Repetitive Sequences in the Crocodilian Mitochondrial Control Region: Poly-A Sequences and Heteroplasmic Tandem Repeats

David A. Ray[1] *and Llewellyn D. Densmore*

Department of Biological Sciences, Texas Tech University, Lubbock

Heteroplasmic tandem repeats in the mitochondrial control region have been documented in a wide variety of vertebrate species. We have examined the control region from 11 species in the family Crocodylidae and identified two different types of heteroplasmic repetitive sequences in the conserved sequence block (CSB) domain—an extensive poly-A tract that appears to be involved in the formation of secondary structure and a series of tandem repeats located downstream ranging from approximately 50 to approximately 80 bp in length. We describe this portion of the crocodylian control region in detail and focus on members of the family Crocodylidae. We then address the origins of the tandemly repeated sequences in this family and suggest hypotheses to explain possible mechanisms of expansion/contraction of the sequences. We have also examined control region sequences from *Alligator* and *Caiman* and offer hypotheses for the origin of tandem repeats found in those taxa. Finally, we present a brief analysis of intraindividual and interindividual haplotype variation by examining representatives of Morelet's crocodile (*Crocodylus moreletii*).

Introduction

The control region of the mitochondrial genome (fig. 1) is usually divided into three domains (Anderson et al. 1981). Domain I, also known as the ETAS domain, has been shown to contain sequences (TAS/ETAS) associated with termination of the newly synthesized H-strand (the D-loop) during replication (Doda, Wright, and Clayton 1981; Brown et al. 1986; Foran, Hixson, and Brown 1988; Sbisa et al. 1997) and is usually highly variable. Domain II, the central conserved domain (CCD), is typically less variable and has been shown to contain several areas of highly conserved sequences, the B, C, D, E, and F boxes first identified by Anderson et al. (1981). The functions of these sequences are unknown. Domain III, the CSB domain, usually contains the origin of H-strand replication (OH) and the D-loop, three short, conserved sequence blocks (CSBs), and the promoters for H-strand and L-strand transcription (HSP and LSP) (Walberg and Clayton 1981; Brown et al. 1986; King and Low 1987; Foran, Hixson, and Brown 1988). Like domain I, this region has been shown to be highly variable and often contains tandemly repeated sequences of varying copy number. Recently, we reported the first analysis of the overall structure of the control region in crocodilians (Ray and Densmore 2002). In that paper, we described the locations of putative ETAS sequences, the B, C, D, E, and F boxes, and the CSB sequences and indicated the presence of two heteroplasmic repeat regions in domain III. In this contribution, we describe these repeat regions in detail and discuss them in terms of their phylogenetic distribution. We also propose mechanisms by which they may have arisen.

Tandemly repeated sequences in the control region of the mitochondrial genome have been documented in a wide range of taxa (reviewed in Lunt, Whipple, and Hyman 1998). The repeats described range in size from small microsatellite-like repeats of 3 bp to long minisatellite repeats of 777 bp and are usually located in the variable flanking regions (the ETAS and CSB domains). This positioning is consistent with the model of replication slippage (Levinson and Gutman 1987), which has been implicated in their formation, expansion, and contraction. Moreover, the association of the ETAS and CSB domains with the initiation and termination of transcription and replication presumably make both of these regions more prone to slippage events.

The most commonly invoked model for explaining the change in number of repeat units in the control region is the illegitimate elongation model of Buroker et al. (1990). This model uses competitive displacement of the H-strand by the D-loop to insert and remove repeats and provides a reasonable explanation for this process in and around the ETAS domain of the control region. Unfortunately, by relying on this competition, the applicability of the model is limited to tandem repeats in the ETAS domain. Two other models have been suggested that are more appropriate to repeated units at the 3′ end of the control region. Broughton and Dowling (1997) offered two models to explain duplications in the control region—the heavy strand and light strand models. Both invoke improper upstream initiation of replication and duplication of the "extra" sequence just before termination to explain the presence of repeats in this part of the genome. By contrast, the model of Mundy, Winchell, and Woodruff (1996) suggests that secondary structure characterizing the neighboring 12S rRNA or tRNA genes contributes to the formation of tandem repeats by causing strand slippage in the nascent heavy strand. Other authors have proposed models of mitochondrial recombination for increasing or decreasing the number of repeats (Lunt and Hyman 1997), but this is considered rare at best (Erye-Walker and Awadalla 2001; Ladoukakis and Zouros 2001).

The study of repeated sequences is important for several reasons. Most obviously, they lend insight into the processes involved in the evolution of the nuclear and mitochondrial genomes. They may also provide information on genome function. For example, in Adelie penguins (*Pygoscelis adeliae*), repeated sequences contain the transcriptional promoters for both the mitochondrial

[1] Present address: Department of Biological Sciences, Louisiana State University, Baton Rouge.

Key words: heteroplasmy, *Crocodylus*, mitochondrial control region, VNTR, Crocodylia.

E-mail: daray@lsu.edu.

Mol. Biol. Evol. 20(6):1006–1013. 2003
DOI: 10.1093/molbev/msg117
© 2003 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038

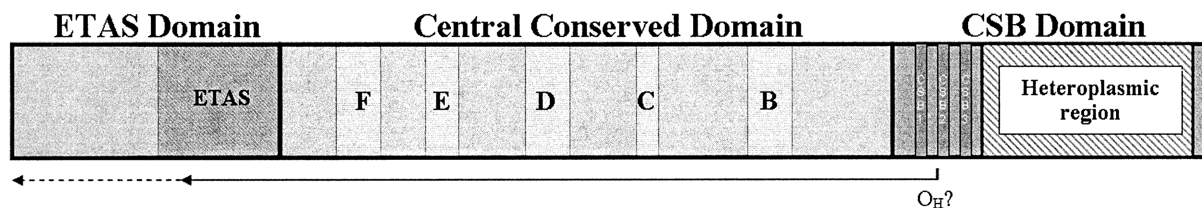


FIG. 1.—Schematic of the crocodylian mitochondrial DNA control region. Major features and the heteroplasmic region discussed in the text are indicated.

heavy and light strands (Ritchie and Lambert 2000). As a result, we should expect to find transcripts of varying sizes possibly requiring differential processing. Several human diseases have been linked to mitochondrial heteroplasmy, and animal model systems may provide insights into the dynamics of transmission and development of these diseases (Tamura et al. 1999; Bianchi, Bianchi, and Richard 2001; Sanchez-Cespedes et al. 2001). Studying tandem repeats in the animal mitochondrial genome may also be useful from a population level perspective. A number of researchers have attempted to employ these types of repeats to study population dynamics (Rand and Harrison 1989; Arnason and Rand 1992; Brown, Beckenbach, and Smith 1992; Hoelzel, Hancock, and Dover 1993; Cesaroni et al. 1997) with mixed results.

