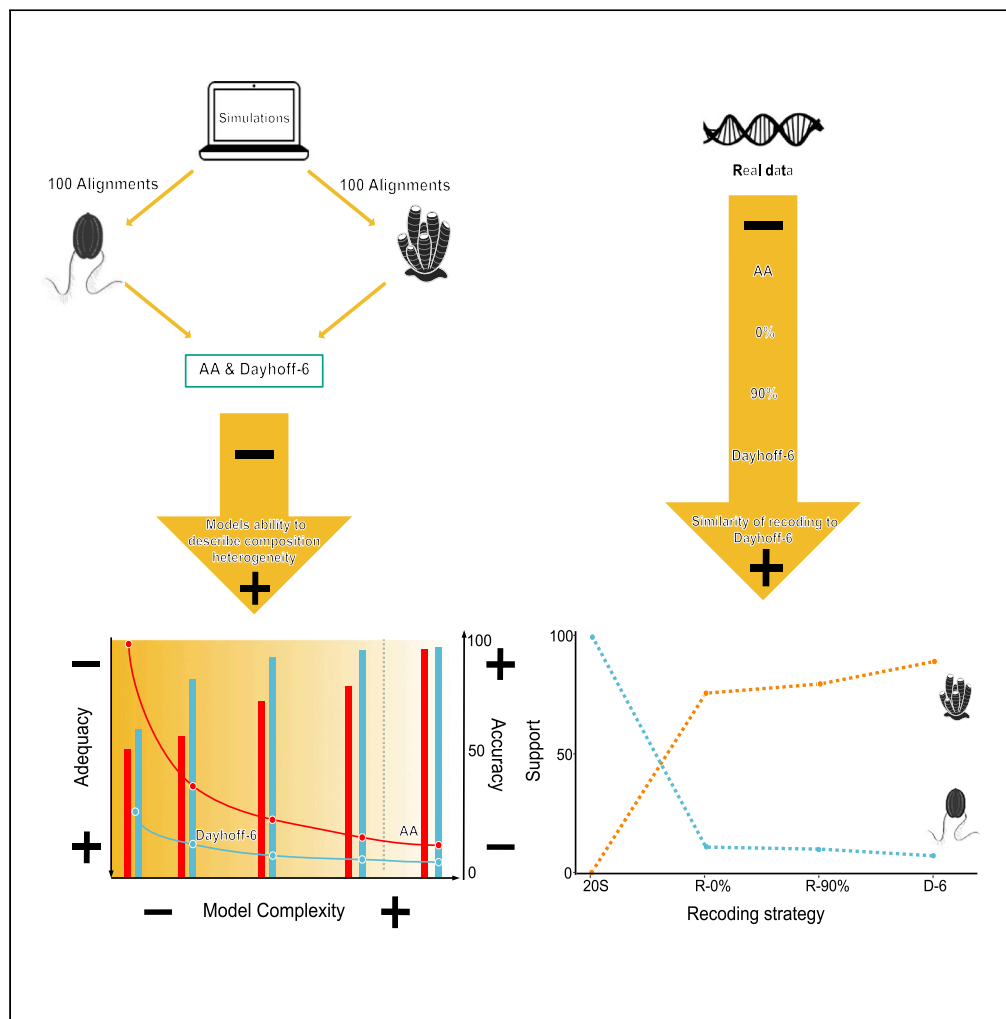


Article

Resolving tricky nodes in the tree of life through amino acid recoding



Mattia Giacomelli,
Maria Eleonora
Rossi, Jesus
Lozano-
Fernandez,
Roberto Feuda,
Davide Pisani

davide.pisani@bristol.ac.uk

Highlights

Recoding improves accuracy when the amino acid alignment cannot be adequately modeled

Recoding increases computational speed and can help achieving a greener phylogenomics

There is no evidence that recoding might induce tree reconstruction artifacts

Comb jellies are unlikely to represent the sister of all the other animals

Article

Resolving tricky nodes in the tree of life through amino acid recoding

Mattia Giacomelli,¹ Maria Eleonora Rossi,² Jesus Lozano-Fernandez,³ Roberto Feuda,⁴ and Davide Pisani^{1,2,5,*}

SUMMARY

Genomic data allowed a detailed resolution of the Tree of Life, but “tricky nodes” such as the root of the animals remain unresolved. Genome-scale datasets are heterogeneous as genes and species are exposed to different pressures, and this can negatively impacts phylogenetic accuracy. We use simulated genomic-scale datasets and show that recoding amino acid data improves accuracy when the model does not account for the compositional heterogeneity of the amino acid alignment. We apply our findings to three datasets addressing the root of the animal tree, where the debate centers on whether sponges (Porifera) or comb jellies (Ctenophora) represent the sister of all other animals. We show that results from empirical data follow predictions from simulations and suggest that, at the least in phylogenies inferred from amino acid sequences, a placement of the ctenophores as sister to all the other animals is best explained as a tree reconstruction artifact.

INTRODUCTION

Genome-scale (i.e., phylogenomic) datasets have enormously increased our understanding of the Tree of Life.^{1–4} Yet, “tricky nodes” remain that are proving difficult to resolve, such as the root of the bacterial, archaeal, plant and animal trees.^{5–9} For example, at the root of the animal tree, it is still unclear whether the sponges (Porifera) or the comb jellies (Ctenophora) are the sister to all other metazoans,^{10–20} with implications reaching areas as different as developmental biology and paleoecology.^{21,22}

Compositional heterogeneity is an important confounding factor in phylogenomics, and the use of evolutionary models, such as the Category (CAT²³) and the Node Discrete Compositional Heterogeneous (NDCH²⁴) model, that can respectively account for the fact that amino acids have different frequencies at different sites (hereafter referred to as across-site compositional heterogeneity) and in different species (hereafter referred to as across-lineage compositional heterogeneity) are gaining traction.^{4,25} However, even these models fail to adequately describe (i.e., fit) the heterogeneity (both across sites and lineages) of datasets designed to resolve tricky nodes in the Tree of Life.¹¹

Amino acid recoding, hereafter simply referred to as recoding (see Figure 1 and STAR Methods) was initially used to reduce the across-lineage compositional heterogeneity and saturation of amino acid datasets.^{26–29} However, recoding has also been shown to reduce across-site compositional heterogeneity¹¹ and this approach has been widely applied, frequently in combination with across-site and/or across-lineage compositionally heterogeneous models, to attempt resolving challenging nodes in the Tree of Life. Other applications of recoding can be found in the literature, e.g., to test inadequate fit of models to nucleotide data.³⁰ To demonstrate the popularity of recodings, a list of more than 100 papers that used these methods has recently been published,³¹ but in a context where it was argued that recoding cannot improve phylogenetic accuracy. Although some arguments against the use of recoding have been rebutted,³² it has recently been suggested that recoding can only strips information from the data.¹³ This claim remains unchallenged, and it is still unclear whether recoding, in some cases, might reduce accuracy.³²

We investigated whether recoding can improve phylogenetic accuracy using simulated alignments¹² and three empirical datasets.^{5,19,20} We found that recoding, in the largest majority of cases, had a positive effect on phylogenetic accuracy in simulations. We show that the extent to which recoding improved accuracy correlated with the extent to which recoding decreased the Z-scores of Posterior Predictive Analyses (PPA – STAR methods) performed to test whether the model used adequately described the data. Contrary

¹Bristol Palaeobiology Group, School of Biological Sciences, University of Bristol, Life Sciences Building, Tyndall Avenue, Bristol, BS8 1TQ, UK

²Bristol Palaeobiology Group, School of Earth Sciences, University of Bristol, Life Sciences Building, Tyndall Avenue, Bristol BS8 1TQ, UK

³Department of Genetics, Microbiology and Statistics, & Biodiversity Research Institute (IRBio), Faculty of Biology, University of Barcelona, Barcelona, Spain

⁴Department of Genetics and Genome Biology, University of Leicester, Leicester, UK

⁵Lead contact

*Correspondence:

davide.pisani@bristol.ac.uk
<https://doi.org/10.1016/j.isci.2022.105594>



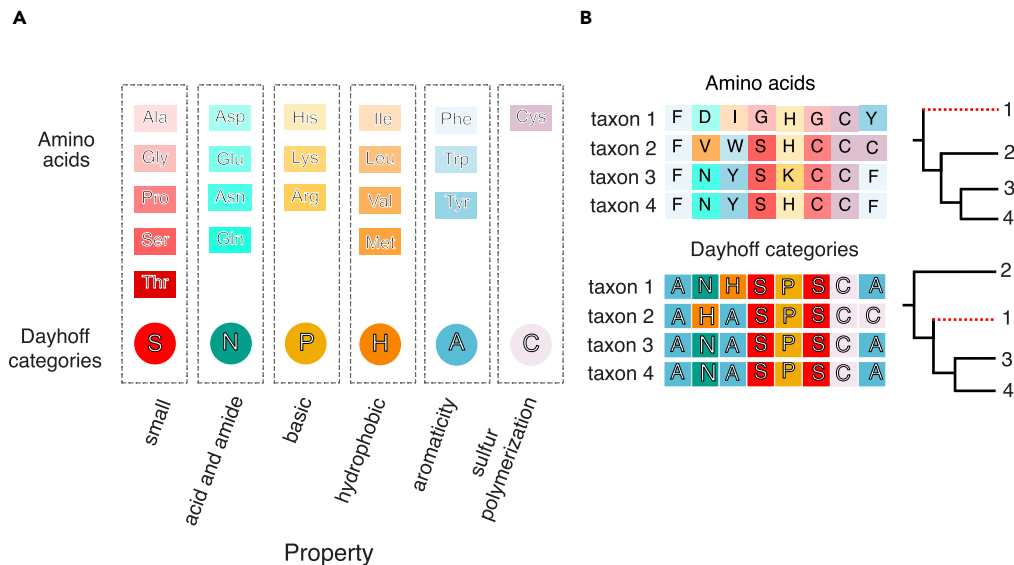


Figure 1. Recoding data

The six-bin Dayhoff-6 recoding scheme,³³ see [STAR methods](#), is the most widely used recoding strategy, and is the one primarily tested in this study. Dayhoff-6 recoding partitions amino acids into six differently sized bins (see [STAR methods](#)), based on how frequently they are expected to exchange with each other.

(A) The bins of the Dayhoff-6 scheme and the biochemical properties of the amino acids in each bin.

(B) An exemplar amino acid dataset and its Dayhoff-6 recoded representation. Dayhoff-6 recoding is achieved by replacing, in multiple sequence alignments, one letter amino acid codes with one letter codes representing the bin where the considered amino acid is clustered.

to previous conjectures,¹³ we found no evidence that recoding might cause the emergence of artifactual (i.e., incorrect but highly supported) relationships, and we therefore conclude that, at the least in our simulations, recoding is conservative because, at worst, it seems to reduce support for true clades. We show that results from three empirical datasets^{5,19,20} compare well with those from simulated data, suggesting that recoding can help resolve tricky nodes in the Tree of Life, and that a placement of the ctenophores as the sister of all the other animals, in analyses of amino acid sequence data, is best explained as a tree reconstruction artifact. Finally, we note that recoding substantially improves computational efficiency. Accordingly, we suggest that recoding has potential as a ready available tool that can be added to the list of techniques available to achieve a greener phylogenomics.^{34,35}

RESULTS

Recoding improves accuracy with phylogenomic scale datasets

We studied whether recoding improved phylogenetic accuracy using 200, simulated, across-site compositional heterogeneous alignments of 30,000 sites and 20 taxa.¹² A graphical representation of our complete analytical pipeline is reported in [Figure S1](#). The parameters of the substitution model used to simulate the data were estimated from a published superalignment¹⁸ under two alternative trees ([Figure S2](#)). The first tree represents a metazoan phylogeny where the ctenophores are the sister of all the other animals – Ctenophora-sister hypothesis.^{10,17,19,20,36} The second assumes that the sponges are the sister of all the other animals – Porifera-sister hypothesis.^{11,14–16,18} When data are simulated under Ctenophora-sister two long branches (Ctenophora and the outgroups) are on one side of the internal, target branch ([Figure S2](#)), and two short branches (Porifera and the branch leading to the other animals) are found on the other side. Ctenophora-sister therefore represents the case when the simulation target is a tree in the “Farris zone”.¹² When Porifera-sister is the simulation target, a short and a long branch are found on both sides of the internal, target branch ([Figure S2](#)). Porifera-sister represents the case when the simulation target is a tree in the “Felsenstein zone”.¹² One hundred datasets simulated under each topology ([STAR methods](#) for details) were analyzed in a Bayesian framework as both amino acid and Dayhoff-6 ([Figure 1](#) and [STAR methods](#)) recoded data. Sensitivity tests were performed to evaluate whether our experimental choices affected our results – [STAR methods](#) and [Figure S1](#).

We started our analyses using a CAT-based²³ across-site compositionally heterogeneous model with 10 site-frequency categories. This model was combined with a Poisson (F81³⁷) process that assumes that all amino acids have the same exchangeability rate, and a Gamma distribution³⁸ to account for across-site heterogeneity of replacement rate – [STAR methods](#). This is the relatively simple nCAT10-F81 + G model, which will be hereafter referred to as nCAT10. We then increased the number of site-frequency categories to 30, 60, 120 and 240 (defining nCAT30-F81 + G to nCAT240-F81 + G models; hereafter referred to as nCAT30 to nCAT240) to progressively better account for across-site compositional heterogeneity. After that, we performed analyses using the well-known CAT-F81 + G (hereafter CAT²³) model which estimates the number of site-frequency categories necessary to best describe the across-site compositional heterogeneity of the data during tree search. We then tested a General Time Reversible (GTR + G; hereafter GTR) model that, differently from the CAT-based models described above, does not account for across-site compositional heterogeneity but models the variation in exchange rate among amino acids.³⁹ Finally, we tested the CAT-GTR + G (hereafter CAT-GTR) model, which combines the properties of the CAT and GTR models and accounts for both across-site compositional heterogeneity and the heterogeneity of the amino acid replacement process.

