Six-state Amino Acid Recoding is not an Effective Strategy to Offset Compositional Heterogeneity and Saturation in Phylogenetic Analyses

Alexandra M. Hernandez¹,² and Joseph F. Ryan¹,²

¹Whitney Laboratory for Marine Bioscience, 9505 Ocean Shore Boulevard, St. Augustine, FL, 32080, USA
²Department of Biology, University of Florida, 220 Bartram Hall, P.O. Box 118525, Gainesville, FL, 32611, USA

Corresponding Author:
Joseph F. Ryan
9505 Ocean Shore Boulevard, St. Augustine, FL, 32080, USA
904-201-8426
joseph.ryan@whitney.ufl.edu

© The Author(s) 2021. Published by Oxford University Press, on behalf of the Society of Systematic Biologists. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.
ABSTRACT

Six-state amino acid recoding strategies are commonly applied to combat the effects of compositional heterogeneity and substitution saturation in phylogenetic analyses. While these methods have been endorsed from a theoretical perspective, their performance has never been extensively tested. Here, we test the effectiveness of 6-state recoding approaches by comparing the performance of analyses on recoded and non-recoded datasets that have been simulated under gradients of compositional heterogeneity or saturation. In our simulation analyses, non-recoding approaches consistently outperform 6-state recoding approaches. Our results suggest that 6-state recoding strategies are not effective in the face of high saturation. Further, while recoding strategies do buffer the effects of compositional heterogeneity, the loss of information that accompanies 6-state recoding outweighs its benefits. In addition, we evaluate recoding schemes with 9, 12, 15, and 18 states and show that these consistently outperform 6-state recoding. Our analyses of other recoding schemes suggest that under conditions of very high compositional heterogeneity, it may be advantageous to apply recoding using more than 6 states, but we caution that applying any recoding should include sufficient justification. Our results have important implications for the more than 90 published papers that have incorporated 6-state recoding, many of which have significant bearing on relationships across the tree of life.

KEYWORDS

six-state amino acid recoding, Dayhoff 6-state recoding, S&R 6-state recoding, compositional heterogeneity, substitution saturation
Compositional heterogeneity and substitution saturation are major challenges to phylogenetic inference. Compositional heterogeneity stems from the tendency of genes or organisms to have unequal proportions of amino acids (Collins et al. 1994; Foster and Hickey 1999). These unequal amino acid frequencies are caused by mutational and selective pressures acting at the nucleotide level (Singer and Hickey 2000; Knight et al. 2001), as well as differences in translational efficiency (Akashi and Eyre-Walker 1998). The combination of evolutionary and biological processes results in different amino acid compositions across taxa on the tree. Consequently, challenges to phylogenetic analyses arise when distantly related taxa share sequence similarities due to homoplasy (convergence), rather than descent from a common ancestor (Foster and Hickey 1999; Tarrio et al. 2001).

Similarly, phylogenetic reconstruction artifacts emerge under substitution saturation of amino acids. Substitution saturation occurs when there have been multiple amino acid substitutions at the same site washing out the evolutionary signal (Ho and Jermiin 2004). Like compositional heterogeneity, sequence saturation can lead to long branch attraction, driving unrelated taxa to group together in a clade due to homoplasy (Felsenstein 1978; Hendy and Penny 1989).

There is a large body of research on the conditions for state aggregation (or lumpability) in modeling character data such as DNA or amino acids (Kemeny and Snell, 1960; Courtois, 1977). Based on this foundation, matrix recoding has been proposed as a solution for both compositional heterogeneity and substitution saturation (Blanquart and Lartillot 2006; Susko & Roger 2007). Under matrix recoding methods, nucleotides or amino acids are lumped into groups based on function (Blanquart and Lartillot 2006). For example, under the RY nucleotide recoding strategy, purines (i.e., A and G) are coded with the character R and pyrimidines (i.e., T
and C) are coded with the character Y (Woese et al. 1991; Phillips et al. 2001). In this recoding scenario, only transversion events are meaningful in a phylogenetic analysis. A similar recoding strategy has been implemented for amino acids, the most well-known being Dayhoff 6-state recoding. In Dayhoff 6-state recoding, chemically related amino acids that frequently replace each other are pooled together into six groups based on similar substitution scores in the Dayhoff (or PAM250) matrix (Dayhoff et al. 1978): AGPST, DENQ, HKR, ILMV, FWY, and C (Embley et al. 2003a; Hrdy et al. 2004). Thus, only amino acid changes between categories, and not within categories, are considered substitutions. Since the introduction of Dayhoff 6-state recoding, several other 6-state amino acid recoding strategies based around other scoring matrices have been developed. For example, S&R 6-state recoding (Susko and Roger 2007; Feuda et al. 2017) is based on the JTT matrix (Jones et al. 1992) and KGB 6-state recoding (Kosiol et al. 2004; Feuda et al. 2017) is based on the WAG matrix (Whelan and Goldman 2001).

Authors have increasingly been applying 6-state recoding to phylogenetic analyses. To date, there are at least 91 phylogenetic studies that have implemented 6-state amino acid recoding strategies, with the highest number of studies published in 2019 (Table 1). Several of these studies have proposed controversial topologies based on results from recoded matrices with deep implications across the tree of life (e.g., Rodríguez-Ezpeleta and Embley 2012; Feuda et al. 2017; Laumer et al. 2018; Puttick et al. 2018; Marlétaz et al. 2019). For example, the relationships of non-bilaterian animals have a major influence on how we understand the origin and evolution of key animal innovations (e.g., true epithelia, the gut, neural and muscle cell types), and recent papers using 6-state recoding have major implications on how these relationships are viewed (Feuda et al. 2017; Laumer et al. 2018). While amino acid recoding has
been considered from a theoretical perspective (Davidson et al. 2002; Embley et al. 2003a; Hrdy et al. 2004), and there have been comparisons between different recoding strategies (Susko and Rogers, 2007), there has not been extensive empirical testing of the widely applied 6-state recoding approaches. Historically, simulation has been an effective strategy for empirically testing the performance of phylogenetic approaches (Kuhner and Felsenstein 1994; Swofford et al. 2001; Zwickl and Hillis 2002; Kubatko and Degnan 2007; Huang and Knowles 2016). In this study, we simulate datasets with a gradient of either compositional heterogeneity or saturation and compare the performance of maximum-likelihood analyses on 6-state recoded datasets to the same analyses on non-recoded datasets. We also run a subset of these analyses using 9-, 12-, 15-, and 18-state recoding schemes and compare these results to those achieved with 6-state recoded and non-recoded matrices.

