The Complete Mitochondrial DNA Sequence of the Horseshoe Crab *Limulus polyphemus*

Dennis V. Lavrov, Jeffrey L. Boore, and Wesley M. Brown

Department of Biology, University of Michigan

We determined the complete 14,985-nt sequence of the mitochondrial DNA of the horseshoe crab *Limulus polyphemus* (Arthropoda: Xiphosura). This mtDNA encodes the 13 protein, 2 rRNA, and 22 tRNA genes typical for metazoans. The arrangement of these genes and about half of the sequence was reported previously; however, the sequence contained a large number of errors, which are corrected here. The two strands of *Limulus* mtDNA have significantly different nucleotide compositions. The strand encoding most mitochondrial proteins has 1.25 times as many A's as T's and 2.33 times as many C's as G's. This nucleotide bias correlates with the biases in amino acid content and synonymous codon usage in proteins encoded by different strands and with the number of non–Watson-Crick base pairs in the stem regions of encoded tRNAs. The sizes of most mitochondrial protein genes in *Limulus* are either identical to or slightly smaller than those of their *Drosophila* counterparts. The usage of the initiation and termination codons in these genes seems to follow patterns that are conserved among most arthropod and some other metazoan mitochondrial genomes. The noncoding region of *Limulus* mtDNA contains a potential stem-loop structure, and we found a similar structure in the noncoding region of the published mtDNA of the prostriate tick *Ixodes hexagonus.* A simulation study was designed to evaluate the significance of these secondary structures; it revealed that they are statistically significant. No significant, comparable structure can be identified for the metastriate ticks *Rhipicephalus sanguineus* and *Boophilus microplus.* The latter two animals also share a mitochondrial gene rearrangement and an unusual structure of mt-tRNA(C) that is exactly the same association of changes as previously reported for a group of lizards. This suggests that the changes observed are not independent and that the stem-loop structure found in the noncoding regions of *Limulus* and *Ixodes* mtDNA may play the same role as that between *trnN* and *trnC* in vertebrates, i.e., the role of lagging strand origin of replication.

Introduction

Metazoan mitochondrial DNA (mtDNA) is typically a circular molecule between 14 and 18 kb in size that encodes 37 genes: 13 protein genes (subunits 6 and 8 of the F_0 ATPase [*atp6* and *atp8*], cytochrome *c* oxidase subunits 1–3 [*cox1–cox3*], cytochrome *b* [*cob*], and NADH dehydrogenase subunits 1–6 and 4L [*nad1–nad6* and *nad4L*]), 2 ribosomal RNA genes (small- and large-subunit rRNAs [*rrnS* and *rrnL*]), and 22 tRNA genes (designated by the one-letter code, with the two *L* and two *S* tRNAs differentiated by anticodon sequences[tag/taa and gct/tga, respectively]) (Wolstenholme 1992). Of the ~100 complete metazoan mtDNA sequences that have been published, only about one third are from taxa other than Vertebrata (Boore 1999). Among these, the phylum Arthropoda is best represented if judged by the number of sequences (12). However, the taxonomic sampling within Arthropoda is extremely biased: 7 of 12 sequenced mtDNAs are from the class Insecta, and 5 of those are from a single order (Diptera). The class Cheliceriformes is currently represented by two complete mtDNA sequences, those of the ticks *Ixodes hexagonus* and *Rhipicephalus sanguineus* (Black and Roehrdanz 1998). In addition, the complete gene arrangement, but only about half of the sequence, was determined for the horseshoe crab *Limulus polyphemus* (Staton, Daehler, and Brown 1997) and the cattle tick *Boophilus microplus* (Campbell and Barker 1999). Our original intention was to complete the *Limulus* mtDNA sequence. However, in that process we found a large number of errors in the published sequence, prompting us to resequence the entire mtDNA (GenBank accession number AF216203). In all, we identified 155 errors, all corrected here, and we also resolved the identities of nucleotides at 54 positions that were undetermined in the previous study. The complete *Limulus* mtDNA sequence allows us to analyze nucleotide composition and codon usage patterns and to identify structural features that may be involved in regulating mtDNA replication and/or gene expression.

Limulus polyphemus is one of the five extant species of Xiphosura (horseshoe crabs), one of the two extant major lineages of chelicerates (the other lineage, Arachnida, includes spiders, scorpions, ticks, and mites, among others). Originally thought to be crustaceans (hence the common name), xiphosurans were recognized as aquatic chelicerates late in the 19th century (Lankester 1881). The fossil record of horseshoe crabs goes back to the Devonian, and modern-looking horseshoe crabs first appear in the mid-Mesozoic (Størmer 1952). Their apparently slow rate of morphological change since has led to their being dubbed ''living fossils'' (Fisher 1984) and regarded as a keystone group for studies of evolution and of arthropod phylogeny.

Materials and Methods

An mtDNA preparation from the horseshoe crab *L. polyphemus* was a gift from John Avise. The same DNA preparation was used by Staton, Daehler, and Brown (1997), whose results we used to design oligonucleotide primers matching the sequences within *cox1*, *nad5*, *cob*, and *rrnS*: COX1-F1 (5′-GTATAGCTCACGCAG-

Key words: *Limulus polyphemus*, mitochondrial genome, evolution, codon bias, secondary structure.

Address for correspondence and reprints: D. V. Lavrov, Department of Biology, University of Michigan, 830 North University Avenue, Ann Arbor, Michigan 48109-1048. E-mail: dlavrov@umich.edu.

Mol. Biol. Evol. 17(5):813–824. 2000
© 2000 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038

Downloaded from https://academic.oup.com/mbe/article/17/5/813/981126 by guest on 23 April 2024

GAGCCTCA-3′), COX1-R1 (5′-GTCAAGTCTACT-GAGGCTCCTGC-3′), NAD5-F1 (5′-GAGGAGAGT-GAATAGGACCCCAA-3′), NAD5-R1 (5′-ACACCT TGGGGTCCTATTCACTC-3′), COB-F1 (5′-CGAG-TAATTCATGCAAACGGAGC-3′), COB-R1 (5′-TCGTCCTACGTGAAGATAAAGGC-3′), srRNA-F1 (5′-ATCTGCTCTGTAATCGATGGTCC-3′), and srRNA-R1 (5′-ACGAGGACCATCGATTACAGAGC-3′). We then amplified the whole *Limulus* mitochondrial genome in four overlapping fragments ranging in size from 3 to 5 kb using Perkin Elmer's XL PCR kit and COX1-F1-NAD5-F1, NAD5-R1-COB-R1, COB-F1-srRNA-F1, and srRNA-R1-COX1-R1 primer pairs. The PCR cycling parameters were according to the XL PCR kit manual with the exception that an annealing step was added to each cycle with the temperature decreasing from 68°C to 60°C (−0.5°C per cycle) during the first 16 cycles and remaining at 60°C during the subsequent 21 cycles. Each PCR reaction yielded a single band when visualized with ethidium bromide staining after electrophoresis in a 1% agarose gel. Reaction products were purified by three serial passages through Ultrafree (30,000 nominal molecular weight limit) columns (Millipore) and used as templates in dye-terminator cycle-sequencing reactions according to the supplier's (Perkin Elmer) instructions. Both strands of each amplification product were sequenced by primer walking using an ABI 377 automated DNA sequencer.

Sequences were produced and assembled using Sequencing Analysis and Sequence Navigator software (ABI) and analyzed with MacVector, version 6.5 (Oxford Molecular Group), and Wisconsin Package, version 10.0 (Genetics Computer Group [GCG], Madison, Wis.), programs. Protein and ribosomal genes were identified using the MacVector Internet Blast Search function with default parameters. Transfer RNA genes were recognized by eye as sequences with potential tRNA secondary structure and specifically identified by their anticodon sequence. The 5′ ends of protein genes were inferred to be at the first legitimate in-frame start codon (ATN, GTG, TTG, GTT; Wolstenholme 1992) in the open reading frame (ORF) that was not located within the upstream gene encoded on the same strand. The two exceptions were *nad4* and *atp6*, each of which has been previously demonstrated to overlap with its upstream gene (*nad4L* and *atp8*, respectively) in many mtDNAs (Wolstenholme 1992). An unusual start codon (TTA) was inferred for *cox1* based on the sequence similarity between nucleotides present upstream from the first legitimate initiation codon in the ORF and those at the 5′ end of the *Drosophila cox1*. Regardless of the actual initiation codon, all proteins were assumed to start with formyl-Met, as has been demonstrated for other mitochondrial systems (Smith and Marcker 1968; Fearnley and Walker 1987).

