

Phylogenetic analysis of arthropods using two nuclear protein-encoding genes supports a crustacean + hexapod clade

Jeffrey W. Shultz^{1*} and Jerome C. Regier²

¹*Department of Entomology, University of Maryland, College Park, MD 20742, USA (js314@umail.umd.edu)*

²*Center for Agricultural Biotechnology, University of Maryland Biotechnology Institute, College Park, MD 20742, USA (regier@glue.umd.edu)*

Recent phylogenetic analyses using molecular data suggest that hexapods are more closely related to crustaceans than to myriapods, a result that conflicts with long-held morphology-based hypotheses. Here we contribute additional information to this debate by conducting phylogenetic analyses on two nuclear protein-encoding genes, elongation factor-1 α (EF-1 α) and the largest subunit of RNA polymerase II (Pol II), from an extensive sample of arthropod taxa. Results were obtained from two data sets. One data set comprised 1092 nucleotides (364 amino acids) of EF-1 α and 372 nucleotides (124 amino acids) of Pol II from 30 arthropods and three lobopods. The other data set contained the same EF-1 α fragment and an expanded 1038-nucleotide (346-amino-acid) sample of Pol II from 17 arthropod taxa. Results from maximum-parsimony and maximum-likelihood analyses strongly supported the existence of a Crustacea + Hexapoda clade (Pancrustacea) over a Myriapoda + Hexapoda clade (Atelocerata). The apparent incompatibility between the molecule-based Pancrustacea hypothesis and morphology-based Atelocerata hypothesis is discussed.

Keywords: molecular systematics; phylogeny; Arthropoda; elongation factor-1 α ; RNA polymerase II

1. INTRODUCTION

Phylogenetic analyses of the major arthropod groups using nuclear ribosomal genes often suggest that hexapods are more closely related to crustaceans than to myriapods (Field *et al.* 1988; Patterson 1989; Turbeville *et al.* 1991; Friedrich & Tautz 1995; Wheeler 1998; M. Friedrich, personal communication), a result supported by mitochondrial gene order (Boore *et al.* 1998) and protein-encoding nuclear genes (Regier & Shultz 1997). These findings contradict the long-accepted view that hexapods and myriapods form a group, Atelocerata, and have been criticized for their inconsistency with morphological evidence (Edgecombe 1998; Wägele 1996; Kraus 1998). However, evidence cited in favour of Atelocerata generally consists of lists of traditional morphological similarities rather than rigorous phylogenetic analyses or new characters. Other workers have been more accepting of a crustacean + hexapod clade, and some have provided lists of supportive morphological similarities, mostly drawn from neuroanatomy and morphogenesis (Averof & Akam 1995; Dohle 1997, 1998; Kutsch & Breidbach 1994; Osorio *et al.* 1995, 1997; Popadić *et al.* 1996). Again, however, these characters tend to be granted significance in the absence of phylogenetic analysis (Nilsson & Osorio 1998; Whitington & Bacon 1998).

Here we address the crustacean–hexapod–myriapod problem by analysing nucleotide (nt) and inferred amino acids (aa) from two nuclear genes, elongation factor-1 α (EF-1 α) and the largest subunit of RNA polymerase II (Pol II) from 30 arthropod and three lobopod species. We also analyse a recently expanded sample of Pol II (1038 nt, 346 aa), both alone and with EF-1 α , for 17 arthropods.

For both the 33- and 17-taxon studies, combined analyses of EF-1 α and Pol II reconstructed hexapods and crustaceans as a clade, Pancrustacea, using both maximum-parsimony (MP) and maximum-likelihood (ML) methods. Our results also supported the monophyly of Arthropoda, Pancrustacea, Chelicerata and Myriapoda, but did not provide compelling resolution within Pancrustacea or among Pancrustacea, Myriapoda and Chelicerata. We discuss possible reasons for the difficulty in resolving deep relationships within Arthropoda using molecular data as well as the apparent inconsistency of molecular and morphological characters in resolving arthropod phylogeny.

2. MATERIAL AND METHODS

(a) *Taxon sampling and specimen preservation*

Analyses were based on 36 arthropod species drawn from Hexapoda, Crustacea, Myriapoda and Chelicerata and three non-arthropods, that is, a tardigrade and two onychophorans (figure 1). Specimens were either alive until frozen at -85°C or stored in 100% ethanol at ambient temperature for up to one year before final storage at -85°C .

(b) *Data sets, polymerase chain reaction, sequencing and sequence assembly*

Sequences were partitioned into a 33-taxon data set (1092 nt of EF-1 α , 372 nt of Pol II) derived from 30 arthropods and three lobopods and a 17-taxon data set (1092 nt of EF-1 α , 1038 nt of Pol II) comprising arthropods only. Due to difficulty in amplifying Pol II, the 17-taxon set was not a strict subset of the 33-taxon set (figures 1 and 2). Combined analysis of all 39 taxa was considered premature given the substantial amount of missing data. Protocols for amplifying and sequencing EF-1 α cDNA (1093 nt excluding terminal primer sequences) have been

*Author for correspondence.

