Genetics and Molecular Research

Analysis of the mitochondrial *COI* gene and its informative potential for evolutionary inferences in the families Coreidae and Pentatomidae (Heteroptera)

H.V. Souza[1], S.R.C. Marchesin[2] and M.M. Itoyama[1]

[1]Laboratório de Citogenética e Molecular de Insetos, Departamento de Biologia, Instituto de Biociências, Letras e Ciências Exatas, Universidade Estadual Paulista, São José do Rio Preto, SP, Brasil
[2]Universidade Paulista, Campus Juscelino Kubitschek São José do Rio Preto, SP, Brasil

Corresponding author: H.V. Souza
E-mail: souzahv@gmail.com

Genet. Mol. Res. 15 (1): gmr.15017428
Received August 11, 2015
Accepted October 7, 2015
Published February 5, 2016
DOI http://dx.doi.org/10.4238/gmr.15017428

ABSTRACT. The mitochondrial cytochrome oxidase subunit 1 (*COI*) gene is one of the most popular markers used for molecular systematics. Fragments of this gene are often used to infer phylogenies, particularly the region near the 5'-end, which is used by the DNA Barcoding Consortium. With a growing number of sequences being deposited in the DNA barcoding database, there is an urgent need to understand the evolution of this gene and its evolutionary relationship among species; it is also important to analyze the informative potential of the gene for phylogenetic inferences for each group used. In this study, the *COI* gene was divided into three distinct regions: a 5'-region, a central region, and a 3'-region. The nucleotide composition of these regions was analyzed, and their potential for making informative phylogenetic inferences using species in the families Coreidae and Pentatomidae (Heteroptera) was assessed. It was found that the same

region in the *COI* gene may present different behaviors for each family analyzed, and that using additional regions from the same gene may even prejudice the analysis.

Key words: COI; Mitochondrial DNA; Heteroptera; Phylogeny; Molecular systematics

INTRODUCTION

The mitochondrial cytochrome oxidase subunit 1 (*COI*) gene is one of the most popular markers used in molecular systematics. Portions of this gene are often used to infer phylogenies. In addition, *COI* is currently the focus of considerable interest, especially its 5' portion, which is used by the DNA Barcoding Consortium (Hebert et al., 2003; Stoeckle, 2003). This region is comprised of approximately 640 nucleotides (Folmer et al., 1994) and has been used as a unique identification code for many species. Additionally, this region is used to facilitate the correct identification of specimens and the discovery of new species (Moritz and Cicero, 2004).

Several studies indicate that this segment could be effectively used as a DNA barcode to identify various species of animals, such as butterflies or fish (Hebert et al., 2003; Ward et al., 2009). An increasing number of sequences are being deposited in the DNA barcoding database (Monaghan et al., 2006); thus, there is an urgent need to understand the evolutionary pattern of this gene among species. It is also important to analyze the informative potential of the gene for inferring phylogenies for each group studied. These studies have even suggested changing the segment used for the DNA barcoding initiative; such suggestions include using the region encoding the *COI-COII* genes or using an additional region of the same gene, such as the I3-M11 portion in the species in which this gene is conserved. Other studies have suggested maximizing the rate of nucleotide divergence or its evolutionary consistency rate to improve accuracy of the DNA barcoding as a marker of mtDNA divergence and the robustness of the phylogenetic signal. Ideally, a region should be chosen to simultaneously maximize all of these factors; however, some tradeoffs will be inevitable (Erpenbeck et al., 2006; Roe and Sperling, 2007).

Nucleotide substitution patterns may play an important role in obtaining information about phylogenetic relationships and the population structure of different organisms, and such variations may show increased diversification among independent lineages and increasingly divergent taxa (Galtier et al., 2005).

In addition, molecular characteristics originally used in the identification or classification of a species, such as those deposited in DNA barcoding databases, are often subsequently incorporated in studies of phylogenetic approaches and/or phylogeography. Therefore, it is important to study how variations in phylogenetic information result in changes at more inclusive taxonomic levels (Galtier et al., 2005).

Phylogenetically informative characteristics gradually accumulate in the form of polymorphisms in diverging lineages, and the more divergent the lineage, the more these characteristics lead to increased phylogenetic information (Galtier et al., 2005). However, simultaneously with this, the probability of multiple mutations at a particular site increases, resulting in the loss of informative characteristics (saturation) for a given taxonomic level, leading to erroneous evolutionary interpretations (Galtier et al., 2005; Schneider, 2007).

Among the more inclusive taxa, the *COI* gene is considered conserved among Metazoa (Jacobs et al., 1988) and has been used to analyze phylogenetic relationships within the

Arhynchobdellida order (Borda and Siddall, 2004) and Pteriomorphia subclass (Matsumoto, 2003). Nevertheless, it is often used to distinguish between families and less inclusive taxonomic lineages (Pollock et al., 1998; Damgaard and Sperling, 2001).

In the Heteroptera suborder, the *COI* gene, along with other genes and morphological characteristics, has been used to infer the evolution of some taxa, such as the infraorder (Pentatomomorpha), superfamily (Pentatomidae), family (Anthocoridae), and genus (*Limnogonus*) (Muraji et al., 2000; Li et al., 2005; Grazia et al., 2008; Damgaard et al., 2010).

In this study, the *COI* gene was divided into three distinct regions: a region near the 5'-end, which is used in the DNA barcoding consortium; a central region; and a region near the 3'-end, an underexplored region of the gene. Each of these regions was analyzed for nucleotide composition and informative potential for phylogenetic inferences using species in the families Coreidae and Pentatomidae (Heteroptera).

MATERIAL AND METHODS

In this study, we analyzed the species *Anasa bellator*, *Anisoscelis foliacea*, *Athaumastus haematicus*, *Catorhintha guttula*, *Chariesterus armatus*, *Dallacoris obscura*, *Dallacoris pictus*, *Hypselonotus fulvus*, *Leptoglossus gonagra*, *Leptoglossus zonatus*, *Sphictyrtus fasciatus*, and *Zicca annulata*, all belonging to the family Coreidae; and the species *Antiteuchus tripterus*, *Chlorocoris complanatus*, *Dichelops melacanthus*, *Edessa meditabunda*, *Euschistus heros*, *Loxa deducta*, *Mormidea v-luteum*, *Oebalus poecilus*, *Oebalus ypsilongriseus*, *Piezodorus guildinii*, *Platycarenus umbraculatus*, *Proxys albopunctulatus,* and *Thyanta perditor* belonging to the family Pentatomidae (Table 1).

