Dating the arthropod tree based on large-scale transcriptome data

Peter Rehma, Janus Borner, Karen Meusemann, Björn M. von Reumont, Sabrina Simon, Heike Hadrys, Bernhard Misof, Thorsten Burmester

1. Introduction

Dating of diversification and speciation events is a major aim of evolutionary studies. For a long time, fossil remains were the prime source of such time estimates. The fossil record, however, is far from complete and in many cases the taxonomic assignment of fossil specimens is uncertain (Benton and Donoghue, 2007). DNA and protein sequences provide a complementary source of information for the inference of life history. Although there is an ongoing debate whether such a molecular clock approach is actually valid (Graur and Martin, 2004), many studies have obtained reasonable time estimates for a broad range of taxa (for review, see: Benton and Ayala, 2003; Hedges and Kumar, 2003).

In theory, molecular clock calculations have the power to be more precise than fossil dates because latter usually are underestimates. At best fossils provide an approximation to the oldest member of the taxon in question (cf. Benton and Ayala, 2003). In fact, sequence-derived dates tend to be older than the fossil dates (Hedges and Kumar, 2003). This is particularly true for deep divergence times. For example, the first conclusive fossil evidence for crown group bilaterians dates ~550–530 mya (Benton and Donoghue, 2007), but molecular estimates suggest an emergence of bilaterians between 1300 and 670 mya (e.g., Blair and Hedges, 2005; Lynch, 1999; Otsuka and Sugaya, 2003; Peterson et al., 2008). The discrepancy of molecular and fossil dates, and among different molecular clock approaches can be attributed to insufficient data, wrong taxonomic assignment or dating of fossils, and, most importantly, to rate heterogeneity among lineages over time and between genes (e.g., Adachi and Hasegawa, 1995; Benton and Donoghue, 2007; Bromham et al., 1998; Graur and Martin, 2004).

The undisputed fossil record of the phylum Arthropoda dates back to the early Cambrian period (Budd and Jensen, 2000; Budd and Telford, 2009; Edgecombe, 2010). The identity of possible representatives of arthropods from the earlier Ediacaran period is questionable (Nielsen, 2001). Based on fossils and geological
considerations, Benton and Donoghue (2007) assumed an earliest
date of 581 mya for the divergence of Arthropoda and Nematoda.
Molecular clock analyses, however, usually support much older
time estimates that range from 1200 to 625 mya for the origin of
Arthropoda (Blair, 2009; Blair et al., 2005; Douzery et al., 2004;
Hausdorf, 2000; Lee, 1999; Sanders and Lee, 2010; Wang et al.,
1999). Due to the lack of sequence data from important taxa, cal-
culations of internal divergence times within arthropods are
sparse, with the exception of the insects (e.g., Gaunt and Miles,
2002; Regier et al., 2004, 2005).
Here we analyze the divergence times of major arthropod taxa
based on a superalignment spanning 37,476 amino acid positions,
which had been derived from Expressed Sequence Tags (ESTs)
(Meusemann et al., 2010). This is – to the best of our knowledge
– the largest dataset that has ever been used for molecular clock
studies.

2. Materials and methods

2.1. Sequence data and phylogenetic tree

In a previous study (Meusemann et al., 2010), 775 orthologous
genes from 214 euarthropods, three onychophorans, two tardi-
grades, eight nematodes, three annelids and three mollusks, de-
erived from EST data and selected genomes were identified with the
HaMStR approach (Ebersberger et al., 2009). Single multiple
protein alignments were generated with MAFFT L-INSI (Katoh
and Toh, 2008). Alignment masking was conducted with ALISCORE
and ALICUT (Kück et al., 2010; Misof and Misof, 2009). An optimal
subset of data was selected by MARE 01-alpha (MAtrix REDuction;
http://mare.zfmk.de). The finally selected superalignment spans
37,476 amino acid positions, comprised 129 genes and 117 taxa,
including 101 arthropods (available at TreeBase, http://www.tree-
base.org, under study accession no. S10507). A Bayesian phyloge-
etic tree was inferred with PhyloBayes (Lartillot et al., 2009).
For details, refer to Meusemann et al. (2010).

