the D-loop region, data from mouse cell mitochondria indicate that most of the 7S DNA molecules do not participate in DNA replication but rather are synthesized and then rapidly lost from the D-loop[21].

The ~1,100-base pair region of the human mitochondrial genome between the $tRNA^{Pro}$ and $tRNA^{Phe}$ genes contains no large open reading frames in either strand. Furthermore, except for a small 7S RNA[15], no large stable polyadenylated transcript has been found to map in this region. Unless RNA splicing is invoked it is extremely unlikely that this part of the genome codes for a major protein species. In the region covered by the D-loop, several blocks of nucleotide sequence homology exist between the human, bovine (S.A. *et al.*, unpublished results) and rat mitochondrial genomes[22]. However, the DNA separating these blocks is rather variable in length and exhibits no significant homology. The region between the 5′ end of the 7S DNA and the $tRNA^{Phe}$ gene seems to be one of the least conserved regions of the entire genome. This was initially detected by electron microscopy of heteroduplexes of sheep and goat mtDNAs[23] and has been confirmed by the DNA sequence comparisons. Very little sequence homology between the human, rat[22] or bovine mitochondrial genomes can be detected in this region. The size of the region also seems to vary much more than do intergenic spaces in other parts of the mtDNA, being 385 nucleotides in the human placenta, 291 nucleotides in rat liver and (based on alignment with the human and rat sequence) ~180 nucleotides in beef heart. This region may contain control signals for transcription of the mitochondrial genome.

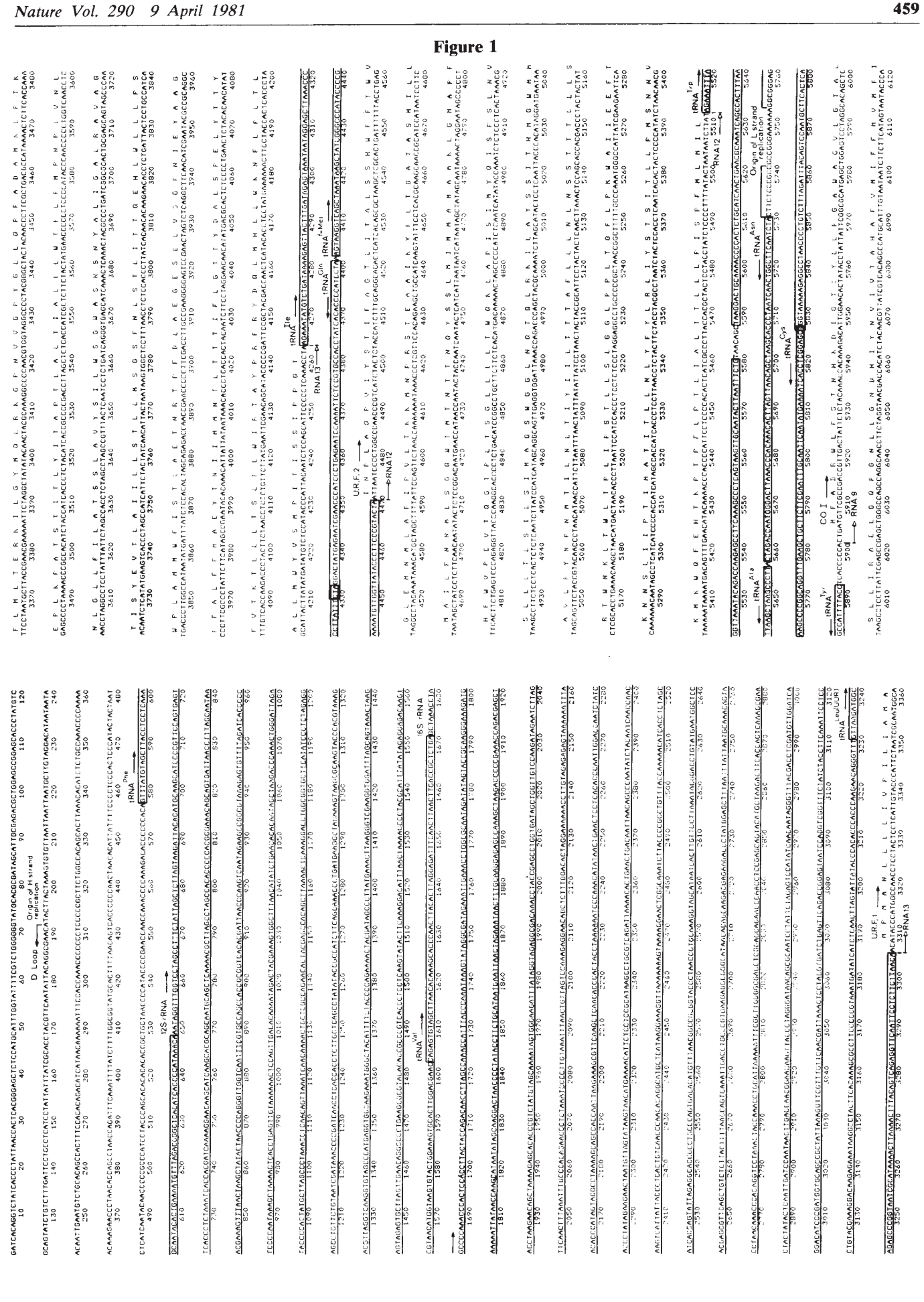
L-strand synthesis during mtDNA replication initiates two-thirds of the way around the genome from the region of H-strand synthesis[24]. In mouse cell mitochondria the DNA sequence immediately surrounding the L-strand origin is a simple inverted repeat that may form a hairpin loop structure when the L-strand origin is exposed by H-strand displacement synthesis during DNA replication[12]. An analogous and highly homologous presumptive L-strand origin exists in the human mtDNA sequence (nucleotides 5,730–5,763) within a cluster of five tRNA genes. Interestingly, the differences between the human and mouse inverted repeats occur symmetrically, so that if secondary structure did exist, full base-pairing in the stem region would be maintained.

