Available online at www.sciencedirect.com
ScienceDirect
Molecular Phylogenetics and Evolution xxx (2007) xxx–xxx
MOLECULAR PHYLOGENETICS AND EVOLUTION
www.elsevier.com/locate/ympev

The effect of model choice on phylogenetic inference using mitochondrial sequence data: Lessons from the scorpions

Martin Jones [a,*], Benjamin Gantenbein [b], Victor Fet [c], Mark Blaxter [a]

[a] *Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3JT, UK*
[b] *AO Research Institute, Clavadelerstrasse 8, Davos Platz CH-7270, Switzerland*
[c] *Department of Biological Sciences, Marshall University, Huntington, WV 25755-2510, USA*

Received 25 April 2006; revised 14 November 2006; accepted 14 November 2006

Abstract

Chelicerates are a diverse group of arthropods, with around 65,000 described species occupying a wide range of habitats. Many phylogenies describing the relationships between the various chelicerate orders have been proposed. While some relationships are widely accepted, others remain contentious. To increase the taxonomic sampling of species available for phylogenetic study based on mitochondrial genomes we produced the nearly complete sequence of the mitochondrial genome of the scorpion *Mesobuthus gibbosus*. Mitochondrial gene order in *M. gibbosus* largely mirrors that in *Limulus polyphemus* but tRNA secondary structures are truncated. A recent analysis argued that independent reversal of mitochondrial genome strand-bias in several groups of arthropods, including spiders and scorpions, could compromise phylogenetic reconstruction and proposed an evolutionary model that excludes mutational events caused by strand-bias (Neutral Transitions Excluded, NTE). An arthropod dataset of six mitochondrial genes, when analyzed under NTE, yields strong support for scorpions as sister taxon to the rest of Chelicerata. We investigated the robustness of this result by exploring the effect of adding additional chelicerate genes and taxa and comparing the phylogenies obtained under different models. We find evidence that (1) placement of scorpions arising at the base of the Chelicerata is an artifact of model mis-specification and scorpions are strongly supported as basal arachnids and (2) an expanded chelicerate dataset finds support for several proposed interordinal relationships (ticks plus mites [Acari] and spiders plus whip spiders plus whip scorpions [Araneae + Pedipalpi]). Mitochondrial sequence data are subject to systematic bias that is positively misleading for evolutionary inference and thus extreme methodological care must be taken when using them to infer phylogenies.

Keywords: Phylongeny; Scorpiones; Chelicerata; Mitochondrion; Strand-bias; NTE

1. Introduction

1.1. Chelicerates

The chelicerates (subphylum Chelicerata) are a diverse group of arthropods including many organisms of medical (*Ixodes scapularis*, a Lyme disease vector), economic (*Rhipicephalus* [formerly *Boophilus*, see Barker and Murrell, 2004] *microplus*, a cattle tick) and evolutionary (Xiphosura, horseshoe crabs) interest. Sea spiders (class: Pycnogonida) are sometimes considered chelicerates but several alternative hypotheses have not been ruled out (Dunlop and Arango, 2005). The phylogenetics of chelicerates has been studied at many levels, from the species (Dobson and Barker, 1999) to the phylum (Mallatt et al., 2004). Major problems are the relationships between the various chelicerate orders, for which many solutions have been proposed but on which a consensus is yet to be reached (Weygoldt, 1998). In particular, the phylogenetic position of the Scorpiones, a key order in the arachnid phylogeny, is highly disputed (Weygoldt, 1998; Dunlop and Braddy, 2001; Dunlop and Webster, 1999; Giribet et al., 2002; Wheeler and Hayashi, 1998). Scorpions

* Corresponding author. Fax: +44 131 650 7489.
E-mail address: martin.jones@ed.ac.uk (M. Jones).

1055-7903/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.ympev.2006.11.017

have been variously proposed to be basal (Hypothesis 1, Fig. 1a) or derived (Hypothesis 3, Fig. 1c) arachnids (reviewed in Wheeler and Hayashi, 1998). Although scorpions are a key taxon for the understanding of chelicerate evolution, current sequence datasets are limited in extent. In particular, scorpion mitochondrial genomes have not been fully exploited (Gantenbein and Largiadèr, 2003). Fourteen complete mitochondrial genomes from Acari (ticks and mites), three from spiders, and one from a New World scorpion (*Centruroides limpidus*: Buthidae) have been sequenced. Here, we describe the mitochondrial genome of the East Mediterranean scorpion *Mesobuthus gibbosus* (Buthidae) and use it to further explore chelicerate relationships.

1.2. Multigene datasets

Hypotheses regarding the relationships between chelicerate orders have been largely based on morphology (Weygoldt and Paulus, 1979; Giribet et al., 2002; Wheeler and Hayashi, 1998; Regier and Shultz, 2001). Nuclear ribosomal RNA genes have been used for wider arthropod and chelicerate phylogeny (Mallatt and Giribet, 2006; Wheeler and Hayashi, 1998). Here, we examine the utility of mitochondrial sequence data for addressing questions of chelicerate phylogeny. Recent data suggest that large datasets, comprising many genes, can resolve problematic phylogenies with a high degree of confidence (Philippe et al., 2005a; Rokas et al., 2003). By combining the information from multiple gene sequences, clades can be recovered that are not recovered under analysis of any of the individual genes ('hidden support', described in de Queiroz and Gatesy, 2006). Different genes with different evolutionary rates may give strong phylogenetic signals at different depths in a phylogenetic tree (Giribet, 2002). Thus, by including multiple genes in a phylogenetic study, one could obtain a tree that any single gene would be unable to resolve, as advocated earlier by the total evidence proponents (e.g., Kluge, 1989; Nixon and Carpenter, 1996). The use of multiple genes for phylogenetics comes with its own set of difficulties, the most significant of which are computational complexity, and the need for evolutionary models that describe the variation between genes (Pupko et al., 2002). As the number of genes (and hence the number of characters) included in a multiple sequence alignment grows, so does the time required to evaluate the likelihood or parsimony score of a corresponding phylogenetic tree and hence the time required to execute tree search algorithms. The choice of evolutionary model, always a critical issue in phylogenetic reconstruction (Lemmon and Moriarty, 2004), is particularly important where multiple genes are involved. If the genes evolve under different evolutionary constraints, a single model of DNA evolution may not accurately describe the history of all characters in the alignment, and separate models and parameters may have to be assigned to each gene. Additionally, if some gene sequences are unavailable for some taxa, the alignment may have an appreciable proportion of missing data which may adversely affect the robustness of the tree (Wiens, 2003).

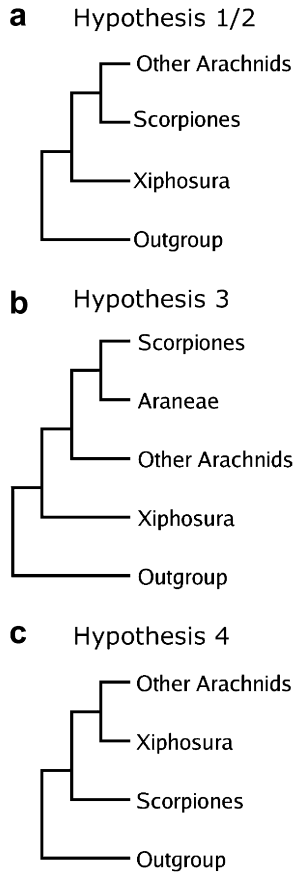

Fig. 1. Summary trees showing various phylogenetic hypotheses regarding the placement of Scorpiones. Scorpiones are shown (a) as sister taxon to other arachnids (b) as sister taxon to other chelicerates (c) as derived arachnids forming a sister taxon to Araneae.

1.3. Strand-bias

In a single gene phylogenetic study involving few characters, the accuracy of the tree is limited by the amount of phylogenetic signal present. In contrast, in a multigene study, the accuracy of the tree is more likely to be limited by systematic errors. Phenomena such as differing substitution rates between lineages (Brinkmann et al., 2005; Felsenstein, 1978), differing patterns of rates between lineages (Gadagkar and Kumar, 2005; Philippe et al., 2005b) and compositional bias (Galtier and Gouy, 1995) have all been found to lead to tree reconstruction artifacts. Model choice has been shown to be a key factor in overcoming difficulties associated with analysis of biased sequences (Lemmon and Moriarty, 2004; Posada and Buckley, 2004; Sullivan and Swofford, 2001). Additionally, differing patterns of evolution between genes can cause problems in phylogenetic reconstruction under models that fail to take inter-gene differences into account (Nylander et al., 2004).

