Compositional Heterogeneity and Phylogenomic Inference of Metazoan Relationships

Maximilian P. Nesnidal,[1] Martin Helmkampf,†,[1] Iris Bruchhaus,[2] and Bernhard Hausdorf*,[1]

[1]Zoological Museum, University of Hamburg, Hamburg, Germany

[2]Bernhard Nocht Institute for Tropical Medicine, Hamburg, Germany

†Present address: School of Life Sciences, Arizona State University.

***Corresponding author:** E-mail: hausdorf@zoologie.uni-hamburg.de.

Associate editor: Andrew Roger

Research article

Abstract

Compositional heterogeneity of sequences between taxa may cause systematic error in phylogenetic inference. The potential influence of such bias might be mitigated by strategies to reduce compositional heterogeneity in the data set or by phylogeny reconstruction methods that account for compositional heterogeneity. We adopted several of these strategies to analyze a large ribosomal protein data set representing all major metazoan taxa. Posterior predictive tests revealed that there is compositional bias in this data set. Only a few taxa with strongly deviating amino acid composition had to be excluded to reduce this bias. Thus, this is a good solution, if these taxa are not central to the phylogenetic question at hand. Deleting individual proteins from the data matrix may be an appropriate method, if compositional heterogeneity among taxa is concentrated in a few proteins. However, half of the ribosomal proteins had to be excluded to reduce the compositional heterogeneity to a degree that the CAT model was no longer significantly violated. Recoding of amino acids into groups is another alternative but causes a loss of information and may result in badly resolved trees as demonstrated by the present data set. Bayesian inference with the CAT–BP model directly accounts for compositional heterogeneity between lineages by introducing breakpoints along the branches of the phylogeny at which the amino acid composition is allowed to change but is computationally expensive. Finally, a neighbor joining tree based on equal input distances that consider pattern and rate heterogeneity showed several unusual groupings, which are most likely artifacts, probably caused by the loss of information resulting from the transformation of the sequence data into distances. As long as no more efficient phylogenetic inference methods are available that can directly account for compositional heterogeneity in large data sets, using methods for reducing compositional heterogeneity in the data in combination with methods that assume a stationary amino acid composition remains an option for controlling systematic errors in tree reconstruction that result from compositional bias. Our analyses indicated that the paraphyly of Deuterostomia in some analyses is the result of systematic errors that also affected the relationships of Entoprocta and Ectoprocta.

Key words: Metazoa, phylogenomics, compositional bias.

Introduction

The advent of phylogenomics based on large expressed sequence tag (EST) or genome projects resulted in alignments that are a magnitude larger than previously available sequence data sets and promised the resolution of the relationships of metazoan phyla (Philippe et al. 2005; Philippe and Telford 2006; Baurain et al. 2007). Taxon sampling has been improved so that genomic scale data are now available from most metazoan phyla (Hausdorf et al. 2007; Brinkmann and Philippe 2008; Dunn et al. 2008; Helmkampf et al. 2008; Struck and Fisse 2008; Witek et al. 2008; Hejnol et al. 2009). The strongly increasing amount of available data reduces the influence of random errors on phylogenetic inference. Nevertheless, many internal nodes are still poorly supported, and different analyses produce contradictory topologies. One reason for such incongruent outcomes may be systematic errors resulting from violations of the assumptions of the models used for tree reconstruction (Delsuc et al. 2005).

Most models of protein evolution assume that the amino acid composition is stationary. Violations of this assumption may result in incorrect topological estimation. Compositional heterogeneity (Lockhart et al. 1994; Foster and Hickey 1999; Foster 2004; Jermiin et al. 2004; Phillips et al. 2004; Collins et al. 2005) or a combination of compositional heterogeneity and rate heterogeneity among lineages (Ho and Jermiin 2004) are common problems in this respect. Lartillot and Philippe (2008) noted that the assumption of compositional homogeneity made by conventional protein models is strongly violated in the metazoan phylogenomic data set they examined.

We investigated different strategies to reduce the potential influence of compositional heterogeneity on phylogenomic analyses of metazoan relationships. We evaluated three approaches for reducing compositional heterogeneity in a ribosomal protein data set concerning their effectiveness in reducing violations of the model of protein evolution and their influence on the phylogenetic information content of the alignments. These approaches were the

Downloaded from https://academic.oup.com/mbe/article/27/9/2095/1009607 by guest on 20 March 2024

© The Author 2010. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. All rights reserved. For permissions, please e-mail: journals.permissions@oxfordjournals.org

exclusion of taxa with strongly deviating amino acid composition, the recoding of amino acids in groups, and the exclusion of the proteins with the most heterogeneous amino acid composition between taxa from the alignment. Furthermore, we applied two methods that consider compositional heterogeneity directly in the phylogenetic reconstruction, namely Bayesian inference analysis with the CAT–BP model (Blanquart and Lartillot 2008) and employing distance methods using equal input distances (Tamura and Kumar 2002).

Materials and Methods

Extraction and Alignment of Ribosomal Protein Sequences

Amino acid sequences of ribosomal proteins from 48 metazoans were retrieved from NCBI's RefSeq database, from gene model data sets derived from recent genome projects, or from EST data processed as previously described (Hausdorf et al. 2007, with the addition of a second clustering step to improve contig assembly). Slow-evolving taxa were selected instead of fast-evolving ones whenever possible (e.g., *Paraplanocera*, *Xiphinema*). These data were surveyed by the TBlastN algorithm based on a query set of 78 human cytoplasmic ribosomal protein sequences acquired from the Ribosomal Protein Gene Database (http://ribosome.med.miyazaki-u.ac.jp, excluding *rps4y* and *rpl41*, which are redundant or too short, respectively). All hits with an *e* value lower than 1×10^{-10} were again queried against the human ribosomal protein sequences by employing the genewisedb algorithm (score cutoff <50) as implemented in the Wise2 package (Birney et al. 2004). This was done to receive protein translations corrected for frameshift errors due to sequencing inaccuracy. Generally, the longest sequence was taken of each gene and taxon. The resulting nonredundant gene sets were individually aligned by the L-INS-i algorithm implemented in MAFFT (Katoh et al. 2002; Katoh and Toh 2008) and edited by Gblocks (Castresana 2000) using low stringency parameters. The final alignment, spanning 11,544 amino acid positions, was attained by concatenating all single alignments and has been deposited at TreeBASE (http://www.treebase.org, accession number S10436). Alignments with reduced taxon sets were attained by removing taxa from the final complete alignment.