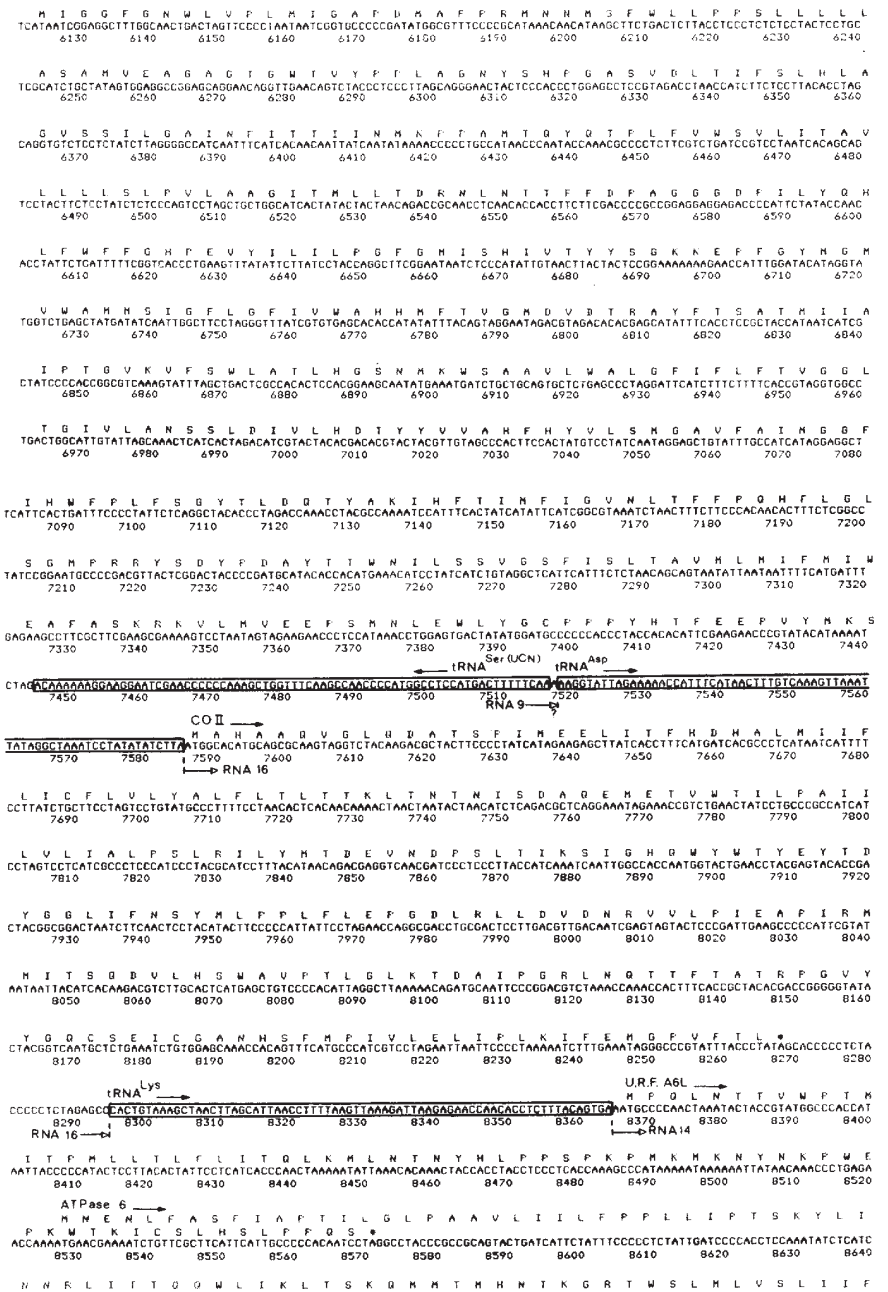
Ribosomal RNA genes

Eperon *et al.*[16] were able to align the DNA sequence with the sequences of the 5′ ends [15,25] of the 12S and 16S rRNAs and to show that there were no noncoding bases between the $tRNA^{Phe}$ gene and the 12S rRNA gene, and between the $tRNA^{Val}$ gene

Fig. 1 The DNA sequence of human mitochondria (pages 459–461). The sequence of the L strand is numbered arbitrarily from the *Mbo*I-5/7 boundary[56] in the D-loop region. The sequence is displayed using the program TRANMT of Staden (unpublished) and shows the position of the genes for the tRNAs (boxed) and the rRNAs (underlined). The unidentified reading frames (URF 1–6) and identified protein coding genes are translated into the one-letter amino acid code (termination codons are denoted by an asterisk). The genes for cytochrome *c* oxidase subunits I–III are abbreviated COI–III and cyt *b* denotes the gene for cytochrome *b*. Which strand contains the sense sequence of the genes is denoted by → (L strand) and ← (H strand). The ends of the presumptive mRNAs transcribed from the H strand as determined by RNA sequence analysis[17,18] are also shown below the sequence. The 5′ end of the nascent H-strand ('7S') DNA found in the D-loop is at nucleotide 191 (ref. 10). The origin of L-strand synthesis during DNA replication[12] is indicated by overlining of the possible base-pairing regions. The DNA sequence was derived mainly from a single human placenta mtDNA preparation[56] but some regions were determined on HeLa cell mtDNA. Restriction enzyme fragments of placental mtDNA were cloned in pBR322[56]. Single-stranded template DNA was prepared by the exonuclease III procedure[57] and by cloning in bacteriophage M13 (refs 58–61). The DNA sequence was determined by the chain termination method of Sanger *et al.*[62,63]. The sequence was analysed using the computer programs of Staden (refs 31, 55, 64 and unpublished). Nucleotides 10, 934–5, 14,272 and 14,365 were ambiguous and their identity assumed to be the same as in the bovine mtDNA sequence (S.A., M.H.L.B., A.R.C., I.C.E., F.S., I.G.Y., unpublished). The base composition of the L strand is 24.7% T, 31.2% C, 30.9% A and 13.1% G.

