Amino Acid Properties, Substitution Rates, and the Nearly Neutral Theory

Jennifer E. James [1,2,*], Martin Lascoux [2]

[1]Department of Cell and Molecular Biology, SciLifeLab, Uppsala University, Uppsala, Sweden
[2]Department of Ecology and Genetics, Uppsala University, Uppsala, Sweden

*Corresponding author: E-mail: Jennifer.james@icm.uu.se.

Accepted: February 12, 2025

Abstract

Do the properties of amino acids affect their rates of substitution? The neutral theory predicts that greater selective constraint leads to slower rates of evolution; similarly, we expect amino acids that are more different from each other to have lower rates of exchange because such changes are most likely to affect protein structure and function. Here, we test these predictions, using substitution rates estimated from empirical amino acid exchangeability matrices. To measure degree of amino acid difference, we focused on two physicochemical properties, charge and size, uncorrelated metrics that are known to have important implications for protein structure and function. We find that for both charge and size, amino acid pairs with large differences had lower rates of substitution. We also found that amino acids that differed in both properties had the lowest rates of substitution, suggesting that both physicochemical properties are under selective constraint. Mutation properties, such as the number of mutations or the number of transitions as opposed to transversions separating amino acid pairs, were also important predictors of substitution rates. The relationship between amino acid substitution rates and differences in their physiochemical properties holds across several taxonomically restricted datasets. This finding suggests that purifying selection affects amino acid substitution rates in a similar manner across taxonomic groups with different effective population sizes.

Key words: substitution rates, amino acids, selective constraint, mutations, physicochemical properties.

Significance

The amino acid composition of a protein determines its overall structure and function, because amino acids vary in physicochemical properties such as size, charge, and hydrophobicity. Over time, mutations result in exchanges between different amino acids. It has been suggested that substitution rates between more physicochemically different amino acids are lower, because such substitutions are more disruptive to protein structure and are therefore more likely to be harmful and to be removed by selection. We test this hypothesis, finding that amino acids that are more different do have lower rates of exchange over evolutionary time, and that this result is general to different taxonomic groups. This indicates the importance amino acid properties have to understanding patterns of molecular evolution.

Introduction

The neutral theory highlights the importance of purifying selection and conservation as forces in molecular evolution. Kimura (1983) stressed that more important proteins will evolve more slowly than less important proteins due to purifying selection removing variants that cause amino acid change. This logic can easily be extended to variants that do cause amino acid change: under Kimura's theory, we expect rates of exchange between amino acids that are dissimilar to be lower, as exchanges between more dissimilar amino acids are likely to cause large changes in protein structure and function. Investigating rates of exchange

between different amino acid pairs therefore provides an interesting test of the neutral theory.

It is now clear that rates of substitution between different pairs of amino acids vary. For example, Smith (2003) defined amino acid substitutions as either "radical" (physicochemically different) or "conservative" (physicochemically similar) and found that radical substitution rates were lower than conservative rates for a dataset of *Drosophila* genes, which is evidence that radical substitutions are subject to stronger negative selection than conservative mutations. More recently, Weber and Whelan (2019) used several multispecies alignments to investigate radical and conservative substitution rates. While they found evidence that radical substitutions were removed by purifying selection more frequently than conservative substitutions in large populations, they did not find evidence for a difference in rates in species with small populations (mammals, birds and vertebrates). Similarly, Zou and Zhang (2019) estimated amino acid exchangeabilities from codon models between 90 pairs of closely related species, finding substantial variation across taxa. These findings suggest that patterns of amino acid substitutions might be better explained by the nearly neutral theory, which posits that there is a fraction of new mutations of such weak effect that their fixation probabilities are determined by both selection and drift (Ohta 1973). The rate of fixation of such mutations is expected to depend on the effective population size of species (Ohta 1992; Gillespie and Ohta 1996). The implication of the above findings is that while amino acid properties are under selection, selection is only effective in populations with large sizes, and as such there is no detectible difference between radical or conservative substitution rates in taxa with small effective sizes. However, by contrast, Chen et al. (2019) used a dataset of 9 pairs of species and found that substitution rates were strongly related across taxa rather than varying among species with different population sizes, and that rate were negatively correlated to the difference in physicochemical properties between amino acid pairs. There is therefore some contention as to whether amino acid substitution rates vary among taxonomic groups, or if they vary systematically with differences in amino acid physicochemical properties. There are also some methodological issues with past studies. For example, in studies that group amino acid substitutions as either "radical" or "conservative," it is not clear how radical or conservative substitution should be defined. Some amino acid substitutions are more "radical" than others, as some involve exchanges between amino acids that differ in many properties.

There are also several methodological difficulties to estimating substitution rates. First, a general requirement of many methods is a phylogenetic tree, which may differ across the genome due to incongruities between gene trees and the species tree, resulting in biases in substitution rate estimates (Mendes and Hahn 2016; Carruthers et al. 2022). Second, accurate substitution rate estimation requires the careful alignment of orthologous genomic regions that must be either long enough or divergent enough to contain sufficient substitutions for inference without mutation saturation, i.e. the occurrence of multiple substitutions at the same site along the genome. Some judgment is therefore required as to whether a dataset is likely to provide estimates of substitution rates that are reasonably unbiased, and care must be taken to account for factors such as heterogeneous codon usage among sites, heterogeneous base composition over time and heterogeneous substitution rates over time, all of which can result in systematic error in substitution rate estimation (Del Amparo et al. 2021). The problem of mutation saturation is expected to increase when evolution is rapid, or if the timescale under consideration is long, leading to systematic underestimates of substitution rates. This has been demonstrated using datasets of viral genomes, which showed a decay of the transition:transversion ratio with time as expected if mutation saturation had occurred. There was a decay in the transition:transversion ratio with time even for datasets with short divergence times and small numbers of substitutions, in which there would be little a priori reason to expect mutation saturation (Duchêne et al. 2015). It is also known that substitution rates can be estimated poorly when not accounting for among-site substitution rate heterogeneity. Rate heterogeneity among sites has been modeled in the past using a gamma distribution, however, this approach has been shown to be biased (Ferretti et al. 2024), and therefore tools such as IQ-Tree incorporate a distribution-free rate model for among-sit heterogeneity, as implemented in ModelFinder (Kalyaanamoorthy et al. 2017).

