Protein evolution is structure dependent and non-homogeneous across the tree of life*

Akanksha Pandey†
Department of Biology
University of Florida
Gainesville, FL 32611, USA
aakanksha.vit@gmail.com

Edward L. Braun†
Department of Biology
University of Florida
Gainesville, FL 32611, USA
ebraun68@ufl.edu

ABSTRACT

Protein sequence evolution is a complex process that varies across the tree of life and among-sites within proteins. Comparing evolutionary rate matrices for specific taxa ('clade-specific models') can reveal this variation and provide information about the basis for changes in the patterns of protein evolution over time. However, clade-specific models can only provide this information if the variation among taxa exceeds the variation among proteins. We showed this to be the case by demonstrating that clade-specific model fit could distinguish among proteins from the four taxa that we examined (vertebrates, plants, oomycetes, and yeasts). Model fit classified proteins correctly by clade of origin >70% of the time. A relatively small number of dimensions can explain differences among models. If model parameters are averaged across all sites ~80% of the variance among models reflects clade; for models that consider protein structure ~50% of the variance reflected relative solvent accessibility and ~25% reflected clade. Relaxed purifying selection in taxa with smaller long-term effective population sizes appears to explain much of the among clade variance. Relaxed selection on solvent-exposed sites was correlated with the degree of change in amino acid side-chain volume for substitutions; other differences among models were more complex. Beyond the information they reveal about protein evolution, our clade-specific models also represent tools for phylogenomic inference. Availability: model files are available from https://github.com/ebraun68/clade_specific_prot_models.

CCS CONCEPTS

• Molecular evolution • Population genetics • Molecular Structural Biology • Bioinformatics

KEYWORDS

Substitution matrix, Purifying selection, Effective population size, Protein structure, Relative solvent accessibility

*Availability: model files, code, and additional data are available from github: https://github.com/ebraun/clade_specific_prot_models

†Address correspondence to Edward L. Braun

ACM-BCB'20, September, 2020, Virtual Meeting

DOI: 10.1145/3388440.3412473

ACM Reference format:

Akanksha Pandey and Edward L. Braun. 2020. Protein evolution is structure dependent and non-homogeneous across the tree of life. In *Proceedings of ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM-BCB'20). ACM, New York, NY, USA, 11 pages*

1 Introduction

In phylogenetics, tree topology and/or branch lengths are typically the parameters of interest. However, amino acid rate matrices, which have been studied since the dawn of computational biology as a field [1], also provide information about the process of evolution. Patterns of amino acid substitution vary across the tree of life [2], [3] and among proteins [4]. It has long been appreciated [5] that the accumulation of substitutions over evolutionary time reflects two processes: 1) the rate at which novel mutations enter populations; and 2) the impact of drift and selection on the fate of those mutations. This paradigm suggests the patterns of protein evolution will vary across the tree of life; after all, the rate and spectrum of mutations and strength of selection (the latter reflecting, in large part, variation in effective population size, N_e) varies across the tree [6], [7]. The sensitivity of ratio of radical to conservative amino acid substitutions to N_e [2], [8] suggests variation in the strength of selection is likely to be especially important for establishing the patterns of protein evolution.

Using the radical to conservative substitution rate ratio to examine changes in the pattern of sequence evolution is complicated by the challenge of defining radical (i.e., non-conservative) amino acid changes. Zuckerkandl and Pauling ([9], p. 129) recognized that the "…inadequacy of *a priori* views on [amino acid substitution] conservatism and nonconservatism is patent" in the very earliest days of molecular evolution and that problem remains unsolved. Many studies divide residues into two categories (e.g., polar/non-polar or small/large) and treat between-category substitutions as radical [8]. That idea can be extended by using continuous values to describe the physicochemical characteristics of the amino acids instead of binary classification [4], but that still relies on the use of prespecified amino acid characteristics. Assessing changes in the process of protein sequence evolution without *a priori* assumptions would be desirable.

The general time-reversible model for amino acids (GTR_{20}) might provide a practical way to address this question. The GTR_{20}

instantaneous rate matrix (or $\boldsymbol{Q}$ matrix) can be decomposed into a symmetric rate ($\boldsymbol{R}$) matrix with 189 free parameters reflecting the 'exchangeability' of each pair of amino acids and a diagonal matrix ($\boldsymbol{\Pi}$) with 19 free amino acid equilibrium frequency parameters [10]. Of course, the time-reversibility assumption that limits the number of free parameters is inappropriate when protein evolution has changed across the tree. After all, postulating that the model changes over time (i.e., non-homogeneity) intrinsically renders models non-time-reversible. However, we can avoid this problem by estimating GTR_{20} parameters for clades with a limited taxonomic scope and then comparing the clade-specific parameter estimates. If the deviations from time-reversibility for the underlying model of protein evolution are sufficiently limited within clades, then comparisons among clades should reveal the ways that protein evolution has changed across the tree of life.

Using GTR_{20} parameter estimates to understand shifts in the process of sequence evolution presents several challenges. Previous studies [3], [11] indicate that we will find differences among clades. However, it is reasonable to expect substantial variation among proteins [4] and among sites within proteins [12], [13]. There are two ways this heterogeneity could confound our ability to use GTR_{20} parameter estimates to understand patterns of protein evolution across the tree of life. First, a high degree of variation among individual proteins might obscure variation among clades (Fig. 1). Second, simply optimizing GTR_{20} model parameters on a large protein dataset will yield average exchangeability estimates for all sites. Patterns revealed by comparing these 'averaged' parameter estimates could be confusing if there is substantial variation among-sites within proteins. These factors make it important to find ways to examine the impact of these sources of variation on any conclusions we reach regarding differences among taxa.

One way to assess fine-scale variation in protein evolution (i.e., variation among individual proteins) would be to estimate GTR_{20} parameters for individual proteins and compare them. This is not practical; GTR_{20} has too many parameters to obtain accurate parameter estimates using individual protein alignments. However, it is possible to estimate model parameters using a relatively large training dataset and then classify the proteins in an independent validation dataset. Hereafter, we will call the GTR_{20} matrices estimated as part of this study 'models' because they are analogous to the empirical models that are often used in protein phylogenetics, such as the PAM [14], JTT [15], WAG [16], and LG [17]; we call those models (and similar models that are trained on diverse datasets) 'standard empirical models.' Using clade-specific and standard empirical models to classify individual proteins will allow us to establish the part of the parameter space shown in Fig. 1 that best describes the large-scale patterns of protein evolution. If the variance among individual proteins exceeds the variation among clades (lower portion of Fig. 1) clade-specific models should be a poor classifier. In contrast, if variation among clades exceeds the variation among proteins (upper portion of Fig 1), clade-specific models as classifiers should be good classifiers (i.e., the best-fitting model for validation set proteins will be the model generated from that clade). Finally, the number of times that model fit fails as a classifier will increase as the variation among proteins increases. It should be possible to establish the specific parameters that vary among clades and determine whether they are consistent with predictions regarding the expected differences among clades in the strength of selection, assuming there is sufficient variation among clades.

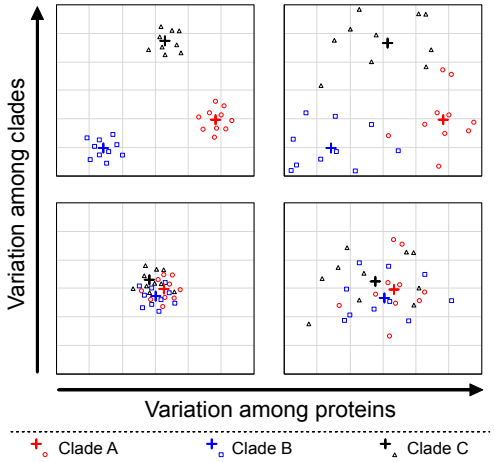

Figure 1: Possible patterns of variation in patterns of protein sequence evolution, both among proteins and among clades. Conceptual illustration showing the relationships among protein models where the underlying models are presented after some type of dimension reduction. Crosses indicate models generated using the training data (i.e., average parameter estimates); smaller circles, squares, and triangles indicate individual proteins. The number of dimensions necessary to summarize the GTR_{20} rate matrices is unclear; we are showing two dimensions to illustrate the idea underlying our analytical framework.

