

Points of View

Syst. Biol. 53(6):978–989, 2004
Copyright © Society of Systematic Biologists
ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150490888877

Identifying and Removing Fast-Evolving Sites Using Compatibility Analysis: An Example from the Arthropoda

DAVIDE PISANI

Department of Zoology, The Natural History Museum, Cromwell Road, London SW7 5BD, United Kingdom; E-mail: d.pisani@nhm.ac.uk

Felsenstein (1978) first recognized that long branch attraction (LBA) can seriously affect the accuracy of phylogenetic reconstruction. Although Felsenstein specifically addressed this problem in the cases of parsimony and clique analyses, it is now known that LBA can affect any tree reconstruction method, including maximum likelihood (ML) and Bayesian approaches. However, LBA is only a problem for distance, ML, and Bayesian methods when the assumed substitution model is underparameterized, i.e., when it is unrealistically simple (Swofford et al., 2001; Lemmon and Moriarty, 2004). LBA should therefore be avoidable by analyzing the data using ML, Bayesian, or distance methods under the best-fitting substitution model (providing this is a good approximation of the true substitution model). However, ML and (although to a much lesser extent) Bayesian analyses are time consuming, whereas many widely used implementations of distance methods (for example Kumar et al., 2001) do not allow the specification of complex substitution models. Accordingly, estimating the relationships of fast-evolving species still represents one of the most serious problems of molecular phylogenetics.

Strategies for dealing with LBA that do not necessarily rely on the use of probabilistic methods or complex evolutionary models have been suggested. These strategies can be of special utility when parsimony or distance methods are used. These include (1) increasing the taxon sampling (Hendy and Penny, 1989; Hillis, 1996; Rannala et al., 1998; Pollock et al., 2002; Poe, 2003); (2) optimal outgroup selection (Wheeler, 1990); and (3) sampling strategies specifically targeting slowly evolving species (e.g., Aguinaldo et al., 1997). Unfortunately, none of these strategies is universally applicable. For example, increasing the taxon sampling can (in some cases) exacerbate LBA (Kim, 1996; Poe and Swofford, 1999; Poe, 2003), phylogenetic uncertainty can prevent the selection of adequate outgroups, and for certain groups, it is possible that no slowly evolving species can be identified.

An alternative approach to countering LBA that does not necessarily rely on the use of complex substitution models is to identify and remove fast evolving sites, which are expected to contribute substantially to it (e.g., Brinkmann and Philippe, 1999; Hirt et al., 1999). Particu-

larly, Brinkmann and Philippe (1999) proposed a simple parsimony-based method, christened slow-fast (SF), for identifying (and then removing) fast evolving sites from an alignment. SF, as well as other methods that will not be considered in detail here (e.g., Hirt et al., 1999), can be especially useful when taxon sampling is limited, if close outgroups are unavailable, and in all cases when fast evolving species are included in the investigation.

Here, the use of alternative, compatibility-based methods (see Felsenstein, 2003; Semple and Steel, 2003; Pisani, 2002; Wilkinson, 2001; Meacham and Estabrook, 1985, for an introduction) for identifying fast-evolving sites is proposed and illustrated using arthropod data in a taxonomic congruence (Miyamoto and Fitch, 1995) context. Unlike the parsimony-based SF, the methods here proposed are topology independent, allowing for their application in cases where SF cannot be applied (see below).

IDENTIFYING FAST-EVOLVING SITES USING PARSIMONY

Brinkmann and Philippe's (1999) SF method works as follows. The aligned sequences are partitioned into n nonoverlapping monophyletic groups based on a priori background knowledge. For each of the n groups, parsimony analysis is carried out and the length of each position determined. Finally, for each position in the alignment, the sum of its lengths on each of the n trees is taken as a proxy of its evolutionary rate. Characters most likely to be fast evolving will be highlighted by their higher parsimony scores.

SF can prove useful in many situations, and it has been used to address important phylogenetic problems (e.g. Philippe et al., 2000; Brochier and Philippe, 2002). However, it requires subsets of monophyletic taxa to be known and defined a priori. It is not difficult to envision cases where this knowledge is unavailable or where the monophyletic subsets are misspecified. The full implications of misspecification for SF need to be investigated, for example through simulations, nevertheless, because the length of a character is a topology-dependent variable, it would generally lead to character lengths being erroneously calculated and putatively fast evolving sites misidentified. In any case, even if the sequences

were correctly partitioned, fast evolving sites will be accurately highlighted only to the extent that the within monophyletic group phylogenies are correctly resolved. Otherwise, character lengths may be erroneously calculated and a variable proportion of fast-evolving sites misidentified. In conclusion, the limitations of SF are a consequence of the method being topology dependent.

IDENTIFYING FAST-EVOLVING SITES USING COMPATIBILITY

Character compatibility was first introduced by Le Quesne (1969). Briefly, two characters are compatible if they can be mapped on the same tree without homoplasy (see Felsenstein, 2003, Semple and Steel, 2003, for details), otherwise they are incompatible.

Given a set of aligned sequences, for each site in the alignment, it is possible to calculate its incompatibility score, that is, the number of positions with which the specified site is incompatible. Because fast-evolving sites have lost most (if not all) of their phylogenetic information, they are expected to show more incompatibilities than slowly evolving ones and high incompatibility scores could therefore be used to highlight, and thus remove, potentially fast-evolving sites.

Incompatibility scores are calculated, for each character in a data set, without reference to any specific phylogenetic hypothesis, therefore, compatibility-based methods of character selection are topology independent, and the potential problems of misspecification of monophyletic groups and phylogenetic inaccuracy within subsets do not pertain.

Wilkinson (1992) and Meacham (1994) independently introduced compatibility-based randomization tests analogous to the parsimony PTP test (Archie, 1989; Faith and Cranston, 1991), but used to evaluate the quality of individual characters rather than entire matrices. Their methods differ only in trivial details (Wilkinson, 2001) and only Wilkinson's (1992) method will be considered here. It provides a test of the null hypothesis that a character is no less incompatible with the other characters in the data than is a random, phylogenetically uninformative, character. Expectations under the null model are determined for a given character by repeatedly randomly permuting (shuffling) the assignment of its character states across the taxa and counting the number of other (unpermuted) characters that it is incompatible with. The test statistic used, named the Le Quesne Probability (LQP), is the probability of a random character having as low or lower incompatibility with the rest of the data than does the original character (Wilkinson, 2001; see also Wilkinson and Nussbaum, 1996). Characters with high LQP have notably high incompatibility so that LQP values can be used to highlight fast-evolving sites.

When removing characters from an alignment, there is an important caveat to be considered: fast evolving sites could still convey some phylogenetic signal (*sensu* Pisani and Wilkinson, 2002). Accordingly, caution should be taken when removing sites (even if fast evolving), and character removal should be limited (hopefully) only to

outliers. However, the threshold at which character deletion would not improve phylogenetic accuracy is data dependent and so can only be found experimentally, by the sequential deletion of sites of increasingly good quality (e.g., sites with decreasing LQP values, lower incompatibility scores, or shorter lengths). Although protocols to optimize site removal can be outlined, these protocols cannot avoid a certain amount of subjectivity. In any case, LQP values offer an advantage over incompatibility and parsimony scores because the LQP of a character reflects both its compatibility with the other characters in the data and its performance with reference to a fully noisy character. Accordingly, LQP values provide a simple way of monitoring the risk of deleting useful information from the data whereas trying to remove noise. For example, deletion of characters with LQP of 1 is likely to result in an improved signal to noise ratio for the whole data set, whereas deletion of characters of decreasing LQP (e.g., LQP = 0.4) will not necessarily improve the signal to noise ratio of the data set, and could result in the loss of potentially important phylogenetic information.