Phylogenetic Analyses and Evaluation of Model Violation Caused by Compositional Heterogeneity

We performed Bayesian inference analyses with the CAT model that adjusts for site-specific amino acid frequencies (Lartillot and Philippe 2004) as implemented in PhyloBayes version 3.1c (http://megasun.bch.umontreal.ca/People/lartillot/www/index.htm). Eight independent chains were run for each analysis. The number of points of each chain, the number of points that were discarded as burn-in, and the largest discrepancy observed across all bipartitions (maxdiff) are listed in supplementary table S1 (Supplementary Material online). Taking every tenth sampled tree, a 50% majority rule consensus tree was computed using all chains.

We evaluated in how far the assumptions of the CAT model are violated by using posterior predictive tests. In posterior predictive tests, the observed value of a given test statistic on the original data is compared with the distribution of the test statistic on data replicates simulated under the reference model using parameter values drawn from the posterior distribution (every tenth sampled tree). The reference model is rejected for that statistic if the observed value of the test statistic deviates significantly. We used two test statistics measuring compositional heterogeneity implemented in PhyloBayes. One measures the compositional deviation of each taxon by summing the absolute differences between the taxon-specific and global empirical frequencies over the 20 amino acids. This test statistic indicates which taxa deviate significantly but raises a multiple-testing issue. Alternatively, the maximum deviation over the taxa is used as a global statistic.

To check the results of the Bayesian analyses with the CAT model, we performed maximum likelihood analyses using RAxML, version 7.2.4 (Stamatakis 2006) with the LG model (Le and Gascuel 2008) or the MULTICAT model (for recoded data). Confidence values were computed by bootstrapping (Felsenstein 1985) (100 replications).

Approaches for Reducing the Potential Impact of Compositional Bias

Three approaches to reduce compositional heterogeneity in the data set were applied. First, we excluded the taxa with the most strongly deviating amino acid composition as indicated by the posterior predictive tests and repeated the Bayesian inference analysis as described.

Second, we recoded the amino acid data into groups. We used the six groups of amino acids (AGPST, C, DENQ, FWY, HKR, and ILMV) that tend to replace one another (Dayhoff et al. 1978), as has been done by Embley et al. (2003). Susko and Roger (2007) developed an algorithm for constructing bins of amino acids in order to minimize compositional heterogeneity for a given alignment by minimizing the maximum chi-squared statistic for a taxon of the data set. We used the program minmax-chisq (http://www.mathstat.dal.ca/tsusko/software.cgi) to obtain these min-max chi-squared bins for the ribosomal protein data set. In order to lose as little information as possible, we chose the largest number of bins for which the minimum *P* value is larger than 0.1, which indicates that compositional homogeneity cannot be rejected for this set of bins according to the chi-square test.

The third strategy consisted in evaluating the compositional bias in each of the 78 ribosomal proteins separately by performing Bayesian inference analysis with the CAT model and posterior predictive tests using the global test statistic. Then, we excluded the proteins from the concatenated data set for which the CAT model is significantly violated according to posterior predictive tests. In addition, we ordered the proteins according to the *Z* scores of the

Downloaded from https://academic.oup.com/mbe/article/27/9/2095/1009607 by guest on 20 March 2024

global test statistic and excluded the third respectively the half of the proteins with the highest *Z* scores.

As alternative to the approaches for reducing compositional heterogeneity in the data set, we applied two phylogeny reconstruction methods that account for compositional heterogeneity. First, we performed a Bayesian analysis with the CAT–BP model (Blanquart and Lartillot 2008) as implemented in nhPhyloBayes (http://www.lirmm.fr/mab/blanquart/), which accounts for compositional heterogeneity between lineages by introducing breakpoints along the branches of the phylogeny at which the amino acid composition is allowed to change. In nhPhyloBayes, the number of components in the mixture has to be fixed. In the PhyloBayes analysis with the complete data set, the mean number of categories was 40.4 ± 14.9. Thus, we did not change the default in nhPhyloBayes (50 categories). Nine independent chains were run for each analysis. The number of points that were discarded as burn-in was determined for each chain separately by checking when the posterior probabilities of each run reach stationarity. Taking every tenth sampled tree, a 50% majority rule consensus tree was computed using all chains that sampled trees in the same high posterior probability range.

Second, we constructed a neighbor joining tree (Saitou and Nei 1987) based on equal input distances considering pattern and rate heterogeneity (Tamura and Kumar 2002) as implemented in MEGA version 4.1 (Tamura et al. 2007). For comparison, we calculated also a neighbor joining tree based on Jones, Taylor, and Thorton (JTT) distances (Jones et al. 1992) with rate variation among sites. Both analyses were calculated with $\alpha = 0.607$ as determined in the maximum likelihood analysis. Confidence values were computed by bootstrapping (Felsenstein 1985) (1,000 replications).

Results and Discussion

Strategies for Reducing Compositional Heterogeneity in Sequence Data

A posterior predictive test based on a PhyloBayes analysis of the complete data set including 11,544 amino acid positions derived from 78 ribosomal proteins of 48 metazoan taxa (fig. 1*A*) confirmed the observation of Lartillot and Philippe (2008) that the assumption of compositional homogeneity made by most protein models is strongly violated in metazoan phylogenomic data (table 1; supplementary table S2, Supplementary Material online). Thus, there is a risk of observing artifacts resulting from compositional bias. We applied three approaches to reduce compositional heterogeneity of the data set, namely excluding the taxa with the most strongly deviating amino acid composition according to the posterior predictive test, recoding of amino acids in groups with similar properties (Embley et al. 2003; Rodríguez-Ezpeleta et al. 2007) and in bins that minimize compositional heterogeneity (Susko and Roger 2007), and excluding proteins with a deviating amino acid composition.

The test statistic for individual taxa indicated that the amino acid composition of 18 taxa is significantly deviating (not considering the multiple-testing issue). When these taxa were excluded from the calculations (supplementary fig. S1, Supplementary Material online), a posterior predictive test indicated that the CAT model is no longer significantly violated (table 1; supplementary table S2, Supplementary Material online). We also excluded subsets including only the 6 and the 12 taxa with the smallest *P* values (supplementary figs. S2 and S3, Supplementary Material online). According to posterior predictive tests, both strategies proved to be sufficient to prevent significant model violation (table 1; supplementary table S2, Supplementary Material online). This shows that it is not necessary to exclude all taxa that have a significantly deviating amino acid composition to reduce the heterogeneity to a degree the model is no longer significantly violated.