All sequences in non-recombining animal mitochondrial genomes are tightly linked and, in some respects, behave as a single gene. It might therefore be supposed that mitochondrial gene sets would be affected by biases to a similar

Please cite this article in press as: Jones, M. et al., The effect of model choice on phylogenetic inference using mitochondrial sequence data: Lessons from the scorpions, Mol. Phylogenet. Evol. (2007), doi:10.1016/j.ympev.2006.11.017

degree, and hence show fewer differences between genes. Recently, strand-bias in mitochondrial genomes has been shown to have a misleading effect on phylogenetic reconstruction (Hassanin, 2006; Hassanin et al., 2005). In typical metazoan mitochondrial genomes one strand shows a compositional bias towards A and C nucleotides (positive AT and CG skew). This has been proposed to be caused by the asymmetry of the replication and transcription processes. In some arthropod species, comparisons of orthologous genes reveal a reversed bias (towards T and G nucleotides), which is believed to be driven by inversion of the mitochondrial control region (Reyes et al., 1998; Tanaka and Ozawa, 1994). In phylogenetic studies, some taxa with reverse strand-bias have been shown to cluster together artefactually when analyzed under standard evolutionary models (Hassanin, 2006). Hassanin et al. specified a novel recoding scheme (Neutral Transitions Excluded, NTE) which, by excluding neutral or quasi-neutral transitions, eliminates changes driven by both normal and reverse strand-bias. When analyzed under the NTE model, a dataset of six mitochondrial protein-coding genes for 71 arthropod species yielded a phylogeny with Chelicerata monophyletic and Scorpiones as sister taxon to all other chelicerates (Hassanin, 2006, Hypothesis 2, Fig. 1). To investigate the robustness of this result, and the utility of the NTE model for chelicerate phylogeny using mitochondrial data, we used a bioinformatics pipeline to mine publicly available sequence data and assemble a dataset of aligned chelicerate mitochondrial genes, including species for which only a few mitochondrial gene sequences were available. Bayesian phylogenetic analysis was performed on subsets of these sequences using a variety of evolutionary models.

2. Materials and methods

2.1. Mesobuthus mitochondrion

The sequenced mtDNA is of a specimen of *M. gibbosus* collected by V.F. from Visitsa, Mt. Pilion, Eastern Thessaly, Greece. Fresh DNA was extracted using the Puregene extraction kit (Gentra). About 200–400 ng total genomic DNA was used for 50 µl PCR. Amplicons were checked for single bands on agarose gels, and subsequently purified as previously described (Gantenbein and Largiadèr, 2003). The three overlapping fragments were sequenced with the primer walking technique. PCR primers and internal sequencing primers are given in Supplementary Material. Amplification was performed using the expand-long template PCR system from Roche and the following cycling profile: 94 °C for 2 min, 10 cycles (94 °C for 15 s, 55 °C for 1 min, 68 °C for 4 min), and 20 cycles (94 °C for 15 s, 54 °C for 1 min, 68 °C for 4 min + 15 s every cycle), final extension at 68 °C for 15 min. PCR products were subcloned into the plasmid vector pCR2.1 using the TOPO Cloning Kit (Invitrogen). At least four positive clones per transformation were picked and sequenced with M13 and internal sequencing primers using the primer walking technique.

Sequencing reactions were carried out using Big Dye according to the manufacturer's protocol and were analyzed on an automated sequencer (either ABI377XL or ABI3730). Only sequences with <5% ambiguities were considered for analysis. A total of 96 sequences were aligned to produce the final contig (51 sequences on the forward strand and 45 on the reverse strand). Five percent of the bases were sequenced only once and the average sequence coverage was ~5-fold.

A consensus sequence was built using Seqman from the Dnastar package (using a minimum sequence identity of 80%) and imported into Jellyfish for annotation. The protein-coding genes were identified by the presence of an open reading frame and by similarity of inferred amino acid sequences to those of other complete chelicerate mitochondrial genomes available from public databases. Abbreviated stop codons of "T"-type (which are completed by post-transcriptional polyadenylation) were assumed if they matched with the boundary of a downstream gene or were found less that 20 bp before the adjacent gene.

The boundaries of the small ribosomal (12S rRNA) and the large ribosomal (16S rRNA) RNA genes were determined by prediction of the secondary RNA structure using the Vienna RNAserver (Hofacker, 2003) and subsequent comparison with other arthropods (Wuyts et al., 2004). Additionally, a multiple sequence alignment of the scorpion mitochondrial genome with that of 13 other chelicerates was performed to further confirm the gene boundaries (Mallatt et al., 2004; Masta and Boore, 2004).

For the identification of tRNAs the program tRNAscan-SE 1.21 (Lowe and Eddy, 1997) was used, following the protocol of Masta and Boore (2004) and relaxing search parameters (Cove cut-off to >0.1). The N-tRNA gene was manually annotated by finding instances of the appropriate codon and aligning the surrounding sequence to N-tRNA from the scorpion *Centruroides limpidus*. tRNA structures were drawn using the graphic view option implemented in tRNAscan-SE. For gene abbreviations we followed Masta and Boore (2004).

2.2. Bioinformatics

A bioinformatics pipeline (Jones and Blaxter, 2006; available from http://www.nematodes.org/bioinformatics/), utilising Perl, BioPerl (Stajich et al., 2002) and a number of other tools, was used to carry out the processes of sequence extraction, consensus building, alignment, taxon selection and phylogenetic reconstruction. At each stage of the analysis, output was stored in a relational database for retrieval in the next stage.

2.3. Sequence extraction

All GenBank records for chelicerate species were obtained through NCBI ENTREZ (http://www.ncbi.nlm.nih.gov/entrez/, 23/6/2005) and the sequences of interest (corresponding to the genes ATP6, ATP8,

Please cite this article in press as: Jones, M. et al., The effect of model choice on phylogenetic inference using mitochondrial sequence data: Lessons from the scorpions, Mol. Phylogenet. Evol. (2007), doi:10.1016/j.ympev.2006.11.017

COX1, COX2, COX3, CYTB, ND1, ND2, ND3, ND4, ND4L, ND5, ND6, RNA_12S, RNA_16S, RNA_SSU, and RNA_LSU) extracted based on annotation. Sequences of interest were also obtained for the following outgroup arthropods: *Drosophila melanogaster* (Hexapoda, Diptera), *Triatoma dimidiata* (Hexapoda, Heteroptera), *Triops cancriformis* (Crustacea, Notostraca), *Daphnia pulex* (Crustacea, Cladocera), *Scutigera coleoptrata* (Myriapoda, Scutigeromorpha). EST sequences (those derived from expressed sequence tags, e.g., Kenyon et al., 2003) were processed separately: a BLAST (Altschul et al., 1997) database was assembled containing complete translated protein sequences for the protein-coding genes of interest from a range of chelicerate species, and filtered at 80% redundancy using BLASTCLUS. To screen EST sequences for genes of interest, each EST was used as the query sequence in a BLASTX search of this database. If the EST sequence had a hit in the database with an *E*-value of less than $1e^{-8}$, it was assigned the same gene name as that of the hit. A similar screening protocol was used to identify mitochondrial ribosomal RNA (rRNA) genes, using a database of complete rRNA gene sequences from a range of chelicerate species in a BLASTN search.

2.4. Consensus building

For each of the genes of interest, a consensus sequence was built for each species using a rule-based approach. Sequences derived from fully sequenced genomes were preferred over sequences extracted on the basis of annotation, which were, in turn, preferred over EST sequences extracted on the basis of BLAST similarity to known sequences. If multiple sequences of a single type were available (e.g., there were two annotated sequences present for a gene in a particular species), phrap (Gordon et al., 1998) was used to generate a consensus sequence, with the longest consensus sequence being preferred if multiple consensus sequences were generated. Consensus sequences that were very short (<30 bases), contained long mononucleotide/dinucleotide runs (>12 bases) or produced incomplete translations were removed.

2.5. Alignment

For each gene of interest, a multiple sequence alignment was generated using POA (Lee et al., 2002). In preliminary analyses the local alignment strategy used in POA proved to be a good solution for aligning sequences with missing data. For protein-coding genes, the alignment was carried out in two stages. Firstly, the translated protein sequences for all genome-derived and annotated sequences were aligned, then the nucleotide sequences aligned with TRANALIGN from the EMBOSS package (Olson, 2002) using the protein alignment as a guide. Secondly, any EST-derived sequences were aligned to this nucleotide alignment using the POA profile alignment option. Aligned DNA sequences were stored in the database. Complete DNA alignments for each individual gene were analyzed using Modeltest (Posada and Crandall, 1998). The model selected by Akaike information criterion (AIC) was, in each case, either General Time Reversible (GTR) or Tamura–Nei (TrN), with gamma rate variation between sites (G) and a proportion of invariant sites (I) selected in all cases.