When Ctenophora-sister is the simulation target (i.e. the “Farris zone” case), the true tree was recovered with a frequency close to 100% irrespective of model and data type, i.e. Amino acids (AA) and recoded (Rec) data ([Figure 2A](#)). Differently, when Porifera-sister is the simulation target (i.e., the “Felsenstein zone” case), model and data type have a strong impact on accuracy ([Figure 2A](#)). To account for this topology-dependency, for each data type and model we calculated its Total Accuracy (TA_{AA} and TA_{Rec}) as the percentage of accurately inferred trees under both topologies with a minimal posterior probability of 0.5 (see [STAR methods](#) for our definition of accuracy and [Figure 2B](#)). We used Posterior Predictive Analysis (PPA) of character state Diversity (PPA-Div – [STAR methods](#) for details and interpretation) to quantify how well different models describe the across-site compositional heterogeneity of the amino acid and recoded data.

The extent to which recoding improved accuracy and reduced PPA-Div scores seem related ([Figures 2A, 2B, and 2C](#)). The CAT-based model with the highest PPA-Div_{AA} scores (indicating the poorest fit to the amino acid data) was nCAT10 (Average PPA-Div_{AA} ~ 17; [Figure 2C](#)). This model achieved much lower PPA-Div_{Rec} scores (Average PPA-Div_{Rec} ~ 2.5) and emerged as the model for which recoding most strongly improved accuracy (TA_{AA} = 57%; TA_{Rec} = 89.5%; [Figures 2A and 2B](#) – a 32.5% improvement). At the other end of the spectrum, CAT-GTR achieved the lowest PPA-Div_{AA} scores (Average PPA-Div_{AA} = 0.6 – [Figure 2C](#)) indicating that this model adequately described the simulated amino acid data. CAT-GTR achieved PPA-Div_{Rec} scores comparable to those achieved by the amino acid data (Average PPA-Div_{Rec} ~ -0.9), and emerged as the model that, with recoded data, had the worst (despite still very good) relative performance (TA_{AA} = 98%; TA_{Rec} = 95%; [Figure 2B](#)). CAT, which for our simulated datasets estimated on average 277 site-frequency categories (SD = 28.23), and nCAT240 were also found to have very low PPA-Div_{AA} scores ([Figure 2C](#)), indicating that they adequately described the amino acid data. As in the case of CAT-GTR, the estimated PPA-Div_{Rec} scores of CAT and nCAT240 ([Figure 2C](#)) were comparable to those obtained using amino acid data, and as in the case of CAT-GTR, recoding did not improve accuracy under these two models ([Figures 2A and 2B](#)). The other CAT-based models (nCAT30 to nCAT120) had PPA-Div_{AA} and PPA-Div_{Rec} scores falling between the extremes defined by nCAT10 on one side, and CAT-GTR, CAT, and nCAT240 on the other ([Figure 2C](#)). For these models, accuracy improved upon recoding ([Figures 2B and 2D](#)), with the improvement becoming less marked as the ability of the model to describe the simulated amino acid data increased, and PPA-Div_{AA} and PPA-Div_{Rec} scores became more similar ([Figures 2A, 2B, and 2C](#)). PPA-Div scores for GTR (Average PPA-Div_{AA} = 43.6; Average PPA-Div_{Rec} = 30.2) indicated that this model fits both amino acid and recoded data poorly. However, PPA scores achieved by GTR when the data are recoded are lower than those achieved by the amino acid data. As expected, also under GTR, recoding improved accuracy. However, the improvement was only 7% ([Figure 2B](#)), and it was achieved in a context where an artifactual tree was still more likely to be inferred than the target tree when Porifera-sister was assumed to be true. This result is consistent with the poor fit of GTR to both the amino acid and recoded data.

To clarify the relationship between changes in PPA-Div Z-scores and accuracy, for each model we compared δ PPA-Div (|PPA-Div_{AA}| minus |PPA-Div_{Rec}|) and δ TA (TA_{Rec} minus TA_{AA}). In PPA analyses smaller absolute Z-scores usually identify a better absolute fit of the model to the data,⁴⁰ and we found a strong

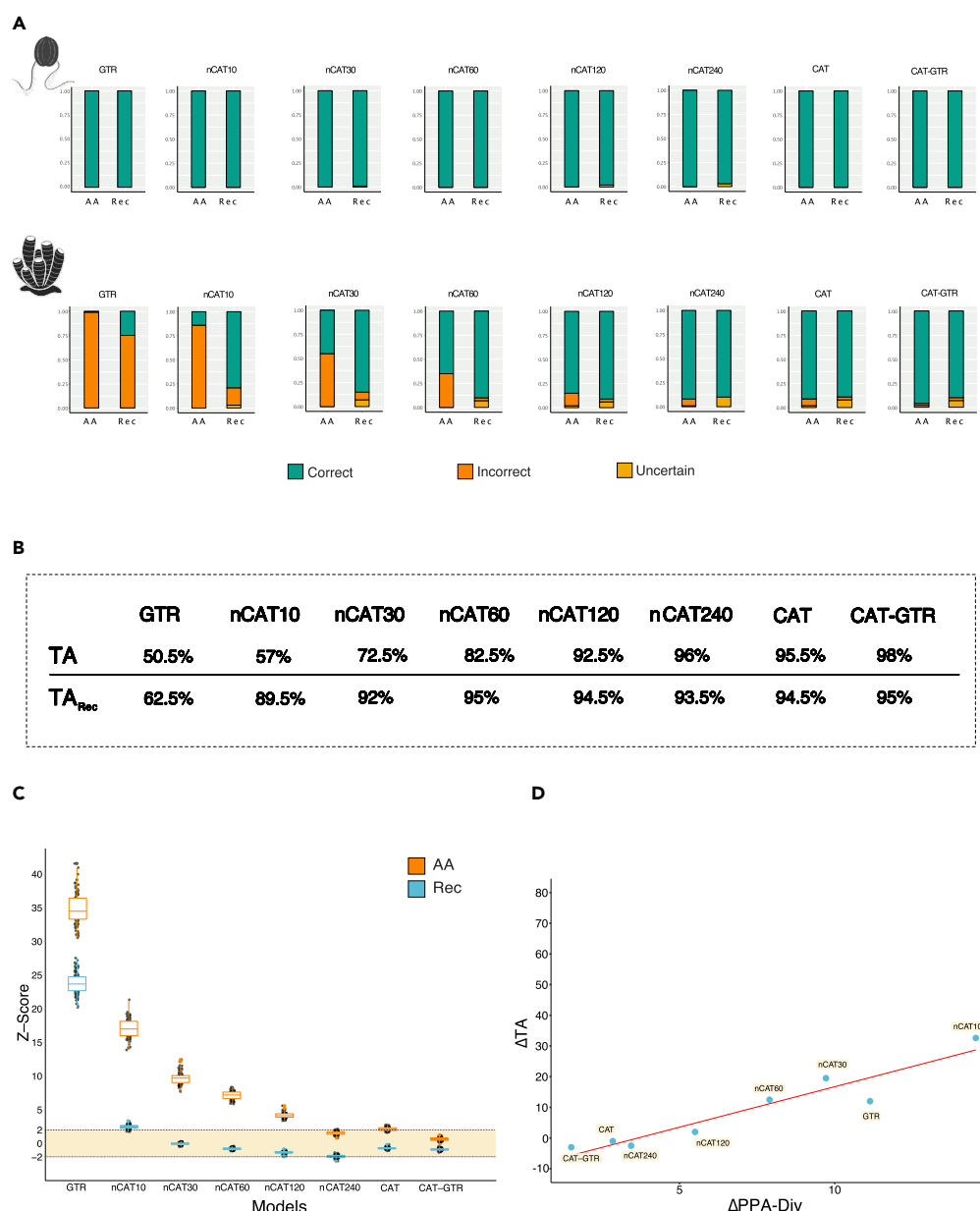


Figure 2. Recoding the data improves accuracy when the model fails to fit the amino acid alignment

(A) Accuracy of amino acids and recoded data as models that can account for more across-sites compositional heterogeneity are used.

(B) Table summarizing the Total Accuracy (TA) for amino acids and recoded data under each model. TA is calculated (from the values in A) as the percentage of accurate trees (see [STAR methods](#)) under both Porifera- and Ctenophora-sister.

(C) Change in the fit (expressed as Z-scores) of the model to the data (estimated using PPA-Div) as models that can account for more across-sites compositional heterogeneity are used. In Orange amino acid datasets; in Blue recoded datasets.

(D) Correlation between the difference in Z-scores achieved by each considered model on the amino acid and recoded datasets (δ PPA-Div), against the difference in TA achieved before and after recoding (δ TA). See [Figures S4–S7](#) for sensitivity tests showing that our conclusions would not have changed if we used Maximum Likelihood instead of Bayesian analyses, if we run our Bayesian analyses 8,000 more generations (convergence was achieved before 1000 cycles), if we used a more stringent threshold to define success (PP = 0.95 instead of PP = 0.5), and if we used alternative recoding schemes (SR6 and KGB6 see [STAR methods](#) for details) instead of Dayhoff-6.

correlation ($R^2 = 0.90$) indicating that recoding seems to improve accuracy when PPA-Div_{AA} is large (i.e., the model does not adequately describe the across-site compositional heterogeneity of the amino acid data), and PPA-Div_{rec} is small (the model seems to adequately describe the across-site compositional heterogeneity of the recoded data). Notably, δTA was negative (i.e., amino acids were more accurate than recoding – Figure 2D) for models that already adequately described the across-site compositional heterogeneity of the amino acid alignments (CAT, nCAT240, and CAT-GTR – see Figure 2C). These results are consistent with the hypothesis that recoding increased accuracy when it improved the fit of the data to the model. This behavior can be expected because, when a model has a good absolute fit to the amino acid data, there should be no model violations left for the recoding to mask. Therefore, with our simulated data, we would expect recoding to improve accuracy only for GTR, nCAT10, nCAT30, nCAT60 and nCAT120. This is because according to PPA-Div_{AA} scores these models do not fit the amino acid data well. However, lower PPA-Div scores are also expected when the size of the state space of a dataset is reduced, which is the case when 20-state amino acid data are recoded using the 6-state Dayhoff-6 scheme (Figure 1 and STAR methods). Although we address this problem in the next section using random recodings, to be conservative (given the size difference between the amino acid and recoded alphabets), we only conclude that our results are consistent with the hypothesis that recoding improved accuracy when it improved the fit of the data to the model but cannot prove it. However, the strong correlation between $\delta PPA\text{-Div}$ and δTA in Figure 2D clearly indicates that recoding the data invariably had a positive effect on the accuracy of models that did not adequately describe the across-site compositional heterogeneity of the amino acid alignments.

Recoding is expected to reduce saturation.^{11,29} We estimated saturation levels (STAR methods) for our amino acid and recoded data, finding recoded alignments to be less saturated. However, the level of saturation observed for the recoded data under nCAT30, nCAT60, nCAT120, nCAT240, CAT and CAT-GTR was statistically indistinguishable – the distributions overlapped (Figure S3). Accordingly, the reduced saturation of the recoded data cannot explain differences in TA observed for these models, suggesting that a reduction in saturation is not the main mechanism by which recoding improves accuracy.

We investigated whether our methodological assumptions might have affected our results. We tested the effect of using Maximum Likelihood instead of Bayesian analysis, the effect of increasing the number of generations in our Bayesian analyses, the use of alternative 6-state recoding strategies, and the use of a more stringent criterion (higher support) to define accuracy (see STAR methods and Figures S4–S7), finding no significant difference.

We recorded the time needed to complete analyses of amino acid and recoded data, finding that recoding improves computational efficiency and significantly reduces the time needed to complete the analysis (Figure 3). This result suggests that recoding has significant potential as a ready available tool for “green phylogenomics”.^{34,35} This is because, by speeding up analyses, recoding reduces emissions without compromising phylogenetic accuracy (Figure 2D).

Randomization experiments suggest that recoding reduce across-site compositional heterogeneity and improves the fit of the data to the model

Recoding is based on the theoretical prediction that there should be more within-bin (i.e. same amino acid category) than across-bin (i.e. different amino acid category) substitutions (STAR methods and Figures 1 and S8), and that the within-bin substitutions contribute more lineage-specific and site-specific compositional heterogeneity than across-bin substitutions (see STAR methods). Randomly generated recoding schemes^{13,29} can be defined (see STAR methods for details). Such schemes should be expected to reduce compositional heterogeneity (both across-sites and lineages) less than empirical recodings that account for biochemical and biophysical properties of amino acids such as Dayhoff-6 (Figure 1 and STAR methods). This is because random recodings will not necessarily mask substitutions between amino acids that replace frequently among each other due to their similar biochemical and biophysical properties¹³ – STAR methods and Figure S8. Accordingly, we can expect that randomly recoded datasets should fit models worse than datasets recoded using Dayhoff-6 (the recoding used in our study). Furthermore, trees inferred from randomly recoded datasets should be expected to have higher tree length (random recodings will mask less substitutions) and be less accurate than trees inferred from datasets recoded using Dayhoff-6.

