J Mol Evol (1997) 44:145–158

JOURNAL OF MOLECULAR EVOLUTION

Evolution of the Mitochondrial Cytochrome Oxidase II Gene in Collembola

Francesco Frati,[1] Chris Simon,[2] Jack Sullivan,[3] David L. Swofford[3]

[1] Dipartimento di biologia Evolutiva, Università di Siena, via P.A. Mattioli 4, 53100, Siena, Italy
[2] Department of Ecology and Evolutionary Biology, U-43, University of Connecticut, 75 North Eagleville Road, Storrs, CT 06269-3043, USA
[3] Laboratory of Molecular Systematics, MSC, MRC-534, Smithsonian Institution, Washington, DC 20560, USA

Received: 12 January 1996 / Accepted: 10 August 1996

Abstract. The sequence of the mitochondrial COII gene has been widely used to estimate phylogenetic relationships at different taxomonic levels across insects. We investigated the molecular evolution of the COII gene and its usefulness for reconstructing phylogenetic relationships within and among four collembolan families. The collembolan COII gene showed the lowest A + T content of all insects so far examined, confirming that the well-known A + T bias in insect mitochondrial genes tends to increase from the basal to apical orders. Fifty-seven percent of all nucleotide positions were variable and most of the third codon positions appeared free to vary. Values of genetic distance between congeneric species and between families were remarkably high; in some cases the latter were higher than divergence values between other orders of insects. The remarkably high divergence levels observed here provide evidence that collembolan taxa are quite old; divergence levels among collembolan families equaled or exceeded divergences among pterygote insect orders. Once the saturated third-codon positions (which violated stationarity of base frequencies) were removed, the COII sequences contained phylogenetic information, but the extent of that information was overestimated by parsimony methods relative to likelihood methods. In the phylogenetic analysis, consistent statistical support was obtained for the monophyly of all four genera examined, but relationships among genera/families were not well supported. Within the genus *Orchesella*, relationships were well resolved and agreed with allozyme data. Within the genus *Isotomurus*, although three pairs of populations were consistently identified, these appeared to have arisen in a burst of evolution from an earlier ancestor. *Isotomurus italicus* always appeared as basal and *I. palustris* appeared to harbor a cryptic species, corroborating allozyme data.

Key words: Collembola — mtDNA — Cytochrome *c* oxidase subunit II — Among-site rate variation — A + T bias — Phylogeny — Likelihood models

Introduction

Collembola represent an interesting group of arthropods for several reasons. They first appeared prior to 400 Myr ago but, if the fossil record and their current diversity are an accurate indicator, they did not experience the remarkable radiation of other, more diverse insect orders. Nevertheless, their phylogenetic position is critical to the understanding of the evolution of insect orders since they constitute one of the earliest hexapod lineages. Several issues regarding the relationships among the basal hexapods have been debated for some time (Kristensen 1981; Bitsch 1994), including the question of whether the Collembola should remain in the Insecta or be elevated to the rank of class (Anderson 1973; Manton 1973, 1979).

Within Collembola, systematic relationships are also unclear, and additional taxonomic information is needed at several levels. Traditionally, the 20 families and over 1,000 species of Collembola have been divided into the two well-differentiated suborders Arthropleona and

Correspondence to: F. Frati; e-mail: FRATI@UNISI.IT

Symphypleona, but a new suborder, Neelipleona, has been proposed (Massoud 1971). Agreement over this taxonomic scheme, however, is not universal, and the distribution of collembolan genera at the suborder and the family level is still a matter of debate.

At the species level, Collembola have a variety of characteristics which make them a good model to study population genetics and speciation processes. They are small, strictly soil-dwelling organisms whose dispersal capability is limited by the absence of wings, among other factors. Furthermore, species often comprise geographically isolated populations among which gene flow can be very low (Fanciulli et al. 1991; Frati et al. 1992). The prolonged geographic isolation, resulting in genetically divergent populations, should tend to promote speciation.

Systematic analysis of Collembola at the species level is often complicated by the lack of strong morphological characters that distinguish the species, and the degree of variability which affects some of the diagnostic characters (for instance, pigmentation patterns and chaetotaxy). The genera *Orchesella* and *Isotomurus* are two clear examples of this problem. Many species within these genera can be distinguished only by pigmentation patterns and these patterns can exhibit high levels of intra- and interpopulation variability. Thus, genetic markers, such as allozymes or DNA sequences, are a valuable tool with which to interpret the systematic significance of morphological variation. These approaches have proven useful in the study of collembolan population genetics (Frati et al. 1992), species diagnosis (Carapelli et al. 1995a), and phylogenetic reconstruction (Frati et al. 1995; Carapelli et al. 1995b).

The utility of mitochondrial DNA sequences for inferring phylogenetic relationships is now well established (Simon et al. 1994, review). Of the several genes exploited for this purpose, the mitochondrial cytochrome *c* oxidase subunit II (COII) is perhaps the most frequently used. A considerable amount of information concerning its evolution is now available for various groups, including rodents (Brown and Simpson 1982), primates (Disotell et al. 1992; Adkins and Honeycutt 1994), hominoids (Ruvolo et al. 1991), and several orders of pterygote (winged) insects (Liu and Beckenbach 1992; Beckenbach et al. 1993; Willis et al. 1992; Brower 1994; Brown et al. 1994; Sperling and Hickey 1994). Prior to our study, no basal insect orders had been examined.

Although the COII gene has proven useful for inferring phylogenetic relationships among relatively closely related insect species and genera, it has been less informative with respect to pterygote orders because the majority of amino acid replacement sites, about 60%, do not appear to be free to vary, and those sites that can vary appear to be near saturation (Liu and Beckenbach 1992; Simon et al. 1994). This gene is therefore unlikely to be useful for even older apterygote insect orders. Studies of the molecular evolution of the COII gene are also of interest to evolutionary biologists because they can provide information useful for improving phylogenetic analysis (Simon et al. 1994).

Cytochrome *c* oxidase is a transmembrane complex of polypeptides composed of three mitochondrially encoded subunits (I–III) and four subunits encoded by the nuclear genome. The biochemical processes catalyzed by the cytochrome *c* oxidase complex are quite well known (Capaldi 1990; Iwata et al. 1995), and it has been possible to correlate the degree of conservation of certain COII residues with the functional cores of the molecule (Capaldi et al. 1983; Millett et al. 1983, Capaldi 1990; Adkins and Honeycutt 1994). Other properties of the COII gene that have been examined from a phylogenetic perspective in a few insect orders include variation in evolutionary rate (Crozier et al. 1989) and A + T content (Liu and Beckenbach 1992; Jermiin and Crozier 1994).

The present paper examines the evolution and phylogenetic usefulness of COII gene sequences in the apterygote insect order Collembola. Several questions are addressed, including: What is (1) the overall pattern of evolution of nucleotide and amino acid sequences among Collembola? (2) the degree of A + T content bias in relation to the evolutionary trend observed in pterygote insects? (3) the degree of conservation of specific residues with respect to the functional domains of the molecule? and (4) the utility of the COII gene for phylogenetic reconstruction at different taxonomic levels within Collembola? To address these questions, data were gathered for conspecific populations, congeneric species, and genera from four different families.

Materials and Methods

Sequence data were collected from 17 specimens representing a total of 11 morphological species of arthropleonan Collembola. They comprised four genera, each of which belongs to a different family. Intraspecific variability was examined for four of the species (Table 1).

Genomic DNA was extracted from single individuals of each population following the methods described in Simon et al. (1991), with slight modifications due to the small size of the specimens. In brief, single individuals (either living, or frozen, or alcohol-preserved) were ground in homogenizing buffer and incubated overnight in sodium dodecylsulfate (SDS) and proteinase K. DNA was then isolated by phenol-chloroform extraction followed by ethanol precipitation. The whole COII gene was amplified with the two flanking primers TL2-J-3037-modified (5′-AATATGGCAGATTAGTGCA-3′) and TK-N-3785-modified (5′-GTTTAAGAGACCAGTACTT-3′) (Simon et al. 1994) using the following profile: denaturation at 94°C (1 min 10 s), annealing at 45°C (1 min 10 s), and extension at 72°C (1 min 30 s). Double-stranded polymerase chain reaction (PCR) products were run on a 1% low-melting-point agarose gel, the band was excised from the gel, and the DNA was purified by phenol-chloroform extraction and ethanol precipitation. Purified DNA was sequenced by the double-stranded technique of Hsiao (1993) using both of the amplification primers and two additional internal primers: a version of C2-N-3389 (Simon et al. 1994) modified to match Collembola with six bases added to the 3′ and eight bases removed from the 5′ end (5′-TCAATATCATTGATGTCC-3′) and a shortened version of C2-N-3661 (Simon et al. 1994: 5′-GCTCCACAAATTTCTGAACA-3′).

