Estimation and Comparison of HIV-Specific Substitution Matrices

Jia Jin Kee

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Science

University of Washington

2014

Committee:

Paul Edlefsen

Peter Gilbert

Program Authorized to Offer Degree:

Biostatistics - Public Health

University of Washington

Abstract

Estimation and Comparison of HIV-Specific Substitution Matrices

Jia Jin Kee

Chair of the Supervisory Committee:
Paul Edlefsen, PhD
Affiliate Assistant Professor
Department of Biostatistics, University of Washington

Amino acid substitution matrices are commonly used for sequence alignment, phylogenetic inference and sequence comparison. Empirical organism-specific substitution matrices constructed using only sequence data from a particular organism are thought to lead to more accurate analyses. In HIV research, the standard substitution matrices are the between- and within-host matrices estimated using HIV sequences introduced by Nickle et al. (2007). This thesis focuses on constructing more granular HIV-specific matrices (clade-specific matrices and gene-specific matrices) and comparing the matrices in a way that accounts for error in matrix estimation. Using standard errors of parameter estimates predicted from a two-part linear model, the analyses indicate statistically significant difference between HIV clade B and HIV clade C matrices as well as between HIV Env gene and HIV Gag gene matrices upon performing Bonferroni-corrected comparisons of 189 estimates of amino acid exchangeability parameters.

Acknowledgements

I am massively indebted for the tremendous input and patient guidance provided by my thesis advisor, Paul Edlefsen. I would also like to express my gratitude to Peter Gilbert for being a member of my thesis committee. I would like to thank Craig Magaret and Ted Holzman for imparting knowledge on amino acid substitution matrices and for providing technical support. I would like to thank Ted Holzman specifically for his immense assistance with navigating the HyPhy program. I am also appreciative of the curated HIV datasets provided by Morgane Rolland and the advice on running the HyPhy program communicated by Sergei Pond.

Table of Contents

List of Equations

List of Figures

List of Tables

1 INTRODUCTION

1.1 Markov Chain Models of Amino Acid Evolution

Discrete-space continuous-time Markov chain models are commonly used to model the evolution of amino acids at each site along an amino acid sequence. A Markov chain describes how one state transitions into another in such a way that the future state depends only on the present state but not the past states (this is also known as the Markov property). In the context of a Markov chain that models the evolution of amino acids, the states that make up the discrete space are the 20 amino acids while time is modeled continuously. Each amino acid site is assumed to be evolving independently and the same Markov chain model describes the point mutation process of ancestral amino acids at any site. A continuous-time Markov chain is characterized by an instantaneous rate matrix Q, where q_{ij} gives the instantaneous rate of change from amino acid i to amino acid j. A constraint is placed on the diagonals of the Q matrix such that $q_{ii} = -\sum_{i \neq j} q_{ij}$ so that each row of Q sums to zero. Typically, when modeling sequence evolution, the Markov chain is assumed to be time reversible, which means that the chain looks the same going backward and forward in time. Under the assumption of time reversibility, the Q matrix can be written as the product of a symmetric matrix R multiplied by a diagonal matrix D (Equation 1).

Equation 1: Instantaneous rate matrix Q

$$
\begin{aligned}
Q &= \{q_{ij}\} \\
&= \{r_{ij}\pi_j\} = \begin{bmatrix} - & r_{1,2}\pi_2 & r_{1,3}\pi_3 & \cdots & r_{1,20}\pi_{20} \\ r_{2,1}\pi_1 & - & r_{2,3}\pi_3 & \cdots & r_{2,20}\pi_{20} \\ r_{3,1}\pi_1 & r_{3,2}\pi_2 & - & \cdots & r_{3,20}\pi_{20} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{20,1}\pi_1 & r_{20,2}\pi_2 & r_{20,3}\pi_3 & \cdots & - \end{bmatrix} \\
&= \begin{bmatrix} - & r_{1,2} & r_{1,3} & \cdots & r_{1,20} \\ r_{2,1} & - & r_{2,3} & \cdots & r_{2,20} \\ r_{3,1} & r_{3,2} & - & \cdots & r_{3,20} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{20,1} & r_{20,2} & r_{20,3} & \cdots & - \end{bmatrix} \begin{bmatrix} \pi_1 & 0 & 0 & \cdots & 0 \\ 0 & \pi_2 & 0 & \cdots & 0 \\ 0 & 0 & \pi_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \pi_{20} \end{bmatrix} = R \cdot D
\end{aligned}
$$

The r_{ij} in the amino acid exchangeability matrix R gives the exchangeability between amino acid i and amino acid j. Since time reversibility is imposed, $r_{ij} = r_{ji}$. Therefore, there are $\binom{20}{2} = 190$ exchangeability parameters. By convention, an identifiability constraint is also placed on the Q matrix; it is scaled to provide meaningful branch lengths (fixing $-\sum_i \pi_i q_{ii} = 1$). Doing so removes one of the 190 exchangeability parameters and leaves us with 189 free exchangeability parameters. On the other hand, the π_j in the diagonal of the D matrix gives the stationary frequency of an amino acid j. Proportions of amino acids observed in the datasets being analyzed are usually taken as estimates of the π_j's.

From the instantaneous rate matrix Q, one can obtain the transition probability matrix P via Equation 2. Each cell of the P matrix $p_{ij}(t)$ gives the probability of change from amino acid i to amino acid j for any time $t > 0$ where t is measured in terms of expected number of substitutions per site.

Equation 2: Transition probability matrix P

$$P(t) = e^{Qt}$$

The different matrices discussed above are collectively known as amino acid substitution matrices. They are used for aligning protein sequences, inferring phylogenetic trees and comparing sequences such as sieve analysis (Altschul, 1991; Edlefsen, Gilbert & Rolland, 2013). There are several common ways to estimate amino acid substitution matrices. There are methods that do not condition on a phylogenetic tree such as the BLOSUM method (Henikoff & Henikoff, 1992) and other methods that condition on an estimated phylogenetic tree such as the Dayhoff (1978) and HyPhy (Pond & Muse, 2005) methods. The Dayhoff method is a counting method that counts amino acid transitions on a phylogenetic tree and gives a PAM_t matrix, which is the integral of the instantaneous rate matrix over a discrete time t that is needed to observe t% expected divergence in the sequence. The HyPhy method is an approximate maximum likelihood method that estimates amino acid exchangeability parameters and branch lengths condition on a neighbor-joining tree topology. The HyPhy approach follows the advice put forth by Whelan and Goldman (2001) that conditioning on near-optimal tree topologies estimated using a different evolutionary model would only negligibly bias estimated parameters. The first author of the HyPhy

program that implements the HyPhy method (Pond & Muse, 2005) was also consulted at one point to get at the issue of conditioning on an incorrect phylogenetic tree. He responded that “much of the phylogenetic literature has been built around the assumption that parameter inference is robust to all but drastic errors in the tree.” (S. Pond, personal communication, May 12, 2014). All in all, both Dayhoff and HyPhy methods estimate a phylogenetic tree with a fixed incorrect matrix and report matrix parameters estimated conditionally on the tree but do not reevaluate the tree to jointly estimate the tree and matrix.

In sequence alignment, the substitution matrices used are usually of a transformed variety called scoring matrices (Mount, 2007). These are log-odds matrices derived from PAM or BLOSUM matrices. Each cell of a scoring matrix represents the odds ratio that a target amino acid would descend from a query amino acid within the context of the evolutionary process represented by the matrix, as opposed to a null “chance” model (that the target and query amino acids are aligned by chance).

1.2 HIV-Specific Amino Acid Substitution Matrices

In HIV research, the current de facto standard substitution matrices being used are the within-host and between-host substitution matrices introduced by Nickle et al. (2007) (see also “PhyML Explanation”, LANL website 2013). Using a training dataset of 3387 HIV-1 clade B sequences from 48 subjects (for estimation of the within-host matrix) and a training dataset of 1026 HIV-1 clade B sequences (one per subject) (for estimation of the between-host matrix) that were available in the laboratory of Jim Mullins at the time, Nickle and colleagues estimated these matrices using the HyPhy program. In particular, the training dataset for estimation of the between-host matrix is made up of amino acid sequence data from multiple HIV genes, including 39 Gag sequences and 241 Env sequences (107 Env gp120 and 134 Env gp41). A separate tree is constructed using sequence data of each gene and amino acid exchangeability parameters were jointly estimated with the tree lengths on all trees by HyPhy. Because the within- and between-host matrices were estimated using only HIV-1 sequences, they are thought to reflect sequence evolution in HIV more accurately than other matrices, allow for better reconstruction of HIV alignments and phylogenies and improve power of HIV sieve analyses (Gilbert, Wu & Jobes, 2008). In addition, Nickle et al. (2007) showed that the within-host matrix and between-host matrix differed (though this claim

was not supported by a statistical analysis) and argued that this makes sense because HIV undergoes different types of immune pressure when replicating in a host's body versus when infecting a new host.