When the amino acid sequences of the ribosomal proteins were recoded using the six Dayhoff groups of amino acids that tend to replace one another (Dayhoff et al. 1978) (supplementary fig. S4, Supplementary Material online), a posterior predictive test indicated that the CAT model is also no longer significantly violated (table 1; supplementary table S2, Supplementary Material online). However, the test statistic for the individual taxa indicated that the amino acid composition of 11 taxa is still significantly deviating. If these taxa were excluded from the calculations in addition to recoding (supplementary fig. S5, Supplementary Material online), the global *Z* score is further decreased (table 1; supplementary table S2, Supplementary Material online). Although fewer taxa were excluded than in the analysis of the unrecoded data set excluding significantly deviating taxa, the *Z* score indicated that the reduction of compositional heterogeneity is stronger than in this analysis (table 1).

Alternative to the Dayhoff groups of amino acids, we determined bins of amino acids that minimize compositional heterogeneity for the ribosomal protein data set with the method described by Susko and Roger (2007). Whereas the minimum *P* values for 18 or more bins are smaller than 0.05 (supplementary table S3, Supplementary Material online), the minimum *P* value for 17 minmax chi-squared bins (AI, RK, N, D, C, Q, ES, G, H, L, M, F, P, T, W, Y, and V) is 0.13, which indicates that compositional homogeneity cannot be rejected for these bins according to the chi-square test. However, a posterior predictive test showed that the CAT model is still significantly violated (table 1; supplementary table S2, Supplementary Material online) if the amino acid sequences of the ribosomal proteins were recoded using these bins. The global *Z* score indicated that the compositional heterogeneity is even slightly stronger than in the unrecoded data set (table 1). This contradiction might be explained by the fact that the chi-square test does not consider correlation due to relatedness of the taxa on a tree or by the biasing effect of invariable sites on this test (Foster 2004; Jermiin et al. 2004). The topology of the resulting tree is identical to that based on the complete unrecoded data set (fig. 1*A*). In addition, we recoded the data set with 12 and 8 minmax chi-squared bins (supplementary

Downloaded from https://academic.oup.com/mbe/article/27/9/2095/1009607 by guest on 20 March 2024

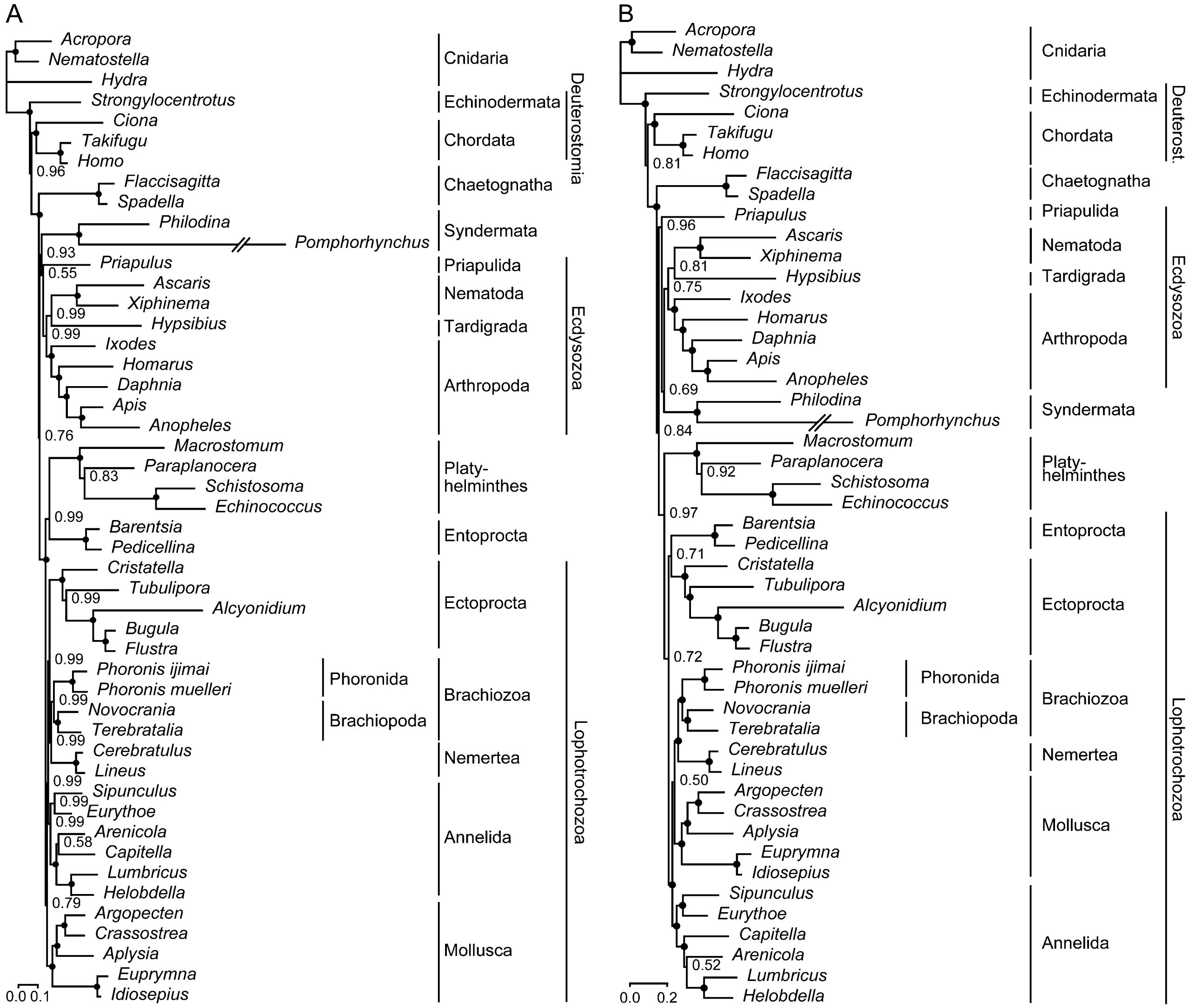

FIG. 1. Bayesian inference reconstructions based on 11,544 amino acid positions derived from 78 ribosomal proteins of 48 taxa. Bayesian posterior probabilities are shown to the right of the nodes; posterior probabilities equal to 1.0 are indicated by black circles. (*A*) Using the CAT model. (*B*) Using the CAT–BP model, which directly accounts for compositional heterogeneity between lineages. Because the topologies have not converged, the confidence values are not rigorous posterior probabilities in this case.

figs. S6 and S7, Supplementary Material online). Despite that the minimum *P* value for 12 bins (ARK, NQ, D, CG, ES, HT, IV, LM, F, P, W, and Y) being 0.62, a posterior predictive test revealed a continuing model violation (table 1; supplementary table S2, Supplementary Material online) if the amino acid sequences of the ribosomal proteins were recoded using these bins. The minimum *P* value for 8 bins (ARCK, NDP, QES, GLMY, HAT, IV, F, and W) amounted to 0.68, which is the highest minimum *P* value for any number of bins found with minmax-chisq. A posterior predictive test indicated that the CAT model is no longer significantly violated (table 1; supplementary table S2, Supplementary Material online) if the ribosomal protein sequences were recoded using these bins.