EXAMPLE: AN ELONGATION FACTOR 1 α PHYLOGENY OF ARTHROPODA

Materials and Methods

DNA sequences of the Elongation Factor 1 α (EF-1 α) gene for a variety of arthropods were retrieved from Genbank (see Appendix for the accession numbers), and an alignment scoring 47 species and 866 nucleotide positions (gaps excluded) was created using Clustal X (Thompson et al., 1994) default options. The data set used in the analyses has been deposited (and is available for download) in TreeBase (www.treebase.org). Taxon sampling was designed to allow the testing of two phylogenetic hypotheses: (1) the monophyly of Hexapoda (contrast Nardi et al., 2003 and Delsuc et al., 2003) and (2) the phylogenetic relationships of the Myriapoda (compare Giribet et al., 2001, with Frederick and Tautz, 1995; Hwang et al., 2001; Cook et al., 2001; Nardi et al., 2003; Delsuc et al., 2003; Pisani et al., 2004; Mallatt et al., 2004; Negrisolo et al., 2004). Some vertebrates (human, salmon, and zebrafish) were selected as outgroups. Relatively distant outgroups were deliberately selected in an attempt to exacerbate LBA.

For the complete data set (all nucleotide positions), departure from homogeneity in base composition across taxa was tested using the χ^2 test (e.g., Negrisolo et al., 2004), and the best-fitting substitution model was selected using Modeltest (Posada and Crandall, 1998). Phylogenetic analyses were performed using Bayesian and ML methods under the best fitting substitution model selected by Modeltest (GTR+ Γ +I). Neighbor-joining (NJ) analyses were also performed. In order to allow LBA, the NJ analyses were carried out using an underparameterized substitution model, which was the gamma-corrected Kimura two-parameter (K2P) model, and transversions only were considered when calculating distances for the NJ analyses. All analyses were performed with gaps removed.

The strict consensus (Sokal and Rohlf, 1981) of the Bayesian and ML trees was calculated and the NJ tree compared with it and with what is generally accepted about the phylogeny of Arthropoda (Compare for example Giribet et al., 2001, with Hwang et al., 2001). This was done to evaluate to what extent the NJ tree differed from the Bayesian and likelihood trees, and to what extent the ML, Bayesian, and NJ trees matched what is generally accepted about the phylogeny of this group.

LQP values were then obtained for each position in the alignment, putatively fast-evolving sites (sites with high LQP) sequentially eliminated, and the retained data reanalyzed using Bayesian, ML, and NJ analyses. As above, NJ analyses were performed assuming an underparameterized (gamma-corrected K2P) model, whereas Bayesian and ML analyses were performed under the best-fitting substitution model, which was reestimated after removing the characters with high LQP.

To test the performance of the compatibility-based method, SF was also implemented and the results obtained using the two methods compared. To implement SF the sequences were partitioned a priori in three monophyletic groups (Pancrustacea, Chelicerata, and Myriapoda). Phylogenetic uncertainty and taxon-sampling limitations did not allow the definition of less inclusive monophyletic groups. Sites highlighted as fast evolving using SF were excluded and the remaining data reanalyzed using NJ, ML, and Bayesian analyses. As in the other cases, the NJ analyses were performed assuming a gamma-corrected K2P model, whereas the ML and Bayesian analyses were performed under the (reestimated) best-fitting substitution model.

To monitor the consequences associated with the deletion of putatively fast-evolving sites, characters were incrementally removed (in steps of 0.1) according to their LQP, and starting with those with the highest values, i.e., those with LQP included between 1 and 0.9. Phylogenetic analyses were carried out after each set of characters was removed and changes in the tree topology monitored. Removed characters were subjected to the PTP test to evaluate whether they conveyed clustering information, and phylogenetic analysis of the removed characters was performed to visualize their information content.

To evaluate whether the exclusion of putatively fast-evolving sites improved phylogenetic estimation, the likelihood of the NJ trees obtained after each set of characters was deleted and of the strict consensus of the Bayesian and ML trees was calculated. This was done (under the best fitting substitution model) considering (1) all nucleotide positions and (2) each set of retained characters. To evaluate whether the observed topological changes were significant, the Shimodaira-Hasegawa (SH) test (see Felsenstein, 2003, for details) was used to compare the NJ trees with each other and with the strict consensus of the ML and Bayesian trees. The likelihood calculations for the SH test were always performed considering all nucleotide positions, under the best fitting substitution model.

Support for the nodes in the NJ trees was evaluated using the interior branch length test (1000 replicates; see Nei

and Kumar, 2000, for details). Deletion of fast-evolving sites could have a potentially deleterious impact on the branch lengths of the recovered tree. As pointed out by Kim (1996), LBA can be exacerbated when long terminal branches (i.e., fast-evolving species) follow short internal branches. Accordingly, it is possible to conjecture that the application of methods for the removal of fast-evolving sites, if unevenly affecting different part of the tree, potentially, could exacerbate LBA. The interior branch test allows monitoring the significance of the branch lengths for the recovered tree, thus making it possible to evaluate whether site removal could be exacerbating LBA. Support for the nodes in the strict consensus of the ML and Bayesian trees are expressed as their posterior probabilities.

NJ analyses and the interior branch test were implemented using MEGA 2.1 (Kumar et al., 2001). The program DNALQP (which is part of the software package PICA 4.0; Wilkinson, 2001) was used to calculate LQP values. PAUP* (Swofford, 1998) was used to implement the χ^2 test for homogeneity in base composition across taxa, the PTP test (1000 replicates with heuristic search and the multiple trees option turned off), to carry out the ML and parsimony analyses (100 replicates with heuristic search and random sequence addition), to infer the gamma parameter (α) values used in the NJ analyses, to calculate the likelihood of the inferred trees, and to implement the SH test (RELL option with 100,000 replicates). Bayesian analyses were performed using MrBayes 3.0 (Ronquist and Huelsenbeck, 2003). For each Bayesian analysis 2,000,000 generations were run, sampling every 1000 generations. The burn in period of each analysis was estimated plotting the likelihood of the sampled trees.

Results

Diagnosing LBA.—The ML and Bayesian analyses of the data (all nucleotide positions) yielded trees differing only in trivial details, and their strict consensus is reported in Figure 1. This tree is generally consistent with the “known” phylogeny of Arthropoda, for example, supporting monophyletic Pancrustacea, Chelicerata, Myriapoda, Branchiopoda, and Malacostraca. However, it does not support a monophyletic Hexapoda (in accordance with Nardi et al., 2003), although, interestingly, the springtail *Tomocerus* sp. clusters with the “primitive” insects *Metajapyx subterraneus* and *Ctenolepisma lineata* (see Fig. 1). In accordance with Giribet et al. (2001), this tree supports a sister-group relation between Myriapoda and Pancrustacea (i.e., it supports the Mandibulata Hypothesis).

The NJ analysis of the data (all nucleotide positions) allowed for a possible case of LBA to be highlighted in the form of an obviously misplaced taxon, the crustacean *Speleonectes tulumensis* (Remipedia), stemming at the base of the tree (Fig. 2). However, LBA is not the only possible explanation for the misplacement of *S. tulumensis*: other possible explanations are base composition bias and paralogy (see Gribaldo and Philippe, 2002, for a review).