2.6. CG skew

For each gene, third codon position bases were extracted from sequences from species in the D1 dataset. CG skew was calculated from 5′ to 3′ in each gene for those bases using the following formula: where C and G represent the proportion of C and G nucleotides, respectively. Skew was calculated both before and after NTE recoding. Under the NTE recoding scheme, all third codon position bases are recoded as purine (R) or pyrimidine (Y). For NTE-recoded sequences, the proportion of C and G nucleotides was calculated as half the proportion of Y and R nucleotides, respectively.

2.7. Phylogenetic analysis

Multiple sequence alignments suitable for phylogenetic analysis were generated by specifying a subset of genes and a subset of taxa and extracting the corresponding pre-aligned sequences from the database. Multiple sequence alignments are available from the corresponding author on request. The alignments were analyzed using MrBayes 3.1 (Ronquist and Huelsenbeck, 2003). Several different analyses were carried out on the D1 and D2 datasets (Table 2). By default, a GTR + I + G model (nst = 6; rates = invgamma) was applied to all alignments. For those models where the NTE scheme was used, bases were recoded according to Hassanin, Leger and Deutsch (2005) and third codon position bases were assigned to a separate partition with two substitution types (nst = 2). For those models where the alignment was partitioned by gene base frequencies, substitution rates, alpha parameters and proportions of invariant sites were unlinked across partitions, and a rate multiplier was used to allow rate variation between partitions.

Analyses were carried out in MrBayes 3.1.2 using default MCMCMC parameters (two independent runs with one cold chain and three heated chains per run) for 1,000,000 generations. Convergence of split frequencies and flattening of likelihood scores were inspected to ensure stationarity and convergence between runs and the first 100,000 generations (10%) discarded as burn-in. Trees sampled after the burn-in period were summarized to give the 50% majority rule consensus tree. To calculate Bayes factor support for the hypothesis corresponding to the optimal tree, analyses were run using parameter estimates from the initial analysis with the following taxa constrained to be monophyletic: Hypothesis 2—*Ixodes hexagonus, Mastigoproctus giganteus, Limulus polyphemus, Argiope bruennichi, Phrynus* sp., *Leptotrombidium pallidum, Heptathela hangzhouensis,*

Please cite this article in press as: Jones, M. et al., The effect of model choice on phylogenetic inference using mitochondrial sequence data: Lessons from the scorpions, Mol. Phylogenet. Evol. (2007), doi:10.1016/j.ympev.2006.11.017

Cario capensis; Hypothesis 3—*Argiope bruennichi, Heptathela hangzhouensis, Mesobuthus gibbosus, Euscorpius flavicaudis.*

3. Results

3.1. Genome annotation and tRNA scanning

We have sequenced nearly all of the mitochondrial genome of *M. gibbosus* (15,681 bp). The sequence has been deposited in the EMBL sequence database under accession AJ716204. The genome is AT-rich, having a GC content of 31.6%. About 300 bp of the *M. gibbosus* mitochondrial genome were not sequenced due to highly repetitive and AT-rich DNA motifs in the putative D-loop (control region). Most protein-coding genes in the *M. gibbosus* mitochondrial genome are initiated with the start codon ATG. However, alternative initiation codons ATN are used for the genes *atp8*, *nad1*, *nad2*, *nad4*, *nad5* and *nad6*. For termination, both stop codons (TAA, TAG) and the incomplete stop codon "T" (*coxII*, *coxIII*, *nad3*, *nad4*, *nad5* and *nad6*) are used. Initial analysis of the genome sequence using tRNAscan (Lowe and Eddy, 1997) with default parameters identified eleven tRNA genes (A, F, I, K, M, R, T, S1, P, W and Y). Using the variant model for nematode tRNA genes and reducing stringency settings identified 10 additional tRNA genes. The tRNA-N gene could only be identified manually. Four of the tRNA genes conspicuously lack a TΨC arm (a phenomenon previously described in spiders (Masta and Boore, 2004) and nematodes (Wolstenholme et al., 1987) and five had extremely short aminoacyl receptor stems with atypical base-pairing (see Supplementary Materials for predicted structures of all *M. gibbosus* mitochondrial tRNAs).

Mitochondrial genome gene order has been used as a phylogenetic marker (Boore and Brown, 1998, 2000). The gene order of the *M. gibbosus* mitochondrial genome (see Supplementary Material for a map of the genome) is identical to that of the recently sequenced New World buthid scorpion, *Centruroides limpidus* (Dávila et al., 2005). Gene order is more conserved between *M. gibbosus* and the horseshoe crab *Limulus polyphemus* (three tRNA gene rearrangements) than between *L. polyphemus* and the spider *Habronattus oregonensis* (eight tRNA gene rearrangements).

3.2. Phylogenetic analysis

We examined a subset (D1) consisting of 12 protein-coding genes for species from the chelicerate orders considered in Hassanin (2006). We then examined a second subset (D2) containing sequences from additional chelicerate orders, which incorporated a significant proportion of missing data. In all analyses of D1 and D2, Pycnogonida was placed within the ticks and mites. Placement of Pycnogonida with Acari has been suggested previously (reviewed in Dunlop and Arango, 2005) but is unexpected, since morphological evidence has been found to place pycnogonids as sister taxon to all other chelicerates or to all other arthropods (e.g., Maxmen et al., 2005; see Giribet et al., 2005). Owing to the unexpectedness of this result, and the relatively small amount of sequence data available for Pycnogonida (~40% in D1, ~30% in D2) we regard this placement as tentative until confirmed by future analyses and have therefore ignored Pycnogonida when describing relationships in the following sections.

3.3. Dataset D1: mostly complete mitochondrial genomes

The D1 dataset consisted of 12 protein-coding genes for 11 chelicerate species and five outgroup species (Table 1). To keep the analysis computationally tractable, we used only a subset of the taxa for which mitochondrial genomes are available, selecting taxa to obtain the widest taxonomic coverage. In preliminary analyses, all orders were found to be monophyletic (with the exception of Acari because *Varroa destructor* behaves erratically under phylogenetic analysis due to its extremely high AT content [Navajas et al., 2002]) and inclusion of multiple representatives of each order did not alter the results. In order to investigate the effect of model, recoding scheme and codon position weighting on the phylogeny, six analyses were carried which are given in Table 2, along with the summary trees supported under each model (Fig. 2). The analyses differ in (1) the inclusion or exclusion of the third bases of codons (2) the use of the NTE recoding scheme and (3) permitting each gene to have different model parameters (partitioning).

Different phylogenies were obtained from Bayesian analysis of the D1 dataset under the different analyses (Fig. 2; hypotheses in Fig. 1). The three analyses in which third codon position bases were excluded yielded similar trees (Hypothesis 2) with support for scorpions as sister taxon to other chelicerates (rendering Arachnida paraphyletic). The analysis using the GTR + I + G model with partitioning (Fig. 2a) gave a tree with the remaining arachnids divided into two clades; one of spiders and related orders ((Araneae + Uropygi) + Amblypygi) and one of mites and ticks (Ixodida + Trombidiformes). Using the NTE recoding scheme (Fig. 2b), an identical tree was recovered with the exception of rearrangements within the two major arachnid clades. In the analysis using the NTE recoding scheme with partitioning (Fig. 2c), the tree was identical to that obtained under the NTE analysis except for the relationships within the clade of mites and ticks, which were identical with those of the GTR partitioned tree, and the reduced support for scorpions as sister taxon to other chelicerates (75% compared to 95%).

