J Mol Evol (1995) 40:382–391

JOURNAL OF MOLECULAR EVOLUTION

Evolution and Structural Conservation of the Control Region of Insect Mitochondrial DNA

De-Xing Zhang, Jacek M. Szymura,* Godfrey M. Hewitt

School of Biological Sciences, University of East Anglia, Norwich, NR4 7TJ United Kingdom

Received: 14 June 1993 / Accepted: 15 September 1993

Abstract. The control regions of mitochondrial DNA of two insects, *Schistocerca gregaria* and *Chorthippus parallelus,* have been isolated and sequenced. Their sizes are 752 bp and 1,512 bp, respectively, with the presence of a tandem repeat in *C. parallelus.* (The sequences of the two repeats are highly conserved, having a homology of 97.5%.) Comparison of their nucleotide sequences revealed the presence of several conserved sequence blocks dispersed through the whole control region, showing a different evolutionary pattern of this region in these insects as compared to that in *Drosophila.* A highly conserved secondary structure, located in the 3′ region near the small rRNA gene, has been identified. Sequences immediately flanking this hairpin structure rather than the sequences of this structure themselves are conserved between *S. gregaria/C. parallelus* and *Drosophila,* having a sequence consensus of "TATA" at 5′ and "GAA(A)T" at 3′. The motif "G(A)nT" is also present in the 3′ flanking sequences of mammalian, amphibian, and fish mitochondrial L-strand replication origins and a potential plant mitochondrial second-strand-replication origin, indicating its universal conservation and functional importance related to replication origins. The stem-and-loop structure in *S. gregaria/C. parallelus* appears to be closely related to that found in *Drosophila* despite occupying a different position, and may be potentially associated with a second-strand-replication origin. This in turn suggests that such a secondary structure might be widely conserved across invertebrates while their location in the control region may be variable. We have looked for such a conserved structure in the control regions of two other insects, *G. firmus* and *A. mellifera,* whose DNA sequences have been published, and their possible presence is discussed.

Mitochondrial control regions characterized to date in five different insect taxa (*Drosophila, G. firmus, A. mellifera, S. gregaria,* and *C. parallelus*) may be classed into two distinct groups having different evolutionary patterns. It is observed that tandem repetition of regions containing a probable replication origin occurred in some species from disjunct lineages in both groups, which would be the result of convergent evolution. We also discuss the possibility of a mechanism of "parahomologous recombination by unequal crossing-over" in mitochondria, which can explain the generation of such tandemly repeated sequences (especially the first critical repetition) in the control region of mtDNA, and also their convergent evolution in disjunct biological lineages during evolution.

Key words: *Schistocerca gregaria* — *Chorthippus parallelus* — Mitochondrial DNA — A + T-rich regions — Tandem repeats — Secondary structure — Replication origin — Convergent evolution — Parahomologous recombination — Unequal crossing-over

Introduction

Animal mitochondrial DNA (mtDNA) is a favorite molecule for evolutionary studies due to several special char-

* *Present address:* Department of Comparative Anatomy, Jagiellonian University, Ingardena 6, 30-060 Krakow, Poland
Correspondence to: G.M. Hewitt

acteristics. For example, in metazoan organisms the mitochondrial genome size is relatively small, ranging from 14.3 kb to 39.3 kb (Gray 1989); many species have mtDNAs of 15.7–19.5 kb (Brown 1985) and thus are particularly tractable; the complete sequence of the whole genome is known in several organisms both vertebrate and invertebrate; organization and functions of genes in the genome have been well characterized; most importantly, the mitochondrial genome seems to lack any recombination (Brown 1985), and is much more variable than nuclear DNA. The control region (also called "the D-loop region" in vertebrates and "the A + T–rich region" in some invertebrates) of mtDNA contains the origin of replication and has been shown to be the most variable region in both vertebrates and invertebrates. (For review, see Simon 1991.) In insects, this region is the only major noncoding region in the mitochondrial genome, being rich in adenine and thymine in most of the insects so far studied. It is responsible for a large part of the variation in mitochondrial genome in both DNA sequence and size. Three types of size variations have been observed in this region: (1) insertions/deletions of a few nucleotides (e.g., Monnerot et al. 1990); (2) variation in copy number of tandemly repeated sequences (e.g., Solignac et al. 1986; Rand and Harrison 1989; Monforte et al. 1993); and (3) extensive length variation of a variable domain (e.g., Solignac et al. 1986). All these variations also occur in vertebrates in the corresponding D-loop region (e.g., Brown and DesRosiers 1983; Mignotte et al. 1990; Saccone et al. 1991), and tandem repeat size variation sometimes leads to heteroplasmy of mitochondrial molecules. (For review see Moritz et al. 1987.)

In vertebrates, the regions containing the mitochondrial replication origins have been well studied. Comparison of DNA sequences from different species revealed several conserved sequence blocks and secondary structures associated with the replication origins (Walberg and Clayton 1981; Brown et al. 1986; Saccone et al. 1987, 1991). This also showed that the secondary structure associated with the L-strand (light strand) replication origin is widely conserved (Anderson et al. 1982; Clayton 1982; Wong et al. 1983; Johansen et al. 1990). However, very little sequence information is available in invertebrates. Up to now, the complete sequence of the mitochondrial control region has been reported in only three distinct insect species groups—the fruit fly *Drosophila* (Clary and Wolstenholme 1985, 1987; Monnerot et al. 1990), the cricket *Gryllus firmus* (Rand and Harrison 1989), and the honeybee *Apis mellifera* (Crozier and Crozier 1993). Studies carried out in different *Drosophila* species showed that the mitochondrial A + T–rich region can be divided into two distinct domains—one conserved domain located in the 5′ region which is very similar across different *Drosophila* species such as *D. yakuba, D. virilis, D. teissieri, D. obscura,* and *D. ambigae;* and one variable domain including the rest of the sequence which is highly variable both in sequence and in length (Monnerot et al. 1990; Monforte et al. 1993). It has been observed that the conserved domain is tandemly repeated in the A + T–rich of *D. tristis* (Monforte et al. 1993), and very probably the same has happened in *D. melanogaster, D. mauritiana, D. simulans,* and *D. sechellia* (Solignac et al. 1986). A conserved secondary structure has been identified in the conserved domain and is inferred to be implicated in the origin of replication (Clary and Wolstenholme 1987; Monforte et al. 1993). As *Drosophila, G. firmus,* and *A. mellifera* are phylogenetically distant, with their nucleotide sequences deeply diverged, no conserved sequence blocks could be identified in their mitochondrial control region. So whether such a conserved secondary structure as found in *Drosophila* exists in other invertebrate organisms is unknown.