The other type of fine-scale variation, variation among sites within proteins, is more difficult to examine. Patterns of protein evolution are complex [18] and the best way to extract information about the patterns of molecular evolution while still acknowledging variation within proteins remains unclear. However, selection to maintain protein structure, which has a fundamental role in maintaining protein function, is likely to play a major role in the overall structure of amino acid substitution matrices [19]. The relative solvent accessibility (RSA) of individual amino acids is one of the most important determinants of the patterns of sequence evolution for globular proteins [20], [21]. This suggests it should be possible to subdivide proteins into solvent exposed (high RSA) and buried (low RSA) sites before estimating substitution matrix parameters for various clades. This would add another dimension to the parameter space shown in Fig. 1 (i.e., a dimension describing variation among sites within proteins). It also makes it necessary to use a mixture model as a classifier (i.e., a model with two 'sub-models' where the site likelihoods are calculated as a weighted mixture of both sub-model matrices). However, using these exposed/buried ('XB') mixture models is a straightforward extension of the idea of using models as a classifier to determine which part of parameter space best describes the large-scale patterns of protein evolution.

Herein, we examine the extent to which models of protein sequence evolution exhibit clade-specific features using six

eukaryotic datasets selected to exhibit differences in the strength of selection. These clades selected for this study included vertebrates (expected to have small long-term N_e), plants (expected to have intermediate long-term N_e), and microbial eukaryotes (expected to have large long-term N_e). We focused on eukaryote datasets to limit the impact of horizontal gene transfer on parameter estimates; the high rate of horizontal gene transfer in prokaryotes [22] could distort estimates. We added a seventh dataset with a broad sample of eukaryotes; multiple changes in the underlying model of sequence evolution are likely to have occurred for the taxa in that dataset (unless the lower part of Fig. 1 is the best description of protein evolution). This 'all Euk' dataset was included to assess the impact of using a dataset that has experienced changes in the rate matrix. We then used the new models as classifiers to assess among-protein variation and examine the way models differ, examining parameter differences among clades, among sites that were grouped by RSA, and for the combination of RSA and clade.

2 Methods

We generated 14 new models of protein evolution (seven based on all sites and seven XB mixture models) by selecting proteins from seven datasets (Table 1). One training dataset [23] included non-coding data; for that dataset we extracted the coding exons, added orthologous sequences from 117 avian genome assemblies (using the Reddy *et al.* [24] pipeline to extract data from genomes), re-aligned the data using MAFFT v.7.130b [25], and translated the data to yield amino acid alignments. All training datasets were concatenated; we have made the data files for this project available in Zenodo (http://doi.org/10.5281/zenodo.3964471).

Table 1. Training datasets selected for this study

Clade	# Proteins/Sites	# Taxa	Model	Citation
Birds (1)	250/109,969	48	JTT	[26], [27]
Birds (2)	250/161,112	317	HIVb	[23]
Mammals	249/238,319	116	HIVb	[28]
Plants (1)	310/80,315	46	JTT	[29]
Oomycetes	277/83,312	17	LG	[30]
Yeasts	200/81,802	343	LG	[31]
All Euk (1)	248/58,469	149	LG	[32]

We estimated model parameters using IQ-TREE [33] v. 1.6.10, as implemented in CIPRES science gateway [34]. Before conducting full model optimization, we identified the best-fitting standard empirical model for each training dataset using the -m TEST option with AIC_c as the decision criterion. The best-fitting standard empirical model varied among clades (Table 1), but the rate heterogeneity parameters for all best-fit models included both invariant sites and Γ-distributed rates. Thus, we used GTR_{20}+I+Γ to estimate the new clade-specific models. We fixed the tree topology and among-sites rate heterogeneity parameters (the Γ-distribution shape parameter and proportion of invariant sites) based on the analysis using the standard empirical model before optimizing the other model parameters (the exchangeabilities, equilibrium amino acid frequencies, and branch lengths) by maximum likelihood. Based on the AIC_c the GTR_{20}+I+Γ model had a better fit to the training data than the best-fitting empirical model (see supplementary file in github and Zenodo) in all but one case (the 'all Euk' all sites model; see Results and Discussion for details). The new clade-specific models are available from github in a format usable by IQ-TREE and PAML [35].

We selected six validation datasets (Table 2; available from Zenodo). In most cases, the datasets in Table 1 had enough alignments to divide them into training and validation sets. However, we used all genes in the plant and 'all Euk' training datasets (we refer to datasets with a broad sample of eukaryotes as 'all Euk' datasets). In those two cases, we selected another dataset with comparable taxa ('plant 2' and 'all Euk 2') to use as the validation set. We eliminated overlaps between these two training and validation datasets by removing proteins from the validation dataset if they had a BLAST [36] *E*-value ≤ 10^{-40}. We identified the best-fitting model for each validation set protein using IQ-TREE with the -mset option; we tested the seven clade-specific models and the 18 standard empirical models implemented in IQ-TREE. For comparison, we conducted the same analyses using the proteins in training datasets.

Table 2. Validation datasets selected for this study

Clade	# Proteins	# Taxa	Citation
Birds (1)	200	48	[26], [27]
Mammals	200	116	[28]
Plants (2)	200	107	[37]
Oomycetes	150	15	[30]
Yeasts	200	343	[31]
All Euk (2)	149	104	[38]

To reduce within-protein heterogeneity, we separated globular proteins into exposed and buried sites. First, we used TMHMM [39] to identify and remove transmembrane proteins. Then the globular proteins were subdivided based on site RSA using the https://github.com/aakanksha12/Structural_class_assignment_pipeline pipeline. That pipeline generates a weighted consensus sequence that is then used as input for ACCpro [40] from the SCRATCH-1D suite [41]. ACCpro assigns each residue to one of the two categories: exposed or buried, with the latter defined as <25% RSA. After subdividing the data into exposed and buried residues the data for all proteins were concatenated, resulting in 14 data matrices (one exposed dataset and one buried dataset for the seven datasets in Table 1). There many aspects of protein structure that can be considered in of models of sequence evolution, ranging from the straightforward use of RSA that we considered in the study, to secondary structure [17], [21], [42]–[44], residue-residue interactions [43], to details like amino acid torsion angles [45]. We focused on RSA because that aspect of protein structure has a substantial impact on the rate matrix (see Fig. 4 in Pandey and Braun [21]) and it made it straightforward to construct training datasets of sufficient size for this study.

Rate matrices were estimated for all 14 exposed- and buried-site training datasets as described above. The exposed and buried rate matrices were then combined to create seven XB mixture models,

in which the site likelihoods are weighted averages over both alternative (exposed and buried) matrices. We identified the best-fitting mixture model for individual proteins in the validation datasets using 'fit_mixture_model.pl' (available from github), which examines the fit of eight models: the seven XB models we generated and the Le *et al.* [10] EX2 model (which is also an exposed/buried mixture model). We tested four versions of each model that differed in their treatment of rate heterogeneity (no rate heterogeneity vs. Γ-distributed rates) and equilibrium amino acid frequencies (observed frequencies vs maximum likelihood estimates [+FO]). As above, we also assessed model fit for all proteins in training datasets. The XB models are available from github in a nexus file that can be used by IQ-TREE.

We analyzed the GTR_{20} exchangeability (***R***) matrices for standard empirical models and the new models generated in two ways. First, we clustered Euclidean distances among matrices by neighbor-joining [46]. Second, we used principal component analysis (PCA) to explore differences among models. We normalized the matrix elements (i.e., the 190 exchangeability values for each pair of amino acids) to sum to one for both analyses. The Euclidean distances were calculated by treating the normalized ***R*** matrix as a vector with 190 values. We used three normalized vectors for the PCAs: 1) vectors with all 190 elements; 2) vectors of 75 elements limited to amino acid exchanges possible given single nucleotide change (1-nt interchanges); and 3) vectors of 101 elements for the amino acid exchanges possible given two nucleotide changes (2-nt interchanges). We used JMPPro version 12.2 (SAS Institute Inc.) with default settings for the PCA.

We compared the 1-nt exchangeabilities for our clade-specific models to matrices that describe amino acid properties. First, we used a symmetric version of the Yampolsky and Stoltzfus [47] EX matrix, which describes the impact of mutations in laboratory mutagenesis studies (note: this EX matrix is unrelated to the Le *et al.* [10] EX2 model, as stated above that EX2 model is an XB-type model). Lower EX matrix values indicate that mutating wild-type amino acid i to amino acid j typically results in more severe phenotypic changes in the laboratory. The EX_s matrix (EX matrix-symmetric) was produced by averaging the EX matrix values for i to j mutations and j and i mutations and then normalizing the matrix to assign the most experimentally exchangeable amino acid pair (I and V) a value of one. Second, we compared clade-specific model exchangeabilities to matrices that capture differences in amino acid side-chain volume and polarity, which were obtained from Braun [4]. All of these comparisons used Spearman's rank correlations with two-tailed tests for significance.

