Higher-level phylogeny of mosquitoes (Diptera: Culicidae): mtDNA data support a derived placement for *Toxorhynchites*

ANDREW MITCHELL, FELIX A. H. SPERLING and DONAL A. HICKEY

Insect Syst.Evol.

Mitchell, A., Sperling, F. A. H. & Hickey, D. A.: Higher-level phylogeny of mosquitoes (Diptera: Culicidae): mtDNA data support a derived placement for *Toxorhynchites*. *Insect Syst. Evol.* 33: 163-174. Copenhagen, July 2002. ISSN 1399-560X.

We assess the potential of complete coding sequences of the mitochondrial *cytochrome oxidase* genes (*COI* and *COII*) and the intervening tRNA-Leucine gene for use in mosquito higher-level systematics, and apply this data to an outstanding question: the phylogenetic affinities of *Toxorhynchites*. Traditionally placed in its own subfamily and regarded as sister group to Culicinae, recent morphological data instead have suggested that this distinctive genus belongs well within the Culicinae. Published molecular systematic studies seemingly conflict with this new morphological data or are ambiguous. The mitochondrial gene data that we present show good potential for elucidating suprageneric relationships in Culicidae, and strongly support the placement of *Toxorhynchites* well within the Culicinae. Reexamination of published data sets suggests that there is no substantive conflict among data sets on this issue.

Andrew Mitchell, School of Molecular and Cellular Biosciences, University of Natal, Private Bag X01, Scottsville, 3209, South Africa (mitchella@nu.ac.za).
Felix A. H. Sperling, Department of Biological Sciences, University of Alberta, Edmonton, Alberta, T6G 2E9, Canada (felix.sperling@ualberta.ca).
Donal A. Hickey, Department of Biology, University of Ottawa, Ottawa, Ontario, K1N 6N5, Canada (dhickey@uottawa.ca).

Introduction

Mosquito taxonomy has received intense attention as a direct result of the immense medical importance of many mosquitoes as vectors of disease-causing organisms. The realisation that accurate delimitation of species boundaries and an in-depth knowledge of geographic variation are essential for vector control has resulted in a focus on species-level taxonomy in Culicidae. The higher-level phylogeny of Culicidae is much more poorly understood. Among the reasons for this is the importance of nomenclatural stability in this important group, which understandably has led to conservative nomenclature. One could also argue, however, that the widespread importance of mosquitoes is in itself a strong motivation for providing an up-to-date classification that more closely reflects phylogeny. Nevertheless, the earlier phenetic work has largely pre-empted later cladistic studies, and most systematists have followed the classification of Edwards (1932) with little modification. Rigorous phylogenetic methodology has only recently been applied to the problem of relationships among subfamilies by Harbach & Kitching (1998). However, their novel results have yet to gain wide acceptance. This is largely because of apparent conflict between their data and published data sets, both morphological and molecular. In this study we assess the potential of mitochondrial DNA sequence data for resolving contentious issues in mosquito higher-level systematics by examining the phylogenetic placement of *Toxorhynchites*.

Recent classifications of Culicidae recognize three subfamilies: Anophelinae, with 3 genera, Culicinae, with 34 genera arranged in 10 tribes, and Toxorhynchitinae, with a single genus, *Toxorhynchites*. The distinctiveness of *Toxorhynchites* was recognized early on, because the species are among the largest and most spectacular of all mosquitoes. *Toxorhynchites* females do not take blood meals, and oviposit in small, natural plant containers where the larvae are predatory (Steffan & Evenhuis 1985). *Toxorhynchites* species have even been used as biological control agents of other mosquitoes, though with limited success (Steffan

& Evenhuis 1981). Despite these differences, a few authors considered the apparent peculiarity of the Toxorhynchitinae to have been overemphasized. Belkin (1962), for example, treated the group as a tribe of equal rank to the ten tribes of Culicinae, and considered them closely related to the Sabethini. This hypothesis was supported by Harbach & Kitching's (1998) morphology-based cladistic analysis of all mosquito genera. Molecular data, in contrast, have seemingly supported the traditional placement of *Toxorhynchites* (Besansky & Fahey 1997) or have been equivocal (Miller et al. 1997).

Which of these alternative placements for *Toxorhynchites* is better supported? We present sequence data of the complete mitochondrial *cytochrome oxidase* genes, *COI* and *COII*, and the intervening tRNA-Leucine gene (a total of 2.3kb), in order to assess the utility of these genes for mosquito higher-level phylogenetics, and to provide additional data bearing on this problem.

Materials and methods

Abbreviations. – *CO*, *cytochrome oxidase*; GTR, general time reversible; ML, maximum likelihood; MP, maximum parsimony; mp, most parsimonious.

Taxon sampling. – Taxa sampled and sources of material are listed in Tab. 1. Specimens were collected in the wild or taken from laboratory colonies and then either frozen live at -70°C or preserved a few days in 95% ethanol before being frozen at -70°C. Identifications of wild-collected material were provided by D. M. Wood (Agriculture and Agri-Food Canada, Ottawa, Canada). Our taxon sample included eleven mosquito species, including four anophelines, five culicines (from four tribes), and two toxorhynchitines. The outgroup comprised two species, including a representative of Chaoboridae, most likely the sister group of Culicidae (Oosterbroek & Courtney 1995) or at least very closely related (Pawlowski et al. 1996), and a more distantly related culicomorph (a blackfly, Simuliidae). Published sequences were obtained from GenBank for *Anopheles quadrimaculatus* (Mitchell et al. 1993) and *A. gambiae* (Beard et al. 1993). All other sequences are new data.

Laboratory protocols. – For most specimens, DNA was extracted from individual specimens using a phenol/chloroform procedure as described by Sperling et al. (1994). For two specimens (*Sabethes cyaneus* and *Toxorhynchites* sp.) DNA was extracted using the Qiagen DNeasy Tissue Kit (Qiagen Inc., Valencia, California). PCR was accomplished using the protocol of Sperling et al.

Table 1. Taxon sampling.