An alternative approach is to consider exchangeability rates modeled at the amino acid level, rather than at the codon level. Amino acid substitution models are often estimated empirically from large datasets of multiple sequence alignments, with the resulting production of 20 by 20 matrices of exchangeability rates between amino acid pairs. Substitution matrices such as the Dayhoff, JTT, LG, and WAG matrices are widely used in phylogenetic analyses of protein sequences and are implemented in software tools including MEGA (Tamura et al. 2021), PAML (Yang 2007), and RAxML (Stamatakis 2014). Despite their wide use in phylogenetic methods, these models are often not considered in the context of the evolutionary insights they provide. These models can be computationally challenging to run and, due to the large number of parameters to be estimated, necessitate large datasets to prevent model over-fitting. However, recent advances in the field, including the development of user-friendly interfaces for estimating amino acid substitution models and the increasing availability of large sequence datasets, has made this approach increasingly tractable.

In this study, we used empirical models of amino acid substitution, incorporating recent advances in the field in

order to address Kimura's prediction that rates of exchange between more different amino acids will be lower. If rates of exchange between amino acid pairs are affected by the physicochemical differences between them, this suggests the action of purifying selection. We explicitly include mutational processes in our analyses, including GC content, the transition-transversion ratio, average amino acid frequencies, and number of mutational steps, to tease out the relative importance of mutational versus selective forces (Dagan et al. 2002). We also ensured that we accounted for amino acid composition when conducting our statistical analyses, an issue that has typically been overlooked in past studies. We investigated whether patterns in substitution rates differ across different timespans, or different taxonomic groups. We might expect difference in groups that vary in N_e according to the nearly neutral theory.

Results

In this study, we investigated the relationship between amino acid substitution rates and differences in their physiochemical properties. It is not appropriate to consider all amino acid physicochemical properties together in statistical analyses, because many such properties are related, resulting in multicollinearity; for example, polarity and hydrophobicity are strongly correlated (e.g. correlation between contact energy and the Kyte-Doolittle index of hydrophobicity, Pearson's $R = -0.82$, $P < 2E\text{-}16$), as are molecular weight and volume (Pearson's $R = 0.91$, $P < 2E\text{-}16$) (the full correlation matrix between amino acid properties considered in this study is shown in supplementary fig. S1, Supplementary Material online). We therefore considered a number of physicochemical properties, taken from the online database AAIndex (Kawashima et al. 2008), analyzing the differences between amino acid pairs in their properties using a principal component analysis (PCA) (Fig. 1). The principal components (PCs) in a PCA are linear combinations of the original variables that explain the most variance in the data. Our PCA analysis split the amino acid physicochemical properties into charge/hydrophobicity associated metrics, and size-associated metrics: properties such as hydrophobicity,

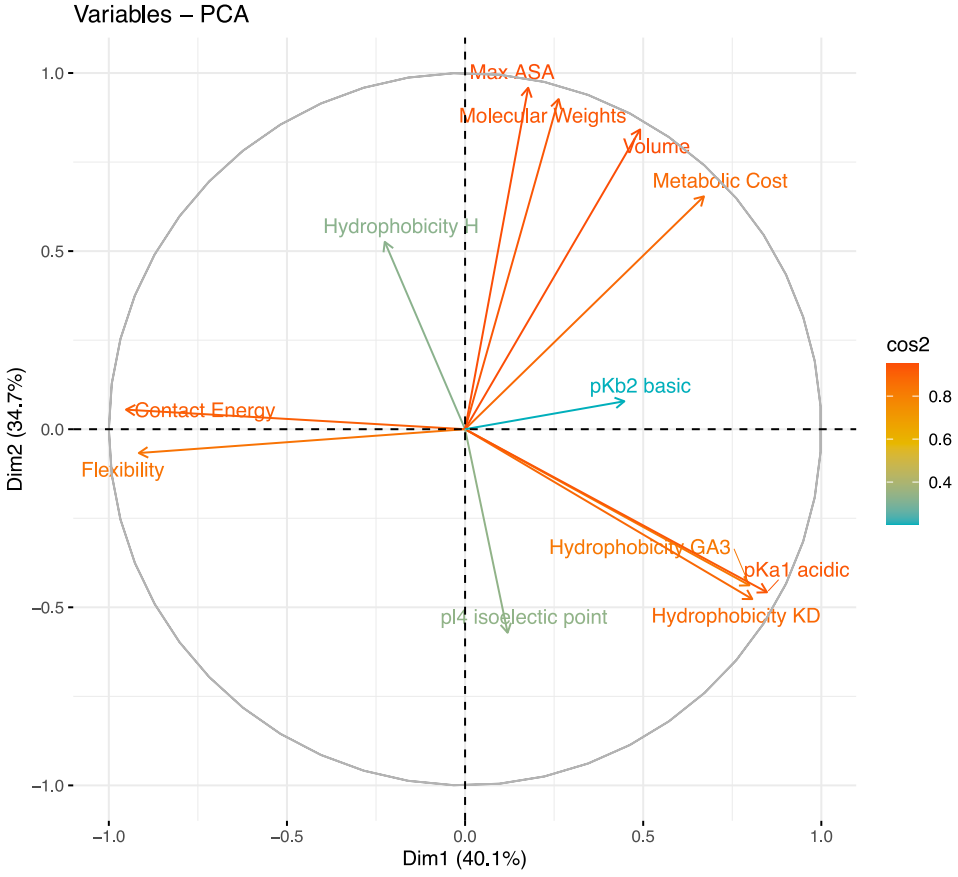

Fig. 1. PCA of the difference in physicochemical properties between amino acid pairs, showing the variables and their contributions to the first two principal components. Variables are labeled on their respective arrows, with colors of text and arrows indicating $\cos^2$, the representation of the variable on the principal component.

