Mitochondrial genomes suggest that hexapods and crustaceans are mutually paraphyletic

Charles E. Cook^{1,*}, Qiaoyun Yue^{2,†} and Michael Akam¹

¹Department and Museum of Zoology, University of Cambridge, Downing Street, Cambridge CB2 3EJ, UK

²Pedobiologia, Shanghai Institute of Entomology, Chinese Academy of Sciences, 225 Chongqing(S), Shanghai 200025, People's Republic of China

For over a century the relationships between the four major groups of the phylum Arthropoda (Chelicerata, Crustacea, Hexapoda and Myriapoda) have been debated. Recent molecular evidence has confirmed a close relationship between the Crustacea and the Hexapoda, and has included the suggestion of a paraphyletic Hexapoda. To test this hypothesis we have sequenced the complete or near-complete mitochondrial genomes of three crustaceans (*Parhyale hawaiiensis*, *Squilla mantis* and *Triops longicaudatus*), two collembolans (*Onychiurus orientalis* and *Podura aquatica*) and the insect *Thermobia domestica*. We observed rearrangement of transfer RNA genes only in *O. orientalis*, *P. aquatica* and *P. hawaiiensis*. Of these, only the rearrangement in *O. orientalis*, an apparent autapomorphy for the collembolan family Onychiuridae, was phylogenetically informative.

We aligned the nuclear and amino acid sequences from the mitochondrial protein-encoding genes of these taxa with their homologues from other arthropod taxa for phylogenetic analysis. Our dataset contains many more Crustacea than previous molecular phylogenetic analyses of the arthropods. Neighbour-joining, maximum-likelihood and Bayesian posterior probabilities all suggest that crustaceans and hexapods are mutually paraphyletic. A crustacean clade of Malacostraca and Branchiopoda emerges as sister to the Insecta *sensu stricto* and the Collembola group with the maxillopod crustaceans. Some, but not all, analyses strongly support this mutual paraphyly but statistical tests do not reject the null hypotheses of a monophyletic Hexapoda or a monophyletic Crustacea. The dual monophyly of the Hexapoda and Crustacea has rarely been questioned in recent years but the idea of both groups' paraphyly dates back to the nineteenth century. We suggest that the mutual paraphyly of both groups should seriously be considered.

Keywords: Arthropoda; Hexapoda; mitochondria; genome; phylogeny; Pancrustacea

1. INTRODUCTION

The relationships among the four subphyla of the Arthropoda have been the subject of intense research for over a century. Until recently the prevailing opinion grouped myriapods and hexapods together as the Atelocerata (equal to Tracheata), grouped the Crustacea with the Atelocerata to form the Mandibulata and placed the chelicerates basal among the arthropods (Brusca & Brusca 2003). However, during the past decade molecular studies have consistently associated hexapods with the crustaceans to form a group now termed the Pancrustacea or Tetraconata (Friedrich & Tautz 1995; Boore *et al.* 1998; García-Machado *et al.* 1999; Wilson *et al.* 2000; Cook *et al.* 2001; Deutsch 2001; Dohle 2001; Hwang *et al.* 2001; Mallatt *et al.* 2004) and this view is now widely accepted, although some recent morphological studies still support the Atelocerata (Koch 2001; Kraus 2001; Bitsch & Bitsch 2004). Molecular studies have sometimes placed the hexapods within the crustaceans, often as the sister

group to the branchiopods or the malacostracans, but most suffer from limited sampling among crustacean lineages. A recent analysis of complete mitochondrial sequences strongly supported a division of the hexapods into two unrelated groups: one comprising the Collembola and the other the Insecta *sensu stricto* (i.e. the Microcorphya (bristletails), the Zygentoma (firebrats, silverfish and relatives) and the Pterygota (all of the winged insects; Nardi *et al.* 2003a), with the Collembola at the base of the Pancrustacea and sister to a crustacean and insect clade). However, this study attracted some methodological criticism (Delsuc *et al.* 2003; Nardi *et al.* 2003b) and included only two crustaceans. Since then an analysis using nuclear ribosomal RNA genes has supported a monophyletic Hexapoda within a paraphyletic Crustacea (Mallatt *et al.* 2004).

Morphological diversity among the crustaceans is greater than that of the other arthropod subphyla, perhaps greater than that of any other group of animals. The most recent formal classification divides the subphylum into six classes, Branchiopoda, Cephalocarida, Malacostraca, Maxillopoda, Ostracoda and Remipedia, containing 849 families and over 52 000 described species with many times that number surely not yet described (Martin & Davis 2001). Previous molecular studies of arthropod relationships have not adequately sampled from the

*Author and address for correspondence: Charles E. Cook, Natural Environment Research Council, British Antarctic Survey, Biological Sciences Division, High Cross, Madingley Road, Cambridge CB3 0ET, UK (cecook@alum.exeter.edu).

†Present address: School of Biological Sciences, University of East Anglia, Norwich NR4 7TJ, UK.

diversity of the Crustacea. We present here a phylogenetic analysis of the arthropods with a mitochondrial dataset that includes many more crustacean taxa than any previous study. Our analysis uses the complete set of mitochondrial protein-encoding genes. Similar approaches have been used successfully to address relationships among arthropods, mammals and birds (García-Machado *et al.* 1999; Arnason *et al.* 2002; Nardi *et al.* 2003a; Phillips & Penny 2003; Harrison *et al.* 2004; Negrisolo *et al.* 2004).

2. MATERIAL AND METHODS

(a) *Samples and sequencing*

Complete mitochondrial sequences were determined for the three pancrustacean species *Triops longicaudatus* (Leconte 1846) (Crustacea: Branchiopoda: Notostraca), GenBank AY639934; *Squilla mantis* (Linnaeus 1758) (Crustacea: Malacostraca: Hoplocarida), Genbank AY639936; and *Thermobia domestica* (Packard 1873) (Hexapoda: Zygentoma), GenBank AY639935. Almost-complete mitochondrial sequences were determined for the three additional pancrustaceans; *Parhyale hawaiiensis* (Dana 1853) (Crustacea: Malacostraca: Amphipoda), GenBank AY639937; *Onychiurus orientalis* Stach 1964 (Hexapoda: Collembola), GenBank AY639939; and *Podura aquatica* Lameere 1895 (Hexapoda: Collembola), GenBank AY639938. The collembolan sequences include all 13 protein-encoding mitochondrial genes. The *P. hawaiiensis* sequence includes 12 out of 13 protein-encoding genes—only a partial fragment of the *nad1* gene was recovered despite repeated attempts to amplify and clone this fragment.