Taxon	Source[a]	GenBank Accession No.[b]
Culicidae		
Anophelinae		
Anopheles (Anopheles) earlei	Ottawa, Canada; D. M. Wood	AF425843
Anopheles (Anopheles) quadrimaculatus	Mitchell et al. (1993)	NC000875
Anopheles (Cellia) gambiae	Beard et al. (1993)	NC002084
Anopheles (Cellia) stephensi	LC, University of Maryland; V. Kulasekara	AF425844
Culicinae		
Aedes atropalpus	LC, University of Ottawa; T. Arnason	AF425845
Aedes aegypti	LC, University of Ottawa; D. Hickey	AF425846
Culex tarsalis	LC, Queens University; V. Walker	AF425847
Culiseta impatiens	Ottawa, Canada; D. M. Wood	AF425848
Sabethes cyaneus	LC, University of Notre Dame; N. Besansky	AF425840
Toxorhychitinae		
Toxorhynchites rutilus	York Co., Pennsylvania; D. M. Wood	AF425849
Toxorhynchites sp.	Vietnam; D.C. Currie	AF425850
Chaoboridae		
Chaoborus sp.	Ottawa, Canada; D. M. Wood	AF425842
Simuliidae		
Cnephia dakotensis	Ottawa, Canada; D. M. Wood	AF425841

[a]LC = Laboratory colony
[b]The complete alignment is also available from TreeBase (http://www.herbaria.harvard.edu/treebase/).

(1994), using the end primers TY-J-1460 (=K698: 5' -TAC AAT TTA TCG CCT AAA CTT CAG CC -3') and TK-N-3782 (=Eva: 5'- GAG ACC ATT ACT TGC TTT CAG TCA TCT -3') and a variety of internal primers originally developed by us from a comparison of the GenBank sequences of *Drosophila yakuba*, *Anopheles gambiae*, and *A. quadrimaculatus*. For three species, *Cnephia dakotensis*, *Chaoborus* sp., and *Sabethes cyaneus*, the above end primers did not give amplification products. For these taxa we developed primers farther out from the *COI/COII* region. The upstream primer used for *Cnephia* and *Sabethes* was TW-J-1305 (=K912: 5'- GTT AAA TAA ACT AAT AGC CTT CAA AGC TG -3') and for *Chaoborus* was TY-J-1460 I (=K911: 5'- TTA CAA TTT CTT ACT TAA GTT CAG CC -3'). The downstream primer used for both *Cnephia* and *Chaoborus* was TD-N-3862 (5'- GGC CGT CTG ACA AAC TAA TGT TAT -3') from Simon et al. (1994). For most specimens PCR fragments were sequenced directly using the Taq DyeDeoxy™ Terminator Cycle Sequencing System (Applied Biosystems, Inc.) and fractionated on an ABI 373 automated DNA sequencer. For two specimens (*Sabethes cyaneus* and *Toxorhynchites* sp.) PCR fragments were sequenced directly using the DYEnamic™ ET terminator cycle sequencing kit (Amersham Pharmacia Biotech; Cleveland, Ohio) and fractionated on an ABI 377 automated DNA sequencer.

Data analysis. – Sequences were assembled into contiguous arrays using SeqEd (Applied Biosystems, Inc.; Foster City, California) and Sequencher (GeneCodes Corp.; Ann Arbor, Michigan). Alignment of the protein-coding sequences and most of the intervening tRNA region was trivial, owing to the absence of insertions or deletions, and therefore it was performed by hand using the software Se-Al (Rambault 1996). The aligned DNA sequences were translated to amino acids using the insect mitochondrial genetic code, in MacClade 3.03 (Maddison & Maddison 1992). Alignment-ambiguous sites were excluded prior to data analysis for both the nucleotide and amino acid data sets. The aligned data set is deposited in Treebase (http://www.herbaria.harvard.edu/treebase/).

The data from published and new molecular data sets were not combinable because of differences in taxon sampling, therefore separate phylogenetic analyses were performed on all data sets. All phylogenetic analyses of molecular data were performed using identical protocols. All phylogenetic analyses were performed using PAUP* versions 4.0b4a and 4.0b8 Altivec (Swofford 1999). Incongruence length difference (ILD) tests (Farris et al. 1994) consisted of 1000 heuristic search repetitions, each with 10 random addition sequences of taxa. A branch-and-bound maximum parsimony (MP) search was performed on our mtDNA data set. Further MP analyses on our data, and all other MP analyses, consisted of 1000 random addition sequences of taxa performed on all nucleotide sequences and, in the case of the *COI* and *COII* genes, also on the inferred amino acid sequences (740 characters). Bootstrap analyses employed 1000 repetitions, each with 10 random addition sequences of taxa. Constraint searches were carried out for each data set to determine the number of additional steps needed to accommodate alternative phylogenetic hypotheses. Bremer support (or 'decay') indices (Bremer 1988) were calculated for the morphological data set (Harbach & Kitching 1998).

To assess the effects of biased base composition on phylogeny reconstruction we performed distance analyses based on the LogDet (Lockhart et al. 1994) and compared these trees with those derived using the general time reversible (GTR) model.

Maximum likelihood (ML) analyses were performed on all molecular data sets, and followed an identical protocol. Following Frati et al. (1997), 16 models (four substitution models permuted with four rate-distribution models) were evaluated on the most parsimonious (mp) trees for each data set. Likelihood ratio tests were used to determine which models had significantly higher likelihood scores. The best model was then chosen from the set of models with the best score, with preference given to the model with the fewest free parameters. This model was selected for further analyses, with all parameters estimated once from the data on the mp trees, and fixed subsequently. Heuristic searches performed under the ML criterion used 100 random addition sequences for taxa with TBR branch swapping. Bootstrap analyses employed 1000 repetitions, each with a single random addition sequence of taxa. As for MP, constraint searches were carried out for each data set to determine the ML score for alternative phylogenetic hypotheses. Kishino-Hasegawa tests (Kishino & Hasegawa 1989), Shimodaira-Hasegawa tests (Shimodaira & Hasegawa 1999), Templeton

Table 2. Base frequencies (%) for the 877 variable sites.

Taxon	A	C	G	T
Cnephia dakotensis	33.9	14.5	11.7	40.0
Chaoborus sp.	33.9	15.3	8.1	42.8
Anopheles quadrimaculatus	39.0	13.3	9.6	38.1
Anopheles earlei	39.0	9.9	8.3	42.8
Anopheles gambiae	37.9	12.3	9.2	40.6
Anopheles stephensi	38.8	10.5	8.1	42.6
Aedes atropalpus	37.7	10.4	6.7	45.2
Aedes aegypti	35.8	13.8	5.9	44.5
Culex tarsalis	35.8	8.3	7.5	48.3
Culiseta impatiens	36.7	11.3	7.0	45.0
Sabethes cyaneus	34.6	9.7	3.5	52.2
Toxorhynchites rutilus	39.2	11.3	3.0	46.5
Toxorhynchites sp.	39.1	9.4	2.5	49.0
Mean	37.0	11.5	7.0	44.4

tests (Templeton 1983) and winning site tests were performed to assess the significance of differences among trees resulting from constrained and unconstrained searches.