Table 1. List of collembolan species used in this study

	Populations[a]	Code
Entomobryiomorpha		
Family Entomobryidae		
Orchesella villosa	SIE, AMI	Ovi
Orchesella ranzii	GIG	Ora
Orchesella dallaii	MAR	Oda
Orchesella cincta	BSR	Oci
Orchesella flavescens	CAN	Ofl
Family Isotomidae		
Isotomurus palustris	GER, RAD, GIG, CIR	Ipa
Isotomurus maculatus	SIE	Ima
Isotomurus unifasciatus	SIE, RAD	Iun
Isotomurus italicus	GIG	Iit
Poduromorpha		
Family Onychiuridae		
Tetrodontophora bielanensis	PSB	Tbi
Family Neanuridae		
Thaumanura ruffoi	SIE, BSR	Tru

[a] Key to locality names: AMI: Mt. Amiata; BSR: Bocca Serriola; CAN: Cansiglio; CIR: Circeo; GER: Gerfalco; GIG: Giglio island; MAR: Marmore; PSB: Passo S. Boldo; RAD: Radi; SIE: Siena

Sequences from 17 individuals were aligned with the published sequences of the dragonfly (*Sympetrum stiolatum;* Liu and Beckenbach 1992) and *Drosophila yakuba* (Clary and Wolstenholme 1985) using the multiple alignment program CLUSTAL V (Higgins et al. 1992); alignments were confirmed by visual inspection using both nucleotides and amino acids.

Basic sequence statistics and evolutionary distances were computed using MEGA (Kumar et al. 1993) and test version 4.0d42 of PAUP* written by D.L.S. Amino acid sequences were inferred using the *Drosophila* mitochondrial code (de Bruijn 1983) and a codon usage table was also constructed.

A number of phylogenetic analyses were performed using PAUP*. Exact searches under the parsimony criterion were performed using the branch-and-bound algorithm, retaining all equally parsimonious trees (MULPARS). Heuristic searches under the maximum-likelihood and minimum-evolution (Kidd and Cavalli-Sforza 1971; Rzhetsky and Nei 1992) criteria were performed by stepwise-addition using ten random-addition sequences followed by tree bisection-reconnection (TBR) branch rearrangement. Branch lengths were constrained to be nonnegative in the minimum-evolution analysis. LogDet/paralinear distances (Lockhart et al. 1994; Lake 1994) were used for the distance (minimum-evolution) analysis because these distances appear not to suffer from the inflation of variance associated with distances derived under other complex models (Waddell 1995; Swofford et al. 1996). Model parameters for the maximum-likelihood search were fixed to values estimated (by maximum-likelihood) on trees found by the parsimony and distance analyses. Nodal support was estimated by bootstrap analysis. In the likelihood analyses, model parameters were fixed in each of the replicates to values estimated for the original data.

The nucleotide sequences reported in this paper have been deposited in the EMBL, GenBank, and DDBJ Nucleotide Sequence Databases under accession numbers X80688-9, X95725, X95782-94, and X95894. The alignment of nucleotide sequences has been deposited in the EMBL Nucleotide Sequence Database under the accession number DS25208.

Results and Discussion

Alignment

An unambiguous alignment (except for the 3′-end) of the amino acid sequences was obtained using CLUSTAL V. In *Orchesella,* the COII gene is the same length (690 bp) as other insects (Liu and Beckenbach 1992). The sequence of *Isotomurus* was easily alignable with *Orchesella* except for the 3′-end, where, in *Isotomurus,* no standard termination codon was present in the corresponding position. Due to the difficulty of aligning the 3′-end of the gene and to the incompleteness of some of the sequences, the last 18 bases were not used in the sequence analysis.

The alignment revealed the presence of two indels between amino acids 120 and 133: The first is one codon long (*Tetrodontophora bielanensis* is missing a codon) and the second is four codons long. (*Thaumanura ruffoi* has four codons not found elsewhere.) These four unique codons were also found in the partially sequenced *Bilobella aurantiaca* (data not shown) and may represent a synapomorphy of the family Neanuridae to which both genera belong. Both indels are located in a variable region of the gene, which also displays indels in other insect sequences (Liu and Beckenbach 1992). This variability suggests that this region has minor importance for the function of the gene product and may be capable of accepting more insertions or deletions as well as amino acid replacements.

Ignoring the indels, 57% of the nucleotide positions (381/669) were variable among the four collembolan families. As expected, most of the variable sites were in third codon positions, while only 15 of them were located in second codon positions (Table 2). Only four codons were conserved across all the specimens. It appeared that almost all of the third codon positions are free to vary; this conclusion was supported by our analysis of among-site rate variation discussed below.

Table 2. Distribution of variable sites in the 17 collembolan sequences[a]

	Variable sites	
Codon position	A	B
1st	0.279	0.478
2nd	0.154	0.263
3rd	0.567	0.969

[a] A: Fraction of total variable sites that are found in each codon position; B: Fraction of sites in each codon position that are variable

Nucleotide Composition

Table 3 shows the nucleotide composition of the sequences, all of which are rich in A + T, as expected for an insect mitochondrial gene (Brown 1985; Clary and Wolstenholme 1985; Crozier and Crozier 1993; Simon et al. 1994). However, average A + T content in Collembola (64%) appears considerably lower than that measured in other insects. In particular, a clear evolutionary trend toward increasing A + T content is evident (Fig. 1),

Table 3. Table of nucleotide content at the three codon positions

Specimen	All sites					1st codon position				
	A	T	C	G	Total[a]	A	T	C	G	Total[a]
O. villosa–SIE	31.3	36.8	19.2	12.8	672	30.4	29.0	19.6	21.0	224
O. villosa–AMI	31.3	37.1	18.9	12.8	672	30.4	30.4	18.3	21.0	224
O. ranzii	29.8	34.1	21.9	14.3	672	28.1	29.5	20.1	22.3	224
O. flavescens	30.5	36.5	19.5	13.5	672	31.3	29.5	18.8	20.5	224
O. cincta	29.0	33.9	22.6	14.4	672	29.0	26.8	20.5	23.7	224
O. dallaii	30.1	38.1	19.3	12.5	672	29.9	30.8	17.9	21.4	224
I. unifasciatus–SIE	31.0	30.2	23.5	15.3	672	27.2	24.6	21.0	27.2	224
I. unifasciatus–RAD	31.3	30.5	23.1	15.2	672	27.7	24.1	21.0	27.2	224
I. maculatus	28.3	31.0	24.3	16.5	672	28.6	25.4	20.1	25.9	224
I. palustris–GER	29.0	32.4	22.9	15.6	672	28.1	25.9	20.1	25.9	224
I. palustris–RAD	28.3	31.7	23.5	16.5	672	28.1	25.0	20.5	26.3	224
I. palustris–GIG	30.4	31.3	22.3	16.1	672	27.7	26.3	19.2	26.8	224
I. palustris–CIR	31.3	31.4	22.2	15.2	672	38.1	26.3	19.2	26.3	224
I. italicus	30.8	29.8	24.0	15.5	672	26.8	25.4	20.1	27.7	224
Te. bielanensis	34.8	36.9	18.1	10.2	669	34.1	28.7	17.9	19.3	223
Th. ruffoi–SIE	29.8	35.2	24.0	11.0	684	34.6	27.2	21.1	17.1	228
Th. ruffoii–BSR	30.0	35.1	24.0	11.0	684	34.6	27.2	21.1	17.1	228
Mean	30.4	33.6	22.0	14.0	11445	29.7	27.2	19.8	23.3	3815
Compositional bias[b]			0.187					0.092		
Test of homogeneity[c]			$\chi^2 = 136.087$ $P < 0.001$					$\chi^2 = 40.750$ $Pn < 0.908$		

Specimen	2nd codon position					3rd codon position				
	A	T	C	G	Total[a]	A	T	C	G	Total[a]
O. villosa–SIE	26.8	37.5	23.2	12.5	224	36.6	43.8	14.7	4.9	224
O. villosa–AMI	27.2	37.5	23.2	12.1	224	36.2	43.3	15.2	5.4	224
O. ranzii	26.8	37.1	23.7	12.5	224	34.4	35.7	21.9	8.0	224
O. flavescens	27.2	37.9	22.8	12.1	224	33.0	42.0	17.0	8.0	224
O. cincta	27.2	38.8	22.8	11.2	224	30.8	36.2	24.6	8.5	224
O. dallaii	26.8	37.5	24.1	11.6	224	33.5	46.0	16.1	4.5	224
I. unifasciatus–SIE	25.0	37.5	25.4	12.1	224	40.6	28.6	24.1	6.7	224
I. unifasciatus–RAD	25.0	37.5	25.4	12.1	224	41.1	29.9	22.8	6.3	224
I. maculatus	25.4	37.5	24.6	12.5	224	30.8	29.9	28.1	11.2	224
I. palustris–GER	25.0	37.5	25.4	12.1	224	33.9	33.9	23.2	8.9	224
I. palustris–RAD	25.0	37.5	25.4	12.1	224	31.7	32.6	24.6	11.2	224
I. palustris–GIG	25.9	37.5	24.6	12.1	224	37.5	29.9	23.2	9.4	224
I. palustris–CIR	25.9	37.5	24.6	12.1	224	39.7	30.4	22.8	7.1	224
I. italicus	25.4	37.9	24.6	12.1	224	40.2	25.9	27.2	6.7	224
Te. bielanensis	29.1	39.0	22.4	9.4	223	41.3	43.0	13.9	1.8	223
Th. ruffoi–SIE	24.6	38.6	25.0	11.8	228	30.3	39.9	25.9	3.9	228
Th. ruffoii–BSR	24.6	38.6	25.0	11.8	228	30.7	39.5	25.9	3.9	228
Mean	26.1	37.8	24.2	11.9	3815	35.4	35.9	21.8	6.8	3815
Compositional bias[b]			0.185					0.285		
Test of homogeneity[c]			$\chi^2 = 13.168$ $P < 0.999$					$\chi^2 = 2000.303$ $P < 0.001$		

[a] Total number of sites considered
[b] Compositional bias is measured according to Irwin et al. (1991)
[c] Homogeneity test conducted using PAUP*

confirming an earlier observation by Jermiin and Crozier (1994). The presence of such an evolutionary trend would reinforce their hypothesis of directional mutational pressure acting in the hexapod lineage. Some evidence suggests that this trend may have reversed in some holometabolous insect lineages (Jermiin and Crozier 1994).