flexibility, and contact energy, all of which are highly correlated (supplementary fig. S1, Supplementary Material online), contributed to PC1 (40%), while size metrics such as maximum accessible solvent area, molecular weight, and volume were the major contributors to PC2 (35%). There was little additional variance explained from further PCA axes, with PC1 and 2 together explaining 74% of the variance in the data. In our further analysis, we therefore used these first two principal components as fixed effects in general linear models, with PC1 representing difference in charge and PC2 representing difference in size between amino acids.

We take as our estimates of exchangeability rates the nontime reversible *Q*-matrices from Minh et al. (2021), calculated over the Pfam dataset (El-gebali et al. 2019), and the time reversible matrices from Dang et al. (2022), which were calculated over the Pfam dataset in addition to several taxonomically restricted datasets. Such matrices reflect the structure of the genetic code, mutational biases, and selection pressure, and also use distribution-free rate models to account for among-site heterogeneity in substitution rates. We therefore conducted models while taking into account the number of mutations that separate pairs of amino acids, the approximate transition to transversion ratio per amino acid, and the average difference in codon GC content between amino acid pairs. For our main results, we used the best-fitting, time-reversible Q matrix calculated over the Pfam dataset (Minh et al. 2021; Dang et al. 2022), and as such, our main results should be general to a range of taxonomic groups and proteins.

We conducted linear mixed effect models of exchangeability rates, including amino acid composition as a random effect to account for the statistical nonindependence in the data. We found that pairs of amino acids with large differences in physiochemical properties have low rates of exchangeability, and low substitution rates (Table 1). This fits the predictions of the neutral theory: exchanges between amino acids with very different physiochemical properties are more likely to affect protein structure, and are therefore more likely to be removed by purifying selection, as opposed to exchanges between physicochemically similar amino acids. Both difference in amino acid size and difference in amino acid charge had a significant negative relationship with substitution rate ($P < 2e\text{-}16$ for both PC1 and PC2 in our model, Fig. 2), and also significantly interacted with each other, such that if the difference in one property was large, the relationship between the difference in the other property and exchangeability rate was more negative (an interaction term significantly improved model fit, $P = 4e\text{-}14$). This suggests that exchanges between amino acids that differ in both size and charge may be the most disruptive to protein function.

We found that average difference in codon GC composition does not affect the fit of our models (model comparison $P = 0.40$). Therefore, although local GC composition undoubtedly affects mutation rates and amino acid composition, it does not appear to influence amino acid exchangeability rates over long evolutionary timescales. In contrast, the number of mutations separating amino acid pairs significantly negatively impacts rates of exchange between them (shown in color scales of Fig. 1, mutational steps in general linear model $P < 2e\text{-}16$). Similarly, amino acid pairs separated by more transversion mutations have lower exchangeability rates (average difference in purine and pyrimidine content in general linear model $P = 0.019$).

It is not surprising that amino acid exchanges that require multiple mutations occur at lower rates, and that this factor explains a considerable proportion of variance in our model (the proportion of variance explained by marginal effects including all parameters but GC is 0.68, while if mutational parameters are instead included as random effects, the proportion of variance explained by marginal effects drops to 0.35). We further investigated how the number of mutations separating amino acid pairs changes the relationship between rates of substitution and mutational and physicochemical differences by independently considering amino acids separated by different numbers of mutational steps. Even considering amino acid pairs separated by the same number of mutations, we find negative relationships between amino acid substitution rates and differences in their physiochemical properties (see Fig. 3 and supplementary table S1, Supplementary Material online). For amino acids separated by one mutation, we find that physicochemical property differences are the primary driver of the relationship, with other mutational parameters not significantly improving model fit if included. This is likely because, although not statistically significant as an interaction term in our general linear models, there is a relationship between transition:transversion differences and mutational steps. Transition:tranversion differences are on average higher, and have higher

Table 1 Fixed effects of general linear model results

…	Estimate	SE	Degrees of freedom	*t*-value	*P*
Intercept	0.076	0.084	105.00	0.90	0.373
PC1, Charge	−0.436	0.021	342.94	−20.46	<2.0E-16***
PC2, Size	−0.289	0.026	357.33	−11.20	<2.0E-16***
Mutational steps slope	−0.422	0.026	350.50	−16.15	<2.0E-16***
Transition: transversion difference slope	−0.202	0.086	371.25	−2.36	0.019*
PC1:PC2 interaction	0.083	0.011	341.77	7.86	4.9E-14***

Shown is the best model, including an interaction term between PC1 and 2, and excluding difference in GC content. '*' indicates a statistical significance at the 0.05 level, '***' indicates statistical significance at the 0.005 level.

