

An extension of testing 6-state recoding strategies to address substitution bias and compositional heterogeneity

Principal Investigator: Joseph Ryan and Alexandra Hernandez

Draft or Version Number: v.1.0

25 May 2018

LIST OF ABBREVIATIONS

JTT	Jones-Thornton-Taylor (a model for amino acid substitution)
PAM	Point accepted mutation (a model for amino acid substitution)
RFD	Robinson-foulds distance
SICB	Society for Integrative and Comparative Biology
S&R-6	Susko & Roger 6-state recoding
RAxML	Randomized Axelerated Maximum Likelihood
TOPD	TOPological Distance
FMTS	From Multiple to Single
GTR	general time-reversible model
LG	Le & Gascuel (a model for amino acid substitution)
ML	Maximum-likelihood

1 INTRODUCTION: BACKGROUND INFORMATION AND SCIENTIFIC RATIONALE**1.1 BACKGROUND INFORMATION**

Dayhoff, JTT, and LG matrices are 20-state amino acid replacement models used to score amino acid substitutions in phylogenetic analyses. Recently, recoding techniques have been employed to address difficulties that these models have dealing with compositional heterogeneity and substitution saturation (Susko & Roger 2007). Dayhoff recoding (i.e., Dayhoff-6) specifically recodes amino acids from Dayhoff matrices according to 6 groups of chemically related amino acids that frequently replace one another (Hrdy et al. 2004), while JTT recoding (i.e., S&R-6) is a 6-state recoding strategy based off binning experiments on the JTT model by Susko & Roger (2007).

1.2 RATIONALE

The principle of using recoding to address substitution saturation and compositional heterogeneity is appealing from a theoretical perspective but has never been tested empirically. Evidence from our analysis presented at the SICB 2018 Conference showed that under all simulations in the study, Dayhoff-6 recoding performed worse than the PAM250 (Dayhoff) model. These preliminary results raised doubts about the benefits of using recoding approaches. However, we received feedback that our analyses did not directly address compositional heterogeneity and that since we used the same model for simulation and testing, we did not consider problems stemming from poor model fit. The proposed analyses herein aim to address these concerns.

1.3 OBJECTIVES

This is a three-part study that will serve as an extension to our project investigating 6-state recoding strategies (https://github.com/josephryan/Hernandez_Ryan_2018_RecodingSim). The objectives are 1) determine how recoding performs compared to a model that was not used for simulation (i.e. does recoding improve results when the non-recoded model fit is poor?) 2) verify the performance of recoding strategies using inferred model parameters from the data under a range of saturation levels 3) determine if recoding addresses problems with compositional heterogeneity.

2 STUDY DESIGN AND ENDPOINTS

2.1 Effect of model fit on recoding vs. non-recoding using LG model

2.1.1. Reconstruct topologies in RAXML (Stamatakis 2014) using the LG model to estimate trees for the data generated in the previous analysis via simulations in Seq-Gen (Rambaut & Grassly 1997) (https://github.com/josephryan/Hernandez_Ryan_2018_RecodingSim). These previous datasets were simulated over the Chang et al. (2015) phylogeny under both the Dayhoff and JTT models with branch-length scaling factors (-s) set to 1, 5, 10, 15, and 20.

```
raxmlHPC -p 420 -m PROTGAMMALG -n Chang.1.LG -s Chang.PAM.1.phy
```

2.1.2 Use the program TOPD/FMTS (Puigbó et al. 2007) to compute Robinson-Foulds distances (RFDs) between LG-generated trees and the tree from Chang et al. (2015) that was used to simulate the data. We will compare these RFDs to those from comparisons between the trees estimated using recoded amino acids.

```
cat Chang_orig_phylobayes.tre RAXML_best* > Chang_all_trees.tre
```

```
perl topd_v4.6.pl -f Chang_all_trees.tre -m split -r no
```