Species were collected in São Jose do Rio Preto, SP, Brazil region (20° 47' 13" S, 49° 21' 38 " W) and were fixed in absolute ethanol. To obtain DNA fragments, thoracic muscles of the collected specimens were used according to the protocol described by Bargues and Mas-Coma (1997). - = not analyzed.

Two pairs of primers were used, based on their descriptions in the literature (Table 2). Due to the absence of specific primers for the 3'-end, three other primers were designed (V-VII) (Table 2) with the Geneious v.5.0 program using the complete sequences of the *COI* gene from the *Halyomorpha halys* and *Nezara viridula* species in the family Pentatomidae (accession Nos. NC_013272 and NC_011755) and *Hydaropsis longirostris* species in the family Coreidae (accession No. NC_012456).

Seven primers, described in Table 2, were used for the analysis of the *COI* gene. The annealing region of each primer is shown in Figure 1. These regions were named Region 1, which is near the 5'-end and contains approximately 650 bp; Region 2 (intermediate), which contains approximately 450 bp; and Region 3, which is near the 3'-end and contains nearly 400 bp (Figure 1).

It was not possible to amplify fragments for region 3 in the following species: *Leptoglossus gonagra* and *Hypselonotus fulvus* in the family Coreidae, and *Chlorocoris complanatus*, *Mormidea v-luteum*, and *Platycarenus umbraculatus* in the family Pentatomidae. Thus, for these mentioned species only regions 1 and 2 were analyzed.

Amplification of the gene regions was performed in a Veriti® 96-Well Thermal Cycler in a 10 µL-total reaction volume with 100 ng DNA, 0.2 mM each primer pair, 100 mM dNTP, 10 mM Tris-HCl, 1.5-3 mM $MgCl_2$, and 0.5 U Platinum® DNA polymerase (Invitrogen).

Sequencing reactions were performed using an ABI 3730 DNA Analyzer sequencer (Applied Biosystems). The sequencing reactions were performed using the BigDye® Terminator v3.1 Cycle Sequencing Kit. The sequences were analyzed using the Sequencing Analysis 5.3.1 software with the KB base caller.

Table 1. Classification of species in the families Coreidae and Pentatomidae and the CBI database accession Nos. used in this study.

Family	Sub-family	Tribe	Species	Accession Nos.		
				Região 1	*Região 2*	*Região 3*
Coreidae	Coreinae	Anisoscelidini	*Anisoscelis foliacea*	JQ037895	JQ218470	JQ218512
			Leptoglossus gonagra	JQ037902	JQ218477	-
			Leptoglossus zonatus	JQ037903	JQ218478	JQ218500
		Acanthocerini	*Athaumastus haematicus*	JQ037896	JQ218471	JQ218495
		Chariesterini	*Chariesterus armatus*	JQ037898	JQ218473	JQ218497
		Coreini	*Anasa bellator*	JQ031214	JQ218469	JQ218494
			Catorhintha guttula	JQ037897	JQ218472	JQ218496
			Hypselonotus fulvus	JQ037901	JQ218476	-
			Sphictyrtus fasciatus	JQ037904	JQ218479	JQ218501
			Zicca annulata	JQ037905	JQ218480	JQ218502
		Leptoscelini	*Dallacoris obscura*	JQ037899	JQ218474	JQ218498
			Dallacoris pictus	JQ037900	JQ218475	JQ218499
Pentatomidae	Discocephalinae	Discocephalini	*Antiteuchus tripterus*	JQ218456	JQ218481	JQ218503
			Platycarenus umbraculatus	JQ218466	JQ218491	-
	Edesinae	Edessini	*Edessa meditabunda*	JQ218459	JQ218484	JQ218507
	Pentatominae	Carpocorini	*Dichelops melacanthus*	JQ218458	JQ218483	JQ218506
			Euschistus heros	JQ218460	JQ218485	JQ218504
			Mormidea v-luteum	JQ218462	JQ218487	-
			Oebalus poecilus	JQ218463	JQ218488	JQ218509
			Oebalus ypsilongriseus	JQ218464	JQ218489	-
		Chlorocorini	*Chlorocoris complanatus*	JQ218457	JQ218482	-
			Loxa deducta	JQ218461	JQ218486	JQ218508
		Pentatomini	*Proxys albopunctulatus*	JQ218467	JQ218492	JQ218505
			Thyanta perditor	JQ218468	JQ218493	JQ218511
		Piezodorini	*Piezodorus guildinii*	JQ218465	JQ218490	JQ218510

Table 2. Primers used in the analysis of the cytochrome oxidase I gene in species of the families Coreidae and Pentatomidae (Heteroptera).

Region			Sequence (5'→3')	Reference
1	I	Forward	TTTCAACAAATCATAAAGATATTGG	Folmer et al. (1994)
	II	Reverse	TAAACTTCAGGGTGACCAAAAAATCA	Folmer et al. (1994)
2	III	Forward	AGCAGGAATTTCATCAATTTT	Muraji et al. (2000)
	IV	Reverse	CTGTAAATATGTGATGTGCTC	Muraji et al. (2000)
3	V	Forward	RTTGGAGGATTAACAGGAGTAA	Present study
	VI	Forward	TTATGAATGTTCTGMTGGNGG	Present study
	VII	Reverse	GGAGTAATTCTAGCCAACTC	Present study

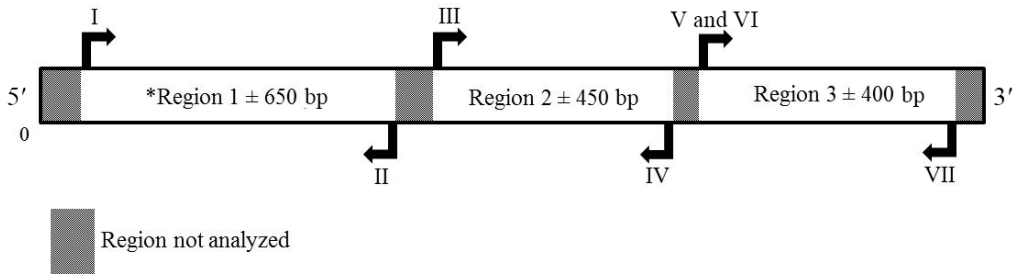

Figure 1. Regions of the cytochrome oxidase I gene, which have an average length of 1.574 bp. Primers I-VII are described in Table 2. *Region used for DNA barcoding.

