Phylogenetic analysis of Biomphalaria tenagophila (Orbigny, 1835) (Mollusca: Gastropoda)

Liana K Jannotti-Passos[1]/+, Jeronimo C Ruiz[2], Roberta L Caldeira[3], Silvane MF Murta[2], Paulo Marcos Z Coelho[4], Omar S Carvalho[3]

[1]Moluscário Lobato Paraense [2]Laboratório de Parasitologia Celular e Molecular [3]Laboratório de Helmintologia e Malacologia Médica [4]Laboratório de Esquistossomose, Instituto de Pesquisas René Rachou-Fiocruz, Av. Augusto de Lima 1715, 30190-001 Belo Horizonte, MG, Brasil

Mitochondrial DNA of Biomphalaria tenagophila, *a mollusc intermediate host of* Schistosoma mansoni *in Brazil, was sequenced and characterised. The genome size found for* B. tenagophila *was 13,722 bp and contained 13 messenger RNAs, 22 transfer RNAs (tRNA) and two ribosomal RNAs (rRNA). In addition to sequencing, the mitochondrial DNA (mtDNA) genome organization of* B. tenagophila *was analysed based on its content and localization of both coding and non-coding regions, regions of gene overlap and tRNA nucleotide sequences. Sequences of protein, rRNA 12S and rRNA 16S nucleotides as well as gene organization were compared between* B. tenagophila *and* Biomphalaria glabrata, *as the latter is the most important* S. mansoni *intermediate host in Brazil. Differences between such species were observed regarding rRNA composition. The complete sequence of the* B. tenagophila *mitochondrial genome was deposited in GenBank (accession EF433576). Furthermore, phylogenetic relationships were estimated among 28 mollusc species, which had their complete mitochondrial genome deposited in GenBank, using the neighbour-joining method, maximum parsimony and maximum likelihood bootstrap.* B. tenagophila *was positioned at a branch close to* B. glabrata *and Pulmonata molluscs, collectively comprising a paraphyletic group, contrary to Opistobranchia, which was positioned at a single branch and constituted a monophyletic group.*

Key words: *Biomphalaria tenagophila* - mitochondrial genome - phylogenetic

The phylum Mollusca constitutes the second largest animal phylum and encompasses seven classes, namely: Gastropoda, Polyplacophora, Cephalopoda, Bivalvia, Scaphopoda, Aplacophora and Monoplacophora (Brusca & Brusca 2007). The phylum has a long geological history and approximately 35,000 species are known. *Biomphalaria* (Gastropoda) molluscs are known to have originated during the Jurassic era and endured a wide range of environmental changes, resulting in the development of survival mechanisms such as self-fecundation, diapause, aestivation, capacity to survive into deep trenches and high prolificacy (Paraense 1955), making them currently adapted to different environmental conditions.

Biomphalaria tenagophila is an important species concerning schistosomiasis transmission in Southern Brazil. This species has been the focus of investigations, as it comprises susceptible (Paraense & Corrêa 1978) populations and one population that is resistant to *Schistosoma mansoni*, which is from the Biological Reserve of Taim in Rio Grande do Sul, Brazil.

These molluscs have already been challenged by different *S. mansoni* strains and variable number of miracidia, yet remained resistant to infection (Santos et al. 1979, Bezerra et al. 2003, Martins-Souza et al. 2003, Rosa et al. 2004, 2006, Coelho & Bezerra 2006). Several studies have been carried out in order to gain a better understanding of resistance traits, including molecular approaches using nuclear DNA regions (Pires et al. 1997, Spatz et al. 1999, Barbosa 2001, Rosa et al. 2005, Barbosa et al. 2006), but little is known about the genes related with the resistance phenotype in this population (Santos et al. 1979, de Freitas et al. 1985, Rosa 2002).