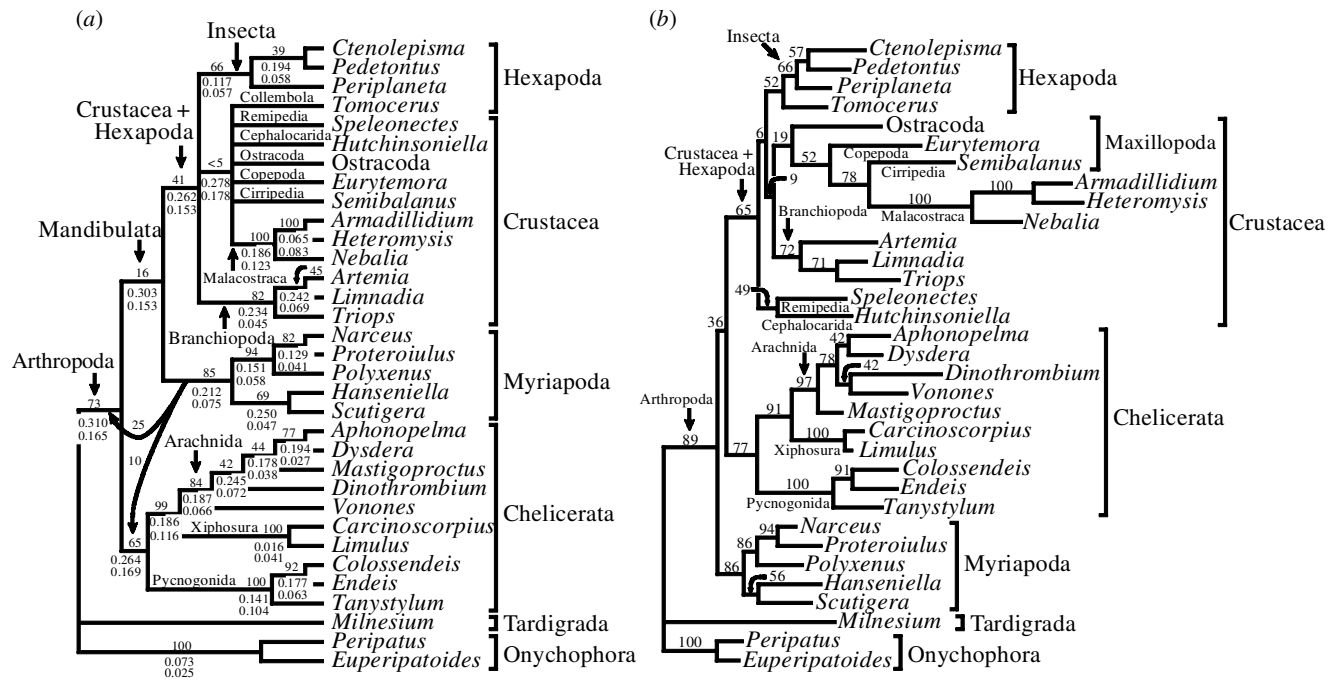


Figure 1. Phylogenetic analyses of combined EF-1 α + Pol II data for 33 taxa. (a) Strict consensus of six MP trees (consistency index (CI) = 0.5045, retention index (RI) = 0.5541, tree length = 1340, characters = 488, MP-informative characters = 195) based on analysis of amino acids and assuming equally weighted transformations. Bootstrap values are placed above branches. GenBank accession numbers (EF-1 α ; Pol II) follows. HEXAPODA: *Ctenolepisma lineata* (AF063405; AF138973), *Eumesocampa fragilis* (AF137388; AF138978–AF138980), *Machilooides banksi* (AF137390; AF138990–AF138992), *Metajapyx subterraneus* (AF137389; AF138987, AF138988), *Pedetontus saltator* (U90056; U90041), *Periplaneta americana* (U90054; U90040), *Tomocerus* sp. (U90059; AF139011, AF139012). CRUSTACEA: Malacostraca: *Armadillidium vulgare* (U90046; AF138970), *Heteromysis formosus* (AF063410; AF138983), *Nebalia hessleri* (AF063413; AF138996); Maxillopoda: *Semibalanus balanoides* (AF063404; AF138971, AF138972), *Ostracoda* (AF063414; AF138997–AF138999), *Eurytemora affinis* (AF063408; AF138977). Branchiopoda: *Artemia salina* (X03349; U10331), *Limnadia lenticularis* (AF063412; AF138989), *Triops longicaudatus* (U90058; U90043), *Remipedia*: *Speleonectes tulumensis* (AF063416; AF139008–AF139010), *Cephalocarida*: *Hutchinsoniella macracantha* (AF063411; AF138984–AF138986). MYRIAPODA: Diplopoda: *Narceus americanus* (U90053; U90039), *Proteroiulus fuscus* (AF063415; AF139000), *Polyxenus fascicularis* (U90055; AF139001, AF139002). Symphyla: *Hanseniella* sp. (U90049; AF138982), *Scutigera coleoptrata* (U90057; U90042). CHELICERATA: Arachnida: *Aphonopelma chalcodes* (U90045; U90035), *Dysdera crocata* (U90047; U90036), *Mastigoproctus giganteus* (U90052; U90038), *Dinothrombium pandorae* (U90048; AF138976), *Vonones ornata* (U90060; U90044), *Nipponopsalis abei* (AF137391; AF138993–AF138995), Xiphosura: *Carcinoscorpius rotundicauda* (AF063407; AF138975), *Limulus polyphemus* (U90051; U90037), Pycnogonida: *Colossendeis* sp. (AF063406; AF138974), *Endeis laevis* (AF063409; AF138981), *Tanystylum orbiculare* (AF063417; AF139013, AF139014). Outgroups: Tardigrada: *Milnesium tardigradum* (AF063419; AF139016). Onychophora: *Peripatus* sp. (AF137395; AF139017), *Euperipatoides rowelli* (AF137394; AF139015).

described elsewhere (Regier & Shultz 1997). However, our strategy for amplifying Pol II differed from our previous studies due to difficulty in amplifying the desired cDNA fragment across all taxa. This included the development of a new set of primers. The new Pol II primers are as follows: Pol15F, ACW GCH GAR ACH GSK TAY ATY CA [2318]; Pol14F, YTK ATH AAR GCT ATG GA [2342]; Pol8F, ATG ATH TGG AAY GYN CAR AA [2711]; Po30F, CAR ATG ACI YTN AAY ACI TTY CAY T [3061]; Pol17R, TTY TGN GCR TTC CAD ATC AT [2692]; Po28R, ART GRA AIG TRT TNA RIG TCA TYT G [3037]; Po27R, GCN CCA ACC ATY TCN CC [2986]; Po32R, CCY TGN ARI GTC ATR TC [3619]; Po23R, TTN TCI GCR TTR TCR TC [3418]. Degenerate positions are indicated by ambiguity codes (Dixon *et al.* 1985) with 'I' representing inosine. Primers labelled 'F' are forward primers and bind to the anti-sense strand; primers labelled 'R' are reverse-complement primers. Numbers in brackets at the 3'-end of each primer refer

to its position relative to Pol II (sense strand) from *Artemia salina* (GenBank accession no. U10331). Actual primers included a M13 sequence (not shown) at the 5'-end to facilitate automated sequencing.

Primers Pol15F and Pol17R were used to amplify a 373-nt region (excluding primer regions) for use in the 33-taxon analysis. A 1042-nt Pol II sequence was obtained from specimens in the 17-taxon analysis by amplifying three contiguous or adjacent fragments. Fragments 15/17, 18/28 and 30/32 were amplified by reverse transcriptase-polymerase chain reaction (RT-PCR), gel isolated, re-amplified with one nested primer (Pol14F/Pol17R, Pol8F/Po27R, Po30F/Po23R), and gel isolated. Faint bands were re-amplified with M13 primers and gel isolated prior to sequencing. cDNAs were amplified using a touchdown protocol (annealing conditions: 25 cycles from 55 °C to 45 °C followed by 12 cycles at 45 °C). Re-amplifications followed a standard three-step protocol (annealing