MATERIALS & METHODS

Reproducibility and Transparency Statement

Custom scripts, command lines, and data used in these analyses are available in GitHub (https://github.com/josephryan/Hernandez_Ryan_2021_Recoding) and Zenodo (DOI:10.5281/zenodo.4660589). To maximize transparency and minimize confirmation bias, all analyses were pre-planned using phylotocol (DeBiasse and Ryan 2018) and pre-registered using the Center for Open Science’s pre-registration platform (https://osf.io/smj6k/ and https://osf.io/6ubgj/). Prior to the initial submission of this manuscript, we made four changes to the original plan outlined in our phylotocol. Details of changes and all versions of our phylotocol are available in our GitHub repository (see section 5 "Amendment History" in the phylotocol). Briefly, our changes included (1) adding tests of compositional heterogeneity to our original plan.
to test saturation, (2) incorporating P4 after realizing that Seq-Gen was not well suited for testing compositional heterogeneity, (3) adding deep splits evaluation criteria, and (4) adding statistical tests and testing alternative Dayhoff strategies. The latest version of our phylotocol includes all of these changes along with the additional analyses we made in response to reviews of our manuscript by 3 reviewers (prior to running new analyses).

**Overview of Empirical Datasets Employed**

The following methods can be divided into two main analyses: compositional heterogeneity and saturation. Both analyses employ empirical data from the following papers: Chang et al. (2015) hereafter “Chang,” and Feuda et al. (2017) hereafter “Feuda.” The topologies from Chang and Feuda are based on the same dataset which is made up of 51,940 amino acid positions from 78 taxa representing a wide range of animals and 9 non-animal outgroups. Feuda extensively applied 6-state amino acid recoding to this dataset in a reanalysis of the Chang study, which did not use recoding.

For the compositional heterogeneity analysis, we use several hypothetical 20-taxon symmetrical trees which consist of 4 clades (named clade-A, clade-B, clade-C, and clade-D) made up of 5 taxa each (Fig. 1a), and apply global parameters estimated from the Chang dataset. For the saturation analysis, we use the topologies reported in Chang and Feuda. More details on these analyses are provided below.

**Testing 6-state Recoding Performance on Compositional Heterogeneity**

We used the script comphet.pl (available in our GitHub repository) to simulate amino acid data in P4 (Foster 2004) on four hypothetical 20-taxon balanced trees (Fig. 1a). We chose
P4 because it specializes in simulating data in which amino acid (or nucleotide) composition varies across the tree. Using the amino acid rates of substitution estimated from the Chang dataset, we simulated sequences that were 1,000 amino acids in length under the GTR model. To introduce compositional heterogeneity, we used a balanced tree and generated one set of amino acid frequencies for clade-A and clade-C and a different set of frequencies for clade-B and clade-D. We paired amino acids by starting with the order of the 20 amino acids as they are commonly used as input to standard phylogenetic programs (i.e., A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V), divided them in half (i.e., [A-I] and [L-V]), and paired the two groups (i.e., (A,L), (R,K), (N,M) (D,F), (F,P), (Q,S), (E,T), (G,W), (H,Y), (I,V)). For clade-B and clade-D, we used the amino acid frequencies estimated from the Chang dataset (Table S1). For clade-A and clade-C we added X to the amino acid in each of the 10 frequency pairs that had the lowest frequency in the Chang dataset and subtracted X from the other, where X is the inflation parameter (i.e., 0.1, 0.5, 0.9) multiplied by the lowest frequency of the pair.

For example, the Chang frequencies for the amino acids R and K are 0.063 and 0.080 respectively. These frequencies were used for clade-B and clade-D without adjustment. To determine the increment value X under the inflation parameter 0.1, we multiplied the frequency of R, which is the lowest of the pair, by 0.1 (X=0.0063). We then added X to the Chang frequency of R (0.063 + 0.0063) and subtracted X from the Chang frequency of K (0.080 - 0.0063). We rounded these values to 3 decimal places (because P4 requires frequencies to add up to 1 and the sum of non-rounded frequencies was often slightly above or below 1) for a final set of frequencies of R = 0.069 and K=0.074. See pseudocode in the supplementary material or the CompHet.pm module in our GitHub repository for the code used to implement this strategy. See supplementary materials for comparisons of results using 1,000 random pairing strategies that
show that the paring strategy described in the previous paragraph does not bias the results in favor of non-recoding. We recoded each simulated dataset with both Dayhoff 6-state recoding and S&R 6-state recoding (Table S2), and then reconstructed maximum-likelihood trees of these recoded datasets using the GTR multi-state model and of the non-recoded datasets using the Dayhoff and JTT models in RAxML (Stamatakis 2014). We calculated Robinson-Foulds distances (Robinson and Foulds 1981) between each of the resultant 48,000 phylogenies and the trees used for simulation using TOPD/FMTS (Puigbo et al. 2007). We also scored trees based on deep splits, a custom metric (see the is_mono.pl script in the GitHub repository) that evaluates the monophyly of the clade that includes clade-A and clade-B (this evaluation, by definition, also includes the monophyly of the clade that includes clade-C and clade-D). The rationale for this metric rather than Robinson-Foulds distances is that it focused on errors that were most likely due to convergent amino acid compositions (i.e., the pulling together of compositionally homogeneous but unrelated clades or tips). We evaluated deep split accuracy for each combination of model, recoding type (including no recoding), and level of applied compositional heterogeneity (i.e., inflation parameter). We performed chi-squared tests to compare the number of incorrect trees between non-recoding and recoding approaches. To correct for multiple chi-squared testing, we applied the Bonferroni correction at which $\alpha = 0.002$.