Figure 1





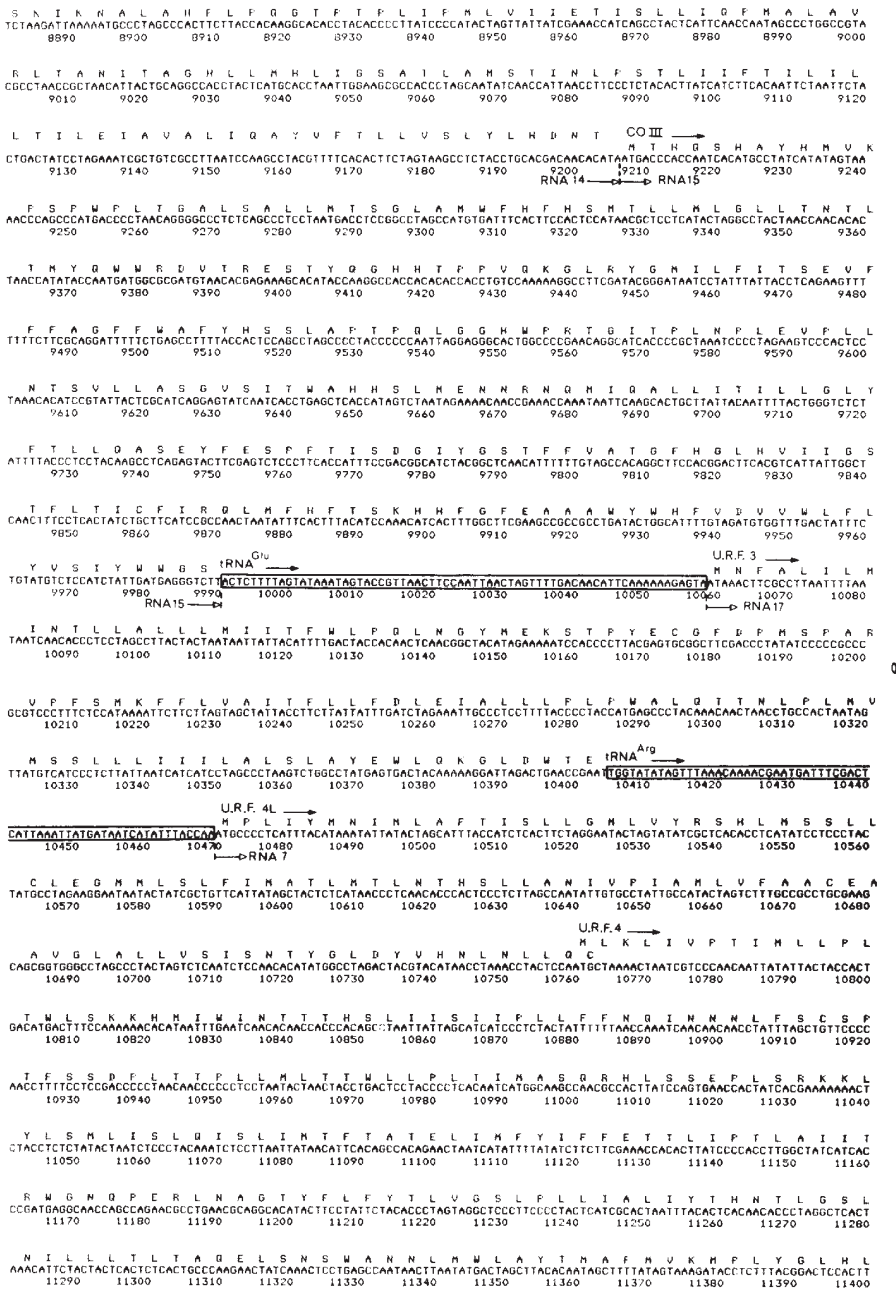
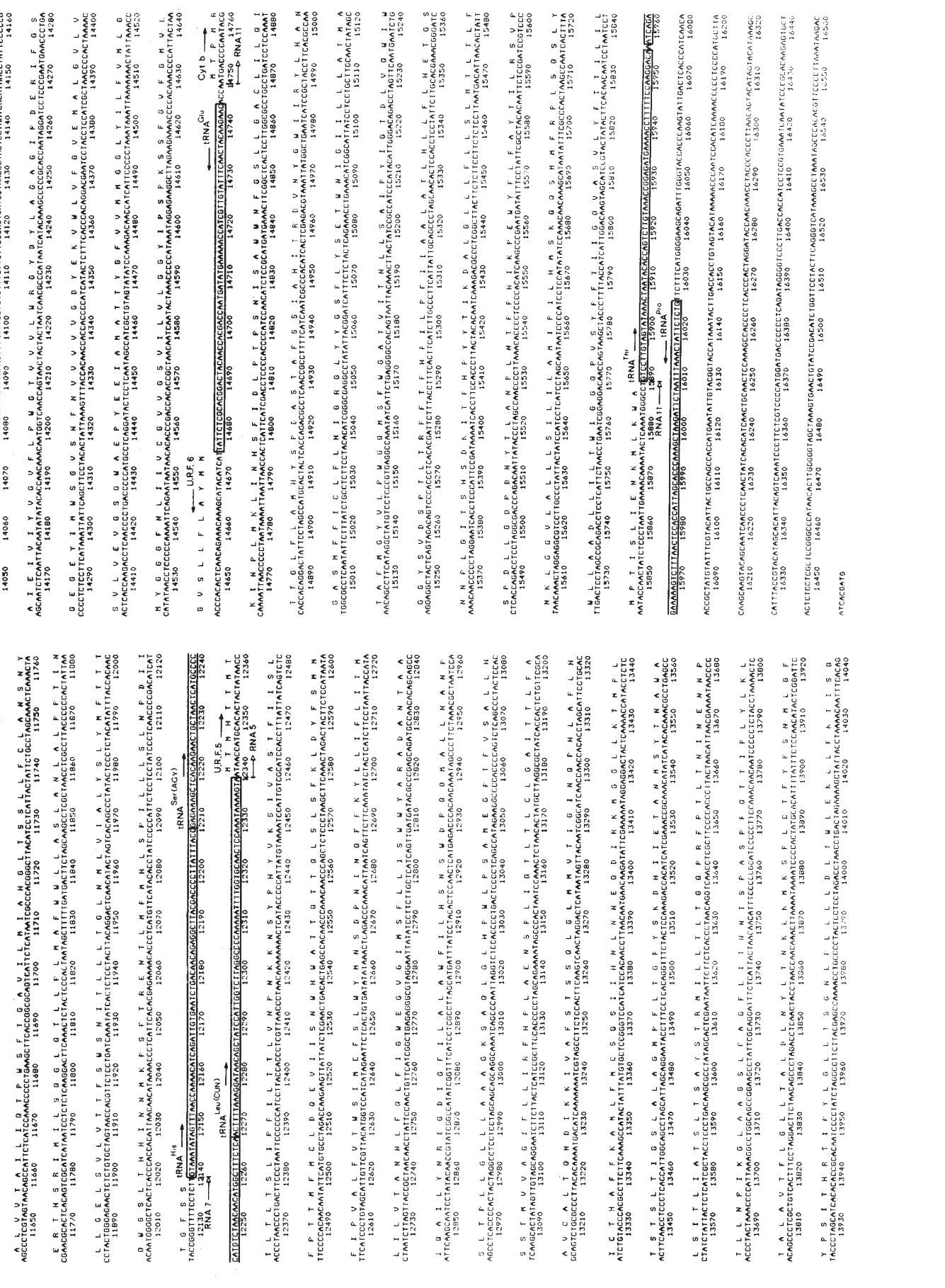


Figure 1

Figure 1



and the 16S rRNA gene. However, the 3′ ends of the rRNA genes could only be predicted to be very close to the 5′ ends of the flanking $tRNA^{Val}$ and $tRNA^{Leu}$ genes. It has since been shown that the mouse small subunit rRNA gene abuts the 3′ $tRNA^{Val}$ gene[26,27] and alignment of a hamster large subunit rRNA sequence with the murine gene sequence suggests that there are also no noncoding bases between the large subunit rRNA gene and the 3′ $tRNA^{Leu}$ gene[28]. Thus, the 12S rRNA and 16S rRNA and the flanking tRNAs in a primary transcript could be released by cleavage precisely at the 5′ and 3′ ends of the tRNAs without the need for further processing.