2.2 Effect of model fit on recoding vs. non-recoding using data simulated w/ estimated model

2.2.1 Simulate the evolution of amino acids along the phylogeny produced in Chang et al. (2015) using Seq-Gen and apply model parameters inferred from the data (i.e., amino acid frequency, alpha parameter, and transition rates taken from the main ML tree of Chang et al. 2015). We asked the authors of Chang et al. (2015) for their RAXML output files, but they did not have these available, so we will reestimate these values using the exact RAXML run that was used in the original study. The branch-length parameter (-s) for simulations will be set to 1, 5, 10, 15, and 20.

```
seq-gen -z420 -mGENERAL -r[rates] -f[frequency] -a[alpha] -n1000 -s1 -or  
Chang_orig_phylobayes.tre > Chang.mismatch.1.phy
```

```
seq-gen -z420 -mGENERAL -r[rates] -f[frequency] -a[alpha] -n1000 -s5 -or  
Chang_orig_phylobayes.tre > Chang.mismatch.5.phy
```

```
seq-gen -z420 -mGENERAL -r[rates] -f[frequency] -a[alpha] -n1000 -s10 -or  
Chang_orig_phylobayes.tre > Chang.mismatch.10.phy
```

```
seq-gen -z420 -mGENERAL -r[rates] -f[frequency] -a[alpha] -n1000 -s15 -or  
Chang_orig_phylobayes.tre > Chang.mismatch.15.phy
```

```
seq-gen -z420 -mGENERAL -r[rates] -f[frequency] -a[alpha] -n1000 -s20 -or  
Chang_orig_phylobayes.tre > Chang.mismatch.20.phy
```

divide.pl (For all 5 commands above, divide the 1000 datasets outputted by Seq-Gen into separate phylip files)

```
perl divide.pl Chang.mismatch.1.phy Chang.mismatch.1
```

2.2.2 Convert simulated sequences to Dayhoff-6 recoded datasets using the script chunkify2.pl. Chunkify2.pl also generates scripts to perform maximum-likelihood analyses in RAXML for each non-recoded dataset under the Dayhoff model and for each recoded dataset under the MULTIGAMMA multi-state model with GTR.

*Note: Chunkify2.pl requires the user to input data on servers being used and processors available for each server (this is hard-coded in the script and should be altered when reproducing results on another machine).

```
perl chunkify2.pl
```

To execute the scripts generated from chunkify2.pl for maximum-likelihood analyses of both recoded and non-recoded datasets:

```
ls -l servername* | perl -ne 'chomp; print "sh $_ &\n";' | sh
```

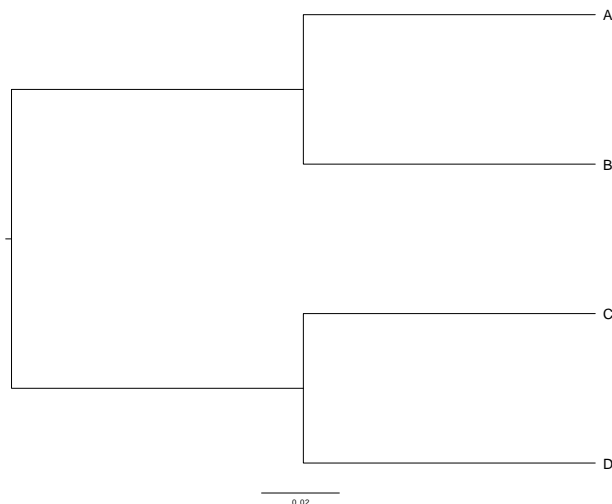
2.2.3 Concatenate trees from section 2.2.2 into one tree file along with the true tree used for simulation, then calculate RFDs in TOPD/FMTS to compare differences between trees. Box plots will be used to visually compare RFDs and t-tests will be used to determine if there are significant differences in RFDs between non-recoded and recoded trees in R (R Core Team 2017). The script compare_trees_to_sim.pl.v04 performs these analyses.

*Note: The directories in which to perform these analyses are hardcoded and should be altered when reproducing results on another machine.

```
perl compare_trees_to_sim.pl.v04
```

2.3 Effect of compositional heterogeneity

2.3.1 Simulate the evolution of amino acids 1,000,000 times along a four-taxa bifurcating tree (comp_het.tre) (shown below) using the model parameters inferred from the data in section 2.2.1.



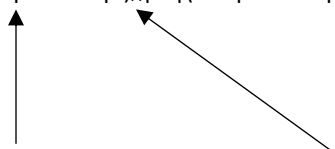
```
seq-gen -z420 -mGENERAL -r[rates] -f[frequency] -a[alpha] -n1000000 -s1 -or  
comp_het.tre > comp_het.phy
```

Divide the 1,000,000 datasets outputted by Seq-Gen into separate phylip files.

```
perl divide.pl comp_het.phy comp_het_sim
```