With the exceptions of *nad4L* and *atp8*, just noted, the protein gene terminus was inferred to be at the first in-frame stop codon encountered, unless that codon was located within the sequence of a downstream gene encoded on the same strand. Otherwise, a truncated stop codon (T or TA) adjacent to the beginning of the downstream gene was designated as the termination codon and assumed to be completed by polyadenylation to a complete TAA stop codon after transcript processing. The 5′ end of *rrnS* was inferred from sequence similarity to the 5′ end of *Drosophila yakuba rrnS* and from its potential to form a characteristic secondary structure. The 3′ end of *rrnS* and the 5′ and 3′ ends of *rrnL* were assumed to be adjacent to the 5′ end of *trnV*, the 3′ end of *trnV*, and the 5′ end of *trnL(tag)*, respectively. By applying the same rules to the published sequences of the two tick mtDNAs, we found that the reassignment of several initiation and termination codons was warranted (see *Results*).

The inferred amino acid sequences for each gene from *L. polyphemus*, *I. hexagonus*, *R. sanguineus*, and *D. yakuba* were aligned using the CLUSTAL W program (Thompson, Higgins, and Gibson 1994) within MacVector, version 6.5 (gap penalty = 5; extension penalty = 1; no gap separation distance; all other options at default settings). These alignments were used to calculate percentage amino acid identity for the homologous genes. Amino acid and codon usage on the different *Limulus* mtDNA strands were compared using χ^2 analyses of contingency tables; when a 2×2 contingency table was used, the Yates correction for continuity was applied (Yates 1934). To illustrate the quantitative difference, the odds ratio (OR) was calculated as the ratio of a particular amino acid (group of amino acids, codon, group of codons) to all other amino acids (codons) for one strand divided by the same ratio for the second strand.

The stem-loop structures in noncoding regions were found using the Stemloop program in the Wisconsin Package, version 10.0 (GCG), with default settings. To evaluate their significance, we devised a method to estimate the probability of observing a secondary structure of an equal or greater length or with an equal or greater number of hydrogen bonds. The Shuffle and Stemloop programs from the Wisconsin Package were combined in a short script to randomly reorder the nucleotides in the noncoding regions and then to identify potential secondary structures in the shuffled sequence. This simulation and analysis was repeated 1,000 times. The probability for each proposed secondary structure in the actual sequence to be observed by chance alone was calculated as the frequency of simulations that produced a secondary structure of an equal or greater length or with an equal or greater number of hydrogen bonds.

Results and Discussion

Genome Size and Structure

The size of *Limulus* mtDNA is 14,985 bp, in agreement with the estimate of Saunders, Kessler, and Avise (1986), but about 1 kb smaller than that of Staton, Daehler, and Brown (1997), due to their 1-kb overestimate of the size of the large noncoding region. Most genes are either immediately contiguous or overlapping (fig. 1). Aside from the large noncoding region, only 23 noncoding nucleotides are present. Protein genes account for 73.8% of the genome (11,058 bp), rRNA genes for

Downloaded from https://academic.oup.com/mbe/article/17/5/813/981126 by guest on 23 April 2024

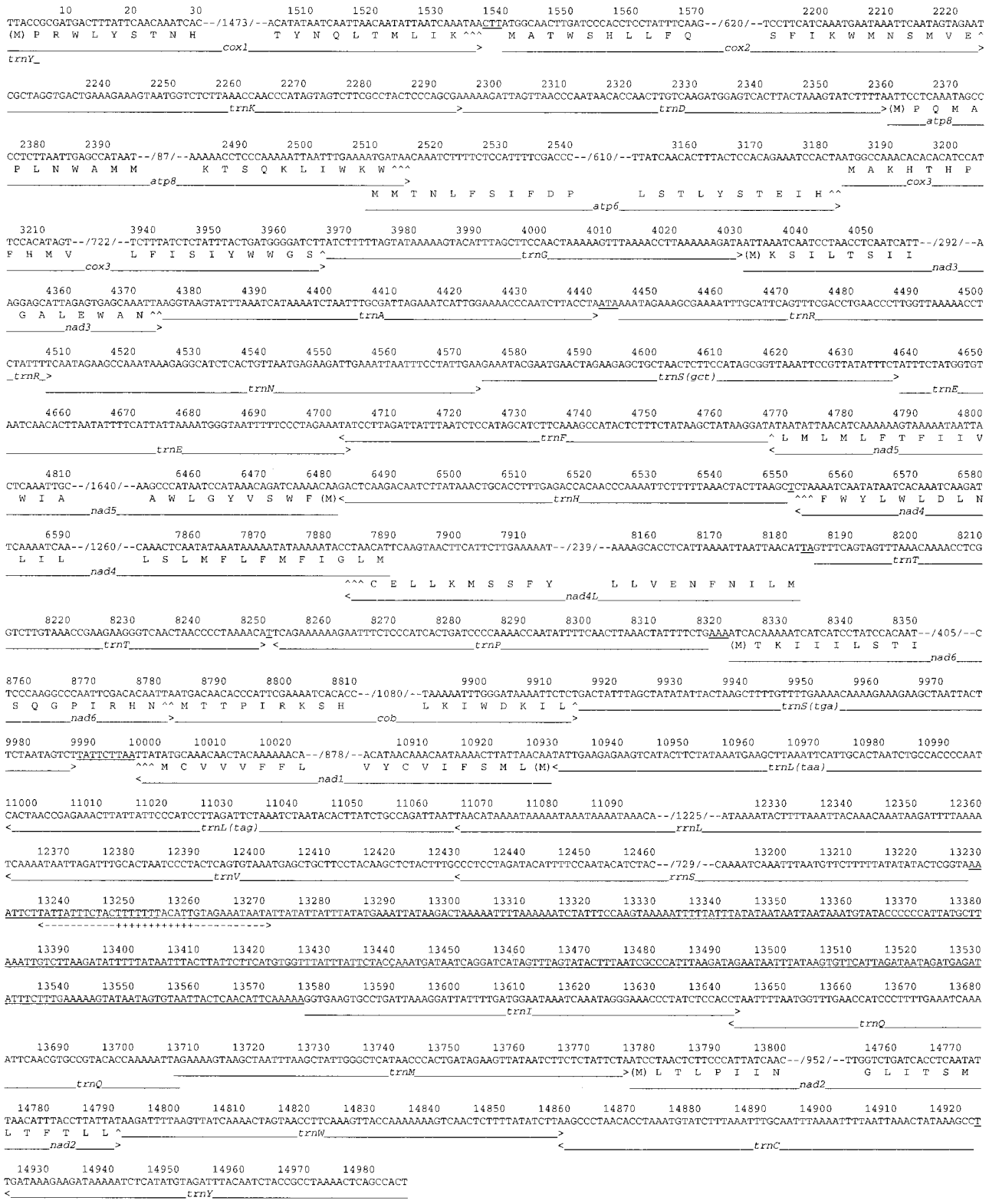

FIG. 1.—A partly schematic representation of the complete mtDNA sequence of *Limulus polyphemus.* Numbers within the slash marks indicate omitted nucleotides. Noncoding nucleotides, including the large noncoding region, are underlined. Minus signs (for stem nucleotides) and plus signs (for loop nucleotides) mark the segment of the large noncoding region with a potential to form a stem-loop structure. A dart (>) marks the last nucleotide for each gene and indicates the direction of transcription; nucleotides participating in termination codons of protein-coding genes are underscored with carets (^). All genes are inferred to initiate with formyl-methionine regardless of the actual start codon; M's are placed in parentheses to indicate presumed nonconformity to the genetic code. Both ends of *rrnL* and the 3′ end of *rrnS* are inferred to be immediately adjacent to their neighboring tRNA genes. The 5′ end of *rrnS* was inferred by sequence comparison with *Drosophila rrnS.*

Downloaded from https://academic.oup.com/mbe/article/17/5/813/981126 by guest on 23 April 2024

Table 1
Comparison of Mitochondrial Protein Genes of the Horseshoe Crab (*Limulus polyphemus*) with Two Ticks (*Ixodes hexagonus* and *Rhipicephalus sanguineus*) and the Fruit Fly (*Drosophila yakuba*)

