Molecular Phylogenetics and Evolution
Vol. 20, No. 1, July, pp. 136–148, 2001
doi:10.1006/mpev.2001.0956, available online at http://www.idealibrary.com on IDEAL®

Elongation Factor-2: A Useful Gene for Arthropod Phylogenetics

Jerome C. Regier*,[1] and Jeffrey W. Shultz†

*Center for Agricultural Biotechnology, University of Maryland Biotechnology Institute, Plant Sciences Building, College Park, Maryland 20742; and †Department of Entomology, University of Maryland, Plant Sciences Building, College Park, Maryland 20742

Received September 26, 2000; revised January 24, 2001; published online June 6, 2001

Robust resolution of controversial higher-level groupings within Arthropoda requires additional sources of characters. Toward this end, elongation factor-2 sequences (1899 nucleotides) were generated from 17 arthropod taxa (5 chelicerates, 6 crustaceans, 3 hexapods, 3 myriapods) plus an onychophoran and a tardigrade as outgroups. Likelihood and parsimony analyses of nucleotide and amino acid data sets consistently recovered Myriapoda and major chelicerate groups with high bootstrap support. Crustacea + Hexapoda (= Pancrustacea) was recovered with moderate support, whereas the conflicting group Myriapoda + Hexapoda (= Atelocerata) was never recovered and bootstrap values were always <5%. With additional nonarthropod sequences included, one indel supports monophyly of Tardigrada, Onychophora, and Arthropoda relative to molluscan, annelidan, and mammalian outgroups. New and previously published sequences from RNA polymerase II (1038 nucleotides) and elongation factor-1α (1092 nucleotides) were analyzed for the same taxa. A comparison of bootstrap values from the three genes analyzed separately revealed widely varying values for some clades, although there was never strong support for conflicting groups. In combined analyses, there was strong bootstrap support for the generally accepted clades Arachnida, Arthropoda, Euchelicerata, Hexapoda, and Pycnogonida, and for Chelicerata, Myriapoda, and Pancrustacea, whose monophyly is more controversial. Recovery of some additional groups was fairly robust to method of analysis but bootstrap values were not high; these included Pancrustacea + Chelicerata, Hexapoda + Cephalocarida + Remipedia, Cephalocarida + Remipedia, and Malaocostraca + Cirripedia. Atelocerata (= Myriapoda + Hexapoda) was never recovered. Elongation factor-2 is now the second protein-encoding, nuclear gene (in addition to RNA polymerase II) to support Pancrustacea over Atelocerata. Atelocerata is widely cited in morphology-based analyses, and the discrepancy between results derived from molecular and morphological data deserves greater attention.

Key Words: Arthropoda; elongation factor-1α; elongation factor-2; molecular systematics; Pancrustacea; RNA polymerase II.

INTRODUCTION

The conceptual framework for understanding organismal diversity of arthropods will remain incomplete and controversial as long as robustly supported phylogenetic relationships are lacking. This is illustrated by the current debate on the phylogenetic placement of hexapods. The morphology-based Atelocerata hypothesis maintains that hexapods share a common terrestrial ancestor with myriapods, but the molecule-based Pancrustaea hypothesis maintains that hexapods share a common aquatic ancestor with crustaceans. These alternative hypotheses are sometimes portrayed as being strongly supported by two different kinds of data, but a more nuanced interpretation may be necessary. In particular, recent parsimony-based studies of morphological characters recover Atelocerata (Wheeler, 1998; Edgecombe *et al.,* 2000), but node support for this clade is very low (decay index = 1; BP = 68% in Edgecombe *et al.,* 2000). Similarly, ribosomal sequences usually do not recover Pancrustacea when taxon sampling is high (Giribet and Ribera, 2000; Spears and Abele, 1998; Wheeler, 1998; but see Eernisse, 1998), although the overall set of relationships appears closer to Pancrustacea than to Atelocerata. One study based on combined 18S and 28S rDNA (Friedrich and Tautz, 1995) reconstructed Pancrustacea with high bootstrap support but included only two crustaceans and specifically excluded a "long-branch" hexapod (*Drosophila melanogaster*). Further, relevant phylogenetic signal was contributed primarily by 28S rDNA and not by 18S rDNA (Regier and Shultz, 1997). The nuclear genes encoding ubiquitin (Wheeler *et al.,* 1993), histone H3 (Colgan *et al.,* 1998), snRNA U3 (Colgan *et al.,* 1998), and elongation factor-1α (Regier and Shultz, 1998) recovered neither Pancrustacea nor Atelocerata. However, recent studies of Pol II[2] and Pol II + EF-1α sampled a wide range of

[1] To whom correspondence should be addressed. E-mail: regier@glue.umd.edu.

[2] Abbreviations used: EF-1α, elongation factor-1α; EF-2, elongation factor-2; GTR, general time-reversible; nt, nucleotide; nt1, first

1055-7903/01 $35.00

arthropods and recovered Pancrustacea with strong nodal support (up to 100% BP) (Shultz and Regier, 2000). Further support for Pancrustacea has come from studies of mitochondrial gene order, in which a single leucyl–tRNA rearrangement was proposed as a synapomorphy for Pancrustacea (Boore *et al.,* 1998). In addition to the Pancrustacea/Atelocerata controversy, there are other contested higher-level arthropod groupings, e.g., the monophyly of Chelicerata (e.g., Shultz and Regier, 2000; Dunlop and Selden, 1998), Myriapoda (e.g., Regier and Shultz, 2000; Kraus, 1998), Crustacea (e.g., Edgecombe *et al.,* 2000; Giribet and Wheeler, 1999), and Mandibulata (e.g., Edgecombe *et al.,* 2000; Giribet and Ribera, 2000; Shultz and Regier, 2000).

If these controversies are to be resolved, then additional evidence is needed. Toward this goal, the present study examines higher-level arthropod relationships in light of newly generated sequences encoding elongation factor-2. Like Pol II and EF-1α sequences analyzed previously (Regier and Shultz, 1997, 1998; Shultz and Regier, 2000), EF-2 has a highly conserved protein sequence whose evolutionary changes provide signal across deep phylogenetic splits (Friedlander *et al.,* 1994). Additionally, all three genes have now been sequenced for the same ingroup and outgroup taxa, enabling a direct comparison of individual gene utility and a combined analysis with 4029 nucleotide characters per taxon (see also Baker and DeSalle, 1997; Mitchell *et al.,* 2000; Wiegmann *et al.,* 2000). Of particular note is that EF-2 by itself provides strong support for Myriapoda, modest support for Pancrustacea, weak but consistent support for Chelicerata, and low to no support for Crustacea. In combined analyses, Myriapoda, Pancrustacea, and Chelicerata are strongly supported, but support for Crustacea remains very low. This study illustrates the power of analyzing multiple genes, separate from generating larger data sets from more taxa.

MATERIALS AND METHODS

Specimen Preservation, Taxon Sampling, and the Data Set

Specimens either were alive until frozen at −85°C or were stored in 100% ethanol at ambient temperature for up to 1 year prior to final storage at −85°C. Seventeen arthropod and 2 nonarthropod taxa (Giribet *et al.,* 1996; Eernisse, 1998; Nielsen, 1998) were sampled for EF-2 (1899 nt each), EF-1α (1092 nt each), and Pol II (1038 nt each). All 19 EF-2 and 2 Pol II sequences are new to this study. Species names, higher classification, and GenBank Accession Nos. (EF-2; EF-1α; Pol II) are as follows: *Tomocerus* sp. (Hexapoda: Collembola. AF240830; U90059; AF139011, AF139012), *Eumesocampa frigilis* (Hexapoda: Diplura. AF240818; AF137388; AF138978, AF138979, AF138980), *Machiloides banksi* (Hexapoda: Microcoryphia. AF240822; AF137390; AF138990, AF138991, AF138992), *Artemia salina* (Crustacea: Branchiopoda. AF240815; X03349; U10331), *Hutchinsoniella macracantha* (Crustacea: Cephalocarida. AF240820; AF063411; AF138984, AF138985, AF138986), *Semibalanus balanoides* (Crustacea: Cirripedia. AF240817; AF063404; AF138971, AF138972), *Armadillidium vulgare* (Crustacea: Malacostraca. AF240816; U90046; AF138970), "ostracod" (Crustacea: Maxillopoda. AF240825; AF063414; AF138997, AF138998, AF138999), *Speleonectes tulumensis* (Crustacea: Remipedia. AF240829; AF063416; AF139008, AF139009, AF139010), *Mastigoproctus giganteus* (Chelicerata: Arachnida: Thelyphonida. AF240823; U90052; U90038), *Nipponopsalis abei* (Chelicerata: Arachnida: Opiliones. AF240824; AF137391; AF138993, AF138994, AF138995), *Limulus polyphemus* (Chelicerata: Xiphosura. AF240821; U90051; U90037), *Endeis laevis* (Chelicerata: Pycnogonida. AF240819; AF063409; AF138981, AF240882, AF240883), *Tanystylum orbiculare* (Chelicerata: Pycnogonida. AF240831; AF063417; AF139013, AF139014), *Scolopendra polymorpha* (Myriapoda: Chilopoda. AF240828; AF137393; AF139006, AF139007), *Polyxenus fasciculatus* (Myriapoda: Diplopoda. AF240826; U90055; AF139001, AF139002), *Scutigerella* sp. (Myriapoda: Symphyla. AF240827; AF137392; AF139003, AF139004, AF139005), *Peripatus* sp. (Onychophora. AF240835; AF137395; AF139017, AF240892), and *Milnesium tardigradeum* (Tardigrada. AF240883; AF063419; AF139016, AF240887, AF240888). Five multiply sampled arthropod groups—Arthropoda, Hexapoda, Euchelicerata, Pycnogonida, and Arachnida—were designated "test clades" based on their wide acceptance among morphological and molecular systematists. Recovery of test clades was one criterion used to assess a gene's utility.

