Turkish Journal of Biology

Volume 35 | Number 6 Article 9

1-1-2011

Utilization of cytochrome oxidase I in Cephus pygmeus (L.) (Hymenoptera: Cephidae)

ERTAN MAHİR KORKMAZ

MAHİR BUDAK

HASAN HÜSEYİN BAŞIBÜYÜK

Turk J Biol
35 (2011) 713-726

doi:10.3906/biy-1003-65

Utilization of cytochrome oxidase I in *Cephus pygmeus* (L.) (Hymenoptera: Cephidae)

Ertan Mahir KORKMAZ[1], Mahir BUDAK[2], Hasan Hüseyin BAŞIBÜYÜK[2]
[1]Department of Molecular Biology and Genetics, Faculty of Science, Cumhuriyet University, 58140 Sivas - TURKEY
[2]Department of Biology, Faculty of Science, Cumhuriyet University, 58140 Sivas - TURKEY

Received: 11.03.2010

Abstract: The mitochondrial cytochrome oxidase I (COI) gene region of *Cephus pygmeus* (L.) (Hymenoptera: Cephidae) was characterized at the nucleotide and amino acid levels. The partial sequence (687 bp) of the gene was determined and compared with other cephids and insect species at both the nucleic acid and amino acid sequence levels. The examined fragment was placed at an equal distance to both ends of the gene and consisted of 13 structural regions. The 5' end of the partial COI gene was more informative than the 3' end at the intraspecific level. Pairwise sequence divergences ranged from 0.15% to 1.17% at the intraspecific level, while the rate varied from 8.14% to 15.32% in comparison with other *Cephus* species. The percentage of nucleotide composition at each codon position and the rate of nucleotide substitutions were variable and displayed a bias toward A + T. A strong bias of transitional substitutions was observed at the intraspecific level, but there was a lower rate at the interspecific level. The codon usage patterns and the extent of the substitutions at synonymous codons were also investigated at the amino acid level. These results suggest that the characterized region is informative for estimating both intra- and interspecific relationships due to its possession of both completely conserved and variable regions.

Key words: Cephidae, COI, substitution bias, codon usage

Cephus pygmeus (L.) (Hymenoptera: Cephidae)'da sitokrom oksidaz I'in kullanımı

Özet: *Cephus pygmeus* (L.) (Hymenoptera: Cephidae)'ta mitokondrial sitokrom oksidaz I gen bölgesi nükleotid ve amino asit düzeyinde karakterize edildi. Genin kısmi DNA dizisi (687 bp) belirlendi ve hem nükleik asit hem de amino asit düzeyinde bazı diğer cephid ve böcek türleri ile karşılaştırıldı. Genin araştırılan kısmı her iki uca eşit uzaklıkta olup 13 yapısal bölge içermektedir. Tür içi düzeyde kısmi COI geninin 5' ucu 3' ucundan daha bilgi vericidir. Nükleotid farklılaşma oranı tür içi düzeyde % 0,15-1,17 arasında iken, bu oran diğer *Cephus* türleri ile karşılaştırıldığında % 8,14-15,32 arasında gözlenmektedir. Her bir kodon pozisyonundaki nükleotid kompozisyonunun yüzdesi ve nükleotid yer değişimlerinin oranı değişken olup A + T yönünde bir eğilim sergilemektedir. Transisyonel değişimler tür içinde güçlü bir eğilim sergilerken, bu eğilim türler arasında daha düşük düzeydedir. Ayrıca, amino asit düzeyinde de kodon kullanım örüntüleri ve sinonim kodonlardaki yer değişimlerin derecesi araştırıldı. Bu sonuçlar karakterize edilen bölgenin hem tümüyle korunmuş hem de değişken bölgeler içermesi nedeniyle tür içi ve türler arası ilişkiyi belirlemede bilgi verici bir belirteç olabileceğini önermektedir.

Anahtar sözcükler: Cephidae, COI, nükleotid yerdeğişim eğilimi, kodon kullanımı

Introduction

Cephus pygmeus (L.) (Hymenoptera: Cephidae), the wheat stem sawfly, is widely distributed in Europe, Asia, northern Africa, and North America (1-6). This Holarctic species is also observed as a very abundant and widespread species in central Anatolia (7). Many widespread species exhibit considerable phenotypic variation as a result of local adaptations and experienced historical events (8,9). The wide distribution of *C. pygmeus* appears to have resulted in intraspecific variation in some morphological structures and color patterns (7). These variable characteristics are mainly claw shape, body size, and color patterns of the abdomen, forewings, and legs. The available identification keys are mainly based on the use of such characteristics and therefore fail to distinguish the species. Consequently, using morphological characteristics based on the taxonomy of *C. pygmeus* for identification is problematic; investigation of genetic markers may be utilized in studying both the phylogeny and the taxonomy of the species.

Mitochondrial genes are particularly well suited as a genetic marker for phylogenetic studies at the intra- and interspecific level (9-12). Mitochondrial protein coding genes generally evolve fast and reach saturation rapidly, possibly due to a deficient mismatch repair system and an A + T-rich base composition (13,14). The usefulness of genes for reconstructing a nucleotide-based phylogeny can be influenced by the substitution properties of those genes (15). The structural and functional characterization of such genes may reveal a more accurate evolutionary pattern specific to a related taxon.

Here, the mitochondrial cytochrome oxidase subunit I (COI) gene was selected for molecular characterization of *C. pygmeus* and compared with some other cephid species. The selected COI gene region was also utilized as a DNA-based bioidentification system for animals at the global level (16). The COI gene has proved to be useful for molecular phylogenetic analysis of different taxonomic levels because of its relatively large size, its possession of both highly conserved and variable regions, and its heterogeneous evolutionary rate (11,15,17,18).

We found that synonymous nucleotide substitutions at codon sites of the gene provided variation within species. The percentage of nucleotide composition and the rate of nucleotide substitutions at each codon position were variable and displayed a bias toward A + T. A strong bias of transitional substitutions was also observed at the intraspecific level. We concluded that the characterized gene region is informative for estimating relationships at both the intra- and interspecific levels.

Material and methods

Samples

The voucher specimens used in this study were collected from localities presented in Table 1 and Figure 1 and were deposited in the Entomological Collection of Cumhuriyet University, Sivas (ECCUS). Gene sequences of 24 specimens were investigated: 20 sequences of *C. pygmeus*, 2 sequences representing the genus *Cephus* (*C. pulcher* and *C. sareptanus*), and 2 sequences representing other genera (*Trachelus* and *Calameuta*) of the tribe Cephini. The sequences of *Apis mellifera* and *Drosophila melanogaster* deposited in GenBank were also included in the analyses for comparison (19,20).