In this paper, we report the sequences and characterization of the mitochondrial A + T–rich region in two distinct orthopteran insect species—the desert locust *Schistocerca gregaria* and the meadow grasshopper *Chorthippus parallelus.* Comparison of these sequences allows us to identify several conserved sequence blocks and a highly conserved secondary structure. We show also that although the evolutionary pattern of the A + T–rich region in these two insects is different from that in *Drosophila* species, the conserved secondary structure may be equivalent and implicated in replication origin, indicating its possible conservation across invertebrates. This is the first report of the presence of such a conserved secondary structure in the mitochondrial control region in insects other than *Drosophila* species. In addition, we look for such a conserved structure in the published sequences of *G. firmus* and *A. mellifera,* and discuss their possible presence.

Using these and other published data, we discuss possible evolutionary processes in the mitochondrial control region in insects and a mechanism called "parahomologous recombination by unequal crossing-over" to explain the generation of tandemly repeated sequences in the mitochondrial control region during evolution.

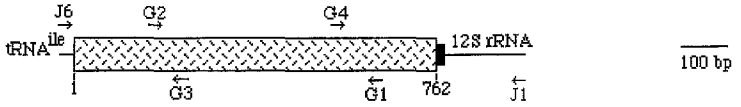
Materials and Methods

Insect Samples. The desert locust *Schistocerca gregaria gregaria* and the meadow grasshopper *Chorthippus parallelus parallelus* were frozen in liquid nitrogen after collection and stored at −80°C until DNA was extracted.

DNA Isolation. Frozen leg from single individuals of *S. gregaria* was used to isolate total DNA for PCR. This was ground to powder in liquid nitrogen and submerged in extraction buffer containing 150 mM NaCl, 100 mM Tris · Cl (pH 8.0), 50 mM EDTA, and 1% SDS. This mixture was extracted with phenol/chloroform and DNA was precipitated with ethanol as described by Sambrook et al. (1989).

Two methods have been used to isolate mtDNA from *C. parallelus.* (1) CsCl gradient method described by Solignac (1991). A population sample of grasshoppers was used to prepare pooled mtDNA of high quality. (2) Minipreparation method. MtDNA from a single individual was extracted as follows. One grasshopper was ground in cold buffer

A. ***S. gregaria***



B. ***C. parallelus***

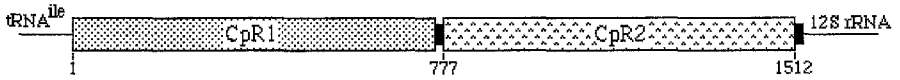


Fig. 1. Schematic map showing the mitochondrial A + T–rich regions of *S. gregaria* **(A)** and *C. parallelus* **(B).** The *large shadowed boxes* represent the control region of *S. gregaria* and individual repeat units in the control region of *C. parallelus.* The *small black boxes* at the end of the large shadowed boxes indicate the 12-bp sequence of the 12S rRNA gene, which is repeated once in *C. parallelus* flanking the second tandem repeat. *J6, G2, G4, J1, G1,* and *G3* in **(A)** represent primers used to amplify and sequence the control region of *S. gregaria; arrows* indicate their directions (5′–3′). *CpR1* and *CpR2* in **(B)** represent the first and second unit of the tandem repeats in *C. parallelus,* respectively. *tRNA^{ile},* tRNA^{ile} gene; *12S rRNA,* the small rRNA gene. *Numbers* below the shadowed boxes indicate their terminal nucleotide position.

containing 250 mM sucrose, 30 mM Tris · Cl (pH 7.5), and 10 mM EDTA, followed by a spin of 1,000*g*–2,500*g* to remove the nuclei and cellular debris. Supernatant obtained was centrifuged at 10,500*g* to pellet mitochondria. The pellet was then resuspended in a buffer (50 mM Tris · Cl, 50 mM EDTA, pH 8.5). To this mixture one-tenth volume of 10% SDS was added to lyse mitochondria. Proteins were removed by adding potassium acetate (one-tenth volume of 5 M) and centrifugation. After treatment with RNase, proteinase K, and phenol/chloroform extractions, mtDNA was precipitated with ethanol.

Cloning of C. parallelus *mtDNA.* MtDNA prepared using the mini-preparation method was digested with *Hin*dIII and cloned into the plasmid pUC18 (Boehringer Mannheim) by the shotgun method. MtDNA prepared by CsCl gradient method was radioactively labeled and used as probe to screen clones containing mtDNA fragments. Positive recombinants were further checked by hybridization with *Locusta migratoria* (McCracken et al. 1987) and *Caledia captiva* (Marchant 1988) cloned mtDNA fragments. One of the positive clones has a 6.3-kb *Hin*dIII insert, which contains the mitochondrial COII, COI, ND2, small rRNA, several tRNA genes, and the control region. A 3.6-kb *Xba*I fragment containing the control region and flanking genes from this clone has been further subcloned into pUC18 and sequenced.

Amplification of the Control Region of S. gregaria. A PCR amplification method (Erlich 1989; Innis et al. 1990) was used to isolate the mitochondrial control region of *S. gregaria.* Primers for PCR in conserved regions have been designed by comparing the sequences of *C. parallelus* and *Drosophila* (Clary and Wolstenholme 1985, 1987). Primer J6 (GGTAATCCTTTAATCAGGCACTCC) lies in the 3′ end of tRNA^{ile} gene; Primer J1 (CGTATAACCGCGGCTGCTGGC), antisense to J6, is located in the 5′ region of the small rRNA gene (Fig. 1). The mitochondrial control region, together with parts of the flanking tRNA^{ile} gene and small rRNA gene of *S. gregaria,* was amplified using the above primer pair. A single fragment of about 1 kb was obtained. Double-stranded PCR was carried out in a 50-μl reaction containing 1.5 mM $MgCl_2$, 50–200 μM dNTP, primers at 0.06 μM each, ~1 ng DNA, and 2 units of Taq polymerase (Promega) in 1 × reaction buffer (50 mM KCl, 10 mM Tris-HCl, 0.1% Triton X-100, pH 9.0 at 25°C, Promega); 25 cycles were performed in a DNA thermal cycler 480 (Perkin Elmer Cetus), each consisting of melting at 95°C for 40 s, annealing at 58°C for 1 min, and extension at 72°C for 1 min. Single-stranded PCR was performed under the same conditions described above, except that 2 μl of the double-stranded PCR product was used as template, and only one primer was added at 1.5 μM in a 100-μl reaction (annealing temperature may be higher than that for double stranded PCR). Excess nucleotides and primers in PCR products were removed using the concentrator Centricon-30 (Amicon).