Procedures for RT-PCR amplification, nested PCR reamplification, and DNA sequencing have been described, along with primers for EF-1α and Pol II (see Shultz and Regier, 2000). Primer sequences (5′ → 3′) for amplification of the EF-2 cDNA are herein listed (20*F*: ATG GTN AAY TTY ACN GTI GA [20]; 95F: GCN CAY GTN GAY CAY GGI AA [95]; 436*R*: TCN GTY TGN ACR CAN ACI CC [412]; 455*F*: GGN GTN TGY GTN CAR ACI GA [431]; 691*R*: GTR AAN GCC CAN CCR TG [664]; 707*F*: CAY GGN TGG GCN TTY AC [680]; 1216*R*: TAC ATC ATN ARI GGN CC [1132]; 1232*F*: GGN CCN YTI ATG ATG TA [1148]; 1390*R*:

codon position; nt1noLR, nt1 data subset in which any characters that encode a leucine or arginine residue for any taxon are excluded at that homologous position for all taxa; nt1LR, nt1 data subset which includes any characters that encode a leucine or arginine residue for any taxon plus all other homologous characters for other taxa; nt2, second codon position; nt3, third codon position; PCR, polymerase chain reaction; Pol II, RNA polymerase II (largest subunit); RT-PCR, reverse transcription/polymerase chain reaction.

CCC ATC ATN ARI ATN GT [1306]; 1640*R*: CAY TGN ACC ATN GGR TC [1549]; 1655*F*: GAY CCN ATG GTN CAR TG [1565]; 1672*R*: CCN GCD ATD ATR TGY TCI CC [1582]; 2080*F*: CAR TAY YGI AAY GAR ATI AAR GA [1988]; 2089*R*: GCC CAY TGR AAN CCI GCN AC [1996]; 2108*F*: GTN GCN GGI TTY CAR TGG GC [2015]; and 2293*R*: TCN GGR CAY TGD ATY TC [2200]). For the primer names, *F* identifies a forward primer and *R* a reverse primer. For the primer sequences, R = A and G; Y = C and T; D = A, G, and T; N = A, C, G, and T; I = inosine. The numbers in brackets identify the 3′ nucleotide for the homologous position in the published EF-2 cDNA sequence from *Drosophila melanogaster* (GenBank Accession No. X15805) relative to the first coding nucleotide. The data set for phylogenetic analysis includes the region between 95*F* and 2089*R* minus the 5′-most nucleotide, and there are no missing data. In brief, the EF-2 cDNA was initially amplified by RT-PCR with primer pairs that yielded three overlapping fragments, namely, 20*F* or 95*F*/1390*R*,707*F*/1672*R*, and 1655*F*/2089*R* or 2293*R*. These fragments were then reamplified by PCR with internal primers and sequenced directly. Faint bands were reamplified from M13 sites present in all primers (not shown in the above primer sequences). Other primers listed were used for nested PCR reamplifications. Occasionally, taxon-specific primers were synthesized to amplify across gaps in sequence. DNA sequencing reactions were fractionated and analyzed on Applied Biosystems automated DNA sequencers.

The PREGAP and GAP4 programs within the Staden package (Staden *et al.,* 1999) were used to edit and assemble contigs. The Genetic Data Environment software package (version 2.2, Smith *et al.,* 1994) was used to manually align assembled sequences and to construct nucleotide data matrices for phylogenetic analysis. No indels were required to align the EF-1α and Pol II sequences across all 19 taxa. For EF-2, there were only three regions of short indels across the 19 taxa, a 12-nt insertion unique to *Speleonectes* (positioned between nt 193 and 194 in *Artemia salina,* GenBank Accession No. AF240815), a 3-nt insertion unique to *Milnesium* (positioned between nt 499 and 500 in *Artemia salina*), and, relative to the outgroup taxa *Milnesium* and *Peripatus,* a 3-nt deletion in all arthropods except *Endeis* (no indel), *Scutigerella* (no indel), *Polyxenus* (3-nt insertion), and *Scolopendra* (3-nt insertion) (3-nt insertions positioned between nt 712 and 713 in *Artemia salina*). These indels, accounting for 0.16–0.63% of the total EF-2 sequence, were removed from the data set for phylogenetic analysis of sequence data.

EF-2 sequences from the polychaete annelid *Nereis virens* (GenBank Accession No. AF240834) and the polyplacophoran mollusk *Chaetopleura apiculata* (Accession No. AF240832) were also generated for this study but, within the context of our broad but minimal taxon sampling scheme, were too divergent to yield reliable rooting (see also Shultz and Regier, 2000). However, these sequences plus the 19 analyzed in this report were manually aligned in Genetic Data Environment with 5 more EF-2 sequences already deposited in GenBank (the green alga *Chlorella kessleria* (Accession No. M680645), the nematode *Caenorhabditis elegans* (Accession No. M86959), three mammals (the hamster *Cricetulus griseus:* Accession No. M13708; *Homo sapiens:* Accession No. X51466; and the rat *Rattus norvegicus:* Accession No. Y07504), and an insect (*Drosophila melanogaster:* Accession No. X15805)). Across this expanded alignment, eight regions of indels were identified. Their positions are nt 288–301 (region I), nt 591–604 (region II), nt 726–727 (region III), nt 807–817 (region IV), nt 879–880 (region V), nt 921–922 (region VI), nt 1425–1426 (region VII), and nt 1662–1663 (region VIII), all relative to the first coding nucleotide in *Drosophila melanogaster.*

We also aligned the 19 EF-1α sequences analyzed in this study with *Drosophila melanogaster* (Hexapoda, GenBank Accession No. X06869) and 13 additional nonarthropod sequences. These additional taxa are *Euperipatoides rowelli* (Onychophora, Accession No. AF137394), *Onchocerca volvulus* (Nematoda, Accession No. M64333), *Chaetopleura apiculata* (Mollusca, Accession No. U90062), *Mytilis edulis* (Mollusca, Accession No. AF063420), *Acmaea testudinalis* (Mollusca, Accession No. U90061), *Nereis virens* (Annelida, Accession No. U90064), *Enchytraeus* sp. (Annelida, Accession No. AF063409), *Tubifex tubifex* (Annelida, Accession No. AF063422), *Hirudo medicinalis* (Annelida, Accession No. U90063), *Phascolopsis gouldii* (Sipuncula, Accession No. AF063421), *Homo sapiens* (Mammalia, Accession No. X03558), *Rattus norvegicus* (Mammalia, Accession No. X63561), and *Cricetulus longicaudatus* (Mammalia, Accession No. D00522). Across this expanded alignment, one indel region was identified (positioned between nt 660 and 673 relative to first coding nucleotide in *Drosophila melanogaster*).

For Pol II, a similarly broad sampling of taxa revealed no indels outside of two in *Caenorhabditis elegans.*

Sequence data sets consisted of nucleotide or amino acid characters, the latter conceptually translated with MacClade software (version 3.08; Maddison and Maddison, 1992) from nucleotide sequences. EF-2, EF-1α, and Pol II sequences were analyzed separately and in combination. Additionally, genes were partitioned by codon position (nt1/nt2/nt3). For some analyses, nt1 was further partitioned into nt1noLR/nt1LR to identify a nt1 data set, namely, nt1noLR, for which synonymous change was unlikely.