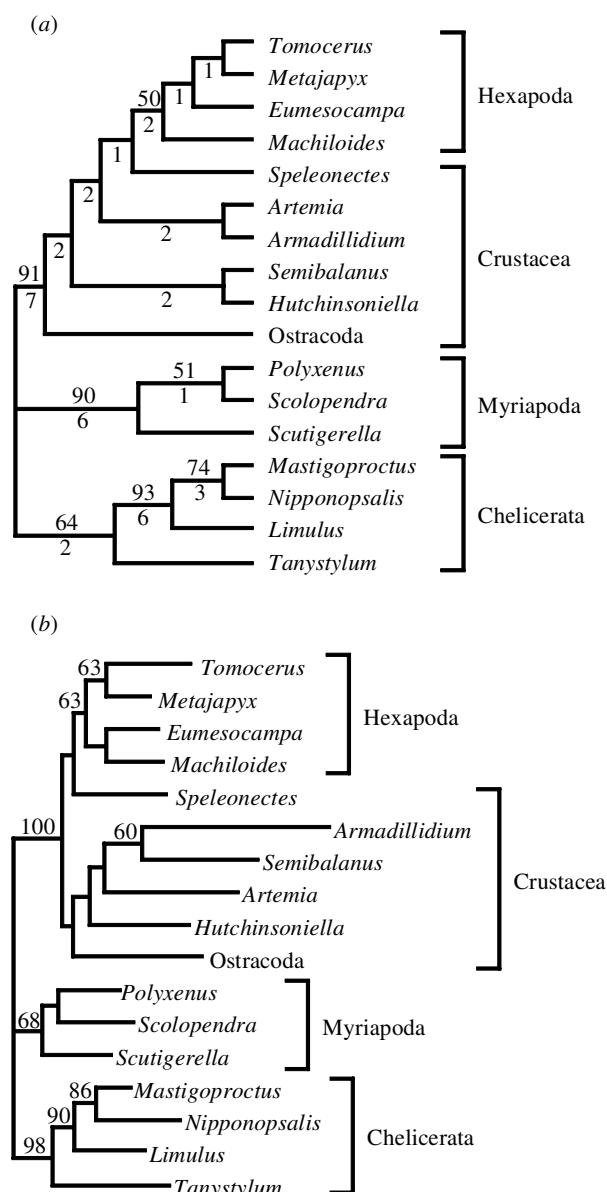


Figure 2. Phylogenetic analyses of combined EF-1 α + Pol II data for 17 taxa. (a) Single most-parsimonious tree (CI = 0.6197, RI = 0.4301, tree length = 1094, characters = 710, MP-informative characters = 201) based on analysis of amino acids and assuming equally weighted transformations. Bootstrap values greater than 50% are placed above branches. Decay indices are placed below branches. (b) Topology with highest likelihood (lnL = -10078.27138) based on analysis of nt1 + nt2 (three parameters per gene). Branch lengths are proportional to inferred number of changes. Bootstrap values are placed above branches. Terminal taxa are identified by genus only. The tree is unrooted.

conditions: 22 cycles at 50 °C). Sequencing reactions were fractionated and analysed on Applied Biosystems–Perkin-Elmer automated DNA sequencers.

Automated DNA-sequencer chromatograms were edited and contigs were assembled using TED and XDAP in the Staden software package (Dear & Staden 1991). Sequences from multiple species were aligned and nucleotide data sets were constructed using the Genetic Data Environment software package (v. 2.2; Smith *et al.* 1994). Optimal alignments of EF-1 α and Pol II required no indels. Amino-acid data sets were

constructed using MacClade, v.3.07 (Maddison & Maddison 1992).

(c) Data analysis

MP analyses were conducted on the 33-taxon data set and on the 17-taxon data set for EF-1 α and Pol II, separately and combined, using PAUP*4.0 test versions (D. Swofford, Smithsonian Institution) and v. β 1 (Swofford 1998). Nucleotides and amino acids were treated as unordered characters. Nucleotides were analysed with and without nt3, which appeared to be multiply substituted (see §3). MP analyses of amino acids were also conducted using a 'protpars' step matrix (Felsenstein 1993) in which serine codons differing at nt3 were coded separately. Analysis employed a heuristic search using tree-bisection and reconnection (TBR) branch swapping with random taxon addition (100 replications). Calculation of bootstrap percentages (Felsenstein 1985) also employed a heuristic search (1000 replications) using TBR branch swapping with random taxon addition (ten addition sequences–replicate). Decay or Bremer support indices (Bremer 1988; Donoghue *et al.* 1992) were also calculated. We assessed possible conflict between the amino-acid signal in EF-1 α and Pol II using the incongruence length difference test (Farris *et al.* 1995) implemented as the partition homogeneity test in PAUP*4.0, with 1000 random bi-partitions each analysed by TBR branch swapping on ten random addition sequences.

ML analyses of nucleotide data sets, with and without nt3, were performed with PAUP*4.0 under a general time-reversible (GTR) model (Rodriguez *et al.* 1990). The GTR model was selected over others based on a likelihood-ratio test (Huelsenbeck & Rannala 1997). Briefly, likelihood scores for a constrained topology were calculated assuming various models (general reference, Swofford *et al.* 1996; GTR, Rodriguez *et al.* 1990; SYM, Zharkikh 1994; HKY85, Hasegawa *et al.* 1985; K2P, Kimura 1980; JC, Jukes & Cantor 1969). For pairwise comparisons of models, differences in likelihood scores were doubled and this statistic was tested using a χ^2 -distribution with n degrees of freedom, where n is the number of parameters that differ between substitution models. The GTR model was significantly preferred over all other models (p -values < 0.005). Among-site rate variation was accommodated by estimating the proportion of sites assumed to be invariable (Hasegawa *et al.* 1985) and by assigning separate rate categories to each of the three codon positions, or two in cases where nt3 was excluded. To partition synonymous and non-synonymous change more completely, we also partitioned nt1 sites into two separate rate categories, those that encode two or more leucine or arginine residues across all represented species and those that did not; only leucine and arginine codons undergo single-nucleotide, synonymous substitutions at nt1. This approach yielded a four-parameter-per-gene model overall, or a three-parameter-per-gene model with nt3 excluded. Likelihood-ratio tests showed that bi-partitioning of nt1 yielded a significant improvement (p -values < 0.005). As a first step in the ML search, likelihood parameters were optimized using MP trees derived from nucleotides (nt3 excluded) and/or amino acids. Nearest-neighbour-interchange (NNI) branch swapping was then performed and new likelihood parameters were estimated from the most likely topology. We then performed TBR branch swapping on the resulting tree and re-estimated the likelihood parameters. These parameters were used for a heuristic search with NNI branch swapping and 100 random taxon additions. After optimization on the resulting ML topology, the likelihood parameters from the overall best tree

Table 1. *Support provided for selected arthropod groups by a 33-taxon sample of EF-1 α and Pol II using different analytical methods*

(Recovery of a group by a particular method (ML, maximum likelihood; MP, maximum parsimony) is indicated by an asterisk (*) and support is indicated by bootstrap percentages when such values are greater than 50%. Taxa indicated by superscript 'T' are test clades; i.e. clades that are generally accepted and strongly supported by morphological or other characters. Other abbreviations: aa, amino acids; nt, nucleotides; par, parameter; = wt, equal weights; protpars, protein parsimony; Ch, Chelicerata; Cr, Crustacea; Hx, Hexapoda; My, Myriapoda.)

