Elongation Factor-1α Occurs as Two Copies in Bees: Implications for Phylogenetic Analysis of EF-1α Sequences in Insects

Bryan N. Danforth and Shuqing Ji

Department of Entomology, Cornell University

We report the complete sequence of a paralogous copy of elongation factor-1α (EF-1α) in the honeybee, *Apis mellifera* (Hymenoptera: Apidae). This copy differs from a previously described copy in the positions of five introns and in 25% of the nucleotide sites in the coding regions. The existence of two paralogous copies of EF-1α in *Drosophila* and *Apis* suggests that two copies of EF-1α may be widespread in the holometabolous insect orders. To distinguish between a single, ancient gene duplication and parallel, independent fly and bee gene duplications, we performed a phylogenetic analysis of hexapod EF-1α sequences. Unweighted parsimony analysis of nucleotide sequences suggests an ancient gene duplication event, whereas weighted parsimony analysis of nucleotides and unweighted parsimony analysis of amino acids suggests the contrary: that EF-1α underwent parallel gene duplications in the Diptera and the Hymenoptera. The hypothesis of parallel gene duplication is supported both by congruence among nucleotide and amino acid data sets and by topology-dependent permutation tail probability (T-PTP) tests. The resulting tree topologies are also congruent with current views on the relationships among the holometabolous orders included in this study (Diptera, Hymenoptera, and Lepidoptera). More sequences, from diverse orders of holometabolous insects, will be needed to more accurately assess the historical patterns of gene duplication in EF-1α.

Introduction

Elongation factor-1α (EF-1α) is a nuclear protein-coding gene involved in the GTP-dependent binding of charged tRNAs to the acceptor site of the ribosome during translation (Maroni 1993, pp. 126–134). In *Drosophila*, EF-1α occurs as two copies, EF-1α F1 and EF-1α F2, which are expressed at different times during development (Hovemann et al. 1988). EF-1α genes have been characterized in other animals, including brine shrimp (*Artemia*; Lenstra et al. 1986), mice (Rao and Slobin 1986; Roth et al. 1987), humans (Brands et al. 1986), and honeybees (*Apis mellifera*; Walldorf and Hovemann 1990). Because of the conserved nature of the amino acid sequence among these disparate organisms, EF-1α has been identified as a potentially useful gene for studies of higher-level phylogenetic relationships, especially in insects (Friedlander, Regier, and Mitter 1992, 1994; Brower and DeSalle 1994; Mitchell et al. 1997; Belshaw and Quicke 1997). Amino acid sequences of EF-1α have recently been used to resolve evolutionary relationships among early eukaryotes (Hasegawa et al. 1993; Kamaishi et al. 1996) and among arthropod classes (Regier and Shultz 1997).

Contrary to an earlier report of a single copy of EF-1α in honeybees (Walldorf and Hovemann 1990), we have identified and characterized an additional copy present in representatives of all major bee families surveyed. The two copies in bees, as in *Drosophila*, differ in intron position and in nucleotide sequence. We have completely characterized the sequence of the paralogous copy in *Apis mellifera* and report here its intron/exon structure, sequence, and relationship to other EF-1α sequences reported for insects.

Key words: elongation factor-1α, insects, *Apis mellifera*, intron/exon evolution, gene duplication, phylogeny.

Address for correspondence and reprints: Bryan N. Danforth, Department of Entomology, Comstock Hall, Cornell University, Ithaca, New York 14853-0901. E-mail: bnd1@cornell.edu.

Mol. Biol. Evol. 15(3):225–235. 1998

The existence of two paralogous copies of EF-1α in a diverse array of bees, ants (T. Schultz, personal communication), and flies raises the possibility that two copies are widespread in the Holometabola. This has important implications for using EF-1α in higher-level phylogenetic studies of insects, where paralogous copies may be confused.

Materials and Methods

Fresh, frozen, and alcohol-preserved specimens were used in the initial DNA extractions, and all gave satisfactory results. Specimens were briefly frozen in liquid nitrogen and ground in individual 1.5-ml Eppendorf tubes in the presence of 2 × CTAB extraction buffer and 100 μg of proteinase K. Tubes were incubated for 2 h at 55°C and homogenates were extracted with chloroform-isoamylalcohol, digested for 30 min in the presence of 10 μg RNase, and then extracted again with phenol-chloroform-isoamylalcohol and chloroform-isoamylalcohol, in that order. The DNA was precipitated with 2.5 volumes of ice-cold ethanol and 0.1 volume 3 M sodium acetate, washed once in 80% ethanol, and resuspended in 50 μl Tris-EDTA (pH 7.6) buffer.

PCR primers for amplification of EF-1α (see below) were based on a comparison of the published sequences in *Apis*, *Drosophila*, and heliothine moths. Additional, apoid-specific, primers were developed based on preliminary sequencing results. These primers worked effectively as PCR and sequencing primers in most species of bees tested.

Characterization of the upstream (5′) sequence in *Apis mellifera* was accomplished by cassette-ligation-mediated PCR (Isegawa et al. 1992). This procedure involves completely digesting total genomic DNA (or cDNA) with appropriate restriction enzymes and then ligating onto the ends of the restriction fragments roughly 50 bp double-stranded DNA cassettes with overhanging ends complementary to those generated by each restriction enzyme (Takara LA PCR in vitro cloning kit

Table 1
PCR Primers Used in Cassette-Ligation-Mediated PCR and RT-PCR

	Positions
Takara cassette-specific primers[a]	
Takara C1: 5′-GTA CAT ATT GTC GTT AGA ACG CGT A-3′	
Takara C2: 5′-GCG TAA TAC GAC TCA CTA TAG GGA GA-3′	
Apis-specific forward primers	
Apis F2-forward 1: 5′-CAT CGT TAT GCT TGT GCC AAG-3′	2,081–2,101
Apis F2-forward 2: 5′-GCT TTC CAA GAA TTT CCG CCT T-3′	2,121–2,142
Apis F2-forward 3: 5′-TCG TAA TGG AAA GAC AAC TGA AGA-3′	2,024–2,047
Apis-specific reverse primers	
F2-reverse 1: 5′-A ATC AGC AGC ACC TTT AGG TGG-3′	1,598–1,619
Apis F2-reverse 1: 5′-AGC AAC ATA ACC ACG ACG TAA TTC-3′	1,556–1,579
Apis F2-reverse 2: 5′-GAA ATC TCT GTG TCC AGG AGC ATC-3′	632–655
Apis F2-reverse 3: 5′-ACG TTT CGA ATT TCC ACA AAG C-3′	587–608
Apis F2-reverse 4: 5-TAG CGT TGC TCT CGT GCG AG-3′	2,347–2,366

[a] Slightly modified from those provided by the manufacturer.

[cat. no. TAK-RR015]; available from PanVera, Madison, Wis.). Following ligation, the DNA is precipitated with ethanol and resuspended in a small volume of water (5 μl). The cassette-ligated restriction fragments are then used as template for one round of PCR using one primer specific to the target sequence and another primer specific to the cassette (see table 1 for primers used). A second round of PCR using another set of primers nested slightly inside of the first typically produces a single PCR product that can be sequenced directly or cloned into a T/A vector (Promega, Madison, Wis.).

To characterize the downstream (3′) sequences of the gene, we used RT-PCR (Access RT-PCR System, Promega) with forward primers listed in table 1 (*Apis* F2-forward 1, *Apis* F2-forward 2, and *Apis* F2-forward 3) and a poly-T reverse primer. RNA was extracted from adult worker bees using the Ultra-spec RNA isolation system (Biotecx Labs, Houston, Tex.).