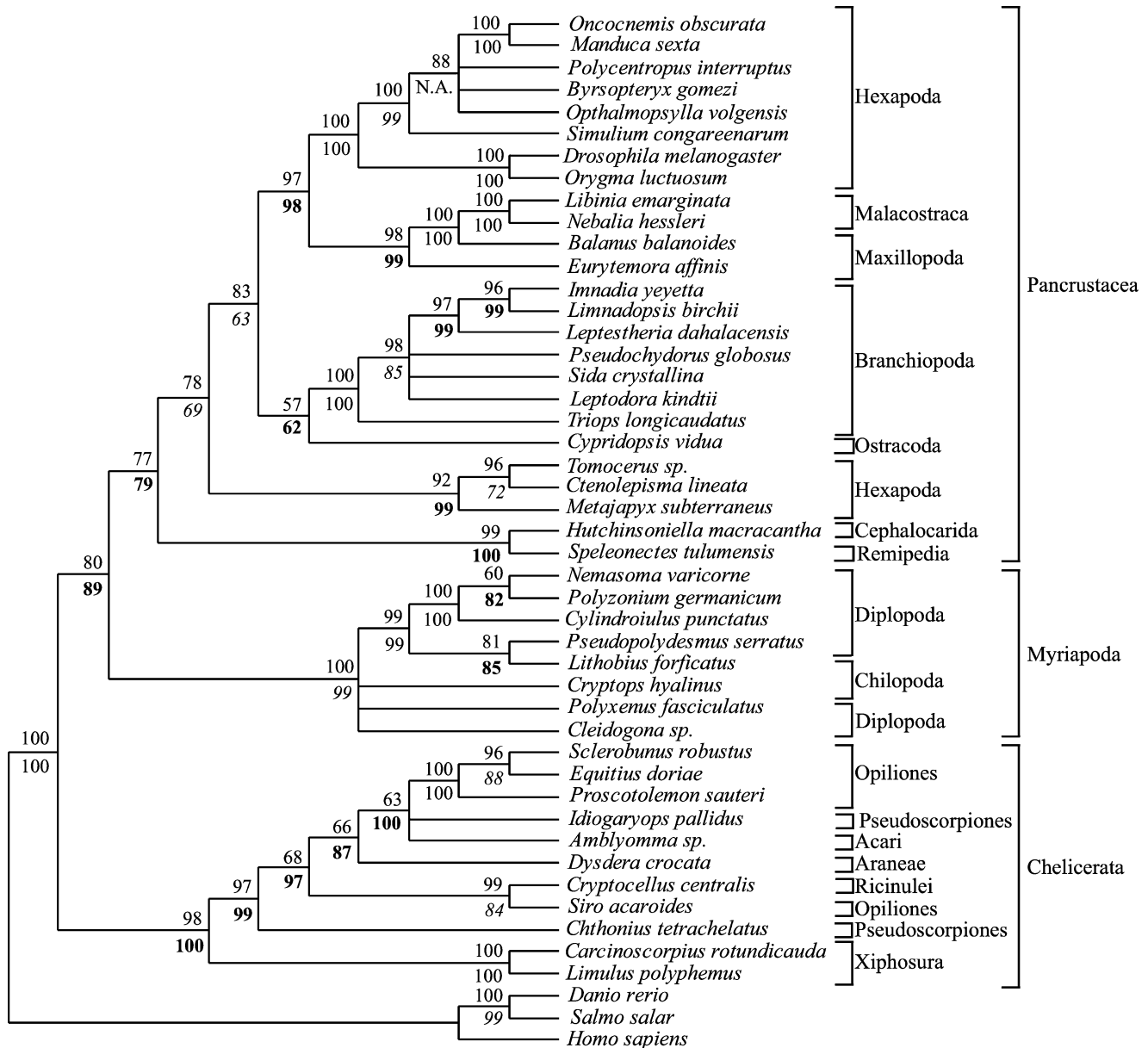


FIGURE 1. Strict consensus of the maximum likelihood and Bayesian trees. Numbers above the nodes represent the posterior probabilities (all sites). Numbers below the nodes are the “revised” posterior probabilities calculated after excluding all the sites with LQP ≥ 0.5 (172 sites excluded on a total of 866). In bold: posterior probabilities that increased after excluding the sites with high LQP. In italics: posterior probabilities that decreased after excluding the sites with high LQP. N.A.: not applicable. This node was not supported in the Bayesian analysis performed after excluding the sites with LQP ≥ 0.5 .

The χ^2 test for homogeneity in base composition could not highlight significant heterogeneity across taxa ($P = 0.927$). Evidence for the presence of multiple copies of the EF-1 α gene within Arthropoda exists (Hovemann et al., 1988; Danforth and Ji, 1998; Hedin and Maddison, 2001). However, based on the paralog intron-structure (Hedin and Maddison, 2001) and phylogenetic analyses (Danforth and Ji, 1998), these EF-1 α paralogs appear to be the result of independent duplications in specific lineages rather than of ancient duplication events (Hedin and Maddison, 2001). Accordingly, neither base composition bias nor paralogy can explain the

misplacement of *S. tulumensis* in Figure 2, leaving LBA as the most likely explanation. This is further confirmed by the ML and Bayesian analyses (see Fig. 1) where *S. tulumensis* nests within Pancrustacea as the sister group of the crustacean *Hutchinsoniella macracantha*. These results were to be expected if the misplacement of *S. tulumensis* in Figure 2 was due to LBA, which was avoided in the Bayesian and ML analyses implementing a better fitting substitution model.

The NJ tree in Figure 2 also shows several other obviously misplaced taxa. The “primitive” insects *M. subterraneus* and *C. lineata* are nested outside Pancrustacea,