A similar trend of decreasing G content was observed in third codon positions (Fig. 1). This bias, however, seems to be common among the mitochondrial genes of many metazoans (Jermiin and Crozier 1994). Again, *Tetrodontophora* had the lowest-observed third codon position G content among Collembola with only 1.8%. The presence of such a bias is apparently correlated with the elimination of tRNAs having CNN as an anticodon, which might have resulted from the economization of mitochondrial genome size (Osawa et al. 1992).

The distribution of bias in base composition was not

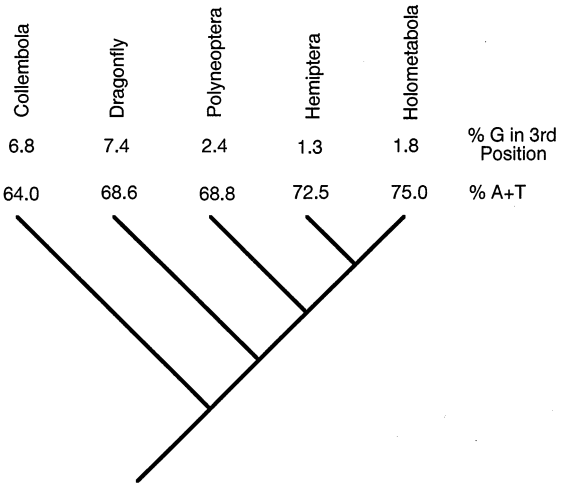

Fig. 1. Trends in overall A + T content and %G at third-codon positions among insect COII sequences.

Table 4. Index of compositional bias (Irwin et al. 1991) in different genes and species

		Codon position		
Gene	Taxon	1st	2nd	3rd
COII	Collembola (15 spp.)	0.092	0.185	0.285
	pterygote insects (13 spp.)[a]	0.175	0.233	0.490
	primate (25 spp.)[b]	0.065	0.185	0.292
Cyt b	*Tetraponera rufoniger*[c]	0.189	0.267	0.372
	mammals (18 spp.)[d]	0.076	0.221	0.401
ND1	*Tetraponera rufoniger*[c]	0.394	0.262	0.442
ND6	*Tetraponera rufoniger*[c]	0.540	0.270	0.413

[a] Liu and Beckenbach (1992)
[b] Adkins and Honeycutt (1994)
[c] Jermiin and Drozier (1994)
[d] Irwin et al. (1991)

uniform with respect to the three codon positions (Table 3). It was highest at third, intermediate at second, and lowest at first codon positions. The same pattern was observed among 13 species of pterygote insects and in other animal species for different mitochondrial genes, whereas a different distribution of the bias was observed in the ant ND1 and ND6 genes (Table 4). Homogeneity of base frequencies across taxa was strongly rejected by a chi-squared test of independence (Table 3). This heterogeneity is apparently due only to variation at third positions, however, as there was no evidence of heterogeneity at first or second positions considered separately. Note that this test ignores correlation due to phylogeny and therefore tends to reject the null hypothesis too easily, so that failure to reject can safely be taken as evidence of homogeneity. However, for third positions, rejection of the null hypothesis is so strong that it cannot reasonably be attributed to this bias of the test.

Inferred Amino Acid Sequences

Due to the truncation of the 3′-end of the gene in some species and to the presence of the indels described above, the total length of the amino acid sequences studied here is 224 residues in *Orchesella* and *Isotomurus*, 223 in *Tetrodontophora*, and 228 in *Thaumanura*. The most common amino acids were leucine (9.9 to 12.7%), isoleucine (8.0 to 13.2%), and serine (7.6 to 11.0%), while the rarest was cysteine (0.9 to 1.8%). Excluding the two indels, 114 out of 223 sites were variable (51.1%); these were not equally distributed across the whole length of the polypeptide chain, but appear to be concentrated in certain regions. These highly variable regions included the beginning of the molecule (amino acids 1–8), the central indel-containing region (amino acids 120–133), and the C-terminal part (amino acids 210–224). This terminal region, however, included a conserved glutamic acid at position 212. However, some remarkably conserved windows were present, including those spanning amino acids 14–26, 76–91, 101–113, 133–140, 156–166, 169–189, and 193–209. For some of these regions a clear relationship between conservation and biochemical properties of the gene product has been postulated, as discussed below.

Codon Usage

Codon usage in the collembolan COII gene reflected a bias against codons ending with C and especially G (Table 5). AGG was the only codon which was never used in the collembolan COII gene. However, a preference for triplets ending with A was displayed by the sixfold degenerate codon family encoding leucine. A similar observation was made both in pterygote insects (Liu and Beckenbach 1992) and in primates (Adkins and Honeycutt 1994).

Relative synonymous codon usage (RSCU) values (shown in Table 5) give an estimate of the preference for alternative synonymous codons (Sharp et al. 1986). The codon with highest RSCU was AAC (coding for asparagine) but all the remaining codons with RSCU greater than 2 had A or T in third position. The high frequency with which AAC was used with respect to the synonymous AAU (which would be expected to occur more frequently as a result of A + T bias) suggested that, at least in some cases, codon usage was not simply related to nucleotide content but appeared to be influenced by selection for specific tRNA molecules. The mitochondrially encoded tRNA for asparagine, in fact, has the anticodon GUU (Clary and Wolstenholme 1985).

Favored initiation codons among collembolan mitochondrial COII genes were found to be ATC or ATT. ATA was used only in *Tetrodontophora*. ATG was not used in these collembolan species, even though it is the most frequently used in primates (Disotell et al. 1992; Adkins and Honeycutt 1994) and all other pterygote in-

Table 5. Codon usage in collembolan species sequenced in this study based on *Drosophila* mitochondrial genetic code (de Brujin 1983)[a]

UUU (F)	10.8 (1.42)	UCU (S)	7.0 (2.43)	UAU)Y)	4.1 (0.94)	UGU (C)	1.3 (0.54)
UUC (F)	4.4 (0.58)	UCC (S)	3.0 (1.04)	UAC (Y)	4.4 (1.00)	UGC (C)	1.5 (0.64)
UUA (L)	11.3 (2.76)	UCA (S)	6.1 (2.12)	UAA (*)	0.0 (0.00)	UGA (W)	4.5 (1.89)
UUG (L)	1.1 (0.26)	UCG (S)	0.5 (0.18)	UAG (*)	0.0 (0.00)	UGG (W)	0.9 (1.00)
CUU (L)	4.2 (1.77)	CCU (P)	4.6 (1.92)	CAU (H)	2.8 (0.63)	CGU (R)	1.9 (0.34)
CUC (L)	1.1 (0.26)	CCC (P)	2.7 (1.00)	CAC (H)	3.4 (1.00)	CGC (R)	1.5 (1.06)
CUA (L)	6.3 (0.25)	CCA (P)	3.4 (0.69)	CAA (Q)	7.0 (1.43)	CGA (R)	1.9 (0.34)
CUG (L)	0.6 (0.25)	CCG (P)	1.4 (0.68)	CAG (Q)	1.3 (0.94)	CGG (R)	0.6 (0.29)
AUU (I)	15.6 (2.81)	ACU (T)	5.8 (1.65)	AAU (N)	7.0 (1.26)	AGU (S)	0.9 (1.45)
AUC (I)	7.1 (1.00)	ACC (T)	2.4 (0.70)	AAC (N)	6.1 (3.03)	AGC (S)	1.6 (0.57)
AUA (M)	7.2 (1.00)	ACA (T)	5.5 (1.63)	AAA (K)	3.8 (1.00)	AGA (S)	0.9 (0.33)
AUG (M)	1.2 (0.35)	ACG (T)	1.2 (0.21)	AAG (K)	0.4 (0.55)	AGG (S)	0.0 (0.00)
GUU (V)	3.9 (1.00)	GCU (A)	3.2 (0.96)	GAU (D)	4.9 (1.14)	GGU (G)	2.6 (1.16)
GUC (V)	2.2 (0.53)	GCC (A)	2.4 (0.72)	GAC (D)	4.3 (1.01)	GGC (G)	1.0 (0.43)
GUA (V)	6.0 (1.47)	GCA (A)	4.7 (1.40)	GAA (E)	7.2 (1.70)	GGA (G)	3.6 (1.57)
GUG (V)	1.9 (0.58)	GCG (A)	0.6 (0.15)	GAG (E)	1.9 (0.84)	GGG (G)	1.8 (1.00)

[a] Estimates of observed average codon frequencies; standard amino acid identification signal-letter codes and numerical RSCU values are given in parentheses

sects (Liu and Beckenbach 1992). While ATA is known to code for methionine in mitochondrial genes of yeast, insects, and vertebrates, ATC and ATT should code for isoleucine, and their presence as initiation codons is unusual. Fearnley and Walker (1987), however, demonstrated that ATT is translated as methionine when found as an initiation codon in the human mitochondrial ND2 gene and the same could be true for insect mitochondrial genes. This position-dependent dual coding function of ATT might apply to ATC as well, but this has not been demonstrated.