Protein	No. of Amino Acids[a]				% Amino Acid Identity				Predicted Initiation and Termination Codons in *L. polyphemus*	
	Limulus	*Ixodes*	*Rhipicephalus*	*Drosophila*	*Limulus*/*Ixodes*	*Limulus*/*Rhipicephalus*	*Limulus*/*Drosophila*	*Ixodes*/*Rhipicephalus*		
ATP6	224	221*	221	224	50.0	45.5	60.9	54.8	ATG (−7)[b]	TA(A)[c] (0)
ATP8	51	51	52	53	22.6	30.2	34.0	34.0	ATT (0)[d]	TAA (−7)
COX1	511	512	512	512	80.5	76.8	79.3	83.0	TTA (−5)	TAA (3)
COX2	228	225	224*	228	64.9	58.3	68.0	73.3	ATG (3)	T (0)[d]
COX3	261	259*	259*	262	54.4	56.7	68.7	63.7	ATG (0)	T (0)
COB	377	360*	358	378	60.2	57.0	70.6	71.1	ATG (0)	T (0)
NAD1	310	314*	312*	324	48.9	42.2	56.8	51.3	TTG (0)[d]	TAA (8)
NAD2	338	323	318*	341	35.7	27.9	43.4	41.2	ATC (0)	T(AA) (0)
NAD3	114	111	113*	117	54.3	38.8	55.6	50.4	ATT (0)	TA(A) (0)
NAD4	445	440*	437*	446	45.7	42.4	51.0	53.2	ATG (−7)	TAG (1)
NAD4L ...	99	91	91	96	39.4	31.3	49.5	45.1	ATG (2)	TAG (−7)
NAD5	571	556*	552	573	43.6	38.4	48.9	46.8	TTG (0)[d]	T (0)
NAD6	153	141	149	174	31.0	26.6	37.9	32.5	ATC (3)	TA(A) (0)

[a] The numbers of amino acids in *Drosophila* proteins were taken from Clary and Wolstenholme (1985); those in *Ixodes* and *Rhipicephalus* proteins were inferred from the reported sequences (Black and Roehrdanz 1998). The initiation and termination codons for several genes in the latter sequences were reassigned using the criteria described in *Materials and Methods.* The numbers of amino acids in these genes are marked with asterisks.

[b] The numbers in parentheses after initiation and termination codons are the numbers of noncoding nucleotides upstream and downstream of a gene. Negative numbers indicate that the genes are overlapping.

[c] Nucleotides in parentheses indicate a potential for a complete termination codon overlapping the downstream gene.

[d] These initiation/termination codons differ from ones inferred by Staton, Daehler, and Brown (1997).

14.0% (2,095 bp), tRNA genes for 9.8% (1,466 bp), and noncoding DNA for 2.5% (371 bp). Since *cox1* and *trnY* overlap by five nucleotides (fig. 1), those 5 bp were counted twice in the above calculation. The mitochondrial gene arrangement of *Limulus* and its phylogenetic implications were reported previously (Boore et al. 1995; Staton, Daehler, and Brown 1997). The *Limulus* mitochondrial gene arrangement appears to be primitive for arthropods and differs only by the position of *trnL(taa)* from the derived arrangement shared between insects and crustaceans (Boore et al. 1995; Boore, Lavrov, and Brown 1998). The arrangement found in the prostriate tick *Ixodes* is identical to that of *Limulus*; however, metastriate ticks *Rhipicephalus* and *Boophilus* share a major gene rearrangement that clearly represents a derived state (Black and Roehrdanz 1998; Campbell and Barker 1998, 1999).

Base Composition

The A+T content of *L. polyphemus* mtDNA is 67.6%, lower than that of two other chelicerates, the ticks *Ixodes* (72.7%) and *Rhipicephalus* (78.0%) (Black and Roehrdanz 1998), and of all other reported arthropod mtDNAs except *Artemia* (64.4%; Valverde et al. 1994) and *Daphnia* (62.2%; Crease 1999). The strand encoding most of the proteins has the following nucleotide composition: A = 5,622 (37.5%), T = 4,503 (30.1%), C = 3,400 (22.7%), and G = 1,460 (9.7%); hereafter, we will refer to this strand as the α strand, and to the other strand as the β strand. One interesting feature of *Limulus* mtDNA is the significant base-compositional difference between the two strands. This difference can be measured as GC- and AT-skews, where GC-skew = (G − C)/(G + C) and AT-skew = (A − T)/(A + T) (Perna and Kocher 1995). For the *Limulus* α strand, GC-skew = −0.40 and AT-skew = 0.11. An absolute GC-skew this high has not previously been reported for an arthropod mtDNA but is a prominent feature of mammalian mtDNAs (Reyes et al. 1998). The AT-skew is also more extreme for *Limulus* mtDNA than for all other published arthropod mtDNAs except that of *Locusta* (Flook, Rowell, and Gellissen 1995), for which it is 0.18. Reflecting their base-compositional differences, there is a significant bias between the two strands in both the amino acid composition and the synonymous codon usage pattern of the encoded proteins, and also in the number of non–Watson-Crick base pairs in the stem regions of encoded tRNAs (discussed below). Although the exact mechanism responsible for creating strand asymmetry in mtDNA is unknown, one possibility is that it is created by the differential accumulation of mutations on the strand that is displaced and in single-stranded form during much of a replication cycle (reviewed in Reyes et al. 1998).

Protein Genes

Size and Sequence Similarity

The sizes of most protein genes in *Limulus* mtDNA are either identical to or slightly smaller than their *Drosophila* counterparts, the only exception being *nad4L*, which overlaps with the downstream *nad4* in *Limulus* and in many other animals, but not in *Drosophila* (table 1). The sizes of all protein genes except *nad1*, *cox1*, and *atp8* are larger in *Limulus* than in the ticks *Ixodes* and *Rhipicephalus.* Sequence comparisons among *Limulus*, these two tick species, and *Drosophila* reveal *cox1* as the most conserved and *atp8* and *nad6* as the least conserved genes (table 1), an order commonly observed among arthropods (e.g., Crease 1999). The previously reported, unusually small size of *Limulus atp8* and its

Downloaded from https://academic.oup.com/mbe/article/17/5/813/981126 by guest on 23 April 2024

very low amino acid identity to the homologous gene in *Drosophila* (Staton, Daehler, and Brown 1997) was an artifact of sequencing errors; use of the corrected sequence yields values close to those typical for other metazoans (table 1).

Translation Initiation and Termination Signals

An ATG, ATT, ATC, or TTG initiation codon was inferred for all *Limulus* protein genes except *cox1* (table 1), which appears to use the nonstandard start codon TTA (Staton, Daehler, and Brown 1997). The fifth codon upstream from this TTA is an in-frame stop codon, and the first conventional start codon (ATT) is 36 nt downstream from it. Unusual start codons for *cox1* have also been inferred in other species (e.g., ATAA in *Drosophila* [Clary and Wolstenholme 1983], TCG in two species of mosquito [Beard, Hamm, and Collins 1993], and CTG in *Balanoglossus* [Castresana et al. 1998]). It is unclear why this is the case, especially since the sequence of *cox1* is usually the most conserved of metazoan mitochondrial genes.

The use of ATG as a start codon is limited to the protein genes encoded immediately downstream of either another protein gene (*cox1-cox2, atp8-atp6, atp6-cox3, nad6-cob, nad4L-nad4*)—and in all such cases, the downstream protein gene has an ATG start codon—or a tRNA gene encoded on the opposite strand (*trnT-nad4L*). This is especially striking in the cases of four genes encoded on the α strand (*cox2, atp6, cox3,* and *cob*) in which, out of a total of 137 Met codons, only 8 are ATGs. When a protein gene is encoded immediately downstream from a tRNA gene on the same strand, (*atp8, nad1, nad2, nad3,* and *nad5*), it always starts with a start codon other than ATG, and there are no intervening nucleotides between the two genes (table 1).

Two different patterns are observed when two protein genes are adjacent and on the same strand. The first is exemplified by the overlapping *atp8-atp6* and *nad4L-nad4* gene pairs. The ATG start codon for the downstream gene in each pair is within the coding sequence of the upstream gene, and in both cases the genes overlap by 7 nt. For *nad4L* and *nad4*, 7-nt overlaps have also been reported for *Locusta* (Flook, Rowell, and Gellissen 1995), two species of *Anopheles* (Mitchell, Cockburn, and Seawright 1993; Beard, Hamm, and Collins 1993), *Lumbricus* (Boore and Brown 1995), *Balanoglossus* (Castresana et al. 1998), and numerous vertebrate mtDNAs (Wolstenholme 1992). A 7-nt overlap between *nad4L* and *nad4* can also be inferred for each of the tick sequences reported (Black and Roehrdanz 1998; Campbell and Barker 1999), although the overlaps were not recognized by those authors. In most of the cases listed above, as well as in some invertebrates in which these two genes do not overlap (e.g., *Drosophila*; Clary and Wolstenholme 1985; Garesse 1988), the amino acid sequence at the inferred NH_2 terminus of NAD4 is well conserved. A 7-nt overlap is also present between *atp8* and *atp6* in all arthropod mtDNA sequences published except that of *Apis* (Crozier and Crozier 1993), where the overlap is 19 nt, but in vertebrate mtDNAs, the overlap between these genes ranges from 2 to 46 nt (reviewed in Wolstenholme 1992).

