# Evidence for Multiple Reversals of Asymmetric Mutational Constraints during the Evolution of the Mitochondrial Genome of Metazoa, and Consequences for Phylogenetic Inferences

ALEXANDRE HASSANIN,[1] NELLY LÉGER,[2] AND JEAN DEUTSCH[3]

[1]*Muséum National d'Histoire Naturelle, Département Systématique et Evolution, UMR 5202—Origine, Structure, et Evolution de la Biodiversité, Case Postale N°51, 55, rue Buffon, 75005 Paris, France; E-mail: hassanin@mnhn.fr*
[2]*Université Pierre et Marie Curie (Paris 6), UMR 7138—Systématique, Adaptation, Evolution, Batiment B, 7ème étage, 7, quai Saint Bernard, 75252 Paris Cedex 05, France*
[3]*Université Pierre et Marie Curie (Paris 6), UMR 7622—Biologie du Développement, 9, quai St Bernard, Case 24, 75252 Paris Cedex 05, France*

*Abstract.*—Mitochondrial DNA (mtDNA) sequences are commonly used for inferring phylogenetic relationships. However, the strand-specific bias in the nucleotide composition of the mtDNA, which is thought to reflect asymmetric mutational constraints, combined with the important compositional heterogeneity among taxa, are known to be highly problematic for phylogenetic analyses. Here, nucleotide composition was compared across 49 species of Metazoa (34 arthropods, 2 annelids, 2 molluscs, and 11 deuterosomes), and analyzed for a mtDNA fragment including six protein-coding genes, i.e., *atp6, atp8, cox1, cox2, cox3,* and *nad2.* The analyses show that most metazoan species present a clear strand asymmetry, where one strand is biased in favor of A and C, whereas the other strand has a reverse bias, i.e., in favor of T and G. The origin of this strand bias can be related to asymmetric mutational constraints involving deaminations of A and C nucleotides during the replication and/or transcription processes. The analyses reveal that six unrelated genera are characterized by a reversal of the usual strand bias, i.e., *Argiope* (Araneae), *Euscorpius* (Scorpiones), *Tigriopus* (Maxillopoda), *Branchiostoma* (Cephalochordata), *Florometra* (Echinodermata), and *Katharina* (Mollusca). It is proposed that asymmetric mutational constraints have been independently reversed in these six genera, through an inversion of the control region, i.e., the region that contains most regulatory elements for replication and transcription of the mtDNA. We show that reversals of asymmetric mutational constraints have dramatic consequences on the phylogenetic analyses, as taxa characterized by reverse strand bias tend to group together due to long-branch attraction artifacts. We propose a new method for limiting this specific problem in tree reconstruction under the Bayesian approach. We apply our method to deal with the question of phylogenetic relationships of the major lineages of Arthropoda. This new approach provides a better congruence with nuclear analyses based on 18S rRNA gene sequences. By contrast with some previous studies based on mtDNA sequences, our data suggest that Chelicerata, Crustacea, Myriapoda, Pancrustacea, and Paradoxopoda are monophyletic. [Arthropoda; asymmetry; genome; long-branch attraction artifact; mitochondria; molecular evolution; mutations; phylogeny; strand bias.]

The mitochondrial (mt) genome varies extensively in size and gene content across diverse eukaryotic groups, but its structure is surprisingly uniform among metazoans (Boore, 1999; Taanman, 1999). A typical metazoan mtDNA is a circular and double-stranded molecule of 14 to 18 kb, and encodes 37 genes: 13 protein subunits of the enzymes of oxidative phosphorylation (subunits 6 and 8 of the ATPase [*atp6* and *atp8*], cytochrome c oxidase subunits 1 to 3 [*cox1* to *cox3*], apocytochrome *b* [*cob*], and NADH dehydrogenase subunits 1 to 6 and 4L [*nad1* to *nad6* and *nad4L*]); two rRNA of the mitochondrial ribosome (small and large subunit rRNAs [*rrnS* and *rrnL*]); and 22 for tRNAs necessary for the translation of the proteins encoded by the mtDNA (Attardi, 1985; Taanman, 1999). It has a very compact gene organization, with no introns, generally few noncoding nucleotides between genes, in some cases short overlaps of genes, and the presence of only one major noncoding region, named the control region, which contains the main regulatory elements for the initiation of replication and transcription.

The most remarkable feature of mtDNA is the strand-specific bias in nucleotide composition. In mammals, one strand is G rich, whereas the other strand is G poor, and because they show different buoyant densities in a cesium chloride gradient, they are respectively called heavy (H) and light (L) strands (Anderson et al., 1981). This strand bias is particularly evident at fourfold degenerate sites of protein-coding genes, where patterns of substitutions are unaffected by selection: one strand is rich in A and C nucleotides whereas the other is rich in T and G (Tanaka and Ozawa, 1994; Perna and Kocher, 1995; Reyes et al., 1998). The underlying mechanism that leads to the strand bias has been generally related to replication, because this process has long been assumed to be asymmetric in the mtDNA and could therefore affect the occurrence of mutations between the two strands (Clayton, 1982; Tanaka and Ozawa, 1994; Reeyes et al., 1998). These hypotheses have, however, been questioned by recent experiments suggesting that replication is not asymmetric because of the double-stranded state of both strands during the DNA synthesis (Yang et al., 2002).

Sequences of the mt genome have been widely used for inferring phylogenetic relationships between highly divergent lineages. In particular, they have been extensively used for deciphering interrelationships between the four main groups of the phylum Arthropoda, i.e., (1) Crustacea (crabs, shrimps, etc.), (2) Hexapoda (insects, proturans, springtails, and diplurans), (3) Myriapoda (centipedes, millipedes, and their kin), and (4) Chelicerata (horseshoe crabs, arachnids, and pycnognids) (Brusca and Brusca, 2003). The analyses of mtDNA sequences have revealed several unexpected results with huge consequences for the interpretation of morphological characters: (i) Crustacea have been found paraphyletic, with Malacostraca being more closely related to Hexapoda than Branchiopoda (Garcia-Machado et al., 1999; Wilson et al., 2000; Nardi et al., 2001; Hwang et al., 2001; Nardi et al., 2003); (ii) Hexapoda have been found

paraphyletic, with Insecta allied with Crustacea rather than with Collembola (Nardi et al., 2003); (iii) Chelicerata and Myriapoda have been found para- or polyphyletic (Nardi et al., 2003; Delsuc et al., 2003); and (iv) Hwang et al. (2001) have suggested that Myriapoda share more affinities with Chelicerata while most morphological studies propose to group Myriapoda either with Pancrustacea into the clade Mandibulata (e.g., Snodgrass, 1938), or with Hexapoda into the clade Atelocerata (e.g., Snodgrass, 1938; Cisne, 1974).

The usefulness of mtDNA as a marker for highly divergent lineages remains controversial (e.g., Curole and Kocher, 1999). Two main characteristics of the mt genome are expected to be problematic for reconstructing the phylogeny of arthropods: mutational saturation and heterogeneity in nucleotide composition among taxa. The first arthropods probably arose in ancient Precambrian seas over 600 million years ago (Brusca and Brusca, 2003). As a consequence, mutational saturation due to multiple hits is a major problem in tree reconstruction, and with mt sequences, saturation is all the more important because the mt genome typically evolves much more rapidly than the nuclear genome (Li, 1997; Burger et al., 2003). The mt genomes of arthropods are also characterized by a strong compositional bias, but in contrast to the mammalian mtDNA, which is A+C rich, it is particularly rich in A and T nucleotides (e.g., Garcia-Machado et al., 1999; Wilson et al., 2000; Dotson and Beard, 2001; Shao et al., 2001; Machida et al., 2002). This heterogeneity in nucleotide composition among metazoan lineages can lead to incorrect phylogenetic inferences because unrelated taxa with similar base compositions may be erroneously grouped together (Tarrío et al., 2001; Rosenberg and Kumar, 2003).

In the present work, nucleotide composition was analyzed in a mtDNA fragment, including the six protein-coding genes *atp6*, *atp8*, *cox1*, *cox2*, *cox3*, and *nad2* for 34 arthropods and 15 species belonging to five other phyla. This fragment was chosen because the arrangement of these six genes is conserved in most arthropod species. Our analyses confirm that most metazoan species present a clear strand asymmetry, where one strand is biased in favor of A and C, whereas the other strand has a reverse bias, i.e., in favor of T and G. The origin of this strand bias is related to asymmetric mutational constraints involving deaminations of A and C nucleotides during the replication and/or transcription processes. Six unrelated genera are however characterized by a reversal of the usual strand bias, i.e., *Argiope* (Araneae), *Euscorpius* (Scorpiones), *Tigriopus* (Maxillopoda), *Branchiostoma* (Cephalochordata), *Florometra* (Echinodermata), and *Katharina* (Mollusca). We suggest that asymmetric mutational constraints have been independently reversed in these six genera, through an inversion of the control region, i.e., the region that contains most regulatory elements for replication and transcription of the mtDNA.

By using the same data matrix, we also studied the effect of strand-bias on phylogenetic inferences. We show that reversals of asymmetric mutational constraints have

dramatic consequences on phylogenetic inferences, as taxa characterized by reverse strand bias tend to group together due to long-branch attraction artifacts. We propose a new method for limiting this specific problem in tree reconstruction under the Bayesian approach. We apply our method to the issue of phylogenetic relationships between the major lineages of Arthropoda to test the validity of our method. We show that this new approach provides a better congruence with nuclear analyses based on 18S rRNA (*18S*) gene sequences.

MATERIAL AND METHODS

*Taxonomic Sampling and DNA Alignments*

The taxonomic sample comprises 49 species (Table 1). It has been chosen for inferring phylogenetic relationships among the major arthropod lineages by using mtDNA and 18S rDNA sequences. For the 18S rDNA analyses, we sought to choose a taxonomic sampling as close as possible to the 49 taxa used in the mtDNA analyses (Table 1). The ingroup is the phylum Arthropoda, represented by 34 species with 13 Insecta, 2 Collembola, 7 Crustacea, 3 Myriapoda, and 9 Chelicerata. The outgroup includes 15 genera belonging to five different Metazoan phyla, i.e., Annelida, Chordata, Echinodermata, Hemichordata, and Mollusca. Five species of chelicerates were specially sequenced for this study: one pycnogonid, i.e., *Endeis spinosa*, and four arachnids, i.e., *Argiope bruennichi* (Araneae), *Euscorpius flavicaudis* (Scorpiones), *Mastigoproctus giganteus* (Uropygi), and *Phrynus* sp. (Amblypygi). The protocols used for mtDNA extraction and sequencing are given elsewhere (Hassanin, submitted).

Two different DNA alignments were performed manually with Se-Al v2.0a11 (Sequence Alignment Editor Version 2.0 alpha 11; Andrew Rambaut, software available at http://evolve.zoo.ox.ac.uk/): the first one includes six protein-coding genes of the mt genome, i.e., *atp8* and *atp6*, *cox1* to *cox3*, and *nad2*; the second one corresponds to the 18S rRNA gene. All regions involving ambiguity for the position of the gaps were excluded from the analyses to avoid erroneous hypotheses of primary homology. The reduced alignment of mt sequences consists of 3948 nucleotides (nt), and the one of *18S* sequences includes 1463 nt. They are available upon request to AH.

Two criteria were used for the choice of the taxonomic sample: (1) highly divergent mtDNA sequences, such as those produced for *Apis* (NC_001566), *Thrips* (NC_004371), or *Varroa* (NC_004454), were not included to facilitate protein alignments in order to retain more characters for the analyses; and (2) taxa, for which the 18S rRNA gene was not available in the databases, were also excluded (e.g., *Bombyx mori*).

*Analyses of the Nucleotide Composition for mtDNA Sequences*

For each of the 49 mt sequences, the nucleotide percentages were calculated at the synonymous third positions for three groups of codons: (1) the NNN group

TABLE 1.    Taxonomic sampling.

| NCBI classification | | | mtDNA taxa | Mitochondrial gene order type | Accession numbers | | 18S rDNA taxa |
|---|---|---|---|---|---|---|---|
| | | | | | mtDNA | 18S rDNA | |
| **Arthropoda** | | | | | | | |
| Chelicerata | Pycnogonida | | *Endeis spinosa* | Unknown | AY731173 | AF005441 | |
| | Uropygi | | *Mastigoproctus giganteus* | Unknown | AY731174 | AF005446 | |
| | Amblypygi | | *Phrynus sp.* | Unknown | AY731172 | F005445 | *Paraphrynus* |
| | Araneae | | *Argiope bruennichi* | Unknown | AY731171 | AF005447 | *Nesticus* |
| | Scorpiones | | *Euscorpius flavicaudis* | Unknown | AY731175 | X77908 | *Androctonus* |
| | Acari | | *Ornithodoros moubata* | *Limulus* | NC_004357 | L76355 | |
| | | | *Rhipicephalus sanguineus* | *Rhipicephalus* | NC_002074 | L76342 | |
| | | | *Ixodes hexagonus* | *Ixodes* | NC_002010 | L76351 | *I. cookei* |
| | Xiphosura | | *Limulus polyphemus* | *Limulus* | NC_003057 | L81949 | |
| Myriapoda | Chilopoda | | *Lithobius forficatus* | *Lithobius* | NC_002629 | AF000773 | *L. variegatus* |
| | Diplopoda | Spirobolida | *Narceus annularus* | *Narceus* | NC_003343 | AF062969 | *Spirobolus* |
| | | Spirostreptida | *Thyropygus sp.* | *Narceus* | NC_003344 | AY210829 | *Orthoporus* |
| Crustacea | Malacostraca | Dendrobranchiata | *Penaeus monodon* | *Drosophila* | NC_002184 | AF186250 | *P. vannamei* |
| | | Pleocyemata | *Pagurus longicarpus* | *Pagurus* | AF150756 | AF436018 | |
| | | Pleocyemata | *Panulirus japonicus* | *Drosophila* | NC_004251 | U19182 | *P. argus* |
| | Maxillopoda | | *Tigriopus japonicus* | *Tigriopus* | AB060648 | AF363306 | *T. californicus* |
| | Branchiopoda | Phyllopoda | *Daphnia pulex* | *Drosophila* | NC_000844 | AF014011 | |
| | | Phyllopoda | *Triops cancriformis* | *Drosophila* | NC_004465 | AF144219 | *T. longicaudatus* |
| | | Sarsostraca | *Artemia franciscana* | *Artemia* | NC_001620 | AJ238061 | |
| Insecta | Coleoptera | Elateriformia | *Pyrocoelia rufa* | *Drosophila* | NC_003970 | AF451941 | *Duliticola* |
| | | Cucujiformia | *Tribolium castaneum* | *Drosophila* | NC_003081 | X07801 | *Tenebrio* |
| | | Cucujiformia | *Crioceris duodecimpunctata* | *Drosophila* | NC_003372 | AF267426 | *C. asparagi* |
| | Diptera | Muscomorpha | *Drosophila melanogaster* | *Drosophila* | NC_001709 | M21017 | |
| | | Muscomorpha | *Ceratitis capitata* | *Drosophila* | NC_000857 | AF096450 | |
| | | Muscomorpha | *Chrysomya chloropyga* | *Chrysomya* | NC_002697 | AF322424 | *Melinda* |
| | | Nematocera | *Anopheles quadrimaculatus* | *Anopheles* | NC_000875 | L78065 | *A. albimanus* |
| | Hemiptera | | *Triatoma dimidiata* | *Drosophila* | NC_002609 | AJ243328 | |
| | Lepidoptera | | *Antheraea pernyi* | *Antheraea* | NC_004622 | AF286273 | *Hemileuca* |
| | | | *Ostrinia nubilalis* | *Antheraea* | NC_003367 | X89491 | *Galleria* |
| | Orthoptera | | *Locusta migratoria* | *Locusta* | NC_001712 | AF370793 | |
| | Phthiraptera | | *Heterodoxus macropus* | *Heterodoxus* | NC_002651 | AY077759 | *H. calabyi* |
| | Thysanura | | *Tricholepidion gertschi* | *Drosophila* | AY191994 | AF370789 | |
| Collembola | | | *Gomphiocephalus hodgsoni* | *Drosophila* | AY191995 | Z26765 | *Hypogastrura* |
| | | | *Tetrodontophora bielanensis* | *Tetrodontophora* | NC_002735 | AY037171 | *Onychiurus* |
| **Outgroup** | | | | | | | |
| Annelida | Oligochaeta | | *Lumbricus terrestris* | *Lumbricus* | NC_001673 | AJ272183 | |
| | Polychaeta | | *Platynereis dumerilii* | *Platynereis* | NC_000931 | Z83754 | *Nereis* |
| Mollusca | Polyplacophora | | *Katharina tunicata* | *Katharina* | NC_001636 | AY145380 | *Ischnochiton* |
| | Cephalopoda | | *Loligo bleekeri* | *Loligo* | NC_002507 | AY145383 | *Loligo pealei* |
| Chordata | Cephalochordata | | *Branchiostoma floridae* | *Branchiostoma* | NC_000834 | M97571 | |
| | Craniata | | *Bos taurus* | *Bos* | NC_001567 | M10098 | *Homo* |
| | | | *Petromyzon marinus* | *Petromyzon* | NC_001626 | M97575 | |
| | | | *Myxine glutinosa* | *Bos* | NC_002639 | M97574 | |
| Hemichordata | | | *Balanoglossus carnosus* | *Balanoglossus* | NC_001887 | D14359 | |
| Echinodermata | Eleutherozoa | Asteroidea | *Asterina pectinifera* | *Asterina* | NC_001627 | AB084551 | |
| | | | *Pisaster ochraceus* | *Pisaster* | NC_004610 | AF088804 | *Heliaster* |
| | | Echinoidea | *Arbacia lixula* | *Arbacia* | NC_001770 | Z37514 | |
| | | | *Paracentrotus lividus* | *Arbacia* | NC_001572 | AF279215 | *Psammechinus* |
| | | | *Strongylocentrotus purpuratus* | *Arbacia* | NC_001453 | L28055 | |
| | Pelmatozoa | Crinoidea | *Florometra serratissima* | *Florometra* | NC_001878 | AF088803 | *Dorometra* |

includes all fourfold degenerate codons at third position; (2) the NNR group includes all twofold degenerate codons with a purine (A or G) at third position; and (3) the NNY group includes all twofold degenerate codons with a pyrimidine (C or T) at third position. Because of variations in the mt genetic code of the Metazoa (Knight et al., 2001; Yokobori et al., 2001), the composition of NNN, NNR and NNY groups varies between Cephalochordata, Echinodermata + Hemichordata, Vertebrata, and other phyla of Metazoa (Annelida, Arthropoda, and Mollusca). The NNN group consists of the nine codons A, G, L2, P, R, S1, S2, T, and V, except for Chordata because of exclusion of S2; the NNR group comprises the six codons E, K, L1, M, Q and W, except for Echinodermata and Hemichordata because of exclusion of K and M; and the NNY group includes the eight codons C, D, F, H, I, N, S2, and Y, except for Annelida, Arthropoda, Cephalochordata, and Mollusca because of exclusion of S2, as well as for Echinodermata and Hemichordata because of exclusion of I, N and S2.

All the six protein-coding genes here examined (i.e., *atp6* and *atp8*, *cox1* to *cox3*, and *nad2*) are located on the same strand except for four genera: *Asterina*, *Florometra*, and *Pisaster,* for which *nad2* is inverted, and *Heterodoxus*, for which *atp6* and *atp8* are inverted. Because of these gene inversions, the nucleotide composition was arbitrarily examined for the strand containing the coding sequence of the *cox1* to *cox3* genes, which was constant in all species. For instance, the frequency of adenine at fourfold degenerate third codon positions was determined as follows for *Asterina*: the number of fourfold degenerate third codon positions ($N_1$) and the frequency of Adenine ($F_A$) were calculated in the sequence including *cox1* to *cox3*, *atp6*, and *atp8*; the number of fourfold degenerate third codon positions ($N_2$) and the frequency of Thymine ($F_T$) were caculated in the *nad2* gene; and the frequency of adenine in the complete mtDNA fragment was deduced by adding $(F_A N_1)/(N_1 + N_2)$ with $(F_T N_2)/(N_1 + N_2)$.