Herein, we describe tandem repeats in the mitochondrial control region of 11 species in the family Crocodylidae, including interindividual and intraindividual variation in *Crocodylus moreletii*. We also comment on similar repeat units in the control region of *Caiman crocodilus*, *Alligator mississippiensis*, and *A. sinensis*.

Materials and Methods

Blood from *Osteolaemus tetraspis tetraspis* and ten species of Crocodylus (*C. acutus*, *C. intermedius*, *C. moreletii*, *C. palustris*, *C. porosus*, *C. niloticus*, *C. mindorensis*, *C. rhombifer*, *C. siamensis*, and *C. cataphractus*) was obtained by puncture of the caudal or dorsal sinus (Gorzula, Arocha-Pinango, and Salazar 1976; Bayliss 1987). Total genomic DNA was then isolated for all taxa using either the SDS-Urea preparation of White et al. (1998) or the Gentra Puregene isolation kit (Gentra Systems, Minneapolis, Minn.). Both methods produce high-molecular-weight DNA appropriate for amplification using standard PCR protocols. Cesium chloride gradient preparations of mtDNA (White et al. 1998) from selected animals were also used to ensure that we were not amplifying nuclear translocations of the D-loop. We obtained the same sequences in either case.

The primers CR2L (5′-CGT TAT ACA TAT TAC TCT TTA ATT AGG CCC CC-3′) and 12SH1 (5′-GTT GAG CAG TAG CTA ATA ATA AGG TCA GGA-3′) were used to amplify the 3′ end of the D-loop in 100 μl reactions under the following conditions: 10 mM Tris-HCl (pH 8.3), 50 mM KCl, 0.1% Triton 100, 1 mM $MgCl_2$, 0.15 μM of each dNTP, 0.2 μM of each primer, and 1 U *Taq* polymerase. A thermal profile of 3 cycles of 95°C for 1 min, 58°C for 1 min, 72°C for 1.5 min, followed by 30 cycles of 94°C for 15 sec, 59°C for 20 sec, and 65°C for 2 min was used. This was followed by a final elongation of 20 min at 72°C. The reduced elongation temperature was used to try and increase the fidelity of the enzyme while allowing the addition of a terminal A nucleotide for TA cloning provided by *Taq* polymerase.

PCR products were checked for heteroplasmy and polymorphism on a 2.0% agarose gel stained with ethidium bromide. Amplification products were cleaned using the Qiagen PCR purification kit (Qiagen Inc., Valencia, Calif.) after which, the fragments were inserted into the pCR-4 cloning vector using the TOPO-TA cloning kit from Invitrogen (Carlsbad, Calif.). Plasmids were chemically transformed into TOP10 (Invitrogen) competent cells using the protocols provided. At least 10 colonies from each individual were screened for inserts via PCR using the primers M13F and M13R under standard conditions. Colonies were then selected on the basis of size differences. The number of colonies sequenced from each individual was dependent on the number of different size fragments observed, but in each case, at least two colonies were selected. Plasmids from selected colonies were isolated using the Qiagen miniprep kit and sequenced using an ABI 310 genetic analyzer employing the sequencing primers M13F and M13R. Sequences were analyzed and edited using the program BioEdit (Hall 1999). Analysis for secondary structure was performed with RNAstructure3.71 (Mathews et al. 1999) using the default settings. Sequences from each clone examined have been deposited with GenBank (accession numbers AY138864 to AY138894).

Most products from the region under investigation ranged between 550 and 750 bp in length. This sometimes presented a problem for sequencing analysis because the ABI 310 genetic analyzer typically yields readable sequences of only 400 to 500 bp. The presence of the repeated sequences did not allow for the use of internal primers for either PCR or sequencing. We were, therefore, forced to use information on fragment size from gel electrophoresis and sequence overlap to determine the sequence for the entire fragment. Fragments exceeding 800 bp (seen primarily in *C. palustris* and *C. porosus*) had little overlap when sequenced from both directions, and, therefore, base identity in the central regions was less certain.

While size variation in the tandem repeat region among and within individuals was observed in several species of the Crocodylidae (*Crocodylus intermedius*, *C. palustris*, *C. niloticus*, *C. mindorensis*, *C. siamensis*, and *C. porosus*) via direct sequencing, we chose to concentrate our efforts on a more detailed analysis of repeats in Morelet′s crocodile (*C. moreletii*). This was accomplished by modifying the primer 12SH1 to include a HEX label and amplifying the tandem repeat region from 19 individuals of *C. moreletii*. After amplification, PCR products

were analyzed on an ABI 310 genetic analyzer using Genescan software (ABI).

Results

By convention, most mitochondrial DNA sequences are discussed using the light (L) strand as a reference. Unless otherwise stated, directional phrases such as ''5′ from'' and ''upstream from'' will always refer to descriptions of L-strand sequences.