All the *Orchesella* species terminated at the 230th codon with the presumed stop codon TAA or TAG (de Bruijn 1983). Among *Isotomurus*, only *I. unifasciatus* and *I. maculatus* were sequenced across this region and they both displayed the codon TTA (leucine) at the corresponding position. *Thaumanura* had TTC (phenylalanine) in the corresponding position, but had a TAA terminator three codons upstream and another TAG terminator five codons downstream. In all these cases, however, the presence of a T, coupled with polyadenylation (Anderson et al. 1981; Clayton 1984), might be enough to mark the termination of the gene as observed in the mitochondrial genome of other insects (Clary and Wolstenholme 1985; Liu and Beckenbach 1992; Crozier and Crozier 1993; Mitchell et al. 1993) as well as in mammalian mitochondrial genes (Bibb et al. 1981; Andersone et al. 1982).

Functional Aspects of Amino Acid Conservation

The degree of conservation of regions of the polypeptide chain and even of single residues depends on the importance of these regions or residues in the structure and function of the molecule. The conservation of the COII gene region spanning residues 62–85 (Capaldi et al. 1983) was correlated with the interaction with helices in other subunits of the complete cytochrome *c* oxidase molecule and its localization within the inner mitochondrial membrane (Iwata et al. 1995). This stretch of amino acids is believed to form a membrane-inserted helix in which there are 15 extremely conserved residues located on one side. Eleven of these 15 amino acids were also conserved in the Collembola we examined. The four that were variable appeared to involve replacements with amino acids having similar chemical and structural properties (A-S-T, I-V, A-G, I-L) and which could therefore be defined as conservative replacements (French and Robson 1983).

Subunit II of cytochrome *c* oxidase is also known to be involved in the ligation of two copper atoms to each molecule (Iwata et al. 1995). It has been proposed that the ligation of Cu involves cysteine and histidine residues (Stevens et al. 1982). Four amino acids in the C-terminal part of COII are believed to play this role: His-161, Cys-196, Cys-200, and His-204 of the bovine protein sequence (Millett et al. 1983; Capaldi 1990). Interestingly, the same four residues are conserved in Collembola, a snail (Lecanidou et al. 1994), several primates (Adkins and Honeycutt 1994), and all pterygote insects studied to date (Liu and Beckenbach 1992; Willis et al. 1992; Beckenbach et al. 1993; Mitchell et al. 1993; Sperling and Hickey 1994; Sperling et al. 1994, 1995). In bacterial COII (Iwata et al. 1995), the same residues participate in the formation of the copper center.

Another important function of COII is to provide a cleft for the binding of cytochrome *c*. Such a cleft appears to involve specific carboxylate groups capable of interacting with a ring of lysines on the cytochrome *c* molecule (Millett et al. 1983; Iwata et al. 1995). Several carboxylate residues are present in the COII polypeptide chain and it was demonstrated with protection experi-

ments (in bovines) that Asp-112, Glu-114, Asp-158, and Glu-198 are directly involved in cytochrome *c* binding (Millett et al. 1983; Capaldi 1990). While Asp-112, Asp-158, and Glu-198 are well conserved in Collembola, pterygote insects, and primates (Asp-112 only accepts a conservative change to Glu in *I. palustris*–CIR), Glu = 114 does not appear either in insects or in primates. The poor conservation of this residue suggests that it may play a less critical role in cytochrome *c* binding. Possible candidates to substitute for the function of Glu-114 are Glu-109 (always conserved in insects and primates, as well as in bovines), Glu-117 (conserved in Collembola but changed to serine in the flea, aspartic acid in the wasp, and different nonconservative residues among primates), or Asp-119 (conserved in Collembola but changed to asparagine in the wasp and some primates). Perhaps the function associated with Glu-114 in bovines is replaced by different adjacent carboxylate groups in other taxa.

Glu-89, Phe-188, Glu-202, and Trp-226 were conserved across all collembolans studied. All of these amino acids have twofold degenerate codons, but none of them is known to play a fundamental role in the function of the gene product. However, Glu-89 and Glu-202 are conserved in all insects (Liu and Beckenbach 1992) and primates (Adkins and Honeycutt 1994), suggesting an important role in structure or function.

Genetic Divergence

Evolutionary distances (mean number of substitutions per site) between pairs of taxa were estimated according to the HKY85 model of nucleotide substitution (Hasegawa et al. 1985) assuming a Γ-distribution of rates across sites (shape parameter = 0.3121) using the approximate method of P.O. Lewis and D.L. Swofford (see below for the rationale underlying the selection of this model). The observed proportion of nucleotide differences and HKY85 distances are shown in Table 6 and summarized for comparisons involving different taxonomic levels in Table 7.

Intraspecific variability was generally low (Table 6), with the exception of comparisons between two pairs of populations of *I. palustris* that differed by a raw sequence divergence of 0.179–0.188, a level of differentiation equivalent to interspecific comparisons within *Isotomurus.* These data confirm an earlier conclusion based on allozyme data (Frati et al. 1995) that cryptic species may be present within *I. palustris* as currently defined. Amino acid sequences were extremely conserved in intraspecific comparisons, with no differences observed between the two specimens of *Th. ruffoi,* and only one replacement observed between pairs of *O. villosa* and *I. unifasciatus* specimens. A single amino acid replacement was also observed between *I. palustris* from RAD and GER, but these specimens differed by eight to ten replacements from GIG and CIR individuals, providing further evidence for the existence of cryptic species within *I. palustris*.

Sequence divergence between congeneric species was extraordinarily high (Tables 6 and 7), typically above 0.19 (uncorrected) and 0.40 (HKY85-corrected); these levels of divergence are much higher than those observed in the COII gene among species of other genera of insects. Willis et al. (1992) found species of the hymenopteran genus *Apis* to diverge by not more than 0.10 (uncorrected) while genetic distance between species of the *Drosophila obscura* group (Beckenbach et al. 1993) ranged from 0.02 to 0.115 (Jukes-Cantor corrected). Even lower uncorrected distance values have been observed within several genera of Lepidoptera, including *Yponomeuta* (Sperling et al. 1995), *Choristoneura* (Sperling and Hickey 1994), and *Heliconius* (Brower 1994). Interspecific amino acid divergences averaged 0.039 (approximately 8.7 total replacements per sequence pair) among *Isotomurus* species, whereas divergences averaged 0.086 among the five species of *Orchesella* (about 19.3 total replacements per sequence pair). Thus, congeneric collembolan species appear to be more differentiated at the amino acid level than congeneric species of the *Drosophila obscura* group (Beckenbach et al. 1993).

For comparative purposes, we calculated HKY85 distances within the pterygote insect orders examined by Liu and Beckenbach (1992); these were 1.083 (within Hymenoptera), 1.070 (within Orthoptera), and 1.700 (within Coleoptera). Among Collembola, generally comparable estimates were obtained in *Orchesella*–*Isotomurus* comparisons (0.850–1.259) but estimates were as high as 2.859 between *Isotomurus* and *Thaumanura*. Differences in A + T bias among orders (with collembolans being especially low) further distort these distances, but many of the distance values between the collembolan species and the dragonfly or *Drosophila* sequences are lower than distances between collembolan species from different superfamilies. In addition, the range of HKY85 distances for interordinal comparisons in the taxa examined by Liu and Beckenbach (1992; 0.527–2.909) encompasses the range of similarly corrected distances among Collembolan families (0.850–2.859).

Amino acid sequences also appear to be saturated at this level of comparison. Uncorrected values range from a low average divergence of 20.85% between *Orchesella* and *Isotomurus* species to a high of 38.66% between *Isotomurus* and *Thaumanura*. (Fig. 3). Comparisons of collembolan species with either the dragonfly (average 38.5% divergence) and *Drosophila* (average 31.7% divergence) fall well within the broad range of divergence values (21–48%) calculated by Liu and Beckenbach (1992) among pterygote insect orders (Fig. 2).