Data Analysis

Maximum-parsimony (MP) analyses of nucleotide and amino acid data sets were conducted with PAUP*4.0b2(PPC) (Swofford, 1998) and unordered character transformations. Most analyses assumed

equally weighted character transformations, both with and without nt3. However, stepmatrices were incorporated into "6-parameter-parsimony-analysis-with-ln-weighting" of nucleotide data (Cunningham, 1997a) and into "Protpars" analysis of amino acid data (implemented in MacClade, see Maddison and Maddison, 1992). For Protpars analysis, serine codons that differed at nt1 were coded separately. Analysis consisted of a heuristic search using TBR branch swapping with random sequence addition (100 sequence-addition replicates). Bootstrap analysis (1000 bootstrap replications) was identical except for 10 sequence-addition replicates per bootstrap replication. Decay indices/Bremer support values (Bremer, 1994) were obtained from MP analysis with the "load constraints" and "enforce topological constraints" options after a constraint tree was written. To test for conflict across data sets, the incongruence length difference test (Farris *et al.,* 1995) was performed, with the partition homogeneity test in PAUP*4.0 with 1000 random bipartitions, and each was analyzed by TBR branch swapping on 10 random sequence-addition replicates. The test of Kishino and Hasegawa (1989) was implemented with the "Tree Scores" option in PAUP*4.0 to test for the significance of differences in fit of the amino acid data set to its MP topology and to the MP topology in which Crustacea were constrained to be monophyletic.

The minimum and maximum number of character changes that map to each node, their degree of homoplasy, and whether or not they were the sole changes for that character (i.e., binary characters) were determined with PAUP*4.0b2. The frequency of particular character changes within the amino acid data set when mapped across the MP tree was calculated in MacClade.

Neighbor-joining analysis (Saitou and Nei, 1987) of total nucleotide data was performed with PAUP*4.0b2(PPC) with Logdet correction (Lockhart *et al.,* 1994) and under the assumption that 25% of all sites were invariable.

Maximum-likelihood (ML) analysis of amino acid data was performed with the *protml* program within the MOLPHY software package (version 2.2; Adachi and Hasegawa, 1994) and the empirical transition matrix compiled by Jones *et al.* (1992). All parsimony trees which were within 1% of the MP tree length (i.e., within 24 steps of tree length 2400, total number of parsimony trees examined = 64,887) were read into *protml,* and their likelihood scores were calculated. The most likely parsimony tree (length = 2411) had a log-likelihood score of −17612.91. By comparison, the MP tree (length = 2400) had a log-likelihood score of −17643.31.

Maximum-likelihood analyses of nucleotide data sets, with and without nt1LR and nt3, were performed with PAUP*4.0b2 under a general time-reversible (GTR) model of nucleotide sequence evolution (Rodriguez *et al.,* 1990). The GTR model was selected over four others based on a likelihood ratio test (Huelsenbeck and Rannala, 1997; see Shultz and Regier, 2000, for details of implementation; see Swofford *et al.,* 1996, for a general discussion of models). The GTR model also proved superior when four different estimates of among-site rate heterogeneity were incorporated. The first estimate was to assign distinct rate categories to designated character partitions; this is called the *GTR + ssr* ("site-specific rates") approach. The second estimate was to fit total character change to a gamma distribution by optimization for the shape parameter α and with the assumption of four distinct rate categories; this is called the $GTR + \Gamma$ approach. The third estimate was to assume rate homogeneity for all sites except those estimated to be invariant; this is called the $GTR + I$ approach. The fourth estimate combined the last two into the $GTR + \Gamma + I$ approach.

Three data sets (all nt, nt1 + nt2, nt1noLRnt2) were partitioned in various ways and optimized by ML to a common topology (Fig. 1B). Across all data sets, the highest log-likelihood scores result when character change is fitted to a gamma distribution, i.e., 11.66, 9.42, and 9.42% increases for the all-nt, nt1 + nt2, and nt1noLRnt2 data sets, respectively, relative to no partitioning. Further modest but significant increases in log-likelihood scores can result from the additional estimation of invariant sites, i.e., 0.34 and 0.17% for the all-nt and nt1noLRnt2 data sets, respectively, but no increase for the nt1 + nt2 data set. Because fitting a gamma distribution is computationally intensive, we have also tested the effect of preassigning rate categories, which is less computationally intensive. For the all-nt data set, the largest increase (9.65%) in log-likelihood score results from the assigning of a separate rate category to nt3. A further increase (1.17–2.76%) occurs by the partitioning of nt1 into nt1LR and nt1noLR, but some of this increase is nullified if nt1LR is combined with nt3. Assigning nt1noLR and nt2 to separate categories has no significant effect on the log-likelihood score and is not preferred because an extra parameter is required. Separating characters by gene leads to an additional increase in log-likelihood score, although the overall magnitude is relatively small, i.e., 0.04–0.20%. Estimating invariant sites alone by ML for the all-nt data set is slightly more effective than partitioning by codon position and much more effective for data sets that are missing nt3.

As a first step in the ML search, likelihood parameters were optimized with MP trees derived from amino acids. NNI branch swapping was then performed and new likelihood parameters were optimized on the most likely topology. TBR branch swapping was conducted on the new tree and likelihood parameters were reoptimized. These parameters were then used as input for a heuristic search with NNI branch swapping and 100 random sequence-addition replicates, followed by parameter reoptimization. These parameters were used in bootstrap analyses

(typically, 500 replications for ML, GTR+ssr model), each based on a heuristic search with NNI branch swapping and 10 random sequence-addition replicates.

The parametric bootstrap was implemented as a means of testing for long-branch attraction between *Armadillidium* and *Semibalanus*. To do this, 100 simulated nucleotide data sets (length = 1343 characters each) were constructed with Seq-Gen (version 1.1; Rambaut and Grassly, 1997) under a ML, GTR+Γ model. To select ML parameters (i.e., branch lengths, base frequencies, α), the original nt1noLR + nt2 data set from all three genes combined (= 1343 nt) was optimized on the shortest-length, MP, amino-acid topology that did not group Malacostraca and Cirripedia (length = 8 steps beyond MP tree length of 1200). MP analyses with equally weighted character transformations were then executed (parameters: 100 random sequence additions, TBR branch swapping), and the number of times that Malacostraca + Cirripedia was recovered was expressed as a percentage of the total.

Three ML parameters were separately estimated for the nt1noLR + nt2 and nt3 data sets from EF-2, from EF-1α, and from Pol II. The first parameter was relative rate, which was estimated by preassignment of six rate categories (i.e., nt1noLR + nt2 and nt3 for each of the three genes) and the fitting of total character change to the ML topology obtained from analysis of the nt1 + nt2 data set (GTR + ssr model with rate categories preassigned as nt1noLR/nt1LR/nt2 for each gene). The second parameter was α, the shape parameter, which was calculated with a GTR + Γ+I model and fitted to the previously mentioned ML topology. The third parameter was *I*, the percentage of invariant sites, which was calculated with a GTR model in which all noninvariant sites were assigned a single rate and fitted to the previously mentioned ML topology. The fraction of observed constant sites was calculated directly from the data.

Percentage differences of all pairwise combinations of EF-2, EF-1α, and Pol II amino acid and nt3 data sets were calculated in PAUP*4.0b2. Average differences were plotted relative to individual nodes of the MP tree obtained from analysis of amino acids under equally weighted character transformations. Differences were calculated by the averaging of all values across the basal dichotomy within a particular clade. Base frequencies and a χ^2 test for their homogeneity were calculated by gene and by data set with PAUP*4.0b2.

RESULTS

Phylogenetic Analysis and Assessment of Node Support—EF-2 and Its Comparison with EF-1α and Pol II

MP analysis of EF-2 amino acids under equal weighting recovered Arachnida, Euchelicerata, Myriapoda, and Pycnogonida with strong node support (i.e., BP >75%) and Pancrustacea with moderate support (BP = 70%) (Fig. 1A, Table 1). No other groups had BP values above 50%, although the "test clades" Arthropoda and Hexapoda were also recovered. Similar results were obtained with other analytical approaches, including ML (Table 1). At least for MP, inclusion of nt3 lowered node support and number of test clades recovered (see below). Similar analyses of EF-1α and Pol II have been presented previously (Shultz and Regier, 2000; Table 1).

For several taxonomic groups, levels of BP support vary dramatically across the three genes, although in no case do different genes strongly support conflicting groups. For example, Myriapoda is strongly supported by EF-2, moderately supported by EF-1α, but poorly supported by Pol II. Pancrustacea is strongly supported by Pol II, moderately by EF-2, but poorly by EF-1α. Hexapoda is most strongly supported by EF-1α, weakly by EF-2, and poorly by Pol II. Arthropoda is moderately to strongly supported by EF-1α and Pol II but weakly by EF-2. In general, it appears that no one gene uniformly has high BP values for taxonomic groups strongly supported by another gene. Only Pycnogonida and Euchelicerata receive moderate to strong support from all three genes.

