Complete Sequence of the Mitochondrial DNA of the Annelid Worm *Lumbricus terrestris*

Jeffrey L. Boore and Wesley M. Brown

Department of Biology, University of Michigan, Ann Arbor, Michigan 48109-1048

Manuscript received December 11, 1994
Accepted for publication June 8, 1995

**ABSTRACT**

We have determined the complete nucleotide (nt) sequence of the mitochondrial genome of an oligochaete annelid, the earthworm *Lumbricus terrestris*. This genome contains the 37 genes typical of metazoan mitochondrial DNA (mtDNA), including ATPase8, which is missing from some invertebrate mtDNAs. ATPase8 is not immediately upstream of ATPase6, a condition found previously only in the mtDNA of snails. All genes are transcribed from the same DNA strand. The largest noncoding region is 384 nt and is characterized by several homopolymer runs, a tract of alternating TA pairs, and potential secondary structures. All protein-encoding genes either overlap the adjacent downstream gene or end at an abbreviated stop codon. In *Lumbricus* mitochondria, the variation of the genetic code that is typical of most invertebrate mitochondrial genomes is used. Only the codon ATG is used for translation initiation. *Lumbricus* mtDNA is A + T rich, which appears to affect the codon usage pattern. The DHU arm appears to be unpaired not only in tRNA*(AGN)*, as is typical for metazoans, but perhaps also in tRNA*(UCN)*, a condition found previously only in a chiton and among nematodes. Relating the *Lumbricus* gene organization to those of other major protostome groups requires numerous rearrangements.

---

**THE mitochondrial genome of multicellular animals consists of a closed circular DNA molecule except among some cnidarians, where it consists of one or two linear molecules (WARRIOR and GALL 1985; BRIDGE et al. 1992). Its usual size range is from 14 to 17 kb, but variations as large as 40 kb are known (MORITZ et al. 1987; SNYDER et al. 1987; WOLSTENHOLME 1992). Despite this size variation, there is little variation in gene content, since in all cases analyzed the larger size is due either to variation in the length of a noncoding region (BROWN 1985; HARRISON 1989; LAROCHE et al. 1990) or to iteration of some portion of the mitochondrial DNA (mtDNA) (MORITZ and BROWN 1986, 1987; WALLIS 1987; ZEVERING et al. 1991; AVEVEDO and HYMAN 1993; FULLER and ZOORUS 1993; STANTON et al. 1994).

Metazoan mtDNAs ordinarily contain 36 or 37 genes: 2 for ribosomal RNAs (r-rRNA) and 1-rRNA), 22 for tRNAs and 13 for subunits of multimeric proteins of the inner mitochondrial membrane [cytochrome oxidase subunits I–III (COI–III), cytochrome b apoenzyme (Cytb), ATP synthase subunits 6 and (usually) 8 (ATPase6 and ATPase8) and NADH dehydrogenase subunits 1–6 and 4L (ND1–6, ND4L)]. In addition, there is usually at least one sequence of variable length that does not encode any gene. In vertebrates (MONToya et al. 1982; Bogenhagen et al. 1985; King and Low 1987; Foran et al. 1988; Clayton 1991, 1992) and insects (Clary and Wolstenholme 1985a) these noncoding sequences are known to include elements that regulate mtDNA replication and transcription.

arrangements generally remain unchanged over long periods of evolutionary time, (3) stability of gene arrangements is best explained by a lack of genetic recombination and by the infrequency of rearrangements that maintain functional genomes, rather than by invoking selection for or against any particular gene order, (4) the great number of potential gene arrangements makes it very unlikely that different taxa would independently arrive at identical arrangements, and (5) with a few possible exceptions, gene arrangements seem relatively stable within major groups but variable between them. These characteristics make the comparison of mitochondrial gene arrangements a potentially powerful means for inferring phylogenetic relationships among major metazoan groups (Brown 1985; Möritz et al. 1987; Jacobs et al. 1988b; Smith et al. 1993; Boore and Brown 1994b; Boore et al. 1995).

In addition, many features of the molecular biology of mtDNA have been shown to vary and can be compared among metazoans (alternative translation start codons, unusual tRNA and rRNA structures, genetic code variations, features of mtDNA replication and transcription, etc.) (see Brown 1985; Wolstenholme 1992; Boore and Brown 1994b). While comparisons of these and other aspects of the nuclear genome are possible, they are currently more accessible from the ships among many major metazoan groups. To this end, the worm, protostome phyla, Arthropoda and Mollusca.

**Materials and Methods**

Mitochondrial DNA was isolated from the common earthworm, L. terrestris, and purified by cesium chloride-ethidium bromide centrifugation. A detailed restriction endonuclease cleavage map was constructed for this mtDNA using eight enzymes. DNA fragments from the enzymatic digests were resolved electrophoretically in both 1% agarose and 3.5% acrylamide gels, which allows visualization and accurate size estimation of fragments from a few bp to ~15 kb.

The methods we used to clone and sequence mtDNA and to identify individual mitochondrial gene sequences are described in Boore and Brown (1994a). The entire Lumbricus mtDNA was cloned by linearizing it at its unique BglII site at position 5900 (see Figure 1 and Appendix) and inserting it as a unit into the BamHI site of bacteriophage lambda EMBL3 DNA (BamHI cleavage leaves a BglII-compatible overhang). BamHI and SalI codigestion produces five DNA fragments from the mtDNA portion of this recombinant phage, of ~8.0, 4.3, 1.0, 0.93 and 0.76 kb. Since ligation of the compatible BglII and BamHI ends of the phage fails to regenerate either site, the terminal portions of the mtDNA insert (0.93 and 1.0 kb) were liberated by cleaving the SalI sites that flank the BamHI site of the vector. Each of these fragments was cloned into pbScript as in Boore and Brown (1994a). Cleavage maps of the recombinant plasmids, designated L.0.0, I.4.3, L.i.0, L.0.93 and L.0.76, respectively, were constructed for 20 other restriction enzymes. Further subcloning yielded a total of 14 additional clones (Figure 1).