Despite the high apparent level of multiple nucleotide substitutions and amino acid replacements, it is obvious that collembolan families and species are extremely di-

Table 6. Genetic distance estimation[a]

	1	2	3	4	5	6	7	8	9
1 Ovi-SIE	—	0.021	0.200	0.202	0.200	0.142	0.272	0.271	0.294
2 Ovi-AMI	0.023	—	0.203	0.205	0.200	0.143	0.274	0.272	0.297
3 Ora-GIG	0.483	0.503	—	0.209	0.226	0.190	0.300	0.300	0.300
4 Ofl-CAN	0.462	0.480	0.487	—	0.199	0.202	0.286	0.284	0.281
5 Oci-BSR	0.454	0.454	0.588	0.507	—	0.224	0.286	0.284	0.294
6 Oda-MAR	0.254	0.260	0.424	0.488	0.588	—	0.291	0.290	0.303
7 Iun-SIE	0.868	0.885	1.086	0.977	0.982	1.023	—	0.010	0.175
8 Iun-RAD	0.850	0.867	1.083	0.958	0.961	1.002	0.011	—	0.176
9 Ima-SIE	1.133	1.181	1.169	0.911	1.133	1.259	0.363	0.368	—
10 Ipa-GER	0.928	1.017	1.086	0.767	1.012	0.999	0.421	0.419	0.379
11 Ipa-RAD	0.948	1.038	1.095	0.772	1.029	1.020	0.399	0.398	0.394
12 Ipa-GIG	0.918	0.936	0.971	1.090	0.977	0.966	0.426	0.432	0.447
13 Ipa-CIR	0.968	1.006	1.043	1.099	0.887	0.982	0.384	0.396	0.394
14 Iit-GIG	0.988	0.971	0.954	0.797	0.697	0.989	0.477	0.458	0.438
15 Tbi-PSB	1.455	1.494	1.562	1.443	1.563	1.731	1.526	1.426	1.846
16 Tru-SIE	1.389	1.343	1.669	1.442	1.661	1.597	2.041	2.039	2.859
17 Tru-BSR	1.433	1.385	1.691	1.482	1.641	1.618	2.014	2.012	2.822
18 Dragonfly	1.846	1.852	1.652	1.853	1.994	1.543	2.222	2.266	2.252
19 Dyakuba	1.340	1.266	1.693	1.379	1.864	1.327	1.953	1.909	2.092

	10	11	12	13	14	15	16	17	18	19
1 Ovi-SIE	0.281	0.284	0.283	0.284	0.284	0.311	0.324	0.327	0.351	0.315
2 Ovi-AMI	0.288	0.291	0.284	0.287	0.283	0.312	0.321	0.324	0.351	0.311
3 Ora-GIG	0.296	0.299	0.288	0.297	0.284	0.324	0.339	0.341	0.342	0.333
4 Ofl-CAN	0.263	0.265	0.296	0.297	0.265	0.314	0.323	0.324	0.345	0.317
5 Oci-BSR	0.284	0.286	0.290	0.280	0.250	0.329	0.336	0.335	0.351	0.341
6 Oda-MAR	0.287	0.290	0.288	0.287	0.284	0.329	0.336	0.338	0.336	0.315
7 Iun-SIE	0.194	0.190	0.193	0.184	0.205	0.318	0.368	0.366	0.360	0.351
8 Iun-RAD	0.194	0.190	0.194	0.187	0.202	0.314	0.368	0.366	0.362	0.350
9 Ima-SIE	0.181	0.184	0.194	0.185	0.197	0.330	0.381	0.380	0.360	0.350
10 Ipa-GER	—	0.019	0.182	0.179	0.188	0.303	0.378	0.377	0.357	0.354
11 Ipa-RAD	0.021	—	0.188	0.181	0.185	0.312	0.381	0.380	0.363	0.365
12 Ipa-GIG	0.373	0.397	—	0.105	0.205	0.320	0.359	0.357	0.360	0.356
13 Ipa-CIR	0.361	0.365	0.170	—	0.196	0.317	0.368	0.366	0.357	0.348
14 Iit-GIG	0.403	0.389	0.454	0.415	—	0.309	0.353	0.353	0.362	0.359
15 Tbi-PSB	1.228	1.333	1.428	1.438	1.319	—	0.330	0.332	0.368	0.338
16 Tru-SIE	2.647	2.644	1.904	2.181	1.820	1.556	—	0.003	0.396	0.360
17 Tru-BSR	2.614	2.610	1.880	2.154	1.832	1.577	0.003	—	0.395	0.359
18 Dragonfly	2.105	2.282	2.239	2.137	2.216	2.373	2.838	2.798	—	0.251
19 Dyakuba	2.071	2.356	2.032	1.831	2.087	1.754	2.271	2.241	0.685	—

[a] Uncorrected percentages of nucleotide divergence are given above the diagonal. HKY85 + Γ distances, calculated for all sites, are given below the diagonal, with rates assumed to follow a gamma distribution with $\alpha = 0.3121$

vergent. This high divergence is an interesting feature that could be explained in two ways. First, the COII gene could be evolving intrinsically faster in Collembola than in other insect orders. For example, different relative rates of evolution of the COII gene were observed by Crozier et al. (1989) in bee vs fly lineages. Increased mutational rate, however, seems an unlikely explanation because factors typically associated with increased mutational rate, such as high metabolic rate and short generation time (Rand 1994, review), either do not characterize Collembola or are not unique to Collembola.

The alternative explanation for the high level of divergence among collembolan species is that higher genetic divergence could simply be due to the age of the species and families. The first Collembola are believed to have appeared as early as 400 Myr ago, represented by *Rhyniella praecursor*, which is considered the first known fossil insect (Massoud 1967; Walley and Jarzembowski 1981). Although we have no information on the age of the families or species of Collembola we examined, this possibility seems the more likely of the two.

Our findings of extreme genetic divergence among collembolan families appear even more striking considering that all species we examined belong to the suborder Arthropleona, which is well differentiated from the other suborder Symphypleona. Both these suborders appear to be monophyletic based on morphology. These data thus provide some support for the suggestions of Anderson

Table 7. Average uncorrected sequence divergence values (above the diagonal) and HKY85 + Γ distances (below the diagonal) among genera[a]

	Orchesella	*Isotomurus*	*Tetrodontophora*	*Thaumanura*	*D. yakuba*	Dragonfly
Orchesella	0.196	0.286	0.320	0.331	0.322	0.346
	0.459	(0.250–0.303)	(0.311–0.329)	(0.321–0.341)	(0.311–0.341)	(0.336–0.351)
Isotomurus	0.985	0.190	0.315	0.369	0.354	0.360
	(0.697–1.259)	0.412	(0.303–0.330)	(0.353–0.381)	(0.348–0.365)	(0.357–0.363)
Tetrodontophora	1.541	1.443	—	0.331	0.338	0.368
	(1.443–1.731)	(1.228–1.846)		(0.330–0.332)		
Thaumanura	1.527	2.255	1.567	—	0.360	0.396
	(1.343–1.691)	(1.820–2.859)	(1.556–1.577)		(0.359–0.360)	(0.395–0.396)
D. yakuba	1.478	2.041	1.754	2.256	—	0.251
	(1.266–1.864)	(1.831–2.356)		(2.241–2.271)		
Dragonfly	1.790	2.215	2.373	2.818	0.685	—
	(1.543–1.994)	(2.105–2.286)		(2.798–2.838)		

[a] Ranges are indicated in parentheses. For genera where more than one species is availabale (*Orchesella* and *Isotomurus*), the same estimates are averaged among species and given on the diagonal (uncorrected: above; HKY85 + Γ—corrected: below)

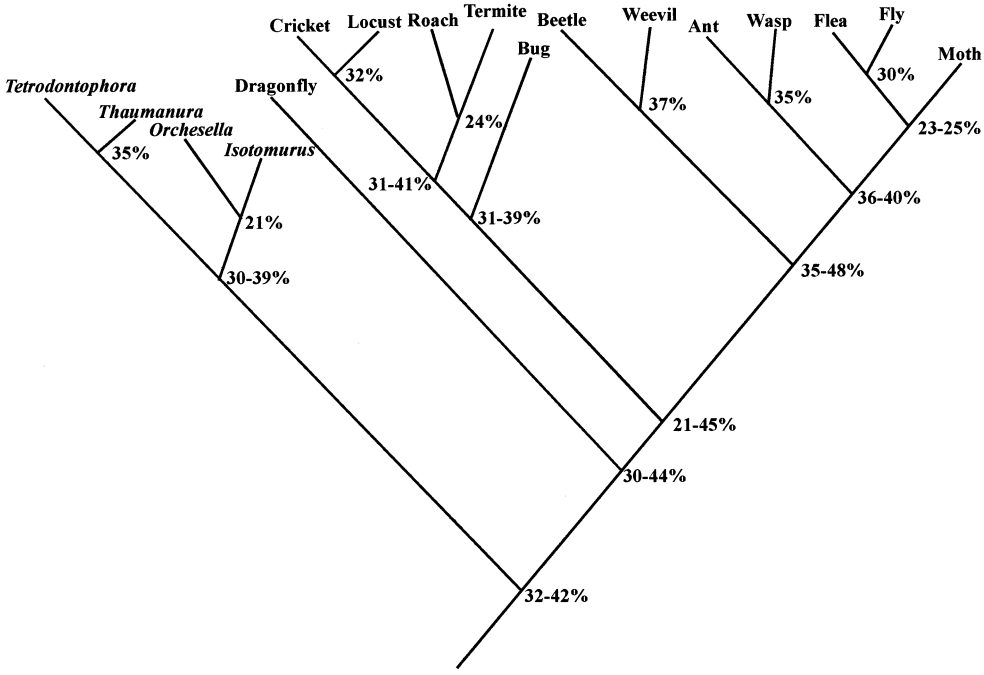

Fig. 2. COII amino acid divergence from pairwise comparisons of species from different insect orders (this work; and Liu and Beckenbach 1992) superimposed on an insect phylogenetic tree (Hennig 1981). Modified from Simon et al. 1994).