Results

COI-COII data

Alignment and data partitions. – *Sabethes cyaneus* shows a different order of genes in this region of the mitochondrial genome. While all other taxa examined in this study show the same gene order as *Anopheles gambiae* (Beard et al., 1993), i.e., tRNA-Tyr, *COI*, tRNA-Leu, *COII*, tRNA-Lys, in *S. cyaneus* the tRNA-Tyr is replaced by tRNA-Trp. Thus the first 39 sites in the *S. cyaneus* sequence could not be aligned with other taxa and were deleted from the alignment. This non-alignable region included six nucleotides corresponding to the first two codons of the *COI* gene in the other taxa.

The complete alignment consisted of 2345 sites. Based on comparison with the mitochondrial genome of *Anopheles gambiae* (Beard et al. 1993) the *COI*, tRNA-Leu, and *COII* genes are found at nucleotide positions 24-1559, 1572-1639, and 1647-2330, respectively. The flanking regions, sites 1-23 (tRNA-Tyr) and 2331-2345 (tRNA-Lys), were almost invariable in the ingroup, but *Cnephia dakotensis* had a large insert in each region, therefore these sites were excluded from the data set. Also excluded were the following sites, which could not be unambiguously aligned: 1560-1571, 1619-1627 and 1640-1646. In total, 66 characters were excluded from all analyses, leaving 2279 sites in the final data set.

An incongruence length difference test performed on three partitions, *COI*, tRNA-Leu, and *COII*, was nonsignificant ($p = 0.65$), therefore all data were combined for analysis.

Base composition. – Chi-square tests for homogeneity of base composition among taxa were performed for various partitions of the data set. For the entire data set (2279 sites) the null hypothesis of homogeneity was rejected ($p = 0.019$). Excluding invariable sites gave a highly significant test result ($p << 0.001$). Tab. 2 shows base composition of the 877 variable sites. The same trends were seen for the *COI* and *COII* genes considered separately but not for the tRNA-Leu gene, which was homogeneous even on exclusion of invariable sites. Base composition was also determined by codon position for the two *CO* genes combined. Only third codon positions showed significantly heterogeneous base composition when invariable sites were counted ($p << 0.001$), and only second codon positions showed homogeneous base composition when invariable sites were excluded ($p = 0.817$).

Table 3. Summary statistics for data partitions mapped onto mp tree for all data (Fig. 3).

	Combined data	CO I			CO II			tRNA-Leu	Amino acids
		nt.1	nt. 2	nt. 3	nt.1	nt. 2	nt. 3		
No. characters	2279	512	512	512	228	228	228	59	740
No. (%) variable	877 (38)	137 (27)	47 (9)	399 (78)	76 (33)	26 (11)	181 (79)	11 (19)	181 (24)
No. (%) inform.[a]	620 (27)	91 (18)	26 (5)	304 (59)	50 (22)	18 (8)	129 (57)	2 (3)	114 (15)
No. steps	2279	301	73	1187	154	37	512	15	360
CI, excl. uninf.[b]	0.45	0.49	0.62	0.43	0.52	0.68	0.41	0.67	0.70
RI	0.37	0.52	0.73	0.28	0.57	0.82	0.27	0.60	0.73

[a]Number of parsimony-informative characters
[b]Consistency index, excluding parsimony-uninformative characters

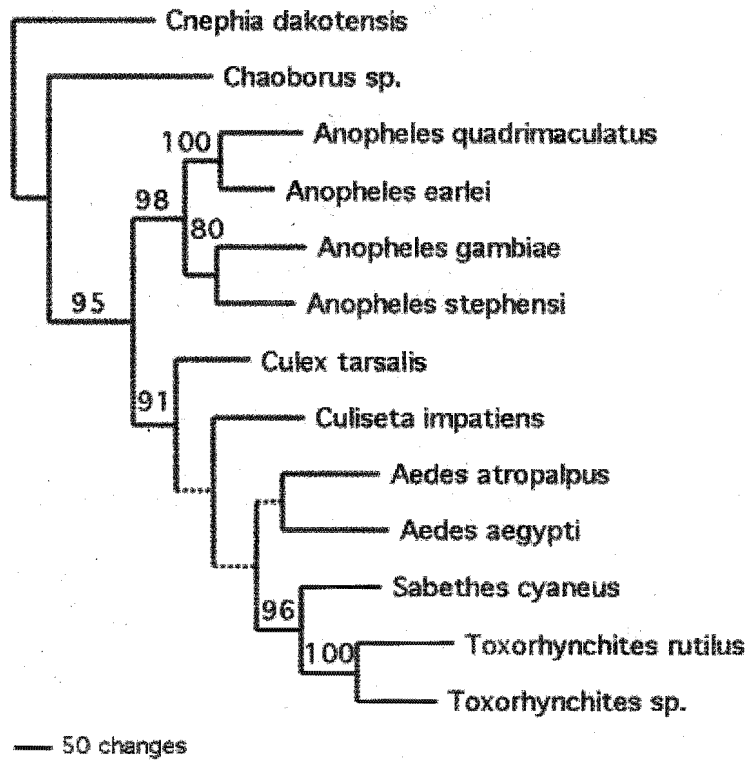

Figure 1. One of three most parsimonious (mp) trees from equally-weighted analysis of all sites, *COI-COII* nucleotides. Length = 2279 steps, CI (excl. uninf.) = 0.45, RI = 0.37. Numbers above branches are bootstrap values. Branches that collapse in the strict consensus tree are indicated by dashed lines.

Maximum parsimony (MP) analyses. – Equally-weighted parsimony analysis of all sites resulted in three most parsimonious (mp) trees of 2279 steps (Fig. 1). Data partitions were mapped onto the mp tree to assess variation in data quality (Tab. 3). Codon positions of *COI* and *COII* had very similar percentages of variable and parsimony-informative characters, and similar consistency indices (CIs) and retention indices (RIs). The tRNA-Leu gene was less variable. The combined *COI/COII* amino acid data had the highest CI and RI of all partitions.

In the ingroup (Culicidae) the mean uncorrected pairwise sequence divergence for all sites was 13.5% with the values ranging from 7.9% between *Anopheles earlei* and *A. quadrimaculatus* to 17.3% between *A. quadrimaculatus* and *Toxorhynchites rutilus*. For ingroup/outgroup comparisons the mean divergence was 18.0%.