The three amplified gene regions were aligned with the ClustalW program (Thompson et al., 1997) and analyzed in the MEGA5 program (Tamura et al., 2011). Distance analyses were performed using the PAUP4 program (Swofford, 2000), and the mean, standard deviation and minimum and maximum values were determined using the Minitab® Statistical Software.

Neighbor-joining phylogenetic inferences were performed using the K2P distance model in the MEGA5 program (Tamura et al., 2011), and Bayesian inferences were performed with the MrBayes v3.2 program (Ronquist et al., 2012). Four unrooted phylogenetic relationships were generated and analyzed for each family while taking into account the analysis of the nucleotides in regions 1, 2, and 3.

RESULTS

The complete sequences of the cytochrome oxidase I genes found in the literature include 1533, 1541, and 1536 bp for *Hydaropsis longirostris* (Coreidae) (accession No. NC_012456), *Halyomorpha halys* (accession No. NC_013272), and *Nezara viridula* (Pentatomidae) (accession No. NC_011755), respectively, with an average of 1537 bp for each of the three species. In the analysis of the three gene regions, 20 sequences (one for each species) were obtained with an average of 1443 bp (93.9% of the gene) belonging to the *Anasa bellator*, *Anisoscelis foliacea*, *Athaumastus haematicus*, *Catorhintha guttula*, *Chariesterus armatus*, *Dallacoris obscura*, *Dallacoris pictus*, *Hypselonotus fulvus*, *Leptoglossus gonagra*, *Leptoglossus zonatus*, *Sphictyrtus fasciatus*, and *Zicca annulata* species in the family Coreidae and the *Antiteuchus tripterus*, *Chlorocoris complanatus*, *Edessa meditabunda*, *Euschistus heros*, *Loxa deducta*, *Mormidea v-luteum*, *Oebalus poecilus*, *Oebalus ypsilongriseus*, *Platycarenus umbraculatus*, and *Thyanta perditor* species in the family Pentatomidae (Table 1). After the alignment was performed, 1379 nucleotides, including 625, 386, and 368, belonged to regions 1, 2, and 3, respectively, in the cytochrome oxidase I gene in the family Coreidae, and 1378 nucleotides, including 621, 393, and 364 belonged to regions 1, 2, and 3, respectively, in the family Pentatomidae.

When we performed an analysis of the nucleotide variation in the 3 codon sites, the 3rd codon position demonstrated conservation ranging from 6.2 to 10.9% for these sites in the family Coreidae (Table 3) and from 5.5 to 9.6% in the family Pentatomidae (Table 4). In the 1st and 2nd positions, the conservation levels ranged from 41.8 to 51.5% in the family Coreidae (Table 3) and from 40.9 to 53.6% in the family Pentatomidae (Table 4).

Table 3. Analysis of the nucleotide composition, amino acid residues, and divergences obtained for the three analyzed regions of the cytochrome oxidase I gene in species belonging to the family Coreidae.

	Region 1			Region 2			Region 3		
	Conserved	Variables		Conserved	Variables		Conserved	Variables	
Nucleotide composition									*Total*
1st Position	164 (42.1%)	44 (18.7%)		106 (43.4%)	25 (17.6%)		99 (41.8%)	24 (18.3%)	
2nd Position	201 (51.5%)	8 (3.4%)		123 (50.4%)	5 (3.5%)		112 (47.3%)	11 (8.4%)	
3rd Position	25 (6.4%)	183 (77.9%)		15 (6.2%)	112 (78.9%)		26 (10.9%)	96 (73.3%)	
Total	390 (62.4%)	235 (37.6%)		244 (63.2%)	142 (36.8%)		237 (64.4%)	131 (35.6%)	1379
Amino acid residues			Total			Total			
Total	181 (87.4%)	26 (12.6%)	207	114 (89.1%)	14 (10.9%)	128	97 (79.5%)	25 (20.5%)	122
AT%	Mean	Min.-max.		Mean	Min.-max.		Mean	Min.-max.	
1st Position	54.2	51.2-58.7		56.6	53.5-58.9		67.7	64.2-69.9	
2nd Position	54.9	54.5-55.5		59.5	58.9-61.2		64.7	63.4-65.9	
3rd Position	81.7	71.5-88.0		85.3	78.9-93.0		83.5	75.4-91.8	
Total	63.6			67.1			71.9		
Divergence (K2P distance)									
Mean	19.9%			17.4%			17.3%		
Standard deviation	3.2%			3.2%			3.8%		
Min.-max.	0.3-24.2%			0.3-23.4%			0.5-24.1%		

Mean = average, Min.-max. = minimum and maximum.

Table 4. Nucleotide composition analysis, amino acid residues, and divergences obtained for the three analyzed regions of the cytochrome oxidase gene I of the species belonging to family Pentatomidae.