DNA Sequencing. The 3.6-kb *Xba*I mitochondrial insert of *C. parallelus* in the recombinant pUC18 clone has been progressively deleted by Exonuclease III using the Nested Deletion Kit from Pharmacia and then sequenced using the AutoRead Sequencing Kit (Pharmacia) on Pharmacia A.L.F. DNA Sequencer.

The sequence of the mitochondrial control region of *S. gregaria* has been determined by directly sequencing single-stranded PCR products using the Sequenase DNA Sequencing Kit (USB) manually or using the AutoRead Sequencing Kit on A.L.F. DNA Sequencer. New primers have been designed from the sequences obtained using PCR primers J1 and J6 in order to sequence the whole control region: primer G1 (AATGACCACAACAACTTCTC), G2 (CATCTTACCATTATCAA), G3 (TAAAAACATAAGTAGC), and G4 (GTGAAAAGAAAGATT). Their locations are shown in Fig. 1. At least two independent PCR products have been used to sequence each segment of the control region.

Computer Analysis. The GCG package of the SERC Daresbury Laboratory has been used to make sequence comparisons, multiple sequence alignment, and secondary structure analysis (Programs BESTFIT, PILEUP, and FOLDRNA). The software DNASTAR (Beta 0.94, 1992) has also been used for primary sequence analysis. The multiple sequence alignment in Fig. 2 is a modified and combined version of that obtained using the GCG package and DNASTAR. Secondary structures shown in Fig. 4C were modified from results obtained by the GCG package.

Results and Discussion

The desert locust *Schistocerca gregaria* and meadow grasshopper *Chorthippus parallelus* have been classified in the insect order *Orthoptera,* suborder *Caelifera,* and superfamily *Acrididae* (*Locustidae*). *S. gregaria* is in the subfamily *Cyrtacanthacridinae* and *C. parallelus* in *Gomphoceridinae.* The cricket *Gryllus firmus* is in the suborder *Ensifera* of the *Orthoptera. Drosophila* is in the order *Diptera* and honeybee *A. mellifera* in the order *Hymenoptera.* The above five species groups are in the

Fig. 2. Conserved sequence blocks in the control regions of *S. gregaria* and *C. parallelus*. Sequences are shown in the direction 5′–3′, *S.g., S. gregaria; CpR1* and *CpR2*, the two tandemly repeated units in *C. parallelus* (Fig. 1). The eight conserved sequence blocks, *A, B, C, D, E1, E2, F,* and *G* are indicated; within each block, nucleotides identical in all three sequences are top-marked with *asterisks*. The 12-nucleotide sequences written in *lowercase in brackets* in S.g. and CpR2 are of the small rRNA gene, and are repeated once at the end of CpR1.

same subclass *Pterygota* of the class *Insecta* (Uvarov 1966; Borror et al. 1989).

Primary Structure and Conserved Sequence Blocks in the Control Regions of S. gregaria *and* C. parallelus

The control region of *S. gregaria* comprises 762 bp, compared to 1,512 bp in *C. parallelus*. The unusual length of the latter sequence is due to the presence of a tandemly repeated sequence of the original A + T–rich region (CpR1 and CpR2 in Fig. 1). The A + T content of these regions is 86.8% and 85.1% in *S. gregaria* and *C. parallelus,* respectively, justifying the name "A + T–rich region." In both sequences no open reading frames containing more than 40 codons can be found in any of the six possible frames (if ATG, ATT, and ATA are used as translation initiation codons, and TAA and TAG as stop codons. Clary and Wolstenholme 1985). Comparison of the sequences of *S. gregaria* and the two repeated sequences (CpR1 and CpR2) of *C. parallelus* is given in Fig. 2. The two repeats in *C. parallelus* A + T–rich region are highly homologous (97.5% similarity); the repeat CpR2, has a 41-bp deletion at its 5′ end as compared to CpR1, and is flanked by a 12-bp direct repeat of the small rRNA gene origin (Figs. 1 and 2). Sequence similarity between *S. gregaria* and CpR1 of *C. parallelus* is 67.7%. Eight conserved sequence blocks have been identified between *S. gregaria* and *C. parallelus:* blocks A, B, C, D, E1, E2, F, and G (Fig. 2). It is worth noting that these conserved blocks are spread through the whole A + T–rich region, producing a frequency of nucleotide changes that is broadly the same over the whole region. This indicates a different evolutionary pattern for this region in *S. gregaria* and *C. parallelus* as compared to that in *Drosophila* species, where the A + T–rich region contains two distinct domains, one conserved domain located near to the $tRNA^{ile}$ gene and one variable domain. (See below.)

Among these blocks, block A of 33/34 nt is characterized by a run of nucleotide T's, located immediately downstream of, or very close to, the $tRNA^{ile}$ gene. This block, despite its richness in A + T (87.9% and 94.1% in *S. gregaria* and *C. parallelus,* respectively), is well conserved between these two subfamilies (88.2% similarity); its location near the 5′ end with a run of T's flanked by one purine on each side can also be traced in the A + T–rich region of *Drosophila* species (Fig. 3A; Clary and Wolstenholme 1987; Monforte et al. 1993). This may indicate some functional importance for this block.

Block B is characterized by the core sequence "5′-TTAATATATTACATTT-3′," which is absolutely conserved between the two subfamilies. A similar sequence is also present in the same relative location in *Drosophila* mitochondrial A + T–rich region (Fig. 3B), suggesting its possible implication in transcription or replication control. These sequences, having a consensus "5′-A . . . TAA . T . ATTTA . . TT . . . ATA . . ACATTT-3′" (Fig. 3B), resemble the template stop signals for D-loop synthesis in human and mouse mtDNA (Clayton 1982) and CSB1 block identified in mammalian mitochondrial D-loop regions (Walberg and Clayton 1981; Saccone et al. 1991), and they share the sequence "5′-ACAT-3′."