Phylogenetic analyses of the individual ribosomal proteins and posterior predictive tests using the global test statistic (supplementary table S4, Supplementary Material online) demonstrated that the amino acid composition of 7 of the 78 proteins violates the assumptions of the CAT model. However, using only the concatenated sequences of the 71 proteins (in total 10,469 amino acid positions) that match the model assumptions for phylogenetic inference (supplementary fig. S8, Supplementary Material online) proved not to be effective in preventing model misspecification (table 1; supplementary table S2, Supplementary Material online). Despite the reduction of compositional heterogeneity by excluding significantly deviating proteins, the test statistic for the individual taxa indicated that the amino acid composition of 16 taxa is still significantly biased. If these taxa were excluded from the calculations in addition to the exclusion of proteins (supplementary fig. S9, Supplementary Material online), posterior predictive testing revealed that the CAT model is no longer significantly violated (table 1; supplementary table S2, Supplementary Material online). Alternative to excluding only those proteins for which the CAT model is

Downloaded from https://academic.oup.com/mbe/article/27/9/2095/1009607 by guest on 20 March 2024

Table 1. Results of Posterior Predictive Tests Indicating the Ability of Different Approaches to Reduce Compositional Bias in Phylogenetic Inference of Metazoan Relationships.

Approach	Remaining Taxa	Remaining Positions in Alignment	Z Score	P Value	Number of Taxa with Significantly Deviating Amino Acid Composition
Original data set	48	11,544	4.59	0.00	18
Exclusion of the 18 taxa with a significantly differing amino acid composition	30	11,544	0.24	0.37	2
Exclusion of the 6 taxa with the most strongly differing amino acid composition	42	11,544	−0.02	0.47	10
Exclusion of the 12 taxa with the most strongly differing amino acid composition	36	11,544	0.05	0.45	4
Recoding using 6 Dayhoff groups	48	11,544	1.47	0.08	11
Recoding using 6 Dayhoff groups and exclusion of taxa with a significantly differing amino acid composition	37	11,544	−0.20	0.52	1
Recoding using 17 minmax chi-squared bins	48	11,544	5.37	0.00	20
Recoding using 12 minmax chi-squared bins	48	11,544	3.56	0.01	18
Recoding using 8 minmax chi-squared bins	48	11,544	1.48	0.08	12
Exclusion of proteins for which the CAT model is significantly violated	48	10,469	4.21	0.00	16
Exclusion of proteins for which the CAT model is significantly violated and exclusion of taxa with a significantly differing amino acid composition	32	10,469	−0.72	0.76	0
Exclusion of the third of the proteins with the highest heterogeneity	48	7,671	2.26	0.01	10
Exclusion of the half of the proteins with the highest heterogeneity	48	5,710	1.49	0.08	11

violated, we excluded the third respectively the half of the proteins with the highest Z scores of the global test statistic. If the third of the proteins with the highest Z scores is excluded from the data set (supplementary fig. S10, Supplementary Material online), the CAT model is still significantly violated (table 1; supplementary table S2, Supplementary Material online). However, if only the 39 proteins with the lowest Z scores were used for the phylogenetic analyses (supplementary fig. S11, Supplementary Material online), the model is no longer violated (table 1; supplementary table S2, Supplementary Material online).

Thus, the posterior predictive tests demonstrated that the compositional heterogeneity in the ribosomal protein data set could be reduced to a degree that the assumptions of the used model were no longer significantly violated with all three applied approaches. Depending on the data set and the main purpose of an analysis, it may be decided whether it is more appropriate to exclude the taxa or genes with the most deviating amino acid composition or to recode the amino acids into bins. Excluding taxa with strongly deviating amino acid composition has the disadvantage that the phylogenetic relationships of such taxa cannot be inferred. If this concerns all representatives of a taxon of interest, no inferences can be made about the relationships of that taxon. In the ribosomal protein data set, all representatives of Platyhelminthes and Syndermata have a significantly deviating amino acid composition so that the relationships of these phyla could not be determined when all taxa with a significantly deviating amino acid composition are excluded (supplementary fig. S1, Supplementary Material online). However, if only the six most strongly deviating taxa were excluded, Platyhelminthes and Syndermata are still represented in the data set and the model is no longer violated (supplementary fig. S2, Supplementary Material online). Excluding taxa may be the preferential option if only a few taxa have a strongly deviating amino acid composition and if these are not absolutely necessary for the question to be solved. Deleting proteins may be an appropriate method, if compositional heterogeneity is concentrated in a few proteins. In the present data set, this was not the case and about half of the ribosomal proteins had to be excluded to reduce compositional heterogeneity to a degree that the model used was no longer significantly violated (table 1). Probably, deleting proteins will rarely be an effective method because compositional bias is usually a genome-wide phenomenon.

In contrast, recoding may be the most appropriate approach if important or many taxa and many proteins have a strongly deviating amino acid composition. However, recoding the ribosomal protein sequences with the six Dayhoff groups of amino acids led to strong loss of information resulting in a large polytomy within Lophotrochozoa and a reduction of the posterior probabilities of several branches of the inferred phylogeny (supplementary fig. S4, Supplementary Material online). Using more bins for recoding conserves more information. Unfortunately, the method described by Susko and Roger (2007) does not guarantee that the bins that reduce compositional heterogeneity most effectively are determined. Nevertheless, it might be a helpful tool for exploring which bins reduce compositional bias so that the used model is no longer violated and that still more information is conserved than

Downloaded from https://academic.oup.com/mbe/article/27/9/2095/1009607 by guest on 20 March 2024

with recoding using the six Dayhoff groups. The efficiency of bins found with this method in reducing compositional heterogeneity can be checked with other approaches like posterior predictive tests.

We performed maximum likelihood analyses to check whether the described results of the Bayesian analyses might represent idiosyncrasies of the CAT model. A maximum likelihood analyses with the LG model (Le and Gascuel 2008) of the complete data set resulted in a tree (supplementary fig. S12, Supplementary Material online) that is similar to the result of the corresponding Bayesian analyses (fig. 1*A*) but shows monophyletic Deuterostomia. A maximum likelihood analyses with the data set recoded using the six Dayhoff groups of amino acids resulted in a similar topology (supplementary fig. S13, Supplementary Material online). However, in agreement with the corresponding Bayesian analyses, the support for the monophyly of Deuterostomia increased and the support for Entoprocta + Platyhelminthes decreased in comparison with the tree based on the unrecorded data set.