Phylogenetic Analysis and Assessment of Node Support—Combined Data

The partition homogeneity test quantifies conflict between data sets, although there is not a generally accepted definition for a significant result (Cunningham, 1997b). Cunningham has argued for combining data when *P* values are $\geq$0.01. By this criterion, EF-2 can be combined with EF-1α and with Pol II for all data types (amino acids, all nt, nt1 + nt2, nt1noLR + nt2). In contrast, *P* values for combined EF-1α and Pol II data sets are less than 0.01 (but greater than the minimum possible value of 0.001) for all data sets except nt1noLR + nt2, which is 0.010. This "conflict" does not reside in any one set of taxonomic relationships. For example, constraint of chelicerates, hexapods, myriapods, and arthropods to be monophyletic raises the score to 0.026; constraint of Pancrustacea to be monophyletic raises the score to 0.013; and removal of the outgroup taxa or constraint of *Artemia* + ostracod to be monophyletic yields marginally significant scores of 0.007 and 0.006, respectively. We have chosen to analyze combined data in the absence of strong evidence to the contrary.

Amino acid data and nucleotide data from EF-2 + EF-1α + Pol II have been analyzed by MP (see Fig. 1B) and ML, and, when practical, bootstrap percentages have been calculated (Table 1). Analyses of nt1noLR + nt2, nt1 + nt2, and amino acids recover all five test clades, almost always with BP >80%. The one exception is MP analysis of amino acids with a Protpars step matrix, which fails to recover Hexapoda, the most

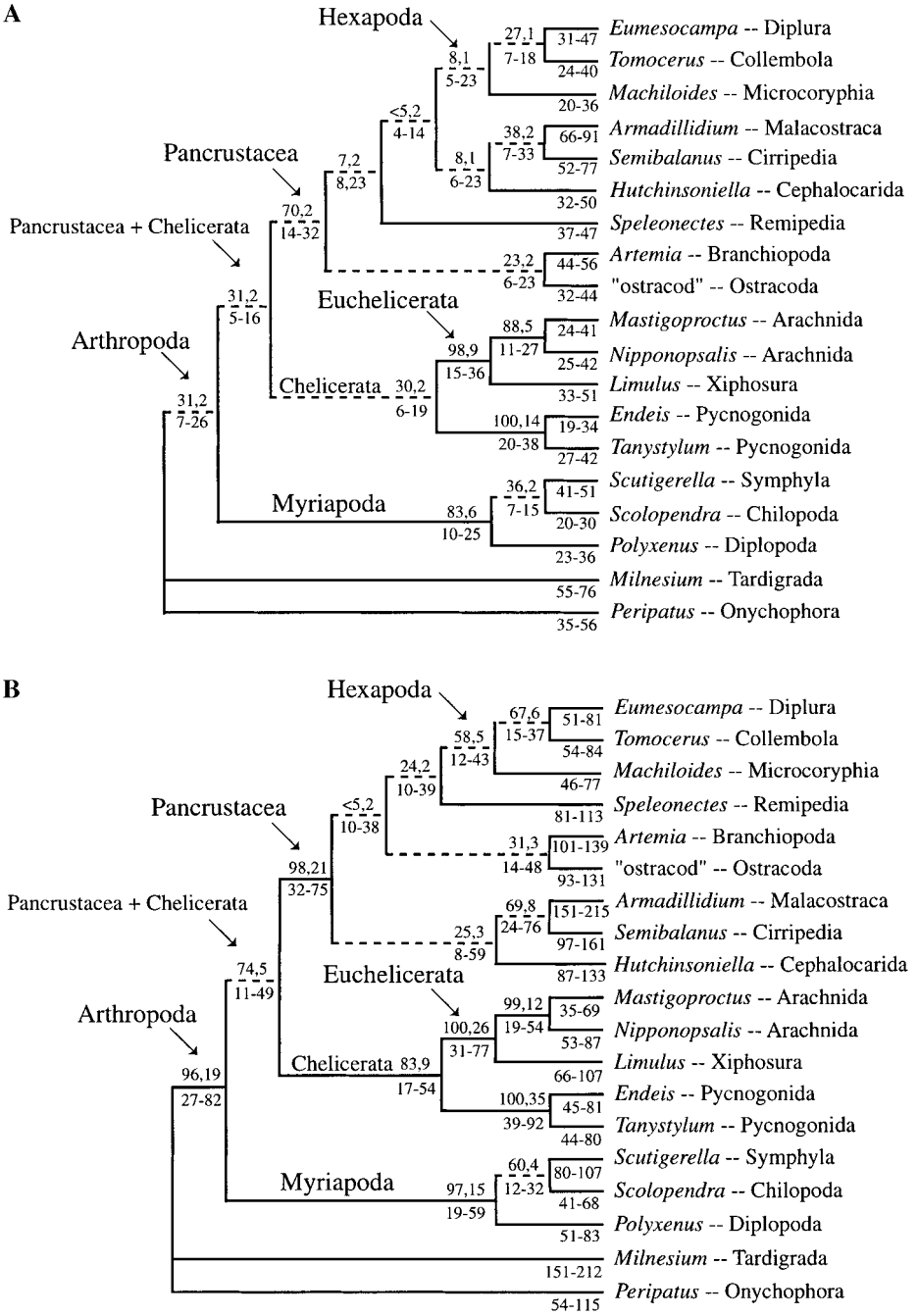

FIG. 1. Maximum-parsimony topologies of arthropod taxa. Above branches, BP values followed by decay indices. Below branches, minimum and maximum numbers of character changes that map to that particular node. Dashed lines represent groups recovered with <75% BP support; recovery of these groups may be sensitive to analytical method and the particular data set (see Table 1). To the right of the tree are listed the taxa (generic names only) followed by their higher-order classification. Traditionally defined Crustacea (not labeled) consists of Pancrustacea minus Hexapoda. (A) Parsimony analysis based on analysis of EF-2 amino acid sequences. All transformations were weighted equally. Number of MP trees = 1; tree length = 1080; Consistency Index = 0.5750; Retention Index = 0.4039; number of parsimony-informative characters = 182. (B) Parsimony analysis based on combined analysis of EF-2 + Pol II + EF-1α amino acid sequences. All transformations were weighted equally. Number of MP trees = 1; tree length = 2400; Consistency Index = 0.5875; Retention Index = 0.4072; number of parsimony-informative characters = 404.

weakly supported test clade with all methods and data sets. Inclusion of nt3 in the data set causes BP values to decrease for Arthropoda, Arachnida, and Hexapoda but not for Pycnogonida and Euchelicerata. ML analysis of total nucleotides recovers four test clades; MP analysis recovers two; and neighbor-joining with a Log-

TABLE 1

Bootstrap Support Values for Arthropod Groups Based on Combined and Separate Analyses of EF-2, EF-1α, and Pol II[a]

	EF-2 + EF-1α + Pol II										
	ML					MP					
Dataset (No. partitions)	GTR(ssr) nt1*2 (6)	GTR(ssr) nt12 (9)	GTR(ssr) nt123 (9)	GTR(Γ + I) nt1*2	Protml aa	=wt nt1*2	6-ppm nt1*2	=wt nt12	=wt nt123	=wt aa	Protpars aa
Arthropoda (√)	87*	87*	<5	19/23*	*	91*	86*	80*	<5	96*	93*
Pancrustacea + Chelicerata	59*	68*	<5	17/23*	*	35*	44	67*	<5	74*	68*
Pancrustacea + Myriapoda (= Mandibulata)	13	6	<5	0/23		17	17*	<5	<5	<5	8
Chelicerata + Myriapoda	<5	5	<5	0/23		16	9	6	<5	10	7
Chelicerata + Crustacea (= Schizoramia)	<5	<5	<5	0/23		<5	<5	<5	<5	<5	<5
Pancrustacea (Crustacea + Hexapoda)	95*	97*	67*	20/23*	*	94*	91*	85*	<5	98*	96*
Atelocerata (Myriapoda + Hexapoda)	<5	<5	<5	0/23		<5	<5	<5	<5	<5	<5
Crustacea	<5	<5	<5	3/23		<5	<5	<5	<5	<5	<5
Myriapoda	99*	94*	<5	23/23*	*	97*	86*	80*	<5	97*	92*
Chelicerata	75*	86*	39*	21/23*	*	62*	72*	83*	<5	83*	80*
Hexapoda (√)	66*	87*	71*	18/23*	*	58*	72*	82*	46	58*	53
Hexapoda + Cephalocarida + Remipedia	41*	36*	7	11/23	*	39*	45*	20	<5	6	15
Hexapoda + Remipedia	13	14	<5	5/23		15	8	13	<5	24*	18
Cephalocarida + Remipedia	58*	57*	34	18/23*	*	52*	75*	33	<5	14	28
Pancrustacea − Ostracoda	33*	39*	<5	4/23*	*	31*	17	34	<5	12	37*
Malacostraca + Cirripedia + Branchiopoda	51*	33*	<5	15/23*	*	55*	60*	11	<5	12	33
Malacostraca + Cirripedia	73*	61*	<5	16/23*	*	81*	76*	54*	<5	69*	75*
Pycnogonida (√)	100*	100*	100*	23/23*	*	100*	100*	100*	100*	100*	100*
Euchelicerata (Xiphosura + Arachnida) (√)	100*	100*	100*	23/23*	*	100*	100*	100*	92*	100*	100*
Arachnida (√)	95*	97*	61*	20/23*	*	97*	98*	92*	36	99*	96*