All DNA sequencing was done with an ABI 373A automated sequencer, using the PCR primers as sequencing primers (end primers). This procedure gave good sequencing results for up to 700 bp. Sequencing was done in both directions.

Phylogenetic analyses of nucleotide sequences, amino acid sequences, and intron positions were performed using test versions of PAUP*4 (PAUP versions 4.0d54, 4.0d56, 4.0d57, and 4.0d59; D. Swofford, personal communication; see Swofford [1993] for details on earlier versions of the program). For parsimony analyses of nucleotides, amino acids, and intron presence/absence, we used either the exhaustive search option or heuristic search with TBR branch swapping, random addition sequence for taxa, and 500 replicates per search. For bootstrap analysis (Felsenstein 1985), we used 500 replicates. In order to evaluate the extent to which the data significantly support overall tree topologies and specific monophyletic groupings, we used the permutation tail probability (PTP) test and the topology-dependent permutation tail probability (T-PTP) test (Archie, 1989*a*, 1989*b*; Faith 1990, 1991; Faith and Cranston 1991).

A maximum-likelihood analysis of nucleotide data (Felsenstein 1981, as described in Swofford et al. 1996) was implemented in PAUP*4. Nucleotide frequencies were determined empirically, and we used both the Hasegawa, Kishino, and Yano (1985) model and the Felsenstein (1984) two-parameter model for unequal base frequencies. Both the transition : transversion ratio and the shape parameter of the gamma distribution (α) were determined empirically within the analysis by maximum likelihood.

A maximum-likelihood analysis of the amino acid sequences was performed using the program PUZZLE, version 4.0 (Strimmer and Von Haeseler 1996). Amino acid frequencies were determined empirically, rate heterogeneity was estimated from the data by maximum likelihood using a mixed model (one invariable and four gamma distribution rates) or gamma distributed rates, and the model of substitution was based on Dayhoff, Schwartz, and Orcutt (1978). Only one outgroup can be selected in PUZZLE, so the results are not necessarily comparable to those obtained by PAUP*4.

We used the Kishino and Hasegawa (1989) test as implemented in PHYLIP, version 3.572c (Felsenstein 1993), to evaluate the statistical significance of the alternative tree topologies obtained under parsimony and maximum likelihood.

MacClade (Maddison and Maddison 1992) was used to map characters on trees and to investigate alternative tree topologies.

GenBank accession numbers for previously published sequences used in the this study are listed in *Acknowledgments.* The sequence of the *Apis mellifera* F2 copy (described herein) was submitted to GenBank under accession number AF015267.

Results

Evidence that Two Copies of EF-1α Are Present in All Major Bee Families

Degenerate PCR primers initially developed based on comparisons of *Apis* (Walldorf and Hovemann 1990), *Drosophila* (Hovemann et al. 1988), and heliothine moths (Cho et al. 1995) nonspecifically amplified two paralogous copies in bees. These primers were EF1-For3 (5′-GGN GAC AA[C/T] GTT GG[T/C] TTC AAC G-3′; the 5′ end corresponds to position 1496 in *Apis mellifera* [Walldorf and Hovemann 1990]) and Cho10 (5′-

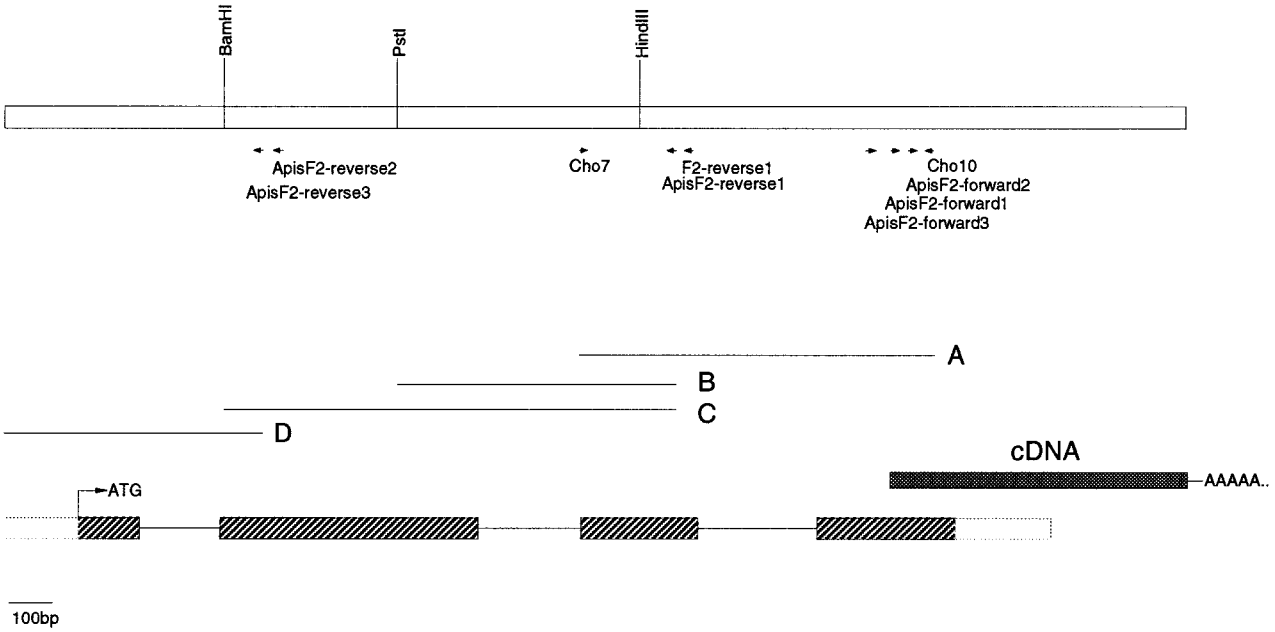

FIG. 1.—Intron/exon structure of the newly characterized copy of EF-1α (EF-1α F2) in *Apis mellifera*. Bars indicate exons, lines indicate introns, and dashed lines indicate noncoding sequence. Restriction sites are shown for *Bam*H1, *Eco*RI, *Hin*dIII, *Pst* I, and *Sau*3A. Primer sites are indicated by arrows.

AC [A/G]GC [A/G/C]AC [G/T]GT [T/C]TG [A/T/C]C[T/G] CAT GTC-3′; the 5′ end corresponds to position 1887 in *Apis mellifera* [Walldorf and Hovemann 1990] and partially matches the sequence of primer rcM4 in Cho et al. [1995]; fig. 1). Together, these primers were expected to produce a single 392-bp PCR product corresponding to the EF-1α sequence reported by Walldorf and Hovemann (1990). However, in all bees tested, representing most major bee families (including Colletidae, Andrenidae, Halictidae, and Apidae), we obtained two bright bands, even at high annealing temperatures (>64°C). One band corresponded to the expected 392-bp PCR product, whereas a larger (roughly 600-bp) band was also obtained. Following gel purification and sequencing of these two PCR products, we confirmed that the larger PCR product represented a paralogous copy of EF-1α with an approximately 200-bp intron located between the two primer sites. Following this initial discovery, we determined the complete sequence of the paralogous copy in *Apis mellifera* (as described below) in order to compare the two copies in bees with copies previously reported from other insect orders.

