Genetics and Molecular Biology, 31, 1 (suppl), 293-302 (2008)

www.sbg.org.br

Research Article

The complete mitochondrial genome of the pirarucu (*Arapaima gigas*, Arapaimidae, Osteoglossiformes)

Tomas Hrbek[1,2] and Izeni Pires Farias[1]

[1]*Laboratório de Evolução e Genética Animal, Instituto de Ciências Biológicas, Manaus, Universidade Federal do Amazonas, AM, Brazil.*
[2]*Biology Department, University of Puerto Rico, Rio Piedras, San Juan, Puerto Rico.*

Abstract

We sequenced the complete mitochondrial genome of the pirarucu, *Arapaima gigas*, the largest fish of the Amazon basin, and economically one of the most important species of the region. The total length of the *Arapaima gigas* mitochondrial genome is 16,433 bp. The mitochondrial genome contains 13 protein-coding genes, two rRNA genes and 22 tRNA genes. Twelve of the thirteen protein-coding genes are coded on the heavy strand, while *nad6* is coded on the light strand. The *Arapaima* gene order and content is identical to the common vertebrate form, as is codon usage and base composition. Its control region is atypical in being short at 767 bp. The control region also contains a conserved ATGTA motif recently identified in the Asian arowana, three conserved sequence blocks (CSB-1, CBS-2 and CBS-3) and its 3' end contains long series of di- and mono-nucleotide microsatellite repeats. Other osteoglossiform species for which control region sequences have been published show similar control region characteristics.

Key words: pirarucu, mitogenomics, Amazon basin.

Received: August 22, 2006; Accepted: July 10, 2007.

Introduction

Comparing complete animal mitochondrial genome sequences is becoming an increasingly common method of phylogenetic reconstruction and of modeling genome evolution. Mitochondrial genomes from over 300 vertebrate species, with a large concentration on teleost fishes, have now been sequenced (Boore, 1999; Curole and Kocher, 1999; Inoue *et al.*, 2001; Miya *et al.*, 2001; Miya *et al.*, 2003). Mitochondrial sequences have proven to be of great utility in molecular phylogenetic studies, providing large number of phylogenetically informative characters, and complete genome sequences have provided valuable insights into a number of deep-level phylogenetic questions (*e.g.*, Boore and Brown, 1998; Mindell *et al.*, 1998; Naylor and Brown, 1998; Miya *et al.*, 2001; Inoue *et al.*, 2003a; Miya *et al.*, 2003; Brinkmann *et al.*, 2004).

Further information comes from the gene order of mitochondrial genomes. While the gene order among vertebrates is highly conserved (*e.g.*, Inoue *et al.*, 2001, 2003a; Miya *et al.*, 2003), and most animal mitochondrial genomes contain the same 37 intron-less genes (Brown, 1985; Boore, 1999), there are some well documented exceptions such as the gene order of birds (Mindell *et al.*, 1998; Haring *et al.*, 2001), squamate reptiles (Macey *et al.*, 1997a; Macey *et al.*, 1997b), teleost fishes (Miya and Nishida, 1999; Inoue *et al.*, 2003b), and mammals (Pääbo *et al.*, 1991). These rearrangement to date have not been shown to be homoplaseous, and thus provide high quality phylogenetic information (Boore *et al.*, 1997; Pereira, 2000; Morrison *et al.*, 2002) similar to SINE and LINE data.

The mitochondrial genome is typically compact at ~16 kb with few, if any, intergenic spacers. The two non-coding regions which usually represent less than 5% of the total genome size are the control region which contains the heavy-strand replication origin and is involved in regulating of transcription and replication (Clayton, 1982; Shadel and Clayton, 1997), and the light-strand replication origin (Wong and Clayton, 1985). While the structural properties of the control region are important in transcription and replication, actual sequence of nucleotides is relatively free to vary (Shadel and Clayton, 1997) making the control region a popular candidate for population-level and phylogeographic studies (Avise, 2004).

Complete mitochondrial genomes are available for a number of the species of the order Osteoglossiformes, but not for *Arapaima gigas*. A peculiarity of all of the osteoglossiform genomes deposited in GenBank (AB043025 and AB043068, Inoue *et al.*, 2001), with the exception of

Send correspondence to Tomas Hrbek. Biology Department, University of Puerto Rico, Rio Piedras, San Juan, Puerto Rico. E-mail: hrbek@uprrp.edu.

Scleropages formosus, is that they are all missing the control region. In their conservation genetic study, Hrbek *et al.* (2005) were also unable to amplify the control region of majority of the individuals used in the study in spite of designing specific, highly stringent primers. Those individuals that amplified often produced only a weak product; amplification of different individuals resulted in different sized products, and some individuals also showed multiple bands. These results pointed to the possible presence of repeats and secondary structures that would prevent efficient amplification and sequencing of this region. Multiple PCR bands suggested possible mtDNA heteroplasmy. Consultation with the authors of the *Osteoglossum* and *Pantodon* mitochondrial genomes revealed that they also were unable to efficiently amplify and sequence the control region; only the 5' portion of the control regions are deposited, and are characterized by a large series of tandem repeats. In a recent publication characterizing the complete mitochondrial genome of the Asian arowana *Scleropages formosus*, Yue *et al.* (2006) observed tandem repeats in the control region, and also mitochondrial heteroplasmy. Therefore, we sequenced the complete mitochondrial genome of *Arapaima gigas*, including the control region, in order to characterize the genome and assess its potential phylogenetic utility, and that of its genes and gene regions.

Material and Methods

Laboratory protocols

The tissue sample used in this analysis was obtained from a specimen captured in a participatively managed fishery area north of the city of Santarém. A white muscle tissue sample was collected and preserved in 95% ethanol and transported to laboratory. Total genomic DNA was extracted using Qiagen spin-column according to the manufacturer's protocol.

Polymerase Chain Reaction (PCR) amplification was performed on total genomic DNA. Negative controls were performed for all reactions. PCR was performed in 50 µL reaction volumes containing 23.6 µL of ddH_2O, 3.4 µL of 10 mM $MgCl_2$, 5.0 µL of 10x buffer (200 mM Tris-HCl [pH 8.8], 20 mM $MgSO_4$, 100 mM KCl, 100 mM $(NH_4)_2SO_4$, 1% Triton® X-100, 1 mg/mL nuclease-free BSA), 5.0 µL of each primer (2 µM), 4.0 µL dNTP mix (10 mM), 8 units of KlenTaqLA DNA Polymerase, and 2 µL of DNA template (approximately 50 ng/µL).