**Testing 6-state Recoding Performance on Saturation**

We used Seq-Gen (Rambaut and Grass 1997) to simulate the evolution of amino acids on the Chang and Feuda trees (incorporating both topology and branch-length estimates). We chose Seq-Gen because it has a branch length scaling factor parameter that allows for straight-forward
introduction of saturation into simulations. We confirmed that increasing the branch length scaling factor parameter in Seq-Gen linearly increased levels of saturation (Fig. S1) using the script seq-gen_saturation_test.pl (available in the accompanying GitHub repository). Next, we performed 1,000 simulations per combination of tree (Chang and Feuda), branch length scaling factor parameter (1–20), and model of amino acid substitution (either Dayhoff or JTT) for a total of 80,000 datasets. We simulated an additional 1,000 datasets on the Chang topology for a subset of branch length scaling factor parameters (1, 5, 10, 15, 20) under the GTR model using the amino acid rates of substitution, amino acid frequencies (up to three decimal places as in our P4 analysis), and gamma rate heterogeneity estimated from the Chang dataset with maximum-likelihood (see shell script run_seqgen_estimated_model.sh for detailed parameters), bringing the grand total to 85,000 datasets. Each dataset included 1,000 amino acid columns.

For simulations performed on the Chang tree, we increased the branch length scaling factor parameter from 1 to 20 in increments of 1. The Feuda tree was produced from recoding the Chang dataset (Feuda et al. 2017), and because trees produced from recoded data have substantially fewer substitutions and therefore shorter branch lengths, we multiplied each branch length on the Feuda tree by 2.6 (based on our calculation that the sum of branch lengths in the recoded tree was 2.6 times shorter than the sum of branch lengths in the non-recoded Chang tree).

We performed maximum-likelihood analyses with RAxML for each set of sequences produced from simulations over the Chang and Feuda topologies. For the datasets simulated with Dayhoff and JTT substitution models, we reconstructed trees using the generating model, the 6-state recoding scheme derived from that model, and for a subset of branch length scaling factor parameters (1, 5, 10, 15, 20) we also reconstructed trees using LG, a sub-optimal model in this
context, as it was not the model used for the simulations. For the datasets simulated with the GTR substitution model, we generated trees using Dayhoff and Dayhoff 6-state recoding. We produced 180,000 phylogenies in total to test saturation. To test the performance of each recoding (or non-recoding) scheme, we calculated Robinson-Foulds distances between the topology used for simulation (i.e., Chang or Feuda) and the reconstructed trees generated from simulated sequences using TOPD/FMTS. We used a t-test to determine if there were significant differences in Robinson-Foulds distances between recoded and non-recoded datasets for each branch length scaling factor. To correct for multiple t-tests, we applied the Bonferroni correction at which $\alpha = 0.0009$.

**Testing Alternative Recoding Strategies on Compositional Heterogeneity**

To test the effect of the number of states on recoding, we developed alternative Dayhoff 9-, 12-, 15-, and 18-state recoding strategies. The first step in these analyses was to determine the optimal amino acid binning strategy for each number of tested states. Since the number of possible bins for each state is finite, ideally, we would use an exhaustive algorithm to identify the binning scheme that maximizes the sum of intra-bin substitution scores originating from the log odds matrix for PAM 250 (Dayhoff et al. 1978). Unfortunately, as pointed out by Susko and Roger (2007), the number of possible bins is very large (e.g., there are roughly $1.5 \times 10^{13}$ choices of bins under an 8-state recoding strategy) and an exhaustive algorithm is computationally intractable. Instead, we calculated the sum of intra-bin scores using the PAM 250 log odds matrix (see score.pl in our GitHub repository) for several binning schemes that incorporated subsets of the Dayhoff 6-state recoding bins and chose the best-scoring binning strategies from this set (Table S3). We also compared our best binning strategies to those proposed in Susko and...
Roger (2007) using the PAM 250 log odds matrix to calculate intra-bin substitution scores, and in all cases the scores we generated were higher, except for one which had an equal score (not entirely surprising given that the Susko and Roger bins were optimized for JTT recoding).

We compared the binning schemes that scored the highest for each recoding strategy (Table 2) against the Dayhoff and Dayhoff 6-state recoded matrices by testing their performance under reasonably high levels of compositional heterogeneity. We recoded the data that we simulated for the compositional heterogeneity analysis (data simulated with inflation parameter 0.5 using the hypothetical tree 0.002 (Fig. 1a)) using our Dayhoff 9-, 12-, 15-, and 18-state recoding strategies and reconstructed maximum-likelihood trees in RAxML. As in the main compositional heterogeneity analysis outlined above, we calculated deep splits scores (using the script is_mono.pl), to test the monophyly of the clade that included clade-A and clade-B and the clade that included clade-C and clade-D. We also performed a chi-squared test to compare the number of incorrect trees produced under Dayhoff-18 recoding (see Results for rationale) to those produced under non-recoding. To correct for multiple chi-squared testing, we applied the Bonferroni correction at which $\alpha = 0.017$.

