Hox genes and the phylogeny of the arthropods

Charles E. Cook, M. Louise Smith, Maximilian J. Telford, Alberto Bastianello* and Michael Akam

The arthropods are the most speciose, and among the most morphologically diverse, of the animal phyla. Their evolution has been the subject of intense research for well over a century, yet the relationships among the four extant arthropod subphyla – chelicerates, crustaceans, hexapods, and myriapods – are still not fully resolved. Morphological taxonomies have often placed hexapods and myriapods together (the Atelocerata) [1, 2], but recent molecular studies have generally supported a hexapod/crustacean clade [2–9]. A cluster of regulatory genes, the *Hox* genes, control segment identity in arthropods, and comparisons of the sequences and functions of *Hox* genes can reveal evolutionary relationships [10]. We used *Hox* gene sequences from a range of arthropod taxa, including new data from a basal hexapod and a myriapod, to estimate a phylogeny of the arthropods. Our data support the hypothesis that insects and crustaceans form a single clade within the arthropods to the exclusion of myriapods. They also suggest that myriapods are more closely allied to the chelicerates than to this insect/crustacean clade.

Address: University Museum of Zoology, Downing Street, Cambridge CB2 3EJ, United Kingdom.

Current address: *Università degli studi di Padova, Dipartimento di Biologia, viale Colombo, Padova 3-35121, Italy.

Correspondence: Charles E. Cook
E-mail: ccook@phillips.exeter.edu

Received: **26 February 2001**
Revised: **26 March 2001**
Accepted: **26 March 2001**

Published: **15 May 2001**

Current Biology 2001, 11:759–763

0960-9822/01/$ – see front matter

The discrepancy between traditional taxonomy and the growing weight of molecular evidence suggests that many of the morphological characters used to build arthropod phylogenies have been subject to convergence, particularly between insects and myriapods. However, it remains difficult to build well-resolved molecular trees, probably because the diversification of the arthropods was rapid and ancient; crown group arthropods are already present in the early Cambrian 520 million years ago [11]. The few characters that provide relevant phylogenetic signals are masked by 500 million years of noise.

The *Hox* genes are an ancient family of developmental regulatory genes that, in arthropods, are differentially expressed along the anterior/posterior axis of the body to define tagmosis and many finer details of segment organization [12]. *Hox* genes have been conserved since the early divergence of the bilaterian animals, and *Hox* gene sequences have proven useful for resolving deep phylogenies [10].

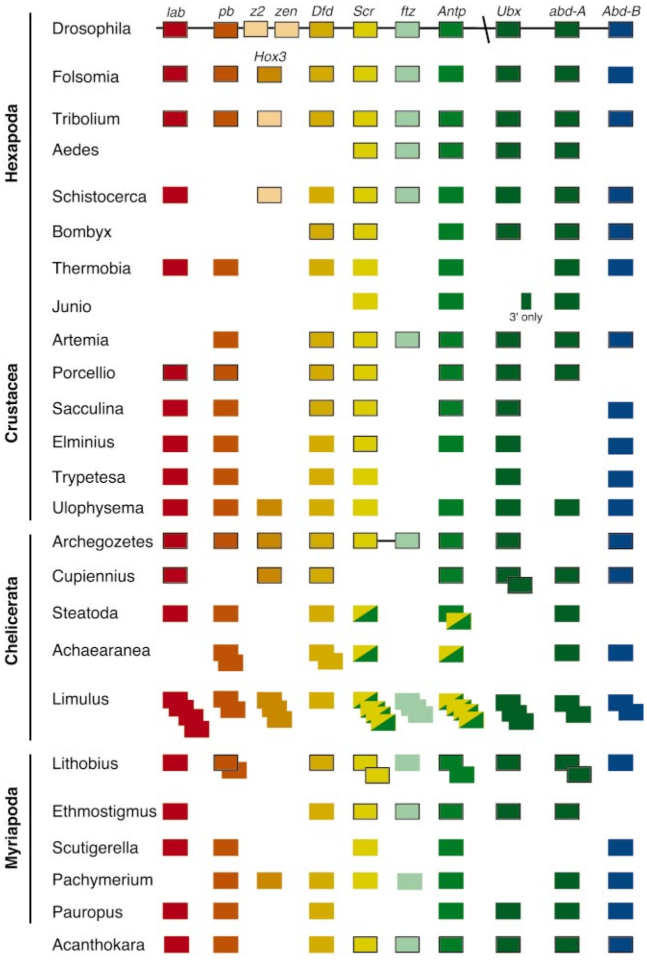
The *Hox* genes of insects can be assigned to one of ten classes on the basis of "signature" amino acid residues in and around the homeodomain [10]. We report here the sequences of 10 *Hox* genes from a basal hexapod, *Folsomia candida* (Collembola), one gene assignable to each of these 10 classes, and 13 *Hox* genes from the centipede *Lithobius forficatus*. These 13 genes are together assignable to nine of the ten insect *Hox* classes. We did not identify a *zen*/*Hox3* class gene in *L. forficatus*, but four of the other classes are represented by two distinct *Lithobius* genes.