To assure fidelity of priming, we used a touch-down PCR method. The temperature profile consisted of 1) preheating at 68 °C for 60 s, 2) denaturation at 93 °C for 10 s, 3) annealing at 55-50 °C for 35 s, 4) extension at 68 °C for 7 min, and 5) a final extension at 68 °C for 10 min. Steps 2-4 were repeated 25 times; in the first 9 cycles, the annealing temperature was decreased by 0.5 °C until 50 °C annealing temperature was reached. Using this methodology we amplified three overlapping segments.

PCR products were evaluated on a 1% agarose gel, and then purified with Qiagen spin-columns. The sequencing strategy employed the 'primer walking' methodology. We sequenced each amplified fragment with the two amplification primers, and upon obtaining the sequence information we performed additional sequencing reactions with internal primers available in the laboratory or specifically designed primers until sequence data were obtained for the complete fragment. Many of the additional primers used in this study were derived from primers published in Miya and Nishida (1999); see Table 1 for primers.

Cycle sequencing PCR followed manufacturer's recommended protocol for DYEnamic ET Dye Terminator mix (GE Healthcare); primer annealing temperature was at 50 °C and we used ~ 30 ng of purified PCR product. Cycle sequencing PCR products were precipitated using a mixture of 70% ethanol and 175 mM ammonium acetate. Precipitated DNA product was resuspended in Hi-Di Formamide, and resolved on a MegaBACE 1000 automatic DNA analysis system (GE Healthcare) using the manufacturer's recommended settings.

The complete genome was re-sequenced, providing a 2x genomic coverage. It was then aligned and annotated against the mitochondrial genome of *Osteoglossum bicirrhosum* (GenBank# AB043025).

Data analysis

Orthologous protein-coding regions were aligned in Clustal W (Thompson *et al.*, 1996), and alignment was confirmed by conceptually translating protein-coding DNA regions into amino-acid sequences in BioEdit (Hall, 1999). Alignments of ribosomal and transfer RNAs were constructed in Clustal W (Thompson *et al.*, 1996) and manually adjusted, if necessary, to conform to secondary structural models (Kumazawa and Nishida, 1993; Ortí *et al.*, 1996; Wang and Lee, 2002; Waters *et al.*, 2002). Codon usage frequencies, and amino acid composition of the genome was inferred in the program MEGA 3.1 (Kumar *et al.*, 2004). Mitochondrial gene regions were tested for an anti-G bias characteristic of the mitochondrial DNA genes, but not of the nuclear genome, to support our conclusion that we have collected genuine mitochondrial DNA data (Zhang and Hewitt, 1996). Hairpins in the control regions were inferred using the software mFold (Zuker, 2003) implemented on the website www.idtdna.com.

Results

The total length of the *Arapaima gigas* mitochondrial genome is 16,433 bp. The genome sequence is deposited in GenBank under the accession number EF523611. The *Arapaima* gene order and content (Figure 1 and Table 2) is identical to the ancestral vertebrate state (*e.g.*, Inoue *et al.*, 2001, 2003a; Miya *et al.*, 2003). The genome codes for one

Table 1 - Primers used in the amplification and sequencing of the complete mitochondrial genome of *Arapaima gigas*. The primer designations correspond to their 3' position in the human mitochondrial genome (Anderson *et al.*, 1981) by convention. H and L designate the heavy and the light strand, respectively. Many of the primers reported for the first time in this study are used in ongoing studies in our laboratory, or were derived from primers published in Miya and Nishida (1999).

Location	Primer	Region	Primer sequence	Source
	amplification			
L1090	12Sa	12S	5'-AAACTGGGATTAGATACCCCACTA-3'	(Hrbek and Larson, 1999)
H8516	ATPr.1	ATP8/6	5'-CTTAGTGTCATGGTCAGTTTCA-3'	(Hrbek *et al.*, 2005)
L8537	ATPf.1	ATP8/6	5'-TGAAACTGACCATGACACTAAG-3'	(Hrbek *et al.*, 2005)
H15149	CBr.4	cytb	5'-CCTCARAAGGATATYTGTCCTCA-3'	This study
L15995b	Prof.1	$tRNA^{Pro}$	5'-CTCYCACCCCTGACTCCCAAAG-3'	This study
H693	12Sr.5	12S	5'-GGCGGATACTTGCATGT-3'	This study
	sequencing			
L185b	Dlf.4	D-loop	5'-GGCATTTGGTTCCTATTTCAGG-3'	This study
L617	Phef.1	$tRNA^{Phe}$	5'-AAGCATAACATTGAAGATG-3'	This study
H693	12Sr.5	12S	5'-GGCGGATACTTGCATGT-3'	This study
L1090	12Sa	12S	5'-AAACTGGGATTAGATACCCCACTA-3'	(Hrbek and Larson, 1999)
H1067	12Sr.4	12S	5'-TAGTGGGGTATCTAATCCCAGTTT-3'	This study
L1579	12Sf.2	12S	5'-AAGTCGTAACATGGTAAGTGYAC-3'	This study
H1782	16Sr.3	16S	5'-TTTCATCTTTCCCTTGCGGTAC-3'	(Hrbek and Larson, 1999)
H2001	16Sr.6	16S	5'-AACCAGCTATCACCAGGCTCG-3'	This study
L2021	16Sf.3	16S	5'-CGAGCCTGGTGATAGCTGGTT-3'	This study
H2493	16Sr.5	16S	5'-GATGTTTTTGGTAAACAGG-3'	This study
L2510	16sf.5	16S	5'-GCCCTGTTTACCAAAAACAT-3'	This study
L3002	16Sf.2	16S	5'-TACGACCTCGATGTTGGATCAGG-3'	(Hrbek *et al.*, 2005)
H3058	16Sr.4	16S	5'-CCGGTCTGAACTCAGATCACGT-3'	This study
L3079	16Sf.1	16S	5'-ACGTGATCTGAGTTCAGACCG-3'	(Hrbek *et al.*, 2005)
L3317	Leu	$tRNA^{Leu}$	5'-CCGGCCAATGCAAAAGACCTAA-3	This study
L3416	ND1f.4	ND1	5'-CATACAACTRCGAAAAGGRCC-3'	This study
L3899	ND1f.8	ND1	5'-GAAACAAACCGAGCCCCYTT-3'	This study
L4280	ILEr.4	$tRNA^{Ile}$	5'-ACTGTATCAAAGTGGYCCTT-3'	(Hrbek *et al.*, 2005)
L4299	ILEf.1	$tRNA^{Ile}$	5'-AAGGRTTACTTTGATAGAGT-3'	(Hrbek and Larson, 1999)
H4364	GLNr.2	$tRNA^{Gln}$	5'-GGAAGCACTARGAGTTTTGA-3'	This study
L4437	METf.6	$tRNA^{Met}$	5'-AAGCTTTYGGGCCCATACC-3'	(Macey *et al.*, 1997a)
L4882	ND2f.14	ND2	5'-TGACAAAARCTAGCCCC-3'	(Hrbek and Larson, 1999)
H4980	ND2r.6	ND2	5'-ATTTTTCGTAGTTGGGTTTGRTT-3'	This study
H5540	TRPr.5	$tRNA^{Trp}$	5'-TTTAAAGCTTTGAAGGC-3'	(Hrbek and Larson, 1999)
L5550	TRPf.1	$tRNA^{Trp}$	5'-CTAARAGCCTTCAAAGC-3'	This study
H5934	CO1r.1	CO1	5'-AGRGTGCCAATGTCTTTGTGRTT-3'	(Macey *et al.*, 1997a)
L6190	CO1f.1	CO1	5'-GCATTTCCGCGAATAAATAA-3'	This study
L6717	CO1f.1	CO1	5'-TACATRGGAATRGTATGAGC-3'	This study
H7414	CO1r.2	CO1	5'-GAAAAGCAGGTTCTTCAAATG-3'	This study
H7985	CO2r.2	CO2	5'-TCGGTGATCTACTTCTAATAGACG-3'	This study
L8106	CO2f.1	CO2	5'-TGGGTGTTAAAATAGATGC-3'	(Hrbek *et al.*, 2005)
H8516	ATPr.1	ATP8/6	5'-CTTAGTGTCATGGTCAGTTTCA-3'	(Hrbek *et al.*, 2005)