To verify further that no portion of the mtDNA was lost in cloning, a fragment of ~500 bp spanning the BglII cloning site was amplified from Lumbricus mtDNA by PCR, using primers designed on the basis of the flanking sequences. The sequence of the PCR fragment was determined and verified as identical to that obtained from using the sequence of the two flanking plasmid clones. For PCR amplification, we employed 30 cycles of (30 sec at 94°C, 1 min at 55°C and 2 min at 72°C). Single-stranded templates for sequencing were generated by using unequal ratios of the amplification primers.

Open reading frames (ORFs) were determined using the program MacVector® (IBI, version 3.0) and the genetic code for invertebrate mtDNA. ORFs were identified based on the similarity of their inferred amino acid sequences to those of the corresponding mitochondrial genes of D. yakuba (Clary and Wolstenholme 1985a) and Katharina tunicata (Boore and Brown 1994a). In the case of ND4L, ND6 and ATPase8, identity was further confirmed by a comparison of hydrophilicity profiles (Kyte and Doolittle 1982) (see Figure 3). Transfer RNA genes were identified genetically by their potential to be folded into characteristic mitochondrial tRNA structures and specifically by their anticodon sequences. Ribosomal RNAs were identified by their similarity in sequence and potential secondary structure to the mitochondrial tRNAs of D. yakuba (Clary and Wolstenholme 1985b).

Sequence of the L. terrestris mitochondrial genome has been deposited in GenBank under accession number U24570.

**Results and Discussion**

**Genome Composition:** The mtDNA of the annelid worm L. terrestris is 14,998 bp in size. It contains no introns and, in keeping with its small size, intergenic regions are few and short; most of its genes abut directly or overlap.

Lumbricus mtDNA is 61.6% A + T, at the low end of the range for invertebrates. By contrast, mosquito (Mitchell et al. 1993) and Drosophila (Clary and Wolstenholme 1985a) mtDNAs are ~77% A + T, Apis mtDNA is 84.9% A + T (Crozier and Crozier 1993) and Katharina mtDNA is 69% A + T (Boore and Brown 1994a). A + T richness is similar to the sequenced portions of Mytilus mtDNA (62%) (Hoffmann et al. 1992). Both the coding and noncoding DNA strands have approximately equal proportions of purines and pyrimidines (0.087 on a scale where 0.0 is an equal ratio of pyrimidines and purines and 1.0 represents all purines or all pyrimidines on one strand).

**Gene Content and Organization:** Lumbricus mtDNA contains genes for 13 proteins (ND1–6, ND4L, COI–III, ATPase6, ATPase8 and Cytb) for 2 rRNAs (s-
**Lumbricus terrestris**

![Figure 1](image)

**FIGURE 1.**—The mitochondrial genome of the common earthworm, *L. terrestris*, with subcloning and sequencing strategy. Genes are designated by their products as abbreviated in the text except that A6 and A8 refer to subunits 6 and 8 of the mitochondrial ATP synthase, and tRNA genes are identified by the one letter code for the amino acid they specify; two tRNAs each for leucine and serine are further differentiated by the codon each recognizes. UNK (unknown) refers to the longest unassigned DNA. All genes are transcribed from left to right on this figure. Arrows below with upward-facing barbs mark sequence obtained with synthetic oligonucleotide primers designed to the *Lumbricus* sequence. Only cleavage sites for the restriction enzymes that were employed in cloning are mapped onto the scale bar (scale in nucleotides) according to the following abbreviations: A, BamHI; B, BclI; C, EcoRI; D, PstI; E, EcoRV; F, AvaI; PA, XbaI; T, SalI; X, XhoI.

**L.0.76** and for the 22 tRNAs typical of other metazoan mtDNAs. All genes are transcribed from the same DNA strand.

*Lumbricus* mtDNA contains an ORF corresponding to ATPase 6, which is not present in the mtDNA of the bivalve *Mytilus* (Hoffmann et al. 1992) and in those of nematodes (Wolstenholme et al. 1987; Okimoto et al. 1992). The separation of ATPase 8 from ATPase 6 in *Lumbricus* mtDNA is unusual; this has been reported otherwise only in the mtDNAs of land snails (Lecanidou et al. 1994; Terrett et al. 1994). In all other metazoans where it is found, it immediately precedes ATPase 6, which it often overlaps in another reading frame.

In two cases *(tRNA^6^-tRNA^8^, and tRNA^8^-tRNA^6^)* tRNA genes appear to overlap by two nt. Functionally, this overlap might be resolved in several ways: transcription from different promoters, alternative transcript processing resulting in alternating messages of full length and relaxed constraints on tRNA structure, such that some tRNAs are functional despite being deficient for some terminal nt. This last possibility appears unlikely because all *Lumbricus* mitochondrial tRNA genes, including these four, conform to the usual expectation of a matched seven member amino-acyl stem.

