Investigating the ability of 6-state recoding strategies to minimize the effects of saturation in phylogenetic analyses of amino-acid matrices

Principal Investigators: Joseph Ryan and Alexandra Hernandez

Draft or Version Number: v.1.0

6 March 2018

LIST OF ABBREVIATIONS

JTT	Jones-Thornton-Taylor (a model for amino acid subsitution)
PAM	Point accepted mutation (a model for amino acid substitution)
SD	Split distance
OSF	Open Science Foundation
SICB	Society for Integrative and Comparative Biology
S&R6	Susko & Roger 6-state recoding
RAxML	Randomized Axelerated Maximum Likelihood
TOPD	TOPological Distance
FMTS	From Multiple to Single
GTR	general time-reversible model

NOTE: Version 1.0 of this phylotocol will be uploaded as a pre-registration to the Open Science Framework (OSF) website and our GitHub site for this project. If changes are required, we will make them to our GitHub site in realtime and create an ammendment to our OSF pre-registration through the OSF website.

1 INTRODUCTION: BACKGROUND INFORMATION AND SCIENTIFIC RATIONALE

1.1 BACKGROUND INFORMATION

Models of protein evolution are used to score amino acid substitutions in sequence alignments or phylogenetic analyses. They predict the probability of one amino acid changing to another. The Dayhoff and JTT matrices are examples of these 20-state amino acid replacement models. However, these models do not account for compositional heterogeneity and substitution saturation (Susko & Roger 2007). Recently, recoding techniques have been employed to address these problems when exploring distant phylogenetic relationships. 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. 2005), while JTT recoding (i.e., S&R-6) is a 6-state recoding strategy based off of binning experiments on the JTT model by Susko & Roger (2007).

1.2 RATIONALE

The principle of using recoding to address substitution saturation has never been directly tested. 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 raise doubts about the benefits of using recoding approaches.

1.3 OBJECTIVES

The objective of this study is to test the performance of recoding under a range of saturation levels and determine if it is appropriate for deep phylogenetic questions. To do this we will perform simulations to evaluate the effect of two different types of recoding (i.e., Dayhoff-6 recoding and S&R6 recording) applied to simulated data matrices that are evolved using the model that corresponds to the basis of the particular recoding strategy (i.e. PAM250 for Dayhoff-6 recoding and JTT for S&R6 recoding). Sequences will be simulated on the topologies from Chang et al. (2015) and Feuda et al. (2017). These are two different topologies based on the same dataset, which includes a wide range of animals and a few closely related outgroups.

2 STUDY DESIGN AND ENDPOINTS

2.1 Simulate the evolution of amino acids along the phylogeny produced in Chang et al. (2015) using the PAM250 model (i.e. the model on which Dayhoff-6 recoding is based), JTT model (i.e., the model on which S&R6 recoding is based), and different intervals of branch lengths (units of substitutions per site). We will vary the branch-length parameter (-s) in seq-gen from 1 to 20 by increments of 1.