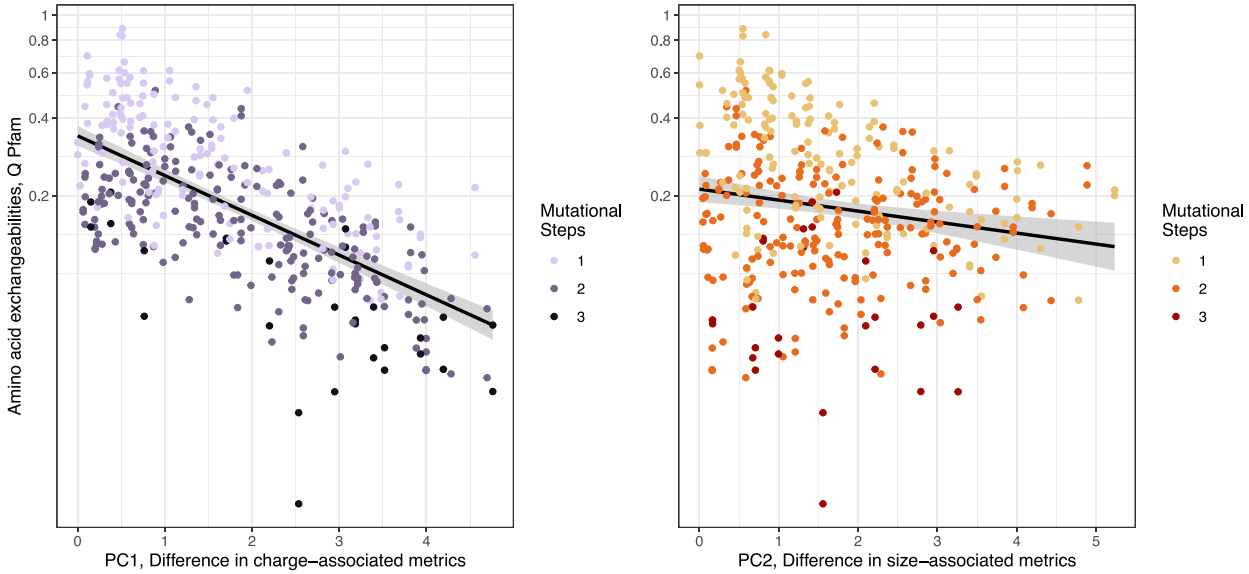

Fig. 2. Relationship between amino acid differences in physicochemical properties (PC1, charge associated metrics, left, and PC2, size-associated metrics, right), and substitution rates, for Pfam data. Points are colored in a gradient from dark to light representing most to least mutations separating amino acid pairs. Black lines are to aid the eye only, as the plots show nonindependent data. These lines are linear regression slopes for exchangeability rate ~ difference in physiochemical property.

variances, for amino acids that are separated by more mutational steps (Welch's ANOVA $P = 0.001$, supplementary fig. S2, Supplementary Material online). It is also possible that we lack power to detect the comparatively small effect of transition:transversion difference in subsets of the Pfam substitution rate dataset (AIC values with and without GC content and transition:transversion difference are 81.29 and 81.94 respectively, ANOVA $P = 0.098$). The fact that rates of substitution for amino acids separated by one mutation remain strongly predicted by the difference in their physicochemical properties is particularly striking because amino acids that are more similar tend to be separated by only a single mutational step (Osawa and Jukes 1989; Freeland and Hurst 1998), which we confirm is also true for the physicochemical properties considered in this study, although the difference is only significant when comparing our first principal component, associated with the difference in charge/hydrophobicity between pairs of amino acids (one/two mutations mean difference in PC1, charge: 1.55, 1.88, *t*-test $P = 0.011$, one/two mutations mean difference in PC2, size: 1.63, 1.72, *t*-test $P = 0.5$). For amino acid pairs separated by two mutations, the relationship between substitution rate and difference in charge and difference in size was negative and significant, with mutational parameters not improving model fit ($P = 0.46$), a very similar pattern to that observed by amino acid separated by one mutation. For amino acid pairs separated by 3 mutations, only the negative relationship between substitution rate and difference in charge related metrics was statistically supported, likely due to the smaller size of this dataset. The primary effect that number of mutation steps has in our models is to change the intercept, such that the greater the number of mutation steps separating amino acid pairs, the lower the intercept (Fig. 2).

These results were calculated using the time-reversible exchange matrix for the Pfam dataset. They are therefore representative of long evolutionary timescales and are general across a whole range of species. We also repeated these analyses using more taxonomically restricted datasets, to investigate whether these patterns hold. We found results to be very similar when considering yeast, plants, birds, insects, and mammal datasets, and when using exchangeability rates as estimated using time nonreversible matrices (model results shown in supplementary table S2, Supplementary Material online). The relationship between exchangeability rates and the difference in these 2 physicochemical properties holds across all the datasets we tested. It is interesting that differences in amino acid physicochemical properties remain predictive of their substitution rates for both highly diverged sequences (e.g. the Pfam dataset) and for clade-specific datasets. Average difference in GC content was not related to exchangeability rate in any datasets, while number of mutational steps explained a large proportion of the variance in exchangeability rates across all datasets. We see some evidence that mutational factors may have a larger impact on amino acid substitution rates in small N_e groups than in large N_e groups, with mutational factors explaining a greater proportion of the variance in substitution rates in mammals, birds and plants than in yeast and insects (supplementary table S2,