3 Results and Discussion

3.1 Vertebrate and non-vertebrate models form two clusters in model space

Clustering of Euclidean distances among models along with midpoint rooting revealed two distinct clusters in the model space (Fig. 2a). The first cluster comprises the bird and mammal models and three standard empirical models trained using viral data. The second includes all of the other models estimated for this project along with all other standard empirical models. These results corroborate the hypothesis that patterns of sequence evolution vary across the tree of life and they further suggest that models trained using vertebrate data are especially distinctive.

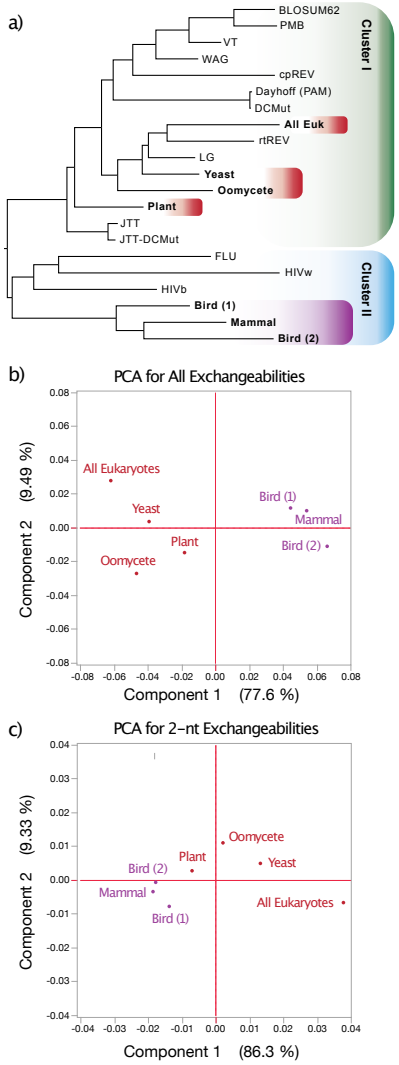

Figure 2: Cluster analysis and PCAs of clade-specific models of sequence evolution. (a) 'Tree of models' generated by neighbor-joining of Euclidean distances among exchangeability (***R***) matrices for the new clade-specific models (bold) and standard empirical models. Plots showing the first two PCs calculated using (b) all exchangeabilities and (c) 2-nt exchangeabilities. The proportion of the variance explained by each PC is listed alongside each axis.

The strong separation between the vertebrate models and the other clade-specific models was also evident in a PCA of the 190 exchangeability parameters of these models (Fig. 3b). PC1 and PC2 were both significant, but PC1 explained most of the variation and it separated the models into vertebrate and non-vertebrate models. Perhaps surprisingly, the three vertebrate models (two of which were estimated using bird data) appeared to be about as distinct from each other as the non-vertebrate models. The PCA for the 1-nt exchangeability values was quite similar, probably reflecting the

fact that the largest exchangeability values are those possible with a single substitution (the plot and the exchangeability values are available from github). In contrast, PCA of 2-nt exchangeabilities (Fig. 3c) revealed a different pattern; in that analysis PC1 also explained most of the variance but the plant model fell between the vertebrate models and the models for microbial eukaryotes (i.e., the yeast and oomycete). The vertebrate models were closer to each other than they were in the 1-nt PCA and the 'all Euk' model was located even further from vertebrates than the yeast and oomycete models. The latter finding probably reflects the fact that microbial eukaryotes dominate that dataset.

3.2 The best-fitting model for most individual proteins is the appropriate clade-specific model

The best-fitting model for individual proteins in each validation dataset was one of novel clade-specific models (Table 3); the only exception was the 'all Euk' validation dataset where the LG model [17] had the best fit more than 60% of the time (compared to 31.5% for the new 'all Euk' model). The best-fitting models for the vertebrate validation datasets were split among the three vertebrate models (Table 3). The results for individual proteins in the training data were virtually identical (available from github); the exception was the 'all Euk' training data where the new 'all Euk' model had the best fit for 93.5% of proteins. These results indicate that the average patterns of protein evolution for each clade provide substantial information regarding the patterns of substitution within those clades and further suggests that idiosyncratic differences among proteins play a limited role in model fit (i.e., our results are consistent with the top portion of the model space shown Fig. 1).

Table 3. Percentage of times clade-specific models (bold) he best-fitting model (rows) for individual proteins in each validation dataset (columns). '—' indicates the models was not recovered in analyses of the specified validation dataset.

Best model	Birds	Mammals	Plants	Oomycetes	Yeast	All Euk
Bird (1)	**25**	2	1	—	—	—
Bird (2)	24	21	—	—	—	—
Mammal	27.5	**65**	—	—	—	—
Plant	6.5	4	**70.5**	4.7	2.5	—
Oomycete	0.5	—	2	**84**	1.5	2
Yeast	1	—	3	2	**88**	4.7
All Euk	—	—	0.5	2	2	**31.5**
LG	1.5	0.5	5.5	5.3	4.5	60.4
JTT	8	6	16.5	—	—	—
WAG	—	—	0.5	1.3	—	1.3
DCMut	—	—	0.5	—	—	—
BLOSUM	0.5	0.5	—	—	—	—
PMB	1	—	—	—	—	—
VT	1.5	—	—	—	—	—
HIVb	2	—	—	—	—	—
FLU	0.5	0.5	—	—	—	—
mt models	0.5	0.5	—	0.7	1.5	—

3.3 Variation among structural environments is stronger than variation among clades

Clustering the matrices from the exposed/buried model XB models with standard empirical models and the two matrices from the EX2 model [10] revealed three relatively distinct clusters (Fig. 3a). All exposed models formed a divergent cluster on one side of the midpoint root; the deepest split within the exposed models was between vertebrates and non-vertebrates (the EX2 exposed component nested within non-vertebrates). The results for the buried components were similar; the bird and mammal buried components formed a cluster that was distinct from the second group that included the buried components of all non-vertebrate models and the EX2 model. Both of those buried clusters were nested within groups of standard empirical models (none of the latter were structure aware). The vertebrate buried components formed a cluster sister to three viral models (HIVb/HIVw [48] and FLU [49]). Thus, matrices for the structural models exhibited two levels of separation: 1) the separation between the exposed and buried clusters; and 2) the separation between the taxonomic groups (vertebrates vs. all other taxa).

PCA of all 190 exchangeability parameters (Fig. 3b) and of the 1-nt exchangeabilities (available from github) for the structural subsets of these datasets revealed similar patterns. PC1 explained ~50% of the variance and it separated the exposed and buried models. PC2 explained slightly more than 25% of the variance and it separated the models by clade in a manner consistent with the models based on all data. The exposed 'bird 1' model, based on alignments from Jarvis et al. [26], was especially distinctive. In contrast, PCA of 2-nt exchangeabilities was less informative; most models clustered near the center of a score plot of the first two PCs (which together explain 82.2% of the variance among models, see Fig. 3c). The exposed and buried sub-models of the 'all Euk' XB model were the most distinctive, with higher values of PC1 than any other sub-models from the same structural environment. The major similarity between the 2-nt PCA and the others is that the vertebrate exposed sub-models for vertebrates were more spread out than the buried sub-models for the same taxa, with the exposed 'bird 1' model being especially distinctive.

3.4 The best-fitting XB model for most individual proteins is the appropriate clade-specific model

Classification of validation set proteins using clade-specific XB mixture models similar to those obtained using the all-sites models (Table 4). In the majority of cases, the best-fitting XB modelsfor the plant, oomycete, and yeast validation data were the appropriate clade-specific models (>75% in all cases). Likewise, the best-fitting models for proteins in the vertebrate validation sets were almost always models trained using vertebrate data (>80% in both cases). As we observed with the all-sites models, the novel 'all Euk' XB model was not the best-fitting model for the 'all Euk' validation set; it was the EX2 model [10] instead. The results for analyses of individual proteins in the training set (available from github) were similar. These results indicate that the clade-specific single matrix models and the clade-specific can both act as classifiers when used with proteins from specific clades.