A. Block A-like sequences

```
S. gregaria      atttaatatataaatcgaaaGTTTTTTTGaaattgtttt
C. parallelus    atttaatatataaatcgaaaGTTTTTTTGtaaaataag
D. virilis       aaacccgtctATTTTTTTTTTTTTTTTTTTTTGtacttta
D. yakuba        aaaactcATCTTTTTTTTTTTTTTTTTTTAttatt
D. teissieri     aaaactcATCTTTTTTTTTTTTTTTTTTTTTAttatt
D. obscura       tattccATTCTTTTTTTTTTTTTTTTTTTTAttcta
D. ambigua       tattctcATTCTTTTTTTTTTTTTTTTTTTAttcta

Consensus        ......RTTTTTTTT(TTTTTTTTTTTTTT)R......
```

B. Block B-like sequences

```
S. gregaria      atatatAtaaTAAaTaATTTAtaTTaatATAttACATTTaattga
C. parallelus    acttttAataTAAaTg-TTTAatTTaatATAttACATTTgtttga
D. virilis       tttaatAttaTAAtT-ATTTAaaTTa-aATAatACATTTtagtaa
D. yakuba        tttaa-AaatTAAtTgATTTAtaTTa-gATAacACATTTtaggaa
D. teissieri     tttaa-AaatTAAtTgATTTAtaTTa-gATAacACATTTtaggaa
D. obscura       aataa-AataTAAtTaATTTAttTTttaATAatACATTTtagtat
D. ambigua       aaaaa-ActtTAAtTaATTTAttTTttaATAacACATTTtagtat

Consensus        ......A...TAA.T.ATTTA..TT...ATA..ACATTT......
```

Fig. 3. *S. gregaria/C. parallelus* blocks A- and B-like sequences in the A + T–rich regions of *Drosophila* species. *Letters in lowercase* and *dots* in the consensus sequences indicate nonconserved nucleotides. **A** Block A–like sequences. The number of nucleotide T's in brackets in the consensus sequence is variable between species; *R,* purine nucleotides. **B** Block B–like sequences. *Bold letters in uppercase* represent conserved nucleotides which are given in the consensus sequence. *D. yakuba* and *D. virilis* sequences are from Clary and Wolstenholme (1985, 1987), *D. teissieri* from Monnerot et al. (1990), and *D. obscura* and *D. ambigua* from Monforte et al. (1993).

Block C and block D are characterized by their extreme richness in A + T and by their location downstream of the conserved block B and upstream of two other important blocks E1 and E2 (see below), and they are well conserved in these two subfamilies.

Block F of 21 nt is identical in the two subfamilies. It is worth noting that this sequence block, "5′-ATATAA-TAGAGAAGTTGTTGT-3′," has a high G content (24%) and spans the junction between a A/T-rich 5′ stretch and G/C-rich 3′ section. Adjacent to the 5′ A/T-rich stretch is another G/C-rich region, so block F is a core surrounded by G/C-rich sequences. These unusual characteristics might involve some as-yet-unknown functional role.

Block G is a short sequence lying near the 3′ end close to the small rRNA gene, having a conserved sequence pattern of "5′-TTTTCTwTAAwTATTTGAwTC-3′" (w = A or T).

Block E1 of 18 nt is identical, and block E2, 22/24 nt in length, is highly conserved (91.7% similarity) in the two subfamilies. In fact, block E1 is a part inverse repeat of block E2; the sequences containing these two blocks can form a stem and loop (or hairpin) secondary structure (Fig. 4A, and see below for further discussion).

Conserved Secondary Structure in S. gregaria *and* C. parallelus *and Its Possible Conservation in Invertebrates*

As mentioned above, the sequence containing the two blocks E1 and E2 can form a stem and loop (or hairpin) secondary structure (Fig. 4A). In *S. gregaria,* the stem of this highly conserved secondary structure is formed by a perfect match of 17 nucleotide pairs, and the terminal loop is 12 nt. In *C. parallelus,* the corresponding stem is formed by 16 nucleotide pairs with only one mismatch, the terminal loop is 14 nt. It is interesting that in *C. parallelus,* the first 13 nucleotide pairs in the stem are almost identical (one mismatch) in sequence to that in *S. gregaria,* while the remaining pairs are completely different—i.e., there are 3 pairs in *C. parallelus* with 1 A-T pair and 2 C-G pairs, while in *S. gregaria* nucleotide substitutions have resulted in 4 pairs with 1 C-G pair and 3 A-T pairs. The extra pair in the latter may compensate partly the effect of having one less C-G pair. In contrast to the conservation in the stems, the sequences of the terminal loop are highly divergent, indicating that the loop region sequence in the conserved secondary structure has less functional importance. In addition, this conserved secondary structure remains almost intact in the tandemly repeated copy in *C. parallelus* (Fig. 4A). If the sequences of the terminal loops in the two repeated units (CpR1 and CpR2) are compared, a difference of 7.1% (1 insertion/deletion event over the 14 nt) can be observed. This makes up 5.6% of the overall difference between CpR1 and CpR2.

While the sequences of the loop region are highly divergent between the two subfamilies, it is noteworthy that the sequences immediately flanking the hairpin structure (5′ and 3′) are highly conserved (Fig. 4A and D). The 5′ flanking sequences are A + T–rich ("... TTATA"), while the 3′ flanking sequences appear to contain a "GAAAGAATA" motif (Fig. 4A, and see below) and are more conserved than the 5′ ones. Again, these patterns are strictly conserved in the second repeat in *C. parallelus.*

It has recently been shown that a conserved hairpin secondary structure is probably universally present in the A + T–rich region of *Drosophila* species (Fig. 4B; Clary and Wolstenholme 1987; Monforte et al. 1993). By S1 nuclease assays, Monforte et al. (1993) demonstrated in vitro the formation of just such a secondary structure in *D. ambigua.* The hairpin structures are important because the replication of circular DNA molecules has been shown to initiate within or close to them (Zannis-Hadjopoulos et al. 1988). Furthermore it is known that in mammalian mtDNA, the replication origin for the synthesis of the light strand (L-strand) is associated with a stem and loop secondary structure, which is located between the $tRNA^{asn}$ and $tRNA^{cys}$ genes and highly conserved in mammalian, amphibian, and fish mtDNA (Clayton 1982; Anderson et al. 1982; Wong et al. 1983; Johansen et al. 1990). In *Drosophila,* from data obtained by electron microscopy and DNA sequence analysis, Clary and Wolstenholme (1987) have suggested that the conserved hairpin-forming sequence could be the site of initiation of the second strand synthesis. This may well be true considering that a potential second-strand-replication origin (oriB) also seems to be associated with