1.3 Research Questions

The argument that HIV within-host and between-host substitution matrices would lead to more accurate analyses suggests that using even more granular HIV-specific substitution matrices could further improve inference of alignments, trees, and sieve effects. In a fast evolving virus like HIV, the different HIV genes may have their own distinct pattern of evolution depending on how conserved the genes are and the manner of the evolutionary constraints on them. For example, even though both Env and Gag genes code for core structural proteins in HIV, Gag is relatively conserved compared to Env (Foley, 2000). While Env is thought to be the primary target of antibodies, Gag is a primary site of T cell recognition. It is therefore conceivable that proteins encoded by the two genes have different evolutionary patterns. In addition, different HIV clades are predominant in different geographical regions and the different immune pressures imposed by their human hosts can lead to the clades evolving differently. HIV clade B is the predominant clade in Europe and the Americas whilst HIV clade C is more common in the Southern African region (Buonaguro, Tornesello & Buonaguro, 2007). Clade C HIVs also share a more recent common ancestor than clade B (Travers et al., 2004), so the substitution matrix of clade C sequences may reflect more recent evolutionary pressures than the matrix describing the longer evolution of clade B sequences.

To robustly justify the usage of more granular HIV-specific substitution matrices, the first thing to determine is whether the matrices are different at all, beyond what would be expected by chance due to sampling variation. In this thesis I performed statistical comparisons of HIV amino acid exchangeability matrices estimated by HyPhy across two classes of HIV-specific matrices: clade-specific matrices and gene-specific matrices. To my knowledge, this work describes the first formal comparison of substitution matrices that attempts to account for error in estimation of matrix due to sampling variability. As described in detail below, I constructed a two-part linear model for standard errors of amino acid exchangeability estimates using a training set of 1618 HIV clade B Gag amino acid sequences and I used the model to predict standard errors in actual datasets. To determine if gene-specific matrices are different across

genes, I compared exchangeability matrices estimated by HyPhy using HIV clade C Env gene and HIV clade C Gag gene curated amino acid sequence data, respectively. To determine if clade-specific matrices are different, I compared exchangeability matrices estimated by HyPhy using HIV clade B Gag gene and HIV clade C Gag gene curated amino acid sequence data, respectively. The results are three new empirical substitution matrices for use in HIV sequence analysis. My analyses showed that gene-specific matrices as well as clade-specific matrices are statistically significantly different across clades and genes, respectively.

2 METHODS

2.1 Datasets

A training set of 1618 HIV-1 clade B Gag amino acid sequences of 643 residues each was used to construct a two-part linear model for the standard errors of amino acid exchangeability parameters estimated by the HyPhy program. The sequence alignments were obtained from the LANL database based on its 2012 distribution and were filtered to ensure that they have at least a Hamming distance of 20. Partitioned (sampled without replacement) datasets of three sample sizes were generated: thirty-two partitioned datasets each containing 50 sequences, sixteen partitioned datasets each containing 100 sequences and eight partitioned datasets each containing 200 sequences. Ensuring a minimum Hamming distance of 20 and obtaining partitioned datasets are steps taken to guard against oversampling sequences from individual subjects, which would bias the estimation of evolutionary parameters. Note that if we sampled with replacement, the estimated tree would effectively treat identical sequences as degenerate and the effective sample size would be diminished.

The datasets used to estimate the clade-specific and gene-specific substitution matrices were also obtained from the LANL database, filtering for the correct open reading frame, excluding hypermutated sequences (Rose & Korber, 2000) and ensuring a minimum inter-sequence distance to ensure that each subject is represented at most once: the HIV-1 clade B Gag dataset contains 440 cDNA sequences collected from US, the HIV-1 clade C Gag dataset contains 678 cDNA sequences collected from South Africa and the HIV-1 clade C Env dataset contains 579 cDNA sequences collected from South Africa. Morgane Rolland provided these datasets, aligned and curated. Excluding hypermutated sequences and sequences with incorrect reading frame ensure the sequences are from viable viruses. Prior to analyses, the cDNA sequences were translated into amino acid sequences with a codon-aware script provided by Paul Edlefsen. Additionally, five hypervariable regions in the protein sequence encoded by the Env gene were identified using amino acid patterns that precede and follow these regions in the HIV-1 reference sequence (HXB2) and removed because the correct reading frame in the hypervariable regions is not known. Ted Holzman performed the steps of identifying the hypervariable regions and

excising them from the sequences, in a manner ensuring that the same regions were extracted from these datasets as had been extracted from those used by Nickle et al. (2007).

In addition, to serve as a negative control experiment, I compared exchangeability matrices under approximations of the null hypothesis. In principle, simple random resampling of the data with altered gene or clade labels is insufficient to reflect the null hypothesis of no difference because of the non-independence of the sequences (as reflected by the structure in the phylogenetic tree). Nevertheless, I considered three comparisons to attempt to obtain approximations to the correct null phylogenetic tree. Firstly, the 440 HIV clade B Gag sequences and 678 clade C Gag sequences were pooled using a transitive alignment (software provided by Paul Edlefsen and Cindy Molitor) with the HXB2 reference sequence as the common sequence; I then randomly partitioned the pooled data into two datasets each containing 559 sequences. Secondly, I randomly partitioned the HIV clade B sequence data into two datasets. Thirdly, two datasets were constructed by putting an equal number of randomly selected unique HIV clade B Gag and HIV clade C Gag sequences in each dataset: each dataset has 220 HIV clade B Gag sequences and 220 clade C Gag sequences. In all three comparisons, amino acid exchangeability matrices were estimated for the two datasets using HyPhy and the 189 amino acid exchangeability parameters were compared with Bonferroni correction.

2.2 Exchangeability, Instantaneous Rate, PAM and Scoring Matrices for Dayhoff and HyPhy

In the initial stage of this project, there was concern that HyPhy would be too time-consuming to use for parameter estimation due to the likelihood optimization procedure. To overcome the potential time limitation, I ported the Dayhoff method implemented in the Darwin programming language (Dessimoz, Gannarozzi & Schneider, 2011) into the R programming language and used it to estimate an amino acid mutation matrix. The porting was done mainly to better understand the Dayhoff method and partly because I am more familiar with the R programming language. The algorithm takes a substitution count matrix as its input. I obtained the substitution count matrix generated by HyPhy after its initial tree-building stage, which is not symmetric since it is based on transitions observed on a neighbor-joining tree of an amino acid sequence dataset. Two versions of analyses were performed using the original non-symmetric

HyPhy count matrix and the symmetrized HyPhy count matrix (see Section 2.4). The motivation for symmetrizing the HyPhy count matrix prior to putting it into the Dayhoff algorithm is because in the Dayhoff setting, it is not known which is the ancestral amino acid in an amino acid transition (Dessimoz, Gannarozzi & Schneider, 2011). Nonetheless, the Dayhoff algorithm (Dessimoz, Gannarozzi & Schneider, 2011) does not impose any symmetry requirement on the input count matrices. An arbitrary pseudocount of 1 is added to each cell of the count matrix to prevent gaining infinity values in the derived scoring matrix. A substitution count matrix (either the original non-symmetric matrix or the symmetrized matrix) with pseudocounts added is taken as the input to the Dayhoff algorithm to estimate the amino acid mutation matrix, which is a count matrix normalized by the sum of each column. Assuming time reversibility, the Dayhoff mutation matrix is symmetrized. The symmetrized Dayhoff mutation matrix is considered the equivalent of the amino acid exchangeability matrix estimated by HyPhy. Because the HyPhy program has a module that derives PAM and log odds matrices given the amino acid exchangeability matrix, I used the module to derive a PAM1 matrix and the corresponding scoring matrix (log odds 1 matrix) given the symmetrized Dayhoff mutation matrix.