Transcriptional mapping analysis in several species of animals has demonstrated the presence of bicistronic transcripts for both the *atp8-atp6* and the *nad4L-nad4* gene pairs (e.g., Ojala et al. 1980; Berthier et al. 1986). In the case of *atp8-atp6*, it was also demonstrated that both of these genes are fully translated into proteins (Fearnley and Walker 1986). Recently, Taanman (1999) has suggested that the *nad4L* and *atp8* mRNAs, if single, may be too short to be translated efficiently. However, in the annelid *Lumbricus* and in several mollusk mtDNAs, *atp8* not only is separated from *atp6*, but is also flanked by two tRNA genes (Boore 1999) and therefore is likely to be translated from a single mRNA. Therefore, while there may be some selective advantage in having *atp8-atp6* and *nad4L-nad4* adjacent, this does not seem to be imperative for all animals.

The second pattern is exemplified by the *atp6-cox3* and *nad6-cob* gene pairs. In these pairs, the incomplete stop codon (TA) of the upstream gene directly abuts the ATG start codon of the downstream gene, creating the appearance of a complete TAA stop codon. The same pattern was observed for the *atp6-cox3* and, interestingly, the *nad4L-nad4* gene pairs in *D. yakuba* and *D. melanogaster*, for the *atp6-cox3* and *nad6-cob* gene pairs in *Anopheles gambiae* and *Anopheles quadrimaculatus*, and for the *atp6-cox3* genes in *Artemia franciscana*, *Balanoglossus carnosus*, *Homo sapiens*, and many other metazoans.

The only pair of adjacent protein genes in *Limulus* that does not follow either of these patterns is *cox1-cox2*. There are three intervening nucleotides between the TAA stop codon of *cox1* and the ATG start codon of *cox2*. In the cases of several other arthropods, intervening nucleotides were also reported between the *atp6-cox3* pair and/or the *nad6-cob* pair (most notably in *Ixodes* and *Rhipicephalus*). While this demonstrates that the two patterns described above are not universal, their frequent presence among this phylogenetically diverse group of metazoans suggests their involvement in some underlying mechanisms of gene expression in animal mtDNA, such as mRNA translation and/or processing.

Only five protein genes in *Limulus* are inferred to have complete termination codons (table 1). In two (*atp8* and *nad4L*), the termination codons are entirely within the coding sequence of a downstream protein gene (see above). In four of eight genes inferred to have truncated stop codons (*atp6, nad3, nad6,* and *nad2*), the next one or two nucleotides of the downstream gene completes a stop codon. In both *atp6* and *nad6*, these codons could be functional if the *atp6-cox3* and *nad6-cob* pairs were translated from bicistronic transcripts. In *nad3* and *nad2*, the complete stop codons overlapping the downstream genes could be used as such if translation occurred before the tRNAs adjacent to the 3′ ends of these genes were cleaved from the common transcript (Ojala, Montoya, and Attardi 1981). The presence of several different RNAs that can map exactly with two or more adjacent gene sequences has been demonstrated in *Drosophila* (Berthier et al. 1986). In the four other

Downloaded from https://academic.oup.com/mbe/article/17/5/813/981126 by guest on 23 April 2024

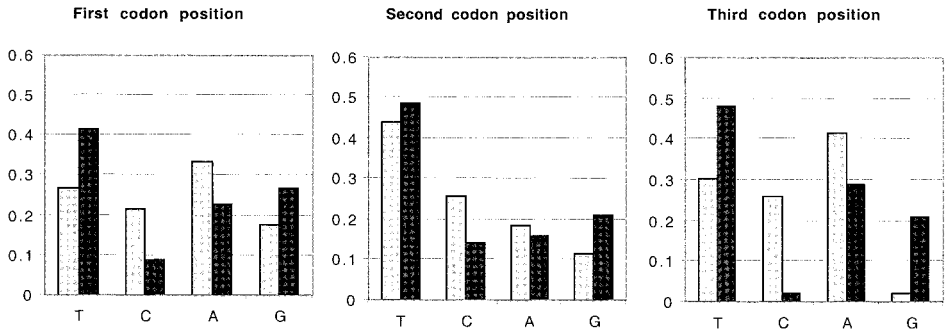

FIG. 2.—Comparison of the nucleotide composition of the first, second, and third codon positions of protein genes encoded by each strand of mtDNA. α strand nucleotide composition is shown by lighter shaded bars and β strand nucleotide composition is shown by darker shaded bars.

genes inferred to have truncated stop codons, the closest in-frame complete stop codon is from 17 to 136 nt downstream of the truncated stop codon. In all of these cases, a solitary T at the inferred 3′ end of a gene directly abuts the 5′ end of a downstream tRNA gene, and the mRNA is probably polyadenylated to form a UAA stop codon after the downstream tRNA is cleaved from the polycistronic transcript.

One start and two stop codons reported here differ from those inferred by Staton, Daehler, and Brown (1997) due to several frameshift sequencing errors made in their study; those errors are corrected in table 1. Based on the rules described in *Materials and Methods,* we also inferred that *nad1* starts at the TTG codon adjacent to *trnL(taa)* and is therefore extended by 6 nt from that reported by Staton, Daehler, and Brown (1997).

Codon Usage

Since the two strands of *Limulus* mtDNA have very different nucleotide compositions, the patterns of codon usage in proteins encoded on the two strands were analyzed separately. The frequencies of nucleotides in all three codon positions for each of the strands are presented in figure 2, from which two patterns are evident. First, when the two strands are compared, the number of T's and G's is smaller and, consequently, the number of A's and C's is greater in the α strand than in the β strand for all codon positions. Second, the relative frequencies of nucleotides for all codon positions are different between the two strands. In the α strand, the frequencies of nucleotides are A > T > C > G at first and third codon positions and T > C > A > G at second codon positions. In the β strand, the order is T > G > A > C at first and second codon positions and T > A > G > C at third positions. Selection may explain the high frequency of T at second positions in both strands, since codons with T in that position specify hydrophobic (nonpolar) amino acids, which are essential for the membrane-associated proteins encoded by mtDNA. The greatly reduced frequency of G at third codon positions in the α strand and C at third codon positions in the β strand probably reflects the mutational pattern in the genome, since nucleotides at third codon positions are under the least selective pressure. At all three codon positions, the differences in nucleotide frequencies on the two strands are statistically significant (χ^2 = 212, 115, and 748 for first, second, and third codon positions, respectively; $P < 0.001$), suggesting that both amino acid composition and synonymous codon usage should differ between the two strands.

Indeed, we found that amino acid frequencies differ between the two strands of *Limulus* mtDNA, as revealed by χ^2 analysis of 20 × 2 contingency tables (χ^2 = 250, df = 19, $P < 0.001$). However, when we grouped amino acids according to the chemical nature of their side chains (nonpolar, polar, acidic, or basic) and compared the frequency of each group in each strand, we found no significant differences for three of the four groups (table 2). Only the group of basic amino acids was significantly different in frequency between the strands (χ^2 = 11.03, df = 1, $P < 0.001$). We also compared individual amino acid frequencies in each strand and found significant differences for 9 of the 20 amino acids (table 2). Serine and leucine are each encoded by two different codon families in metazoan mtDNA. In both cases, the different families occurred at significantly different frequencies on the two strands, despite the fact that the overall frequencies of each amino acid were not significantly different between them. To quantitatively illustrate these findings, we computed the ORs for each codon family, amino acid, or group of amino acids (table 2).