(1973) and Manton (1973) that Collembola may deserve elevation to the rank of class (with some or all collembolan families being raised to the rank of order).

Molecular Systematics

The information contained in these data provides a well-supported phylogenetic hypothesis for most of the species examined and is valuable for understanding the phylogenetic usefulness of the COII gene at different taxonomic levels within Collembola. At the family level, phylogenetic estimates from DNA data can be compared with those based on morphology. At the species level, in addition to morphology, DNA data can be compared with previously collected allozyme data for the genera *Orchesella* and *Isotomurus*.

Because third codon positions violate the assumption of stationary nucleotide composition and appear to be saturated, we have restricted discussion of phylogenetic analyses to first and second positions. In this data set (446 bp), 165 sites were variable (137 parsimony-informative). Initial trees were obtained form minimum-evolution (ME) analysis using LogDet distances, which produced one tree, and maximum parsimony (MP) analysis using equal weights, which produced six equally parsimonious trees (Fig. 3A), all different from the ME tree. The six MP trees differed from each other only in reso-

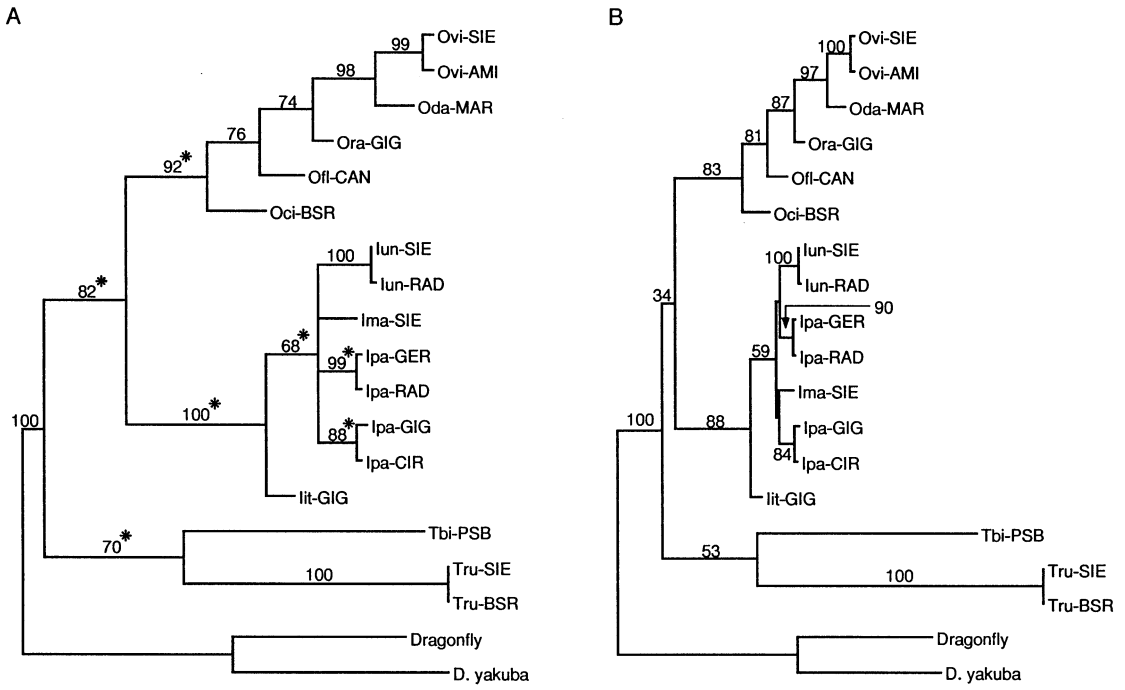

Fig. 3. **A** Strict consensus of six equally parsimonious trees derived using equal weights. **B** Maximum likelihood tree under the HKY85 + Γ model of evolution. In both trees, numbers above branches represent bootstrap values. (*Asterisks* indicate overestimation of support by parsimony compared to maximum likelihood.)

lution among some *Isotomurus* species (Fig. 3A). Parsimony analysis using a step matrix to down-weight transitions by a factor 2:1 resulted in three trees, a subset of the six equal-weight trees. Each of these seven trees (six MP trees plus the ME tree) was then evaluated according to the maximum-likelihood (ML) criterion under an array of models, optimizing all model parameters separately for each tree. The models were the Jukes-Cantor (JC: Jukes and Cantor 1969), Kimura two-parameter (K2P: Kimura 1980), HKY85 (Hasegawa et al. 1985), and general time-reversible (GTR, equals REV of Yang 1994) models. For each substitution model, the possibility of among-site rate variation was accommodated in four ways: (1) all sites assumed to evolve at the same rate; (2) some proportion of sites assumed to be invariable, with rates at the remaining sites assumed to be equal (e.g., Hasegawa et al. 1985); (3) rates assumed to follow a discrete approximation to the gamma distribution (Yang 1994); and (4) a mixture of invariable sites and gamma-distributed rates (e.g., Gu et al. 1995). Thus, 16 models (4 substitution models ×4 rate-distribution models) were tested on each tree.

For almost all models, parsimony tree #6 had the best likelihood score; the only exceptions were a few equal-base-frequency models, which are clearly inappropriate for these data. The likelihood score (Fig. 4) under the most general and parameter-rich model, GTR + I + Γ (−2,647.15848), is only slightly better than the HKY85 + Γ model (−2,650.26075). Because the latter model is nested within the former one, a likelihood-ratio test (e.g., Yang et al. 1995) can be used to assess whether simpler models provide an adequate fit to the data in comparison to the more general model. In this case, the observed difference in likelihood scores results in a likelihood-ratio test statistic of $2(3.102) = 6.204$. The GTR + I + Γ model requires estimation of five more free parameters from the data than HKY85 + Γ; therefore this test statistic can be compared to the χ^2 distribution with 5 *df*; the resulting *P* value is not significant ($0.5 > P > 0.1$). Because the two models were not significantly different and because the fewer number of parameters estimated by the simpler model improves the variance of the estimates, we used the HKY85 + Γ model to conduct a heuristic maximum-likelihood search (ten random input orders with TBR branch swapping), with parameters fixed to the values estimated above ($\alpha = 0.3291$; transition/transversion [TI/TV] ratio = 1.464). This search also selected a topology identical to parsimony tree #6 as the ML tree, and we are therefore reasonably confident that this is the ML tree under what we believe to be the most appropriate reconstruction model for these data (HKY85 + Γ).

All analyses of the first and second codon positions support the monophyly of each genus, unite *Isotomurus* (Isotomuridae) with *Orchesella* (Entomobryidae), and unite *Thaumanura* (Neanuridae) with *Tetrodontophora* (Onychiuridae). These results support the current interfamilial classification. Within *Isotomurus,* all analyses indicate *I. italicus* is basal, and *I. palustris* appears either poly- or paraphyletic, but no analysis provides strong support for relationships among *Isotomurus* species. These data do not reliably resolve the position of *I. maculatus,* but allozyme data (Frati et al. 1995) and the

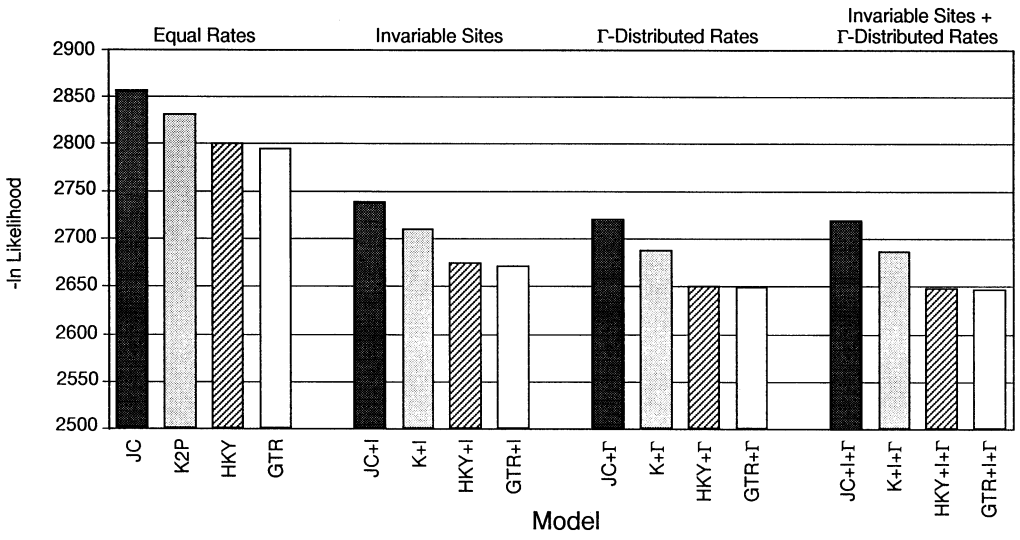

Fig. 4. Model choice. Likelihood scores for sixteen models were evaluated on parsimony tree #6. The HKY85 + Γ model seems to be the best compromise between goodness of fit and parameter economization.

sequence of a fragment of the large nuclear rDNA gene (Carapelli et al. 1995b) suggest a sister-species relationship with *I. unifasciatus.* Within the genus *Orchesella, O. dallaii* and *O. villosa* are consistently sister species, *O. ranzii* appears to be the sister species to those, and *O. cincta* is basal. The basal position of *O. cincta* is also suggested by allozyme data (Frati et al. 1992).