When third codon position bases were included, each analysis recovered a different tree. Under the GTR + 3p analysis (Fig. 2d) scorpions were no longer sister taxon to all other chelicerates. A clade of scorpions and the spider *Argiope bruennichi* was recovered, with a (Uropygi + *Heptathela hangzhousensis*) clade as a sister group (summary tree c with the exception of paraphyletic

Please cite this article in press as: Jones, M. et al., The effect of model choice on phylogenetic inference using mitochondrial sequence data: Lessons from the scorpions, Mol. Phylogenet. Evol. (2007), doi:10.1016/j.ympev.2006.11.017

Table 1
List of taxa included in datasets D1 and D2

Taxon	Order[a]	NCBI TXID[b]	Ref. Seq. ID[c]	Citation[d]	Dl (12,669 characters)[e]		D2 (15,950 characters)[f]	
					No. chars.[g]	% Present[h]	No. chars.[g]	% Present[h]
Unidentified Opilioacarid	Opilioacarida	150113	—	—	—	—	257	1.61%
Unidentified Allothyrid	Holothyrida	91335	—	—	—	—	539	3.38%
Steganacarus magnus	Oribatida	52000	—	—	—	—	387	2.43%
Camisia horrida	Oribatida	240610	—	—	—	—	597	3.74%
Sarcoptes scabei	Astigmata	197185	—	—	—	—	7394	46.36%
Opilio parietinus	Opiliones	121214	—	—	—	—	478	3.00%
Ixodes hexagonus	Ixodida	34612	NC_002010	Black and Roehrdanz (1998)	11820	93.30%	14449	90.59%
Mastigoproctus giganteus	Uropygi	58767	—	—	3756	29.65%	4150	26.02%
Limulus polyphemus	Xiphosura	6850	NC_003057	Lavrov, Boore and Brown (2000)	12108	95.57%	14692	92.11%
Argiope bruennichi	Araneae	94029	—	—	4857	38.34%	4863	30.49%
Phrynus sp.	Amblypygi	309714	—	—	5154	40.68%	5165	32.38%
Endeis spinosa	Pycnogonida	136194	—	—	5019	39.62%	5030	31.54%
Mesobuthus gibbosus	Scorpiones	123226	NC_006515	This work	12090	95.43%	14331	89.85%
Euscorpius flavicaudis	Scorpiones	100976	—	—	4926	38.88%	6100	38.24%
Leptotrombidium pallidum	Trombidiformes	279272	NC_007177	Shao et al. (2005)	11487	90.67%	13776	86.37%
Heptathela hangzhouensis	Araneae	216259	NC_005924	Qiu et al. (2005)	11997	94.70%	14472	90.73%
Carios capensis	Ixodida	176285	NC_005291	—	11922	94.10%	14444	90.56%
Scutigera coleoptrata[i]	Scutigeromorpha	29022	NC_005870	Negrisolo et al. (2004)	11979	94.55%	14496	90.89%
Triops cancriformis[i]	Notostraca	194544	NC_004465	Umetsu et al. (2002)	12063	95.21%	14613	91.62%
Daphnia pulex[i]	Diplostraca	6669	NC_000844	Crease (1999)	12135	95.78%	14727	92.33%
Triatoma dimidiata[i]	Hemiptera	72491	NC_002609	Dotson and Beard (2001)	12003	94.74%	14578	91.40%
Drosophila melanogaster[i]	Diptera	7227	NC_001709		12123	95.69%	14723	92.31%

[a] The order to which each species belongs (NCBI taxonomy).
[b] The Taxonomy ID (TXID) assigned to the species by the NCBI GenBank database.
[c] The Ref. Seq. ID (where applicable) for the complete mitochondrial genome sequence.
[d] The citation (where applicable) for the complete mitochondrial genome sequence.
[e] Dataset Dl includes the following genes: ATP6, C0X1, C0X2, C0X3, CYTB, ND1, ND2, ND3, ND4, ND4L, ND5, ND6.
[f] Dataset D2 includes the following genes: ATP6, ATP8, C0X1, C0X2, C0X3, CYTB, ND1, ND2, ND3, ND4, ND4L, ND5,ND6, RNA_12S, RNA_16S.
[g] No. chars.: the number of aligned characters in each species.
[h] % present: the proportion of bases in the complete alignment present in each species.
[i] Outgroup.

Table 2
Details of models of sequence evolution used in phylogenetic analyses of the Dl and D2 datasets

Analysis name[a]	Model of base change	Codon positions	Recoded	Partitioned	Dl tree[g]	D2 tree[g]
GTRp	GTR + G + F	1,2	n	y[e]	4	4
NTE	NTE + G + I[c]	1,2	y[d]	n	4	—
NTEp	NTE + G + I[c]	1,2	y[d]	y[f]	4	—
GTR + 3p	GTR + G + I[b]	1,2,3	n	y[e]	3	—
NTE + 3	NTE + G + I[c]	1,2,3	y[d]	n	4	—
NTE + 3p	NTE + G + I[c]	1,2,3	y[d]	y[f]	1/2	1/2

[a] Name used to refer the analysis in the text. Codon positions 1 and 2 are included in all analyses. Third codon positions are only included in analyses that contain +3. Analyses in which the dataset was partitioned by gene end in *p*.
[b] General Time Reversible (GTR) model applied to all bases (nst = 6 in MrBayes). All models include gamma rate variation and a proportion of invariant sites.
[c] Neutral Transitions Excluded (NTE) model; GTR applied to first and second codon position bases (nst = 6) and two substitution type model applied to third codon position bases (nst = 2). All models include gamma rate variation and a proportion of invariant sites.
[d] Bases are recoded according to the NTE scheme of Hassanin, 2005. All third codon position bases and a subset of first and second codon position bases are recoded as purine/pyrimidine (R/Y), eliminating phylogenetic signal caused by neutral and nearly-neutral transitions.
[e] The alignment is partitioned by gene. Each partition has independent base frequencies, transition rate matrix, gamma parameter and proportion of invariant sites. A rate multiplier is used to permit rate variation between partitions.
[f] First and second codon position bases are partitioned by gene and all third codon position bases form a separate partition. Each partition has independent base frequencies, transition rate matrix, gamma parameter and proportion of invariant sites. A rate multiplier is used to permit rate variation between partitions.
[g] The hypothesis (see Section 1 and Fig.1) supported by each dataset under the analysis.

Please cite this article in press as: Jones, M. et al., The effect of model choice on phylogenetic inference using mitochondrial sequence data: Lessons from the scorpions, Mol. Phylogenet. Evol. (2007), doi:10.1016/j.ympev.2006.11.017

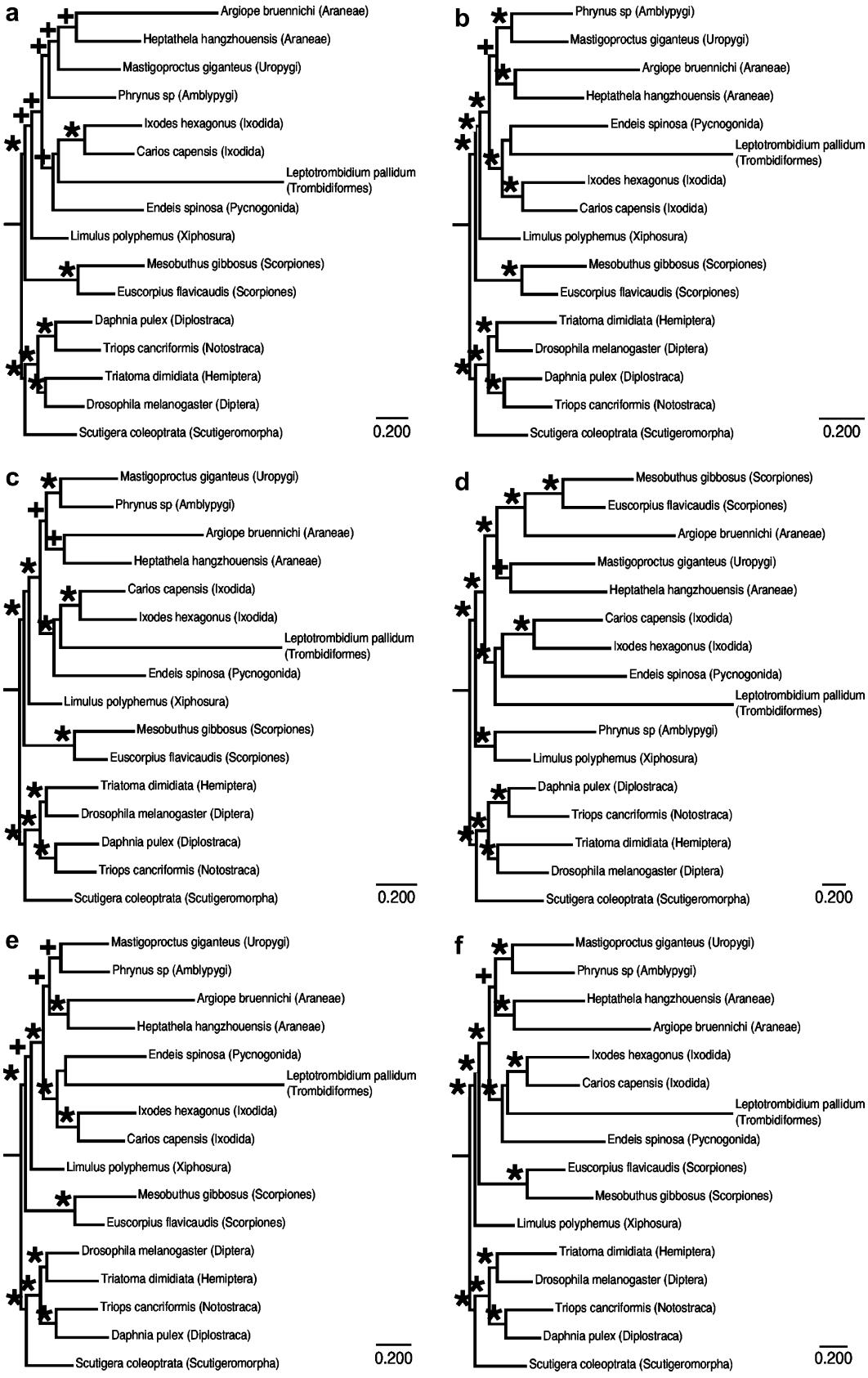

Fig. 2. Phylogeny reconstructed from the D1 dataset using Bayesian analysis under the following models: (a) GTRp (b) NTE (c) NTEp (d) GTR + 3p (e) NTE + 3 (f) NTE + 3p (see Table 2 for details of models). The order to which each species belongs is given in parentheses. The scale bar shows the branch length associated with 0.200 expected changes per site. Labels on branches indicate the level of support: (∗), posterior probability = 1.00; (+), posterior probability = 0.9–0.99. Posterior probabilities of >0.9 are considered significant; lower posterior probabilities are not shown.