The strand bias in nucleotide composition was analyzed at third positions of NNN, NNR, and NNY codons by comparing the frequencies of complementary nucleotides, i.e., A (%) versus T (%), and C (%) versus G (%). A statistical test was used for testing the null hypothesis of strand symmetry, i.e., A (%) = T (%) or C (%) = G (%). For instance, the comparison between A and T frequencies was done by using the following formula: $U = |F_A - F_T|/\sqrt{[F(1 - F)(1/N_1 + 1/N_2)]}$, where $F_A$ and $F_T$ are the observed frequencies of adenine and thymine, $N_1$ and $N_2$ are the numbers of codons used for calculating respectively $F_A$ and $F_T$, and F is the weighted average, ie., $F = [(F_A N_1) + (F_T N_2)]/(N_1 + N_2)$. According to this test, if U is superior to 1.96, the null hypothesis of strand symmetry is rejected at confidence level 0.05 (95%). The strand bias was then described by skewness (Lobry, 1995; Perna and Kocher, 1995), which measures on one strand the relative number of As to Ts (AT skew = $[A - T]/[A + T]$) and Cs to Gs (CG skew = $[C - G]/[C + G]$). AT skews were considered to be statistically significant only when adenine and thymine frequencies are significantly different. Similarly, CG skews were considered to be statistically significant only when cytosine and guanine frequencies are significantly different.

Nucleotide composition was also analyzed at nonsynonymous sites by comparing the frequencies of codons that are fourfold degenerate at third positions, that differ at a single nonsynonymous position (first or second). In order to examine a high number of sites for statistics, only codons that code for easily interchangeable amino acid residues were compared (Naylor and Brown, 1997, 1998; Hassanin et al., 1998). Three pairs of codons were therefore compared: (1) ACN versus GCN, which only differ at first position, and code respectively for the amino acids T and A; (2) CTN versus GTN, which only differ at first position, and code respectively for L2 and V amino acids; and (3) GCN versus GTN, which only differ at second position, and code respectively for A and V amino acids. For instance, the relative frequencies of ACN and GCN codons were calculated as follows: (1) the original data matrix was transformed by replacing all codons, except

those of interest, by question marks; (2) the base frequencies were estimated under PAUP 4.0b10 by selecting only informative first codon positions, and after exclusion of *atp6*, *atp8*, and *nad2* genes owing to their inversion in *Heterodoxus*, *Asterina*, *Florometra*, or *Pisaster*.

### Reconstruction of Ancestral Mitochondrial Genome Organizations

In order to reconstruct the ancestral mitochondrial genome organization for several taxa of interest, each of the 44 complete mt genomes (i.e., all taxa except *Argiope*, *Endeis*, *Eusorpius*, *Mastigoproctus*, and *Phrynus*) was described by a matrix including 74 characters, corresponding to the 3′ and 5′ ends of each of the 37 mt genes. For each character, the states were coded by determining the 3′ or 5′ end of the neighboring genes.

Ancestral gene arrangements were inferred by MP analyses. Because gene rearrangements may be homoplastic due to the limited number of genes, we used a constraint tree analysis for inferring ancestral genomes. Heuristic searches were performed under PAUP 4.0b10 (Swofford, 2003), using 100 replicates of random stepwise addition of taxa, and by keeping only trees compatible with a constraint-tree named "Taxa," where all taxa listed in Table 1 were considered as being monophyletic, as well as Lophotrochozoa (i.e., Annelida + Mollusca), Deuterostomia, and Protostomia, and where Echinodermata and Hemichordata were assumed to be sister-groups as suggested by the literature (Bromham and Degnan, 1999; Cameron et al., 2000).

For each internal node of interest, the ancestral character-states were inferred by using either Acctran (accelerated transformation) or Deltran (delayed transformation) optimizations. Then, we performed a consensus sequence where character-states were coded as ambiguous ("?" in Appendix 1) in case of conflicting inferences between Acctran and Deltran optimizations. In a final procedure, the consensus sequences were used for reconstructing circular ancestral mt genomes. Some ambiguities in the ancestral sequences were resolved by this last procedure (see Results).

### Phylogenetic Analyses

Phylogenetic analyses were performed using maximum parsimony (MP), Bayesian, and maximum likelihood (ML) methods. MP analyses were carried out under PAUP 4.0b10 (Swofford, 2003). The MP tree was found by heuristic searches using default options but 100 replicates of random stepwise addition of taxa. Bootstrap proportions (BPs) were obtained after 1000 replicates by using 10 replicates of random stepwise addition of taxa. Bayesian analyses were conducted under MrBayes v3.0b4 (Huelsenbeck and Ronquist, 2001). The Bayesian approach combines the advantages of defining an explicit model of molecular evolution and of obtaining a rapid approximation of posterior probabilities of trees by use of Markov chain Monte Carlo (MCMC) (Huelsenbeck et al., 2001). MODELTEST 3.06 (Posada and Crandall, 1998) was used for choosing the model

of DNA substitution that best fits our data. The selected likelihood model was the General Time Reversible model (Yang, 1994) with among-site substitution rate heterogeneity described by a gamma distribution and a fraction of sites constrained to be invariable (GTR+I+$\Gamma$4). Two different variants of the model were used for the mt analyses: (i) a single GTR+I+$\Gamma$4 model for all sites; and (ii) a new method, named "Neutral Transitions Excluded," which codes purines by R and pyrimidines by Y at all third codon positions, at first positions of CTN (L2) and TTN (F and L1) codons, and at first and second positions of ACN (T), ATN (I and M), GCN (A), and GTN (V) codons. We used a GTR+I+$\Gamma$4 model for first and second codon positions, and a two-state substitution model I+$\Gamma$4 for third codon positions. All Bayesian analyses were done with five independent Markov chains run for 1,000,000 Metropolis-coupled MCMC generations, with tree sampling every 100 generations and a burn-in of 1000 trees. The analyses were run twice using different random starting trees to evaluate the convergence of the likelihood values and posterior clade probabilities (Huelsenbeck et al., 2002). BPs were also obtained under the ML method by using the program SEQBOOT in the PHYLIP package Version 3.6b (Felsenstein, 2004) for generating 100 bootstrapped data sets, and by analysing the latters with PHYML (Guindon and Gascuel, 2003).

## RESULTS

### Nucleotide Composition at Synonymous Third Codon Positions

The nucleotide compositions at synonymous third codon positions of the mt fragment including the six coding genes *atp6* and *atp8*, *cox1* to *cox3*, and *nad2* are indicated in Table 2 for each of the 49 taxa examined.

At twofold degenerate third codon positions, the analyses of NNR codons show that all taxa except *Tigriopus* have more adenine than guanine, and the analyses of NNY codons show that most taxa have more thymine than cytosine, with some exceptions like *Asterina*, *Balanoglossus*, *Bos*, and *Phrynus*. The comparisons between NNR and NNY codons reveal that the highest percentages were found for adenine, with the exception of seven genera, which exhibit higher values for thymine: *Argiope* (T = 94%; A = 77%), *Artemia* (T = 76%; A = 70%), *Branchiostoma* (T = 74%; A = 64%), *Euscorpius* (T = 91%; A = 64%), *Florometra* (T = 99%; A = 76%), *Katharina* (T = 88%; A = 66%), and *Tigriopus* (T = 74%; A = 47%).

At fourfold degenerate third codon positions, all taxa, except *Katharina* and *Tigriopus*, have more adenine than guanine, and all taxa have more thymine than cytosine, with the exceptions of *Asterina*, *Balanoglossus*, *Bos*, *Narceus*, and *Phrynus*. For most species, the highest percentage was found for Adenine, but it was not the case for 14 taxa: the highest percentage was found for Cytosine for *Balanoglossus*, and for thymine for *Antheraea*, *Artemia*, *Branchiostoma*, *Ceratitis*, *Daphnia*, *Euscorpius*, *Florometra*, *Gomphiocephalus*, *Heterodoxus*, *Katharina*, *Panulirus*, *Penaeus*, and *Tigriopus*. All taxa without exception are A+T rich rather than G+C rich. However, a very high A+T content (i.e., >90%) was found in *Florometra* and most insects, in particular for Lepidoptera, Diptera, Orthoptera (*Locusta*), and Phthiraptera (*Heterodoxus*), whereas the lowest values of A+T content (i.e., <75%) were found in annelids, *Katharina*, myriapods, *Phrynus*, several crustaceans (*Panulirus*, *Tigriopus*, and branchiopods), and all deuterostomes but *Florometra*.

### Strand Asymmetry in the Nucleotide Composition of mtDNA Sequences

*Strand compositional bias at synonymous third codon positions.*—For determining which taxa are characterized by a strand bias in nucleotide composition, the frequencies of complementary nucleotides (A versus T, or C versus G) were compared at synonymous third codon positions in order to know whether the hypothesis of strand symmetry is rejected or not at a confidence level of 0.05 (95%). At twofold degenerate third positions, the hypothesis of strand symmetry is rejected for all taxa, except *Artemia* and *Heterodoxus* (Table 2; underlined values of $AT_2$ and $CG_2$ skews). At four fold degenerate third positions, the hypothesis of strand symmetry is rejected for all taxa: cytosine and guanine frequencies are significantly different for all taxa, except *Artemia*, *Crioceris*, *Drosophila*, and *Locusta* (Table 2; underlined values of $CG_4$ skew), but the latter taxa present a significant difference between adenine and thymine frequencies (Table 2; underlined values of $AT_4$ skew). The strand bias is therefore conspicuous at both two- and fourfold degenerate third codon positions in all taxa, except in *Artemia* and *Heterodoxus*, for which there is no evidence for strand asymmetry.

*Evidence for global reversals of strand compositional bias.*—For each of the 49 taxa, AT and CG skews were calculated for twofold degenerate third codon positions (Table 2; $AT_2$ and $CG_2$ skews) and fourfold degenerate third codon positions (Table 2; $AT_4$ and $CG_4$ skews). Note that AT and CG skews are statistically significant only if the null hypothesis of symmetry, i.e., A (%) = T (%) or C (%) = G (%), is rejected. By considering only significant values of skew (underlined values in Table 2), it appears that most taxa are characterized by positive values for AT and CG skews, indicating that they present a strand compositional bias characterized by an excess of A relative to T nucleotides and of C relative to G nucleotides. However, eight taxa are characterized by significant negative values for $AT_2$, $AT_4$, $CG_2$, and/or $CG_4$ skews, implying that they present a reverse strand compositional bias, i.e., characterized by an excess of T relative to A nucleotides and of G relative to C nucleotides: *Argiope*, *Artemia*, *Branchiostoma*, *Euscorpius*, *Florometra*, *Heterodoxus*, *Katharina*, and *Tigriopus*. Because only one skew is significant for *Artemia* ($AT_4$) and *Heterodoxus* ($CG_4$), it cannot be definitively concluded that the strand bias is reversed for these two taxa. By contrast, the reverse bias is obvious for the other six genera: $AT_2$, $CG_2$, and $CG_4$ skews are significant and negative for *Argiope* and *Branchiostoma*, whereas all the four skews are significant and negative for *Euscorpius*, *Florometra*, *Katharina*, and *Tigriopus*.

TABLE 2. Nucleotide composition at synonymous and nonsynonymous sites.

| | Third codon positions (POS 3) | | | | | | | | | | | | | | POS 1 | | | | POS 2 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | NNR group | | NNY group | | Skew | Skew | NNN group | | | | | | Skew | Skew | ACN (vs GCN) | | CTN (vs GTN) | | GCN (vs GTN) | |
| | A % | Sites | T % | Sites | AT2 | CG2 | A % | C % | G % | T % | Sites | A+T % | AT4 | CG4 | % | Sites | % | Sites | % | Sites |
| Endeis | 88.50 | 339 | 65.49 | 397 | +0.15 | +0.50 | 46.86 | 13.24 | 5.92 | 33.97 | 574 | 80.84 | +0.16 | +0.38 | 66.67 | 21 | 41.18 | 17 | 50.00 | 10 |
| Mastigoproctus | 91.32 | 288 | 72.09 | 430 | +0.12 | +0.53 | 44.37 | 19.16 | 4.54 | 31.93 | 595 | 76.30 | +0.16 | +0.62 | 50.00 | 24 | 52.17 | 23 | 57.14 | 14 |
| Phrynus | 90.37 | 218 | 41.19 | 420 | +0.37 | +0.72 | 40.00 | 32.69 | 5.22 | 22.09 | 670 | 62.09 | +0.29 | +0.72 | 60.00 | 25 | 63.64 | 22 | 36.36 | 11 |
| Argiope | 77.10 | 358 | 94.24 | 382 | -0.10 | -0.60 | 44.34 | 2.78 | 10.39 | 42.49 | 539 | 86.83 | +0.02 | -0.58 | 47.06 | 17 | 11.11 | 27 | 17.39 | 23 |
| Euscorpius | 63.50 | 326 | 91.34 | 381 | -0.18 | -0.62 | 25.00 | 5.94 | 13.99 | 55.07 | 572 | 80.07 | -0.38 | -0.40 | 42.86 | 14 | 20.59 | 34 | 25.00 | 20 |
| Ornithodoros | 94.02 | 301 | 67.04 | 443 | +0.17 | +0.69 | 48.51 | 14.19 | 3.33 | 33.98 | 571 | 82.49 | +0.18 | +0.62 | 73.91 | 23 | 60.00 | 10 | 80.00 | 5 |
| Rhipicephalus | 93.18 | 337 | 85.93 | 469 | +0.04 | +0.35 | 47.74 | 7.66 | 2.36 | 42.24 | 509 | 89.98 | +0.06 | +0.53 | 73.33 | 15 | 37.50 | 8 | 42.86 | 14 |
| Ixodes | 92.76 | 290 | 63.38 | 467 | +0.19 | +0.67 | 45.44 | 18.96 | 4.65 | 30.95 | 559 | 76.39 | +0.19 | +0.61 | 76.92 | 13 | 50.00 | 12 | 40.00 | 5 |
| Limulus | 96.58 | 263 | 53.79 | 396 | +0.28 | +0.86 | 45.58 | 21.65 | 2.59 | 30.18 | 656 | 75.76 | +0.20 | +0.79 | 56.52 | 23 | 57.69 | 26 | 50.00 | 14 |
| Lithobius | 87.27 | 267 | 65.90 | 393 | +0.14 | +0.46 | 45.50 | 21.07 | 5.80 | 27.63 | 655 | 73.13 | +0.24 | +0.57 | 51.85 | 27 | 52.00 | 25 | 46.67 | 15 |
| Narceus | 91.50 | 247 | 52.63 | 380 | +0.27 | +0.70 | 44.48 | 31.40 | 2.76 | 21.37 | 688 | 65.84 | +0.35 | +0.84 | 37.04 | 27 | 35.71 | 28 | 57.90 | 19 |
| Thyropygus | 88.00 | 285 | 59.38 | 389 | +0.19 | +0.55 | 47.58 | 22.78 | 4.21 | 25.43 | 641 | 73.01 | +0.30 | +0.69 | 50.00 | 34 | 48.28 | 29 | 55.56 | 18 |
| Penaeus | 90.82 | 294 | 74.29 | 385 | +0.10 | +0.47 | 40.88 | 9.59 | 6.29 | 43.24 | 636 | 84.12 | -0.03 | +0.21 | 47.22 | 36 | 32.26 | 31 | 50.00 | 18 |
| Pagurus | 94.36 | 266 | 73.91 | 391 | +0.12 | +0.64 | 52.58 | 7.75 | 2.58 | 37.08 | 658 | 89.67 | +0.17 | +0.50 | 53.33 | 30 | 35.29 | 34 | 38.10 | 21 |
| Panulirus | 74.44 | 266 | 58.11 | 370 | +0.12 | +0.24 | 34.90 | 18.26 | 11.34 | 35.49 | 679 | 70.40 | -0.01 | +0.23 | 34.29 | 35 | 34.29 | 35 | 62.50 | 24 |
| Tigriopus | 47.30 | 296 | 73.76 | 343 | -0.22 | -0.34 | 25.22 | 10.39 | 28.04 | 36.35 | 674 | 61.57 | -0.18 | -0.46 | 31.82 | 22 | 36.11 | 36 | 42.86 | 21 |
| Daphnia | 68.47 | 241 | 56.77 | 384 | +0.09 | +0.16 | 29.28 | 22.61 | 15.94 | 32.17 | 690 | 61.45 | -0.05 | +0.17 | 47.22 | 36 | 43.24 | 37 | 61.11 | 18 |
| Triops | 88.09 | 277 | 68.67 | 399 | +0.12 | +0.45 | 43.66 | 15.96 | 7.20 | 33.18 | 639 | 76.84 | +0.14 | +0.38 | 46.88 | 32 | 46.43 | 28 | 55.56 | 18 |
| Artemia | 70.00 | 280 | 76.13 | 398 | -0.04 | -0.11 | 31.05 | 14.65 | 13.85 | 40.45 | 628 | 71.50 | -0.13 | +0.03 | 22.73 | 22 | 46.67 | 30 | 52.17 | 23 |
| Pyrocoelia | 96.40 | 333 | 82.41 | 432 | +0.08 | +0.66 | 61.82 | 7.27 | 2.73 | 28.18 | 550 | 90.00 | +0.37 | +0.45 | 63.16 | 19 | 50.00 | 12 | 71.43 | 7 |
| Tribolium | 88.89 | 261 | 56.65 | 406 | +0.22 | +0.59 | 56.79 | 15.74 | 4.17 | 23.30 | 648 | 80.09 | +0.42 | +0.58 | 56.25 | 32 | 60.00 | 20 | 72.73 | 11 |
| Crioceris | 92.70 | 315 | 80.10 | 412 | +0.07 | +0.46 | 48.13 | 7.82 | 6.29 | 37.76 | 588 | 85.88 | +0.12 | +0.11 | 56.67 | 30 | 38.89 | 18 | 50.00 | 10 |
| Drosophila | 98.26 | 345 | 90.32 | 413 | +0.04 | +0.70 | 52.07 | 2.51 | 2.51 | 42.91 | 557 | 94.97 | +0.10 | 0.00 | 44.44 | 27 | 21.05 | 19 | 47.06 | 17 |
| Ceratitis | 97.52 | 323 | 85.00 | 400 | +0.07 | +0.72 | 45.61 | 4.56 | 2.03 | 47.80 | 592 | 93.41 | -0.02 | +0.38 | 39.29 | 28 | 38.89 | 18 | 42.86 | 21 |
| Chrysomya | 93.98 | 332 | 76.66 | 407 | +0.10 | +0.59 | 50.52 | 1.91 | 0.35 | 47.22 | 576 | 97.74 | +0.03 | +0.69 | 48.28 | 29 | 25.00 | 20 | 40.00 | 20 |
| Anopheles | 95.85 | 337 | 84.90 | 404 | +0.06 | +0.57 | 55.75 | 6.27 | 2.61 | 35.37 | 574 | 91.12 | +0.22 | +0.41 | 48.49 | 33 | 27.78 | 18 | 43.75 | 16 |
| Triatoma | 91.76 | 279 | 58.03 | 417 | +0.23 | +0.67 | 54.44 | 17.45 | 3.23 | 24.88 | 619 | 79.32 | +0.37 | +0.69 | 51.72 | 29 | 62.50 | 24 | 53.33 | 15 |
| Antheraea | 97.40 | 308 | 90.06 | 483 | +0.04 | +0.59 | 42.18 | 9.92 | 2.29 | 45.61 | 524 | 87.79 | -0.04 | +0.63 | 59.09 | 22 | 50.00 | 8 | 40.00 | 10 |
| Ostrinia | 98.80 | 334 | 88.65 | 467 | +0.05 | +0.81 | 61.09 | 5.06 | 1.75 | 32.10 | 514 | 93.19 | +0.31 | +0.49 | 65.22 | 23 | 12.50 | 8 | 44.44 | 9 |
| Locusta | 96.65 | 328 | 71.83 | 394 | +0.15 | +0.79 | 72.68 | 4.38 | 2.36 | 20.57 | 593 | 93.25 | +0.56 | +0.30 | 48.28 | 29 | 42.86 | 14 | 50.00 | 14 |
| Heterodoxus | 91.82 | 379 | 88.64 | 440 | +0.02 | +0.16 | 46.61 | 2.05 | 4.52 | 46.82 | 487 | 93.43 | 0.00 | -0.38 | 71.43 | 7 | 22.22 | 18 | 20.00 | 10 |
| Tricholepidion | 90.67 | 300 | 70.65 | 385 | +0.12 | +0.52 | 58.10 | 14.76 | 6.98 | 20.16 | 630 | 78.25 | +0.48 | +0.36 | 60.71 | 28 | 52.38 | 21 | 50.00 | 16 |
| Gomphiocephalus | 90.97 | 299 | 78.69 | 413 | +0.07 | +0.40 | 42.45 | 8.29 | 5.14 | 44.11 | 603 | 86.57 | -0.02 | +0.23 | 56.00 | 25 | 26.32 | 19 | 42.11 | 19 |
| Tetrodontophora | 97.20 | 286 | 70.71 | 420 | +0.16 | +0.83 | 43.84 | 14.12 | 4.60 | 37.44 | 609 | 81.28 | +0.08 | +0.51 | 73.68 | 19 | 47.37 | 19 | 46.67 | 15 |
| Lumbricus | 80.69 | 233 | 53.60 | 375 | +0.20 | +0.41 | 40.45 | 22.07 | 11.03 | 26.45 | 707 | 66.90 | +0.21 | +0.33 | 42.11 | 38 | 48.78 | 41 | 68.18 | 22 |
| Platynereis | 82.13 | 263 | 60.31 | 383 | +0.15 | +0.38 | 41.20 | 20.30 | 12.78 | 25.71 | 665 | 66.92 | +0.23 | +0.23 | 47.50 | 40 | 42.42 | 33 | 57.90 | 19 |
| Katharina | 66.07 | 336 | 88.00 | 411 | -0.14 | -0.48 | 21.80 | 4.16 | 22.63 | 51.41 | 601 | 73.21 | -0.40 | -0.69 | 32.00 | 25 | 19.23 | 26 | 44.00 | 25 |
| Loligo | 91.86 | 295 | 67.15 | 350 | +0.16 | +0.60 | 40.03 | 18.45 | 3.46 | 38.06 | 607 | 78.09 | +0.03 | +0.68 | 57.58 | 33 | 53.85 | 26 | 52.94 | 17 |
| Branchiostoma | 63.75 | 309 | 74.00 | 377 | -0.07 | -0.16 | 33.07 | 7.47 | 23.53 | 35.93 | 629 | 69.00 | -0.04 | -0.52 | 35.29 | 34 | 13.79 | 29 | 36.00 | 25 |
| Bos | 88.51 | 261 | 42.97 | 375 | +0.35 | +0.66 | 48.15 | 27.03 | 5.61 | 19.20 | 677 | 67.36 | +0.43 | +0.66 | 61.29 | 31 | 59.38 | 32 | 46.67 | 15 |
| Petromyzon | 94.38 | 267 | 57.07 | 408 | +0.25 | +0.77 | 45.17 | 20.65 | 2.68 | 31.50 | 673 | 76.67 | +0.18 | +0.77 | 52.27 | 44 | 57.14 | 28 | 71.43 | 21 |
| Myxine | 87.36 | 269 | 51.47 | 375 | +0.26 | +0.59 | 36.11 | 27.00 | 9.26 | 27.63 | 637 | 63.74 | +0.13 | +0.49 | 60.00 | 20 | 51.85 | 27 | 47.06 | 17 |
| Balanoglossus | 76.38 | 127 | 27.92 | 240 | +0.46 | +0.51 | 32.46 | 35.07 | 7.56 | 24.90 | 767 | 57.37 | +0.13 | +0.65 | 34.88 | 43 | 66.67 | 36 | 60.00 | 20 |
| Asterina | 89.44 | 161 | 38.43 | 255 | +0.40 | +0.71 | 43.17 | 27.77 | 7.05 | 22.01 | 695 | 65.18 | +0.32 | +0.60 | 51.52 | 33 | 58.82 | 34 | 61.11 | 18 |
| Pisaster | 91.06 | 179 | 54.69 | 256 | +0.25 | +0.67 | 39.70 | 20.75 | 4.51 | 35.04 | 665 | 74.74 | +0.06 | +0.64 | 50.00 | 32 | 45.16 | 31 | 55.56 | 18 |
| Arbacia | 77.22 | 180 | 62.60 | 246 | +0.10 | +0.24 | 37.90 | 18.81 | 8.46 | 34.84 | 686 | 72.74 | +0.04 | +0.38 | 30.56 | 36 | 46.88 | 32 | 56.00 | 25 |
| Paracentrotus | 79.62 | 157 | 46.94 | 245 | +0.26 | +0.44 | 46.54 | 21.02 | 7.05 | 25.39 | 709 | 71.93 | +0.29 | +0.50 | 30.56 | 36 | 52.50 | 40 | 56.00 | 25 |
| Strongylocentrotus | 67.58 | 182 | 48.97 | 243 | +0.16 | +0.22 | 36.07 | 25.37 | 9.38 | 29.18 | 682 | 65.25 | +0.11 | +0.46 | 33.33 | 33 | 44.83 | 29 | 47.37 | 19 |
| Florometra | 76.34 | 224 | 98.89 | 271 | -0.13 | -0.91 | 15.14 | 1.83 | 4.49 | 78.54 | 601 | 93.68 | -0.68 | +0.42 | 30.00 | 20 | 9.68 | 31 | 40.74 | 27 |