Studies using mitochondrial DNA have been pursued, as they provide further knowledge on extra-nuclear genetic information (Avise 1986). The mitochondrial DNA (mtDNA) molecule is a double-closed circle coding 13 messenger RNAs (mRNA), 22 transfer RNAs (tRNA) and two ribosomal RNAs (rRNA). This molecule contains non-coding regions with different sizes, including transcription start and replication sites (Wolstenholme 1992). Fractions of these non-coding regions may be duplicated, resulting in mtDNA molecules of different sizes in some organisms. In molluscs, mtDNA is also a circular molecule in which mitochondrial genes (Boore & Brown 1994) and sequence order (Lecanidou et al. 1994) are organized in varied patterns compared with mtDNA from other organisms. Moreover, mtDNA size varies within the phylum Mollusca. Among the *Biomphalaria* species, *Biomphalaria glabrata* mtDNA is the only completely sequenced mitochondrial genome (DeJong et al. 2004).

In the present paper, the *B. tenagophila* mtDNA genome was completely sequenced and characterized, and genome organization was analysed concerning its content and localization of both coding and non-coding regions, as well as gene overlapping regions. Nucleotide

Financial support: FAPEMIG, FIOCRUZ
+ Corresponding author: passoslk@cpqrr.fiocruz.br
Received 30 January 2009
Accepted 18 December 2009

sequences of 12S and 16S subunits of the ribosomal RNA and gene organization in *B. tenagophila* and *B. glabrata* were compared. In addition, the phylogenetic relationship was estimated among 28 mollusc species, for all of which the complete mitochondrial genomes were available in GenBank, through the neighbour-joining method (NJ), maximum parsimony (MP) and maximum likelihood (ML).

MATERIALS AND METHODS

Molluscs and DNA extraction - The *B. tenagophila* cephalopodal region was used as a DNA source. Specimens of this species were sent from the Biological Reserve of Taim originating from the colony at Federal University of Minas Gerais Interdepartmental Group for Schistosomiasis and then kept under laboratory conditions at the Lobato Paraense Mollusc Rearing Sector of the René Rachou Research Center-Fiocruz. This fragment was incubated in a buffer extraction solution (50 mM Tris-HCl -a GIBCO pH 8.0; 50 mM EDTA, 100 mM NaCl, SDS 0.5%) added to 20 µg/mL proteinase K for 12 h at 37ºC. The snail DNA extraction was carried out using kit Wizard® (Promega), according to the manufacturer's instructions.

Polymerase chain reaction (PCR) - Universal primers [LCO and HCO (Folmer et al. 1994)], 16Sar and 16Sbr (Palumbi 1996), 12Sai and 12Sbi (Simon et al. 1990), Cyb1 - 5'AAAGGAAGAACTAAGCCTTTAGG 3' (designed from *B. glabrata* sequence) and CB3 (Kocher et al. 1989) were used to amplify part of the regions of COI, 16S, 12S and Cytb, respectively. These regions were directly sequenced from PCR products, after enzyme purification with ExoSap (GE Healthtech).

PCR amplifications were performed in a final volume of 10 µL. Each reaction mixture was devised as follows: 1 ng template DNA, 5 pmoles of each primer, 200 µM each nucleotide (dNTP- PROMEGA) and 0.8 units of *Taq*DNA polymerase High-fidelity (Invitrogen) in a buffer solution containing 10 mM Tris-HCl, pH 8.5, 50 mM KCl, 1.5 mM $MgCl_2$.

Samples were amplified in a thermocycler (Programmable Thermal Controller, MJ Research Inc, model PTC-100) according to the following cycle program: an initial cycle consisting of a denaturation step of 3 min at 95°C, an annealing step of 1 min at 50ºC and an extension step of 30 s at 72ºC followed by 34 cycles of 1 min at 95°C (denaturation), 1 min at 50°C (annealing) and 1 min at 72°C (extension), and a final extension step of 7 min at 72ºC. After obtaining nucleotide sequences, specific primers were designed (Table I).

These primers amplified four *B. tenagophila* mitochondrial DNA fragments: COI-16S, 16S-Cyb, Cyb-12S and 12S-COI. This ensured that every mtDNA nucleotide was included.