Triops longicaudatus were hatched and grown to adulthood from a 'Triops World' kit purchased from Netfysh (<http://www.netfysh.com>, Haxby, York, UK). The eggs contained in the kit were supplied from North America but their precise provenance is unavailable. A single *S. mantis* was purchased from the fish market at Iraklio, Crete by M. Averof, frozen and then shipped to Cambridge. Two individuals of *P. hawaiiensis* were obtained from a laboratory culture maintained by C. Extavour in the Museum of Zoology, Cambridge. The museum culture was founded with individuals from the culture described by Gerberding *et al.* (2002). The collembolans *O. orientalis* and *P. aquatica* were collected from the Shanghai Botanic Garden and identified by Qiaoyun Yue. Three individuals of *T. domestica* were obtained from a laboratory culture in the Museum of Zoology, Cambridge.

DNA was extracted from all samples using the Qiagen genomic DNA buffer set and Qiagen genomic-tip 20 filters. For *T. longicaudatus*, *S. mantis*, *T. domestica* and *P. hawaiiensis* DNA was extracted from single individuals. However, for all but *S. mantis* we used DNA from two (*Ph*) or three (*Tl*, *Td*) individuals to generate the sequences that we report herein. For *O. orientalis* and *P. aquatica* we extracted DNA from pools of 10–20 individuals collected from a single location.

For each taxon we amplified short 300–800 bp segments of the mitochondrial genome using degenerate primers designed to match specific arthropod mitochondrial genes (primer sequences available on request). Amplification parameters were generally not stringent (50–55 °C annealing temperatures), although additional amplifications with higher annealing temperatures of up to 65 °C were attempted when multiple PCR products were observed. These fragments were cloned into the T-Easy vector (Promega) and at least three

clones sequenced for each using the ABI 'Big Dye' technology. Species-specific primers were then designed from these short fragments and used to amplify longer 2–6 kb segments of the mitochondrial genome from each taxon (Roche Expand long template system). Long PCR products were cloned into the T-Easy vector and at least three clones from each were picked. Each clone was sequenced using a primer-walking strategy initiated with vector-specific primers. Sequences were assembled using SEQUENCHER (v. 3.1; GeneCodes Corp.). Sequences were annotated by comparison with other published mitochondrial sequences. Some transfer RNA (tRNA) genes were identified using the tRNAscan server at <http://www.genetics.wustl.edu/eddy/tRNAscan-SE/> (Lowe & Eddy 1997). Other tRNA genes were identified by comparison with published arthropod tRNA gene sequences.

(b) *Phylogenetic analysis*

Complete mitochondrial genomes from 48 arthropod species were available when this analysis commenced (Electronic Appendix, table A1). We extracted from each of our six new sequences, and from the publicly available sequences, the nucleotide and putative amino acid regions for each of the 13 mitochondrial protein-encoding genes and then created an alignment for each gene separately. We performed multiple alignments with CLUSTALW (Chenna *et al.* 2003) using three combinations of opening/extension gap penalties: 10/0.2 (default), 12/4 and 5/1. Conserved regions were identified for each alignment using the program GBLOCKS (Castresana 2000) with block parameters set as follows: the minimum number of sequences for conserved and flank positions was 38 (out of 49), the maximum number of contiguous non-conserved positions was eight (default), the minimum length of a block was 8 and the number of allowed gap positions was zero. We then compared the GBLOCKS-identified conserved regions for all three alignments, identified regions that were identical in all three alignments and concatenated these blocks to form an alignment for that gene. This rigorous alignment and block identification procedure ensured that the regions included in the alignment were reliably homologous. The *atp8* gene is very short (*ca.* 50 amino acid residues) and not well conserved so it was not used for further analyses. For each of the remaining 12 protein-encoding genes we then used CODONALIGN (<http://www.sinauer.com/hall/>) to create a corresponding alignment of the nucleotides that encoded those amino acids. We then concatenated all 12 alignments for both the amino acids and the nucleotides to create two datasets for phylogenetic analysis. The original 12 datasets contained 3997 amino acid residues with 54 taxa, of which we retained 2429 (7287 nucleotides) or 61% after elimination of poorly aligned regions (see Electronic Appendix, table A2). We note that we have constructed this alignment for the purpose of estimating phylogenetic relationships between the major Pancrustacea taxa and this necessitated eliminating sequence regions which may be useful in analysis at lower taxonomic levels. Thus, an alignment constructed only using sequences from the Insecta, or from the Malacostraca, would include more nucleotides and might provide a stronger phylogenetic signal within those taxa than the alignment we use here.

Among the 54 taxa in our dataset were seven flies (Diptera), six butterflies and moths (Lepidoptera), three beetles (Coleoptera) and four decapod crustaceans (Decapoda). Preliminary phylogenetic analyses using

maximum-likelihood (ML) distance matrices in PAUP* (Swofford 1998) and quartet puzzling (TREE-PUZZLE; Strimmer & von Haeseler 1996) clearly showed that all four of these groups were always monophyletic in every analysis. To reduce computational time for later analyses we eliminated all but one each of the Diptera, Lepidoptera and Coleoptera and all but two of the Decapoda. In addition, we observed in the nucleotide and amino acid alignments that nine taxa, the honeybee (*Apis mellifera*), the Hymenoptera (*Melipona bicolor*), the wallaby louse (*Heterodoxus macropus*), *Thrips imaginis* and five ticks (*Ixodes hexagonus*, *Ixodes persulcatus*, *Ornithodoros moubata*, *Rhipicephalus sanguineus* and *Varroa destructor*), were missing some motifs in many genes that were clearly common to all other taxa. These regions were sufficiently divergent in these taxa that we could not be confident of sequence homology between the divergent taxa and all other taxa. Additionally, in previous phylogenetic analyses these nine taxa have all been reported to have long branches and to have unstable phylogenetic positions (Crozier & Crozier 1993; Shao *et al.* 2001; Nardi *et al.* 2003a; Shao & Barker 2003; Negrisoló *et al.* 2004). We therefore excluded these taxa from the analysis. This left 30 taxa, including one chelicerate, three myriapods, four collembolans, 14 crustaceans and eight insects, for which we had one amino acid and one nucleotide dataset.