The three mp trees from the equally-weighted analysis recovered as monophyletic the Anophelinae, *Anopheles (Anopheles)*, *Anopheles (Cellia)*, and *Toxorhynchites* (Fig. 1). *Aedes* was monophyletic in only one of the three trees. Anophelinae was basal within Culicidae, and Culicinae was paraphyletic with respect Toxorhynchitinae. *Toxorhynchites* was firmly supported as sister-group to *Sabethes* within the Culicinae. A search conducted using logarithmic-weighting produced a single tree which differed only in that *Culex* was the most basal culicine, followed by *Culiseta*, then a monophyletic *Aedes*, though there was no bootstrap support for *Aedes* monophyly.

For the inferred amino acid data set, a single mp tree of 360 steps was recovered (Fig. 2A), which differed from Fig. 1 in two respects: (i) subgenus *Anopheles (Cellia)* was paraphyletic with respect to *Anopheles (Anopheles)*, and (ii) there was strong support for the monophyly of *Aedes*. Further analyses were performed on the amino

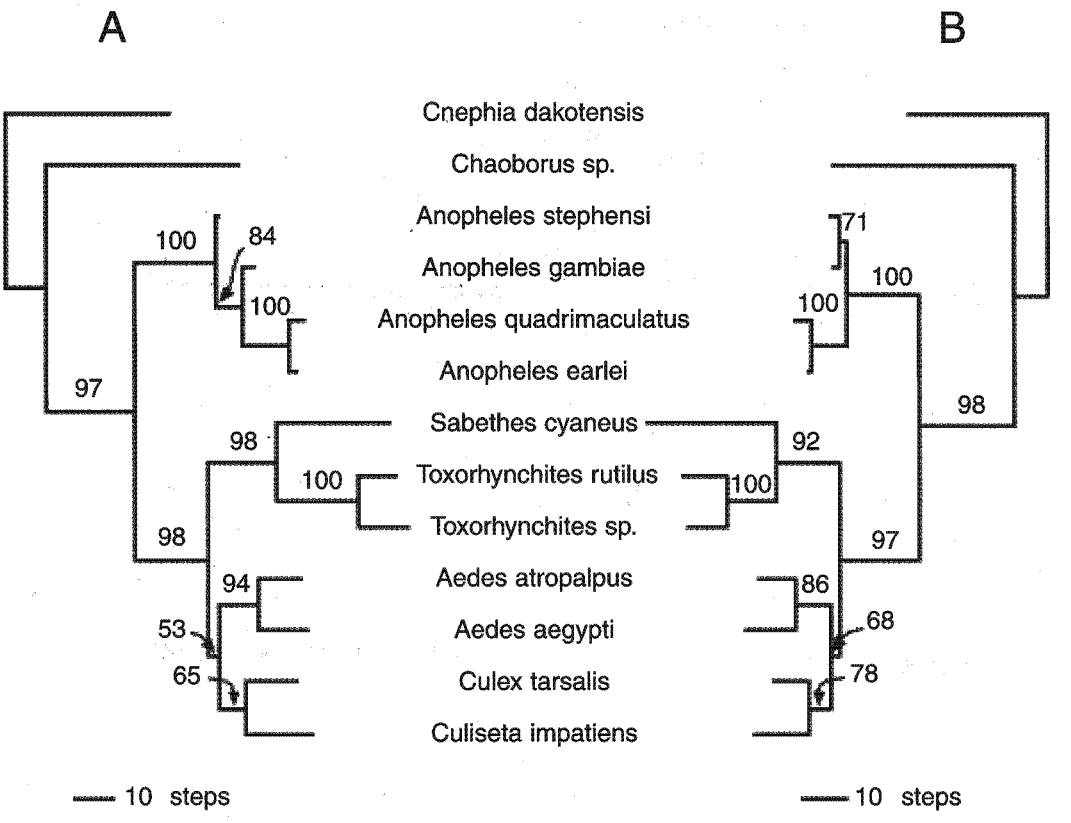

Figure 2. Most parsimonious (mp) trees derived using inferred amino acids for *COI* and *COII*. Numbers above branches are bootstrap values. (A) All amino acids coded as separate character states, length = 360 steps, CI (excl. uninf.) = 0.70, RI = 0.73. (B) Isoleucine, leucine and valine coded as equivalent character states, length = 279 steps, CI (excl. uninf.) = 0.71, RI = 0.76.

acids with leucine, isoleucine, and valine and coded as equivalent character states (see Discussion). The resulting mp tree (Fig. 2B) recovered *Anopheles (Cellia)* as monophyletic with 70% bootstrap support, otherwise the topology was identical to Fig. 2A.

Minimum evolution (ME) analyses. – Unlike other distance measures, the LogDet model (Lockhart et al., 1994) is considered to be relatively insensitive to base composition bias effects. ME analyses were therefore performed using both LogDet and general time reversible (GTR; Lanave et al., 1984) models. Both trees recovered similar relationships to the mp trees, with only weakly supported differences in the relationships among basal culicines. *Toxorhynchites* was always strongly supported as sister-group to *Sabethes*, as in the MP analyses. Bootstrap support for the monophyly of subgenus *Anopheles (Cellia)* was very strong under the LogDet model (97%) but only moderate (78%) weak under the GTR model, as in the parsimony analyses.

Maximum likelihood (ML) analyses. – Likelihood ratio tests showed that the most appropriate substitution model for this data set was the general time reversible model (Lanave et al., 1984), with invariable sites and gamma-distributed rates, i.e., GTR + I + Γ. For this model the mp tree in Fig. 1 had a likelihood score of -ln L = 12,387.81, significantly better than all other models tested. (The next best model was GTR + Γ; 2Δ ln L = 9.05, 1 d.f., $p < 0.005$). Heuristic searches performed under this model produced a single ML tree with a score of -ln L = 12,386.63 (Fig. 3).