2.2. Bayesian estimates of divergence times

The Bayesian phylogenetic majority rule consensus tree
(Meusemann et al., 2010) was used as input for molecular clock
estimates. The program PhyloBayes 3.2d was applied to calculate
divergence times and 95% confidence intervals within a Bayesian
framework (Lartillot et al., 2009). Three relaxed clock models, the
log-normal autocorrelated clock model (LOG) (Thorne et al.,
1998), the ‘CIR’ process (CIR) (Cox et al., 1985; see also Lepage
et al., 2006) and uncorrelated gamma multipliers (UGM) (Drum-
mond et al., 2006), were used under a uniform prior on divergence
times for 50,000 cycles with a burn-in of 20,000 cycles. The CIR
process is similar to the Ornstein–Uhlenbeck model, but with
superior properties (Aris-Brosou and Yang, 2003; Lepage et al.,
2006, 2007). Rates across sites were modeled assuming a discrete
gamma distribution with four categories and with a Dirichlet pro-
cess. Bayes factors were estimated using thermodynamic integra-
tion as implemented in PhyloBayes (Lepage et al., 2007) with
100,000 generations and a burn-in of 10,000. The three relaxed
clock models were compared with the unconstrained model.
Cross-validation of the models was performed by dividing the
alignment into eight subsets (seven learning sets and one test
set). Ten repetitions were run, as specified in PhyloBayes.

In a first approach, we tested the effects of gamma distributed
priors for the root node on our results. To evaluate the impact of
the priors, we defined different means and standard deviations
(SD) of the prior distribution: mean 1000 mya (SD 1000 myr/
500 myr) and 750 mya (SD 750 myr/325 myr), respectively. In
addition, a uniform root prior was assumed with a maximum limit
of 5000 mya imposed by PhyloBayes. All analyses were also run
under the priors (i.e. with no data) to assess whether the prior dis-
tribution was sufficiently wide (i.e. non-informative). The results
were compared with those obtained when the data were analyzed.

To assess the impact of missing data, all amino acid positions in
the concatenated alignment were sorted according to their taxon
coverage. Only the 50% of positions with the highest taxon cover-
age were used in a separate molecular clock analysis with the same
settings as described above. The effect of substitution rates was
tested by dividing the complete superalignment (129 genes) into
three subsets, each containing 43 genes with i lowest, ii intermedi-
ate, and iii highest substitution rates. Genes were assigned to these
categories according to the mean PAM distance of all possible se-
quence pairs within each alignment. To avoid artifacts due to missing
data, only taxa for which sequences of all genes are present were
selected for the assessment of pairwise distances: Apis mellifera,
Bombus mori, Daphnia pulex, Drosophila melanogaster, and
Trichoplax adhaerens. Positions with gaps were ignored. All three
subsets were analyzed in separate runs according to the procedure
described above.

2.3. Calibration of the molecular clock

Seven calibration points were evenly distributed throughout
the phylogenetic tree, including one calibration point within the
outgroup (Table 1; Supplemental Table S1). We aimed to cover dif-
ferent regions of the tree and to include calibrations for deep nodes
as well as for shallow nodes. To avoid a distortion of the time esti-
mates by systematic misplacement of fossil calibration points, we
used fossils with reliable systematic placement. Numerical ages
were obtained from the International Stratigraphic Chart 2009
(http://www.stratigraphy.org), assuming the minimum age of the
respective stage interval for calibration points 1–4 and 7 (Supple-
mental Table S1). The minimum age of calibration point 5 was
dated according to the minimal age of the Namurian A/E1 (Du-
sar, 2006) and calibration point 6, for which a minimum and a
maximum age was obtained from Benton and Donoghue (2007).
Each of the settings described above was run with seven calibra-
tion points for each dataset. In addition, the complete dataset
was analyzed with six calibration points (omitting calibration point 1
within the outgroup).