Composition bias and mutation saturation, particularly in third codon positions, are commonly observed for arthropod mitochondrial datasets that include taxa with deep divergence times (Shao *et al.* 2001; Nardi *et al.* 2003a; Shao & Barker 2003; Negrisoló *et al.* 2004). Recent work on mammalian and avian mitochondrial genomes has shown that recoding of the third positions of codons from ACGT to purines (R) and pyrimidines (Y) can recover some signal from saturated or biased third position data (Phillips & Penny 2003; Harrison *et al.* 2004). Delsuc *et al.* (2003) specifically advocate this approach for arthropod mitochondrial sequences. Accordingly, we also prepared two additional nucleotide datasets in which either third positions or first and third positions were recoded as purines and pyrimidines.

We tested saturation at all codon positions (both ACGT and RY encoded for first and third positions) by graphing ML distances against simple *p*-distances for first, second and third positions (Electronic Appendix, figure A1). We observed, as did Negrisoló *et al.* (2004), almost complete saturation in third positions. We also used TREE-PUZZLE to perform likelihood mapping tests for first, second and third positions separately as well as for various combined datasets. We found, again as did Negrisoló *et al.* (2004), that first and second positions contained substantial phylogenetic signal but that third positions did not (Electronic Appendix, figure A2).

We then examined the amino acid dataset and the various nucleotide datasets for evidence that composition bias might be affecting the analytical result. Hassanin *et al.* (2005) note that inversions of the mitochondrial control region can affect composition bias in genes that are not affected by the rearrangement and that taxa affected in this way may be attracted to each other in a phylogenetic analysis. In our dataset only two taxa, *Hutchinsoniella macracantha* and *Tigriopus japonicus*, have detectable inversions in the control region and these two taxa are not close together in our trees, so we conclude that control region inversion is not affecting our results. We also tallied which taxa failed the *X*-squared composition bias test implemented in TREE-PUZZLE (Strimmer & von Haeseler 1996, 1997) in the various

datasets (Electronic Appendix, table A3). Finally, for each dataset we calculated a distance matrix using nucleotide compositions as the basis for the calculation (i.e. sequences with similar composition biases are assigned smaller 'distances') and then constructed a neighbour-joining tree from this distance matrix (as implemented at <http://www.nhm.ac.uk/zoology/external/p4.htm>; trees not shown; Lockhart *et al.* 1994). For each tree we then noted which taxa had passed and which had failed the TREE-PUZZLE compositional bias test and examined the trees to see if taxa that failed the test were grouping together. If there is composition bias then these same taxa will also group in standard phylogenetic analyses. We observed no systematic relationship between taxa that failed the composition bias test, their positions on the composition-based distance trees or their positions on the trees resulting from standard phylogenetic analyses, so we conclude that composition bias, while present, is not systematically affecting the phylogenetic analyses of our data.

We initially used a nucleotide dataset that only included first and second positions for ML analyses based upon these results. However, the addition of RY-encoded third positions did slightly increase the phylogenetic signal in the likelihood mapping analysis, while also decreasing the number of taxa that failed the nucleotide composition-bias test in TREE-PUZZLE, so we therefore repeated the analysis on a dataset that included RY-encoded third positions. We did not do further analyses using RY-encoded first positions, which clearly reduced the phylogenetic signal, or on the ACGT-encoded third positions as these were clearly saturated.

Phylogenies were estimated as follows. First, we used the neighbour-joining method as implemented in PAUP* (v. 4.0b10) (Swofford 1998) with the default ML distance parameters to generate a starting tree. We then used a likelihood ratio test as implemented in MODELTEST (Posada & Crandall 1998) and identified a general time-reversible model (GTR+I+ Γ) as optimal for each dataset. Using this model, and the same tree, we estimated the likelihood when the number of rate categories varied between 1 and 10, and used a χ^2 -test to determine when increasing the number of categories ceased to significantly improve the ML estimate. Eight rate categories were optimal for both datasets.

To find the best estimate of phylogeny for any given dataset we would like to calculate the ML value for every possible tree. For our datasets with 30 taxa an exhaustive search of tree space was not possible due to computational limits. To optimize our estimate of the phylogeny for each dataset we used the neighbour-joining tree as the starting point for an iterative search using PAUP* (Collins & Wiegmann 2002a,b; Telford *et al.* 2003). First, we estimated the variables in the GTR+I+ Γ_8 model (8 rate categories, gamma shape parameter, proportion of invariant sites, GTR substitution-rate matrix and nucleotide frequencies) for the neighbour-joining tree and then used these variables in a ML search using the nearest neighbour interchange branch swapping algorithm to identify trees near the starting tree with 'better' estimated ML values. Second, we re-estimated the variables in the GTR+I+ Γ_8 model (gamma shape parameter, proportion of invariant sites, GTR substitution-rate matrix and nucleotide frequencies) for the tree found by the nearest neighbour interchange search. We then used these values in another nearest neighbour interchange search of tree space and re-estimated the parameters for the new tree. When the likelihood value stabilized we repeated the search in tree space using the tree bisection and reconnection

ancestral pancrustacean (including *Squilla mantis*, *Thermobia domestica* and *Triops longicaudatus*)