Underlined values of skew are statistically significant (see Materials and Methods). Significant negative values of skew are highlighted in grey.
NNR group: all twofold degenerate codons with a purine (A or G) at third position.
NNY group: all twofold degenerate codons with a pyrimidine (C or T) at third position.
NNN group: all fourfold degenerate codons at third position.
AT2 and CG2 skews were calculated for twofold degenerate codons. AT4 and CG4 skews were calculated for fourfold degenerate codons.
AT skew = [A (%) − T (%)]/[A (%) + T (%)] CG skew = [C (%) − G (%)]/[C (%) + G (%)].
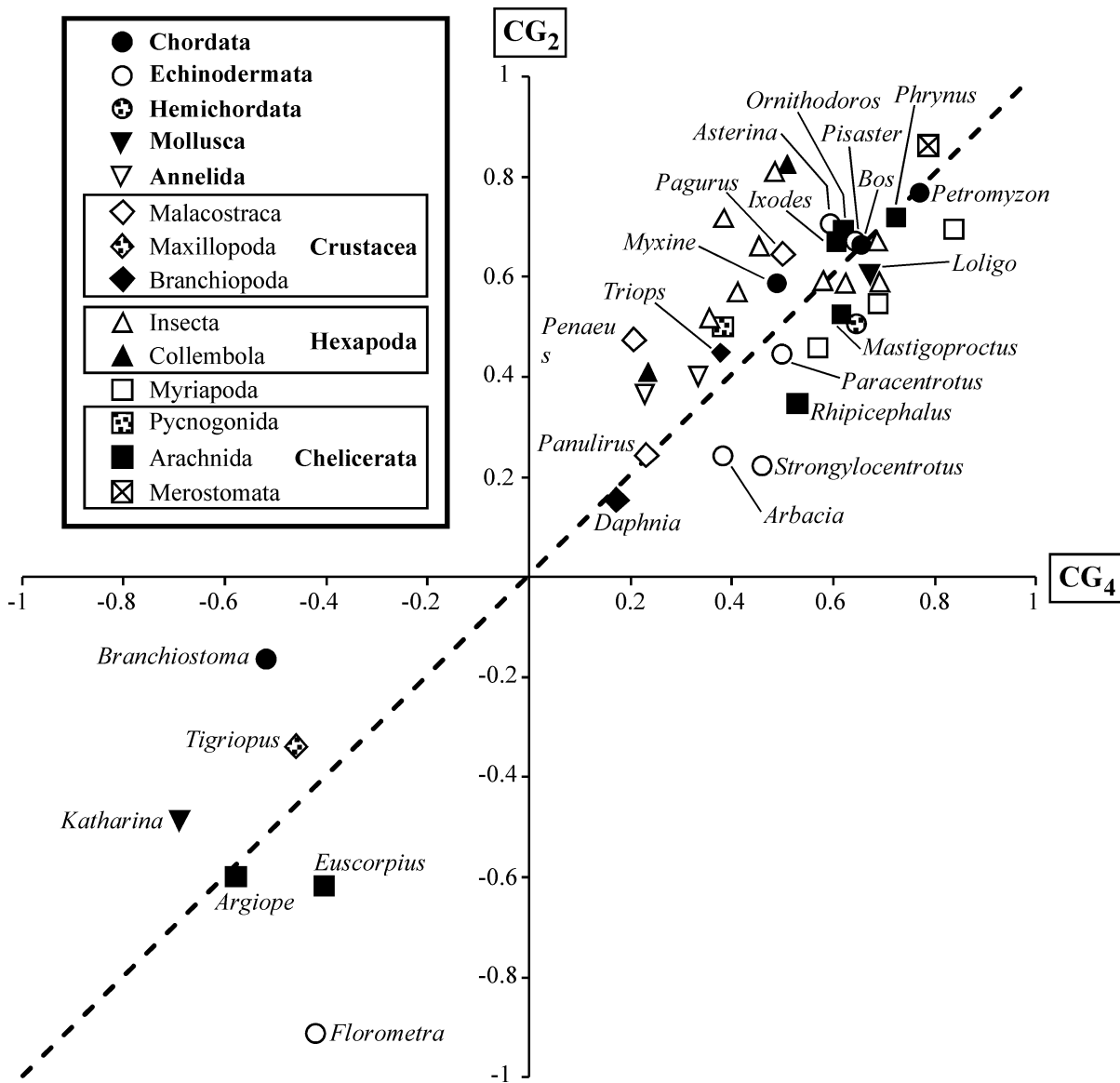
FIGURE 1. CG skews calculated for each taxa at four- and twofold degenerate third codon positions, in abscissa and ordinate, respectively.

The comparisons between statistically significant AT and CG skews (underlined values of skew in Table 2) reveals that absolute values are always higher for CG than for AT skews, with the exceptions of *Florometra* and *Tricholepidion*, for which the $AT_4$ skew is higher than the $CG_4$ skew. In addition, statistically significant values are more numerous for CG than for AT skews. These comparisons suggest therefore that CG skews are the best indicators of strand asymmetry.

The $CG_2$ skews were plotted against the $CG_4$ skews for all species presenting significant values for both CG skews, i.e., all taxa expect *Artemia, Crioceris, Drosophila, Heterodoxus*, and *Locusta* (Fig. 1). All species fall into two groups: the first one includes the six genera with a reverse strand bias, i.e., presenting a negative skew for both two- and fourfold degenerate sites: *Argiope, Branchiostoma, Euscorpius, Florometra, Katharina*, and *Tigriopus*; and the sec-

ond one includes all other species, which are characterized by a positive skew for both two- and fourfold degenerate sites. Interestingly, most points are close to the $y = x$ straight line. This result suggests that two- and fourfold degenerate third codon positions are similarly affected by strand compositional bias. Because transversions are synonymous at fourfold degenerate third codon positions, but would result in amino acid changes in twofold degenerate third codon positions, this result implies that the strand bias is mainly generated by mutations corresponding to transitions rather than transversions.

*Detection of the reverse strand bias at nonsynonymous positions.*—To test whether the reverse strand bias observed for six taxa at synonymous sites (i.e., *Argiope, Branchiostoma, Euscorpius, Florometra, Katharina,* and *Tigriopus*) is also observed at nonsynonymous sites, we compared the frequencies between codons that differ

at a single non-synonymous position (first or second) and that code for similar amino acids (Table 2). Because distant taxa are expected to present important differences in the genetic code and selective constraints, codon frequencies were only compared between closely related taxa. For this reason, the comparisons were limited to Arachnida for *Argiope* and *Euscorpius*, to Mollusca for *Katharina*, to Chordata for *Branchiostoma*, and to Echinodermata for *Florometra*. The case of *Tigriopus* was not treated because its phylogenetic position within Pancrustacea remains ambiguous. When compared to their closely related taxa, *Argiope*, *Euscorpius*, *Branchiostoma*, *Florometra*, and *Katharina*, exhibit very atypical codon frequencies: (1) They are biased against ACN over GCN, when codons specifying for T and A amino acids are compared: 47% and 42% for *Argiope* and *Euscorpius*, respectively, versus 50% to 77% for other arachnids; 35% for *Branchiostoma* versus 52% to 61% for other chordates; 30% for *Florometra* versus 31 to 52% for Eleutherozoa; and 32% for *Katharina* versus 58% for *Loligo*. (2) They are biased against CTN over GTN when codons specifying for L2 and V amino acids are compared: 11 and 21% for *Argiope* and *Euscorpius*, respectively, versus 38% to 64% for other arachnids; 14% for *Branchiostoma* versus 52% to 59% for other chordates; 10% for *Florometra* versus 45% to 53% for Eleutherozoa; and 19% for *Katharina* versus 54% for *Loligo*. (3) They are biased against GCN over GTN, when codons specifying for A and V amino acids are compared: 17% and 25% for *Argiope* and *Euscorpius*, respectively, versus 36% to 80% for all other arachnids; 36% for *Branchiostoma* versus 47% to 71% for other chordates; 41% for *Florometra* versus 47% to 61% for Eleutherozoa; and 44% for *Katharina* versus 53% for *Loligo*. The results suggest therefore that *Argiope*, *Branchiostoma*, *Euscorpius*, *Florometra*, and *Katharina* present a reverse strand bias, which can be observed not only at synonymous positions but also at nonsynonymous positions.

*Strand asymmetry and gene inversion.*—By assuming that the two mtDNA strands evolve under opposite asymmetric mutational constraints, a gene inversion is expected to produce a reversal of mutational patterns and with time, mutations are expected to completely reverse the strand compositional bias at synonymous positions. In other words, two genes encoded by two opposite strands are expected to have reverse strand biases. This assumption was confirmed by analyzing the nucleotide composition of *Asterina*, *Florometra*, and *Pisaster*. These three species present a clear strand bias (see underlined values of skew in Table 2), and are characterized by an inversion of nad2 with respect to the other genes: *atp6* and *atp8* and *cox1* to *cox3*. As expected, the analyses of two- and fourfold degenerate third codon positions indicate that *nad2* presents a reverse bias (Table 3): for *Asterina* and *Pisaster*, AT and CG skews are negative in *nad2*, but positive in *atp6* and *atp8* and *cox1* to *cox3* genes; for *Florometra*, AT and CG skews are positive in *nad2*, but negative in *atp6* and *atp8* and *cox1* to *cox3* genes. In the case of *Florometra*, the trends are reversed because of the global reversal of strand asymmetry (see above). These results clearly indicate that genes encoded by different strands are affected by reversed asymmetric mutational constraints.

### Ancestral Mitochondrial Genome Organizations

The mt genome organization was studied by MP analysis using the matrix of 74 characters shown in Appendix 1. Of 74 total characters, 71 are parsimony-informative. By keeping only trees compatible with the constraint-tree named "Taxa" (see Material and Methods), 38 equiparsimonious trees of 589 steps were found (CI = 0.90; RI = 0.88). The strict consensus of the 38 trees is identical to the constraint-tree (not shown). Each of these 38 trees was used for determining the sequence of character-states for the common ancestors of Chelicerata, Branchiopoda, Insecta, Pancrustacea, Mollusca, Chordata, Echinodermata, Eleutherozoa, and Asteroidea. Each ancestral sequence of 74 states presented in Appendix 1 is a consensus of the 76 ancestral sequences deduced from the analyses of each of the 38 MP trees by using either Acctran or Deltran optimizations. The deduced ancestral

TABLE 3.   Nucleotide composition of genes encoded by opposite strands.

| | NNR group | | NNY group | | Skew AT2 | Skew CG2 | NNN group | | | | | Skew AT4 | Skew CG4 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | A % | sites | T % | Sites | | | A % | C % | G % | T % | Sites | | |
| *Asterina atp6-8 + cox1-3* | 90.30 | 134 | 37.22 | 223 | +0.42 | +0.73 | 43.04 | 28.15 | 6.55 | 22.26 | 611 | +0.32 | +0.62 |
| nad2 | 46.88 | 32 | 85.19 | 27 | −0.29 | −0.56 | 20.24 | 10.71 | 25.00 | 44.05 | 84 | −0.37 | −0.40 |
| *Pisaster atp6-8 + cox1-3* | 91.28 | 149 | 56.16 | 219 | +0.24 | +0.67 | 38.84 | 21.30 | 4.60 | 35.26 | 587 | +0.05 | +0.64 |
| nad2 | 45.95 | 37 | 90.00 | 30 | −0.32 | −0.69 | 33.33 | 3.85 | 16.67 | 46.15 | 78 | −0.16 | −0.62 |
| *Florometra atp6-8 + cox1-3* | 75.61 | 205 | 98.79 | 248 | −0.13 | −0.91 | 13.99 | 2.08 | 4.35 | 79.58 | 529 | −0.70 | −0.35 |
| nad2 | 100.00 | 23 | 84.21 | 19 | +0.09 | +1.00 | 70.83 | 5.56 | 0.00 | 23.61 | 72 | +0.5 | +1.00 |

Underlined values of skew are statistically significant (see Materials and Methods). Significant negative values of skew are highlighted in grey.
NNR group: all twofold degenerate codons with a purine (A or G) at third position.
NNY group: all twofold degenerate codons with a pyrimidine (C or T) at third position.
NNN group: all fourfold degenerate codons at third position.
AT2 and CG2 skews were calculated for twofold degenrate codons. AT4 and CG4 skews were calculated for fourfold degenrate codons.
AT skew = [A (%) − T (%)]/[A (%) + T (%)].
CG skew = [C (%) − G (%)]/[C (%) + G (%)].

organization of Chelicerata is exactly the same that of *Limulus*; those of Branchiopoda, Insecta, and Pancrustacea are identical to that of *Drosophila*; and the one of Asteroidea is identical to that of *Pisaster* (not shown). For Chordata, Echinodermata, Eleutherozoa, and Mollusca, the states of several characters were found to be different between Acctran and Deltran optimizations. Several ambiguities were however solved after taking into account the circularity of the mtDNA genome. For instance, in the case of Eleutherozoa, the states of characters 10 and 29 were found ambiguous by MP analysis (Eleutherozoa-MP-A, Appendix 1): they correspond respectively to the 3′ end of the *rrnL* gene (3rL, Appendix 1), and to the 5′ end of the *cox1* gene (5c1, Appendix 1). After genome reconstruction, the states of these two characters were found unambiguous (Eleutherozoa-GR-U, Appendix 1), because the only way to produce a circular genome is to join the 3′ end of the *rrnL* gene with the 5′ end of the *cox1* gene. The deduced arrangement is identical to the one observed in *Arbacia* and all other Echinoidea.

### Nucleotide Composition and Mitochondrial Gene Order Organization

All taxa with a reverse strand bias display an unusual gene order organization of the mt genome, and interestingly, the position of the control region is not conserved by comparison with their close relatives.

In *Florometra*, the control region is located between *T*- and *D*-tRNA genes, whereas it is between *T*-tRNA and *rrnS* in Asteroidea, or between *T*- and *P*-tRNA genes in Echinoidea. However, all echinoderms have in common a genomic fragment, including *F*-tRNA, *rrnS*, *Q*-tRNA, *T*-tRNA and the control region (CR), where all genes are 5′ → 3′ oriented. The fragment [*F*-*rrnS*-*Q*-*T*-CR] is oriented as the *cox1* to *cox3* genes in the respective common ancestors of Echinodermata and Eleutherozoa (Fig. 2). By contrast, its orientation is inverted with respect to the *cox1* to *cox3* genes in *Florometra*, indicating without any ambiguity that an inversion of the control region occurred in the lineage leading to *Florometra*. We suggest that this event is responsible for the reverse strand bias observed in this genus.