	Region 1			Region 2			Region 3		
	Conserved	Variables		Conserved	Variables		Conserved	Variables	
Nucleotide composition									Total
1st Position	149 (40..9%)	58 (22.6%)		104 (41.8%)	26 (18.0%)		93 (42.5%)	29 (20.0%)	
2nd Position	195 (53.6%)	12 (4.7%)		125 (50.2%)	7 (4.9%)		105 (47.9%)	16 (11.0%)	
3rd Position	20 (5.5%)	187 (72.7%)		20 (8.0%)	111 (77.1%)		21 (9.6%)	100 (69.0%)	
Total	364 (58.6%)	257 (41.4%)		249 (63.4%)	144 (36.6%)		219 (60.2%)	145 (39.8%)	1378
Amino acid residues			Total			Total			
Total	160 (77.7%)	46 (22.3)%)	206	109 (83.8%)	21 (16.2%)	130	84 (70.0%)	36 (30.0%)	120
AT%	Mean	Min.-max.		Mean	Min.-max.		Mean	Min.-max.	
1st Position	55.0	53.4-56.9		59.2	56.6-61.5		67.2	66.1-68.6	
2nd Position	55.2	54.6-56.0		61.8	60.8-63.1		64.0	61.2-65.8	
3rd Position	82.1	74.4-88.9		83.0	77.9-90.8		82.4	75.2-87.6	
Total	64.1			68.0			71.2		
Divergence (K2P distance)									
Mean	19.3%			16.3%			18.8%		
Standard deviation	3.2%			2.6%			3.4%		
Min.-max.	13.9-27.2%			10.4-21.4%			9.8-26.5%		

Mean = average, Min.-max. = minimum and maximum.

Regarding the variable sites, 17.6 to 18.7% (Coreidae) and 18.0 to 22.6% (Pentatomidae) of these sites were located within the 1st position of the codon and 3.4 to 8.4% (Coreidae) and 4.7 to 11.0% (Pentatomidae) were in the 2nd codon position (Tables 3 and 4). In the 3rd codon position, the variations ranged from 73.3 to 78.9% (Coreidae) and 69.0 to 77.1% (Pentatomidae) for informative sites (Tables 3 and 4).

In this analysis, we found that region 3 was the most conserved in the family Coreidae, with 64.4% of the sites demonstrating conservation, whereas in the family Pentatomidae, region 2 demonstrated the highest number of conserved sites with 63.4% of sites showing conservation. Region 1 had the highest percentage of variability in both families: 37.6% in Coreidae and 41.4% in Pentatomidae.

The percentage of conserved amino acid residues was 87.4, 89.1, and 79.5% for gene regions 1, 2, and 3, respectively, in the family Coreidae (Table 3), and 77.7, 83.8, and 70.0%, respectively, in the family Pentatomidae (Table 4). Thus, the most conserved gene region in both families was region 2 with regard to the amino acid residues.

At the third position, there was an average frequency of AT nucleotides of 81.7, 85.3, and 83.5% for regions 1, 2, and 3, respectively, in the family Coreidae (Table 3) and 82.1, 83.0, and 82.4% for regions 1, 2, and 3, respectively, in the family Pentatomidae (Table 4). In the 1st and 2nd positions of region 1, the AT percentage was 54.2 and 54.9%, respectively, for the family Coreidae and 55.0 and 55.2%, respectively, for the family Pentatomidae. In region 2, the AT values were 56.6 and 59.5% for the 1st and 2nd positions, respectively, in the family Coreidae, and 59.2 and 61.8% for the 1st and 2nd positions, respectively, in the family Pentatomidae. In region 3, frequencies of 67.7 and 64.7% in Coreidae and frequencies of 67.2 and 64.0% in the family Pentatomidae were found for the 1st and 2nd positions, respectively. We found that region 1 demonstrated a lower average AT content compared with the other two regions, with contents of 63.6 and 64.1% for the Coreidae and Pentatomidae families, respectively, while the third region demonstrated the highest average percentage of AT nucleotides in both of the analyzed families with percentages of 71.9 and 71.2% for the Coreidae and Pentatomidae families, respectively.

The average divergences obtained using distance matrices (K2P) were 19.9, 17.4, and 17.3 for regions 1, 2, and 3, respectively, for the family Coreidae (Table 3) and 19.3, 16.3, and 18.8% for the family Pentatomidae (Table 4).

In a neighbor-joining phylogenetic analysis of the family Coreidae, we observed paraphyly of the Coreini and Anisoscelidini tribes when analyzing the first and second regions of the cytochrome oxidase 1 gene (Figures 2A and B) and both of the concatenated regions (Figure 2C). When regions 1, 2, and 3 were concatenated, there was paraphyly of only the Anisoscelidini tribe (Figure 2D). There was inconsistent clustering for all analyses except for the *Dallacoris* genus, which demonstrated high bootstrap levels. When the first and second regions were concatenated, we observed an evolutionary scenario similar to that demonstrated for region 1, and there was a decrease in the bootstrap rate for most branches (Figure 2C). Concatenation of region 3 with regions 1 and 2 led to a slight increase in the bootstrapping of some of the branches, and the *Anisoscelis foliacea* species was clustered with *Athaumastus haematicus*, recovering the monophyly of the Coreini tribe (Figure 2D).

A Bayesian inference of the family Coreidae paraphyly of the Coreini and Anisoscelidini tribes was visualized for all analyses and involved region 1 (Figure 3A), region 2 (Figure 3B), concatenated regions 1 and 2 (Figure 3C), and concatenated regions 1, 2 and 3 (Figure 3D). In these four studied inferences, different evolutionary scenarios were formed when they were compared with each other, and only the branch involving species for the *Dallacoris* and *Leptoglossus* genera remained unchanged in all the analyses.