DNA analysis

Specimens were stored in 99% ethanol at –20 °C. The total genomic DNA was isolated from a single leg of each specimen utilizing the method described by Sambrook and Russell with slight alterations (21). In brief, the ethanol was removed by washing 3 times for 30 min each in 10 mM Tris-HCl (pH 8.0) containing 100 mM NaCl and 1 mM $MgCl_2$. The tissue was then homogenized in 500 µL of sterile salt homogenizing buffer (400 mM NaCl, 10 mM Tris-HCl, pH 8.0; and 2 mM EDTA, pH 8.0). Next, 10% SDS (2% final concentration) and proteinase K (100 µg/mL final concentration) were added, mixed well, and incubated overnight at 37 °C. The homogenate was then extracted with phenol-chloroform followed by an overnight salt-ethanol precipitation at –20 °C. The DNA was dissolved in 1× TE buffer (10 mM Tris-HCl, pH 8.0; 1 mM EDTA, pH 8.0) and stored at –20 °C. For polymerase chain reaction (PCR), 2 µL of each DNA solution was used directly. The DNA was amplified using conserved COI primers

Figure 1. The locations of sampling sites. Haplotype localities marked as (▲) haplogroup 1, (●) haplogroup 2, and (□) haplogroup 3.

Table 1. List of species (specimens) sequenced, their locality information, and GenBank accession numbers.

Species	Specimens		Locality	Accession No.
Cephus pygmeus	Hap1	YozAkd	39°49′N, 36°18′E	HQ644304
	Hap2	CorAla	39°55′N, 34°56′E	HQ644312
	Hap3	NigUluk	37°35′N, 34°25′E	HQ644313
	Hap4	ErzYal	39°47′N, 39°21′E	HQ644314
	Hap5	AksGz	38°15′N, 34°24′E	HQ644315
	Hap6	AnkEmir	38°53′N, 31°03′E	HQ644316
	Hap7	KrsKz	39°01′N, 34°17′E	HQ644317
	Hap7	KrsTuz	38°15′N, 33°30′E	HQ644317
	Hap7	NigCam	38°00′N, 35°02′E	HQ644317
	Hap8	Nev	38°27′N, 34°42′E	HQ644318
	Hap9	NigAzat	38°10′N, 34°31′E	HQ644319
	Hap10	NigGl	37°59′N, 35°04′E	HQ644305
	Hap11	NigPoz	37°32′N, 34°56′E	HQ644306
	Hap12	SivHal	39°42′N, 36°49′E	HQ644307
	Hap13	SivKang	39°14′N, 37°23′E	HQ644308
	Hap13	SivKaras	39°16′N, 37°55′E	HQ644308
	Hap14	SivSc	39°45′N, 36°44′E	HQ644309
	Hap15	SivZar	39°57′N, 37°53′E	HQ644310
	Hap16	SivKamp	39°42′N, 37°01′E	HQ644311
	Hap16	SivMerk	39°49′N, 37°10′E	HQ644311
Cephus pulcher	C.pulcher		38°43′N, 37°17′E	HQ644303
Cephus sareptanus	C.sareptanus		40°09′N, 37°54′E	HQ644320
Trachelus tabidus	T.tabidus		40°09′N, 32°31′E	HQ644302
Calameuta haemorrhoidalis	Cl.haemorrhoidalis		39°27′N, 32°58′E	HQ644301

with the following sequence: COI–s1859, 5'–GGAACIGGATGAACWGTTTAYCCICC–3' and COI–a2590, 5'–TCCTATTGATARWACATARTGRAAATG–3' (22,23). Amplification conditions were as follows: denaturation for 5 min at 94 °C, followed by 37 cycles of denaturation at 94 °C for 30 s, annealing at 59 °C for 45 s, extension at 72 °C for 30 s, and a 5 min final extension at 72 °C. The PCR products were run in 1% agarose gel and then the correct bands were purified using a DNA gel extraction kit (Fermentas) according to the manufacturer's instructions (Figure 2). The purified amplification products were sent for sequencing in both directions (İontek, İstanbul). Chromatographs were checked by eye, and, if necessary, were corrected. Sequences were aligned using MEGA (ver. 4.0) (24). Finally, sequences of all specimens were deposited in GenBank (Table 1).

Data analysis

Molecular characterization of the sequences was examined at the level of both nucleotides and amino acids. Nucleotides, with the Jukes-Cantor correction (25) and haplotype diversity measures providing information about the level of intraspecific variation, were generated by computing the indices implemented in DnaSP software, ver. 4.90.1 (26). Patterns and rates of nucleotide substitution were also determined to calculate pairwise distances and to generate trees. Substitution saturation in the COI sequence was analyzed using the saturation index described by Xia et al. (27) and executed in the program DAMBE (28). The observed saturation index (I_{ss}) was compared with the critical index value ($I_{ss.c}$) to detect the unsaturated sequences usable for phylogenetic analyses. If I_{ss} is not smaller than $I_{ss.c}$, severe substitution saturation is suggested (29). In order to investigate the nucleotide-based genetic distance between the haplotypes, phenetic trees were constructed under the UPGMA of MEGA ver. 4.0 (29). In addition to nucleotide-based distance trees, maximum likelihood and maximum parsimony trees were built using PAUP*4.0b10 (30). The nucleotide sequences were later translated into amino acid sequences on the basis of the genetic code for *Drosophila* mtDNA (31). The types and rates of amino acid substitution and the codon usage bias were also quantified by applying the DnaSP program. The aligned amino acid sequence was divided into 13 regions based on the study by Lunt et al. (17) using TMHMM Server ver. 2.0.

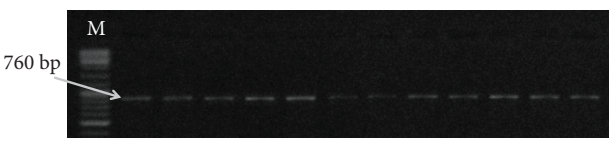

Figure 2. The COI gene amplification products separated by 1% agarose gel electrophoresis. Lane M: 100-bp DNA molecular weight marker. Other lanes: amplified PCR products (760 bp).

Results and discussion

Nucleotide characterization

Diversity indices

A complete alignment of the partial COI gene sequences from the 20 *C. pygmeus* specimens resulted in a fragment containing 687 base pairs, among which 16 nucleotide positions were variable and including a total of 17 mutations. The analyzed sequences correspond to a functional mitochondrial gene region because of the presence of singular peaks in each chromatograph, the absence of indel and premature stop codons, and the presence of the highest nucleotide substitutions at codon position 3 (9). The studied COI gene fragment was at an almost equal distance from both ends. The 5' end of the COI gene was more informative than the 3' end at the intraspecific level (22,32-34). This is due to synonymous substitution in degenerative codon positions and it provides sufficient variation for the investigation of relationships at the intraspecific level as well as among closely related species (35). Among 20 specimens, 16 haplotypes were detected; most were found in only 1 individual, except for haplotypes 7, 13, and 16. Most haplotypes exhibited 1 or 2 nucleotide differences from each other. Haplotypes 12 and 15 were different from haplotype 16 in 8 nucleotide positions. However, this was considerably less than the difference between *C. pygmeus* and *C. pulcher* (78-83 nucleotides). Overall haplotype diversity (*H*d) was 0.974 ± 0.02. Pairwise sequence divergences ranged from 0.14% to 1.17% and overall nucleotide diversity (*P*i) was 0.0055 ± 0.0016 (Table 2). The rate varied from 8.14% to 15.32% in comparison with other *Cephus* species in the present study. The presence of very low divergence between haplotypes

Table 2. Genetic distance estimation: number of base substitutions per site in the analysis between sequences. All results are based on pairwise analysis of sequences. Analyses were conducted using the Jukes-Cantor method (25) in MEGA 4.0.