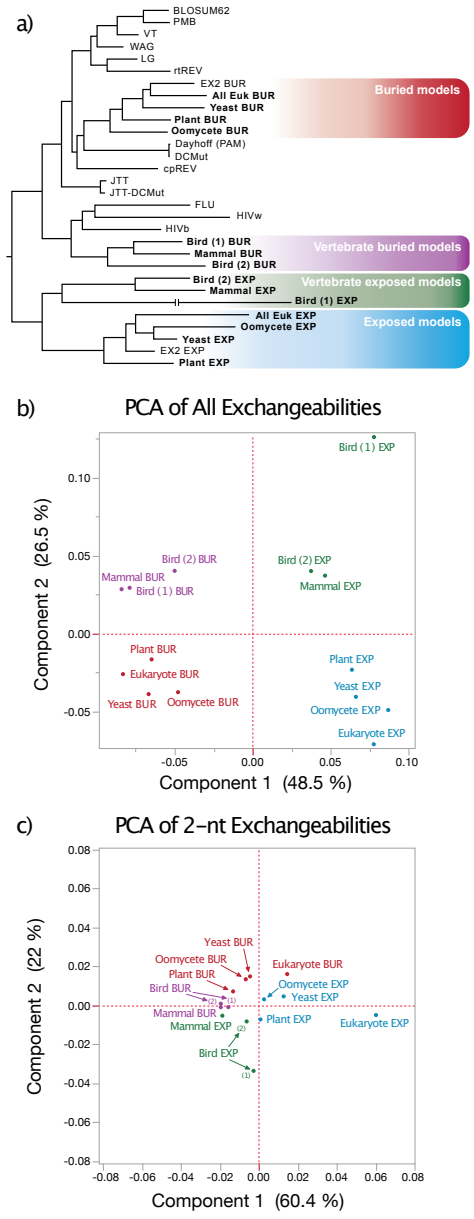

Figure 3: Cluster analysis and PCAs of clade-specific models of evolution that incorporate protein structure (see left). (a) 'Tree of models' generated by neighbor-joining of Euclidean distances among exchangeability matrices for new XB models (bold), the components of the EX2 model, and standard empirical models. The "bird (1) EXP" branch is presented with a break to indicate that it was a long branch; in this figure was shortened for readability. Plots showing calculated using (b) all exchangeabilities and (c) 2-nt exchangeabilities. The proportion of the variance explained by each PC is listed alongside each axis

Table 4. Percentage of times clade-specific XB models (rows) were the best-fitting model for individual proteins in each validation dataset (columns).

Best-fit model	Birds	Mammals	Plants	Oomycetes	Yeast	All Euk
Bird (1)	**26.5**	5	—	—	—	—
Bird (2)	24.5	18	—	—	—	—
Mammal	32	**68.5**	—	—	—	—
Plant	12	8	**94.5**	4.7	1.5	—
Oomycete	1.5	—	1	**76**	5	6.7
Yeast	2.5	—	1	4	**79**	2
All Euk	—	—	—	3.3	1	**33.5**
EX2	1	0.5	3	12	13.5	57.7

3.5 Exchangeabilities for different structural environments follow a 'rule of opposites'

The highest exchangeabilities for exposed sites involved pairs of hydrophobic residues; when exchangeabilities for all six models were averaged the three highest values corresponded to I-V, F-Y, and I-M. In contrast, the highest exchangeabilities for the buried environment involved polar pairs (in this case, the three highest values were R-K, D-E, and Q-H). This pattern may seem surprising; after all, it has long been appreciated that polar residues are common in solvent exposed environments whereas hydrophobic residues dominate the buried sites [50]. We call the observation that the most exchangeable amino acids in each structural environment are the less common amino acids in that environment the 'rule of opposites.'

The 'rule of opposites' allows us to differentiate between two alternative hypotheses to explain the relationship between exchangeabilities and amino acid frequencies. One might postulate that the amino acids that are rare in a specific structural environment would have very low exchangeabilities because those amino acid amino acids would be necessary for specific functions. Alternatively, one might postulate that exchanges between pairs of rare amino acids are common as long as the physicochemical nature of the amino acid is conserved. These results corroborate the second hypothesis and further suggest that at least some rare amino acids are especially exchangeable.

The exposed and buried sub-models of the XB models could be separated into a vertebrate and a non-vertebrate cluster (along PC2 in Fig. 5, panels a and b). Specific elements that separate the models were evident among the largest exchangeability values. Although the largest element for the exposed sub-models was I-V in both the vertebrate and plant/microbial groups, the next three elements differed. For vertebrates the next two elements involved exchanges between cysteine and aromatic residues (C-W and C-Y) whereas the plant/microbial models involved much more physicochemically-similar pairs (F-Y and L-M). Despite these differences, both groups conform to the rule of opposites (cysteine and aromatic residues are uncommon in solvent exposed environments; data available from github). In contrast, the top two exchangeabilities for the buried model were identical for the vertebrate and plant/microbial buried sub-models, although there were certainly a number of additional differences.

3.6 Differences in the strength of selection likely explain differences among clade-specific models

It should be possible to gain insights into the basis for the differences among-clade specific models by comparing changes in

amino acid properties to differences in 1-nt exchangeabilities. Comparing all-sites model exchangeability differences for vertebrates vs. plants and microbes revealed a correlation with experimental amino acid exchangeabilities (i.e., values in the EX_s matrix) and differences in side-chain volume. The EX_s correlation was negative (Spearman's correlation; r_S = -0.29309, P = 0.01071) whereas the correlation with changes in amino acid side chain volume (hereafter, Δ volume) was positive and even stronger (r_S = 0.39197, P = 0.00051). The directions of both correlations were consistent with the hypothesis that the major difference between vertebrate and non-vertebrate models is relaxed selection against slightly deleterious mutations in vertebrates (presumably due to their lower long-term N_e). In contrast, exchangeability differences were not correlated with Δ polarity (r_S = -0.0405, P = 0.73009). These results suggest changes in side-chain volume is the primary property subject to differential selection between vertebrates and plants/microbial eukaryotes.

A similar pattern was evident for the exposed sub-model of the clade specific XB models. Specifically, exchangeability differences for surface residues were correlated with EX_s (r_S = -0.34316, P = 0.00258) and Δ volume (r_S = 0.45707, P = $4x10^{-5}$); there was no correlation with Δ polarity (r_S = -0.09805, P = 0.40265). Weaker (and non-significant) correlations were evident for the buried sites (r_S = -0.20979, P = 0.07085 for EX_s; r_S = 0.30234, P = 0.00838 for Δ volume; r_S = -0.03851, P = 0.7429 for Δ polarity). However, buried sub-models for vertebrate and plants/microbes did exhibit some specific differences: 1-nt interchanges with higher buried-site exchangeabilities in vertebrates included A-T, R-H, M-V, R-Q, and C-Y whereas exchangeabilities with comparable elevation in the plant/microbial XB buried sub-models were F-Y, S-T, A-S, and N-H. Although there was no unifying physicochemical property for either set of pairs, we note that two of the pairs with elevated relative exchangeabilities in vertebrates (R-Q and C-Y) fall into different Dayhoff groups (i.e., the six groups shown in Fig. 84 of Dayhoff *et al.* [14]) and the pairs that are in the same Dayhoff group are relatively distinctive (e.g., R and H are both basic but they differ in shape, size, and even their charge at physiological pH). In contrast, two of the exchangeabilities elevated in the plant/microbial buried sub-models (F-Y and S-T) are physicochemically similar and only one (N-H) would change the Dayhoff group. Thus, these results are also consistent with the hypothesis that the lower long-term N_e of vertebrates has reduced the effectiveness of selection against slightly deleterious substitutions.

The hypothesis that among-clade differences in the patterns of protein sequence evolution reflects the strength of purifying selection raises several issues. First, the observation that our new vertebrate models clustered with models trained using viral data (HIVb/HIVw [48] and FLU [49]) may seem puzzling if N_e is a major factor in establishing model differences; after all, viruses are microbes so one might assume their long-term N_e would be very large. Second, our failure to find a correlation between differences in exchangeabilities and Δ polarity may seem surprising given the important role that polarity appears to play in models of protein evolution (cf. Braun [4]). However, neither of these issues actually provide evidence against our hypothesis. Drift actually appears to play an important role in viral evolution [51]. The combined effects of strong background selection and population bottlenecks expected during progression within hosts and transmission among hosts [52] are expected to reduce N_e for viruses. The absence of a correlation between Δ polarity and differences in exchangeabilities is also expected given if there is very strong selection against changes in polarity. Exchangeabilities are only expected to exhibit a correlation with 'Δ property' when selection against changes in the property are weak enough for drift to dominate in low N_e clades and selection to dominate in high N_e clades. Stronger purifying selection would result in selection dominating regardless of N_e, eliminating the correlation. Thus, neither of those observations are problems for the hypothesis that differences in the strength of purifying selection due to differences in N_e lead to differences among the clade-specific models.