	EF-2						
	ML		MP				
Dataset (No. partitions):	GTR(ssr) nt12 (3)	GTR(Γ + I) nt1*2	=wt nt1*2	=wt nt12	=wt nt123	=wt aa	Protpars aa
Arthropoda (√)	<5		<5	6	<5	31*	5
Pancrustacea + Chelicerata	<5		<5	12	<5	31*	17
Pancrustacea + Myriapoda (=Mandibulata)	<5		<5	<5	<5	<5	<5
Chelicerata + Myriapoda	<5		<5	<5	<5	<5	<5
Chelicerata + Crustacea (=Schizoramia)	<5		<5	<5	<5	<5	<5
Pancrustacea (Crustacea + Hexapoda)	65*	*	48	58	<5	70*	73*
Atelocerata (Myriapoda + Hexapoda)	<5		<5	<5	<5	<5	<5
Crustacea	<5		<5	<5	<5	<5	<5
Myriapoda	88*	*	94*	65*	<5	83*	78
Chelicerata	<5		<5	7	5	30*	24
Hexapoda (√)	36*		22	27*	23	8*	24
Hexapoda + Cephalocarida + Remipedia	23*		17	23*	<5	11	26
Hexapoda + Remipedia	<5		<5	<5	<5	8	7
Cephalocarida + Remipedia	60*	*	42	53*	8	21	39
Pancrustacea − Ostracoda	<5		11	15	<5	8	22
Malacostraca + Cirripedia + Branchiopoda	6		17	8	<5	<5	20
Malacostraca + Cirripedia	16		34	21	<5	38*	46*
Pycnogonida (√)	100*	*	100*	100*	100*	100*	100*
Euchelicerata (Xiphosura + Arachnida) (√)	100*	*	93*	99*	68*	98*	98*
Arachnida (√)	81*	*	83*	77*	20	88*	88*

TABLE 1—*Continued*

	EF-1α							Pol II						
	ML		MP					ML		MP				
Dataset (No. partitions):	GTR(ssr) nt12 (3)	GTR(Γ+I) nt1*2	=wt nt1*2	=wt nt12	=wt nt123	=wt aa	Protpars aa	GTR(ssr) nt12 (3)	GTR(Γ+I) nt1*2	=wt nt1*2	=wt nt12	=wt nt123	=wt aa	Protpars aa
Arthropoda (√)	69*	*	79*	63*	19	63*	76*	77*	*	88*	63*	<5	78*	81*
Pancrustacea + Chelicerata	8	*	<5	<5	<5	7	<5	9		<5	7	<5	17*	17
Pancrustacea + Myriapoda (= Mandibulata)	<5		<5	<5	<5	<5	<5	<5		<5	<5	<5	<5	<5
Chelicerata + Myriapoda	<5		<5	<5	<5	<5	6	<5		7	11	<5	7	12
Chelicerata + Crustacea (=Schizoramia)	<5		<5	<5	<5	<5	<5	<5		<5	<5	<5	<5	<5
Pancrustacea (Crustacea + Hexapoda)	6		<5	<5	<5	<5	<5	88*	*	88*	75*	<5	87*	91*
Atelocerata (Myriapoda + Hexapoda)	18		17	14	<5	11	16	<5		<5	<5	<5	<5	<5
Crustacea	<5		<5	<5	<5	<5	<5	<5		<5	<5	<5	<5	<5
Myriapoda	62*	*	76*	64*	<5	66*	74*	<5		<5	5	<5	9	5
Chelicerata	21		12	20	<5	17	12	76*		72*	44*	11	22	66*
Hexapoda (√)	76*	*	60*	81*	23	46	69*	16*		<5	18	<5	23	<5
Hexapoda + Cephalocarida + Remipedia	<5		<5	<5	<5	<5	<5	<5		<5	<5	<5	<5	<5
Hexapoda + Remipedia	<5		9	<5	<5	<5	<5	6*		<5	<5	<5	<5	<5
Cephalocarida + Remipedia	19		16	16	8	<5	9	8		8	<5	<5	<5	<5
Pancrustacea − Ostracoda	<5		<5	<5	<5	<5	<5	<5		<5	<5	<5	9	6
Malacostraca + Cirripedia + Branchiopoda	26	*	20	12	<5	<5	10	<5		8	<5	<5	10	<5
Malacostraca + Cirripedia	61*	*	62*	65*	<5	33	50*	10		10	<5	<5	7	8
Pycnogonida (√)	100*	*	100*	100*	100*	99*	100*	100*	*	100*	99*	63*	99*	100*
Euchelicerata (Xiphosura + Arachnida) (√)	84*	*	79*	77*	68*	94*	91*	65*	*	71*	36*	30*	73*	73*
Arachnida (√)	93*	*	87*	90*	20	89*	83*	33*		59*	8	19	39*	69*

[a] GTR(ssr), GTR model with preassigned site-specific rates; GTR (Γ + I), GTR model with character change fitted to a gamma distribution (1 parameter) with invariable sites separately estimated; =wt, character state transformations are equally weighted; 6-ppm, 6-parameter parsimony; nt1*2 (6), nt1noLR/nt2 by gene (2 partitions/gene or 6 total); nt12 (3), nt1noLR/nt1LR/nt2 for EF-2, EF-1α, or Pol II; nt12 (9), nt1noLR/nt1LR/nt2 by gene (3 partitions/gene or 9 total); √, test clade; *, identifies clades present in ML topology or in the MP strict consensus topology. For the combined nt1*2 data set, only 23 bootstrap replications were performed with the ML, GTR (Γ + I) model, and the BP results are expressed as a fraction of the total. For the combined aa data set, BP analysis was not performed with the Protml model. For individual-gene, nt1*2 data sets, BP analyses were not performed with the GTR(Γ+I) model.

Det distance measure recovers three (Table 1, unpublished observations).

Five other groups—Pancrustacea, Myriapoda, Chelicerata, Pancrustacea + Chelicerata, and Malacostraca + Cirripedia—are recovered with nucleotide and amino acid data sets and under ML and MP conditions (Table 1). Pancrustacea receives strong BP support—from 85 to 98% depending on the data set and analytical method, as long as nt3 characters are excluded. However, even with nt3, Pancrustacea is still recovered by ML (BP, 67%). Myriapoda is also strongly supported with BP values up to 99%, although recovery is sensitive to inclusion of nt3. Chelicerata receives substantial BP support (up to 86%) and is recovered by ML even when nt3 is included. BP support for Pancrustacea + Chelicerata and Malacostraca + Cirripedia is lower with maximal values of 74 and 81%, respectively. Group recovery is sensitive to inclusion of nt3.

The terminal branches leading from the common malacostracan/cirripedian node are the two longest arthropod branches on the tree (e.g., see Fig. 1), raising the possibility of long-branch attraction. To test this, we used the parametric bootstrap (Huelsenbeck and Hillis, 1996), generating and analyzing simulated data

sets based on parameters optimized to a topology in which Malacostraca and Cirripedia are not together (see Materials and Methods). MP analysis of approximately 18% of the simulated data sets placed Malacostraca + Cirripedia as a group, arguing that long-branch attraction cannot be excluded as an explanation for their grouping. Whether Malacostraca and Cirripedia are also clustered because of phylogenetic signal is not tested by the parametric bootstrap. More broadly, all crustacean (i.e., Pancrustacea minus Hexapoda) terminal taxa have long branches with relatively low BP values uniting their generally unstable groupings (see Table 1). In this interesting but problematical region of the tree, Cephalocarida + Remipedia (up to 75% BP) and Pancrustacea minus Ostracoda (up to 39% BP) are often recovered. Interestingly, Crustacea is never recovered, although the test of Kishino and Hasegawa (1989) demonstrates that the amino acid data set does not significantly discriminate between the MP result (Fig. 1) and an otherwise identical MP analysis in which crustaceans are constrained to be monophyletic.

Molecular Evolution of EF-2, EF-1α, and Pol II

Each of the three genes displayed unequal base frequencies in all character partition categories (nt1, nt1noLR, nt1LR, nt2, nt3), although nt3 displayed the least bias overall. Despite this, nt3 and nt1LR characters were strongly nonhomogeneous ($P < 0.001$ except $P = 0.021$ for nt1LR from EF-1α) across the 19 taxa, unlike nt2 and nt1noLR characters. Nonhomogeneity suggests lineage-specific shifts in base frequency, and we note that these shifts correlate with characters that undergo synonymous change.