The high concentration of nascent transcripts from this region of the genome has led to the suggestion that there is a transcription attenuator in this region[20]. This is supported by the observation that the hamster large subunit rRNA ends and is polyadenylated at any of four nucleotides within the last five nucleotides of the gene[28], which contrasts with the precise termini of the 12S rRNA and mRNAs[29,30] which are presumed to be generated by processing of a primary transcript. It is possible that the 12S rRNA is also adenylated at the 3′ terminus after transcription and processing: a single addition has been proposed for the hamster small subunit rRNA by comparison with the murine gene sequence[30].

The sequences of the rRNA genes were interpreted to provide support for a model in which mt ribosomes did not have sequence-dependent signals for recognition of the correct initiation codon but rather initiated at the first initiation codon in the mRNA[16]. This is consistent with the known 5′-end sequences of the mRNAs[18] although not with the observation of single transcripts covering the URF A6L/ATPase 6 and URF 4L/URF 4 reading frames (Fig. 2) if the expected proteins are expressed.

The tRNA genes

Twenty-two tRNA genes have been identified in the human mtDNA sequence either by visual inspection or using the program TRNA[31]. Equivalent genes have been located in the bovine mtDNA sequence and most of these have been confirmed by direct sequencing of bovine mitochondrial tRNAs (B. A. Roe, E. Y. Chen, P. W. Armstrong and J. F. H. Wong, unpublished). We have been unable to substantiate a previous report[8] which tentatively identified a gene for $tRNA_m^{Met}$. Apart from the $tRNA_m^{Met}$, these 22 tRNAs are sufficient to read all codons using a mechanism unique to mitochondrial systems[7–9]. No mammalian mt tRNAs have been shown to be imported from the cytoplasm (ref. 32 and B. A. Roe and E. Y. Chen, unpublished results). It had previously been assumed that the minimum number of tRNAs required to read all codons would be 32 for the 'universal' genetic code using G:U or I:A/C/U wobble[33]. In human mitochondria the tRNAs for the two-codon families, that is, those genetic code boxes with two codons for one amino acid, have G:U wobble anticodons. However, only one tRNA is found for each of the four-codon families[8] (see Table 1). These tRNAs have a U in the first position of the anticodon although each must read all four codons in its respective family. This is probably accomplished by U:N wobble[8], although a two-out-of-three mechanism might operate[34]. Although the modification pattern of the U in the first position of the anticodon of mammalian mt tRNAs is now known, it is likely that a similar situation exists to that found in *Neurospora* mitochondria[7], whereby the tRNAs for the four-codon families have an unmodified U, but when the U is required for reading the two-codon families it is modified. This modification presumably prevents such tRNAs from reading codons ending in U or C.

Two species of mammalian mt methionyl tRNAs have been identified, a $tRNA_f^{Met}$ and a $tRNA_m^{Met}$ (refs 32, 35). Hybridization of these tRNAs to mtDNA showed partial additivity, indicating the existence of two separate genes[32]. In the DNA sequences we have only been able to identify one $tRNA^{Met}$ gene, presumably coding for the initiator tRNA, based on its anticodon sequence CAT and the close homology of its anticodon arm with other initiator tRNAs (see ref. 36). The tRNA anticodon sequence CAU should normally be specific for AUG but as discussed below we believe that AUA and possibly AUU are also initiation codons in mammalian mitochondria. To recognize AUA (and AUU?) as well as AUG, the C in the first position of the anticodon would probably have to be modified. The AUA-specific $tRNA^{Ile}$ of *Escherichia coli* and of bacteriophage T4 have modified C residues in the first position of the anticodon[37,38], and hence represent examples of specific recognition of AUA by a tRNA with a CAU anticodon. Although potential $tRNA_m^{Met}$-like structures are found in the human and bovine sequences, none of these is particularly convincing nor are they conserved in the corresponding sequence. Also, extensive sequence analysis of the bovine mt tRNAs has not revealed any candidate for a $tRNA_m^{Met}$ (B. A. Roe, E. Y. Chen, P. W. Armstrong and J. F. H. Wong, personal communication). If the gene exists in the mt genome it may have an unusual structure that is preventing its detection, for example, like that of $tRNA^{Ser}_{AGU/C}$ which completely lacks the D arm[39]. The other possibility is that one gene can give rise to both a $tRNA_f^{Met}$ and a $tRNA_m^{Met}$ by differential modification, although the hybridization data of Aujame and Freeman[32] suggest otherwise. We are at present trying to clarify this problem by hybridization of labelled methionyl tRNA to restriction fragments of mtDNA.

The sequences of the mammalian mt tRNAs are unusual in that all, except $tRNA^{Leu}_{UUR}$, lack some or all of the following features found in other tRNAs: (1) the universal sequence G–T–ψ–C–R–A, (2) the constant seven-base length of the 'TψC' loop which in mammalian mt tRNAs varies between three and nine bases, and (3) the constant bases A_{14}, G_{15} and $G_{18}G_{19}$ and the connections with U_8 and U_{48} (yeast $tRNA^{Phe}$ numbering system, see ref. 40). Thus, the mt tRNAs are apparently stabilized by fewer tertiary interactions. The most extreme case is that of $tRNA^{Ser}_{AGU/C}$, which lacks the D arm[39]. Also, considering the similarity of homologous cytoplasmic tRNA species from different animals it is surprising that substantial variations are found when specific tRNAs from human and bovine mitochondria are compared. A compilation and comparison of these tRNAs will be published elsewhere. These missing or different tertiary interactions may mean that the mammalian mitochondrial tRNAs have more freedom than cytoplasmic tRNAs to evolve in certain areas, perhaps reflecting a different interaction with the ribosome. The role of the tRNAs in processing of the primary transcript is discussed below.