The thermocycler program involved 30 cycles: an initial cycle containing a denaturation step of 1 min at 94°C and an annealing and extension step of 15 min at 68ºC, followed by 29 cycles of 15 s at 95°C (denaturation) and 15 min at 68°C (annealing and extension) [Long PCR (Cheng et al. 1994)].

Molecular cloning of PCR products (COI-16S, 16S-Cyb, Cyb-12S, 12S-COI) - Molecular fragments generated by PCR were cloned in different plasmids according to size: the 2,200 bp fragment (COI-16S) was cloned in TOPO TA cloning (Invitrogen - Life Technologies), the 2,800 bp fragment (Cyb-12S) was cloned in pGEM-T EASY (Promega) and the fragments of 4,500 bp (16S-Cyb e 12S-COI) in TOPO XL PCR cloning (Invitrogen - Life Technologies). The protocol followed the manufacturer's instructions. The TOPO 10 F` strain from *Escherichia coli* bacteria was transformed with recombinant plasmids.

Genome assembly and sequencing analysis - The shotgun sequencing reads obtained from 521 DNA templates, including those obtained using the 54 primers specifically designed for primer walking were analysed using PHRED/PHRAP/CONSED (Ewing et al. 1998, Ewing & Green 1998, Gordon et al. 1998) in order to read sequencer trace data, write the base calls and quality values, assemble shotgun DNA sequence data and to visually inspect, edit and finish the alignments. A PHRED quality value of 20, which corresponds to an error probability of 1/100 in the base call at each position of the read and a minimal overlap of 50 bp with 92% identity were used as initial constraints in the clustering for genome assembly.

Tandem Repeat Finder software was used to find and analyse repeats in the *B. tenagophila* assembled genome. Repeat Masker (AFA Smit & P Green, unpublished observations) together with Repbase databank (http://www.girinst.org) were used to identify known repeats and mask low complexity regions in the mollusc genome.

Analyses of the *B. tenagophila* assembled genome using different algorithms such as BLAST and FASTA were carried out using different databases including: the protein database from National Center for Biotechnology Information (NCBI), the database of Clusters of Orthologous Groups of proteins (http://www.ncbi.nlm.nih.gov/COG/), a simple modular architecture research tool (SMART) (http://smart.emblheidelberg.de/), protein families database of alignments and HMMs (http://www.sanger.ac.uk/Software/Pfam/) and Gene Ontology (http://www.geneontology.org/). Locally compiled mollusca and ribosomal RNA DBs were built by parsing public databases.

The annotation and graphical output of the *B. tenagophila* genomic sequence was performed with an ARTEMIS (http://www.sanger.ac.uk/Software/Artemis/) and in-house developed PERL scripts to analyse and format the results. All *in silico* searches and analyses described above, together with specific pipelines and PERL scripts developed during this work, were run in a local server using in-house copies of databases and software.

The search for tRNA genes in genomic sequences was performed using a combinatorial approach using *ab initio* predictions made by tRNAscan-SE software together with similarity searches against a locally compiled tRNA database.

The visualisation of comparative genome analysis between *B. tenagophila* and *B. glabrata* was performed using ACT software (http://www.sanger.ac.uk/Software/).

TABLE I

Specific primers

Primers	Sequences
16SL1	5'GCACAGTCAAGGTACTGCGGCCATTTAAAAAAAATATCACTGG
16SL2	5'AATCTCAAATGTATCCTGCCCAGTGATATTTTTTAAATGG
COIL2	5'GGCCTGGTTGGAACTGGGTTATCTTTACTTATTCGTTTAGAGC
COIR4	5'CCAAGCTCTAAACGAATAAGTAAAGATAACCCAGTTCCAACCAG
CybLF	5'AAAGGAAGAACTAAGCCTTTAGGAGATTTAAACCATTTGG
CybLR	5'TTAAATGGGATTTTTTCCAAATGGTTTAAATCTCCTAAAGG
12SLR	5'ATAATTAATGAATTTCTTTGATTTTGTGCACACATTGCCC
12SLF	5'AAAGAAATTCATTAATTATATATTTTTTAGTTACTTTTAAGTCC

Phylogenetic inference and genome analysis - Putative gene products of the assembled genome were compared to the available complete proteomes of other mitochondrial genomes together and with general protein DBs such as to the non-redundant protein data set from NCBI in order to investigate the evolutionary relationships among them.