taxon	EF1- α + Pol II						EF-1 α				Pol II			
	ML	ML	MP	MP	MP	MP	ML	ML	MP	MP	ML	ML	MP	MP
	all nt 4 par	Ntl+2 3par	all nt = wt	ntl+2 = wt	aa = wt	aa prot- pars	all nt 4 par	ntl+2 3 par	ntl+2 = wt	aa = wt	all nt 4 par	ntl+2 3 par	ntl+2 = wt	aa = wt
Arthropoda	75*	89*	*	88*	73*	75*	59*	75*	74*	63*	*	*	*	*
Pancrustacea	*	65*	—	61*	*	*	—	—	—	—	57*	62*	65*	72*
Atelocerata	—	—	—	—	—	—	*	—	—	—	—	—	—	—
Mandibulata	—	—	—	—	*	—	—	—	—	—	—	—	—	—
Cr + Hx + Ch	*	*	—	—	—	*	—	—	—	—	—	—	—	—
My + Ch	—	—	—	—	—	—	—	—	—	—	—	—	*	*
Crustacea	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Malacostraca ^T	100*	100*	100*	100*	100*	100*	100*	100*	100*	100*	95*	98*	100*	99*
Eumalacostraca ^T	100*	100*	100*	100*	100*	100*	100*	100*	100*	98*	97*	100*	100*	99*
Branchiopoda ^T	70*	72*	—	81*	83*	83*	60*	67*	83*	88*	—	—	—	—
Maxillopoda	—	—	—	—	—	—	—	—	—	—	—	*	—	—
Hexapoda ^T	63*	56*	—	—	—	*	59*	62*	73*	56*	—	—	—	—
Insecta ^T	68*	65*	*	—	66*	75*	77*	74*	52*	61*	—	—	—	—
Myriapoda	75*	86*	56*	72*	85*	76*	53*	58*	63*	82*	*	*	—	—
Diplopoda ^T	97*	86*	79*	68*	94*	95*	67*	62*	—	92*	87*	55*	—	—
Chelicerata	55*	77*	—	83*	65*	*	—	—	—	—	*	57*	64*	*
Pycnogonida ^T	100*	100*	100*	100*	100*	100*	100*	100*	100*	97*	100*	100*	99*	84*
Xiphosura ^T	100*	100*	100*	100*	100*	100*	100*	100*	100*	100*	100*	98*	99*	98
Euchelicerata ^T	97*	91*	61*	81*	99*	98*	91*	91*	86*	96*	*	—	—	57*
Arachnida ^T	93*	97*	57*	94*	84*	84*	87*	94*	92*	85*	*	*	*	*
Tetrapulmonata ^T	—	—	—	—	*	—	—	—	—	59	—	—	—	—
Araneae ^T	—	*	—	—	77*	52	—	*	—	89*	—	—	—	—

(figure 1) were used in bootstrap analyses (100–1000 replications). Each replication was based on a heuristic search with NNI branch swapping and ten random sequence additions.

ML analyses for the combined 33-taxon amino-acid data were performed using the *protml* program in MOLPHY (v.2.2 and v.2.3; Adachi & Hasegawa 1994), which incorporates empirical transition matrices (Dayhoff *et al.* 1978; Jones *et al.* 1992). To do this, the 1244 amino-acid trees within three steps of the MP tree (1340 steps) were read into *protml*, and their likelihood scores were computed.

We used the test of Kishino & Hasegawa (1989) to determine whether alternative trees differed significantly in how well they fit the data. The trees included those proposed by previous workers and those generated in the course of this study. Only fully dichotomous trees were compared. In cases where a hypothesis specified only a few nodes (e.g. Crustacea + Hexapoda), relationships among the specified taxa were constrained, and unconstrained relationships were re-optimized.

Relative rates of non-synonymous substitution in EF-1 α and Pol II were estimated by ML using total estimated change for nt2 on optimal trees obtained from combined analysis. Relative rates were determined separately for the 33- and 17-taxon sets. The likelihood model was as described above. Percentage differences of all pairwise combinations of EF-1 α and of Pol II amino-acid sequences were calculated in PAUP*4.0. Average differences

were plotted on the MP tree obtained from analysis of amino acids of EF-1 α and Pol II. Differences were calculated by averaging all values across the basal dichotomy within a particular clade. Base compositions were calculated by gene and by codon position using PAUP*4.0.

3. RESULTS

(a) *Recovery of test clades*

Clades widely accepted by systematists were designated 'test clades' (tables 1 and 2). We identified 12 test clades for the 33-taxon set and three for the 17-taxon set. We explored the ability of EF-1 α and Pol II, separately and combined, to recover test clades using ML and MP analyses under a variety of weighting schemes (tables 1 and 2). ML analysis of ntl + nt2 and MP analysis of amino acids recovered the most test clades. EF-1 α recovered as many test clades as the best combined analysis (33 taxa: ML analysis of ntl + nt2, MP analysis of amino acids; 17 taxa: ML and MP analyses of ntl + nt2). Bootstrap support for many test clades was 70% or higher, except Hexapoda (both data sets), Insecta, Tetrapulmonata and Araneae. Test-clade recovery for Pol II was generally lower than for EF-1 α and the combined data.

Table 2. Support provided for selected arthropod groups by a 17-taxon sample of EF-1 α and Pol II using different analytical methods

(Recovery of a group by a particular method (ML or MP) is indicated by an asterisk and support is indicated by bootstrap percentages when values are greater than 50%. Taxa indicated by superscript 'T' are test clades; i.e. clades that are generally accepted and strongly supported by morphological or other characters. Other abbreviations as in table 1.)

taxon	EF-1 α + Pol II						EF-1 α				Pol II			
	ML	ML	MP	MP	MP	MP	ML	ML	MP	MP	ML	ML	MP	MP
	all nt 4 par	ntl+2 3 par	all nt = wt	ntl+2 = wt	aa = wt	aa prot- pars	all nt 4 par	ntl+2 3 par	ntl+2 = wt	aa = wt	all nt 4 par	ntl+2 3 par	ntl+2 = wt	aa = wt
Pancrustacea	—	100*	—	94*	91*	97*	—	—	—	—	—	95*	94*	98*
Atelocerata	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Cr + My	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Crustacea	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Hexapoda ^T	—	63*	—	—	50*	*	—	60*	*	—	—	—	—	—
Myriapoda	—	68*	—	57*	90*	81*	—	76*	72*	77*	—	—	—	*
Chelicerata	68*	98*	—	92*	64*	55*	—	—	—	—	58*	81*	*	—
Euchelicerata ^T	93*	90*	52*	76*	93*	94*	80*	86*	89*	90*	75*	—	—	79*
Arachnida ^T	79*	86*	62*	64*	74*	82*	80*	89*	80*	81*	—	—	—	*

(b) Base compositions, pairwise differences, substitution rates and gene combination

Base compositions of EF-1 α and of Pol II were found to be biased when analysed by codon position, although only nt3 was significantly non-homogeneous (χ^2 -test, $p = 0.000$). For EF-1 α , base compositions by codon position differed less than 10% across data sets. For Pol II, across-data-set differences were larger (nt2: 38% difference for C; nt3: 21% difference for C). This difference appeared to represent different the amounts of Pol II sequence rather than different taxon samples (J. W. Shultz and J. C. Regier, unpublished observations).