I later discovered that running HyPhy is not extremely time-consuming. On the Fred Hutchinson's servers, it usually takes two to three days to run HyPhy on datasets whose sizes range from 50 to about 700 sequences. 189 exchangeability parameters of the amino acid exchangeability matrix (the Isoleucine <-> Leucine exchangeability parameter is constrained to be 1) were estimated from HyPhy according to the procedure outlined in Nickle et al. (2007) using a version of the script provided in that publication, updated by Ted Holzman to work with the current version of HyPhy.

2.3 Standard Errors of Estimates of HyPhy Amino Acid Exchangeabilities

Nickle et al. (2007) did not show that the difference between within-and between- host models is beyond the variation that would be expected by chance. Also, none of the most commonly used substitution matrices (BLOSUM, PAM, WAG) has been evaluated from the perspective of sampling variation and statistical inference with uncertainty. Since HyPhy is a maximum likelihood estimation program, and since there is clearly variation sample-to-sample in both the tree and matrix estimation steps (as is shown

below), this thesis sought to provide statistical analysis to support that any observed difference across estimated matrices reveals a true difference beyond what could be explained by estimation error alone.

The HyPhy program performs an imputation procedure on all raw estimated amino acid exchangeabilities so that none of the maximum likelihood estimates are zero (Equation 3). However, confidence intervals and standard errors via the Fisher scoring method are available for the raw estimated exchangeabilities only.

Equation 3: HyPhy parameter estimate imputation procedure

For non-zero raw exchangeability estimates: imputed estimate = raw estimate * A

For zero-valued raw exchangeability estimates: $imputed\ estimate = \frac{1}{[(f_i+f_j)*B]}$

(Note: A and B are different scaling constants; f_i and f_j are proportions of amino acids i and j observed in the dataset being analyzed)

Fortunately, both the raw and imputed estimates are outputs of the HyPhy program, so I was able to figure out the scaling constants to compute standard errors for the imputed estimates. For an imputed estimate that is not zero to begin with, its standard error is obtained by multiplying the standard error of the raw estimate with a scaling constant of (raw estimate / imputed estimate). The scaling constant is different for each dataset. On the other hand, for an imputed estimate that is initially zero, its standard error is the same as the standard error of the raw estimate.

In this work standard errors inferred from HyPhy were used to compare the matrices, but not as the primary method due to several shortcomings. In particular, in large datasets, HyPhy would sometimes (1) return confidence intervals with an upper bound of 10000 but with an erroneous negative estimated variance, (2) give very small standard errors or (3) terminate due to an error of trying to perform LU decomposition on singular matrices. These three aberrations were observed when exchangeability matrices for HIV clade C Env (579 sequences), HIV clade B Gag (440 sequences) and HIV clade C Gag (678 sequences) datasets were estimated. In addition, in all datasets regardless of sample size, HyPhy would return the same standard error for all estimates if I asked HyPhy for the confidence intervals of all

exchangeability parameters simultaneously. This bug does not occur if confidence intervals were requested for smaller subsets of parameters simultaneously, but the number of parameters that one can request confidence intervals for simultaneously before the bug appears is different from dataset to dataset.

Another issue is that HyPhy gives exchangeability estimates condition on a fixed phylogenetic tree. In principle, there should be some error when estimating the tree. However, by conditioning on the tree, the error in estimation of the tree is not accounted for in the standard error of the exchangeability estimates given by HyPhy. To investigate this issue, I compared HyPhy standard errors of exchangeability estimates from one partitioned dataset containing 50 sequences to the empirical standard errors obtained across size 50 datasets. The comparison was also carried out for HyPhy standard errors from one partitioned dataset containing 200 sequences and empirical standard errors obtained across size 200 datasets. A paired t-test was performed to determine whether the differences between HyPhy's and empirical standard errors are similar or not at the two sample sizes. A significance level of 0.05 was used for the paired t-test.

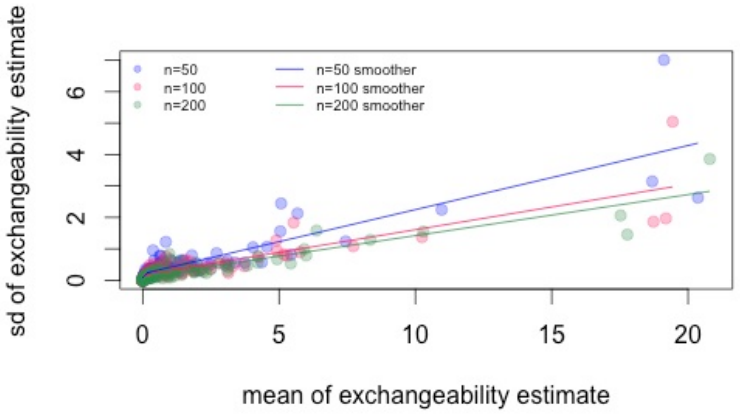
2.4 Construction of Linear Model for Standard Errors of Estimates of Evolutionary Parameters

To perform statistical comparison between cell values of two matrices, some measure of variability of the estimated cell values is needed. This thesis considers a linear model that models the standard errors of estimates of evolutionary parameters using the inverse of the square root of the number of sequences, n. In principle across a broad range of consistent statistical estimators, the standard error should decrease proportionally with the inverse of root n and eventually approach zero when n is very large. Therefore, it makes sense theoretically to expect this model to extrapolate beyond the range of sample sizes at which we can estimate it, though I am aware of the dangers of extrapolation and in particular the potential for anti-conservatism if they are underestimated and have taken further steps to evaluate the reliability of these estimates.

In the training data, empirical errors of estimates of the amino acid exchangeability parameters (i.e. standard deviations across parameters estimated from random partitions of the data) seem to be

larger for exchangeability parameters with larger estimated values. Figure 1 shows the plot of empirical standard errors of estimates of amino acid exchangeability parameters estimated by HyPhy against the average parameter estimates (averaged across partitioned datasets of a particular sample size for sample sizes n=50, 100 and 200).

Figure 1: Empirical standard error of estimate of amino acid exchangeability parameter versus the mean of the corresponding parameter estimate; darker shades correspond to more data points



In light of this observation, I considered modeling the standard errors as a linear function of the evolutionary parameter estimate and the inverse of root n, with an interaction between the parameter estimate and the inverse of root n (Equation 4). The interaction term is used to account for possible difference in the rate of change in standard error with inverse of root n for cells with different estimated parameter values.

Equation 4: Modeling standard errors as a linear function of inverse of root n, parameter estimate and interaction term between the two

$$E(SE) = \beta_0 + \beta_1 * 1/\sqrt{n} + \beta_2 * est + \beta_3 * est * 1/\sqrt{n}$$

Versions of the model in Equation 4 with and without intercept term were compared using a chi-square test to confirm that the intercept is not significantly different from its theoretically expected value of zero. If inverse of root n is zero, then n is infinity, and there should be no error under the assumption that the model is asymptotically consistent.

Versions of the aforementioned model with and without intercept were fitted on empirical standard errors of estimated cell values of the Dayhoff mutation matrix, HyPhy amino acid exchangeability matrix, PAM1 matrix and log odds 1 matrix of Dayhoff and HyPhy, respectively. The empirical standard errors of the cell values of all Dayhoff matrices do not appear to be asymptotically consistent (Tables 1 and 2). The standard errors of cell values appear to be decreasing as inverse of root n increases. In other words, the standard errors seem to be decreasing as sample size n decreases. A possible reason for the non-consistency of standard errors could be due to the Dayhoff mutation matrix not scaled such that $-\sum_i \pi_i q_{ii} = 1$ whereas the scaling is already in place for substitution matrices estimated by HyPhy. In models modeling standard errors of parameter estimates of all Dayhoff matrices, the intercept term is statistically significantly different from zero.