Consistent with the strand bias in nucleotide composition (fig. 2), there is a corresponding bias in amino acid usage. In most (but not all) cases, amino acids encoded by AC-rich codons are used preferentially on the α strand, while those encoded by GT-rich codons are used preferentially on the β strand. To analyze this further, we divided all amino acids (except leucine and serine, which are each encoded by two families of codons) into three groups based on the nucleotide compositions of their codons—"AC-rich," with A and/or C in both first and second codon positions (H, K, N, P, Q, T); "GT-rich," with G and/or T in both first and second codon positions (C, F, G, V, W); and "neutral" (A, D, E, I, M, R, Y)—and compared their frequencies on two strands using the χ^2 test of contingency tables. There were 1.22 times as many amino acids specified by AC-rich codons as by GT-rich codons on the α strand, but

Downloaded from https://academic.oup.com/mbe/article/17/5/813/981126 by guest on 23 April 2024

Table 2
Amino Acid Compositions of Protein Genes

		α Strand Proteins[a]		β Strand Proteins[b]		Both Strands			
		No.	%	No.	%	No.	%	OR[c]	χ^2 Test[d]
Nonpolar									
Alanine	GCN	113	5.01	63	4.43	176	4.78	1.14	0.52
Isoleucine	ATY	267	11.83	81	5.69	348	9.46	2.22	37.68***
Leucine	Total	327	14.49	240	16.87	567	15.41	0.84	3.60
	CTN	223	9.88	36	2.53	259	7.04	0.29	106.43***
	TTR	104	4.61	204	14.34	308	8.37	4.22	70.95***
Methionine	ATR	137	6.07	77	5.41	214	5.82	1.13	0.58
Phenylalanine	TTY	167	7.40	164	11.52	331	8.99	0.61	17.65***
Proline	CCN	119	5.27	33	2.32	152	4.13	2.34	18.49***
Tryptophan	TGR	68	3.01	43	3.02	111	3.02	1.00	0.01
Valine	GTN	91	4.03	131	9.21	222	6.03	0.41	40.31***
Total		1,289	57.11	832	58.47	2,121	57.64	0.95	0.60
Polar									
Asparagine	AAY	101	4.47	46	3.23	147	3.99	1.40	3.20
Cysteine	TGY	9	0.40	40	2.81	49	1.33	0.14	36.84***
Glutamine	CAR	47	2.08	20	1.41	67	1.82	1.49	1.87
Glycine	GGN	113	5.01	124	8.71	237	6.44	0.55	19.30***
Serine	Total	234	10.37	153	10.75	387	10.52	0.96	0.10
	AGN	37	1.64	71	4.99	108	2.93	1.56	10.54**
	TCN	197	8.73	82	5.76	279	7.58	0.32	33.22***
Threonine	ACN	155	6.87	24	1.69	179	4.86	4.30	49.51***
Tyrosine	TAY	64	2.84	57	4.01	121	3.29	0.70	3.40
Total		723	32.03	464	32.61	1,187	32.26	0.97	0.11
Acidic									
Aspartate	GAY	39	1.73	21	1.48	60	1.63	1.17	0.21
Glutamate	GAR	44	1.95	43	3.02	87	2.36	0.64	3.90*
Total		83	3.68	64	4.50	147	3.99	0.81	1.32
Basic									
Arginine	CGN	38	1.68	24	1.69	62	1.68	1.00	0.02
Histidine	CAY	65	2.88	13	0.91	78	2.12	3.22	15.33***
Lysine	AAR	59	2.61	26	1.83	85	2.31	1.44	2.06
Total		162	7.18	63	4.43	225	6.11	1.67	11.03***
Grand total		2,257		1,423		3,680			

[a] ATP6, ATP8, COX1, COX2, COX3, COB, NAD2, NAD3, NAD6.

[b] NAD1, NAD4, NAD4L, NAD5.

[c] OR = odds ratio, the proportion of an amino acid (or a group of amino acids) to all other amino acids on the α strand over the same proportion on the β strand.

[d] χ^2 test of the difference in the frequency of an amino acid or a group of amino acids on the two strands. One, two, and three asterisks indicate the probabilities, $P < 0.05$, $P < 0.01$, $P < 0.001$, respectively, that these differences would be observed by chance. No asterisk indicates $P > 0.05$.

3.1 times as many amino acids specified by GT-rich codons as by AC-rich codons on the β strand, while the proportion of neutral amino acids was similar on both strands (41.4% and 35.5% for α and β strands, respectively). The overall difference in the frequencies of these three groups of amino acids on two strands was highly significant ($\chi^2 = 164$, df = 2, $P < 0.001$). Moreover, any protein encoded by the α strand had a lower proportion of amino acids specified by GT-rich codons than by AC-rich codons than any protein on the β strand (the numbers range between 0.33 and 1.19 and between 2.76 and 3.61, respectively). Obviously, since the α and β strands encode different proteins, the observed codon usage differences between the strands may be due to the specific amino acid requirements of these proteins, and not to strand nucleotide composition. However, the observation that every protein on the β strand has a ratio of GT- to AC-rich amino acids at least 2.3 times as large as any protein on the α strand argues against this, as does the pattern of synonymous codon usage (described below), and the nucleotide composition of tRNA and rRNA genes described in a later section.

All amino acids in arthropod mtDNA are specified by either a two- or a four-codon family, or by a combination of two such families. In all cases, when an amino acid is specified by a two-codon family, both members of the family end with either a purine (A or G) or a pyrimidine (T or C). Since the α and β strands of *Limulus* mtDNA are AC- and GT-rich, respectively, we expected to see different frequencies of usage of the two classes on the two strands and, indeed, we found in all such cases that those frequencies were significantly different and in accordance with the base composition bias of the two strands (table 3). Likewise, the pattern of codon usage for four-codon families is also significantly different between the two strands and is consistent with the overall strand nucleotide composition. However, when the frequencies of individual codons in a family were compared between strands, we found several cases in which they were not significantly different. Those cases are underlined in table 3 and may be due to some other constraints on codon usage, such as dinucleotide bias (Karlin and Burge 1995) or selection. To quantitatively illustrate the differences in the number of codons

Downloaded from https://academic.oup.com/mbe/article/17/5/813/981126 by guest on 23 April 2024

Table 3
Percentage and Number of Codons in Codon Families on Each Strands of *Limulus* mtDNA

	α Strand Proteins[a]				β Strand Proteins[b]					
Amino Acid	NNT	NNC	NNA	NNG	NNT	NNC	NNA	NNG	OR[c]	χ^2 Test[d]
Nonpolar										
Ala (GCN) . . .	30.1 (34)	31.9 (36)	38.1 (43)	0.0 (0)	66.7 (42)	4.8 (3)	19.0 (12)	9.5 (6)	7.44	41.4***
Ile (ATY)	60.7 (162)	39.3 (105)	—	—	97.5 (79)	2.5 (2)	—	—	25.60	37.9***
Leu (CTN) . . .	36.3 (81)	24.2 (54)	38.1 (85)[e]	1.3 (3)	69.4 (25)	2.8 (1)	25.0 (9)	2.8 (1)	4.30	16.9***
Leu (TTR) . . .	—	—	98.1 (102)	1.9 (2)	—	—	60.2 (124)	39.8 (82)	33.73	48.3***
Met (ATR) . . .	—	—	94.2 (129)	5.8 (8)	—	—	54.5 (42)	45.5 (35)	13.44	45.7***
Phe (TTY) . . .	52.7 (88)	47.3 (79)	—	—	96.3 (158)	3.7 (6)	—	—	23.64	80.3***
Pro (CCN) . . .	37.0 (44)	21.0 (25)	41.2 (49)	0.8 (1)	75.8 (25)	6.1 (2)	9.1 (3)	9.1 (3)	9.21	26.3***
Trp (TGR) . . .	—	—	92.6 (63)	7.4 (5)	—	—	72.1 (31)	27.9 (12)	4.88	7.1**
Val (GTN) . . .	33.0 (30)	14.3 (13)	49.5 (45)	3.3 (3)	45.8 (60)	0.8 (1)	28.2 (37)	25.2 (33)	4.30	40.2***
Polar										
Asn (AAY) . .	57.4 (58)	42.6 (43)	—	—	97.8 (45)	2.2 (1)	—	—	33.36	22.7***
Cys (TGY) . . .	33.3 (3)	66.7 (6)	—	—	92.5 (37)	7.5 (3)	—	—	24.67	13.4***
Gln (CAR) . . .	—	—	97.9 (46)	2.1 (1)	—	—	30.0 (6)	70.0 (14)	107.33	33.4***
Gly (GGN) . . .	16.8 (19)	16.8 (19)	54.0 (61)	12.4 (14)	17.7 (22)	1.6 (2)	38.7 (48)	41.9 (52)	3.59	37.0***
Ser (AGN) . . .	10.8 (4)	18.9 (7)	70.3 (26)	0.0 (0)	25.4 (18)	1.4 (1)	56.3 (40)	16.9 (12)	6.04	19.6***
Ser (TCN) . . .	25.4 (50)	32.0 (63)	42.1 (83)	0.5 (1)	73.2 (60)	2.4 (2)	18.3 (15)	6.1 (5)	10.95	73.0***
Thr (ACN) . . .	24.5 (38)	24.5 (38)	49.7 (77)	1.3 (2)	62.5 (15)	8.3 (2)	29.2 (7)	0.0 (0)	4.79	14.7**
Tyr (TAY) . . .	46.9 (30)	53.1 (34)	—	—	100.0 (57)	0.0 (0)	—	—	∞	39.5***
Acidic										
Asp (GAY) . .	43.6 (17)	56.4 (22)	—	—	90.5 (19)	9.5 (2)	—	—	12.29	10.6**
Glu (GAR) . . .	—	—	97.7 (43)	2.3 (1)	—	—	48.8 (21)	51.2 (22)	45.05	24.3***
Basic										
Arg (CGN) . . .	7.9 (3)	15.8 (6)	73.7 (28)	2.6 (1)	45.8 (11)	4.2 (1)	25.0 (6)	25.0 (6)	20.64	24.0***
His (CAY) . . .	38.5 (25)	61.5 (40)	—	—	92.3 (12)	7.7 (1)	—	—	19.20	10.5**
Lys (AAR) . . .	—	—	94.9 (56)	5.1 (3)	—	—	38.5 (10)	61.5 (16)	29.87	30.0***

[a] ATP6, ATP8, COX1, COX2, COX3, COB, NAD2, NAD3, NAD6.
[b] NAD1, NAD4, NAD4L, NAD5.
[c] OR = odds ratio, the proportion of GT- to AC-rich codons on the β strand over the same proportion on the α strand.
[d] χ^2 test of the difference in the frequencies of codons in codon families on two strands. For two-codon families, df = 1; for four-codon families, df = 3. Two or three asterisks indicate the probabilities, $P < 0.01$, $P < 0.001$, respectively, that these differences would be observed by chance.
[e] The frequencies of all codons except those underlined differ significantly between the two strands.

ending with A/C and those ending with G/T, we computed the OR for each codon family (table 3).