For each node on an MP topology (Fig. 1B), we have calculated average pairwise differences across the basal sister clades for nt3 nucleotides and amino acids from each gene (data not shown), to contrast the frequency of synonymous and nonsynonymous substitutions. Across all nodes on the tree and for all genes, values at nt3 are greater than or equal to 50% (exception: the two pycnogonid species for EF-2 and EF-1α), with a maximum observed value of 73% between *Armadillidium* and *Semibalanus* for Pol II. These high values approach the theoretical maximum of 75% under a Jukes and Cantor (1969) model of substitution, an overestimate for our data set given the observed bias in base frequency, and indicate that multiple, overlapping substitutions have been frequent and are likely to be highly homoplasious. In support of this, pairwise values at nt3 do not consistently increase with taxonomic depth, a feature that would generally be expected for a data set with a strong phylogenetic signal.

Pairwise differences at amino acids tend to increase with taxonomic depth (exception: *Armadilidium* + *Semibalanus* + *Artemia* group) from the lowest values of 8% to the highest of 28% between outgroup taxa. Generally, values for EF-2 and EF-1α tend to be similar and somewhat lower than those for Pol II. These values are well removed from the theoretical maximum of 95% under the assumption that all amino acid changes are equally likely, and they support the hypothesis that amino acid changes (and nonsynonymous changes generally) still contain recoverable phylogenetic signal.

Gene-specific rates of nonsynonymous and synonymous substitution have been estimated by ML from the nt1noLR + nt2 and the nt3 character partitions, respectively. Nonsynonymous substitutions occur at similar rates in EF-2 and EF-1α, whereas those in Pol II occur about 20% faster. Rates of synonymous substitutions are almost identical across genes and are approximately 11- to 12-fold faster than nonsynonymous substitutions, although this may be a substantial underestimate given the considerable homoplasy in the nt3 data sets (see Regier *et al.,* 1998). Similar studies demonstrate that the nt1LR and nt1 characters evolve at average rates that are intermediate to nt1noLR + nt2 and nt3 characters, consistent with their admixture of characters undergoing nonsynonymous and synonymous changes.

Gene-specific parameters that estimate among-site rate heterogeneity have also been estimated by ML. The nt3 data sets have few observed constant sites or ML-estimated invariant sites (i.e., $\leq$3%), consistent with a preponderance of rapid, synonymous change. In contrast, the nt1noLR + nt2 data sets, which encode only nonsynonymous changes, contain between 60% (for Pol II) and 68% (for EF-2 and EF-1α) ML-estimated invariant sites, consistent with their highly conserved protein sequences. The observed percentage of constant sites is only 1–2% higher. α is an indicator of among-site rate heterogeneity across a set of characters, and it differs across genes. For the nt1noLR + nt2 data set, α values were 0.59, 0.34, and 0.31 for EF-2, EF-1α, and Pol II, respectively. For the nt3 data set, α values were 0.50, 1.09, and 0.53, respectively. However, given uncertainty over the accuracy of partitioning very slowly evolving from truly invariant characters (Sullivan *et al.,* 1999), the safest result is simply to emphasize that patterns of among-site rate heterogeneity are gene specific.

Analysis of Indels as Possible Phylogenetic Characters

Eight regions of indels (identified as I–VIII in Materials and Methods) were revealed when EF-2 sequences were aligned across 25 taxa. Three (indels VI–VIII) are unique to the alga; one (indel III) is unique to mammals; one (indel V) is unique to the mollusk. The evolutionary histories of indel regions I, II, and IV all involve multiple insertion/deletion events, and their complete decipherment is uncertain. For indel region I, the data are most parsimoniously

	EF-2 INDEL REGION I	EF-2 INDEL REGION IV
Arthropoda: Hexapoda		
Drosophila	----------KECK--	--DNK---
Eumesocampa	----------KDCK--	--DNK---
Tomocerus	----------KDSN--	--DNK---
Machiloides	----------KDVK--	--DNK---
Arthropoda: Crustacea		
Speleonectes	------DIDGKLEK--	--DNK---
Artemia	----------KETK--	--DNK---
ostracod	----------KAVK--	--DNQ---
Armadillidium	----------SNED--	--DNE---
Semibalanus	----------KDND--	--DNV---
Hutchinsoniella	----------GDGR--	--GYK---
Arthropoda: Chelicerata		
Mastigoproctus	----------KGID--	--GYK---
Nipponopsalis	----------KGVN--	--GFK---
Limulus	----------KGEK--	--GYK---
Endeis	----------KETN--	-aEYK---
Tanystylum	----------KDSN--	--GFK---
Arthropoda: Myriapoda		
Scutigerella	----------KNSK--	-qDYK---
Scolopendra	----------KETK--	TGEFK---
Polyxenus	----------KTQK--	GPDFK---
Tardigrada:		
Milnesium	----------AGHR--	-aDHV---
Onychophora:		
Peripatus	----------IDNK--	-eDYK---
Nematoda:		
Caenorhabditis	TVEVDG----KKEKyn	--ESK---
Mollusca:		
Chaetopleura	--------------tn	--DAQ---
Annelida:		
Nereis	--------------lt	--GRV---
Mammalia:		
Homo	----------------	-pEGKKLP
Rattus	----------------	-pDGKKLP
Cricetulus	----------------	-pDGKKLP
Chlorophyta:		
Chlorella	----------------	-adtck--

FIG. 2. Alignment of conceptually translated amino acid sequences across two indel regions in EF-2. For indel region I, the decision not to overlap the first four residues of *Speleonectes* with the first five residues of *Caenorhabditis* is based on their differing lengths (indicating multiple insertion/deletion events) and on the strong evidence for the monophyly of Arthropoda and Pancrustacea. Lower case letters in the alignments indicate less certain homology assignments.

interpreted as supporting a clade of Arthropoda + Tardigrada + Onychophora, with the position of Nematoda uncertain (Fig. 2). Homology assignments across indel region II are uncertain because there are multiple, 1-nt-long insertions within a short stretch. Indel region IV has a shared 6-nt insertion for *Polyxenus* (Diplopoda) and *Scolopendra* (Chilopoda) that is distinct from *Scutigerella* (Symphyla), from nonmyriapod arthropods, and from nonarthropods (Fig. 2). Independent of the basal arthropod clade (either Chelicerata, Myriapoda, or Pancrustacea), the most parsimonious mapping within the sampled Myriapoda is to group Chilopoda and Diplopoda with Symphyla as its sister group.

For EF-1α, there is only one indel region across Arthropoda, Tardigrada, Onychophora, Mollusca, Annelida, and Sipuncula, but length variation in this region within Mollusca and Annelida is such as to prohibit an unambiguous interpretation of interphylum relationships.

DISCUSSION

Comparing the Molecular Evolution of EF-2, EF-1α, and Pol II

EF-2, EF-1α, and Pol II are low- to single-copy number, protein-encoding, nuclear genes with conservative rates of nonsynonymous nucleotide and amino acid change. These features make them reasonable candidates as deep-taxonomic-level, phylogenetic markers, a presumption that has already been demonstrated within arthropods for EF-1α and Pol II (Regier and Shultz, 1997, 1998; Shultz and Regier, 2000). In this study, the utility of EF-2 has been similarly indicated through recovery of well-supported test clades (Fig. 1A, Table 1) and through concordance with EF-1α and Pol II results (Regier and Shultz, 1997; Shultz and Regier, 2000). By at least one method, EF-2 recovers all test clades (Table 1), although BP support is modest to strong only for Pycnogonida, Euchelicerata, and Arachnida. EF-1α and Pol II also recover all test clades with modest to strong BP support. Even when individual genes and analytical methods do not recover specific test clades, conflicting taxonomic groups are only weakly supported (Table 1 and unpublished observations). Exceptions occur when nt3 is included in the data set, particularly when analyzed by parsimony with equally weighted transformations. Under such conditions, character state changes at nt3 account for more than 72% of total change when fitted to the tree in Fig. 1B and are highly homoplasious (Retention Index = 0.1973 for nt3 versus 0.3987 for nt2), with pairwise differences for nt3 approaching theoretically maximum levels even near the tips of the tree. Furthermore, the base composition at nt3, but not at nt1 or nt2, is nonhomogeneous. We argue that removal of nt3 from our data set is justified until more powerful analytical methods are developed. Of course, this statement will need to be reassessed as taxon sampling increases (see Källersjö *et al.,* 1999). A similar argument can be made for excluding nt1LR. The feature shared by nt3 and nt1LR, but by neither nt1noLR nor nt2, is the ability to undergo synonymous change, and it seems highly likely that it is this feature that makes phylogenetic analysis at deep taxonomic levels so unreliable with these character sets. In contrast, the utility of synonymous change at shallower taxonomic levels has been illustrated for EF-1α (Cho *et al.,* 1995).

