

Comment on “Hexapod Origins: Monophyletic or Paraphyletic?”

Nardi *et al.* (1) suggested, rather cautiously, that hexapods (insects plus collembolans in their data set) might be a diphyletic rather than a monophyletic group. According to their interpretation, collembolans evolved separately from other insects and emerged before crustaceans. This unexpected result has huge consequences for the interpretation of both morphological and developmental evolution in arthropods (2) and therefore deserves further scrutiny—especially from a methodological standpoint.

Nardi *et al.* drew their conclusions from maximum likelihood and Bayesian analyses at the amino acid level of four of the 13 mitochondrial proteins for both the original 35-taxon data set and a 15-taxon subset. However, phylogenetic analyses of amino acids carry several potential caveats. First, the currently available models of mitochondrial amino acid substitution are based on empirically deduced matrices from mammalian-dominated sequence databases. Second, the maximum likelihood analysis used in (1) does not model the variation of rate across sites, which is known to be one of the most important parameters of the likelihood model (3). Third, bias in nucleotide composition also affects the amino acid composition of the gene product, thereby causing potential problems for phylogenetic reconstruction (4).

Some of these pitfalls might be avoided by analyzing nucleotide sequences for which more realistic models of sequence evolution and powerful reconstruction methods are available. In particular, we have recently shown that in the case of mammalian complete mitochondrial genomes (5), it is possible to deal with saturation and base

composition heterogeneity by recoding nucleotides as purines (R) and pyrimidines (Y). This approach provided a solution to longstanding controversies concerning the position of the root of the mammalian tree (5).