Complete Sequence for *Apis mellifera*

For honeybees, we initially obtained sequences of the two paralogous copies by using, in combination, primers Cho7 (5′-CA[A/G] GAC GTN TA[T/C] AA[A/G] AT[T/C] GG-3′; the 5′ end corresponds to position 1115 in *Apis mellifera* [Walldorf and Hovemann 1990] and partially matches the sequence of primer M51.9 in Cho et al. [1995]; fig. 1) and Cho10 (see above; fig. 1). These primers produced two bands that we could gel-purify and sequence directly. The smaller band corresponded exactly to the published *Apis* sequence (herein called the F1 copy; Walldorf and Hovemann 1990), while the larger band differed substantially from the published sequence, primarily at third-position sites (herein called the F2 copy; fragment A in fig. 1).

We then used cassette-ligation-mediated PCR to obtain the upstream portions of the F2 copy. In our initial digests of honeybee DNA, we used four restriction enzymes: *Bam*HI, *Eco*RI, *Hin*dIII, and *Pst* I. In the initial round of screening, we used *Apis*-specific reverse primers F2-reverse 1 and *Apis* F2-reverse 1 (table 1 and fig. 1). Following the second round of PCR, the *Pst* I digest yielded a roughly 700-bp PCR product (B in fig. 1), and the *Bam*HI digest yielded a roughly 1,200-bp PCR product (C in fig. 1). Direct sequencing of these PCR products indicated that they were identical in sequence (where they overlapped) and overlapped broadly with our known target sequence (the Cho7/Cho10 PCR product).

In a second round of screening, we designed new *Apis*-specific primers (*Apis* F2-reverse 2 and *Apis* F2-reverse 3; table 1 and fig. 1) and, on the second round of PCR amplification, obtained a single roughly 655-bp PCR product from the *Eco*R1-digested DNA (D in fig. 1). This fragment extended well upstream of the start codon in the *Apis* F2 copy.

Using RT-PCR, we obtained a cDNA fragment using forward primers *Apis* F2-forward 3 and *Apis* F2-forward 1 with a poly-T reverse primer (fig. 1). Having obtained the cDNA sequence, we designed a reverse primer downstream of the stop codon (*Apis* F2-reverse 4) and, using *Apis* F2-forward 3 as a forward primer, obtained a PCR product from genomic DNA that confirmed the cDNA sequence for the 3′ end of the gene. This allowed us to completely characterize the intron/exon structure of the newly described copy.

Table 2
Base Composition for the 11 Sequences Included in this Study

	A	C	G	T	*P*
Overall	0.260 ± 0.0228	0.247 ± 0.0413	0.259 ± 0.0265	0.234 ± 0.0364	0.000***
First position	0.300 ± 0.0113	0.172 ± 0.0138	0.379 ± 0.010	0.149 ± 0.0076	0.999 NS
Second position	0.325 ± 0.0055	0.245 ± 0.0077	0.160 ± 0.0041	0.270 ± 0.0021	1.0 NS
Third position	0.158 ± 0.0633	0.323 ± 0.1109	0.234 ± 0.0782	0.285 ± 0.103	0.000***

NOTE.—Data are presented as mean ± standard deviation. Significance values were based on a chi-square test (df = 30) for homogeneity across taxa (as implemented in PAUP*4).

In all, we sequenced a 2,762-bp portion of the F2 copy in the honeybee, which includes the entire 1,386-bp coding sequence.

Comparison of the *Apis* F2 Copy and the Previously Reported Sequences in Insects

We aligned the complete coding sequence of the *Apis* F2 copy with the complete coding sequences of the published *Apis* F1 copy, the two complete *Drosophila* sequences (F1 and F2), and the complete *Artemia* sequence (Lenstra et al. 1986), as well as with partial sequences for three basal noctuid moths (Cho et al. 1995), a cockroach (*Periplaneta americana*), a bristletail (*Pedetontus saltator*), and a collembolan (*Tomocerus* sp.; the latter three sequences are from Regier and Shultz 1997) using MegAlign in the Lasergene software package (DNASTAR, Madison, Wis.). Alignments were unambiguous in the protein-coding regions, and only two one-codon indels were observed (alignments are available from the authors).

Base Composition

Base compositions for the 10 hexapod and 1 crustacean sequences were similar to those observed by Mitchell et al. (1997) for noctuid moths. For the 11 sequences included, there was a weak but statistically significant base compositional bias (table 2). However, as in the data set of Mitchell et al. (1997), the nucleotide composition varied significantly by site (table 2), with A and G most common in first positions, A and T most common in second positions, and third positions most variable in base composition. The sequence of the newly characterized *Apis* F2 copy was slightly A/T-biased relative to the seven other sequences (0.296, 0.178, 0.233, 0.293 [ACGT]; table 2), primarily due to A/T-bias in the third positions. First and second positions generally conformed to the overall pattern for the 10 other sequences. Using either uncorrected or LogDet distances (Lockhart et al. 1994) gave the same estimates of sequence divergence among taxa ($r = 0.996$***), suggesting that base-compositional bias is not a significant problem in this data set.

Sequence Divergence

Table 3 shows the uncorrected nucleotide and amino acid divergences among the 11 sequences. The two *Apis* copies differ from each other by 25.0% overall, with most differences confined to third positions; divergences were 7.5%, 4.1%, and 63.2% for first, second, and third positions, respectively. Nucleotide sequence divergence between the two *Drosophila* copies was 18.6% overall. Figure 2 shows the relationship between sequence divergence at third positions and that at first and second positions for the seven insect and *Artemia* sequences. Figure 2 includes comparisons between paralogous copies as well as among orthologous loci.

Intron Position

The F2 copy has three introns located at the following positions: 144/145, 753/754, and 1029/1030. None of the three *Apis* F2 introns corresponds in location to either of the two *Apis* F1 introns (fig. 3).

In order to determine if the two copies of EF-1α in *Apis* could be interpreted as homologs of the two copies of EF-1α in *Drosophila*, we examined intron position as a criterion of similarity (the moth sequences lack introns [Cho et al. 1995]). Alignment of coding regions revealed that intron positions are shared between

Table 3
Uncorrected Pairwise Divergences Between Amino Acid (below diagonal) and Nucleotide (above diagonal) Sequences (excluding introns)

	Artemia	*Pedetontus*	*Periplaneta*	*Tomocerus*	*Dros* F1	*Dros* F2	*Apis* F1	*Apis* F2	*Basilodes*	*Trichoplusia*	*Spodoptera*
Artemia	—	0.23060	0.23118	0.24773	0.23123	0.25774	0.26698	0.23033	0.21520	0.21149	0.21141
Pedetontus	0.12637	—	0.18152	0.21003	0.24149	0.24872	0.27363	0.21607	0.20834	0.21620	0.21475
Periplaneta	0.10714	0.05495	—	0.21886	0.20841	0.23251	0.26642	0.19321	0.18627	0.20076	0.19596
Tomocerus	0.12912	0.07967	0.07143	—	0.22776	0.25781	0.25578	0.20101	0.21475	0.21189	0.22373
Dros F1	0.11905	0.09341	0.09615	0.12088	—	0.18642	0.21950	0.23990	0.15254	0.15188	0.14413
Dros F2	0.11255	0.08791	0.08516	0.11813	0.09307	—	0.23094	0.26711	0.21327	0.19995	0.20166
Apis F1	0.10846	0.09066	0.09066	0.08516	0.10195	0.09328	—	0.24964	0.23845	0.23696	0.23866
Apis F2	0.11280	0.07692	0.04945	0.08242	0.08026	0.08243	0.07375	—	0.20834	0.21126	0.21947
Basilodes	0.10437	0.06319	0.05495	0.10165	0.06068	0.6796	0.08495	0.05340	—	0.06935	0.05887
Trichoplusia	0.09709	0.06319	0.05769	0.09890	0.06553	0.7039	0.08495	0.05583	0.00728	—	0.06532
Spodoptera	0.10194	0.06319	0.05769	0.09890	0.06311	0.7039	0.08252	0.05583	0.00728	0.00728	—