Clustal W (Thompson et al. 1994) and MAFFT (Katoh & Toh 2008) were used to perform a global multiple sequence alignment between the structural features annotated in the predicted genomes and proteomes of *B. tenagophila* and *B. glabrata*. Analysis reports showing the statistics for the multiple alignments were obtained through the use of Alistat software (Krogh et al. 1994).

The methods used for phylogenetic reconstruction, using protein amino acid sequences (except for ATP8), were MP, NJ and ML. The MP and NJ were performed using algorithms provided in Phylogenetic Analysis Using Parsimony (PAUP v. 4.0) (Swofford 2002). MP analyses were carried out through heuristic searches by rearranging the branches. This method involves moving branches in order to achieve new topologies through a new reorganization. This procedure was performed by sectioning the tree into two pieces and connecting them to other parts of the tree in a process named "tree bisection and reconnection". In total, 10,000 random replications were created. This process was carried out with the characters with unroots and unweights and the gaps introduced in the alignment process being treated as data errors and topological restrictions were not applied. In order to test the reliability degree of the phylogenetic trees, the bootstrap method (Felsenstein 1985) was used with 10,000 pseudo-replications. For genetic distance analysis, NJ (Saitou & Nei 1987) was generated by the model Kimura-2 parameter, including 10,000 pseudo-replication bootstraps. The ML method, using the JTT model (Whelan & Goldman 2001) of amino acid evolution with Gama variation rate, was carried out with PHYML software (Guindon & Gascuel 2003) by aligning protein amino acid sequences (except for ATP8). In order to test the reliability degree of phylogenetic trees, bootstrap values (Felsenstein 1985) with 100 replications were used. As the outgroup, the mollusc species *Katharina tunicata* (Polyplacophora) was included in phylogenetic analyses.

Table II shows molluscs used in the phylogenetic study.

RESULTS

mtDNA content - *B. tenagophila* mtDNA was completely sequenced and characterised. The genome size encompasses 13,722 bp. This genome codes for two rRNAs, 13 protein genes and a set of 22 tRNAs. The proportion of AT was 75.78%. Nine non-coding regions ranging from 1-54 bp consisted of a total of 110 bp. The largest non-coding regions were located between the NAD1 and NAD4L genes and between COIII and $tRNA^{Ile}$, consisting of 54 and 42 bp, respectively. Nineteen overlapping regions were found in *B. tenagophila* mtDNA and the largest overlap was found between NAD4L and Cytb genes comprising 29 bp.

The tRNAscan software identified six out of the 22 tRNA (Asp, Asn, Thr, Val, Ala, Pro). The other tRNAs were identified by nucleotide alignment with its correspondent in *B. glabrata* and *B. tenagophila* tRNA had an average size of 63 bp.

The genes found in the mitochondrial genome coding for proteins were analysed based on the mitochondrial genetic code of invertebrates. Most protein-coding genes were located on the plus strand (NAD5, NAD1, NAD4L, Cytb, COII, NAD4, NAD2, COI, NAD6). Genes coding the ATP6, ATP8, NAD3 and COIII proteins were located on the minus strand.

Regarding protein size, COII, NAD2, NAD4 and NAD4L were longer in *B. tenagophila* than in *B. glabrata*. The protein with the largest size difference between the two species proteins was NAD1, which was 11 additional amino acids shorter in *B. glabrata*. The sizes of NAD1 and NAD6 were smaller in *B. tenagophila* than in *B. glabrata*. For ATP8, ATP6, NAD2, NAD4 and COIII, the termination codon is likely completed by adding an A during mRNA transcription. The least conserved protein was ATP8, with 60% identity in the two species and the most conserved protein was COI, with 98% identity between these two species.