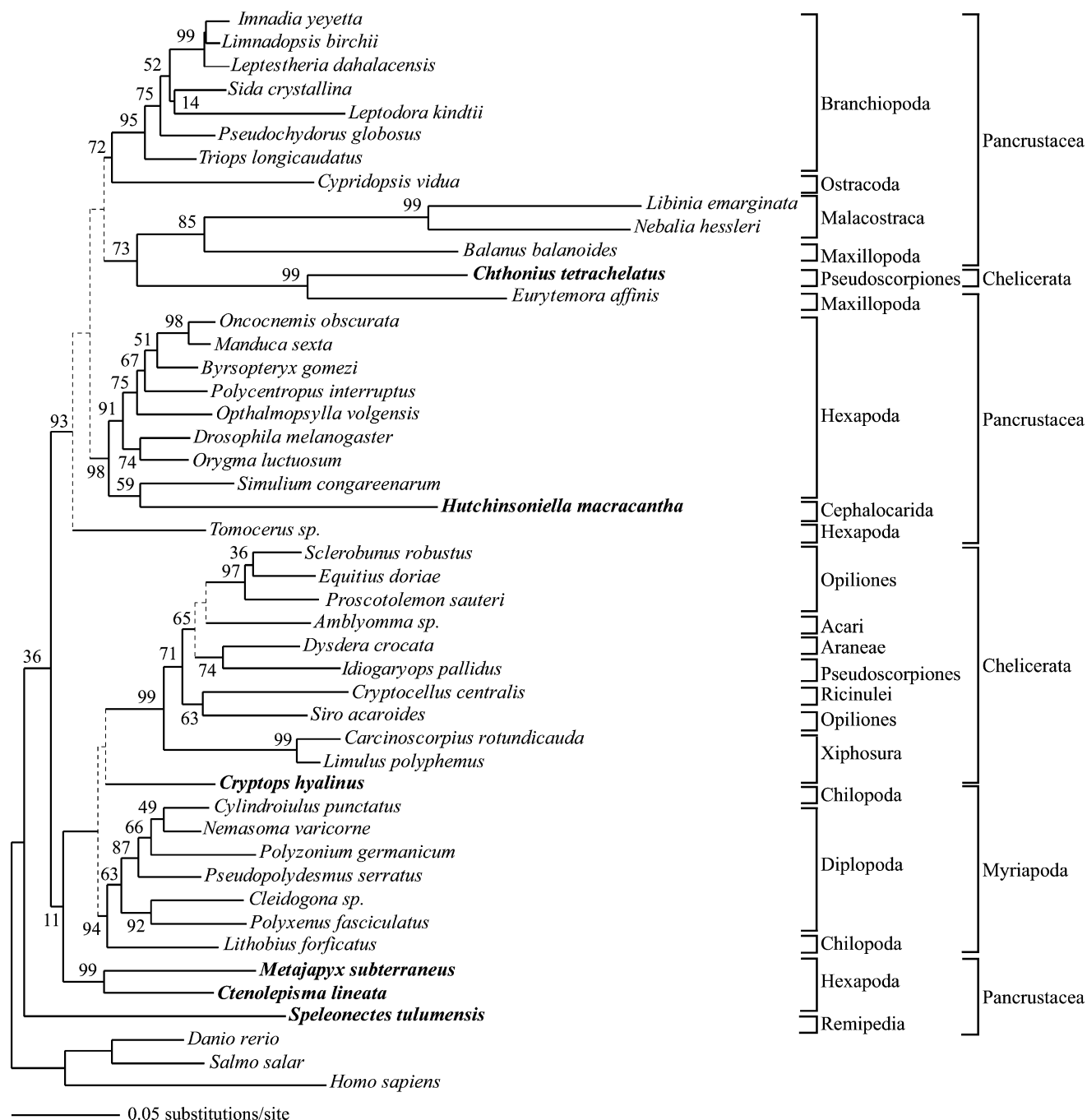


FIGURE 2. Neighbor-joining tree. Tree obtained from the analysis of all 866 sites ($\alpha = 0.24$). Numbers at the nodes represent support as expressed using the interior branch test. Dashed lines represent relationships unsupported in the tree summarizing the results of the interior branch test. In bold: obviously misplaced taxa.

the centipede (Chilopoda) *Cryptops hyalinus* is outside Myriapoda, the crustacean *Hutchinsoniella macracantha* is within Insecta, and the chelicerate *Chthonius tetrachelatus* is within Pancrustacea. Other interesting aspects of this tree are (1) the basal position of the springtail *Tomocerus* sp. that is nested outside Hexapoda (in accordance with Nardi et al., 2003) and, most importantly, (2) the pairing of the Chelicerata and Myriapoda. The latter aspect of the tree in Figure 2 contradicts the re-

sults of the ML and Bayesian analyses (compare Figs. 1 and 2) and provides support for the Myriochelata (Pisani et al., 2004), or Paradoxopoda (Mallatt et al., 2004), hypothesis. However, the likelihood of the NJ tree in Figure 2 is much lower than that of the strict consensus of the ML and Bayesian trees (Table 1), and the SH test suggests this tree fits the data significantly worse than the tree in Figure 1 does ($P = 0.0092$). Concluding, the NJ tree in Figure 2 is likely to be a poor reflection of

TABLE 1. Likelihood of the recovered trees. Likelihood values have all been calculated under the best-fitting substitution model (reestimated after each set of characters was excluded). Tree 1: Strict consensus of the Bayesian and ML trees (all nucleotide positions). Tree 2: NJ tree of all nucleotide positions. Tree 3: Suboptimal LQP tree (see text). Tree 4: The best LQP tree (see text). Tree 5: The best SF tree (see text). In bold: best NJ trees. An asterisk indicates the best trees overall.

Data Set	Log likelihoods				
	Tree 1	Tree 2	Tree 3	Tree 4	Tree 5
All data	-16513.66130*	-16595.47120	-16581.03254	-16531.51774	-16538.18100
Character with LQP \geq 0.6 excluded	-11794.60946*	-11911.73404	-11877.11806	-11812.16499	-11832.47369
Character with LQP \geq 0.5 excluded	-10069.06375*	-10181.10148	-10145.04290	-10083.85923	-10105.12040
Characters with Length \geq 9 excluded	-6376.14738*	-7320.65609	-7309.39463	-7242.93566	-7238.02008

the true phylogenetic relationships of these EF-1 α sequences.

Coping with LBA using LQP.—Deletion of putatively fast evolving sites (sites with high LQP) resulted in substantial changes of the recovered NJ trees, but did not result in any substantial change of the topology of the recovered ML and Bayesian trees. However, the posterior probabilities of many of the nodes recovered by the ML and Bayesian analyses varied when sites with high LQP were removed (see Fig. 1).

A first important change of the recovered NJ tree topology was observed after all the characters with LQP \geq 0.6 (i.e., the first 131 worst performing characters) were excluded (Fig. 3). This tree will be hereafter referred as “the suboptimal LQP tree” (see text below). Further site removal, down to all the sites with LQP \geq 0.5 (i.e., the first 172 worst performing characters) resulted in other notable changes in the recovered NJ tree (Fig. 4). The tree in Figure 4 will be hereafter referred as “the best LQP tree.” This is because its likelihood (calculated after the characters removed to infer it were reintroduced) is greater than that of any other NJ tree recovered (see also Table 1). Additional exclusion of increasingly better performing characters (with LQP $<$ 0.5) led only to the deterioration of the results (i.e., appearance of nonsensical clades, large polytomies, and decreasing likelihood of the recovered trees), suggesting important phylogenetic information was being removed and character deletion should have stopped.

It is immediately evident that removal of characters with high LQP improved the accuracy of the recovered NJ trees. In fact, even in the suboptimal LQP tree the crustacean *S. tulumensis* nested within Pancrustacea as the sister group of the crustacean *H. macracantha*. This is consistent with the results of the Bayesian and ML analyses of the complete data set (compare Figs. 1 and 3), and with the results of Giribet et al. (2001). Furthermore, in the suboptimal LQP tree, the “primitive” insects *M. subterraneus* and *C. lineata* are also nested within Pancrustacea, whereas the centipede *C. hyalinus* is clustered at the base of Myriapoda. In this tree, the only taxon that is still grossly misplaced, therefore, is the chelicerate *C. tetrachelatus*. The likelihood of the suboptimal LQP tree is greater than that of the NJ tree obtained from the analysis of the complete data set (see also Table 1), and is exceeded only by that of the best LQP tree. Nonetheless, the SH test indicates the suboptimal LQP tree does not fit the data significantly better than the

NJ tree obtained from the analysis of the complete data set ($P = 0.2605$).

Exclusion of other putatively fast evolving sites (i.e., all the sites with LQP \geq 0.5; see above) eventually resulted in the recovery of the best LQP tree, in which there are no grossly misplaced taxa. The likelihood of this tree is greater than that of the suboptimal LQP tree (and of any other NJ-LQP tree), suggesting that the exclusion of extra sites with high LQP (down to all the sites with LQP \geq 0.5) further improved phylogenetic estimation. This is further confirmed by the SH test, suggesting that the best LQP tree fits the data significantly better than both the NJ tree obtained from the analysis of the complete data set ($P = 0.0402$), and the suboptimal LQP tree ($P = 0.031$). Most importantly, the same test cannot reject the null hypothesis that the best LQP tree and the strict consensus of the Bayesian and likelihood trees fit the data equally well ($P = 0.2158$).

The best LQP tree is topologically very dissimilar from the suboptimal LQP tree (compare Figs. 3 and 4). The suboptimal LQP tree suggests Myriapoda to be the sister group of Chelicerata (i.e., it supports Myriochelata), whereas the best LQP tree suggests Myriapoda to be the sister group of Pancrustacea (i.e., it supports Mandibulata).

The PTP test suggested clustering information was being removed together with the putatively fast evolving sites ($P < 0.001$). However, the phylogenetic trees recovered from the analyses of the removed characters, except for suggesting some obvious group (e.g., Xiphosura, that is, the horseshoe crabs), were mainly nonsensical. Therefore, it is possible to conclude that the signal loss associated with the removal of the putatively fast-evolving sites is expression of the “residual” signal associated with the better supported groups, which is strong enough to be conserved even in the worst performing characters.

SF analysis.—Implementing SF, the sites in the alignment were partitioned into groups having equal length, the worst performing ones having a length of 21 steps. Interestingly, the fast-evolving characters identified using SF are sometimes different (and generally differently ranked) when compared with those highlighted in the LQP analysis.