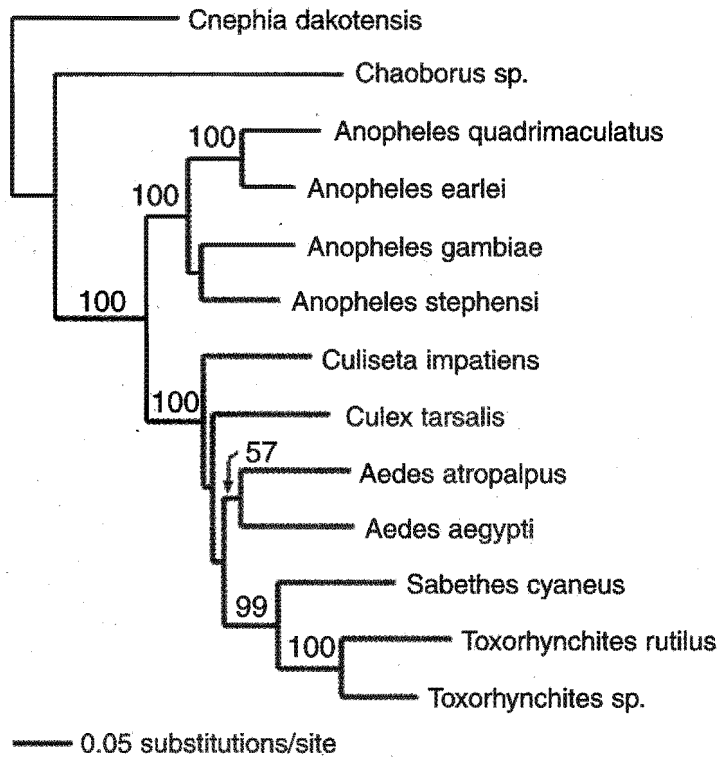

Figure 3. Maximum likelihood (ML) tree for *COI-COII* nucleotides, obtained under the GTR + I + Γ model, -ln L = 12,386.63. Numbers above branches are bootstrap values. Model parameters: Nucleotide frequencies: A=0.31837, C=0.13402, G=0.13385, T=0.41376; Substitution rate matrix: A/C=3.023, A/G=18.37, A/T=21.36, C/G=5.678, C/T=60.29, G/T=1; Proportion of invariable sites: 0.47238; Γ- distribution shape parameter (α)=0.88003.

A search was performed with the Culicinae constrained to be monophyletic, and the ML tree obtained had a score of ln L = 12,414.93. This tree was compared to all the parsimony and likelihood trees obtained for this data set and described previously. The Shimodaira-Hasegawa test (S-H test) was used in addition to the Kishino-Hasegawa test (K-H test) because of recent concerns over the bias of the former test (Goldman et al., 2000). The tree in which Culicinae was constrained to be monophyletic was the only tree with a significantly lower likelihood than the ML tree (K-H test, p = 0.019 and S-H test, p = 0.027).

Data set of Besansky & Fahey (1997)

Besansky & Fahey (1997) analysed data from the nuclear, protein-encoding *white* gene for 13 species of mosquitoes and two chaoborid midges. Excluding third codon positions and applying successive approximations character weighting (Farris 1969) the authors obtained a mp tree which was congruent with the traditional classification of Culicidae, i.e., Anophelinae was basal and Toxorhynchitinae was sister group to a monophyletic Culicinae.

Besansky & Fahey's (1997) alignment was produced using the PILEUP program in the GCG Sequence Analysis Package (Genetic Computer Group 1994), using the default settings. The authors noted an alignment-ambiguous region between positions 117 and 189, but decided to keep this region in their data set because it had little effect on tree topology. They also coded the gaps within this region as additional two-state characters at the end of the data set. On examination of the inferred amino acid sequences it appeared to us that *Toxorhynchites* was too diver-

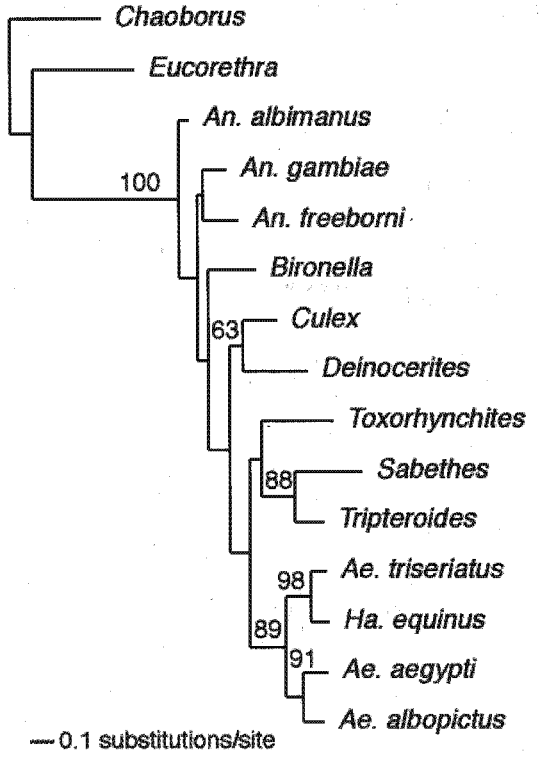

Figure 4. Maximum likelihood (ML) tree for the *white* gene data set (Besansky & Fahey 1997) obtained under HKY85 + I + Γ model, -ln L = 5,682.76. Numbers above branches are bootstrap values under ML criterion.

gent in this region, as well as in the flanking regions, to be aligned reliably with the other sequences. Given the lack of clear homology for these characters, we tested the effects of excluding the alignment-ambiguous region from analyses, and not adding binary characters coding for the gaps to the data set.

For our reanalysis we excluded from the data set a block of 99 alignment-ambiguous nucleotides, corresponding to positions 100-198 in Besansky & Fahey's (1997) alignment (their fig. 3). We repeated the parsimony analysis on this reduced data set, excluding third codon positions. We found 2 mp trees of 383 steps, the strict consensus of which was identical to Besansky & Fahey's fig. 5. However, applying successive approximations character weighting (SACW) to this reduced data set resulted in a different topology to that obtained when Besansky & Fahey (1997) applied SACW to the full data set (their fig. 6). Rather than being placed as sister group to the Culicinae, *Toxorhynchites* was placed within the Culicinae, specifically within the Sabethini as sister group to *Sabethes*. There was substantial support for this placement, albeit under SACW: a 97% bootstrap value for *Toxorhynchites* being sister group to *Sabethes*, and a 99% bootstrap value for the monophyly of (Sabethini + *Toxorhynchites*).

We determined the most appropriate ML model for this reduced data set, including third codon positions, to be that of Hasegawa et al. (1985), with invariable sites and gamma-distributed rates, i.e., HKY85 + I + Γ. Under this model we found a single ML tree (Fig. 4) in which Anophelinae was the most basal group, although paraphyletic with respect to all other mosquitoes, and *Toxorhynchites* was placed well within Culicinae, as sister group to Sabethini. Pairwise Kishino-Hasegawa tests revealed no significant differences ($p = 0.84$ and $p = 0.60$, respectively) between the ML tree and trees resulting from searches in which either Anophelinae was constrained to be monophyletic, or *Toxorhynchites* was constrained to be the sister group to Culicinae.