We also report three new *Hox* sequences (*labial*, *Ultrabithorax*, and *Abdominal-B*) from the oribatid mite *Archegozetes longisetosus* (Chelicerata); short PCR fragments of five *Hox* genes from a symphylan (Myriapoda), *Scutigerella immaculata* and seven from a pauropod (Myriapoda) *Pauropus* species (Figure 1); a short *proboscipedia* fragment from the branchiopod crustacean *Artemia franciscana*; and, from the grasshopper *Schistocerca gregaria*, short fragments of the genes *labial*, *Deformed*, and *Ultrabithorax*, which have not previously been described. All sequences have been deposited in GenBank with accession numbers AF318494-AF318499, AF335458–AF335464, AF361326–AF3613335, AF362084–AF362097, AF363015–AF363018, and AJ309283–AJ309285.

We have aligned these sequences with a data set containing most of the published arthropod *Hox* gene sequences. We use these combined data to examine the relationships among the arthropod taxa by phylogenetic analysis using amino acid sequences.

The *Hox* genes of all arthropods can be assigned to the same classes as those of insects. Figure 1 depicts these assignments for most of the published arthropod *Hox* genes (taxa for which three or fewer genes are known are not shown). Allowing for the incompleteness of the data, most arthropod taxa appear to contain a single representative of each *Hox* gene class, which together presumably

Figure 1



Arthropod *Hox* genes. Arthropods for which at least four *Hox* genes have been reported are shown. *Drosophila melanogaster*, with the best-characterized *Hox* cluster, is shown at the top. Other taxa are arranged by subphylum. Lines connecting genes indicate known linkage relationships. Black borders around genes indicate that all or most of the sequence of the 60 residue homeobox region has been reported. Boxes without borders represent short fragments only. Box colors indicate homology within columns. Some short fragments are identifiable as central class genes (i.e., *Antennapedia*-like) but cannot be classified as any one gene. These are shown in two colors. Gene duplications are shown as additional, slightly offset boxes in each row. For taxon abbreviations and sequence sources, see Table S3 in the Supplementary material.

comprise a single *Hox* cluster as they do in insects. We use the *Drosophila* gene names to refer to the orthologous genes in all arthropods.

Gaps in Figure 1 represent genes that have not been found, not missing genes. To our knowledge, it has not been conclusively demonstrated that any arthropod taxon is missing any *Hox* cluster gene, though this may be the case for genes of the *abd-A* class in cirripedes [13]. Gene or cluster duplications have occurred in some lineages, but these have occurred subsequently to the radiation of the four subphyla.

