Complete Sequence of a Sea Lamprey (*Petromyzon marinus*) Mitochondrial Genome: Early Establishment of the Vertebrate Genome Organization

Woo-Jai Lee and Thomas D. Kocher

Department of Zoology and Center for Marine Biology, University of New Hampshire, Durham, New Hampshire 03824

Manuscript received May 26, 1994
Accepted for publication October 26, 1994

ABSTRACT

The complete nucleotide sequence of a sea lamprey (*Petromyzon marinus*) mitochondrial genome has been determined. The lamprey genome is 16,201 bp in length and contains genes for 13 proteins, two rRNAs, 22 tRNAs and two major noncoding regions. The order and transcriptional polarities of protein-coding genes are basically identical to those of other chordate mtDNAs, demonstrating that the common mitochondrial gene organization of vertebrates was established at an early stage of vertebrate evolution. The two major noncoding regions are separated by two tRNA genes. The first region probably functions as the control region because it contains distinctive conserved sequence blocks (CSB-II and III) common to other vertebrate control regions. The central conserved domain observed in other vertebrate control regions is not found in the lamprey, suggesting that it is a recently evolved functional domain in vertebrates. Noncoding segments are not found in the expected position of the origin of replication for the second strand, suggesting either that one of the tRNA genes has a dual function or that the second noncoding region may function as the second-strand origin. The base composition at the wobble positions of fourfold degenerate codon families is highly biased toward thymine (32.7%). Values of GC- and AT-skew are typical of vertebrate mitochondrial genomes.

LAMPREYS, among the earliest diverged vertebrates, have a unique evolutionary history in the 550 million years since they separated from the main vertebrate lineage. As one of the earliest diverged vertebrates, they attract particular attention from evolutionary biologists, because they provide information important to understanding of the early evolution of vertebrates (STOCK and WHITT 1992; FOREY and JANVIER 1993). Mostly because of the lack of distinct characters, the systematics of lamprey species is not well established (HUBBS and POTTER 1971). Approximately 40 species are distributed in coastal drainages throughout the Northern Hemisphere, and four species are found in temperate areas of the Southern Hemisphere (MOYLE and CECH 1988). Sea lampreys spend 3–7 years as ammocoetes before migrating to oceans or lakes to spend ≤2 years as adults feeding on fishes.

The patterns of mitochondrial genome organization may be informative for the phylogeny of distantly related taxa because the rate of gene rearrangement is much slower than nucleotide substitutions (BROWN 1985; SMITH *et al.* 1993). Each phylum shows a common basic gene order despite minor relocations of genes in some taxa (HOFFMANN *et al.* 1992). In nematodes, the gene order of two species (*Caenorhabditis elegans* and *Ascaris suum*) differs only in the transposition of the A + T rich region (OKIMOTO *et al.* 1992). In insects, 11 tRNA genes have been moved between the genomes of honey bee and *Drosophila yakuba,* but most keep the same transcriptional polarity (GARESSE 1988; CROZIER and CROZIER 1993). Exceptions to the conservation of gene order within phyla are found in echinoderms. Partial sequences demonstrate that brittle stars and sea stars share a nearly identical order, which is different from that of sea urchin and sea cucumber (SMITH *et al.* 1993). The change of gene order can be explained by the simple inversion of a 4.6-kb segment. Vertebrate mtDNAs also show a highly conserved gene order from bony fish to human, although minor rearrangements have been found in chicken and marsupial mtDNAs (DESJARDINS and MORALIS 1990; JANKE *et al.* 1994). The bird genome shows a transposition of genes for *ND6* and *tRNA-Glu* relative to the genes for *CYT b, tRNA-Thr* and *tRNA-Pro.* In the marsupial genome, rearrangements have occurred within the cluster of tRNA genes near the replication origin of the second strand.

As more sequence data accumulate, the control region appears to be the most enigmatic portion of the animal mitochondrial genome. Although the functions of some sequence elements in the control region of mammalian mtDNA have been described (CLAYTON 1982), many questions about the function and evolution of the region still remain unanswered. In particular, the function of the central conserved region found in vertebrates, but not other deuterostomes, is not known.

Corresponding author: Thomas D. Kocher, Department of Zoology, Spaulding Life Science Building, University of New Hampshire, Durham, NH 03824. E-mail: tdk@christa.unh.edu

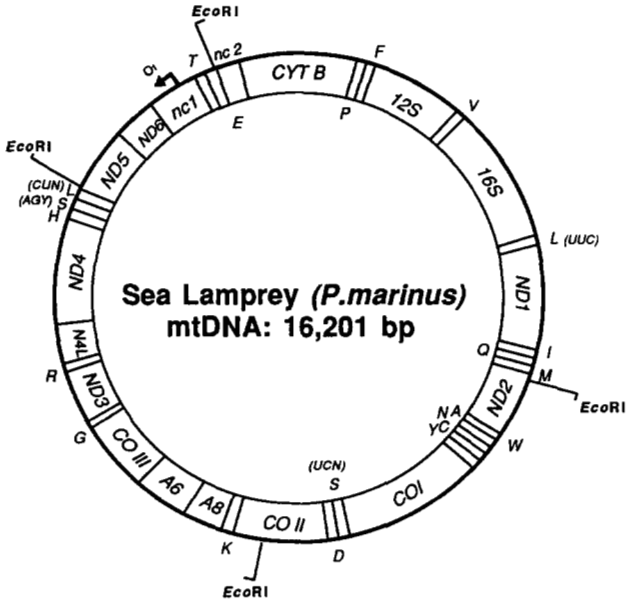

FIGURE 1.—Gene map of a sea lamprey mitochondrial DNA. The positions of 13 protein-coding genes and two major noncoding regions are shown using abbreviations given in the text. All protein-coding genes except for *ND6* are encoded on the first strand (outer) with clockwise transcriptional polarity. tRNA genes are represented by a one-letter amino acid code located either outside or inside circles according to the coding strand. The codon families of each serine and leucine tRNA are presented in parentheses. Those labeled outside the circle are encoded on the first strand with clockwise transcriptional polarity. NC1 and NC2, noncoding regions 1 and 2, respectively. The four *Eco*RI restriction sites used for cloning are shown.

In this paper, we present the complete nucleotide sequence of a sea lamprey mitochondrial genome, including the pattern of base composition and comparisons of gene arrangement with other vertebrate and invertebrate genomes. This sequence will help elucidate the evolution of mitochondrial gene order in early vertebrates and will facilitate systematic studies of lampreys worldwide.

MATERIALS AND METHODS

mtDNA isolation and cloning: Adult sea lampreys were collected from fish ladders on the Cocheco River at Dover, NH, during the spawning season of 1992 (from May to June). Purified mitochondrial DNA was obtained from the fresh liver and eggs using slight modifications of the protocols described by LANSMAN *et al.* (1981) and DOWLING *et al.* (1990). The low-speed centrifugations were performed at 550 × *g* instead of 700–800 × *g* and a CsCl-ethidium bromide gradient at a density of 1.50 g/ml, instead of 1.55 g/ml, was used in ultracentrifugation. The isolated mtDNA was digested with *Eco*RI, resulting in four fragments (3, 3.5, 4 and 6 kb) covering the whole genome.

Each of the *Eco*RI fragments was cloned in pBluescript II SK and amplified using *Escherichia coli* XL-1 Blue (Stratagene). Nested sets of deletions were constructed from the four recombinant DNAs with the Erase-A-Base system kit (Promega). After checking the sizes of the inserts on a 1.0% agarose gel, overlapping clones were prepared and sequenced.

DNA sequencing: All nucleotide sequences were obtained from double-stranded plasmid DNA. Plasmid DNA from either a standard alkaline lysis miniprep (SAMBROOK *et al.* 1989) or Magic miniprep kit (Promega) was used for Taq DyeDeoxy Terminator cycle-sequencing reaction (Applied Biosystems Inc.). Extra dye terminators were removed from the cycle sequencing reaction with a spin column containing 5% Sephadex. Finally, the entire product of cycle sequencing was dried and resuspended in 4 μl of formamide-EDTA buffer, denatured at 90° and loaded on an automated DNA sequencer (Applied Biosystems 373A). The sequence obtained from each subclone averaged 350 bp and overlapped the next clone for 100–150 bp. There was no sequence variation observed within the overlapping regions in any clones.

After determining the entire nucleotide sequence, we designed four pairs of primers on the ends of the fragments with the help of a computer program (Primemate, DNAStar Inc.) to see if there were any missing *Eco*RI fragments. The PCR products were each ~300–400 bp in length, overlapping the junctions of two neighboring fragments. The sequences of the PCR products confirmed that the four *Eco*RI-generated fragments covered the entire sea lamprey mitochondrial genome.

Sequence analyses: The sequences of subcloned plasmids were aligned with SeqEd (Applied Biosystems), and the entire nucleotide sequence was further analyzed with ESEE (CABOT and BECKENBACH 1988) and GCG (Genetic Computer Group, version 7.0). The locations of 13 protein-coding genes were determined by comparisons of DNA or amino acid sequences of other mitochondrial genomes. The 22 tRNA genes and 2 rRNA genes also were identified by sequence homology, secondary structure and/or anticodon sequences. The rRNA sequences were compared by COMPARE and DOTPLOT, the secondary structures were folded by FOLD and SQUIGGLES and the comparisons of multiple sequences were made with PILEUP programs in the GCG package.