Data set of Harbach & Kitching (1998)

All trees produced by Harbach & Kitching (1998) placed *Toxorhynchites* well within the Culicinae, although there was substantial variation among trees in the tribal-level relationships within Culicinae. We calculated Bremer support (or 'decay') indices for this data set. Using PAUP*4.0b2a (Swofford 1999) we found 130460 trees of length ≤ 351 steps, i.e., up to one step longer than the mp tree. The strict consensus of these trees had seven of 36 nodes resolved within the ingroup (i.e., all other nodes had decay indices of 1). The most strongly supported clade was Anophelinae, with a decay index of eight. The Culicinae clade, which contained *Toxorhynchites*, had a decay index of five, although it cost only three steps to place *Toxorhynchites* as sister-group to the Culicinae. Kishino-Hasegawa tests, Templeton tests and winning site tests confirmed that this difference in tree lengths was not statistically significant ($p > 0.5$ for all tests).

Discussion

For the mtDNA data set, we noted heterogeneity of base composition in first and third codon posi-

tions. An apparent phylogenetic trend in base composition is discernible in Tab. 2: The proportion of guanine present decreases as one ascends the tree, starting at 11.7% in the outgroup, decreasing to 8-9% in Anophelinae, 5-7% in Culicinae other than *Sabethes*, and finally 2-3% in *Sabethes* and *Toxorhynchites*. Base composition biases are known to affect phylogenetic inferences. The question of whether base composition is a reflection of phylogeny or is driving the homoplasious association of taxa can be addressed by comparing distance trees derived under the LogDet model (Lockhart et al. 1994) with trees derived under conventional models. The LogDet model is robust to among-taxon variation in base composition, therefore if both trees give the same pattern of relationships, one can be reasonably confident that base composition bias is not driving the results. This was indeed the case for this data set, with LogDet and GTR-model distance trees recovering the same pattern of relationships.

The full amino acid data set recovered erroneous relationships within *Anopheles*, placing *Anopheles (Cellia) gambiae* as sister group to *Anopheles (Anopheles)*, with 84% bootstrap support. In our experience (e.g., Mitchell et al. 1997) amino acid characters can reach local saturation before their corresponding nucleotide sequences. This is most likely due to loss of the phylogenetic information in synonymous substitutions and convergence among amino acids (Simmons 2000). Naylor & Brown (1997) noted for a data set comprising complete mitochondrial genome sequences that leucine, isoleucine and valine produced the poorest fits of all the amino acids. We noted many homoplasious substitutions among these amino acids in our data set therefore, in an attempt to trace the source of discrepancy between the amino acid and nucleotide trees, we coded L, I and V as equivalent character states and repeated the analyses. The mp tree derived by this method (Fig. 2B) recovered the expected relationships within *Anopheles*, with moderate support for *Anopheles (Cellia)* monophyly.

Interestingly, for the distance analyses bootstrap support for the monophyly of *Anopheles (Cellia)* was higher (97%) under the LogDet model than under the GTR model (78%), suggesting a possible role played by base composition bias in the poor performance of the full amino acid data set in recovering this clade. Indeed, Foster & Hickey (1999) described how base composition bias can affect phylogeny reconstruction even when it is based on the inferred amino acid sequences. Similarly, among taxon base composition bias might account for the low bootstrap support (47%) for the monophyly of *Anopheles (Cellia)* under the ML criterion.

The convergence of parsimony, distance and likelihood methods on very similar tree topologies gives us confidence in these results. Within Culicidae, our *COI-COII* data set appears to be providing useful phylogenetic signal for reconstructing suprageneric relationships and potentially for intrageneric relationships as well. In contrast, Foley et al. (1998) found that the *COII* gene did not provide adequate resolution of deeper relationships within *Anopheles*. Indeed, plots of uncorrected versus ML-corrected pairwise distances for third codon position sites of *COII* in both data sets (not shown) indicate that these sites are approaching saturation. However, the situation is different for first and second codon positions. This is illustrated by a comparison of uncorrected pairwise sequence divergence values for each codon position of the *COII* gene for Foley et al.'s (1998) ingroup (*Anopheles*) versus our ingroup (all mosquitoes): divergence values at first, second and third codon positions peak at 8.0%, 2.6%, and 32.3%, respectively, for the 34 *Anopheles* sequences, and at 14.5%, 7.5%, and 34.2% for the 11 mosquito sequences in our data set. Third codon position divergences level off at a little more than 30% in both data sets and therefore appear to be at or near local saturation. We note, however, that this does not necessarily preclude their phylogenetic utility at this level (e.g., Orti & Meyer 1996; Baker et al. 2001). In contrast, first and second codon position divergences continue to increase when taxon sampling is extended from *Anopheles* to include the other mosquito genera in our data set, so these sites clearly have not reached saturation in Foley et al.'s data set. Thus it would appear that the greater utility of this gene region for reconstructing suprageneric rather than intrageneric relationships stems largely from the increased first and second codon position divergences (although we note that *COII* third position divergences increase again to 43% when ingroup/outgroup comparisons are made for our data set).

Comparison of the summary statistics for our data (Tab. 3) with those for Foley et al.'s data (Tab. 4) provides further evidence supporting our argument for the phylogenetic utility of this gene

region at deeper levels within Culicidae. For Foley et al.'s data, 79% of the total tree length comes from third codon position sites, which have an RI of 0.448, and 21% of the total tree length comes from first and second positions, with RIs of 0.511 and 0.644, respectively. In our data set, third codon positions of the *COII* gene (RI = 0.27) contribute only 71% of the *COII* tree length while first and second positions (with RIs = 0.66 and 0.81, respectively) contribute 29% of the total tree length. Thus in our data set, which bears on deeper relationships, more data comes from the slower-evolving first and second codon positions. These sites also are less homoplasious in our data set than in Foley et al.'s (1998) data set.

Another factor potentially accounting for the greater phylogenetic information content of our data set versus that of Foley et al. (1998) is the greater number of characters in our data set, which also included the entire *COI* gene, which is more than twice the length of *COII*. In combination, the *COI*, *COII*, and tRNA-Leu genes show much potential for studies of mosquito higher-level phylogeny at many levels. It is therefore still possible that a combined *COI* and *COII* data set will prove informative of the deeper splits in *Anopheles* as well.