Table 1: Regression results with Dayhoff matrices obtained from non-symmetric HyPhy count matrices

	Coef. Estimate* (P-Value)		R^{2}*	
	Inverse Root n	Intercept	Multiple	Adjusted
Mutation Matrix	-0.0017 (< 0.001)	0.000103 (0.000114)	0.61	0.61
Instantaneous Rate Matrix	-4.08×10^{-5} (1.33×10^{-6})	4.34×10^{-6} (6.03×10^{-7})	0.69	0.69
PAM1 Matrix	-4.09×10^{-4} (2.11×10^{-6})	5.29×10^{-5} (3.12×10^{-9})	0.70	0.70
Log Odds 1 Matrix	-1.70 (0.6723)	0.98 (0.0152)	0.029	0.024

*From model with intercept

Table 2: Regression results with Dayhoff matrices obtained from symmetrized HyPhy count matrices

	Coef. Estimate* (P-Value)		R^2*	
	Inverse Root n	Intercept	Multiple	Adjusted
Mutation Matrix	-1.34×10^{-3} (1.21×10^{-6})	1.12×10^{-4} (5.22×10^{-5})	0.64	0.64
Instantaneous Rate Matrix	-4.70×10^{-5} (1.15×10^{-7})	-4.64×10^{-6} (3.75×10^{-7})	0.71	0.71
PAM1 Matrix	-4.87×10^{-4} (7.54×10^{-8})	5.76×10^{-5} (7.93×10^{-10})	0.71	0.71
Log Odds 1 Matrix	-2.16 (0.6048)	1.04 (0.0137)	0.027	0.022

*From model with intercept

I had at first observed this same behavior with the standard errors for a few of the amino acid exchangeability parameters estimated by HyPhy (Table 3), before settling on a partitioning, rather than a subsampling, approach to creating the training datasets. That had lead me to despair that the error in estimating the tree was insurmountable, in violation of the argument put forth by Whelan and Goldman that conditioning on the tree contributes has negligible influence on the resulting inference. The observed non-decreasing-errors behavior of the Dayhoff approach even with properly partitioned datasets may also reflect error in estimation of the tree, which is perhaps overcomed by HyPhy (which re-estimates branch lengths but keeps the tree topology fixed) but not by Dayhoff (which ignores branch lengths altogether).

Table 3: Non-consistency of standard errors of amino acid exchangeability parameters estimated by HyPhy in datasets sampled with replacement

AA Exchangeability	n=50		n=500	
	HyPhy SE	Empirical SE	HyPhy SE	Empirical SE
Alanine <-> Asparagine	0.14	0.17	0.058	0.24
Alanine <-> Threonine	0.37	1.40	0.12	2.24
Aspartic Acid <-> Glutamic Acid	1.04	3.23	0.14	6.23

The standard errors model that I ultimately used is built using empirical standard errors of cell values from HyPhy amino acid exchangeability matrices (data points corresponding to the Isoleucine <-> Leucine exchangeability parameter are excluded because it is a non-free parameter). The regression results (Table 4) seem to indicate consistency of the standard errors of estimates of the exchangeability parameters. Comparison of the versions of the model with and without intercept indicates that the intercept term is not statistically significant (p-value = 0.67). I also checked that there is no scaling problem with the estimates of the exchangeability parameters (there is no hidden constant scalar multiple that differs across training sets of different sizes). The parameter estimates (averaged across partitioned datasets of a particular sample size) at the three different sample sizes (n = 50, 100, 200) are similar, on average.

Table 4: Regression results of versions of model with and without intercept fitting empirical standard errors of estimates of HyPhy amino acid exchangeability parameters

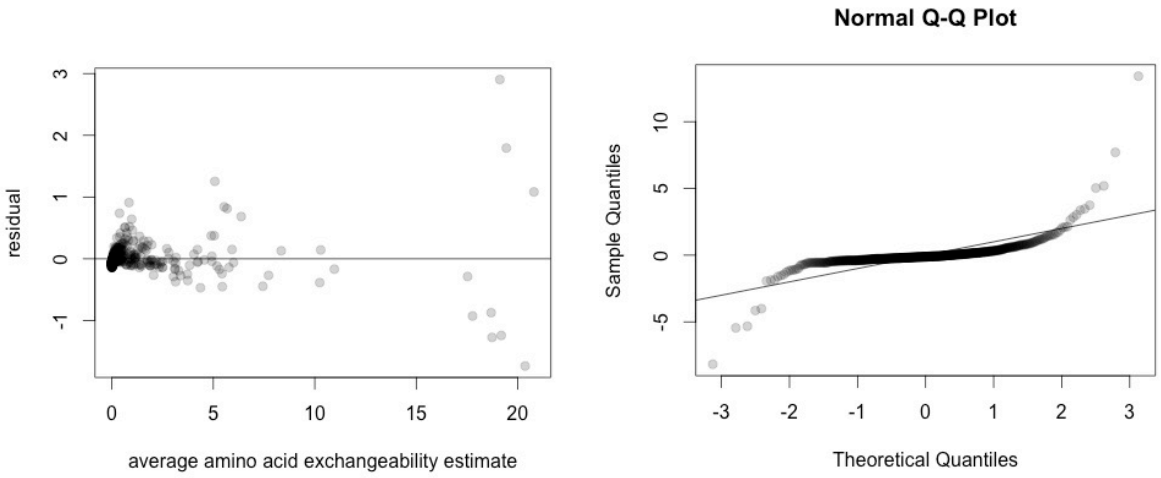
	Model With Intercept (R^2: Multiple=0.79; Adjusted=0.78)		Model Without Intercept (R^2: Multiple=0.83; Adjusted=0.83)	
	Coef. Estimate	P-value	Coef. Estimate	P-value
Inverse Root n	0.79	0.035924	0.95	$< 2x10^{-16}$
Rate Estimate	0.05	0.000373	0.053	8.3×10^{-5}
Interaction Term	1.12	$< 2x10^{-16}$	1.098	$< 2x10^{-16}$
Intercept	0.019	0.643296	-	-

Besides that, the model without intercept looks reasonable as the points in the residual versus average estimated parameter (obtained by averaging estimates of amino acid exchangeability parameters across all partitioned datasets of a particular sample size) plot are centered around y=0 (Figure 2a). The normality assumption seems to be reasonable as well (Figure 2b). Based on empirical standard errors of the parameter estimates, three amino acid exchangeability parameters appear to have larger estimated parameter values and empirical standard errors of parameter estimates than the rest: Phenylalanine <-> Tyrosine (F-Y), Lysine <-> Arginine (K-R) and Isoleucine <-> Valine (I-V). Interestingly, these

exchangeabilities are transitions between two amino acids of the same physicochemical property: F-Y is a transition between two aromatic amino acids, K-R is a transition between two basic amino acids and I-V is a transition between two hydrophobic amino acids. This leads to some observed potential heteroscedasticity in the linear model used (Figure 2a), which is in a region with poor resolution of the model.

Figure 2: Diagnostics plots of the model in Equation 4 without intercept: (a) Residual versus average HyPhy amino acid exchangeability estimate (averaged across partitioned datasets of a particular sample size) plot; black line corresponds to y=0 line (b) QQ plot of standardized residuals; black line corresponds to x=y line

(a) (b)



Since the goal of this work is to identify a conservative standard error model, overestimating the standard errors of these larger-valued rates is preferred over underestimating the standard errors, which may happen in an improperly specified homoscedastic model. Alternative models were evaluated and the most conservative model that gives the largest estimated standard errors was used for the primary comparison of matrices. A two-part linear model for standard errors of HyPhy amino acid exchangeability estimates appears to be the most conservative model. In both parts of the model, the intercepts do not appear to be significantly different from zero (p-value for first part of model = 0.67; p-value for second part

of model = 0.52). A test of heteroscedasticity (Fox & Weisberg, 2014) was carried out on the two-part model without intercept and the test indicates heteroscedasticity in the first part of the two-part model (p-value < 0.001). The first part of the model accounting for heteroscedasticity in the data is used to model standard errors of exchangeability parameter estimates other than the F-Y, I-V and K-R exchangeabilities and was built using the corresponding data points. There is no difference in the point estimates of regression coefficients between the regression analysis assuming heteroscedasticity and the regression robust to violation of the aforementioned assumption, only the statistical significance of the coefficients changes. Table 5 presents regression results accounting for heteroscedasticity of the first part of the two-part model. On the other hand, the second part of the model is used to model standard errors of the estimates of the F-Y, I-V and K-R exchangebilities and was built using only data points corresponding to those parameters. Because there are only nine data points available, this part of the model does not include an interaction term between the estimated parameter value and inverse of root n to avoid overfitting. Table 5 presents regression results of the second part of the two-part model. From the residual versus average rate estimate plots for the two parts of the model (Figure 3a and Figure 4a), it appears that both parts fit the data reasonably well. The normality assumptions for both parts also seem to be reasonably satisfied (Figure 3b and Figure 4b). The final two-part linear model for predicting of standard errors of parameter estimates is presented in Equation 5.