Cl.haemorrhoidalis		0.0160	0.0156	0.0178	0.0160	0.0158	0.0159	0.0159	0.0160	0.0158	0.0157	0.0159	0.0160	0.0156	0.0160	0.0161	0.0157	0.0160	0.0155	0.0159
T.tabidus	0.1608		0.0135	0.0159	0.0169	0.0166	0.0164	0.0168	0.0169	0.0164	0.0167	0.0167	0.0164	0.0166	0.0164	0.0167	0.0167	0.0169	0.0162	0.0162
C.pulcher	0.1396	0.1223		0.0114	0.0148	0.0146	0.0141	0.0148	0.0146	0.0148	0.0148	0.0148	0.0143	0.0146	0.0142	0.0147	0.0147	0.0149	0.0147	0.0139
C.sareptanus	0.1698	0.1608	0.0793		0.0156	0.0155	0.0154	0.0156	0.0154	0.0155	0.0156	0.0157	0.0156	0.0155	0.0155	0.0156	0.0155	0.0157	0.0155	0.0154
Hap1	0.1554	0.1771	0.1240	0.1396		0.0024	0.0035	0.0020	0.0032	0.0031	0.0025	0.0027	0.0032	0.0028	0.0029	0.0023	0.0021	0.0014	0.0034	0.0039
Hap2	0.1536	0.1734	0.1223	0.1378	0.0044		0.0026	0.0014	0.0028	0.0028	0.0020	0.0024	0.0028	0.0015	0.0031	0.0020	0.0024	0.0020	0.0024	0.0035
Hap3	0.1554	0.1680	0.1172	0.1361	0.0087	0.0044		0.0031	0.0035	0.0034	0.0028	0.0031	0.0025	0.0026	0.0028	0.0035	0.0031	0.0033	0.0033	0.0023
Hap4	0.1554	0.1753	0.1240	0.1396	0.0029	0.0014	0.0058		0.0031	0.0031	0.0025	0.0028	0.0032	0.0021	0.0034	0.0014	0.0028	0.0014	0.0029	0.0039
Hap5	0.1554	0.1771	0.1223	0.1378	0.0073	0.0058	0.0087	0.0073		0.0027	0.0020	0.0025	0.0032	0.0024	0.0035	0.0034	0.0025	0.0028	0.0032	0.0038
Hap6	0.1572	0.1716	0.1240	0.1396	0.0073	0.0058	0.0087	0.0073	0.0058		0.0019	0.0023	0.0031	0.0024	0.0034	0.0035	0.0023	0.0028	0.0031	0.0038
Hap7	0.1536	0.1753	0.1240	0.1396	0.0044	0.0029	0.0058	0.0044	0.0029	0.0029		0.0014	0.0025	0.0014	0.0028	0.0029	0.0014	0.0021	0.0024	0.0032
Hap8	0.1554	0.1753	0.1258	0.1413	0.0058	0.0044	0.0073	0.0058	0.0044	0.0044	0.0014		0.0028	0.0020	0.0030	0.0031	0.0019	0.0024	0.0029	0.0035
Hap9	0.1572	0.1698	0.1189	0.1378	0.0073	0.0058	0.0044	0.0073	0.0073	0.0073	0.0044	0.0058		0.0028	0.0014	0.0035	0.0028	0.0029	0.0035	0.0020
Hap10	0.1519	0.1734	0.1223	0.1378	0.0058	0.0014	0.0044	0.0029	0.0044	0.0044	0.0014	0.0029	0.0058		0.0031	0.0026	0.0020	0.0025	0.0019	0.0035
Hap11	0.1554	0.1698	0.1172	0.1361	0.0058	0.0073	0.0058	0.0087	0.0087	0.0087	0.0058	0.0073	0.0014	0.0073		0.0037	0.0025	0.0032	0.0038	0.0024
Hap12	0.1572	0.1734	0.1223	0.1413	0.0044	0.0029	0.0073	0.0014	0.0087	0.0087	0.0058	0.0073	0.0087	0.0044	0.0102		0.0031	0.0019	0.0032	0.0042
Hap13	0.1519	0.1753	0.1223	0.1378	0.0029	0.0044	0.0073	0.0058	0.0044	0.0044	0.0014	0.0029	0.0058	0.0029	0.0044	0.0073		0.0025	0.0027	0.0035
Hap14	0.1572	0.1771	0.1258	0.1413	0.0014	0.0029	0.0073	0.0014	0.0058	0.0058	0.0029	0.0044	0.0058	0.0044	0.0073	0.0029	0.0044		0.0032	0.0037
Hap15	0.1483	0.1698	0.1223	0.1378	0.0087	0.0044	0.0073	0.0058	0.0073	0.0073	0.0044	0.0058	0.0087	0.0029	0.0102	0.0073	0.0058	0.0073		0.0041
Hap16	0.1572	0.1662	0.1156	0.1344	0.0102	0.0087	0.0044	0.0102	0.0102	0.0102	0.0073	0.0087	0.0029	0.0087	0.0044	0.0117	0.0087	0.0087	0.0117	

is possibly due to a high level of migration and a narrow sampling area. However, this conclusion needs to be confirmed by a greater sample size and more diverse geographic locations.

Nucleotide composition

The percentage of nucleotide composition at each codon position and the rate of nucleotide substitutions were variable (Table 3). The mean frequency of nucleotides was biased toward A + T (T 37.1%, C 15.6%, A 32.2%, and G 15.1%). The A + T content at codon positions 3, 2, and 1 were 88.7%, 59.3%, and 60.4%, respectively. The nucleotide G had the lowest (2.1%) and A the highest content (48.8%) at codon position 3. High frequencies of T and C were observed at position 2, probably reflecting selective constraints arising from the membrane-spanning structure of the protein due to hydrophobic amino acids, which are generally encoded by codons containing C and T at position 2 (36). These characteristics are also found in other cephid species (Table 3). The nucleotide bias in the COI gene is similar to those of other reported members of Hymenoptera (37-41). Although the A + T content appeared relatively homogenous for all positions among cephid species, the frequency of bases in codon position 3 were heterogeneous (Table 3). Such a pattern in nucleotide composition indicates that the mutational range of the COI gene for cephid species varies over shorter time scales (36). This gene is also phylogenetically informative because of easily observed differences in the amount of bias among all positions (15).