Table 2 Characteristics of human mt genes and comparison with bovine mt genes

Gene	Protein length	Molecular weight	% Amino acid conservation	Initiation and termination codons			
				Human		Bovine	
URF1	318	35,600	75.79	ATA	*	ATG	*
URF2	347	38,900	62.82	ATT	*	ATA	*
COI	513	57,000	91.23	ATG	AGA	ATG	TAA
COII	227	25,500	72.69	ATG	TAG	ATG	TAA
URF A6L	68	7,900	51.52	ATG	TAG?	ATG	TAA?
ATPase 6	226	24,800	77.88	ATG	*	ATG	*
COIII	261	30,000	86.97	ATG	*	ATG	*
URF3	115	13,200	73.91	ATA	*	ATA	*
URF 4L	98	10,700	73.47	ATG	TAA?	ATG	TAA?
URF4	459	51,400	74.07	ATG	*	ATG	*
URF5	603	66,600	69.49	ATA	TAA	ATA	TAA
URF6	174	18,600	62.64	ATG	AGG	ATG	TAA
Cytochrome *b*	380	42,700	78.10	ATG	*	ATG	AGA

The predicted length and molecular weight of the expected products of the human mt genes and unidentified reading frames were calculated from the DNA sequence using the computer programs of Staden (ref. 55 and unpublished results). Also shown is the percentage identical amino acids conserved between the human and bovine sequences calculated using the program AACHAN (R. Staden, unpublished). The predicted initiation and termination codons of the human and bovine genes are shown; * indicates that the termination codon is created post-transcriptionally by polyadenylation (see text).

Protein coding genes

Thirteen significant reading frames have been identified in the human mtDNA sequence, of which five have been assigned to

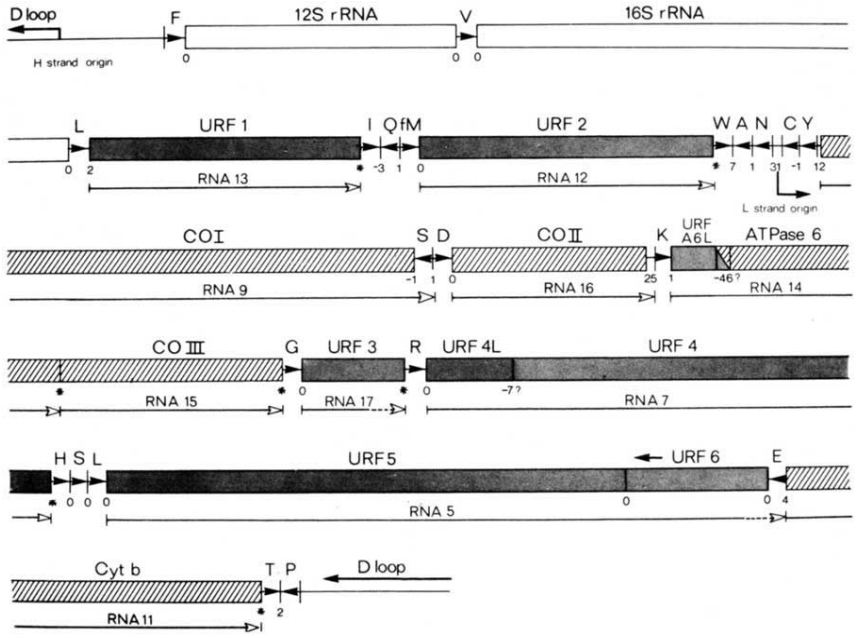


Fig. 2 Organization of the genome. A schematic representation of the sequence shown in Fig. 1. The tRNA genes are identified by the one-letter amino acid code and their sense sequence by either → (L strand) or ← (H strand). Unidentified reading frames 1–5 (URF1–5) and the identified protein genes for cytochrome *c* oxidase subunits I–III (COI–III), ATPase 6 and cytochrome *b* (cyt *b*) all have their sense sequence on the L strand; URF6, however, is H-strand coded. The number of noncoding bases between genes is given below the junctions of the genes. A minus number denotes an overlap and an asterisk indicates that there are no noncoding bases at that point and that the termination codon for the preceding gene is created in the mRNA by polyadenylation (see text). The numbered RNAs are the presumptive mRNAs transcribed from the H strand which were aligned with the DNA sequence by hybridization to restriction fragments and by sequence analysis of their 5′ and 3′ ends (see accompanying articles[17,18] for a more detailed transcription map showing the positions of the L-strand transcripts and the precursors to the H-strand transcripts).

known mitochondrially synthesized proteins. Only reading frames of at least 40 amino acids conserved with at least 50% amino acid homology in the bovine sequence were considered. The gene for cytochrome *c* oxidase subunit I (COI) was identified from an N-terminal amino acid sequence of the bovine protein (J. E. Walker, personal communication) and confirmed by comparison with the yeast mt COI gene. COII was identified from the complete amino acid sequence of the bovine protein[41] and COIII by homology with the yeast gene[42] and by an N-terminal sequence of the bovine protein (G. Buse, personal communication). The genes for ATPase 6 and cytochrome *b* were identified by homology with the corresponding yeast genes[43,44] and also by amino acid sequences of bovine cytochrome *b* (W. Machleidt, personal communication). We have found no homology between the bovine amino acid sequence of ATPase 9 (E. Wachter, personal communication) and any of the reading frames, so that, as in *Neurospora*[45], this protein must be nuclear coded; in yeast it is mitochondrially coded[46,47].

In regions representing identified genes the respective human and bovine sequences are highly homologous. As shown in Table 2, there is ~80% amino acid conservation, which is 5% higher than the nucleotide conservation, reflecting the degeneracy of the genetic code. Over half the codons for the conserved amino acids are different in the third (degenerate) position or, in the case of the leucine codons UUR and CUN, in the first and third positions.

We have found eight unidentified reading frames (URFs) that are conserved in human and bovine mtDNA. On average, these show ~70% nucleotide and amino acid conservation (Table 2), again with approximately half of the conserved amino acids having changed codons. This type of homology is unlikely to have occurred by chance: for example, if the reading frames are compared out of phase in the +1 and +2 frames, the amino acid homology, including all the termination codons, is 35.4% and the percentage of conserved amino acids with changed codons drops to 4.5%. Therefore, we believe that most, if not all, of these reading frames represent genes for mammalian mitochondrial proteins. Further evidence in support of this is provided by the mapping and sequence analysis of the polyadenylated presumptive mRNAs reported in the accompanying articles[17,18]. These RNAs correspond exactly to the reading frames of URF1, URF2, COI, COII, URF A6L and ATPase 6, COIII, URF3, URF 4L and URF4, URF5 and the antisense of URF6, and cytochrome *b* as shown in Figs 1 and 2. In addition, there are large L-strand transcripts covering the H-strand-coded URF6, and the number and sizes of these reading frames are in fairly good agreement with the band pattern obtained by gel electrophoresis of HeLa cell mt proteins produced in the presence of emetine, an inhibitor of cytoplasmic protein synthesis (J. E. Walker, personal communication).