In all members of Crocodylidae a distinct poly-A tract, beginning approximately 30 bp downstream from the presumed CSB-3 sequence (Ray and Densmore 2002) was observed. The poly-A region varied among sequenced clones but ranged from 53 bp (*Crocodylus palustris*) to 83 bp (*C. rhombifer*). In some clones, the poly-A region was followed by a shorter but distinct tract of nucleotides dominated by thymine residues. This pattern was characteristic of New World crocodiles (*C. acutus*, *C. moreletii*, *C. rhombifer*, and *C. intermedius*) and several sequences obtained from *C. niloticus* and *C. mindorensis* (the Nile and Philippine crocodiles, respectively). Tandemly repeated sequences immediately followed the poly-A or poly-T tracts in all clones and made up a large portion (40% to 64%) of the CSB domain.

Tandem Repeat Primary Structure

Within Crocodylidae, the sequences of repeats were similar and appear to be derived from the same ancestral sequence. They show a high A/T bias (> 78%) with runs of up to 8 nt each. In all Old World species examined, the longest single repeat was 63 bp (*Crocodylus siamensis*). In New World species, the longest single repeat was 74 bp (*C. acutus*). Generally, Old World species showed less variation in repeat length as the sequence proceeded from 5′ to 3′. With the exception of *C. siamensis*, these species exhibited a pattern in which the repeats varied by only 1 or 2 nt among fully repeated sequences. The main array (three to 10 repeats) was usually followed by a truncated version of the full sequence (10 to 38 bp).

New World crocodile species tended to be highly variable along the entire length of the repeat region. These species typically showed a pattern of shortening as they proceeded from 5′ to 3′ along the light strand. For example, in one individual of *Crocodylus moreletii*, clone 3557.3 (fig. 2), the initial sequence (copy 1) is 70 bp in length. Copy 1 is imperfectly repeated in the adjacent sequence as a 70-bp sequence with seven nucleotide substitutions and two indels. Copies 3, 4, and 5 are perfectly repeated sequences of 51 bp each. They differ from copy 2 by one substitution and a single large deletion. The final copy of the repeated sequence is truncated to 33 bp and is a perfect

```
5′ - AAAAAAAAAAAAAAAGAAAAAAAAAAAAAAAAAAAGAGAGAGAAAGAAAAACACAAAGGAAAAACAGGAAAAA
     TTTTTTAAATTTTTTTGTTTTTTTTAAACAAA
     -TTATCAACCTAGGCTAAAATAGGAAAAAATTTTTAAAAAAATTTTAAAAAATTTTAAAAAAAATTAAAAA
     TTTATTAACCTAGGCTAAAATAGGAAGAA-TTTTTAAAAAATTTAAAAAAAATTAAAAAAAAAATTAAAAA
     TTTATTAACCTAGGCTAAAATAGGAAAAAATTTTTT--------------------AAAAAAAATTAAAAA
     TTTATTAACCTAGGCTAAAATAGGAAAAAATTTTTT--------------------AAAAAAAATTAAAAA
     TTTATTAACCTAGGCTAAAATAGGAAAAAATTTTTT--------------------AAAAAAAATTAAAAA
     TTTATTAACCTAGGCTAAAATAGGAAAA - 3′
```

FIG. 2.—Sequence of clone 3557-3. Note the presence of the poly-A region followed by a shorter poly-T tract. ''–'' represents a gap introduced to ease visualization of repeated units. The ''TAGG'' motif is identified by bold type.

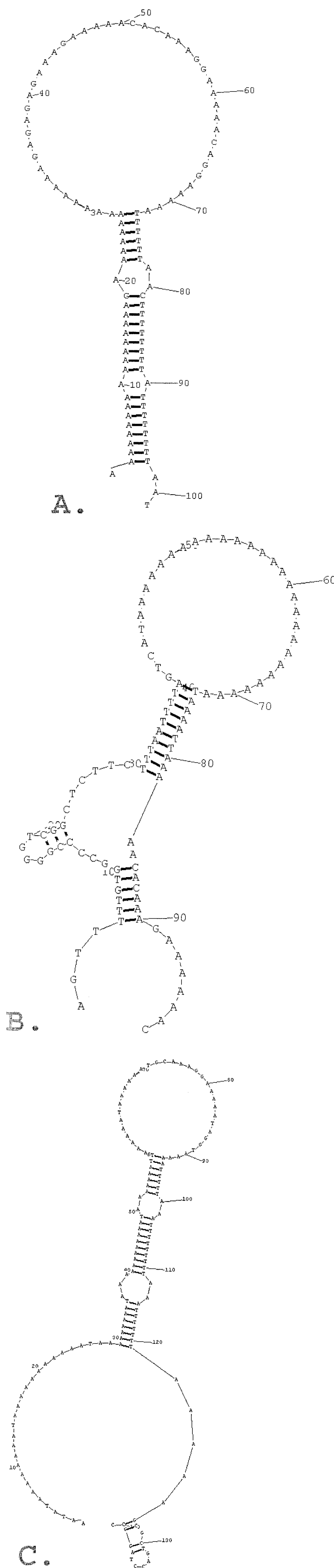
Table 1
Copy Number 1 of Each Tandem Array As Identified in *Crocodylus* spp., *Osteolaemus tetraspis tetraspis*, *Alligator mississippiensis*[a], *Alligator sinensis*, and *Caiman crocodilus*[b]