Nucleotide substitutions

The distribution of polymorphic sites for all cephid species shows that the majority of substitutions

Table 3. Table of nucleotide content at all sites and 3 codon positions.

Specimen	All sites				Codon position 1			
	T(U)	C	A	G	T-1	C-1	A-1	G-1
Cl.haemorrhoidalis	34.8	18.6	31.6	15.0	28	14.0	31.0	27.5
T.tabidus	36.7	15.9	32.8	14.7	28	13.1	32.3	26.6
C.seraptanus	37.0	16.2	32.2	14.7	30	12.2	31.4	26.6
C.pulcher	33.3	18.6	33.5	14.6	27	14.4	31.9	26.6
C.pygmeus (means of Hap1-16)	37.1	15.6	32.2	15.1	29	12.2	31.4	27.5
Mean	36.9	15.9	32.2	15.0	29	12.4	31.5	27.4
Drosophila	41.3	13.5	29.3	15.9	31	12.2	25.8	31.0
Apis	41.6	12.4	33.8	12.2	33	11.4	32.8	23.1

Specimen	Codon position 2				Codon position 3			
	T-2	C-2	A-2	G-2	T-3	C-3	A-3	G-3
Cl.haemorrhoidalis	42	25.8	16.6	15.3	34	16.2	47.2	2.2
T.tabidus	42	25.3	17.0	15.3	40	9.2	48.9	2.2
C.seraptanus	43	24.9	16.6	15.7	38	11.4	48.5	1.7
C.pulcher	42	25.8	16.6	15.3	31	15.7	52.0	1.7
C.pygmeus (means of Hap1-16)	43	25.3	16.3	15.5	40	9.2	48.7	2.1
Mean	43	25.3	16.4	15.6	39	9.9	48.8	2.1
Drosophila	43	26.6	15.3	15.3	50	1.7	46.7	1.3
Apis	45	22.3	18.8	13.5	47	3.5	49.8	0.0

are at synonymous positions. Within species, 15 segregating sites are silent, with substitutions occurring at codon position 3 (88.2%); the remaining 2 sites harbor replacement substitutions at codon position 1 [position 16 (G ↔A)] and codon position 2 [position 122 (G ↔ A)]. Synonymous substitutions in *Cephus* and Cephini were also identified at 111 of the 116 (95.6%) and 182 of the 192 (94.7%) polymorphic sites, respectively. The vast majority of synonymous substitutions were also found at codon position 3, with a rate of 90.09% for *Cephus* and 92.8% for Cephini. Positions 1 and 2 are relatively more conserved in comparison with position 3. However, many replacement substitutions occur in codon position 1 (80.0% for *Cephus* and 30.0% for Cephini), and position 2 is highly conserved (Table 4). The accumulation of most substitutions at codon position 3 shows strong functional constraints over the gene, but not the amino acid composition (42). The maximum likelihood estimation revealed the following substitution pattern: 49.69% were from G ↔ A while 2.77% were from C ↔ G. The bias in the proportion of (T + C)/(A + G) is well explained in the various protein-coding genes (43). The proportions of pyrimidines (T, C) to purines (A, G) in the COI sequences deviate from a 50:50 ratio. The proportion of (T + C)/(A + G) is 52.8%/47.2% for all 3 codon positions, 41%/59% for position 1, 68.1%/31.9% for position 2, and 49.2%/50.9% for position 3. The deviation in nucleotide composition probably relates to the strand asymmetry in the replication process (13) and can also reflect the hydrophobic properties of the gene product.

The overall transition-to-transversion ratio (*R*) was 2.922 within species and was higher than those observed in other cephid species (*R* = 0.772). The index of substitution saturation within species, I_{ss}, was much smaller (0.324) than $I_{ss.c}$ (0.736), indicating that the sequences were not significantly saturated or informative for phylogenetic analysis (27). A similar result was also observed among all cephid species analyzed here. Scatter plots of pairwise sequence divergence versus the total number of transitions and transversions within and among species (Figure 3ab) also suggest slight saturation with a strong bias toward transitions within species and decline of the bias among species as a result of increased genetic distance. This aspect may be explained by the fast saturation of transitional substitutions due to biases in base composition and substitution types within species. Therefore, the use of transversional substitutions rather than all substitutions may be more useful in the investigation of phylogenetic relationships within species. Transversional substitutions provide more phylogenetic information since multiple substitutions at the site of a gene may potentially obscure the information (15). This was confirmed by the bootstrap consensus UPGMA (44) trees based on all substitutions and only transversions (Figures

Table 4. Percentage of synonymous and replacement substitutions, and the number of polymorphic, informative, and singleton sites occurring in different codon positions.

		Synonymous substitutions (%)			**Replacement substitutions (%)**			**Polymorphic sites**			**Informative sites**			**Singleton sites**		
Nucleotide codon positions	**Number of bases**	*C.pygmeus*	*Cephus*	Cephini	*C.pygmeus*	*Cephus*	Cephini	*C.pygmeus*	*Cephus*	Cephini	*C.pygmeus*	*Cephus*	Cephini	*C.pygmeus*	*Cephus*	Cephini
Total sequence	687	—	—	—	—	—	—	*16*	144	191	*7*	41	107	*9*	103	84
1	229	0	5.67	7.42	0.43	2.18	1.74	1	13	23	0	2	11	1	15	14
2	229	0	0	0	1.31	0.43	0.87	1	0	6	1	1	1	0	4	3
3	229	6.55	57.2	79.47	0	0.43	0.87	14	131	162	6	38	95	8	84	67
Amino acids	229	—	—	—	—	—	—	2	8	11	1	2	4	1	6	7