Table 5: Regression results of the two-part model without intercept term

	First Part of Model*		Second Part of Model (R^2: Multiple=0.83; Adjusted=0.78)	
	Coef. Estimate*	Corrected P-value*	Coef. Estimate	P-value
Inverse Root n	0.92799	$< 2x10^{-16}$	22.92669	0.303
Rate Estimate	0.07136	0.03451	0.04512	0.71
Interaction Term	0.97613	0.008939	-	-

*From regression analysis accounting for heteroscedasticity in data

Equation 5: Two-part linear model for standard errors of estimates of HyPhy amino acid exchangeability parameters (part 1 is for all amino acid exchangeabilities except for F-Y, K-R and I-V; part 2 is for those exchangeabilities only)

$$Part\ 1:\ E(SE) = \ 0.92799 * 1/\sqrt{n}\ \ + 0.07136 * est + \ 0.97613 * est * 1/\sqrt{n}$$

$$Part\ 2:\ E(SE) = \ 22.92669 * 1/\sqrt{n} + 0.04512 * est$$

Figure 3: Diagnostics plots of the first part of the two-part model: (a) Residual versus average HyPhy amino acid exchangeability estimate (averaged across partitioned datasets of a particular sample size) plot; black line corresponds to y=0 line (b) QQ plot of standardized residuals; black line corresponds to x=y line

(a) (b)

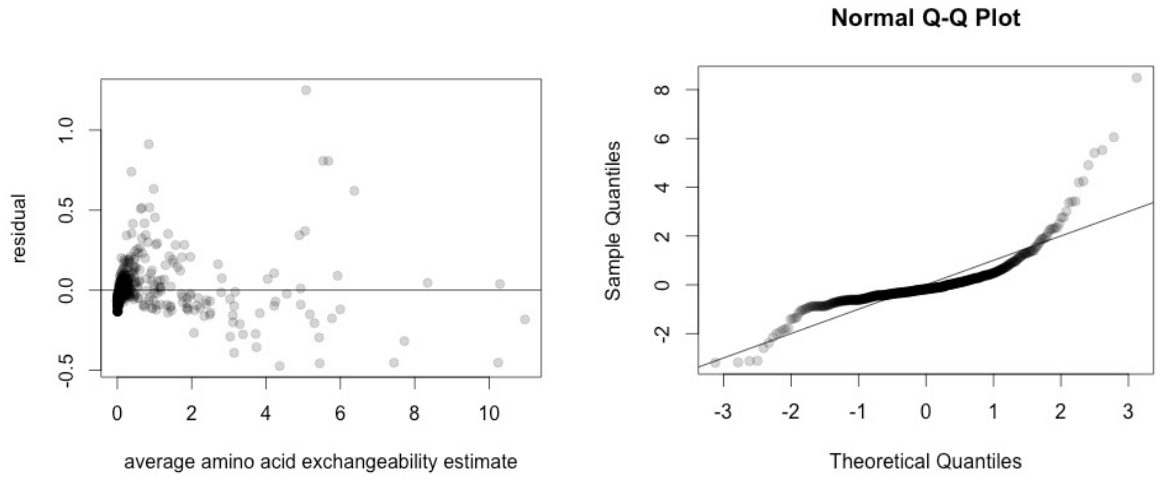
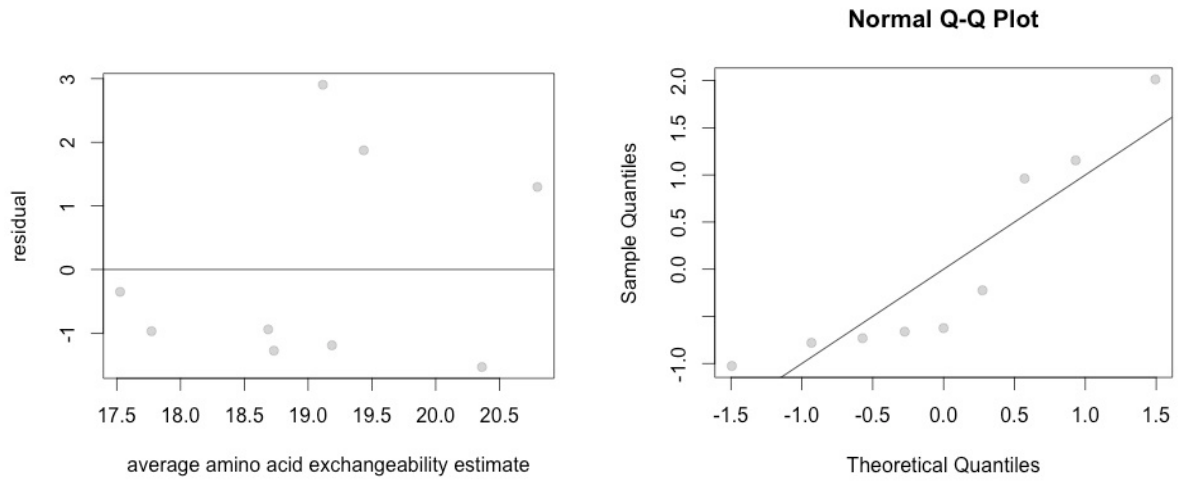


Figure 4: Diagnostics plots of the second part of the two-part model: (a) Residual versus average HyPhy amino acid exchangeability estimate (averaged across partitioned datasets of a particular sample size) plot; black line corresponds to y=0 line (b) QQ plot of standardized residuals; black line corresponds to x=y line

(a) (b)



The two-part model is built on the training partitions of HIV clade B Gag amino acid sequence data. To determine whether the model is applicable to data on a different HIV gene and clade, empirical standard errors of HyPhy amino acid exchangeability estimates across ten HIV clade C Env partitioned datasets each containing 57 sequences were compared to the standard errors predicted by the two-part model. The results seem to indicate that the two-part linear model is applicable to parameter estimates from a different HIV clade and gene.

There are other alternative approaches one might employ to obtain standard errors of parameter estimates. One alternative is to estimate standard errors via bootstrapping. However, bootstrapping sequences is not appropriate in this context because sampling the same sequences would provide no extra information and create a degenerate phylogenetic tree. Besides that, running HyPhy on many bootstrapped datasets would take a very long time, and is thus impractical. Another alternative is to obtain sandwich estimators of the standard errors of the rate estimates via delta method. However, no delta method is readily available for phylogenetic and substitution matrix inference in the HyPhy model. The

amino acid sequences are very high-dimensional data; thus deriving a high-dimensional multivariate delta method for this problem is beyond the scope of this thesis.

2.5 Comparison of HyPhy Amino Acid Exchangeability Matrices Using Predicted Standard Errors

Bonferroni correction is used for multiplicity adjustment of comparing 189 amino acid exchangeability parameter estimates between two matrices (excluding the Isoleucine <-> Leucine exchangeability that is constrained to 1 in HyPhy). The per-test significance level used is 0.05 divided by the number of matrix cell values tested. If any individual hypothesis is rejected, the overall null that the two matrices compared are the same is rejected conservatively at a 5% type-1 error rate. The formula of the Z statistics to calculate a p-value is shown in Equation 6. The standard error (denominator) used in the Z statistics is the square root of the sum of the squared standard errors of the estimates of exchangeability parameters in the two matrices being compared predicted by the two-part linear model in Equation 5 (first part or second part of the model is used depending on the estimate of amino acid exchangeability parameter being compared). To determine if HIV gene-specific amino acid exchangeability matrices are different, exchangeability matrix estimated by HyPhy using a HIV clade C Env dataset was compared to exchangeability matrix estimated using a HIV clade C Gag dataset. On the other hand, to determine whether HIV clade-specific amino acid exchangeability matrices are different, exchangeability matrix estimated by HyPhy using a HIV clade B Gag dataset was compared to exchangeability matrix estimated by HyPhy using a HIV clade C Gag dataset.