Characters identified as putatively fast evolving by SF had been sequentially deleted down to the worst 175 (i.e., down to all the characters of length 10 or higher), that is, in a number comparable to that of the characters excluded to obtain the best LQP tree. NJ analysis of the

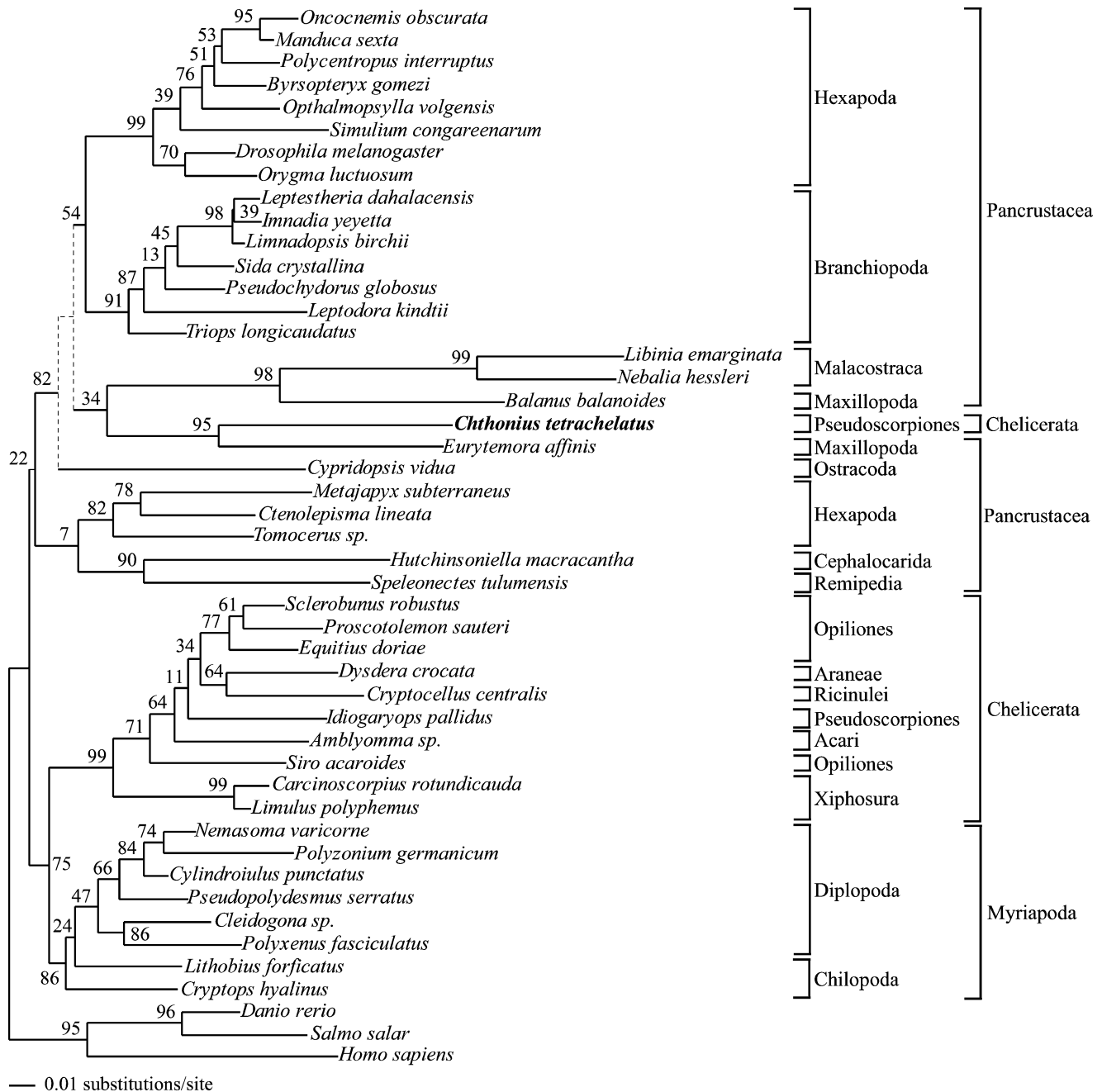


FIGURE 3. Neighbor-joining tree. Tree obtained after excluding the sites with $LQP \geq 0.6$, 131 on a total of 866 ($\alpha = 0.18$). This is the suboptimal LQP tree (see text). Numbers at the nodes represent support as expressed using the interior branch test. Dashed lines represent relationships unsupported in the tree summarizing the results of the interior branch test. In bold: obviously misplaced taxa.

remaining characters resulted in a tree still showing misplaced *S. tulumensis* and *H. macracantha* (not shown). The additional exclusion of all the characters of length 9 (for a total of 196), however, resulted in a tree topologically very similar to the best LQP tree (compare Figs. 4 and 5). The tree in Figure 5 is the SF tree of maximum likelihood (see Table 1) and will be referred hereafter as the best SF tree. Additional character removal (down to all the characters of length 5 or more) did not lead to any potential improvement of the results, and to a

generalized decrease of the likelihood of the recovered trees.

The likelihood of the best LQP tree and of the best SF tree are very similar (see Table 1) and the SH test cannot reject the null hypothesis that the two trees fit the data equally well ($P = 0.7186$). Besides, as in the case of the best LQP tree, the SH test cannot reject the null hypothesis that the best SF tree and the strict consensus of the Bayesian and ML trees of Figure 1 fit the data equally well ($P = 0.1586$).