In previous work on EF-1 α amino acids in arthropods and outgroups, we observed that the largest pairwise distances (>12%) were associated with unstable and weakly supported groupings (Regier & Shultz 1998). The corresponding pairwise differences for Pol II amino acids were about twice as high (figure 1a), which is consistent with likelihood estimates indicating that non-synonymous substitution rate in Pol II is higher than that inferred for EF-1 α . For nucleotides, pairwise differences at nt3 (33-taxon data set) were generally greater than 50% for both genes (range = 35–66% for EF-1 α and 38–76% for Pol II), suggesting high levels of homoplasy at this site. The decision to combine multiple genes within one analysis should consider the possibility of conflicting phylogenetic signals (Cannatella *et al.* 1998). Reliable criteria for recognizing incompatibility of data sets have not been established, but at least two lines of evidence indicated that combining EF-1 α and Pol II was appropriate. First, incongruence length difference tests conducted on the 33-taxon data did not reveal significant levels of conflict ($p > 0.05$) between EF-1 α and Pol II amino acids when test clades were constrained to be monophyletic. Second, unconstrained topologies based on ML analysis of EF-1 α and of Pol II nucleotides for the 17-taxon data did not differ significantly (Kishino–Hasegawa test, $p = 0.1789$ when EF-1 α ntl + nt2 data were fitted to trees of highest likelihood, $p = 0.7706$ when all nucleotides from Pol II were fitted to trees of highest

likelihood). However, other comparisons revealed potential conflict when the cut-off for significance was $p = 0.050$ (Cunningham 1997). Still, given the uncertainty about cut-off levels for identifying significant conflict, we chose to analyse EF-1 α and Pol II, both separately and in combination, using a variety of analytical approaches.

(c) Analysis of 33-taxon data sets

For combined data, methods that recovered the most test clades also tended to recover Arthropoda, Myriapoda, Crustacea + Hexapoda and Chelicerata (table 1 and figure 1). Malacostraca + Maxillopoda was recovered by ML analysis of ntl + nt2 (figure 1), by MP analysis of amino acids in three out of six MP trees (figure 1; J. W. Shultz and J. C. Regier, unpublished observations), and in the amino-acid MP tree with the highest likelihood score (not shown). Except for Crustacea + Hexapoda, relationships among Crustacea, Hexapoda, Myriapoda and Chelicerata were ambiguous. Alternative groupings proposed by previous workers, including Mandibulata (= Crustacea + Hexapoda + Myriapoda) (Snodgrass 1938) and Schizoramia (= Crustacea + Chelicerata) (Cisne 1974), were not recovered or had low bootstrap values (figure 1, tables 1 and 2). Relationships among crustacean classes were also unresolved, except perhaps the grouping of Malacostraca and a paraphyletic Maxillopoda, even though lower-level clades (e.g. Branchiopoda, Malacostraca) and higher-level clades (e.g. Crustacea + Hexapoda) were supported. Significantly, neither Atelocerata nor Crustacea were supported as monophyletic groups by any analytical method.

For EF-1 α , MP analysis of amino acids resulted in a strict consensus tree less resolved than that obtained with the combined data, despite high recovery of test clades (J. W. Shultz and J. C. Regier, unpublished observations, table 1). When MP and ML analyses of amino acids and nucleotides, respectively, were compared, the major differences were found across groups in which internal pairwise sequence differences were the greatest (figure 1a)

and bootstrap percentages (BP) were the lowest (table 1, J. W. Shultz and J. C. Regier, unpublished observations). Thus, relationships among the major arthropod clades as recovered by EF-1 α were unstable, as were class relationships within Crustacea and placement of Pycnogonida. In contrast, groups with lower internal divergence such as Arachnida, Euchelicerata, Pycnogonida, Myriapoda, Diplopoda, Hexapoda, Insecta and Branchiopoda were recovered. Two groups with high internal sequence divergence (Malacostraca and Arthropoda) were also recovered, but this may have reflected the relatively greater distance between these groups and their nearest relatives (Tardigrada and Onychophora for Arthropoda). ML analysis of total nucleotides was the only approach that recovered Atelocerata, but bootstrap support was only 6%. EF-1 α alone provided virtually no support for a Crustacea + Hexapoda clade (<5% BP) (table 1), but neither did it offer a well-supported alternative.

Pol II supported a Crustacea + Hexapoda clade (72% BP by MP analysis using amino acids and 57% BP by ML analysis using nucleotides) (table 1). Like EF-1 α , Pol II did not moderately or strongly support any group not also found in the combined analysis (figure 1). Groups weakly supported by Pol II but moderately to strongly supported by EF-1 α included Branchiopoda, Hexapoda, Insecta, Myriapoda, Xiphosura, Euchelicerata, Arachnida and Araneae (table 1).

(d) *Analysis of 17-taxon data sets*

For combined data, methods that recovered the most test clades (ML analysis of ntl + nt2, MP of amino acids) also recovered Crustacea + Hexapoda, Myriapoda and Chelicerata (table 2, figure 2). BP values for Crustacea + Hexapoda were greater than 90% for amino acids and ntl + nt2; Crustacea + Hexapoda was not recovered when nt3 was included in the data set. Inclusion of nt3 also reduced support for Hexapoda Myriapoda and Chelicerata, but only weakly affected support for Euchelicerata and Arachnid. Myriapoda was strongly supported by MP analysis of amino acids (90% BP), and Chelicerata was strongly supported by ML analysis of ntl + nt2 (98% BP). Due to difficulty in obtaining Pol II in outgroup taxa, we could neither resolve relationships among Crustacea + Hexapoda, Myriapoda and Chelicerata nor evaluate support for Mandibulata or Schizoramia. However, results from MP analysis of amino acids and ML analysis of ntl + nt2 were inconsistent with a monophyletic Atelocerata (Kishino–Hasegawa test, $p = 0.020$ for amino acids, 0.0007 for ntl + nt2).