Seq-Gen PAM (Rambaut & Grassly 1997)

```
seq-gen -mPAM -z420 -n1000 -s1.0 -a1.0 -or Chang_orig_phylobayes.tre  > Chang.PAM.1.phy

seq-gen -mPAM -z420 -n1000 -s2.0 -a1.0 -or Chang_orig_phylobayes.tre  > Chang.PAM.2.phy

seq-gen -mPAM -z420 -n1000 -s3.0 -a1.0 -or Chang_orig_phylobayes.tre  > Chang.PAM.3.phy

seq-gen -mPAM -z420 -n1000 -s4.0 -a1.0 -or Chang_orig_phylobayes.tre  > Chang.PAM.4.phy

seq-gen -mPAM -z420 -n1000 -s5.0 -a1.0 -or Chang_orig_phylobayes.tre  > Chang.PAM.5.phy

seq-gen -mPAM -z420 -n1000 -s6.0 -a1.0 -or Chang_orig_phylobayes.tre  > Chang.PAM.6.phy

seq-gen -mPAM -z420 -n1000 -s7.0 -a1.0 -or Chang_orig_phylobayes.tre  > Chang.PAM.7.phy

seq-gen -mPAM -z420 -n1000 -s8.0 -a1.0 -or Chang_orig_phylobayes.tre  > Chang.PAM.8.phy

seq-gen -mPAM -z420 -n1000 -s9.0 -a1.0 -or Chang_orig_phylobayes.tre  > Chang.PAM.9.phy

seq-gen -mPAM -z420 -n1000 -s10.0 -a1.0 -or Chang_orig_phylobayes.tre  > Chang.PAM.10.phy

seq-gen -mPAM -z420 -n1000 -s11.0 -a1.0 -or Chang_orig_phylobayes.tre  > Chang.PAM.11.phy

seq-gen -mPAM -z420 -n1000 -s12.0 -a1.0 -or Chang_orig_phylobayes.tre  > Chang.PAM.12.phy

seq-gen -mPAM -z420 -n1000 -s13.0 -a1.0 -or Chang_orig_phylobayes.tre  > Chang.PAM.13.phy

seq-gen -mPAM -z420 -n1000 -s14.0 -a1.0 -or Chang_orig_phylobayes.tre  > Chang.PAM.14.phy

seq-gen -mPAM -z420 -n1000 -s15.0 -a1.0 -or Chang_orig_phylobayes.tre  > Chang.PAM.15.phy

seq-gen -mPAM -z420 -n1000 -s16.0 -a1.0 -or Chang_orig_phylobayes.tre  > Chang.PAM.16.phy

seq-gen -mPAM -z420 -n1000 -s17.0 -a1.0 -or Chang_orig_phylobayes.tre  > Chang.PAM.17.phy

seq-gen -mPAM -z420 -n1000 -s18.0 -a1.0 -or Chang_orig_phylobayes.tre  > Chang.PAM.18.phy

seq-gen -mPAM -z420 -n1000 -s19.0 -a1.0 -or Chang_orig_phylobayes.tre  > Chang.PAM.19.phy

seq-gen -mPAM –z 420 -n1000 -s20.0 -a1.0 -or Chang_orig_phylobayes.tre  > Chang.PAM.20.phy
```

Seq-Gen JTT (Rambaut & Grassly 1997)

Same 20 commands as Seq-Gen PAM, but all occurrences of PAM are replaced with JTT

divide.pl (For all 40 commands above, divide the 1000 datasets outputted by Seq-Gen into separate phylip files)

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

2.2 Generate maximum-likelihood trees for simulated sequences using the models PAM250, Dayhoff-6 recoding, JTT, and S&R6.

RAxML (Stamatakis 2014)

applyRAxML2AllFilesInDirectory.pl (script distributed with RAxML applies commands to all files in a directory)

2.2a Perform maximum-likelihood analysis using the PAM250 model.

```
perl applyRAxML2AllFilesInDirectory.pl /dir "raxmlHPC -p 420 -m PROTGAMMADAYHOFF"
```

2.2b We will convert sequences to a dayhoff-6 recoded dataset using the script dayhoff6recode_fasta.pl and perform a maximum-likelihood analysis using RAxML's MULTIGAMMA multi-state model with GTR.

```
ls -1 *.phy | perl -ne 'chomp; m/(.*).phy/; print "perl phy2fa.pl $_ > $1.fa\n";' > phy2fastacmd.sh
```

```
sh ./phy2fastacmd.sh
```

```
ls -1 *.fa | perl -ne 'chomp; m/(.*).fa/; print "perl dayhoff6recode_fasta.pl $_ > $1_recoded.fa\n";' >
dayhoffrecodecmd.sh
```

```
sh ./dayhoffrecodecmd.sh
```

```
ls -1 *recoded.fa | perl -ne 'chomp; m/(.*_recoded).fa/; print "fasta2phy.pl $_ > $1.phy\n";' > recode2phycmd.sh
```

```
sh ./recode2phycmd.sh
```

```
perl  applyRAxML2AllFilesInDirectory.pl /dir "raxmlHPC -p 420 -m MULTIGAMMA -K GTR"
```