Table 1 (cont.)

Location	Primer	Region	Primer sequence	Source
L8537	ATPf.1	ATP8/6	5'-TGAAACTGACCATGACACTAAG-3'	(Hrbek *et al.*, 2005)
L9158	ATP6f.1	ATP6	5'-GCMGTAGCTATTATTCAAGC-3'	(Hrbek *et al.*, 2005)
H9264	CO3r.2	CO3	5'-GAGGAGAGCRGCRGATGCCCC-3'	(Hrbek *et al.*, 2005)
L9514	CO3f.1	CO3	5'-TTCTGAGCCTTCTTYCA-3'	This study
L10038	GLYf.1	$tRNA^{Gly}$	5'-CTTCCAATTATTTAATCTTG-3'	This study
L10765	ND4f.1	ND4	5'-TTAAATCTCTTACAATGCTA-3'	This study
L11414	ND4f.2	ND4	5'-GACTACCAAAAGCCCAYGTAGA-3'	This study
H11534	ND4r.1	ND4	5'-GCTATAACAATAAAGGGGTA-3'	This study
L12070	ND4f.3	ND4	5'-CACATTCACGAGAACACCTTCTCATA-3'	This study
L12321	Leuf.1	$tRNA^{Leu}$	5'-GGAACCAAAAACTCTTGGTGCAA-3'	This study
L12809	ND5f.6	ND5	5'-ATATCYTTCCTCCTAATTGGTTGATG-3'	This study
L13562	ND5f.3	ND5	5'-TCAYACCTAAAYGCTTCAGCCCT-3'	This study
H13600	ND5r.2	ND5	5'-AAGAAGATTACCCGGAAGCTGTA-3'	This study
L13991	ND5f.7	ND5	5'-GGACAAACCATAGCGTCTCAACT-3'	This study
H14080	ND5r.1	ND5	5'-AGGTAGGTTTTAATTAGACC-3'	This study
L14725	GLUf.3	$tRNA^{Glu}$	5'-GGCACGAAAAACCGCCGTTG-3'	This study
H15149	CBr.4	cytb	5'-CCTCARAAGGATATYTGTCCTCA-3'	This study
L15513c	CBf.2	cytb	5'-CTRGGAGACCCNGAAAACTT-3'	This study
L15923	THRf.4	$tRNA^{thr}$	5'-AACACAAAGCATCGGTCTTGTAA-3'	This study
H15976	PROr.2	$tRNA^{Pro}$	5'-TAGCTTTGGGAGTTAAGGGTGGG-3'	This study
L15995	Prof.1	$tRNA^{Pro}$	5'-CTCYCACCCCTGACTCCCAAAG-3'	This study

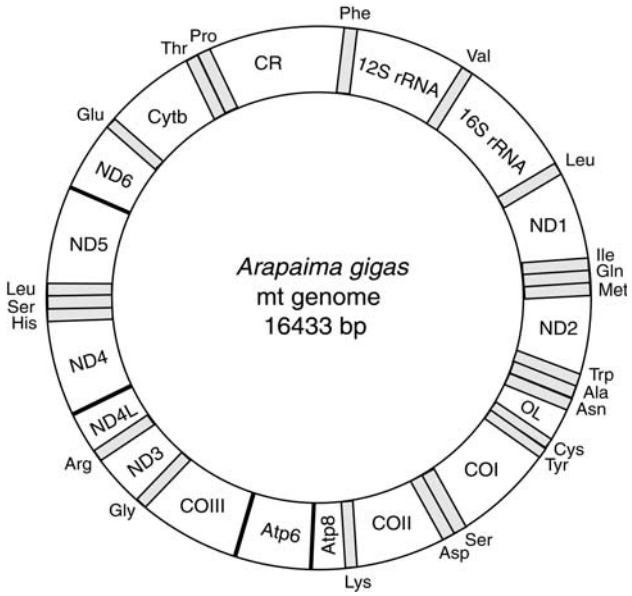

Figure 1 - Schematic map of the complete mitochondrial genome of *Arapaima gigas*.

subunit of the Cytochrome *b* (*cob*) which forms part of the ubiquinol cytochrome c oxidoreductase complex; three subunits of the Cytochrome oxidase (*cox*) which form part of the cytochrome c oxidase complex; seven subunits of the NADH dehydrogenase (*nad*) which form part of the nicotinamide adenine dinucleotide ubiquinone oxidoreductase complex; and two subunits of ATP synthase (*atp*). It also contains the small (*rrnS*) subunit and the large (*rrnL*) subunit ribosomal RNA genes and 22 tRNA genes (*trn*). A noncoding control region located between *trnP* and *trnF* genes contains the origin of heavy strand replication (O_H), and the light strand replication origin (O_L) is found between the tRNAs genes *trnN* and *trnC*.

The reference *Arapaima gigas* individual that we sequenced had a relatively short control region of 787 bp. A schematic representation is shown in Figure 2. Similar to typical vertebrate mitochondrion, this non-coding region contains the heavy-strand replication origin (O_H) and can be divided into three different domains (Brown *et al.*, 1986). Domain I is only 147 nucleotides long and contains a 23 bp thermo-stable hairpin suggested to be involved in the regulation of replication of the mitochondrial genome (Buroker *et al.*, 1990); however, it does not appear to contain the termination associated sequence (Doda *et al.*, 1981). Domain II, the central conserved block, extended from nucleotide 148 to 386. Domain III contained three conserved sequence blocks (CSB-1 at position 461-486; CSB-2 at position 577-593; and CSB-3 at position 620-637). Similar to the results reported by Broughton (2001), the CSB-1 was the least conserved, while CSB-3 was the most highly conserved of the three blocks. In do-

Table 2 - Gene organization of the *Arapaima gigas* mitochondrial genome.