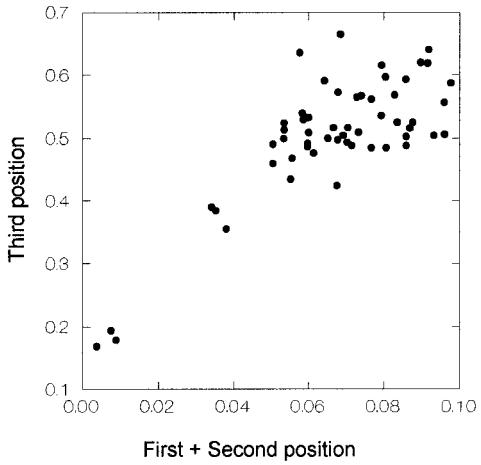

FIG. 2.—Sequence divergence (uncorrected) at third positions plotted against first and second position divergence for all pairwise comparisons among the seven sequences included in this study.

the *Drosophila* and *Apis* copies, but the positions are uninformative as to the historical relationships among copies (fig. 3). The *Drosophila* F2 copy shares a single intron position with the *Apis* F1 copy (position 823/824), but also shares an intron position with the *Apis* F2 copy (position 1029/1030). Neither of these shared introns can be taken as evidence of an ancient shared duplication being responsible for the existence of two paralogous copies in *Apis* and *Drosophila.*

With the inclusion of *Artemia,* which shares some introns with the insect sequences, we were able to identify seven possible intron locations and, thus, we coded intron presence/absence as a character (fig. 3 and table 4). Because the sequences from the cockroach, the bristletail, and the collembolan were based on cDNA sequences (89% of the complete data set; data missing on two of seven intron positions [Regier and Shultz 1997]), they were excluded from the analysis.

Table 4
Data Matrix for Analysis of Intron Presence/Absence

	Intron No.						
Taxon	1	2	3	4	5	6	7
Artemia	1	1	0	0	1	1	0
Apis F1	0	0	0	1	0	0	1
Apis F2	1	0	1	0	1	0	0
Dros F1.	0	0	0	0	0	0	0
Dros F2.	0	0	0	1	1	0	0
Basilodes	0	0	0	0	0	0	—
Trichoplusia	0	0	0	0	0	0	—
Spodoptera	0	0	0	0	0	0	—

NOTE.—Intron positions shown in Figure 3.

We analyzed intron position in a parsimony analysis with *Artemia* as the outgroup and obtained 29 equally parsimonious trees. All trees required eight steps and had a consistency index (CI) of 0.8750 and a retention index (RI) of 0.75. Five of the seven intron positions were congruent with all tree topologies (all five had a CI of 1.0). Some of these characters were autapomorphies (e.g., intron 7 is unique to *Apis* F1), while others (characters 1, 2, and 6) support basal nodes within the tree. Two characters were incongruent with each other (introns 4 and 5) and require different interpretations of intron insertion/deletion.

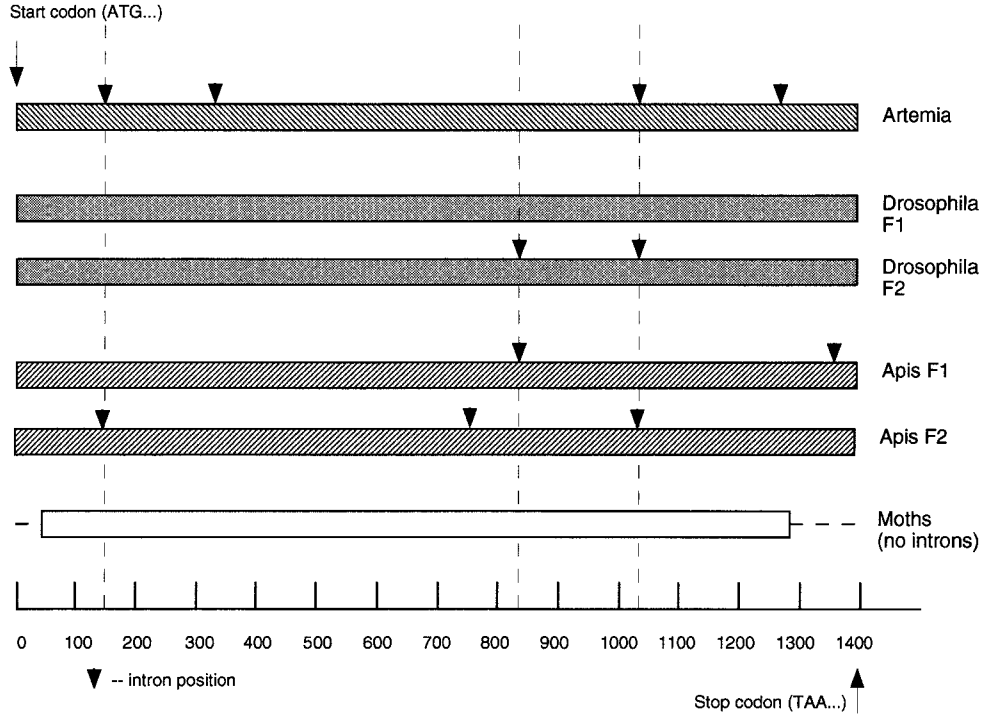

FIG. 3.—Comparisons of intron locations in the two *Drosophila* copies, the two *Apis* copies, the intronless moth sequences, and the *Artemia* sequence. The map shows only the coding region, starting with the start codon and ending with the stop codon. Intron locations are indicated by triangles. Introns that correspond exactly in position are indicated by vertical dashed lines.

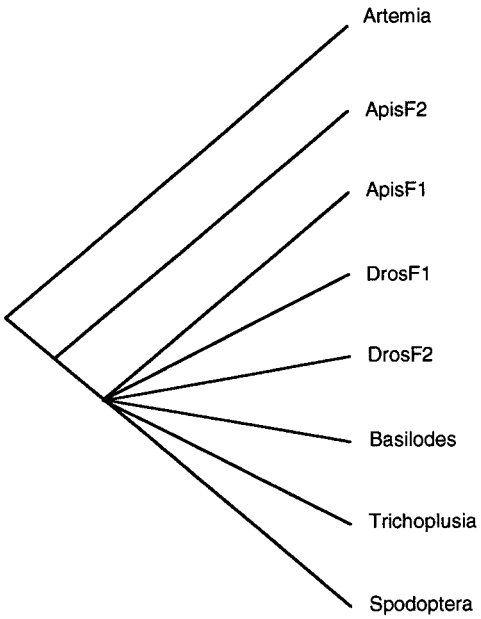

FIG. 4.—Parsimony analysis of intron positions. Data set is eight taxa by seven characters. Introns were coded as present/absent based on figure 3. This tree is the strict consensus of 29 equally parsimonious trees.

A strict consensus tree of the 29 equally parsimonious trees is shown in figure 4. Based on the consensus tree, the *Apis* F2 copy is basal relative to all other insect sequences. Monophyly of the holometabolan sequences is supported by the loss of introns 2 and 6, and *Apis* F2 is excluded from the clade including all other sequences from the Holometabola by virtue of its retention of intron 1 (which is lost in all other Holometabola).