Miller et al. (1997) presented data from the nuclear 18S and 5.8S rRNA genes to address phylogenetic relationships of the Culicomorpha. Their taxon sample of four mosquito species included representatives of each of the three subfamilies of Culicidae, thus their data is relevant to the placement of *Toxorhynchites* within the family. Miller et al.'s ML tree placed *Toxorhynchites* as sister group to *Culex*, with *Aedes* sister group to this pair (65% bootstrap support for these three taxa forming a clade) and *Anopheles* basal. Bootstrap support for the placement of *Toxorhynchites* within the Culicinae was weak. Our reanalysis of this data set found that placement of *Toxorhynchites* in the traditional position, as sister-group to all Culicinae, could not be excluded as a significantly less likely hypothesis on the basis of Kishino-Hasegawa tests ($p = 0.97$).

Analyses of the *white* gene data set (Besansky & Fahey 1997) from which third codon positions had been excluded were entirely dependent on inclusion/exclusion of the alignment-ambiguous region and associated binary 'gap characters.' Exclusion of the alignment-ambiguous sites resulted in strong support for the placement of *Toxorhynchites* well within the Culicinae. ML analyses of this data set also place *Toxorhynchites* well within the Culicinae, as sister-group to the Sabethini. However, the phylogenetic signal for this placement was not strong under the ML criterion, and placement of *Toxorhynchites* in the traditional position could not be ruled out.

Krzywinski et al. (2001) examined the higher-level relationships of Anophelinae using mitochondrial DNA (ND5 and *cyt* b genes) and nuclear ribosomal DNA sequences. While the *cyt* b data saturated rapidly and proved uninformative except at the very lowest levels, ND5 and 28S rDNA both were informative of anopheline phylogeny, particularly in combination. Unfortunately (for our purposes) this study did not include non-mosquito outgroups, making it difficult to draw firm conclusions about the phylogenetic affinities of *Toxorhynchites*. Instead the trees were rooted on *Uranotaenia*, presumably because Harbach & Kitching (1998) hypothesized that *Uranotaenia* and *Aedeomyia* are basal to all other Culicinae. These two taxa both are placed basally in Krzywinski et al.'s (2001) ML tree, but *Toxorhynchites* is well separated from them, instead being placed within the Culicinae. This placement is in agreement with both Harbach & Kitching (1998) and our *COI-COII* data.

Krzywinski et al.'s (2001) ND5 data contained much phylogenetic signal despite the small size of the region utilized (525 bp) and the fact that this is one of the fastest-evolving genes in the mitochondrial genome (Clary & Wolstenholme 1985; cited in Krzywinski et al. 2001). We can be confident therefore that the 2.3 kb of more slowly evolving *COI-COII* data presented in this paper will be informative at these levels, and probably considerably deeper.

Harbach & Kitching's (1998) mp trees place *Toxorhynchites* within the Culicinae. However, given that it costs just three extra steps to enforce a sister-group relationship between Culicinae and Toxorhynchitinae, we suggest that this traditional hypothesis has not been strongly refuted by their data set considered on its own.

Considered alone, only our *COI-COII* data set had sufficient resolving power to distinguish statistically between a basal placement for *Toxorhynchites* as sister group to the Culicinae, and a more derived placement well within the Culicinae (the preferred hypothesis), based on Kishino-Hasegawa/Shimodaira-Hasegawa tests and non-

parametric tests. Bootstrap values supporting the latter, more derived placement were very high for our data set, i.e., ≥96% under all phylogenetic methods used. In addition, none of our reanalyses of published data, using either likelihood or parsimony methods, recovered *Toxorhynchites* in the more basal position. Concordance among multiple, independent lines of evidence is probably the most powerful tool at our disposal for corroboration of phylogenetic hypotheses (Miyamoto & Cracraft 1991). That all five data sets examined support a derived placement of *Toxorhynchites* within Culicinae gives us confidence in this phylogenetic hypothesis.

Conclusions. – Both the new data presented here and the previously published DNA sequence data support the placement of *Toxorhynchites* within the Culicinae, and the *COI-COII* data does so with the greatest confidence. Combining data from these different sources could improve phylogenetic signal (Mitchell et al. 2000) and lead to a more decisive result, but this cannot be done in this case because existing studies (which apparently were commenced at about the same time) have sampled different taxa and different genes. We recommend that future DNA sequencing efforts aiming to resolve mosquito subfamily and tribal relationships conclusively should capitalize on the published data, where appropriate, and choose their taxa in such a way that combined data sets can be assembled (Caterino et al. 2001). Taxon sampling incompatibilities notwithstanding, all existing DNA sequence data is in broad agreement with the morphology-based study of Harbach & Kitching (1998) regarding the placement of *Toxorhynchites*. Following these authors, Toxorhynchitinae should be reduced to tribal status, although further taxon sampling is needed to establish where they are placed within Culicinae. It appears that complete sequences of the *COI*, *COII*, and tRNA-Leu genes will be effective in further resolving the higher-level relationships of Culicidae in general, and the phylogenetic placement of *Toxorhynchites* within the Culicinae in particular.

Acknowledgements

We thank D. M. Wood for species identifications, T. Arnason, N. Besansky, D. C. Currie, V. Kulasekara, V. Walker and D. M. Wood for providing specimens, J. Leibovitz for excellent technical assistance, and anonymous reviewers for constructive criticism. F. A. H. S. thanks J. Huelsenbeck for productive preliminary discussion. Research funding was provided by NSERC Canada Operating Grants to F. A. H. S. and D. A. H.

References

Baker, R. H., Wilkinson, G. S. & DeSalle, R. (2001) Phylogenetic utility of different types of molecular data used to infer evolutionary relationships among stalk-eyed flies (Diopsidae). *Syst. Biol.* 50: 87-105.

Beard, C. B., Hamm, D. M. & Collins, F. H. (1993) The mitochondrial genome of the mosquito *Anopheles gambiae*: DNA sequence, genome organization, and comparisons with mitochondrial sequences of other insects. *Insect Mol. Biol.* 2: 103-124.

Belkin, J. N. (1962) *The mosquitoes of the South Pacific (Diptera, Culicidae)*. Vol. 1: ix + 608 pp. University of California Press, Berkeley.

Besansky, N. J., & Fahey, G. T. (1997) Utility of the *white* gene in estimating phylogenetic relationships among mosquitoes (Diptera: Culicidae). *Mol. Biol. Evol.* 14: 442-454.

Bremer, K. (1988) The limits of amino-acid sequence data in angiosperm phylogeny reconstruction. *Evolution* 42: 795-803.

Caterino, M. S., Cho, S. & Sperling, F. A. H. (2000) The current state of insect molecular systematics: a thriving tower of Babel. *Ann. Rev. Ent.* 45: 1-54.