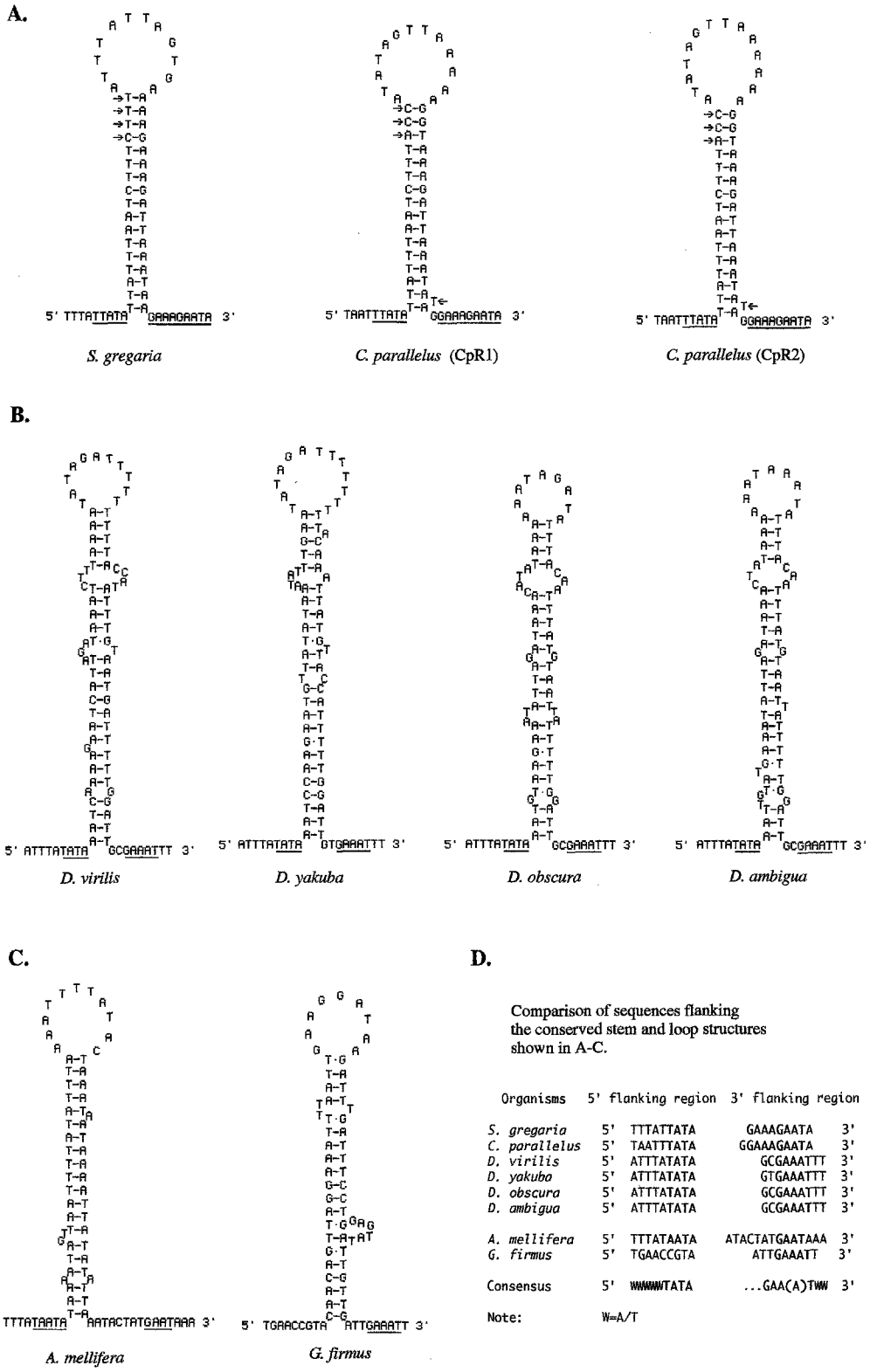


Fig. 4. Possible conserved secondary structures in the mitochondrial control regions of *S. gregaria* and *C. parallelus* **(A),** *Drosophila* species **(B),** *A. mellifera* and *G. firmus* **(C),** and the conserved flanking sequences of these secondary structures **(D).** In **A, B,** and **C,** the *underlined sequences* in the flanking regions of the stem-and-loop structures are the conserved motifs, which are summarized in **D.** In **A,** *arrows* indicate the differences of the stem regions between *S. gregaria* and *C. parallelus.* The secondary structures of *Drosophila* species are from Monforte et al. (1993) with slight modification; the sequences shown in **C** correspond to nucleotides 15,581–15,660 (complementary) in Crozier and Crozier (1993) for *A. mellifera* and 619–694 in Rand and Harrison (1989) for *G. firmus. D. teissieri* also has a similar secondary structure with conserved flanking sequences (data not shown) in its A + T–rich region (Monnerot et al. 1990); please refer to Monforte et al. (1993).

Table 1. 3′ sequences flanking the stem-and-loop structures associated or potentially associated with second-strand-replication origins in mitochondrial DNA

Organisms	3′ flanking sequences (5′–3′)	Sequence sources
S. gregaria	stem-**GAAAGAAT**ATAAT	This paper
C. parallelus	stem-G**GAAAGAAT**AATTT	This paper
Drosophila	stem-GY**GAAAT**TTTTTATT	Refer to Figs. 3 and 4
G. morhua	stem-A**GAT**AGATGCTCGCTG	Johansen et al. 1990
X. laevis	stem-TA**GAAT**GAAGCTC	Reviewed in Brown 1985
Rat	stem-TA**GAT**TGAAGCCA	Reviewed in Brown 1985
Mouse	stem-A**GAT**TGAAGCCA	Reviewed in Brown 1985
Cow	stem-TA**GAT**TGAAGCCA	Reviewed in Brown 1985
Human	stem-TA**GAT**TGAAGCCA	Reviewed in Brown 1985
P. hybrida	stem-**GAAAGAAAAT**TCTT	de Haas et al. 1991
Conserved motif	stem .. **G(A)nT**	
Chicken	$tRNA^{cys}$-GTA**GGC**AGAAGCCA	Desjardins and Morais 1990
Japanese quail	$tRNA^{cys}$-TA**GAC**AGAAGCTA	Desjardins and Morais 1991

a stem-and-loop secondary structure in plant mtDNA (de Haas et al. 1991) and the mammalian L-strand-replication origin even shares primary sequence homologies with some nonmitochondrial systems (Clayton 1982).