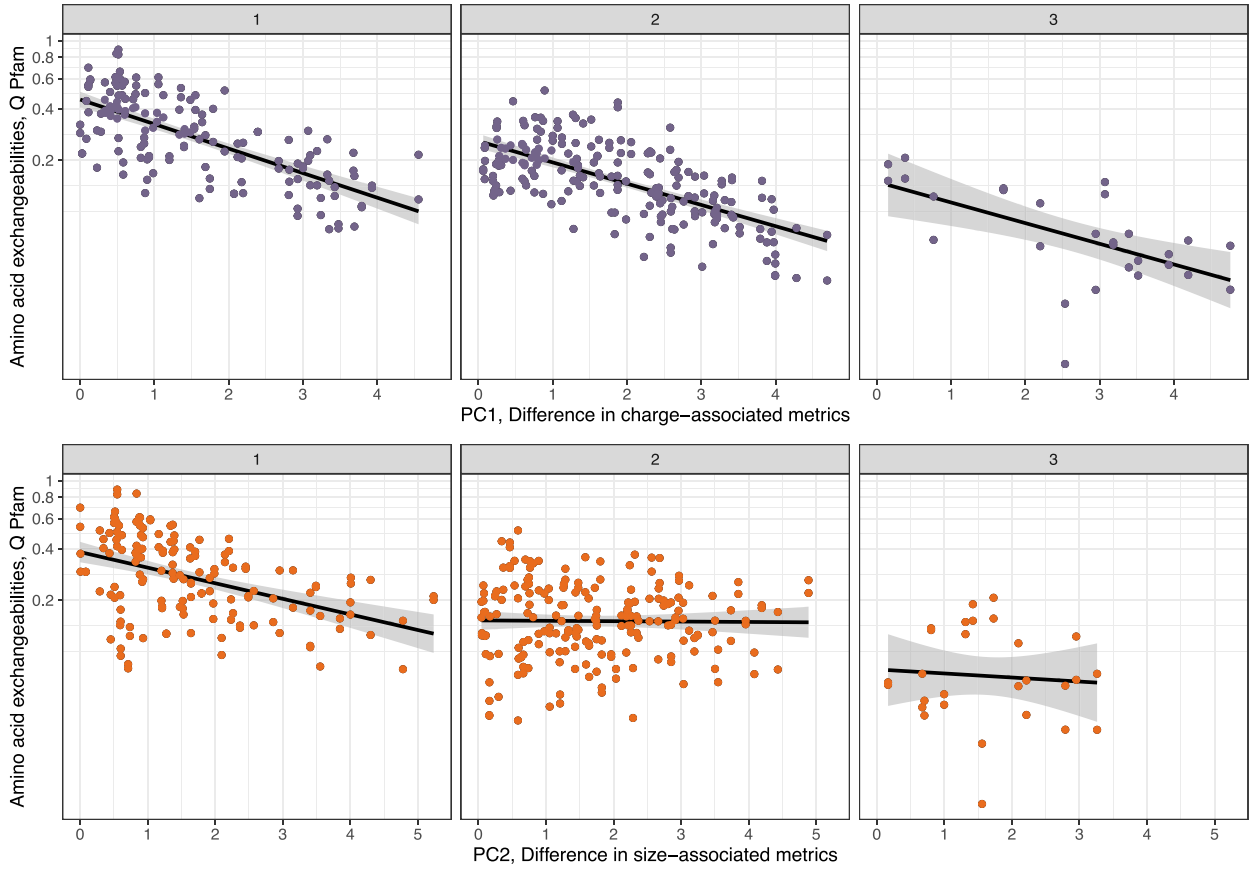

Fig. 3. Relationship between amino acid differences in physicochemical properties (PC1, charge associated metrics, top, and PC2, size-associated metrics, bottom), and substitution rates, for Pfam data. Data are separated into panels depending on the number of mutational steps (1, 2, or 3) between amino acid pairs. Black lines are to aid the eye only, as the plots show nonindependent data. These lines are linear regression slopes for exchangeability rate ~ difference in physiochemical property.

Supplementary Material online). This may be because the rates of substitutions, which require more than one mutational step are dependent on the rate of new mutations entering the population, which depends on census size.

Discussion

Overall, we find a relationship between the difference in physicochemical properties between pairs of amino acids and the exchangeability rates between them, such that more different amino acids have lower rates of exchange. Additionally, while differences in amino acid exchangeability rates are known to exist across taxonomic groups, as evidenced by studies that find substantially better model fits for phylogenetic analyses when using taxon-specific exchangeability rate matrices (Minh et al. 2021; Del Amparo and Arenas 2023), we find that a relationship between differences in physicochemical properties and substitution rates holds across all taxonomic groups considered. Our results suggest that, on average, there is evolutionary constraint acting at sites within proteins, such that changes with a small effect at the amino acid site level are more common. The exact shape of the relationship between amino acid substitution rates and their chemical differences has historically been under some dispute, with Kimura fitting an exponential curve to the data in his analysis (Kimura 1983) while Gillespie, using more modern estimates of substitution rates and focusing on amino acids separated by a single mutation, found the polynomial distribution to be a better fit to the data, suggesting that the most frequent substitutions were between amino acids with small, but not the smallest, chemical differences (1991). Our findings are more similar to those of Kimura, and thus are in line with the neutral theory explanation for this pattern: that nonconservative amino acid substitutions, which are more likely to disrupt protein structure and function, occur less frequently.

Our findings are a result of estimating amino acid substitution rates over many proteins and species. It is possible that there are biases or unconsidered factors in these datasets that could result in the observed relationship between amino acid physicochemical differences and substitution rates. For example, if particular pairs of amino acids are

more often found in young proteins with high rates of turnover and high substitution rates, or if particular pairs of amino acids are often associated with rapidly evolving active sites, this would generate a relationship between the amino acid physicochemical differences and substitution rates, without the differences in physicochemical properties being causal. However, this is unlikely to be responsible for our results because we observe a trend with substitution rate for both charge/hydrophobicity associated metrics and size-associated metrics, properties that are not correlated to each other (supplementary fig. S1, Supplementary Material online), i.e. pairs of amino acids can be different in terms of charge/hydrophobicity but not size, and vice versa. Therefore, particular pairs of very exchangeable amino acids are probably not driving both trends.