Calibration point 1: the minimum age of the divergence of Mol-
lusca and Annelida is defined by small helcionelloids of the genus
Oelandiella from the pre-Tommotian (Cambrian, Purella Biozone,
Nemakit-Daldybian) period 528 mya (Gubanov and Peel, 1999;
Khomentovsky and Karlova, 1993). Calibration point 2: evidence
for euarthropods is provided by Rusophycus-like trace fossils from
the early Tommotian 521 mya (Crimes, 1987). This calibration
point is confirmed by recently described Lower Cambrian euar-
thropods fossils (Chen, 2009), which derives from the Maotianshan
Shale (Qiongzhusian) dating 521–515 mya. Calibration point 3:
the oldest unambiguous myriapod fossil is the millipede Codiedes-
mus ericotopusus (Wilson and Anderson, 2004) from the Cowie For-
mation, Silurian. At that time, millipedes and centipedes had
separated, thus providing data for Cretaceous–Tertiary (Cenozoic)
nickels (i.e. C. ericotopusus) provides a minimal age for the diver-
gence of Diplopoda (millipedes) and Chilopoda (centipedes) at the
transition from Wenlock to Ludlow 418.7 mya (base of Ludfordian,
Ludlow). Calibration point 4: the split between Entognatha and
Ectognatha (true insects) dates to the early Devonian (Pragian) per-
iod 404.2 mya, delimited by the first entognath fossil of the sprintail
Rhyiella precursura (Whalley and Jarzembowski, 1981).
Calibration point 5: the minimum date for the split between pale-
smanida and palaeopteran lineages is provided by an insect wing
from the Upper Sleisian Basin, Czech Republic (Béthoux and Nel,
2005), which dates to the Lower Carboniferous 324.8 mya and has
been
assigned to Archaeorthoptera based on its venation (Prokop et al., 2005). Calibration point 6: the oldest fossil of a clade including Culicomorpha and Brachycera (Diptera) is assigned to Archaeorthoptera based on its venation (Prokop et al., 2005). Calibration point 6: the oldest fossil of a clade including Culicomorpha and Brachycera (Diptera) is assigned to Archaeorthoptera based on its venation (Prokop et al., 2005). Calibration point 6: the oldest fossil of a clade including Culicomorpha and Brachycera (Diptera) is assigned to Archaeorthoptera based on its venation (Prokop et al., 2005). Calibration point 6: the oldest fossil of a clade including Culicomorpha and Brachycera (Diptera) is assigned to Archaeorthoptera based on its venation (Prokop et al., 2005). Calibration point 6: the oldest fossil of a clade including Culicomorpha and Brachycera (Diptera) is assigned to Archaeorthoptera based on its venation (Prokop et al., 2005). Calibration point 6: the oldest fossil of a clade including Culicomorpha and Brachycera (Diptera) is assigned to Archaeorthoptera based on its venation (Prokop et al., 2005). Calibration point 6: the oldest fossil of a clade including Culicomorpha and Brachycera (Diptera) is assigned to Archaeorthoptera based on its venation (Prokop et al., 2005). Calibration point 6: the oldest fossil of a clade including Culicomorpha and Brachycera (Diptera) is assigned to Archaeorthoptera based on its venation (Prokop et al., 2005). Calibration point 6: the oldest fossil of a clade including Culicomorpha and Brachycera (Diptera) is assigned to Archaeorthoptera based on its venation (Prokop et al., 2005). Calibration point 6: the oldest fossil of a clade including Culicomorpha and Brachycera (Diptera) is assigned to Archaeorthoptera based on its venation (Prokop et al., 2005). Calibration point 6: the oldest fossil of a clade including Culicomorpha and Brachycera (Diptera) is assigned to Archaeorthoptera based on its venation (Prokop et al., 2005). Calibration point 6: the oldest fossil of a clade including Culicomorpha and Brachycera (Diptera) is assigned to Archaeorthoptera based on its venation (Prokop et al., 2005).
Bayesian time estimates were robust to changes in root priors, which determine age distributions and corresponding standard deviations for the root node. In a first approach, a uniformly distributed prior with an upper limit of the root of 5000 mya was assumed, as imposed by PhyloBayes (lane “–” in Supplemental Tables S2 and S3). To evaluate the impact of specific root priors, we defined different age means and standard deviations of the prior distribution: (a) 1000 mya and 1000 myr standard deviation, (b) 1000/500, (c) 750/750, and (d) 750/325. For all nodes, time estimates show low variation and differed – on average – less than 2% from the mean age with different root priors (standard deviation averaged over all nodes).