For EF-1 α , MP analysis revealed strong support for Myriapoda, Euchelicerata, and Arachnida and modest support for Hexapoda (table 2). Other relationships were unresolved or supported by decay indices of one step and bootstrap values below 50% (figure 2). With regard to the grouping of Crustacea and Hexapoda, the amino-acid data did not fit significantly better on the unconstrained MP topology (tree length = 1742) than on the MP topology in which the Crustacea + Hexapoda group was constrained to be monophyletic (tree length = 1749) (Kishino–Hasegawa test, $p = 0.3079$). ML analysis of ntl + nt2 recovered a clade including hexapods and all crustaceans except *Speleonectes* (21% BP). In summary, relationships recovered from phylogenetic analysis of

EF-1 α were either consistent with those derived from the combined data and from the 33-taxon analysis or were only weakly supported by the data. For Pol II alone, analysis of amino acids and of ntl + nt2 (but not nt3) strongly supported a Crustacea + Hexapoda clade (>94% BP, table 2). Myriapoda and Chelicerata and their subgroups were recovered by some, but not all, methods and data sets. Hexapoda was not recovered and relationships among crustacean classes were unstable.

4. DISCUSSION

(a) *Molecular evidence and the Pancrustacea concept*

The possibility that hexapods are more closely related to crustaceans than to myriapods was suggested by the earliest molecular systematic studies using small subunit (18S) ribosomal genes (Field *et al.* 1988; Patterson 1989). The result was inconclusive, however, because the analyses included only four arthropod taxa (*Artemia*, *Drosophila*, *Spiroboles* and *Limulus*), three of which appeared to have highly divergent sequences. Specifically, it was suggested that the fruit fly *Drosophila* and brine shrimp *Artemia* were grouping due to ‘long-branch attraction’ (Felsenstein 1988). Even when these problems were partially addressed by including more arthropods and replacing *Drosophila* and *Artemia* with a beetle (*Tenebrio*) and a crayfish (*Procambarus*), respectively, analysis of 18S rDNA still recovered a crustacean + hexapod clade under a variety of analytical schemes (Turbeville *et al.* 1991; see also Wheeler *et al.* 1993).

However, it is unclear whether the crustacean + hexapod clade is robust to expanded taxon sampling. In their analysis of 18S rDNA from 23 crustaceans, seven hexapods and other groups, Spears & Abele (1998) did not recover a crustacean + hexapod clade unless several singly represented, ‘long-branch’ crustacean taxa (Cephalocarida, Mystacocarida and Remipedia) were excluded, and then bootstrap support was only 51%. Another study that included 28S as well as 18S rDNA showed 100% bootstrap support for a crustacean + hexapod clade, although only ten arthropods were sampled (Friedrich & Tautz 1995). Again, the result depended on exclusion of certain ‘problematic’ taxa. Our reanalysis of these data showed that high bootstrap support for a crustacean + hexapod clade resides in the 28S gene (100% by MP analysis), not the 18S gene (20% BP). M. Friedrich and D. Tautz (personal communication) have recently used ML and parametric bootstrap analyses to determine if their results could be attributed to long-branch attraction, and their results corroborate their previous findings.

We can now add the substantial support of Pol II for a crustacean + hexapod clade. Pol II differs from 18S rDNA in that support is present even when highly divergent and singly represented clades, such as Cephalocarida or Remipedia, are included. Furthermore, bootstrap support approaches 100% with increased sampling of Pol II (tables 1 and 2), although support for a crustacean + hexapod clade diminishes when nt3 is included. Given the high pairwise divergences at nt3 (typically >50%), the non-homogeneous base composition at nt3, the reduced recovery of test clades when nt3 is included (tables 1 and 2), and the widespread recognition that non-synonymous changes and amino-acid changes are more

conservative than synonymous changes, it is important to emphasize the robustness of our results for recovery of a crustacean + hexapod clade with nt3 excluded.

From the viewpoint of experimental design, it is noteworthy that, in contrast to Pol II, EF-1 α does not provide support for a crustacean + hexapod clade, despite its greater overall recovery of well-established clades (table 1) and its lower overall rate of non-synonymous substitution. Recently, molecular systematists have discussed the relative merits of increased taxon sampling and increased character sampling in improving phylogenetic resolution (e.g. Graybeal 1998). Our study indicates that gene selection can also be important because different molecules can support different parts of a tree without conflicting in other parts. As yet, there is no simple recipe for identifying suitable genes, despite progress (Graybeal 1994; Friedlander *et al.* 1994; Brower & DeSalle 1998).

Although recent molecular analyses of arthropod phylogeny have tended to recover a crustacean + hexapod clade, they also share an inability to recover the placement of myriapods and chelicerates and relationships within Crustacea and Hexapoda. It may be premature to ascribe any special significance to this observation, but two possible explanations are apparent. First, the spectrum of evolutionary rates represented by these data may not be appropriate for resolving phylogenetic events at the relevant divergence times. This explanation suggests that progress in arthropod phylogeny may depend more on the use of new molecular markers than expanded taxon sampling of established markers. Second, the ability to resolve phylogenetic events with any character system may be positively correlated with the time elapsing between cladogenetic events, and a persistent inability to resolve certain relationships may reflect accelerated rates of cladogenesis, such as may have occurred during the Cambrian 'explosion' or during the invasion of freshwater and terrestrial habitats.

(b) *Morphology and the Atelocerata versus Pancrustacea problem*

Many recent analyses that compare morphological and molecular data suggest that phylogenetic hypotheses that are well supported by molecular evidence rarely conflict with those that are well supported by morphological evidence (Omland 1997). Rather, molecular evidence is generally consistent with compelling morphology-based results or illuminates areas where morphology is problematic (Moritz & Hillis 1996). In an apparent departure from this trend, findings from various molecular studies seem to conflict with certain long-held views on the relationships among mandibulate arthropods (i.e. Crustacea, Hexapoda and Myriapoda). Specifically, systematists have traditionally recognized hexapods and myriapods as a monophyletic group, Atelocerata (Kingsley 1894; Snodgrass 1938; Cisne 1974; Manton 1977; Boudreaux 1979; Weygoldt 1986; Wills *et al.* 1994; Kraus 1998; etc.). Yet, molecular studies tend to support a clade, Pancrustacea, that encompasses crustaceans and hexapods and excludes myriapods. This result is particularly striking because it has emerged from studies of independent sets of molecular characters, including nuclear protein-encoding genes (Regier & Shultz 1997; present study), nuclear

ribosomal nucleotides (Friedrich & Tautz 1995) and mitochondrial gene order (Boore *et al.* 1998). If the Pancrustacea hypothesis is eventually shown to be correct, some workers may attribute the inconsistency between morphological and molecular data to shortcomings of morphological data. However, this conclusion would not be appropriate. It is important when considering this issue to separate the phylogenetic use of morphology from the philosophies and methods used in discovering morphological characters and in deriving phylogenetic hypotheses from them.