The partition homogeneity test provides additional evidence that the signals from each gene, while not identical, do not strongly conflict. This is clearest in pairwise comparisons of EF-2 and Pol II and of EF-2 and EF-1α. P values for comparisons of EF-1α and Pol II are marginal (e.g., $P = 0.010$ for the nt1noLR + nt2 data set), but Cunningham (1997b) has argued for combination in such cases.

The observation that BP support levels for some test

(and nontest) groups vary widely by gene illustrates the benefit of analyzing multiple genes in combination (Baker and DeSalle, 1997; Mitchell *et al.,* 2000). Given the magnitude of variation in some cases, this benefit must be in addition to any which may result simply from increased data set size, which is approximately doubled with the addition of EF-2 to EF-1α and Pol II (from 2130 nt to 4029 nt per taxon). A reasonable hypothesis would be that genes with similar overall rates of substitution, such as EF-2, EF-1α, and Pol II, can still differ in their degree of among-site rate heterogeneity. If correct, evolutionary models that incorporate distinct features of character subsets (beyond the already highly effective partitioning of synonymous and nonsynonymous change) could further improve the predictive ability of phylogenetic analysis (e.g., Goldman *et al.,* 1998).

Resolved and Unresolved Issues in Arthropod Phylogeny

Recent, strictly molecular analyses support to differing degrees the grouping of Crustacea and Hexapoda (= Pancrustacea) to the exclusion of Myriapoda, arguing against Atelocerata (= Myriapoda + Hexapoda), a grouping that is currently undergoing reevaluation by morphologists in light of the molecular results (see Zrzavý *et al.,* 1998). Our previous studies showed that Pol II provided strong support for Pancrustacea and that EF-1α did not strongly refute this hypothesis (Shultz and Regier, 2000). In the current studies, EF-1α grouped all pancrustaceans except the remipede (data set: nt1noLR + nt2; analytical method: ML, GTR, Γ + I model). Also, constraining EF-1α amino acids alone to the combined-data, MP topology (Fig. 1) adds only 1.2% to the overall tree length (versus 2.6% for Pol II and 0.4% for EF-2). Analyzed in combination and for most of the same arthropod taxa (minus outgroup taxa) as in the current study, EF-1α + Pol II yielded BP support values for Pancrustacea of 91–100%, depending on the data set and analytical method. However, given the biological significance of such a large and nontraditional taxonomic grouping and the fact that our evidence resided essentially with Pol II alone, additional evidence in the form of sequence data from EF-2 was sought. By itself, EF-2 recovers Pancrustacea when nucleotides are analyzed by ML or when amino acids are analyzed by MP, with BP support up to 73% (Table 1). In combination, EF-2 + EF-1α + Pol II consistently and strongly recover Pancrustacea by all methods for which nt3 is excluded (Table 1). Pancrustacea is recovered by ML analysis of total nucleotides, albeit with reduced BP support. Finally, Pancrustacea is supported by nonhomplasious state changes at five characters, one of which is unique across the entire data set (a leucine/cysteine transformation) and two of which are binary. Given these results, we now consider Pancrustacea to be a well-supported group. Whether there is "strong" conflict between molecular and most morphological studies is currently difficult to judge given their different approaches to identifying homology.

Other than the seven test clades and Pancrustacea, three groups received BP support greater than 70% in at least one analysis (Table 1). First, Malacostraca + Cirripedia was recovered with BP support up to 81%, although a parametric bootstrap analysis could not eliminate long-branch attraction as a possible explanation. Traditionally, Cirripedia has been placed with Ostracoda and several other groups not sampled to form the class Maxillopoda, which is often considered to be paraphyletic (see Spears and Abele, 1998), a result consistent with our results. Second, Pancrustacea + Chelicerata was recovered with BP support up to 74%. This is noteworthy because chelicerates are sometimes thought to be basally divergent among extant arthropods, in part because their presumed sistergroup, Mandibulata (i.e., (Crustacea + (Myriapoda + Hexapoda)), is thought to be well supported (Kukalová-Peck, 1998). We consider our data and that of others to be inconclusive, particularly without greater sampling of nonarthropods. However, it is striking that our molecular data provide much less support for alternative hypotheses (e.g., Mandibulata, Chelicerata + Myriapoda, Schizoramia) (Table 1). Third, Cephalocarida + Remipedia was recovered with BP support up to 75%. The grouping of Malacostaca, Cirripedia, and Branchiopoda (BP, up to 60%) and of Cephalocarida, Remipedia, and Hexapoda (BP, up to 45%) are additional hypotheses modestly supported by molecular data.

Solving Difficult Problems in Higher-Level Systematics

Recently, there has been much discussion as to how to improve phylogenetic results, usually within the framework of deciding whether it is better to sample more taxa or to increase the data set (e.g., Graybeal, 1998). Our results suggest that sampling more types of data (by sampling multiple genes) may have added benefit relative to simply getting more data of the same type. This conclusion follows from the observation that BP values for particular groups can vary widely across genes and in ways that do not simply reflect the total number of characters (Table 1). This strategy of sampling multiple genes may be particularly useful when the goal is to resolve a large number of deep-taxonomic-level nodes.

We have provided evidence that, whereas gene combination generally improves node support relative to single-gene analyses, a strict adherence to "total evidence" is not necessarily the best approach. For example, we have argued that the removal of nt3 (and, to a lesser extent, nt1LR) appears justified, based on several explicit criteria, most importantly, recovery of test clades. As another example, the few unambiguous, nonhomoplasious character state changes identified in

the total data set provide confirmatory support for specific clades, some of which receive only modest BP support (Table 1).

Other aspects of our overall analytical approach may also require justification. For example, we compared results with different optimization criteria, different models of nucleotide substitution, and different character subsets because all strategies are imperfect when presented with real data. Although we have used methods that distinguish among some options, it remains difficult to select a "preferred" method. Additionally, and more significantly, discrepancies across methods and data sets require explanation. The observation that all of our strongly supported and some of the moderately supported groups are recovered across a wide range of analytical approaches implies that, for these groups, the data set is not highly homoplasious.

Last, the search for higher-order characters within molecular data is derived from the reasonable presumption that homoplasy should be less frequent for such characters. In this regard, studies of mitochondrial gene order have identified characters that unite groups across Metazoa, including Pancrustacea (Boore *et al.,* 1998; but see Masta, 2000 and Le *et al.,* 2000). Indels are another higher-order character. Although quite rare across all three genes in this study, the evolution of informative indels can still be difficult to interpret because they have undergone multiple insertion/deletion events (Fig. 2). However, indel region I in EF-2 most parsimoniously unites Arthropoda with Onychophora and Tardigrada, consistent with their inclusion in Panarthropoda (Giribet *et al.,* 1996; Eernisse, 1998; Nielsen, 1998). Within Myriapoda, indel region IV groups Chilopoda and Diplopoda to the exclusion of Symphyla, consistent with a previous molecular study (Regier and Shultz, 2000). Further sampling of EF-2 within Myriapoda will be needed to test the robustness of this observation.

ACKNOWLEDGMENTS

We thank Ms. Zaile Du for expert technical assistance. This work was supported by the National Science Foundation (DEB-9629791, DEB-9981970), the Maryland Agricultural Experiment Station, and the Center for Agricultural Biotechnology.

REFERENCES

Adachi, J., and Hasegawa, M. (1994). Programs for molecular phylogenetics. Version 2.2. Institute of Statistical Mathematics, Tokyo.

Baker, R. H., and DeSalle, R. (1997). Multiple sources of character information and the phylogeny of Hawaiian drosophilids. *Syst. Biol.* **46**: 654–673.

Boore, J. L., Lavrov, D. V., and Brown, W. M. (1998). Gene translocation links insects and crustaceans. *Nature* **392**: 667–668.

Bremer, K. (1994). Branch support and tree stability. *Cladistics* **10**: 295–304.

Cho, S., Mitchell, A., Regier, J. C., Mitter, C., Poole, R. W., Friedlander, T. P., and Zhao, S. (1995). A highly conserved nuclear gene for low-level phylogenetics: Elongation factor-1α recovers morphology-based tree for heliothine moths. *Mol. Biol. Evol.* **12**: 650–656.

Colgan, D. J., McLauchlan, A., Wilson, G. D. F., Livingston, S. P., Edgecombe, G. D., Macaranas, J., and Gray, M. R. (1998). Histone H3 and U2 snRNA DNA sequences and arthropod molecular evolution. *Aust. J. Zool.* **46**: 419–437.