Thus it is interesting to know whether the conserved hairpin-forming sequences identified in *S. gregaria* and *C. parallelus* are equivalent to that found in *Drosophila* species (Fig. 4B). There is no significant sequence similarity in these sequences between *S. gregaria/C. parallelus* and *Drosophila;* the loop sequences in *S. gregaria/C. parallelus* are longer on average (12–14 nucleotides) compared to those in *Drosophila* (9, 11, or 13 nucleotides); and the stem regions in the two Orthoptera are much shorter (17/16 pairs) than those in the fruitflies (22–24 pairs), but the former have a much more perfect match than the latter (Fig. 4A and B). Despite of these differences, the conserved secondary structures in these organisms are very similar (Fig. 4A and B). This can be seen not only from the conformation of the stem and loop structures itself but also from several other features such as the similarities of sequences flanking them (Fig. 4D) and their relative locations in the control regions. (1) It is clear that the 5′ flanking sequences are all rich in A + T, having a consensus sequence "TATA"; the 3′ flanking sequences share a sequence consensus "GAA(A)T." It is worth pointing out here that the motif "G(A)nT" ("n" varies from one to four) is also conserved in the 3′ flanking sequences of stem-and-loop structures associated with L-strand replication origins in mammalian, amphibian, and fish mtDNA and a potential second-strand-replication origin (oriB) of *Petunia hybrida* mtDNA (Table 1), while when the equivalent L-strand replication origins are absent in the corresponding regions in chicken and Japanese quail (Desjardins and Morais 1990, 1991; and see below) this motif is also missing (Table 1). These conserved motifs might function as recognition signals for some specific trans-active factors to interact with the hairpin structure. (2) We can also see that two conserved sequence blocks, A and B, in *S. gregaria/C. parallelus* (Figs. 2 and 3) both have homologues lying in similar locations in *Drosophila.* Moreover, in *S. gregaria* and *C. parallelus,* the conserved sequence blocks C and D which are almost pure A/T sequences in *S. gregaria* are located between the block B and the conserved hairpin structure; a similar conserved A + T–rich sequence is also present in a similar location in the A + T–rich region of *Drosophila* (data not shown; see Fig. 5 in Monforte et al. 1993).

So it seems that these conserved secondary structures in *S. gregaria/C. parallelus* and *Drosophila* are closely related despite the distant phylogenetic relationship of these organisms, and they may well have a similar function such as a second-strand-replication origin. Furthermore, this suggests that such a secondary structure in the mitochondrial control regions may be widely conserved in invertebrates, indicating its early occurrence in evolution. We have therefore looked in the published sequences of mitochondrial control regions of the honeybee *A. mellifera* (Crozier and Crozier 1993) and the cricket *G. firmus* (Rand and Harrison 1989), and were able to detect some hairpin-forming sequences of similar size and possibly location in both insects.

In *A. mellifera,* since the sequence is extremely rich in A + T (96%) (Crozier and Crozier 1993), and contains several runs of ATs, more than a dozen stem-and-loop-forming sequences can be found (data not shown). Among these candidates, we found that one shown in Fig. 4C as the most closely homologous in location, size, and sequence similarity, particularly the conserved status of the 5′ flanking sequence "TAATA" and the 3′ "GAAT" (Fig. 4C,D). Of course, identifying such a structure from primary sequence data will only be possible by comparing the sequences of the A + T–rich region between *A. mellifera* and a related hymenopteran species; the sequences able to form such a conserved structure would be much more conserved than the rest. It has been suspected that the duplicated region between the $tRNA^{leu}$ and COII genes in *A. mellifera* mitochon-

drial DNA may contain an additional, or replacement, origin of replication (hairpin structure) (Cornuet et al. 1991). However, this duplicated region seems highly variable even within the same species, and is often reduced or sometimes absent in other bees (Cornuet et al. 1991; Crozier and Crozier 1993), making this latter possibility less likely.

In *G. firmus,* the corresponding control region is much less A + T–rich. It consists of three almost identical tandem repeats with the last one having a 26-nt deletion at the 3′ end; the full length of one repeated unit is 220 bp. Rand and Harrison (1989) have proposed that the cruciform structure-forming dyad symmetric sequences "GGGGGCATGCCCCC" present in all the repeats could be important for the functioning of the replication origin in this region. Alternatively, we were able to detect a sequence segment in each repeat unit forming a stem-and-loop structure homologous to those found in *Drosophila* and *S. gregaria/C. parallelus* (Fig. 4C, note "GAAAT" in the 3′ flanking sequence and "CGTA" in the 5′ flanking sequence which can be a derived form of "TATA"). As in *C. parallelus,* each repeat unit contains such a hairpin-forming sequence located in their center region, 98 nt upstream of its 3′ end (62 nt for the last repeat). Again this speculation needs to be confirmed by further studies using a related species. Nevertheless, these observed characteristics in *A. mellifera* and *G. firmus* are distinctly suggestive of mtDNA control regions.

Evolution of the Control Region of Mitochondrial DNA in Insects

The sequence of mitochondrial A + T–rich region characterized to date in five different insect taxa (*Drosophila, G. firmus, A. mellifera, S. gregaria,* and *C. parallelus*) can be classed into two distinct groups according to their primary structures. *Drosophila* species fall into the first group and their A + T–rich regions, as observed by Monnerot et al. (1990) and Monforte et al. (1993), contain two different domains: one conserved domain of 417/438 bp adjacent to $tRNA^{ile}$ gene which is highly conserved among *Drosophila* species and contains the highly conserved secondary structure potentially associated with a replication origin; and one variable domain including the rest of the region, which is highly variable both in nucleotide sequence and length. It is worth noting that in some *Drosophila* species, such as *D. tristis* (Monforte et al. 1993) and probably *D. melanogaster, D. mauritiana, D. simulans, D. sechellia* (Solignac et al. 1986), tandem repetition occurred producing A + T–rich regions that contain more than one conserved domain.