3.7 Broadly sampled training data distorts model parameter estimates

Most empirical models have used as much training datasets as possible to reduce the variance of model parameter estimates. However, some studies have reported that estimates of parameters describing the amino acid substitution process exhibit time dependence [53]–[55]. The results of Benner *et al.* [53], who estimated log-odds matrices using many pairs of aligned sequences selected to fall within certain ranges of divergence, are especially interesting. They highlighted eight specific amino acid pairs; the log-odds scores for the first set (which we will call 'type A pairs' hereafter) have higher values when they are estimated using divergent sequence pairs whereas the second set (hereafter, 'type B pairs') have lower log-odds scores they were estimated using divergent sequence pairs. Type A pairs (F-W, W-Y, C-M, and C-V) are similar amino acids (mean EX_s = 0.5258) that are encoded by codons that differ by at least two nucleotides. Type B pairs (C-W, R-C, C-Y, and R-W) are dissimilar amino acids (mean EX_s = 0.3547) encoded by codons that differ by a single nucleotide. These observations led Benner *et al.* [53] to conclude that "the genetic code influences accepted point mutations strongly at early stages of divergence, while the chemical properties of the side chains dominate at more advanced stages" (where 'advanced stages' refers to long evolutionary timescales).

We included the 'all Euk' training dataset to assess the impact of estimating model parameters using highly diverged sequences. Our exchangeability parameter estimates exhibited a pattern similar to the pattern observed by Benner *et al.* [53] for log-odds scores; the mean exchangeability for type A pairs in the clade-specific models ranged from 15.4% of the 'all Euk' value (for W-Y) to 17.6% (for C-V). We observed similar patterns for both XB sub-models, with the mean exchangeabilities for type A pairs ranging from 13.2% of the 'all Euk' value (for buried site W-Y interchanges) to 22.1% (for exposed site W-Y interchanges). As expected, we observed the opposite pattern for type B pairs. When we normalized the 'all Euk' type B exchangeabilities to the maximum for that pair in any clade specific model we found values that ranged from an absolute minimum of 3.4% (for R-W interchanges in the exposed environment) to 24.8% (for R-C interchanges in the exposed environment).

Kosiol and Goldman [56] pointed out that apparent time-dependence must represent a problem associated with parameter

estimation; after all, long-term substitution patterns ultimately reflect the accumulation of substitutions over many short periods of time. Thus, Benner *et al.* [53] log-odds score estimates should not exhibit time dependence if the accumulation of amino acid substitutions can be modeled as a time-homogeneous Markov process. Kosiol and Goldman [56] resolved this paradox by showing that a time-homogeneous Markov model for nucleotides can appear non-Markovian when the data are aggregated into the encoded amino acids. In fact, they even demonstrated apparent time-dependence of log-odds scores for the type A and B pairs qualitatively similar to the Benner *et al.* [53] patterns (although there were only two pairs, C-M and C-V, that were similar in quantitative terms). Exchangeabilities for the LG model, which was trained using a taxonomically diverse dataset [17], also exhibited the pattern of high values for type A pairs and (to a lesser degree) low values for type B pairs. Three of the type A pairs (F-W, W-Y, and C-M) in the LG model had values ~60% of the 'all Euk' comparable values and they were higher than the comparable values for any clade-specific model (see models on github). This suggests the LG model may be subject to a 'time-dependency' effect similar to our 'all Euk' model, albeit one that is not as extreme.

The fact that type B pairs involve physicochemically-dissimilar amino acids that require a single nucleotide substitution for interchanges creates an additional complexity. They are exactly the type of substitutions expected to accumulate at an elevated rate in taxa with a lower long-term N_e, so the observation that vertebrate models always had the highest type B exchangeabilities (see models files available from github) is not surprising. The surprise is the high values for the type A substitutions, which involve interchanges of similar amino acids that require multiple nucleotide substitutions. Since some type B exchanges represent intermediates for type A substitutions (e.g., the only two-step pathway for F-Y involves C as an intermediate) one might expect selection against these disfavored intermediates to reduce type A exchangeabilities. If type A exchangeabilities were only elevated in the 'all Euk' model one might dismiss them as purely artifactual. However, type A exchangeabilities are also elevated in the clade-specific microbial (oomycete and yeast) models. This suggests they warrant further examination.

Of course, type A exchangeabilities are even higher in the 'all Euk' model (e.g., W-Y was the highest 2-nt exchangeability, and it had the ninth highest value of the 190 exchangeabilities). The very high type exchangeabilities in the 'all Euk' rate matrix probably reflect artifacts (a combination of the aggregation effects described by Kosiol and Goldman [57] and violations of the time reversibility assumption at the timescale of all eukaryotes) superimposed on the features of more typical microbial models (after all, most taxa in the 'all Euk' training data are microbial). These issues could also explain the poor performance of 'all Euk' models in cross-validation (Tables 3 and 4) and the observation that the all sites 'all Euk' training data was the only case where fit of GTR_{20}+I+Γ model was not better (based on the AIC_c) than the fit best empirical model (see Methods). Regardless, it seems reasonable to speculate that biases in the 'all Euk' models could have an impact on other uses of those rate matrices, like phylogenetic estimation.

3.8 Does GTR_{20} provide a useful framework to examine variation among clades?

We examined differences in amino acid exchangeabilities among taxa by comparing GTR_{20} model parameters. Superficially, the GTR_{20} model might not appear to provide the most natural framework for testing the hypothesis that exchangeabilities vary among clades (Fig. 1). After all, GTR_{20} assumes time reversibility and changes in amino acid exchangeabilities among taxa violate that assumption. This suggests that a model that relaxes the time reversibility assumption might should be used. The general Markov model (GMM) represents just such a model. However, the amino acid GMM requires estimation of 380 free parameters per branch, unlike the nucleotide GMM which only requires 12 free parameters [58]. Moreover, the GMM cannot be used with among-sites rate variation (except for a +invariant sites version [59]) and rate variation is ubiquitous in protein evolution [12]. Thus, we assumed there is less variation within each clade than there is among clades, estimated exchangeability parameters using the GTR_{20}+I+Γ model, and compared those parameter estimates. The assumption that the variation among clades exceeds variation within clades seems justified given their performance in with the cross-validation data (Tables 3 and 4).

One might take the arguments against the GMM even further by asking whether the GTR_{20} model is itself actually too parameter rich. The GTR_{20} model dimension could be reduced by restricting subsets of exchangeabilities to be equal, an approach used in analyses of nucleotide data [60]. However, it is impractical for protein models because the number of possible GTR submodels is a Bell [61] number. There are 203 possible GTR_4-type models (202 submodels and GTR_4 itself) but there are $>10^{250}$ GTR_{20} submodels, rendering thisapproach impractical for protein data (unless one makes *a priori* assumptions regarding the appropriate restrictions). Alternatively, one might use parameters based on specific amino acid properties, an approach suggested by Braun [4]. Although the Braun models can be used to examine differences among clades, they require *a priori* assumptions regarding the most important amino acid properties. A major goal in this study was to avoid *a priori* assumptions regarding the amino acid properties that contribute to the differences among models.

Zou and Zhang [3] proposed a codon model that provides an alternative framework; specifically, they extended the Yang et al. [62] codon model by replacing the single ω parameter (the ratio of non-synonymous to synonymous substitutions) with a vector of 75 ω parameters (one for each 1-nt exchange). This model has fewer free parameters than the GTR_{20}+I+Γ model; a Zou-Zhang-type model with a transition-transversion parameter and equilibrium codon frequencies calculated using the product of the nucleotide frequencies for first, second, and third codon positions has 85 free parameters (GTR_{20}+I+Γ has 210 parameters). However, the lower dimension of the Zou-Zhang model reflects the assumption that ω=0 is 2-nt and 3-nt interchanges. Making that assumption would have made it impossible to detect the relatively high type A exchangeabilities in some models. Allowing instantaneous doublet and triplet changes improves the fit of codon models in other contexts [57], so failing to include free ω parameters for 2-nt and 3-nt interchanges might be problematic. However, extending the Zou-Zhang model to include those interchanges increases the

number of free parameters by requiring the addition of 115 ω parameters and two parameters for the doublet and triplet substitution rates. Since the dimension of the modified Zou-Zhang model is similar to that for the GTR_{20}+I+Γ model it should be clear that codon models do not have a clear advantage with respect to the number of free parameters.