Our study primarily focusses on how purifying selection shapes patterns of molecular evolution. However, the focus of evolutionary biology on natural selection over other evolutionary processes, particularly mutation, has been questioned by several authors who claim that mutation has been overlooked as a force driving evolution. It has been argued that mutational processes can in themselves be adaptive, such that mutation bias could explain the origins of adaptation independently of, or in addition to, natural selection (Nei 2013; Laland et al. 2015). Stoltzfus and McCandlish (2017) cite the enrichment of parallel adaptive mutations for transitions in support of this view, arguing that, as transition mutations are more common, this provides evidence for a role of mutation bias in adaptation. However, adaptive mutations could be enriched in transitions for reasons other than mutation bias, such as selection for genomic GC content (Svensson and Berger 2019). It has also been argued that patterns of diversity across the *A. thaliana* genome suggest genes subject to strong purifying selection have lower mutation rates, which would be evidence for the role of mutation bias in patterns of evolution (Monroe et al. 2022). These claims have proved controversial, with a number of authors providing alternative interpretations of the data, including the action of purifying selection during development (Liu and Zhang 2022; Veitia 2022; Charlesworth and Jensen 2023; Majic and Payne 2023). Our study, while not a direct test of the idea, does not particularly support the mutation-driven evolution perspective. We found that the relationship between substitution rates and physicochemical property differences between amino acid pairs is similar regardless of the number of mutational steps separating them, although substitution rates are highest between amino acid pairs separated by smaller numbers of mutations. This suggests that there is natural selection acting on the physicochemical properties of amino acids, affecting rates of amino acid exchange over evolutionary time (Kimura, 1983). However, it is clear that natural selection interacts with mutational forces to influence patterns of substitution.

Patterns of amino acid substitutions are determined by a combination of mutational forces, selection and drift, all of which may differ across species groups. In support of this, clade-specific differences in amino acid substitution rates have been found by a number of researchers. For example, Pandey and Braun (2020) estimated clade-specific amino acid exchangeability rate matrices in order to investigate how much model variance could be explained by species clade. They found that, for the majority of proteins, the clade-specific model fit was able to correctly assign proteins to their clade. Models of amino acid exchangeability have also been shown to cluster in terms of similarity in a manner akin to the branches of the tree of life, meaning that in general more closely related species have more similar amino acid exchangeability rate matrices (Scolaro and Braun 2023). Likewise, we found that the two most closely related species sets in our datasets, mammals and birds, did have more similar amino acid exchangeability matrices, and more similar substitution rate matrices, than other species sets (supplementary tables S1 and S2, Supplementary Material online). Further development of time nonreversible and taxon-specific models of amino acid substitution may shed light on the biological underpinnings of the observed variation in substitution patterns across taxa.

Given a specific nearly neutral hypothesis for the observed relationship between amino acid substitution rates and differences in their physicochemical properties, we might also expect specific differences among clades in terms of this relationship. In species with large effective population sizes, we might expect substitution rates among amino acid pairs that differ in their physicochemical properties to be even lower than in species with small effective population sizes due to stronger purifying selection, resulting in a steeper relationship between substitution rates and physicochemical property differences. This pattern was observed in some previous studies (Weber and Whelan 2019; Zou and Zhang 2019).

On a related note, in a recent study, McShea et al. (2024) investigated differences in amino acid compositions and flux rates across mammalian species, comparing rodents and lagomorphs, which have large effective population sizes, to primates, which have small effective population sizes. The authors found that high N_e mammals show stronger selection for smaller, cheaper to synthesize amino acids.

Additionally, the authors showed that at this taxonomic scale there are differences in selective preferences even between amino acid pairs that are physicochemically similar. However, differences among clades in substitution rate trends are not particularly apparent in our results, possibly because of the broad taxonomic scale considered in this study (which contrasts birds, insects, mammals, yeast, and plants). The relationship between substitution rates and the physicochemical properties of amino acids was similar among datasets, despite some differences in

substitution rate matrices among taxonomic groups. It could be that, even for those groups in our dataset with small effective population sizes, selection against large changes in the physicochemical properties of amino acids is sufficiently strong on proteins overall that genetic drift does not have a large effect on substitution rates. It may be that N_e is sufficiently large across all studied taxa that the effect of further increasing N_e is small.

The further refining of amino acid substitution models is an important area of future research. In particular, while variation in substitution rates across sites is known to exist and is modeled when estimating amino acid exchangeabilities, it is possible that not only rates but also patterns of substitution vary across sites within proteins. For example, while solvent exposed sites are known to evolve more rapidly (Goldman et al. 1998), Pandey and Braun (2020) found that patterns of substitution also vary on a finer scale across such sites, after estimating amino acid substitution models for exposed and buried residues separately. The influence of protein structure on patterns of molecular evolution was also explored by Perron et al. (2019), who, using protein sequences for which structural data was available, estimated substitution models while also incorporating information on the rotational configuration of amino acid side chains. The resulting models had an expanded number of possible exchangeabilities, as each amino acid could be exchanged not only with every other amino acid, but also one of three side chain conformational states, expanding the 20 by 20 amino acid exchangeability matrix to a 55 by 55 exchangeability matrix. The models performed well when fitted to both real and simulated data and resulted in biological insight: the authors found evidence that side chain configuration was often conserved when an amino acid exchange occurred, and that highly exchangeable amino acids often had similar side chain geometries. More recently, Ferreiro et al. (2024) implemented an ABC approach to model substitutions while incorporating information on the likely selective constraints acting on proteins (including amino acid physicochemical properties), and allowing for site-dependent evolution.

Our results, and the ongoing developments of such methods, highlight the importance of incorporating knowledge of amino acid properties and protein structure into models of molecular evolution.

Materials and Methods

We analyzed amino acid exchange matrices calculated using QMaker (Minh et al. 2021) and nQMaker (Dang et al. 2022), implemented in IQ-TREE (Minh et al. 2020). Both QMaker and nQMaker estimate an amino acid exchange rate matrix using maximum likelihood, while also modeling variability in evolutionary rates across sites using a probability distribution-free method implemented in ModelFinder (Kalyaanamoorthy et al. 2017). The QMaker amino acid models of substitution rates that we used are estimated using a maximum likelihood procedure that makes three assumptions: that substitution rates are constant over time (homogeneous), that amino acid frequencies are at equilibrium (stationary), and that rates of exchange between 2 amino acids are equal, i.e. the rate of substitution from A to C is equal to that of C to A (reversible). nQMaker relaxes the time-reversibility assumption, producing time nonreversible matrices that do not obey detailed balance, that is, fluxes between amino acid pairs do not need to have the same magnitude.