2.3.2 Use the script `bdac.pl` to assign scores to amino acids (from 1 to 20 in the following order: ACDEFGHIKLMNPQRSTVWY) for each taxon from the simulated alignments. The script will calculate a `bdac` index for each dataset, which is the total score for taxa B+D subtracted from taxa A+C. In other words, the `bdac` index is the score of taxa $((B+D)-(A+C))$. The 1,000 amino acid alignments with the highest `bdac` indices will be chosen as a way of selecting taxa with the greatest compositional bias (taxa A and C will be biased with amino acids in the lower alphabet range and differ considerably from taxa B and D, which will be biased with amino acids in the upper alphabet range). To ensure that taxa A+C and B+D are compositionally different from each other, we will compute amino acid frequencies (using the same script) and sum the total difference in amino acid frequencies between taxa A vs C and B vs D. We call this sum of differences a compositional heterogeneity index, or `comp-het` index for short. We will sum these 1,000 `comp-het` indices and use this sum as a test statistic in a Monte Carlo simulation. The simulation will generate 1,000 sums of 1,000 `comp-het` indices that are randomly selected from the original 1,000,000 datasets and calculate the number of these sums that are greater than our test statistic. If our test statistic is not significant (i.e. we identify more than 50 sums greater than our test statistic), we will continue to perform step 2.3.1 using an order of magnitude greater number of simulated datasets until we reach significance.

$$\text{comp-het index} = |(FreqA - FreqC)_A| + |(FreqA - FreqC)_C| + \dots |(FreqA - FreqC)_V|$$



Frequency for taxa A for amino acid Alanine

```
perl bdat.pl
```

2.3.3 Once we have determined that amino acid composition differs between taxa A+C and B+D we will use the four sequences resulting from each of the simulations as starting sequences to simulate four clades of 5-taxa each. We will then concatenate these four 5-taxa datasets into a single dataset. Conceptually, the end result will be to connect each of the four clades into a single tree (see figure 1).

```
seq-gen -z420 -mGENERAL -r[rates] -f[frequency] -a[alpha] -n1000 -s1 -k  
comp_hetA.phy -or comp_hetA.tre > comp_het_Asim.phy
```

```
seq-gen -z420 -mGENERAL -r[rates] -f[frequency] -a[alpha] -n1000 -s1 -k  
comp_hetB.phy -or comp_hetB.tre > comp_het_Bsim.phy
```

```
seq-gen -z420 -mGENERAL -r[rates] -f[frequency] -a[alpha] -n1000 -s1 -k  
comp_hetC.phy -or comp_hetC.tre > comp_het_Csim.phy
```

```
seq-gen -z420 -mGENERAL -r[rates] -f[frequency] -a[alpha] -n1000 -s1 -k  
comp_hetD.phy -or comp_hetD.tre > comp_het_Dsim.phy
```