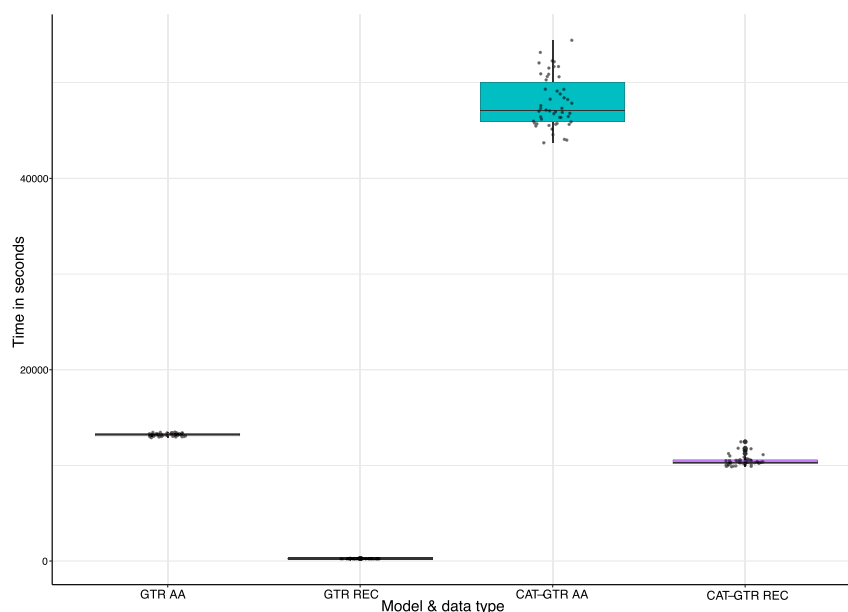


Figure 3. Recoding as a tool in green phylogenomics

Time taken to complete 50 GTR and CAT-GTR analyses of amino acid and recoded datasets. Recoded analyses are invariably completed in a shorter time, with the difference becoming significantly more marked when using the complex CAT-GTR model. Given the high accuracy of recoded analyses this result suggests that recoding could play a significant role in the development of “green phylogenomics”.

We used 200 6-state random recoding schemes based on the binning-structure of Dayhoff-6 (Figure 1 and STAR methods for details) to test whether within-bin substitutions are more abundant and contribute more compositional heterogeneity than across-bin substitutions, and to test whether Dayhoff-6 improves fit and accuracy more than random recodings. It should be noted that (I) randomly recoded datasets are still expected to infer meaningful trees, as they only mask a random subset of substitutions within an alignment, with the remaining substitutions being unchanged (Figure S8 for a visual representation of this phenomenon). (II) There is only a limited number of ways in which 20 amino acids can be allocated to six bins of the same size of those of Dayhoff-6, and random recoding schemes are therefore expected to frequently retain some random similarity to Dayhoff-6 (see STAR methods). (III) Because the size of the character state space of randomly generated recodings and Dayhoff-6 is the same, PPA-Div scores from datasets recoded using Dayhoff-6 and our 200 random recodings schemes should be comparable – differently from the case of PPA-Div scores of recoded and amino acid data (see previous section). Based on points 1 to 3 above, we can predict that if recoding works as expected, random recodings of greater residual similarity to Dayhoff-6 should mask more substitutions, achieve greater accuracy, and fit the data better than random recodings that retained less residual similarity to Dayhoff-6. We generated 100 random recoding schemes with 90% residual amino acid similarity to Dayhoff-6 (90%-recodings), and 100 random recoding schemes with no residual amino acid similarity to Dayhoff-6 (0%-recodings) – Figure S8 and STAR methods.

Trees inferred from Dayhoff-6 recoded data had an average tree length of 49,683.18 parsimony steps. Trees inferred from 90%-recodings were 72,390.12 steps long on average, and trees inferred from 0%-recodings had an average tree length of 88,046.08 steps. These results indicate that Dayhoff-6 masks ~39% and ~23% more substitutions than 0%- and 90%-recodings (Figure 4A), confirming that amino acids that are classified in the same bin by Dayhoff-6 substitute more frequently than amino acids that are placed in different bins. PPA-Div scores obtained from Dayhoff-6 recoded datasets (Average PPA-Div_{Rec} ~ 2.5; SD = 0.33) are significantly better (the distributions do not overlap) than PPA-Div scores from 0%-recoded datasets (Average PPA-Div_{0%} ~ 6; SD = 1.2), whereas PPA-Div scores from 90%-recoded datasets (Average PPA-Div_{90%} ~ 4; SD = 0.87) overlap with both PPA-Div_{Rec} and PPA-Div_{0%} (Figure 4B). Finally, the extent to which random recodings improved accuracy (TA_{0%} = 81%; TA_{90%} = 83.5%) was found to depend on their similarity to Dayhoff-6 (TA_{Rec} = 89.5%) – Figure 4C. These results show that Dayhoff-6 performs better than random

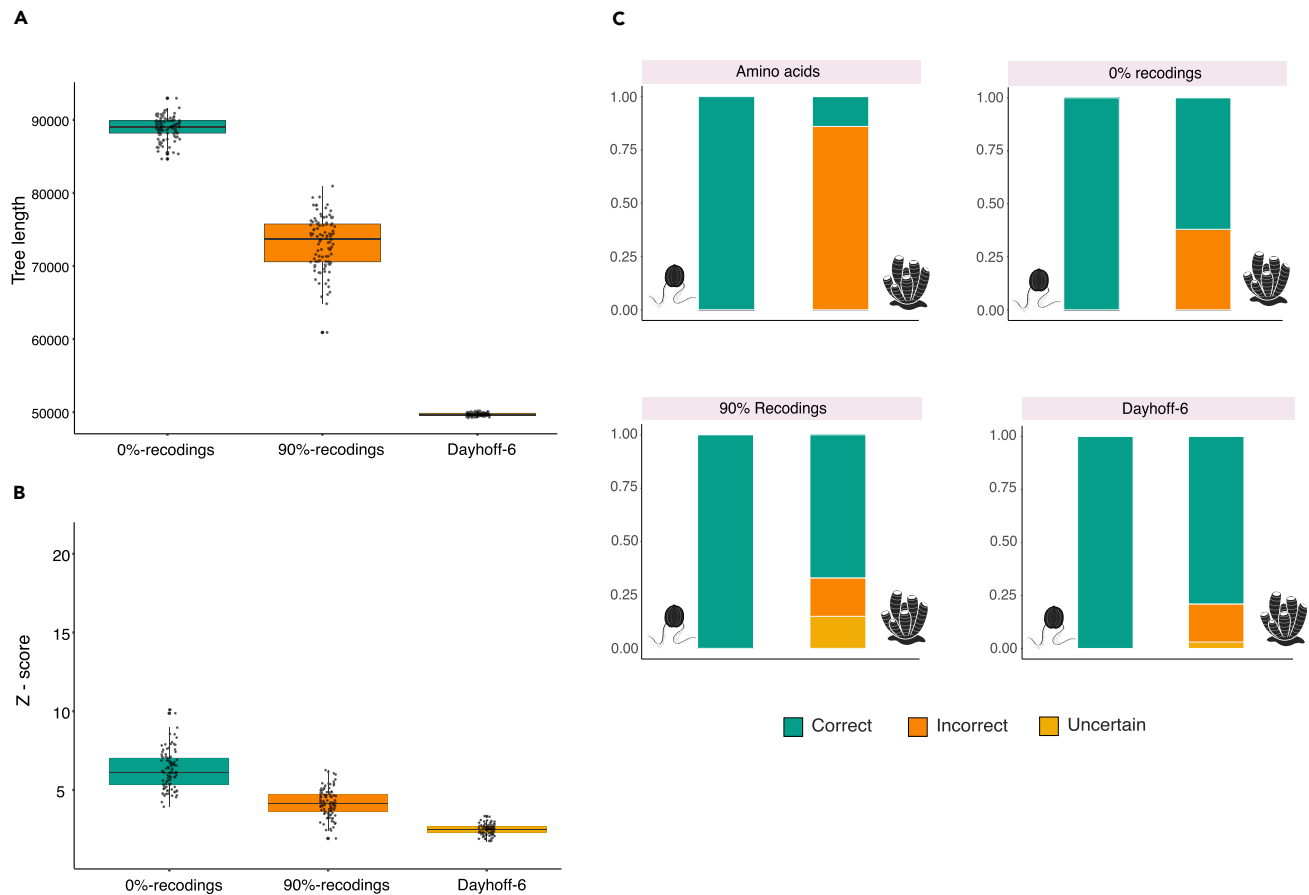


Figure 4. Dayhoff-6 outperforms random recoding schemes

(A) Boxplot representing the distribution of tree lengths for 0%-recoded, 90%-recoded, and Dayhoff-6 recoded dataset.

(B) Comparison of PPA-Div scores for 0%-recoded, 90%-recoded, and Dayhoff-6 recoded datasets. The figure indicates that PPA-Div scores of Dayhoff-6 recoded data are significantly better than PPA-Div scores from 0%-recoded data – the distributions do not overlap.

(C) A comparison of the TA values achieved by amino acids data, 0%-recoded, 90%-recoded and Dayhoff-6 recoded data, indicating that Dayhoff-6 outperforms both amino acids and randomly recoded data.

recodings and that Dayhoff-6 recoded data have a better fit to the model than randomly recoded data. These results suggest that Dayhoff-6 helps improving the fit of the data to the model.

The performance of recoding improves as more sites and taxa are used

Our random recoding experiments reject the view¹³ that recoding merely strip information from the data as this would be incompatible with the results in Figure 4. However, every recoding strategy, including the widely accepted recoding of nucleotides into amino acids, incurs a cost in the form of an information loss⁴¹ (STAR methods). We estimated (STAR methods) that the information lost from our simulated alignments when the data were recoded from amino acids to Dayhoff-6 was ~49%, and that this loss was homogeneously distributed across the branches of the target topology (see Figure S9). Accordingly, we hypothesized that alignment length can be expected to have an impact on the performance of recoded datasets, as longer alignments will have more substitutions (everything else being equal) and should therefore be expected to be more informative. To test this hypothesis, we analyzed alignments with 20 taxa and 1,000, 5,000, 10,000 and 30,000 sites. We found that with 1,000 site alignments, amino acids performed best, but in a context where both data types had very poor accuracy ($TA_{AA-1000} = 26\%$ and $TA_{Rec-1000} = 11\%$, Figure 5). With 5,000 site alignments, accuracy improved for both data types with amino acids doing marginally better ($TA_{AA-5000} = 59.5\%$ and $TA_{Rec-5000} = 58.5\%$, Figure 5). These results agree with what was observed in previous studies.^{31,32} Finally, with alignments of 10,000 and 30,000 sites, we found recoded data to outperform amino acids ($TA_{AA-10,000} = 63\%$,

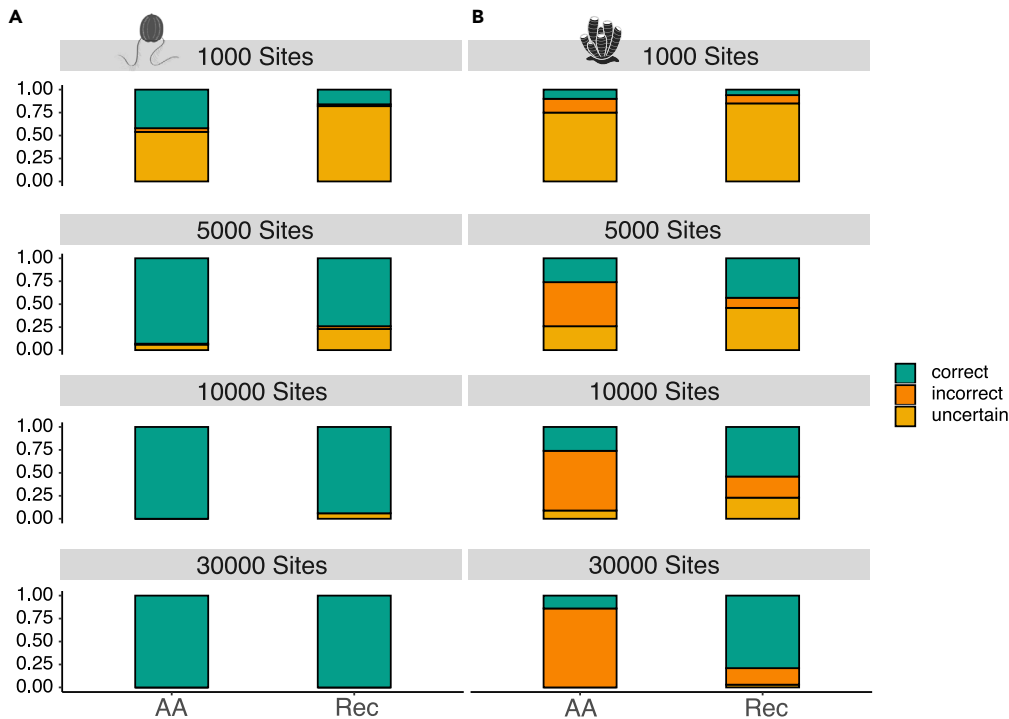


Figure 5. Accuracy of recoded data increases with alignment size

(A) Accuracy of amino acid and recoded datasets of 1,000, 5,000, 10,000 and 30,000 sites analyzed under nCAT10, when the generating tree assumes Ctenophora-sister to be true.

(B) Success rate of amino acid and Dayhoff-6 datasets of 1,000, 5,000, 10,000 and 30,000 sites analyzed under nCAT10, when the generating tree assumes Porifera-sister to be true. Analyses performed in Phylobayes. In Green: Correct trees; Dark Orange: Incorrect Trees; Light Orange: Uncertain trees.