Equation 6: Formula of Z statistics

$$Z = \frac{est_1 - est_2}{\sqrt{se_1^2 + se_2^2}}$$

(Note: est_1 and est_2 are parameter estimates in the two matrices being compared; se_1 and se_2 are the corresponding predicted standard errors)

3 RESULTS

3.1 Comparison of HyPhy Standard Errors and Empirical Standard Errors of Parameter Estimates

From Figure 5, it appears that HyPhy standard errors of estimates of the exchangeability parameters in one partitioned dataset of size 200 are similar to the empirical standard errors across partitioned datasets of size 200, on average. HyPhy standard errors are quite similar to empirical standard errors across partitioned datasets of size 50 as well. However, there are two outliers at sample size 50. It appears that HyPhy standard errors for the Phenylalanine <-> Tyrosine (F-Y) and the Cysteine <-> Histidine (C-H) exchangeability parameters estimated from a partitioned dataset of size 50 are quite large compared to the empirical standard errors. Table 6 summarized the HyPhy and empirical standard errors of the estimates of those two parameters.

Figure 5: Difference between empirical and HyPhy standard errors for 189 estimates of amino acid exchangeability parameters in partitioned datasets of size 50 and size 200

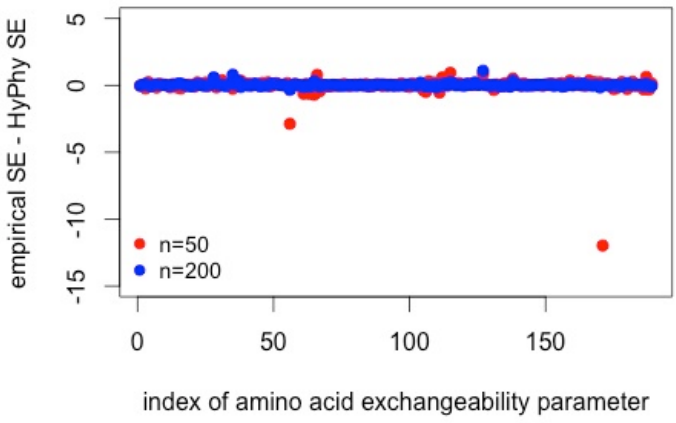


Table 6: HyPhy and empirical standard errors across partitioned HIV clade B Gag datasets of size 50 for estimates of F-Y and C-H exchangeability parameters

AA Exchangeability	HyPhy SE	Empirical SE
F-Y	19	7
C-H	4	0.95

Excluding data points corresponding to the F-Y and C-H exchangeabilities, the paired t-test comparing the standard error difference at n=50 and standard error difference at n=200 does not indicate any significant difference (p-value = 0.6926). Although there is lack of sufficient evidence to reject the null hypothesis that HyPhy standard errors are not underestimated, there are several reasons to distrust HyPhy's estimates of standard errors as discussed in Section 2.3. Hence, this work focuses on using standard errors predicted from the two-part linear model for comparison of HyPhy amino acid exchangeability matrices. In addition, standard errors from the linear model can be used on other transformations of the exchangeability matrices whilst HyPhy standard errors are only for the estimated parameters of the exchangeability matrices.

3.2 Applicability of Standard Error Model to a Different HIV Clade and Gene

To determine whether the two-part linear model for standard errors is applicable to a different HIV clade and gene, I compared the empirical standard errors of HyPhy amino acid exchangeability parameter estimates across ten HIV clade C Env partitioned datasets each containing 57 sequences to the standard errors predicted by the linear model. From the histogram of differences between empirical and predicted standard errors in Figure 6, it appears that most empirical and predicted standard errors of parameter estimates are similar since the mode of the histogram is around 0. It is noteworthy that predicted standard errors for the estimates of the F-Y, I-V and K-R exchangebility parameters are larger than the corresponding empirical standard errors. However, since the goal here is to have a conservative model of standard error, it is acceptable to have predicted standard errors that are larger than the empirical ones.

Table 7 summarizes the predicted and empirical standard errors for the estimates of the three amino acid exchangeability parameters.

Figure 6: Histogram of differences between predicted standard errors of estimates of HyPhy exchangeability parameters and empirical standard errors across ten partitioned HIV clade C Env datasets each containing 57 sequences

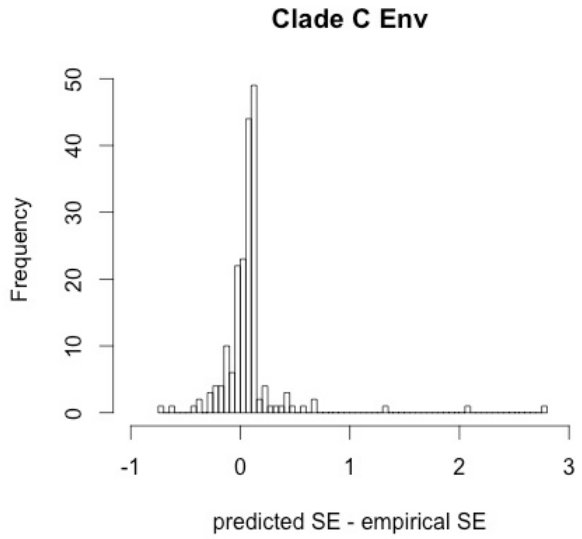


Table 7: Predicted and empirical standard errors across partitioned HIV clade C Env datasets of size 57 for estimates of F-Y, I-V and K-R amino acid exchangeability parameters

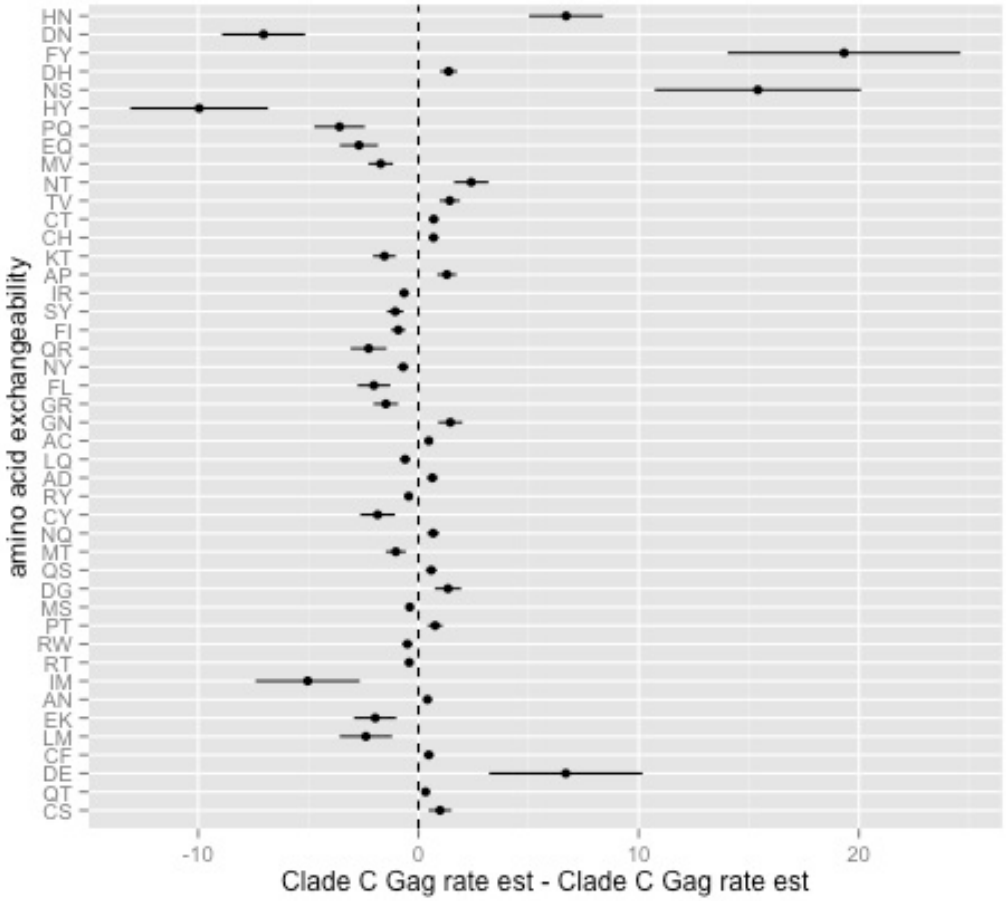
AA Exchangeability	Predicted SE	Empirical SE
F-Y	3.4	2.1
I-V	3.4	0.6
K-R	3.6	1.5

3.3 Comparison of HIV-Specific Amino Acid Exchangeability Matrices

To determine whether HIV gene-specific matrices are different, exchangeability matrix estimated by HyPhy using HIV clade C Env amino acid sequence data and exchangebility matrix estimated by HyPhy using HIV clade C Gag data were compared. Using standard errors predicted by the two-part linear