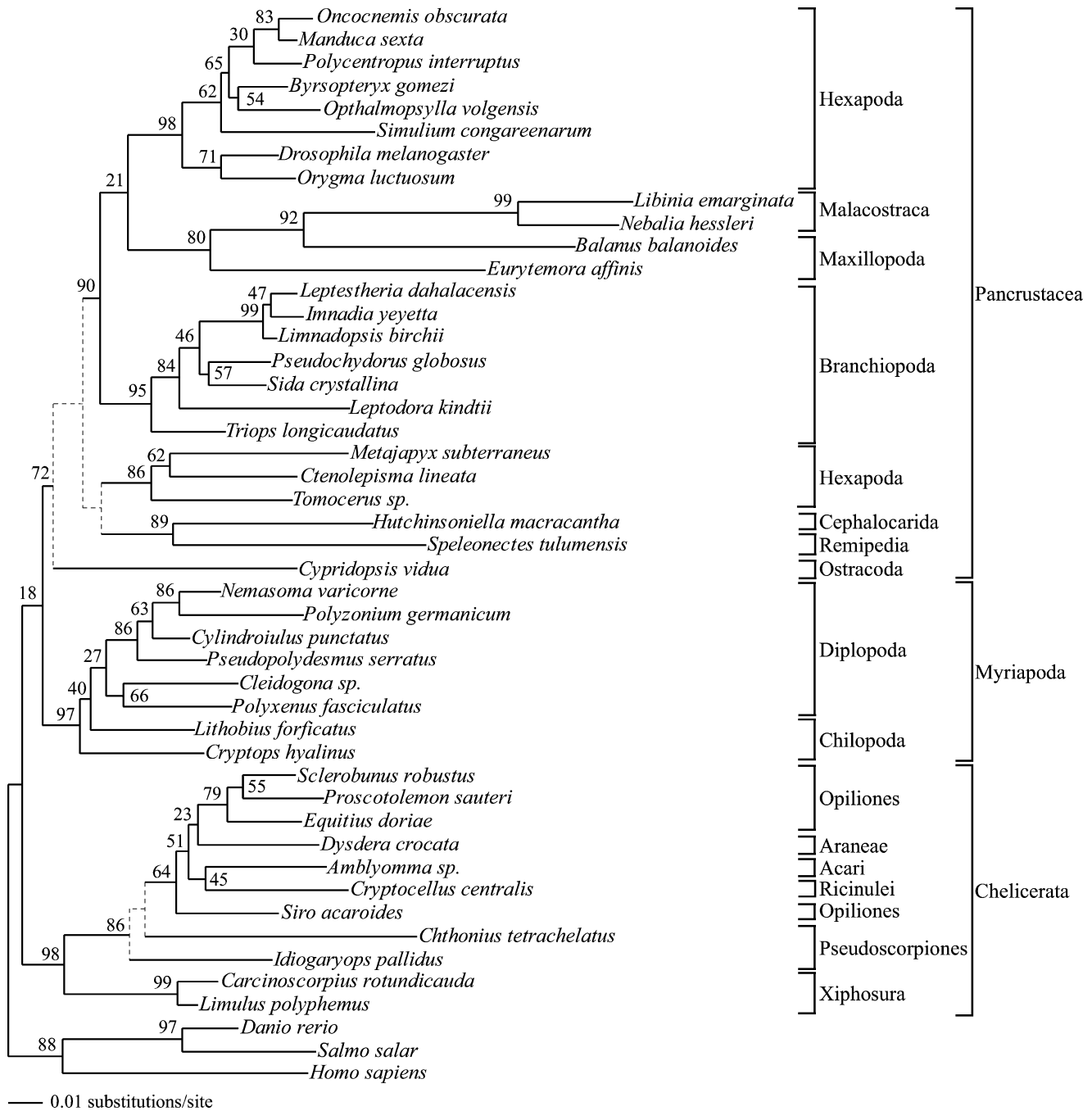


FIGURE 4. Neighbor-joining tree. Tree obtained after excluding the sites with LQP ≥ 0.5 , 172 on a total of 866 ($\alpha = 0.19$). This is the best LQP tree (see text). Numbers at the nodes represent support as expressed using the interior branch test. Dashed lines represent relationships unsupported in the tree summarizing the results of the interior branch test.

DISCUSSION

Phylogenetic Implications

Two competing hypotheses have been proposed for the relationships among Pancrustacea, Myriapoda and Chelicerata. These are (1) the Mandibulata hypothesis (see for example Giribet et al., 2001) and (2) the Myriochelata (or Paradoxopoda) hypothesis. Recent molecular phylogenetic analyses seem to support Myriochelata

(see Frederich and Tautz, 1995; Hwang et al., 2001; Cook et al., 2001; Kusche and Burmester, 2001; Nardi et al., 2003; Delsuc et al., 2003; Pisani et al., 2004; Mallatt et al., 2004; Negrisolo et al., 2004). However, morphological and combined morphological and molecular analyses (e.g., Giribet et al., 2001) support Mandibulata. Everything considered, the results here presented seems to favor Mandibulata. In fact, although the interior branch test indicates Mandibulata to be poorly supported in

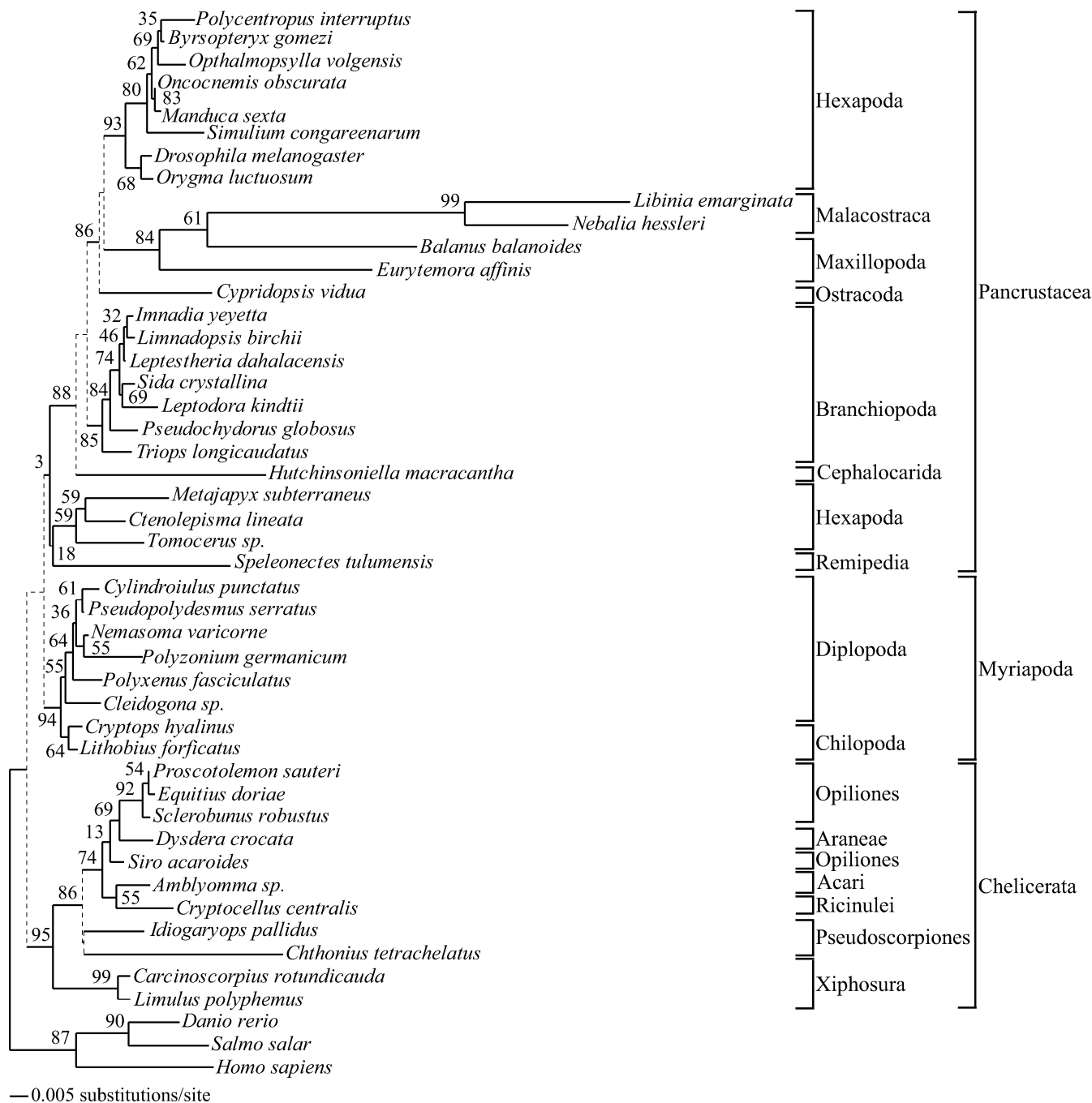


FIGURE 5. Neighbor-joining tree. Tree obtained after removing the sites identified (using SF) as having length ≥ 9 steps, 196 on a total of 866 ($\alpha = 0.20$). This is the best SF tree (see text). Numbers at the nodes represent support as expressed using the interior branch test. Dashed lines represent relationships unsupported in the tree summarizing the results of the interior branch test.

both the best LQP and SF trees, their congruence with the strict consensus of the ML and Bayesian trees (see Figs. 1, 4, and 5) suggests LBA is not being exacerbated by character deletion. Besides, the posterior probability of Mandibulata is relatively high and it increases when characters of high LQP are removed (Fig. 1).

Monophyly of insects is not even supported in the best LQP tree (in accordance with Nardi et al., 2003).

However, interestingly, a clade including the springtail *Tomocerus* sp. and the insects *M. subterraneus* and *C. lineata* is recovered in the suboptimal LQP tree, in both the best LQP and SF trees, and in the strict consensus of the Bayesian and ML trees (compare Figs. 1, 3, 4, and 5). Accordingly, this data set leaves the question of the monophyly/paraphyly of Hexapoda substantially unresolved.