Please cite this article in press as: Jones, M. et al., The effect of model choice on phylogenetic inference using mitochondrial sequence data: Lessons from the scorpions, Mol. Phylogenet. Evol. (2007), doi:10.1016/j.ympev.2006.11.017

Araneae). A clade of ticks and mites was recovered. A clade consisting of Xiphosura + Amblypygi was sister taxon to all other chelicerates. Under the NTE + 3 analysis (Fig. 2e) the tree was identical to that found under NTE (summary tree a) although with negligible support for scorpions as sister taxon to all other chelicerates. Under the *NTE* + 3p analysis (Fig. 2f) Arachnida was monophyletic, with Scorpions a sister taxon to the other arachnids, and Xiphosura as sister taxon to the other chelicerates (Hypothesis 1). Two major clades of arachnids were recovered; one of spiders and related orders (Araneae + (Uropygi + Amblypygi)) and one of mites and ticks (Ixodida + Trombidiformes). Analyses under NTE + 3p when the tree was constrained to support Hypothesis 2 or Hypothesis 3 yielded harmonic mean log-likelihood estimates of −81,004.98 and −80,996.41, respectively, compared with an estimate of −80,722.24 for the optimal tree.

To test whether the data supported use of partitioned model, we estimated harmonic mean log-likelihoods for two pairs of analyses that varied only in the model; NTE versus NTEp and NTE + 3 versus NTE + 3p. In each case the partitioned model was preferred: NTE: −53885 ≪ NTEp: −53193; NTE + 3: −82056 ≪ NTE + 3p: −80721.

When CG skew was examined at third codon positions, the species in the D1 dataset fell into two groups (Fig. 4). For the majority of species the mitochondrial genes had a characteristic pattern of positive and negative skew. Three species (*M. gibbosus, A. bruennichi, E. flavicaudis*) had a pattern that was the reversed relative to other species in the genes for which data was available. After NTE recoding, no such pattern was discernible.

3.4. Dataset D2: including single mitochondrial genes for some taxa

The D2 dataset consisted of thirteen protein-coding genes and two rRNA genes for 17 chelicerate species and five outgroup species (Table 1). The number and percentage of characters present varied between taxa, ranging from 257 (1.61%) characters present for an unidentified Opilioacarid (Opilioacaridae) to 14723 (92.31%) characters present for *D. melanogaster* (outgroup). Taxa with complete mitochondrial genomes, such as *D. melanogaster* had ~8% missing characters due to minor differences in sequence length between taxa, resulting in end gaps. We included species with a range of numbers of characters present to investigate the effects of missing data on the phylogenetic placement of 'neglected' taxa.

Under the NTE + 3p analysis (Fig. 3), Xiphosura was robustly placed arising as sister taxon to the remaining chelicerates, with Scorpiones a sister taxon to the remaining arachnids (Hypothesis 1) excluding Pycnogonida (see above). Relationships within the arachnids were generally poorly resolved, although some groups were strongly supported: (Ixodida + Holothyrida) and (Uropygi + Amblypygi). In the GTRp analysis Hypothesis 2 was recovered (not shown).

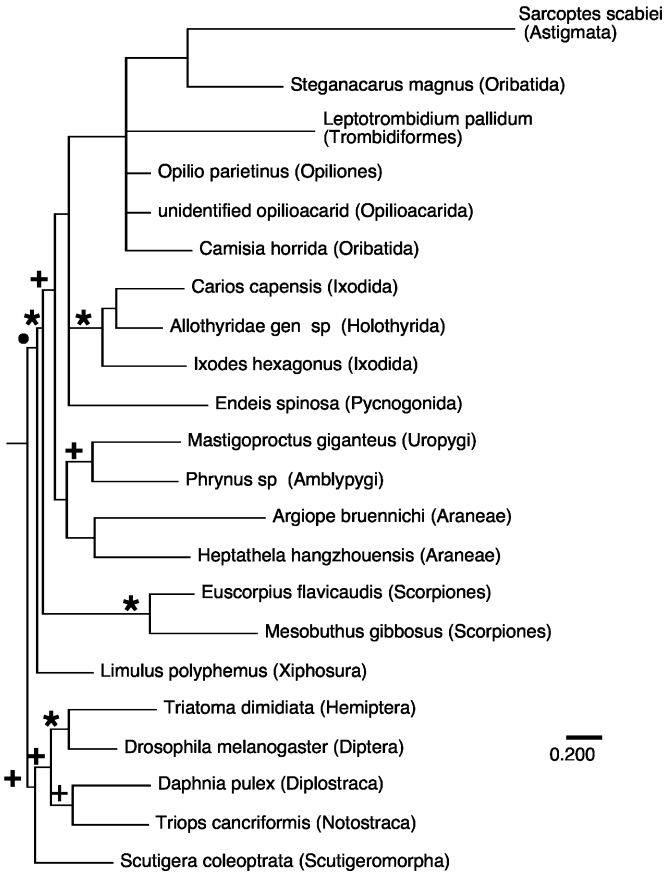

Fig. 3. Phylogeny reconstructed from the D2 dataset using Bayesian analysis under the NTE + 3p model (see Table 2 for details of models). The order to which each species belongs is given in parentheses. The scale bar shows the branch length associated with 0.200 expected changes per site. Labels on branches indicate the level of support: (∗), posterior probability = 1.00; (+), posterior probability = 0.9–0.99; and (●), posterior probability = 0.80–0.89. Posterior probabilities of >0.9 are considered significant. Posterior probabilities of less than 0.80 are not shown.

4. Discussion

4.1. Mesobuthus gibbosus mitochondrion

The *M. gibbosus* mitochondrial genome is ~16 kb in size and encodes the usual metazoan complement of 13 protein genes and two ribosomal RNA genes. The order of these genes is the same as observed in another scorpion, *Centruroides limpidus* (Dávila et al., 2005), and in the horseshoe crab, *Limulus polyphemus* (Lavrov et al., 2000). It has previously been proposed that the gene organization observed in *L. polyphemus* is plesiomorphic for arthropods in general. The tRNA genes were difficult to identify, due to reductions in the TΨC or DHU arms in 14 genes and reduction in the lengths of the aminoacyl acceptor stem regions. Reduced tRNAs (compared to the standard metazoan model) have been observed previously in other chelicerates, such as spiders (Masta and Boore, 2004; Qiu et al., 2005); millipedes (Lavrov et al., 2002), and nematodes (Ohtsuki et al., 1998; Watanabe et al., 1994). In *Lithobius forficatus*, a centipede,

Please cite this article in press as: Jones, M. et al., The effect of model choice on phylogenetic inference using mitochondrial sequence data: Lessons from the scorpions, Mol. Phylogenet. Evol. (2007), doi:10.1016/j.ympev.2006.11.017

RNA editing has been described that stabilizes reduced tRNA secondary structure (Lavrov et al., 2000), but whether this also occurs in chelicerates is unknown at present. The arrangement of tRNAs on the mitochondrial genome shows four differences compared with *L. polyphemus* (possibly the result of three rearrangements of the genes for tRNAs I and D individually and [P,T] as a pair).