Species	Sequence
C. acutus (n = 1)	TTATCAA**CCTAGGCTAAAATAGG**AAAAAATTTTTAAAAAAAATTTTAAAAAATTTTAAAAAAAAATTAAAAA [6]
C. cataphractus (n = 2)	**AAAATAGG**TAAAAATTTTTAAATTTTTTTTTAAATTTTTTTAAAAAGCGCTGA**CCTAGGCC** [7, 8]
C. intermedius (n = 4)	AA**CCTAGGCTAAAATAGG**AAAAAATTTTTAAAAAAATTTTAAAAAATTTTAAAAAAAAATTAAAAATTTATT [4–8]
C. mindorensis (n = 1)	TTTTAAACAAATTATTAA**CCTAGGCCAAAATAGG**AAAAAATTTTAAAAAAATTTTAAAAAAA [6, 8]
C. moreletii (n = 3)	TTATCAA**CCTAGGCTAAAATAGG**AAAAATTTTTTAAAAAAATTTTAAAAAAAATTTTAAAAAAAATTTAAAAAT [5, 6]
C. niloticus (n = 2)	AAAAATTATTAA**CCTAGGCTAAAATAGG**AAAAATTTTTTAAAAAAATTTAAAAAAAAAATTT [7, 8]
C. palustris (n = 3)	**AGG**AAAAAATTTTTAAAAAAATTTAAAAAAATTTTAAAAAAAATTATTAA**CCTAGGCTAAAAT** [7, 8, 9, 11]
C. porosus (n = 1)	**AGG**AAAAAATTTTTAAAAAAATTTAAAAAAATTTTAAACAAATTATTAA**CCTAGGCTAAAAT** [8, 9]
C. rhombifer (n = 1)	TTTTTTTTAAAAATTATTAA**CCTAGGCTAAAATAGG**AAAAAATTTTAAAAAATTTTTTTAAAAATTTTAAAAAA [5]
C. siamensis (n = 1)	AAGAAAAAAAATTTTAAAAAAATTTAAAAAAATTTTAAACAAATTATCAA**CCTAGGCCAGAATAGG** [6]
O. t. tetraspis (n = 1)	AAAATTTTTAATTTTTTTTTAAAAAATTATCAA**CCTAGGGCAAAATAGG**AAGAATTTTCTAAA [5]
A. mississippiensis	ATAATTTTATATTAATATAACT [3] ATAATTTTATATTATAGGGCC [15]
A. sinensis	ATTATAGGGCCATAAATTTAT [7] TATTATAGGGCCATAAATTTAC [6] AGCCATAAACTTATATTATAG [4]
Ca. Crocodilus	TTGGCAAAAACCCCTACATTTACTAAAGATTATATTTTTTATATACATATAATTACCCTATAGACATATAGAAA-AAACTATAAATAATATTAAAATTAACACATCCAATTACCTTGTAAAACAGCATTATATAATTAAAA-TTTTATATTTTTATTGTTAAAATCTTACATTTTTATTGTTAAAATCTTACATTTTTATTGTTAAAATTAACAAT-TCATACCCACTACGCCCATTAAAATTTATACTTTTACTAATGTAAAACCAACTAAACCTTAACTTCCCT-TCTA [4]

NOTE.—All sequences proceed 5′ to 3′. Numbers in brackets indicate repeat number(s) found in various individuals. *A. mississippiensis* and *A. sinensis* each have multiple examples of repeated sequences. The nested repeat of *Caiman crocodilus* discussed in the text in is underlined. The number of individuals examined for each species is indicated in parentheses.

[a] Janke and Arnason 1997.

[b] Janke et al. 2001.



repeat of the first 35 bp of copies 3, 4, and 5. This pattern is repeated in the other New World species studied, whereas in the Old World species the progressive shortening was either not observed or was seen only in some clones.

A sequence motif (CCTAGGSYAAAATAGG [table 1]) was identified in all members of *Crocodylus* and in *Osteolaemus* and will hereafter be referred to as the ''TAGG'' motif. It is characterized by an inverted repeat (underlined) that forms a stable secondary structure (see below). *Osteolaemus tetraspis tetraspis* mtDNA also contained this motif with one consistent substitution (C→G) in the seventh position.

Secondary Structure

In all individuals examined, at least some portion of the poly-A tracts formed a stem-loop structure based on one of three scenarios. First, the poly-A tract may form a stable hairpin with the poly-T tract immediately adjacent (an almost perfect inverted repeat of the poly-A tract) and in the 3′ direction. This is seen in sequence from *C. moreletii* in figure 3*A*. In the second scenario, illustrated using sequence from *C. palustris* (fig. 3*B*), the poly-A region forms a stem and loop with a series of T nucleotides found 5′ of the adenosine tract. Finally, the poly-A tract may form a secondary structure with thymine residues within the first tandem repeat, as in *C. cataphractus* (fig. 3*C*).

All of the tandemly repeated sequences themselves formed stable stem-loop structures (fig. 4). If one rearranges a single repeat from the various species so that the TAGG motif is at the extreme 3′ end, all of the respective secondary structures are found to be similar. In each case, at least two stem-loop structures are formed, the first is made up of inverted repeats of poly-A and poly-T tracts. A middle stem-loop, also made up of A/T nucleotides was sometimes present in *C. rhombifer*, in the longest version of the *C. acutus* repeat, and in *C. cataphractus*. The final structure was made up of the TAGG motif and was identical in all taxa except *Osteolaemus* (differences are described above) and *C. cataphractus* and *C. siamensis* (which contain a final A-T pair in the stem and a C-T substitution in the loop).

Repeat Units in Other Crocodilians

Complete control region sequences for three members of Alligatoridae were also available from GenBank, *Alligator mississippiensis* (Y13113 [Janke and Arnason 1997]), *Alligator sinensis* (AF511507 [Yiquan Wang, personal communmication]), and *Caiman crocodilus* (AJ404872 [Janke et al. 2001]). In this study, we made no attempt to characterize possible heteroplasmic variation within these species but will discuss characteristics of the analogous control region sequences. The aforementioned poly-A tracts are conspicuously missing in all three

←

FIG. 3.—The most stable secondary structures (as predicted using RNAStructure 3.71) from the three poly-A tract patterns in species of the Crocodylidae as discussed in the text. Free energy calculations are −12.1 kcal/mol, −6.8 kcal/mol, and −12.5 kcal/mol, for (*A*), (*B*), and (*C*), respectively.