Phylogenetic Inference Methods Accounting for Compositional Heterogeneity

An alternative to reducing compositional heterogeneity in the data is using phylogeny reconstruction methods that directly account for compositional heterogeneity. One approach is using maximum likelihood or Bayesian analysis with models that consider nonstationary sequence evolution. Several such models have been proposed (Foster 2004; Blanquart and Lartillot 2006, 2008; Dutheil and Boussau 2008). We used the program nhPhyloBayes that implements the CAT–BP model (Blanquart and Lartillot 2008) for an automatic tree search. The CAT–BP model accounts for compositional heterogeneity between lineages by introducing breakpoints along the branches of the phylogeny at which the amino acid composition is allowed to change. This makes the calculations computationally expensive.

We started nine chains with the complete ribosomal protein data set. The runs showed two different behaviors: Whereas the number of breakpoints at which the amino acid composition changes varied between 0 and 5 (with a mean number between 0.18 and 0.48) in four of the nine runs, it suddenly increased to 38–58 breakpoints in the other chains at different times and remained in that range for the rest of the run (supplementary fig. S14*A*, Supplementary Material online). These long burn-ins indicate a lack of efficiency of the Markov chain Monte Carlo (MCMC) sampling algorithm. In the CAT–BP model, a conservative prior on the number of breakpoints N is used to avoid potential dominant effects of the prior on the posterior. Because the prior on N is conservative, a high observed N (as in our analysis) indicates that there is compositional bias in the data. The high number of breakpoints in the latter chains reflects the result of the posterior predictive test that 18 taxa belonging to several different clades have amino acid compositions that significantly deviate from the assumptions of the CAT model (supplementary table S2, Supplementary Material online). Although the high N induces a significant improvement of the model fit, the sudden increases of the number of breakpoints were accompanied by equally abrupt decreases of the likelihood (supplementary fig. S14*B*, Supplementary Material online). Nevertheless, the posterior probabilities increased (supplementary fig. S14*C*, Supplementary Material online). There are two possible interpretations of this behavior. From a frequentist's point of view, one may argue that the analysis went wrong. However, in the Bayesian logic, it is accepted that priors have an influence on the results and one accepts the solution with the highest posterior probability, even if it is different from the maximum likelihood solution.

The five chains with many breakpoints and high posterior probabilities converged with regard to most parameters. However, the topologies of these chains did not yet converge. This concerns especially the relationships of Ectoprocta and Entoprocta. In four of the five chains and in the consensus tree (fig. 1*B*), Bryozoa including Ectoprocta and Entoprocta is monophyletic (posterior probabilities 0.94–1.00), but in one chain, Entoprocta is the sister group of Platyhelminthes (posterior probability 1.00) and Ectoprocta is the sister group of Kryptrochozoa + Annelida + Mollusca (posterior probability 1.00). These two topologies are local optima. The lack of efficiency of the mixing behavior prevents the MCMC algorithm to escape from one optimum to the other. Thus, the obtained phylogeny should be interpreted with caution. Deuterostomia is paraphyletic in all chains.

Second, we constructed a neighbor joining tree based on equal input distances considering pattern and rate heterogeneity (Tamura and Kumar 2002). These distances were designed as an improvement in comparison to LogDet distances (Lockhart et al. 1994). There are several probable artifacts in the resulting tree (supplementary fig. S15, Supplementary Material online), including paraphyly of Deuterostomia, polyphyly of Ecdysozoa and a sister group relationship between Platyzoa, that is, Syndermata + Platyhelminthes, and Nematoda + Tardigrada. The same topology is recovered in a neighbor joining analysis based on JTT distances with rate variation among sites (supplementary fig. S16, Supplementary Material online). Thus, the use of equal input distances considering pattern and rate heterogeneity could not alleviate artifacts found with a model that does not consider heterogeneous amino acid compositions. The observed artifacts are, at least partly, the result of the loss of information resulting from the transformation of sequence data into distances.

A phylogenetic inference method that takes compositional heterogeneity into account is preferable to methods for reducing compositional heterogeneity in the data because reducing compositional heterogeneity results always in a loss of information. Unfortunately, the current implementation of the CAT–BP model is computationally so expensive that the topologies of the different runs did not converge even after several months. The much faster neighbor joining method using equal input distances produced artifacts. nhPhyloBayes may be a good option for

Downloaded from https://academic.oup.com/mbe/article/27/9/2095/1009607 by guest on 20 March 2024

controlling systematic errors in tree reconstruction that result from compositional bias for smaller data sets. However, for large data sets, using methods for reducing compositional heterogeneity in combination with inference methods that assume a stationary amino acid composition remains an option until faster hardware and more efficient algorithms have been developed.

Systematic Errors Affecting Phylogenetic Inference of Metazoan Relationships Based on Ribosomal Protein Sequences

Aside from evaluating methods for reducing compositional bias in phylogeny reconstruction, the identification of systematic errors in the reconstruction of metazoan phylogeny based on ribosomal protein sequences was another purpose of this study. Systematic errors may be indicated by a conflict between highly supported splits in trees calculated using different modifications of the same data set or calculated with different methods. A comparison of the trees resulting from the different analyses of the ribosomal protein data set (fig. 1, supplementary figs. S1–S16, Supplementary Material online) revealed that two cases are affected by such conflicts, namely the monophyly of Deuterostomia and the relationships of Ectoprocta and Entoprocta. The relationships of these groups found in the different analyses are summarized in table 2.