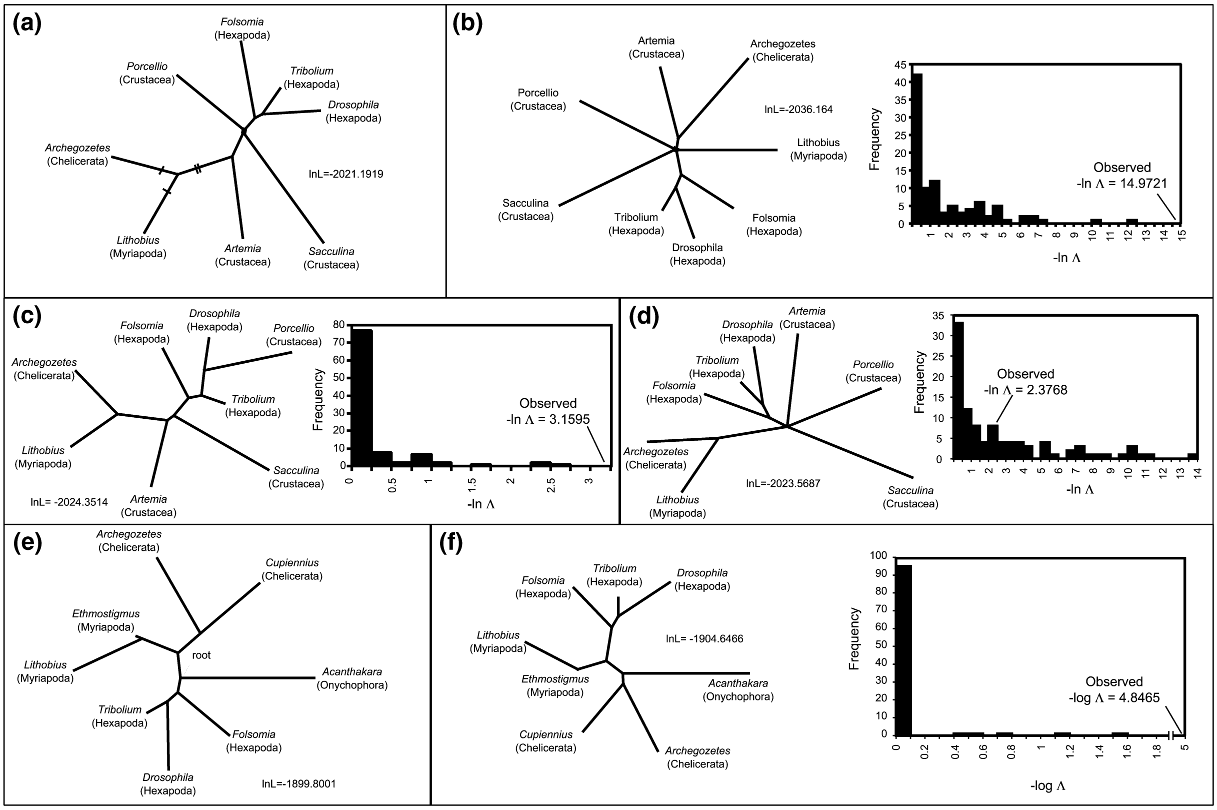
Most of the sequence variation in the arthropod *Hox* genes occurs in the regions flanking the homeodomain rather than in the homeodomains themselves. For many of the taxa shown in Figure 1, only short fragments of part of the homeodomain have been identified. These short sequences have few variable sites and are therefore not useful for phylogenetic analysis. We identified ten taxa for which all or most of the homeodomain and flanking regions have been reported for at least four genes (for alignments see Figure S1 in the Supplementary material available with this article on the internet) and chose those for further phylogenetic analysis.