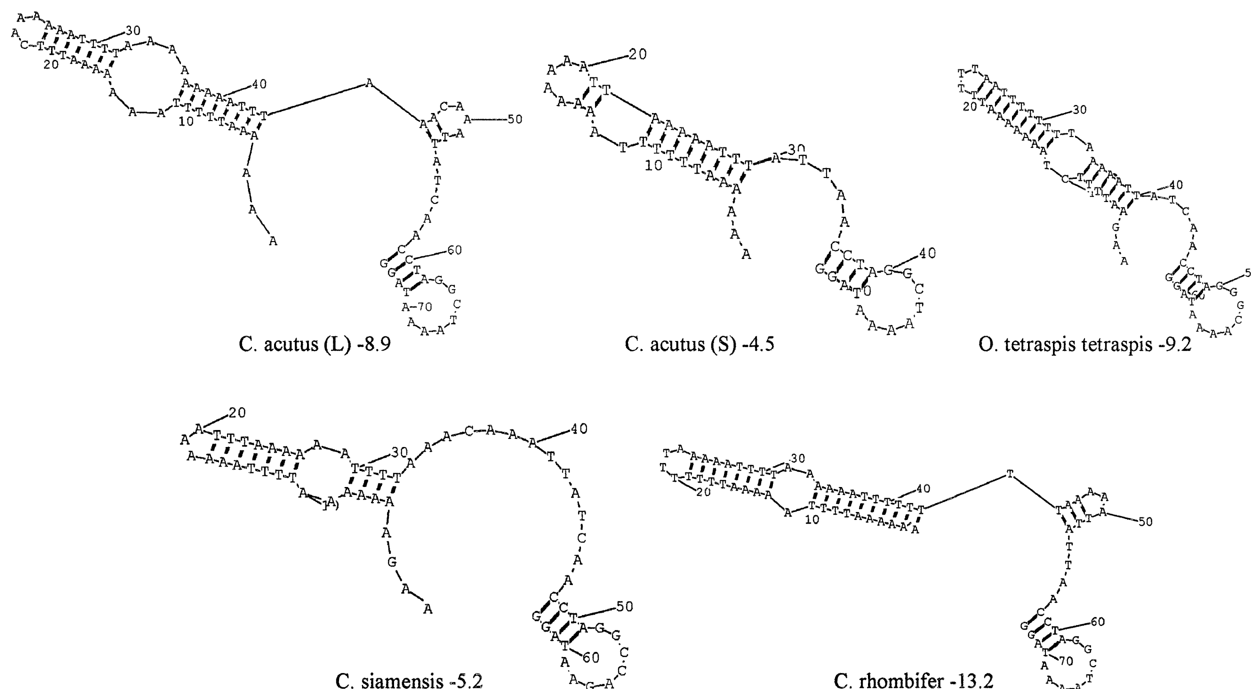


FIG. 4.—Representative secondary structures from the tandem repeats found in the mitochondrial control region of Crocodylidae. Estimated free energy values (kcal/mol) of each structure are provided. (L) and (S) refer to complete and truncated versions of a repeated sequence, respectively.

alligatorid species examined, but these taxa did exhibit tandemly repeated sequences at approximately the same positions as members of Crocodylidae. *Caiman crocodilus* had the longest single repeat, 287 bp. Repeat units in the two *Alligator* species were the shortest observed in any crocodilian and ranged from 9 bp to 22 bp. Within the array of *Caiman crocodilus*, there is a smaller sequence that is itself repeated 2.7 times in three of the four larger repeats and 1.7 times in the fourth (table 1). The tandem array in *Caiman* was mentioned in Janke et al. (2001), but no detailed descriptions of the region were provided. Eighty-seven percent of the *Ca. crocodilus* CSB domain sequence consisted of tandem repeats. *Alligator mississippiensis* and *A. sinensis* CSB domains were composed of 62% and 55% repeated sequences, respectively. Examples of the tandem repeats from each species examined are presented in table 1. We were unable to amplify the heteroplasmic region apparent in Crocodylidae and Alligatoridae in the either of the two gharial species (see Ray and Densmore 2002).

Table 2
Frequencies of Different Size Classes of Haplotypes Identified in 29 Individuals (32 Haplotypes) of *Crocodylus moreletii*

Haplotype Size Classes (bp)	Haplotype Frequencies
536	0.0313
591	0.2188
592	0.2813
593	0.0625
643	0.0313
644	0.0313
645	0.2813
646	0.0313
678	0.0313

Intraindividual and Interindividual Variation

Nine size classes ranging in size from 536 bp to 678 bp were observed among the 29 *Crocodylus moreletii* sampled (table 2). Three heteroplasmic individuals were identified, cm2664, cm2465, and cm3433. All three contained two haplotypes, 592/645 bp in cm2664, 593/645 bp in cm2465, and 536/678 bp in cm3433.

Discussion

Repeat Generation and Evolution

In the cases described here, variations on the model of Mundy, Winchell, and Woodruff (1996) seem appropriate for identifying the location of the original sequence duplication. This model concentrates on tandem repeats at the 3′ end of the control region, downstream of the conserved sequence blocks. However, it relies heavily on secondary structures downstream (3′) of the repeated region to generate the initial copies. Although there are a few small stem-loop structures downstream of the final repeat copies in the control region of Crocodylidae, none seem large enough to cause the kind of slippage required by their model. The poly-A and poly-T tracts immediately 5′ of the tandemly repeated sequences, however, are good candidates as a source for the original duplication.

As noted in figure 3, the poly-A portions located 5′ of the tandem repeats form stable secondary structures in all species considered. This being the case, the model of Mundy, Winchell, and Woodruff (1996) is applicable if one assumes that the nascent light strand (not the heavy

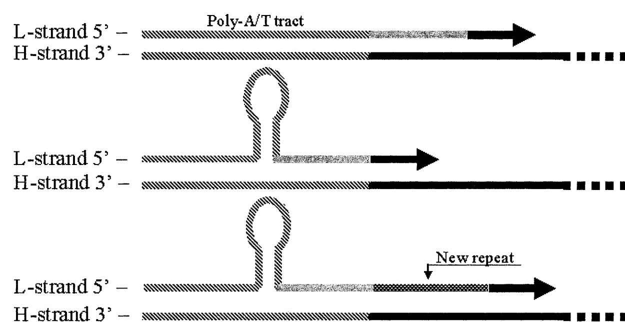


FIG. 5.—A variation of the Mundy, Winchell, and Woodruff (1996) model that may explain the initial formation of the tandem repeats in Crocodylidae. The nascent L-strand is subjected to secondary structure formation shortly after replication of the poly-A/T tract, causing the neighboring sequence to slip and be replicated twice.

strand as described in their model) is allowed to ''slip'' immediately after this region is copied (fig. 5). This need only happen once, after which the standard model of slipped-strand mispairing can explain expansions and contractions of the tandem array (see below). Assuming this is the correct model, the original duplication would be 3′ of and adjacent to the poly-A tract.