For the other taxa concerned by a reverse asymmetry, it is not possible to know exactly what gene rearrangements occurred, but an inversion of the control region is highly probable because its position has changed by comparison with closely related taxa. The control region of *Branchiostoma* is flanked by *nad5* and *G*-tRNA (Boore et al., 1999), whereas it is located between *P*- and *F*-tRNA genes in the ancestral genome of Chordata (Appendix 1). The control region of *Katharina* could be either in the largest unassigned sequence of 424 nt between *D*-tRNA and *cox2*, or eventually in the second largest unassigned sequence of 141 nt between *E*-tRNA and *cox3* (Boore and Brown, 1994). Although the position of the control region could not be inferred in *Loligo*, due to the presence of multiple large noncoding regions (Tomita et al., 2002), it is clear that it is not positioned as in *Katharina* because *D*-tRNA, *cox2*, *E*-tRNA, and *cox3* are differently arranged in

*Loligo*. The control region of *Tigriopus* is located between *W*-tRNA and *cox1* genes (Machida et al., 2002), whereas it is found between *rrnS* and *I*-tRNA in the ancestral genomes of Crustacea and Pancrustacea, with *rrnS* inverted with respect to *I*-tRNA. In *Tigriopus*, *rrnS* and *I*-tRNA present a different location and are in the same orientation. For *Argiope* and *Euscorpius*, the position of the control region is not known because the mtDNA has not been entirely sequenced.

*Artemia* and *Heterodoxus* are the sole taxa that do not exhibit a clear strand bias. Interestingly, both display an unusual gene order organization of the mt genome, with a control region not positioned as observed in their close relatives. The mt genome of *Artemia* is very similar to the one inferred for the common ancestor of Branchiopoda. However, its control region is not placed between *rrnS* and *I*-tRNA, but between *rrnS* and *M*-tRNA, and the *I*-tRNA gene is inverted and positioned between *W*- and *Q*-tRNA genes (Garesse et al., 1997). The arrangement of genes in the mt genome of *Heterodoxus* is very different from the one reconstructed for the common ancestor of Insecta. In particular, its control region is not positioned between *rrnS* and *I*-tRNA, because it could be either in the largest unassigned sequence of 73 nt between *atp8* and *Q*-tRNA, or eventually in the second largest unassigned sequence of 47 nt between *cox2* and *nad3* (Shao et al., 2001).

### Phylogenetic Analyses

*Evidence for long-branch attraction artifacts.*—The mtDNA data matrix including 3948 nt characters and 49 taxa was first analyzed by the MP method. The most-parsimonious tree of 37,369 steps obtained (Fig. 3) is characterized by very high levels of homoplasy (CI = 0.1868 and RI = 0.3038). Taking into account the background knowledge in metazoan classification and phylogeny, seven taxa present odd positions. The louse *Heterodoxus* finds its place within a group of chelicerates, although this grouping is not supported. Six genera are grouped together in spite of their known distant relationships (box in Fig. 3): *Argiope* (Chelicerata, Araneae), *Euscorpius* (Chelicerata, Scorpiones), *Tigriopus* (Crustacea), *Katharina* (Mollusca), *Branchiostoma* (Chordata), and *Florometra* (Echinodermata). Interestingly, these six genera exhibit a very unusual base composition by comparison with other metazoans. They present a strand compositional bias characterized by an excess of thymine relative to adenine, and of guanine relative to cytosine. This bias is the reverse of what is observed in most other taxa where adenine is in excess relative to Thymine, and where cytosine is in excess relative to guanine (Table 2).

The tree performed with Bayesian inferences using the GTR+I+Γ4 model is more in agreement with what we know about Metazoan phylogeny (Fig. 4). Several taxa, which were found para- or polyphyletic in the MP analysis, are now monophyletic: Annelida (Bayesian posterior probability: $PP_B = 1$; $BP_{ML} = 100$), Arthropoda ($PP_B = 1$; $BP_{ML} = 50$), Chordata ($PP_B = 0.97$; not found with ML), Echinodermata ($PP_B = 1$; $BP_{ML} = 100$), Mollusca
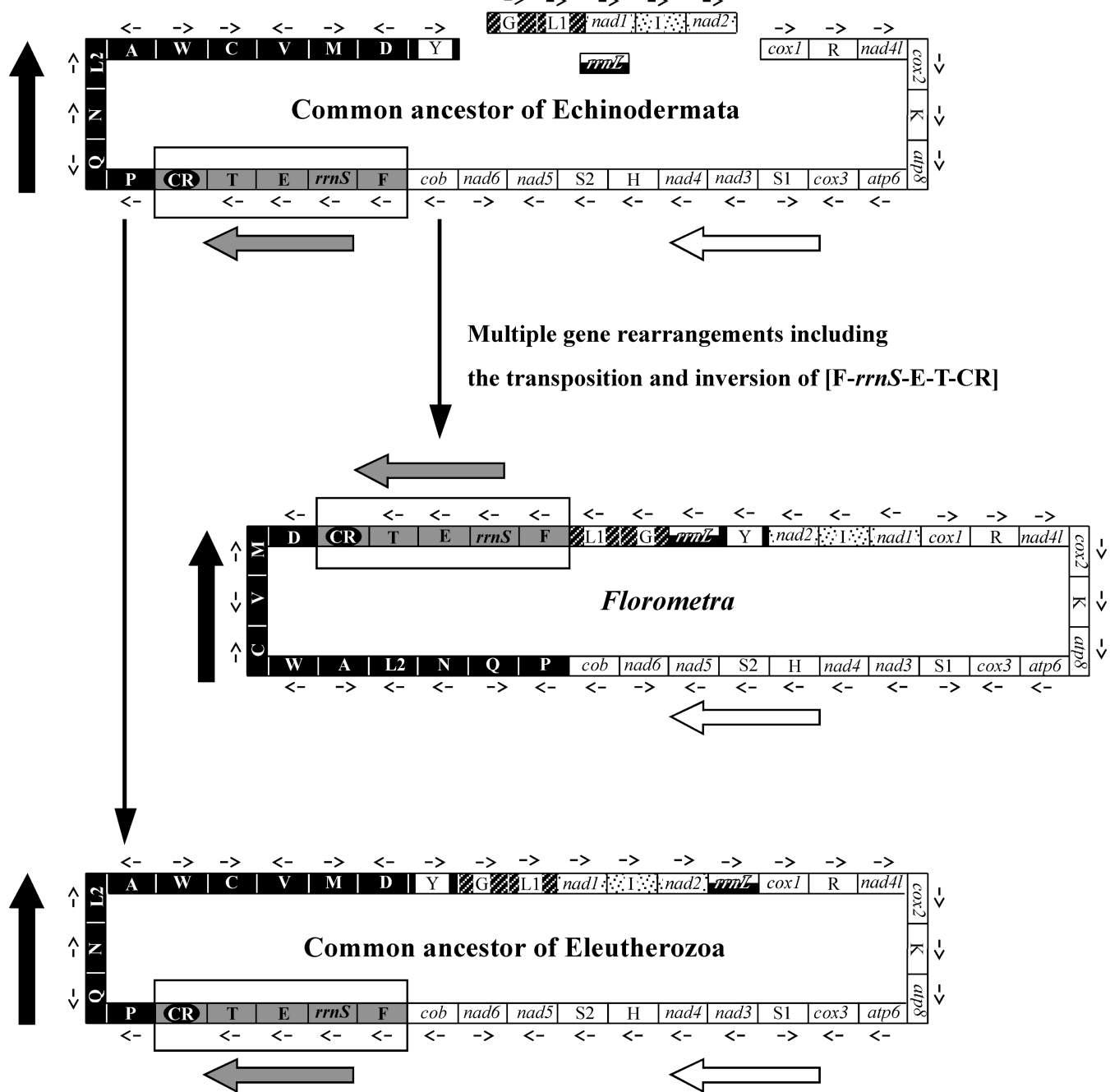
FIGURE 2.   Inversion of the control region during the evolution of Echinoderms. Small arrows indicate the $5' \rightarrow 3'$ orientation of the genes. Large arrows indicate the relative orientations of the three major fragments conserved in all Echinoderms: (1) the black fragment contains 10 tRNA genes (P, Q, N, L2, A, W, C, V, M, and D); (2) the white fragment includes five tRNA genes (R, K, S2, H, and S1) and 11 protein-coding genes (*cox1*, *nad4l*, *cox2*, *atp8*, *atp6*, *cox3*, *nad3*, *nad4*, *nad5*, *nad6*, and *cob*); and (3) the grey fragment includes the control region (CR), three tRNA genes (T, E, and F), and the *rrnS* gene. In *Florometra*, the grey fragment has been inverted with respect to the white fragment. This implies that the control region, which belongs to the grey fragment, has been inverted in *Florometra* with respect to the white fragment, which is the one used for analyzing the strand bias in nucleotide composition.

($PP_B = 1$; $BP_{ML} = 52$), Lophotrochozoa ($PP_B = 1$; $BP_{ML} = 52$), Myriapoda ($PP_B = 0.66$; $BP_{ML} = 37$), Deuterostomia/Protostomia ($PP_B = 1$; $BP_{ML} = 79$), Hexapoda ($PP_B = 0.55$; not found with ML), and Insecta ($PP_B = 0.88$; not found with ML). On the other side, Arachnida, Chelicerata, Crustacea, and Pancrustacea remain polyphyletic due to the grouping of three unrelated genera (box in

Fig. 4; $PP_B = 0.99$; $BP_{ML} = 68$): *Argiope* (Chelicerata, Araneae), *Euscorpius* (Chelicerata, Scorpiones), and *Tigriopus* (Crustacea). Each of these three latter genera is associated with a very long branch, suggesting that they are grouped together due to a long-branch attraction artifact. More generally, all taxa with reversed strand bias, i.e., *Argiope*, *Branchiostoma*, *Euscorpius*, *Florometra*,
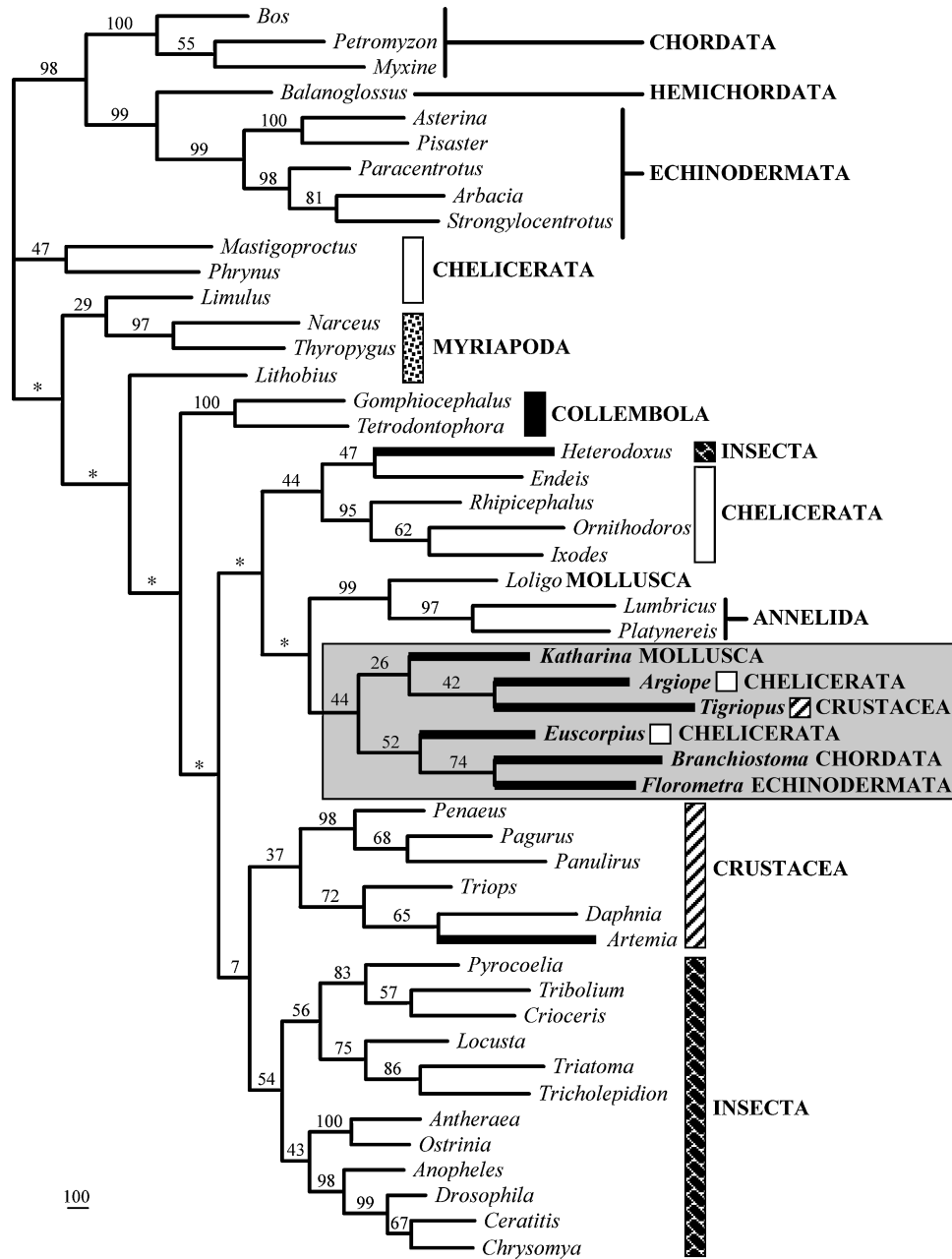
FIGURE 3. Most-parsimonious tree obtained with all the 49 taxa. Bold lines indicate branches of the taxa, for which asymmetric mutational constraints had been reversed during their evolutionary history, and taxa enclosed into the box are characterized by a completely reverse strand bias. Asterisks indicate that the node was not retrieved by the bootstrap analysis.

*Katharina*, and *Tigriopus*, are long-branched in comparison with their close relatives. *Artemia* and *Heterodoxus* are also associated with very long branches, as well as most chelicerates.

*Exclusion of the taxa with reversed asymmetric mutational constraints.*—We have also performed phylogenetic analyses on a reduced taxa sampling, including only the taxa in which the six genes studied, i.e *atp6* and *atp8*, *cox1* to *cox3*, and *nad2*, are transcribed on the same strand characterized by an excess of adenine relative to thymine and of cytosine relative to guanine as in other taxa. The Bayesian

tree performed by using the GTR+I+Γ4 model (Fig. 5A) indicates that several taxa, which were previously found polyphyletic, are now monophyletic: Chelicerata ($PP_B = 1$; $BP_{ML} = 55$), Crustacea ($PP_B = 1$; $BP_{ML} = 97$), and Pancrustacea ($PP_B = 1$; $BP_{ML} = 97$).

For comparison, we performed a Bayesian analysis on the basis of the complete sequences of the 18S rRNA gene (Fig. 5B). The *18S* tree is similar to the mtDNA tree, but some nodes are in conflict: (1) the squid *Loligo* appears with a long branch as the sister-group of the arthropods ($PP_B = 0.97$; $BP_{ML} = 64$), whereas mtDNA sequences
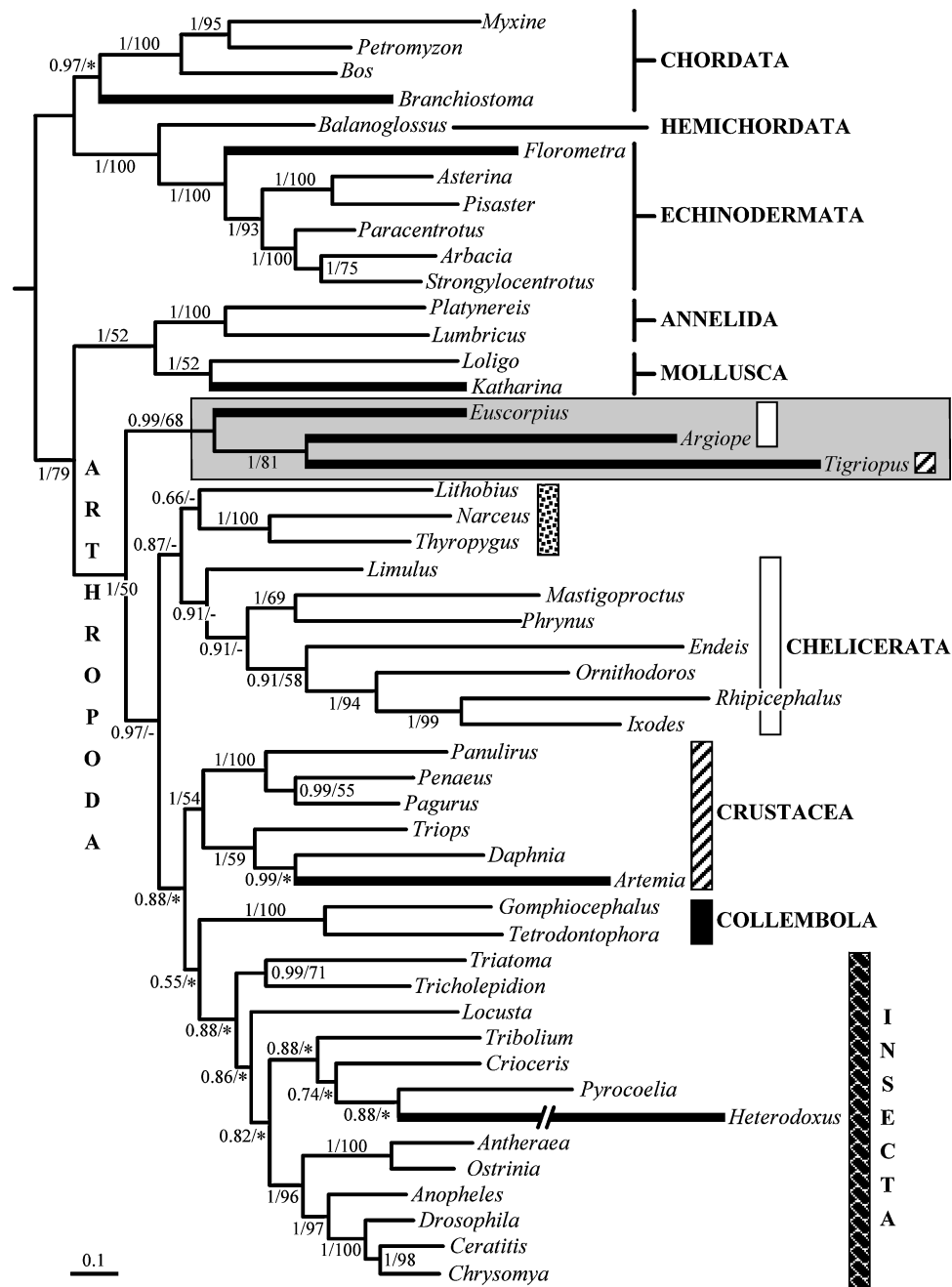
FIGURE 4. Bayesian tree performed with all the 49 taxa. The model used is the one selected by MODELTEST 3.06, i.e., GTR+I+Γ4. Bold lines indicate branches of the taxa, for which asymmetric mutational constraints had been reversed during their evolutionary history. Note that the branch length of *Heterodoxus* is three times as long as represented in the tree. The values indicated on the branches correspond to the posterior probabilities (to the left of the slash) obtained with the Bayesian analysis, and to the bootstrap proportions (BP) obtained with the maximum likelihood analysis (to the right of the slash). Dash indicates that the node was not supported by a BP value superior to 50. Asterisk indicates that an alternative hypothesis was supported by a BP value greater than 50.