A feature of many older morphology-based studies was their aim to demonstrate relationships through subjective evaluation of specific characters rather than through testing phylogenetic hypotheses within a generally accepted analytical framework. Specifically, the goal of many pre-parsimony arthropod systematists was to erect prioritized lists of similarities to somehow gauge phylogenetic affinities or to propose speculative evolutionary scenarios and to erect trees consistent with the resulting transformation. The current practice of determining the states of many characters in each of many representative taxa emerged recently with the widespread acceptance and computerization of matrix-based parsimony analysis. Because the heyday of comparative anatomy ended long before this analytical innovation, the literature is replete with diverse anatomical descriptions, each focusing on a different organ system in a different set of taxa. Consequently, it is now difficult or impossible to use existing knowledge to fill species-by-character matrices that encompass a significant range of phylogenetic and morphological diversity.

The problem is exemplified by the current debate on the validity of Atelocerata. Arthropod systematists have tended to support this group with a traditional list of similarities, including the presence of ectodermal Malpighian tubules, tracheal systems, post-antennal sense organs, coxal vesicles, anterior tentorial apodemes and absence of claw levator muscles, second antennae and mandibular palpi (Dohle 1997, 1998; Kraus & Kraus 1996; Kraus 1998; Wägele 1996). This list is impressive, but the inductive approach it represents ignores such complicating factors as homoplasy within Atelocerata and morphological diversity within Crustacea (e.g. vestigial second antennae and absence of mandibular palpi in several crustacean lineages (McLaughlin 1980; Schram 1986)). We know of no parsimony-based morphological analysis that has acknowledged and accommodated homoplasy within a meaningful spectrum of crustacean, hexapod and myriapod species or has actively sought possible 'ateloceratan synapomorphies' among crustaceans and chelicerates. The absence of such studies has led to the impression that Atelocerata is supported by a compelling array of time-tested morphological characters. This impression, coupled with a modern aversion to comparative morphological work, has led some systematists to import traditional views wholesale into MP analyses (e.g. Eernisse *et al.* 1992; Wheeler *et al.* 1993; Wills *et al.* 1994; Wheeler 1998), thereby reinforcing the perception that morphology-based analyses of arthropod phylogeny are empirically robust and analytically rigorous.

Given the incompatibility of the existing morphological literature with rigorous application of matrix-based

analytical methods, an obvious tactic for assessing the phylogenetic use of morphology would be to conduct original, exhaustive anatomical surveys of a relevant spectrum of representative arthropods and outgroups. Given a matrix comparable in completeness to that routinely generated for molecular sequence data, the relative contributions of morphology and molecules to resolving arthropod phylogeny could be assessed rigorously and objectively.

We thank Markus Friedrich for information on unpublished research, Nobuo Tsurusaki for providing *Nipponopsalis*, and Diane Nelson for identifying *Milnesium*. This work was supported by the Maryland Agricultural Experiment Station, the Center for Agricultural Biotechnology (University of Maryland Biotechnology Institute), and National Science Foundation grants DEB-9629791 and IBN-9733777.

REFERENCES

- Adachi, J. & Hasegawa, M. 1994 *Programs for molecular phylogenetics*, v. 2.2. Tokyo: Institute of Statistical Mathematics.
- Averof, M. & Akam, M. 1995 Insect–crustacean relationships: insights from comparative developmental and molecular studies. *Phil. Trans. R. Soc. Lond.* **B347**, 293–303.
- Boore, J. L., Lavov, D. V. & Brown, W. M. 1998 Gene translocation links insects and crustaceans. *Nature* **392**, 667–668.
- Boudreaux, H. B. 1979 *Arthropod phylogeny with special reference to insects*. New York: Wiley.
- Bremer, K. 1988 The limits of amino-acid sequence data in angiosperm phylogenetic reconstruction. *Evolution* **42**, 795–803.
- Brower, A. V. Z. & DeSalle, R. 1998 Patterns of mitochondrial versus nuclear DNA sequence divergence among nymphalid butterflies: the utility of *wingless* as a source of characters for phylogenetic inference. *Insect Mol. Biol.* **7**, 73–82.
- Cannatella, D. C., Hillis, D. M., Chippindale, P. T., Weigt, L., Rand, A. S. & Ryan, M. J. 1998 Phylogeny of frogs of the *Physalaemus pustulosus* species group, with an examination of data incongruence. *Syst. Biol.* **47**, 311–335.
- Cisne, J. L. 1974 Trilobites and the origin of arthropods. *Science* **186**, 13–18.
- Cunningham, C. W. 1997 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.
- Dayhoff, M. O., Schwartz, R. M. & Orcutt, B. C. 1978 *Atlas of protein sequence and structure*, vol. 5, suppl. 3. Washington, DC: National Biomedical Research Foundation.
- Dear, S. & Staden, R. 1991 A sequence assembly and editing program for efficient management of large projects. *Nucl. Acids Res.* **19**, 3907–3911.
- Dixon, H. B. F., Bielka, H., Cantor, C. R., Liébecq, C., Sharon, N., Velick, S. F. & Vliegenthart, J. F. G. 1985 Nomenclature for incompletely specified bases in nucleic acid sequences: [NC-IUB] Recommendations 1984. *Mol. Biol. Evol.* **3**, 99–108.
- Dohle, W. 1997 Are the insects more closely related to the crustaceans than to the myriapods? *Entomol. Scand. Suppl.* **51**, 7–16.
- Dohle, W. 1998 Myriapod–insect relationships as opposed to an insect–crustacean sister group. In *Arthropod relationships* (ed. R. A. Fortey & R. H. Thomas), pp. 305–315. London: Chapman & Hall.
- Donoghue, M. J., Olmstead, R. G., Smith, J. F. & Palmer, J. D. 1992 Phylogenetic relationships of Dipsacales based on rbcL sequences. *Annls Missouri Bot. Gard.* **79**, 333–345.
- Edgecombe, G. D. 1998 Devonian terrestrial arthropods from Gondwana. *Nature* **394**, 172–175.
- Eernisse, D. J., Albert, J. S. & Anderson, F. E. 1992 Annelida and Arthropoda are not sister taxa: a phylogenetic analysis of spiralian metazoan morphology. *Syst. Biol.* **41**, 305–330.
- Farris, J. S., Källersjö, M., Kluge, A. G. & Bult, C. 1995 Constructing a significance test for incongruence. *Syst. Biol.* **44**, 570–572.
- Felsenstein, J. 1985 Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**, 783–791.
- Felsenstein, J. 1988 Phylogenies from molecular sequences: inferences and reliability. *A. Rev. Genet.* **22**, 521–565.
- Felsenstein, J. 1993 *PHYLIP (phylogeny inference package)*. Seattle, WA: Department of Genetics, University of Washington.
- Field, K. G., Olsen, G. J., Lane, D. J., Giovannoni, S. J., Ghiselin, M. T., Raff, E. C., Pace, N. R. & Raff, R. A. 1988 Molecular phylogeny of the animal kingdom. *Science* **239**, 748–753.
- Friedlander, T. P., Regier, J. C. & 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. & Tautz, D. 1995 Ribosomal DNA phylogeny of the major extant arthropod classes and the evolution of myriapods. *Nature* **376**, 165–166.
- Graybeal, A. 1994 Evaluating the phylogenetic utility of genes: a search for genes informative about deep divergences among vertebrates. *Syst. Biol.* **43**, 174–193.
- Graybeal, A. 1998 Is it better to add taxa or characters to a difficult phylogenetic problem? *Syst. Biol.* **47**, 9–17.
- Hasegawa, M., Kishino, H. & Yano, T. 1985 Dating of the human–ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22**, 160–174.
- Huelsenbeck, J. P. & Rannala, B. 1997 Phylogenetic methods come of age: testing hypotheses in an evolutionary context. *Science* **276**, 227–232.
- Jones, D. T., Taylor, W. R. & Thornton, J. M. 1992 The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.* **8**, 275–287.
- Jukes, T. H. & Cantor, C. R. 1969 Evolution of protein molecules. In *Mammalian protein metabolism* (ed. H. N. Munro), pp. 21–132. New York: Academic Press.
- Kimura, M. 1980 A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16**, 111–120.
- Kingsley, J. S. 1894 The classification of the Arthropoda. *Am. Nat.* 118–135.
- Kishino, H. & Hasegawa, M. 1989 Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. *J. Mol. Evol.* **29**, 170–179.
- Kraus, O. 1998 Phylogenetic relationships between higher taxa of tracheate arthropods. In *Arthropod relationships* (ed. R. A. Fortey & R. H. Thomas), pp. 295–303. London: Chapman & Hall.
- Kraus, O. & Kraus, M. 1996 On myriapod/insect interrelationships. *Mém. Mus. Natn. Hist. Nat.* **169**, 283–290.
- Kutsch, W. & Breidbach, O. 1994 Homologous structures in the nervous systems of Arthropoda. *Adv. Insect Physiol.* **24**, 1–113.
- McLaughlin, P. A. 1980 *Comparative morphology of Recent Crustacea*. San Francisco, CA: W. H. Freeman & Co.
- Maddison, W. P. & Maddison, D. R. 1992 *MacClade: analysis of phylogeny and character evolution*, v. 3. Sunderland, MA: Sinauer Associates.
- Manton, S. M. 1977 *The Arthropoda: habits, functional morphology, and evolution*. London: Clarendon.
- Moritz, C. & Hillis, D. M. 1996 Molecular systematics: context and controversies. In *Molecular systematics*, 2nd edn (ed. D. M. Hillis, C. Moritz & B. K. Mable), pp. 1–13. Sunderland, MA: Sinauer Associates.