Standard deviations of the mean time estimates were lower with modeling the Dirichlet process (1.1%; root 1.8%) than with gamma distributed rates across sites (2.1%; root 2.2%). The exclusion of the outgroup calibration point, i.e. the minimum age of 521 mya for the split between Mollusca and Polychaeta, resulted in slightly (4.1%) younger divergence time estimates (not shown). The mean age of the root, i.e. the origin of Ecdysozoa, was about 5.7% younger using the reduced set of calibration points. Averaged over all calculations, the ecdysozoan divergence dated to 573 mya.

### Effect of substitution rates on time estimates

To assess the effect of variations of substitution rates among genes, the full 129 gene dataset was subdivided into three data subsets with 43 genes each. The mean PAM distances of the genes in subset i ranged from 0.01 to 0.17 (slowest substitution rate), in subset ii from 0.17 to 0.30 (intermediate rate) and in subset iii from 0.31 to 0.80 (fastest rate). Averaged over data subsets of the

### Table 2

<table>
<thead>
<tr>
<th></th>
<th>LOG Γ</th>
<th>LOG D</th>
<th>CIR Γ</th>
<th>CIR D</th>
<th>UGM Γ</th>
<th>UGM D</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOG Γ</td>
<td>0.004984</td>
<td>0.027751</td>
<td>0.028321</td>
<td>0.109461</td>
<td>0.109718</td>
<td></td>
</tr>
<tr>
<td>LOG D</td>
<td>0.007357</td>
<td>0.010987</td>
<td>0.003706</td>
<td>0.112200</td>
<td>0.112359</td>
<td></td>
</tr>
<tr>
<td>CIR Γ</td>
<td>−0.009793</td>
<td>−0.002589</td>
<td>0.0089166</td>
<td>0.089328</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CIR D</td>
<td>−0.009723</td>
<td>−0.004126</td>
<td>−0.08827</td>
<td>0.001537</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UGM Γ</td>
<td>−0.009071</td>
<td>−0.094367</td>
<td>−0.085352</td>
<td>−0.082959</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UGM D</td>
<td>−0.090971</td>
<td>−0.094367</td>
<td>−0.085352</td>
<td>−0.082959</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fig. 1. Mean divergence times of major ecdysozoan taxa averaged over all estimations under the log–normal autocorrelated clock model shown in Supplemental Table S2. Gray bars indicate 95% mean confidence intervals (see Supplemental Table S3). Calibrated nodes are marked with an asterisk (see Supplemental Table S1 for calibration points). mya, million years ago.
autocorrelated LOG model, molecular clock analyses resulted in dates that were only $\sim0.1\%$ younger compared to the complete dataset (Supplemental Table S6). In subset i the estimated dates were on average 2.9% younger (standard deviation 8.7%) as compared to the complete dataset. Subsets ii and iii showed on average slightly older dates (1.8% and 1.5%, respectively; standard deviation 6.5% and 6.9%, respectively). Most divergence times in the data subsets agree very well with the values derived from the complete dataset. Only two splits showed notable deviations. The split between Odonata and Ephemeroptera displayed high variation between subsets under all molecular clock settings, with subset i resulting in 22% younger dates and subsets ii and iii having 4% older dates. The results of the intermediate subset ii were similar with that of the entire dataset. The variations of divergence times of Isoptera and ‘Blattodea’ were even higher, with the intermediate subset ii resulting in 29% younger dates, whereas the slowly and fast evolving subsets showed older dates (subset i: 32%; subset iii: 35%).