Two of the H-strand transcripts contain more than one reading frame, RNA14 (URF A6L and ATPase 6) and RNA7 (URF 4L and URF4). These present an anomaly and it is not known whether the reading frames are separated by a further processing event of these RNAs. The same organization is found in bovine mtDNA, making it unlikely that the breaks in the reading frames are due to sequence errors. A possible explanation is that these frames contain short intervening sequences which, when excised, leave only a single reading frame. However, using nuclease S_1 protein experiments, Ojala *et al.*[48] found no evidence for intervening sequences in the mtDNA regions coding for the polyadenylated RNAs.

Initiation codons

Although all the identified genes (except ATPase 6) and some of the URFs have an AUG initiation codon within the first six bases of their mRNA start, others do not. The first ATG in bovine URF2 is 320 codons downstream from the RNA start in a reading frame of 347 codons. Also, in human and bovine URF3 the first AUG is, respectively, 89 and 87 codons downstream in a reading frame of 115 codons. Ignoring the cases of ATPase 6 and URF4, which are the second reading frames in their presumptive mRNAs, and also the case of human URF2, which is discussed separately below, all reading frames that do not have an AUG near the 5′ end of the RNAs have an AUA in an equivalent position. Thus, if these reading frames are translated completely, as their homology suggests they are, AUA must function as an initiation codon as well as AUG. AUA has already been shown to code for internal methionine as well as AUG[2,3]. In the case of human URF2 there is an AUU in the same position as the AUA in bovine URF2. Thus, to translate completely this reading frame we must argue that, like AUA, AUU can also be an initiation codon in mitochondrial genes.

Termination of translation

Over half of the presumptive mRNAs do not have UAA or UAG termination codons at the end of the reading frame (Table 2). In three cases, human COI and URF6 and bovine cytochrome *b*, termination probably occurs at AGA and AGG codons. These have been predicted to be termination codons and not arginine codons as in the universal genetic code[8]. This is based on the observation that only CGN arginine codons are used in all the reading frames, and AGA and AGG are only found at the ends of reading frames at the junction with the next

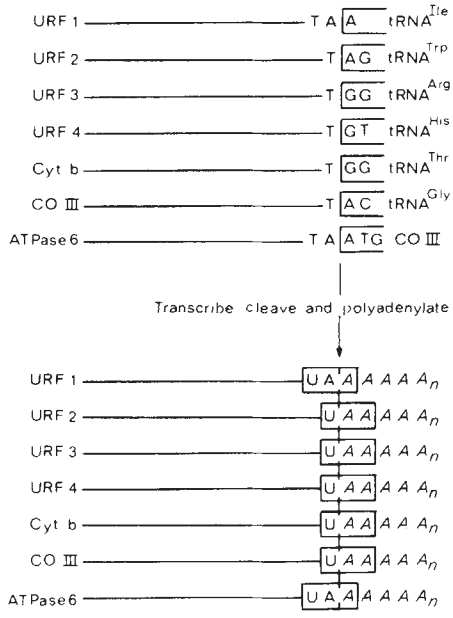


Fig. 3 The transcriptional processing and polyadenylation model for termination of translation.

gene or where the reading frame continues into the next gene. Also, no gene coding for a tRNA that could translate the codons AGA/G has been found. Finally, the composition of a C-terminal peptide fragment of bovine cytochrome *b* (W. Machleidt, personal communication) agrees with the prediction from the DNA sequence that this gene terminates in AGA.

To explain the lack of termination codons in most of the reading frames, we have proposed a model for the post-transcriptional creation of UAA termination codons in the mRNAs[49]. This model was based on the observation that if the tRNAs were precisely cleaved out of a primary transcript by an RNase P type enzyme(s)[50,51], mRNAs from those reading frames with no termination codon would be left with either a U or UA at their 3′ ends in phase with the reading frame. Thus, if these RNAs were then polyadenylated, a UAA termination codon would be created in-phase as shown in Fig. 3. Ojala *et al.*[17], in the accompanying article, have confirmed this hypothesis by sequence analysis of the 3′ ends of these mRNAs. This mechanism also provides the termination codon for ATPase 6. ATPase 6 and COIII have overlapping termination and initiation codons, that is UAAUG, and because the cleavage site is between the two A residues[18] the termination codon is destroyed by RNA processing of the primary transcript (Figs 1, 2). However, as there is no tRNA gene in this region, the signal in the RNA sequence for processing is not known.

Transcriptional processing

As discussed above, comparisons of the human DNA sequence with the polyadenylated RNA sequences[17,18] confirm the earlier predictions that the tRNA coding regions are involved in processing of the primary transcripts[15,16,49,52,53]. It is now clear that in the case of L-strand-coded tRNA genes the H-strand transcript is cleaved precisely at the 5′ and 3′ termini of the tRNA genes. Cleavage also seems to occur where the H-strand transcript has antisense, that is, H-strand-coded, tRNA genes. For example, cleavage of the H-strand transcript occurs between the antisense tRNA^{Glu} gene and the cytochrome *b* gene (see Figs 1, 2). In both cases the cleavage point is a few bases downstream from the antisense tRNA gene. An additional cleavage point is between the ATPase 6 and COIII genes, where there is no known tRNA structure. Thus, except for this latter case it is reasonable to assume that in the primary H-strand transcript L-strand-coded tRNA sequences can fold up and be recognized by an RNase P-like activity. Antisense (H-strand-coded) tRNA coding regions also seem to be recognized by the RNA processing enzymes, although cleavage points have only been found on the 3′ side. With the exception the 12S rRNA and the tRNAs, the RNAs are then polyadenylated as discussed above; the tRNAs are matured by addition of CCA at their 3′ terminus. At some stage base modification of the rRNAs and tRNAs must also occur. It is likely that the H-strand-coded genes in the L-strand primary transcript are similarly processed but this has been less well studied.

Mitochondrial origins

In conclusion, it is now clear that the mammalian mitochondrial genetic system cannot generally be classified as either prokaryote-like or eukaryote-like. The mammalian mt genetic code is different from the so-called universal genetic code and is read in a unique fashion by a minimal set of mitochondrial tRNAs. The mt tRNAs generally lack a number of features previously found to be invariant among tRNAs from all other sources, and the mammalian mt rRNAs have distinctive structures only distantly related to all previously determined rRNA sequences[16] Similarly, the economy with which the large initial mitochondrial H-strand transcripts seem to be processed is unprecedented. It is also striking that mammalian mitochondria are very different from other mitochondria. In yeast mitochondria, for example, not only is there a slightly different genetic code, but also, the genes are widely spaced and in a different order, and in some cases they contain intervening sequences[44,54] These radical differences make it difficult to draw conclusions regarding mitochondrial evolution. Some form of endosymbiosis, involving the colonization of a primitive eukaryotic cell by a respiring bacteria-like organism, is an attractive hypothesis to explain the origins of mitochondria. However, the endosymbiont may have been no more closely related to current prokaryotes than to eukaryotes. It must also be borne in mind when making comparisons that the mechanism of selection and the selective pressures experienced by a captive cellular organelle may be quite different from those that affect a free-living organism.