Deuterostomia (echinoderms, hemichordates, and chordates) is usually considered as one of the best supported metazoan groups (Hennig 1979; Ax 1995; Zrzavý et al. 1998; Nielsen 2001; Peterson and Eernisse 2001; Halanych 2004; Hejnol et al. 2009). However, their monophyly has recently been questioned by Lartillot and Philippe (2008). In the Bayesian trees calculated with the complete data set (fig. 1A), the data set recoded using 17 minmax chi-squared bins, the data set recoded using 12 minmax chi-squared bins (supplementary fig. S6, Supplementary Material online), the data set including only the 71 not significantly deviating ribosomal proteins (supplementary fig. S8, Supplementary Material online), and the data set including only the 52 ribosomal proteins (supplementary fig. S10, Supplementary Material online) with the least heterogeneity between lineages, paraphyly of Deuterostomia is strongly supported (posterior probabilities $\geq$0.95) in most cases because the representative of the echinoderms, *Strongylocentrotus*, is sister to all other representatives of Bilateria. However, this topology is not in concordance with the result obtained by Lartillot and Philippe (2008). In contrast to our findings, echinoderms (and hemichordates) are more closely related to other Bilateria in their tree calculated with the CAT model. Both posterior predictive tests of Lartillot and Philippe (2008) and our own posterior predictive tests showed that the assumptions of the CAT model are significantly violated by both phylogenomic data sets (table 1; supplementary table S2, Supplementary Material online). The situation is different in the calculations based on the data sets in which the model assumptions are not violated with the exception of the analysis based on the 39 ribosomal proteins with the least heterogeneity between lineages (supplementary fig. S11, Supplementary Material online). The support for the paraphyly of Deuterostomia decreases with increasing number of exclusions of taxa with the strongly deviating amino acid composition (posterior probability 0.94, if only the 6 most strongly deviating taxa were excluded, supplementary fig. S2, Supplementary Material online; 0.72, if the 12 most strongly deviating taxa were excluded, supplementary fig. S3, Supplementary Material online). In the tree calculated with the 71 protein data set from which taxa with a deviating amino acid composition have been excluded (supplementary fig. S9, Supplementary Material online), the support for Bilateria exclusive of Echinodermata also decreased (posterior probability 0.90), and in the tree based on the data set recoded using the Dayhoff groups (supplementary fig. S4, Supplementary Material online) or using eight minmax chi-squared bins (supplementary fig. S7, Supplementary Material online), the monophyly of Deuterostomia is well supported (posterior probability 0.96). This indicates that the paraphyly of Deuterostomia was caused by a systematic error. Thus, there is no reason to suppose that deuterostomy was ancestral (Lartillot and Philippe 2008).

A further case concerns the relationships of Entoprocta and Ectoprocta (table 2). These two taxa have been assigned to different subgroups of bilaterians based on the differences in cleavage patterns, larval types, and body cavities. The coelomate ectoprocts share a ciliated tentacular feeding apparatus around the mouth opening called lophophore and radial cleavage with Phoronida and Brachiopoda and were classified with them as Lophophorata (=Tentaculata). This group was long considered the sister or the stem group of Deuterostomia (Hennig 1979; Emig 1984; Ax 1995). The acoelomate entoprocts show spiral cleavage and have trochophora-type larvae and either were included in Trochozoa (Ax 1995; Zrzavý et al. 1998; Peterson and Eernisse 2001) or were classified with other acoelomata phyla in Platyzoa or as sister group of Platyzoa (Halanych 2004; Passamaneck and Halanych 2006). However, recent phylogenomic analyses (Hausdorf et al. 2007; Helmkampf et al. 2008; Struck and Fisse 2008; Witek et al. 2008; Hejnol et al. 2009) and ribosomal DNA (rDNA) analyses (Baguñà et al. 2008; Paps et al. 2009) confirmed the view of Nielsen (1985, 2001) and Cavalier-Smith (1998) that Ectoprocta and Entoprocta form a monophylum, Bryozoa (=Polyzoa, also including Cycliophora).

A sister group relationship between the acoelomate Entoprocta and Platyhelminthes is strongly supported by the analyses with data sets for which the assumptions of the CAT model are significantly violated (table 1; supplementary table S2, Supplementary Material online), namely the trees calculated with the complete data set (posterior probability 0.99; fig. 1A), the data set including the 71 not significantly deviating ribosomal proteins (posterior probability 0.95; supplementary fig. S8, Supplementary Material online), and the data set including the 52 least heterogeneous ribosomal proteins (posterior probability 0.98; supplementary fig. S10, Supplementary Material online). Again, the situation is different in some of the analyses

Downloaded from https://academic.oup.com/mbe/article/27/9/2095/1009607 by guest on 20 March 2024

Table 2. Phylogenetic Differences between Different Approaches to Reduce Compositional Bias.

Method	Data Set	Deuterostomia	Echinodermata Sister to Rest of Bilateria	Ectoprocta + Entoprocta	Entoprocta + Platyhelminthes	Lophophorata
Bayesian (CAT model)	Original data set	—	0.96	—	0.99	—
Bayesian (CAT model)	Exclusion of the 18 taxa with a significantly differing amino acid composition	?	?	—	?	0.99
Bayesian (CAT model)	Exclusion of the 6 taxa with the most strongly differing amino acid composition	—	0.94	—	—	0.65
Bayesian (CAT model)	Exclusion of the 12 taxa with the most strongly differing amino acid composition	—	0.72	—	?	0.73
Bayesian (CAT model)	Recoding using 6 Dayhoff groups	0.96	—	—	—	—
Bayesian (CAT model)	Recoding using 6 Dayhoff groups and exclusion of taxa with a significantly differing amino acid composition	?	?	0.98	—	—
Bayesian (CAT model)	Recoding using 17 minmax chi-squared bins	—	1.00	—	0.96	—
Bayesian (CAT model)	Recoding using 12 minmax chi-squared bins	—	0.91	—	0.99	—
Bayesian (CAT model)	Recoding using 8 minmax chi-squared bins	0.96	—	—	1.00	0.79
Bayesian (CAT model)	Exclusion of proteins for which the CAT model is significantly violated	—	0.95	—	0.95	—
Bayesian (CAT model)	Exclusion of proteins for which the CAT model is significantly violated and exclusion of taxa with a significantly differing amino acid composition	—	0.90	—	—	0.99
Bayesian (CAT model)	Exclusion of the third of the proteins with the highest heterogeneity	—	0.97	—	0.98	—
Bayesian (CAT model)	Exclusion of the half of the proteins with the highest heterogeneity	—	0.86	—	0.94	—
Bayesian (CAT + BP model)	Original data set	—	0.81	0.71*	—	—
Maximum likelihood (LG model)	Original data set	0.90	—	—	0.88	—
Maximum likelihood (MULTICAT model)	Recoding using 6 Dayhoff groups	0.97	—	—	0.52	—
Neighbor joining (equal input distances with pattern and rate heterogeneity)	Original data set	—	—	0.36	—	—
Neighbor joining (JTT distances)	Original data set	—	—	0.51	—	—

NOTE.—If a group is monophyletic, the posterior probability, respectively, the bootstrap support is given.
*Posterior probabilities in four of five chains 0.94–1.00.