Although the parsimony and minimum-evolution analyses suggest a relationship between *Orchesella* and *Isotomurus,* there is little support for the relationship in the ML bootstrap analysis. All three analyses provide strong to moderate support for the relationship between *Thaumanura* and *Tetrodontophora.* It now appears that the COII gene is quite useful for resolving closely related taxa, but it seems too variable to give unequivocal information at the family level in the taxa we studied. This conclusion is similar to that of Liu and Beckenbach (1992), who found that the COII gene did not help in resolving phylogenetic relationships among orders of pterygote insects.

Patterns of Nucleotide Substitution

Excessive among-site rate variation can exert a major impact on the estimation of relationships among taxa in that ignoring it causes all methods of phylogenetic analysis to become biased due to the failure to adequately correct for superimposed substitutions (e.g., Kuhner and Felsenstein 1994; Yang et al. 1994). In addition, data sets with much rate heterogeneity can be more susceptible to the misleading effects of nonrandom noise than data sets with less rate heterogeneity (Sullivan et al. 1995). There is strong among-site rate variation in the data when all sites are examined together, when first and second sites are examined together, and when each of these is examined separately (α = 0.321, 0.329, 0.416, 0.301, respectivley). There is no evidence for rate variation among third-position sites (α > 300; Table 8), however this shape parameter must be interpreted cautiously because third codon positions violate the assumption of stationarity of nucleotide composition.

Table 8. Likelihood-ratio test for among-site rate variation in various partitions of the COII data. $\chi^2 = 2(\ln L_{\text{Gamma}} - \ln L_{\text{Single-Rate}})$

	ln likelihood			
Subset	SingleRate	Gamma	alpha	χ^2
All data	−7,144.315	−6,552.353	0.312	1,183.924*
1st & 2nd	−2,800.354	−2,650.261	0.329	300.186*
1st	−1,705.446	−1,618.676	0.416	171.540*
2nd	−1,034.455	−991.299	0.301	86.321*
3rd	−3,546.822	−3,546.822	>300	0.000 NS

* $P < 0.001$

Consistent with theoretical predictions (e.g., Wakeley 1996), many previous studies (reviewed in Simon et al. 1994) have shown that the observed TI/TV ratio changes with respect to the evolutionary distance of the taxa compared, with more transitions than transversions usually observed in comparisons among closely related taxa. In comparisons between more distantly related taxa, however, transversions tend to outnumber transitions due to multiple substitutions. Not surprisingly, this phenomenon is also evident in Collembola (Fig. 5). The bias toward transitions is most evident in comparisons between conspecific populations and is weaker between congeneric species. As the sequences become more randomized due to multiple substitutions between more distantly related taxa, transversions tend to outnumber transitions because there are twice as many kinds of transversions.

In addition to the transition bias, a pyrimidine bias

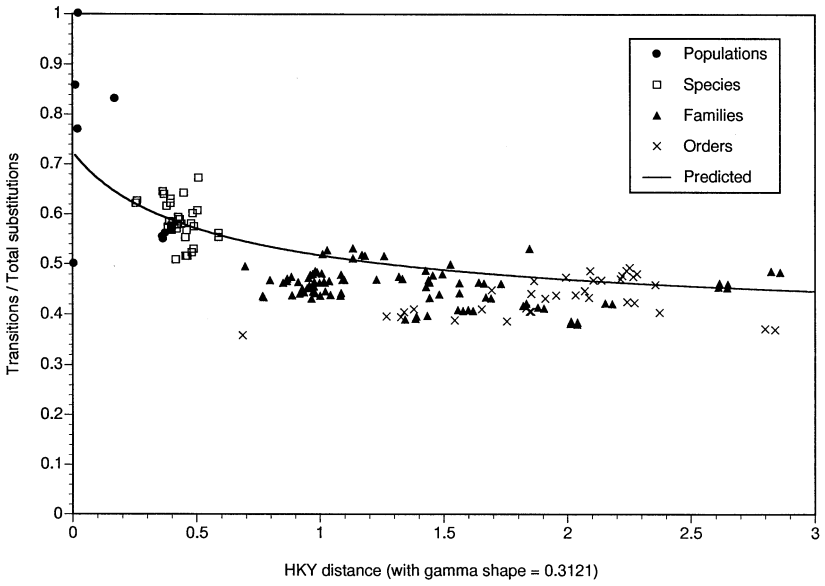

Fig. 5. Proportion of transitions observed (number of observed transitions over number of observed substitutions) vs HKY85 + Γ distances, plotted by taxonomic categories among Collembola. ''Orders'' refers to comparisons between collembolan specimens and dragonfly and *D. yakuba.* The curve represents the expected proportion of observed transitions under the HKY85 + Γ model. The deviation from expected is probably the result of third-codon positions evolving according to a very different process than first- and second-codon positions (Table 7).

was observed in transitional changes. The average number of T-C substitutions in each pairwise comparison was about twice as high (53.7) as the A-G substitutions (24.4). This bias, also reported for *Drosophila* (Garesse 1988; Tamura 1992) and mammals (Brown et al. 1982), is primarily due to the unequal base composition in the sequences rather than to a more fundamental difference in the (instantaneous) rates of substitution. Were the latter true, the general time-reversible model (GTR), which allows all six substitution types to have distinct rate parameters, would have yielded significantly higher likelihood scores relative to the HKY85 model, which only allows transitions and transversions to have different rates. The biased nucleotide composition in turn probably relates to the strand asymmetry in the replication process of mitochondrial DNA (Tamura 1992).

Acknowledgments. We thank A. Carapelli, R. Dallai, P.P. Fanciulli, N. Goldman, K. Holsinger, D. Penny, D. Pollock, and Z. Yang for assistance and valuable discussion. G. Spicer, A. Beckenbach, and an anonymous reviewer provided useful comments. This work was partly supported by grants from the Italian MURST (40% and 60%) and CNR to F.F., by a grant from the US National Science Foundation to C.S., and by a Smithsonian Institution Molecular Evolution Fellowship to J.S.

References

Adkins RM, Honeycutt RL (1994) Evolution of the Primate cytochrome *c* oxidase subunit II gene. J Mol Evol 38:215–231

Anderson DT (1973) Embryology and phylogeny in anellids and arthropods. Pergamon Press, Oxford

Anderson S, Bankier AT, Barrell BH, de Bruijn MHL, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, Schreier PH, Smith AJH, Staden R, Young IG (1981) Sequence and organization of the human mitochondrial genome. Nature 290:457–465

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

Beckenbach AT, Wei YW, Liu H (1993) Relationships in the *Drosophila obscura* species group, inferred from mitochondrial cytochrome oxidase II sequences. Mol Biol Evol 10:619–634

Bibb MJ, Van Ettenn RA, Wright CT, Walberg MW, Clayton DA (1981) Sequence and gene organization of mouse mitochondrial DNA. Cell 26:167–180

Bitsch J (1994) The morphological groundplan of Hexapoda: critical review of recent concepts. Ann Soc Entomol Fr (NS) 30:103–129

Brower AVZ (1994) Phylogeny of *Heliconius* butterflies inferred from mitochondrial DNA sequences (Lepidoptera: Nymphalidae). Mol Phylog Evol 3:159–174

Brown WM (1985) The mitochondrial genome of animals. In: MacIntyre RJ (ed) Molecular evolutionary genetics. Plenum, New York, pp 95–130

Brown GG, Simpson MV (1982) Novel features of animal mtDNA evolution as shown by sequences of two rat cytochrome oxidase subunit II genes. Proc Natl Acad Sci USA 79:3246–3250

Brown JM, Pellmyr O, Thompson JM, 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

Brown WM, Prager EM, Wang A, Wilson AC (1982) Mitochondrial DNA sequences of primates: tempo and mode of evolution. J Mol Evol 18:225–239

de Bruijn MHL (1983) *Drosophila melanogaster* mitochondrial DNA, a novel gene organization and genetic code. Nature 304:234–241

Capaldi RA (1990) Structure and function of cytochrome *c* oxidase. Annu Rev Biochem 59:569–596

Capaldi RA, Malatesta F, Darley-Usmar VM (1983) Structure of cytochrome *c* oxidase. Biochem Biophys Acta 726:135–148

Carapelli A, Fanciulli PP, Frati F, Dallai R (1995a) The use of genetic markers for the diagnosis of sibling species in the genus *Isotomurus* (Insecta, Collembola). Boll Zool 62:71–76

Carapelli A, Frati F, Fanciulli PP, Dallai R (1995b) Genetic differentiation of six sympatric species of *Isotomurus* (Collembola, Isotomidae); is there any difference in their microhabitat preference? Eur J Soil Biol 31:87–99