Of these ten taxa, only the three hexapods *D. melanogaster*, *F. candida*, and *T. castaneum* are represented by complete or almost complete sequences for all ten genes. Other taxa are represented by data for some genes only. Consequently, we could not construct a single data set containing all genes and all taxa. Instead, we assembled several different data sets, each of which had the potential to resolve a particular phylogenetic question. Each data set comprised concatenated amino acid sequences of taxa for which complete or nearly complete homeodomain sequences were available for the same set of genes. For each additional taxon included, it was usually necessary to reduce the number of genes selected. Each data set was tested for its ability to produce well-supported trees.

No data set that included good sampling of all four arthropod subphyla allowed us to include a sufficient number of genes to resolve well-supported trees, but two more limited data sets did prove informative; we used data set 1, with a chelicerate, a myriapod, three crustaceans, and three hexapods, to test crustacean/hexapod relationships, and we used data set 2, with an onychoporan, two chelicerates, two myriapods, and three hexapods, to test chelicerate, myriapod, and hexapod relationships. Results from phylogenetic analyses of these two data sets are shown in Figure 2.

Figure 2a shows an unrooted maximum-likelihood tree for data set 1. Although the tree is unrooted, only three possible locations (marked) for a root are biologically plausible. Other roots would imply improbable relationships, for example involving splitting the crustaceans into two major clades or grouping crustaceans with chelicerates and myriapods and thereby leaving hexapods as basal in the tree. Furthermore, analysis of data set 2 supports one of these three roots, which is indicated by a double bar on Figure 2a. We conclude that the root for this tree is on one of the three branches shown and that these data therefore support a monophyletic group containing hexapods and crustaceans. This tree also shows crustaceans

Figure 2



Phylogenetic analyses of concatenated amino acid sequences from various arthropods. Taxa are identified by genus. For full species names and sequence sources, see Figure S3 in the Supplementary material. **(a)** A PAML maximum-likelihood tree for data set 1 including three hexapods, three crustaceans, one myriapod, and one chelicerate with 431 amino acid residues from the following six genes: *Dfd, Scr, ftz, Antp, Ubx,* and *Abd-B*. Parameters were optimized with a likelihood ratio test. The model used allowed each gene to evolve at a separate rate, and it had a single gamma rate distribution for the entire data set. Bars across branches represent possible positions for a root. A double bar indicates a root supported by analysis of data set 2. The next-best 13 trees rearranged the three crustacean lineages and the three hexapod lineages relative to each other, but all maintained the hexapods-within-crustaceans topology. **(b)** Parametric bootstrap results and best tree in which myriapods and hexapods form a monophyletic group. This tree was used as the null hypothesis for the generation of 100 artificial data sets for parametric bootstrapping, and for each data set we calculated a test statistic, $-\ln \Lambda$, by finding the difference between lnL of the best tree for that data set and the lnL of the best tree conforming to the null hypothesis. The values were binned (X axis) and tallied (Y axis) as shown. All of the 100 values of the test statistic are less than that for the original data set, so the null hypothesis is rejected. **(c)** The best tree in which hexapods are not monophyletic. This tree was used as the null hypothesis for the generation of 100 data sets as above, and a frequency plot of the test statistic, $-\ln \Lambda$, for this data set is shown. All of the 100 values of the test statistic are less than that for the original data set, so the null hypothesis is rejected. **(d)** The best tree in which hexapods and crustaceans are separate monophyletic lineages. This tree was used as the null hypothesis for the generation of 100 artificial data sets for parametric bootstrapping, and a frequency plot is shown. The value of $-\ln\Lambda$ for the real data is exceeded by 25% of the artificial data sets, so the null hypothesis cannot be rejected in this case. **(e)** The maximum-likelihood tree for dataset 2 including two chelicerates, two myriapods, three hexapods, and an onychophoran, with 445 amino acid residues from the six genes *lab, Dfd, Scr, Antp, Ubx, abdA,* and *AbdB*. Parameters were optimized with a likelihood ratio test. The model used allowed each gene to evolve at a separate rate and had a single gamma rate distribution for the entire data set. The position where the outgroup, *Acanthokara* (Onychophora), joins the tree is marked as "root." **(f)** The best tree in which myriapods and hexapods form a monophyletic group. This tree was used as the null hypothesis for the generation of 100 artificial data sets for parametric bootstrapping, and a frequency plot is shown. All 100 of the data sets had $-\ln \Lambda$ values below that for the original data set, thus the null hypothesis is rejected.

as paraphyletic with respect to hexapods; we test this relationship below.