Overall, the intron presence/absence data provide little insight into the historical relationships among genes except to suggest that the *Apis* F2 copy retains a number of primitive attributes.

Parsimony Analysis of Nucleotide and Amino Acid Sequences

To determine whether the presence of two copies in *Apis* (and other bees) represents a gene duplication separate from that in *Drosophila,* or whether the paralogous copies in these two taxa can be traced to a single ancestral gene duplication, we performed a series of parsimony analyses on the nucleotide and amino acid sequences of the coding regions. For all parsimony analyses, we used *Artemia, Periplaneta, Pedetontus,* and *Tomocerus* as outgroups. The ingroup (the Holometabola) was constrained to be monophyletic, both because there is substantial evidence of holometabolan monophyly (Boudreaux 1979; Kristensen 1991; Whiting et al. 1997) and because the gene duplication(s) appear to be restricted to two holometabolan orders: Diptera and Hymenoptera. To test for significant signal heterogeneity among the three nucleotide positions, we ran the partition homogeneity test (Huelsenbeck et al. 1996) as implemented in PAUP*4 and found that there is no significant signal heterogeneity among positions ($P = 0.10$ for 100 replicates). All tree statistics are listed in table 5.

Figure 5*a* shows the tree resulting from an initial unweighted analysis of 1,392 nucleotide positions (equal to the entire coding region), 435 of which were parsimony-informative. Branch lengths are indicated by numbers along each branch of the tree, and the bootstrap support values are shown in brackets for each node. In subsequent analyses, we downweighted third positions by 2:2:1, 5:5:1, and 10:10:1 (first : second : third positions). Figure 5*b–d* shows the tree topologies obtained under these weighting schemes. While altered weighting schemes produced slightly different tree topologies, the trees obtained are all significantly different from random trees obtained in permuted data sets (table 5). Tree topology appears to be stable to downweighting of third positions after 5:5:1 downweighting, because the tree topologies shown in figure 5*c* and *d* are the same.

While we obtained different tree topologies with different weighting schemes, there are nodes that appear repeatedly in many trees. The trees based on the unweighted analysis and 2:2:1 downweighting of third positions are congruent with the tree obtained in the analysis of intron positions in that *Apis* F2 appears basal with respect to the other holometabolan sequences. This result is strongly supported (bootstrap values of 87%–99%, depending on the weighting scheme). Both trees (fig. 5*a* and *b*) support the view that there may have been an ancestral gene duplication, with *Drosophila* F2 and *Apis* F1 copies being homologs.

Trees based on heavy downweighting of third positions (5:5:1 and 10:10:1) imply a different hypothesis of gene duplication. Both trees (fig. 5*c* and *d*) suggest that independent, parallel gene duplication events underly the existence of two copies in the flies and the Hymenoptera. Both trees show high levels of bootstrap

Table 5
Descriptive Tree Statistics for Parsimony Analyses

	CI[a]	Steps	No. of Trees	PTP[b] *P* Value	Tree Topology
Nucleotide data					
Unweighted	0.5207	1,607	2	<0.001 (131)	Figure 5*a*
2:2:1 weighting	0.5169	1,923	1	<0.001 (161)	Figure 5*b*
5:5:1 weighting	0.5184	2,834	1	<0.001 (218)	Figure 5*c*
10:10:1 weighting	0.5241	4,324	1	<0.001 (362)	Figure 5*d*
Amino acid data	0.5738	189	1	<0.001 (3)	Figure 6*a*

[a] Consistency index excluding uninformative characters.

[b] The PTP was implemented in PAUP*4 with 1,000 replicates. The numbers in parentheses indicate the difference (in steps) between the shortest tree obtained in the 1,000 permuted replicates and the observed shortest tree obtained in the original analysis of unpermuted data. The larger the value in parentheses, the more deviation there is between the observed data set and the ''best'' permuted data set.

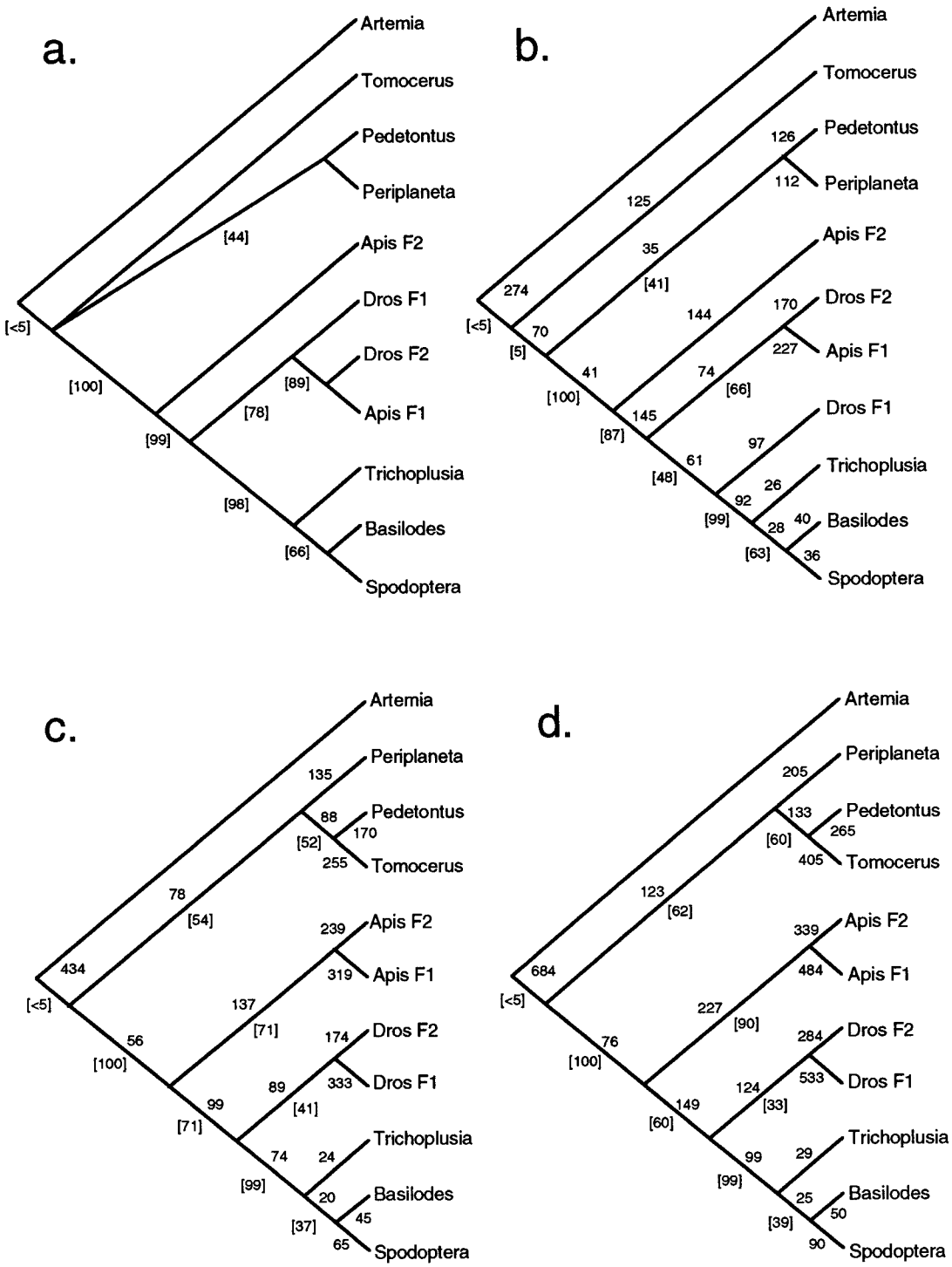

FIG. 5.—Parsimony analyses of nucleotide sequences for the 10 hexapods plus *Artemia* (1,392 total nucleotide positions, 435 parsimony informative). In all analyses, *Artemia, Tomocerus, Pedetontus,* and *Periplaneta* were selected as outgroups. Tree statistics are given in table 5. Numbers along branches are numbers of steps, and numbers in brackets are bootstrap values for 500 replicates. *a,* Based on unweighted analysis of nucleotide data. *b,* Based on 2:2:1 (first : second : third positions) weighting. *c,* Based on 5:5:1 weighting. *d,* Based on 10:10:1 weighting. The topologies shown in *c* and *d* are the same.

support for the monophyly of the *Apis* paralogous copies (71% and 90%, respectively).