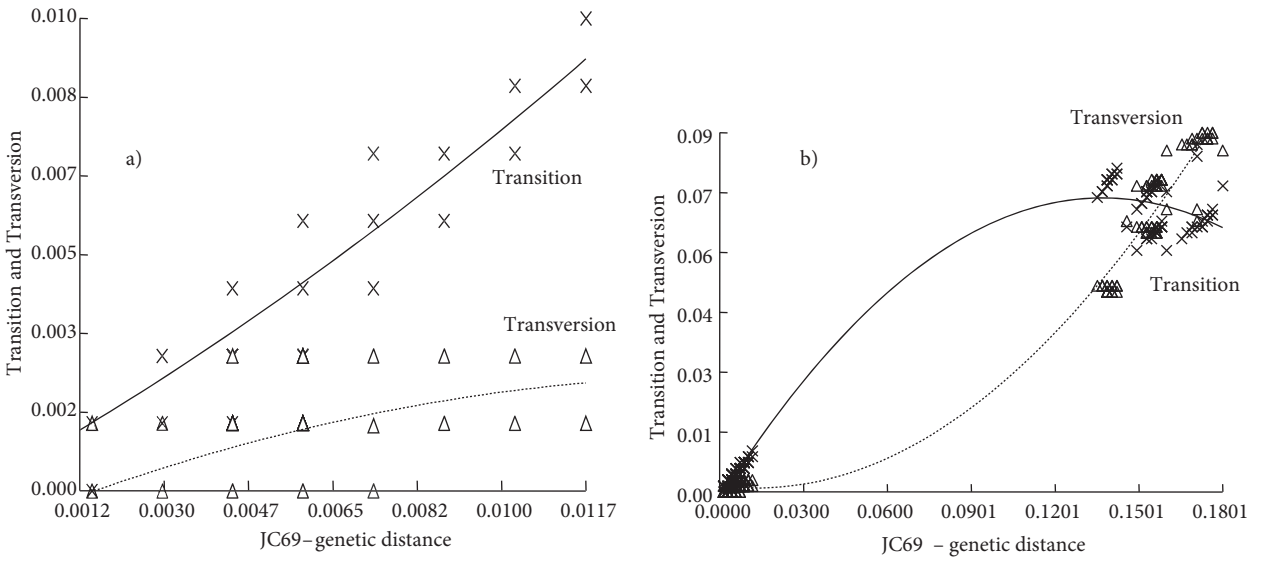

Figure 3. Graphs illustrating skew of transitions and transversions versus genetic distance (JC69) in a) *C. pygmeus* and b) other representatives of the subfamily Cephini at the level for the COI gene (transversions Δ, transitions ×).

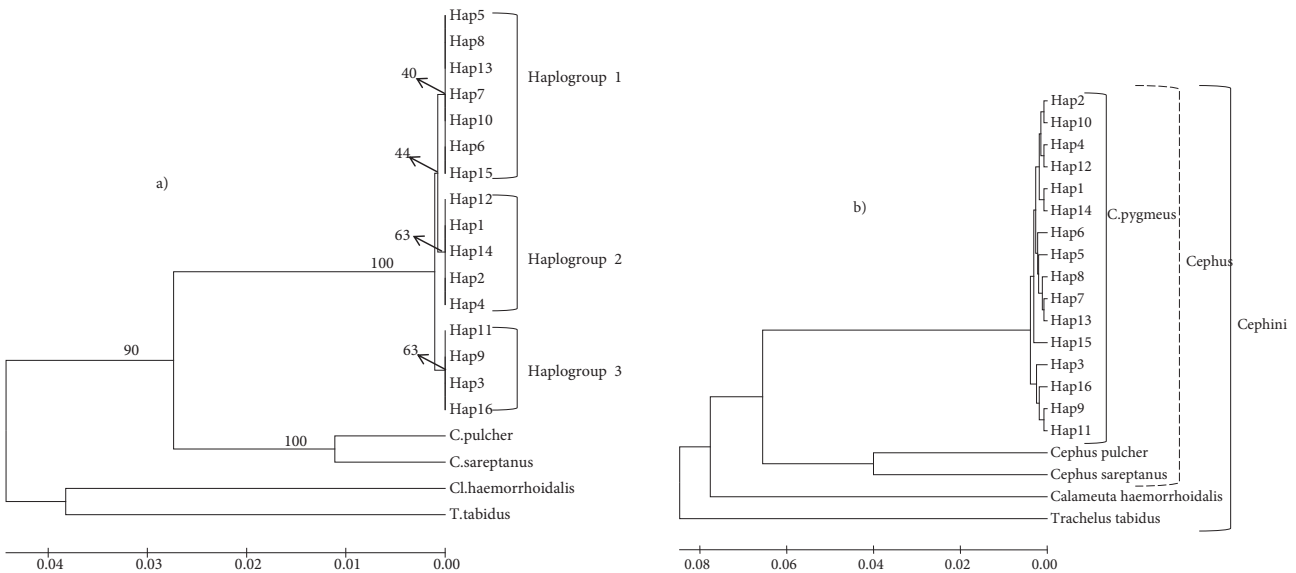

Figure 4. Trees constructed by the UPGMA method with the use of Kimura 2 parameters considering a) only transversion differences with bootstrap values presented and b) all nucleotide substitutions. Bootstrap analysis was performed on 500 replicates.

4a and 4b). As shown in Figure 4a, the use of only transversional substitutions resulted in 3 different clusters with an indication of a geographical pattern (Figure 1). However, the inclusion of all substitutions obscures the intraspecific haplotype grouping (Figure 4b). The maximum likelihood analysis generated only 1 tree with the same topology (see Figure 4a). However, the maximum parsimony analysis resulted in 15 trees with a step length of 261 (CI = 0.7854, HI = 0.2146, RI = 0.7282, RC = 0.5719). These cladograms showed the same topologies at the specific level, but there were slight differences in the haplotype grouping that are not presented here. The recovered haplogroups were not completely mutual with intraspecific morphological variations. We therefore suggest that the differentiation observed in the COI gene region of *C. pygmeus* may not be related to intraspecific morphological variations.

Amino acid characterization

Structural and functional characterization of gene products as well as nucleotide state may be important indicators of intra- and interspecific analysis (15,17). Hence, the partial COI gene of *C. pygmeus* and other cephid species were analyzed at the amino acid level and compared with those of *Apis* and *Drosophila*. The codon usage pattern was primarily investigated, as the nucleotide biases and genome compositional constraints are also reflected in the codon usage (45,46). Nucleotide evolution in functional genes also depends on codon composition since the selection of nucleotide substitutions operates at the codon level rather than the nucleotide level.

Codon usage and codon bias

Codon usage in the partial COI gene of *C. pygmeus* and other cephid species exhibits a bias against codons ending with C and particularly G, except Asn (GAY) and Asp (AAY), whose most frequently used codons ended with cytosine (Table 5). The strong bias for the COI gene is observed in many hymenopteran species and a general trend is toward the usage of high AT codons in Hymenoptera (39). In the present study, hydrophobic amino acids were more frequently encoded by UUU (F), UUA (L), AUU (I), and AUA (M) (Table 5). The 4-fold degenerate codon family presents a particular bias towards adenine at codon position 3, except for Pro (CCN). In addition, a preference for triplets ending with A is shown by the 6-fold degenerate codon family encoding leucine. A similar state is found in pterygote insects (46,47). Relative synonymous codon usage (RSCU) values indicate the usage of alternative synonymous codons (Table 5) (48). Selection among alternative synonymous codons may be caused by an increase of translation speed and a decrease of amino acid misincorporation, and can result in nucleotide bias (45-47). Therefore, the determination of synonymous codon usage and amino acid composition is a useful tool to explore changes in the mutation-selection pressure. Here, the codon with the highest RSCU was UUA (coding for leucine), with the exception of *Calameuta* and various others among the cephid

Table 5. Codon usage in Cephidae species sequenced based on *Drosophila* mitochondrial genetic code (31) *Estimates of observed average codon frequencies, standard amino acid identification signal-letter codes, and numerical RSCU values are given in parentheses.