Figure 1. Comparison of mitochondrial gene orders for the taxa shown. The ancestral pancrustacean gene order was also found for *Squilla mantis*, *Thermobia domestica* and *Triops longicaudatus*. The style of the figure is adapted from fig. 1 of Lavrov *et al.* (2004). Genes are not drawn to scale. Protein genes (large boxes) are abbreviated as in the text and tRNA genes are abbreviated using single-letter codes. The striped area on the left of the ancestral sequence represents a large non-coding region. Genes are transcribed from left to right except when underlined. Shaded boxes indicate genes whose positions have changed when compared with the ancestral pancrustacean gene order. Arrows with subscript letters indicate positions where a tRNA gene (identified by the letter) is located in the ancestral pancrustacean sequence but not in the sequence shown. All of the gene order changes we report involve tRNA genes; there were no translocations of protein-encoding genes. Note that the sequences of *Onychiurus orientalis*, *Podura aquatica* and *Parhyale hawaiiensis* are incomplete so the positions of some genes in these taxa are still unknown.

branch-swapping algorithm, recalculated the parameters for the tree bisection and reconnection tree and repeated nearest neighbour interchange searches, parameter optimizations and tree bisection and reconnection searches until the likelihood score was stationary.

In no trees were either the crustaceans or hexapods (Collembola and Insecta *sensu stricto*) monophyletic. To test the robustness of these results we found the best tree for the dataset using the iterative method described above under three constraints: that the Hexapoda are monophyletic, that the Crustacea are monophyletic and that both hexapods and crustaceans are monophyletic. We then performed a Shimodaira–Hasegawa test, using resampling estimated by log-likelihood optimization with 1000 bootstrap replicates as implemented in PAUP*, to compare the likelihoods of the best trees found under these null hypotheses with the overall best tree for each dataset (Shimodaira & Hasegawa 1999; Goldman *et al.* 2000).

We also used MRBAYES v. 3.0 (Huelsenbeck & Ronquist 2001) to estimate phylogenies by Bayesian inference for each dataset. We used a GTR+I+ Γ_8 model with first, second and third positions assigned separate partitions. For the dataset with RY-encoded third positions the number of substitution types was set at one rather than six (i.e. there are only two mutation rates, R to Y and Y to R, rather than six A to G, G to A, etc.).

Four chains were run with 200 000 generations, and the ML estimate and topology of every 100th tree was stored. A graph of the likelihood values showed that these reached a plateau after approximately 40 000 generations or 20% into the run. We used the last 1000 of the stored trees (the final 50%) to calculate consensus trees for each analysis. The frequency with which each branch of the tree is represented on the consensus tree represents a posterior probability of the likelihood of that branch. Note that such values, when calculated using Bayesian inference, are reported to over-estimate branch support in the tree (Simmons *et al.* 2004).

We also performed phylogenetic analyses on the amino acid dataset that corresponded to the coding sequences of the nucleotide datasets. For this analysis we used MRBAYES to estimate phylogenies. We used the WAG (Whelan & Goldman 2001) model of amino acid substitution, also with eight gamma categories. Four chains were run for 100 000 generations and the ML estimate and topology of every 50th tree was stored. A graph of likelihood values showed that these reached a plateau after 20 000 generations or 20% into the run. We used the last 1000 of the stored trees (the final 50%) to calculate a consensus tree for this analysis.

3. RESULTS

(a) Gene order

In a series of papers, Boore, Brown and Lavrov (Boore *et al.* 1995, 1998; Lavrov *et al.* 2002, 2004) have shown that rearrangements in the mitochondrial gene order are useful as characters in studying hexapod and crustacean relationships. Gene orders for the newly sequenced taxa are shown in figure 1. The mitochondrial gene order in *S. mantis*, *T. domestica* and *T. longicaudatus* is identical to the presumed ancestral pancrustacean gene order (Lavrov *et al.* 2004). Of the sequence that we determined for *O. orientalis* tRNA-serine(UCN) (*trnS2*; hereafter all tRNA genes abbreviated using single letter codes) differs from the ancestral gene order in that it is translocated from between *cob* and *nad4* (gene abbreviations as in Lavrov *et al.* 2004) to the position occupied by *trnQ* between *trnI* and *trnM* in the ancestral gene arrangement. This same translocation was observed in another collembolan, *Tetradontophora bielaniensis* (Nardi *et al.* 2001). Both species are assigned to the family Onychiuridae within the Collembola, so this character is (given the available data) autapomorphic for the family. The gene order in *P. aquatica* differs from the ancestral gene order in the region between *nad2* and *cox1*. In the ancestral pancrustacean this region contains three tRNA genes with the gene order *nad2-trnW-trnC-trnY-cox1*

Figure 2. Estimate of the maximum likelihood of crustacean and hexapod phylogeny from mitochondrial protein-encoding nucleotide sequences of 30 taxa. This tree is derived from a dataset in which third positions were recoded as purines (R) and pyrimidines (Y). An analysis using a dataset with only first and second positions gave a topologically identical tree with very similar branch lengths. The chelicerate *Limulus polyphemus* was assigned as the outgroup. The consensus tree from the Bayesian analysis of the third RY dataset was topologically identical to this tree. The consensus tree from the Bayesian analysis of the dataset with first and second positions only was identical, except for the placement of *Speleonectes tulumensis*, which was grouped with *Vargula hilgendorffii*. The first two numbers at a node represent Bayesian posterior support probabilities for that node for the RY encoded dataset and for the first and second positions only dataset. The third and fourth numbers at the node indicate support for that node in a non-parametric bootstrap analysis of neighbour-joining trees calculated from maximum-likelihood distances for the same two datasets. Slashes with no number to the right indicate lack of support for that branch in an analysis.

(italics indicate that the gene is encoded on the opposite strand). In *P. aquatica* the gene order is *nad2-trnC-trnW*; *trnY* is not present while *trnC* and *trnW* have exchanged positions. Similar changes in the positions of one or two tRNA genes have been reported for a number of other pancrustacean taxa (Lavrov *et al.* 2004). Of the new sequences we present here, that of *P. hawaiiensis* differs most from the ancestral sequence with at least 12 tRNA translocations. Of the gene order changes, we observe that only the sequence of *O. orientalis*, supporting the collembolan family Onychiuridae, is phylogenetically informative.