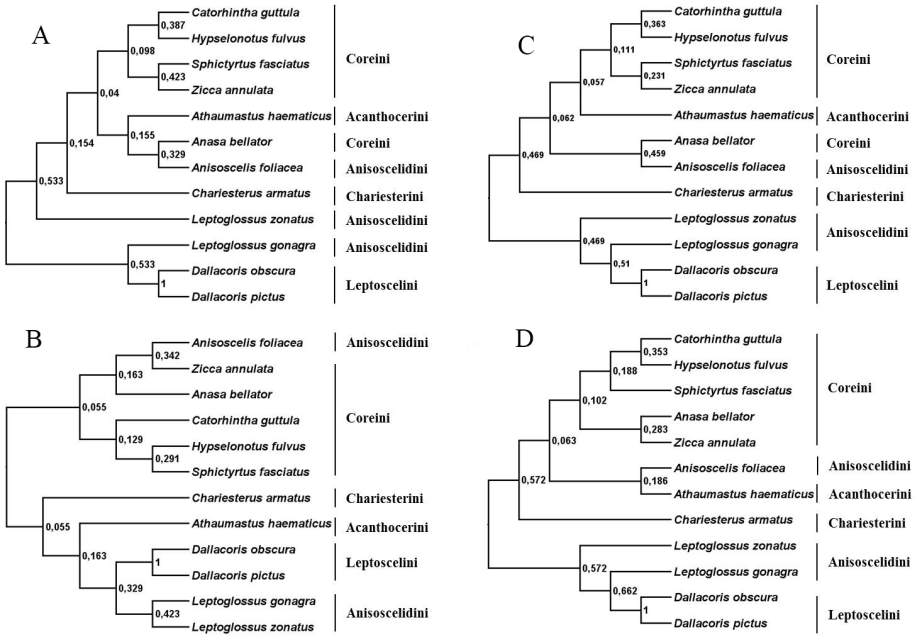

Figure 2. Neighbor-joining phylogenetic analysis of regions of the cytochrome oxidase 1 gene in species in the Coreidae family, using Kimura 2-parameter distance. **A.** region 1; **B.** region 2; **C.** concatenated regions 1 and 2; and **D.** concatenated regions 1, 2 and 3.

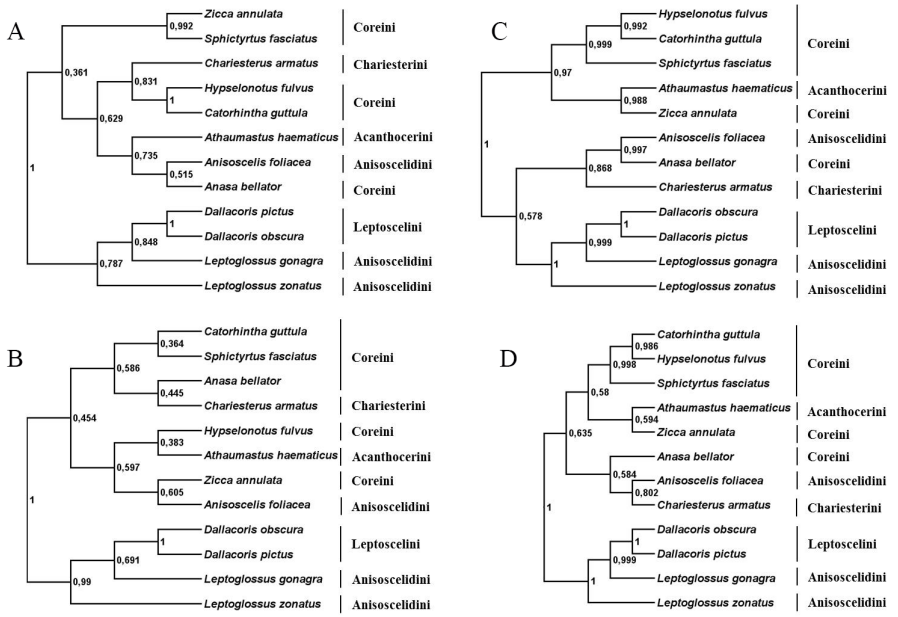

Figure 3. Bayesian inference using nucleotides from regions of the cytochrome oxidase 1 gene in species of the Coreidae family. **A.** region 1; **B.** region 2; **C.** concatenated regions 1 and 2; and **D.** concatenated regions 1, 2 and 3.

A neighbor-joining phylogenetic analysis of the cytochrome oxidase gene region 1 in the family Pentatomidae revealed paraphyletic clusters containing the tribes Chlorocorini and Discocephalini (Figure 4A). When analyzing the region 2 of the same gene, similar to the tribes Chlorocorini and Discocephalini, the tribes Carpocorini and Pentatomini formed paraphyletic clusters. Additionally, the *Oebalus* genus did not form a monophyletic cluster (Figure 4B). When concatenating these two regions, it was observed that the tribes Discocephalini, Pentatomini, and Chlorocorini were paraphyletic (Figure 4C), and the Carpocorini tribe became paraphyletic as well (Figure 4D). In all analyses, clusters were formed with low consistency in the branches, with the exception of the genus *Oebalus*, when analyzing the three concatenated gene regions.

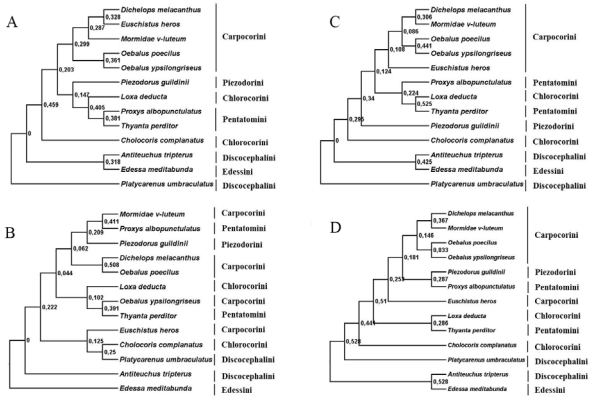

Figure 4. Neighbor-joining phylogenetic analysis of regions of the cytochrome oxidase 1 gene in species of the Pentatomidae family using Kimura 2-parameter distance. **A.** region 1; **B.** region 2; **C.** concatenated regions 1 and 2; and **D.** concatenated regions 1, 2 and 3.

For the Bayesian inference of Pentatomidae species using region 1 of the gene, it was found that only the tribe Carpocorini was paraphyletic (Figure 5A). However, only region 2 led to paraphyly for the tribes Carpocorini, Chlorocorini, Pentatomini, and Discocephalini (Figure 5B). When regions 1 and 2 were concatenated, paraphyly remained for these tribes (Figure 5C); this paraphyly also remained following the addition of region 3 in the analysis (Figure 5D).