Any method to estimate amino acid exchangeabilities (including our approach and the Zou-Zhang approach) will depend on the quality of the multiple sequence alignments used to estimate parameters. We do not believe this is a source of error for several reasons. Most of our training and validation datasets (Tables 1 and 2) were used without modification and the source publications used different alignment pipelines. We did use our own pipeline for the bird (2) dataset (see Methods), but the larger point is that the alignment strategies differed among datasets. We would not expect our models to be good classifiers (as shown in Tables 3 and 4) or to have interpretable exchangeability parameters our results were strongly dependent on alignments. Additional evidence that our conclusions are robust to the alignment pipeline can be found in a set of clade-specific rate matrices described in a preprint [63] posted after this study was completed. Those Minh et al. [63] models are comparable to our all sites models (Fig. 2 and Table 3) and they clustered with the appropriate all sites models from this study (i.e., cluster analysis yielded a 'tree-of-models' similar to Fig. 2a and models based on data from the same clades formed clusters within the two major clusters; see information on github). Also, as we noted in section 3.7, the behavior of type A interchanges was evident in previous studies [53], [64] that used different alignment methods. Finally, we note that Zou and Zhang [3] also found that among-clade differences in amino acid exchangeabilities and they used yet another different pipeline. Although it is difficult to rule out the possibility of modest biases that reflect the details of the sequence alignment methods (or other aspects of the analytical pipeline), it is unlikely the large-scale patterns revealed by this study are trivial effects of alignment errors.

4 Conclusions

Efforts to estimate models of protein sequence evolution began in the very earliest days of computational biology; the first version of the PAM matrix was estimated over 50 years ago using a mere 814 substitutions from 11 protein families [1]. However, analyses of empirical models have provided little information about the processes governing protein evolution beyond the relatively straightforward conclusion that most amino acid exchanges involve physicochemically-similar amino acids. However, that was a conclusion that Dayhoff and Eck [65] reached (in very general terms) by examining the first version of the PAM matrix. On the other hand, efforts to develop models of protein evolution from first principles [19], [66], [67] remain impractical for phylogenetic analyses, especially in the phylogenomic era when hundreds or thousands of protein alignments are analyzed (e.g., the studies in Tables 1 and 2). The continued development of empirical models (e.g., the WAG [16] and LG [17] models) has provided models that can be used in that framework. It has not escaped our attention that our clade-specific models can also be used to improve phylogenomic analyses. The HIVb/HIVw [48] and FLU [49] models were generated to improve analyses of proteins from those viruses; our clade-specific models should improve phylogenetic estimation for specific taxa. Moreover, our clade-specific XB models should further improve model fit (and tree estimation) by accommodating variation among taxa and variation among-sites within proteins due to protein structure. All of our models are available from github and can be implemented in programs, such as IQ-TREE [33], that are used in many phylogenomic studies.

Although our models may be valuable for phylogenomic inference, the primary goal of this effort was to learn about the various ways that protein evolution has changed over time. Many efforts to understand the ways that evolutionary models change over time have assumed a single model for all sites within proteins. For practical reasons, they have also reduced the model dimension by using on a single parameter, like the radical to conservative substitution rate ration, with radical vs. conservative substitutions defined in a binary manner [2], [8]. Although this basic approach has been extended to a limited number of parameters by considering the physicochemical properties of amino acids [4], it is difficult to 'cast a wide net' in order to learn the ways that the process of amino acid has changed over time. Herein, we have estimated parameters that describe protein evolution in various clades using a simple framework (the GTR_{20} model, combined with among-sites rate heterogeneity) that does not presuppose an important role for any specific amino acid property. In doing so we found that there is substantial variation among clades in their model and that this variation among clades is evident both for amino acids located on the surface of proteins and for residues buried in the interior of proteins. We also found evidence that vertebrates are more tolerant of substitutions that change amino side-chain volume than plant/microbial models; however, this was only evident in only for models that describe the evolution of solvent exposed residues. We also showed that training empirical models using sequences sampled from taxa that were sampled too broadly (i.e., the 'all Euk' training data) can lead to distorted parameter estimates. Finally, we found that most proteins from a specific taxon were clustered in model space and that a relatively simple hypothesis – patterns of substitution reflect the strength of purifying selection, which differs among taxa due to differences among taxa in their long-term N_e – can explain many of the observed differences among taxa.

ACKNOWLEDGMENTS

We are grateful to Rebecca Kimball, Gordon Burleigh, Gavin Naylor, Emily Sessa, and Tamer Kahveci for helpful discussions and to the anonymous reviewers for constructive criticism. This work was supported in part by the U.S. National Science Foundation (grant DEB-1655683 to E.L.B. and Rebecca Kimball).

REFERENCES

[1] M. O. Dayhoff, R. V. Eck, and C. M. Park, "A model of evolutionary change in proteins," in *Atlas of Protein Sequence and Structure*, vol. 4, M. O. Dayhoff, Ed. Silver Springs, MD: National Biomedical Research Foundation, 1969, pp. 75–84.

[2] C. C. Weber and S. Whelan, "Physicochemical amino acid properties better describe substitution rates in large populations.," *Mol. Biol. Evol.*, vol. 36, no. 4, pp. 679–690, Apr. 2019.

[3] Z. Zou and J. Zhang, "Amino acid exchangeabilities vary across the tree of life.," *Sci. Adv.*, vol. 5, no. 12, p. eaax3124, Dec. 2019.

[4] E. L. Braun, "An evolutionary model motivated by physicochemical properties of amino acids reveals variation among proteins.," *Bioinformatics*, vol. 34, no. 13, pp. i350–i356, Jul. 2018.

[5] M. Kimura, "DNA and the neutral theory.," *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, vol. 312, no. 1154, pp. 343–354, Jan. 1986.

[6] W. Sung, M. S. Ackerman, S. F. Miller, T. G. Doak, and M. Lynch, "Drift-barrier hypothesis and mutation-rate evolution.," *Proc. Natl. Acad. Sci. USA*, vol. 109, no. 45, pp. 18488–18492, Nov. 2012.

[7] M. G. Behringer and D. W. Hall, "The repeatability of genome-wide mutation rate and spectrum estimates.," *Curr. Genet.*, vol. 62, no. 3, pp. 507–512, Aug. 2016.

[8] B. Nabholz, N. Uwimana, and N. Lartillot, "Reconstructing the phylogenetic history of long-term effective population size and life-history traits using patterns of amino acid replacement in mitochondrial genomes of mammals and birds.," *Genome Biol. Evol.*, vol. 5, no. 7, pp. 1273–1290, 2013.

[9] E. Zuckerkandl and L. Pauling, "Evolutionary divergence and convergence in proteins," in *Evolving genes and proteins*, Elsevier, 1965, pp. 97–166.

[10] S. Q. Le, N. Lartillot, and O. Gascuel, "Phylogenetic mixture models for proteins.," *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, vol. 363, no. 1512, pp. 3965–3976, Dec. 2008.

[11] S. Huzurbazar, G. Kolesov, S. E. Massey, K. C. Harris, A. Churbanov, and D. A. Liberles, "Lineage-specific differences in the amino acid substitution process.," *J. Mol. Biol.*, vol. 396, no. 5, pp. 1410–1421, Mar. 2010.

[12] J. Echave, S. J. Spielman, and C. O. Wilke, "Causes of evolutionary rate variation among protein sites.," *Nat. Rev. Genet.*, vol. 17, no. 2, pp. 109–121, Feb. 2016.

[13] A. G. Meyer and C. O. Wilke, "Integrating sequence variation and protein structure to identify sites under selection.," *Mol. Biol. Evol.*, vol. 30, no. 1, pp. 36–44, Jan. 2013.

[14] M. O. Dayhoff, R. M. Schwartz, and B. C. Orcutt, "A model of evolutionary change in proteins," in *Atlas of Protein Sequence and Structure*, vol. 5, M. O. Dayhoff, Ed. Silver Springs, MD: National Biomedical Research Foundation, 1978, pp. 345–352.