Gene/element	Strand	Position	Size	Start	Stop
trnF	H	1..68	69	-	-
rrnS	H	69..1022	954	-	-
trnV	H	1023..1093	71	-	-
rrnL	H	1094..2779	1686	-	-
trnL (TAA)	H	2780..2853	74	-	-
nad1	H	2858..3829	972	ATG	TAA
trnI	H	3831..3902	72	-	-
trnQ	L	3903..3972	70	-	-
trnM	H	3972..4040	69	-	-
nad2	H	4041..5085	1045	ATG	T*
trnW	H	5086..5154	69	-	-
trnA	L	5156..5224	69	-	-
trnN	L	5226..5298	73	-	-
O_L	L	5299..5334	36	-	-
trnC	L	5335..5401	67	-	-
trnY	L	5402..5472	71	-	-
cox1	H	5474..7030	1557	GTG	AGA
trnS (TGA)	L	7026..7097	72	-	-
trnD	H	7101..7172	72	-	-
cox2	H	7177..7867	691	ATG	T*
trnK	H	7868..7941	74	-	-
atp8	H	7943..8110	168	ATG	TAA
atp6	H	8101..8784	684	ATG	TAA
cox3	H	8784..9568	785	ATG	T*
trnG	H	9569..9639	71	-	-
nad3	H	9640..9988	349	ATG	T*
trnR	H	9989..10058	70	-	-
nad4L	H	10059..10355	297	ATG	TAA
nad4	H	10349..11738	1390	ATG	T*
trnH	H	11739..11807	69	-	-
trnS (GCT)	H	11808..11875	68	-	-
trnL (TAG)	H	11876..11948	73	-	-
nad5	H	11949..13790	1842	ATG	TAA
nad6	L	13787..14317	531	ATG	TAA
trnE	L	14318..14385	68	-	-
cob	H	14390..15530	1141	ATG	T*
trnT	H	15531..15602	72	-	-
trnP	L	15603..15666	64	-	-
O_H (control region)	H	15667..16433	767	-	-

* = TAA stop codon is completed by the addition of 3' Adenine residues to the mRNA.

main III, a 14 unit AT repeat is present from position 678-705 and shortly thereafter it is followed by mono-nucleotide adenine and thiamine repeats. Repeat sequences are 5 A residue, 11 T residue, 6 T residue and 9 A residue mono-nucleotide repeats separated by short non-repeat sequence regions. As expected, the control region is heavily biased against guanine with a composition of 0.342 (A), 0.217 (C), 0.132 (G), and 0.309 (T). The control region for O_L contains a highly conserved hairpin loop with a perfectly complementary bases-pairing stem (CCTCCGCCT/AGGCGGGAGG). The secondary structure of the O_L has been suggested to regulate light-strand replication (Wong and Clayton, 1985).

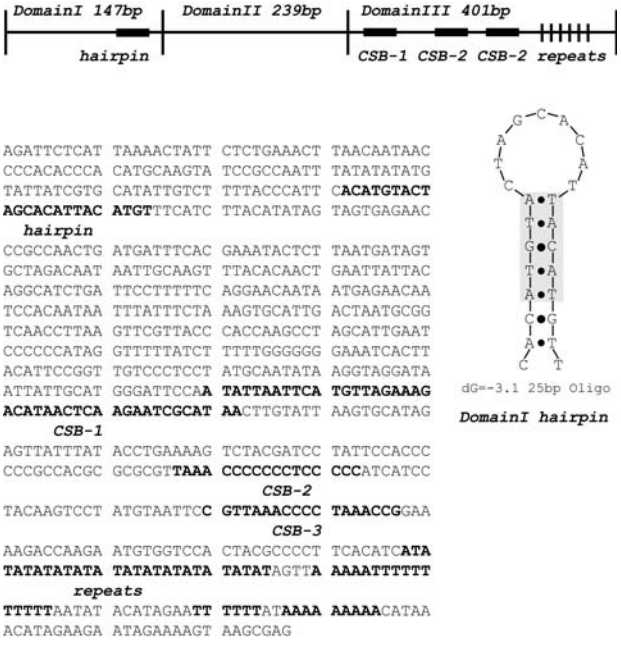

Figure 2 - Schematic map characterizing of the control region of *Arapaima gigas*. Indicated are the three domains, a potentially regulatory hairpin in domain 1, the three conserved sequence blocks in domain 3, and a series of repeats in domain 3.

With the exception of *cox1* which starts with GTG, all protein-coding genes begin with the ATG start codon; stop codons include 12 TAA, six of which are incomplete, and one AGA (Table 2). Incomplete stop codons are common in mitochondrial genes, and TAA stop codons are created via posttranscriptional polyadenylation of the 3' end of the mRNA (Ojala *et al.*, 1981). Reading frames of three pairs of genes, *atp8-atp6*, *nad4L - nad4* and *nad5-nad6* overlap by several nucleotides, a pattern which is also common in other vertebrate mitochondrial genomes. Some genes are separated by up to five nucleotide non-coding spacers (Table 2).

Codon usage in the 13 coding genes consisted of 28.4% A, 26.1% C, 13.8% G and 30.7% T bases. These values were similar to those observed in other osteoglossiform fishes and show a strong anti-G bias. The anti-G bias was especially pronounced in the third position of the 12 heavy-strand encoded genes which consisted of 41.7% A, 28.1% C, 3.8% G and 26.4% T bases (Figure 3; Table 3). The rank order of nucleotide usage frequency at the third codon position is the same as in *Osteoglossum bicirrhosus* and *Pantodon buchholzi*; however, in *Scleropage formosus* the rank order of A and C is reversed. The most frequently