4.2. Evolutionary models

The reversals of strand-bias documented by Hassanin (2006) clearly show how systematic error can limit the usefulness of multigene phylogenetic analysis. Multiple independent reversals of strand-bias in mitochondrial genomes and mitochondrial genes cause erroneous clustering of distantly related arthropod taxa when alignments are subjected to phylogenetic analysis under a standard General Time Reversible (GTR) model. In their work, Hassanin and colleagues found evidence that strand-bias is mainly generated by transitions (Hassanin, 2006; Hassanin et al., 2005). They proposed a recoding scheme, Neutral Transitions Excluded (NTE) which removes the effect of strand-bias by recoding bases at neutral and nearly-neutral positions as purines and pyrimidines (R/Y coding). Here we have investigated a wider range of analyses, examining the effects of GTR versus NTE, the effects of inclusion or exclusion of third codon position bases, and the effects of a partitioned versus an unpartitioned model.

A notable result is the contribution of phylogenetic signal from third codon position bases. Phylogenetic analyses of dataset D1 in the three analyses that include only first and second codon position bases yield similar trees with scorpions as sister taxon to all other chelicerates (Hypothesis 2). Support for scorpions as sister taxon to all other chelicerates decreases under the NTEp analysis which includes a more realistic partitioned model, suggesting that this placement may be artifactual. In contrast, the three analyses that included third codon position bases gave trees with radically different placement of Scorpiones, and correspondingly different hypotheses regarding chelicerate evolution. In the GTR + 3p analysis, inclusion of third codon position bases yields in a clade of scorpions and *Argiope bruennichi*, a spider (Hypothesis 3). Hassanin (2006) showed that *A. bruennichi* and *Euscorpius flavicaudis* both have a reversed pattern of third position CG skew relative to other arthropods. Analysis of the D1 dataset reveals that *A. buennichi* and *M. gibbosus* both show a reversed CG skew pattern in third codon position bases relative to other species in D1 (Fig. 4) which is removed by NTE recoding, suggesting that the (Scorpiones + *A. bruennichi*) clade is an artifact of mitochondrial strand-bias. This proposition is supported by the results of the NTE + 3 analysis, in which the portion of the third codon position signal due to mitochondrial strand-bias is excluded and scorpions again are supported as sister taxon to the other chelicerates (Hypothesis 2). A final striking result is the change in the tree when the partitioned analysis is used (NTE + 3p). Here, Xiphosura is identified as sister taxon to other chelicerates and Arachnida is monophyletic (ignoring Pycnogonida, see Section 3), with scorpions the sister taxon to other arachnids (Hypothesis 1).

To test the significance of support for the favored hypothesis, we can use the harmonic mean log likelihood of each hypothesis, calculated from the MrBayes runs, to estimate Bayes Factors as described in Kass and Raftery (1995). The Bayes Factors for [Hypothesis 1 vs. Hypothesis 2] and [Hypothesis 1 vs. Hypothesis 3] were 565.48 and 548.34, respectively, indicating very strong support for Hypothesis 1 (Bayes Factor values of >20 are considered to indicate strong support). In the same way, we calculated Bayes Factor support for a partitioned over a non-partitioned model for the NTE and NTE + 3 analyses. The Bayes Factors were 1383.9 and 2670.62, respectively, indicating extremely strong support for the partitioned model in both pairs of analyses.

We explain these results by suggesting that two potential sources of bias are present in the D1 dataset. Firstly, third codon position bases are strongly affected by mitochondrial strand-bias, an effect minimized by using the NTE recoding scheme. Secondly, evolutionary model parameters vary between genes, leading to erroneous results if an underparameterized, unpartitioned model is used. The commonly observed inversions of mitochondrial regions will lead genes to acquire a strand bias that is the opposite of that of the genome as a whole. It is notable that robust support for the phylogenetic position of scorpions was only recovered from mitochondrial DNA sequences under the appropriate model. These findings emphasize the importance of model choice in phylogenetic analysis, and specifically the importance, when using multiple genes, of removing any sources of systematic bias and adequately allowing for differences in evolutionary models between genes.

4.3. D2 dataset

Analysis of the D2 dataset using GTR + 3p and NTE + 3p analyses showed the same pattern as we observed for the D1 dataset (Table 2): the analyses gave Hypothesis 2 and Hypothesis 1, respectively. The poor resolution of the phylogeny recovered under the most realistic analysis, NTE + 3p (Fig. 3) can probably be attributed to the large proportion of missing data, coupled with a loss of phylogenetic signal due to the NTE recoding scheme. While NTE removes erroneous phylogenetic signal due to strand-bias, it will also inevitably remove genuine phylogenetic signal.

In summary, our results endorse a traditional picture of the relationships between chelicerate orders, excluding Pycnogonida. It is important to note that some orders, in particular members of the proposed clade Dromopoda (Opiliones, Scorpiones, Solifugae) are unrepresented due to lack of sequence data. We cannot rule out the possibility that the addition of these taxa to the tree would change the phylogenetic conclusions. With these caveats, our findings support an uncontroversial view of chelicerate evolution,

Please cite this article in press as: Jones, M. et al., The effect of model choice on phylogenetic inference using mitochondrial sequence data: Lessons from the scorpions, Mol. Phylogenet. Evol. (2007), doi:10.1016/j.ympev.2006.11.017

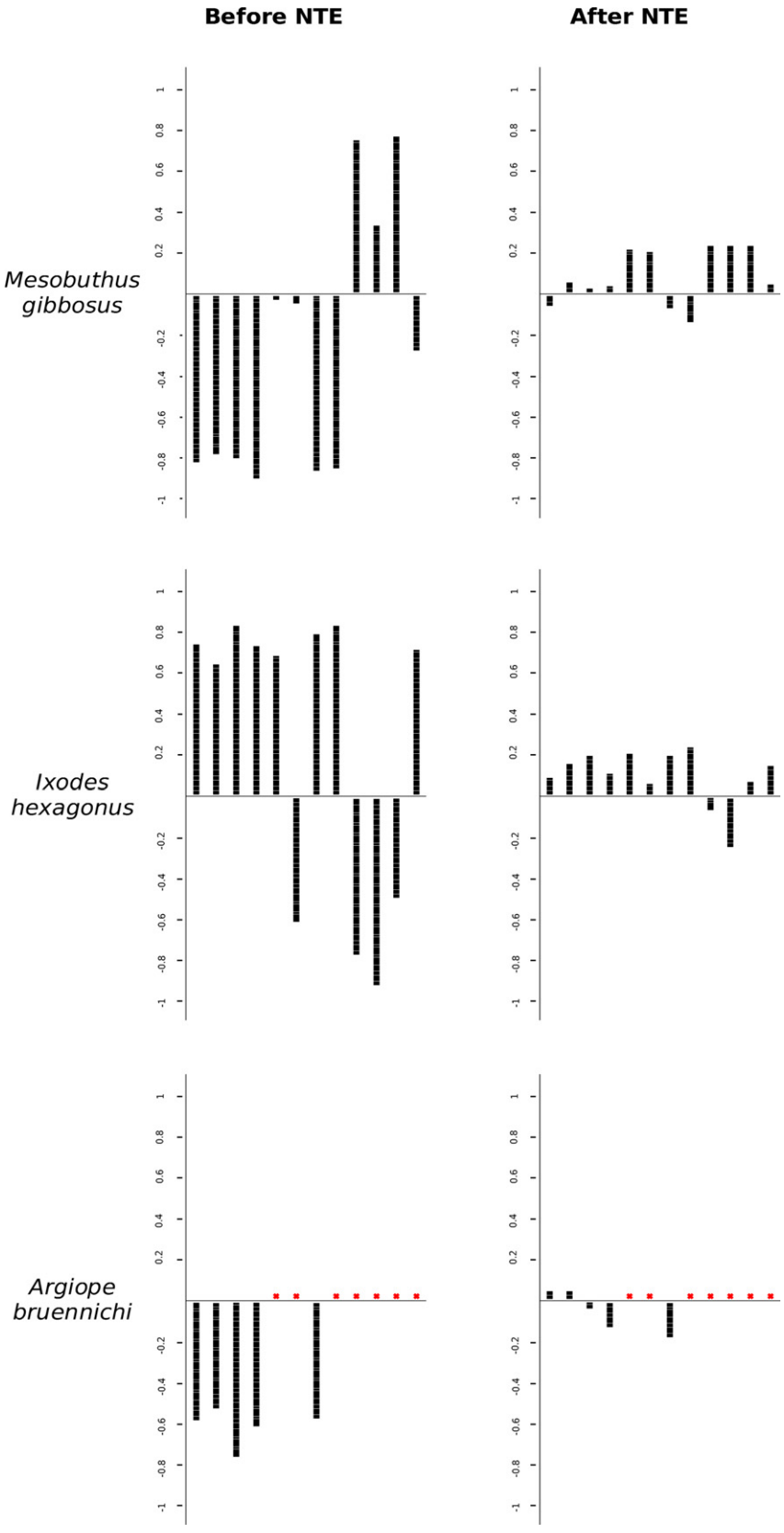

Fig. 4. Graphs showing CG skew (see Section 2 for details) for three representative taxa from the D1 dataset. Each graph shows the skew at third codon position bases for the following genes in order: ATP6, COX1, COX2, COX3, CYTB, ND1, ND2, ND3, ND4, ND4L, ND5, ND6. Graphs on the right represent sequences before NTE recoding; graphs on the left, after NTE recoding. Missing data are indicated by red X's. *E. flavicaudis* showed the same pattern as *M. gibbosus* and *A. bruennichi*; all other taxa showed the same pattern as *I. hexagonus.*.