$TA_{Rec-10,000} = 74\%$, and $TA_{AA-30,000} = 57\%$, $TA_{Rec-30,000} = 89.5\%$; Figure 5), consistent with the results of Foster et al.³² With reference to taxon sampling (Figure S10), we found that accuracy increases as more taxa are included in the dataset, as it is common when the data are affected by attraction artifacts.^{42,43}

Recoding of empirical datasets can help resolve tricky nodes

The target trees used in our simulations represent the phylogenetic uncertainty at the root of the animal tree (Ctenophora-sister versus Porifera-sister hypothesis). The alignments we used were simulated under a CAT-LG + G model (hereafter CAT-LG) that was parameterized using values estimated from a dataset designed to study early animal relationships.¹⁸ We thus applied the knowledge gained from our simulations to three, different (i.e. independently assembled) datasets that were also designed to address the same problem^{5,19,20} – hereafter referred to as Whelan2015, Whelan2017, and Laumer2019. To make sure results from the simulated and the empirical data were comparable, we evaluated whether the across-site compositional heterogeneity of the three empirical datasets could have been generated by a process that could be described using a model that is not too dissimilar to CAT-LG. To do so we used PPA-Div_{AA} to test how well CAT-LG can describe the across-site compositional heterogeneity of Whelan2015, Whelan2017, and Laumer2019. PPA-Div_{AA} indicates that CAT-LG adequately describes the across-site compositional heterogeneity of Laumer2019 (PPA-Div_{Laumer2019} = 1.55), but not that of Whelan2015 and Whelan2017 (PPA-Div_{Whelan2017} = 4.69; PPA-Div_{Whelan2015} = 4.64). However, PPA-Div_{AA} scores for Whelan2015 and Whelan2017 are rather small (compare with those in Figure 2C) suggesting that models that are not too different from CAT-LG should be able to adequately describe Whelan2015 and Whelan2017.

We then analyzed Laumer2019, Whelan2015, and Whelan2017 as amino acid alignments, and using 0%-recodings, 90%-recodings, and Dayhoff-6. For all analyses the nCAT60 model was used. This was done to maintain comparability between our results and those of Li et al.¹³, who conjectured that recoding only strip signal

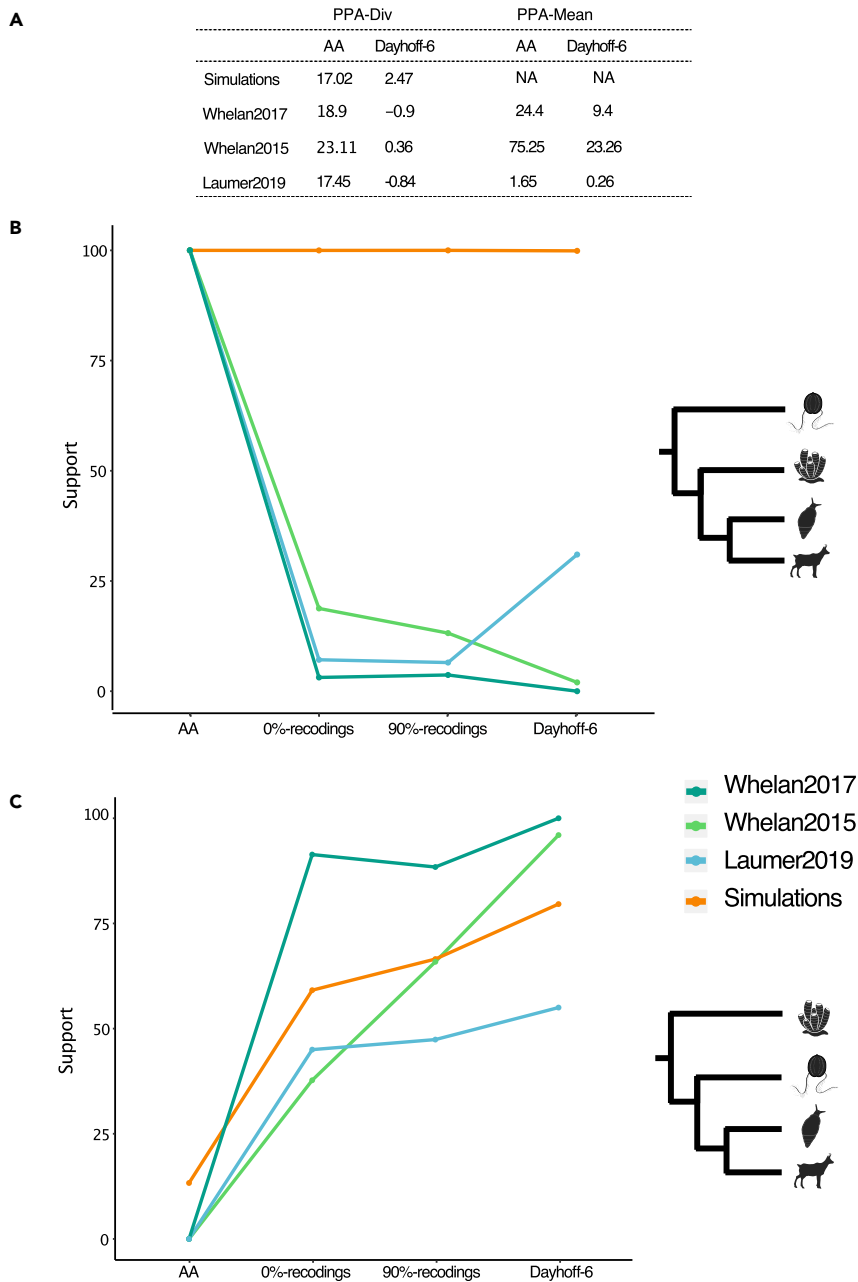


Figure 6. Results from empirical datasets follow predictions from simulations

(A) PPA-Div and PPA-Mean scores for all three empirical datasets when the alignments are analyzed as amino acids and Dayhoff-6 recoded data. Average PPA scores for our simulated data are also reported (under nCAT10), indicating that PPA scores achieved for the simulated data under nCAT10 are comparable to those achieved under nCAT60 for the empirical datasets. (B) Support for Ctenophora-sister as different data types (amino acids, 0%-recodings, 90%-recodings, and Dayhoff-6) are used. (C) Support for Porifera-sister as different data types (amino acids, 0%-recodings, 90%-recodings, and Dayhoff-6) are used. In Orange: reference support values obtained for the two considered clades in our simulations. Light Green: Whelan2015; Dark Green Whelan2017; Blue Laumer2019. Note: Support values for 0%-recodings and 90%-recodings represent average values calculated over all random recoding generated.

from the data. For all datasets we estimated PPA_{AA} and PPA_{Rec} , as well as support values for both Porifera- and Ctenophora-sister for all four data types. We used both PPA-Div and PPA-Mean (STAR methods) to test whether nCAT60 adequately describes the heterogeneity of these datasets, as empirical data usually display compositional heterogeneity across both sites and taxa. $PPA-Div_{AA}$ (Figure 6A) indicates that the across-site

compositional heterogeneity of the three amino acid datasets is described very poorly by nCAT60. PPA-Mean_{AA} (Figure 6A) indicates that across-lineage compositional heterogeneity is unlikely to be a problem for Laumer2019, but nCAT60 describes very poorly the across-lineage compositional heterogeneity of Whelan2015 and Whelan2017. PPA-Div_{Rec} and PPA-Mean_{Rec} indicate that nCAT60 adequately describes the across-site compositional heterogeneity of the three recoded datasets and that also when using Dayhoff-6 across-lineage compositional heterogeneity should not be a problem for Laumer2019. For Whelan2015 and Whelan2017, PPA-Mean_{Rec} indicates that recoding reduced across-lineage compositional heterogeneity, but the fit was still poor. nCAT60 is not an across-lineage compositional homogeneous model. Accordingly, its inability to model the across-lineage compositional heterogeneity of the recoded versions of Whelan2015 and Whelan2017 was expected, given the high PPA-Mean_{AA} scores achieved by these datasets (Figure 6A). However, the across-lineage compositional heterogeneity of these datasets is still much better described on the recoded data.

For all three datasets support for Ctenophora-sister was maximal (PP = 1) when the data were analyzed as amino acids (Figure 6B). In the cases of Whelan2015 and Whelan2017 support for Ctenophora-sister progressively decreased, becoming minimal (PP_{Whelan2015} = 0.02 and PP_{Whelan2017} = 0) when the data are Dayhoff-6-recoded. In the case of Laumer2019 minimal support for Ctenophora-sister was achieved using 90%-recodings (PP = 0.06). Support for this clade increased to PP = 0.31 under Dayhoff-6, but PP = 0.31 is still within the range of what is expected for random resolutions. For all three datasets, support for Porifera-sister followed the opposite trend, being minimal (PP = 0) for amino acid datasets, and maximal (PP_{Laumer2019} = 0.55; PP_{Whelan2015} = 0.96; PP_{Whelan2017} = 1) under Dayhoff-6 (Figure 6C). The lower discriminatory power of the recoded version of Laumer2019 (Ctenophora-sister PP = 0.31 and Porifera-sister PP = 0.55) may be a consequence of this dataset being less informative than Whelan2017 and Whelan2015. Laumer2019 had an average of ~5 substitutions per site (see STAR methods) while Whelan2015 and Whelan2017 had an average of ~7 substitutions per site, and recoding can be expected to work best with more informative datasets that better absorb the loss of information caused by its application (see above). Irrespective of that, the trend observed for the changes in support levels for Ctenophora-sister and Porifera-sister, are comparable across the three datasets. We plotted (Figures 6B and 6C) average support values observed for our simulated data under nCAT10 (the model that in simulations achieved PPA-Div_{AA} scores comparable to those obtained using nCAT60 with the empirical data). For all three datasets, changes in support values for Ctenophora-sister strongly disagree with what is observed in simulations where Ctenophora-sister is true. On the other hand, changes in support levels for Porifera-sister follow the pattern observed in simulations where Porifera-sister was true.

Our simulations suggest that when the model used does not adequately describe the compositional heterogeneity of the amino acid data (as in the case of the three considered empirical datasets), recoding improves phylogenetic accuracy. Li et al.¹³ stated that if recoding was a valid approach to improve phylogenetic accuracy, it should be expected that if a clade is true it should be recovered in analyses using an empirical recoding such as Dayhoff-6, but not in analyses using random recodings. Our random recoding simulations show that this conjecture is only partially correct. It is correct that Dayhoff-6 outperforms random recodings, but random recodings still improve phylogenetic accuracy, a result that is not novel,²⁹ and the performance of random recodings depends on their similarity to Dayhoff-6 (see above), a result that is not unexpected. Accordingly, we reformulate Li et al.¹³ conjecture to state that: within the limit of utility of recoding methods (e.g., when the model does not describe the compositional heterogeneity of the amino acid data well – see above), true clades should be expected to achieve progressively better support as random recodings more similar to Dayhoff-6 are used. For artifactual clades, the opposite trend should be expected. Analyses of our three empirical datasets consistently find that support values for Porifera-sister follow the trend expected for true clades while support values for Ctenophora-sister follow the trend expected for artifactual clades.

DISCUSSION

Recoding is justified

The use of recoding has been justified based on its predicted ability to mask substitutions that are *a-priori* expected to contribute more compositional heterogeneity and saturation – i.e., within-bin substitutions.¹¹ Using our simulated data and random recoding we could confirm that: (I) within-bin substitutions are more frequent than across-bin substitutions, as Dayhoff-6 removes more substitutions than random recodings. (II) Recoding seems effective at improving fit and reducing across-site compositional heterogeneity as PPA-Div scores of alignments recoded using Dayhoff-6 are better than those obtained using random

recodings. (III) Recoding improves accuracy when there is a sizable difference in Z-score between the recoded and the amino acid data (Figures 2C and 2D). Although this difference in Z-scores can be partially attributed to the smaller size of the recoded alphabet, random recoding experiments suggest that as we mask within-bin substitutions (and recoding amino acid data masks within-bin substitutions – Figures S8 and S9), across-site compositional heterogeneity decreases while fit and accuracy increase (Figure 4). We therefore suggest that recoding can improve phylogenetic analyses in two ways. First, by making the parameter space smaller, recoding makes model parametrization easier and computationally more efficient, given the unmasked data. Second, by simplifying the data (masking sites with a high proportion of multiple substitutions), recoding improves the ability of the model to describe the compositional heterogeneity of the data. That is, recoding can improve the fit of the data to the model, which is crucial to improve accuracy.^{40,44}

A previous investigation attempted comparing recoding and random recoding schemes and concluded that recoding did not perform better than random.¹³ However, that study used only four random recoding schemes, and may have lacked statistical power. Our analyses used 200 random recoding schemes of different similarity to Dayhoff-6 (0%- and 90%-recodings), and robustly establishes that both real and to a lesser extent random recodings can improve accuracy, as even 0%-recodings outperform amino acids ($TA_{0\%} = 81\%$; $TA_{AA\sim} = 57\%$ – compare Figures 2 and 4) in our simulations. This result might seem counterintuitive, and its full interpretation is outside of the scope of this study. However, it should not surprise that 0%-recodings infer valid trees, given that random recoding schemes do not randomize the data, they randomly mask a subset of substitutions, with the unmasked substitutions being shared with the amino acid data (Figure S8). We hypothesize that the improved accuracy observed with 0%-recodings might be a consequence of the fact that a 6-state model is more easily parametrized (given the recoded data) than a 20-state model (given the amino acid data), a testable hypothesis.