A correlation between nucleotide bias and codon usage in mtDNA was also reported by Foster, Jermiin, and Hickey (1997), who compared the amino acid contents in proteins from a very AT-rich and a relatively AT-poor mtDNA. A later study by this group demonstrated that such a bias can affect both DNA-based and protein-based phylogenetic reconstructions (Foster and Hickey 1999). The present study adds an intriguing dimension to the problem, since genes encoded by different strands of *Limulus* mtDNA have different nucleotide compositions and therefore may have different biases in phylogenetic reconstructions. The different patterns of codon usage between the two strands of *Limulus* mtDNA also explain the observation made by Staton, Daehler, and Brown (1997) that the lysine codon AAA is used over five times as frequently as the (presumably) more stably pairing AAG codon in *Limulus* protein genes. When codon usage for lysine was compared separately for each strand, we found that AAA codons were over 15 times as frequent as AAG codons in the α strand but that AAG codons are 1.5 times as frequent as AAA codons in the β strand, even though the average percentage of G's is still smaller than that of A's in the latter strand. It appears that while there might be some selective advantage to using the AAG codon, it is greatly overpowered by the compositional bias between the two strands.

Transfer RNA and Ribosomal RNA Genes

There are 22 potential tRNA genes in *Limulus*, as there are in most other published metazoan mtDNAs. The sequences and structures of these genes were described by Staton, Daehler, and Brown (1997). We found sequencing errors in six of them: tRNA(D), tRNA(E), tRNA(R), tRNA(N), tRNA(I), and tRNA(V); the corrections are shown in bold in figure 3. The corrected sequences improve the potential secondary structures of these tRNA genes and also eliminate the ''difficult case'' of a 2-nt overlap between *trnR* and *trnN*, noted in the previous study, reducing it to a 1-nt overlap (fig. 1). The latter can be resolved by polyadenylation, as has been demonstrated for some mitochondrial tRNAs (Yokobori and Pääbo 1997).

The tRNA genes are 68.6% AT-rich, close to the overall A+T composition of the genome. Thirteen of the genes are encoded by the α strand and have an AT-skew of 0.10 and a GC-skew of −0.06. Nine tRNA genes are encoded by the β strand and have an AT-skew of −0.08 and a GC-skew of 0.42. The compositional difference between these two sets of tRNAs, most evident in the measurements of GC-skew, results in the

Downloaded from https://academic.oup.com/mbe/article/17/5/813/981126 by guest on 23 April 2024

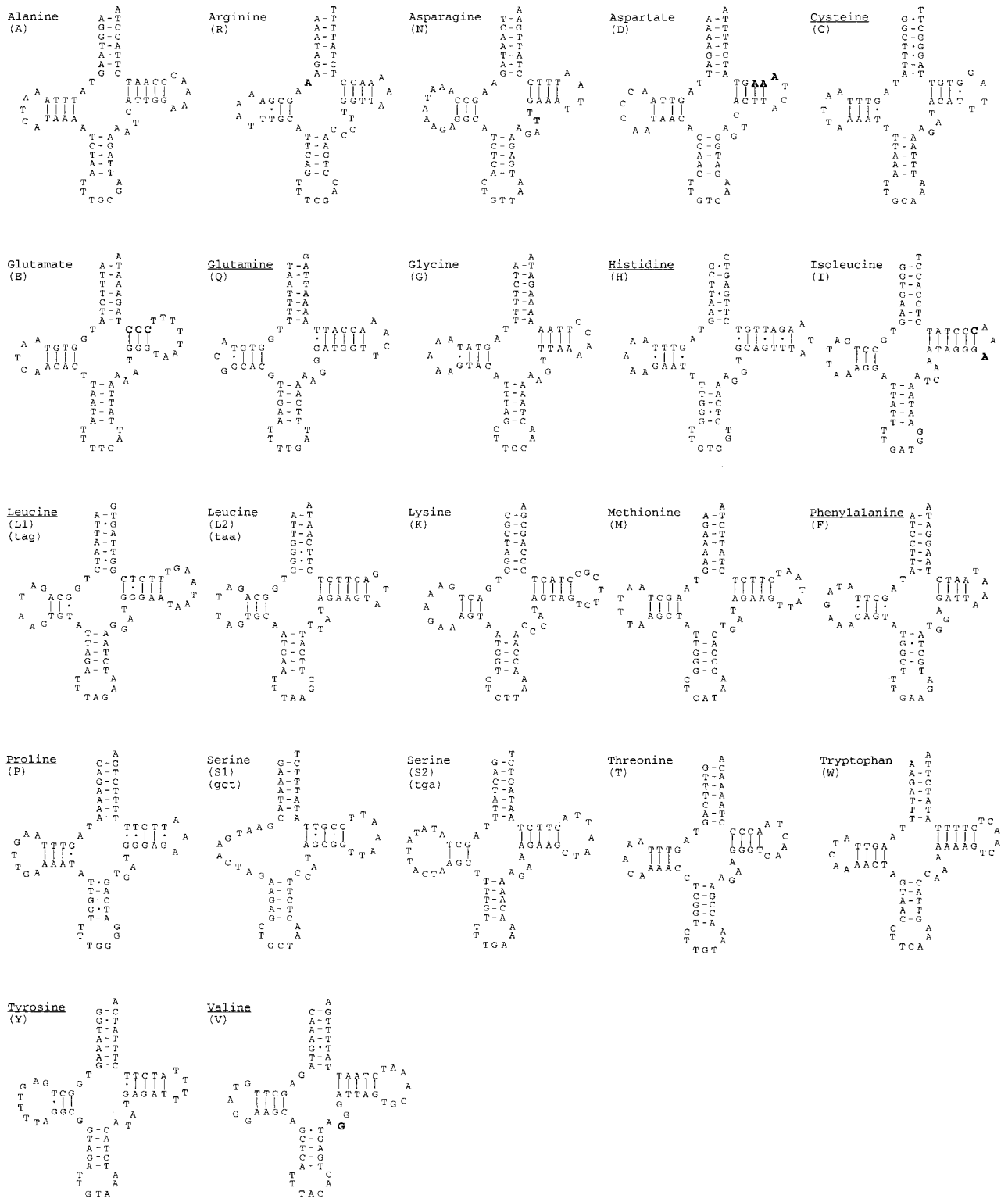

FIG. 3.—The potential secondary structures of the 22 inferred tRNAs. Those encoded by the β strand are underlined. AT/GC pairs are indicated by dashes, and GT pairs are indicated by dots. The differences from the previously reported sequences are in bold type.

drastically different frequencies of GT (=GU) pairs in the inferred tRNA structures (fig. 3). There are 35 GT pairs in the inferred stem regions of tRNAs encoded by the β strand (an average of 3.89 pairs per tRNA), but only three among the tRNAs encoded by the α strand (an average of 0.23 per tRNA). On the other hand, there are nine mismatches in stem regions of α strand–encoded tRNAs, but only one in β strand–encoded tRNAs (fig. 3).