agree with the monophyly of Lophotrochozoa as *Loligo* is associated with annelids (PP$_B$ = 1; BP$_{ML}$ = 100); (2) branchiopods (*Daphnia* and *Triops*) occupy a basal position within Pancrustaceans (PP$_B$ = 0.91; BP$_{ML}$ = 32), whereas they are grouped with other crustaceans in the mtDNA tree (PP$_B$ = 1; BP$_{ML}$ = 97); (3) the orthopteran *Locusta* and the hemipteran *Triatoma* are grouped together with a long branch for *Triatoma* (PP$_B$ = 0.91; BP$_{ML}$ = 37), whereas mtDNA sequences, oddly as well, group *Triatoma* with the zygentoman *Tricholepidion* (PP$_B$ = 1; BP$_{ML}$ = 80); (4) pterygotes, represented by *Triatoma*, *Locusta*, Coleoptera, Lepidoptera, and Diptera, appear monophyletic (PP$_B$ = 0.91; not found with ML), whereas they are not in the mtDNA analysis due to the odd placement of *Triatoma*; (5) *Drosophila* is associated with Calliphoridae (*Melinda*) (PP$_B$ = 0.96; BP$_{ML}$ = 86), whereas it is
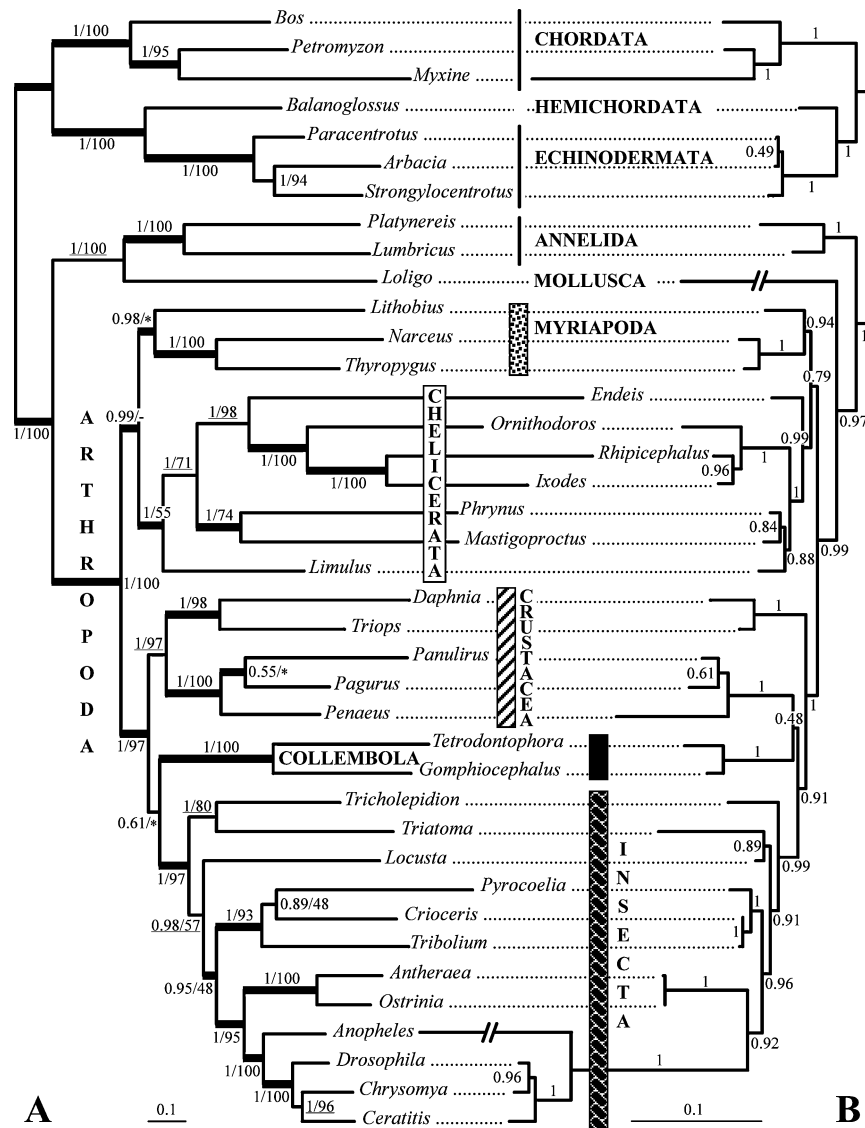
FIGURE 5.    Bayesian trees performed by excluding taxa with reversed asymmetric mutational constraints in the mitochondrial genome. The analyses were done by excluding 10 genera: all the 8 genera with reversed asymmetric mutational constraints, i.e., *Argiope*, *Artemia*, *Branchiostoma*, *Euscorpius*, *Florometra*, *Heterodoxus*, *Katharina*, and *Tigriopus*, and all species with an inverted protein-coding gene, i.e., *Asterina* and *Pisaster*, for which *nad2* is inverted. The model used is the GTR+I+Γ4. The values indicated on the branches correspond to the posterior probabilities (to the left of the slash) obtained with the Bayesian analysis, and to the bootstrap proportions (BPs) obtained with the maximum likelihood analysis (to the right of the slash). Dash indicates that the node was not supported by a BP value superior to 50. Asterisk indicates that an alternative hypothesis was supported by a BP value greater than 50. Bold lines of the mitochondrial tree (A) indicate nodes retrieved in the Bayesian tree performed with 18S rRNA sequences (B), whereas underlined values indicate nodes with posterior probabilities superior to 0.90 that are not congruent with the *18S* tree. Note that the branch length of *Anopheles* and *Loligo* is twice as long as represented in the *18S* tree.

sister-group of a clade composed of *Ceratitis* and Calliphoridae (*Chrysomya*) in the mtDNA tree (PP$_B$ = 1; BP$_{ML}$ = 96); (6) Euchelicerates (*Limulus* + Arachnida) are monophyletic (PP$_B$ = 1; BP$_{ML}$ = 95), whereas mtDNA sequences support their paraphyly due to the placement of *Endeis* (pycnogonid) as a sister-group of acarids (PP$_B$ = 1; BP$_{ML}$ = 98).

*A new method for limiting the misleading effect of strand bias reversals.*—In a third data set, we excluded from the original data set only those taxa where one or two genes are inverted with respect to the other genes in the segment of the mt genome of interest, i.e. *Hetero-*

*doxus*, in which *atp6* and *atp8* are inverted, and *Asterina*, *Florometra* and *Pisaster*, in which *nad2* is inverted. In order to take into account taxa presenting a reverse strand bias with respect to the great majority of taxa, we propose to use a modified matrix where all neutral and quasineutral transitions are excluded. Neutral transitions are all synonymous transitions, i.e., all transitions at third codon positions, and transitions at first positions of Leucine codons (TTR and CTN). Quasineutral transitions are nonsynonymous transitions involving easily interchangeable amino acid residues (Naylor and Brown, 1997; Hassanin et al., 1998), i.e., ACN ↔ GCN
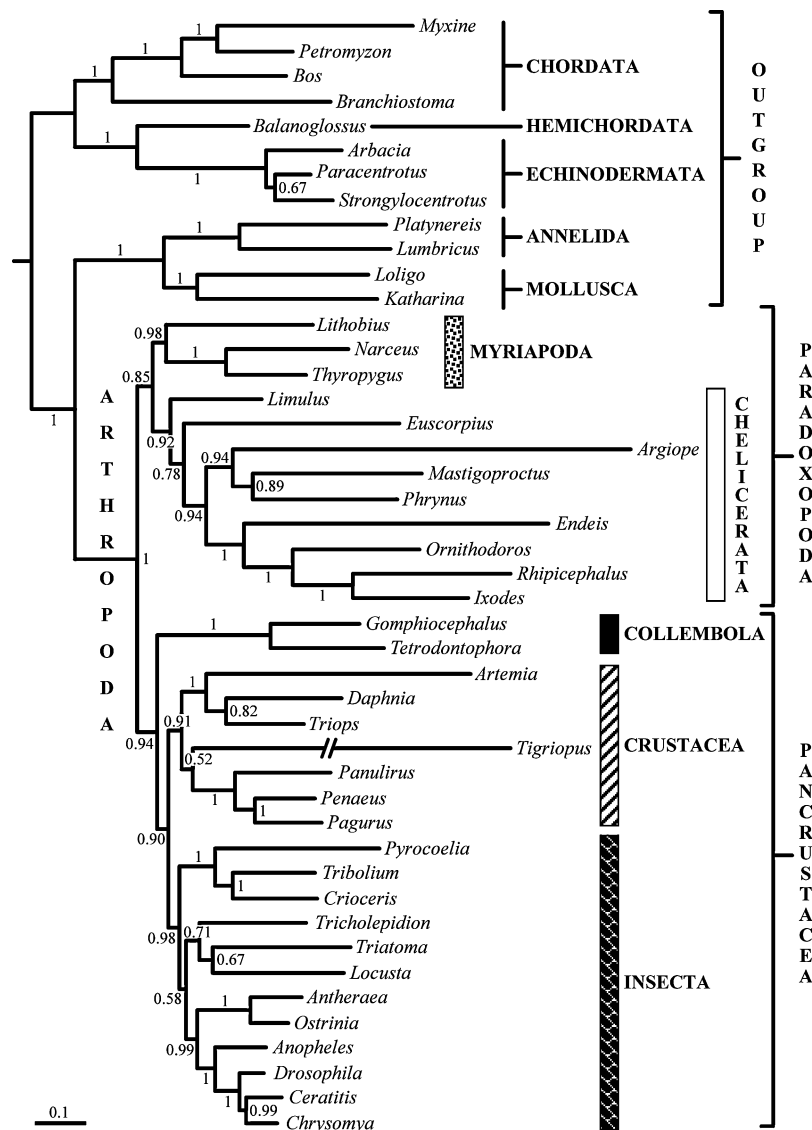
FIGURE 6. Bayesian tree obtained by using the "Neutral Transitions Excluded" model. The analyses were done excluding *Asterina*, *Heterodoxus*, and *Pisaster*, because one or two genes are inverted in these genera. The Bayesian tree was obtained using mtDNA sequences only, with the "Neutral Transitions Excluded" model, which implies to code purines by R and pyrimidines by Y at all third codon positions, at first positions of CTN and TTN codons, and at first and second positions of ACN, ATN, GCN, and GTN codons, and to apply a GTR+I+Γ4 model for first and second codon positions, and a two-state substitution model + I+Γ4 for third codon positions. The values indicated on the branches correspond to posterior probabilities. Note that the branch length of *Tigriopus* is twice as long as represented in the tree.

(T ↔ A), ATN ↔ GTN (I/M ↔ V), CTN ↔ TTY (L2 ↔ F), ACN ↔ ATN (T ↔ I/M), and GCN ↔ GTN (A ↔ V). In this method, that we call "Neutral Transitions Excluded," purines are coded by R and pyrimidines by Y at all third codon positions, at first positions of CTN (L2) and TTN (L1 and F) codons, and at first and second positions of ACN (T), ATN (I and M), GCN (A), and GTN (V) codons. The obtained tree (Fig. 6) is very similar to the one performed with only 39 taxa, i.e., excluding all taxa presenting a reverse strand bias (Fig. 5A). Some of the latter, i.e., *Argiope*, *Euscorpius*, *Tigriopus*, are still associated with a long branch with respect to their close relatives, but they do not group together as previously shown in Fig. 4: *Argiope* and *Euscorpius* fall with other Chelicerates ($PP_B = 0.92$), whereas *Tigriopus* is enclosed with other Crusatceans ($PP_B = 0.91$). A major difference concerns Hexapoda that dot not appear monoyphyletic because Collembola are sister-group of the clade uniting Crustacea with Insecta ($PP_B = 0.90$).

DISCUSSION

*Strand-Specific Compositional Bias*

At sites under little or no selective constraints, such as fourfold degenerate codon positions, all mutations are neutral, or nearly so, and have an equal probability of being fixed in the population. Thus, substitutions at these sites are expected to reflect the underlying rates and

patterns of mutation (Kimura, 1983). According to Wu and Maeda (1987), asymmetry in mutation rate and/or mutation pattern between the two DNA strands should be reflected in nucleotide compositions of neutral sites as well. If patterns of substitutions are symmetric, the equilibrium frequencies of nucleotides are expected to be the same for both strands. In other words, the frequency of adenine should equal the frequency of thymine on the same strand. Similarly, the frequency of cytosine and guanine should be the same. The nucleotide composition at fourfold degenerate sites is given in Table 2 for each of the 49 taxa here examined. The results show that the symmetry does not hold. Despite some differences in base frequencies, all species of Metazoa, except *Artemia* and *Heterodoxus* (but see below), present an important strand asymmetry in the nucleotide composition since one strand is characterized by a positive skew, i.e., A (%) > T (%) and C (%) > G (%), whereas the other strand is characterized by a negative skew, i.e., T (%) > A (%) and G (%) > C (%), simply because of base complementarity. Since this bias is also detectable at non-synonymous sites, this confirms that it is in effect at all positions of the mt genome. Hence, we can define a positive strand, which is characterized by positive AT and CG skews, and a negative strand, which is characterized by negative AT and CG skews. In mammals, the positive and negative strands correspond to the previously named L (light) and H (heavy) strands, respectively.

### What Asymmetric Mechanism Generates the Strand Compositional Bias?

The strand bias is the consequence of asymmetric patterns of change where certain substitutions are more common than their complements, thereby generating inequalities between the frequencies of the complementary bases A/T and C/G (Wu and Maeda, 1987; Lobry, 1995; Sueoka, 1995). In theory, two mechanisms can bias the occurrence of mutations between the two strands: replication and transcription (Francino and Ochman, 1997). Both result in asymmetric patterns of mutations because one strand remains transiently in single-stranded state and is therefore more exposed to DNA damage than the other strand, which is paired with the nascent DNA during replication or the nascent RNA during transcription.

Concerning mtDNA replication, two models have been proposed in mammals: the "strand-displacement model" implies that the H strand is in transient single-stranded state during DNA synthesis, whereas the "stranded-coupled model" considers that the two strands are always double-stranded (Bogenhagen and Clayton, 2003).

According to the "strand-displacement model," mtDNA replication is an asymmetric process, due to the presence of two distinct replication origins (Robberson et al., 1972; Clayton, 1982; Bogenhagen and Clayton, 2003). The H strand replication origin ($O_H$) is located in the main noncoding region of the mtDNA, called control region or D-loop, and the L strand replication origin ($O_L$) is located about 11 kb downstream of the $O_H$ (between

the N- and C-tRNA genes). MtDNA replication starts at $O_H$, with the production of a triple-stranded structure because of the elongation of the nascent H strand, which displaces the parental H strand. When the displacement exposes $O_L$ as a single-strand template, the synthesis of the L strand starts at the opposite direction. Because the replication is very slow, requiring about 2 hours (Clayton, 1982), the parental H strand remains single-stranded for a long time, i.e., until paired by the newly synthesized L strand. In contrast, the parental L strand never remains single stranded in any phase of replication. As a consequence of its single-stranded state, the H strand is supposed to be more exposed to mutations than the L strand (Tanaka and Ozawa, 1994). This model is supported by experiments that have revealed that the rate of spontaneous deaminations of A and C nucleotides are higher in single-stranded DNA than in double-stranded DNA (Sancar and Sancar, 1988; Frederico et al., 1990). In addition, a significant positive correlation has been determined in mammals between the duration of the single-stranded state of the parental H strand (Dssh) and the frequency of cytosine on the L strand. Similarly, negative tendencies have been evidenced between Dssh and the frequencies of guanine and thymine on the L strand (Tanaka and Ozawa, 1994; Reeyes et al., 1998). If the model proposed for mammals can be generalized to other metazoans, it could take into account for the strand-specific compositional bias.

According to the "strand-coupled model," the replication of mtDNA proceeds, principally, perhaps exclusively, by a strand-coupled mechanism: both DNA strands are fully double-stranded, and the newly synthesised L strand involves extensive ribonucleotide incorporation (Yang et al., 2002). As a final step in the replication process, ribonucleotides would be replaced by deoxynucleotides through the POLG, which is known to possess a reverse transcriptase activity (Yang et al., 2002).

Transcription is clearly asymmetric because it can introduce biases in the patterns of mutation on the two strands: while RNA is being synthesized on the transcribed strand of DNA, the nontranscribed DNA strand remains transiently single stranded. Several experiments on *Escherichia coli* have shown that transcription biases the mutational patterns between the transcribed and nontranscribed strands by exposing the nontranscribed strand to DNA damage. For instance, transcription causes approximately fourfold increase in the frequency of cytosine → uracil deaminations in the nontranscribed strand (Beletskii and Bhagwat, 1996). In the mitochondria of mammals, both strands are however symmetrically transcribed over their entire length, starting from two promoters, which are located in the control region. However, the L strand, which is for the most part noncoding in mammals, is transcribed two or three times more frequently than the H strand (Attardi, 1985). Therefore, transcription can be considered as an asymmetric process, and the negative H strand is expected to be more prone to deamination and transcription-coupled repair mutations due to its single-stranded state during transcription of the L strand.

To conclude, the compositional bias in favor of a high A+C content on the positive L strand could be related to high levels of deaminations of A and C on the negative H strand, but additional experiments are needed to know what asymmetric process is directly involved: replication, transcription, or both of them.

*Mutational Processes Involved in Strand Asymmetry*

For all taxa, except *Heterodoxus* (but see below), similar trends were found for both two- and fourfold degenerate third codon positions (Table 2 and Fig. 1). This suggests that the strand-specific compositional bias is the consequence of asymmetric patterns of substitutions involving transitions rather than transversions. Two major asymmetric mutational patterns can be therefore considered: (1) more $A^-T^+ \rightarrow G^-C^+$ than $G^-C^+ \rightarrow A^-T^+$ transitions; and (2) more $C^-G^+ \rightarrow T^-A^+$ than $T^-A^+ \rightarrow C^-G^+$ transitions. Spontaneous deaminations of A and C nucleotides on the negative H strand would explain the strand bias (Tanaka and Ozawa, 1994; Reyes et al., 1998): deamination of adenine on the negative strand would explain the low percentage of $A^-T^+$ pairs because it yields a base, hypoxanthine, that pairs with cytosine rather than thymine (Lindahl, 1993), producing a $A^-T^+ \rightarrow G^-C^+$ transition; similarly, deamination of cytosine on the negative strand would explain the low percentage of $C^-G^+$ pairs because it yields a base, uracil, that pairs with adenine instead of guanine (Lindahl, 1993), producing a $C^-G^+ \rightarrow T^-A^+$ transition. If both deaminations of A and C nucleotides accumulated at similar rates in single- and double-stranded DNA molecules, we expect to observe the following patterns at synonymous positions of the positive L strand: A (%) > G (%); C (%) > T (%); A (%) = C (%); and G (%) = T (%). Such patterns are not observed since adenine is more frequent than cytosine, and thymine is more frequent than guanine (Table 2). Assuming that deamination is the main process involved in the observed compositional bias, the patterns can, however, be explained by differences in the rates of deaminations, firstly, between single and double stranded DNAs, and secondly, between A and C nucleotides (Fig. 7). This model is supported by previous reports showing that deaminations of adenine occur at 2% to 3% of the rate of deaminations of cytosine (Lindahl, 1993; Gilbert et al., 2003), and that the rate of deaminations are slower in double-stranded DNA than in single-stranded DNA (Sancar and Sancar, 1988; Frederico et al., 1990). The nucleotide composition observed at synonymous sites of the positive L strand is in perfect agreement with the model (Fig. 7): adenine is more frequent than thymine due to higher rates of deamination for cytosine in the single-stranded "negative H strand" (dCs) than in the double-stranded "positive L strand" (dCD); similarly, cytosine is more frequent than guanine due to higher rates of deamination for adenine in the single-stranded "negative H strand" (dAs) than in the double-stranded "positive L strand" (dAD); and, as expected with slower rates of deaminations for A than C nucleotides, the frequency of guanine is lower than that
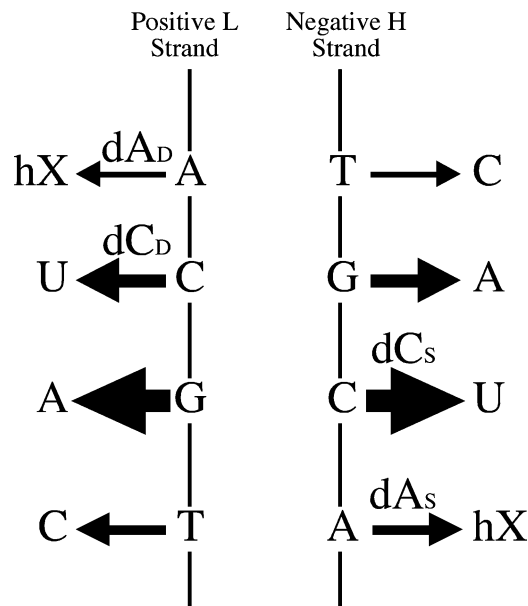


FIGURE 7. Deaminations of adenine and cytosine in the mitochondrial DNA. The positive L strand is characterized by positive AT and CG skews (i.e., A % > T % and C % > G %), whereas the negative H strand is characterized by negative AT and CG skews (i.e., T % > A % and G % > C %). Deaminations may take place in the single stranded "negative H strand" as well as in the double stranded "positive L strand": cytosine (C) into uracil (U) and adenine (A) into hX (hypoxanthine). Thickness of the arrows indicates the rate of deamination: thin and thick arrows are used for slow and fast rates, respectively. Deaminations of A and C nucleotides on the double-stranded "positive L strand" are indicated by dAD and dCD whereas deaminations of A and C nucleotides on the single-stranded "negative H strand" are indicated by dAs and dCs.

of adenine. However, the relative frequencies of C and T nucleotides are highly variable among taxa. In particular, four taxa have more C than T nucleotides in both two- and fourfold degenerate third codon positions, i.e., *Asterina*, *Balanoglossus*, *Bos*, and *Phrynus* (Table 2). These important variations of cytosine and thymine frequencies suggest that the rates of deaminations have changed during the evolutionary history of metazoans. This hypothesis is corroborated by experimental evidence showing that the rates of deaminations are not constant between eukaryotes and bacteria: deaminations of cytosine are 40-fold higher in *Saccharomyces cerevisiae* than in *Escherichia coli* (Impellizzeri et al., 1991). Similarly, an increase in the rates of adenine deamination in the single stranded "negative H strand" (dAs) may explain the high percentages of cytosine observed for *Asterina*, *Balanoglossus*, *Bos*, and *Phrynus*.