Next, we translated the nucleotide sequences into amino acid sequences and analyzed the resulting data set. The amino acid data set consisted of 464 amino acid positions, 43 of which were parsimony-informative. In an analysis of the 11 sequences, we obtained a single tree topology that is congruent with the trees obtained with weighted nucleotide data (fig. 6*a*). In this tree topology, as in those obtained with weighted nucleotide data, the two *Apis* copies are sister taxa, and the two *Drosophila* copies are sister taxa.

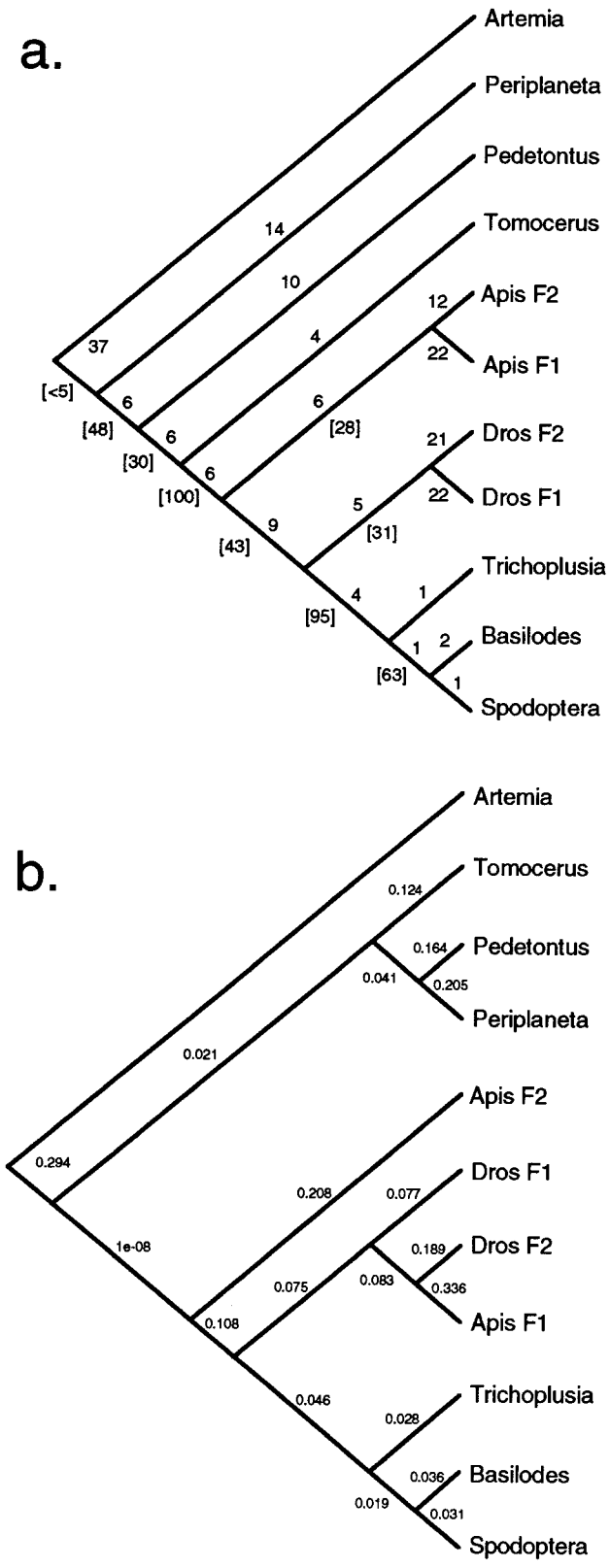

FIG. 6.—*a*, Parsimony analysis of amino acid data (464 total amino acid positions, 43 parsimony informative). Numbers as in figure 5. This tree topology is the same as those in fig. 5*c* and *d*. *b*, Maximum-likelihood analysis of nucleotide data. Nucleotide frequencies were determined empirically within the model, and we used both the Hasegawa, Kishino, and Yano (1985) model and the Felsenstein (1984) two-parameter model for unequal base frequencies (tree shown here). Both the transition : transversion ratio (ts/tv = 2) and the shape parameter of the gamma distribution (α = 0.240948) were determined empirically. Numbers along branches indicate branch lengths. Log likelihood = −8066.404.

Combining nucleotide or amino acid sequence data with the data matrix for intron position did not alter the tree topologies obtained when nucleotides and amino acids were analyzed alone (figs. 5*a–d*) but did lengthen the trees from 8 to 10 steps.

Maximum-Likelihood Analyses of Nucleotide and Amino Acid Sequences

We investigated the phylogenetic relationships among the insect nucleotide sequences by maximum likelihood. We obtained the same tree whether we used the Hasegawa, Kishino, and Yano (1985) model or the Felsenstein (1984) two-parameter model (fig. 6*b*). In both analyses, we obtained the same tree topology as in the analysis of unweighted nucleotide data (fig. 5*a*). We performed a maximum-likelihood analysis of the amino acid data using PUZZLE, version 4.0, and obtained a nearly unresolved tree, except for support for the monophyly of the moth sequences plus *Drosophila* F1. This tree is congruent with the topology shown in figure 5*b* (α = 0.12; log likelihood = −2,448.12). Using the Kishino and Hasegawa (1989) test, we determined that the tree topologies obtained in the unweighted analysis of nucleotides (fig. 5*a*) and in the maximum-likelihood analysis (fig. 6*b*) were not significantly different from those obtained in the weighted parsimony analyses (figs. 5*b–d*).

Assessment of Alternative Hypotheses of Gene Duplication

In order to assess whether the data significantly support the hypothesis of parallel gene duplication or a single common ancestral duplication, we applied the T-PTP test (Archie 1989*a*, 1989*b*; Faith 1990, 1991; Faith and Cranston 1991) with three alternative test phylogenies (table 6). The only test phylogeny for which there is significant support is the hypothesis of parallel gene duplication (fig. 5*c* [and *d*]). The two test phylogenies consistent with a hypothesis of ancestral gene duplication were not significantly different from the alternative topologies.

Discussion

While normally viewed as a single-copy gene in insects (Friedlander, Regier, and Mitter 1992, 1994), the presence of two copies of EF-1α in distantly related holometabolous orders (Boudreaux 1979, pp. 234–261; Kristensen 1991) raises the possibility that two copies are widespread in the Holometabola. This could result either from an ancient gene duplication that occurred before the divergence of flies and bees or from parallel gene duplications in the ancestors of these two groups.