Please cite this article in press as: Jones, M. et al., The effect of model choice on phylogenetic inference using mitochondrial sequence data: Lessons from the scorpions, Mol. Phylogenet. Evol. (2007), doi:10.1016/j.ympev.2006.11.017

with Xiphosura a sister taxon to other chelicerates and Arachnida monophyletic. Within the Arachnida, Scorpiones are sister taxon to the remaining arachnids which are divided into two high-level clades: ticks plus mites (Acari) and spiders plus whip scorpions (Araneae + Pedipalpi). Under this scheme Scorpiones may occupy a key role in the understanding of arachnid diversity—again, in the absence of all its putative sister groups according to most modern treatments of chelicerate phylogeny (e.g., Shultz, 1990; Wheeler and Hayashi, 1998; Giribet et al., 2002). Morphological, genomic and developmental characters that are shared between scorpions and other arachnids presumably represent the ancestral arachnid state, allowing us to polarize such characters on the chelicerate phylogeny. However, the position of most of the controversial arachnids (e.g., Palpigradi, Pseudoscorpiones, Solifugae, Opiliones and Ricinulei) remain untested.

As has been shown in many previous analyses, large datasets comprising multiple genes can be effective at resolving phylogenies in situations where single genes are insufficient, provided care is taken to avoid systematic bias (Philippe et al., 2005a; Rokas et al., 2003). For multigene mitochondrial data, particular attention must be paid to normal and reverse strand-bias and to parameter differences between genes. Taxa can also be robustly placed using partial datasets. In D1, for example, *A. bruennichi* is robustly placed despite having only 33% of characters present, and in D2, *E. flavicaudis* is robustly placed using only 38% of characters. It is likely that, in both of these cases, the confidence with which the taxa are placed can be attributed to two factors; a large absolute number of characters (4857 coding nucleotides in the case of *A. bruennichi*) and the presence of a closely related species with a near-complete alignment in the dataset. In general, species not fulfilling these criteria are unlikely to be robustly placed by phylogenetic analysis of partial datasets. The sole pycnogonid species, *Endeis spinosa*, had ~40% of characters present in the D1 dataset, but had no closely related species in the alignment and was consistently placed in a clade of mites. More sequence data or a wider taxonomic sampling from the pycnogonids would be necessary to investigate this surprising result.

Given the large historical interest in the use of mitochondrial genes as phylogenetic markers, particularly in the arthropods (Cameron et al., 2006; Cook et al., 2005; Nardi et al., 2003), and the tendency towards the use of large, multiple gene datasets for phylogenetic reconstruction, the effect of model choice is a pertinent issue for those interested in molecular systematics. Strand-bias is one of several types of compositional bias; indeed, it is likely that the composition of most gene sequences is influenced by multiple biasing factors. Our findings, though directly pertaining to only a small set of taxa, suggest that mitochondrial gene sequences may be actively misleading. Care should be taken, when using any multiple gene dataset, to avoid explore the potential for systematic bias. In the case of mitochondrial genes, strand-bias should be of particular concern and the previous use of mitochondrial genomes in resolving deep phylogenies requires critical reevaluation.

Acknowledgments

This work was supported by a BBSRC PhD Studentship to M. J. Phylogenetic analyses were carried out on a condor compute cluster funded by NERC. B.G. was supported by an SNF-IHP Grant (83-EU-065528), the Royal Society (R-36579) and the Basler Stiftung für experimentelle Zoologie (c/o Eidgenössisches Tropeninstitut). We would like to thank three reviewers for their helpful comments.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ympev.2006.11.017.

References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

Barker, S.C., Murrell, A., 2004. Systematics and evolution of ticks with a list of valid genus and species names. Parasitology 129 (Suppl), S15–S36.

Black 4th, W.C., Roehrdanz, R.L., 1998. Mitochondrial gene order is not conserved in arthropods: prostriate and metastriate tick mitochondrial genomes. Mol. Biol. Evol. 15, 1772–1785.

Boore, J., Brown, W., 1998. Big trees from little genomes: mitochondrial gene order as a phylogenetic tool. Curr. Opin. Genet. Dev. 8, 668–674.

Boore, J., Brown, W., 2000. Mitochondrial genomes of *Galathealinum*, *Helobdella*, and *Platynereis*: sequence and gene arrangement comparisons indicate that Pogonophora is not a phylum and Annelida and Arthropoda are not sister taxa. Mol. Biol. Evol. 17, 87–106.

Brinkmann, H., Giezen, M., Zhou, Y., Raucourt, G., Philippe, H., 2005. An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. Syst. Biol. 54, 743–757.

Cameron, S., Barker, S., Whiting, M., 2006. Mitochondrial genomics and the new insect order Mantophasmatodea. Mol. Phylogenet. Evol. 38, 274–279.

Cook, C., Yue, Q., Akam, M., 2005. Mitochondrial genomes suggest that hexapods and crustaceans are mutually paraphyletic. Proc. Biol. Sci. 272, 1295–1304.

Crease, T.J., 1999. The complete sequence of the mitochondrial genome of *Daphnia pulex* (Cladocera: Crustacea). Gene 233, 89–99.

Dávila, S., Piñero, D., Bustos, P., Cevallos, M., Dávila, G., 2005. The mitochondrial genome sequence of the scorpion *Centruroides limpidus* (Karsch 1879) (Chelicerata; Arachnida). Gene 360, 92–102.

de Queiroz, A., Gatesy, J., 2006. The supermatrix approach to systematics. Trends Ecol. Evol. 22, 34–41.

Dobson, S., Barker, S., 1999. Phylogeny of the hard ticks (Ixodidae) inferred from 18S rRNA indicates that the genus *Aponomma* is paraphyletic. Mol. Phylogenet. Evol. 11, 288–295.

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

Dunlop, J.A., Arango, C.P., 2005. Pycnogonid affinities: a review. J. Zoo. Syst. Evol. Res. 43, 8–21.

Dunlop, J., Braddy, S., 2001. Scorpions and their sister-group relationships. In: Fet, V., Selden, P.A. (Eds.), Scorpions 2001 *In Memoriam*

Gary A Polis. The British Arachnological Society, Burnham Beeches, UK, pp. 1–24.

Dunlop, J., Webster, M., 1999. Fossil evidence, terrestrialization and arachnid phylogeny. J. Arachnol. 27, 86–93.

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

Gadagkar, S., Kumar, S., 2005. Maximum likelihood outperforms maximum parsimony even when evolutionary rates are heterotachous. Mol. Biol. Evol. 22, 2139–2141.

Galtier, N., Gouy, M., 1995. Inferring phylogenies from DNA sequences of unequal base compositions. Proc. Natl. Acad. Sci. USA 92, 11317–11321.

Gantenbein, B., Largiadèr, C., 2003. The phylogeographic importance of the Strait of Gibraltar as a gene flow barrier in terrestrial arthropods: a case study with the scorpion *Buthus occitanus* as model organism. Mol. Phylogenet. Evol. 28, 119–130.

Giribet, G., 2002. Current advances in the phylogenetic reconstruction of metazoan evolution. A new paradigm for the Cambrian explosion? Mol. Phylogenet. Evol. 24, 345–357.

Giribet, G., Edgecombe, G., Wheeler, W., Babbitt, C., 2002. Phylogeny and systematic position of Opiliones: a combined analysis of chelicerate relationships using morphological and molecular data. Cladistics 18, 5–70.

Giribet, G., Richter, S., Edgecombe, G.D., Wheeler, W.C., 2005. The position of crustaceans within Arthropoda—evidence from nine molecular loci and morphology. Crust. Issues 16, 307–352.

Gordon, D., Abajian, C., Green, P., 1998. Consed: a graphical tool for sequence finishing. Genome Res. 8, 195–202.

Hassanin, A., 2006. Phylogeny of Arthropoda inferred from mitochondrial sequences: strategies for limiting the misleading effects of multiple changes in pattern and rates of substitution. Mol. Phylogenet. Evol. 38, 100–116.