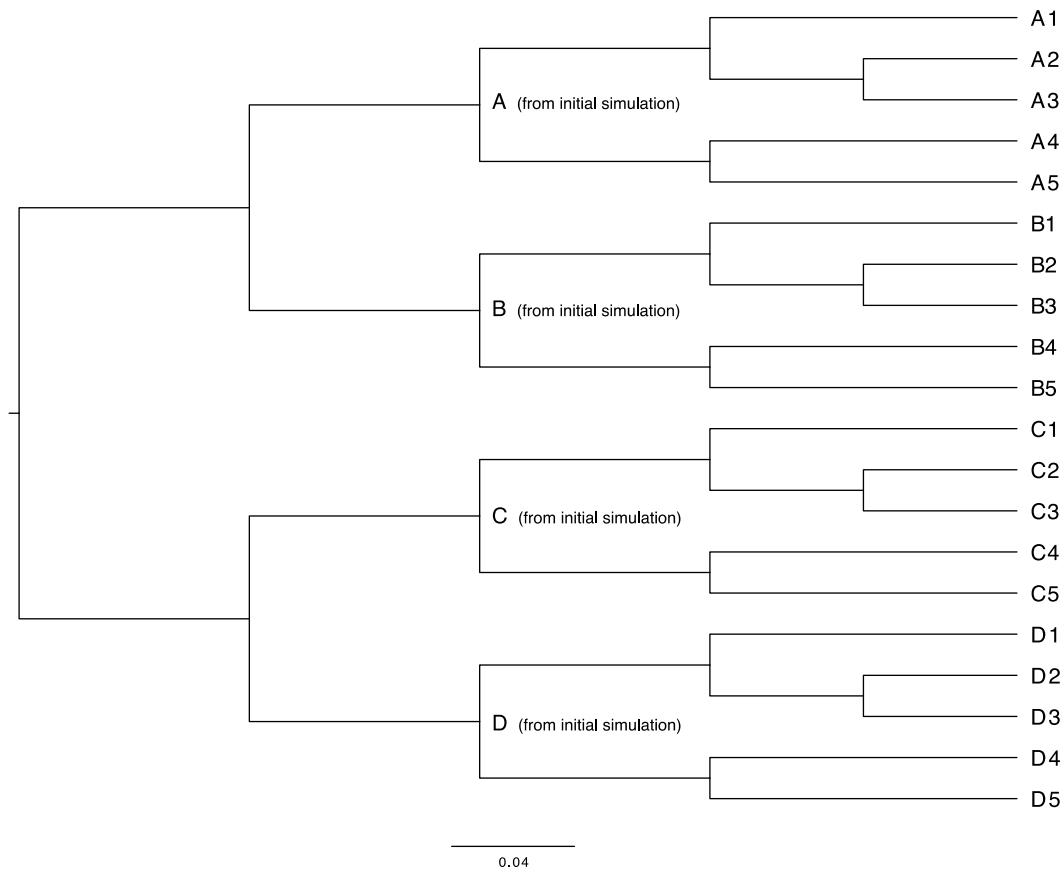


Figure 1 - Conceptual tree for our simulation of compositional heterogeneity. It's a 2-part process where A, B, C, and D are simulated first and selected based on A and C having amino acid letters that occur earlier in the alphabet and B and D having letters that occur later in the alphabet. These selected sequences are then used as the starting sequences for simulating 20 total sequences (5 each on each subclade).

2.3.4 In order to confirm that these datasets are compositionally heterogeneous, we will compute the comp-het indices for the 1,000 twenty-taxa datasets and use the sum of these indices as a test statistic in a Monte Carlo simulation. For the simulation, we will generate 1,000 twenty-taxa datasets and compute a sum of comp-het indices for these data 1,000 times. We will generate a p-value by dividing the number of sums that are greater than our test statistic by 1,000. If our test statistic is not significant, we will rerun steps starting at section 2.3.1 using an order of magnitude greater number of simulated datasets until we reach significance.

2.3.5 We will recode the 1,000 20-taxa datasets generated in section 2.3.3 using Dayhoff-6 recoding and perform maximum-likelihood analyses in RAxML for the recoded sets as well as for the non-recoded datasets under the Dayhoff model using the script `chunkify_comp.pl` (performs the same as `chunkify2.pl` but with parameters specifically set for this dataset).

```
perl chunkify_comp.pl
```

To execute the scripts generated from `chunkify_comp.pl` for maximum-likelihood analyses of both recoded and non-recoded datasets:

```
ls -l servername* | perl -ne 'chomp; print "sh $_ &\n";' | sh
```

2.3.6 We will use the script `compare_comp_trees.pl` to calculate RFDs in TOPD/FMTS to compare differences between recoded and non-recoded trees against the true tree used for simulation. Box plots will be generated to visualize RFD results and t-tests will be used to determine if there are significant differences in RFDs between non-recoded and recoded trees in R.

*Note: The directories in which to perform these analyses are set in the script and should be edited when reproducing results on another machine.

```
perl compare_comp_trees.pl
```

3 WORK COMPLETED SO FAR W DATES

We have not performed any additional tests after completing the work from the Hernandez_Ryan_2018_RecodingSim repo.

4 LITERATURE REFERENCES

- Chang, E. S., Neuhof, M., Rubinstein, N. D., Diamant, A., Philippe, H., Huchon, D., & Cartwright, P. (2015). Genomic insights into the evolutionary origin of Myxozoa within Cnidaria. *Proceedings of the National Academy of Sciences of the United States of America*, 112(48), 14912–7. <https://doi.org/10.1073/pnas.1511468112>
- Hrdy, I., Hirt, R. P., Dolezal, P., Bardonová, L., Foster, P. G., Tachezy, J., & Martin Embley, T. (2004). Trichomonas hydrogenosomes contain the NADH dehydrogenase module of mitochondrial complex I. *Nature*, 432(7017), 618–622. <https://doi.org/10.1038/nature03149>
- Puigbo, P., Garcia-Vallve, S., & McInerney, J. O. (2007). TOPD/FMTS: a new software to compare phylogenetic trees. *Bioinformatics*, 23(12), 1556–1558. <https://doi.org/10.1093/bioinformatics/btm135>
- R Development Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>
- Rambaut, A., & Grass, N. C. (1997). Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Bioinformatics*, 13(3), 235–238. <https://doi.org/10.1093/bioinformatics/13.3.235>
- Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9), 1312–1313. <https://doi.org/10.1093/bioinformatics/btu033>
- Susko, E., & Roger, A. J. (2007). On Reduced Amino Acid Alphabets for Phylogenetic Inference. *Molecular Biology and Evolution*, 24(9), 2139–2150. <https://doi.org/10.1093/molbev/msm144>

5 PHYLOTOCOL AMENDMENT HISTORY

Version	Date	Significant Revisions