3.5. Effect of missing data on time estimates

By removing 50% of the amino acid positions with the lowest taxon coverage, the relative amount of missing data in the alignment decreased from 51.0% to 38.5%. The effect of data reduction on the resulting divergence time estimates was minimal (Supplemental Table S7). Mean divergence times were 1.5% older than the results obtained from the complete dataset. Similar to the results from the subdivided datasets, the splits between Odonata–Ephemeroptera and Isoptera–‘Blattodea’ showed high variance. Under the autocorrelated clock models, divergence times of Odonata and Ephemeroptera were 7.7% older, whereas divergence time estimates of Isoptera and ‘Blattodea’ were 5.5% younger compared to the complete dataset.

4. Discussion

Molecular clock analysis has become a powerful tool based on a data source largely independent from the fossil record for the inference of divergence times of organisms. Still, there is much discrepancy between time estimates of different studies (e.g., Douzery et al., 2004; Peterson et al., 2008; Pisani et al., 2004). Factors that influence the outcome of molecular clock calculations are the sampling size, the selection of taxa and genes, rate heterogeneity, the suitability of the dating method, and the accuracy of calibration points. Simultaneous analyses of a large number of orthologous genes and application of multiple fossil calibration points provide more reliable estimates of divergence times if rate heterogeneity is considered (Thorne and Kishino, 2002; Yang and Yoder, 2003).

Due to the limited availability of orthologous genes, studies on large multi-genus datasets were usually restricted to only few taxa (e.g., Aris-Brosou and Yang, 2003; Blair and Hedges, 2005; Blair et al., 2005; Douzery et al., 2004; Gaunt and Miles, 2002; Gu, 1998; Lynch, 1999; Peterson et al., 2008; Regier et al., 2005; Wang et al., 1999). An alternative source of sequence data, which had not been applied to a molecular clock approach so far, is provided by EST data.

4.1. Applicability of EST data for molecular clock analyses

Our supermatrix with 129 orthologous genes and 117 taxa (Meusemann et al., 2010) is – to the best of our knowledge – the largest dataset that has ever been used for molecular clock studies (cf. Aris-Brosou and Yang, 2003; Blair and Hedges, 2005; Blair et al., 2005; Douzery et al., 2004; Gaunt and Miles, 2002; Gu, 1998; Lynch, 1999; Peterson et al., 2008; Regier et al., 2005; Wang et al., 1999). Our analyses are expected to provide more reliable estimates than the inference from few genes due to rate homogenization (Battistuzzi et al., 2010; Thorne and Kishino, 2002; Yang and Yoder, 2003). In addition, the essentially stochastic nature of ESTs further is expected to reduce sampling bias caused by the selection of specific genes.

A potential drawback of our approach may be the fragmentary nature of ESTs, which is reflected by 51% missing data in the concatenated superalignment (Meusemann et al., 2010). It has been demonstrated that large datasets are less sensitive to missing data (Philippe et al., 2004). In fact, when we reduced the amount of missing data by removing positions with low coverage, thereby increasing data density, only minimal changes in divergence time calculations were observed (Supplemental Table S7). Only two splits showed notable variation (Odonata–Ephemeroptera and Isoptera–‘Blattodea’). Therefore, molecular clock analyses of our EST-dataset were essentially robust to the effect of missing data.