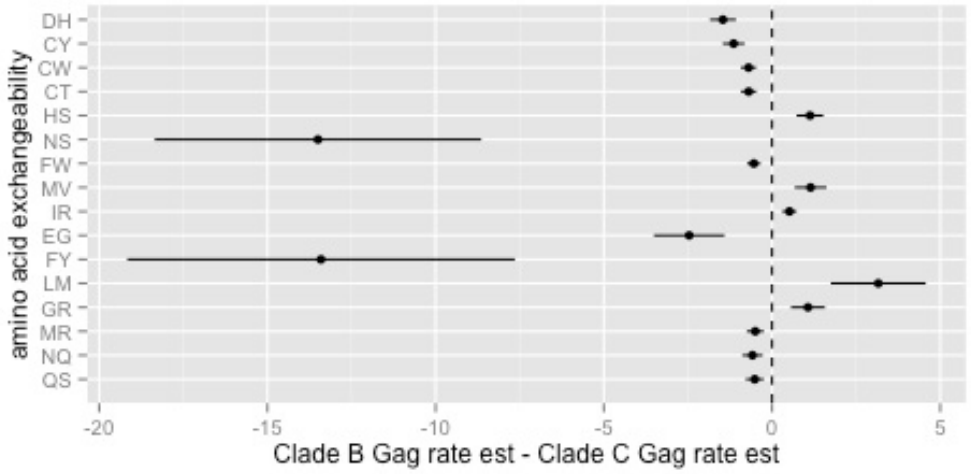
model, 44 out of 189 exchangeability parameter estimates are significantly different between the two matrices at a significance level of 0.05/189. Figure 7 shows 95% confidence intervals for the differences of the exchangeability estimates for exchangeabilities that are significantly different between the two matrices compared. On the other hand, 80 out of 183 parameter estimates are significantly different between the two matrices at a significance level of 0.05/183 when using standard errors provided by HyPhy. There are 6 estimates of amino acid exchangebilities whose standard errors that are degenerate or that HyPhy failed to estimate in one of the two datasets.

Figure 7: 95% confidence intervals of difference of amino acid exchangeability estimates for exchangeabilities that are significantly different between HIV clade C Env and HIV clade C Gag HyPhy matrices; exchangeabilities are ordered from most significant on top to least significant at the bottom



To determine whether HIV clade-specific matrices are different, exchangeability matrix estimated by HyPhy using HIV clade B Gag data and exchangeability matrix estimated by HyPhy using HIV clade C Gag data were compared. Using standard errors predicted by the two-part linear model, 16 out of 189 exchangeability parameter estimates are significantly different between the two matrices at a significance level of 0.05/189. Figure 8 shows 95% confidence intervals for the differences of the exchangeability estimates for exchangeabilities that are significantly different between the two matrices compared. On the other hand, 75 out of 165 parameter estimates are significantly different between the two matrices at a significance level of 0.05/165 when using standard errors provided by HyPhy. There are 24 estimates of amino acid exchangebilities whose standard errors that are degenerate or that HyPhy failed to estimate in one of the two datasets.

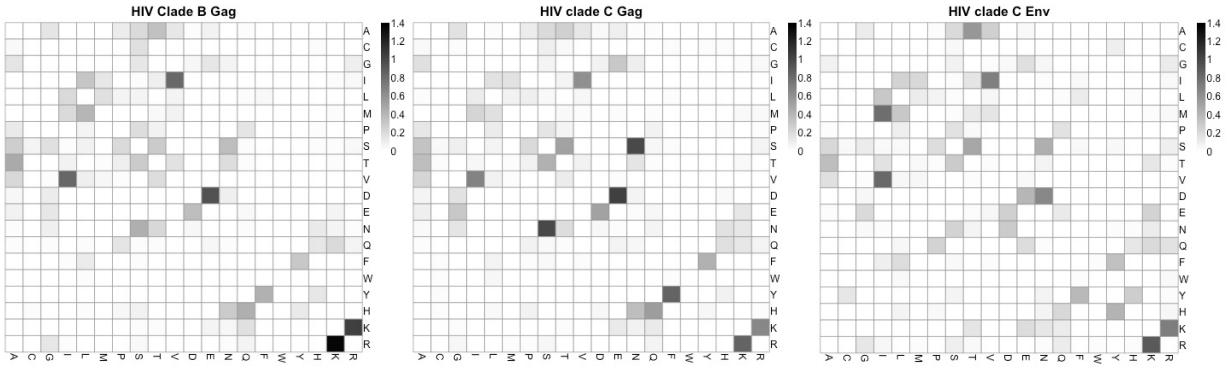
Figure 8: 95% confidence intervals of difference of amino acid exchangeability estimates for exchangeabilities that are significantly different between HIV clade B Gag and HIV clade C Gag HyPhy matrices; exchangeabilities are ordered from most significant on top to least significant at the bottom



In general, it appears that using standard errors from the two-part linear model gives a more conservative comparison of the HyPhy amino acid exchangability matrices. Fewer estimates of exchangeability parameters are declared to be significantly different for the comparisons of gene-specific matrices and clade-specific matrices.

Figure 9 shows visual summaries of instantaneous rate matrices for the HIV clade B Gag, HIV clade C Gag and HIV clade C Env datasets. Each instantaneous rate matrix is obtained by multiplying each element in the amino acid exchangeability matrix with the frequency of amino acid observed in the dataset being analyzed corresponding to the matrix column (Equation 1). The three transitions that have previously been identified as interesting in Section 2.4 seem to have relatively large instantaneous rate estimates in the three instantaneous matrices based on datasets analyzed in this work, especially for the Isoleucine <-> Valine (I-V) and Lysine <-> Arginine (K-R) transitions and less so for the Phenylalanine <-> Tyrosine (F-Y) transition. Among the three matrices, the estimate value for the K-R transition is largest in the HIV clade B Gag matrix. The estimate value for the I-V transition is slightly larger in the HIV clade B Gag matrix compared to the other two matrices. F-Y transition has the largest estimated value in the HIV clade C Gag matrix and much less so in the HIV clade B Gag and HIV clade C Env matrices. Other transitions with large estimates are transitions involving amino acids with the same physicochemical property (e.g. the Aspartic Acid <-> Glutamic Acid transition where both amino acids are acidic, the Isoleucine <-> Methionine transition where both amino acids are hydrophobic) and transitions involving amino acids whose codons are different by one nucleotide (e.g. the Alanine <-> Threonine transition and the Serine <-> Asparagine transition).

Figure 9: Instantaneous rate matrices estimated by HyPhy using HIV clade B Gag, HIV clade C Gag and HIV clade C Env sequence data, respectively



3.4 Comparison of Amino Acid Exchangeability Matrices Under the Null

As a sanity check, the HyPhy amino acid exchangeability matrices were compared under approximate "null" hypotheses using standard errors predicted by the two-part linear model. Three comparisons were performed to attempt to get at the correct null. Firstly, two datasets were obtained by randomly partitioning the pooled dataset of 440 HIV clade B Gag sequences and 678 HIV clade C Gag sequences. The exchangeability matrices estimated by HyPhy using the two datasets were compared via comparing the 189 amino acid exchangeability parameter estimates. There is one exchangeability (Cysteine <-> Tyrosine) with significantly different parameter estimates between the two matrices. The second comparison was done on two datasets formed by randomly partitioning the HIV clade B Gag dataset and there is one significantly different exchangeability estimate (Cysteine <-> Histidine). Thirdly, two matrices estimated from using datasets containing equal number of clade B Gag and clade C Gag sequences were compared. Surprisingly, there are eleven amino acid exchangeability parameters with significantly different estimates between the two matrices. In all three comparisons, it is still unclear whether the matrices were compared under the right "null". This issue is further discussed in Section 4.2.