The base composition and codon frequency were obtained with COMPOSITION and CODONFREQUENCY programs in GCG. Base frequency statistics were calculated with the formulas in N. T. PERNA and T. D. KOCHER (unpublished results): $\%GC$ = proportion of $G + C$ out of the total base number. "Skew" indicates the difference of base frequencies between two strands. GC − Skew = $(G - C)/(G + C)$ and AT − Skew = $(A - T)/(A + T)$. Confidence intervals were calculated using a program supplied by N. PERNA. Similarity of base composition was tested using the method of RZHETSKY and NEI (1995).

RESULTS AND DISCUSSION

Genome content: The gene content of the sea lamprey mitochondrial genome includes 13 proteins, 22

FIGURE 2.—Complete nucleotide sequence of lamprey mitochondrial genome. The sequence is derived from the first strand in the direction of transcription (→). Beginning and end of each gene is indicated (|). The amino acid translations presented below the nucleotide sequence were derived using the mammalian mitochondrial genetic code. Stop codons are designated (*). Numbering begins from the first nucleotide of *CYT b*. *CYT b*, cytochrome *b*; *COI-III*, cytochrome *c* oxidase subunits I-III; *ATP6* and *ATP8*, ATPase subunits 6 and 8; *ND1–4L*, NADH dehydrogenase subunits 1–4L.