As all other metazoan mtDNAs sequenced, *Limulus* mtDNA contains genes for both small and large ribo-

Downloaded from https://academic.oup.com/mbe/article/17/5/813/981126 by guest on 23 April 2024

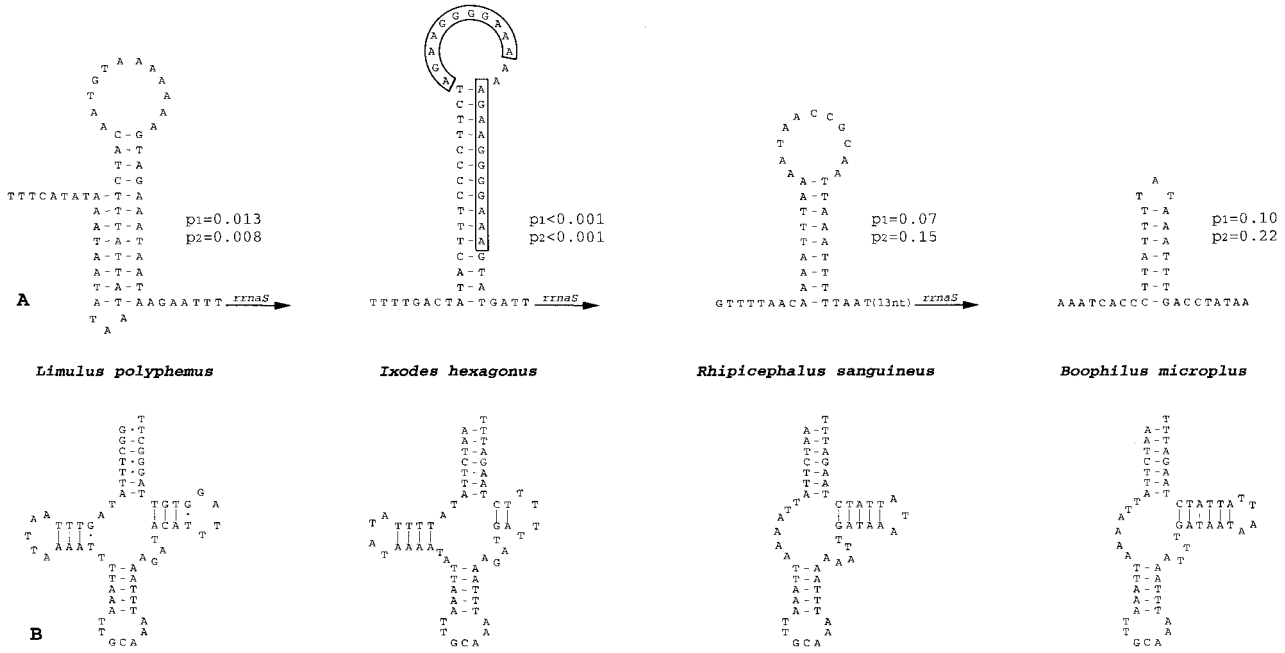

FIG. 4.—The best potential stem-loop structures present in the large noncoding regions of *Limulus polyphemus, Ixodes hexagonus, Rhipicephalus sanguineus,* and *Boophilus microplus* (*A*) and the change in secondary structure of tRNA(C) in *R. sanguineus* and *B. microplus* compared with those of *L. polyphemus* and *I. hexagonus* (*B*). The probability of finding, by chance, a secondary structure with a stem of equal or greater length is indicated by p_1, and that of finding a secondary structure with an equal or greater number of pair bonds is indicated by p_2. A potential alternative secondary structure is shown for *Limulus.* An alternative secondary structure can be also formed in *Ixodes*, where 11 nt in the loop region are identical to those forming the stem (the boxed sequences). The secondary structure of *Boophilus microplus* tRNA(C) differs from one proposed by Campbell and Barker (1999).

somal subunit RNAs (*rrnS* and *rrnL*). Both genes are encoded by the β strand and are separated by *trnV,* which is identical to their arrangement in many other metazoans. The size of the inferred *rrnS* is 799 nt, and its G+C content is 30.3%. The sequence is well conserved, especially at the 3′ end, where 82 out of the last 100 nt are identical to those of *Drosophila.* The size of the inferred *rrnL* is 1,296 nt, and its G+C content is 29.0%. The AT- and GC-skews for *rrnS* are −0.15 and 0.33, respectively, and those for *rrnL* are −0.15 and 0.45, respectively.

Noncoding Region

The largest noncoding region in *Limulus* mtDNA is 348 bp long. It is located between *rrnS* and *trnI* and is 81.3% A+T, significantly more AT-rich than the rest of the genome ($\chi^2 = 30.1$; $P < 0.001$). Within it, a 36-nt sequence close to *rrnS* has the potential to form a stem-loop structure with a 12-bp stem and a 12-nt loop (fig. 4*A*). A similar structure can be folded in the noncoding region of another chelicerate, the prostriate tick *Ixodes*; however, the best potential secondary structures found in the noncoding regions of the metastriate ticks *Rhipicephalus* and *Boophilus* (fig. 4*A*) differ from these in several respects. They have shorter stems (9 bp and 7 bp, respectively), with the one in *Rhipicephalus* composed entirely of AT pairs in short homopolymer runs. Simulation studies (see *Materials and Methods*) demonstrated that the probabilities of observing stems of this length by chance are 0.07 and 0.12, and the probabilities of observing this number of pair bonds are 0.15 and 0.22 for *Rhipicephalus* and *Boophilus,* respectively. The loops of these structures also lack the run of A's present in both *Limulus* and *Ixodes* (fig. 4*A*) and in several myriapod (unpublished data) stem-loop structures. In addition to the shared lack of significant secondary structures in the noncoding regions, both *Rhipicephalus* and *Boophilus* share a major mtDNA rearrangement and a drastically changed tRNA(C) structure that lacks the D-arm (fig. 4*B*).

Since both *Limulus* and *Ixodes* have the same gene arrangement (which is also primitive for Arthropoda), a stem-loop structure in the same portion of the noncoding region, and a canonical structure for tRNA(C), it is reasonable to assume that these features are primitive for ticks. Therefore, the shared changes observed in the metastriate ticks *Rhipicephalus* and *Boophilus* are derived and constitute synapomorphies for this group (Metastriata) (Black and Roehrdanz 1998; Campbell and Barker 1999). Each of these changes by itself is quite rare among arthropods: the major rearrangement was the first reported for this group, and the tRNA structure with the D-arm replacement loop was previously reported only for tRNA(S). Therefore, the probability of all three of them happening independently in the same lineage is quite low, and it is likely that the combination of these changes resulted from a single event, or from several related events in the mitochondrial genome. This conclusion is supported by the observation that exactly the same combination of changes is found in an entirely

Downloaded from https://academic.oup.com/mbe/article/17/5/813/981126 by guest on 23 April 2024

different group of animals, the iguanid lizards (Macey et al. 1997). In these animals, there is a unique rearrangement, with *trnI* and *trnQ* switched in order relative to the typical vertebrate arrangement. In addition, the stem-loop structure usually present in vertebrate mtDNA between *trnN* and *trnC* and shown to be the origin of light-strand replication (O_L) (e.g., Martens and Clayton 1979; Tapper and Clayton 1981) is missing, and tRNA(C) has a D-arm replacement loop in place of a D-arm. The fact that the combination of changes is exactly the same in two groups of animals may also suggest a similar function for the stem-loop in *Limulus* and *Ixodes* to that in vertebrates; i.e., this secondary structure may be the origin of lagging strand replication. Consistent with this idea, the origin of lagging strand replication in *D. melanogaster* mtDNA was mapped to the AT-rich noncoding region (Goddard and Wolstenholme 1978). The change in the structure of tRNA(C) is especially interesting in light of the observation that the origin of some lagging strands was mapped to this tRNA gene in human mtDNA (Tapper and Clayton 1981). It is possible, therefore, that the loss of the D-arm in tRNA(C), observed in both metastriate ticks and iguanid lizards, may allow this structure to function more efficiently as the origin of lagging strand replication.

An interesting feature of the sequence of stem-loop structures in both *Limulus* and *Ixodes* is its ability to form an alternative secondary structure that is less stable than the main structure (fig. 4*A*). In *Limulus*, the sequence adjacent to the 5′ end of the main secondary structure is complementary to 8 nt in the stem of the main stem-loop structure. The 11 nt in the loop of the *Ixodes* secondary structure are identical to those forming the stem and can form an alternative stem of 8 bp (fig. 4*A*).

Conclusions

An unusual feature of the *L. polyphemus* mitochondrial genome is a significant compositional bias between the two strands. The absolute value of GC-skew reported here is the highest among all published arthropod mitochondrial genomes and is similar to the values observed for mammalian mtDNAs. The AT-skew is also more extreme than those in most other arthropod mitochondrial sequences. The compositional bias between the two strands correlates with, and most likely causes, the biases in amino acid content and synonymous codon usage in proteins encoded by different strands. This may have an effect on phylogenetic reconstruction based on DNA and protein sequence comparisons (Foster and Hickey 1999), especially since genes encoded by the two strands have opposite biases. The nucleotide bias also correlates with the number of non–Watson-Crick base pairs in the stem regions of encoded tRNAs.