Cunningham, C. (1997a). Is congruence between data partitions a reliable predictor of phylogenetic accuracy? Empirically testing an iterative procedure for choosing among phylogenetic methods. *Syst. Biol.* **46**: 464–478.

Cunningham, C. (1997b). Can three incongruence tests predict when data should be combined? *Mol. Biol. Evol.* **14**: 733–740.

Dunlop, J. A., and Selden, P. A. (1998). The early history and phylogeny of the chelicerates. *In* "Arthropod Relationships" (R. A. Fortey and R. H. Thomas, Eds.), pp. 221–235. Chapman & Hall, London.

Edgecombe, G. D., Wilson, G. D. F., Colgan, D. J., Gray, M. R., and Cassis, G. (2000). Arthropod cladistics: Combined analysis of histone H3 and U2 snRNA sequences and morphology. *Cladistics* **16**: 155–203.

Eernisse, D. J. (1998). Arthropod and annelid relationships re-examined. *In* "Arthropod Relationships" (R. A. Fortey and R. H. Thomas, Eds.), pp. 43–56, Chapman & Hall, London.

Farris, J. S., Källersjö, M., Kluge, A. G., and Bult, C. (1995). Constructing a significance test for incongruence. *Syst. Biol.* **44**: 570–572.

Friedlander, T. P., Regier, J. C., and Mitter, C. (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.

Friedrich, M., and Tautz, D. (1995). Ribosomal DNA phylogeny of the major extant arthropod classes and the evolution of myriapods. *Nature* **376**: 165–167.

Giribet, G., Carranza, S., Baguña, Ruitort, M., and Ribera, C. (1996). First molecular evidence for the existence of a Tardigrada + Arthropoda clade. *Mol. Biol. Evol.* **13**: 76–84.

Giribet, G., and Ribera, C. (2000). A review of arthropod phylogeny: New data based on ribosomal DNA sequences and direct character optimization. *Cladistics* **16**: 204–231.

Giribet, G., and Wheeler, W. (1999). The position of arthropods in the animal kingdom: A search of a reliable outgroup for internal arthropod phylogeny. *Mol. Phylogenet. Evol.* **13**: 132–143.

Goldman, N., Thorne, J. L., and Jones, D. T. (1998). Assessing the impact of secondary structure and solvent accessibility on protein structure. *Genetics* **149**: 445–458.

Graybeal, A. (1998). Is it better to add taxa or characters to a difficult phylogenetic problem? *Syst. Biol.* **47**: 9–17.

Huelsenbeck, J. P., and Hillis, D. M. (1996). Parametric bootstrapping in molecular phylogenetics: Applications and performance. *In* "Molecular Zoology: Advances, Strategies, and Protocols" (J. D. Ferraris and S. R. Palumbi, Eds.), pp. 19–45. Wiley–Liss, New York.

Huelsenbeck, J. P., and Rannala, B. (1997). Phylogenetic methods come of age: Testing hypotheses in an evolutionary context. *Science* **276**: 227–232.

Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.* 8: 275–287.

Jukes, T. H., and Cantor, C. R. (1969). Evolution of protein molecules. *In* "Mammalian Protein Metabolism" (H. N. Munro, Ed.), pp. 21–132. Academic Press, New York.

Källersjö, M., Albert, V. A., and Farris, J. S. (1999). Homoplasy increases phylogenetic structure. *Cladistics* **15**: 91–93.

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

Kraus, O. (1998). Phylogenetic relationships between higher taxa of tracheate arthropods. *In* "Arthropod Relationships" (R. A. Fortey and R. H. Thomas, Eds.), pp. 295–303. Chapman & Hall, London.

Kukalová-Peck, J. (1998). Arthropod phylogeny and 'basal' morphological structures. *In* "Arthropod Relationships" (R. A. Fortey and R. H. Thomas, Eds.), pp. 249–268. Chapman & Hall, London.

Le, T. H., Blair, D., Agatsuma, T., Humair, P.-F., Campbell, N. J. H., Iwagami, M., Littlewood, D. T. J., Peacock. B., Johnston, D. A., Bartley, J., Rollinson, D., Herniou, E. A., Zarlenga, D. S., and McManus, D. P. (2000). Phylogenies inferred from mitochondrial gene orders—A cautionary tale from parasitic flatworms. *Mol. Biol. Evol.* **17**: 1123–1125.

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

Maddison, W. P., and Maddison, D. R. (1992). MacClade: Analysis of phylogeny and character evolution. Version 3.0. Sinauer, Sunderland, MA.

Masta, S. E. (2000). Mitochondrial sequence evolution in spiders: Intraspecific variation in tRNAs lacking the TΨC arm. *Mol. Biol. Evol.* **17**: 1091–1100.

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

Nielsen, C. (1998). The phylogenetic position of the Arthropoda. *In* "Arthropod Relationships" (R. A. Fortey and R. H. Thomas, Eds.), pp. 11–22. Chapman & Hall, London.

Rambaut, A., and Grassly, N. C. (1997). Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput. Appl. Biosci.* **13**: 235–238.

Regier, J. C., Fang, Q. Q., Mitter, C., Peigler, R. S., Friedlander, T. P., and Solis, M. A. (1998). Evolution and phylogenetic utility of the *period* gene in Lepidoptera. *Mol. Biol. Evol.* **15**: 1172–1182.

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

Regier, J. C., and Shultz, J. W. (1998). Molecular phylogeny of arthropods and the significance of the Cambrian "explosion" for molecular systematics. *Am. Zool.* **38**: 918–928.

Regier, J. C., and Shultz, J. W. (2001). A phylogenetic analysis of Myriapoda (Arthropoda) using two nuclear protein-encoding genes. *Zool. J. Linn. Soc.*, in press.

Rodriguez, F., Oliver, J. L., Marin, A., and Medina. J. R. (1990). The general stochastic model of nucleotide substitution. *J. Theor. Biol.* **142**: 485–501.

Saitou, N., and Nei, M. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**: 406–425.

Shultz, J. W., and Regier, J. C. (2000). Phylogenetic analysis of arthropods using two nuclear protein-encoding genes supports a crustacean + hexapod clade. *Proc. R. Soc. Lond. B* **267**: 1011–1019.

Smith, S. W., Overbeek, R., Woese, C. R., Gilbert, W., and Gillevet, P. M. (1994). The genetic data environment and expandable GUI for multiple sequence analysis. *Comput. App. Biosci.* **10**: 671–675.

Spears, T., and Abele, L. G. (1998). Crustacean phylogeny inferred from 18S rDNA. *In* "Arthropod Relationships" (R. A. Fortey and R. H. Thomas, Eds.), pp. 169–187. Chapman & Hall, London.

Staden, R., Beal, K. R., and Bonfield, J. K. (1999). The Staden package, 1998. *In* "Bioinformatics Methods and Protocols" (S. Misener and S. Krawetz, Eds.). Humana Press, Totowa, NJ.

Sullivan, J., Swofford, D. L., and Naylor, G. J. P. (1999). The effect of taxon sampling on estimating rate heterogeneity parameters of maximum-likelihood models. *Mol. Biol. Evol.* **16**: 1347–1356.

Swofford, D. L. (1998). PAUP*, 4.0 beta version. Sinauer, Sunderland, MA.

Swofford, D. L., Olsen, G. J., Waddel, P. J., and Hillis, D. M. (1996). Phylogenetic inference. *In* "Molecular Systematics" (D. M. Hillis, C. Moritz, and B. K. Mable, Eds.), pp. 407–514. Sinauer, Sunderland, MA.

Wheeler, W. C. (1998). Sampling, groundplans, total evidence and the systematics of arthropods. *In* "Arthropod Relationships" (R. A. Fortey and R. H. Thomas, Eds.), pp. 87–96. Chapman & Hall, London.

Wheeler, W. C., Cartwright, P, and Hayashi, C. Y. (1993). Arthropod phylogeny: A combined approach. *Cladistics* **9**: 1–39.

Wheeler, W. C., and Hayashi, C. Y. (1998). The phylogeny of the extant chelicerate orders. *Cladistics* **14**: 173–192.

Wiegmann, B. M., Mitter, C., Regier, J. C., Friedlander, T. P., Wagner, D. M., and Nielsen, E. S. (2000). Nuclear genes resolve Mesozoic-aged divergences in the insect order Lepidoptera. *Mol. Phylogenet. Evol.* **15**: 242–259.

Zrzavý, J., Hypša, V., and Vlášková, M. (1998). Arthropod phylogeny: taxonomic congruence, total evidence and conditional combination approaches to morphological and molecular data sets. *In* "Arthropod Relationships" (R. A. Fortey and R. H. Thomas, Eds.), pp. 97–107. Chapman & Hall, London.