2.2c Perform a maximum-likelihood analysis using the JTT model.

```
perl applyRAxML2AllFilesInDirectory.pl /dir "raxmlHPC -p 420 -m PROTGAMMAJTT"
```

2.2d Convert sequences to an S&R6 recoded dataset using the script s&r6recode_fasta.pl and perform a maximum-likelihood analysis using RAxML's MULTIGAMMA multi-state model with GTR.

```
ls -1 *.phy | perl -ne 'chomp; m/(.*).phy/; print "perl phy2fa.pl $_ > $1.fa\n";' > phy2fastacmd.sh
```

```
sh ./phy2fastacmd.sh
```

```
ls -1 *.fa | perl -ne 'chomp; m/(.*).fa/; print "perl s&r6recode_fasta.pl $_ > $1_recoded.fa\n";' >
s&r6recodecmd.sh
```

```
sh ./s&r6recodecmd.sh
```

```
ls -1 *recoded.fa | perl -ne 'chomp; m/(.*_recoded).fa/; print "fasta2phy.pl $_ > $1.phy\n";' > recode2phycmd.sh
```

```
sh ./recode2phycmd.sh
```

```
perl applyRAxML2AllFilesInDirectory.pl /dir "raxmlHPC -p 420 -m PROTGAMMAJTT"
```

2.3 Concatenate trees from 2.2 into one tree file along with the true tree used for simulation, then perform boot split distance calculations in TOPD/FMTS. Lower SD values will indicate similarity between the true tree and reconstructed trees, higher SD values indicate larger differences between the true tree and reconstructed trees.

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

TOPD/FMTS (Puigbò et al. 2007)

```
perl  topd_v4.6.pl -f Chang_all_trees.tre -c reference -m split
```

Perform all analyses above on Feuda et al. (2017) topology.

Since the Feuda et al. (2017) phylogeny was produced from recoding the Chang et al. (2015) dataset, all branch lengths are reduced by 2.6 substitutions per site compared to the Chang et al. (2015) phylogeny. To account for this, the branch length parameter will range from 2.6 to 52 by increments of 2.6.

3 EXPECTED OUTCOMES

It is expected that at minimal levels of saturation, non-recoded matrices will outperform recoded matrices (since non-recoded matrices were used to generate simulated sequence data. If, as saturation increases, trees from recoded matrices steadily reduce the gap in accuracy vs. the trees from the non-recoded datasets, and eventually “overtake” the non-recoded analyses in terms of accuracy, recoding would be considered a legitimate strategy to reduce the affects of saturation. If recoded datasets are always suboptimal regardless of saturation levels, and especially if there is not a clear trend in recoding at least steadily approaching the accuracy of non-recoded datasets, it would suggest that recoding is a suboptimal strategy, and that results based on this strategy should be reexamined. Based on our preliminary findings, we hypothesize that recoding will not improve tree reconstruction regardless of saturation levels.

4 WORK COMPLETED SO FAR W DATES

We simulated amino acids along the phylogenies produced in Chang et al. (2015) and Feuda et al. (2017) using the PAM250 model with branch lengths = 0.5, 1, 1.5, 2, 2.5, and 3 in Seq-Gen. We then performed maximum-likelihood analyses on each set of simulated sequences using RAxML with the PAM250 and Dayhoff-6 models. We compared the resulting reconstructed topologies to the trees we used for simulation by calculating SD values in TOPD/FMTS. In all cases the PAM250 reconstructions performed better than the dayhoff-6 reconstructions. More details on this analysis including a phylotocol are provided in the 01-PRELIMINARY_TESTS directory of the DayhoffRecodingTests repo.