[15] D. T. Jones, W. R. Taylor, and J. M. Thornton, "The rapid generation of mutation data matrices from protein sequences," *Bioinformatics*, vol. 8, no. 3, pp. 275–282, 1992.

[16] S. Whelan and N. Goldman, "A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach.," *Mol. Biol. Evol.*, vol. 18, no. 5, pp. 691–699, May 2001.

[17] S. Q. Le and O. Gascuel, "An improved general amino acid replacement matrix.," *Mol. Biol. Evol.*, vol. 25, no. 7, pp. 1307–1320, Jul. 2008.

[18] C. O. Wilke, "Bringing molecules back into molecular evolution.," *PLoS Comput. Biol.*, vol. 8, no. 6, p. e1002572, Jun. 2012.

[19] G. Parisi and J. Echave, "Structural constraints and emergence of sequence patterns in protein evolution.," *Mol. Biol. Evol.*, vol. 18, no. 5, pp. 750–756, May 2001.

[20] G. C. Conant and P. F. Stadler, "Solvent exposure imparts similar selective pressures across a range of yeast proteins.," *Mol. Biol. Evol.*, vol. 26, no. 5, pp. 1155–1161, May 2009.

[21] A. Pandey and E. L. Braun, "Phylogenetic analyses of sites in different protein structural environments result in distinct placements of the metazoan root.," *Biology (Basel)*, vol. 9, no. 4, Mar. 2020.

[22] S. M. Soucy, J. Huang, and J. P. Gogarten, "Horizontal gene transfer: building the web of life.," *Nat. Rev. Genet.*, vol. 16, no. 8, pp. 472–482, Aug. 2015.

[23] R. O. Prum, J. S. Berv, A. Dornburg, D. J. Field, J. P. Townsend, E. M. Lemmon, and A. R. Lemmon, "A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing.," *Nature*, vol. 526, no. 7574, pp. 569–573, Oct. 2015.

[24] S. Reddy, R. T. Kimball, A. Pandey, P. A. Hosner, M. J. Braun, S. J. Hackett, K.-L. Han, J. Harshman, C. J. Huddleston, S. Kingston, B. D. Marks, K. J. Miglia, W. S. Moore, F. H. Sheldon, C. C. Witt, T. Yuri, and E. L. Braun, "Why do phylogenomic data sets yield conflicting trees? Data type influences the avian tree of life more than taxon sampling.," *Syst. Biol.*, vol. 66, no. 5, pp. 857–879, Sep. 2017.

[25] K. Katoh, G. Asimenos, and H. Toh, "Multiple alignment of DNA sequences with MAFFT.," *Methods Mol. Biol.*, vol. 537, pp. 39–64, 2009.

[26] E. D. Jarvis, S. Mirarab, A. J. Aberer, B. Li, P. Houde, C. Li, S. Y. W. Ho, B. C. Faircloth, B. Nabholz, J. T. Howard, A. Suh, C. C. Weber, R. R. da Fonseca, J. Li, F. Zhang, H. Li, L. Zhou, N. Narula, L. Liu, G. Ganapathy, B. Boussau, Md. S. Bayzid, V. Zavidovych, S. Subramanian, T. Gabaldón, S. Capella-Gutiérrez, J. Huerta-Cepas, B. Rekepalli, K. Munch, M. Schierup, B. Lindow, W. C;.Warren, D. Ray, R. E Green, M. Bruford, X. Zhan, A. Dixon, S. Li, N. Li, Y. Huang, E. P. Derryberry, M. F. Bertelsen, F. Sheldon, R. T. Brumfield, C. Mello, P. V. Lovell, M. Wirthlin, J. A. Samaniego, A. M. V. Velazquez, A. Alfaro-Núñez, P. F. Campos, T. Sicheritz-Ponten, A. Pas, T .Bailey, P. Scofield, M. Bunce, D. Lambert, Q. Zhou, P. Perelman, A. C. Driskell, G. Ruby, B. Shapiro, Z. Xiong, Y. Zeng, S. Liu, Z. Li, B. Liu, K. Wu, J. Xiao, X. Yinqi, Q. Zheng, Y. Zhang, H. Yang, J. Wang, L. Smeds, F. E. Rheindt, M. Braun, J. Fjeldså, L. Orlando, K. Barker, K. A. Jønsson, W. Johnson, K.-P. Koepfli, S. O'Brien, D. Haussler, O. A. Ryder, C. Rahbek, E. Willerslev, G. R. Graves, T. C. Glenn, J. McCormack, D. Burt, H. Ellegren, P. Alström, S. V. Edwards, A. Stamatakis, D. P. Mindell, J. Cracraft, E. L. Braun, T. Warnow, Wang J., M. T. P. Gilbert, and G. Zhang, "Whole-genome analyses resolve early branches in the tree of life of modern birds.," *Science*, vol. 346, no. 6215, pp. 1320–1331, Dec. 2014.

[27] E. D. Jarvis, S. Mirarab, A. J. Aberer, B. Li, P. Houde, C. Li, S. Y. W. Ho, B. C. Faircloth, B. Nabholz, J. T. Howard, A. Suh, C. C. Weber, R. R. da Fonseca, A. Alfaro-Núñez, N. Narula, L. Liu, D. Burt, H. Ellegren, S. V. Edwards, A. Stamatakis, D. P, Mindell, J. Cracraft, E. L. Braun, T. Warnow, Wang J., M. T. P. Gilbert, G. Zhang, and Avian Phylogenomics Consortium, "Phylogenomic analyses data of the avian phylogenomics project.," *Gigascience*, vol. 4, p. 4, Feb. 2015.

[28] E. J. P. Douzery, C. Scornavacca, J. Romiguier, K. Belkhir, N. Galtier, F. Delsuc, and V. Ranwez, "OrthoMaM v8: a database of orthologous exons and coding sequences for comparative genomics in mammals.," *Mol. Biol. Evol.*, vol. 31, no. 7, pp. 1923–1928, Jul. 2014.

[29] Z. Xi, L. Liu, J. S. Rest, and C. C. Davis, "Coalescent versus concatenation methods and the placement of *Amborella* as sister to water lilies.," *Syst. Biol.*, vol. 63, no. 6, pp. 919–932, Nov. 2014.

[30] M. S. Ascunce, J. C. Huguet-Tapia, A. Ortiz-Urquiza, N. O. Keyhani, E. L. Braun, and E. M. Goss, "Phylogenomic analysis supports multiple instances of polyphyly in the oomycete peronosporalean lineage.," *Mol. Phylogenet. Evol.*, vol. 114, pp. 199–211, Jun. 2017.

[31] X.-X. Shen, D. A. Opulente, J. Kominek, X. Zhou, J. L. Steenwyk, K. V. Buh, M. A. B. Haase, J. H. Wisecaver, M. Wang, D. T. Doering, J. T. Boudouris, R. M. Schneider, Q. K. Langdon, M. Ohkuma, R. Endoh, M. Takashima, R.-I. Manabe, N. Čadež, D. Libkind, C. A. Rosa, and A. Rokas, "Tempo and mode of genome evolution in the budding yeast subphylum.," *Cell*, vol. 175, no. 6, pp. 1533-1545.e20, Nov. 2018.

[32] J. F. H. Strassert, M. Jamy, A. P. Mylnikov, D. V. Tikhonenkov, and F. Burki, "New phylogenomic analysis of the enigmatic phylum Telonemia further resolves the eukaryote tree of life.," *Mol. Biol. Evol.*, vol. 36, no. 4, pp. 757–765, Apr. 2019.

[33] L.-T. Nguyen, H. A. Schmidt, A. von Haeseler, and B. Q. Minh, "IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies.," *Mol. Biol. Evol.*, vol. 32, no. 1, pp. 268–274, Jan. 2015.

[34] M. A. Miller, W. Pfeiffer, and T. Schwartz, "Creating the CIPRES Science Gateway for inference of large phylogenetic trees," in *2010 Gateway Computing Environments Workshop (GCE)*, 2010, pp. 1–8.

[35] Z. Yang, "PAML 4: phylogenetic analysis by maximum likelihood.," *Mol. Biol. Evol.*, vol. 24, no. 8, pp. 1586–1591, Aug. 2007.

[36] C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden, "BLAST+: architecture and applications.," *BMC Bioinformatics*, vol. 10, p. 421, Dec. 2009.