The sequence of the single large noncoding region of *Limulus* mtDNA has the potential to form a statistically significant secondary structure with a 7-nt poly-A run in the loop region. A comparable structure was found in the noncoding region of the prostriate tick *Ixodes*, but not in the metastriate ticks *Rhipicephalus* and *Boophilus*. In addition, the latter two animals share a gene rearrangement and an altered structure of tRNA(C), exactly the same association of changes as previously reported for a group of lizards. Based on these observations, we suggest that the stem-loop structure in the noncoding region of *Limulus* and *Ixodes* plays the same role as the stem-loop structure between *trnN* and *trnC* in vertebrates; i.e., it is the lagging strand origin of replication.

Acknowledgments

This work was supported by NSF grants DEB 9807100 (to W.M.B. and J.L.B.) and DEB 9972712 (to W.M.B. and D.V.L.). We thank John Avise for the *Limulus* mtDNA, and Kevin Helfenbein and two anonymous reviewers for helpful comments on the manuscript.

LITERATURE CITED

Beard, B. C., D. M. Hamm, and F. H. Collins. 1993. The mitochondrial genome of the mosquito *Anopheles gambiae*: DNA sequence, genome organization and comparisons with mitochondrial sequences of other insects. Insect Mol. Biol. **2**:103–124.

Berthier, F., M. Renaud, S. Alziari, and R. Durand. 1986. RNA mapping on *Drosophila* mitochondrial DNA: precursors and template strands. Nucleic Acids Res. **14**:4519–4533.

Black, W. C. IV, and R. L. Roehrdanz. 1998. Mitochondrial gene order is not conserved in arthropods: prostriate and metastriate tick mitochondrial genomes. Mol. Biol. Evol. **15**:1772–1785.

Boore, J. L. 1999. Animal mitochondrial genomes. Nucleic Acids Res. **27**:1767–1780.

Boore, J. L., and W. M. Brown. 1995. Complete sequence of the mitochondrial DNA of the annelid worm *Lumbricus terrestris*. Genetics **141**:305–319.

Boore, J. L., T. M. Collins, D. Stanton, L. L. Daehler, and W. M. Brown. 1995. Deducing the pattern of arthropod phylogeny from mitochondrial DNA rearrangements. Nature **376**:163–165.

Boore, J. L., D. V. Lavrov, and W. M. Brown. 1998. Gene translocation links insects and crustaceans. Nature **392**:667–668.

Campbell, N. J. H., and S. C. Barker. 1998. An unprecedented major rearrangement in an arthropod mitochondrial genome. Mol. Biol. Evol. **15**:1786–1787.

———. 1999. The novel mitochondrial gene arrangement of the cattle tick, *Boophilus microplus*: fivefold tandem repetition of a coding region. Mol. Biol. Evol. **16**:732–740.

Castresana, J., G. Feldmaier-Fuchs, S. Yokobori, N. Satoh, and S. Pääbo. 1998. The mitochondrial genome of the hemichordate *Balanoglossus carnosus* and the evolution of deuterostome mitochondria. Genetics **150**:1115–1123.

Clary, D. O., and D. R. Wolstenholme. 1983. Genes for cytochrome c oxidase subunit I, URF2, and three tRNAs in *Drosophila* mitochondrial DNA. Nucleic Acids Res. **11**: 6859–6872.

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

Crease, T. J. 1999. The complete sequence of the mitochondrial genome of *Daphnia pulex* (Cladocera: Crustacea). Gene **233**:89–99.

Downloaded from https://academic.oup.com/mbe/article/17/5/813/981126 by guest on 23 April 2024

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.

FEARNLEY, I. M., and J. E. WALKER. 1986. Two overlapping genes in bovine mitochondrial DNA encode membrane components of ATP synthase. EMBO J. **5**:2003–2008.

———. 1987. Initiation codons in mammalian mitochondria: differences in genetic code in the organelle. Biochemistry **26**:8247–8251.

FISHER, D. C. 1984. The Xiphosurida: archetypes of bradytely? Pp. 196–213 *in* N. ELDREDGE and S. M. STANLEY, eds. Living fossils. Springer, New York.

FLOOK, P. K., C. H. F. ROWELL, and G. GELLISSEN. 1995. The sequence, organization, and evolution of the *Locusta migratoria* mitochondrial genome. J. Mol. Evol. **41**:928–941.

FOSTER, P. G., and D. A. HICKEY. 1999. Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. J. Mol. Evol. **48**:284–290.

FOSTER, P. G., L. S. JERMIIN, and D. A. HICKEY. 1997. Nucleotide composition bias affects amino acid content in proteins coded by animal mitochondria. J. Mol. Evol. **44**:282–288.

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

GODDARD, J. M., and D. R. WOLSTENHOLME. 1978. Origin and direction of replication in mitochondrial DNA molecules from *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA **75**:3886–3890.

KARLIN, S., and C. BURGE. 1995. Dinucleotide relative abundance extremes: a genomic signature. Trends Genet. **11**: 283–290.

LANKESTER, E. R. 1881. *Limulus,* an arachnid. Q. J. Microsc. Sci. **23**:504–548, 609–649.

MACEY, J. R., A. LARSON, N. B. ANANJEVA, and T. J. PAPENFUSS. 1997. Evolutionary shifts in three major structural features of the mitochondrial genome among iguanian lizards. J. Mol. Evol. **44**:660–674.

MARTENS, P. A., and D. A. CLAYTON. 1979. Mechanism of mitochondrial DNA replication in mouse L-cells: localization and sequence of the light-strand origin of replication. J. Mol. Biol. **135**:327–351.

MITCHELL, S. E., A. F. COCKBURN, and J. A. SEAWRIGHT. 1993. The mitochondrial genome of *Anopheles quadrimaculatus* species A: complete nucleotide sequence and gene organization. Genome **36**:1058–1073.

OJALA, D., C. MERKEL, R. GELFAND, and G. ATTARDI. 1980. The tRNA genes punctuate the reading of genetic information in human mitochondrial DNA. Cell **22**:393–403.

OJALA, D., J. MONTOYA, and G. ATTARDI. 1981. tRNA punctuation model of RNA processing in human mitochondria. Nature **290**:470–474.

PERNA, N. T., and T. D. KOCHER. 1995. Patterns of nucleotide composition at fourfold degenerate sites of animal mitochondrial genomes. J. Mol. Evol. **41**:353–358.

REYES, A., C. GISSI, G. PESOLE, and C. SACCONE. 1998. Asymmetrical directional mutation pressure in the mitochondrial genome of mammals. Mol. Biol. Evol. **15**:957–966.

SAUNDERS, N. C., L. G. KESSLER, and J. C. AVISE. 1986. Genetic variation and geographic differentiation in mitochondrial DNA of the horseshoe crab, *Limulus polyphemus*. Genetics **112**:613–627.

SMITH, A. E., and K. A. MARCKER. 1968. N-formylmethionyl transfer RNA in mitochondria from yeast and rat liver. J. Mol. Biol. **38**:241–243.

STATON, J. L., L. L. DAEHLER, and W. M. BROWN. 1997. Mitochondrial gene arrangement of the horseshoe crab *Limulus polyphemus* L.: conservation of major features among arthropod classes. Mol. Biol. Evol. **14**:867–874.

STØRMER, L. 1952. Phylogeny and taxonomy of fossil horseshoe crabs. J. Paleontol. **26**:630–640.

TAANMAN, J. W. 1999. The mitochondrial genome: structure, transcription, translation and replication. Biochim. Biophys. Acta **1410**:103–123.

TAPPER, D. P., and D. A. CLAYTON. 1981. Mechanism of replication of human mitochondrial DNA. Localization of the 5′ ends of nascent daughter strands. J. Biol. Chem. **256**: 5109–5115.

THOMPSON, J. D., D. G. HIGGINS, and T. J. GIBSON. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**:4673–4680.

VALVERDE, J. R., B. BATUECAS, C. MORATILLA, R. MARCO, and R. GARESSE. 1994. The complete mitochondrial DNA sequence of the crustacean *Artemia franciscana*. J. Mol. Evol. **39**:400–408.

WOLSTENHOLME, D. R. 1992. Animal mitochondrial DNA: structure and evolution. Int. Rev. Cytol. **141**:173–216.

YATES, F. 1934. Contingency tables involving small numbers and the χ^2 test. J. R. Stat. Soc. **1**(Suppl.):217–235.

YOKOBORI, S., and S. PÄÄBO. 1997. Polyadenylation creates the discriminator nucleotide of chicken mitochondrial tRNA(Tyr). J. Mol. Biol. **265**:95–99.

STEPHEN PALUMBI, reviewing editor

Accepted January 28, 2000

Downloaded from https://academic.oup.com/mbe/article/17/5/813/981126 by guest on 23 April 2024