UUU(F)	16.0(1.60)	UCU(S)	6.0(2.85)	UAU(Y)	3.2(1.26)	UGU(C)	0.0(0.00)
UUC(F)	4.1(0.41)	UCC(S)	1.1(0.50)	UAC(Y)	1.9(0.74)	UGC(C)	0.0(0.00)
UUA(L)	23.0(4.76)	UCA(S)	6.0(2.87)	UAA(*)	0.0(0.00)	UGA(W)	5.0(2.00)
UUG(L)	0.0(0.00)	UCG(S)	0.0(0.00)	UAG(*)	0.0(0.00)	UGG(W)	0.0(0.00)
CUU(L)	1.0(0.21)	CCU(P)	4.8(2.11)	CAU(H)	6.1(1.53)	CGU(R)	0.0(0.00)
CUC(L)	0.0(0.00)	CCC(P)	0.0(0.00)	CAC(H)	1.9(0.48)	CGC(R)	0.0(0.00)
CUA(L)	5.0(1.03)	CCA(P)	3.3(1.44)	CAA(Q)	2.0(2.00)	CGA(R)	3.0(4.00)
CUG(L)	0.0(0.00)	CCG(P)	1.0(0.44)	CAG(Q)	0.0(0.00)	CGG(R)	0.0(0.00)
AUU(I)	23.0(1.84)	ACU(T)	5.0(1.05)	AAU(N)	3.1(0.89)	AGU(S)	0.0(0.00)
AUC(I)	2.0(0.16)	ACC(T)	1.0(0.21)	AAC(N)	3.9(1.11)	AGC(S)	0.0(0.00)
AUA(M)	12.1(1.85)	ACA(T)	13.1(2.74)	AAA(K)	4.3(2.00)	AGA(S)	3.7(1.77)
AUG(M)	1.0(0.15)	ACG(T)	0.0(0.00)	AAG(K)	0.0(0.00)	AGG(S)	0.0(0.00)
GUU(V)	5.0(1.82)	GCU(A)	7.3(1.71)	GAU(D)	4.1(1.01)	GGU(G)	7.0(1.17)
GUC(V)	0.0(0.00)	GCC(A)	1.5(0.35)	GAC(D)	4.0(0.99)	GGC(G)	0.0(0.00)
GUA(V)	6.0(2.18)	GCA(A)	8.2(1.94)	GAA(E)	3.0(2.00)	GGA(G)	14.2(2.36)
GUG(V)	0.0(0.00)	GCG(A)	0.0(0.00)	GAG(E)	0.0(0.00)	GGG(G)	2.9(0.48)

species. All of the remaining codons with RSCU greater than 2 have T or, particularly, A in codon position 3. UUA being used with respect to UUG may suggest that codon usage in cephids is simply related to nucleotide content rather than selection for certain tRNA molecules, due to the mitochondrially encoded tRNA for leucine being $tRNA^{Leu(UUR)}$ and having anticodon TAA (49).

The extent of the substitutions at synonymous codons is measured by 3 algorithms (Table 6). Because the G + C contents at all codon positions are close to each other, *C. pygmeus* haplotypes exhibit a similar codon usage bias. The ratios were also compared with those of *Apis* and *Drosophila*, which showed higher codon bias than all cephid specimens. A relationship between the base composition of codon

Table 6. Estimates of codon bias in the partial COI gene of the analyzed specimens.

Specimen	ENC	CBI	SChi2	G + C2	G + C3s	G + Cc
Drosophila	28.699	0.852	0.513	0.419	0.031	0.294
Apis	27.702	0.839	0.406	0.358	0.035	0.246
Cl.haemorrhoidalis	33.085	0.614	0.290	0.410	0.183	0.336
T.tabidus	31.252	0.718	0.281	0.406	0.114	0.306
C.seraptanus	32.932	0.630	0.289	0.406	0.131	0.309
C.pulcher	34.303	0.715	0.385	0.410	0.175	0.332
Mean of haplotypes	32.791	0.676	0.282	0.408	0.114	0.307
C.pygmeus						
Hap1	32.820	0.676	0.283	0.410	0.114	0.307
Hap2	32.679	0.679	0.287	0.406	0.114	0.306
Hap3	32.422	0.684	0.283	0.406	0.109	0.304
Hap4	32.810	0.673	0.282	0.406	0.118	0.307
Hap5	32.635	0.680	0.284	0.410	0.114	0.306
Hap6	33.410	0.673	0.269	0.410	0.118	0.309
Hap7	33.008	0.670	0.278	0.410	0.118	0.309
Hap8	33.144	0.663	0.273	0.410	0.122	0.310
Hap9	32.790	0.677	0.286	0.410	0.109	0.306
Hap10	32.867	0.672	0.282	0.406	0.118	0.307
Hap11	32.612	0.683	0.297	0.410	0.105	0.304
Hap12	32.466	0.683	0.284	0.406	0.114	0.306
Hap13	32.849	0.675	0.285	0.410	0.114	0.307
Hap14	32.952	0.670	0.278	0.410	0.118	0.309
Hap15	32.680	0.671	0.289	0.406	0.118	0.307
Hap16	32.405	0.687	0.281	0.410	0.100	0.303

ENC: Effective number of codons, CBI: codon bias index, SChi2: scaled chi-square, G + C2: G + C content at codon position 2, G + C3s: G + C content (synonymous) at codon position 3, G + Cc: G + C content at codon positions. Scaled chi-square computed using Yates' correction.

usage and amino acid occurrence was observed. This relationship was assessed by calculating the number of G + C-rich codons (Pro, Ala, Arg, and Gly) and A + T-rich codons (Phe, Ile, Met, Tyr, Asn, and Lys) and then calculating their ratios (50). In *C. pygmeus*, this ratio was 0.71, which is similar to that found in other cephids (0.67-0.70), higher than that of *Apis mellifera* (0.46), and lower than that of *Drosophila* (0.89).

Amino acid composition and structure

At the intraspecific level, 2 amino acid substitutions were identified: the first at codon position 6 from alanine to threonine, and the next at codon position 41 from serine to lysine (Table 7). Amino acid substitution in codon position 6 occurred between hydrophobic and polar groups, while amino acid substitution in codon position 41 resulted in the replacement of 1 polar group with another. The amino acid changes occurred between the external loop E2 and internal loop I2 regions. A total of 5 amino acid substitutions were identified among *Cephus* species, 3 of which were the replacement of 1 hydrophilic amino acid with another (Table 7). Changes between serine and threonine or to serine were relatively frequently observed among cephid species (Table 7). These changes appear to be evolutionary conservative, as they have similar

Table 7. Variable positions in 229 amino acids of the COI gene for 16 haplotypes found in *Cephus pygmeus* and other cephid specimens.