The second group comprises the A + T–rich regions of *S. gregaria* and *C. parallelus* and perhaps *G. firmus.* Unlike the first group, the A + T–rich regions of this group cannot be divided into distinct conserved and variable domains. As shown in *S. gregaria* and *C. parallelus,* the A + T–rich regions seem equally conserved (or variable) along their sequence. Length variation occurred, if at all, at the 5′ end rather than at the 3′ end of the A + T–rich region. For example, the largest insertion/deletion between *S. gregaria* and *C. parallelus* has occurred in the $tRNA^{ile}$ adjacent end (Fig. 2). Interestingly, tandem repetition has also occurred in this group in *C. parallelus* and *G. firmus,* producing A + T–rich regions containing more than one possible origin of replication.

So the A + T–rich regions in *S. gregaria/C. parallelus* have a different evolutionary pattern from that in *Drosophila.* The absolute location of the highly conserved secondary structures in these regions is also different in *S. gregaria/C. parallelus* compared to that in *Drosophila,* although in both groups they lie in the region which has less length variation (i.e., the 5′ region in the first group and the 3′ region in the second). In the two Orthoptera the conserved secondary structures are located near to the small rRNA gene while in *Drosophila* it is near to the $tRNA^{ile}$ gene. (In fact, the conserved domain in *Drosophila* resembles a condensed or compact A + T–rich region of *S. gregaria;* the preservation of the stem-and-loop structures and several conserved sequence blocks, together with their relative locations in the sequences, indicate a basic similarity between them.) It will be interesting to know in which group *A. mellifera* falls, and if the hypothetical stem-and-loop structure we proposed exists in it, or if an equivalent one can be found in a similar location to either *Drosophila* or *S. gregaria.* If such a secondary structure indeed reveals the second strand (L-strand) replication origin as inferred in *Drosophila* (Clary and Wolstenholme 1987), the above observation would indicate that its location may be variable from one species group to another. This may well prove true because in galliform birds, such as chicken (*Gallus gallus domesticus*) and Japanese quail (*Coturnix japonica*), it has been reported that an L-strand replication origin equivalent to the conserved mammalian and amphibian sequences cannot be found in the corresponding regions (the 30-nt or so noncoding regions between the gene for $tRNA^{asn}$ and $tRNA^{cys}$, which are absent from these organisms) (Desjardins and Morais 1990, 1991). Whether an equivalent secondary structure can be found elsewhere in chicken and Japanese quail mitochondrial genome will be important for testing both the universal conservation and variable location of L-strand replication origin in metazoan mtDNA. Since the primary sequence homologies between the highly conserved hairpin structures in *Drosophila* and *S. gregaria/C. parallelus* are very low, the primary structure seems not so important, and is thus much less conserved during evolution.

As mentioned above, direct tandem repeats of the conserved domain in the A + T–rich region are observed in some *Drosophila* species. This situation also occurs in

the cricket *G. firmus* (Rand and Harrison 1989), where the corresponding control region consists of three tandem repeats with copy-number variation within species. In the present paper, we show that the A + T–rich region of *C. parallelus* is made up of two tandem repeats of the original A + T–rich region as found in *S. gregaria.* In other closely related species pairs, such as *D. yakuba* with *D. melanogaster* and *D. obscura* with *D. tristis,* the tandem-repeat pattern exists in one and not the other. The occurrence of such tandem repetitions in the control region in dispersed phylogenetic positions strongly suggests a convergent evolution. Furthermore, the occurrence of tandem repetition in the control region poses a most interesting question. As repeated units all contain the highly conserved stem-and-loop structure (potential replication origin, e.g., in *C. parallelus*) and are evolving concertedly (Solignac et al. 1986), tandem repetition thus results in a control region containing more than one potential origin of replication. Therefore it is interesting to ask how replication control functions in this situation and whether all repeated units are implicated in this control. Analysis of published data suggests that DNA replication seems to originate in only one of the repeated region (one or other near the center) in *Drosophila,* thereby suggesting certain mechanisms involved in this process.

The mechanism that could generate such repetition in mitochondria is unknown at present. It is very possible that the original control region contained only one "repeat" unit, as is the case in *S. gregaria.* Several mechanisms have been proposed by different authors (e.g., Rand and Harrison 1989; Cornuet et al. 1991; Monforte et al. 1993), however, difficulties arise when using them to explain the generation of the first critical repetition, or the convergent evolution of this phenomenon in a number of widely spread species in disjunct biological lineages. Recombination has been thought less likely in animal mtDNA (Brown 1985); however, it may be a suitable mechanism to produce the first repetition of tandem repeats in a mitochondrial control region, especially via parahomologous recombination by unequal crossing-over. It has already been suggested that parahomologous recombination by unequal crossing-over theoretically can generate repeats from nonrepetitious DNA (Smith 1976). What this mechanism requires is local base-pairing between regions with reasonable homology (not necessarily extensive stretches of homology). This is possible especially when the nucleotide composition of a sequence is strongly biased, such as the mitochondrial A + T–rich regions in insect. When the control region is concerned in unequal crossing-over, one of the two recombinant products will lack the control region and thus some necessary control elements for transcription and replication, and should be eliminated by the cell. Although there is no direct evidence yet for the existence of recombination in animal mtDNA, phenomena such as the presence of oligomeric mtDNA in mammalian cells indirectly support its occurrence (Clayton 1982). Indeed the widespread presence of tandemly repeated sequences in the mtDNA in animals (for examples, see Densmore et al. 1985; Moritz and Brown 1986; Solignac et al. 1986; Boyce et al. 1989; Rand and Harrison 1989; Buroker et al. 1990; La Roche et al. 1990; Mignotte et al. 1990) may itself be an indication of the existence of recombination in mitochondria. Actual DNA sequences of the whole tandem repeat-containing control region are only available in *G. firmus* (Rand and Harrison 1989) and *C. parallelus* (in this paper), and more sequence data will certainly reveal useful information on the mechanisms generating these repeats.

Acknowledgments. This work has been supported by grants C-91283 from UNDP (FAO) BIO2CT-920476 from EU and GR/E93428 from SERC (UK).