In their unprocessed form, tRNA sequences (or structures) have been suggested to serve as signals that mediate the precise cleavages that generate gene-specific messages from the polycistronic mitochondrial transcript (Battey and Clayton 1980; Ojala et al. 1980, 1981). In cases where protein-encoding genes are not separated by intervening tRNA genes, sequences capable of forming stem-and-loop structures may serve a similar purpose. These have been found at such gene boundaries in the mtDNAs of mouse, human and cow (see Bibb et al. 1981), *D. yakuba* (Clary and Wolstenholme 1985a) and *D. melanogaster* (de Bruijn 1983), a sea urchin (Cantatore et al. 1987), the nematodes *C. elegans* and *Ascaris suum* (Okimoto et al. 1992) and the chiton Katharina (Boore and Brown 1994a). In *Lumbricus* mtDNA there are three junctions between protein-encoding genes that lack intervening tRNAs: ND2-COI, ND6-Cytb and ND4L-ND4. The 100-nt sequences centered on each of these junctions were searched for all potential secondary structures with a stem of 5 nt or more and a loop size of 0–50 nt. Although several such potential structures were noted, none were in a constant position relative to the start or stop codons of flanking genes.
Gene arrangement compared with those of other protostomes: The superphylum Protostomia contains three major phyla: Annelida, Arthropoda and Mollusca. This is the first report of a mitochondrial gene arrangement for an annelid. The gene arrangements of several arthropod and molluscs have been published and are available for comparison. Representing arthropods, complete mitochondrial gene arrangements have been reported for four insects and for one crustacean: D. yakuba (CLARY and WOLSTENHOLME 1985a), D. melanogaster (GARESSE 1988), the mosquito Anopheles quadrivittatus (MITCHELL et al. 1993), the honeybee Apis mellifera (CROZIER and CROZIER 1993) and the brine shrimp Artemia franciscana (VALVERDE et al. 1994). These arthropod arrangements are quite similar; in the two Drosophilidae species they are identical, and those of mosquito, honeybee and brine shrimp differ, respectively, from Drosophila’s by 3, 11 and 2 tRNA gene positions.

Discounting differences in the positions of tRNA genes, which are numerous, a minimum of six rearrangements is required to interconvert the mitochondrial gene arrangements of several arthropods between these three major phyla: Annelida, Arthropoda and Mollusca. The arthropod arrangements are quite similar; in the two Drosophilidae species they are identical, and those of mosquito, honeybee and brine shrimp differ, respectively, from Drosophila’s by 3, 11 and 2 tRNA gene positions.

Gene arrangement for an annelid. The gene arrangement of several arthropod species they are identical, and those of mosquito, honeybee and brine shrimp differ, respectively, from Drosophila’s by 3, 11 and 2 tRNA gene positions.

Representing molluscs, complete mitochondrial gene arrangements have been reported for the chiton K. tunica (class Polyplacophora) (BOORE and BROWN 1994a), the gastropod O. purpura, is very similar to Katharina’s (T. COLLINS, L. DAEHLER and W. BROWN, unpublished data). As evidence of how unique the Mytilus gene arrangement is, it shares only two gene boundaries with Lumbricus and Cardiaลา (BOORE and BROWN 1994a) and the brine shrimp P. purpura, is very similar to Katharina’s (T. COLLINS, L. DAEHLER and W. BROWN, unpublished data), the arrangement found in pulmonate gastropods appears to be as unique as that of Mytilus. No pair of genes has the same relative arrangement in Lumbricus and Cardiaลา, and only one gene pair is identically arranged in Lumbricus and Albinaria: tRNA-1-rRNA-tRNA-ND1, a sharing that is common among metazoan mtDNAs.

Initiation and termination of translation: Each of the 13 protein-encoding genes is initiated by an ATG codon. This is unusual among animal mtDNAs, which usually employ two to several alternative translation initiation codons (ANDERSON et al. 1981, 1982; BIBB et al. 1981; OJALA et al. 1981; CLARY and WOLSTENHOLME 1985a; OKIMOTO et al. 1990; HOFFMANN et al. 1992; WOLSTENHOLME 1992; BOORE and BROWN 1994a, b).

Unlike the easily inferred initiation points, the termination points of the protein-encoding genes in Lumbricus mtDNA are difficult to infer. None of these genes has a TAA or TAG (stop) codon that does not overlap the adjacent downstream gene. Abbreviated stop codons (T or TA), which are converted to TAA codons by polyadenylation after transcript cleavage, are common in animal mtDNAs (OJALA et al. 1980; ANDERSON et al. 1981; OJALA et al. 1981; reviewed in WOLSTENHOLME 1992). In Lumbricus mtDNA, all protein genes may end with abbreviated stop codons, although a few could end with complete stop codons if an overlap with a directly adjacent downstream gene is allowed (Table 1; Appendix). Overlapping genes could be resolved by differential transcript processing, with alternate cleavages resulting in some complete messages for both genes, or by translation of a bicistronic mRNA when both genes encode proteins. Overlapping genes occur commonly in metazoan mtDNAs, but the mechanism(s) used to resolve them or their products are unknown. ND4L and ND4 appear to overlap by 5–7 nt, depending on whether ND4L ends with T, TA or TAA. As
Annelid Mitochondrial Genome

**Katharina tunicata**

![Diagram of mitochondrial genome arrangement for Katharina tunicata](image1)

**Lumbricus terrestris**

![Diagram of mitochondrial genome arrangement for Lumbricus terrestris](image2)

**Drosophila yakuba**

![Diagram of mitochondrial genome arrangement for Drosophila yakuba](image3)

**FIGURE 2.**—The mitochondrial genome arrangements of an annelid, *L. terrestris*, a chiton, *K. tunicata* (Boore and Brown 1994a), and an insect, *D. yakuba* (Clary and Wostenholme 1985a). Each has been arbitrarily linearized at the gene for COI, and the rearrangements needed to interconvert two of the three pairings are shown. Translocations are indicated by arrows connecting homologous genes or blocks of genes (blocks are designated by a heavy bar); inversions are indicated similarly, with the addition of a circular arrow. Only rearrangements involving rRNA and protein-encoding genes are shown; differences in tRNA gene positions are numerous but not depicted. All genes of Lumbricus mtDNA are transcribed from left to right, as are all genes of Katharina and Drosophila mtDNA other than those designated by underlining to signify leftward transcription. Abbreviations are as noted for Figure 1 except A + T, which refers to the large A + T-rich region of Drosophila mtDNA.

shown in the APPENDIX, the first three amino acids of ND4 (met-leu-lys) in Lumbricus are the same as in Drosophila, suggesting that the initiation point inferred for ND4 is correct. Similarly, Lumbricus ND4L ends with lys-cys, as does ND4L in Katharina, suggesting that the inferred termination point of ND4L is also correct. By contrast, use of the first available abbreviated stop codon upstream of the one inferred for ND4L would result in truncation of its protein by seven amino acids.