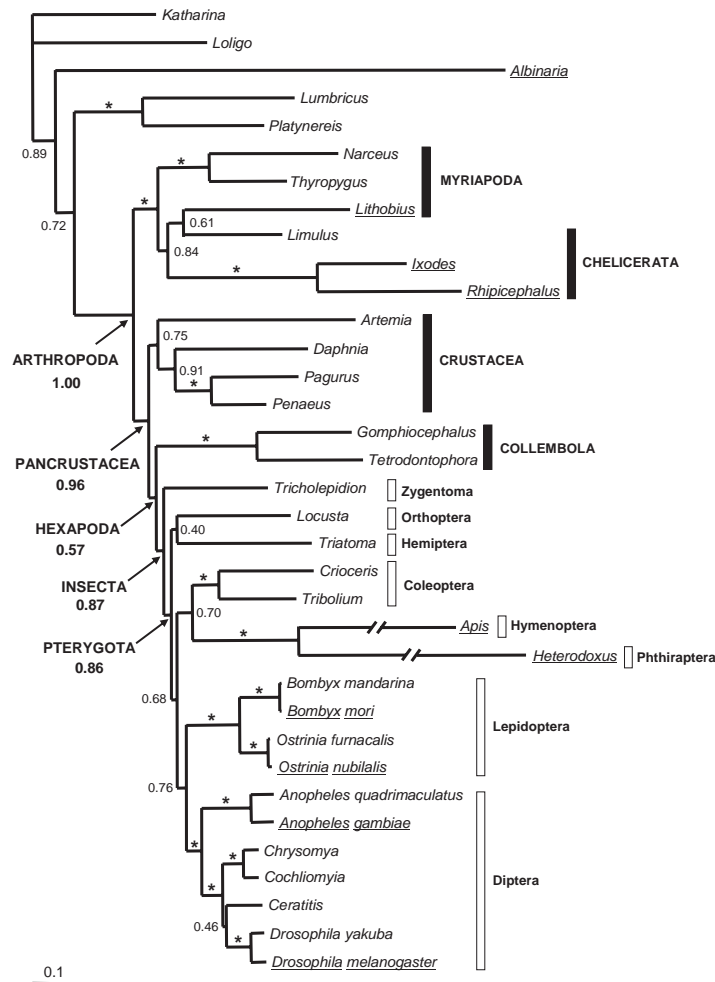


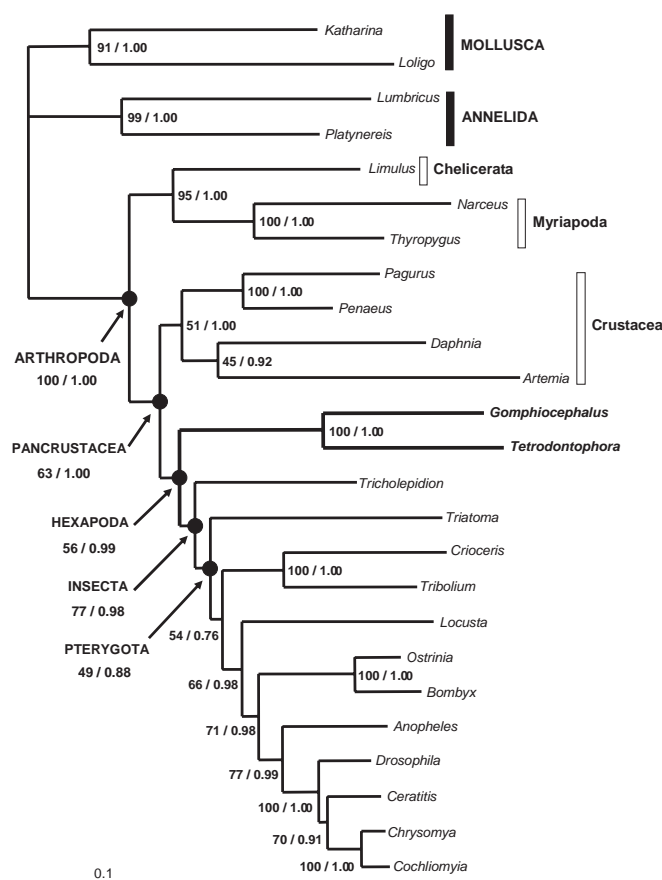
Fig. 1. Bayesian 50% majority rule consensus tree with associated branch lengths obtained using nucleotide sequences of *cox1*, *cox2*, *cox3*, and *cytb* (3750 sites) corresponding to the 35-taxon data set in (1). The first and third codon positions were RY-coded, whereas second codon positions were kept as nucleotides. MrBayes version 3.0b4 (12) was used to perform a partitioned-likelihood Bayesian search in which three independent substitution models were attributed to each codon position—a two-state substitution model + I + Γ for RY-coded first and third codon positions, and a GTR + I + model for second codon position nucleotides. Four incrementally heated Metropolis-coupled Markov chain Monte Carlo (MCMC) were run for 500,000 generations, sampling trees and parameters every 10 generations. The consensus tree was obtained from the 35,000 trees sampled after the initial burn-in period. Values at nodes indicate Bayesian posterior probabilities (* = 1.00). Note that the terminal branch lengths leading to the bee (*Apis*) and louse (*Heterodoxus*) have been reduced by a factor of three. Underlined taxa are not included in the 25-taxon data set.

Applying this strategy to nucleotides from the original Nardi *et al.* data set strongly suggests that by correcting for different artefacts it is possible to extract a useful historical signal. Unlike Nardi *et al.*, we were able to place the honeybee (*Apis*) and louse (*Heterodoxus*) within Insecta (Fig. 1). The artefactual position of these taxa as sister-groups of ticks in (1) was explained as being a consequence of high shared AT nucleotide composition in the mitochondrial genome sequences. From our results, base composition heterogeneity seems to be more easily accommodated in phylogenetic reconstructions using nucleotides. More importantly, our analysis conforms to classical views of arthropod phylogeny: Collembolans are a sister group of insects, and these monophyletic hexapods group with crustaceans into Pancrustacea (Fig. 1). One remaining problem with this tree concerns the paraphyly of myriapods induced by the nesting of the centipede (*Lithobius*) inside chelicerates.