Hassanin, A., Leger, N., Deutsch, J., 2005. Evidence for multiple reversals of asymmetric mutational constraints during the evolution of the mitochondrial genome of metazoa, and consequences for phylogenetic inferences. Syst. Biol. 54, 277–298.

Hofacker, I., 2003. Vienna RNA secondary structure server. Nucleic Acids Res. 31, 3429–3431.

Jones, M., Blaxter, M., 2006. TaxMan: a Taxonomic database Manager. BMC Bioinformatics 7, 536.

Kass, R., Raftery, A., 1995. Bayes Factors. J. Am. Stat. Assoc. 90, 773–795.

Kenyon, F., Welsh, M., Parkinson, J., Whitton, C., Blaxter, M., Knox, D., 2003. Expressed sequence tag survey of gene expression in the scab mite *Psoroptes ovis*—allergens, proteases and free-radical scavengers. Parasitology 126, 451–460.

Kluge, A.G., 1989. A concern for evidence and a phylogenetic hypothesis of relationships among *Epicrates* (Boidae, Serpentes). Syst. Zool. 38, 7–25.

Lavrov, D., Boore, J., Brown, W., 2000. The complete mitochondrial DNA sequence of the horseshoe crab *Limulus polyphemus*. Mol. Biol. Evol. 17, 813–824.

Lavrov, D., Boore, J., Brown, W., 2002. Complete mtDNA sequences of two millipedes suggest a new model for mitochondrial gene rearrangements: duplication and nonrandom loss. Mol. Biol. Evol. 19, 163–169.

Lee, C., Grasso, C., Sharlow, M.F., 2002. Multiple sequence alignment using partial order graphs. Bioinformatics 18, 452–464.

Lemmon, A.R., Moriarty, E.C., 2004. The importance of proper model assumption in bayesian phylogenetics. Syst. Biol. 53, 265–277.

Lowe, T., Eddy, S., 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 25, 955–964.

Mallatt, J., Giribet, G., 2006. Further use of nearly complete 28S and 18S rRNA genes to classify Ecdysozoa: 37 more arthropods and a kinorhynch. Mol. Phylogenet. Evol. 40, 772–794.

Mallatt, J.M., Garey, J.R., Shultz, J.W., 2004. Ecdysozoan phylogeny and Bayesian inference: first use of nearly complete 28S and 18S rRNA gene sequences to classify the arthropods and their kin. Mol. Phylogenet. Evol. 31, 178.

Masta, S., Boore, J., 2004. The complete mitochondrial genome sequence of the spider *Habronattus oregonensis* reveals rearranged and extremely truncated tRNAs. Mol. Biol. Evol. 21, 893–902.

Maxmen, A., Browne, W., Martindale, M., Giribet, G., 2005. Neuroanatomy of sea spiders implies an appendicular origin of the protocerebral segment. Nature 437, 1144–1148.

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

Navajas, M., Le Conte, Y., Solignac, M., Cros-Arteil, S., Cornuet, J., 2002. The complete sequence of the mitochondrial genome of the honeybee ectoparasite mite *Varroa destructor* (Acari: Mesostigmata). Mol. Biol. Evol. 19, 2313–2317.

Negrisolo, E., Minelli, A., Valle, G., 2004. Extensive gene order rearrangement in the mitochondrial genome of the centipede *Scutigera coleoptrata*. J. Mol. Evol. 58, 413–423.

Nixon, K.C., Carpenter, J.M., 1996. On simultaneous analysis. Cladistics 12, 221–241.

Nylander, J.A.A., Ronquist, F., Huelsenbeck, J.P., Nieves-Aldrey, J.L., 2004. Bayesian phylogenetic analysis of combined data. Syst. Biol. 53, 47–67.

Ohtsuki, T., Kawai, G., Watanabe, K., 1998. Stable isotope-edited NMR analysis of *Ascaris suum* mitochondrial tRNAMet having a TV-replacement loop. J. Biochem. (Tokyo) 124, 28–34.

Olson, S.A., 2002. EMBOSS opens up sequence analysis. European Molecular Biology Open Software Suite. Brief Bioinform. 3, 87–91.

Philippe, H., Lartillot, N., Brinkmann, H., 2005a. Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. Mol. Biol. Evol. 22, 1246–1253.

Philippe, H., Zhou, Y., Brinkmann, H., Rodrigue, N., Delsuc, F., 2005b. Heterotachy and long-branch attraction in phylogenetics. BMC Evol. Biol. 5, 50.

Posada, D., Buckley, T.R., 2004. Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. Syst. Biol. 53, 793–808.

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

Pupko, T., Huchon, D., Cao, Y., Okada, N., Hasegawa, M., 2002. Combining multiple data sets in a likelihood analysis: which models are the best? Mol. Biol. Evol. 19, 2294–2307.

Qiu, Y., Song, D., Zhou, K., Sun, H., 2005. The mitochondrial sequences of *Heptathela hangzhouensis* and *Ornithoctonus huwena* reveal unique gene arrangements and atypical tRNAs. J. Mol. Evol. 60, 57–71.

Regier, J.C., Shultz, J.W., 2001. Elongation factor-2: a useful gene for arthropod phylogenetics. Mol. Phylogenet. Evol. 20, 136–148.

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

Rokas, A., Williams, B.L., King, N., Carroll, S.B., 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425, 798–804.

Ronquist, F., Huelsenbeck, J., 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19, 1572–1574.

Shao, R., Mitani, H., Barker, S.C., Takahashi, M., Fukunaga, M., 2005. Novel mitochondrial gene content and gene arrangement indicate illegitimate inter-mtDNA recombination in the chigger mite, *Leptotrombidium pallidum*. J. Mol. Evol. 60, 764–773.

Stajich, J.E., Block, D., Boulez, K., Brenner, S.E., Chervitz, S.A., Dagdigian, C., Fuellen, G., Gilbert, J.G.R., Korf, I., Lapp, H., Lehvaslaiho, H., Matsalla, C., Mungall, C.J., Osborne, B.I., Pocock, M.R., Schattner, P., Senger, M., Stein, L.D., Stupka, E., Wilkinson, M.D., Birney, E., 2002. The Bioperl toolkit: Perl modules for the life sciences. Genome Res. 12, 1611–1618.

Sullivan, J., Swofford, D.L., 2001. Should we use model-based methods for phylogenetic inference when we know that assumptions about among-site rate variation and nucleotide substitution pattern are violated? Syst. Biol. 50, 723–729.

Please cite this article in press as: Jones, M. et al., The effect of model choice on phylogenetic inference using mitochondrial sequence data: Lessons from the scorpions, Mol. Phylogenet. Evol. (2007), doi:10.1016/j.ympev.2006.11.017

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

Umetsu, K., Iwabuchi, N., Yuasa, I., Saitou, N., Clark, P.F., Boxshall, G., Osawa, M., Igarashi, K., 2002. Complete mitochondrial DNA sequence of a tadpole shrimp (*Triops cancriformis*) and analysis of museum samples. Electrophoresis 23, 4080–4084.

Watanabe, Y., Tsurui, H., Ueda, T., Furushima, R., Takamiya, S., Kita, K., Nishikawa, K., Watanabe, K., 1994. Primary and higher order structures of nematode (*Ascaris suum*) mitochondrial tRNAs lacking either the T or D stem. J. Biol. Chem. 269, 22902–22906.

Weygoldt, P., 1998. REVIEW Evolution and systematics of the Chelicerata. Exp. Appl. Acarol. 22, 63–79.

Weygoldt, P., Paulus, H., 1979. Untersuchungen zur Morphologie, Taxonomie und Phylogenie der Chelicerata. Zeitschrift für Zoologische Systematik und Evolutionsforschung. Z. Zool. Syst. Evol. 17, 177–200.

Wheeler, W.C., Hayashi, C.Y., 1998. The phylogeny of the extant chelicerate orders. Cladistics 14, 173–192.

Wiens, J.J., 2003. Missing data, incomplete taxa, and phylogenetic accuracy. Syst. Biol. 52, 528–538.

Wolstenholme, D., Macfarlane, J., Okimoto, R., Clary, D., Wahleithner, J., 1987. Bizarre tRNAs inferred from DNA sequences of mitochondrial genomes of nematode worms. Proc. Natl. Acad. Sci. USA 84, 1324–1328.

Wuyts, J., Perrière, G., Peer, Y., 2004. The European ribosomal RNA database. Nucleic Acids Res. 32, D101–D103.

Please cite this article in press as: Jones, M. et al., The effect of model choice on phylogenetic inference using mitochondrial sequence data: Lessons from the scorpions, Mol. Phylogenet. Evol. (2007), doi:10.1016/j.ympev.2006.11.017