We used the amino acid exchangeability matrices available online from IQ-TREE, as calculated over the Pfam dataset (version 31, El-gebali et al. 2019), and the matrices available for taxonomically restricted datasets (which included birds, insects, mammals, plants, and yeast). To convert available exchangeability matrices to substitution rate matrices, we multiplied each exchangeability rate by the frequency of the appropriate amino acid in the dataset, converting the exchangeability rate matrix to the substitution rate matrix, Q. We use time nonreversible models for taxonomically restricted datasets, and time-reversible Q matrix in our analysis of the Pfam dataset, as previous studies showed that this model provides a better fit to these datasets (Minh et al. 2021; Dang et al. 2022). It is important to note that while the taxonomically restricted Q-matrices were calculated using datasets of sequences present in their respective taxonomic group, most proteins are not specific to a particular clade, but instead are common to many species across the tree of life.

We took amino acid physicochemical properties from AAindex (Kawashima et al. 2008) an online database of amino acid indices, choosing properties of particular relevance to protein structure. We also considered a number of mutational amino acid properties. We calculated the average GC content over all codons per amino acid, and then estimated the transition:transversion ratio over all codons per amino acid. As we do not know the nucleotide composition of all amino acid substitutions in the datasets we used, this is an approximation. We first calculated the proportion of pyrimidine bases to total bases for all codons encoding for a single amino acid. We then calculated the difference in this proportion between all amino acid pairs, to estimate an average transition to transversion ratio between every amino acid pair.

We also included the minimum number of mutational steps separating each amino acid. For all physicochemical and mutational amino acid properties, we calculated the absolute difference between amino acid pairs, creating a matrix of values per property. We then conducted statistical analyses on the data, after first log transforming amino acid substitution rates to better meet the assumptions of general linear models (supplementary fig. S3, Supplementary

Material online). We conducted general linear models including amino acid composition as random effects in our model, to account for the fact that amino acids appear multiple times among all pairs. We conducted all data manipulation and statistical analyses in R.

Supplementary Material

Supplementary material is available at *Genome Biology and Evolution* online.

Acknowledgments

This work was supported by a SCAS Natural Sciences Fellowship and by a grant from the Wenner-Gren Foundation. This work was also supported by the SciLifeLab and Wallenberg Data Driven Life Science Program (grant: KAW 2024.0159).

Data Availability

All code and data tables used to conduct the analyses are available at: https://github.com/j-e-james/AminoAcidSubstitutions.

Literature Cited

Carruthers T, Sun M, Baker WJ, Smith SA, De Vos JM, Eiserhardt WL. The implications of incongruence between gene tree and Species tree topologies for divergence time estimation. Syst Biol. 2022:71(5):1124–1146. https://doi.org/10.1093/sysbio/syac012.

Charlesworth B, Jensen JD. Population genetic considerations regarding evidence for biased mutation rates in Arabidopsis thaliana. Mol Biol Evol. 2023:40(2):msac275. https://doi.org/10.1093/molbev/msac275.

Chen Q, Lan A, Shen X, Wu CI. Molecular evolution in small steps under prevailing negative selection: a nearly universal rule of codon substitution. Genome Biol Evol. 2019:11(10):2702–2712. https://doi.org/10.1093/gbe/evz192.

Dagan T, Talmor Y, Graur D. Ratios of radical to conservative amino acid replacement are affected by mutational and compositional factors and may not be indicative of positive darwinian selection. Mol Biol Evol. 2002:19(7):1022–1025. https://doi.org/10.1093/oxfordjournals.molbev.a004161.

Dang CC, Minh BQ, McShea H, Masel J, James JE, Vinh LS, Lanfear R. nQMaker: estimating time nonreversible amino acid substitution models. Syst Biol. 2022:71(5):1110–1123. https://doi.org/10.1093/sysbio/syac007.

Del Amparo R, Arenas M. Influence of substitution model selection on protein phylogenetic tree reconstruction. Gene. 2023:865:147336. https://doi.org/10.1016/j.gene.2023.147336.

Del Amparo R, Branco C, Arenas J, Vicens A, Arenas M. Analysis of selection in protein-coding sequences accounting for common biases. Brief Bioinform. 2021:22(5):bbaa431. https://doi.org/10.1093/bib/bbaa431.

Duchêne S, Ho SY, Holmes EC. Declining transition/transversion ratios through time reveal limitations to the accuracy of nucleotide substitution models. BMC Evol Biol. 2015:15(1):36. https://doi.org/10.1186/s12862-015-0312-6.

El-gebali S, Mistry J, Bateman A, Eddy SR, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A, Sonnhammer ELL, et al. The Pfam protein families database in 2019. Nucleic Acids Res. 2019:47(D1):D427–D432. https://doi.org/10.1093/nar/gky995.

Ferreiro D, Branco C, Arenas M. Selection among site-dependent structurally constrained substitution models of protein evolution by approximate Bayesian computation. Bioinformatics. 2024:40(3):btae096. https://doi.org/10.1093/bioinformatics/btae096.

Ferretti L, Golubchik T, Di Lauro F, Ghafari M, Villabona-Arenas J, Atkins KE, Fraser C, Hall M. Biased estimates of phylogenetic branch lengths resulting from the discretised Gamma model of site rate heterogeneity. bioRxiv 2024:606208. https://doi.org/10.1101/2024.08.01.606208, preprint: not peer reviewed.

Freeland SJ, Hurst LD. The genetic code is one in a million. J Mol Evol. 1998:47(3):238–248. https://doi.org/10.1007/PL00006381.