ND3, ND5, ATPase6 and COII end with complete stop codons that, in each case, overlap an adjacent tRNA gene by 1–2 nt. Thus, although it is possible to infer their 5' termini with confidence, it is not possible to do so for their 3' termini.

Six protein-encoding genes have been inferred to end with abbreviated stop codons because extending their reading frames to complete stop codons requires hypothesizing significant overlaps with downstream genes. Whereas (see APPENDIX) the abbreviated stop codon inferred for ATPase8 abuts directly and the 3' end of its protein aligns well with those of other species, the first complete stop codon is 20 nt inside RNA2 and its use would result in a protein six amino acids longer. As inferred here, ND1 abuts tRNA16 directly, producing a protein 16 amino acids shorter than the one inferred by Clary and Wostenholme (1985a) for Drosophila. There is a complete downstream stop codon inside tRNA16 that, if used, would generate a protein 14 amino acids longer. Although this protein would be similar in length to ND1 in Drosophila, there is no significant sequence similarity at the 3' end, and use of this stop codon would require ND1 to overlap tRNA16 by 44 nt. Similarly, if ND2 is constrained to terminate at a complete instead of an abbreviated stop codon, the ND2
protein would be extended by seven amino acids. Again, this would yield a protein similar in length to ND2 in Drosophila, but without significant 3' sequence similarity, and would require ND2 to overlap COI by 23 nt. The abbreviated stop codons for ND4, ND6 and COIII directly abut adjacent genes and, if used, would result in proteins whose carboxyl ends have significant sequence similarity to their homologues in Drosophila. If the reading frames of ND4, ND6 and COIII are constrained to extend to complete stop codons, respective overlaps of 14, 38 and 10 nt result. Finally, if COI ends at the abbreviated stop that directly abuts tRNA\textsuperscript{am}, its corresponding protein is two amino acids longer than that of Drosophila but two amino acids shorter than that of Katharina. Constraining COI to end at a complete stop codon would require a 35-nt overlap with tRNA\textsuperscript{am}.

Based on DNA sequence alone, the 3' terminus of Cytb is the most difficult to infer. We have, therefore, designated the abbreviated stop codon that would result in a Cytb protein of length most similar to that of Drosophila and Katharina at the Cytb terminus. The Cytb-containing ORF extends another 56 nt, to a TAA codon in a region that is otherwise unassigned (see APPENDIX). If Cytb actually extends to this TAA codon, only 10 nt would separate it from tRNA\textsuperscript{am}. The 18 additional amino acids that would be encoded have no similarity to the inferred carboxyl end of Cytb of either Drosophila or Katharina.

**Protein-coding genes:** Lumbricus mtDNA encodes the 13 proteins that are typical of most metazoan mtDNAs: COI–III, ND1–6, ND4L, ATPose\textsubscript{6}, ATPase\textsubscript{8} and Cytb. Together, these genes comprise 11,106 nt (74%) of the total mitochondrial genome.

The inferred lengths of most of the proteins are very similar among Lumbricus, Katharina and Drosophila (see Table 2). Two proteins vary significantly in length. In Lumbricus, ND1 is 4.9% shorter than in Drosophila and 2.5% shorter than in Katharina, and ND6 is 10.3% shorter than in Drosophila and 6.0% shorter than in Katharina. No other proteins are inferred to vary by >3% in length among these three species.

Table 2 shows percentages of amino acid identity in pairwise comparisons for the mitochondrial protein-encoding genes of Lumbricus, Katharina and Drosophila, each representative of a major protostome phylum. The most conserved gene is COI, followed by COIII and Cytb, and then by COII. These results are nearly identical to those obtained from comparisons among other distantly related animals, for example, between sea urchin and vertebrate mtDNAs (JACOBS \textit{et al.} 1988a). The consistency of these percentages for each gene in these comparisons is remarkable. The degree of conservation of the amino acid sequences is likely to be related to the intensity of purifying selection on each of these proteins.

A combination of small size and low degree of sequence conservation make the identifications of ATPase\textsubscript{8}, ND4L and ND6 difficult. To confirm the sequence-based identifications, hydrophilicity profiles of the proteins (ROYCE and DOOLITTLE 1982) were compared to those of their homologues in Katharina and Drosophila (Figure 3). Despite a low level of primary sequence conservation, the hydrophilicity profiles among species are very similar, confirming the identification of these genes and suggesting that hydrophilicity is an important factor in patterns of amino acid replacement.

The five most common amino acids among the mitochondrially encoded proteins of Lumbricus are leu (15.1%), ser (9.8%), ile (8.0%), ala (7.5%) and phe (7.0%). The least common amino acids are cys (1.1%), arg (1.7%) and glu (1.8%). Selection for A + T richness appears to affect the codon usage pattern, with 65.3% of codons ending in A or T (Table 1). The frequency of codons with C in the third position is 24.4%, approximately as expected, but codons ending in G, at 10.3%, are very underrepresented. Three of the four least commonly used codons (TCG, CAG, CGG and CGC) end with G.

**Genetic code:** Lumbricus mtDNA appears to share the variations of the mitochondrial genetic code that are typical of other invertebrate mtDNAs (see WOLSTENHOLME 1992), and all codons except the stop codons TAA and TAG are used in its mitochondrial protein genes. For the reasons that follow, it is likely that AGA and AGG code for serine, TGA for tryptophan, and ATA for methionine.


AGA codes for serine in D. yakuba, but its mtDNA has no AGG codons (Clary and Wolstenholme 1985a).