**RESULTS**

*The Efficacy of 6-state Recoding under a Compositional Heterogeneity Gradient*

We simulated data with various levels of compositional heterogeneity by setting the amino acid frequencies of two non-sister 5-taxon clades (e.g., clade-A and clade-C in Figure 1a) to be highly divergent to the amino acid frequencies of the other two non-sister major clades (e.g., clade-B and clade-D in Figure 1a) on a balanced 20-taxon tree. We adjusted the level of compositional heterogeneity by increasing the frequency differences of each amino acid between
the two sets of frequencies by a factor that we call the inflation parameter. We adjusted the impact of introduced compositional heterogeneity by varying the length of the stem branches leading to those four clades (Fig. 1a). We tested the impact of sequence length by generating alignments of length 1,000, 2,000, 3,000, 4,000 and 5,000 (see supplementary material for methods). For each simulated dataset, we generated maximum-likelihood trees using recoding and non-recoding approaches. We scored these trees based on Robinson-Foulds distances from the true tree, as well as on whether a tree recovered the two major 10-taxon clades (i.e., a clade containing all clade-A and clade-B taxa and a clade containing all clade-C and clade-D taxa).

For each tree, we simulated 1,000,000 datasets with no introduced compositional heterogeneity (i.e., inflation parameter set to 0) to generate a null distribution of comp-het indices, to which we compared the compositionally heterogeneous datasets. We reconstructed trees on data simulated over hypothetical tree 0.002 for the first 1,000 out of these 1,000,000 datasets. In our phylogenetic analyses of these 1,000 datasets lacking compositional heterogeneity, recoded datasets performed consistently worse than non-recoded datasets (Fig. S2).

Analyses of non-recoded datasets consistently produced trees that were more accurate than those produced on recoded datasets using both our deep splits metric and Robinson-Foulds distances. Despite changes to the stem branch length on the tree and level of compositional heterogeneity implemented by the inflation parameter, non-recoding methods produced more accurate trees (Fig. 1b; Fig S3). While the performance of the recoding approaches diminished at a slower rate than non-recoding approaches (Fig. 1b) under increasing compositional heterogeneity, non-recoding performed significantly better than recoding in all cases tested, except under the highest level of compositional heterogeneity and shortest stem branch (Table
EVALUATING THE PERFORMANCE OF 6-STATE RECODING

S4; Table S5). We explored how data size impacted the performance of recoding methods in combination with compositional heterogeneity (details of these analyses are described in supplementary material). As sequence length increased, phylogenetic analyses of non-recoded datasets outperformed analyses of recoded datasets, except under the highest level of compositional heterogeneity (Fig. S4; i.e., inflation parameter = 0.9). Additionally, we explored the effect of tree shape and compositional heterogeneity on recoding methods (analyses described in supplementary material). These results were consistent in that non-recoding methods outperformed recoding under all levels of compositional heterogeneity tested (Fig. S6).

To gauge how our simulated data compared to real data in terms of the levels of compositional heterogeneity, we scored real and simulated datasets using the average relative compositional frequency variability (RCFV) score (Kück & Struck 2014). Higher RCFV scores indicate greater variability in amino acid composition across a dataset. We found that the level of compositional heterogeneity (as measured by RCFV) in datasets simulated with the inflation parameter set to 0.9 was substantially higher than the majority of real datasets. The median RCFV score was 0.088 for all datasets simulated under the inflation parameter of 0.9, while the median RCFV score for data from papers in Table 1 was 0.036 (Fig. S5). We reason that our simulated datasets therefore are substantially compositionally heterogeneous since these published datasets, many of which used compositional heterogeneity as justification for the application of recoding, are likely enriched for compositional heterogeneity.

The Efficacy of 6-state Recoding under a Saturation Gradient

We simulated datasets on the Chang and Feuda trees under the Dayhoff and JTT models with increasing levels of saturation. Under all tested levels of saturation, phylogenetic
reconstructions using the Dayhoff and LG models on non-recoded data matrices that were
simulated under the Dayhoff model produced trees with fewer errors on average (as measured by
Robinson-Foulds distances from the true tree) than those that used the Dayhoff 6-state recoded
matrix (Fig. 2a). The results were similar for data simulated under the JTT model, where trees
reconstructed with the JTT and LG models on non-recoded data matrices contained fewer errors
on average across all tested levels of saturation compared to reconstructions with the S&R 6-
state recoded matrix (Fig. 2b). The results were consistent regardless of which tree (i.e., Chang
or Feuda) was used for data simulations (Fig. S7). As saturation increased, the performance of
recoding approaches decreased at a faster rate than non-recoding approaches (Fig. S7). T-tests
performed for each branch length scaling factor parameter showed that Robinson-Foulds
distances were significantly higher for recoded datasets compared to non-recoded datasets (p-
value < 2.2e-16).

We also simulated data under the GTR model using the amino acid rates of substitution,
amino acid frequencies, and gamma rate heterogeneity parameters estimated from the Chang
dataset. Phylogenetic analyses of data simulated under GTR resulted in fewer errors on average
when reconstructed with non-recoded Dayhoff matrices compared to reconstructions with the
Dayhoff 6-state recoded matrices (Fig. 2c). T-tests carried out for each branch length scaling
factor parameter indicated that recoded approaches performed significantly worse than non-
recoded approaches (p-value < 2.2e-16).

Further, we tested the combined effects of sequence length and saturation on the
performance of recoding strategies (see supplementary material for methods). Increases in
sequence length minimized the impact of saturation and reduced errors in phylogenetic
reconstruction for both recoding and non-recoding methods. However, non-recoding methods
performed significantly better on all sequence lengths and levels of saturation, except for on the largest simulated dataset with the lowest level of saturation where results from recoded and non-recoded analyses were equivocal (Fig. S8; Table S6).