*Global Reversals of Asymmetric Mutational Constraints*

The present analyses have shown that six unrelated taxa have a clear reverse strand bias since they are T/G rich rather than A/C rich: *Branchiostoma* within chordates, *Florometra* within echinoderms, *Katharina* within molluscs, *Tigriopus* within crustaceans, and *Argiope* and *Euscorpius* within arachnids (Table 2 and Fig. 1). Because

this reverse strand bias is detected for synonymous as well as nonsynonymous sites, it seems that the phenomenon affects all positions of the mt genes, suggesting that asymmetric mutational constraints have been reversed in these taxa. Two possible scenarios can be proposed for explaining this dramatic change of mutational patterns: (1) inversion of the fragment including the six protein-coding genes with respect to the control region, or reciprocally, (2) inversion of the control region with respect to these six genes. The control region, also called D-loop in vertebrates and "A+T rich" region in some invertebrates, has been shown to be the most variable region of the mtDNA, rendering impossible DNA alignment between distant species (e.g., Mardulyn et al., 2003). It contains the first origin of replication (equivalent to $O_H$ in mammals) and all initiation sites used for transcription (Taanman, 1999). So, whatever the mechanism involved in the asymmetric patterns of mutation, i.e., replication or transcription, the control region appears to be the key region for determining the strand compositional bias. Therefore, an inversion of the control region is expected to produce a global reversal of asymmetric mutational constraints in the mtDNA, resulting with time, in a complete reversal of strand compositional bias. This hypothesis is strongly corroborated by the present analyses of mt gene arrangements during the evolution of Metazoa. In the case of echinoderms, it is clear that an inversion of the control region occurred in the lineage leading to *Florometra* (Fig. 2), explaining why this genus presents a reverse strand bias (Table 2 and Fig. 1). Such an inversion can be also proposed for all other taxa with a reverse strand bias because comparisons with their close relatives reveal that their control region is always differently positioned. An inversion of the control region can be also proposed for *Artemia* and *Heterodoxus*, but in these two genera, the event occurred probably too much recently for observing a complete reversal of strand bias, due to the lack of time for accumulating a sufficient number of mutations. This hypothesis is based on three arguments: (1) one skew is significantly negative for each of these two genera: $AT_4$ for *Artemia* and $CG_4$ *Heterodoxus* (Table 2), suggesting an inversion of the control region relative to the *cox1-3* genes, or reciprocally; (2) the three other values of skew are not significant (Table 2), indicating that the strand bias is not strong, and consequently that the inversion is a recent evolutionary event; and (3) their control region is not positioned as in their close relatives. Additional species closely related to *Artemia* and *Heterodoxus* need however to be analyzed for confirming this hypothesis.

## Phylogenetic Inferences and Reversals of Asymmetry in the mtDNA

The mtDNA sequences have been shown very powerful for inferring relationships at low taxonomic levels, such as relationships between species, genera or even families. However, the usefulness of mtDNA sequences has been questioned for higher taxonomic levels such as relationships between orders, classes, or phyla (Curole

and Kocher, 1999). One explanation is that the phylogenetic signal is obscured by saturation when sequence comparisons involve highly divergent groups. Because the mt genome evolves at much higher rates than the nuclear genome (Li, 1997), multiple hits are more frequent in mtDNA sequences. However, reversals of asymmetric mutational constraints can be another crucial factor for explaining the difficulties encountered by many phylogeneticists for studying deep divergences with mtDNA sequences. Here, we show that asymmetric mutational constraints can be reversed through two different mechanisms: (i) inversion of the control region, which results in a global reversal, and (ii) gene inversion, which results in a local reversal. What could be the consequences of such reversals for phylogenetic inferences? When mutational constraints are reversed, some mutation types, which were frequent, become rare, whereas some other types, which were rare, become frequent. As a consequence, when global reversals of asymmetric mutational constraints occurred independently in several taxa, these taxa are expected to group together due to the long-branch attraction (LBA) phenomenon (Felsenstein, 1978). Here, the long branches do not result from a global acceleration of mutational rates, but they are due to the rapid accumulation of some substitution types, which are rare in other lineages. Although this kind of LBA effect would need to be established in more details using simulation analyses, such as those proposed by Huelsenbeck (1997), it is expected to be particularly misleading for phylogenetic studies. This is exactly what we obtained when using the MP method of tree reconstruction (Fig. 3): all taxa characterized by a reverse strand bias fall together into the same clade in spite of their distant relationships, i.e., *Florometra* (Echinodermata), *Branchiostoma* (Chordata), *Katharina* (Mollusca), *Tigriopus* (Crustacea), *Argiope* (Chelicerata, Araneae), and *Euscorpius* (Chelicerata, Scorpiones). The Bayesian approach seems to be less prone to LBA with *Branchiostoma* located within Chordata, *Florometra* within Echinodermata, and *Katharina* associated with the other representative of the phylum Mollusca (Fig. 4). This result confirms that model-based methods, such as Bayesian and ML analyses, are less sensitive to LBA than MP methods. Indeed, they have the advantage to deal with multiple hits (Swofford et al., 2001), and to take into account heterogeneity of evolutionary rates among sites, a parameter especially important for overcoming LBA (Cunningham et al., 1998). However, any model-based method will be strongly affected when the assumed substitution model is strongly violated (e.g., Swofford et al., 2001; Rosenberg and Kumar, 2003). At present, most models assume that the process of substitution is stationary, i.e., the frequencies of nucleotides remained constant over the period covered by the data. Hence, they cannot manage with reversals of mutational constraints. It is particularly relevant to point out that all the eight genera affected by a reversal of asymmetric mutational constraints during the evolution of their mtDNA have a very long branch (Fig. 4), suggesting that their phylogenetic position should be regarded with caution. Here, the long

branches are not the consequence of accelerated rates of evolution, but they rather reflect the fact that parameters of the model are inaccurate for these taxa. The phylogenetic placements of *Branchiostoma*, *Florometra*, and *Katharina* are in agreement with the traditional morphological classification of metazoans. All other genera affected by a reversal of strand asymmetry occupy an unreliable position in the tree: *Artemia* is the sister-genus of *Daphnia*, whereas *Daphnia* is expected to be associated with *Triops*; *Heterodoxus* is link to *Pyrocoelia*, rendering the Coleoptera paraphyletic; *Argiope*, *Euscorpius*, and *Tigriopus* are united in the same clade although they are not closely related. In addition, we consider that local reversals of mutational constraints, resulting from gene inversions, are also dramatic for phylogenetic inferences. In these cases, the misleading effect on tree topology could be less marked than for global reversals, but the incorporation of these sequences into the analyses may strongly affect the estimation of the parameters of the evolutionary model, and then, tree reconstruction.

Because multiple reversals of asymmetric mutational constraints are expected to considerably mislead phylogenetic inferences based on mtDNA sequences, we recommend specific strategies for improving phylogenetic reconstruction. The first step is to detect taxa for which asymmetric mutational constraints have reversed. To deal with the problem of LBA, one possible solution is then to exclude all these taxa from phylogenetic analyses. The drawback of this radical strategy is that interesting taxa could be removed, limiting the impact of phylogenetic results. As a possible alternative, we propose to use a new method for coding molecular characters, which aims at excluding neutral or quasineutral transitions. There are two main arguments for adopting this "Neutral Transitions Excluded" model: (i) the asymmetric mutational constraints act principally by the way of transitions rather than transversions; and (ii) selected transitions are expected to be less affected by changes in asymmetry than neutral transitions.

### Application to the Phylogeny of Arthropods

Previous analyses based on mtDNA sequences have revealed several unexpected results involving a complete reinterpretation of morphological characters. Numerous studies have concluded that Crustacea are paraphyletic, with Malacostraca being more closely related to Insecta than Branchiopoda (Garcia-Machado et al., 1999; Wilson et al., 2000; Nardi et al., 2001; Hwang et al., 2001; Nardi et al., 2003). Nardi et al. (2003) have also suggested the paraphyly of Hexapoda, with Insecta being more closely related to Crustacea than Collembola. Chelicerata have been found paraphyletic (Delsuc et al., 2003) or polyphyletic (Nardi et al., 2003). Myriapoda have been found paraphyletic (Nardi et al., 2003; Delsuc et al., 2003). Hwang et al. (2001) have proposed a sister-group relationship between Chelicerata and Myriapoda. All of these analyses were performed with several taxa char-

acterized by a reversal of asymmetric mutational constraints: *Artemia*, *Heterodoxus*, and *Katharina*, suggesting possible artifacts in parameter estimations and tree reconstruction.

Here, we have shown that reversals of asymmetric mutational constraints have dramatic consequences for phylogenetic inferences. The detection of these reversals and their management in phylogenetic analyses allowed us to reconcile mtDNA data with traditional morphological hypotheses and molecular analyses based on the 18S rRNA gene (Fig. 5B). Indeed, we retrieved the monophyly of Crustacea, Hexapoda (Insecta + Collembola), Chelicerata, Myriapoda, and Pancrustacea (Crustacea + Hexapoda), when taxa with reversed strand asymmetric mutational constraints were excluded for the analyses (Fig. 5A). In addition, the analyses evidenced strong affinities between Chelicerata and Myriapoda, confirming the monophyly of Paradoxopoda, a taxon recently named by Mallatt et al. (2004) on the basis of 18*S*/28*S* analyses. When the "Neutral Transitions Excluded" model was applied on a largest sample integrating taxa with reversed strand bias, most of these groups were also retrieved as being monophyletic (Fig. 6). The only exception is Hexapoda, which was found to be paraphyletic. In addition, the position of *Artemia*, as sister-group of the clade uniting *Daphnia* and *Triops*, is now in agreement with traditional classifications and molecular studies using nuclear markers, such as EF1$\alpha$ (Braband et al., 2002), as well as 18S and 28S rRNA genes (Mallat et al., 2003). These results suggest that the "Neutral Transitions Excluded" model is useful for phylogenetic inferences by improving both parameter estimations and tree reconstruction. Further applications and simulations are however needed to precise the impact of this coding procedure on tree reconstruction.

### REFERENCES

Anderson, S., A. T. Bankier, B. G. Barrell, M. H. L. De Bruijn, A. R. Coulson, J. Drouin, I. C. Eperon, D. P. Nierlich, B. A. Roe, F. Sanger, P. H. Schreier, A. J. H. Smith, R. Staden, and I. G. Young. 1981. Sequence and organization of the human mitochondrial genome. Nature 290:457–465.

Attardi, G. 1985. Animal mitochondrial DNA: An extreme example of genetic economy. Int. Rev. Cytol. 93:93–145.

Beletskii, A., and A. S. Bhagwat. 1996. Transcription-induced mutations: Increase in C to T mutations in the nontranscribed strand during transcription in *Escherichia coli*. Proc. Natl. Acad. Sci. USA 93:13919–13924.

Bogenhagen, D. F., and D. A. Clayton. 2003. The mitochondrial DNA replication bubble has not burst. Trends Biochem. Sci. 28:357–360.

Boore, J. L. 1999. Animal mitochondrial genomes. Nucleic Acids. Res. 27:1767–1780.

Boore, J. L., and W. M. Brown. 1994. Complete DNA sequence of the mitochondrial genome of the black chiton, *Katharina tunicata*. Genetics 138:423–443.

Boore, J. L., L. L. Daehler, and W. M. Brown. 1999. Complete sequence, gene arrangement, and genetic code of mitochondrial DNA of the cephalochordate *Branchiostoma floridae* (Amphioxus). Mol. Biol. Evol. 16:410–418.

Braband, A., S. Richter, R. Hiesel, and G. Scholtz. 2002. Phylogenetic relationships within the Phyllopoda (Crustacea, Branchiopoda) based on mitochondrial and nuclear markers. Mol. Phylogenet. Evol. 25:229–244.

Bromham, L. D., and B. M. Degnan. 1999. Hemichordates and deuterostome evolution: Robust molecular phylogenetic support for a hemichordate + echinoderm clade. Evol. Dev. 1:166–171.

Brusca, R. C., and G. J. Brusca. 2003. Invertebrates. 2nd edition. Sinauer, Sunderland, Massachusetts.

Burger, G., M. W. Gray, and B. F. Lang. 2003. Mitochondrial genomes: Anything goes. Trends Genet. 19:709–716.

Cameron, C. B., J. R. Garey, and B. J. Swalla. 2000. Evolution of the chordate body plan: New insights from phylogenetic analyses of deuterostome phyla. Proc. Natl. Acad. Sci. USA 97:4469–4474.

Cisne, J. L. 1974. Trilobites and the origin of arthropods. Science 186:13–18.

Clayton, D. A. 1982. Replication of animal mitochondrial DNA. Cell 28:693–705.

Cunningham, C. W., H. Zhu, and D. M. Hillis. 1998. Best-fit maximum likelihood models for phylogenetic inference: Empirical tests with known phylogenies. Evolution 52:978–987.

Curole, J. P., and T. D. Kocher. 1999. Mitogenomics: Digging deeper with complete mitochondrial genomes. TREE 14:394–398.

Delsuc, F., M. J. Phillips, and D. Penny. 2003. Comment on "Hexapod Origins: Monophyletic or Paraphyletic?". Science 301:1482d.

Dotson, E. M. and C. B. Beard. 2001. Sequence and organization of the mitochondrial genome of the Chagas disease vector, *Triatoma dimidiata*. Insect Mol. Biol. 10:205–215.

Felsenstein, J. 1978. Cases in which parsimony or compatability methods will be positively misleading. Syst. Zool. 27:401–410.

Felsenstein, J. 2004. PHYLIP (Phylogeny Inference Package) version 3. 6b. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

Francino, M. P., and H. Ochman. 1997. Strand asymmetries in DNA evolution. Trends Genet. 13:240–245.

Frederico, L. A., T. A. Kunkel, and B. R. Shaw. 1990. A sensitive genetic assay for the detection of cytosine deamination: Determination of rate constant and the activation energy. Biochemistry 29:2532–2537.

Garcia-Machado, E., M. Pempera, N. Dennebouy, M. Oliva-Suarez, J. C. Mounolou, and M. Monnerot. 1999. Mitochondrial genes collectively suggest the paraphyly of Crustacea with respect to Insecta. J. Mol. Evol. 49:142–149.

Garesse, R., J. A. Carrodeguas, J. Santiago, M. L. Pérez, R. Marco, and C. G. Vallejo. 1997. *Artemia* mitochondrial genome: Molecular biology and evolutive considerations. Comp. Biochem. Physiol. 117B:357–366.

Gilbert, M. T., A. J. Hansen, E. Willerslev, L. Rudbeck, I. Barnes, N. Lynnerup, and A. Cooper. 2003. Characterization of genetic miscoding lesions caused by postmortem damage. Am. J. Hum. Genet. 72:48–61.

Guindon, S., and O. Gascuel. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol. 52:696–704.

Hassanin, A., G. Lecointre, and S. Tillier. 1998. The 'evolutionary signal' of homoplasy in protein-coding gene sequences and its phylogenetic consequences for weighting in phylogeny. Comptes Rendus de l'Académie des Sciences, série III 321:611–620.

Hassanin, A. (submitted). Phylogeny of Arthropoda inferred from mitochondrial sequences.

Huelsenbeck, J. P. 1997. Is the Felsenstein zone a fly trap? Syst Biol. 46:69–74.

Huelsenbeck, J. P., B. Larget, R. E. Miller, and F. Ronquist. 2002. Potential applications and pitfalls of Bayesian inference of phylogeny. Syst. Biol. 51:673–688.

Huelsenbeck, J. P., and F. Ronquist. 2001. MrBayes: Bayesian inference of phylogenetic trees. Bioinformatics 17:754–755.

Huelsenbeck, J. P., F. Ronquist, R. Nielsen, and J. P. Bollback. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. Science 294:2310–2314.

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.

Impellizzeri, K. J., B. Anderson, and P. M. Burgers. 1991. The spectrum of spontaneous mutations in a *Saccharomyces cerevisiae* uracil-DNA-glycosylase mutant limits the function of this enzyme to cytosine deamination repair. J Bacteriol. 173:6807–6810.

Kimura, M. 1983. The neutral theory of molecular evolution. Cambridge University Press, Cambridge.

Knight, R. D., S. J. Freeland, and L. F. Landweber. 2001. Rewiring the keyboard: Evolvability of the genetic code. Nature Rev. 2:49–58.

Li, W.-H. 1997. Molecular evolution. Sinauer Associates, Sunderland, Massachusetts.

Lindahl, T. 1993. Instability and decay of the primary structure of DNA. Nature 362:709–715.

Lobry, J. R. 1995. Properties of a general model of DNA evolution under no-strand-bias conditions. J. Mol. Evol. 40:326–330.

Machida, R. J, M. U. Miya, M. Nishida, and S. Nishida. 2002. Complete mitochondrial DNA sequence of *Tigriopus japonicus* (Crustacea: Copepoda). Mar. Biotechnol. 4:406–417.

Mallatt, J. M., J. R. Garey, and J. W. Shultz. 2003. 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. In press.

Mardulyn, P., A. Termonia, and M. C. Milinkovitch. 2003. Structure and evolution of the mitochondrial control region of leaf beetles (Coleoptera: Chrysomelidae): A hierarchical analysis of nucleotide sequence variation. J. Mol. Evol. 56:38–45.

Nardi, F., A. Carapelli, P. P. Fanciulli, R. Dallai, and F. Frati. 2001. The complete mitochondrial DNA sequence of the basal hexapod *Tetrodontophora bielanensis*: Evidence for heteroplasmy and tRNA translocations. Mol. Biol. Evol. 18:1293–1304.

Nardi, F., G. Spinsanti, J. L. Boore, A. Carapelli, R. Dallai, and F. Frati. 2003. Hexapod origins: Monophyletic or paraphyletic? Science 299:1887–1889.

Naylor, G. J. P., and W. M. Brown. 1997. Structural biology and phylogenetic estimation. Nature 388:527–528.

Naylor, G. J. P., and W. M. Brown. 1998. Amphioxus mitochondrial DNA, Chordate phylogeny, and the limits of inference based on comparisons of sequences. Syst. Biol. 47:61–76.

Posada, D., and K. A. Crandall. 1998. MODELTEST: Testing the model of DNA substitution. Bioinformatics 14:817–818.

Perna, N. T., and T. D. Kocher. 1995. Unequal base frequencies and the estimation of substitution rates. Mol. Biol. Evol. 12:359–361.

Reyes, A., C. Gissi, G. Pesole, and C. Saccone. 1998. Asymmetrical directional mutation pressure in the mitochondrial genome of mammals. Mol. Biol. Evol. 15:957–966.