(b) Phylogeny

Figure 2 shows a phylogeny estimated by ML for two nucleotide datasets (one of first and second positions only, one for first, second and RY-encoded third positions) with 30 arthropod taxa including four chelicerates and myriapods, four collembolans, eight insects and 14 crustaceans. This tree displays certain anomalies that lead us to exclude some taxa from a subsequent analysis. However, this full dataset agrees with previous studies in providing strong support for the Pancrustacea as a

monophyletic clade that excludes myriapods. Bayesian support values are quite high for a clade that places the Branchiopoda as sister to a Malacostraca and Cephalocarida clade, and this group as sister to the Insecta, but the Branchiopoda, Malacostraca and Cephalocarida node was not strongly supported in a non-parametric bootstrap analysis by neighbour-joining. Figure 2 shows strong Bayesian support for an (Insecta, (Branchiopoda, Malacostraca)) clade distinct from a (Collembola, some Maxillopoda) clade. There is strong Bayesian support for the four Collembola as sister to a group of crustaceans that includes many maxillopods and the malacostracan *P. hawaiiensis*, but this branch is also not supported by non-parametric bootstrapping. Within the Collembola there is strong support for the clade containing two species assigned to the family Onychiuridae, *O. orientalis* and *T. bielensis*. This provides molecular support (although with only two taxa) for that family and is concordant with the gene rearrangement shared by these two taxa.

The analysis of this 30 taxa dataset supports the separation of the Hexapoda into a clade that includes the Collembola and a second clade comprising the Insecta *sensu stricto*. It also supports paraphyly of the Crustacea.

Figure 3. Estimate of the maximum likelihood of crustacean and hexapod phylogeny from mitochondrial protein-encoding nucleotide sequences of 25 taxa. This tree is derived from a dataset in which third positions were recoded as purines (R) and pyrimidines (Y). An analysis using a dataset with only first and second positions gave a topologically identical tree with very similar branch lengths. The chelicerate *Limulus polyphemus* was assigned as the outgroup. The consensus tree from the Bayesian analysis of the third RY dataset was topologically identical to this tree. The consensus tree for the Bayesian analysis of the dataset with first and second positions only was identical, except that *Vargula hilgendorffii* grouped with the maxillopods with weak support (66%) and the clade of maxillopods, the clade of insects, branchiopods and malacostracans and the clade of collembolans were an unresolved trichotomy. Numbers refer to branch support as Bayesian posterior probabilities and non-parametric bootstrap analyses, as in figure 1.

However, there are some anomalous relationships in the tree that have strong statistical support. For instance, the monophyly of the Malacostraca is well supported by morphological studies (Martin & Davis 2001), yet in figure 2 the malacostracan *P. hawaiiensis* is within a clade of Maxillopoda. In addition, figure 2 places the cirripede *Pollicipes polymerus* at the base of the Pancrustacea, though support for the placement of cirripedes in the Maxillopoda is strong (Martin & Davis 2001). The tree also places the cephalocarid *H. macracanthus* in a position sister to the Malacostraca. This position is in accordance with the classification of Hessler (Hessler 1992) who associated Branchiopoda, Cephalocarida and Malacostraca based upon morphological characters, but is contrary to evidence from mitochondrial gene order that associates *H. macracantha* with the remipede *Armillifer armillatus* and the maxillopod *Argulus americanus*. Such anomalies are often observed in molecular phylogenies of the arthropods, and have been attributed to unequal base composition, to uneven rates of evolution and to 'long branch attraction' (Nardi *et al.* 2003b; Felsenstein 2004).

Within the Insecta the analysis fails to recover some relationships that are otherwise well supported; for instance, the beetle *T. castaneum* is more basal in the tree than is the grasshopper *Locusta migratoria*. We noted above

that this dataset is not optimal for analysis at 'lower' levels in the Pancrustacea so we will not further examine the relationships within the Insecta in this paper, and suggest that such an analysis should be done using a dataset that includes more insect taxa and has been constructed using only insect sequences to maximize the phylogenetic signal retained in the alignment.

As a further test of these relationships we used the amino acid dataset that exactly correlates with the nucleotide dataset used above for a separate phylogenetic analysis (Electronic Appendix, figure A3). In this analysis there was also strong support for an (Insecta, (Branchiopoda, Malacostraca)) clade as separate from a (Collembola, some Maxillopoda) clade. However, placement of four taxa, *Artemia franciscana*, *H. macracantha*, *P. hawaiiensis* and *Speleonectes tulumensis* was different from their placement in the nucleotide analysis shown in figure 2. To eliminate uncertainty in the analysis due to these anomalous taxa we excluded those four, as well as the cirripede *P. polymerus*, from the dataset and repeated the analyses of both nucleotide datasets and of the amino acid dataset with the remaining 25 taxa.

Analyses of the 25 taxon datasets (figures 3 and 4) mirror those for the larger dataset. The Malacostraca and Branchiopoda are again placed as sister groups which are

Figure 4. Estimate of arthropod phylogeny using an amino acid dataset with 25 taxa by Bayesian inference. Numbers represent posterior probabilities.

associated with the Insecta. These relationships are strongly supported by Bayesian posterior probabilities, but not by non-parametric bootstrapping using ML distances and neighbour-joining. The Collembola are again shown as the sister group to a lineage of maxillopod crustaceans but with strong support in only one Bayesian analysis. The maxillopod *Vargula hilgendorfii* (Ostracoda) is at the base of the Pancrustacea in the analyses of nucleotide datasets (figure 3), but there is no statistical support for this placement. In the analysis of the amino acid dataset, *V. hilgendorfii* groups with the other three maxillopod taxa and there is strong Bayesian support for this relationship.