The Effect of Alternative Recoding Strategies on Compositional Heterogeneity

We used the data simulated under inflation parameter 0.5 (mid-level of compositional heterogeneity) using the hypothetical tree 0.002 (short stem branches; Fig. 1a) from the main compositional heterogeneity analysis to test Dayhoff 9-, 12-, 15-, and 18-state recoding strategies and compared the performance of these methods to Dayhoff 6-state recoding and non-recoding. As in the main compositional heterogeneity analysis outlined above, trees were assessed by deep splits to determine if they recovered the two compositionally heterogeneous 10-taxon clades (i.e., a monophyletic group of clade-A and clade-B, and a monophyletic group of clade-C and clade-D). The percentage of trees that passed these criteria increased as the number of Dayhoff states increased with Dayhoff 18-state recoding outperforming all other strategies including the non-recoding approach (Fig. 3). Non-recoding outperformed all other recoding strategies except Dayhoff 12- and 15-state recoding under the highest level of compositional heterogeneity (inflation parameter 0.9; Fig. 3c). We performed a chi-squared test to determine if the differences in numbers of incorrect trees between analyses run with Dayhoff 18-state recoding and those run without recoding were significant. The difference was significant only under the highest level of compositional heterogeneity (p-values for inflation parameters 0.1, 0.5, and 0.9: 0.4314, 0.2183, and 6.622e-06 respectively).

DISCUSSION
The philosophy underlying recoding strategies in phylogenetics is that sacrificing some information is beneficial in cases where homoplasy is high, as is the case when there is substantial heterogeneity in nucleotide or amino acid composition or when datasets are highly saturated. Six-state amino acid recoding has been proposed as a strategy to improve phylogenetic reconstruction in the presence of compositional heterogeneity and saturation (Embley et al. 2003a; Hrdy et al. 2004; Martin et al. 2005). While there have been simulation analyses that compare different binning schemes (Susko and Roger 2007; Nesnidal et al. 2010), there are few if any studies that compare the accuracy of 6-state recoding to non-recoding approaches. In this study, we used simulations under gradients of compositional heterogeneity and saturation to compare the performance of 6-state amino acid recoding strategies. Remarkably, we found that non-recoding approaches outperformed 6-state recoding approaches in all of our comparisons. Our results show that while 6-state recoding seems to be less affected by increases in compositional heterogeneity, it does not overcome the penalty of information loss even under the highest levels of compositional heterogeneity (Fig. 1b). Further, we found that 6-state recoding performs poorly when applied to highly saturated datasets. As such, we conclude that the costs of information loss associated with the 6-state recoding schemes are too great to justify applying these strategies.

We confirm that Dayhoff 6-state recoding is inappropriate for phylogenetic inference and our analyses with S&R 6-state recoding show that limitations extend beyond Dayhoff matrices, as 6-states likely are too few for reliable phylogenetic analysis. It is possible that not all recoding strategies are inappropriate. Specifically, we found that our Dayhoff 9-, 12-, 15-, and 18-state recoding strategies performed better than the standard Dayhoff 6-state recoding approach for all tested levels of compositional heterogeneity (Fig. 3). Dayhoff 18-state recoding performed the
best under all gradients of compositional heterogeneity and may comprise the optimum balance of minimizing compositional heterogeneity while maximizing information retention. However, we do not advocate blindly applying Dayhoff 18-state recoding, especially since significant improvement only occurs under the most extreme compositional heterogeneity setting (0.9), which we show is uncommon in real datasets based on RCFV scores (RCFV scores $\geq 0.1$ occurred in 6 out of 25 sampled publications; Table S7). Nevertheless, conservative recoding approaches under very high levels of compositional heterogeneity may be justified provided that these approaches are properly tested.

Applying a recoding method that is dataset specific may be another tactic to handle compositional heterogeneity or saturation. Susko and Roger (2007) and Nesnidal et al. (2010) applied this strategy by testing several recoding binning schemes informed by their datasets of interest. Tailoring the level and/or type of recoding to the amount of compositional heterogeneity and saturation, perhaps on a column-by-column basis, may be a successful approach, but further testing using such a tailored method would be necessary. Since only a handful of studies have investigated different recoding schemes, it is clear that more analyses are required to gain an understanding of the impact of alternative recoding methods for compositionally heterogeneous and/or saturated datasets.

**Implications**

There are at least 91 publications that use 6-state amino acid recoding, with 2019 seeing more than any year to date (Table 1). Many of these studies have proposed controversial topologies with profound implications across the tree of life including bacteria, archaea, unicellular eukaryotes, fungi, animals, and plants. We have shown that 6-state recoding greatly
reduces information content and therefore often results in suboptimal phylogenetic
reconstructions. We suggest that these datasets should be reevaluated using criteria that assess
the amount of compositional heterogeneity within datasets, and/or reanalyzed using non-
recoding approaches unless extreme levels of compositional heterogeneity are evident. When
applying recoding, it would be beneficial to determine the number of states in a recoding strategy
based on the level of compositional heterogeneity using approaches that we have applied in this
study. Nevertheless, we advocate caution when interpreting results stemming from analyses that
have employed 6-state recoding and contend that published analyses in which 6-state recoding
approaches substantially influenced the conclusions might need to be revisited.

SUPPLEMENTARY MATERIAL

All commands and versions of software used in this study are provided in the supplementary
material. All scripts are available in the following GitHub and Zenodo repositories:
https://github.com/josephryan/Hernandez_Ryan_2021_Recoding

FUNDING

This work was supported by the National Science Foundation under Grant Number 1542597; and
the Graduate Research Fellowship Program to A.M.H. Additional funding to A.M.H. was
provided by the Florida Education Fund Mcknight Doctoral Fellowship Program. The funders
had no role in study design, data collection and analysis, decision to publish, or preparation of
the manuscript.
ACKNOWLEDGEMENTS

We thank Melissa DeBiasse for providing comments on an earlier version of the manuscript and Gordon Burleigh, Christine Schnitzler, Marta Wayne, and Bryan Kolaczkowski for feedback on this project. The authors would like to express their thanks to David Swofford and Gavin Naylor for influential discussions at an early stage of the project. We also thank Michal Kowalewski for providing guidance on statistics employed and Luis Vargas for assistance in troubleshooting scripts. Lastly, we thank UF SACNAS for providing a supportive space for scientific growth and encouragement. The views expressed in this paper do not necessarily reflect the views of those acknowledged.