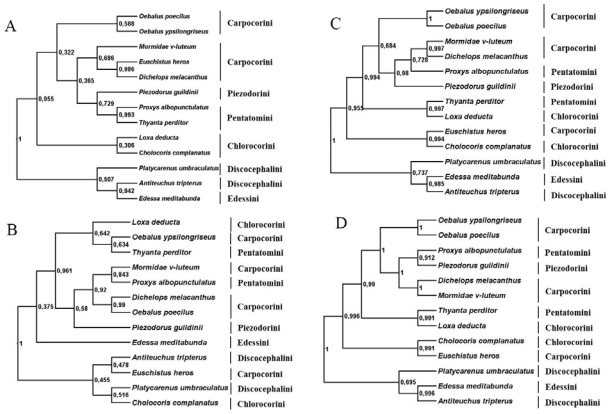

Figure 5. Bayesian inference using nucleotides from regions of the cytochrome oxidase 1 gene in species in the Pentatomidae family. **A.** region 1; **B.** region 2; **C.** concatenated regions 1 and 2; and **D.** concatenated regions 1, 2 and 3.

DISCUSSION

Segments currently used for DNA barcoding can provide different levels of conservation for each group examined. Erpenbeck et al. (2006) suggested the inclusion of an additional *COI* gene region in the DNA barcoding analyses of sponges because the *COI* gene is more conserved in this group, precluding accurate species identification. In our analysis, we found that there is homogeneity when analyzing the conserved sites in three distinct gene regions, ranging from 62.4 to 64.4% for the family Coreidae and from 58.6 to 63.4% for the family Pentatomidae. As expected, the most conserved sites were mainly found in the 1st and 2nd codon positions; the 3rd position had few conserved sites with a maximum homogeneity of 10.9%. In general, by analyzing amino acid residues, we found that there was a high level of conservation in both families, with the second region having a higher conservation level. This finding means that there are few variable sites for evolutionary inference in the Coreidae and Pentatomidae families. Region 3, which is more variable, possessed more informative residues (20.5 and 30.0% for Coreidae and Pentatomidae, respectively). Unfortunately, this variability may be an obstacle for designing primers, which is the case for *Leptoglossus gonagra*, *Hypselonotus fulvus* (Coreidae), *Chlorocoris complanatus*, *Mormidea v-luteum*, and *Platycarenus umbraculatus* (Pentatomidae).

Higher AT nucleotide frequency has been reported to be a common feature for insect mitochondrial DNA (Simon et al., 1994). In our analysis, the AT frequency at the third site of each codon was higher than that in the first two. This difference may be due to the degeneracy of the third codon position, which is characteristic of the genetic code, and this may lead to saturation at this codon position. When analyzing the cytochrome oxidase I gene divided into three regions, we found an increase in the AT percentage at the 1st and 2nd codon positions of region 3 in the Coreidae and Pentatomidae families. For the species belonging to the genus *Ips* (Coleoptera: Scolytidae), as described in the study by Cognato and Sperling (2000), saturation was higher in the third codon of the cytochrome oxidase I gene followed by the 1st and 2nd codons. However, in this genus, the third codon had the most phylogenetically informative sites; therefore, the authors used this position to infer phylogenetic relationships. Nevertheless, nucleotide substitution saturation may decrease the potential of the phylogenetically informative characteristics, thus decreasing phylogenetic consistency (Cognato and Sperling, 2000).

In fact, in our analysis, we found that there is informative potential for evolutionary inferences using the third codon, with a frequency of site variation of approximately 70% in both families. However, the higher AT nucleotide frequency in this codon may lead to inconsistent phylogenetic analysis in higher taxa, and it was found that just over 80% of the nucleotides in this position are thymine and adenine bases. To avoid this bias, some authors preferred to use the translated amino acid code, but when translated into amino acids, phylogenetic analyses can produce unresolved trees, as verified by Cognato and Sperling (2000). On average, region 3 had a higher percentage of AT nucleotides.

Saturation can occur in species that diverged earlier in their evolutionary histories. In a comparison of pairwise divergence among the studied species (K2P distance), it appears that region 1 had a higher average divergence than the other regions, with frequencies of 19.9% for Coreidae and 19.3% for Pentatomidae; however, in general, we found that the average divergence (K2P) in both families for the three regions analyzed did not exceed 20%. Previous studies of various animal groups have reported a minimum distance for interspecific sister species of less than 2% (Klicka and Zink, 1997; Johns and Avise, 1998; Hebert et al., 2004; Ward et al., 2005; Yoo et al., 2006). The average divergence for more inclusive taxa depends on the number of congeners

in the genera analyzed, i.e., the average divergence is lower when there are a greater number of species within the same genus (Jung et al., 2011). Our analyzed species belong to different tribes, and in the case of the family Pentatomidae, they belong to distinct subfamilies (Table 1). However, the differences are less pronounced and may reflect gene divergence at the family level for Heteroptera; more families should be analyzed with respect to the cytochrome oxidase I gene to define real divergence at this taxonomic level.

Although the results from the analysis of the families Coreidae and Pentatomidae were independently analyzed, the results were similar with regard to the nucleotide description and rate of divergence for both families, suggesting that *COI* evolves similarly for both families.

From the phylogenetic analysis obtained by the neighbor-joining method, using region 2 as a marker, phylogenetic inference using the Bayesian inference method yielded less paraphyletic clustering for the family Coreidae; however, the use of region 1 as a marker for phylogenetic inference yielded fewer paraphyletic clusters in the family Pentatomidae. In both cases, the robustness of the branches was not significantly improved when adding gene regions in analysis. In fact, the branches in the analysis demonstrated few bootstraps and were considered consistent, i.e., with frequencies less than 70% (Hillis and Bull, 1993). Likewise, the evolutionary scenarios obtained through Bayesian inference demonstrated paraphyly of the Anisoscelidini and Coreini tribes and consistently low branching, with the exception of the genera *Dallacoris* and *Leptoglossus*.

The results obtained with region 1 for the family Pentatomidae exhibited fewer paraphyletic clusters by the neighbor-joining analysis than by the Bayesian inference. However, Bayesian inference using region 1 formed a paraphyletic cluster only for the tribe Carpocorini, indicating its potential for evolutionary analyses. By adding another region of the same gene in this analysis, paraphyletic clusters were formed with tribes, beyond Carpocorini, demonstrating that, in this case, adding more regions of the same gene may negatively affect the analysis.