Recoding improves accuracy when the alignment is sufficiently long

Recoding incurs a cost in the form of an information loss (which for our simulated datasets was approximately 49%, Figure S9 and STAR methods), and our results indicate that the use of recoding should be limited to phylogenomic-scale datasets that can absorb this cost. Hernandez and Ryan³¹ investigated this problem using 1,000 to 5,000 site alignments and concluded that the use of 6-state recoding schemes (such as Dayhoff-6) cannot improve accuracy, despite finding that recodings with more states had a positive effect on accuracy. When using datasets of 1,000 to 5,000 sites we found the same result: Dayhoff-6 does not improve accuracy, indicating that there is a limit to the use of 6-state recodings that might not be useful in analyses of single genes or of small genomes (e.g. viral genomes). However, when we used phylogenomic-scale alignments (10,000 sites or more) recoding significantly improved accuracy (Figure 5), when the model used to analyze the data was not a good fit to the amino acid alignments (Figure 2D). We interpret these results to suggest that, with phylogenomic scale datasets, when the absolute fit of the model to the amino acid data is poor, the trade-off between information loss and improved fit favors recoding. Differently, when the model already adequately describes the across-site compositional heterogeneity of the amino acid data, there will be no model violations left for the recoding to ameliorate. However, some information will still be lost when the data are recoded, tilting the trade-off balance against the use of recoding. We suggest that the mild deterioration of accuracy observed for nCAT240, CAT and CAT-GTR, and the conclusion of Foster et al.³² that recoding does not always increase accuracy, explained within the context of this trade-off.

Because recoding does not always improve accuracy (see³² and above), a criterion is necessary to guide its application. The correlation in Figure 2D suggests that the comparison of PPA_{AA} and PPA_{Rec} scores provide valuable information to decide whether the use of recoding is appropriate, even if these values do not purely record changes in fit between amino acid and recoded alignments, because they are summary statistics computed on data types with different state spaces.

Simulated datasets are abstractions of empirical data. They can be limited in the number of sites³¹ or taxa,³² and their heterogeneity will not perfectly match that of empirical data. For example, our simulated data are only across-site compositionally heterogeneous, whereas other studies used alignments that were only across-lineage compositionally heterogeneous.^{29,31} Based on our simulations one could erroneously conclude that recoding is redundant when using relatively well fitting models such as nCAT60, CAT and

CAT-GTR. However, absolute goodness-of-fit tests^{40,44} show (see [Figures 6](#) and ¹¹) that even models such as CAT-GTR fail to adequately describe the compositional heterogeneity of empirical datasets designed to address challenging nodes in the Tree of Life. Accordingly, it can be expected that with empirical data there should be substantial scope for recoding to improve accuracy also when the best available site-heterogeneous models are used.

Is Ctenophora-sister a tree reconstruction artifact?

Our simulations suggest that when the alignment is sufficiently long and the model used to analyze the data fails to adequately describe the compositional heterogeneity of the amino acid alignment, recoding should improve accuracy. Based on these conclusions, we moved on to exemplify the use of recoding, and tested three empirical datasets for the root of the animal tree. We chose these datasets because they are published and their taxon sampling compares well with that of the dataset used to parametrize our simulated data. It has been suggested¹³ that if recoding was a valid approach to improve phylogenetic accuracy (and according to our simulations it is), when amino acid and Dayhoff-6 datasets disagree, true clades should be recovered in analyses of Dayhoff-6 recoded dataset, but not in analyses of randomly recoded datasets. Our simulations found that the suggestion of Li et al.,¹³ as formulated, is not valid. Dayhoff-6 recoded datasets achieve greater accuracy than randomly recoded datasets, but, randomly recoded datasets perform better than amino acid datasets, with their performance depending on their residual amino acid similarity to Dayhoff-6 (see above for details). Accordingly, Li et al.'s¹³ conjecture should be rephrased to say that, in cases when recoding can be expected to improve accuracy (and [Figure 6A](#) suggests this should be the case for our empirical datasets), support for true clades should increase as the data are recoded using random recoding schemes that are progressively more similar to Dayhoff-6. For artifactual clades, the opposite trend should be expected. Results from our three empirical datasets consistently show that changes in support for Porifera-sister follow the trend expected for real clades, while support values for Ctenophora-sister follow the trend expected for artifactual clades.

The debate on the root of the animal tree has seen a diversity of analyses and data. Sponges as the sister of all the other animals is favored by morphology and gene content data.^{45–47} Amino acid datasets analyzed using CAT and CAT-GTR tend to switch between Porifera- and Ctenophora-sister¹⁶ depending on outgroup sampling. In these experiments, Porifera-sister is frequently (but not always) supported when close outgroups (not only Choanoflagellata¹⁶), which should reduce attraction artifacts,^{11,16,18–20,48} are used. A common feature of investigations of early animal relationships is that the models used to analyze the data usually fail to a greater (GTR) or lesser (nCAT60/CAT/CAT-GTR) extent to adequately describe the compositional heterogeneity of the relevant alignments (see¹¹ and [Figure 6](#)). When datasets are recoded, CAT-based models (but not GTR) achieve a much better fit (see,¹¹ [Figures 2](#) and [6](#)). These are the conditions when we can expect recoded data to improve accuracy, and CAT-based analyses of recoded data almost invariably reject Ctenophora-sister – see⁴⁹ for an exception.

It has been claimed that analyses of sequence data supporting Porifera-sister represent either CAT-induced overparameterization artifacts, or recoding-induced signal-loss artifacts.¹³ Overparameterization in Bayesian analysis (including analyses that use CAT) has been studied,^{50–53} but no evidence was found that it can cause the emergence of tree reconstruction artifacts. Similarly, our simulations did not find any evidence that recoding can induce the emergence of tree reconstruction artifacts. Differently, recent analyses seem to suggest that the artifactual inference of Ctenophora-sister can also be caused by incomplete lineage sorting.⁵⁴

It has been argued that we should be skeptical of Porifera-sister in analyses of sequence data because it is found only under a narrow range of models that do not fit the data better than models supporting Ctenophora-sister.¹³ However, the size of the model space supporting Ctenophora-sister is not as broad as claimed because models such as WAG, LG, JTT,^{55–57} etc., are just alternatively parameterized GTR matrices. If one were to consider (and explicitly name) alternative parameterizations of CAT, the range of models under which Porifera-sister is supported would similarly expand. Furthermore, none of these alternatively parameterized GTR matrices fits as well as CAT or CAT-GTR, both of which frequently support Porifera-sister,^{16,58} the support for which is not limited to using CAT – contrary to what implied by Li et al.¹³ Although it has been shown that nCAT60 (another across-site compositionally heterogeneous model) could not be distinguished in some Bayesian cross-validations from CAT,¹³ that does not imply that CAT is overparameterized and that nCAT60 should be preferred.⁵⁹ CAT uses more frequency categories but the advantage of Bayesian methods^{23,52} is exactly that they allow the complexity of the model to be tuned

to the dataset being analyzed²³ and more adequately describe it. Modern model selection procedures should not be limited to the identification of the relative best fit model(s).⁴⁰ Absolute goodness-of-fit tests (such as PPA) should also be used to discriminate models, including models of relative best fit.^{40,44} Absolute goodness-of-fit tests indicate that irrespective of its status as a relative co-best fit model, nCAT60 describes the across-site compositional heterogeneity of Whelan2017 ($PPA-Div_{nCAT60} = 18.9$) very poorly, and much worse than CAT ($PPA-Div_{CAT} = 4.13$). This result is consistent with those of Bujaki and Rodrigue,⁵⁹ and indicates that contrary to Li et al.¹³, CAT should be preferred to nCAT60.

The solution to the problem of the root of the animal tree is not going to be found by counting the number of models (or studies) supporting alternative trees as suggested by Li et al.¹³ To understand phylogenetic relationships at the root of the animal tree, we must understand why, under the relative best fit model (be that CAT, CAT-GTR, or nCAT60), support switches from Ctenophora– to Porifera–sister as factors such as data type⁶⁰ and out-group sampling change. One hypothesis is that even current compositionally heterogeneous models struggle to adequately describe these datasets.^{11,16} In such situations, recoding can improve the fit of the data to the model, which is crucial to improve accuracy, and the analysis of recoded alignments under CAT, CAT-GTR, and nCAT60 almost invariably reject Ctenophora-sister independently of outgroup choice.¹¹ We interpret these results to suggest that a placement of the ctenophores as the sister of all the other animals, at the least using sequence data, continues to be better explained as a tree reconstruction artifact.

Conclusions

Phylogenetic accuracy of sequence data crucially depends on the fit of the model to the data.²⁵ Our results suggest that recoding seems to improve the fit of models that fail to adequately describe amino acid alignments, increasing accuracy in simulations. Empirical datasets assembled to resolve tricky nodes in the Tree of Life usually display high levels of compositional heterogeneity, and available models fail to fit them.¹¹ Although the development of models that could adequately fit arbitrary levels of across-site and across-lineage compositional heterogeneity^{48,61} would be welcome, efficient implementations of these models are not yet on the horizon. In the absence of such models, the analysis of recoded data remains a useful tool to study tricky nodes in the Tree of Life. At the least according to our simulations, recoding also has the advantage of drastically reducing computational time, particularly when complex models such as CAT-GTR are used, without compromising accuracy. Accordingly, recoding should be considered as a readily available tool for “green phylogenomics”.

It is not yet clear whether sponges or ctenophores represent the sister of all the other animals. Every data type has its own strengths and weaknesses, and a single line of evidence cannot resolve this controversy. Only concision of multiple lines of evidence,^{4,47,62–64} e.g., sequence data, gene content data, morphology and newly emerging sources of genomic information such as chromosome tectonic,⁶⁵ will ultimately clarify this problem. However, sequence data are the genomic data type that we better understand and for which we have better established models, and arguments based on the fit of amino acid and recoded data suggest that, according to sequence data, a placement of the ctenophores as the sister of all the other animals continue to be better explained as a tree reconstruction artifact.

Limitations of the study

The effect of using recodings strategies with a number of bins different from 6 was not investigated. Our choice of using 6-state recodings was based on the fact that 6-state recodings are much more frequently used in the literature than other recoding schemes. However, we acknowledge that recodings with 2–19 states could be designed and should ideally be tested, as other recodings might be discovered to perform better than 6-state ones. However, even if a recoding scheme with different bin numbers could be shown to outperform the 6-state recodings used here, it will not change the fact that 6-state recoding, in our simulations, performed well. Indeed, we do not expect that our conclusions would change if such recoding strategies were used. Although our simulations were extensive, we only compared the results of our simulations to three empirical datasets that focused on the root of the animal tree. While we acknowledge that further tests on real data should be completed to further generalize our conclusions, at this time we have no reason to believe that our conclusions about the root of the animal tree would change if other empirical datasets were analysed.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- **KEY RESOURCES TABLE**
- **RESOURCE AVAILABILITY**
 - Lead contact
 - Materials availability
 - Data and code availability
- **METHODS DETAILS**
 - Recoding strategies: A description and a logical rationale for their generalisation
 - Datasets
 - Analyses of the simulated datasets
 - Measuring phylogenetic accuracy
 - Measuring model adequacy
 - Measuring saturation
 - Random recodings
 - Estimating the number of substitutions entailed by amino acid and recoded data and approximating information content
 - Empirical datasets
 - Green phylogenomics

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at <https://doi.org/10.1016/j.isci.2022.105594>.

ACKNOWLEDGMENTS

We would like to thank Hervé Philippe, Gert Wörheide, Nicholas Lartillot, Paschalia Kapli, Max Telford, Tom Williams, Peter Foster and two anonymous reviewers for the helpful suggestions. We would also like to thank all the members of the Pisani, Donoghue, Paps and Williams labs (at the University of Bristol), for discussing these ideas. MG and DP were supported by a Marie Skłodowska-Curie grant agreement (764840). MER is supported by a NERC GW4+ PhD studentship. RF is supported by Royal Society University Research Fellowship (UF160226). This work was also supported by a NERC BETR grant (NE/P013678/1), a Gordon and Betty Moore Foundation grant GBMF9741, (<https://doi.org/10.37807/GBMF9741>), and a John Templeton Foundation grant (Grant 62220). The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the John Templeton Foundation. There are no data associated with this article.

AUTHOR CONTRIBUTIONS

Conceptualization: D.P. and M.G.; Methodology: M.G. and D.P.; Data curation: M.G. and D.P.; Writing-original draft: M.G., D.P., and M.E.R.; Writing – review and editing: M.G., D.P., M.E.R., J.L.F., and R.F.; Supervision: D.P.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: December 17, 2021

Revised: September 5, 2022

Accepted: November 11, 2022

Published: December 22, 2022

REFERENCES

1. Metzker, M.L. (2010). Sequencing technologies — the next generation. *Nat. Rev. Genet.* 11, 31–46. <https://doi.org/10.1038/nrg2626>.
2. Dunn, C.W., Giribet, G., Edgecombe, G.D., and Hejnol, A. (2014). Animal phylogeny and its evolutionary implications. *Annu. Rev. Ecol. Evol. Syst.* 45, 371–395. <https://doi.org/10.1146/annurev-ecolsys-120213-091627>.
3. Kapli, P., Yang, Z., and Telford, M.J. (2020). Phylogenetic tree building in the genomic age. *Nat. Rev. Genet.* 21, 428–444. <https://doi.org/10.1038/s41576-020-0233-0>.
4. Tihelka, E., Cai, C., Giacomelli, M., Lozano-Fernandez, J., Rota-Stabelli, O., Huang, D., Engel, M.S., Donoghue, P.C.J., and Pisani, D. (2021). The evolution of insect biodiversity. *Curr. Biol.* 31, R1299–R1311. <https://doi.org/10.1016/j.cub.2021.08.057>.
5. Laumer, C.E., Fernández, R., Lemer, S., Combosch, D., Kocot, K.M., Riesgo, A., Andrade, S.C.S., Sterrer, W., Sørensen, M.V., and Giribet, G. (2019). Revisiting metazoan phylogeny with genomic sampling of all phyla. *Proc. Biol. Sci.* 286, 20190831. <https://doi.org/10.1098/rspb.2019.0831>.