All the analyses of the 25 and 30 taxon datasets suggested that the Collembola form a clade with a group of Maxillopoda (although membership in this group varies), that the Malacostraca and Branchiopoda form a clade and that the Insecta group with the Malacostraca and Branchiopoda. However, the lack of consistent statistical support for all of the key branches in each topology, and the variable placement of some taxa, led us to implement additional statistical testing. We used Shimodaira–Hasegawa tests to further evaluate our results. We found the likelihood of the best tree conforming to that hypothesis and compared this likelihood to that of the best overall tree for each of the four nucleotide datasets and for each of three null hypotheses. The three null hypotheses were (i) monophyly of the Hexapoda, (ii) monophyly of the Crustacea and (iii) monophyly of both Crustacea and Hexapoda (table 1; Shimodaira & Hasegawa 1999; Goldman *et al.* 2000). Our implementation of the Shimodaira–Hasegawa test uses only one topology for each constraint, whereas a more accurate

implementation would use every tree topology compatible with each constraint. As a result, our significance levels are overestimates of the true significance for each constraint. With this caveat, we note that the probability values shown in table 1 range from 0.035 to 0.692, with many being at 0.2 or higher. These values are not particularly low and we would not wish to reject any of the alternative hypotheses based upon these results.

We note that in all of our analyses the two millipedes *Narceus annularis* and *Thyropygus* sp. group together but they do not group with the other myriapod, the centipede *Lithobius forficatus*. Instead, we recover a paraphyletic Myriapoda with the chelicerate *Limulus polyphemus* sometimes grouping with the two millipedes and sometimes with the centipede. We have not further explored the chelicerate and myriapod relationship because our dataset, which does not include a newly available centipede sequence (Negrisolo *et al.* 2004) or an outgroup to the arthropods, is clearly inappropriate for addressing this question. Negrisolo *et al.* (2004) have performed just such an analysis and also recovered a paraphyletic Myriapoda.

4. DISCUSSION

The results of our phylogenetic analyses consistently suggest that the Crustacea and the Hexapoda are paraphyletic with respect to each other. We observe a clade of (Insecta, (Branchiopoda, Malacostraca)) and a second clade of (Collembola, some Maxillopoda). This result is consistent with those of Nardi *et al.* (2003a) who recovered a clade with one malacostracan and one branchiopod as sister to the Insecta, with that clade as sister to the Collembola. There were no other Crustacea

Table 1. Results of Shimodaira–Hasegawa tests of three null hypotheses for each of four nucleotide datasets.

dataset	tree	–ln <i>L</i>	diff. –ln <i>L</i>	probability
30taxa1st2nd	best tree	78 386.073 51	—	—
	hexapods monophyletic	78 411.441 65	25.368 14	0.221
	crustaceans monophyletic	78 425.039 04	38.965 53	0.059
	crustaceans and hexapods monophyletic	78 442.354 95	56.281 44	0.035
30taxa3rdYR	best tree	104 175.124 91	—	—
	hexapods monophyletic	104 205.577 39	30.452 47	0.212
	crustaceans monophyletic	104 189.532 24	14.407 33	0.522
	crustaceans and hexapods monophyletic	104 189.532 24	39.109 93	0.126
25taxa1st2nd	best tree	132 672.466 52	—	—
	hexapods monophyletic	132 692.372 13	19.905 61	0.330
	crustaceans monophyletic	132 691.397 28	18.930 76	0.337
	crustaceans and hexapods monophyletic	132 698.113 50	25.646 98	0.231
25taxa3rdYR	best tree	86 545.911 54	—	—
	hexapods monophyletic	86 571.806 72	25.895 18	0.090
	crustaceans monophyletic	86 564.973 75	19.062 21	0.193
	crustaceans and hexapods monophyletic	86 564.973 75	19.062 21	0.193

in their analysis. It is also consistent with the results of Lavrov *et al.* (2004) who recover an (Insecta, (Branchiopoda, Malacostraca)) clade and a (Collembola, Maxillopoda) clade, although with no statistical support for the latter. Both of these studies also used mitochondrial datasets.

Our results are also broadly similar to those of Mallatt *et al.* (2004) who used mitochondrial ribosomal gene sequences (*rrnS* and *rrnL*) to examine ecdysozoan phylogeny. However, they do show some significant differences. Mallatt *et al.* present a number of trees derived from Bayesian and distance-based analyses. They recover a monophyletic Pancrustacea and paraphyletic Crustacea in all trees, as do we. They also consistently find that the Maxillopoda, represented by three taxa in their dataset, are not monophyletic, again in agreement with our findings. However, their analyses usually group their single collembolan with other hexapods, making hexapods monophyletic with high support values coming particularly from the *rrnL* sequences. In some of their analyses the branchiopods are closer to the hexapods than to the Malacostraca but support for this grouping is not always high. We note that they included only one collembolan and three maxillopods and that the relationship between hexapods and the major crustacean groups was different in their various analyses.

Our analyses have addressed the methodological criticisms aimed at the study of Nardi *et al.* (2003a). They have also broadened sampling of the Collembola and the Crustacea. Analysis of Bayesian posterior probabilities consistently supports the paraphyly of both the Hexapoda and the Crustacea. Therefore, we take this hypothesis seriously. However, it is known that Bayesian posterior probabilities overestimate branch support (Simmons *et al.* 2004). We do not recover strong support for hexapod or crustacean paraphyly using a neighbour-joining approach and we cannot reject alternative hypotheses of hexapod and crustacean monophyly using Shimodaira–Hasegawa testing.

Although the monophyly of the Hexapoda has rarely been questioned by morphological systematists in recent years, this was not always the case. Hansen (1893) noted similarities between the mandibles of a number of pterygote insects and the Malacostraca. He also noted

that the mandibular musculature of the endognathous *Campodea*, *Japyx* and the Collembola are more similar to those of the Crustacea than those of the ectognathous insects. Tillyard (1930) also suggested that there are two distinct and unrelated hexapod groups, the Collembola/Protura (i.e. endognaths) and the Thysanura/Pterygota (i.e. ectognaths) that both descended independently from within the Crustacea. However, he also supported a sister relationship between the Myriapoda and the ectognathous insects.

There is no intrinsic difficulty in obtaining more molecular data to resolve this phylogenetic question; data can be obtained, for example, from multiple nuclear protein-coding genes. It therefore seems wise to await additional molecular data before reopening this old debate.