Another factor that may influence the outcome of a molecular clock approach are differences in substitution rates between genes. The a priori stochastic approach of obtaining ESTs is expected to result in large variations of rates. This is reflected by the PAM distances of the individual proteins ranging from 0.01 to 0.80. However, splitting the dataset into three subsets with different evolutionary rates shows little variation in divergence times between the three estimates (Supplemental Table S6). Only for the two splits mentioned above (i.e. Odonata–Ephemeroptera and Isoptera–‘Blattodea’), the variance was notably large, which may be due to a bias introduced by gene selection. Therefore, we can conclude that – at least if the datasets are large enough – the effect of substitution rate differences in the EST dataset is low.

4.2. Molecular clock models

The Bayesian relaxed clock model approach to the arthropod EST-derived tree was also robust to the choice of priors and parameter settings. Neither the root priors, which specify the age of the root and its standard deviation, nor the model for the site-specific rates (discrete gamma distribution with four rate categories or the Dirichlet process) had significant effects (Table 2; Supplemental Tables S2, S4 and S5). The main factor that actually influenced the time estimates was the applied clock model.

In recent years, a whole range of different molecular clock methods that either rely on a maximum likelihood approach or on Bayesian methods have been proposed (for review, see Lepage et al., 2007). In initial tests, we applied a maximum likelihood, local clock approach, as implemented in the program r8s (Sanderson, 2002) to our data. However, the resulting time estimates were highly dependent on the age of the root and resulted in unreasonable divergence times (data not shown). The dates that derived from the Bayesian models were more consistent and not mutually exclusive, but still showed large differences in the calculated divergence times. We applied three different relaxed Bayesian clock models, two autocorrelated (LOG and CIR) and the uncorrelated UGM model. But which model is correct, i.e. fits best to the data and is thus expected to provide the best time estimate?

Both cross-validation and Bayes factors showed that the autocorrelated clock models were significantly better than UGM. Autocorrelation assumes that adjacent branches in a phylogenetic tree evolve with a similar rate, while in an uncorrelated model the individual rates cluster around the mean. Thus the assumption of autocorrelation of rates in related species appears to be more realistic than averaging rates across the branches. Our results also agree with the study by Lepage et al. (2007), who demonstrated that autocorrelated models outperform uncorrelated models, particularly when the dataset is large. Cross-validation also found that the LOG model was slightly better than the CIR process. Therefore, we discuss below only divergence times averaged over all LOG clock model settings (Table 1), as displayed in Fig. 1.
4.3. Arthropod origins and age of major arthropod taxa

While the fossil record suggests the emergence of the Metazoa during the Cambrian period 542–488 mya (e.g., Chen et al., 2004; Chen, 2009; Conway Morris, 1993; Crimes, 1987; Harvey and Butterfield, 2008; Shu et al., 1996), most molecular clock studies estimated much older dates of up to 1200 mya (e.g., Blair et al., 2005; Feng et al., 1997; Hausdorf, 2000; Lee, 1999; Nei et al., 2001; Peterson et al., 2008; Wang et al., 1999). We obtained notably younger estimates in our studies, which, however, still do not agree with a Cambrian origin of metazoan phyla. For example, we dated the earliest divergence time within the Arthropoda 589 mya, while the first unambiguous arthropod fossils are 521 myr old (Chen, 2009; Crimes, 1987). The gap between a possible Precambrian emergence and the Cambrian metazoan fossils may be explained by a period of cryptic evolution or detection bias, e.g., due to largely unexplored Early Cambrian and Pre-Cambrian Lagerstätten (Benton and Ayala, 2003; Conway Morris, 1993; Forsey et al., 1996; Valentine et al., 1991).