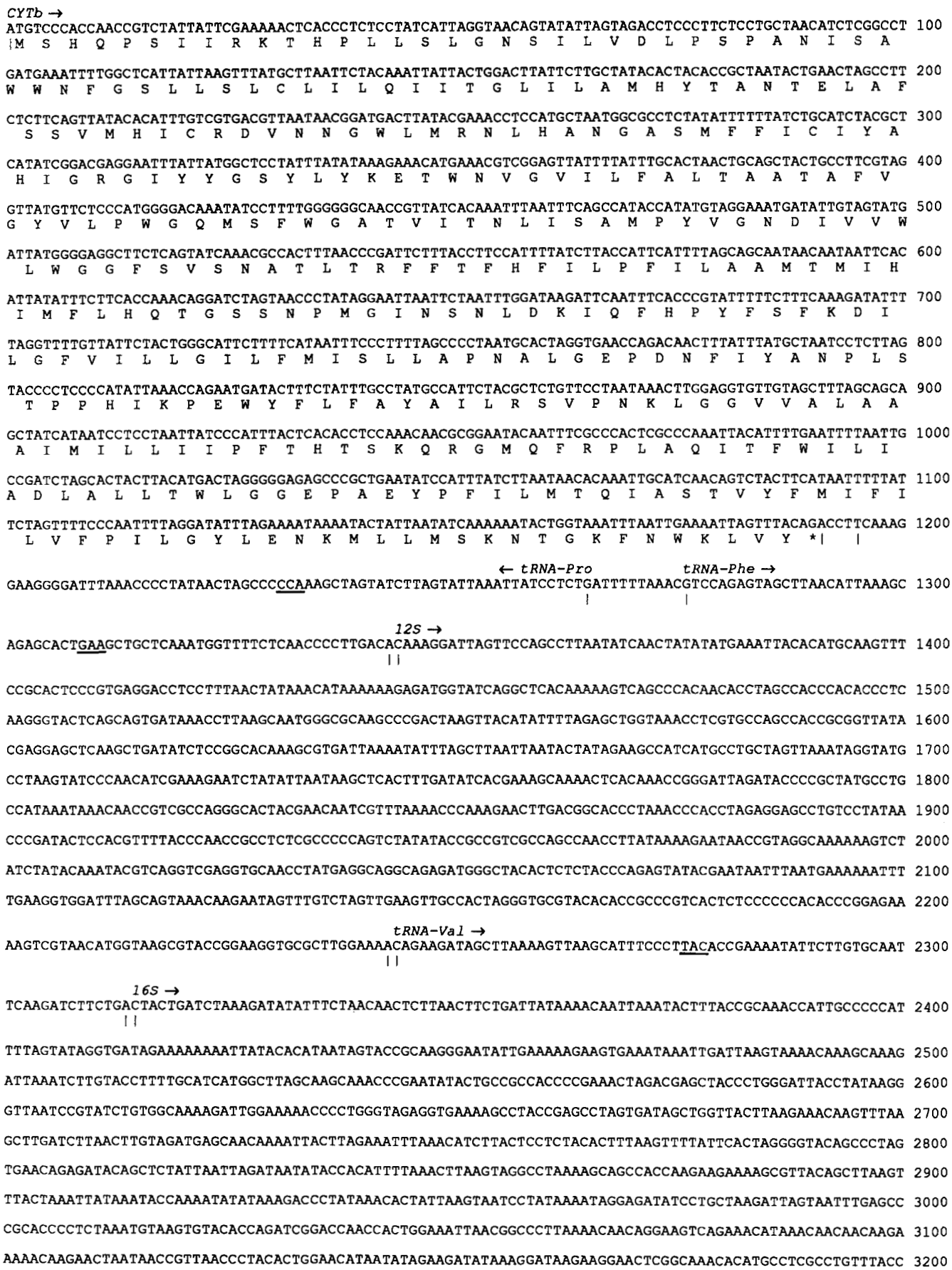

FIGURE 2.—*Continued*

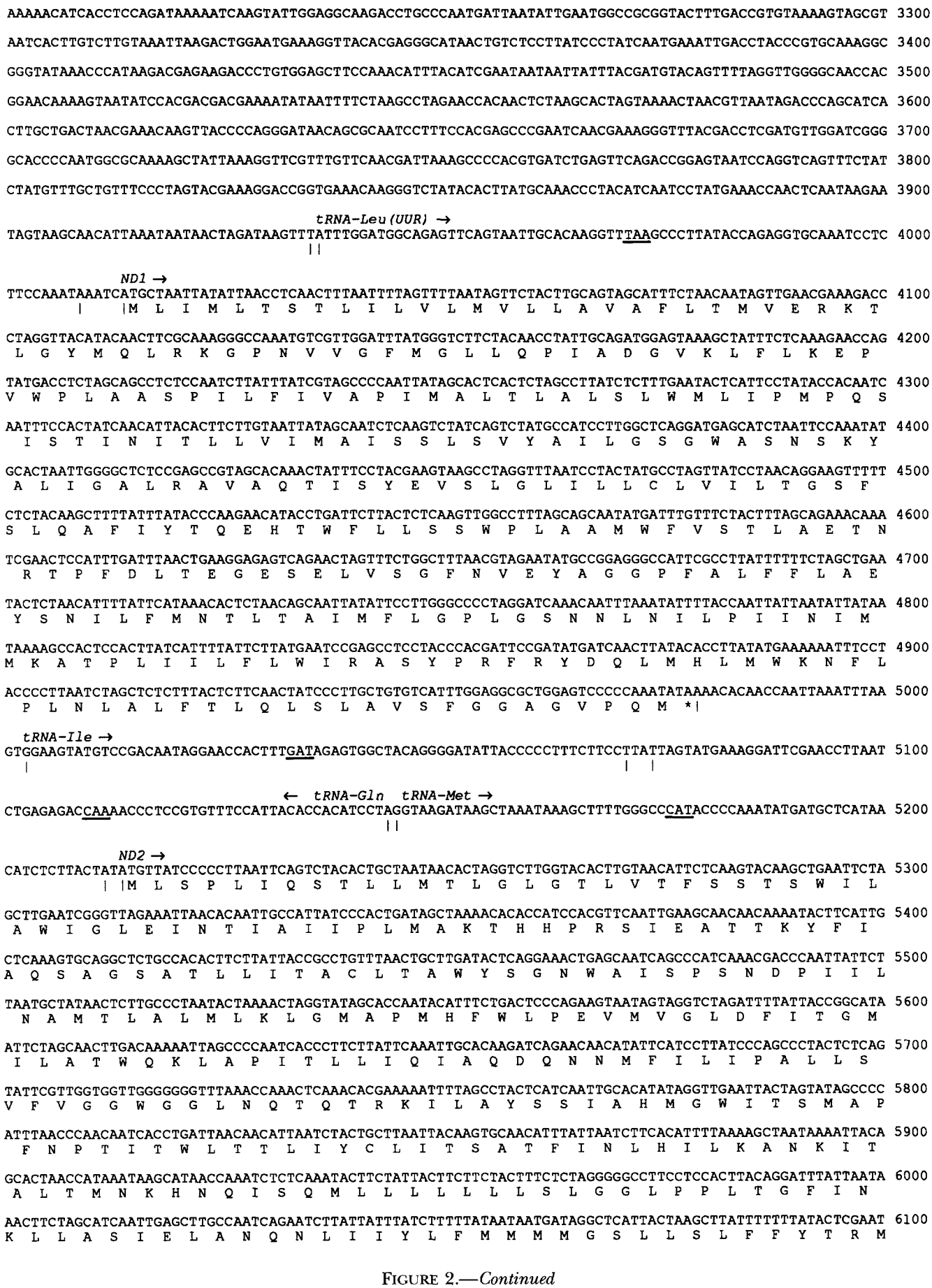

FIGURE 2.—*Continued*

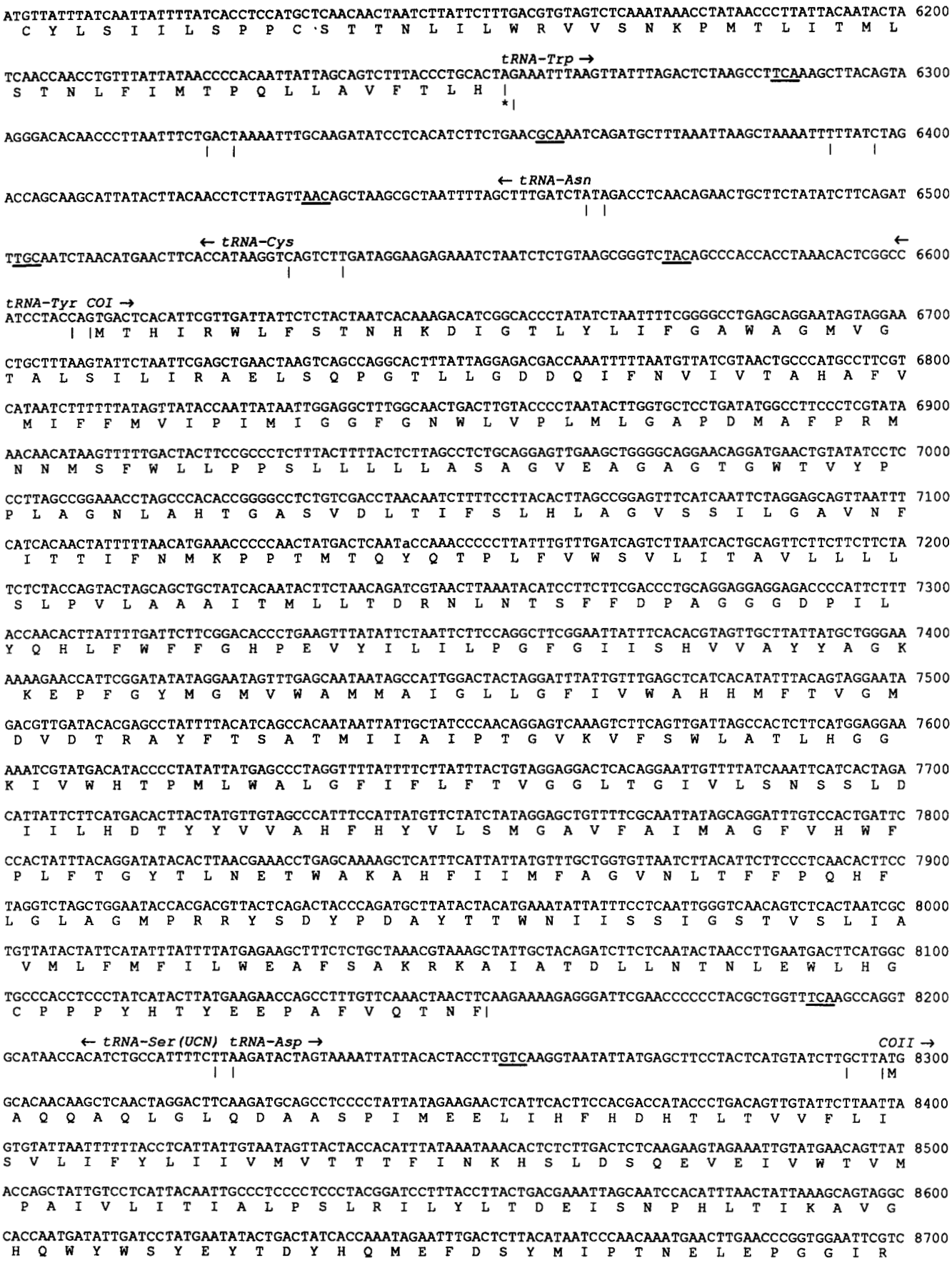

FIGURE 2.—*Continued*

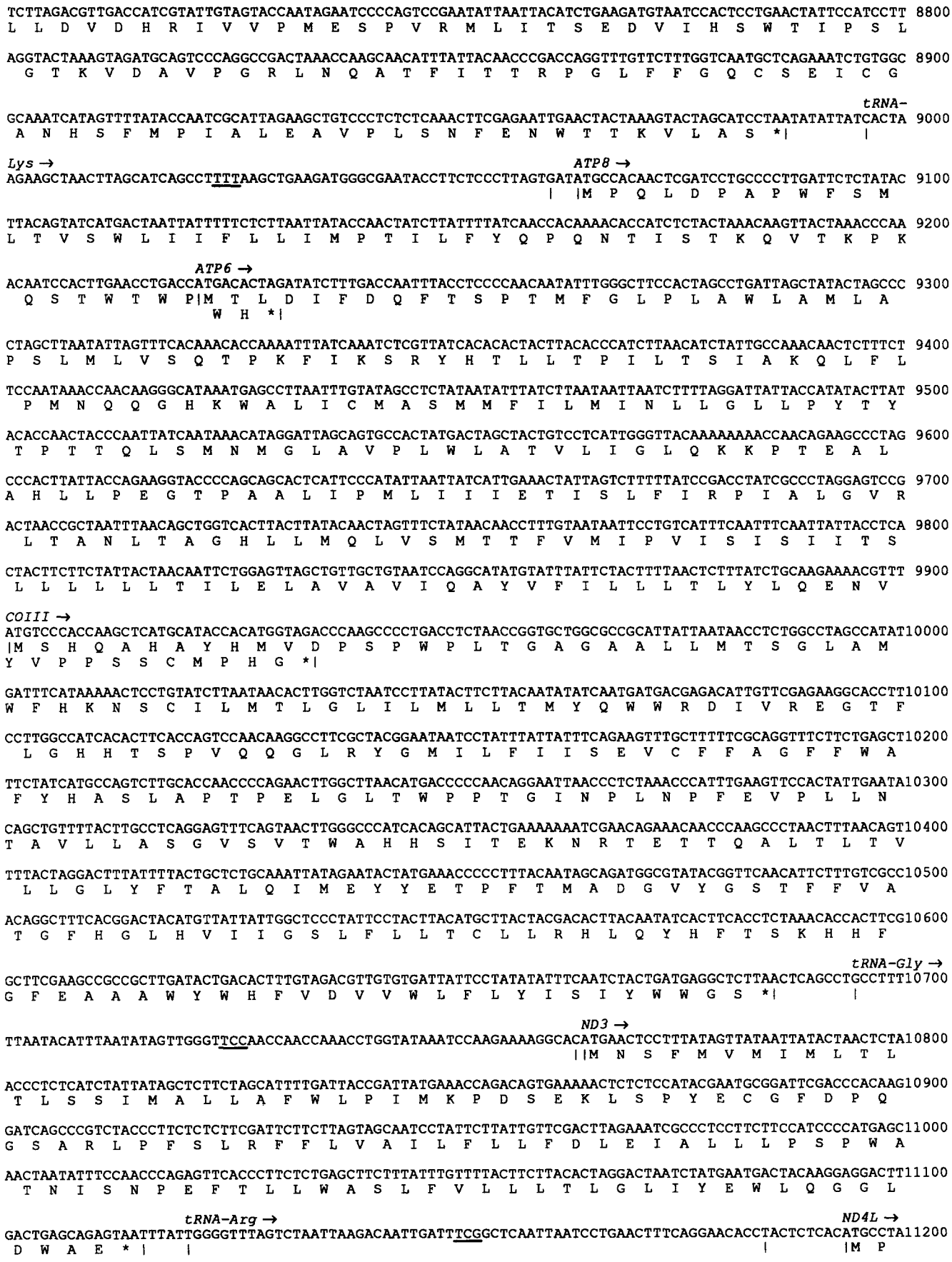

FIGURE 2.—*Continued*

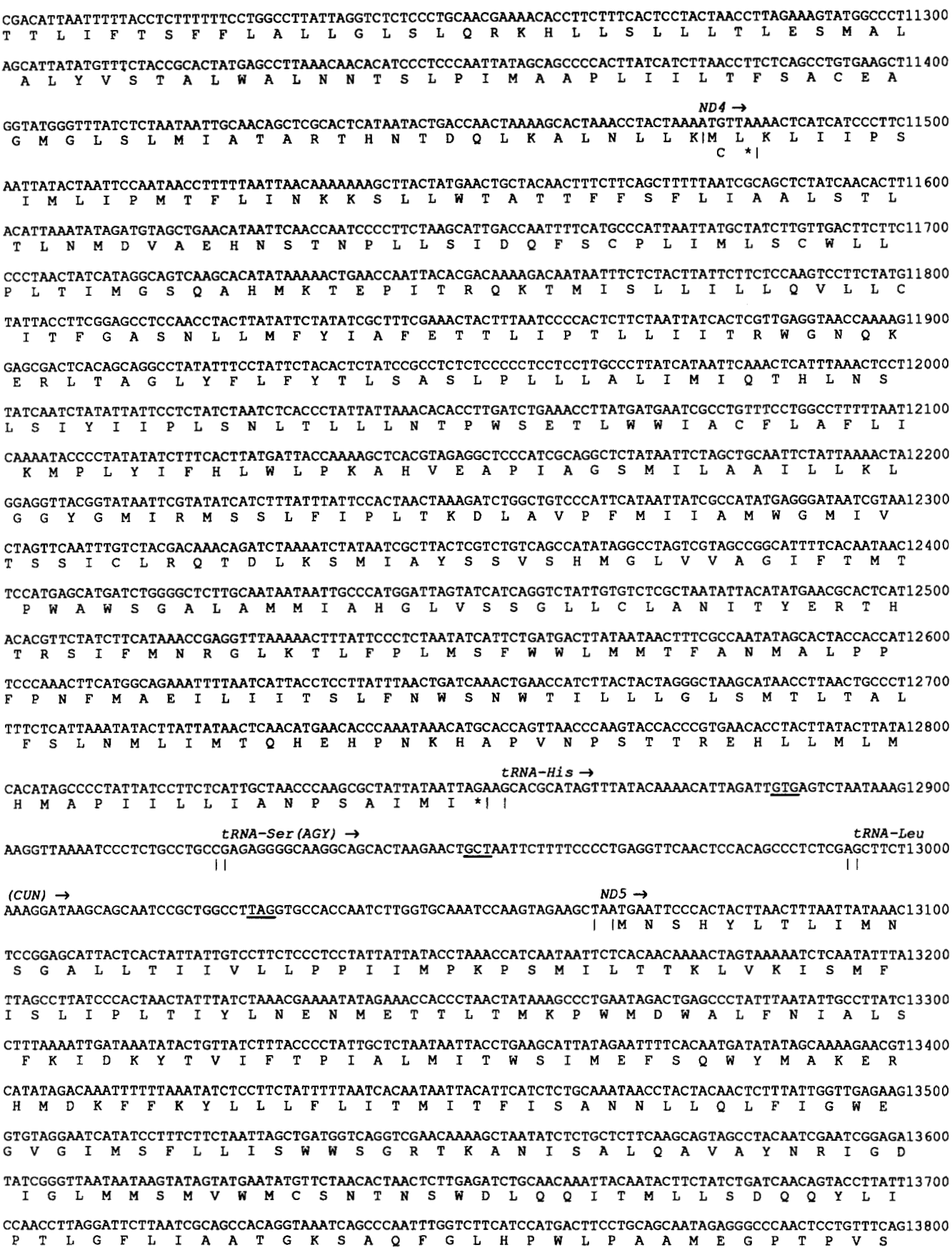

FIGURE 2.—*Continued*

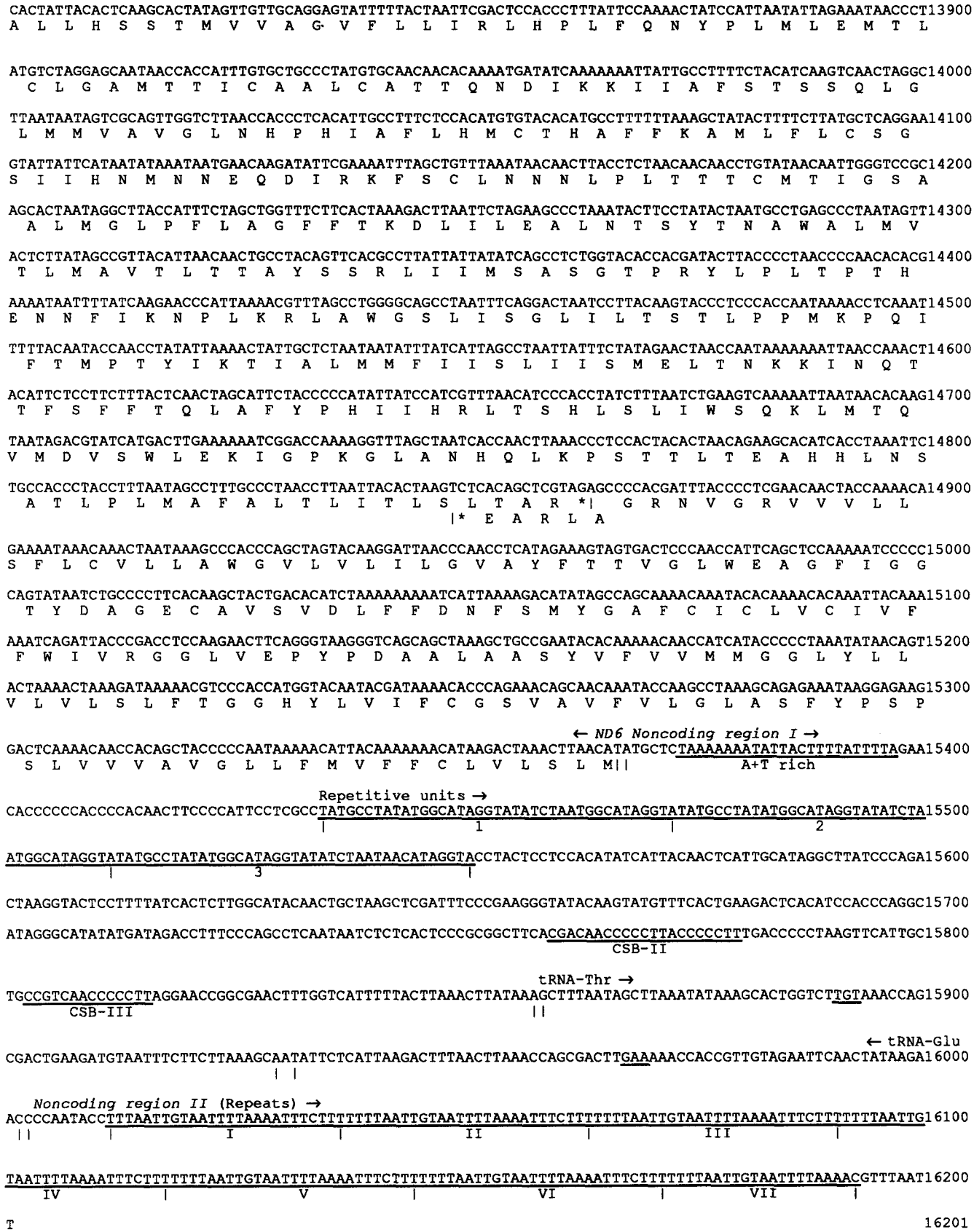

FIGURE 2.—*Continued*

TABLE 1

Location and coding strand of each gene in the sea lamprey mitochondrial DNA

Gene	Location	Strand
CYT b	1–1191	First
tRNA-Pro	1195–1265	Second
tRNA-Phe	1276–1343	First
12S rRNA	1344–2243	First
tRNA-Val	2244–2314	First
16S rRNA	2315–3935	First
tRNA-Leu (UUR)	3936–4009	First
ND1	4014–4979	First
tRNA-Ile	5003–5071	First
tRNA-Gln	5074–5144	Second
tRNA-Met	5145–5212	First
ND2[a]	5214–6257	First
tRNA-Trp	6256–6323	First
tRNA-Ala	6325–6393	Second
tRNA-Asn	6397–6465	Second
tRNA-Cys	6467–6532	Second
tRNA-Tyr	6538–6608	Second
COI[a]	6610–8163	First
tRNA-Ser (UCN)	8154–8224	Second
tRNA-Asp	8226–8294	First
COII	8298–8987	First
tRNA-Lys	8996–9061	First
ATP8[a]	9064–9231	First
ATP6[a]	9222–9935	First
COIII	9901–10686	First
tRNA-Gly	10695–10764	First
ND3	10765–11115	First
tRNA-Arg	11120–11185	First
ND4L[a]	11194–11484	First
ND4	11478–12854	First
tRNA-His	12856–12924	First
tRNA-Ser (AGY)	12925–12994	First
tRNA-Leu (CUN)	12995–13066	First
ND5[a]	13068–14864	First