The above phylogenetic analyses suggest that the gene duplication events occurred independently and in

Table 6
Statistical Evaluation of Alternative Tree Topologies Using the T-PTP Test

	Hypothesis 1	Hypothesis 2	Hypothesis 3	Tree Topology
Nucleotide data				
Unweighted	0.914 (−55)	1.00 (−83)	1.00 (−93)	Figure 5*a*
2:2:1 weighting	0.676 (−52)	0.998 (−94)	1.00 (−105)	Figure 5*b*
5:5:1 weighting	0.050 (−39)*	0.968 (−127)	0.996 (−149)	Figure 5*c*
10:10:1 weighting	0.086 (−84)	0.966 (−237)	0.994 (−284)	Figure 5*d*
Amino acid data	0.046 (−6)*	0.456 (−11)	0.526 (−12)	Figure 6*a*

NOTE.—The T-PTP test was implemented in PAUP*4 with 500 replicates. Three alternative unresolved trees were evaluated by the test. Hypothesis 1 is the hypothesis that the gene duplication occurred independently in bees and in flies: ((*Apis* F1 + *Apis* F2) (moths (*Dros* F1 + *Dros* F2))). Hypotheses 2 and 3 assume that the gene duplication only occurred once and predated the divergence of the Diptera and the Hymenoptera. Hypothesis 2: ((*Apis* F1 + *Dros* F2) (*Apis* F2 + *Dros* F1) + moths); Hypothesis 3: ((*Apis* F1 + *Dros* F1) (*Apis* F2 + *Dros* F2) + moths). The probabilities are those associated with rejecting phylogenetic hypotheses other than those consistent with the constraint tree. Values in parentheses represent the difference (in steps) between the shortest tree that is incompatible with the constraint tree and the shortest tree that is compatible with the constraint tree. Significant *P* values indicate that the trees that are incompatible with the constraint tree are significantly longer than the trees that are compatible with the constraint tree. * = $P < 0.05$.

parallel. While the unweighted nucleotide data suggest the possibility of a single ancestral gene duplication, analyses in which third positions are downweighted and analyses based on amino acid sequence stongly support the view that there were two independent duplications. Furthermore, analysis using the T-PTP test indicates that the only hypothesis that is significantly supported by the data under maximum parsimony is the hypothesis of parallel gene duplication. We consider the tree topologies shown in figures 5*c*, 5*d*, and 6*a* to be the most likely hypothesis of relationships given the current data set. Parallel gene duplication events appear to be more likely than an ancestral gene duplication.

The tree topology obtained with downweighting of third positions and that obtained with amino acids bear a close resemblance to current views on the evolutionary relationships among the three orders of holometabolous insects included in this study: the Hymenoptera, the Diptera, and the Lepidoptera (fig. 7). The Diptera and the Lepidoptera, along with several other orders of insects (including the Mecoptera, the Siphonaptera, the Trichoptera, and, probably, the Strepsiptera [Whiting et al. 1997]) fall into a well-supported monophyletic group that excludes the Hymenoptera (Panorpida *sensu* Boudreaux 1979; Mecopterida *sensu* Kristensen 1991). The grouping of moth and fly sequences in figures 5*c*, 5*d*, and 6*a* is consistent with this view of ordinal relationships. The Hymenoptera are usually considered to be either relatively basal within the Holometabola (Boudreaux 1979) or the sister group to the Panorpida (Kristensen 1991; Whiting et al. 1997). In either case, the hypothesis that the paralogous copies present in *Apis* and other Hymenoptera arose independently of the paralogous copies in the Diptera is consistent with current views of holometabolous ordinal-level relationships (fig. 7).

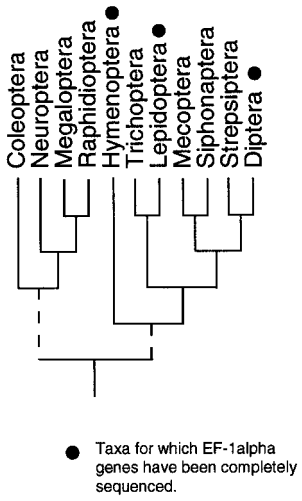

FIG. 7.—Cladogram showing presumed relationships among the holometabolous insect orders (Kristensen 1991; Whiting et al. 1997). While Hymenoptera is shown as the sister group to the Panorpida, there is uncertainty about the relative positions of the Neuropterida (Coleoptera, Neuroptera, Megaloptera, and Raphidioptera) and the Hymenoptera.

The apparent lack of two paralogous copies of EF-1α in moths is consistent with the hypothesis that EF-1α was duplicated in parallel in the Diptera and in the Hymenoptera. If the gene duplication had occurred prior to the divergence of flies and moths, we would expect to find two copies in both orders (fig. 7). That three studies involving EF-1α (Cho et al. 1995; Mitchell et al. 1997; Regier and Shultz 1997) have failed to find more than a single copy of the gene in other orders of hexapods suggests that the duplications may be restricted to the Diptera and the Hymenoptera.

It is possible that two paralogous copies exist in the Lepidoptera and other orders but have been overlooked by previous workers, either because they contain large and frequent introns or because they are present as a pseudogene. This possibility should be considered in future work on EF-1α in other orders. Efforts should be made to exclude the possibility of paralogous copies, and caution should be taken in interpreting results of phylogenetic analyses of highly divergent EF-1α sequences (>18% overall sequence divergence, the minimum divergence observed between paralogs in this study).

We believe that EF-1α may provide a model system for investigating gene duplication within insects as well as for investigating the historical patterns of intron/exon evolution within the context of a cladogram for the insect orders. We are now investigating copy number and intron/exon structure in other holometabolous orders, in-

cluding Neuroptera, Coleoptera, Mecoptera, Lepidoptera, and Trichoptera. More work will be needed to accurately reconstruct the history of intron/exon evolution in insect EF-1α genes.

Acknowledgments

We are grateful to Nan Liu for providing primers and for technical help in working with RNA extractions. Nick Calderone generously provided worker honeybees for RNA extractions. Kelley Tilmon, Mike McDonald, Mike Engel, Ted Schultz, and John Ascher provided helpful comments on earlier versions of this paper. Two anonymous reviewers provided invaluable advice on methods of analysis. This project was supported by a National Science Foundation Research Grant in Systematic Biology (DEB-9508647) to B.N.D.

GenBank accession numbers: *Artemia salina* (X03704, X03705, X03706, X03707, X03708); *Periplaneta americana* (U90054); *Pedetontus saltator* (U90056); *Tomocerus* sp. (U90059); *Apis mellifera* (X52884, X52885); *Drosophila melanogaster* F1 (X06869); *Drosophila melanogaster* F2 (X06870); *Basilodes chrysopis* (U20125); *Spodoptera frugiperda* (U20139); *Trichoplusia ni* (U20140).

LITERATURE CITED

ARCHIE, J. W. 1989*a*. A randomization test for phylogenetic information in systematic data. Syst. Zool. **38**:239–252.

———. 1989*b*. Phylogenies of plant families: a demonstration of phylogenetic randomness in DNA sequence data derived from proteins. Evolution **43**:1796–1800.

BELSHAW, R., and D. L. J. QUICKE. 1997. A molecular phylogeny of the Aphidiinae (Hymenoptera: Braconidae). Mol. Phylogenet. Evol. **7**:281–293.

BOUDREAUX, H. B. 1979. Arthropod phylogeny with special reference to insects. J. Wiley & Sons, New York.