Robberson, D. L., H. Kasamatsu, and J. Vinograd. 1972. Replication of mitochondrial DNA. Circular replicative intermediates in mouse L cells. Proc. Natl. Acad. Sci. USA 69:737–741.

Rosenberg, M. S., and S. Kumar. 2003. Heterogeneity of nucleotide frequencies among evolutionary lineages and phylogenetic inference. Mol. Biol. Evol. 20:610–621.

Sancar, A., and G. B. Sancar. 1988. DNA repair enzymes. Annu. Rev. Biochem. 57:29–67.

Schram, F. 1986. Crustacea. Oxford University Press, New York, Oxford.

Shao, R., N. J. Campbell, and S. C. Barker. 2001. Numerous gene rearrangements in the mitochondrial genome of the Wallaby Louse, *Heterodoxus macropus* (Phthiraptera). Mol. Biol. Evol. 18:858–865.

Snodgrass, R. E. 1938. Evolution of the Annelida, Onychophora and Arthropoda. Smithson. Misc. Collect. 97:1–159.

Sueoka, N. 1995. Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. J. Mol. Evol. 40:318–325.

Swofford, D. L. 2003. 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.

Taanman, J. W. 1999. The mitochondrial genome: Structure, transcription, translation and replication. Biochim. Biophys. Acta 1410:103–123.

Tanaka, M., and T. Ozawa. 1994. Strand asymetry in human mitochondrial DNA mutations. Genomics 22:327–335.

Tarrío, R., F. Rodríguez-Trelles, and F. J. Ayala. 2001. Shared nucleotide composition biases among species and their impact on phylogenetic reconstructions of the Drosophilidae. Mol. Biol. Evol. 18:1464–1473.

Tomita, K., S. Yokobori, T. Oshima, T. Ueda, and K. Watanabe. 2002. The cephalopod *Loligo bleekeri* mitochondrial genome: Multiplied noncoding regions and transposition of tRNA genes. J. Mol. Evol. 54:486–500.

Wilson, K., V. Cahill, E. Ballment, and J. Benzie. 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.

Wu, C.-I., and N. Maeda. 1987. Inequality in mutation rates of the two strands of DNA. Nature 327:169–170.

Yang, M. Y., M. Bowmaker, A. Reyes, L. Vergani, P. Angeli, E. Gringeri, H. T. Jacobs and I. J. Holt. 2002. Biased incorporation of ribonucleotides on the mitochondrial L-strand accounts for apparent strand-asymmetric DNA replication. Cell 111:495–505.

Yang, Z. 1994. Estimating the pattern of nucleotide substitution. J. Mol. Evol. 39:105–111.

Yokobori, S., T. Suzuki, and K. Watanabe. 2001. Genetic code variations in mitochondria: tRNA as a major determinant of genetic code plasticity. J. Mol. Evol. 53:314–326.
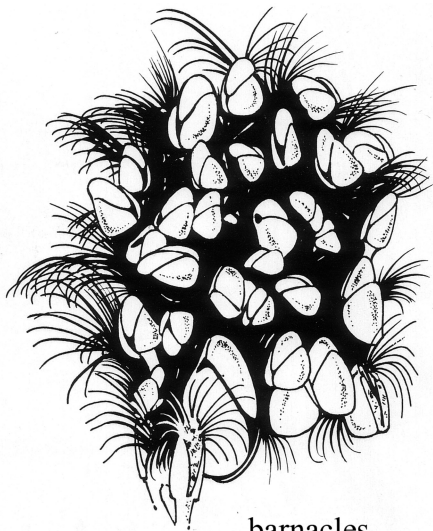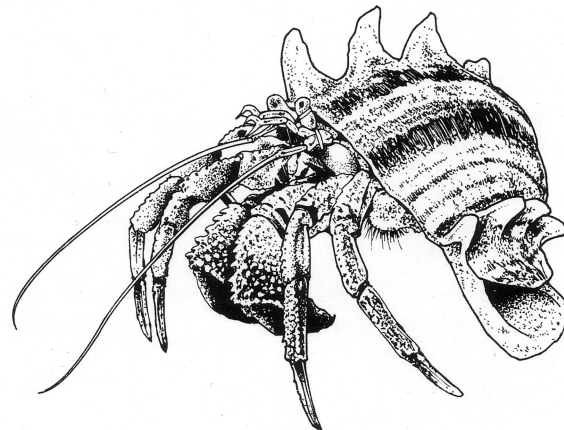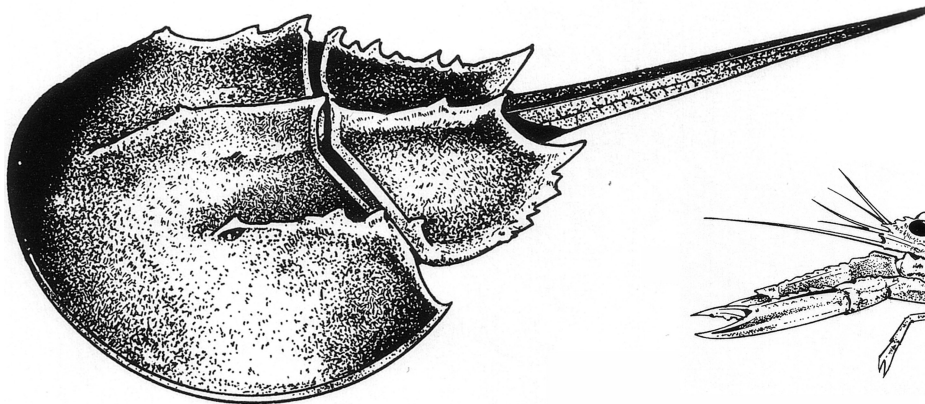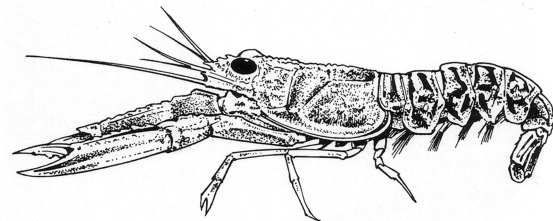
Illustrations of a few of the taxa examined in this study.

APPENDIX 1. Matrix of 74 characters used for inferring ancestral genome organizations.

| | cb | | S1 | | n1 | | L2 | | rL | | V | | rS | | I | | Q | | M | | n2 | | W | | C | | Y | | c1 | | L1 | | c2 | | K | | D | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 | 37 | 38 |
| | 5' | 3' | 5' | 3' | 5' | 3' | 5' | 3' | 5' | 3' | 5' | 3' | 5' | 3' | 5' | 3' | 5' | 3' | 5' | 3' | 5' | 3' | 5' | 3' | 5' | 3' | 5' | 3' | 5' | 3' | 5' | 3' | 5' | 3' | 5' | 3' | 5' | 3' |
| *Drosophila* | 3n6 | 5S1 | 3cb | 3n1 | 3L2 | 3S1 | 3rL | 5n1 | 3V | 5L2 | 3rS | 5rL | CR | 5V | CR | 3Q | 5M | 3I | 5Q | 5n2 | 3M | 5W | 3n2 | 3C | 3Y | 3W | 5c1 | 5C | 5Y | 5L1 | 3c1 | 5c2 | 3L1 | 5K | 3c2 | 5D | 3K | 5a8 |
| *Anopheles* | 3n6 | 5S1 | 3cb | 3n1 | 3L2 | 3S1 | 3rL | 5n1 | 3V | 5L2 | 3rS | 5rL | CR | 5V | CR | 3Q | 5M | 3I | 5Q | 5n2 | 3M | 5W | 3n2 | 3C | 3Y | 3W | 5c1 | 5C | 5Y | 5L1 | 3c1 | 5c2 | 3L1 | 5K | 3c2 | 5D | 3K | 5a8 |
| *Antheraea* | 3n6 | 5S1 | 3cb | 3n1 | 3L2 | 3S1 | 3rL | 5n1 | 3V | 5L2 | 3rS | 5rL | CR | 5V | 3M | 3Q | 5n2 | 3I | CR | 5I | 3M | 5W | 3n2 | 3C | 3Y | 3W | 5c1 | 5C | 5Y | 5L1 | 3c1 | 5c2 | 3L1 | 5K | 3c2 | 5D | 3K | 5a8 |
| *Artemia* | 3n6 | 5S1 | 3cb | 3n1 | 3L2 | 3S1 | 3rL | 5n1 | 3V | 5L2 | 3rS | 5rL | CR | 5V | 3Q | 3Q | 5M | 5I | 5n2 | 5n2 | 3M | 5W | 3n2 | 3I | 3Y | 3M | 5c1 | 5C | 5Y | 5L1 | 3c1 | 5c2 | 3L1 | 5K | 3c2 | 5D | 3K | 5a8 |
| *Chrysomya* | 3n6 | 5S1 | 3cb | 3n1 | 3L2 | 3S1 | 3rL | 5n1 | 3V | 5L2 | 3rS | 5rL | CR | 5c1 | CR | 3Q | 5M | 3I | 5M | 5n2 | 3M | 5W | 3n2 | 3C | 5Y | 3W | 5C | 5c1 | 5Y | 5L1 | 3c1 | 5c2 | 3L1 | 5K | 3c2 | 5D | 3K | 5a8 |
| *Heterodoxus* | 3n1 | 5L1 | 3c3 | 5n1 | 3S1 | 5cb | 3n5 | 3G | 3rS | 5S2 | 5n6 | 3Q | 3a6 | 5rL | 5D | 3F | CR? | 3V | 5W | 5C | 5n4I | 5G | 3P | 5c1 | 3M | 3c1 | 5A | 5c2 | 3T | 3C | 3cb | 3n4 | 3Y | CR? | 5n5 | 5N | 5I | 5n3 |
| *Ixodes* | 3n6 | 5S1 | 3cb | 3n1 | 3L1 | 3S1 | 3rL | 5L1 | 3V | 5L2 | 3rS | 5rL | CR | 5V | CR | 3Q | 5M | 3I | 5Q | 5n2 | 3M | 5W | 3n2 | 3C | 3Y | 3W | 5c1 | 5C | 5Y | 5c2 | 3L2 | 5n1 | 3c1 | 5K | 3c2 | 5D | 3K | 5a8 |
| *Limulus* | 3n6 | 5S1 | 3cb | 3n1 | 3L1 | 3S1 | 3rL | 5L1 | 3V | 5L2 | 3rS | 5rL | CR | 5V | CR | 3Q | 5M | 3I | 5Q | 5n2 | 3M | 5W | 3n2 | 3C | 3Y | 3W | 5c1 | 5C | 5Y | 5c2 | 3L2 | 5n1 | 3c1 | 5K | 3c2 | 5D | 3K | 5a8 |
| *Lithobius* | 3n6 | 5S1 | 3cb | 3n1 | 3L1 | 3S1 | 3rL | 5L1 | 3V | 5L2 | 3rS | 5rL | CR | 5V | 5C | 3Q | 5M | 3I | 5Q | 5n2 | 3M | 5W | 3n2 | 3C | 5I | CR | 5c1 | 3W | 5Y | 5c2 | 3L2 | 5n1 | 3c1 | 5K | 3c2 | 5D | 3K | 5a8 |
| *Locusta* | 3n6 | 5S1 | 3cb | 3n1 | 3L2 | 3S1 | 3rL | 5n1 | 3V | 5L2 | 3rS | 5rL | CR | 5V | CR | 3Q | 5M | 3I | 5Q | 5n2 | 3M | 5W | 3n2 | 3Y | 3Y | 3W | 5c1 | 5C | 5Y | 5L1 | 3c1 | 5c2 | 3L1 | 5D | 3D | 5a8 | 3c2 | 5K |
| *Narceus* | 3n6 | 5S1 | 3cb | 5T | 3L1 | 5P | 3rL | 5L1 | 3V | 5n1 | 3rS | 5rL | CR | 5V | CR | 5M | 3Y | 3T | 3I | 5n2 | 3M | 5W | 3n2 | 3C | 5c1 | 3W | 3F | 5Q | 5C | 5c2 | 3L2 | 5n2 | 3L1 | 5K | 3c2 | 5D | 3K | 5a8 |
| *Pagurus* | 3n6 | 3Y | 5Y | CR | 3rL | 5P | 3c1 | 5n1 | 3V | 5L1 | 3rS | 5rL | 5W | 5V | 3D | 5M | 3C | 3W | 3I | 5n2 | 3M | 5a8 | 5rS | 3Q | 5c1 | 5Q | 5S1 | 3cb | 5C | 5L2 | 3rL | 5n1 | 3n2 | 5Q | 3c2 | 5G | 3A | 5I |
| *Rhipicephalus* | 3n6 | 5S1 | 3cb | 3L2 | 3L1 | 3E | CR' | 3S1 | 3V | 5L1 | 3rS | 5rL | CR | 5V | CR | 3Q | 3F | 3I | 3C | 5n2 | 3M | 5W | 3n2 | 3Y | CR' | 5M | 5c1 | 3W | 5Y | 5c2 | 3c1 | 5n1 | 3c1 | 5K | 3c2 | 5D | 3K | 5a8 |
| *Tetradontophora* | 3n6 | 3n1 | 3I | 5M | 3L2 | 3cb | 3rL | 5n1 | 3V | 5L2 | 3n3 | 5n6 | 3Q | 5V | CR | 5S1 | CR | 5rS | 3S1 | 5P | 3M | 5c2 | 3S2 | 3C | 3Y | 3W | 5c1 | 5C | CR | 5M | 3A | 5n1 | 3n2 | 5L2 | 3rS | 5I | 3P | 5E |
| *Tigriopus* | 3Y | 5H | 3E | 5C | 3L1 | 5Q | 3c2 | 5Y | 3c3 | 5G | 3n3 | 5rL | 3T | 5K | 3K | 5n5 | 3n1 | 5F | 3c1 | 5rS | 3F | 5Q | 3cb | 5a6 | 3S1 | 5A | 3L2 | 5cb | 3n2 | 5N | 3S2 | 5n1 | 3N | 5D | 3I | 5n3 | 3P | 5E |
| *Lumbricus* | 3n6 | 5W | 3n3 | 5n2 | 3L1 | 5I | 3S2 | 5A | 3V | 5L2 | 3rS | 5rL | 3M | 5V | 3n1 | 5K | 3c3 | 5n6 | 3C | 5D | 3S1 | 5C | 3cb | 5a6 | 3n4 | 5M | 3a8 | 5G | 5G | 5N | 3A | 5n1 | 3N | 5G | 3I | 5n3 | 3c2 | 5a8 |
| *Platynereis* | 3n6 | 5W | 3L2 | 5A | 3L1 | 5I | 3rL | 5S1 | 3V | 5L2 | 3rS | 5rL | 3n4 | 5V | 3n1 | 5K | 3c3 | 5n6 | 3a6 | 5D | 3S2 | 5C | 3Q | 5Y | 3n2 | 3c1 | CR | 5a8 | 3C | 5D | 3L2 | 5n1 | 3N | 5G | 3I | 5n3 | 3M | 5c3 |
| *Katharina* | 3n6 | 5S1 | 3cb | 3T | 3Q | 3P | 5c3 | 5H | 3V | 5I | 3rS | 5rL | 3M | 5V | 3N | 5n3 | 3G | 5W | 3C | 5rS | 3S2 | 5c1 | ? | ? | 3Y | 5M | 3W | 5C | 3n2 | 3C | 3G | 5n1 | CR? | 5a8 | 3c3 | 5A | 3c1 | CR? |
| *Loligo* | 3n6 | 5S1 | 3cb | 3n3 | 3L1 | 5P | 3rL | 5L1 | 3V | 5I | 3rS | 5rL | 3W | 5V | ? | 3rL | ? | 5n1 | 5R | 3c2 | 3S2 | 5c1 | 3Q | 5rS | 3Y | 3c1 | 5E | 5C | 3n2 | 3G | 3G | 3T | 3N | 3M | ? | 5S2 | 3A | 5a8 |
| *Bos* | 5E | 5T | 5D | 3c1 | 3L1 | 5I | 3S2 | 5n5 | 3V | 5L1 | 3rS | 5rL | 3F | 5V | 3n1 | 3Q | 5M | 3I | 5Q | 5n2 | 3M | 5W | 3n2 | 3A | 3Y | 5N | 5c1 | 5C | 5Y | 3S1 | 3rL | 5n1 | 3D | 5K | 3c2 | 5a8 | 5S1 | 5c2 |
| *Petromyzon* | 5E | 3P | 5D | 3c1 | 3L1 | 5I | 3S2 | 5n5 | 3V | 5L1 | 3F | 5rL | 3F | 5V | 3n1 | 5M | 5M | 3I | 5Q | 5n2 | 3M | 5N | 3n2 | 3A | 3Y | 5N | 5c1 | 5C | 5Y | 3S1 | 3rL | 5n1 | 3D | 5K | 3c2 | 5a8 | 5S1 | 5c2 |
| *Branchiostoma* | 5E | 5T | 5D | 3c1 | 3L1 | 5I | 3S2 | 5n5 | 3n2 | ? | 3rS | 5rL | 3F | 5V | 3n1 | 5n2 | 5n2 | 3M | 3I | 5n2 | 3Q | 5N | 5N | 3A | 3Y | 5A | 5c1 | 5C | 5Y | 3S1 | 3rL | 5n1 | 3D | 5K | 3c2 | 5a8 | 5S1 | 5c2 |
| *Balanoglossus* | 3n5 | 5E | 5D | 3c1 | 3L1 | 3Q | 3rL | 5L1 | 3V | 5c1 | 5M | 3C | 3F | 5E | 5Q | 5M | 5I | 3n1 | 3I | 3D | 3M | 5N | 5A | 5C | 3Y | 5A | 5c1 | 5C | 5rL | 5R | 3L2 | 5n1 | 3D | 5K | 3c2 | 5a8 | 5S1 | 5c2 |
| *Arbacia* | 5n6 | 5F | 5n3 | 3c3 | 3L1 | 5I | 3N | 3A | 3n2 | CR | 5M | 3C | 3F | 5E | 3n1 | 5n2 | 5N | 3P | 5V | 3D | 3I | 5rL | 5A | 5C | 3W | 3V | 5D | 5G | 5P | 5R | 3G | 5n1 | 3n4I | 5K | 3c2 | 5a8 | 5Y | 3M |
| *Asterina* | 5n6 | ? | 5n3 | 3c3 | 3L1 | 5I | 5N | 3A | 3n2 | CR | 5M | 3C | 3F | 5E | 3n1 | 5n2 | 3N | 3P | 5V | 3D | 3I | 5rL | 5A | 5C | 3W | 3V | 5D | 5G | 5P | 5R | 3G | 5n1 | 3n4I | 5K | 3c2 | 5a8 | 5Y | 3M |
| *Pisaster* | 5n6 | 5F | 5n3 | 3c3 | 3L1 | 5I | 3N | 3A | 3n2 | ? | 5M | 3C | 3F | 5E | 3n1 | 5n2 | 5N | 3P | 5V | 3D | 3I | 5rL | 5A | 5C | 3W | 3V | 5D | 5G | 5P | 5R | 3G | 5n1 | 3n4I | 5K | 3c2 | 5a8 | 5Y | 3M |
| *Florometra* | 5n6 | 5P | 5n3 | 3c3 | 5c1 | 5I | 3N | 3A | 3Y | 5G | 5M | 3C | 3F | 5E | 3n1 | 5n2 | 5N | 3P | 5V | 3D | 3I | 5Y | 5A | 5C | 3W | 3V | 3n2 | 5rL | 5n1 | 5R | 3G | 5F | 3n4I | 5K | 3c2 | 5a8 | 5Y | 3M |
| **Inferred ancestral genome organisations** | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Mollusca MP/GR-A | 3n6 | 5S1 | 3cb | ? | 3L1 | ? | 3rL | 5L1 | 3V | 5L2 | 3rS | 5rL | 3F | 5V | ? | ? | ? | ? | 3C | 5rS | 3S2 | 5c1 | ? | ? | 3Y | 5M | ? | 5C | 3n2 | ? | 3L2 | 5n1 | 3N | ? | ? | ? | ? | 5a8 |
| Chordata MP-A | 5E | 5T | 5D | 3c1 | 3L1 | 5I | 3S2 | 5n5 | 3V | 5L1 | 3rS | 5rL | 3F | 5V | 3n1 | 3Q | 5M | ? | ? | 5n2 | ? | ? | 3A | 3A | 3Y | 5A | 5c1 | 5C | 5Y | 3S1 | 3rL | 5n1 | 3D | 5K | 3c2 | 5a8 | 5S1 | 5c2 |
| Chordata GR-U | 5E | 5T | 5D | 3c3 | 3L1 | 5I | 3S2 | 5n5 | 3V | 5L1 | 3rS | 5rL | 3F | 5V | 3n1 | 5n2 | 5M | ? | 5Q | 5n2 | 3N | 3M | 5N | 3A | 3Y | 5A | 5c1 | 5C | 5Y | 3S1 | 3rL | 5n1 | 3D | 5K | 3c2 | 5a8 | 5S1 | 5c2 |
| Eleutherozoa MP-A | 5n6 | 5F | 5D | 3c1 | 3L1 | 5I | 3N | ? | 3n2 | ? | 5M | 3C | 3F | 5E | 3n1 | 5n2 | 5N | 3P | 5V | 3D | 3I | 5rL | 5A | 5C | 3W | 3V | 5D | 5G | 5rL | 5R | 3G | 5n1 | 3n4I | 5K | 3c2 | 5a8 | 5Y | 3M |
| Eleutherozoa GR-U | 5n6 | 5F | 5n3 | 3c3 | 3L1 | 5I | 3N | 3A | 3n2 | ? | 5M | 3C | 3F | 5E | 3n1 | 5n2 | 5N | 3P | 5V | 3D | 3I | ? | 5A | 5C | 3W | 3V | 5D | 5G | 5R | ? | 3G | ? | 3n4I | 5K | 3c2 | 5a8 | 5Y | 3M |
| Echinodermata MP-A | 5n6 | ? | 5n3 | 3c3 | 3L1 | 5I | 3N | 3A | ? | ? | 5M | 3C | 3F | 5E | 3n1 | 5n2 | 5N | 3P | 5V | 3D | 3I | ? | 5A | 5C | 3W | 3V | ? | ? | ? | ? | 3G | ? | 3n4I | 5K | 3c2 | 5a8 | ? | 3M |
| Echinodermata GR-A | 5n6 | 5F | 5n3 | 3c3 | 5c1 | 5I | 3N | 3A | ? | ? | 5M | 3C | 3F | 5E | 3n1 | 5n2 | 5N | 3P | 5V | 3D | 3I | 5rL | 5A | 5C | 3W | 3V | 5D | ? | 5R | ? | 3G | 5n1 | 3n4I | 5K | 3c2 | 5a8 | 5Y | 3M |