Crocodylus and *Osteolaemus* repeat sequences appear to increase in number in a 5′ to 3′ direction. We suspect this because the ''original'' sequence is imperfectly repeated and there is a tendency in some species (e.g., the New World crocodiles) for the repeats to become more imperfect as they proceed from the 5′ to 3′ direction. This would be expected if the tandem repeats were generated by strand slippage (Levinson and Gutman 1987) of neighboring repeats—as one moves further away from the ''original'' repeat the sequences should become more divergent. The tendency to shorten at the 3′ end was also found when in vitro experiments were performed with rabbit mitochondrial repeats cloned in *E. coli* (Pfeuty, Gueride, and Lecellier 2001). Why this is the case primarily in the New World crocodiles is unknown, although it may be related to the relatively recent radiation of this group of taxa. All recent phylogenetic analyses place the true crocodiles as the most recent branch on the crocodilian tree and the New World crocodiles as the most derived of the *Crocodylus* species (Densmore 1983; Densmore and Owen 1989; Densmore and White 1991; Brochu 2000). In older species, the sequence of the repeat units may have stabilized over the increased evolutionary time.

One of the most commonly invoked models to explain the expansion and contraction of repeated sequences in the control region is that of Buroker et al. (1990). However, their model is not useful for explaining changes in copy number for the repeats described here. Although it has yet to be precisely located, the origin of replication in crocodilian mtDNA is probably not immediately adjacent to the 12S rRNA gene (Ray and Densmore 2002). Thus, the generation of repeats is most likely not due to D-loop/H-strand competition. Instead, slipped-strand mispairing (Levinson and Gutman 1987) could occur during heavy-strand and/or light-strand replication. At these times, one strand of the nascent double helix can presumably slip easily and fold into the proposed secondary structure (fig.

```
5' - TTTTTCTTTTCTCCAACCACAttatattatagggcc
     (ATAATTTTATATTAATATAACT)3
     (ATAATTttatattatagggcc)11
      ATA-TTTTATATTATAGGGCC
     (ATAATTTTATATTATAGGGCC)2
      ATAATTTTATA - 3'
```

FIG. 6.—Sequence (5′ to 3′) of the tandem repeat region in *Alligator mississippiensis* (Janke and Arnason 1997). Chimeric aspects of the repeat units are indicated using lowercase letters and shading. Details are discussed in the text.

3). Depending on which strand slips and is mispaired, arrays can be increased or decreased in size.

The *Alligator mississippiensis* array is interesting in that there are two different sets of tandem arrays adjacent to one another (fig. 6). The first array consists of three identical sequences. The second array is made up of 14 repeats that are identical with the exception of a single-base deletion in repeat number 12. Even more interesting are the chimeric aspects of the region. The first three repeats share the initial 14 bp of sequence (shaded in fig. 6) with the second set of repeats. Furthermore, each repeat in the second array contains a direct copy of the 15-bp nucleotide sequence (lowercase in fig. 6) immediately preceding the first array. These features lend clues as to how such an interesting pattern could have arisen via slipped-strand mispairing, although involving a more complicated mechanism than described above. Slippage during replication could explain the pattern if pairing were to occur between the sequence before the first repeat and the nascent chain of the last sequence of the first tandem array via the common sequence, ATATTATA. *Alligator sinensis* tandem repeats follow a similar heterogeneous pattern (GenBank sequence AF511507).

Caiman crocodilus contains three full-length copies and one truncated version of a 287-bp sequence. The short intervening sequence found between the end of the last tandem repeat and the 12S rRNA gene and the size of the repeated units themselves (> 260 bp) suggest that the model as originally proposed by Mundy, Winchell, and Woodruff (1996) is most appropriate in this case to explain the origin of the repeated sequences. However, the heavy-strand model of Broughton and Dowling (1997) cannot be refuted using the data presented here. Repeat number 4 of the array shares 42% of its sequence with 276 bp at the 5′ end of the 12S rRNA gene in *Caiman*, reinforcing the idea that the 12S gene served as the origin of the repeats.

Intraindividual and Interindividual Variation and Heteroplasmy

Two types of variation among and within individuals were detected. The first type is variation due to single or multiple nucleotide insertion/deletion events in the poly-A tract. The second type is insertion/deletion of entire tandem repeats. For example, in *Crocodylus moreletii* there were nine size classes found. The nine distinctly sized fragments (table 2) can be grouped into classes of 536, 59×, 64×, and 678. The 59× and 64× size classes vary by 2 and 3 bp, respectively. The smaller size differences were due to variation in the poly-A tract, and larger differences resulted

```
2664.2 TGATAAAAAA AAAAAAAAAA AA-G---AAA AAAAAAAAAA AAGAGAGAGA AAGAAAAACA CAAAGGAAAA ACAGGAAAAA TTTTTTAACT
2664.7 .......... .......... ..A.AAA... .......... .......... .......... .......... .......... ..........

2664.2 TTTTTTTATT TTTTTTAATA AATTATCAAC CTAGGCTAAA ATAGGAAAAA TTTTTTAAAA AATTTTAAAA AAATTTTAAA AAAATTTAAA
2664.7 .......... .......... .......... .......... .......... .......... .......... .......... ..........

2664.2 AATTTATTAA CCTAGGCTAA AATAGGAAAA ATTTTTAAAA ATTTTAAAAA ATTTTTTAAA AAAATTTAAA AATTTATTAA CCTAGGCTAA
2664.7 .......... .......... .......... .......... .......... .......... .......... .......... ..........

2664.2 AATAGGAAAA AATTTTTTAA AAAAATTTAA AAATTTATTA ACCTAGGCTA AAATAGGAAA AGATTTTTTA AAAAAATTTA AAAATTTATT
2664.7 .......... .......... .......... ....------ ---------- ---------- ---------- ---------- -----.....

2664.2 AACCTAGGCT AAAATAGGAA AAAATTTTTT AAAAATTTAA AAATTTATTA ACCTAGGCTA AAATAGGAAA ACCTTTAACA CATCAAACTA
2664.7 .......... .......... .......... .......... .......... .......... .......... .......... ..........

2664.2 GGACTGCCAC AGCACCCATA CAGTTATT
2664.7 .......... .......... ........
```