Coping with LBA Using Compatibility-Based Methods

The implementation of both the LQP method suggested here and SF improved phylogenetic accuracy of the NJ analyses. Both methods yielded trees not significantly different from the ML and Bayesian ones, despite the NJ analysis employing an unrealistically simple model. Accordingly, this example shows topology independent, compatibility-based methods can be used to improve the signal to noise ratio of a data set. Furthermore, this example shows the proposed compatibility-based method performed at least as well as SF. Additional investigations need to be performed (for example through simulations) to evaluate better the extent to which compatibility-based methods could improve phylogenetic estimation. However, this example illustrates by means of taxonomic congruence (Miyamoto and Fitch, 1995) that the suggested method can improve phylogenetic accuracy, at least when distance methods are implemented under unrealistically simple substitution models (as it is often the case).

The utility of character removal methods in a ML or Bayesian context (where more complex substitution models are generally implemented) is less obvious. However, it should be pointed out that the available substitution models are all simplifications of the true, but unknown, substitution models (see for example Foster, 2004) and it can thus be conjectured that character removal methods could be useful also in a ML or Bayesian context. It should also be noted that, at least in this example, the exclusion of fast-evolving sites notably increased computational speed. The ML analysis of the complete data set (all positions) lasted more than 300 hours (on a cluster of 6 dual processor nodes), whereas the ML analysis performed after excluding all sites with $LQP \geq 0.5$ could be completed in 42 hours and 46 minutes.

In their simplest implementation (see above), compatibility-based methods for highlighting fast-evolving sites do not need the a priori definition of monophyletic groups of sequences. Furthermore, they do not need phylogenetic trees to be built in order to highlight fast evolving sites. Therefore, they make minimal assumptions about the data (Wilkinson, 1998), and can be implemented when the prior information necessary to implement SF is lacking. However, if enough information is available, the procedure implemented here can be modified, groups of monophyletic sequences defined, and a strict equivalent of SF designed.

All methods of character selection pose the problem of finding an optimal cut-off value under which characters should not be deleted. How to discriminate characters the deletion of which could improve phylogenetic accuracy, therefore, is key. Still, this is the most complex step of any character selection protocol. Because it is obvious that optimal cut-off values must be defined on a case-by-case basis, the threshold discriminating characters that should be deleted from those that should be retained can only be defined experimentally (i.e., by monitoring the effect associated with character deletion). Although this would always imply a certain amount of subjectivity,

some guidelines to monitor the effects of character deletion can be suggested. (1) Groups of characters should be removed sequentially according to their LQP (starting with the characters with a $LQP = 1$ and going down toward lower LQP values). Changes in the tree topology should be examined after each group of characters is removed. (2) Tests to highlight the presence of clustering signal (e.g., the PTP test) should be carried out on the deleted characters. (3) Phylogenetic analysis of the deleted characters should be performed to visualize their information content. (4) The likelihood of each recovered tree should be recorded, and the values compared. (5) Site removal should be stopped if further character deletion results in a significant and systematic deterioration of the results (i.e., appearance of obviously nonsensical clusters and/or substantial loss of resolution, support, or a decrease in the likelihood of the recovered trees). If the PTP (or a similar) test suggests that there is no clustering information in the excluded characters (point 2 above), their elimination should be justifiable. However, deleted characters often convey significant clustering information (Pisani, unpublished results, and above). In such cases, the phylogenetic tree recovered from the analysis of the deleted characters (point 3 above) should be scrutinized in order to evaluate what signal they convey.

CONCLUSIONS

LBA is still one of the major problems of molecular phylogenetics and there is great need of methods that could help coping with it. Both analytical methods and methods focusing on the taxon-sampling process are promising and need to be better investigated. However, because in many cases taxon sampling is limited, analytical methods, for example the methods suggested here or SF, seem particularly important. Still, with few notable exceptions (e.g., Hirt et al., 1999; Brinkmann and Philippe, 1999; this study) they have been overlooked.

It is the hope of the author that this study will stimulate further developments of these methods, as well as new studies specifically investigating areas of applicability of the different approaches, and new, more sophisticated experimental protocols.

ACKNOWLEDGMENTS

The author would like to thanks H. Philippe, F. Delsuc, I. Padovani, M. Wilkinson, F. Thomarat, D. Gower, S. Harris, and J. Cotton for their critical reading of this manuscript and for their helpful comments and suggestions.

REFERENCES

- Aguinaldo, A. M., J. M. Turbeville, L. S. Linford, M. C. Rivera, J. R. Garey, R. A. Raff, and J. A. Lake. 1997. Evidence for a clade of nematodes, arthropods, and other moulting animals. *Nature* 387:489–493.
- Archie, J. W. 1989. A randomization test for phylogenetic information in systematic data. *Syst. Zool.* 38:239–252.
- Brinkmann, H., and H. Philippe. 1999. Archaea sister group of Bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. *Mol. Biol. Evol.* 16:817–825.
- Brochier, C., and H. Philippe. 2002. A non-hyperthermophilic ancestor for Bacteria. *Nature* 417:244.