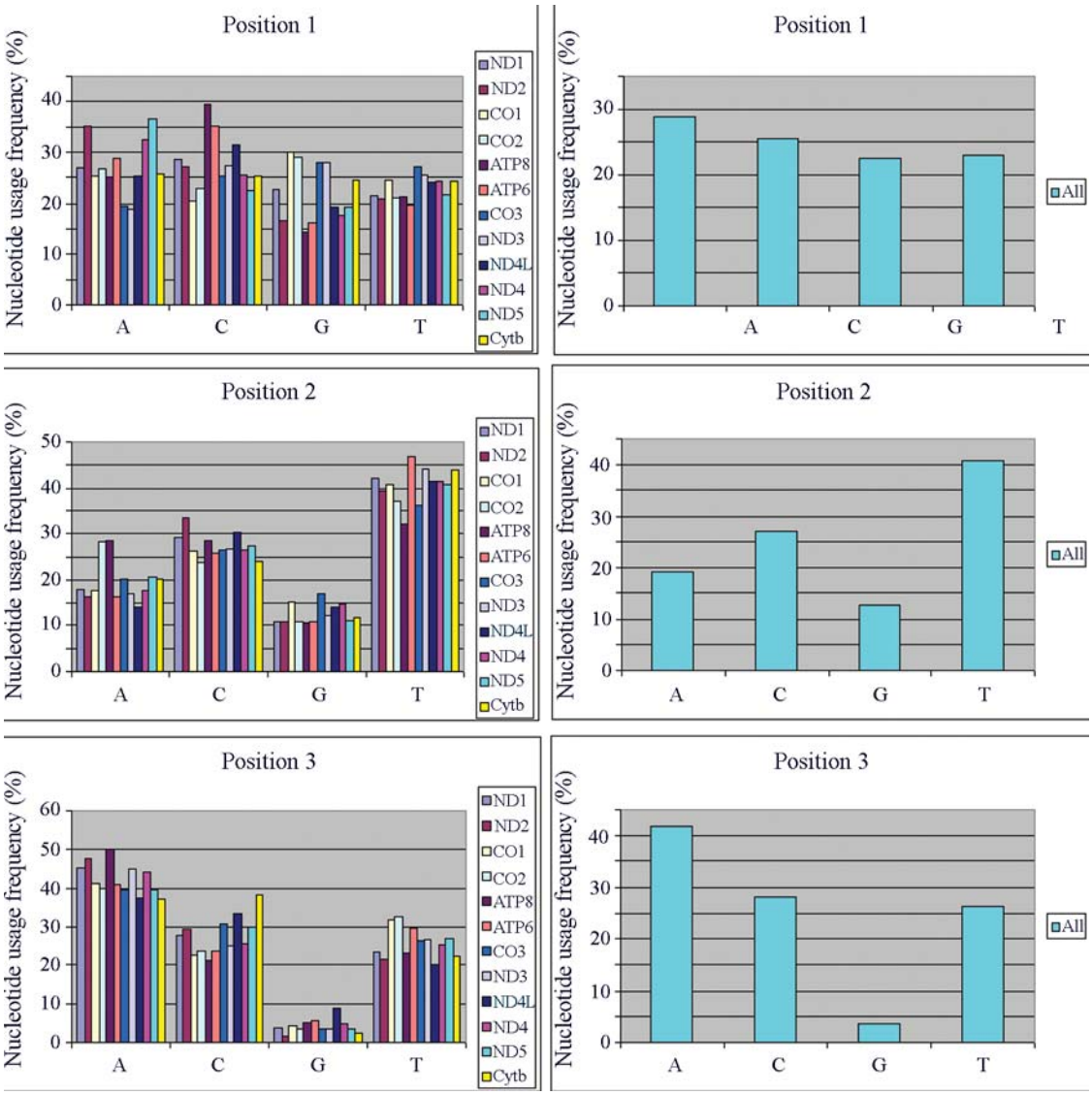

Figure 3 - Nucleotide composition of the 12 mitochondrial genes coded on the heavy strand. Nucleotide composition at first, second and third positions for individual genes is presented on the left side. On the right side, averages across all genes are presented.

encoded amino acids were leucine (16.65%), followed by threonine (8.25%), isoleucine (8.17%) and alanine (7.72%). The least common amino acid was cystein (0.79%) - see Table 4.

All 22 *Arapaima* mitochondrial tRNA genes possess anticodons that match the vertebrate mitochondrial genetic code (Kumazawa and Nishida, 1993). Each tRNA sequence may be folded into a cloverleaf structure with 7 bp in the aminoacyl stem, 5 bp in the TΨC and anticodon stems, and 4 bp in the DHU stem. tRNA stem regions include some non-complementary base pairings, a pattern also commonly observed in other vertebrates (*e.g.*, Kumazawa and Nishida, 1993). The 3' CCA nucleotide tail of mature tRNAs is most likely added post-transcriptionally (Roe *et al.*, 1985).

When the *rrnS* and *rrnL* genes are transcribed into putative rRNAs, both rRNA sequences may also be folded into secondary structures. Stem regions appear to be conserved, whereas loop regions are somewhat more variable relative to other vertebrate sequences. The functional requirement for specific base pairing appears to constrain the evolution of stems relative to some portions of loops (*e.g.*, Sullivan *et al.*, 1995; Ortí *et al.*, 1996; Wang and Lee, 2002).

Discussion

One of the motivations of our study was to sequence the mitochondrial control region, and determine the factors which impeded its efficient use in phylogeographic and population genetic analyses (Hrbek *et al.*, 2005). In spite of designing specific, highly stringent primers, Hrbek *et al.*

Table 3 - Codon usage of the *Arapaima gigas* mtDNA.

Amino acid (anticodon)	Codon group	Usage of codon ending in				Total	%
		A	C	G	T		
Ala (UGC)	GCN	94	108	4	88	294	7.53
Cys (GCA)	TGY	0	19	0	11	30	0.77
Asp (GUC)	GAY	0	37	0	39	76	1.95
Glu (UUC)	GAR	90	0	6	0	96	2.46
Phe (GAA)	TTY	0	120	0	127	247	6.33
Gly (UCC)	GGN	92	63	35	46	236	6.05
His (GUG)	CAY	0	66	0	45	111	2.84
Ile (GAU)	ATY	0	110	0	201	311	7.97
Lys (UUU)	AAR	82	0	5	0	87	2.23
Leu (UAG)	CTN+TTR	367	116	44	107	634	16.24
Met (CAU)	ATR	146	0	40	0	186	4.77
Asn (GUU)	AAY	0	70	0	64	134	3.43
Pro (UGG)	CCN	122	36	7	43	208	5.33
Gln (UUG)	CAR	94	0	94	0	188	4.82
Arg (UCG)	CGN	44	12	4	12	72	1.84
Ser (UGA)	TCN+AGY	89	99	4	60	252	6.46
Thr (UGU)	ACN	139	86	8	81	314	8.05
Val (UAC)	GTN	86	35	12	54	187	4.79
Trp (UCA)	TGR	111	0	9	0	120	3.07
Tyr (GUA)	TAY	0	49	0	64	113	2.90
Stop (UUA)	TAR	5	0	1	0	6	0.15
Stop (UCA)	TGR	1	0	0	0	1	0.03
Total		1562	1026	273	1042	3903	100.00

(2005) were unable to obtain amplification results in the majority of the individuals used in their conservation genetic study. Not all individuals amplified, and those that did, often produced only a weak product often with large size differences among PCR products. PCR amplification of a number of individuals also produced multiple bands suggesting possible mtDNA heteroplasmy. Inoue *et al.* (2001) encountered similar difficulties (pers. com.), and for these reasons did not characterize the control regions of the two osteoglossiform species *Osteoglossum bicirrhosum* (GenBank# AB043025) and *Pantodon buchholzi* (GenBank# AB043068). A recent study of Yue *et al.* (2006) characterized the complete mitochondrial genome of the Asian arowana *Scleropages formosus* and reported tandem repeats in the 5' and 3' ends of the control region, as well as mitochondrial heteroplasmy. These observations suggest that the control region anomalies observed in *Arapaima* may be a general property of the control regions of the fishes of the order Osteoglossiformes.