The mitochondrial genetic code of Lumbricus appears to be like those of other invertebrates, with AGN-specifying serine. AGG occurs 13 times in the protein-encoding genes of Lumbricus mtDNA. When aligned with Drosophila mitochondrial proteins, six of these correspond to serine, three to glycine, and no more
FIGURE 3.—Hydrophilicity profiles (KYTE-DOOLITTLE scale) (KYTE and DOOLITTLE 1982) of the ND4L, ND6 and ATPase8 proteins of L. terrestris, K. tunicata and D. yakuba. The hydrophilicity window was set to seven. Numbers below each profile designate amino acid positions of each protein.

than two to any other amino acid; when aligned with Katharina mitochondrial proteins, five correspond to serine, four to glycine, and no more than one to any other amino acid. Similarly, there are 63 AGAs in Lumbricus mtDNA protein-encoding genes, 31 and 26 of which correspond to serine when aligned to Drosophila- and Katharina-inferred protein sequences, respectively, and no more than seven to any other amino acid in either species. None of the AGG or AGA codons found in Lumbricus protein genes correspond to arginine residues in either Katharina or Drosophila.

TGA, like TCG, codes for tryptophan: TGA is a stop codon in the universal genetic code and in the mtDNA of plants but codes for tryptophan in the mtDNA of fungi (CUMMINGS and DOMENICO 1988), protozoans (ZIAE and SUYAMA 1987; PRITCHARD et al. 1990) and all metazoans examined (see JUKES and OSAWA 1990). The codon TGA occurs 77 times in the protein-encoding genes of Lumbricus mtDNA. When aligned with protein-encoding genes of Katharina, these correspond to TGA codons 37 times, to TGG codons 24 times and to no other codon more than six times, indicating that TGA and TGG both code for tryptophan in Lumbricus. Lumbricus protein-encoding genes contain 24 TGG codons, of which four correspond to TGG codons, 14 to TGA codons and none to more than two of any other codon in Katharina mtDNA. None of the Lumbricus TGA or TGG codons correspond to either TGC or TGT (cysteine) codons of Katharina mtDNA. The frequent correspondence of TGG and TGA codons indicates that they are synonymous, and their lack of correspondence to TGC or TGT codons indicates that the latter codons specify a different amino acid.

ATA codes for methionine: In the universal genetic code, ATT, ATC and ATA code for isoleucine, while ATG alone codes for methionine. In the mitochondria of many organisms, ATA has changed specificity and also codes for methionine. The codon ATA is used 194 times in Lumbricus protein-encoding genes. In comparisons with the homologous proteins of Katharina, 49 of these correspond to methionine (41 to ATA and eight to ATG codons), 20 to isoleucine (16 to ATT and four to ATC), and none to any other amino acid except leucine more than 12 times. Lumbricus ATA codons correspond in position to leucine codons in Katharina mtDNA 54 times. It seems unlikely, however, that ATA codes for leucine in Lumbricus mtDNA, since no tRNA is found with a TAT anticodon, as would be required to uniquely discriminate ATA from other members of the ATN codon family. The high correspondence with leucine is likely to be an artifact, due to the unusually high frequency of leucine codons in Katharina mtDNA (BOORE and BROWN 1994a).

Code evidence from tRNA anticodons: Sharing of the same mitochondrial genetic code among Lumbricus, Mytilus, Katharina and Drosophila is also suggested by the anticodon similarities of their tRNA genes. In this regard, Lumbricus mtDNA is distinctive in having the anticodon TCT, rather than GCT, for tRNAser(AGN). A TCT anticodon might enable this tRNA to discriminate the codons AGA and AGG from the AGN codon family and thus suggest a changed mitochondrial genetic code in annelids. However, we could find no tRNA gene with the GCT anticodon that would be required to specifically recognize AGT and AGC codons. Further, the TCT anticodon is consistent with the presence of a T at the third (wobble) position in all Lumbricus tRNA genes whose products recognize fourfold degenerate codons (discussed below). The mitochondrial tRNAser(AGN) genes of Apis (CROZIER and CROZIER 1993) and Caeno-
The anticodon of Lumbricus mitochondrial tRNA genes is otherwise unexceptional, with only that for tRNAIF differing from those in Drosophila. In Lumbricus, as in Mytilus and Katharina, this anticodon is TTT, whereas in Drosophila it is CTT. This anticodon presumably recognizes the codons AAA and AAG, and a T in the wobble position is expected to do this most efficiently.

Ribosomal RNAs: Lumbricus mtDNA contains the genes for the small and large rRNAs of the mitochondrial ribosomal subunits. The s-rRNA and 1-rRNA genes are 59.6 and 64.9% A+T, respectively, about the same as for the mtDNA overall. Based only on DNA sequence analysis the precise boundaries of these genes remain uncertain, but the assumption that these genes extend to the exact boundaries of their flanking genes gives rRNAs of similar length and good primary sequence match to those of Drosophila and Katharina. The rRNA gene sequences, including all nt delimited by the flanking genes, can be folded into potential secondary structures similar to those proposed for the s-rRNA and 1-rRNA of Drosophila (CLARY and WOLSTENHOLME 1985b) (data not shown). Interpreted in this way, the s-rRNA is 785 nt in length, compared with 826 nt for Katharina and 789 nt for Drosophila, and the 1-rRNA is 1245 nt, compared with 1275 nt for Katharina and 1326 nt for Drosophila. Sizes of the s-rRNA and 1-rRNA of Drosophila and Katharina, and the ERNA in length, are not significantly different from those of other organisms with similar mtDNA size and rRNA gene content. Only C-and T-containing rRNAs are found in these organisms, and these are the only rRNAs that show significant sequence conservation among different species. The s-rRNA of the mitochondrial ribosomal subunits is 59.6% A+T, whereas in Drosophila it is 53.9%. The s-rRNA of the mitochondrial ribosomal subunits is 64.9% A+T, whereas in Drosophila it is 62.1% A+T.