ND6	14849–15367	Second
Noncoding region I	15368–15858	—
tRNA-Thr	15859–15930	First
tRNA-Glu	15932–16002	Second
Noncoding region II	16003–16201	—

[a] Gene overlaps with the following gene.

tRNAs, two rRNAs and two major noncoding regions (Figure 1). As seen in other vertebrate mtDNAs, most genes are encoded on the first, or heavy strand, except for *ND6* and eight tRNA genes. The sequence of all nucleotides of the first strand is presented in Figure 2 and the locations of the genes are shown in Table 1. The length of the lamprey mitochondrial genome is the shortest yet seen in vertebrates because of the small size of the putative control region. The sizes of most of the other genes are similar to other vertebrate mtDNAs. Two exceptions observed are the *CYT b* and *ATP6* genes, which are 50 bp (16 amino acids) and 35 bp longer, respectively, than in humans.

There are two unassigned DNA segments between the *ND6* and *CYT b* genes. The segments are separated by the genes for *tRNA-Thr* and *tRNA-Glu*. The first segment probably corresponds to the control region (D-loop), because it contains sequences normally associated with the replication origin of the heavy strand (see below). The first segment also contains three tandem copies of a 39-bp repeat (Figure 2). The second noncoding region consists almost entirely of seven copies of a 26-bp repeat.