We can be somewhat more confident that the Branchiopoda and the Malacostraca form a clade to the exclusion of at least some of the maxillopodans and that the insects (*sensu stricto*) are related to this clade. Unfortunately there is no consensus as to whether they may be a sister group to the entire clade or derived from within it. This particular question—of primary interest to us when we started this study—remains unresolved.

In our analyses of mitochondrial nucleotide and amino acid datasets a number of crustaceans, including all of the Maxillopoda, are not stable. The group of long-branched crustacean taxa is noteworthy for exhibiting a large number of rearrangements in their mitochondrial genomes. Thus, out of the pancrustacean taxa shown in figure 2, the 18 taxa of Malacostraca (except *P. hawaiiensis*), Branchiopoda, Insecta and Collembola all have relatively short branch lengths and all but one collembolan have the ancestral mitochondrial gene order or a small number of tRNA translocations. By contrast the remaining eight long-branched crustaceans, including the malacostracan *P. hawaiiensis*, all have highly rearranged mitochondrial genomes. We suggest that there may be a relationship between increased sequence mutation rates and an increased likelihood of gene order rearrangements that bears further investigation.

We note that we never recover a clade containing all of the Maxillopoda in our dataset and this may be related to the computational difficulty of correctly placing these

fast-evolving sequences. However, our analysis of composition bias, which may cause long branches, suggests that such bias is not affecting our results. In addition, the relationship between the two longest branch taxa, *A. americanus* and *A. armillatus*, is consistent in all of our analyses and is supported by mitochondrial gene order data (Lavrov *et al.* 2004). As the monophyly of the Maxillopoda is not well supported by morphological or other molecular analyses (Martin & Davis 2001; Mallatt *et al.* 2004), we think it likely that the Maxillopoda are not monophyletic.

This work was supported by BBSRC grant 8/G14526. Q. Y. was supported by The Royal Society and by the National Science Foundation of China grants 30130040 and 30370169.

REFERENCES

- Arnason, U. *et al.* 2002 Mammalian mitogenomic relationships and the root of the eutherian tree. *Proc. Natl Acad. Sci. USA* **99**, 8151–8156.
- Bitsch, C. & Bitsch, J. 2004 Phylogenetic relationships of basal hexapods among the mandibulate arthropods: a cladistic analysis based on comparative morphological characters. *Zool. Scr.* **33**, 511–550.
- Boore, J. L., Collins, T. M., Stanton, D., Daehler, L. L. & Brown, W. M. 1995 Deducing the pattern of arthropod phylogeny from mitochondrial DNA rearrangements. *Nature* **376**, 163–165.
- Boore, J. L., Lavrov, D. V. & Brown, W. M. 1998 Gene translocation links insects and crustaceans. *Nature* **392**, 667–668.
- Brusca, R. C. & Brusca, G. J. 2003 *Invertebrates*. Sunderland, MA: Sinauer Associates.
- Castresana, J. 2000 Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* **17**, 540–552.
- Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T. J., Higgins, D. G. & Thompson, J. D. 2003 Multiple sequence alignment with the CLUSTAL series of programs. *Nucleic Acids Res.* **31**, 3497–3500.
- Collins, K. P. & Wiegmann, B. M. 2002a Phylogenetic relationships and placement of the Empidoidea (Diptera: Brachycera) based on 28S rDNA and EF-1 α sequences. *Insect Syst. Evol.* **33**, 421–444.
- Collins, K. P. & Wiegmann, B. M. 2002b Phylogenetic relationships of the lower Cyclorrhapha (Diptera: Brachycera) based on 28S rDNA sequences. *Insect Syst. Evol.* **33**, 445–456.
- Cook, C. E., Smith, M. L., Telford, M. J., Bastianello, A. & Akam, M. 2001 Hox genes and the phylogeny of the arthropods. *Curr. Biol.* **11**, 759–763.
- Crozier, R. H. & Crozier, Y. C. 1993 The mitochondrial genome of the honeybee *Apis mellifera*: complete sequence and genome organization. *Genetics* **133**, 97.
- Delsuc, F., Phillips, M. & Penny, D. 2003 Comment on 'hexapod origins: monophyletic or paraphyletic'. *Science* **301**, 1482d.
- Deutsch, J. 2001 Are Hexapoda members of the Crustacea? Evidence from comparative developmental genetics. *Annales de la Société Entomologique de France* **37**, 41–49.
- Dohle, W. 2001 Are the insects terrestrial crustaceans? A discussion of some new facts and arguments and the proposal of the proper name 'tetraconata' for the monophyletic unit Crustacea + Hexapoda. *Annales de la Société Entomologique de France* **37**, 85–104.
- Felsenstein, J. 2004 *Inferring phylogenies*. Sunderland, MA: Sinauer Associates, Inc.
- Friedrich, M. & Tautz, D. 1995 Ribosomal DNA phylogeny of the major extant arthropod classes and the evolution of myriapods. *Nature* **376**, 165–167.
- Garcia-Machado, E., Pempera, M., Dennebouy, N., Oliva-Saurez, M., Mounolou, J.-C. & Monnerot, M. 1999 Mitochondrial genes collectively suggest the paraphyly of Crustacea with respect to Insecta. *J. Mol. Evol.* **49**, 142–149.
- Gerberding, M., Browne, W. E. & Patel, N. H. 2002 Cell lineage analysis of the amphipod crustacean *Parhyale hawaiiensis* reveals an early restriction of cell fates. *Development* **129**, 5789–5801.
- Goldman, N., Anderson, J. P. & Rodrigo, A. G. 2000 Likelihood-based test of topologies in phylogenetics. *Syst. Biol.* **49**, 652–670.
- Hansen, H. J. 1893 A contribution to the morphology of the limbs and mouthparts of crustaceans and insects. *Ann. Mag. Nat. Hist. Ser. 6* **7**, 417–434.
- Harrison, G. L. A., McLenachan, P. A., Phillips, M. J., Slack, K. E., Cooper, A. & Penny, D. 2004 Four new avian mitochondrial genomes help get to basic evolutionary questions in the late Cretaceous. *Mol. Biol. Evol.* **21**, 974–983.
- Hassanin, A., Léger, N. & Deutsch, J. 2005 Evidence for multiple reversals of asymmetric mutational constraints during the evolution of the mitochondrial genome of Metazoa, and consequences for phylogenetic inferences. *Syst. Biol.* **54**, 277–298.
- Hessler, R. R. 1992 Reflections on the phylogenetic position of the Cephalocarida. *Acta Zool.* **73**, 315–316.
- Huelsenbeck, J. P. & Ronquist, F. 2001 MrBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**, 754–755.
- Hwang, U. W., Park, C. J., Yong, T. S. & Kim, W. 2001 One-step PCR amplification of complete arthropod mitochondrial genomes. *Mol. Phylogenet. Evol.* **19**, 345–352.
- Koch, M. 2001 Mandibular mechanisms and the evolution of hexapods. *Annales de la Société Entomologique de France* **37**, 129–174.
- Kraus, O. 2001 'Myriapoda' and the ancestry of the Hexapoda. *Annales de la Société Entomologique de France* **37**, 105–128.
- Lavrov, D. V., Boore, J. L. & Brown, W. M. 2002 Complete mtDNA sequences of two millipedes suggest a new model for mitochondrial gene rearrangements: duplication and nonrandom loss. *Mol. Biol. Evol.* **19**, 163–169.
- Lavrov, D. V., Brown, W. M. & Boore, J. L. 2004 Phylogenetic position of the Pentastomida and (pan)crustacean relationships. *Proc. R. Soc. B* **271**, 537–544.
- Lockhart, P. J., Steel, M. A., Hendy, M. D. & Penny, D. 1994 Recovering evolutionary trees under a more realistic model of sequence evolution. *Mol. Biol. Evol.* **11**, 605–612.
- Lowe, T. M. & Eddy, S. R. 1997 tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964.
- Mallatt, J. M., Garey, J. R. & Shultz, J. W. 2004 Ecdysozoan phylogeny and Bayesian inference: first use of nearly complete 28S and 18S rRNA gene sequences to classify the arthropods and their kin. *Mol. Phylogenet. Evol.* **31**, 178–191.
- Martin, J. W. & Davis, G. E. 2001 *An updated classification of the recent Crustacea*. Science Series No. 39. Los Angeles, CA: Natural History Museum of Los Angeles County.
- Nardi, F., Carapelli, A., Fanciulli, P. P., Dallai, R. & Frati, F. 2001 The complete mitochondrial DNA sequence of the basal hexapod *Tetradontophora bielensis*: evidence for heteroplasmy and tRNA translocations. *Mol. Biol. Evol.* **18**, 1293–1304.