Transfer RNAs: Lumbricus mtDNA encodes the 22 tRNA genes that are typical of metazoans (excluding cnidarians). The anticodons are identical with those of the mitochondrial tRNAs of Drosophila, Katharina and Mytilus, except for tRNAase't (AGG) and WAIF (see above). Only G and T occur in the wobble position of Lumbricus mitochondrial tRNA anticodons, other than in tRNAase't. In tRNAs that selectively recognize codons ending in C or T, G is present; in those that either selectively recognize codons ending in A or G or show no selectivity, T is present. In the genetic code of invertebrate mtDNA, there are seven amino acids coded by twofold degenerate codons ending in T or C (an, arg, glu, his, ile, phe, tyr); G is in the wobble position of the anticodon of the WAIF gene. There are six twofold degenerate codon families ending in A or G (coding for gln, glu, leu(UUR), lys, met and trp) and nine fourfold degenerate codon families (coding for gln, glu, leu(CUN), val, ser(UCN), pro, thr, ala, arg, his). Only C-and T-containing rRNAs are found in these organisms, and these are the only rRNAs that show significant sequence conservation among different species. The s-rRNA of the mitochondrial ribosomal subunits is 59.6% A+T, whereas in Drosophila it is 53.9%. The s-rRNA of the mitochondrial ribosomal subunits is 64.9% A+T, whereas in Drosophila it is 62.1% A+T.

Based only on DNA sequence analysis the precise boundaries of these genes remain uncertain, but the assumption that these genes extend to the exact boundaries of their flanking genes gives rRNAs of similar length and good primary sequence match to those of Drosophila and Katharina. The rRNA gene sequences, including all nt delimited by the flanking genes, can be folded into potential secondary structures similar to those proposed for the s-rRNA and 1-rRNA of Drosophila (CLARY and WOLSTENHOLME 1985b) (data not shown). Interpreted in this way, the s-rRNA is 785 nt in length, compared with 826 nt for Katharina and 789 nt for Drosophila, and the 1-rRNA is 1245 nt, compared with 1275 nt for Katharina and 1326 nt for Drosophila. Sizes of the s-rRNA and 1-rRNA of Drosophila and Katharina, and the ERNA in length, are not significantly different from those of other organisms with similar mtDNA size and rRNA gene content. Only C-and T-containing rRNAs are found in these organisms, and these are the only rRNAs that show significant sequence conservation among different species. The s-rRNA of the mitochondrial ribosomal subunits is 59.6% A+T, whereas in Drosophila it is 53.9%. The s-rRNA of the mitochondrial ribosomal subunits is 64.9% A+T, whereas in Drosophila it is 62.1% A+T.
### TABLE 2
Comparisons of the inferred mitochondrial proteins of *L. terrestris*, *K. tunicata* and *D. yakuba*

<table>
<thead>
<tr>
<th>Protein</th>
<th>No. of amino acids</th>
<th>Percent amino acid identity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Lumbricus</td>
<td>Katharina</td>
</tr>
<tr>
<td>ATPase 6</td>
<td>231</td>
<td>290</td>
</tr>
<tr>
<td>ATPase 8</td>
<td>53</td>
<td>53</td>
</tr>
<tr>
<td>COI</td>
<td>513</td>
<td>513</td>
</tr>
<tr>
<td>COII</td>
<td>228</td>
<td>229</td>
</tr>
<tr>
<td>COIII</td>
<td>259</td>
<td>259</td>
</tr>
<tr>
<td>Cytb</td>
<td>379</td>
<td>379</td>
</tr>
<tr>
<td>ND1</td>
<td>308</td>
<td>316</td>
</tr>
<tr>
<td>ND2</td>
<td>335</td>
<td>335</td>
</tr>
<tr>
<td>ND3</td>
<td>117</td>
<td>120</td>
</tr>
<tr>
<td>ND4</td>
<td>447</td>
<td>442</td>
</tr>
<tr>
<td>ND4L</td>
<td>98</td>
<td>100</td>
</tr>
<tr>
<td>ND5</td>
<td>574</td>
<td>571</td>
</tr>
<tr>
<td>ND6</td>
<td>156</td>
<td>166</td>
</tr>
</tbody>
</table>

Inferred size and percentage amino acid identity is compared for each of the 13 mitochondrial protein coding genes (data from Boore and Brown 1994a; Clary and Wolstenholme 1985a). Sequences were aligned pairwise using the PAM250 matrix. Percentages are the number of identical paired amino acids divided by the average length of the protein sequences compared.

ser(AGN) and gly]; T is in the wobble position of the anticodon of all tRNAs recognizing these codons, except for tRNAser. The tRNAser gene has the anticodon CAT; C in the wobble position apparently allows discrimination of ATA and ATG from ATT and ATC, the latter two of which are recognized by the GAT anticodon of tRNAser.

Although mitochondrial tRNAs are known to deviate from many of the features that are invariant in nuclear-encoded tRNAs, those inferred from the Lumbricus mitochondrial genes are surprisingly uniform in a number of these features (Figure 4). For example, the aminoacyl stem is 7 bp; in six of the tRNAs a single mismatch occurs in this stem. Also, the first of the 2 nt separating the aminoacyl stem from the DHU arm must be a T in 17 of the tRNAs and the second is a purine in 20 (an A in 14 and a G in six). The DHU arm has a stem of 3–4 bp and a loop of 3–7 nt, except in tRNAser, which has a stem of 2 bp and a loop of 9, and in tRNAasx, which has a 6-bp stem with one mismatch. The nt separating the DHU arm from the anticodon stem is an A in 13 tRNAs. Each tRNA has a 5-bp anticodon stem; in seven of the tRNAs, this stem contains a single mismatch. The 2 nt preceding the anticodon are always pyrimidines, with the nt nearest the anticodon being a T in 21 cases. The nt following the anticodon is always a purine (an A in 17 and a G in five). The variable arm is 4 nt in 17 tRNAs and ranges from 3 to possibly 6 nt (for one tRNAser(UUC) folding alternative). Finally, the TΨC arm has a loop of 3–7 nt and a stem of 3–5 bp, although two tRNAs are mismatched at the proximate nt pair.