The control region of the reference specimen of *Arapaima gigas* is relatively short at 787 bp. It contains three domains (Brown *et al.*, 1986). Domain I appears to lack the termination associated sequence (Doda *et al.*, 1981), but it does contain a 23 bp thermo-stable hairpin (Figure 2). The hairpin contains the ATGTA/TACAT motif which Yue *et al.* (2006) also observed in other osteoglossiform fishes and in fishes of the genus *Anguilla*. Previously this motif was observed only in mammals (Saccone *et al.*, 1991) and the lungfish (Zardoya and Meyer, 1996). Although not pointed out by Yue *et al.* (2006), the hairpin is actually inverted (TACAT/ATGTA) in the phylogenetically closely related Asian (*Scleropages formosus*) and the silver (*Osteoglossum bicirrhosus*) arowanas, but not in other osteoglossiform species. The inverted repeat appears to be a molecular synapomorphy for the *Scleropages* + *Osteoglossum* clade. Same as in *Scleropages formosus*, domain III of *Arapaima gigas* contains microsatellite repeats; specifically domain III of the reference individuals contains a 14 unit AT repeat followed by mono-nucleotide repeat sequences of A (5x), T (11x), T (6x) and A (9x). Repeats in both the 5' end (domain I) and the 3' end (domain III) of the control region are rare and currently have only been reported in *Scleropages formosus* (GenBank# DQ023143, Yue *et al.*, 2006).

Although only preliminarily characterized, a similar pattern of repeats is observed in the 5' and 3' ends of the control region of *Heterotis niloticus*, the African sister taxon of *Arapaima gigas*. The control regions of *Osteoglossum bicirrhosum* (GenBank# AB043025) and the African *Pantodon buchholzi* (GenBank# AB043068) also

Table 4 - Amino acid usage (%) in the 13 protein coding genes of the *Arapaima gigas* mtDNA.

	ND1	ND2	CO1	CO2	ATP8	ATP6	CO3	ND3	ND4L	ND4	ND5	ND6	Cytb	Avg
Ala	9.91	8.91	8.30	5.22	1.82	7.05	8.43	8.62	9.18	5.83	7.50	9.66	7.37	7.72
Cys	0.00	0.29	0.19	0.87	0.00	0.44	1.15	0.86	3.06	1.08	1.14	1.70	0.79	0.79
Asp	0.93	0.00	2.70	5.65	1.82	0.44	1.92	4.31	1.02	1.30	2.12	1.70	2.89	2.00
Glu	3.41	1.72	2.12	5.22	1.82	1.32	3.83	5.17	2.04	2.38	1.79	3.41	1.58	2.52
Phe	6.19	4.31	9.07	3.91	5.45	5.73	8.81	7.76	7.14	3.67	7.18	5.68	7.89	6.49
Gly	5.26	4.60	9.27	3.48	1.82	4.41	8.05	5.17	5.10	5.18	5.06	14.20	6.32	6.20
His	1.24	2.01	3.47	4.78	10.91	3.08	5.75	0.86	5.10	2.16	2.61	0.00	2.89	2.91
Ile	9.60	8.91	7.53	8.26	3.64	10.13	4.98	6.90	4.08	9.94	9.14	3.41	8.68	8.17
Lys	2.48	2.87	1.74	2.61	3.64	0.44	1.15	0.86	0.00	3.24	3.43	0.57	2.63	2.28
Leu	19.50	19.83	11.39	10.00	10.91	25.11	13.03	20.69	23.47	19.44	15.17	17.61	16.32	16.65
Met	3.41	4.89	5.21	5.22	5.45	3.08	3.45	3.45	5.10	5.62	6.53	5.11	4.21	4.88
Asn	4.33	4.02	2.70	2.17	3.64	3.52	0.38	1.72	2.04	2.81	6.20	1.70	4.74	3.52
Pro	7.12	5.75	5.79	6.52	14.55	5.73	4.98	6.90	3.06	5.62	3.75	2.84	5.53	5.46
Gln	2.17	3.74	1.74	3.48	5.45	4.85	2.68	2.59	3.06	2.38	3.10	0.00	1.58	2.63
Arg	2.48	1.15	1.54	2.61	0.00	2.64	2.3	1.72	2.04	2.38	1.31	2.27	1.84	1.89
Ser	6.50	7.47	5.79	6.96	3.64	4.85	5.75	5.17	10.20	7.34	7.99	5.68	5.79	6.62
Thr	6.50	12.93	7.14	6.96	9.09	10.13	8.43	6.03	11.22	9.29	9.79	1.70	5.53	8.25
Val	3.41	1.44	7.53	9.57	7.27	3.08	6.13	5.17	2.04	2.81	2.77	11.36	6.58	4.91
Trp	2.48	3.16	3.47	2.17	9.09	1.76	4.60	4.31	1.02	4.32	2.12	3.98	2.89	3.15
Tyr	3.10	2.01	3.28	4.35	0.00	2.20	4.21	1.72	0.00	3.24	1.31	7.39	3.95	2.97
Total	323	348	518	230	55	227	261	116	98	463	613	176	380	292.9

contain a large and complex tandem repeats in the 5' end of the control region which corresponds to the domain I hairpin (Yue *et al.*, 2006). However, confirmation of the exact pattern and structure of the control regions of these species is not possible since the central and 3' end portions of the control regions are not available. The control region of the mormyriid *Gnathonemus petersii* lacks large blocks of repeats, and appears to contain all three CSBs. The mormyriids together with hiodontids and notopterids are sister clade to the clade containing the genera *Arapaima*, *Heterotis*, *Sclerophages*, *Osteoglossum* and *Pantodon* (Nelson, 1994; Sullivan *et al.*, 2000), which suggests that the observed control region peculiarities are phylogenetically restricted.