							1	1	2	2
			3	4	4	8	2	4	0	0
	5	6	8	1	9	2	4	2	3	5
Cl.haemorrhoidalis	P	S	V	N	T	T	S	S	W	A
T.tabidus	.	.	M	.	S	.	T	T	Y	S
C.sareptanus	S	.	L	K	S	S	.	T	L	S
C.pulcher	.	.	R	K	S	.	.	T	L	S
Hap1	.	A	.	S	S	.	.	T	L	S
Hap2	.	A	.	K	S	.	.	T	L	S
Hap3	.	A	.	K	S	.	.	T	L	S
Hap4	.	A	.	K	S	.	.	T	L	S
Hap5	.	T	.	S	S	.	.	T	L	S
Hap6	.	A	.	S	S	.	.	T	L	S
Hap7	.	A	.	S	S	.	.	T	L	S
Hap8	.	A	.	S	S	.	.	T	L	S
Hap9	.	A	.	S	S	.	.	T	L	S
Hap10	.	A	.	K	S	.	.	T	L	S
Hap11	.	A	.	S	S	.	.	T	L	S
Hap12	.	A	.	K	S	.	.	T	L	S
Hap13	.	A	.	S	S	.	.	T	L	S
Hap14	.	A	.	S	S	.	.	T	L	S
Hap15	.	A	.	K	S	.	.	T	L	S
Hap16	.	A	.	S	S	.	.	T	L	S

molecular volume and hydrophobicities (39,51). Such changes may favor the increase of the AT content of the gene (31). These amino acid substitutions among species may also be related to adaptation to different host plants or environments. Among *Cephus* species, 3 regions of protein at the amino acid level, from 7 to 37, from 42 to 81, and from 83 to 229, are completely conserved and are similar in other cephid genera.

The aligned COI amino acid sequence consists of 13 regions [6 transmembrane helixes (M4-M9), 4 external loops (E2-E5), and 3 internal loops (I2-I4)] (17). These regions are highly conserved at the intraspecific and interspecific levels within Cephini, but display differences in variability level when compared with those of *Apis* and *Drosophila*. This may be related to the function of the region. The 3 polar amino acids (Thr-73, Thr-80 and Lys-83) necessary for translocation activity in the cytochrome oxidase proton conduction channel at the transmembrane helix M8 region are conserved in *Cephus* species, like all other organisms (17,52). Data generated from mtDNA or nuclear DNA have been utilized for different purposes in recent years (53,54).

A detailed examination of the partial COI gene region in *C. pygmeus* and other cephid species provides a valuable resource as a molecular marker. The analyzed region here is very informative for estimating intra- and interspecific levels as it includes both completely conserved and variable regions. An apparent relationship between the variations in nucleotide composition and changes in the amino acid composition of the gene was presented. Such a relation can occur at the level of nucleotide bias, which may affect the amino acid sequence, or it may have originated from the selection of content of the amino acids (37). We conclude that there is a mutation bias at the nucleotide level, as the differences in A + T content are highest at codon position 3 (Table 3). If the mutational bias had occurred at the amino acid level, then we would have seen the changes at codon positions 1 or 2. The presence of a strong bias for synonymous substitutions both within and among species, and a lower rate of bias along with an increase in substitutions at nonsynonymous sites in genus *Cephus* and tribe Cephini, are attributed to a clear mutational bias at the nucleotide level within species. The compositional bias is also reflected in the extreme codon bias for the COI gene as well as in the types of amino acid substitutions that have occurred during the evolution of the gene (39).

Acknowledgements

This study was partly supported by the Scientific and Technological Research Council of Turkey (TÜBİTAK) and Cumhuriyet University via research grants for projects TOGTAG-2717 and CÜBAP F-147, respectively.

Corresponding author:

Hasan Hüseyin BAŞIBÜYÜK

Department of Biology,

Faculty of Science,

Cumhuriyet University,

58140 Sivas - TURKEY

E-mail: hbbuyuk@cumhuriyet.edu.tr

References

1. Benson RB. The sawflies (Hymen., Symphyta) of Israel. Bull Res Counc of Israel 4: 351-356, 1955.
2. Benson RB. Hymenoptera from Turkey, Symphyta. Bull Br Mus (Nat Hist) Entomol 22: 111-207, 1968.
3. Rashwani A. Introduction to the Major Insect Pests of Wheat and Barley in the Middle East and North Africa, Technical Manual 9. ICARDA. Syria; 1984.
4. Zhelochovtsev AN, Zinovjev AG. Suborder Symphyta (Chalastogastra) –sawflies and horntails. In: Zhelokhovcev AN, Tobias VI, Kozlov MA. eds. Keys to the Insects of the European Part of the USSR, Volume 3, Hymenoptera, Part 6. Leningrad. Nauka; 1988: pp. 7-234.
5. Stange LA. The stem sawflies of Florida (Hymenoptera: Cephidae). Entomology Circular No. 382. Florida Department of Agriculture & Consumer Services. Tallahassee, FL, USA; 1997.

6. Shanower TG, Hoelmer KA. Biological control of wheat stem sawflies: past and future. J Agr Urban Entomol 21: 197-221, 2004.

7. Korkmaz EM, Budak M, Örgen SH et al. New records and a checklist of Cephidae (Hymenoptera: Insecta) of Turkey with a short biogeographical consideration. Turk J Zool 34: 203-211, 2010.

8. Vandewoestijne S, Baguette M, Brakefield PM et al. Phylogeography of *Aglais urticae* (Lepidoptera) based on DNA sequences of the mitochondrial COI gene and control region. Mol Phylogenet Evol 31: 630-646, 2004.

9. Avise JC. Molecular Markers, Natural History, and Evolution. Chapman and Hall. New York; 1994.

10. Avise JC. Phylogeography: The History and Formation of Species. Harvard University Press. Cambridge, MA, USA; 2000.

11. Caterino MS, Cho S, Sperling FAH. The current state of insect molecular systematics: a thriving Tower of Babel. Annu Rev Entomol 45: 1-54, 2000.

12. Simmons RB, Weller SJ. Utility and evolution of cytochrome b in insects. Mol Phylogenet Evol 20: 196-210, 2001.

13. Tamura K. The rate and pattern of nucleotide substitution in *Drosophila* mitochondrial DNA. Mol Biol Evol 9: 814-825, 1992.

14. Moriyama EN, Powell JR. Synonymous substitution rates in *Drosophila*: mitochondrial versus nuclear genes. J Mol Evol 45: 378-391, 1997.

15. Goto SG, Kimura MT. Phylogenetic utility of mitochondrial COI and nuclear Gpdh genes in *Drosophila*. Mol Phylogenet Evol 18: 404-422, 2001.

16. Hebert PD, Cywinska A, Ball SL et al. Biological identification through DNA barcodes. Proc R Soc Lond B 270: 313-321, 2003.

17. Lunt DH, Zhang DX, Szymura JM et al. The insect cytochrome oxidase I gene: evolutionary patterns and conserved primers for phylogenetic studies. Insect Mol Biol 5: 153-165, 1996.