*(Continued on next page)*

297

APPENDIX 1. Matrix of 74 characters used for inferring ancestral genome organizations. (Continued)

| | a8 | | a6 | | c3 | | G | | n3 | | A | | R | | N | | S2 | | E | | F | | n5 | | H | | n4 | | n4l | | T | | P | | n6 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 39 | 40 | 41 | 42 | 43 | 44 | 45 | 46 | 47 | 48 | 49 | 50 | 51 | 52 | 53 | 54 | 55 | 56 | 57 | 58 | 59 | 60 | 61 | 62 | 63 | 64 | 65 | 66 | 67 | 68 | 69 | 70 | 71 | 72 | 73 | 74 |
| | 5′ | 3′ | 5′ | 3′ | 5′ | 3′ | 5′ | 3′ | 5′ | 3′ | 5′ | 3′ | 5′ | 3′ | 5′ | 3′ | 5′ | 3′ | 5′ | 3′ | 5′ | 3′ | 5′ | 3′ | 5′ | 3′ | 5′ | 3′ | 5′ | 3′ | 5′ | 3′ | 5′ | 3′ | 5′ | 3′ |
| *Drosophila* | 3D | 5a6 | 3a8 | 5c3 | 3a6 | 5G | 3c3 | 5n3 | 3G | 5A | 3n3 | 5R | 3A | 5N | 3R | 5S2 | 3N | 5E | 3S2 | 3F | 3n5 | 3E | 3H | 5F | 3n4 | 5n5 | 3n4l | 5H | 5T | 5n4 | 5n4l | 3P | 5n6 | 3T | 5P | 5cb |
| *Anopheles* | 3D | 5a6 | 3a8 | 5c3 | 3a6 | 5G | 3c3 | 5n3 | 3G | 5R | 3R | 5N | 3n3 | 5A | 3A | 3S2 | 5E | 3N | 5S2 | 3F | 3n5 | 3E | 3H | 5F | 3n4 | 5n5 | 3n4l | 5H | 5T | 5n4 | 5n4l | 3P | 5n6 | 3T | 5P | 5cb |
| *Antheraea* | 3D | 5a6 | 3a8 | 5c3 | 3a6 | 5G | 3c3 | 5n3 | 3G | 5A | 3n3 | 5R | 3A | 5N | 3R | 5S2 | 3N | 5E | 3S2 | 3F | 3n5 | 3E | 3H | 5F | 3n4 | 5n5 | 3n4l | 5H | 5T | 5n4 | 5n4l | 3P | 5n6 | 3T | 5P | 5cb |
| *Artemia* | 3D | 5a6 | 3a8 | 5c3 | 3a6 | 5G | 3c3 | 5n3 | 3G | 5A | 3n3 | 5R | 3A | 5N | 3R | 5S2 | 3N | 5E | 3S2 | 5F | 3E | 3n5 | 3H | 5F | 3n4 | 5n5 | 3n4l | 5H | 5T | 5n4 | 5n4l | 3P | 5n6 | 3T | 5P | 5cb |
| *Chrysomya* | 3D | 5a6 | 3a8 | 5c3 | 3a6 | 5G | 3c3 | 5n3 | 3G | 5A | 3n3 | 5R | 3A | 5N | 3R | 5S2 | 3N | 5E | 3S2 | 3F | 3n5 | 3E | 3H | 5F | 3n4 | 5n5 | 3n4l | 5H | 5T | 5n4 | 5n4l | 3P | 5n6 | 3T | 5V | 5P |
| *Heterodoxus* | CR? | 5a6 | 3a8 | 5rS | 5R | 5S1 | 3n2 | 3L2 | 3D | CR? | 5Y | 3N | 5c3 | 5H | 3K | 3A | 3rL | 5F | 3W | 3H | 3S2 | 3I | 5K | 5L2 | 3R | 5R | 3n4l | 3L1 | 5n2 | 5n4 | 3P | 5n4l | 3n6 | 5T | 5V | 5P |
| *Ixodes* | 3D | 5a6 | 3a8 | 5c3 | 3a6 | 5G | 3c3 | 5n3 | 3G | 5A | 3n3 | 5R | 3A | 5N | 3R | 5S2 | 3N | 5E | 3S2 | 3F | 3n5 | 3E | 3H | 5F | 3n4 | 5n5 | 3n4l | 5H | 3T | 5n4 | 5n4l | 5c1 | 3n6 | 5T | 5P | 5cb |
| *Limulus* | 3D | 5a6 | 3a8 | 5c3 | 3a6 | 5G | 3c3 | 5n3 | 3G | 5A | 3n3 | 5R | 3A | 5N | 3R | 5S2 | 3N | 5E | 3S2 | 3F | 3n5 | 3E | 3H | 5F | 3n4 | 5n5 | 3n4l | 5H | 5T | 5n4 | 5n4l | 3P | 5n6 | 3T | 5P | 5cb |
| *Lithobius* | 3D | 5a6 | 3a8 | 5c3 | 3a6 | 5G | 3c3 | 5n3 | 3G | 5A | 3n3 | 5R | 3A | 5N | 3R | 5S2 | 3N | 5E | 3S2 | 3F | 3n5 | 3E | 3H | 5F | 3n4 | 5n5 | 3n4l | 5H | 5T | 5n4 | 5n4l | 3P | 5n6 | 3T | 5P | 5cb |
| *Locusta* | 3K | 5a6 | 3a8 | 5c3 | 3a6 | 5G | 3c3 | 5n3 | 3G | 5A | 3n3 | 5R | 3A | 5N | 3R | 5S2 | 3N | 5E | 3S2 | 3F | 3n5 | 3E | 3H | 5F | 3n4 | 5n5 | 3n4l | 5H | 5T | 5n4 | 5n4l | 3P | 5n6 | 3T | 5P | 5cb |
| *Narceus* | 3D | 5a6 | 3a8 | 5c3 | 3a6 | 5G | 3c3 | 5n3 | 3G | 5A | 3n3 | 5D | 3A | 5N | 3R | 5S2 | 3N | 5E | 3S2 | 5n6 | 3n5 | 5Y | 3H | 5F | 3n4 | 5n5 | 3n4l | 5H | 3P | 5n4 | 5n4l | 3Q | 3n1 | 5n4l | 3E | 5cb |
| *Pagurus* | 3n2 | 5a6 | 3a8 | 5c3 | 3a6 | 5R | 3K | CR | 3K | 5S1 | 3L2 | 5S2 | 3a6 | CR | 3c1 | 5c2 | 3N | 5E | 3S2 | 3F | 3n5 | 5Q | 3H | 5F | CR | 5n5 | 3n4l | 5C | 5T | 5n4 | 5n4l | 5n6 | 3n1 | CR | 3T | 5cb |
| *Rhipicephalus* | 3D | 5a6 | 3a8 | 5c3 | 3a6 | 5G | 3c3 | 5n3 | 3G | 5A | 3n3 | 5R | 3A | 5N | 3R | 5S2 | 3N | 5E | 3S2 | 3F | 3n5 | 3E | 3H | 5F | 3n4 | 5n5 | 3n4l | 5H | 5T | 5n4 | 5n4l | 3P | 5n6 | 3T | 5P | 5cb |
| *Tetrodontophora* | 3D | 5a6 | 3a8 | 5S2 | 3N | 5G | 3c3 | 5T | 3G | 5V | 3A | 5L1 | 3A | 5N | 3N | 5E | 3S2 | 5E | 3S2 | 3F | 3n5 | 3E | 3H | 5F | 3n4 | 5n5 | 3n4l | 5H | 5T | 5n4 | 5n4l | 3P | 5n6 | 3T | 3V | 5n4l |
| *Tigriopus* | 3n4l | 5a6 | 3a8 | 5S2 | 3N | 5rL | 3rL | 5T | 3n5 | 5V | 3C | 5L1 | 3H | 5n4 | 3n4 | 5c3 | 3a6 | 5W | 3D | 5S1 | 3Q | 5n2 | 3I | 5n3 | 3cb | 5R | 3R | 5N | 3n6 | 5a8 | 3G | 5rS | 3M | 5D | 3V | 5n4l |
| *Lumbricus* | 3D | 5Y | 3W | 5R | 3G | 5n6 | 3a8 | 5c3 | 3K | 5S2 | 3L2 | 5S2 | 3a6 | 5H | 3c1 | 5c2 | 3n3 | 5L1 | 3F | 5P | 3n5 | 5G | 3H | 5F | CR | 5n5 | 3n4l | 5H | 3T | 5n4 | 5n4l | 3P | 3E | 3T | 3Q | 3n5 |
| *Platynereis* | 3Y | 5M | 3W | 5R | 3D | 5Q | 3c2 | 5Q | 3I | 5S2 | 3SI | 5R | 3a6 | 5N | 3c1 | 5I | 5n2 | 5n2 | 3F | 5P | 3n5 | 3a6 | 3H | 5F | 3R | 5n5 | 3n4l | 5H | 5T | 5n4 | 5n4l | 3P | 5n6 | 5T | 3Q | 3n5 |
| *Katharina* | 3c2 | 5a6 | 3a8 | 3F | 5E | 5K | 3E | 5L1 | 3I | 5S2 | 3K | 5R | 3a6 | 5N | 3R | 5I | 3K | 5n2 | 5Y | 3P | 3n5 | 3R | 3n4 | 5F | 3L2 | 5n5 | 3L2 | 5H | 3R | 5n4 | 5n4l | 3L1 | 5n1 | 3n1 | 3P | 3n5 |
| *Loligo* | 3D | 5a6 | 3a8 | 3H | 5L2 | 5n3 | ? | 5L1 | 3S1 | 3S1 | ? | 5D | 5M | 3F | 3E | 5c2 | 3K | 5n5 | 5Y | 5N | 3n5 | 3R | 3n4 | 5F | 3L2 | 5n5 | 3L2 | 5H | 3R | 5n4 | 5n4l | 3P | 5n6 | 3T | 3E | 3n5 |
| *Bos* | 3D | 5a6 | 3a8 | 5c3 | 3a6 | 5G | 3c3 | 5n3 | 3G | 5R | 3N | 3W | 3n3 | 5n4l | 3C | 5A | 3H | 5L2 | 5cb | 5n6 | 3n5 | 5rS | 3L2 | 3n6 | 3n4 | 5H | 3R | 5n4 | 3cb | 5n4l | 3P | 3P | 3n6 | 3T | 3E | 3n5 |
| *Petromyzon* | 3K | 5a6 | 3a8 | 5c3 | 3a6 | 5G | 3c3 | 5n3 | 3G | 5R | 3N | 3W | 3n3 | 5n4l | 3C | 5A | 3H | 5L2 | 5cb | 3T | 5P | 5V | 3L2 | 3n6 | 3n4 | 5H | 3R | 5n4 | 3E | 5n4 | CR | 3E | 5F | 3T | 3E | 3n5 |
| *Branchiostoma* | 3K | 5a6 | 3a8 | 5c3 | 3a6 | 5G | CR | 3n6 | 3c3 | 5R | 3N | 3W | 3n3 | 5n4l | 5W | 3n2 | 3H | 5L2 | 5cb | 3cb | 3rS | 5V | 3L2 | CR | 3n4 | 5H | 3R | 5n4 | 3cb | 5n4 | CR | 3P | 5rS | 3cb | 3E | 3G |
| *Balanoglossus* | 3D | 5a6 | 3a8 | 5c3 | 3a6 | 5G | 3c3 | 5n3 | 3G | 5R | 3C | 3W | 3c1 | 5n4l | 5W | 3n2 | 3H | 5n5 | 3cb | 5T | 5n6 | 5rS | 3S2 | 5cb | 3n4 | 5H | 3n4l | 5H | 3E | 5n4 | 3E | CR | 3n6 | CR | 5F | 5P |
| *Arbacia* | 3K | 5a6 | 3a8 | 5c3 | 3a6 | 5G | 3Y | 5L1 | 5S1 | 5n4 | 5W | 3L2 | 3c1 | 5n4l | 5Q | 5L2 | 3H | 5n5 | 3rS | 5T | 3cb | 5rS | 3S2 | 3n6 | 3n4 | 5S2 | 3n3 | 5H | 3R | 5c2 | 3E | CR | CR | 3Q | 3Q | 3n5 |
| *Asterina* | 3K | 5a6 | 3a8 | 5c3 | 3a6 | 5G | 3Y | 5L1 | 5S1 | 5n4 | 5W | 3L2 | 3c1 | 5n4l | 5Q | 5L2 | 3H | 5n5 | 3rS | 5T | 3cb | 5rS | 3S2 | 3n6 | 3n4 | 5S2 | 3n3 | 5H | 3R | 5c2 | 3E | CR | 5c1 | 3Q | 3E | 3n5 |
| *Pisaster* | 3K | 5a6 | 3a8 | 5c3 | 3a6 | 5G | 3Y | 5L1 | 5S1 | 5n4 | 5W | 3L2 | 3c1 | 5n4l | 5Q | 5L2 | 3H | 5n5 | 3rS | 5T | 3cb | 5rS | 3S2 | 3n6 | 3n4 | 5S2 | 3n3 | 5H | 3R | 5c2 | 3E | CR | 5c1 | 3Q | 3E | 3n5 |
| *Florometra* | 3K | 5a6 | 3a8 | 5c3 | 3a6 | 5G | 3rL | 5L1 | 5S1 | 5n4 | 5W | 3L2 | 3c1 | 5n4l | 3H | 5L2 | 3H | 5n5 | 3L1 | 5T | 3L1 | 5rS | 3S2 | 3n6 | 3n4 | 5S2 | 3n3 | 5H | 3R | 5c2 | 3E | CR | 3cb | 3Q | 5cb | 3n5 |
| **Inferred ancestral genome organisations** | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Mollusca MP/GR-A | 3D | 5a6 | 3a8 | 5c3 | 3a6 | 5G | 3c3 | 5n3 | ? | 5S2 | ? | 5R | 3A | 5N | 3R | 5c2 | 3R | 5n5 | 3S2 | ? | 3n5 | 5rS | 3H | 5F | 3n4 | 5n5 | 3n4l | 5H | 5T | 5n4 | 5n4l | ? | 5n6 | ? | 5P | 5cb |
| Chordata MP-A | 3K | 5a6 | 3a8 | 5c3 | 3a6 | 5G | 3c3 | 5n3 | 3G | 5R | 3C | 3W | 3n3 | 5n4l | 5W | 3n2 | 3H | 5L2 | 5cb | 5n6 | 3n5 | 5rS | 3L2 | 3n6 | 3n4 | 5S2 | 3n4l | 5H | 3R | 5n4 | 3cb | 3P | CR | 3T | 3E | 3n5 |
| Chordata GR-U | 3K | 5a6 | 3a8 | 5c3 | 3a6 | 5G | 3c3 | 5n3 | 3G | 5R | 3C | 3W | 3n3 | 5n4l | 5W | 3n2 | 3H | 5L2 | 5cb | 5n6 | 3cb | 5rS | 3L2 | 3n6 | 3n4 | 5S2 | 3n4l | 5H | 3R | 5n4 | 3cb | 3P | CR | 3cb | 3E | 3n5 |
| Eleutherozoa MP-A | 3K | 5a6 | 3a8 | 5c3 | 3a6 | 5G | 3Y | 5L1 | 5S1 | 5n4 | 5W | 3L2 | 3c1 | 5n4l | 5Q | 5L2 | 3H | 5n5 | 3rS | 5T | 3cb | 5rS | 3S2 | 3n6 | 3n4 | 5S2 | 3n3 | 5H | 3R | 5c2 | 3E | CR | CR | 3Q | 5cb | 3E |
| Eleutherozoa GR-U | 3K | 5a6 | 3a8 | 5c3 | 3a6 | 5G | 3Y | 5L1 | 5S1 | 5n4 | 5W | 3L2 | 3c1 | 5n4l | 5Q | 5L2 | 3H | 5n5 | 3rS | 5T | ? | 5rS | 3S2 | 3n6 | 3n4 | 5S2 | 3n3 | 5H | 3R | 5c2 | 3E | CR | CR | 3Q | 5cb | 3n5 |
| Echinodermata MP-A | 3K | 5a6 | 3a8 | 5c3 | 3a6 | 5G | 3Y | 5L1 | 5S1 | 5n4 | 5W | 3L2 | 3c1 | 5n4l | 5Q | 5L2 | 3H | 5n5 | 3rS | 5T | 3cb | 5rS | 3S2 | 3n6 | 3n4 | 5S2 | 3n3 | 5H | 3R | 5c2 | 3E | CR | CR | 3Q | 5cb | 3n5 |
| Echinodermata GR-A | 3K | 5a6 | 3a8 | 5c3 | 3a6 | 5rL | ? | 5L1 | 5S1 | 5n4 | 5W | 3L2 | 3c1 | 5n4l | 5Q | 5L2 | 3H | 5n5 | 3rS | 5T | 3cb | 5rS | 3S2 | 3n6 | 3n4 | 5S2 | 3n3 | 5H | 3R | 5c2 | 3E | CR | CR | 3Q | 5cb | 3n5 |

*Abbreviations:* MP: maximum parsimony; a6 and a8: subunits 6 and 8 of the ATPase; 5: 5′ end; GR: genome reconstruction; c1, c2, c3: cytochrome c oxidase subunits 1–3; 3: 3′ end; A: ambiguous; cb: cytochrome b; U: unambiguous; n1, n2, n3, n4, nd5, nd6, and n4l: NADH dehydrogenase subunits 1–6 and 4L; rL and rS: large and small subunit rRNAs. CR: Control region.

298