There are six cases of gene overlap in the lamprey genome. The largest overlap occurs between the 3′ end of *ATP 6* and *COIII* (35 nucleotides). It is interesting that protein-coding genes, which are immediately followed by tRNA genes encoded on the same strand, do not overlap. The only exception is that two bases of the *tRNA-Trp* gene are used for the stop codon of *ND2*. On the other hand, adjacent protein-coding genes always overlap if no tRNAs are present between them (See Table 1 and Figure 2), and one-nucleotide frameshifts are observed in the overlapping gene pairs. This observation strongly supports the idea that the tRNA genes located between peptide genes function as signals for the processing of transcripts (OJALA *et al.* 1981).

Protein-coding genes: Thirteen peptide genes were identified by their sequence similarity to other vertebrate mtDNAs. A translation using the mammalian mitochondrial code yielded proteins with lengths similar to that of other vertebrates (Table 2), which suggests that the same code is used in the lamprey (Table 3).

The sea lamprey mtDNA uses ATG for translational initiation of 12 protein-coding genes and GTG for *COI*, which is identical to the loach and chicken mt genomes. The use of GTG at the beginning of open reading frames in mitochondrial genomes is not unusual, because the rat and sea urchin mt genomes also use GTG as an initiation codon for *ND1* and *ATP8*, respectively. All sense codons except GCG are used (Table 3). The absence of a GCG codon can be attributed to the low frequency of G at synonymous sites in this genome. Six genes (*CYT b, ND4, ND5, ND6, ATP6* and *COI*) employ AGA as termination codons whereas another six genes (*NDI, ND3, ND4L, ATP8, COII* and *COIII*) use TAA, and *ND2* uses TAG.

Nucleotides are read by two reading frames in four cases of gene overlap (*ATP8/ATP6, ATP6/COIII, ND4L/ND4* and *ND5/ND6*). In human mtDNA, only one mRNA is found for *ATP8/ATP6* and *ND4L/ND4*, respectively (ANDERSON *et al.* 1981). It is possible that the lamprey genome has the same transcriptional process for *ATP8/ATP6* and *ND4L/ND4* genes. Although the structure of transcripts for the other overlapping genes is not known, they also may have one mRNA for each pair.

Base composition: The base composition of mitochondrial genomes varies among animal taxa (BROWN

TABLE 2

Lengths of animal mitochondrial genes

Gene	Species							
	Human	Mouse	Chicken	Frog	Loach	Sea Lamprey	Sea Urchin	Drosophila
Control region	1043	879	1227	2134	896	491	121	1077
12S rRNA	954	975	819	951	989	900	976	867
16S rRNA	1559	1582	1621	1621	1680	1621	1530	1326
Cytb	1141	1144	1140	1140	1141	1191	1157	1137
ND1	956	957	975	970	975	966	969	975
ND2	1042	1036	1038	1039	1047	1044	1059	1025
ND3	346	345	348	342	351	351	351	354
ND4	1378	1378	1377	1384	1383	1377	1380	1339
ND4L	297	294	294	297	297	291	294	290
ND5	1811	1824	1818	1815	1837	1797	1914	1720
ND6	528	519	519	513	522	519	495	525
COI	1541	1545	1548	1549	1551	1554	1554	1536
COII	684	684	684	688	691	690	690	685
COIII	784	784	783	781	768	786	783	789
ATP6	679	681	681	679	684	714	690	674
ATP8	207	204	165	168	168	168	168	162
Total	16,569	16,295	16,775	17,553	16,558	16,201	15,650	16,019

Values expressed as base pairs.

1985). Moreover, the nucleotide composition of the two strands is heterogeneous (BROWN 1983). The wobble positions of fourfold degenerate sites, which may be relatively free of selective constraints on base substitution, probably best reflect the mutational spectrum of the genome. The base composition of fourfold degenerate sites varies greatly among deuterostomes and usually differs from the average composition of each genome (Table 3 and 4). These degenerate sites may be useful for understanding the underlying pattern of evolution of mitochondrial genes (N. T. PERNA and T. D. KOCHER, unpublished results).

We tested whether the base frequencies of lamprey differed from other deuterostome mitochondrial genomes using the method of RZHETSKY and NEI (1995). The lamprey composition is significantly different ($P < 0.01$) from both urchin and frog for both the entire and the third positions of protein-coding genes. The frequency of T (32.7%) is the highest yet observed among deuterostomes (Table 4).

In addition to the intergenomic variation, there is some evidence of intragenomic variations in base composition in the lamprey genome. The composition of fourfold degenerate codons among lamprey protein-coding genes is heterogeneous (Table 5). The range of difference in %GC is from 35.6% (*CYT b*) to 19.9% (*COI*). Even though the base substitutions in the wobble positions do not result in the substitution of amino acids, differences in base frequency may be the result of varying selective constraints from gene to gene. Because mtDNA generally uses a single tRNA for each codon family (except for Leu and Ser), tRNA abundance is not a factor in codon usage. However, codons may have different affinities for the anticodon, ultimately resulting in differences in translational efficiency. Differences in the distribution of bases between the two mitochondrial strands is called "skew" (N. T. PERNA and T. D. KOCHER, unpublished results). The lamprey genome exhibits the typical vertebrate pattern of positive GC-skew and negative AT-skew for genes located on the first strand (Table 5). The opposite pattern is found in the *ND6* gene, which is located on the second strand. The *ND6* gene uses more Gs at the fourfold degenerate sites than protein-coding genes on the first strand (Table 3). On the first strand, GC-skew varies from −1.00 (*ATP8*) to −0.60 (*ND2* and *ATP6*). AT-skew ranges from −0.11 (*ATP8*) to 0.25 (*ND5*). It has been proposed that the extent of GC-skew might be associated with the length of time a DNA segment is single stranded during replication, because single-stranded DNA is more easily subject to deamination (W. K. THOMAS, personal communication). AT-and GC-skews vary in parallel around the lamprey genome (Figure 3), but because the location of the second strand replication origin is not known for the lamprey, it is not possible to relate these patterns to single-stranded exposure.

Transfer RNA genes: Twenty-two tRNA genes can be identified in the sea lamprey mt genome through analyses of sequence similarity and potential secondary structure. The size of lamprey tRNA genes varies from 65 to 74 bases showing high variation in the dihydrouridine (DHU) and T-Ψ-C arms. The DHU arm is the most variable in length. Most DHU arms have 4 bp in