- Cook, C. E., M. L. Smith, M. J. Telford, A. Bastianell, and M. Akam. 2001. Hox genes and the phylogeny of arthropods. *Curr. Biol.* 11:759–763.
- Danforth, B. N. and S. Ji. 1998. Elongation factor-1 α occurs in two copies in bees: Implications for phylogenetic analysis of EF-1 α sequences in insects. *Mol. Biol. Evol.* 15:225–235.
- Delsuc, F., M. Phillips, and D. Penny. 2003. Comment on "Hexapod origins: Monophyletic or Paraphyletic?" *Nature* 301:1482.
- Faith, D. P., and P. S. Cranston. 1991. Could a cladogram this short have arisen by chance alone? On permutation tests for cladistic structure. *Cladistics* 7:1–28.
- Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.* 27:401–410.
- Felsenstein, J. 2003. *Inferring phylogenies*. Sinauer Associates, Sunderland, Massachusetts.
- Foster, P. G. 2004. Modeling compositional heterogeneity. *Syst. Biol.* 53:485–495.
- Frederich, M., and D. Tautz. 1995. Ribosomal DNA phylogeny of the major extant arthropod classes and the evolution of myriapods. *Nature* 376:165–167.
- Giribet, G., G. D. Edgecombe, and W. C. Wheeler. 2001. Arthropod phylogeny based on eight molecular loci and morphology. *Nature* 413:157–161.
- Gribaldo, S., and H. Philippe. 2002. Ancient phylogenetic relationships. *Theor. Pop. Biol.* 61:391–408.
- Hedin, M. C., and W. P. Maddison. 2001. Phylogenetic utility and evidence for multiple copies of Elongation Factor-1 α in the spider genus *Habronattus* (Araneae: Salticidae). *Mol. Biol. Evol.* 18:1512–1521.
- Hendy, M. D., and D. Penny. 1989. A framework for the quantitative study of evolutionary trees. *Syst. Zool.* 38:297–309.
- Hillis, D. M. 1996. Inferring complex phylogenies. *Nature* 383:130–131.
- Hirt, R., J. M. Jr. Logsdon, B. Healy, M. W. Dorey, W. F. Doolittle, and T. M. Embley. 1999. Microsporidia are related to fungi: Evidence from the largest subunit of RNA polymerase II and other proteins. *Proc. Natl. Acad. Sci. USA.* 96:580–585.
- Howemann, B., S. Richter, U. Walldorf, and C. Cziepluch. 1988. Two genes encode related cytoplasmic elongation factors 1 α (EF-1 α) in *Drosophila melanogaster* with continuous and stage specific expression. *Nucleic Acids Res.* 16:3175–3194.
- Hwang, U.-W., M. Friedrich, D. Tautz, C.-J. Park, and W. Kim. 2001. Mitochondrial protein phylogeny joins myriapods with chelicerates. *Nature* 413:154–157.
- Kim, J. 1996. General inconsistency conditions for maximum parsimony: Effects of branch lengths and increasing number of taxa. *Syst. Biol.* 45:363–374.
- Kumar, S., K. Tamura, I. B. Jakobsen, and M. Nei. 2001. MEGA2: Molecular evolutionary genetics analysis software, Arizona State University, Tempe, Arizona.
- Kusche, C., and T. Burmester. 2001. Diplopod hemocyanin sequence and the phylogenetic position of the myriapoda. *Mol. Biol. Evol.* 18:1566–1573.
- Le Quesne, W. J. 1969. A method of selection of characters in numerical taxonomy. *Syst. Zool.* 18:201–205.
- Lemmon, A. R., and E. C. Moriarty. 2004. The importance of proper model assumption in Bayesian phylogenetics. *Syst. Biol.* 53:265–277.
- Mallatt, J. M., J. R. Garey, and J. W. Shultz. 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.
- Meacham, C. A. 1994. Phylogenetic relationships at the basal radiation of angiosperms: Further study by probability of character compatibility. *Syst. Bot.* 19:506–522.
- Meacham, C. A., and G. F. Estabrook. 1985. Compatibility methods in systematics. *Annu. Rev. Ecol. Syst.* 16:431–466.
- Miyamoto, M. M., and W. M. Fitch. 1995. Testing species phylogenies and phylogenetic methods with congruence. *Syst. Biol.* 44:64–76.
- Nardi, F., G. Spinsanti, J. L. Boore, A. Carapelli, R. Dallai, and F. Frati. 2003. Hexapod origins: monophyletic or paraphyletic? *Science* 299:1887–1889.
- Nei, M., and S. Kumar. 2000. *Molecular evolution and phylogenetics*. Oxford University Press, New York.
- Negrisola, E., A. Minelli, and G. Valle. 2004. The mitochondrial genome of the house centipede *Scutigera* and the monophyly versus paraphyly of Myriapods. *Mol. Biol. Evol.* 21:770–780.
- Philippe, H., P. Lopez, H. Brinkmann, K. Budin, A. Germot, J. Laurent, D. Moreira, M. Müller, and H. Le Guyader. 2000. Early branching or fast evolving eukaryotes? An answer based on slowly evolving position. *Proc. R. Soc. Lond. B* 267:1213–1221.
- Pisani, D. 2002. Comparing and combining trees and data in phylogenetic analysis. Ph.D. Thesis, University of Bristol, UK.
- Pisani, D., L. Poling, M. Lyons-Weiler, and S. B. Hedges. 2004. The colonization of land by animals: Molecular phylogeny and divergence times among arthropods. *BMC Biology* 2:1.
- Pisani, D. and M. Wilkinson. 2002. Matrix representation with parsimony, taxonomic congruence and total evidence. *Syst. Biol.* 51:151–155.
- Poe, S. 2003. Evaluation of the strategy of long-branch subdivision to improve the accuracy of phylogenetic methods. *Syst. Biol.* 52:423–428.
- Poe, S., and D. L. Swofford. 1999. Taxon sampling revisited. *Nature* 398:299–300.
- Pollock, D., D. J. Zwickl, J. A. McGuire, and D. M. Hillis. 2002. Increased phylogenetic sampling is advantageous for phylogenetic inference. *Syst. Biol.* 51:664–671.
- Posada, D., and K. A. Crandall. 1998. Modeltest: Testing the model of DNA substitution. *Bioinformatics* 14:817–818.
- Rannala, B., J. P. Huelsenbeck, Z. Yang, and R. Nielsen. 1998. Taxon sampling and the accuracy of large phylogenies. *Syst. Biol.* 47:702–710.
- Ronquist, F., and J. P. Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19:1572–1574.
- Semple, C., and M. Steel. 2003. *Phylogenetics*. Oxford University Press, Oxford, UK.
- Sokal, R. R., and F. J. Rohlf. 1981. Taxonomic congruence in the Lep-*topodomorpha* reexamined. *Syst. Zool.* 30:309–325.
- Swofford, D. L. 1998. PAUP*. Phylogenetic analysis using parsimony (* and other methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.
- Swofford, D. L., P. J. Waddell, J. P. Huelsenbeck, P. G. Foster, P. O. Lewis, and J. S. Rogers. 2001. Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. *Syst. Biol.* 50:525–539.
- Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22:4673–4680.
- Wheeler, W. C. 1990. Nucleic acid sequence phylogeny and random outgroups. *Cladistics* 6:363–367.
- Wilkinson, M. 1992. Consensus compatibility and missing data in phylogenetic inference. Ph.D. Thesis. University of Bristol, UK.
- Wilkinson, M. 1998. Split support and split conflict randomisation tests in phylogenetic inference. *Syst. Biol.* 47:637–695.
- Wilkinson, M. 2001. PICA 4.0: Software and documentation. Department of Zoology, The Natural History Museum, London.
- Wilkinson, M., and R. A. Nussbaum. 1996. On the phylogenetic position of the Uraeotyphlidae (Amphibia: Gymnophiona). *Copeia* 3:550–562.

First submitted 2 December 2003; reviews returned 16 May 2004;

final acceptance 29 August 2004

Associate Editor: Mike Steel

APPENDIX: ACCESSION NUMBERS

Chelicerata:

Xiphosura (Horseshoe Crabs): *Carcinoscorpius rotundicauda* (AF063407), *Limulus polyphemus* (U90051). *Acari* (Ticks): *Amblyomma* sp. (AF240836). *Araneae* (Spiders): *Dysdera crocata* (U90047). *Opiliones* (Daddy Longlegs): *Equitius doriae* (AF240867), *Proscotolemon sauteri* (AF240872), *Siro acaroides* (AF240855), *Sclerobunus robustus* (AF240858). *Ricinulei*: *Cryptocellus centralis* (AF240839). *Pseudoscorpiones*: *Idiogaryops pallidus* (AF240848), *Chthonius tetrachelatus* (AF240841).

Pancrustacea:

Hexapoda (Insects): *Tomocerus* sp. (Springtail; U90059), *Ctenolepisma lineata* (Silverfish; AF063405), *Metajapyx subterraneus* (AF137389), *Simulium congareenarum* (Fly; AF003579), *Orygma luctuosum* (Fly; AY048510), *Drosophila melanogaster* (Fruit Fly; AY089522), *Byrsoteryx gomezi* (Caddisfly; AF436598), *Polycentropus interruptus* (Caddisfly; AF436639), *Oncocnemis obscurata* (Moth; U85685), *Manduca sexta* (Moth; AF234571), *Ophthalmopsylla volgensis* (Flea; AF423846). *Branchiopoda* (Crustaceans): *Triops longicaudatus* (U90058), *Pseudochydorus globosus* (AF526286), *Sida crystallina* (AF526280), *Leptestheria dahalacensis* (AF526291), *Imnadia yeyetta* (AF526289), *Limnadopsis birchii* (AF526290), *Leptodora kindtii* (AF526278). *Ostracoda* (Crustaceans): *Cypridopsis vidua* (AF063414). *Malacostraca* (Crustaceans): *Libinia emarginata* (Spider crab; U90050), *Nebalia hessleri* (AF063413). *Maxillopoda* (Crustaceans): *Balanus*

balanoides (Barnacle; AF063404), *Eurytemora affinis* (AF063408). *Cephalocarida* (Crustaceans): *Hutchinsoniella macracantha* (AF063411). *Remipedia* (Crustaceans): *Speleonectes tulumensis* (AF063416).

Myriapoda:

Chilopoda (Centipedes): *Cryptops hyalinus* (AF240790), *Lithobius forficatus* (AF240799). *Diplopoda* (Millipedes): *Polyxenus fasciculatus* (U90055), *Pseudopolydesmus serratus* (AF240814), *Nemasoma varicorne* (AF240800), *Polyzonium germanicum* (AF240805), *Cylindroiulus punctatus* (AF240792), *Cleidogona* sp. (AF240791).

Vertebrata:

Salmo salar (Salmon; AF321836), *Danio rerio* (Zebrafish; NM131263), *Homo sapiens* (BC000432).