[37] N. J. Wickett, S. Mirarab, N. Nguyen, T. Warnow, E. Carpenter, N. Matasci, S. Ayyampalayam, M. S. Barker, J. G. Burleigh, M. A. Gitzendanner, B. R. Ruhfel, E. Wafula, J. P. Der, S. W. Graham, S. Mathews, M. Melkonian, D. E. Soltis, P. S. Soltis, N. W. Miles, C. J. Rothfels, L. Pokorny, A. Jonathan Shaw, L. DeGironimo, D. W. Stevenson, B. Surek, J. C. Villarreal, B. Roure, H. Philippe, C. W. dePamphilis, T. Chen, M. K. Deyholos, R. S. Baucom, T. M. Kutchan, M. M. Augustin, J. Wang, Y. Zhang, Z. Tian, Z. Yan, X. Wu, X. Sun, G. K.-S. Wong, and J. Leebens-Mack, "Phylotranscriptomic analysis of the origin and early diversification of land plants.," *Proc. Natl. Acad. Sci. USA*, vol. 111, no. 45, pp. E4859-68, Nov. 2014.

[38] G. Lax, Y. Eglit, L. Eme, E. M. Bertrand, A. J. Roger, and A. G. B. Simpson, "Hemimastigophora is a novel supra-kingdom-level lineage of eukaryotes.," *Nature*, vol. 564, no. 7736, pp. 410–414, Nov. 2018.

[39] A. Krogh, B. Larsson, G. von Heijne, and E. L. L. Sonnhammer, "Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes.," *J. Mol. Biol.*, vol. 305, no. 3, pp. 567–580, Jan. 2001.

[40] G. Pollastri, D. Przybylski, B. Rost, and P. Baldi, "Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles.," *Proteins*, vol. 47, no. 2, pp. 228–235, May 2002.

[41] C. N. Magnan and P. Baldi, "SSpro/ACCpro 5: almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity.," *Bioinformatics*, vol. 30, no. 18, pp. 2592–2597, Sep. 2014.

[42] N. Goldman, J. L. Thorne, and D. T. Jones, "Assessing the impact of secondary structure and solvent accessibility on protein evolution.," *Genetics*, vol. 149, no. 1, pp. 445–458, May 1998.

[43] S. S. Choi, E. J. Vallender, and B. T. Lahn, "Systematically assessing the influence of 3-dimensional structural context on the molecular evolution of mammalian proteomes.," *Mol. Biol. Evol.*, vol. 23, no. 11, pp. 2131–2133, Nov. 2006.

[44] S. Q. Le and O. Gascuel, "Accounting for solvent accessibility and secondary structure in protein phylogenetics is clearly beneficial.," *Syst. Biol.*, vol. 59, no. 3, pp. 277–287, May 2010.

[45] U. Perron, A. M. Kozlov, A. Stamatakis, N. Goldman, and I. H. Moal, "Modeling structural constraints on protein evolution via side-chain conformational states.," *Mol. Biol. Evol.*, vol. 36, no. 9, pp. 2086–2103, Sep. 2019.

[46] N. Saitou and M. Nei, "The neighbor-joining method: a new method for reconstructing phylogenetic trees.," *Mol. Biol. Evol.*, vol. 4, no. 4, pp. 406–425, Jul. 1987.

[47] L. Y. Yampolsky and A. Stoltzfus, "The exchangeability of amino acids in proteins.," *Genetics*, vol. 170, no. 4, pp. 1459–1472, Aug. 2005.

[48] D. C. Nickle, L. Heath, M. A. Jensen, P. B. Gilbert, J. I. Mullins, and S. L. Kosakovsky Pond, "HIV-specific probabilistic models of protein evolution.," *PLoS ONE*, vol. 2, no. 6, p. e503, Jun. 2007.

[49] C. C. Dang, Q. S. Le, O. Gascuel, and V. S. Le, "FLU, an amino acid substitution model for influenza proteins.," BMC Evol. Biol., vol. 10, p. 99, Apr. 2010.

[50] C. L. Worth, S. Gong, and T. L. Blundell, "Structural and functional constraints in the evolution of protein families.," *Nat. Rev. Mol. Cell Biol.*, vol. 10, no. 10, pp. 709–720, Oct. 2009.

[51] K. K. Irwin, S. Laurent, S. Matuszewski, S. Vuilleumier, L. Ormond, H. Shim, C. Bank, and J. D. Jensen, "On the importance of skewed offspring distributions and background selection in virus population genetics.," *Heredity*, vol. 117, no. 6, pp. 393–399, Sep. 2016.

[52] S. Gutiérrez, Y. Michalakis, and S. Blanc, "Virus population bottlenecks during within-host progression and host-to-host transmission.," *Curr. Opin. Virol.*, vol. 2, no. 5, pp. 546–555, Oct. 2012.

[53] S. A. Benner, M. A. Cohen, and G. H. Gonnet, "Amino acid substitution during functionally constrained divergent evolution of protein sequences," *Protein Eng.* Des. Sel., vol. 7, no. 11, pp. 1323–1332, 1994.

[54] G. Mitchison and R. Durbin, "Tree-based maximal likelihood substitution matrices and hidden Markov models," *J. Mol. Evol.*, vol. 41, no. 6, Dec. 1995.

[55] T. Müller, R. Spang, and M. Vingron, "Estimating amino acid substitution models: a comparison of Dayhoff's estimator, the resolvent approach and a maximum likelihood method.," *Mol. Biol. Evol.*, vol. 19, no. 1, pp. 8–13, Jan. 2002.

[56] C. Kosiol and N. Goldman, "Markovian and non-Markovian protein sequence evolution: aggregated Markov process models.," *J. Mol. Biol.*, vol. 411, no. 4, pp. 910–923, Aug. 2011.

[57] C. Kosiol, I. Holmes, and N. Goldman, "An empirical codon model for protein sequence evolution.," *Mol. Biol. Evol.*, vol. 24, no. 7, pp. 1464–1479, Jul. 2007.

[58] D. Barry and J. A. Hartigan, "Statistical analysis of hominoid molecular evolution," *Stat. Sci.*, vol. 2, no. 2, pp. 191–207, May 1987.

[59] V. Jayaswal, J. Robinson, and L. Jermiin, "Estimation of phylogeny and invariant sites under the general Markov model of nucleotide sequence evolution.," *Syst. Biol.*, vol. 56, no. 2, pp. 155–162, Apr. 2007.

[60] J. P. Huelsenbeck, B. Larget, and M. E. Alfaro, "Bayesian phylogenetic model selection using reversible jump Markov chain Monte Carlo.," *Mol. Biol. Evol.*, vol. 21, no. 6, pp. 1123–1133, Jun. 2004.

[61] E. T. Bell, "Exponential Numbers," *The American Mathematical Monthly*, vol. 41, no. 7, pp. 411–419, Aug. 1934.

[62] Z. Yang, R. Nielsen, and M. Hasegawa, "Models of amino acid substitution and applications to mitochondrial protein evolution.," *Mol. Biol. Evol.*, vol. 15, no. 12, pp. 1600–1611, Dec. 1998.

[63] B. Q. Minh, C. C. Dang, L. S. Vinh, and R. Lanfear, "QMaker: Fast and accurate method to estimate empirical models of protein evolution," *BioRxiv*, https://doi.org/10.1101/2020.02.20.958819 Feb. 2020.

[64] G. H. Gonnet, M. A. Cohen, and S. A. Benner, "Exhaustive matching of the entire protein sequence database.," *Science*, vol. 256, no. 5062, pp. 1443–1445, Jun. 1992.

[65] M. O. Dayhoff and R. V. Eck, "The chemical meaning of amino acid mutations," in *Atlas of Protein Sequence and Structure*, vol. 4, M. O. Dayhoff, Ed. Silver Springs, MD: National Biomedical Research Foundation, 1969, pp. 85–87.

[66] U. Bastolla, M. Porto, H. E. Roman, and M. Vendruscolo, "A protein evolution model with independent sites that reproduces site-specific amino acid distributions from the Protein Data Bank.," *BMC Evol. Biol.*, vol. 6, p. 43, May 2006.

[67] M. Arenas, H. G. Dos Santos, D. Posada, and U. Bastolla, "Protein evolution along phylogenetic histories under structurally constrained substitution models.," *Bioinformatics*, vol. 29, no. 23, pp. 3020–3028, Dec. 2013.