References

Anderson S, de Bruijn MHL, Coulson AR, Eperon IC, Sanger F, Young IG (1982) Complete sequence of bovine mitochondrial DNA: conserved features of the mammalian mitochondrial genome. J Mol Biol 156:683–717

Borror DJ, Triplehorn CA, Johnson NF (1989) An introduction to the study of insects, 6th ed. Saunders College Publishing, Philadelphia

Boyce TM, Zwick ME, Aquadro CF (1989) Mitochondrial DNA in the bark weevils: size, structure and heteroplasmy. Genetics 123:825–836

Brown GG, DesRosiers LJ (1983) Rat mitochondrial DNA polymorphism: sequence analysis of a hypervariable site for insertions/deletions. Nucleic Acids Res 11:6699–6708

Brown GG, Gadaleta G, Pepe G, Saccone C, Sbisa 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 PJ, Beckenbach AT, Thomas WK, Smith MJ (1990) Length heteroplasmy of sturgeon mitochondrial DNA: an illegitimate elongation model. Genetics 124:157–163

Clary DO, Wolstenholme DR (1985) The mitochondrial DNA molecule of *Drosophila yakuba:* nucleotide sequence, gene organization, and genetic code. J Mol Evol 22:252–271

Clary DO, Wolstenholme DR (1987) *Drosophila* mitochondrial DNA: conserved sequences in the A + T-rich region and supporting evidence for a secondary structure model of the small ribosomal RNA. J Mol Evol 25:116–125

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

Cornuet JM, Garnery L, Solignac M (1991) Putative origin and function of the intergenic region between COI and COII of *Apis mellifera* L. mitochondrial DNA. Genetics 128:393–403

Crozier RH, Crozier YC (1993) The mitochondrial genome of the honeybee *Apis mellifera:* complete sequence and genome organization. Genetics 133:97–117

de Haas JM, Hille J, Kors F, van der Meer B, Kool AJ, Folkerts O, Nijkamp HJJ (1991) Two potential *Petunia hybrida* mitochondrial DNA replication origins show structural and *in vitro* functional homology with the animal mitochondrial DNA heavy and light strand replication origins. Curr Genet 20:503–513

Densmore LD, Wright JW, Brown WM (1985) Length variation and

heteroplasmy are frequent in mitochondrial DNA from parthenogenetic and bisexual lizards (Genus *Cnemidophorus*). Genetics 110:689–707

Desjardins P, Morais R (1990) Sequence and gene organization of the chicken mitochondrial genome. A novel gene order in higher vertebrates. J Mol Biol 212:599–634

Desjardins P, Morais R (1991) Nucleotide sequence and evolution of coding and noncoding regions of a quail mitochondrial genome. J Mol Evol 32:153–161

Erlich HA (ed) (1989) PCR technology. Principles and applications for DNA amplification. Stockton Press, New York

Gray MW (1989) Origin and evolution of mitochondrial DNA. Ann Rev Cell Biol 5:25–50

Innis MA, Gelfand DH, Sninsky JJ, White TJ (eds) (1990) PCR protocols. A guide to methods and applications. Academic Press, San Diego

Johansen S, Guddal PH, Johansen T (1990) Organization of the mitochondrial genome of Atlantic cod, *Gadus morhua.* Nucleic Acids Res 18:411–419

La Roche J, Snyder M, Cook DI, Fuller K, Zouros E (1990) Molecular characterization of a repeat element causing large-scale variation in the mitochondrial DNA of the sea scallop *Placopecten magellanicus.* Mol Biol Evol 7:45–64

Marchant AD (1988) Apparent introgression of mitochondrial DNA across a narrow hybrid zone in the *Caledia captiva* species-complex. Heredity 60:39–46

McCracken A, Uhlenbusch I, Gellissen G (1987) Structure of the cloned *Locusta migratoria* mitochondrial genome: restriction mapping and sequence of its DN-1 (URF-1) gene. Curr Genet 11:625–630

Mignotte F, Gueride M, Champagne AM, Mounolou JC (1990) Direct repeats in the non-coding region of rabbit mitochondrial DNA. Involvement in the generation of intra- and inter-individual heterogeneity. Eur J Biochem 194:561–571

Monforte A, Barrio E, Latorre A (1993) Characterization of the length polymorphism in the A + T-rich region of the *Drosophila obscura* group species. J Mol Evol 36:214–223

Monnerot M, Solignac M, Wolstenholme DR (1990) Discrepancy in divergence of the mitochondrial and nuclear genomes of *Drosophila teissieri* and *Drosophila yakuba.* J Mol Evol 30:500–508

Moritz C, Brown WM (1986) Tandem duplication of D-loop and ribosomal RNA sequences in lizard mitochondrial DNA. Science 233:1425–1427

Moritz C, Dowling TE, Brown WM (1987) Evolution of animal mitochondrial DNA: relevance for population biology and systematics. Ann Rev Ecol Syst 18:269–292

Rand DM, Harrison RG (1989) Molecular population genetics of mtDNA size variation in crickets. Genetics 121:551–569

Saccone C, Attimonelli M, Sbisa E (1987) Structural elements highly preserved during the evolution of the D-loop-containing region in vertebrate mitochondrial DNA. J Mol Evol 26:205–211

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

Sambrook J, Fritsch EF, Maniatis T (1989) Molecular cloning. A laboratory manual, 2nd ed. Cold Spring Harbor Laboratory Press, New York

Simon C (1991) Molecular systematics at the species boundary: exploiting conserved and variable regions of the mitochondrial genome of animals via direct sequencing from amplified DNA. In: Hewitt GM, Johnston AWB, Young JPW (eds) NATO ASI Series, vol. 57. Molecular techniques in taxonomy. Springer-Verlag, pp 33–71

Smith GP (1976) Evolution of repeated DNA sequences by unequal crossover. Science 191:528–535

Solignac M (1991) Preparation and visualization of mitochondrial DNA for RFLP analysis. In: Hewitt GM, Johnston AWB, Young JPW (eds) NATO ASI Series, vol 57. Molecular techniques in taxonomy. Springer-Verlag, Berlin, pp 295–319

Solignac M, Monnerot M, Mounolou JC (1986) Concerted evolution of sequence repeats in *Drosophila* mitochondrial DNA. J Mol Evol 24:53–60

Uvarov B (1966) Grasshoppers and locusts. A handbook of general acridology. Volume 1: anatomy, physiology, development phase polymorphism, introduction to taxonomy. Cambridge University Press, London, pp 397–420

Walberg MW, Clayton DA (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

Wong JFH, Ma DP, Wilson RK, Roe BA (1983) DNA sequence of the *Xenopus laevis* mitochondrial origins and flanking tRNA genes. Nucleic Acids Res 11:4977–4995

Zannis-Hadjopoulos M, Frappier L, Khoury M, Price GB (1988) Effect of anti-cruciform DNA monoclonal antibodies on DNA replication. EMBO J 7:1837–1844