Traditionally, tardigrades have been joined with the arthropods (Brusca and Brusca, 2003). Recent molecular studies suggested a sister group relationship of tardigrades with Cycloneuralia (nematodes and allies) (Bledorn et al., 2009; Lartillot and Philippe, 2008; Meusseman et al., 2010; Roeding et al., 2007). However, this topology has been discussed as an artifact due to long branch attraction (Rota-Stabelli et al., 2011). Previous molecular clock studies estimated an early origin of tardigrades 813–670 mya (Regier et al., 2005; Sanders and Lee, 2010), while we calculated that the divergence of Nemata and Tardigrada took place during the Ediacaran (~575 mya). Although our estimate is in better agreement with the Cambrian fossils, our results suggest that the origin of tardigrades relationships should be considered with caution because of the uncertain tardigrade relationships.

The closest arthropod relative of the myriapods is uncertain. While some molecular studies either suggested a sister group relationship of Myriapoda and Pancrustacea ("Mandibulata"; e.g., Giribet and Ribera, 2000; Regier et al., 2010; Rota-Stabelli et al., 2011), others provided evidence for a common clade of Myriapoda and Chelicerata ("Myriochehata"; Pisani et al., 2004; Roeding et al., 2009). Because Meusseman et al. (2010) recovered Myriochehata in their Bayesian approach, this topology was assumed here although it may be an artifact (Rota-Stabelli et al., 2011). We inferred that Myriochehata and Pancrustacea diverged 562 mya. This is notably younger than previous calculations based on topologies supporting Myriochehata, which ranged from 672–642 mya (Pisani et al., 2004; Regier et al., 2005). The branch that joins Myriapoda and Chelicerata is comparatively short, corresponding to ~15 million years (myr) with a large confidence interval (Fig. 1; Supplementary Table S3). Thus, a rapid divergence of the three clades Myriapoda, Chelicerata and Pancrustacea may explain at least in part the problems associated with the relationships among these taxa.

There is no conclusive myriapod record from the Cambrian, but presence of fossils from putative sister group taxa (Crustacea, Chelicerata) strongly suggests a Cambrian or earlier origin of Myriapoda (Shear and Edgecombe, 2010). While previous studies date the emergence of Myriapoda more than 600 mya (Otsuka and Sugaya, 2003; Pisani et al., 2004; Regier et al., 2005), our estimates of myriapod origin are comparatively young (~556 mya). Within the Myriapoda, our results showed an age for the split of Diplopoda and Chilopoda of ~504 mya, which is slightly older than previous molecular studies (e.g., 442 mya; Pisani et al., 2004) and the fossil record (~420 mya; Edgecombe and Giribet, 2007).

Most studies agree that Pycnogonida (sea spiders) represent the earliest branch within Chelicerata (Meusseman et al., 2010; Regier et al., 2010; Roeding et al., 2009). A larval sea spider from the upper Cambrian (~500 mya) is the oldest fossil evidence for the split of Pycnogonida and Euchelicerata (Walszoeck and Dunlop, 2002). Regier et al. (2005) suggested that this event took place 813–632 mya, but our calculation (~546 mya) is closer to the fossil record. Our estimate for the origin of Xiphosura (horsehoe crabs; ~473 mya) agrees well with the fossil dating (~445 mya (Rudkin et al., 2008).

Traditionally, the divergence of Arachnida and Xiphosura has been considered the first split within the Euchelicerata (Regier et al., 2010; Weygoldt, 1998). However, several molecular studies did not recover monophyletic Arachnida, but suggest a basal position of the Aracni (Meusseman et al., 2010; Roeding et al., 2007, 2009; Sanders and Lee, 2010). Given the low taxon sampling within the arachnids, it must remain uncertain which topology may reflect a true relationship or a long branch attraction phenomenon. This unresolved topology is the most likely explanation for the younger divergence times of the Aracni (~424 mya) and Araneae in other studies (~401–390 mya) (Aris-Brosou and Yang, 2002; Jayaprakash and Hoy, 1999; Sanders and Lee, 2010).