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

Clayton DA (1984) Transcription of the mammalian mitochondrial genome. Annu Rev Biochem 53:573–594

Crozier RH, Crozier YC (1993) The mitochondrial genome of the

honeybee *Apis mellifera:* complete sequence and genome organization. Genetics 133:97–117

Crozier RH, Crozier YC, Mackinlay AG (1989) The CO-I and the CO-II region of honeybee mitochondrial DNA: evidence for variation in insect mitochondrial evolutionary rates. Mol Biol Evol 6: 399–411

Disotell TR, Honeycutt RL, Ruvolo M (1992) Mitochondrial DNA phylogeny of the old-world monkey Tribe Papionini. Mol Biol Evol 9:1–13

Fanciulli PP, Frati F, Dallai R, Rusek J (1991) High genetic divergence among populations of *Tetrodontophora bielanensis* (Insecta, Collembola) in Europe. Rev Ecol Biol Sol 28:165–173

Fearnley IM, Walker JE (1987) Initiation codons in mammalian mitochondria. Differences in genetic code in the organelle. Biochemistry 26:8247–8251

Frati F, Fanciulli PP, Dallai R (1992) Genetic diversity and taxonomy in soil dwelling insects: the genus *Orchesella.* J Hered 83:275–281

Frati F, Carapelli A, Fanciulli PP, Dallai R (1995) The genus *Isotomurus:* where molecular markers help to evaluate the importance of morphological characters for the diagnosis of species. Pol Pismo Entomol 64:41–51

French S, Robson B (1983) What is a conservative substitution? J Mol Evol 19:171–175

Garesse R (1988) *Drosophila melanogaster* mitochondrial DNA: gene organization and evolutionary considerations. Genetics 118:649–663

Gu X, Fu Y, Li W (1995) Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. Mol Biol Evol 546–557

Hasegawa M, Kishino M, Yano T (1985) Dating the human-ape split by a molecular clock of mitochondrial DNA. J Mol Evol 22:160–174

Hennig W (1981) Insect systematics. Wiley, New York

Higgins DG, Bleasby AJ, Fuchs R (1992) CLUSTAL V: improved software for multiple sequence alignment. Comput Appl Biosci 8:189–191

Hsiao TH (1993) Molecular techniques for studying systematics and phylogeny of Chrysomelidae. In: Jolivet P, Cox ML, Petitpierre E (eds) Novel aspects of the biology of Chrysomelidae. Kluwer Academic, Dordrecht, pp 237–248

Irwin DM, Kocher TD, Wilson AC (1991) Evolution of the Cytochrome *b* gene of Mammals. J Mol Evol 32:128–144

Iwata S, Ostermeier C, Ludwig B, Michel H (1995) Structure at 2.8 Å resolution of cytochrome *c* oxidase from *Paracoccus denitrificans.* Nature 376:660–669

Jermiin LS, Crozier RH (1994) 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

Jukes TH, Cantor CR (1969) Evolution of protein molecules. In: Munro HN (ed) Mammalian protein metabolism. Academic, New York, pp 21–132

Kidd KK, Cavalli-Sforza LL (1971) Number of characters examined and error in reconstruction of evolutionary trees. In: Hodson FR, Tautu P (eds) Mathematics in the archeological and historical sciences. Edinburgh University Press, Edinburgh

Kimura M (1980) A simple method for estimating evolutionary rate of base substitution through comparative studies of nucleotide sequences. J Mol Evol 16:111–120

Kristensen NP (1981) Phylogeny of insect orders. Annu Rev Entomol 26:135–157

Kuhner MK, Felsenstein J (1994) A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol Biol Evol 11:459–468

Kumar S, Tamura K, Nei M (1993) MEGA: molecular evolutionary genetic analysis, version 1.01. The Pennsylvania State University, University Park, PA 16802

Lake JA (1994) Reconstructing evolutionary trees from DNA and protein sequences: paralinear distances. Proc Natl Acad Sci USA 91: 1455–1459

Lecanidou R, Douris V, Rodakis GC (1994) Novel features of metazooan mtDNA revealed from sequence analysis of three mitochondrial DNA segments of the land snail *Albinaria turrita* (Gastropoda: Clausiliidae). J Mol Evol 38:369–382

Liu H, Beckenbach AT (1992) Evolution of the mitochondrial cytochrome oxidase II gene among 10 orders of insects. Mol Phylog Evol 1:41–52

Lockhart PJ, Steel MA, Hendy MD, Penny D (1994) Recovering evolutionary trees under a more realistic model of sequence evolution. Mol Biol Evol 11:605–612

Manton SM (1973) Arthropod phylogeny—a modern synthesis. J Zool Lond 171:111–130

Manton SM (1979) Functional morphology and the evolution od hexapod classes. In: Gupta A (ed) Arthropod phylogeny. Van Nostran, New York, pp 387–466

Massoud Z (1967) Contribution à l'étude de *Rhyniella praecursor* Hirst et Maulik 1926, Collembole fossile du Dévonien. Rev Ecol Biol Sol 4:497–505

Massoud Z (1971) Contribution à la connaissance morphologique et systématique des Collemboles Neelidae. Rev Ecol Biol Sol 8:195–198

Millett F, de Jong C, Paulson L, Capaldi RA (1983) Identification of specific carboxylate groups on cytochrome *c* oxidase that are involved in binding cytochrome *c*. Biochemistry 22:546–552

Mitchell SE, Cockburn AF, Sewright JA (1993) The mitochondrial genome of *Anopheles quadrimaculatus* species A: complete nucleotide sequence and gene organization. Genome 36:1058–1073

Osawa S, Jukes TH, Watanabe K, Muto A (1992) Recent evidence for evolution of the genetic code. Microbiol Rev 56:229–264

Rand DM (1994) Thermal habit, metabolic rate and the evolution of mitochondrial DNA. Trends Ecol Evol 9:125–131

Rzhetsky A, Nei M (1992) A simple method for estimating and testing minimum-evolution trees. Mol Biol Evol 9:945–967

Ruvolo M, Disotell TR, Allard MW, Brown WM, Honeycutt RL (1991) Resolution of the African hominoid trichotomy by use of a mitochondrial gene sequence. Proc Natl Acad Sci USA 88:1570–1574

Sharp PM, Tuohy TMF, Mosurski KR (1986) Codon usage in yeasts: cluster analysis clearly differentiates highly and lowly expressed genes. Nucleic Acids Res 14:5125–5143

Simon C, Franke A, Martin A (1991) The polymerase chain reaction: DNA extraction and amplification. In: Hewitt GM, Johnston AWB, Young JPW (eds) Molecular techniques in taxonomy. NATO ASI Series, Springer-Verlag, Berlin, pp 329–355

Simon C, Frati F, Beckenbach AT, 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 Entomol Soc Am 87:651–701

Sperling FAH, Hickey DA (1994) Mitochondrial DNA sequence variation in the spruce budworm species complex (*Choristoneura:* Lepidoptera). Mol Biol Evol 11:656–665

Sperling FAH, Landry J-F, Hickey DA (1994) A DNA-based approach to the identification of insect species used for postmortem interval estimation. J Forensic Sci 39:418–427

Sperling FAH, Landry J-F, Hickey DA (1995) DNA-based identification of introduced ermine moth species in North America (Lepidoptera: Yponomeutidae: *Yponomeuta padella* complex). Ann Entomol Soc Am 88:155–162

Stevens TH, Martin CT, Wang H, Brudvig GW, Scholes CP, Chan SI (1982) The nature of Cu_A in cytochrome *c* oxidase. J Biol Chem 257:12106–12113

Sullivan J, Holsinger KE, Simon C (1995) Among-site rate variation and phylogenetic analysis of 12S rRNA in sigmodontine rodents. Mol Biol Evol 12:988–1001

Swofford DL, Olsen GJ, Waddell PJ, Hillis DM (1996) Phylogenetic

inference. In: Hillis DM, Moritz C, Mable BK (eds) Molecular systematics. 2nd ed. Sinauer, Sunderland, MA, pp 407–514

Tamura K (1992) The rate and pattern of nucleotide substitution in *Drosophila* mitochondrial DNA. Mol Biol Evol 9:814–825

Waddell PJ (1995) Statistical methods of phylogenetic analysis: including Hadamard conjugations, LogDet transforms and maximum likelihood. PhD thesis, Massey University

Wakeley J (1996) The excess of transitions among nucleotide substitutions: new methods of estimating transition bias underscore its importance. Trends Ecol Evol 11:158–162

Walley P, Jarzembowski EA (1981) A new assessment of *Rhyniella*, the earliest known insect, from the Devonian of Rhynie, Scotland. Nature 291:317

Willis LG, Winston ML, Honda BM (1992) Phylogenetic relationships in the honeybee (genus *Apis*) as determined by the sequence of the cytochrome oxidase II region of mitochondrial DNA. Mol Phylog Evol 1:169–178

Yang Z (1994) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol 39:306–314

Yang Z, Goldman N, Friday A (1994) Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. Mol Biol Evol 11:316–324

Yang Z, Goldman N, Friday A (1995) Maximum likelihood trees from DNA sequences: a peculiar statistical estimation problem. Syst Biol 44:384–399