with the data sets in which the assumptions of the used model are not violated. The support for Entoprocta + Platyhelminthes decreased in the analyses based on the 39 least heterogeneous ribosomal proteins (posterior probability 0.94; supplementary fig. S11, Supplementary Material online). When the 6 taxa with the most strongly deviating amino acid composition were excluded from the analysis (supplementary fig. S2, Supplementary Material online) or when the data set with all 48 taxa was recoded with the Dayhoff groups (supplementary fig. S4, Supplementary Material online), Entoprocta is part of a large polytomy within Lophotrochozoa. Finally, Entoprocta and Ectoprocta form a well-supported clade (posterior probability 0.98) in the analysis of the 37 taxa data set recoded using the Dayhoff groups (supplementary fig. S5, Supplementary Material online) supporting the monophyly of Bryozoa. This group has also been found in four of the five chains that attained high posterior probabilities and the consensus tree (fig. 1*B*) of the nhPhyloBayes analysis as well as in the neighbor joining analyses (supplementary figs. S14 and S15, Supplementary Material online), albeit without bootstrap support. Nevertheless, this result is more controversial than the

Downloaded from https://academic.oup.com/mbe/article/27/9/2095/1009607 by guest on 20 March 2024

monophyly of Deuterostomia because the analyses of the unrecoded data sets from which the taxa with the most strongly deviating amino acid composition were excluded (supplementary figs. S1–S3, Supplementary Material online), the analysis of the data set recoded using the eight minmax chi-squared bins (supplementary fig. S7, Supplementary Material online), and the analysis of the 71 protein data set excluding compositionally heterogeneous taxa (supplementary fig. S9, Supplementary Material online) revealed another possibility, namely the monophyly of Lophophorata as suggested by Emig (1984). Moreover, the analyses of the unrecoded data sets excluding the 12 or 18 taxa with the most strongly deviating amino acid composition (supplementary figs. S1 and S3, Supplementary Material online) and the analysis of the 71 protein data set excluding compositionally heterogeneous taxa (supplementary fig. S9, Supplementary Material online) revealed a sister group relationship between Phoronida and Ectoprocta (posterior probability 0.73, 0.98, and 0.93, respectively). This grouping furthermore challenges the monophyly of Brachiozoa including Phoronida and Brachiopoda that has been supported by analyses based on rDNA (Cohen et al. 1998; Cohen 2000; Mallatt & Winchell 2002; Halanych 2004; Cohen & Weydmann 2005; Baguñà et al. 2008; Paps et al. 2009; but see Passamaneck & Halanych 2006), sodium–potassium ATPase α-subunit (Anderson et al. 2004), morphology (Nielsen 2001), a combination of morphological and 18S rDNA data sets (Zrzavý et al. 1998; Giribet et al. 2000; Peterson & Eernisse 2001), and phylogenomic analyses (Helmkampf et al. 2008). Brachiozoa is found in most of the trees (fig. 1*A* and *B*; supplementary figs. S2, S4–S8, S10–S12, S15, and S16, Supplementary Material online) and is also strongly supported in analyses in which the assumptions of the used models are not violated, namely the analyses of the data sets recoded using the Dayhoff groups (supplementary figs. S4 and S5, Supplementary Material online; posterior probability 0.99 and 0.98, respectively) and the analysis with the CAT–BP model (fig. 1*B*; posterior probability 1.00). These contradictory results between analyses of data sets that are in compliance with the assumptions of the used model concerning compositional homogeneity indicate further systematic errors that need to be addressed in future studies.

Supplementary Material

Supplementary figures S1–S16 and supplementary tables S1–S4 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We are grateful to I. Ebersberger, S. Strauss, and A. von Haeseler (Max F. Perutz Laboratories, Center for Integrative Bioinformatics, Vienna) for the processing of the EST data. We also thank S. Blanquart (Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier, Montpellier) for help with nhPhyloBayes and valuable discussions of Bayesian inference, as well as E. Susko (Dalhousie University, Halifax) for help with minmax-chisq. Furthermore, we thank both as well as three anonymous referees for constructive suggestions on the manuscript. This study was funded by the priority program "Deep Metazoan Phylogeny" of the Deutsche Forschungsgemeinschaft (HA 2763/5).

References

Anderson FE, Cordoba AJ, Thollesson M. 2004. Bilaterian phylogeny based on analyses of a region of the sodium-potassium ATPase α-subunit gene. *J Mol Evol.* 58:252–268.

Ax P. 1995. Das System der Metazoa I. Stuttgart (Germany): G. Fischer.

Baguñà J, Martinez P, Paps J, Riutort M. 2008. Back in time: a new systematic proposal for the Bilateria. *Philos Trans R Soc Lond B.* 363:1481–1491.

Baurain D, Brinkmann H, Philippe H. 2007. Lack of resolution in the animal phylogeny: closely spaced cladogeneses or undetected systematic errors? *Mol Biol Evol.* 24:6–9.

Birney E, Clamp M, Durbin R. 2004. GeneWise and Genomewise. *Genome Res.* 14:988–995.

Blanquart S, Lartillot N. 2006. A Bayesian compound stochastic process for modeling nonstationary and nonhomogeneous sequence evolution. *Mol Biol Evol.* 23:2058–2071.

Blanquart S, Lartillot N. 2008. A site- and time-heterogeneous model of amino acid replacement. *Mol Biol Evol.* 25:842–858.

Brinkmann H, Philippe H. 2008. Animal phylogeny and large-scale sequencing: progress and pitfalls. *J Syst Evol.* 46:274–286.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol.* 17:540–552.

Cavalier-Smith T. 1998. A revised six-kingdom system of life. *Biol Rev.* 73:203–266.

Cohen BL. 2000. Monophyly of brachiopods and phoronids: reconciliation of molecular evidence with Linnaean classification (the subphylum Phoroniformea nov.). *Proc R Soc Lond B.* 267:225–231.

Cohen BL, Gawthrop A, Cavalier-Smith T. 1998. Molecular phylogeny of brachiopods and phoronids based on nuclear-encoded small subunit ribosomal RNA gene sequences. *Philos Trans R Soc Lond B.* 353:2039–2061.

Cohen BL, Weydmann A. 2005. Molecular evidence that phoronids are a subtaxon of brachiopods (Brachiopoda: Phoronata) and that genetic divergence of metazoan phyla began long before the early Cambrian. *Organ Diver Evol.* 5:253–273.

Collins TM, Fedrigo O, Naylor GJP. 2005. Choosing the best genes for the job: the case for stationary genes in genome-scale phylogenetics. *Syst Biol.* 54:493–500.

Dayhoff MO, Schwartz RM, Orcutt BC. 1978. A model of evolutionary change in proteins. In: Dayhoff MO, editor. Atlas of protein sequence and structure Vol. 5 Suppl. 3. Washington, DC: National Biomedical Research Foundation. p. 345–352.