4 DISCUSSION

4.1 Summary of Results

In this work, I compared HIV-specific substitution matrices in a way that takes into account sampling errors in estimation of the matrices. I used standard errors of parameter estimates predicted from a two-part linear model as well as standard errors provided by HyPhy to compare two matrices via comparison of 189 estimates of amino acid exchangeability parameters with Bonferroni correction. 189 cell values in the amino acid exchangeability matrix estimated by HyPhy using HIV clade B Gag dataset were compared to the values in the exchangeability matrix estimated by HyPhy using HIV clade C Gag dataset. The same comparison was carried out for the exchangeability matrix estimated using HIV clade C Gag dataset and the matrix estimated using HIV clade C Env dataset. In both comparisons performed using predicted standard errors and HyPhy standard errors, there is statistical evidence that the clade-specific matrices (HIV clade B Gag exchangeability matrix versus HIV clade C Gag exchangeability matrix) are different, as well as the gene-specific matrices (HIV clade C Gag exchangeability matrix versus HIV clade C Env exchangeability matrix). Furthermore, it is known that HIV clade C is more (Foley, 2000) than HIV clade B and that the Gag gene is more conserved than the Env gene (Travers et al., 2004). The results of this work futher support the distinct evolutionary patterns in different HIV clades as well as HIV genes. More broadly, the results motivate the use of HIV-specific matrices that are even more granular than the current standard within- and between-host matrices, which will hopefully allow for more accurate sequence alignment, phylogenetic inference and sequence comparison.

Another interesting finding of this work is that amino acid transitions with large estimated exchangeability parameters and that are harder to estimate are transitions between amino acids of the same physicochemical property. Three such biochemically interesting transitions are identified in our work: Phenylalanine <-> Tyrosine, Isoleucine <-> Valine and Lysine <-> Arginine.

4.2 Limitations

The main limitation with the two-part linear model for standard errors of parameter estimates is the extrapolation to larger sample sizes. One potential way to determine the validity of model extrapolation is by obtaining enough partitioned datasets of a particular HIV clade and gene each containing more than 200 sequences and computing empirical standard errors to be compared against standard errors predicted by the two-part linear model. Unfortunately, there is not enough data to do so.

Besides that, there is the question of how closely the estimated matrices reflect the true evolutionary pattern in HIV. This is because the sequences used to estimate these matrices are mostly HIV sequences from individuals with chronic HIV infections. The issues with estimating HIV-specific substitution matrices are different from other common matrices such as Dayhoff and BLOSUM. Not only is HIV a unique entity with a unique substitution pattern, as a virus (a pseudo-organism), it is also evolving under a set of pressures that are different from so-called organisms. In particular, many HIV sequences are not viable due to errors introduced during the reverse transcription step. Hence some point mutations observed in HIV sequences, especially those from individuals with chronic HIV infections, might not even be accepted by natural selection operating at the scale of transmissible virus. Nonetheless, when using these HIV sequences to estimate substitution matrices, there is no way to know which mutations are accepted point mutations (accepted by natural selection) and which ones are not. Ideally HIV sequences from acutely infected individuals should be used to estimate matrices that would properly model the evolution of HIV viruses that are viable and capable of infecting new hosts. Moreover, the constraints required of a phylogenetic tree are of dubious merit for modeling HIV because of genetic recombination in HIV. If the phylogenetic tree is not the correct model, there will be implications on methods that estimate amino acid substitution parameters that assume a tree structure, such as HyPhy.

It is also somewhat alarming that there is evidence that two matrices are different under approximately null data. This is at least partly due to the fact that the comparison of matrices was not performed under the true null due to sampling non-iid (non identical and independently distributed) data. The sequences are non-iid because they are related through the tree; randomly sampling sequences ignores the tree structure. It is beyond the scope of this thesis to devise a method for sampling from the

phylogenetic tree of HIV sequences under the correct null. On the other hand, it remains possible that the estimates of SE are anti-conservative due to use of an imperfect model, and the overall results showing a large number of differing cells across clade-specific and gene-specific matrices should be interpreted with some caution. Nonetheless, for the comparison of two matrices estimated from two random partitioned datasets of the HIV clade B Gag data each containing 220 sequences, it is noteworthy that the standard errors for the parameter estimates are not extrapolations from the model since partitioned datasets containing at most 200 sequences are part of the training data (and since the standard error is modeled as a linear function of inverse of root n, the difference between the two numbers is even smaller). For this comparison, the observation of one significantly different parameter estimate between the two matrices under the "null" should be mainly due to the problem of sampling non-iid sequences under the "null" rather than problems with the predicted standard errors.

4.3 Future Directions

This work focuses on comparing 189 estimated amino acid exchangeability parameters in the HyPhy exchangeability matrix. In light of significant results under the approximate "null", it is possible that estimates of standard errors are anti-conservative due to the use of an imperfect model. To be more conservative, the upper bound of the 95% prediction interval for the standard error predicted from the model can be used instead.

On the other side of the coin, it would be useful to consider comparison of lower-dimensional summary of a substitution matrix such as the stationary distribution, which would then require only 20 tests instead of 190 tests. Other ideas include comparing the first principal component of the matrix and the trace of the matrix as well as projecting the matrix to transitions involving amino acids of the same physicochemical property and of different physicochemical properties. In addition, it is probably wise to consider what it means for two matrices to be different. Would we consider two matrices as different if only one cell has significantly different values in the matrices? In other words, when do two matrices that are statistically different start to matter practically? This should be answered in future work addressing the

applications of these matrices to determine whether the statistical differences reflect important differences in estimated alignments, phylogenies, and sieve effects.

Besides that, if using more granular substitution matrices would allow for more accurate analyses, one could argue for using substitution matrices that are even more granular than clade- or gene-specific matrices such as region-specific matrices (e.g. different matrices for different variable loops) and context-specific matrices (e.g. different matrices for sequence positions that are antibody contact sites). This increasing granularity would ultimately argue for the use of position-specific substitution matrices, which uses a different matrix for each site along a sequence.

REFERENCES

Altschul, S. F. (1991). Amino acid substitution matrices from an information theoretic perspective. *Journal of molecular biology*, 219(3), 555-565.

Buonaguro, L., Tornesello, M. L., & Buonaguro, F. M. (2007). Human immunodeficiency virus type 1 subtype distribution in the worldwide epidemic: pathogenetic and therapeutic implications. *Journal of virology*, 81(19), 10209-10219.

Dayhoff, M. O., & Schwartz, R. M. (1978). A model of evolutionary change in proteins. In *In Atlas of protein sequence and structure*.

Dessimoz, C., Gannarozzi, G. & Schneider, A. (2011). How to compute mutation and Dayhoff matrices. In *Bio-recipes (Bioinformatics recipes) in Darwin*. Retrieved from http://www.biorecipes.com/Dayhoff/code.html

Edlefsen, P. T., Gilbert, P. B., & Rolland, M. (2013). Sieve analysis in HIV-1 vaccine efficacy trials. *Current opinion in HIV and AIDS*, 8(5), 432-436.

Foley, B. T. (2000). An overview of the molecular phylogeny of lentiviruses. *HIV sequence compendium*, 2000, 35-43.

Fox, J. & Weisberg, S. (2014). Score Test for Non-Constant Error Variance. In *car: Companion to Applied Regression R Reference Manual*. Retrieved from http://cran.r-project.org/web/packages/car/car.pdf

Gilbert, P. B., Wu, C., & Jobes, D. V. (2008). Genome scanning tests for comparing amino acid sequences between groups. *Biometrics*, 64(1), 198-207.

Henikoff, S., & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22), 10915-10919.

Mount, D. W. (2007). Using the basic local alignment search tool (BLAST). *Cold Spring Harbor Protocols*, 2007(7), pdb-top17.

Nickle, D. C., Heath, L., Jensen, M. A., Gilbert, P. B., Mullins, J. I., & Pond, S. L. K. (2007). *HIV-specific probabilistic models of protein evolution*. PLoS One, 2(6), e503.

PhyML Explanation (2013). Retrieved from http://www.hiv.lanl.gov/content/sequence/PHYML/backgroundinfo.html

Pond, S. L. K., & Muse, S. V. (2005). HyPhy: hypothesis testing using phylogenies. In *Statistical methods in molecular evolution (pp. 125-181).* Springer New York.

Rose, P. P., & Korber, B. T. (2000). Detecting hypermutations in viral sequences with an emphasis on G→ A hypermutation. *Bioinformatics*, 16(4), 400-401.

Travers, S. A., Clewley, J. P., Glynn, J. R., Fine, P. E., Crampin, A. C., Sibande, F., Mulawa, D., McInerney, J. & McCormack, G. P. (2004). Timing and reconstruction of the most recent common ancestor of the subtype C clade of human immunodeficiency virus type 1. *Journal of virology*, 78(19), 10501-10506.

Whelan, S., & Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Molecular biology and evolution*, 18(5), 691-699.