TABLE 3

Codon usage in the lamprey mitochondrial genome

Code (AA)	No. (fraction)	Code (AA)	No. (fraction)	Code (AA)	No. (fraction)	Code (AA)	No. (fraction)
Protein-coding genes on the first strand							
TTG (Leu)	5 (0.01)	TGC (Ser)	2 (0.01)	TAG (End)	2 (0.15)	TGG (Trp)	8 (0.07)
TTA (Leu)	183 (0.45)	TCA (Ser)	87 (0.44)	TAA (End)	5 (0.38)	TGA (Trp)	98 (0.84)
TTT (Phe)	122 (0.30)	TCT (Ser)	67 (0.34)	TAT (Tyr)	63 (0.63)	TGT (Cys)	20 (0.60)
TTC (Phe)	96 (0.23)	TCC (Ser)	44 (0.22)	TAC (Tyr)	37 (0.37)	TGC (Cys)	13 (0.39)
CTG (Leu)	14 (0.03)	CCG (Pro)	4 (0.02)	CAG (Gln)	6 (0.06)	CGG (Arg)	1 (0.02)
CTA (Leu)	192 (0.49)	CCA (Pro)	113 (0.57)	CAA (Gln)	92 (0.94)	CGA (Arg)	35 (0.55)
CTT (Leu)	135 (0.34)	CCT (Pro)	49 (0.25)	CAT (His)	47 (0.45)	CGT (Arg)	20 (0.31)
CTC (Leu)	53 (0.13)	CCC (Pro)	32 (0.16)	CAC (His)	58 (0.55)	CGC (Arg)	8 (0.13)
ATG (Met)	24 (0.04)	ACG (Thr)	1 (0.00)	AAG (Lys)	9 (0.09)	AGG (End)	0 (0.00)
ATA (Met)	188 (0.35)	ACA (Thr)	125 (0.41)	AAA (Lys)	90 (0.91)	AGA (End)	5 (0.42)
ATT (Ile)	229 (0.42)	ACT (Thr)	103 (0.34)	AAT (Asn)	76 (0.53)	AGT (Ser)	31 (0.12)
ATC (Ile)	99 (0.18)	ACC (Thr)	73 (0.24)	AAC (Asn)	67 (0.47)	AGC (Ser)	22 (0.09)
GTG (Val)	4 (0.02)	GCG (Ala)	0 (0.00)	GAG (Glu)	10 (0.11)	GGG (Gly)	22 (0.11)
GTA (Val)	62 (0.39)	GCA (Ala)	97 (0.34)	GAA (Glu)	79 (0.89)	GGA (Gly)	82 (0.42)
GTT (Val)	65 (0.40)	GCT (Ala)	83 (0.29)	GAT (Asp)	29 (0.46)	GGT (Gly)	49 (0.25)
GTC (Val)	30 (0.19)	GCC (Ala)	102 (0.36)	GAC (Asp)	34 (0.54)	GGC (Gly)	41 (0.21)
ND6 gene on the second strand							
TTG (Leu)	7 (0.25)	TCG (Ser)	1 (0.10)	TAG (End)	0 (0.00)	TGG (Trp)	2 (0.67)
TTA (Leu)	16 (0.57)	TCA (Ser)	1 (0.10)	TAA (End)	0 (0.00)	TGA (Trp)	1 (0.33)
TTT (Phe)	15 (0.88)	TCT (Ser)	6 (0.60)	TAT (Tyr)	6 (0.75)	TGT (Cys)	7 (1.0)
TTC (Phe)	2 (0.12)	TCC (Ser)	0 (0.00)	TAC (Tyr)	2 (0.25)	TGC (Cys)	0 (0.00)
CTG (Leu)	1 (0.04)	CCG (Pro)	0 (0.00)	CAG (Gln)	0 (0.00)	CGG (Arg)	1 (0.25)
CTA (Leu)	2 (0.07)	CCA (Pro)	0 (0.00)	CAA (Gln)	0 (0.00)	CGA (Arg)	2 (0.50)
CTT (Leu)	2 (0.07)	CCT (Pro)	4 (1.0)	CAT (His)	1 (1.0)	CGT (Arg)	1 (0.25)
CTC (Leu)	0 (0.00)	CCC (Pro)	0 (0.00)	CAC (His)	0 (0.00)	CGC (Arg)	0 (0.00)
ATG (Met)	5 (1.0)	ACG (Thr)	1 (0.25)	AAG (Lys)	0 (0.00)	AGG (End)	0 (0.00)
ATA (Met)	0 (0.00)	ACA (Thr)	0 (0.00)	AAA (Lys)	0 (0.00)	AGA (End)	1 (1.0)
ATT (Ile)	3 (0.50)	ACT (Thr)	3 (0.75)	AAT (Asn)	2 (1.0)	AGT (Ser)	2 (0.20)
ATC (Ile)	3 (0.50)	ACC (Thr)	0 (0.00)	AAC (Asn)	0 (0.00)	AGC (Ser)	0 (0.00)
GTG (Val)	4 (0.13)	GCG (Ala)	0 (0.00)	GAG (Glu)	1 (0.25)	GGG (Gly)	11 (0.58)
GTA (Val)	11 (0.35)	GCA (Ala)	2 (0.13)	GAA (Glu)	3 (0.75)	GGA (Gly)	3 (0.16)
GTT (Val)	14 (0.45)	GCT (Ala)	13 (0.87)	GAT (Asp)	3 (0.75)	GGT (Gly)	3 (0.16)
GTC (Val)	2 (0.06)	GCC (Ala)	0 (0.00)	GAC (Asp)	1 (0.25)	GGC (Gly)	2 (0.11)

Fraction indicates relative codon usage with codon families.

TABLE 4

Base compositions of the third positions of fourfold degenerate codons and the entire first strand in deuterostome mtDNAs

	Fourfold sites (%)				First strand (%)			
Species	A	C	G	T	A	C	G	T
Cow	49.7	27.8	5.3	17.1	33.4	27.2	13.5	27.7
Mouse	54.8	21.3	4.6	20.0	34.5	24.4	12.3	27.2
Rat	52.5	28.7	3.3	15.5	34.1	26.2	12.5	28.7
Frog	45.5	20.7	5.5	28.3	33.1	23.5	13.5	30.0
Lamprey	43.2	20.5	3.5	32.7	32.2	23.8	13.5	30.4
Urchin	38.5	23.9	12.6	25.0	28.7	22.7	18.4	30.2

Only fourfold sites from genes encoded on the first strand were counted. Base frequencies at fourfold degenerate sites are heterogeneous among species ($P < 0.01$) and are more unequal than the overall base composition of the strand.

the stems, but the loops vary in size. For instance, *tRNA-Lys* has only one base for the loop. The *tRNA-Ser(AGY)* has no discernable DHU stem and loop. T-Ψ-C and anticodon arms have 5 bp in the stem regions with only a few mismatches. All anticodons are identical to other vertebrate mtDNAs.

TABLE 5

Compositional statistics for the third position of fourfold degenerate codons

Statistic	Gene													
	Cyt b	*ND1*	*ND2*	*COI*	*COII*	*ATP 8*	*ATP 6*	*COIII*	*ND3*	*ND4L*	*ND4*	*ND5*	Average[a]	*ND6*
Percentage GC	35.6	22.5	20.3	19.9	26.9	28.0	24.5	28.6	22.4	32.0	25.5	23.0	25.7	25.7
GC-skew	−0.74	−0.74	−0.60	−0.74	−0.86	−1.00	−0.60	−0.90	−0.85	−0.67	−0.79	−0.91	−0.78	0.65
AT-skew	0.13	0.24	0.21	0.10	0.19	−0.11	0.22	0.18	0.16	0.21	0.09	0.25	0.15	−0.37

[a] The average is calculated without ND6.

Ribosomal RNA genes: As found in other vertebrate mt genomes, the two ribosomal RNA genes are located between *tRNA-Phe* and *tRNA-Leu (UUR)* and separated by *tRNA-Val*. The lengths of the *12S* and *16S rRNA* genes are 900 bp and 1621 bp, respectively, about average for vertebrates (Table 2). In the *12S rRNA* gene, the size variation among vertebrates is mainly the result of a long segment of indels in the 3′ ends, whereas the *16S rRNA* gene often has a long string of indels in the 5′ end. Several small indels are observed in both genes.

The *12S* gene is more conserved among vertebrates than the *16S* gene. The sequence similarities between lamprey and loach ribosomal RNA genes are 64 and 53% in *12S* and *16S*, respectively.

Noncoding regions and evolution of the control region: The lamprey genome contains two major noncoding regions separated by the tRNA genes, Thr and Glu. The first noncoding region is 491 bp in length and contains a 28 bp A + T rich (93%) region in near the 5′ end. It also contains three copies of a 39-bp repeat (Figure 2).

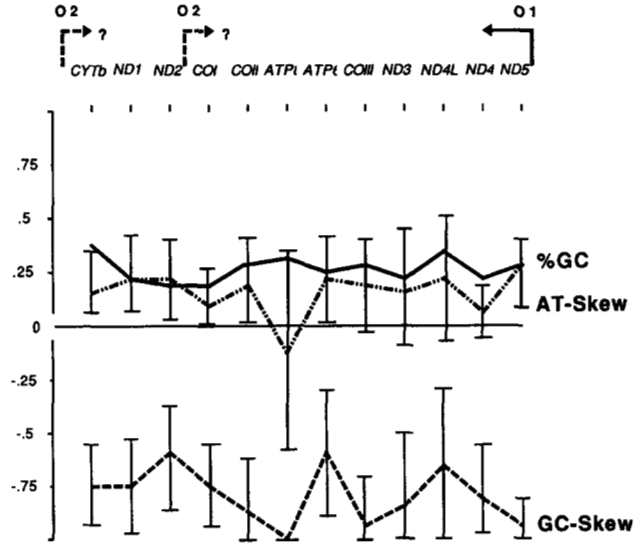

FIGURE 3.—The patterns of %GC, GC-skew and AT-skew in the lamprey mt genome. The genome is characterized by a low GC%, and strong GC-skew. There is no obvious relationship between the composition, AT-skew or GC-skew, and the locations of genes with respect to replicational origins. Bars represent 95% confidence intervals calculated from 1000 bootstrap resamplings.

We identified the conserved sequence block (CSB) -II and -III by comparison of the first noncoding segment to that of mouse (BIBB *et al.* 1981), loach (TZENG *et al.* 1992) and frog (ROE *et al.* 1985). The lamprey CSB-II is identical to that of the loach except for one indel. Because WALBERG and CLAYTON (1981) reported that the CSB-II and -III play critical roles for the process of heavy strand replication of mammalian mtDNAs, we conclude that the first noncoding region is most likely the control region, carrying most of regulatory sequences for replication. If so, the control region of lamprey mtDNA is much shorter than that of any other vertebrate mtDNA.

Unexpectedly, none of the central conserved sequence blocks (A–F) found in mammalian mtDNAs (SOUTHERN *et al.* 1988) can be identified in the putative control region of lamprey mtDNA. Even though the CSB-D block is well conserved from bony fishes to mammals (LEE *et al.* 1995), we could not identify homologous sequences in the lamprey. The absence of the central conserved region in the lamprey suggests that these conserved sequences are a recently evolved domain.

In other vertebrate mtDNAs, tRNAs located in and around the control region are thought to play a role in initiation and termination of D-loop DNA synthesis (SACCONE *et al.* 1985; CANTATORE *et al.* 1988; JACOBS *et al.* 1988). In this regard, we folded the repeats found in the putative control region into a secondary structure. The secondary structure is very stable (Figure 4A), suggesting that the tandem repeat in the first noncoding region may have the same regulatory function as the tRNA-like structure in other vertebrate mtDNAs.