BRANDS, J. H. G. M., J. A. MAASSEN, F. J. VAN-HEMERT, R. AMONS, and W. MÖLLER. 1986. The primary structure of the alpha subunit of human elongation factor 1: structural aspects of guanine nucleotide binding sites. Eur. J. Biochem. **155**:167–172.

BROWER, A. V. Z., and R. DESALLE. 1994. Practical and theoretical considerations for choice of a DNA sequence region in insect molecular systematics, with a short review of published studies using nuclear gene regions. Ann. Entomol. Soc. Am. **87**:702–716.

CHO, S., A. MITCHELL, J. C. REGIER, C. MITTER, R. W. POOLE, T. P. FRIEDLANDER, and S. ZHAO. 1995. A highly conserved nuclear gene for low-level phylogenetics: elongation factor 1-alpha recovers morphology-based tree for heliothine moths. Mol. Biol. Evol. **12**:650–656.

DAYHOFF, M. O., R. M. SCHWARTZ, and B. C. ORCUTT. 1978. A model of evolutionary change in proteins. Pp. 345–352 *in* M. O. DAYHOFF, ed. Atlas of protein sequence and structure. Vol. 5, Suppl. 3. National Biomedical Research Foundation, Silver Spring, Md.

FAITH, D. P. 1990. Chance marsupial relationships. Nature **345**: 393–394.

———. 1991. Cladistic permutation tests of monophyly and nonmonophyly. Syst. Zool. **40**:366–375.

FAITH, D. P., and P. S. CRANSTON. 1991. Could a cladogram this short have arisen by chance alone? On permutation tests for cladistic structure. Cladistics **7**:1–28.

FELSENSTEIN, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. **17**:368–376.

———. 1984. Distance methods for inferring phylogenies: a justification. Evolution **38**:16–24.

———. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution **39**:783–791.

———. 1993. PHYLIP (phylogenetic inference package). Version 3.5c. Distributed by the author, Department of Genetics, University of Washington, Seattle.

FRIEDLANDER, T. P., J. C. REGIER, and C. MITTER. 1992. Nuclear gene sequences for higher level phylogenetic analysis: 14 promising candidates. Syst. Biol. **41**:483–490.

———. 1994. Phylogenetic information content of five nuclear gene sequences in animals: initial assessment of character sets from concordance and divergence studies. Syst. Biol. **43**:511–525.

HASEGAWA, M., T. HASHIMOTO, J. ADACHI, N. IWABE, and T. MIYATA. 1993. Early branchings in the evolution of eukaryotes: ancient divergence of entamoeba that lacks mitochondria revealed by protein sequence data. J. Mol. Evol. **36**:380–388.

HASEGAWA, M., H. KISHINO, and T. YANO. 1985. Dating of the human-ape split by a molecular clock of mitochondrial DNA. J. Mol. Evol. **21**:160–174.

HOVEMANN, B., S. RICHTER, U. WALLDORF, and C. CZIEPLUCH. 1988. Two genes encode related cytoplasmic elongation factors 1alpha (EF-1alpha) in *Drosophila melanogaster* with continuous and stage specific expression. Nucleic Acids Res. **16**:3175–3194.

HUELSENBECK, J. P., J. J. BULL, and C. W. CUNNINGHAM. 1996. Combining data in phylogenetic analysis. Trends Ecol. Evol. **11**:152–158.

ISEGAWA, Y., J. SHENG, Y. SOKAWA, K. YAMANISHI, O. NAKOGOMI, and S. UEDA. 1992. Selective amplification of cDNA sequence from total RNA by cassette-ligation mediated polymerase chain reaction (PCR): application to sequencing 6.5 kb genome segment of hatavirus strain B-1. Mol. Cell. Probes **6**:467–475.

KAMAISHI, T., T. HASHIMOTO, Y. NAKAMURA, F. NAKAMURA, S. MURATA, N. OKADA, K. I. OKAMOTO, M. SHIMIZU, and M. HASEGAWA. 1996. Protein phylogeny of translation elongation factor EF-1α suggests microsporidians are extremely ancient eukaryotes. J. Mol. Evol. **42**:257–263.

KISHINO, H., and M. HASEGAWA. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order of the Hominoidea. J. Mol. Biol. Evol. **29**:170–179.

KRISTENSEN, N. P. 1991. Phylogeny of extant hexapods. Pp. 125–140 *in* I. D. NAUMAN, ed. Insects of Australia. Vol. 1. Cornell University Press, Ithaca, N.Y.

LENSTRA, J. A., A. VAN VLIET, A. C. ARNBERG, F. J. VAN HEMERT, and W. MÖLLER. 1986. Genes coding for the elongation factor EF-1α in *Artemia*. Eur. J. Biochem. **155**:475–484.

LOCKHART, P. J., M. A. STEEL, M. D. HENDY, and D. PENNY. 1994. Recovering evolutionary trees under a more realistic model of sequence evolution. Mol. Biol. Evol. **11**:605–612.

MADDISON, W. P., and D. R. MADDISON. 1992. MacClade version 3: analysis of phylogeny and character evolution. Sinauer, Sunderland, Mass.

MARONI, G. 1993. An atlas of *Drosophila* genes. Oxford University Press, Oxford.

MITCHELL, A., S. CHO, J. C. REGIER, C. MITTER, R. W. POOLE, and M. MATHEWS. 1997. Phylogenetic utility of elongation factor-1α in Noctuoidea (Insecta: Lepidoptera): the limits of synonymous substitution. Mol. Biol. Evol. **14**:381–390.

RAO, T. R., and L. I. SLOBIN. 1986. Structure of the amino-terminal end of mammalian elongation factor Tu. Nucleic Acids Res. **14**:2409.

REGIER, J. C., and J. W. SHULTZ. 1997. Molecular phylogeny of the major arthropod groups indicates polyphyly of the crustaceans and a new hypothesis for the origin of hexapods. Mol. Biol. Evol. **14**:902–913.

ROTH, W. W., P. W. BRAGG, M. V. CORRIAS, N. S. REDDY, J. N. DHOLAKIA, and J. A. WAHBA. 1987. Expression of a gene for mouse eukaryotic elongation factor TU during murine erythroleukemic cell differentiation. Mol. Cell. Biol. **7**: 3929–3936.

STRIMMER, K., and A. VON HAESELER. 1996. Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies. Mol. Biol. Evol. **13**:964–969.

SWOFFORD, D. L. 1993. PAUP: phylogenetic analysis using parsimony. Version 3.1.1. Illinois Natural History Survey, Champaign.

SWOFFORD, D. L., G. J. OLSEN, P. J. WADDELL, and D. M. HILLIS. 1996. Pp. 407–514 *in* D. M. HILLIS, C. MORITZ, and B. K. MAPLE, eds. Molecular systematics. 2nd edition. Sinauer, Sunderland, Mass.

WALLDORF, U., and B. T. HOVEMANN. 1990. *Apis mellifera* cytoplasmic elongation factor 1-alpha (EF-1alpha) is closely related to *Drosophila melanogaster* EF-1alpha. FEBS Lett. **267**:245–249.

WHITING, M. F., J. M. CARPENTER, Q. D. WHEELER, and W. C. WHEELER. 1997. The Strepsiptera problem: phylogeny of the holometabolous insect orders inferred from 18S and 28S ribosomal DNA sequences and morphology. Syst. Biol. **46**: 1–68.

ROSS H. CROZIER, reviewing editor

Accepted December 2, 1997