The low consistencies between the branches may be due to divergence of the species, which cannot recuperate the evolutionary history of the taxonomic family level. Therefore, it is essential to analyze other genes that are informative at the taxonomic level to recover the evolutionary history of these species by a robust method. Indeed, Li et al. (2005) suggested that the *COI* segment alone was not a suitable marker for the molecular phylogeny of Pentatomomorpha.

At the family level, region 1 (corresponding to DNA barcoding) rarely supported a satisfactory resolution (Brown et al., 1994; Miura et al., 1998) and was often demonstrated to be unreliable (Dowton and Austin, 1997; Mardulyn and Whitfield, 1999). While they are more inclusive in taxonomic levels, cytochrome oxidase I sequences are not suitable for resolving relationships (Liu and Beckenbach, 1992; Howland and Hewitt, 1995; Frati et al., 1997).

Similar to region 1, the three regions analyzed in this study were unable to resolve phylogenetic inferences alone in a robust manner for the analyzed families. Thus, this study demonstrates that regions of the same gene may exhibit different behaviors for each family and that the addition of other regions from the same gene may not be necessary or may negatively affect the analysis. The study of these gene segments should be viewed with caution because they are more randomized than informative, requiring analysis with other genes or morphological characteristics to form robust evolutionary scenarios for the reliable analysis of the evolutionary history of these two families.

Conflicts of interest

The authors declare no conflict of interest.

ACKNOWLEDGMENTS

We would like to give special thanks to Dr. Jocélia Grazia of the Department of Zoology of the Federal University of Rio Grande do Sul and Dr. Aline Barcellos of the Museum of Natural Sciences of Porto Alegre State, Brazil, who helped with specimen identification. Research financially supported by the São Paulo Research Foundation - FAPESP (grants #2010/16080-5 and #2011/11054-9), the Foundation for the Development of the São Paulo State University (FUNDUNESP), and the National Council for Scientific and Technological Development (CNPq).

REFERENCES

Bargues MD and Mas-Coma S (1997). Phylogenetic analysis of Lymnaeid snails based on 18S rDNA sequences. *Mol. Biol. Evol.* 14: 569-577.http://dx.doi.org/10.1093/oxfordjournals.molbev.a025794

Borda E and Siddall ME (2004). Arhynchobdellida (Annelida: Oligochaeta: Hirudinida): phylogenetic relationships and evolution. *Mol. Phylogenet. Evol.* 30: 213-225.http://dx.doi.org/10.1016/j.ympev.2003.09.002

Brown JM, Pellmyr O, Thompson JN and Harrison RG (1994). Phylogeny of *Greya* (Lepidoptera: Prodoxidae), based on nucleotide sequence variation in mitochondrial cytochrome oxidase I and II: congruence with morphological data. *Mol. Biol. Evol.* 11: 128-141.

Cognato AI and Sperling FA (2000). Phylogeny of *ips* DeGeer species (Coleoptera: scolytidae) inferred from mitochondrial cytochrome oxidase I DNA sequence. *Mol. Phylogenet. Evol.* 14: 445-460.http://dx.doi.org/10.1006/mpev.1999.0705

Damgaard J and Sperling FAH (2001). Phylogeny of the water strider genus *Gerris* Fabricius (Heteroptera: Gerridae) based on *COI* mtDNA, EF-1α nuclear DNA and morphology. *Syst. Entomol.* 26: 241-254. http://dx.doi.org/10.1046/j.1365-3113.2001.00141.x

Damgaard J, Buzzetti FM, Mazzucconi SA, Weir TA, et al. (2010). A molecular phylogeny of the pan-tropical pond skater genus *Limnogonus* Stål 1868 (Hemiptera-Heteroptera: Gerromorpha-Gerridae). *Mol. Phylogenet. Evol.* 57: 669-677. http://dx.doi.org/10.1016/j.ympev.2010.07.020

Dowton M and Austin AD (1997). Evidence for AT-transversion bias in wasp (Hymenoptera: Symphyta) mitochondrial genes and its implications for the origin of parasitism. *J. Mol. Evol.* 44: 398-405.http://dx.doi.org/10.1007/PL00006159

Erpenbeck D, Hooper JNA and Wörheide G (2006). CO1 phylogenies in diploblasts and the 'Barcoding of Life'- are we sequencing a suboptimal partition? *Mol. Ecol. Notes* 6: 550-553. http://dx.doi.org/10.1111/j.1471-8286.2005.01259.x

Folmer O, Black M, Hoeh W, Lutz R, et al. (1994). DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Mol. Mar. Biol. Biotechnol.* 3: 294-299.

Frati F, Simon C, Sullivan J and Swofford DL (1997). Evolution of the mitochondrial cytochrome oxidase II gene in collembola. *J. Mol. Evol.* 44: 145-158.http://dx.doi.org/10.1007/PL00006131

Galtier N, Enard D, Radondy Y, Bazin E, et al. (2005). Mutation hot spots in mammalian mitochondrial DNA. *Genome Res.* 16: 215-222.http://dx.doi.org/10.1101/gr.4305906 PubMed

Grazia J, Schuh RT and Wheeler WC (2008). Phylogenetic relationships of family groups in Pentatomoidea based on morphology and DNA sequences (Insecta: Heteroptera). *Cladistics* 24: 1-45. http://dx.doi.org/10.1111/j.1096-0031.2008.00224.x

Hebert PDN, Cywinska A, Ball SL and deWaard JR (2003). Biological identifications through DNA barcodes. *Proc. Biol. Sci.* 270: 313-321.http://dx.doi.org/10.1098/rspb.2002.2218

Hebert PDN, Stoeckle MY, Zemlak TS and Francis CM (2004). Identification of Birds through DNA Barcodes. *PLoS Biol.* 2: e312.http://dx.doi.org/10.1371/journal.pbio.0020312