Four mtDNA initiation codons were found in *B. tenagophila*, namely: ATT (in 5 proteins), ATA (in 5 proteins), ATG (in 2 protein) and TTG (in 1 protein). The stop codons were TAA (in 6 proteins) and TAG (in 2 proteins), which have been described for mitochondrial genes (Wolstenholme 1992). Incomplete stop codons were found in five proteins (ATP6, ATP8, COI, COII, NAD3, NAD4).

TABLE II
Molluscs used in the phylogenetic study

Mollusc	Accession (GenBank)	Abbreviation
Lampsilis ornata	AY365193	Lam orn
Haliotis rubra	AY588938	Hal rub
Lophiotoma cerithiformis	DQ284754	Lop cer
Nautilus macromphalus	DQ472026	Nau mac
Lottia digitalis	DQ238599	Lot dig
Ilyanassa obsoleta	DQ238598	Ily obs
Aplysia californica	AY569552	Apl cal
Octopus ocellatus	AB240156	Oct oce
Crassostrea virginica	AY905542	Cra vir
Mytilus trossulus	AY823625	Myt tro
Watasenia scintillans	AB240152	Wat sci
Sepioteuthis lessoniana	AB240154	Sep les
Sepia officinalis	AB240155	Sep off
Roboastra europaea	AY083457	Rob eur
Placopecten magellanicus	DQ088274	Pla mag
Mytilus galloprovincialis	AY497292	Myt gal
Albinaria coerulea	X83390	Alb coe
Todarodes pacificus	AB158364	Tod pac
Octopus vulgaris	AB158363	Oct vul
Biomphalaria glabrata	AY380531	Bio gla
Graptacme eborea	AY484748	Gra ebo
Mytilus edulis	AY484747	Myt edu
Siphonodentalium lobatum	AY342055	Sip lob
Crassostrea gigas	AF177226	Cra gig
Cepaea nemoralis	U23045	Cep nem
Loligo bleekeri	AB029616	Lol ble
Pupa strigosa	AB028237	Pup str
Katharina tunicata	U09810	Kat tun

Ribosomal RNAs 16S and 12S comprised 981 and 708 nucleotides, respectively. Sequence identity found between *B. glabrata* and *B. tenagophila* was 88% for rRNA 16S and 84% for rRNA 12S.

The same mitochondrial gene order was found in both *B. tenagophila* and *B. glabrata* (Fig. 1).

Phylogenomic analysis - mtDNA amino acid sequences of 29 organisms from the Mollusca phylum were aligned (except for ATP8) and used for phylogenomic positioning analysis of *B. tenagophila* and *B. glabrata* in comparison with other molluscs.

Regardless the phylogenetic methodology, the trees showed similar topology, comprising five groups. Group I: MP bootstrap, 94% (Fig. 2), ML bootstrap, 99% (Fig. 3) encompassed six molluscs of the Bivalvia class (Protobranchia sub class, Pteriomorphia super order), one species of the Gastropoda class (Prosobranchia) and one species of the Scaphopoda class (Gadelida). In the NJ tree (Fig. 4), this group clustered the same molluscs when compared with MP and ML trees, except for the species belonging to the Scaphopoda class (Gadelida) that are located in Group III with a bootstrap value of 100%. Group II comprised the Gastropoda class in all tree topologies, including Pulmonata and Opistobranchia sub-classes, which was supported by a bootstrap value of 100%. Group III was located in a basal position in relation to Groups I and II. In MP and ML trees, one species of the Scaphopoda class (Dentaliida) and another species of the Bivalvia class (Protobranchia sub class, Heterodonta super order) were closely positioned. In the NJ tree, three species were closely positioned, being two species of the Scaphopoda class (Dentaliida and Gadelida) and the same species of the Bivalvia class (Protobranchia sub class, Heterodonta super order). This group was supported by the following bootstrap values: 97% (NJ), 94% (ML) and 74% (MP). Group IV included species of the Cephalopoda class in all tree topologies, supported by a high bootstrap value of 100%. Group V comprised species of the Gastropoda class (Protobranchia sub class) in all tree topologies, supported by 100% bootstrap values in both NJ and ML and 82% in MP.