FIG. 7.—Aligned sequences (5′ to 3′) from the tandem repeat regions of two clones from *Crocodylus moreletii* individual 2664. "." represents identical residues and "–" indicates an insertion/deletion event.

from variation in the number of tandem repeats. For example, the repeat region from one individual (cm2664) exhibited two haplotypes derived from both types of variation (fig. 7). First, in the poly-A tract of clone 2664.7 (592 bp), there are two deletions of 1 and 3 A nt, respectively, at the 5′ end of the sequence when compared with the second sequenced clone 2664.2 (645 bp). The second difference is the complete deletion of one tandem repeat (49 bp).

The kinds of variation we observed within and among individuals may be a useful tool for researchers investigating population dynamics in crocodilians, but care must be taken before initiating such studies (Lunt, Whipple, and Hyman 1998). Differences in the numbers and frequencies of repeated sequences in the different tissues of animals have been observed (Casane et al. 1994; Jenuth, Peterson, and Houbridge 1997). All of the DNA isolated for this study was obtained from whole blood and should not present a problem, but researchers attempting to utilize this region for population studies should take care to obtain DNA from identical tissues in all animals or perform tissue comparisons before stating any conclusions.

Finally, it will be interesting to assess the structure of any possible repeated sequences in the two gharials, *Gavialis gangeticus*, and *Tomistoma schlegelli*. These two crocodilians are at the center of a continuing conflict between morphology-based and molecule-based hypotheses about the relationships within the Crocodylia (Brochu and Densmore 2001). Most studies employing morphology of both extant and fossil crocodilians find that the true gharial (*Gavialis*) is an ancient taxon whose ancestors diverged from the rest of the group in the late Cretaceous or early Tertiary. In these studies, the false gharial (*Tomistoma*) is usually found to have an affinity with the family Crocodylidae, a much more recent group. By contrast, all biochemically/molecularly-based analyses to date have found the two gharials to be closely related, usually forming a sister clade to the Crocodylidae (see Brochu and Densmore 2001 for a review). Eventual characterization of any repeats present in the control region of these species may provide evidence for both divergence timing and relative placement of these taxa with regard to the other crocodylian lineages.

Supplementary Materials

All sequences generated for this project have been deposited with GenBank (http://ncbi.nlm.nih.gov) under accession numbers AY138864 to AY138894.

Acknowledgments

We would like to thank the following individuals and institutions for contributing valuable resources and comments to the completion of this work. Both zoos and nonprofit organizations kindly allowed us to collect blood. Several of the Morelet′s crocodile samples were collected with logistical support of Mark and Monique Howells at the Lamanai Field Research Station in Belize. Robert Bradley, Robert Baker, and the Core Facility at Texas Tech University provided material assistance in the generation of sequences for the project. Travis Glenn, Darin Carrol, Federico Hoffman, and Jeff Wickliffe made valuable suggestions about procedure and analyses. Abdel Halim-Salem and Dale Hedges contributed valuable comments to earlier versions of the manuscript. Financial support for this project was provided by the both the Graduate School and Department of Biological Sciences at Texas Tech University (specifically, Carleton Phillips and John Zak). Additional partial support was provided to L.D.D. by the National Geographic Society (#7007–01).

Literature Cited

Anderson, S., A. T. Bankier, B. G. Barrell et al. (14 co-authors). 1981. Sequence and organization of the human mitochondrial genome. Nature **290**:457–465.

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

Bayliss, P. 1987. Survey methods and monitoring within crocodile managementprogrammes. Pp.157–175 *in* G. J. Webb, S. C. Manolis, and P. J. Whitehead, eds. Wildlife management: crocodiles and alligators. Surrey Beatty and Sons, Sydney, Australia.

Bianchi, N. O., M. S. Bianchi, and S. M. Richard. 2001. Mitochondrial genome instability in human cancers. Mutat. Res. **1**:9–23.

Brochu, C. 2000. Phylogenetic relationships and divergence timing of *Crocodylus* based on morphology and the fossil record. Copeia **2000**:657–673.

Brochu, C., and L. Densmore. 2001. Crocodile phylogenetics: a review of current progress. Pp. 3–8 *in* G. Grigg, F. Seebacher, C. Franklin, eds. Crocodilian biology and evolution. Surrey Beatty and Sons, Chipping Norton, Australia.

Broughton R. E., and T. E. Dowling. 1994. Length variation in mitochondrial DNA of the minnow *Cyprinella spiloptera*. **138**:179–190.

Brown, J. R., A. T. Beckenbach, and M. J. Smith. 1992. Mitochondrial length variation and heteroplasmy in populations of white sturgeon (*Acipenser transmontanus*). Genetics **13**: 221–228.

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

Buroker, N. E., J. R. Brown, T. A. Gilbert, P. J. O'Hara, A. T. Beckenbach, W. K. Thomas, and M. J. Smith. 1990. Length heteroplasmy of sturgeon mitochondrial DNA: an illegitimate elongation model. Genetics **124**:157–163.