The association of hexapods and crustaceans, excluding myriapods, is supported by other molecular phylogenies and mitochondrial gene order data. Recent morphological phylogenies, however, group the myriapods and hexapods as sister taxa [14]. We tested the myriapod/hexapod clade by first identifying the best tree in which hexapods and myriapods are sister groups but that excludes crustaceans and chelicerates. We then considered this tree as the null hypothesis in a parametric bootstrapping analysis (Figure

2b). In this analysis, the null hypothesis is rejected with near certainty; these *Hox* gene sequences strongly support a monophyletic lineage of hexapods and crustaceans.

The maximum-likelihood tree for this data set (Figure 2a) suggests that the hexapods are a monophyletic lineage. We tested this result by identifying the best tree in which hexapods are not a monophyletic group and then using this tree as the null hypothesis to generate artificial data sets for parametric bootstrapping (Figure 2c). In this analysis also, the null hypothesis is rejected, and we conclude that these data support the monophyly of the hexapods (specifically Collembola and Pterygota).

The three crustacean taxa in data set 1 belong to three different crustacean subclasses, Branchiopoda, Malacostraca, and Maxillopoda, and represent a wide spectrum of crustacean lineages. In order to test the monophyly of the crustaceans with respect to the hexapods, we identified the best tree that separates the crustaceans and hexapods into two separate monophyletic sister groups, then used this tree as the null hypothesis in a parametric bootstrap test. Results from this analysis are shown in Figure 2d. In this case the null hypothesis that crustaceans and hexapods are two monophyletic lineages cannot be rejected; thus, while our maximum-likelihood tree suggests that hexapods may in fact be a lineage within the Crustacea, confirmation of this result awaits additional data.

Figure 2e shows an unrooted maximum-likelihood tree for data set 2. Although unrooted as shown, the mitochondrial DNA gene order data [4] provide unequivocal evidence that the Onychophora lie outside the arthropod lineage and thus root the arthropods at the base of the Onychophoran branch. When so rooted, the chelicerates, which are represented by two spiders, and myriapods, which are represented by two centipedes, form a monophyletic group that excludes the hexapods. We were unable to include any crustacean sequences in this data set, but previous evidence, as well as the results from analyses of data set 1, support a hexapod/crustacean clade. We tested the robustness of the myriapod/chelicerate clade by identifying the best tree that places the myriapods together with the hexapods and then used this tree as the null hypothesis in a parametric bootstrap test (Figure 2f). For this test the null hypothesis is rejected, so we conclude that the *Hox* gene sequences in our data support the division of the arthropods into two lineages, one including myriapods and chelicerates, and one including crustaceans and hexapods.

When considered together, the analyses of our two data sets therefore suggest that the Arthropoda are divided into two major lineages, one comprising hexapods and crustaceans and another comprising myriapods and chelicerates. This result was unexpected because most taxonomists have considered chelicerates as basal in the arthropod lineage. However, we note that a number of molecular studies have also reported a myriapod/chelicerate clade [5, 6].

The insects were long believed to be most closely related to the myriapods, principally through the common presence of trachea (hence "Tracheata") and malphigian tubules and through the common lack of second antennae. This clade has, however, been repeatedly questioned by molecular studies, with the new consensus being that the hexapods and crustaceans form a single clade. Our data reinforce this result. The corollary of the dismantling of the Tracheata is that malphigian tubules and tracheae of hexapods and myriapods must have evolved convergently, while their secondary antennae were convergently lost. We can also infer that insects must derive not from some homonomous myriapod-like body but rather from an already tagmatized crustacean, with very different implications for the evolution of segmentation.