In two cases, different tRNAs must be charged with the same amino acid: tRNAser(UUC)/tRNAlea(UUR) and tRNAser(AGN)/tRNAser(UUC). It is not presently known whether one or two amino-acyl synthetases are required for each pair of tRNAs. If only one, then some features common to the tRNA pair must be recognized by the synthetase. The two serine tRNAs share several nt within stems in addition to those that are generally conserved among all the tRNAs. Particularly notable is that six of the seven nt of the TΨC loop are identical. No other Lumbricus tRNA gene has an identical sequence, although that of tRNAasx differs by only 1–2 nt and is also similar in having 7 nt in this loop. Significant within-species similarity in the TΨC loops of the paired serine tRNAs has also been noted in the mtDNAs of Katharina (Boore and Brown 1994a), Mytilus (Hoffmann et al. 1992) and Drosophila (Clary and Wolstenholme 1985a).

The 12-nt DHU arm sequences of the two tRNAs coding for leucine are identical, as are those of the 2 nt separating the DHU arm from the amino-acyl stem, the 4-nt variable arm, and several nt within other stems. No other tRNAs match at more than two positions in the variable arm or have significant sequence identity to the DHU arms of these leucine tRNAs. The leucine tRNA pairs in Mytilus, Drosophila and Katharina mtDNA also have significant DHU arm sequence similarity with each other and with those of Lumbricus mtDNA, and in the Mytilus and Drosophila pairs, the variable arm is identical within each species. These similarities, and those noted above for the tRNAser genes, provide an inferential basis for the possibility that each pair is charged by a single synthetase.

Metazoan mtDNA usually encodes only one tRNAser, which presumably performs both initiator and elongator functions. Consistent with such a dual role, mosquito mitochondrial tRNAser contains both the U pre-
FIGURE 4.—Twenty-two mitochondrial tRNA sequences folded into their putative secondary structures. Bars indicate base pairing, with G-T pairs allowed. Dashed bars indicate additional potential base pairings within loops. The tRNA<sub>ser(UCN)</sub> is shown in both a structure with an unpaired DHU arm and an alternative structure with two of three DHU-arm nt paired. The unpaired alternative is truncated relative to the other by 2 nt at each end, marked with arrowheads on the paired alternative. The major tRNA features described in the text are illustrated for tRNA<sub>val</sub>.

Annelid Mitochondrial Genome

Figure 5.—Potential secondary structure at the 3′ end of the largest unassigned DNA sequence. Nucleotides marked with asterisks indicate an additional potential pairing. Numerals flanking the sequence indicate the number assigned to each terminal nt counting from the 5′ end of COI, as shown in the Appendix.

ceding the anticodon that is typical of conventional elongator tRNAs and the three G-C pairs in the anticodon stem that are typical of initiator tRNAs (Dubin and Hsuchen 1984). The tRNA\textsuperscript{as} of Lumbricus mtDNA shares these features, along with a perfect match to all nt of the DHU stem of mosquito, Drosophila, Katharina and bovine mitochondrial tRNA\textsubscript{met}.

In animal mtDNA, tRNA\textsubscript{ser}(AGN) typically has an unpaired DHU arm, while tRNA\textsubscript{ser}(UCN) has the more standard cloverleaf structure (see Garey and Wolstenholme 1989). In the mtDNA of nematodes (Wolstenholme et al. 1987) and of Katharina (Boore and Brown 1994a), tRNA\textsubscript{ser}(UCN) also has an unpaired DHU arm. Lumbricus mtDNA appears to encode yet another example of a tRNA\textsubscript{ser}(UCN) lacking a paired DHU arm, although an alternative structure with two of three paired nt in the DHU arm is possible (Figure 4). The unpaired alternative is truncated by 2 nt at each end, relative to the paired structure, such that it directly abuts tRNA\textsubscript{as} and leaves 1 nt between tRNA\textsubscript{as} and tRNA\textsubscript{met}(UCN), whereas the paired alternative requires overlap of each flanking tRNA by 2 and 1 nt, respectively. This observation, coupled with doubt as to the significance of a 2-or-3 nt match in the DHU arm, suggests that the unpaired alternative may be correct. If so, however, this would be the only tRNA in Lumbricus mtDNA without a paired DHU arm.

Unassigned DNA: The arrangement of genes in Lumbricus mtDNA is unusually compact, even among metazoans. Seven regions contain unassigned nucleotides. Five of these are very short: 1 nt separates the gene pairs tRNA\textsubscript{as}-COII, tRNA\textsubscript{gb}-COIII, tRNA\textsubscript{ser}(UCN)-tRNA\textsubscript{as}(UU\textsubscript{R}) and tRNA\textsubscript{met}(UCN)-ND1, and 2 nt separate tRNA\textsubscript{as} and ATPase6. A 66-nt unassigned region separates Cytb and tRNA\textsubscript{as}, if Cytb terminates on the abbreviated stop codon as predicted (see above). This 66-nt region is unremarkable either for nt composition or potential secondary structure.

The largest region of unassigned DNA is the 384 nt between tRNA\textsubscript{as} and tRNA\textsubscript{as}. It contains five ORFs of 25 or more amino acids, none of which have similarity to any of the mitochondrial proteins of Katharina or Drosophila. It is 64.4% A + T, close to the 61.6% A + T for the genome. Because it is typical of chordate and arthropod mtDNAs to have at least one large noncoding region that contains sequence elements controlling transcription and replication (Montoya et al. 1982; Bogenhagen et al. 1985; Clary and Wolstenholme 1985a; King and Low 1987; Bogenhagen and Romanel 1988; Foran et al. 1988; Clayton 1991), it is possible that the 384-nt sequence in Lumbricus also functions as a control region. However, such an assignment must first be demonstrated experimentally.