The origin of the clade leading to Pancrustacea ("Crustacea" and Hexapoda) was previously estimated between 725 and 565 mya (Burmester, 2001; Otsuka and Sugaya, 2003; Pisani et al., 2004; Regier et al., 2005). Our results (~562 mya) are on the younger side. We estimated the divergence of the clade leading to the crustacean taxa Malacostraca, "Maxillopoda" (Copepoda and Cirripedia) and Branchiopoda, and to the subphylum Hexapoda at ~520 mya in the early Cambrian. This timing is in line with recent findings of a crown group crustacean from the Mount Cap Formation 515–510 mya (Harvey and Butterfield, 2008).

The oldest known hexapods are collemboles (springtails) from the Lower Devonian (~400 mya (Kukalová-Peck, 1991). Previous molecular analyses estimated the split between crustaceans and hexapods (either Malacostraca–Hexapoda or Branchiopoda–Hexapoda) from 492–420 mya, but these studies relied on single or a limited number of genes (Burmester, 2001; Gaunt and Miles, 2002; Otsuka and Sugaya, 2003; Regier et al., 2005; Sanders and Lee, 2010). Based on an alignment of multiple genes, Pisani et al. (2004) proposed that the divergence of Hexapoda and Crustacea took place ~666 mya. Although our estimate (divergence of Branchiopoda and Hexapoda 510 mya) is closer to the hexapod fossil record, there is still a gap of ~100 myr. It must be considered that the true crustacean sister group of the Hexapoda is ambiguous. Recent studies have suggested that the enigmatic Remipedia may represent the closest living crustacean relatives of Hexapoda (Ertas et al., 2009; Regier et al., 2010). Unfortunately, fossil Remipedia are rare and ambiguous, and ESTs are currently not available for dating of the divergence from hexapods.

Within the insects, there is a general discrepancy between molecular time estimates and fossils. For example, we dated the divergence of Archeognatha and Pterygota ~455 mya, while first archeognath fossils derive from ~390 myr old (Labandeira et al., 1988) and first pterygotes from ~325 myr old strata (Prokop et al., 2005). Likewise, the time of the origin of Holometabola was estimated ~391 mya, while the first unambiguous fossils are ~307 mya (Béthoux, 2009). Our time estimates are actually close to previous calculations by (Gaunt and Miles, 2002), but the relatively large gap between molecular and fossil dating requires further investigations.

5. Conclusions

Although any molecular clock calculation for the inference of divergence times embraces problems beyond experimental control, we undertook measures to reduce potential errors to a minimum. The large amount of orthologous sequences from many
arthropod species, and the application of a relaxed Bayesian clock model using evenly distributed calibration points yielded consistent molecular divergence time estimates. Missing data had only minor effect on the estimation of divergence times highlighting the suitability of ESTs for molecular clock analyses. Likewise, selection of three data subsets (from fast, intermediate, or slow evolving genes) and different model priors had only negligible influence. The application of different models (uncorrelated vs. autocorrelated models) had notable effects on divergence time calculations. Along with errors in calibration points, inappropriate data and similar problems, unsuitable models may explain in part the unreasonably early divergence times obtained in some previous molecular clock studies. Our approach resulted in divergence time estimates of the arthropods that are generally in much better agreement with the fossil record.

Acknowledgments

We are grateful to Nicolas Laritolo for his help with PhyloBayes and J.-Wolfgang Wägele for his support. We are grateful to Falko Roeding for his help with the data and initial analyses, and Manfred Menning, Alexander Gubanov, Jacques Venniers and Chales H. Wellman for their support on stratigraphical issues. We thank Jakub Prokop for his advice on insect evolution. We acknowledge the associate editor and two anonymous reviewers for their helpful comments which improved the manuscript. This work has been supported by grants of the Deutsche Forschungsgemeinschaft within the Priority Project 1174 “Deep Metazoan Phylogeny” (Bu 956/8; Mi 649/6; Ha 1947/5; Wa 530/34).

Appendix A. Supplementary material


References