Gillespie JH. The causes of molecular evolution. Oxford: Oxford University Press; 1991.

Gillespie JH, Ohta T. Development of neutral and nearly neutral theories. Theor Pop Biol. 1996:49(2):128–142. https://doi.org/10.1006/tpbi.1996.0007.

Goldman N, Thorne JL, Jones DT. Assessing the impact of secondary structure and solvent accessibility on protein evolution. Genetics. 1998:149(1):448–458. https://doi.org/10.1093/genetics/149.1.445.

Kalyaanamoorthy S, Minh BQ, Wong TKF, Von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods. 2017:14(6):587–589. https://doi.org/10.1038/nmeth.4285.

Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, Kanehisa M. AAindex: amino acid index database, progress report 2008. Nucleic Acids Res. 2008:36(Database):D202–D205. https://doi.org/10.1093/nar/gkm998.

Kimura M. The neutral allele theory of molecular evolution. Cambridge: Cambridge Univeristy Press; 1983.

Laland KN, Uller T, Feldman MW, Sterelny K, Müller GB, Moczek A, Jablonka E, Odling-Smee J. The extended evolutionary synthesis: its structure, assumptions and predictions. Proc. R. Soc. B. 2015:282(1813):20151019. https://doi.org/10.1098/rspb.2015.1019.

Liu H, Zhang J. Is the mutation rate lower in genomic regions of stronger selective constraints? Mol Biol Evol. 2022:39(8):msac169. https://doi.org/10.1093/molbev/msac169.

Majic P, Payne JL. Developmental selection and the perception of mutation bias. Mol Biol Evol. 2023:40(8):msad179. https://doi.org/10.1093/molbev/msad179.

McShea H, Weibel C, Wehbi S, Goodman P, James JE, Wheeler AL, Masel J. The effectiveness of selection in a species affects the direction of amino acid frequency evolution. bioRxiv 526552. https://doi.org/10.1101/2023.02.01.526552, 22 June 2024, preprint: not peer reviewed.

Mendes FK, Hahn MW. Gene tree discordance causes apparent substitution rate variation. Syst Biol. 2016:65(4):711–721. https://doi.org/10.1093/sysbio/syw018.

Minh BQ, Dang CC, Vinh LS, Lanfear R. QMaker: fast and accurate method to estimate empirical models of protein evolution. Syst Biol. 2021:70(5):1046–1060. https://doi.org/10.1093/sysbio/syab010.

Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, Von Haeseler A, Lanfear R, Teeling E. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic Era. Mol Biol Evol. 2020:37(5):1530–1534. https://doi.org/10.1093/molbev/msaa015.

Monroe JG, Srikant T, Carbonell-Bejerano P, Becker C, Lensink M, Exposito-Alonso M, Klein M, Hildebrandt J, Neumann M,

Kliebenstein D, et al. Mutation bias reflects natural selection in Arabidopsis thaliana. Nature. 2022:602(7895):101–105. https://doi.org/10.1038/s41586-021-04269-6.
Nei M. Mutation-driven evolution. Oxford: OUP; 2013.
Ohta T. Slightly deleterious mutant substitutions in evolution. Nature. 1973:246(5428):96–98. https://doi.org/10.1038/246096a0.
Ohta T. The nearly neutral theory of molecular evolution. Annu Rev Ecol Evol Syst. 1992:23(1):263–286. https://doi.org/10.1146/annurev.es.23.110192.001403.
Osawa S, Jukes TH. Codon reassignment (codon capture) in evolution. J Mol Evol. 1989:28(4):271–278. https://doi.org/10.1007/BF02103422.
Pandey A, Braun EL. Protein evolution is structure dependent and non-homogeneous across the tree of life. BCB '20: Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics. 2020:1–11. https://doi.org/10.1145/3388440.3412473.
Perron U, Kozlov AM, Stamatakis A, Goldman N, Moal IH, Pupko T. Modeling structural constraints on protein evolution via Side-chain conformational states. Mol Biol Evol. 2019:36(9):2086–2103. https://doi.org/10.1093/molbev/msz122.
Scolaro GE, Braun EL. The structure of evolutionary model space for proteins across the tree of life. Biology (Basel). 2023:12(2):282. https://doi.org/10.3390/biology12020282.
Smith NGC. Are radical and conservative substitution rates useful statistics in molecular evolution? J Mol Evol. 2003:57(4):467–478. https://doi.org/10.1007/s00239-003-2500-z.
Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014:30(9):1312–1313. https://doi.org/10.1093/bioinformatics/btu033.
Stoltzfus A, McCandlish DM. Mutational biases influence parallel adaptation. Mol Biol Evol. 2017:34(9):2163–2172. https://doi.org/10.1093/molbev/msx180.
Svensson EI, Berger D. The role of mutation bias in adaptive evolution. Trends Ecol Evol. 2019:34(5):422–434. https://doi.org/10.1016/j.tree.2019.01.015.
Tamura K, Stecher G, Kumar S. MEGA11: molecular evolutionary genetics analysis version 11. Mol Biol Evol. 2021:38(7):3022–3027. https://doi.org/10.1093/molbev/msab120.
Veitia RA. Who ever thought genetic mutations were random? Trends Plant Sci. 2022:27(8):733–735. https://doi.org/10.1016/j.tplants.2022.03.003.
Weber CC, Whelan S. Physicochemical amino acid properties better describe substitution rates in large populations. Mol Biol Evol. 2019:36(4):679–690. https://doi.org/10.1093/molbev/msz003.
Yang Z. PAML 4: a program package for phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007:24(8):1586–1591. https://doi.org/10.1093/molbev/msm088.
Zou Z, Zhang J. Amino acid exchangeabilities vary across the tree of life. Sci Adv. 2019:5(12):eaax3124. https://doi.org/10.1126/sciadv.aax3124.

Associate editor: Brian Golding