As noted in (1), the phylogenetic analysis performed on the 35-taxon data set indicates uneven rates of evolution among taxa, making it difficult to draw firm conclusions about relationships between lineages. To test the collembolan position further, Nardi *et al.* reduced the data set to 15 taxa with more homogeneous evolution rates and amino acid compositions. Despite their conservative analysis, they still reported collembolans outside both insects and crustaceans, rendering hexapods diphyletic. However, such a reduced data set is particularly prone to systematic biases from low taxon sampling (6). Although deleting taxa with anomalous rates and base composition can be helpful, care must be taken not to delete taxa that could leave isolated branches and lead to a “long branches attract” phenomenon (7). More specifically, the inclusion of a single out-group can have a strong impact on phylogenetic reconstruction, even in the absence of rate heterogeneity (8). In the case of

TECHNICAL COMMENT

Fig. 2. Maximum likelihood (ML) phylogram obtained using nucleotide sequences of *cox1*, *cox2*, *cox3*, and *cytb* for a 25-taxon data set (3777 sites). The third codon positions were RY-coded, whereas first and second codon positions were kept as nucleotides. PAUP* (13) was used to perform a ML heuristic search under the best fitting GTR + I + Γ model and associated ML estimates of parameters as determined by Modeltest version 3.06 (14). A partitioned-likelihood Bayesian search was carried out with MrBayes (12) using a GTR + I + Γ model for first and second codon position nucleotides and a two-state substitution model + I + Γ for the RY-coded third codon positions, with the same parameter settings as in Fig. 1. Values at nodes indicate ML bootstrap proportions (100 replications)/Bayesian posterior probabilities. The two collembolans are figured in bold.



placental mammal mitogenomics, taxon sampling has been shown to be a major source of phylogenetic error (9), and we found that increasing the number and diversity of taxa produced excellent agreement between nuclear and mitochondrial sequence data (10).

To maximize taxon sampling, we constructed a well-balanced 25-taxon data set designed to break isolated long branches (especially in the outgroup) without adding strong rate heter-

ogeneity. Phylogenetic analyses of this nucleotide data set, including RY-coded third codon positions, produced a tree in which Arthropoda, Pancrustacea, Hexapoda, Insecta, and Pterygota all appear as monophyletic groups, though with variable support (Fig. 2). Moreover, this topology is much more compatible with current views of arthropod phylogeny (11). The probability of randomly selecting a topology compatible with this prior hypothesis is so small

(10) that it provides strong evidence in favor of its veracity. Obviously, additional complete mitochondrial genomes are needed to strengthen the tree further. However, with the data and methods currently available, the hypothesis of a common ancestry for extant hexapods cannot be rejected.

**Frédéric Delsuc
Matthew J. Phillips
David Penny**

*The Allan Wilson Center for Molecular
Ecology and Evolution
Institute of Molecular BioSciences,
Science Tower D
Massey University
Post Office Box 11-222
Palmerston North, New Zealand
E-mail: D.Penny@massey.ac.nz*

References and Notes

1. F. Nardi et al., *Science* **299**, 1887 (2003).
2. R. H. Thomas, *Science* **299**, 1854 (2003).
3. J. Sullivan, D. L. Swofford, *Syst. Biol.* **50**, 723 (2001).
4. P. G. Foster, D. A. Hickey, *J. Mol. Evol.* **48**, 284 (1999).
5. M. J. Phillips, D. Penny, *Mol. Phylogenet. Evol.* **28**, 171 (2003).
6. D. J. Zwickl, D. M. Hillis, *Syst. Biol.* **51**, 588 (2002).
7. M. D. Hendy, D. Penny, *Syst. Zool.* **38**, 297 (1989).
8. B. R. Holland, D. Penny, M. D. Hendy, *Syst. Biol.* **52**, 229 (2003).
9. H. Philippe, *J. Mol. Evol.* **45**, 712 (1997).
10. Y.-H. Lin et al., *Mol. Biol. Evol.* **19**, 2060 (2002).
11. G. Giribet, G. D. Hedgecombe, W. C. Wheeler, *Nature* **413**, 157 (2001).
12. F. Ronquist, J. P. Huelsenbeck, *Bioinformatics* **19**, 1572 (2003).
13. D. L. Swofford, PAUP* version 4.0b10 (Sinauer Associates, Sunderland, MA, 2002).
14. D. Posada, K. A. Crandall, *Bioinformatics* **14**, 817 (1998).
15. F. Nardi and colleagues kindly sent us their amino acid data set. E. Douzery provided helpful comments. Our data sets are available at <http://awcmee.massey.ac.nz/downloads.htm>. This work was supported by a Lavoisier Postdoctoral Grant from the French Ministry of Foreign Affairs to F.D. and by the New Zealand Marsden Fund.

7 May 2003; accepted 15 August 2003