The mitochondrial control region regulates replication of the heavy strand and transcription (see review in Shadel and Clayton, 1997). Together with the conserved sequence blocks whose role is involved in positioning RNA polymerase for transcription and for priming replication (Clayton, 1991; Shadel and Clayton, 1997), an important regulatory element is the termination associated sequence (TAS) normally observed in domain I. TAS appears to act as a signal for termination of D-loop strand synthesis, however, it could not be identified in our reference individual. We speculate that the 23 bp thermo-stable hairpin found in domain I (Figure 2) may take on the role of a signal for termination of D-loop strand synthesis in the absence of TAS. This conclusion is contrary to that of Yue *et al.* (2006) who suggest that the domain I hairpin may be a binding site for proteins involved in replication. Elucidating the role of the domain I hairpin and understanding the consequence of the apparent lack of TAS for mitochondrial replication and for the transcription of mitochondrial genes, if any, will require biochemical and cell molecular studies, however. The second major regulatory region, the light strand replication origin, is found between the genes *trnN* and *trnC*. It is represented by a 35-bp non-coding sequence which may be folded into secondary structure consisting of a perfect 9-bp stem and a 13-bp loop. Secondary structures at the light strand replication origin may act as initiation signals for light strand replication (Wong and Clayton, 1985) and appear to be fully functional.

Protein coding genes are characterized by an anti-G bias which is particularly strong at the third codon position where G is present at only 3.8% frequency (Figure 3; Table 3). The anti-G bias may be due in part to selection against less stable G nucleotides on the light strand, which is exposed as a single strand for a considerable length of time during the asymmetrical replication of mtDNA (Clayton, 1982). A further implication of the model of Clayton (1982) has been pointed out by Reyes *et al.* (1998). The deamination of cytosine into uracil and adenine into hypoxanthine on the heavy strand would lead to a decrease in G content in the light strand, and an increase in G on the heavy strand. The low G content observed in mitochondrial genes may, thus, also be a result of the asymmetrical replication

of the mitochondrial genome. Still further contribution to the anti-G bias may result from the preference for adenine during mRNA transcription, as ATP is generally the most common ribonucleotide available in the mitochondria and, thus, is most efficiently transcribed (Xia, 1996). In contrast to G, the most commonly used nucleotide is A which is also the most commonly used nucleotide in *Osteoglossum bicirrhosum* and *Pantodon buchholzi*, but not in *Scleropages formosus*. Amino acid usage is also similar to that observed in other osteoglossiform fishes, and is heavily biased towards the use of leucine.

It is clear that the control region patterns, or their variations, observed in *Arapaima gigas* are also observed in other osteoglossiform fishes. What is unclear is if these control region characteristics are due to phylogenetic conservatism, or due to homoplasy. No matter what the mechanism, the control region is unlikely to be a suitable phylogenetic marker for phylogeographic and population-level studies due to large stretches of repeats and secondary structures which make amplification and sequencing difficult. Further complications arise due to mitochondrial heteroplasmy potentially caused by slip-strand replication (Macey *et al.*, 1997c) of the domain I hairpin and or of the domain III microsatellite region.

The availability of the complete genome of *Arapaima gigas* will facilitate molecular population studies of both the pirarucu and other osteoglossiform fishes, such as the two species of arowana *Osteoglossum bicirrhosum* and *Osteoglossum ferreirai*, and the aquiculturally important African species *Heterotis niloticus*. The mitochondrial genome is composed of a mosaic of highly conserved and highly variable sections among the evolutionarily divergent *Arapaima gigas* and *Osteoglossum bicirrhosum*. This characteristic greatly facilitates choosing appropriately informative genomic regions for particular questions, as well as primer design for other osteoglossiform species.

Acknowledgments

This research was supported by a research grant from the Fundação de Amparo a Pesquisa do Estado do Amazonas (FAPEAM) to IPF. Permission to conduct fieldwork and to collect tissue samples was granted by IBAMA (License n. 48/2000), and to conduct genetic assessment by CGEN/IBAMA (Deliberation n. 75). We thank Marcelo Crossa for helping us obtain tissue samples, and two anonymous reviewers for suggestions. TH acknowledges FAPEAM and the J. William Fulbright Foundation for financial support.

References

Anderson SA, Bankier T, Barrell BG, de Bruijn MHL, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, *et al.* (1981) Sequence and organization of the human mitochondrial genome. Nature 290:457-465.

Avise JC (2004) Molecular Markers, Natural History and Evolution. 2nd ed. Sinauer Associates, Sunderland, 541 pp.

Boore JL (1999) Animal mitochondrial genomes. Nucleic Acids Res 27:1767-1780.

Boore JL and Brown WM (1998) Big trees from little genomes: Mitochondrial gene order as a phylogenetic tool. Curr Op Gen Dev 8:668-674.

Boore JL, Lavrov D and Brown WM (1997) Mitochondrial gene rearrangements trace metazoan phylogeny. Am Zool 37:70A.

Brinkmann H, Denk A, Zitzler J, Joss JJ and Meyer A (2004) Complete mitochondrial genome sequences of the South American and the Australian lungfish: Testing of the phylogenetic performance of mitochondrial data sets for phylogenetic problems in tetrapod relationships. J Mol Evol 59:834-848.

Broughton RE, Milam JE and Roe BA (2001) The complete sequence of the Zebrafish (*Danio rerio*) mitochondrial genome and evolutionary patterns in vertebrate mitochondrial DNA. Genome Res 11:1958-1967.

Brown GG, Gadaleta G, Pepe G, Saccone C and Sbis E (1986) Structural conservation and variation in the D-loop-containing region of vertebrate mitochondrial DNA. J Mol Biol 192:503-511.

Brown WM (1985) The mitochondrial genome of animals. In: MacIntyre RJ (ed) Molecular Evolutionary Genetics. Plenum Press, New York, pp 95-130.

Buroker NE, Brown JR, Gilbert TA, O'Hara PA, Beckenback AT, Thomas WK and Smith MJ (1990) Length heteroplasmy of sturgeon mitochondrial DNA: An illegitimate elongation model. Genetics 124:157-163.

Clayton DA (1982) Replication of animal mitochondrial DNA. Cell 28:693-705.

Clayton DA (1991) Nuclear gadgets in mitochondrial DNA replication and transcription. Trends Biotech 16:107-111.

Curole JP and Kocher TD (1999) Mitogenomics: Digging deeper with complete mitochondrial genomes. Trends Ecol Evol 14:394-398.

Doda JN, Wright CT and Clayton DA (1981) Elongation of displacement-loop strands in human and mouse mitochondrial DNA is arrested near specific template sequences. Proc Natl Acad Sci USA 78:6116-6120.

Hall T (1999) BioEdit: A user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp Ser 41:95-98.

Haring E, Kruckenhauser L, Gamauf A, Riesing MJ and Pinsker W (2001) The complete sequence of the mitochondrial genome of *Buteo buteo* (Aves, Accipitridae) indicates an early split in the phylogeny of raptors. Mol Biol Evol 18:1892-1904.

Hrbek T and Larson A (1999) The evolution of diapause in the killifish family Rivulidae (Atherinomorpha, Cyprinodontiformes): A molecular phylogenetic and biogeographic perspective. Evolution 53:1200-1216.