We thank G. Attardi for a gift of HeLa cell mtDNA. B.A.R. was supported by NIH grants GM-21405 and GM-25962 and is a recipient of NIH Research Career Development Award GM-00124. P.H.S. was supported by the Deutsche Forschungsgemeinschaft through SFB-74 to B. Müller-Hill as well as a short-term EMBO fellowship. S.A. is a research fellow of the Cystic Fibrosis Foundation.

Received 26 January; accepted 18 February 1981.

1. Borst, P. *Trends biochem. Sci.* **2,** 31–34 (1977).
2. Barrell, B. G., Bankier, A. T. & Drouin, J. *Nature* **282,** 189–194 (1979).
3. Young, I. G. & Anderson, S. *Gene* **12,** 257–265 (1980).
4. Li, M. & Tzagoloff, A. *Cell* **18,** 47–53 (1979).
5. Macino, G., Coruzzi, G., Nobrega, F. G., Li, M. & Tzagoloff, A. *Proc. natn. Acad. Sci. U.S.A.* **76,** 3784–3785 (1979).
6. Fox, T. D. *Proc. natn. Acad. Sci. U.S.A.* **76,** 6534–6538 (1979).
7. Heckman, J. E., Sarnoff, J., Alzner DeWeerd, B., Yyn, S. & RajBhandary, U. L. *Proc. natn. Acad. Sci. U.S.A.* **77,** 3159–3163 (1980).
8. Barrell, B. G. *et al. Proc. natn. Acad. Sci. U.S.A.* **77,** 3164–3166 (1980).
9. Bonitz, S. G. *et al. Proc. natn. Acad. Sci. U.S.A.* **77,** 3167–3170 (1980).
10. Crews, S., Ojala, D., Posakony, J., Nishiguchi, J. & Attardi, G. *Nature* **277,** 192–198 (1979).
11. Gillum, A. M. & Clayton, D. A. *J. molec. Biol.* **135,** 353–368 (1979).
12. Martens, P. A. & Clayton, D. A. *J. molec. Biol.* **135,** 327–351 (1979).
13. Aloni, Y. & Attardi, G. *Proc. natn. Acad. Sci. U.S.A.* **68,** 1757–1761 (1971).
14. Murphy, W., Attardi, B., Tu, C. & Attardi, G. *J. molec. Biol.* **99,** 809–814 (1975).
15. Attardi, G. *et al. in Proc. ICN-UCLA Symp. Extrachromosomal DNA* (1979) (in the press).
16. Eperon, I. C., Anderson, S. & Nierlich, D. P. *Nature* **286,** 460–467 (1980).
17. Ojala, D., Montoya, J. & Attardi, G. *Nature* **290,** 470–474 (1981).
18. Montoya, J., Ojala, D. & Attardi, G. *Nature* **290,** 465–470 (1981).
19. Gillum, A. M. & Clayton, D. A. *Proc. natn. Acad. Sci. U.S.A.* **75,** 677–681 (1978).
20. Brown, W. M., Shine, J. & Goodman, H. M. *Proc. natn. Acad. Sci. U.S.A.* **75,** 735–739 (1978).
21. Bogenhagen, D. & Clayton, D. A. *J. molec. Biol.* **119,** 49–68 (1978).
22. Sekiya, T., Kobayashi, M., Seki, T. & Koike, K. *Gene* **11,** 53–62 (1980).
23. Upholt, W. B. & Dawid, I. B. *Cell* **11,** 571–583 (1977).
24. Berk, A. J. & Clayton, D. A. *J. molec. Biol.* **86,** 801–824 (1974).
25. Crews, S. & Attardi, G. *Cell* **19,** 775–784 (1980).
26. Van Etten, R. A., Walberg, M. W. & Clayton, D. A. *Cell* **22,** 157–170 (1980).
27. Baer, R. & Dubin, D. T. *Nucleic Acids Res.* **8,** 4927–4941 (1980).
28. Dubin, D. T., Timko, K. D. & Baer, R. J. *Cell* (in the press).
29. Cantatore, P. & Attardi, G. *Nucleic Acids Res.* **8,** 2605–2624 (1980).
30. Baer, R. & Dubin, D. T. *Nucleic Acids Res.* **8,** 4927–4941 (1980).
31. Staden, R. *Nucleic Acids Res.* **8,** 817–825 (1980).
32. Aujame, L. & Freeman, K. B. *Nucleic Acids Res.* **6,** 455–469 (1979).
33. Crick, F. H. C. *J. molec. Biol.* **19,** 548–555 (1966).
34. Lagerkvist, U. *Proc. natn. Acad. Sci. U.S.A.* **75,** 1759–1762 (1978).
35. Lynch, D. & Attardi, G. *J. molec. Biol.* **102,** 125–141 (1976).