Delsuc F, Brinkmann H, Philippe H. 2005. Phylogenomics and the reconstruction of the tree of life. *Nat Rev Genet.* 6:361–375.

Dunn CW, Hejnol A, Matus DQ, et al. (18 co-authors). 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature.* 452:745–749.

Dutheil J, Boussau B. 2008. Non-homogeneous models of sequence evolution in the Bio++ suite of libraries and programs. *BMC Evol Biol.* 8:255.

Embley TM, van der Giezen M, Horner DS, Dyal PL, Foster P. 2003. Mitochondria and hydrogenosomes are two forms of the same fundamental organelle. *Philos Trans R Soc Lond B.* 358:191–203.

Emig C. 1984. On the origin of the Lophophorata. *Z Zool Syst Evolutionsforsch.* 22:91–94.

Downloaded from https://academic.oup.com/mbe/article/27/9/2095/1009607 by guest on 20 March 2024

Foster PG. 2004. Modeling compositional heterogeneity. *Syst Biol.* 53:485–495.

Foster PG, Hickey DA. 1999. Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. *J Mol Evol.* 48:284–290.

Felsenstein J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39:783–791.

Giribet G, Distel DL, Polz M, Sterrer W, Wheeler WC. 2000. Triploblastic relationships with emphasis on the acoelomates and the position of Gnathostomulida, Cycliophora, Plathelminthes, and Chaetognatha: a combined approach of 18S rDNA sequences and morphology. *Syst Biol.* 49: 539–562.

Halanych KM. 2004. The new view of animal phylogeny. *Annu Rev Ecol Evol Syst.* 35:229–256.

Hausdorf B, Helmkampf M, Meyer A, Witek A, Herlyn H, Bruchhaus I, Hankeln T, Struck TH, Lieb B. 2007. Spiralian phylogenomics supports the resurrection of Bryozoa comprising Ectoprocta and Entoprocta. *Mol Biol Evol.* 24:2723–2729.

Hejnol A, Obst M, Stamatakis A, et al. (17 co-authors). 2009. Assessing the root of bilaterian animals with scalable phylogenomic methods. *Proc R Soc Lond B.* 276:4261–4270.

Helmkampf M, Bruchhaus I, Hausdorf B. 2008. Phylogenomic analyses of lophophorates (brachiopods, phoronids and bryozoans) confirm the Lophotrochozoa concept. *Proc R Soc Lond B.* 275:1927–1933.

Hennig W. 1979. Wirbellose I (ausgenommen Gliedertiere). Taschenbuch der Speziellen Zoologie, 4th ed. Vol. 2. Jena, Germany: G. Fischer.

Ho SYW, Jermiin LS. 2004. Tracing the decay of the historical signal in biological sequence data. *Syst Biol.* 53:623–637.

Jermiin LS, Ho SYW, Ababneh F, Robinson J, Larkum AWD. 2004. The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated. *Syst Biol.* 53:638–643.

Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci.* 8:275–282.

Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30:3059–3066.

Katoh K, Toh H. 2008. Recent developments in the MAFFT multiple sequence alignment program. *Brief Bioinform.* 9:286–298.

Lartillot N, Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol Biol Evol.* 21:1095–1109.

Lartillot N, Philippe H. 2008. Improvement of molecular phylogenetic inference and the phylogeny of Bilateria. *Philos Trans R Soc B.* 363:1463–1472.

Le SQ, Gascuel O. 2008. An improved general amino acid replacement matrix. *Mol Biol Evol.* 25:1307–1320.

Lockhart PJ, Steel MA, Hendy MD, Penny D. 1994. Recovering evolutionary trees under a more realistic model of sequence evolution. *Mol Biol Evol.* 11:605–612.

Mallatt J, Winchell CJ. 2002. Testing the new animal phylogeny: first use of combined large-subunit and small-subunit rRNA gene sequences to classify the protostomes. *Mol Biol Evol.* 19:289–301.

Nielsen C. 1985. Animal phylogeny in the light of the trochaea theory. *Biol J Linn Soc.* 25:243–299.

Nielsen C. 2001. Animal evolution: interrelationships of the living phyla. 2nd ed. Oxford: Oxford University Press.

Paps J, Baguñà J, Riutort M. 2009. Lophotrochozoa internal phylogeny: new insights from an up-to-date analysis of nuclear ribosomal genes. *Proc R Soc Lond B.* 276:1245–1254.

Passamaneck Y, Halanych KM. 2006. Lophotrochozoan phylogeny assessed with LSU and SSU data: evidence of lophophorate polyphyly. *Mol Phylogenet Evol.* 40:20–28.

Peterson KJ, Eernisse DJ. 2001. Animal phylogeny and the ancestry of bilaterians: inferences from morphology and 18S rDNA gene sequences. *Evol Dev.* 3:170–205.

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

Philippe H, Telford MJ. 2006. Large-scale sequencing and the new animal phylogeny. *Trends Ecol Evol.* 21:614–620.

Phillips MJ, Delsuc F, Penny D. 2004. Genome-scale phylogeny and the detection of systematic biases. *Mol Biol Evol.* 21:1455–1458.

Rodríguez-Ezpeleta N, Brinkmann H, Roure B, Lartillot N, Lang BF, Philippe H. 2007. Detecting and overcoming systematic errors in genome-scale phylogenies. *Syst Biol.* 56:389–399.

Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol.* 4:406–425.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22:2688–2690.

Struck TH, Fisse F. 2008. Phylogenetic position of Nemertea derived from phylogenomic data. *Mol Biol Evol.* 25:728–736.

Susko E, Roger AJ. 2007. On reduced amino acid alphabets for phylogenetic inference. *Mol Biol Evol.* 24:2139–2150.

Tamura K, Dudley J, Nei M, Kumar S. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol.* 24:1596–1599.

Tamura K, Kumar S. 2002. Evolutionary distance estimation under heterogeneous substitution pattern among lineages. *Mol Biol Evol.* 19:1727–1736.

Witek A, Herlyn H, Meyer A, Boell L, Bucher G, Hankeln T. 2008. EST based phylogenomics of Syndermata questions monophyly of Eurotatoria. *BMC Evol Biol.* 8:345.

Zrzavý J, Mihulka S, Kepka P. 1998. Bezděk A. 1998. Tietz D. 1998. Phylogeny of the Metazoa based on morphological and 18S ribosomal DNA evidence. *Cladistics* 14:249–285.

Downloaded from https://academic.oup.com/mbe/article/27/9/2095/1009607 by guest on 20 March 2024