REFERENCES


Analysis Disentangles the Complex Evolutionary History of DNA Replication in Archaea.
Rodriguez-Ezpeletta N., Embley T.M. 2012. The SAR11 Group of Alpha-Proteobacteria Is Not
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., Smith S.A., Putnam N.H., Haddock S.H.D., Dunn C.W.,
Wolfsberg T.G., Mullikin J.C., Martindale M.Q., Baxevanis A.D. 2013. The Genome of the
Ctenophore Mnemiopsis leidyi and Its Implications for Cell Type Evolution. Science.
342:1242592.
Schwentner M., Combosch D.J., Pakes Nelson J., Giribet G. 2017. A Phylogenomic Solution to
Schwentner M., Richter S., Rogers D.C., Giribet G. 2018. Tetracronatan phylogeny with special
focus on Malacostraca and Branchiopoda: highlighting the strength of taxon-specific
Shen X.-X., Todd Hittinger C., Rokas A. 2017. Contentious relationships in phylogenomic
studies can be driven by a handful of genes. Nat. Ecol. Evol. 1:0126.
Shin S., Clarke D.J., Lemmon A.R., Moriarty Lemmon E., Aitken A.L., Haddad S., Farrell B.D.,
Marvaldi A.E., Oberprieler R.G., McKenna D.D. 2017. Phylogenomic Data Yield New and
Simion P., Philippe H., Baurain D., Jager M., Richter D.J., Di Franco A., Roure B., Satoh N.,
Quéinnec É., Ereskovsky A., Lapébie P., Corre E., Delsuc F., King N., Wörheide G.,
Manuel M. 2017. A Large and Consistent Phylogenomic Dataset Supports Sponges as the
mitogenome data in deciphering deep phylogeny of Holometabola. J. Syst. Evol. 54:545–
559.
phylogenies support the monophyly of the three bryophyte groups (Bryophyta Schimp.).
Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of
in phylogenetic estimation and its relevance to the choice between parsimony and likelihood


**Figure and Table Legends**

**Figure 1.** Six-state recoding approaches produce more incorrect trees under various levels of compositional heterogeneity. (a) Trees used for simulations. The value in the name of the tree (e.g., 0.008 in Tree 0.008) denotes the length in substitutions per site of the stem branches of clade-A and clade-B, and stem branches of clade-C and clade-D (highlighted in orange). (b) Percentage of 1,000 trees that did not reconstruct a monophyletic group of taxa from clade-A and clade-B and monophyletic group of taxa from clade-C and clade-D.

**Figure 2.** Six-state recoding approaches produce more errors under increasing levels of saturation. Robinson-Foulds distances were calculated for 1,000 runs for each branch length scaling factor parameter. All data were simulated on the Chang tree. (a) Datasets simulated under the Dayhoff model. (b) Datasets simulated under the JTT model. (c) Datasets simulated under the GTR model using the amino acid rates of substitution, amino acid frequencies, and gamma rate heterogeneity estimated from the Chang dataset.

**Figure 3.** Dayhoff 9-, 12-, 15-, and 18-state recoding produce fewer incorrect trees than Dayhoff 6-state recoding under various levels of compositional heterogeneity. Trees were reconstructed by applying the non-recoded (NR) Dayhoff matrix or alternative Dayhoff recoding strategies (the number of states in the recoding strategy is indicated by digits). Incorrect trees did not include a monophyletic group of taxa from clade-A and clade-B and monophyletic group of taxa from clade-C and clade-D. The Y-axis refers to percentage out of 1,000 trees.

**Table 1.** Publications that use 6-state amino acid recoding. Asterisk indicates the publication included recoding approaches in a main figure to test if this strategy was appropriate.