36. Wrede, P., Woo, N. H. & Rich, A. *Proc. natn. Acad. Sci. U.S.A.* **76,** 3289–3293 (1979).
37. Kuchino, Y., Watanabe, S., Harada, F. & Nishimura, S. *Biochemistry* **19,** 2085–2089 (1980).
38. Fukada, K. & Abelson, J. *J. molec. Biol.* **139,** 377–391 (1980).
39. de Bruijn, M. H. L. *et al. Nucleic Acids Res.* **8,** 5213–5222 (1980).
40. Grauss, D. H., Grüter, F. & Sprinzl, M. *Nucleic Acids. Res.* **6,** r1–r19 (1979).
41. Steffens, G. J. & Buse, G. *Hoppe-Seyler's Z. physiol. Chem.* **360,** 613–619 (1979).
42. Thalenfeld, B. E. & Tzagoloff, A. *J. biol. Chem.* **255,** 6173–6180 (1980).
43. Macino, G. & Tzagoloff, A. *Cell* **20,** 507–517 (1980).
44. Nobrega, F. G. & Tzagoloff, A. *J. biol. Chem.* **255,** 9828–9837 (1980).
45. Sebald, W. & Wachter, E. in *29th Mossbacher Colloquium on Energy Conservation in Biological Membranes* (eds Schäfer, G. & Klingenberg, M.) 228–236 (Springer, Berlin, 1978).
46. Hensgens, L. A., Grivell, L. A., Borst, P. & Bos, J. L. *Proc. natn. Acad. Sci. U.S.A.* **76,** 1663–1667 (1979).
47. Macino, G. & Tzagoloff, A. *J. biol. Chem.* **254,** 4617–4623 (1979).
48. Ojala, D., Merkel, C., Gelfand, R. & Attardi, G. *Cell* **22,** 393–403 (1980).
49. Barrell, B. G. *et al.* in *31st Mosbach Colloquium* (eds Bücher, Th., Weiss, H. & Sebald, W.) (Springer, Berlin, in the press).
50. Robertson, M. D., Altman, S. & Smith, J. D. *J. biol. Chem.* **247,** 5243–5251 (1972).
51. Altman, S. & Smith, J. D. *Nature new Biol.* **233,** 35–39 (1971).
52. Attardi, G. *et al.* in *The Organisation and Expression of the Mitochondrial Genome* (eds Kroon, A. M. & Saccone, C.) (North-Holland, Amsterdam, in the press).
53. Battey, J. & Clayton, D. A. *J. biol. Chem.* **255,** 11599–11606 (1980).
54. Borst, P. & Grivell, L. A. *Cell* **15,** 705–723 (1978).
55. Staden, R. *Nucleic Acids Res.* **4,** 4037–4051 (1977); **5,** 1013–1015 (1978).
56. Drouin, J. *J. molec. Biol.* **140,** 15–34 (1980).
57. Smith, A. J. H. *Nucleic Acids Res.* **6,** 831–848 (1979).
58. Gronenborn, B. & Messing, J. *Nature* **272,** 375–377 (1978).
59. Schreier, P. H. & Cortese, R. *J. molec. Biol.* **129,** 169–172 (1979).
60. Sanger, F., Coulson, A. R., Barrell, B. G., Smith, A. J. H. & Roe, B. A. *J. molec. Biol.* **143,** 161–178 (1980).
61. Anderson, S., Gait, M. J., Mayol, L. & Young, I. G. *Nucleic Acids Res.* **8,** 1731–1743 (1980).
62. Sanger, F., Nicklen, S. & Coulson, A. R. *Proc. natn. Acad. Sci. U.S.A.* **74,** 5463–5467 (1977).
63. Sanger, F. & Coulson, A. R. *FEBS Lett.* **87,** 107–110 (1978).

Distinctive features of the 5′-terminal sequences of the human mitochondrial mRNAs

Julio Montoya, Deanna Ojala & Giuseppe Attardi

Division of Biology, California Institute of Technology, Pasadena, California 91125, USA

The 5′-end proximal sequences of all the putative mRNAs coded for by the heavy strand of HeLa cell mitochondrial DNA have been determined and aligned with the DNA sequence. All these mRNAs start directly at, or very near to, an AUG or AUA triplet, with the exception of one which starts at an AUU. The available evidence indicates that the terminal or subterminal AUGs and AUAs, and possibly also the terminal AUU, are initiator codons for the corresponding polypeptides. In most cases, the individual mRNA coding sequences are flanked on their 5′ side by a tRNA gene, without any intervening nucleotide.

RECENT sequence analysis of human mitochondrial DNA (mtDNA) has revealed an extremely high degree of packing of genetic information in this DNA (ref. 1 and accompanying paper (ref. 26)). In agreement with these observations, a detailed transcription mapping study of HeLa cell mtDNA by the S_1 protection technique has shown that the sequences of the heavy (H) strand are almost completely saturated by the rRNAs, poly(A)-containing RNAs and tRNAs coded for by this strand[2,3]. A particularly intriguing feature of the human mitochondrial gene organization is the frequent juxtaposition, or close proximity, of the individual protein-coding sequences, on their 5′ side, with a tRNA gene or another reading frame. Thus, the gene for subunit II of cytochrome *c* oxidase (COII) has been shown to be immediately contiguous to a $tRNA^{Asp}$ gene without any intervening nucleotide[4]. This unusual gene organization has raised the question of whether these putative mitochondrial mRNAs have, as all eukaryotic mRNA so far analysed[5,6], a 5′ non-coding stretch, which may thus overlap the contiguous tRNA gene or reading frame, or whether they lack completely a leader sequence on their 5′ side.

As a first approach to this question, the 5′-end proximal region of the putative COII mRNA has recently been sequenced and the sequence thus determined, aligned with the COII gene sequence[7]. The results clearly showed that the 5′ end of the COII mRNA corresponds precisely with the first nucleotide of the COII coding sequence. These experiments left open the question of whether the complete lack of 5′ non-coding nucleotides is a general feature of human mitochondrial mRNAs, reflecting stringent constraints in the initiation of mitochondrial protein synthesis or specific rules of RNA processing, or whether, on the contrary, the position of the initiator codon relative to the 5′ end may vary in different mRNAs, possibly depending on the position of the adjacent tRNA gene or reading frame. We report here information bearing on these questions, obtained by sequencing a 5′-end proximal segment of all the H strand-coded, polysome-associated, poly(A)-containing RNAs, which are presumably specific mRNAs, and by aligning the RNA sequences with the mtDNA sequence.

Isolation and 5′-end labelling of mitochondrial mRNAs

To isolate HeLa cell mitochondrial poly(A)-containing RNAs in adequate amounts for sequencing analysis, the micrococcal nuclease procedure for the purification of mitochondrial RNA[8] was scaled up as previously described[7]. Figure 1*a* shows the electrophoretic pattern in a 1.4% agarose–CH_3HgOH slab gel, after ethidium bromide staining, of a large-scale preparation of mitochondrial oligo(dT)-bound RNA from micrococcal nuclease-treated organelles (from ~30 g of cells). One clearly recognizes the characteristic set of components previously

Table 1 5′-Terminal nucleotide analysis of *in vitro* labelled mitochondrial poly(A)-containing RNA species

Poly(A)-containing RNA species	% Of total ^{32}P radioactivity*			
	AMP	CMP	GMP	UMP
5	52.5	8.1	27.6	11.7
7	43.8	2.1	7.8	47.1
9	12.4	0.9	2.0	4.5
11	81.1	3.0	10.2	5.5
12	78.9	2.3	12.7	5.9
13	81.4	2.4	10.0	6.0
14	91.0	3.1	2.2	3.5
15	95.9	0.3	1.6	2.1
17	73.7	2.5	6.4	17.2

The 5′-terminal nucleotide was determined by exhaustive digestion of the *in vitro* labelled RNAs with nuclease P_1 (ref. 12, Calbiochem) and fractionation of the products on polyethyleneimine impregnated-cellulose TLC plates, as previously described[8].

* The values refer to the radioactivity which migrated from the origin. The % of the input radioactivity recovered from the origin varied between 0.1 and 5%.

0028–0836/81/150465—06$01.00