There are several notable features of this 384-nt DNA
sequence. Of the first 34 nt, 24 are in TA pairs, 12 of which are in a single run. Runs of TA pairs have been found in the noncoding portions of the mtDNAs of Caenorhabditis and Ascaris (OKIMOTO et al. 1992), D. yakuba (CLARY and WOLSTENHOLME 1985a) and Katharina (BOORE and BROWN 1994a). Also present are several homopolymer runs, most notably of 12 Gs and 19 As in the 5′ half of this region. By searching for possible stem-and-loop structures with a stem ≥8 nt and a loop of 0–50 nt, a large potential secondary structure could be found in the 3′ half of this region (Figure 5). It is particularly interesting that an alternative pairing is possible between one-half of one stem and part of a large loop (marked by asterisks in Figure 5). It has been suggested that such secondary structures might have regulatory roles (e.g., BROWN et al. 1986; WONG and CLAYTON 1985), but in the Lumbricus case, as in the others, whether the potential structure actually forms or plays any role is purely speculative at present.

Conclusions: Lumbricus terrestris is the first annelid for which the mtDNA sequence and gene organization have been determined. In most respects it is typical of metazoan mtDNAs: it encodes the same set of genes; it contains no introns, few intergenic nucleotides, and a single lengthy noncoding region; it exhibits the same genetic code variations that are seen in other invertebrate mtDNAs; and its tRNA genes are variable in both primary sequence and potential secondary structure. All genes of Lumbricus mtDNA are transcribed from one strand, as is seen in the mtDNA of some, but not all, metazoans.

In Lumbricus mtDNA, ATPase8 and ATPase6 overlap and are translated from a bicistronic primary sequence and potential secondary structure. ATPase8 overlaps ATPase6 in all other species examined except for the mtDNAs of the land snails Albinaria (LECANIDOU et al. 1994) and Cepaea (TERRETT et al. 1994). These genes often overlap in different reading frames; for example, in human mtDNA, ATPase8 and ATPase6 overlap and are translated from a bicistronic transcript with translation initiating alternatively at the 5′ end of the mRNA for ATPase8 or at an internal start codon for ATPase6 (OJALA et al. 1981). It is unknown whether this is also the mode of translation of these two genes in other organisms, although, if so, it could explain their frequent juxtaposition. Obviously, the gene arrangement in Lumbricus mtDNA precludes such a mode.

Comparisons of mitochondrial gene arrangements among protostomes for which there are data indicate that those of Lumbricus, Katharina (BOORE and BROWN 1994a,b), a prosobranch gastropod (Plicopurpura, T. COLLINS, L. DAEHLER and W. BROWN, unpublished data), Drosophila (CLARY and WOLSTENHOLME 1985a; GARESSE 1988), honeybee (CROZIER and CROZIER 1993), mosquito (MITCHELL et al. 1993) and brine shrimp (VALVERDE et al. 1994) can be interconverted by invoking relatively few rearrangement events, if tRNA gene rearrangements are ignored. However, interconversion of the pulmonate gastropod (LECANIDOU et al. 1994; TERRETT et al. 1994) and bivalve mollusc (HOFFMANN et al. 1992) arrangements, or their conversions to any of the arrangements found in the above taxa, require a very large number of rearrangement events. The extensive variability in gene arrangement seen among metazoan mtDNAs suggests that there is little or no selection on gene arrangement per se. Consequently, the paucity of mtDNA rearrangements over very long periods of time (as seen, e.g., within chordates, echinoderms, and arthropods) seems paradoxical. It has been hypothesized (BROWN 1985) that rearrangements are uncommon because genetic recombination is rare in metazoan mtDNA and the only successful events will be those in which breakage and rejoining occur at exact gene termini, a condition necessary for maintaining a functional genome. Despite these constraints, extensive and even radical variation in mitochondrial gene arrangement has occurred over evolutionary time, even within some groups [e.g., compare the arrangements of Mytilus, Katharina and Cepaea (HOFFMANN et al. 1992; BOORE and BROWN 1994a,b; TERRETT et al. 1994) or those of Caenorhabditis/Ascaris and Meloidogyne (OKIMOTO et al. 1991)].

The apparent lack of selection on mitochondrial gene arrangement, coupled with the large number of arrangements that are possible, makes them of potential use for inferring metazoan phylogeny (BOORE and BROWN 1994b; BOORE et al. 1995). Evolutionary relationships among phyla and classes of metazoans are largely unresolved due to difficulties in determining homologous structures among groups with very different body plans, convergence of functionally similar structures that can develop over evolutionary time, and the general lack of transitional forms in the fossil record, especially from strata ancient enough to document early metazoan history. It is our hope that when sufficient numbers of representative mitochondrial gene arrangements have been characterized, shared patterns will be found that will reliably indicate metazoan genealogy.

This work was supported by grants from the National Institutes of Health (GM-30144), the National Science Foundation (BSR-9107306, DEB-9206540), and the H. H. Rackham Graduate School of the University of Michigan. We thank LISA VAWTER for supplying purified mitochondrial DNA, L. LYNN DEAEHLER for expert technical assistance, and R. HOFFMANN, S. J. FUERSTENBERG and T. COLLINS for helpful comments on the manuscript.

Literature Cited

ANDERSON, S., M. H. L. DE BRUIJN, A. R. COULSON, I. C. EPERON, F. SANGER et al., 1982 Complete sequence of bovine mitochon-
Annelid Mitochondrial Genome


Communicating editor: A. G. CLARK

APPENDIX

Partly schematic representation of the complete sequence of the mtDNA of *L. terrestris*, oriented as in the gene map of Figure 1. Gene identities appear above the sequences and are designated as in Figure 1, except that three letter amino acid codes are used to denote the corresponding RNA. The extent of tRNA genes is overlined; protein- and rRNA-encoding genes are underlined, as are tRNA anticodon nts. All genes are transcribed from left to right. Inferred abbreviated stop codons are designated with *; the first in-frame complete stop codon for each protein-encoding gene is designated with (*). In each case, this would require overlap with the adjacent downstream gene (see text). Inferred amino acid sequences of proteins are shown with single letter amino acid codes centered below the second nt of the corresponding codon. Nucleotide numbers refer to the nucleotide below the last digit of the numeral, counting from the 5’ end of *COI*. Slash marks indicate nucleotides internal to genes that are omitted for brevity, with numerals designating the number of omitted nucleotides.