**Table 2.** Best scoring binning schemes optimized on the Dayhoff matrix.
<table>
<thead>
<tr>
<th>Citation</th>
<th>Recoding in main figure</th>
<th>Organismal scope or featured taxon</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benavides et al. (2021)</td>
<td>yes</td>
<td>Gonyleptoidea</td>
</tr>
<tr>
<td>Li et al. (2020)</td>
<td>yes</td>
<td>Bivalves</td>
</tr>
<tr>
<td>Neumann et al. (2020)</td>
<td>yes</td>
<td>Metazoa</td>
</tr>
<tr>
<td>Pandey &amp; Braun (2020)</td>
<td>yes</td>
<td>Metazoa</td>
</tr>
<tr>
<td>Tikhonenkov et al. (2020)</td>
<td>yes</td>
<td><em>Tunicaraptor unikontum</em></td>
</tr>
<tr>
<td>Weinheimer &amp; Aylward (2020)</td>
<td>yes</td>
<td>Caudovirales</td>
</tr>
<tr>
<td>Yan et al. (2020)</td>
<td>yes</td>
<td>Flesh flies</td>
</tr>
<tr>
<td>Cunha &amp; Giribet (2019)</td>
<td>yes</td>
<td>Gastropods</td>
</tr>
<tr>
<td>Laumer et al. (2019)</td>
<td>yes</td>
<td>Animals</td>
</tr>
<tr>
<td>Lawrence et al. (2019)</td>
<td>yes*</td>
<td>Plastids</td>
</tr>
<tr>
<td>Lemer et al. (2019)</td>
<td>yes</td>
<td>Bivalves</td>
</tr>
<tr>
<td>Lozano-Fernandez et al. (2019)</td>
<td>yes</td>
<td>Chelicerates</td>
</tr>
<tr>
<td>Marlétaz et al. (2019)</td>
<td>yes</td>
<td>Spiralia</td>
</tr>
<tr>
<td>Philippe et al. (2019)</td>
<td>yes</td>
<td>Bilateria</td>
</tr>
<tr>
<td>Ballesteros et al. (2019)</td>
<td>no</td>
<td>Palpigradi</td>
</tr>
<tr>
<td>Benavides et al. (2019)</td>
<td>no</td>
<td>Pseudoscorpiones</td>
</tr>
<tr>
<td>Cheng et al. (2019)</td>
<td>no</td>
<td>Zygnematophyceae</td>
</tr>
<tr>
<td>Klinges et al. (2019)</td>
<td>no</td>
<td>Candidatus Aquarickettsia</td>
</tr>
<tr>
<td>Moore et al. (2019)</td>
<td>no</td>
<td>Plastids</td>
</tr>
<tr>
<td>Narayanan Kutty et al. (2019)</td>
<td>no</td>
<td>Calyptratae</td>
</tr>
<tr>
<td>Uribe et al. (2019)</td>
<td>no</td>
<td>Gastropods</td>
</tr>
<tr>
<td>Wolfe et al. (2019)</td>
<td>no</td>
<td>Decapod crustaceans</td>
</tr>
<tr>
<td>Zverkov et al. (2019)</td>
<td>no</td>
<td>Dicyemida and Orthonectida</td>
</tr>
<tr>
<td>Aouad et al. (2018)</td>
<td>yes</td>
<td>Archaea</td>
</tr>
<tr>
<td>Laumer et al. (2018)</td>
<td>yes</td>
<td>Placozoa</td>
</tr>
<tr>
<td>Otero-Bravo et al. (2018)</td>
<td>yes</td>
<td>Pantoea</td>
</tr>
<tr>
<td>Puttick et al. (2018)</td>
<td>yes</td>
<td>Land plants</td>
</tr>
<tr>
<td>Schwentner et al. (2018)</td>
<td>yes</td>
<td>Pancrustacea</td>
</tr>
<tr>
<td>Sousa et al. (2018)</td>
<td>yes</td>
<td>Land plants</td>
</tr>
<tr>
<td>Bennett &amp; Mao (2018)</td>
<td>no</td>
<td>Fulgoroidea symbionts</td>
</tr>
<tr>
<td>Eitel et al. (2018)</td>
<td>no</td>
<td>Placozoa</td>
</tr>
<tr>
<td>Manzano-Marín et al. (2018)</td>
<td>no</td>
<td><em>Cinara strobi</em> symbionts</td>
</tr>
<tr>
<td>Feuda et al. (2017)</td>
<td>yes</td>
<td>Animals</td>
</tr>
<tr>
<td>Szabó et al. (2017)</td>
<td>yes</td>
<td>Pseudococcidae symbionts</td>
</tr>
<tr>
<td>Authors</td>
<td>Result</td>
<td>Domain</td>
</tr>
<tr>
<td>-------------------------</td>
<td>--------</td>
<td>------------------------------</td>
</tr>
<tr>
<td>Williams et al. (2017)</td>
<td>yes</td>
<td>Archaea</td>
</tr>
<tr>
<td>Schwentner et al. (2017)</td>
<td>no</td>
<td>Pancrustacea</td>
</tr>
<tr>
<td>Shin et al. (2017)</td>
<td>no</td>
<td>Curculionoidea</td>
</tr>
<tr>
<td>Simion et al. (2017)</td>
<td>no</td>
<td>Animals</td>
</tr>
<tr>
<td>Yoshida et al. (2017)</td>
<td>no</td>
<td>Tardigrades</td>
</tr>
<tr>
<td>Leliaert et al. (2016)</td>
<td>yes</td>
<td>Viridiplantae</td>
</tr>
<tr>
<td>Zhang et al. (2016)</td>
<td>yes</td>
<td>Roseobacter CHAB-I-5 lineage</td>
</tr>
<tr>
<td>He et al. (2016)</td>
<td>no</td>
<td>Rhizaria</td>
</tr>
<tr>
<td>Song et al. (2016)</td>
<td>no</td>
<td>Holometabola</td>
</tr>
<tr>
<td>Domman et al. (2015)</td>
<td>yes</td>
<td>Plastids</td>
</tr>
<tr>
<td>Luo (2015)</td>
<td>yes</td>
<td>SAR11</td>
</tr>
<tr>
<td>Petitjean et al. (2015)</td>
<td>yes</td>
<td>Archaea</td>
</tr>
<tr>
<td>Borowiec et al. (2015)</td>
<td>no</td>
<td>Animals</td>
</tr>
<tr>
<td>Derelle et al. (2015)</td>
<td>no</td>
<td>Eukaryotes</td>
</tr>
<tr>
<td>Wang &amp; Wu (2015)</td>
<td>no</td>
<td>Mitochondria</td>
</tr>
<tr>
<td>Luo et al. (2014)</td>