Our evidence that chelicerates are allied to myriapods argues against the idea of a clade of mandibulate arthropods (insects/crustaceans and myriapods). Rather, it supports the alternative notion that three taxa sharing a well-defined and complex character — the mandible — might not be monophyletic. This suggests that mandibles might have been present in the common arthropod ancestor and might subsequently have been lost in chelicerates. Alternatively, we must assume the convergent evolution of mandibles in myriapods and in the crustacean/insect clade.

Both of these results reinforce the conclusion that the morphological features traditionally used to infer relationships among the arthropod subphyla make a poor phylogenetic data set. At this depth in the tree, convergence and stochastic change overwhelm whatever phylogenetic signal they contain.

Materials and methods

We amplified short fragments of the *Hox* genes from genomic DNA of *F. candida*, *L. forficatus*, *S. immaculata*, and *Pauropus* sp. by using various combinations of degenerate primers designed to match conserved regions of the *Hox* protein sequences (see Tables S1 and S2 in the Supplementary material). We extended *L. forficatus* and *F. candida* sequences by using inverse PCR (iPCR) [15] or by sequencing phage clones isolated from a genomic library. *F. candida* and *L. forficatus Hox* genes were unambiguously identified by alignment with previously published sequences and the presence of "diagnostic" residues for each gene [10]. Sequences of the homeobox motif and its flanking regions were aligned by eye; alignments were extended into flanking regions only as far as the sequences could be unambiguously aligned (Figure S1 in the Supplementary material).

We report here only phylogenies estimated by maximum likelihood because these allow more specific and accurate models of the evolutionary process to be implemented [16] and because this method allows the testing of alternative tree hypotheses by the use of parametric bootstrapping. We assembled various data sets by concatenating gene sequences from subsets of the total data. The usefulness of each data set for phylogenetic analysis was evaluated by likelihood mapping and quartet-puzzling maximum-likelihood estimation with TREE-PUZZLE [17, 18]. We tested each data set under different models (parameter settings) with PAML [19] and used a likelihood ratio test (LRT) to identify the best model [20]. We then used parametric bootstrapping to evaluate the phylogeny suggested by the best tree with respect to other competing phylogenetic hypotheses [20, 21]. By this method, one takes some other phylogenetic hypothesis as the null hypothesis and calculates a maximum-likelihood value for all trees that conform to the null hypothesis by using the model (parameter settings) identified by the LRT. The difference between the value for this tree and the value for the best overall tree is calculated and used as a test statistic. To assess the significance of this value, one generates artificial data sets (100 for this study) by using the parameter estimates (for this study, tree topology, branch lengths, gamma values, and amino acid frequencies) for the null-hypothesis tree. For each of these artificial data sets, the difference in the maximum-likelihood value for the best tree under the null hypothesis and for the best overall tree are compared. The proportion of the replicates in which this value exceeds the same value calculated from the original data set represents the significance level of the test. A fuller description of this testing is given in the Supplementary material.

Supplementary material

Supplementary material including three tables, a figure, and additional methodological details is available at http://images.cellpress.com/supmat/supmatin.htm.

Acknowledgements

This work was supported by a Biotechnology and Biological Sciences Research Coucil grant to M.E.A. M.J.T is supported by the Wellcome Trust #0605031/Z/00/2. We thank N. Goldman, S. Whelan, and Z. Yang for assistance with the maximum-likelihood analysis and N. Goldman for comments on the manuscript. We also thank R. Thomas for allowing the use of previously unpublished mite *Hox* gene sequences and F. Falciani for for the *A. franciscana* proboscipedia sequence.