Hrbek T, Farias IP, Crossa M, Sampaio I, Porto JIR and Meyer A (2005) Population genetic analysis of *Arapaima gigas*, one of the largest freshwater fishes of the Amazon basin: Implications for its conservation. Anim Conserv 8:297-308.

Inoue JG, Miya M, Tsukamoto K and Nishida M (2001) A mitogenomic prespective on the basal teleostean phylogeny: Re-

solving higher-level relationships with longer DNA sequences. Mol Phylogenet Evol 20:275-285.

Inoue JG, Miya M, Tsukamoto K and Nishida M (2003a) Basal actinopterygian relationships: A mitogenomic perspective on the phylogeny of the "ancient fish". Mol Phylogenet Evol 26:110-120.

Inoue JG, Miya M, Tsukamoto K and Nishida M (2003b) Evolution of the deep-sea Gulper eel mitochondrial genomes: Large-scale gene rearrangements originated within the eels. Mol Biol Evol 20:1917-1924.

Kumar S, Tamura K and Nei M (2004) MEGA3: Integrated software for molecular evolutionary genetics analysis and sequence alignment. Briefings Bioinfor 5:150-163.

Kumazawa Y and Nishida M (1993) Sequence evolution of mitochondrial tRNA genes and deep-branch animal phylogenetics. J Mol Evol 37:380-398.

Macey JR, Larson A, Ananjeva NB, Fang Z and Papenfuss TJ (1997a) Two novel gene orders and the role of light-strand replication in the rearrangement of the vertebrate mitochondrial genome. Mol Biol Evol 14:91-104.

Macey JR, Larson A, Ananjeva NB and Papenfuss TJ (1997b) Evolutionary shifts in three major structural features of the mitochondrial genome among iguanian lizards. J Mol Evol 44:660-674.

Macey JR, Larson A, Ananjeva NB and Papenfuss TJ (1997c) Replication slippage may cause parallel evolution in the secondary structures of mitochondrial transfer RNAs. Mol Biol Evol 14:31-39.

Mindell DP, Sorenson MD and Dimcheff DE (1998) Multiple independent origins of mitochondrial gene order in birds. Proc Natl Acad Sci USA 95:10693-10697.

Miya M, Kawaguchi A and Nishida M (2001) Mitogenomic exploration of higher teleostean phylogenies: A case study for moderate-scale evolutionary genomics with 38 newly determined complete mitochondrial DNA sequences. Mol Biol Evol 18:1993-2009.

Miya M and Nishida M (1999) Organization of the mitochondrial genome of a deep-sea fish, *Gonostoma gracile* (Teleostei, Stomiiformes): First example of transfer RNA gene rearrangements in bony fishes. Mar Biotech 1:416-426.

Miya M, Takeshima H, Endo H, Ishiguro NB, Inoue JG, Mukai T, Satoh TP, Yamaguchi M, Kawaguchi A and Mabuchi K (2003) Major patterns of higher teleostean phylogenies: A new perspective based on 100 complete mitochondrial DNA sequences. Mol Phylogenet Evol 26:121-138.

Morrison CL, Harvey AW, Lavery S, Tieu K, Huang Y and Cunningham CW (2002) Mitochondrial gene rearrangements confirm the parallel evolution of the crab-like form. Proc R Soc London B 269:345-350.

Naylor GJP and Brown WM (1998) Amphioxus mitochondrial DNA, chordate phylogeny, and the limits of inference based on comparisons of sequences. Syst Biol 47:61-76.

Nelson JS (1994) Fishes of the World. 3rd ed. John Wiley and Sons, Inc., New York, 624 pp.

Ojala D, Montoya J and Attardi G (1981) tRNA punctuation model of RNA processing in human mitochondria. Nature 290:470-474.

Ortí G, Petry P, Porto JIR, Jégu M and Meyer A (1996) Patterns of nucleotide changes in mitochondrial ribosomal RNA genes and the phylogeny of piranhas. J Mol Evol 42:169-182.

Pääbo S, Thomas WK, Whitfield KM, Kumazawa Y and Wilson AC (1991) Rearrangements of mitochondrial transfer RNA genes in marsupials. J Mol Evol 33:426-430.

Pereira SL (2000) Mitochondrial genome organization and vertebrate phylogenetics. Genet Mol Biol 23:745-752.

Reyes A, Gissi C, Pesole G and Saccone C (1998) Asymmetrical directional mutation pressure in the mitochondrial genome of mammals. Mol Biol Evol 15:957-966.

Roe BA, Ma D-P, Wilson RK and Wong JF-H (1985) The complete nucleotide sequence of the *Xenopus laevis* mitochondrial genome. J Biol Chem 260:9759-9774.

Saccone C, Pesole G and Sbisá E (1991) The main regulatory region of mammalian mitochondrial DNA: Structure-function model and evolutionary pattern. J Mol Evol 33:83-91.

Shadel GS and Clayton DA (1997) Mitochondrial DNA maintenance in vertebrates. Annu Rev Biochem 66:409-435.

Sullivan JP, Holsinger KE and Simon C (1995) Among-site rate variation and phylogenetic analysis of 12S rRNA in sigmodontine rodents. Mol Biol Evol 12:988-1001.

Sullivan JP, Lavoué S and Hopkins CD (2000) Molecular systematics of the African electric fishes (Mormyroidea, Teleostei) and a model for the evolution of their electric organs. J Exp Biol 203:665-683.

Thompson JD, Higgins DG and Gibson TJ (1996) 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.

Wang H-Y and Lee S-C (2002) Secondary structure of mitochondrial 12S rRNA among fish and its phylogenetic applications. Mol Biol Evol 19:138-148.

Waters JM, Saruwatari T, Kobayashi T, Oohara I, McDowall RM and Wallis GP (2002) Phylogenetic placement of retropinnid fishes: Data set incongruence can be reduced by using asymmetric character state transformation costs. Syst Biol 51:432-449.

Wong TW and Clayton DA (1985) *In vitro* replication of human mitochondrial DNA: Accurate initiation at the origin of light-strand synthesis. Cell 42:951-958.

Xia X (1996) Maximizing transcription efficiency causes codon usage bias. Genetics 144:1309-1320.

Yue GH, Liew WC and Orban L (2006) The complete mitochondrial genome of a basal teleost, the Asian arowana (*Scleropages formosus*, Osteoglossidae). BMC Genomics 7:1-13.

Zardoya R and Meyer A (1996) The complete nucleotide sequence of the mitochondrial genome of the lungfish (*Protopterus dolloi*) supports its phylogenetic position as a close relative of land vertebrates. Genetics 142:1249-1263.

Zhang D-X and Hewitt GM (1996) Nuclear integrations: Challenges for mitochondrial DNA markers. Trends Ecol Evol 11:247-251.

Zuker M (2003) Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res 31:3406-3415.

Associate Editor: Antonio Solé-Cava