<td>yes</td>
<td>Roseobacter</td>
</tr>
<tr>
<td>Fu et al. (2014)</td>
<td>no</td>
<td>Discoba</td>
</tr>
<tr>
<td>Lemieux et al. (2014)</td>
<td>no</td>
<td>Trebouxiophyceae</td>
</tr>
<tr>
<td>Raymann et al. (2014)</td>
<td>no</td>
<td>Archaea</td>
</tr>
<tr>
<td>Luo et al. (2013)</td>
<td>yes</td>
<td>Marine Alphaproteobacteria</td>
</tr>
<tr>
<td>Morgan et al. (2013)</td>
<td>yes</td>
<td>Placental mammals</td>
</tr>
<tr>
<td>Rota-Stabelli et al. (2013)</td>
<td>yes</td>
<td>Pancrustacea</td>
</tr>
<tr>
<td>Hill et al. (2013)</td>
<td>no</td>
<td>Demospongiae</td>
</tr>
<tr>
<td>Kayal et al. (2013)</td>
<td>no</td>
<td>Cnidaria</td>
</tr>
<tr>
<td>Lasek-Nesselquist &amp; Gogarten (2013)</td>
<td>no</td>
<td>3 domains (eukaryotes, archaea, bacteria)</td>
</tr>
<tr>
<td>Ometto et al. (2013)</td>
<td>no</td>
<td>Drosophila suzukii</td>
</tr>
<tr>
<td>Lasek-Nesselquist (2012)</td>
<td>yes</td>
<td>Syndermata</td>
</tr>
<tr>
<td>Rodríguez-Ezepeleta &amp; Embley (2012)</td>
<td>yes</td>
<td>SAR11</td>
</tr>
<tr>
<td>Burki et al. (2012)</td>
<td>no</td>
<td>Plastids</td>
</tr>
<tr>
<td>Derelle &amp; Lang (2012)</td>
<td>no</td>
<td>Eukaryotes</td>
</tr>
<tr>
<td>Heinz et al. (2012)</td>
<td>no</td>
<td>Trachipleistophora hominis</td>
</tr>
<tr>
<td>Nishimura et al. (2012)</td>
<td>no</td>
<td>Mitochondria</td>
</tr>
<tr>
<td>Brochier-Armanet et al. (2011)</td>
<td>yes</td>
<td>Archaea</td>
</tr>
<tr>
<td>Publication</td>
<td>Use 6-state recoding</td>
<td>Clade</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>----------------------</td>
<td>------------------------------</td>
</tr>
<tr>
<td>Williams et al. (2011)</td>
<td>yes</td>
<td>Nucleocytoplasmic large DNA virus</td>
</tr>
<tr>
<td>Matsumoto et al. (2011)</td>
<td>no</td>
<td>Plastids</td>
</tr>
<tr>
<td>Philippe et al. (2011)</td>
<td>no</td>
<td>Xenacoelomorpha</td>
</tr>
<tr>
<td>Wodniok et al. (2011)</td>
<td>no</td>
<td>Streptophyte algae and land plants</td>
</tr>
<tr>
<td>Torruella et al. (2011)</td>
<td>no</td>
<td>Opisthokonta</td>
</tr>
<tr>
<td>Parfrey et al. (2010)</td>
<td>no</td>
<td>Eukaryotes</td>
</tr>
<tr>
<td>Pons et al. (2010)</td>
<td>no</td>
<td>Coleoptera</td>
</tr>
<tr>
<td>Deschamps &amp; Moreira (2009)</td>
<td>yes</td>
<td>Archaeplastida</td>
</tr>
<tr>
<td>Foster et al. (2009)</td>
<td>yes</td>
<td>Eukaryotes</td>
</tr>
<tr>
<td>Masta et al. (2009)</td>
<td>yes</td>
<td>Arachnida</td>
</tr>
<tr>
<td>Cox et al. (2008)</td>
<td>yes</td>
<td>Eukaryotes</td>
</tr>
<tr>
<td>Haen et al. (2007)</td>
<td>no</td>
<td>Hexactinellida</td>
</tr>
<tr>
<td>Andersson et al. (2006)</td>
<td>yes</td>
<td>Eukaryotes</td>
</tr>
<tr>
<td>Fitzpatrick et al. (2006a)</td>
<td>yes</td>
<td>Mitochondria</td>
</tr>
<tr>
<td>Fitzpatrick et al. (2006b)</td>
<td>yes</td>
<td>Fungi</td>
</tr>
<tr>
<td>O’Halloran et al. (2006)</td>
<td>yes</td>
<td>Caenorhabditis elegans</td>
</tr>
<tr>
<td>Delsuc et al. (2006)</td>
<td>no</td>
<td>Chordates</td>
</tr>
<tr>
<td>Wang &amp; Lavrov (2006)</td>
<td>no</td>
<td>Homoscleromorpha</td>
</tr>
<tr>
<td>Martin et al. (2005)</td>
<td>yes</td>
<td>Land plants</td>
</tr>
<tr>
<td>Philip et al. (2005)</td>
<td>no</td>
<td>Eukaryotes</td>
</tr>
<tr>
<td>Hrdy et al. (2004)</td>
<td>yes</td>
<td>Hydrogenosomes</td>
</tr>
<tr>
<td>Embley et al. (2003a)</td>
<td>yes</td>
<td>Hydrogenosomes</td>
</tr>
<tr>
<td>Embley et al. (2003b)</td>
<td>yes</td>
<td>Hydrogenosomes</td>
</tr>
<tr>
<td>Davidson et al. (2002)</td>
<td>yes</td>
<td>Hydrogenosomes</td>
</tr>
</tbody>
</table>

**Table 1.** Publications that use 6-state amino acid recoding. Asterisk indicates the publication included recoding approaches in a main figure to test if this strategy was appropriate.
<table>
<thead>
<tr>
<th>Dayhoff recoding</th>
<th>Binning scheme</th>
</tr>
</thead>
<tbody>
<tr>
<td>9-state</td>
<td>DEHNQ ILMV FY AST KR G P C W</td>
</tr>
<tr>
<td>12-state</td>
<td>DEQ MLIV FY KHR G A P S T N W C</td>
</tr>
<tr>
<td>15-state</td>
<td>DEQ ML IV FY G A P S T N K H R W C</td>
</tr>
<tr>
<td>18-state</td>
<td>ML FY I V G A P S T D E Q N H K R W C</td>
</tr>
</tbody>
</table>

**Table 2.** Best scoring binning schemes optimized on the Dayhoff matrix.