6. Puttick, M.N., Morris, J.L., Williams, T.A., Cox, C.J., Edwards, D., Kenrick, P., Pressel, S., Wellman, C.H., Schneider, H., Pisani, D., and Donoghue, P.C.J. (2018). The interrelationships of land plants and the nature of the ancestral embryophyte. *Curr. Biol.* 28, 733–745.e2. <https://doi.org/10.1016/j.cub.2018.01.063>.
7. Williams, T.A., Szöllösi, G.J., Spang, A., Foster, P.G., Heaps, S.E., Boussau, B., Ettema, T.J.G., and Embley, T.M. (2017). Integrative modeling of gene and genome evolution roots the archaeal tree of life. *Proc. Natl. Acad. Sci. USA* 114, E4602–E4611.
8. Williams, T.A., Cox, C.J., Foster, P.G., Szöllösi, G.J., and Embley, T.M. (2020). Phylogenomics provides robust support for a two-domains tree of life. *Nat. Ecol. Evol.* 4, 138–147. <https://doi.org/10.1038/s41559-019-1040-x>.
9. Derelle, R., Torruella, G., Klimeš, V., Brinkmann, H., Kim, E., Vlček, Č., Lang, B.F., and Eliáš, M. (2015). Bacterial proteins pinpoint a single eukaryotic root. *Proc. Natl. Acad. Sci. USA* 112, E693–E699. <https://doi.org/10.1073/pnas.1420657112>.
10. Dunn, C.W., Hejnal, A., Matus, D.Q., Pang, K., Browne, W.E., Smith, S.A., Seaver, E., Rouse, G.W., Obst, M., Edgecombe, G.D., et al. (2008). Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature* 452, 745–749. <https://doi.org/10.1038/nature06614>.
11. Feuda, R., Dohrmann, M., Pett, W., Philippe, H., Rota-Stabelli, O., Lartillot, N., Wörheide, G., and Pisani, D. (2017). Improved modeling of compositional heterogeneity supports sponges as sister to all other animals. *Curr. Biol.* 27, 3864–3870.e4. <https://doi.org/10.1016/j.cub.2017.11.008>.
12. Kapli, P., and Telford, M.J. (2020). Topology-dependent asymmetry in systematic errors affects phylogenetic placement of Ctenophora and Xenacoelomorpha. *Sci. Adv.* 6, eabc5162. <https://doi.org/10.1126/sciadv.abc5162>.
13. Li, Y., Shen, X.-X., Evans, B., Dunn, C.W., and Rokas, A. (2021). Rooting the animal tree of life. *Mol. Biol. Evol.* 38, 4322–4333. <https://doi.org/10.1093/molbev/msab170>.
14. Nosenko, T., Schreiber, F., Adamska, M., Adamski, M., Eitel, M., Hammel, J., Maldonado, M., Müller, W.E.G., Nickel, M., Schierwater, B., et al. (2013). Deep metazoan phylogeny: when different genes tell different stories. *Mol. Phylogenet. Evol.* 67, 223–233. <https://doi.org/10.1016/j.ympev.2013.01.010>.
15. Philippe, H., Derelle, R., Lopez, P., Pick, K., Borchellini, C., Boury-Esnault, N., Vacelet, J., Renard, E., Houlston, E., Quéinnec, E., et al. (2009). Phylogenomics revives traditional views on deep animal relationships. *Curr. Biol.* 19, 706–712. <https://doi.org/10.1016/j.cub.2009.02.052>.
16. Pisani, D., Pett, W., Dohrmann, M., Feuda, R., Rota-Stabelli, O., Philippe, H., Lartillot, N., and Wörheide, G. (2015). Genomic data do not support comb jellies as the sister group to all other animals. *Proc. Natl. Acad. Sci. USA* 112, 15402–15407. <https://doi.org/10.1073/pnas.1518127112>.
17. Ryan, J.F., Pang, K., Schnitzler, C.E., Nguyen, A.-D., Moreland, R.T., Simmons, D.K., Koch, B.J., Francis, W.R., Havlak, P., et al.; NISC Comparative Sequencing Program (2013). The genome of the ctenophore *mnemiopsis leidyi* and its implications for cell type evolution. *Science* 342, 1242592. <https://doi.org/10.1126/science.1242592>. SA, Smith.
18. Simion, P., Philippe, H., Baurain, D., Jager, M., Richter, D.J., Di Franco, A., Roure, B., Satoh, N., Quéinnec, É., Ereskovsky, A., et al. (2017). A large and consistent phylogenomic dataset supports sponges as the sister group to all other animals. *Curr. Biol.* 27, 958–967. <https://doi.org/10.1016/j.cub.2017.02.031>.
19. Whelan, N.V., Kocot, K.M., Moroz, T.P., Mukherjee, K., Williams, P., Paulay, G., Moroz, L.L., and Halanych, K.M. (2017). Ctenophore relationships and their placement as the sister group to all other animals. *Nat. Ecol. Evol.* 1, 1737–1746. <https://doi.org/10.1038/s41559-017-0331-3>.
20. Whelan, N.V., Kocot, K.M., Moroz, L.L., and Halanych, K.M. (2015). Error, signal, and the placement of Ctenophora sister to all other animals. *Proc. Natl. Acad. Sci. USA* 112, 5773–5778. <https://doi.org/10.1073/pnas.1503453112>.
21. Lenton, T.M., Boyle, R.A., Poulton, S.W., Shields-Zhou, G.A., and Butterfield, N.J. (2014). Co-evolution of eukaryotes and ocean oxygenation in the Neoproterozoic era. *Nat. Geosci.* 7, 257–265. <https://doi.org/10.1038/ngeo2108>.
22. Erwin, D.H. (2020). The origin of animal body plans: a view from fossil evidence and the regulatory genome. *Development* 147, dev182899. <https://doi.org/10.1242/dev.182899>.
23. Lartillot, N., and Philippe, H. (2004). A bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol.* 21, 1095–1109. <https://doi.org/10.1093/molbev/msh112>.
24. Foster, P.G. (2004). Modeling compositional heterogeneity. *Syst. Biol.* 53, 485–495. <https://doi.org/10.1080/10635150490445779>.
25. Williams, T.A., Schrempf, D., Szöllösi, G.J., Cox, C.J., Foster, P.G., and Embley, T.M. (2021). Inferring the deep past from molecular data. *Genome Biol. Evol.* 13, evab067. <https://doi.org/10.1093/gbe/evab067>.
26. Embley, T.M., van der Giezen, M., Horner, D.S., Dyal, P.L., and Foster, P. (2003). Mitochondria and hydrogenosomes are two forms of the same fundamental organelle. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 358, 191–201. <https://doi.org/10.1098/rstb.2002.1190>.
27. Hrdy, I., Hirt, R.P., Dolezal, P., Bardonová, L., Foster, P.G., Tachezy, J., and Embley, T.M. (2004). Trichomonas hydrogenosomes contain the NADH dehydrogenase module of mitochondrial complex I. *Nature* 432, 618–622. <https://doi.org/10.1038/nature03149>.
28. Kosiol, C., Goldman, N., and Buttimore, N.H. (2004). A new criterion and method for amino acid classification. *J. Theor. Biol.* 228, 97–106. <https://doi.org/10.1016/j.jtbi.2003.12.010>.
29. Susko, E., and Roger, A.J. (2007). On reduced amino acid alphabets for phylogenetic inference. *Mol. Biol. Evol.* 24, 2139–2150. <https://doi.org/10.1093/molbev/msm144>.
30. Phillips, M.J. (2009). Branch-length estimation bias misleads molecular dating for a vertebrate mitochondrial phylogeny. *Gene* 441, 132–140. <https://doi.org/10.1016/j.gene.2008.08.017>.
31. Hernandez, A.M., and Ryan, J.F. (2021). Six-state amino acid recoding is not an effective strategy to offset compositional heterogeneity and saturation in phylogenetic analyses. *Syst. Biol.* 70, 1200–1212. <https://doi.org/10.1093/sysbio/syab027>.
32. Foster, P.G., Schrempf, D., Szöllösi, G.J., Williams, T.A., Cox, C.J., and Embley, T.M. (2022). Recoding amino acids to a reduced alphabet may increase or decrease phylogenetic accuracy. *Syst. Biol.* syac042. <https://doi.org/10.1093/sysbio/syac042>.
33. Dayhoff, M., Schwartz, R., and Orcutt, B. (1972). A model of evolutionary change in proteins atlas of protein sequence and structure, 5, pp. 345–352.
34. Álvarez-Carretero, S., Tamuri, A.U., Battini, M., Nascimento, F.F., Carlisle, E., Asher, R.J., Yang, Z., Donoghue, P.C.J., and dos Reis, M. (2022). A species-level timeline of mammal evolution integrating phylogenomic data. *Nature* 602, 263–267. <https://doi.org/10.1038/s41586-021-04341-1>.
35. Kumar, S. (2022). Embracing green computing in molecular phylogenetics. *Mol. Biol. Evol.* 39, msac043. <https://doi.org/10.1093/molbev/msac043>.
36. Hejnal, A., Obst, M., Stamatakis, A., Ott, M., Rouse, G.W., Edgecombe, G.D., Martinez, P., Baguña, J., Bailly, X., Jondelius, U., et al. (2009). Assessing the root of bilaterian animals with scalable phylogenomic methods. *Proc. Biol. Sci.* 276, 4261–4270. <https://doi.org/10.1098/rspb.2009.0896>.
37. Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* 17, 368–376. <https://doi.org/10.1007/BF01734359>.
38. Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA

- sequences with variable rates over sites: approximate methods. *J. Mol. Evol.* 39, 306–314. <https://doi.org/10.1007/BF00160154>.
39. Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. In *Some Mathematical Questions in Biology DNA Sequence Analysis* (American Mathematical Society), pp. 57–86.
40. A Shepherd, D., and Klaere, S. (2019). How well does your phylogenetic model fit your data? *Syst. Biol.* 68, 157–167. <https://doi.org/10.1093/sysbio/syy066>.
41. Laumer, C.E. (2018). Inferring ancient relationships with genomic data: a commentary on current practices. *Integr. Comp. Biol.* 58, 623–639. <https://doi.org/10.1093/icb/icy075>.
42. Holton, T.A., and Pisani, D. (2010). Deep genomic-scale Analyses of the metazoa reject coelomata: evidence from single- and multigene families analyzed under a supertree and supermatrix paradigm. *Genome Biol. Evol.* 2, 310–324. <https://doi.org/10.1093/gbe/evq016>.
43. Hillis, D.M. (1996). Inferring complex phylogenies. *Nature* 383, 130–131.
44. Jermini, L.S., Catullo, R.A., and Holland, B.R. (2020). A new phylogenetic protocol: dealing with model misspecification and confirmation bias in molecular phylogenetics. *NAR Genom. Bioinform.* 2, lqaa041. <https://doi.org/10.1093/nargab/lqaa041>.
45. Pett, W., Adamski, M., Adamska, M., Francis, W.R., Eitel, M., Pisani, D., and Wörheide, G. (2019). The role of homology and orthology in the phylogenomic analysis of metazoan gene content. *Mol. Biol. Evol.* 36, 643–649. <https://doi.org/10.1093/molbev/msz013>.
46. Zhao, Y., Vinther, J., Parry, L.A., Wei, F., Green, E., Pisani, D., Hou, X., Edgecombe, G.D., and Cong, P. (2019). Cambrian sessile, suspension feeding stem-group ctenophores and evolution of the comb jelly body plan. *Curr. Biol.* 29, 1112–1125.e2. <https://doi.org/10.1016/j.cub.2019.02.036>.
47. Juravel, K., Porras, L., Höhna, S., Pisani, D., and Wörheide, G. (2021). Improved resolution of recalcitrant nodes in the animal phylogeny through the analysis of genome gene content and morphology. *Evol. Biol.* <https://doi.org/10.1101/2021.11.19.469253>.
48. Muñoz-Gómez, S.A., Susko, E., Williamson, K., Eme, L., Slamovits, C.H., Moreira, D., López-García, P., and Roger, A.J. (2022). Site- and branch-heterogeneous analyses of an expanded dataset favour mitochondria as sister to known Alphaproteobacteria. *Nat. Ecol. Evol.* 6, 253–262.
49. Borowiec, M.L., Lee, E.K., Chiu, J.C., and Plachetzki, D.C. (2015). Extracting phylogenetic signal and accounting for bias in whole-genome data sets supports the Ctenophora as sister to remaining Metazoa. *BMC Genomics* 16, 987. <https://doi.org/10.1186/s12864-015-2146-4>.
50. Lemmon, A.R., and Moriarty, E.C. (2004). The importance of proper model assumption in bayesian phylogenetics. *Syst. Biol.* 53, 265–277. <https://doi.org/10.1080/10635150490423520>.
51. Huelsenbeck, J., and Rannala, B. (2004). Frequentist properties of bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. *Syst. Biol.* 53, 904–913. <https://doi.org/10.1080/10635150490522629>.
52. Fabreti, L.G., and Höhna, S. (2022). Bayesian inference of phylogeny is robust to substitution model over-parameterization. *Evol. Biol.* <https://doi.org/10.1101/2022.02.17.480861>.
53. Baños, H., Susko, E., and Roger, A.J. (2022). Are profile mixture models over-parameterized? *Evol. Biol.* <https://doi.org/10.1101/2022.02.18.481053>.
54. Lam, J.H., Shivhare, D., Chia, T.W., Chew, S.L., Sinsinbar, G., Aw, T.Y., et al. Next-generation intranasal Covid-19 vaccine: a polymersome-based protein subunit formulation that provides robust protection against multiple variants of concern and early reduction in viral load of the upper airway in the golden Syrian hamster model, Preprint at: *BioRxiv*. <https://doi.org/10.1101/2022.02.12.480188>.
55. Whelan, S., and Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.* 18, 691–699. <https://doi.org/10.1093/oxfordjournals.molbev.a003851>.
56. Le, S.Q., and Gascuel, O. (2008). An improved general amino acid replacement matrix. *Mol. Biol. Evol.* 25, 1307–1320. <https://doi.org/10.1093/molbev/msn067>.
57. Jones, D.T., Taylor, W.R., and Thornton, J.M. (1992). The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.* 8, 275–282. <https://doi.org/10.1093/bioinformatics/8.3.275>.
58. Cannon, J.T., Vellutini, B.C., Smith, J., Ronquist, F., Jondelius, U., and Hejnol, A. (2016). Xenacoelomorpha is the sister group to Nephrozoa. *Nature* 530, 89–93. <https://doi.org/10.1038/nature16520>.
59. Bujaki, T., and Rodrigue, N. (2022). Bayesian cross-validation comparison of amino acid replacement models: contrasting profile mixtures, pairwise exchangeabilities, and gamma-distributed rates-across-sites. *J. Mol. Evol.* 90, 468–475. <https://doi.org/10.1007/s00239-022-10076-y>.
60. C.G.P. McCarthy, P.O. Mulhair, K. Siu-Ting, C.J. Creevey, M.J. O'Connell, Improving orthologous signal and model fit in datasets addressing the root of the animal phylogeny, Preprint at: *BioRxiv*, <https://doi.org/10.1101/2022.11.21.517274>.
61. Crotty, S.M., Minh, B.Q., Bean, N.G., Holland, B.R., Tuke, J., Jermini, L.S., and Haeseler, A.V. (2020). GHOST: recovering historical signal from heterotachously evolved sequence alignments. *Syst. Biol.* 69, 249–264. <https://doi.org/10.1093/sysbio/syzy051>.
62. Campbell, L.I., Rota-Stabelli, O., Edgecombe, G.D., Marchioro, T., Longhorn, S.J., Telford, M.J., Philippe, H., Rebecchi, L., Peterson, K.J., and Pisani, D. (2011). MicroRNAs and phylogenomics resolve the relationships of Tardigrada and suggest that velvet worms are the sister group of Arthropoda. *Proc. Natl. Acad. Sci. USA* 108, 15920–15924. <https://doi.org/10.1073/pnas.1105499108>.
63. McInerney, J.O., O'Connell, M.J., and Pisani, D. (2014). The hybrid nature of the Eukaryota and a consilient view of life on Earth. *Nat. Rev. Microbiol.* 12, 449–455. <https://doi.org/10.1038/nrmicro3271>.
64. Rota-Stabelli, O., Campbell, L., Brinkmann, H., Edgecombe, G.D., Longhorn, S.J., Peterson, K.J., Pisani, D., Philippe, H., and Telford, M.J. (2011). A congruent solution to arthropod phylogeny: phylogenomics, microRNAs and morphology support monophyletic Mandibulata. *Proc. Biol. Sci.* 278, 298–306. <https://doi.org/10.1098/rspb.2010.0590>.
65. Simakov, O., Bredeson, J., Berkoff, K., Marletaz, F., Mitros, T., Schultz, D.T., O'Connell, B.L., Dear, P., Martinez, D.E., Steele, R.E., et al. (2022). Deeply conserved synteny and the evolution of metazoan chromosomes. *Sci. Adv.* 8, eabi5884. <https://doi.org/10.1126/sciadv.abi5884>.
66. Rota-Stabelli, O., Lartillot, N., Philippe, H., and Pisani, D. (2013). Serine codon-usage bias in deep phylogenomics: pancrustacean relationships as a case study. *Syst. Biol.* 62, 121–133. <https://doi.org/10.1093/sysbio/sys077>.
67. Quang, L.S., Gascuel, O., and Lartillot, N. (2008). Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics* 24, 2317–2323. <https://doi.org/10.1093/bioinformatics/btn445>.
68. Nguyen, L.-T., Schmidt, H.A., von Haeseler, A., and Minh, B.Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32, 268–274. <https://doi.org/10.1093/molbev/msu300>.
69. Lartillot, N., Rodrigue, N., Stubbs, D., and Richer, J. (2013). PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Syst. Biol.* 62, 611–615. <https://doi.org/10.1093/sysbio/syt022>.
70. Philippe, H., Poustka, A.J., Chiodin, M., Hoff, K.J., Dessimoz, C., Tomiczek, B., Schiffer, P.H., Müller, S., Domman, D., Horn, M., et al. (2019). Mitigating anticipated effects of systematic errors supports sister-group relationship between xenacoelomorpha and ambulacraria. *Curr. Biol.* 29, 1818–1826.e6. <https://doi.org/10.1016/j.cub.2019.04.009>.
71. Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of

- large phylogenies. *Bioinformatics* 30, 1312–1313. <https://doi.org/10.1093/bioinformatics/btu033>.
72. Lartillot, N., Brinkmann, H., and Philippe, H. (2007). Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evol. Biol.* 7, S4. <https://doi.org/10.1186/1471-2148-7-S1-S4>.
73. Swofford, D.L. (2003). *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods)*. Version 4 (Sinauer Associates, Sunderland, Massachusetts).
74. Lozano-Fernandez, J., Tanner, A.R., Giacomelli, M., Carton, R., Vinther, J., Edgecombe, G.D., and Pisani, D. (2019). Increasing species sampling in chelicerate genomic-scale datasets provides support for monophyly of Acari and Arachnida. *Nat. Commun.* 10, 2295. <https://doi.org/10.1038/s41467-019-10244-7>.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
random_recoding_generator.pl	This study	https://bitbucket.org/bzxdp/giacomelli_et_al_2022_recodings/src/master/Scripts/
compare_recodings.pl	This study	https://bitbucket.org/bzxdp/giacomelli_et_al_2022_recodings/src/master/Scripts/
Recode_to_random.pl	This study	https://bitbucket.org/bzxdp/giacomelli_et_al_2022_recodings/src/master/Scripts/
Phylobayes	N/A	https://github.com/bayesiancook
RAxML	N/A	https://github.com/stamatak
Other		
Simulated datasets	Kapli and Telford (2020) ¹² and This paper	https://bitbucket.org/bzxdp/giacomelli_et_al_2022_recodings/src/master/Simulated_Datasets/
Random recoding schemes	This study	https://bitbucket.org/bzxdp/giacomelli_et_al_2022_recodings/src/master/Random_recodings/
Supplementary file 1 (all results)	This study	https://bitbucket.org/bzxdp/giacomelli_et_al_2022_recodings/src/master/
Table S1 (similarity scores of random recodings and Dayhoff-6)		https://bitbucket.org/bzxdp/giacomelli_et_al_2022_recodings/src/master/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the Lead contact, Davide Pisani at davide.pisani@bristol.ac.uk.

Materials availability

This study did not generate new unique reagents.

Data and code availability

All the data, material and custom scripts used in this study available in Bitbucket: https://bitbucket.org/bzxdp/giacomelli_et_al_2022_recodings/src/master/. Links for individual scripts and data are reported in the [key resources table](#). No new data were generated with this article.

METHODS DETAILS

Recoding strategies: A description and a logical rationale for their generalisation

Nucleotide sequences of protein coding genes can be recoded into codons or amino acid data, and phylogenomic analyses addressing tricky nodes in the Tree of Life usually use data recoded into amino acid sequences. This is because amino acids are expected to be less saturated and heterogeneous than nucleotides due to the degeneracy of the genetic code,⁶⁶ which implies that not all mutations in a nucleotide sequence correspond to an amino acid substitution. Nucleotide sequences representing protein coding genes could also be recoded using “high level recodings”,⁴¹ where codons (or amino acids) are grouped into bins (or categories) based on the chemical and physical properties of the amino acid they code for.

A universe of alphabets can be defined into which nucleotide sequences of protein coding genes could be recoded. These alphabets can have different cardinality, with the largest possible alphabet having 64 states (one for each possible codon), and the smallest 2 states, e.g. a coding strategy where only hydrophobic and hydrophilic amino acids/codons are distinguished (the latter is implemented in Phylobayes as Dayhoff-2). The classic 20-state (amino acid) recoding is therefore just an instance within the universe of possible recodings into which nucleotide sequences of protein coding genes could be translated. Indeed, experiments have been performed, for example, using an alternative 23-state recoding⁶⁶ distinguishing, in addition to the 20 amino acids, the codon classes of the three 6-fold degenerate amino acids. However, these analyses showed the use of a 23-state recoding was not providing advantages over a 20-state recoding suggesting that recodings with more states are not necessarily better for resolving phylogenetic relationships.

Compared to other recodings, 20-state amino acid recording has the advantage that its use can be promptly justified on biological grounds, as it represents the primary structure of protein sequences. It is thus logical and appropriate that we consider amino acid recoding to be the default strategy to recode nucleotide sequences of protein coding genes, and our study should not be interpreted as challenging this view. However, this does not imply that we should assume *a-priori* that amino acid recoding would maximise phylogenetic accuracy over every possible phylogenomic dataset and problem. After all, the use of 20-state amino acid recoding is usually justified on the grounds that, similarly to other recoding strategies such as Dayhoff-6, it masks substitutions (particularly at third codon positions) that tend to be saturated, compositionally heterogeneous, and phylogenetically misleading. Accordingly, if evidence exists that once translated into a set of amino acid sequences a dataset is still compositionally heterogeneous, saturated, or both, the use of a high-order recoding that more efficiently reduces saturation and compositional heterogeneity should be justified. Indeed, as suggested also by Laumer,⁴¹ if we accept the logic by which amino acid data are generated by recoding nucleotide data, and consider that every recoding strategy can be characterised as an approach to recode nucleotide sequences (by directly clustering codons into bins), we have to accept that standard (20-state) amino acid recoding is just one of many, related recoding strategies in which nucleotide datasets of protein coding genes can be transformed.

Dayhoff-6³³ (Figure 1) uses a six state alphabet, and is the most widely used, high-order⁴¹ recoding scheme. However, promising experiments have also been completed using 9-, 12-, 15- and 18-state alphabets.³¹ Dayhoff-6 partitions the 20 amino acids (or their codons) into six differently sized bins: one 5-amino acid bin, two 4-amino acid bins, two 3-amino acid bins and one 1-amino acid bin, based on how frequently different amino acids substitute for each other, i.e. based on their exchangeability rates¹¹ (Figure 1). Other 6-state recoding schemes have been proposed including the 6-bin recoding of Kosiol et al.²⁸ – KGB6, and the 6-bin recoding scheme of Susko and Roger²⁹ – SR6, see also.¹¹ KGB6 and SR6 differ from Dayhoff-6 in the details of the amino acid partitioning rule they implement, but the amino acid content of their bins overlaps.¹¹

When a high-order recoding (i.e. a recoding with less than 20 states) is used, substitutions among amino acids within the same bin are masked through a process of synonymization. This process is expected to reduce compositional heterogeneity and saturation because the amino acids clustered within each bin are usually identified based on the properties that they share. Amino acids within the same bin are therefore expected to interchange among each other more frequently than with amino acids from different bins. Accordingly, it is expected that within-bin substitutions should be more common than across-bin substitutions as the former are less likely to disrupt protein function¹¹; Figure 1.

Our study uses simulated and empirical data (see Figure S1 for a visual summary of our experimental protocol) to investigate whether the use of 6-state recoding (hereafter only referred to as a recoding), can improve the accuracy of phylogenomic analyses. We used the Dayhoff-6 recoding scheme as this is the most common recoding in the literature, but we performed sensitivity tests using KGB6 and SR6 to evaluate whether using different 6-state recodings would have led to different conclusions.

Datasets

We used the dataset of Kapli and Telford¹² which consists of 200 alignments of 30,000 sites and 97 taxa simulated under the CAT-LG + G model, assuming that either the comb jellies (Ctenophora-sister hypothesis; 100 alignments) or the sponges (Porifera-sister hypothesis; 100 alignments) are the sister of all the

other animals (Figure S2). The first topology (Ctenophora-sister) represents the case when the simulation target is in the “Farris zone” (Figure S2), while the second topology (Porifera-sister) represents the case when the simulation target is in the “Felsenstein zone” (Figure S2).¹² Having been simulated using CAT-LG + G, the alignments are across-site, rate (the Gamma component of the model³⁸) and compositionally (the CAT component²³) heterogeneous. The alignments display also heterogeneity of the amino acid replacement rate process (the LG component⁵⁶), see^{4,11} for introductions to the heterogeneity of amino acid substitution models. However, our alignments are not across-lineage compositionally heterogeneous. This is not a problem because Dayhoff-6 reduces both across-site and across-lineage compositional heterogeneity, as well as saturation.¹¹ Lack of across-lineage compositional heterogeneity in our datasets implies that our simulations only demonstrate the behavior of recodings when the datasets are across-site compositionally heterogeneous. However, other studies have tested recodings on across-lineage compositionally heterogeneous datasets^{29,31,32} achieving results comparable to ours, at the least to the extent to which these studies can be compared to our own (e.g. when contrasting results from datasets of comparable size).