The tRNA cluster, located between *ND2* and *COI*, in which the second-strand replication normally is initiated in other vertebrate genomes, lacks a noncoding segment in the lamprey genome. Several intergenic sequences are found in other regions of the molecule, most less than five nucleotides in length. However, there is an unusually long segment (23 bp) between genes for *ND1* and *tRNA-Ile*. This sequence does not fold into a secondary structure, so it is unlikely to func-

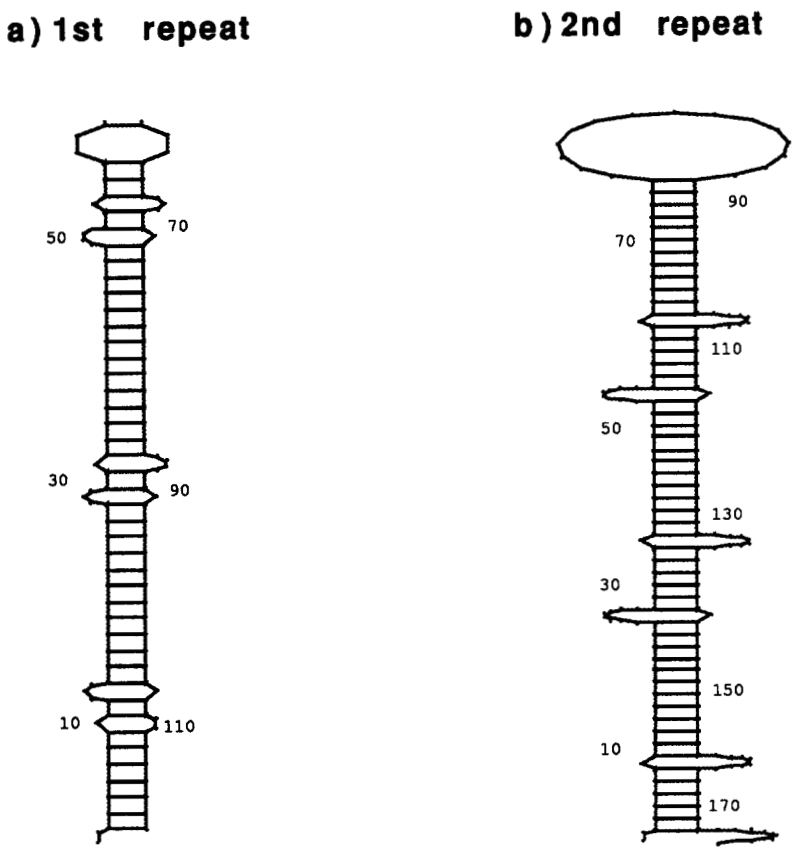

FIGURE 4.—The secondary structures of the tandem repeats in the two noncoding regions. Both structures contain long stems and are highly stable. There are two transitional substitutions among copies of the first repeat, but no variation among copies of the second repeat.

tion as the replication origin of the second strand. It is possible that the repeats in the second of the major noncoding regions function as the second strand origin (O2). In fact, replication of the second strand initiates near the control region in Drosophila mtDNA (CLARY and ATTARDI 1985). As another possibility, one of the tRNA genes between *ND2* and *COI* could play a dual role in translation and replication, as described in CANTATORE *et al.* (1988).

Many studies have reported tandem repeats found in the control region that are capable of forming highly stable secondary structures (WILKINSON and CHAPMAN 1991; ARNASON and RAND 1992). Although the origin of repetitive sequences and the precise mechanisms giving rise to the repeats are unknown, it has been proposed that short tandem repeats might arise by DNA slippage during replication. After the slipped-DNA-strand mispairings, deletion events are commonly observed in the eukaryotic genomes (LEVINSON and GUTMAN 1987). The motif of this repeat is fairly long (39 bp) but contains two submotifs of 12 bp each, suggesting that it might have arisen by simple DNA slippage. The slippage or duplication must have occurred several times to generate the present sequence.

The second noncoding region is 199 bp in length and consists almost entirely of seven copies of an AT-rich 27-bp string. The multiple copies of the repeat are identical and form highly stable secondary structures (Figure 4B), again suggesting they arose by slippage.

Phylogenetic implications of gene order: Because of the slow rate of gene rearrangements, the pattern of mitochondrial genome organization may provide information concerning the topology of deep phylogenetic divergences (BROWN 1985; MORITZ *et al.* 1987). Each phylum has a basic pattern of gene arrangement (Figure 5). In echinoderms and insects, gene order provides useful phylogenetic information at the class level (SMITH *et al.* 1993).

Two major mechanisms have been proposed for changes in mitochondrial gene order. Duplication of segments, followed by internal deletions, is probably a major mechanism (MORITZ *et al.* 1987). Duplications are frequently observed and seem to explain the unique order seen in birds (DESJARDINS and MORAIS 1990). CANTATORE *et al.* (1987) suggested that the rearrangement of tRNA genes also might occur by illicit priming of replication. This mechanism would lead to the accumulation of tRNA genes in a cluster near the replication origin. JACOBS *et al.* (1988) disputed the suggestion that the clustered tRNAs of echinoderms is a derived state. Clustered tRNAs are found in at least four echinoderm classes (SMITH *et al.* 1993) and so are not unique to urchins. Furthermore, the leucine tRNA that CANTATORE *et al.* (1987) have suggested to be recently

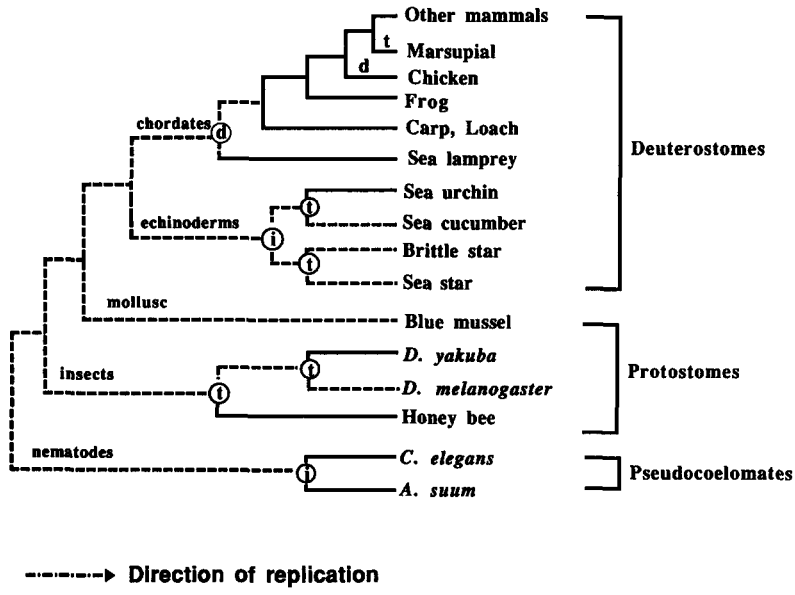

FIGURE 5.—The pattern of mt gene arrangements among animal phyla. The phylogeny is derived from the work of SIDOW and THOMAS (1994). The t, d and i, respectively, represent translocations of tRNA genes, duplication/deletion and inversion events within each phylum. Partially characterized mitochondrial genomes are indicated (- - -).

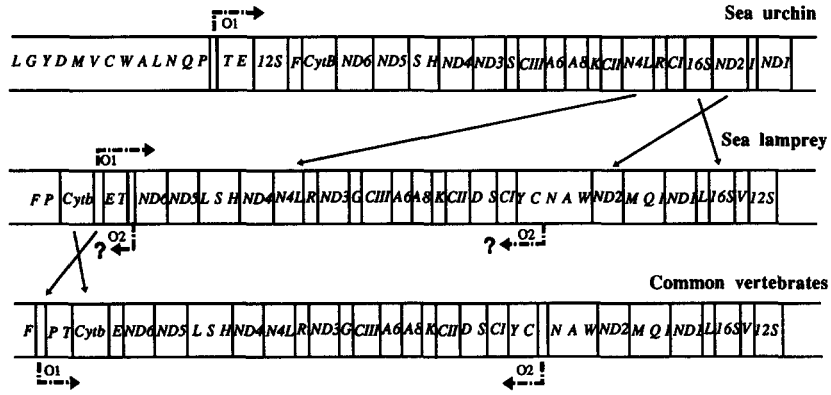

FIGURE 6.—Comparison of gene arrangement among three deuterostomes. Arrows indicate genes in different locations. The movements of tRNAs are not indicated. The gene order of protein genes in the three genomes is colinear except for *ND4L* in the sea urchin mtDNA. Replicational origins of the first and second strands, O1 and O2, respectively are indicated.

inserted near the urchin replication origin, is found in an identical position in the tRNA cluster of sea stars. JACOBS *et al.* (1988) suggest alternative hypotheses, in which tRNAs might have dispersed through the genome on the lineage leading to vertebrates. Neither of these mechanisms explains the inversion of segments observed among echinoderm classes (SMITH *et al.* 1993).