18. Roe AD, Sperling FAH. Patterns of evolution of mitochondrial cytochrome c oxidase I and II DNA and implications for DNA barcoding. Mol Phylogenet Evol 44: 325-345, 2007.

19. Marino A, Mantovani B, Carpana E et al. ND2 and CO1 mitochondrial genes in *Apis mellifera* L.: a molecular approach to Mediterranean populations monitoring. Biologia Evoluzionistica Sperimentale. University of Bologna, Italy; 2002.

20. Lewis DL, Farr CL, Kaguni LS. *Drosophila melanogaster* mitochondrial DNA: completion of the nucleotide sequence and evolutionary comparisons. Insect Mol Biol 4: 263-278, 1995.

21. Sambrook J, Russell DW. Molecular Cloning: A Laboratory Manual, 3rd ed. Cold Spring Harbor Laboratory Press. Cold Spring Harbor, NY, USA; 2001.

22. Simon C, Frati F, Beckenbach A et al. Evolution, weighting and phylogenetic utility of mitochondrial gene sequences and compilation of concerted polymerase chain reaction primers. Ann Entomol Soc Am 87: 651-701, 1994.

23. Normark BB, Jordal BH, Farrel BD. Origin of haplodiploid beetle lineage. Proc R Soc Lond B 266: 2253-2259, 1999.

24. Kumar S, Tamura K, Nei M. Molecular Evolutionary Genetic Analysis, Version 1.0. Pennsylvania State University. University Park, PA, USA; 1993.

25. Jukes TH, Cantor CR. Evolution of Protein Molecules. In: Munro HN. ed. Mammalian Protein Metabolism. Academic Press; 1969: pp. 21-132.

26. Rozas J, Sánchez-DelBarrio JC, Messeguer X et al. DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 19: 2496-2497, 2003.

27. Xia X, Xie Z, Salemi M et al. An index of substitution saturation and its application. Mol Phylogenet Evol 26: 1-7, 2003.

28. Xia X, Xie Z. DAMBE: Software package for data analysis in molecular biology and evolution. J Hered 92: 371-373, 2001.

29. Tamura K, Dudley J, Nei M et al. MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0. Mol Biol Evol 24: 1596-1599, 2007.

30. Swofford, DL. PAUP*: Phylogenetic analysis using parsimony (*and other methods), ver. 4.0b10. Sinauer Associates. Sunderland, MA, USA; 2001.

31. de Brujin MHL. *Drosophila melanogaster* mitochondrial DNA, a novel gene organization and genetic code. Nature 304: 234-241, 1983.

32. Erpenbeck D, Hooper JNA, Wörheide G. CO1 phylogenies in diploblasts and the 'Barcoding of Life' – are we sequencing a suboptimal partition? Mol Ecol Notes 6: 550-553, 2005.

33. Palenko MV, Mukha DV, Zakharov IA. Intraspecific and interspecific variation of the mitochondrial gene of cytochrome oxidase I in ladybirds (Coleoptera: Coccinellidae). Russ J Genet 40: 148-151, 2004.

34. Juan C, Oromi P, Hewitt GH. Phylogeny of the genus *Hegeter* (Tenebrionidae, Coleoptera) and its colonization of the Canary Islands deduced from cytochrome oxidase I mitochondrial DNA sequences. Heredity 76: 392-403, 1996.

35. Johnson SB, Waren A, Vrijenhoek RC. DNA barcoding of *Lepetodrilus limpets* reveals cryptic species. J Shellfish Res 27: 43-51, 2008.

36. Kocher TD, Conroy JA, McKaye KR et al. Evolution of NADH dehydrogenase subunit 2 in east African cichlid fish. Mol Phylogenet Evol 4: 420-432, 1995.

37. Jermiin LS, Crozier RH. The cytochrome b region in the mitochondrial DNA of the ant *Tetraponera rufoniger*: sequence divergence in Hymenoptera may be associated with nucleotide content. J Mol Evol 38: 282-294, 1994.

38. Dowton M, Austin AD. Increased genetic diversity in mitochondrial genes is correlated with the evolution of parasitism in the Hymenoptera. J Mol Evol 41: 958-961, 1995.

39. Dowton M, Austin AD. 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, 1997.

40. Leys R, Cooper SJB, Schwarz P. Molecular phylogeny of the carpenter bees, genus *Xylocopa* (Hymenoptera: Apidae), based on mitochondrial DNA sequences. Mol Phylogenet Evol 17: 407-418, 2000.

41. Danforth BN, Conway L, Ji S. Phylogeny of eusocial *Lasioglossum* reveals multiple losses of eusociality within a primitively eusocial clade of bees (Hymenoptera: Halictidae). Syst Biol 52: 23-36, 2003.

42. Howland DE, Hewitt GM. Phylogeny of the Coleoptera based on mitochondrial cytochrome oxidase I sequence data. Insect Mol Biol 4: 203-215, 1995.

43. Kim I, Cha SY, Yoon MH et al. The complete nucleotide sequence and gene organization of the mitochondrial genome of the oriental mole cricket, *Gryllotalpa orientalis* (Orthoptera: Gryllotalpidae). Gene 353: 155-168, 2004.

44. Sneath PHA, Sokal RR. Numerical Taxonomy. Freeman. San Francisco, CA, USA; 1973.

45. Karlin S, Mrazek J. What drives codon choices in human genes? J Mol Biol 262: 459-472, 1996.

46. Herbeck JT, Novembre J. Codon usage patterns in cytochrome oxidase I across multiple insect orders. J Mol Evol 56: 691-701, 2003.

47. Liu H, Beckenbach AT. Evolution of the mitochondrial cytochrome oxidase II gene among 10 orders of insects. Mol Phylogenet Evol 1: 41-52, 1992.

48. Frati F, Simon C, Sullivan J et al. Evolution of the mitochondrial cytochrome oxidase II gene in Collembola. J Mol Evol 44: 145-158, 1997.

49. Castro LR, Dowton M. The position of the Hymenoptera within the Holometabola as inferred from the mitochondrial genome of *Perga condei* (Hymenoptera: Symphyta: Pergidae). Mol Phylogenet Evol 34: 469-479, 2005.

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

51. French S, Robson B. What is a conservative substitution? J Mol Evol 19: 171-175, 1983.

52. Gennis RB. Site-directed mutagenesis studies on subunit I of the aa3-type cytochrome c oxidase of *Rhodobacter sphaeroides*: a brief review of progress to date. Biochim Biophys Acta 17: 184-7, 1992.

53. Işık K, Goodfellow M. Molecular fingerprinting of some clinically significant *Nocardia* and related strains by restriction polymorphism ribosomal RNA analyses. Turk J Biol 34: 355-365, 2010.

54. Okay S, Özcengiz G. Molecular cloning, characterization, and homologous expression of an endochitinase gene from *Bacillus thuringiensis* serovar *morrisoni*. Turk J Biol 35: 1-7, 2011.