- Nilsson, D.-E. & Osorio, D. 1998 Homology and parallelism in arthropod sensory processing. In *Arthropod relationships* (ed. R. A. Fortey & R. H. Thomas), pp. 333–347. London: Chapman & Hall.
- Omland, K. E. 1997 Correlated rates of molecular and morphological evolution. *Evolution* **51**, 1381–1393.
- Osorio, D., Averof, M. & Bacon, J. P. 1995 Arthropod evolution: great brains, beautiful bodies. *Trends Ecol. Evol.* **10**, 449–454.
- Osorio, D., Bacon, J. P. & Whittington, P. M. 1997 The evolution of arthropod nervous systems. *Am. Sci.* **85**, 244–253.
- Patterson, C. 1989 Phylogenetic relations of major groups: conclusions and prospects. In *The hierarchy of life* (ed. B. Fernholm, K. Bremer & H. Jörnvall), pp. 471–488. Amsterdam: Elsevier Science Publishers.
- Popadić, A., Rusch, D., Peterson, M., Rogers, B. T. & Kaufman, T. C. 1996 Origin of the arthropod mandible. *Nature* **380**, 395.
- Regier, J. C. & 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. & Shultz, J. W. 1998 Molecular phylogeny of arthropods and the significance of the Cambrian ‘explosion’ for molecular systematics. *Am. Zool.* **38**, 918–928.
- Rodriguez, F., Oliver, J. L., Marin, A. & Medina, J. R. 1990 The general stochastic model of nucleotide substitution. *J. Theor. Biol.* **142**, 485–501.
- Schram, F. R. 1986 *Crustacea*. New York: Oxford University Press.
- Smith, S. W., Overbeek, R., Woese, C. R., Gilbert, W. & Gillevet, P. M. 1994 The genetic data environment and expandable GUI for multiple sequence analysis. *Comput. Appl. Biosci.* **10**, 671–675.
- Snodgrass, R. E. 1938 Evolution of the Annelida, Onychophora, and Arthropoda. *Smithson. Misc. Coll.* **97**, 1–159.
- Spears, T. & Abele, L. G. 1998 Crustacean phylogeny inferred from 18S rDNA. In *Arthropod relationships* (ed. R. A. Fortey & R. H. Thomas), pp. 169–187. London: Chapman & Hall.
- Swofford, D. L. 1998 *PAUP**, beta v.4.0. Sunderland, MA: Sinauer Associates.
- Swofford, D. L., Olsen, G. J., Waddell, P. J. & Hillis, D. M. 1996 Phylogenetic inference. In *Molecular systematics*, 2nd edn (ed. D. M. Hillis, C. Moritz & B. K. Mable), pp. 407–514. Sunderland, MA: Sinauer Associates.
- Turbeville, J. M., Pfeifer, D. M., Field, K. G. & Raff, R. A. 1991 The phylogenetic status of arthropods, as inferred from 18S rRNA sequences. *Mol. Biol. Evol.* **8**, 669–686.
- Wägele, J. W. 1996 Arthropod evolution: on informative data and sound methodology. *Trends Ecol. Evol.* **11**, 252.
- Weygoldt, P. 1986 Arthropod interrelationships—the phylogenetic-systematic approach. *Z. Zool. Syst. Evol.* **24**, 19–35.
- Wheeler, W. C. 1998 Sampling, groundplans, total evidence and the systematics of arthropods. In *Arthropod relationships* (ed. R. A. Fortey & R. H. Thomas), pp. 87–96. London: Chapman & Hall.
- Wheeler, W. C., Cartwright, P. & Hayashi, C. Y. 1993 Arthropod phylogeny: a combined approach. *Cladistics* **9**, 1–39.
- Whittington, P. M. & Bacon, J. P. 1998 The organization and development of the arthropod ventral nerve cord: insights into arthropod relationships. In *Arthropod relationships* (ed. R. A. Fortey & R. H. Thomas), pp. 348–362. London: Chapman & Hall.
- Wills, M. A., Briggs, D. E. G. & Fortey, R. A. 1994 Disparity as an evolutionary index: a comparison of Cambrian and Recent arthropods. *Paleobiology* **20**, 93–130.
- Zharkikh, H. 1994 Estimation of evolutionary distances between nucleotide sequences. *J. Mol. Evol.* **39**, 315–329.

As this paper exceeds the maximum length normally permitted, the authors have agreed to contribute to production costs.