References

1. Brusca RC, Brusca GJ: *Invertebrates*. Sunderland, Massachusetts: Sinauer Associates, Inc.; 1990.
2. Giribet G, Ribera C: **A review of arthropod phylogeny: new data based on ribosomal DNA sequences and direct character optimization.** *Cladistics* 2000, **16:**204-231.
3. Boore JL, Collins TM, Stanton D, Daehler LL, Brown WM: **Deducing the pattern of arthropod phylogeny from mitochondrial DNA rearrangements.** *Nature* 1995, **376:**163-165.
4. Boore JL, Lavrov DV, Brown WM: **Gene translocation links insects and crustaceans**. *Nature* 1998, **392:**667-668.
5. Ballard JW, Olsen GJ, Faith DP, Odgers WA, Rowell DM, Atkinson PW: **Evidence from 12S ribosomal RNA sequences that onychophorans are modified arthropods**. *Science* 1992, **258:**1345-1348.
6. Friedrich M, Tautz D: **Ribosomal DNA phylogeny of the major extant arthropod classes and the evolution of myriapods.** *Nature* 1995, **376:**165-167.
7. Giribet G, Carranza S, Baguña J, Riutort M, Ribera C: **First molecular evidence for the existence of a tardigrada + arthropoda clade**. *Mol Biol Evol* 1996, **13:**76-84.
8. Regier JC, Shultz JW: **Molecular phylogeny of the major arthropod groups indicates polyphyly of crustaceans and a new hypothesis for the origin of hexapods.** *Mol Biol Evol* 1997, **14:**902-913.
9. Turbeville JM, Pfeifer DM, Field KG, Raff RA: **The phylogenetic status of arthropods, as inferred from 18S rRNA sequences**. *Mol Biol Evol* 1991, **8:**669-686.
10. deRosa R, Grenier JK, Andreeva T, Cook CE, Adoutte A, Akam M: ***Hox* genes in brachiopods and priapulids and protostome evolution**. *Nature* 1999, **399:**772-776.
11. Budd GE, Jensen S: **A critical reappraisal of the fossil record of the bilaterian phyla.** *Biological Reviews* 2000, **75:**253-295.
12. Akam M: **Arthropods: developmental diversity with a (super) phylum.** *Proc Natl Acad Sci USA* 2000, **97:**4438-4441.
13. Mouchel-Vielh E, Rogolot C, Gibert J-M, Deutsch J: **Molecules of the body plan: the *Hox* genes of Cirripedes (Crustacea)**. *Mol Phylogenet Evol* 1998, **9:**382-389.
14. Edgecombe GD, Wilson GDF, Colgan DJ, Gray MR, Cassis G: **Arthropod cladistics: combined analysis of histone H3 and U2 snRNA sequences and morphology.** *Cladistics* 2000, **16:**155-203.
15. Averof M, Akam M: ***Hom/Hox* genes of Artemia – implications for the origin of insect and crustacean body plans.** *Curr Biol* 1993, **3:**73-78.
16. Whelan S, Lio P, Goldman N: **Molecular Phylogenetics: state of the art methods for looking into the past.** *Trends Genet.* 2001, **17:**262-272.
17. Strimmer K, vonHaeseler A: **Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies**. *Mol Biol Evol* 1996, **13:**964-969.
18. Strimmer K, vonHaeseler A: **Likelihood-mapping: a simple method to visualize phylogenetic content**. *Proc Natl Acad Sci USA* 1997, **94:**6815-6819.
19. Yang Z: **PAML: a program for package for phylogenetic analysis by maximum likelihood.** *Comput Appl Biosci* 1997, **15:**555-556.
20. Huelsenbeck JP, Rannala B: **Phylogenetic methods come of age: testing hypotheses in an evolutionary context.** *Science* 1997, **276:**227-232.
21. Huelsenbeck JP, Crandall KA: **Phylogeny estimation and hypothesis testing using maximum likelihood**. *Annu Rev Ecol Syst* 1997, **28:**437-466.