The lamprey sequence demonstrates that the major rearrangements that distinguish echinoderm and vertebrate mitochondrial genomes did not occur recently on the vertebrate lineage. The lamprey shows several changes in gene order relative to other vertebrates near the noncoding regions (Figure 6). The first major noncoding region is located between *ND6* and *tRNA-Thr* and is probably homologous to the control region of other vertebrates, because it contains sequences similar to CSB-II and CSB-III. The second noncoding region, located between *tRNA-Glu* and *CYT b,* consists almost entirely of repeated sequence. Two tRNA genes (Pro and Phe) that normally flank the control region, are located next to each other, between *CYT b* and *12S,* in the lamprey. At this point, it is impossible to determine whether the lamprey gene order represents an ancestral state for the vertebrate lineage or a uniquely derived state.

Even though the lamprey mtDNA sequence does not allow inference of the complete structure of the ancestral vertebrate gene order, it is clear that most features of vertebrate mt genomes were already well established at an early stage of evolution. Study of other distantly related chordates may shed light on the sequence of rearrangements by which echinoderm and vertebrate mitochondrial gene orders have arisen.

We thank S. SOWER for donation of lamprey tissues, J. CONROY for technical assistance and N. PERNA for help in data analysis. We also thank W. K. THOMAS and W. M. BROWN for their helpful comments during the early stage of analysis, A. RZHETSKY for providing his computer program and two anonymous reviewers for their comments. This work was supported in part by National Science Foundation grant to T.D.K. The complete sequence has been deposited in Genbank with accession No. U11880.

LITERATURE CITED

ANDERSON, S., A. T. BANKIER, B. G. BARRELL, M. H. L. DE BRUIJN, A. R. COULSON *et al.*, 1981 Sequence and organization of the human mitochondrial genome. Nature **290:** 457–465.

ARNASON, E., and D. M. RAND, 1992 Heteroplasmy of short tandem repeats in mitochondrial DNA of Atlantic cod, *Gadus morhua*. Genetics **132:** 211–220.

BIBB, M. J., R. A. ETTEN, C. T. WRIGHT, M. W. WALBERG and D. A. CLAYTON, 1981 Sequence and gene organization of mouse mitochondrial DNA. Cell **26:** 167–180.

BROWN, W. M., 1983 Evolution of animal mitochondrial DNA, pp. 62–88 in *Evolution of Genes and Proteins*, edited by M. NEI and R. K. KOEHN. Sinauer Associates, Sunderland, MA.

BROWN, W. M., 1985 The mitochondrial genome of animals, pp. 95–130 in *Molecular Evolutionary Genetics*, edited by R. J. MACINTYRE. Plenum Press, New York.

CABOT, E. L., and A. T. BECKENBACH, 1989 Simultaneous editing of multiple nucleic acid and protein sequences with ESEE. Comput. Appl. Biosci. **5:** 233–234.

CANTATORE, P., M. GADALETA, M. ROBERTI, C. SACCONE and A. C. WILSON, 198 Duplication and remoulding of transfer RNA genes during the evolutionary rearrangement of mitochondrial genomes. Nature **32:** 853–855.

CANTATORE, P., M. ROBERTI, G. RAINALDI, C. SACCONE and M. N. GADALETA, 1988 Clustering of tRNA genes in Paracentrotus lividus mitochondrial DNA. Curr. Genet. **13:** 91–96.

CLARY, D. O., and G. ATTARDI, 1985 The mitochondrial DNA molecule of *Drosophila yakuba:* nucleotide sequence, gene organization, and genetic code. J. Mol. Evol. **22:** 252–271.

CLAYTON, D. A., 1982 Replication of animal mitochondrial DNA. Cell **28:** 693–705.

CROZIER, R. H., and Y. C. CROZIER, 1993 The mitochondrial genome of the honeybee *Apis mellifera:* complete sequence and genome organization. Genetics **133:** 97–117.

DESJARDINS, P., and R. MORAIS, 1990 Sequence and gene organization of the chicken mitochondrial genome. J. Mol. Biol. **212:** 599–634.

DOWLING, T. E., C. MORITZ and J. D. PALMER, 1990 Nucleic acids II: restriction site analysis, pp. 250–317 in *Molecular Systematics*, edited by D. M. HILLIS and C. MORITZ. Sinauer Associates, Sunderland, MA.

FOREY, P., and P. JANVIER, 1993 Agnathans and the origin of jawed vertebrates. Nature **361:** 129–134.

GARESSE, R., 1988 *Drosophila melanogaster* mitochondrial DNA: gene organization and evolutionary considerations. Genetics **118:** 649–663.

HOELZEL, A. R., 1993 Evolution by DNA turnover in the control region of vertebrate mitochondrial DNA. Curr. Opin. Genet. Dev. **3:** 891–895.

HOFFMANN, R. J., J. L. BOORE and W. M. BROWN, 1992 A novel mitochondrial genome organization for the blue mussel, *Mytilus edulis*. Genetics **131:** 397–412.

HUBBS, C. L., and I. C. POTTER, 1971 Distribution, phylogeny and taxonomy, pp. 1–65 in *The Biology of Lampreys*, edited by M. W. HARDISTY, and I. C. POTTER. Academic Press, New York.

JACOBS, H. T., D. J. ELLIOTT, V. B. MATH and A. FARQUHARSON, 1988 Nucleotide sequence and gene organization of sea urchin mitochondrial DNA. J. Mol. Biol. **202:** 185–217.

JANKE, A., G. FELDMAIER-FUCHS, W. K. THOMAS, A. V. HAESELER and S. PÄÄBO, 1994 The marsupial mitochondrial genome and the evolution of placental mammals. Genetics **137:** 243–256.

LANSMAN, R. A., R. O. SHADE, J. F. SHAPIRA and J. C. AVISE, 1981 The use of restriction endonucleases to measure mitochondrial DNA sequence relatedness in natural populations. III. Techniques and potential applications. J. Mol. Evol. **17:** 214–226.

LEE, W.-J., J. A. CONROY, W. H. HOWELL and T. D. KOCHER, 1995 Structure and evolution of teleost mitochondrial control regions. J. Mol. Evol. (in press).

LEVINSON, G., and G. A. GUTMAN, 1987 Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol. Biol. Evol. **4:** 203–221.

MORITZ, C., T. E. DOWLING and W. M. BROWN, 1987 Evolution of animal mitochondrial DNA: relevance for population biology and systematics. Annu. Rev. Ecol. Syst. **18:** 269–92.

MOYLE, P. B., and J. J. CECH, JR., 1988 *Fishes. An Introduction to Ichthyology*. Prentice-Hall, Englewood Cliffs, NJ.

OJALA, D., J. MONTOYA and G. ATTARDI, 1981 tRNA punctuation model of RNA processing in human mitochondria. Nature **290:** 470–290.

OKIMOTO, R., J. L. MACFARLANE, D. O. CLARY and D. R. WOLSTENHOLME, 1992 The mitochondrial genomes of two nematodes, *Caenorhabditis elegans* and *Ascaris suum*. Genetics **130:** 471–498.

RZHETSKY, A., and M. NEI, 1995 Tests of applicability of several substitution models for DNA sequence data. Mol. Bio. Evol. (in press).

ROE, B. A., D.-P. MA, R. K. WILSON and J. F.-H. WONG, 1985 The complete nucleotide sequence of the *Xenopus laevis* mitochondrial genome. J. Biol. Chem. **260:** 9759–9774.

SACCONE, C. M., ATTIMONELLI and E. SBISÁ, 1985 Primary and higher order structural analysis of animal mitochondrial DNA, pp. 37 in *Achievements and Perspectives of Mitochondrial Research*, Vol. II, edited by E. QUAGLIARIELLO, E. C. SLATER, F. PALMIERI, C. SACCONE and A. M. KROON. Biogenesis. Elsevier, Amsterdam.

SAMBROOK, J., E. F. FRITSCH and T. MANIATIS, 1989 *Molecular Cloning: A Laboratory Manual* (Second ed.). Cold Spring Harbor Laboratory Press, New York, NY.

SIDOW, A., and W. K. THOMAS, 1994 A molecular evolutionary framework for eukaryotic model organisms. Curr. Biol. **4**: 596–603.

SMITH, M. J., A. ARNDT, S. GORSKI and E. FAJBER, 1993 The phylogeny of echinoderm classes based on mitochondrial gene arrangements. J. Mol. Evol. **36:** 545–554.

SOUTHERN, S. O, P. J. SOUTHERN and A. E. DIZON, 1988 Molecular characterization of a cloned dolphin mitochondrial genome. J. Mol. Evol. **28:** 32–42.

STOCK, D. W., and G. S. WHITT, 1992 Evidence from 18S ribosomal RNA sequences that lampreys and hagfishes form a natural group. Science **257:** 787–789.

TZENG, C.-S., C.-F. HUI, S.-C. SHEN and P.C. HUANG, 1992 The complete nucleotide sequence of the *Crossostoma lacustre* mitochondrial genome: conservation and variations among vertebrates. Nucleic Acids Res. **20:** 4853–4858.

WALBERG, M. W., and D. A. CLAYTON, 1981 Sequence and properties of the human KB cell and mouse L cell D-loop regions of mitochondrial DNA. NAR **9:** 5411–5421.

WILKINSON, G. S., and A. M. CHAPMAN, 1991 Length and sequence variation in evening bat D-loop mtDNA. Genetics **128:** 607–617.

Communicating editor: W-H. LI