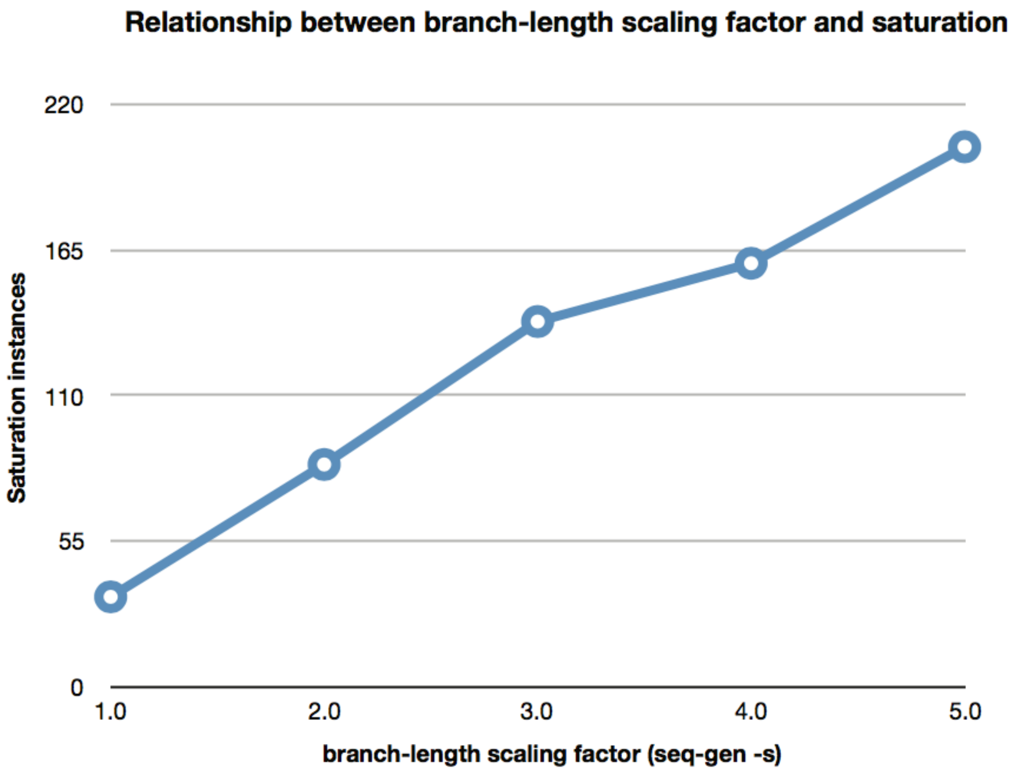

Figure 1 - Relationship between branch-length scaling factor and saturation

We also created a script that shows that increasing the branch length scaling factor parameter increases saturation (Figure 1). The script and test tree are available in the 01-SATURATION_TEST directory of the Hernandez_Ryan_2018_RecodingSim repo.

5 LITERATURE REFERENCES

Chang, E. S., Neuhof, M., Rubinstein, N. D., Diamant, A., Philippe, H., Huchon, D., & Cartwright, P. (2015). Genoinsights 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.

Feuda, R., Dohrmann, M., Pett, W., Philippe, H., Rota-Stabelli, O., Lartillot, N., Wörheide, G., Pisani, D. (2017). Improved Modeling of Compositional Heterogeneity Supports Sponges as Sister to All Other Animals. *Current Biology*, *27*, 3864-3870.

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*, 618-622.

Rambaut A, Grassly NC. 1997. Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Computer Applications in the Biosciences*, *13*(3), 235-238.

Puigbo, P., Garcia-Vallve, S., & McInerney, J. O. (2007). TOPD/FMTS: a new software to compare phylogenetic trees. *Bioinformatics*, *23*(12), 1556-1558.

Susko, E. & Roger, A. J. (2007) On Reduced Amino Acid Alphabets for Phylogenetic Inference. *Molecular Biology and Evolution*, *24*(9), 2139-2150.

6 PHYLOTOCOL AMENDMENT HISTORY

Version	Date	Significant Revisions