Hillis DM and Bull JJ (1993). An Empirical Test of Bootstrapping as a Method for Assessing Confidence in Phylogenetic Analysis. *Syst. Biol.* 42: 182-192. http://dx.doi.org/10.1093/sysbio/42.2.182

Howland DE and Hewitt GM (1995). Phylogeny of the Coleoptera based on mitochondrial cytochrome oxidase I sequence data. *Insect Mol. Biol.* 4: 203-215.http://dx.doi.org/10.1111/j.1365-2583.1995.tb00026.x

Jacobs HT, Elliott DJ, Math VB and Farquharson A (1988). Nucleotide sequence and gene organization of sea urchin mitochondrial DNA. *J. Mol. Biol.* 202: 185-217.http://dx.doi.org/10.1016/0022-2836(88)90452-4

Johns GC and Avise JC (1998). A comparative summary of genetic distances in the vertebrates from the mitochondrial cytochrome b gene. *Mol. Biol. Evol.* 15: 1481-1490.http://dx.doi.org/10.1093/oxfordjournals.molbev.a025875

Jung S, Duwal RK and Lee S (2011). *COI* barcoding of true bugs (Insecta, Heteroptera). *Mol. Ecol. Resour.* 11: 266-270. http://dx.doi.org/10.1111/j.1755-0998.2010.02945.x

Klicka J and Zink RM (1997). The importance of recent ice ages in speciation: a failed paradigm. *Science* 277: 1666-1669.

Li HM, Deng RQ, Wang JW, Chen ZY, et al. (2005). A preliminary phylogeny of the Pentatomomorpha (Hemiptera: Heteroptera) based on nuclear 18S rDNA and mitochondrial DNA sequences. *Mol. Phylogenet. Evol.* 37: 313-326. http://dx.doi.org/10.1016/j.ympev.2005.07.013

Liu H and Beckenbach AT (1992). Evolution of the mitochondrial cytochrome oxidase II gene among 10 orders of insects. *Mol. Phylogenet. Evol.* 1: 41-52.http://dx.doi.org/10.1016/1055-7903(92)90034-E

Mardulyn P and Whitfield JB (1999). Phylogenetic signal in the COI, 16S, and 28S genes for inferring relationships among genera of Microgastrinae (Hymenoptera; Braconidae): evidence of a high diversification rate in this group of parasitoids. *Mol. Phylogenet. Evol.* 12: 282-294.http://dx.doi.org/10.1006/mpev.1999.0618

Matsumoto M (2003). Phylogenetic analysis of the subclass Pteriomorphia (Bivalvia) from mtDNA *COI* sequences. *Mol. Phylogenet. Evol.* 27: 429-440.http://dx.doi.org/10.1016/S1055-7903(03)00013-7

Miura T, Maekawa K, Kitade O, Abe T, et al. (1998). Phylogenetic relationships among subfamilies in higher termites (Isoptera: Termitidae) based on mitochondrial *COII* gene sequences. *Ann. Entomol. Soc. Am.* 91: 515-523. http://dx.doi.org/10.1093/aesa/91.5.515

Monaghan MT, Balke M, Pons J and Vogler AP (2006). Beyond barcodes: complex DNA taxonomy of a South Pacific Island radiation. *Proc. Biol. Sci.* 273: 887-893.http://dx.doi.org/10.1098/rspb.2005.3391

Moritz C and Cicero C (2004). DNA barcoding: promise and pitfalls. *PLoS Biol.* 2: e354. http://dx.doi.org/10.1371/journal.pbio.0020354

Muraji M, Kawasaki K and Shimizu T (2000). Nucleotide sequence variation and phylogenetic utility of the mitochondrial *COI* fragment in anthocorid bugs (Hemiptera: Anthocoridae). *Appl. Entomol. Zool. (Jpn.)* 35: 301-307. http://dx.doi.org/10.1303/aez.2000.301

Pollock DD, Watt WB, Rashbrook VK and Iyengar EV (1998). Molecular phylogeny for Colias butterflies and their relatives (Lepidoptera: Pieridae). *Ann. Entomol. Soc. Am.* 91: 524-531. http://dx.doi.org/10.1093/aesa/91.5.524

Roe AD and Sperling FAH (2007). Patterns of evolution of mitochondrial cytochrome c oxidase I and II DNA and implications for DNA barcoding. *Mol. Phylogenet. Evol.* 44: 325-345.http://dx.doi.org/10.1016/j.ympev.2006.12.005

Ronquist F, Teslenko M, van der Mark P, Ayres DL, et al. (2012). MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* 61: 539-542.http://dx.doi.org/10.1093/sysbio/sys029

Schneider H (2007). Métodos de Análise Filogenética, 3a. Edition. Holos, 200p.

Simon C, Frati F, Beckenbach A, Crespi B, et al. (1994). Evolution, weighting, phylogenetic utility of mitochondrial gene sequences and compilation of conserved polymerase chain reaction primers. *Ann. Entomol. Soc. Am.* 87: 651-701. http://dx.doi.org/10.1093/aesa/87.6.651

Stoeckle M (2003). Taxonomy, DNA and the bar code of life. *Bioscience* 53: 2-3. http://dx.doi.org/10.1641/0006-3568(2003)053[0796:TDATBC]2.0.CO;2

Swofford DL (2000). PAUP* Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.

Tamura K, Peterson D, Peterson N, Stecher G, et al. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* 28: 2731-2739. http://dx.doi.org/10.1093/molbev/msr121

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, et al. (1997). The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* 25: 4876-4882. http://dx.doi.org/10.1093/nar/25.24.4876

Ward RD, Zemlak TS, Innes BH, Last PR, et al. (2005). DNA barcoding Australia's fish species. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 360: 1847-1857.http://dx.doi.org/10.1098/rstb.2005.1716

Ward RD, Hanner R and Hebert PD (2009). The campaign to DNA barcode all fishes, FISH-BOL. *J. Fish Biol.* 74: 329-356. http://dx.doi.org/10.1111/j.1095-8649.2008.02080.x

Yoo HS, Eah JY, Kim JS, Kim YJ, et al. (2006). DNA barcoding Korean birds. *Mol. Cells* 22: 323-327.