The alignments were subsampled to generate datasets with different numbers of characters and taxa. We generated datasets of 1,000, 5,000, 10,000, and 30,000 sites and 20 taxa; as well as datasets of 30,000 sites with 4, 10, and 20 taxa. When subsampling characters from the datasets, we made sure that each character was sampled only once in each subsampled dataset, and that all key animal lineages in the original 97-taxon trees were retained – see Figure S2 for our target topologies. Taxon subsampling was fixed so that all our 20-, 10- and 4-taxon datasets included the same terminals (Figure S2). We decided to use datasets with 20 and 4 taxa to maximise comparability with previous studies.^{31,32}

Analyses of the simulated datasets

A total of 200 datasets of 30,000 sites and 20 taxa (100 assuming Porifera-sister to be true and 100 assuming Ctenophora-sister to be true) were analyzed using a General Time Reversible³⁹ (GTR + G) model, accounting for amino acid replacement rate heterogeneity – GTR, and across-site rate heterogeneity (+G;³⁸). We then used a series of nCAT models where the number of site-frequency categories was fixed *a priori* to 10, 30, 60, 120, and 240 to account for greater across-site compositional heterogeneity. These models should not be confused with the precomputed CAT models⁶⁷ implemented in ML software such as IQtree⁶⁸ and generally referred to as C10, C20, ... Cn. This is because, despite nCAT models have a fixed number of site-frequency categories like the C10, ... Cn models, nCAT models are not precomputed. The nCAT10 to nCAT240 models used a simple Poisson process (F81³⁷), to model amino acid replacement rate, and a Gamma distribution (+G) to account for across-site rate heterogeneity. Accordingly, the full models used in our study were the nCAT10-F81 + G to nCAT240-F81 + G. We then used CAT-F81 + G which estimates the number of site-frequency categories during tree search, and CAT-GTR + G, which combines the properties of CAT + G and GTR + G, modeling across-site compositional and rate heterogeneity and amino acid replacement rate heterogeneity.

Datasets of 1,000 to 10,000 sites and 20 taxa, and datasets of 30,000 sites with 4 or 10 taxa were only analyzed under nCAT10 and used for tests aimed at investigating the effect of changing alignment length and taxon sampling density on our results. The nCAT10 model was selected *a posteriori* for these analyses, because it emerged from the study of the 30,000 sites and 20 taxa alignments that this model achieve the most distinguishable results when applied to our amino acid and recoded alignments.

Bayesian analyses were performed in Phylobayes.⁶⁹ Phylobayes MPI v.18 was used to perform CAT, CAT-GTR and GTR analyses. Phylobayes-MPI-nCAT was used to perform analyses using nCAT models. Bayesian analyses of 1,000 site alignments were run for 1,000 cycles (with a burnin of 500 cycles and a subsampling frequency of 5). All other Bayesian analyses used 2,000 cycles, with a burnin of 1000 cycles and a subsampling frequency of 10. For each dataset a single run was completed, as is customary when multiple datasets are used in jackknife and similar computation intensive experiments.^{18,70} However, to make sure that 2,000 cycles were enough to reach convergence, under nCAT10, we also completed (for each dataset) two more independent runs of 10,000 cycles, testing convergence after 2,000 and 10,000 cycles (using burnins of 1,000 and 3,000 cycles respectively). In addition, to make sure that the use of Bayesian analysis did not bias our results, we performed Maximum Likelihood (ML) analyses in RAxML⁷¹ under GTR + G (30,000 sites and 20 taxa datasets) with 200 bootstrap replicates. For the ML analyses we used GTR + G as it is the only model, among the ones we used, with an ML implementation. All datasets were analyzed as amino

acids and Dayhoff-6 recoded alignments. Analyses of the 30,000 sites and 20 taxa datasets were also performed (under nCAT10) using SR6²⁹ and KGB6²⁸ to test the effect of using alternative 6-state recoding schemes on our results.

Measuring phylogenetic accuracy

Phylogenetic accuracy could simply be defined as the inference of the target (i.e. true) tree. In our case, this would be the successful inference of trees displaying Ctenophora-sister or Porifera-sister, depending on the simulation set. However, accuracy without support is a weak and potentially misleading measure of the utility of a phylogenetic method. We thus define a tree to have been accurately resolved when the target tree is correctly inferred with support (posterior probability in Bayesian analyses and bootstrap in ML) ≥ 0.5 . To make sure that using a 0.5 cutoff to define accuracy did not bias our results, under nCAT10, we also tested the effect of defining as accurate only analyses that inferred the target tree with a posterior probability ≥ 0.95 . In addition to accurately resolved trees, we distinguished incorrect trees (where the wrong tree was inferred with support ≥ 0.5 , or ≥ 0.95), and unresolved trees (where either the correct or the incorrect tree was inferred but with support < 0.5 , or < 0.95).

Measuring model adequacy

Posterior Predictive Analysis (PPA) was used to evaluate whether the considered model could adequately describe the across-site compositional heterogeneity of our simulated and empirical datasets. The test statistic for these analyses is the mean diversity (PPA-Div⁷²) per site (mean number of distinct states per sites), which is computed on the true data, and compared against its null (posterior predictive) distribution. In our analyses of empirical datasets, we also used PPA to test whether the considered models could adequately describe the lineage-specific compositional heterogeneity of the data. For these analyses the test statistic used was the mean deviation measured as the sum over the 20 aminoacids of the absolute differences between the taxon-specific and global empirical frequencies (PPA-Mean) – see Phylobayes manual. The observed value is computed on the true data and compared with its null (posterior predictive) distribution. PPA-Mean has not been calculated for our simulated data because these datasets, having been simulated with a model that is compositionally-homogeneous across-lineage (CAT-LG + G), are not across-lineage compositionally heterogeneous.

The ability of different models to adequately describe the compositional heterogeneity of amino acid and recoded datasets, was expressed using Z-scores, the number of standard deviations separating the heterogeneity observed for the true data from the average heterogeneity of the null (posterior predictive) distribution (Phylobayes manual and¹¹). We interpret Z-scores as follows: a Z score value within the interval 2 to -2 indicates that the hypothesis that the model adequately describes the data cannot be rejected.¹¹ However, these boundaries are set subjectively, as it is clear that the ability to describe the data of models achieving a Z-scores of 1.99 and 2.01 (for example) should not be worth distinguishing. Hence here we use PPA scores only as means to quantify the fit of the different models. As we need to discuss these models and their fit in text, we refer to models for which $Z > |10|$ to be very poor descriptors of the heterogeneity in the data. Models for which: $|5| < Z < |10|$ are deemed to describe the data poorly, and models for which $|3| < Z < |5|$ are described as having a fit that varies from acceptable to fairly poor. Models with $|2| < Z < |3|$ are deemed to be describing compositional heterogeneity in an acceptable way, and finally models for which $|0| < Z < |2|$ adequately model the compositional heterogeneity of the data.

All PPA analyses were performed in Phylobayes MPI⁶⁹ (2,000 cycles with a burnin of 1000 generations, subsampling every 10 generations). For CAT, GTR and CAT-GTR, we used Phylobayes MPI v.1.8. For nCAT10-nCAT20140 we used Phylobayes-nCAT.

Measuring saturation

To test whether recoding reduced saturation, we used PAUPv4.168⁷³ to compute Neighbor Joining (NJ) trees for our amino acid and recoded 30,000 sites and 20 taxa alignments, using observed (P) distances without Gamma correction. After that, tip-to-tip distances were calculated for all the NJ trees and for the corresponding Bayesian trees (under all considered models). Saturation plots were generated, for the amino acids and recoded datasets for each considered model, by comparing corresponding tip-to-tip distances estimated from the NJ tree and the Bayesian tree. For each plot, the regression line (forced through the origin of the axes) was calculated and its gradient (M) was extracted. Saturation (Sat) was

estimated (as it is customary) as $\text{Sat} = 1 - M$.^{14,74} To test whether recoding reduced saturation, Sat scores calculated for all amino acids and recoded datasets (for all tested models) were plotted and compared.

Random recodings

A perl script (Recodings_generator.pl – https://bitbucket.org/bzxdp/giacomelli_et_al_2022_recodings/src/master/) was used to randomly generate recodings based on the Dayhoff-6 amino acid binning scheme: one 5-state bin, two 4-state bins, two 3-state bins and one 1-state bin. This script was used to generate 1,000 recoding schemes. The similarity of these recodings to Dayhoff-6 was then calculated using a second script (Recodings_comparator.pl – https://bitbucket.org/bzxdp/giacomelli_et_al_2022_recodings/src/master/). The latter uses intersections of sets to estimate the percent similarity between the bins of randomly generated recodings and those of a real (reference) recoding – here Dayhoff-6. This experiment found that the random recodings we generated have variable similarity to Dayhoff-6 (average 36.5%, range 0%–80%) see Table S1 – https://bitbucket.org/bzxdp/giacomelli_et_al_2022_recodings/src/master/.

We then generated 100 recodings of 0% similarity to Dayhoff-6 (0%-recodings), and 100 recodings of 90% similarity to Dayhoff-6 (90%-recodings), to evaluate whether the performance of randomly generated recodings depends on their similarity to Dayhoff-6. To generate the 0%- and 90%- recodings, we first generated 20,000,000 recodings using Recodings_generator.pl. This large number of recodings was only generated to make sure to have generated 100 recodings of 90% similarity to Dayhoff-6 as these are rarely generated (Table S1). After that, we used Recodings_comparator.pl to identify the first 100 recodings of specified (either 0% or 90%) similarity to Dayhoff-6.

Each 30,000-site and 20-taxon dataset was randomly associated with one of the randomly generated 0%- and 90%-recodings. PPA-Div scores and accuracy were calculated (under nCAT10) for the 0%-recoded and 90%-recoded datasets. These values were then compared against PPA-Div scores and accuracy values obtained using amino acids and Dayhoff-6.

Estimating the number of substitutions entailed by amino acid and recoded data and approximating information content

To evaluate whether there are more within Dayhoff-6 bin than across Dayhoff-6 bin substitutions, we estimated the tree length of phylogenies inferred using 0%-, 90%, and Dayhoff-6 recoded datasets. The expectation being that if there are more within-bin substitutions, as one would expect if the rationale underpinning the use of recoding is valid, Dayhoff-6 would mask more substitutions than 0%- and 90%-recodings. Accordingly, trees inferred from Dayhoff-6 recoded datasets would have a lower tree length than trees inferred using random recoding schemes. In addition, among the trees inferred using random recoding schemes, the 90%-recoding trees should be expected to have a lower tree length than the 0%-recoding trees, as the 90%-recodings are more similar to Dayhoff-6 than the 0%-recodings and should thus mask more substitutions.

The information content of an alignment can be roughly approximated by the number of substitutions it entails. To approximate the information loss incurred when recoding our simulated datasets, we calculated the tree length of our 30,000 site amino acid alignments and of their recoded version. We then expressed as a percent the difference in average tree length between the recoded and the amino acid datasets. To evaluate how the information loss was distributed across the branches of the phylogeny, we obtained branch lengths (for the simulated amino acids and recoded data) for all the branches in our target tree (this experiment was only done under Porifera-sister), and for each branch we expressed as a percent the difference in length observed when the recoded and amino acid datasets were used. These values were then mapped on the Porifera-sister tree for visual inspection. All parsimony analyses were performed using PAUPv4.168⁷³ – Heuristic search with 1,000 random addition and the multree option turned off.

Empirical datasets

We tested three empirical datasets Whelan-D20²⁰ (in text referred to as Whelan2015), Whelan-strict¹⁹ (in text referred to as Whelan2017), and the non-Bilateria matrix of Laumer et al.⁵ in text referred to as Laumer2019. These datasets were chosen because they addressed the phylogenetic problem represented by our simulated data. We analyzed these datasets as amino acid alignments and as 0%-recoded, 90%-recoded and Dayhoff-6 recoded datasets. All datasets were analyzed using nCAT60.¹³

The three datasets use different outgroups. To rationalise the analyses and make them more comparable, we removed all non Choanoflagellata outgroups from Whelan2015 and Laumer2019. Whelan2017 already include only choanoflagellate outgroups. To evaluate whether the information content of these datasets was comparable, we used parsimony to estimate their tree length, and divided these values by the number of sites in each dataset, approximating in this way the number of substitutions per site of each dataset. We tested whether nCAT60 adequately describes these datasets using PPA-Div and PPA-Mean (see above [measuring model adequacy](#)). As our simulated data were generated under CAT-LG + G, to make sure that results from our simulated and empirical datasets were comparable, we also tested (using PPA-Div) whether CAT-LG + G could describe the across-site compositional heterogeneity of the three empirical datasets. Support values were obtained (under nCAT60) for both Porifera-sister and Ctenophora-sister, under all recoding strategies (0%-recodings, 90%-recodings and Dayhoff-6). These values were plotted and compared against each other. We also plotted support values obtained, from our simulated datasets, when Ctenophora-sister and Porifera-sister were true and compared these values against those obtained from the empirical datasets. For the simulated datasets, we plotted the values obtained under nCAT10. This is because, with the simulated datasets, nCAT10 achieved PPA-Div scores comparable to those achieved by nCAT60 with the empirical datasets, suggesting that nCAT10 fits our simulated amino acid datasets similarly to how nCAT60 fits the empirical datasets.

Green phylogenomics

To estimate whether recoding can be a useful tool for green computing, we obtained the time used by Phylobayes MPI v.1.8, to complete 50 GTR, and CAT-GTR analyses for both amino acid and recoded data, using the timestamps in the trace files. All analyses were parallelised over 28 Intel Xeon Gold 6226R CPU – 2.90 GHz processors.