Casane, D., N. Dennebouy, H. de Rochambeau, J. C. Mounolou, and M. Monnerot. 1994. Genetic analysis of systematic mitochondrial heteroplasmy in rabbits. Genetics **138**: 471–480.

Cesaroni, D., F. Venanzetti, G. Allegrucci, and V. Sbordoni. 1997. Mitochondrial DNA length variation and heteroplasmy in natural populations of the European sea bass, *Dicentrarchus labrax*. Mol. Biol. Evol. **14**:560–568.

Densmore, L. D. 1983. Biochemical and immunological systematics of the order Crocodilia. Pp. 397–465 *in* M. K. Hecht, B. Wallace, and G. T. Prance, eds. Evolutionary biology, Vol. 16. Plenum Press, New York.

Densmore, L. D., and R. D. Owen. 1989. Molecular systematics of the order Crocodilia. Am. Zool. **29**:831–841.

Densmore, L. D., and P. S. White. 1991. The systematics and evolution of the Crocodilia as suggested by restriction endonuclease analysis of mitochondrial and nuclear ribosomal DNA. Copeia **1991**:602–615.

Doda, J. N., C. T. Wright, and D. A. Clayton. 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.

Erye-Walker, A. and P. Awadalla. 2001. Does human mtDNA recombine? J. Mol. Evol. **53**:430–435.

Foran, D. R., J. E. Hixson, and W. M. Brown. 1988. Comparisons of ape and human sequences that regulate mitochondrial DNA transcription and D-loop synthesis. Nucleic Acids Res. **16**:5841–5861.

Gorzula, S., C. L. Arocha-Pinango, and C. Salazar. 1976. A method of obtaining blood by vein puncture from large reptiles. Copeia **1976**:838–839.

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

Hoelzel, A. R., J. M. Hancock, and G. A. Dover. 1993. Generation of VNTRs and heteroplasmy by sequence turnover in the mitochondrial control region of two elephant seal species. J. Mol. Evol. **37**:190–197.

Janke A., and U. Arnason. 1997. The complete mitochondrial genome of *Alligator mississippiensis* and the separation between recent Archosauria (birds and crocodiles). Mol. Biol. Evol. **14**:1266–1272.

Janke, A., D. Frpenbeck, M. Nilsson, and U. Arnason. 2001. The mitochondrial genome of the iguana (*Iguana iguana*) and the caiman (*Caiman crocodylus*): implications for amniote phylogeny. Proc. R. Soc. Lond. B Biol. Sci. **268**: 623–631.

Jenuth, J. P., A. C. S. Peterson, and E. A. Houbridge. 1997. Tissue-specific selection for different mtDNA genotypes in heteroplasmic mice. Nat. Genet. **16**:93–95.

King, T. C., and R. L. Low. 1987. Mapping of control elements in the displacement loop region of bovine mitochondrial DNA. J. Biol. Chem. **262**:6204–6213.

Ladoukakis, E. D., and E. Zouros. 2001. Direct evidence for homologous recombination in mussel (*Mytilus galloprovincialis*) mitochondrial DNA. Mol. Biol. Evol. **18**:1168–1175.

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

Lunt, D. H., and B. C. Hyman. 1997. Animal mitochondrial DNA recombination. Nature **387**:247.

Lunt, D. H., L. E. Whipple, and B. C. Hyman. 1998. Mitochondrial DNA variable number tandem repeats (VNTRs): utility and problems in molecular ecology. Mol. Ecol. **7**:1441–1455.

Mathews, D. H., J. Sabina, M. Zuker, and D. H. Turner. 1999. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J. Mol. Biol. **288**:911–940.

Mundy, N. I., C. S. Winchell, and D. S. Woodruff. 1996. Tandem repeats and heteroplasmy in the mitochondrial DNA control region of the loggerhead shrike (*Lanius ludovicianus*). J. Heredity. **87**:21–26.

Pfeuty, A., M. Gueride, and G. Lecellier. 2001. Expansion/contraction of mammalian mitochondrial DNA repeats in *Escherichia coli* mimics the mitochondrial heteroplasmy. J. Mol. Biol. **341**:709–716.

Rand, D. M., and R. G. Harrison. 1989. Molecular population genetics of mtDNA size variation in crickets. Genetics **121**: 551–569.

Ray, D. A., and L. D. Densmore. 2002. The crocodilian mitochondrial control region: general structure, conserved sequences and evolutionary implications. J. Exp. Zool. (Mol. Dev. Evol.) **294**:334–345.

Ritchie P. A., and D. M. Lambert. 2000. A repeat complex in the mitochondrial control region of Adelie penguins from Antarctica. Genome **43**:613–618.

Sanchez-Cespedes, M., P. Parrella, S. Nomoto et al. (11 coauthors). 2001. Identification of a mononucleotide repeat as a major target for mitochondrial DNA alterations in human tumors. Cancer Res. **61**:7015–7019.

Sbisa, E., F. Tanzariello, A. Reyes, G, Pesole, and C Saccone. 1997. Mammalian mitochondrial D-loop region structural analysis: identification of new conserved sequences and their functional and evolutionary implications. Gene **205**: 125–140.

Tamura, G., S. Nishizuka, C. Maesawa, Y. Suzuki, T. Iwaya, K. Sakata, Y. Endoh, and T. Motoyama. 1999. Mutations in mitochondrial control region DNA in gastric tumours of Japanese patients. Eur. J. Cancer **35**:316–319.

Walberg, M. W., and D. A. Clayton. 1981. Sequence and properties of the human KB cell and mouse L cell D-loop regions of mitochondrial DNA. Nucleic Acids Res. **9**:5411–5421.

White, P. S., O. L. Tatum, H. Tegelstrom, and L. D. Densmore. 1998. Mitochondrial DNA isolation separation and detection of fragments. Pp 65–102 *in* A. R. Hoelzel, ed. Molecular genetic analysis of populations: a practical approach. Oxford University Press, New York.

Axel Meyer, Associate Editor

Accepted March 4, 2003