- Nardi, F., Spinsanti, G., Boore, J. L., Carapelli, A., Dallai, R. & Frati, F. 2003a Hexapod origins: monophyletic or paraphyletic? *Science* **299**, 1887–1889.
- Nardi, F., Spinsanti, G., Boore, J. L., Carapelli, A., Dallai, R. & Frati, F. 2003b Response to comment on 'Hexapod origins: monophyletic or paraphyletic'. *Science* **301**, 1482e.
- Negrisoló, E., Minelli, A. & Valle, G. 2004 The mitochondrial genome of the house centipede *Scutigera* and the monophyly versus paraphyly of myriapods. *Mol. Biol. Evol.* **21**, 770–780.
- Phillips, M. J. & Penny, D. 2003 The root of the mammalian tree inferred from whole mitochondrial genomes. *Mol. Phylogenet. Evol.* **28**, 171–185.
- Posada, D. & Crandall, K. A. 1998 MODELTEST: testing the model of DNA substitution. *Bioinformatics* **14**, 817–818.
- Shao, R. & Barker, S. C. 2003 The highly rearranged mitochondrial genome of the plague thrips, *Thrips imaginis* (Insecta: Thysanoptera): convergence of two novel gene boundaries and an extraordinary arrangement of rRNA genes. *Mol. Biol. Evol.* **20**, 362–370.
- Shao, R., Campbell, N. J. & Barker, S. C. 2001 Numerous gene rearrangements in the mitochondrial genome of the wallaby louse, *Heterodoxus macropus* (Phthiraptera). *Mol. Biol. Evol.* **18**, 858–865.
- Shimodaira, H. & Hasegawa, M. 1999 Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol. Biol. Evol.* **16**, 1114–1116.
- Simmons, M. P., Pickett, K. M. & Miya, M. 2004 How meaningful are Bayesian support values? *Mol. Biol. Evol.* **21**, 188–199.
- Strimmer, K. & von Haeseler, A. 1996 Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies. *Mol. Biol. Evol.* **13**, 964–969.
- Strimmer, K. & von Haeseler, A. 1997 Likelihood-mapping: a simple method to visualize phylogenetic content. *Proc. Natl Acad. Sci. USA* **94**, 6815–6819.
- Swofford, D. L. 1998 *PAUP* phylogenetic analysis using parsimony (*and other methods)*. Sunderland, MA: Sinauer Associates.
- Telford, M. J., Lockyer, A. E., Cartwright-Finch, C. & Littlewood, D. T. J. 2003 Combined large and small subunit ribosomal RNA phylogenies support a basal position of the acoelomorph flatworms. *Proc. R. Soc. B* **270**, 1077–1083. (doi:10.1098/rspb.2003.2342.)
- Tillyard, R. J. 1930 The evolution of the class Insecta. *Pap. Proc. R. Soc. Tas.* 1–89.
- Whelan, S. & Goldman, N. 2001 A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.* **18**, 691–699.
- Wilson, K., Cahill, V., Ballment, E. & Benzie, J. 2000 The complete sequence of the mitochondrial genome of the crustacean *Penaeus monodon*: are malacostracan crustaceans more closely related to insects than to branchiopods? *Mol. Biol. Evol.* **17**, 863–874.

The supplementary Electronic Appendix is available at <http://dx.doi.org/10.1098/rspb.2004.3042> or via <http://www.journals.royalsoc.ac.uk>.

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