Edwards, F. W. (1932) Fam. Culicidae. *Genera Insect.* 194: 1-258. Tervueren, Bruxelles.

Farris, J. S. (1969) A successive approximations approach to character weighting. *Syst. Zool.* 18: 374-385.

Farris, J. S., Källersjö, M. Kluge, A. G. & Bult, C. (1994) Testing the significance of incongruence. *Cladistics* 10: 315-319.

Foley, D. H., Bryan, J. H. Yeates, D. & A. Saul. (1998) Evolution and systematics of *Anopheles*: insights from a molecular phylogeny of Australasian mosquitoes. *Mol. Phylog. Evol.* 9: 262-275.

Foster, P. G. & Hickey, D. A. (1999) Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. *J. Mol. Evol.* 48: 284-290.

Frati, F., Simon, C. Sullivan, J. & Swofford, D. L. (1997) Evolution of the mitochondrial cytochrome oxidase II gene in Collembola. *J. Mol. Evol.* 44: 145-158.

Genetic Computer Group (1994) *Program manual for the Wisconsin package. Version 8*. Madison, Wisconsin.

Goldman, N., Anderson, J. P. & Rodrigo, A. G. (2000) Likelihood-based tests of topologies in phylogenetics. *Syst. Biol.* 49: 652-670.

Harbach, R. E. & Kitching, I. J. (1998) Phylogeny and classification of the Culicidae (Diptera). *Syst. Ent.* 23: 327-370.

Hasegawa, M., Kishino, H. & Yano, T. (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* 21: 160-174.

Huelsenbeck, J. P. (1995) Performance of phylogenetic methods in simulation. *Syst. Biol.* 44: 17-48.

Kishino, H. & Hasegawa, M. (1989) Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. *J. Mol. Evol.* 29: 170-179.

Krzywinski, J., Wilkerson, R. C. & Besansky, N. J.

(2001) Evolution of mitochondrial and ribosomal gene sequences in Anophelinae (Diptera: Culicidae): Implications for phylogeny reconstruction. *Mol. Phylog. Evol.* 18: 479-487.

Lanave, C., Preparata, J., Saccone, C. & Serio, G. (1984) A new method for calculating evolutionary substitution rates. *J. Mol. Evol.* 20: 86-93.

Lockhart, P. J., Steel, M. A., Hendy, M. D. & Penny, D. (1994) Recovering evolutionary trees under a more realistic model of sequence evolution. *Mol. Biol. Evol.* 11: 605-612.

Maddison, W. P. & Maddison, D. R.. (1992) *MacClade: Analysis of phylogeny and character evolution. Version 3.03*. Sinauer Associates, Sunderland, Massachusetts.

Miller, B. R., Crabtree, M. B. & Savage, H. M. (1997) Phylogenetic relationships of the Culicomorpha inferred from 18S and 5.8S ribosomal DNA sequences (Diptera: Nematocera). *Insect Mol. Biol.* 6: 105-114.

Mitchell, A., Cho, S., Regier, J. C., Mitter, C., Poole, R. W. & Matthews, M. M. (1997) Phylogenetic utility of elongation factor-1? in Noctuoidea (Insecta: Lepidoptera): the limits of synonymous substitution. *Mol. Biol. Evol.* 14: 381-390.

Mitchell, A., Mitter, C. & Regier, J. C. (2000) More taxa or more characters revisited: combining data from nuclear, protein-encoding genes for phylogenetic analyses of Noctuoidea (Insecta: Lepidoptera). *Syst. Biol.* 49: 202-224.

Mitchell, S. E., Cockburn, A. F. & Seawright, J. A. (1993) The mitochondrial genome of *Anopheles quadrimaculatus* species A: complete nucleotide sequence and gene organization. *Genome* 36: 1058-1073.

Miyamoto, M. M. & Cracraft, J. (1991) Phylogenetic inference, DNA sequence analysis, and the future of molecular systematics. Pp. 3-17 *in* Miyamoto, M. M. & Cracraft, J.: *Phylogenetic analysis of DNA sequences*. Oxford University Press, New York.

Naylor, G. J. P. & Brown, W. M. (1997) Structural biology and phylogenetic estimation. *Nature* 388: 527-528.

Oosterbroek, P., & Courtney, G. (1995) Phylogeny of the nematocerous families of Diptera (Insecta). *Zool. J. Linn. Soc.* 115: 267-311.

Orti, G. & Meyer, A. (1996) Molecular evolution of ependymin and the phylogenetic resolution of early divergences among euteleost fishes. *Mol. Biol. Evol.* 13: 556-573.

Pawlowski, J., Szadziewski, R. Kmieciak, D. Fahrni, J. & Bittar, G. (1996) Phylogeny of the infraorder Culicomorpha (Diptera: Nematocera) based on 28S RNA gene sequences. *Syst. Ent.* 21: 167-178.

Rambault, A. (1996) Se-Al: Sequence alignment editor, version 1.0. http://evolve.zoo.ox.ac.uk/Se-Al/Se-Al.html

Shimodaira, H. & Hasegawa, M. (1999) Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol. Biol. Evol.* 16:1114-1116.

Simmons, M. P. (2000) A fundamental problem with amino acid sequence characters for phylogenetic analyses. *Cladistics* 16: 274-282.

Simon, C., Frati, F., Beckenbach, A., Crespi, B., Liu, H. & Flook, P. (1994) Evolution, weighting, and phylogenetic utility of mitochondrial gene sequences and a compilation of conserved polymerase chain reaction primers. *Ann. ent. Soc. Am.* 87: 651–701.

Sperling, F. A. H., Anderson, G. S. & Hickey, D. (1994) A DNA-based approach to the identification of insect species used for postmortem interval estimation. *J. foren. Sci.* 39: 418-427.

Steffan, W. A. & Evenhuis, N. L. (1981) Biology of *Toxorhynchites*. *Ann. Rev. Ent.* 26: 159-181.

Steffan, W. A. & Evenhuis, N. L. (1985) Classification of the subgenus *Toxorhynchites* (Diptera: Culicidae). *J. med. Ent. Honolulu* 22: 421-446.

Swofford, D. L. (1999) *PAUP*. Phylogenetic analysis using parsimony (*and other methods). Version 4*. Sinauer Associates, Sunderland, Massachusetts.

Templeton, A. R. (1983) Convergent evolution and nonparametric inferences from restriction fragment and DNA sequence data. Pp. 151-179 *in* Weir, B.: *Statistical analysis of DNA sequence data*. Marcel Dekker, New York.

Revised manuscript accepted October 2001.