The effectiveness of selection in a species affects the direction of amino acid frequency evolution

Hanon McShea[1], Catherine Weibel[2,3], Sawsan Wehbi [4], Jennifer E. James[2,5], Joanna Masel[2]

[1]Department of Earth System Science, Stanford University
[2]Department of Ecology & Evolutionary Biology, University of Arizona
[3]Department of Applied Physics, Stanford University
[4]Graduate Interdisciplinary Program in Genetics, University of Arizona
[5]Department of Ecology and Genetics, Uppsala University

Abstract

Nearly neutral theory predicts that the effectiveness of natural selection will vary with species' neutral genetic diversity, or effective population size (N_e). This relationship allows comparisons among species with different effective population sizes to reveal subtle selective preferences for a variety of traits. We investigate one such trait – amino acid composition – and find that the effectiveness of selection, measured as the Codon Adaptation Index of Species (CAIS), can predict amino acid frequencies across vertebrates. In addition to amino acid frequency (an evolutionary outcome), we also measure amino acid flux (an evolutionary process) on the branches separating mouse and rat, and compare it to flux on the branches separating human and chimpanzee. We decompose flux into a shared component that reflects a variety of universal artifacts, and a species-specific component capturing any effects of N_e. The shared component of flux is correlated with structural disorder propensity. The species component is correlated with selective preference as measured by CAIS, i.e., the two methods agree on which amino acids are preferred under more effective selection, with the CAIS method providing higher statistical resolution, and producing a correlation with random peptide fitness effects. Within highly exchangeable pairs of amino acids, arginine is strongly preferred over lysine, and valine over isoleucine. Interestingly, these preferences match how thermophiles tend to differ from mesophilic relatives, as expected from theories of marginal protein stability at mutation-selection-drift balance.

Introduction

Tomoko Ohta (1973) proposed that there is evolutionarily significant variation among species in the ability of selection to purge weakly deleterious mutations. According to Ohta's nearly neutral theory of evolution, species with small "effective population sizes" (N_e) are less able to purge slightly deleterious mutations and will thus retain more of them. By estimating N_e from neutral genetic diversity, Lynch and Conery (2003) used the nearly neutral theory to explain why species with smaller effective population sizes have more bloated genomes. Species with larger effective population sizes also show stronger synonymous codon usage bias due to selection (Akashi, 1996; Galtier et al., 2018; Subramanian, 2008; Weibel et al., 2023). Here we ask whether differences among species in amino acid frequencies can also be explained by the nearly neutral theory of evolution, and if so, what this tells us about the biophysical basis of intrinsic selective preferences.

We use two complementary approaches to test, for the first time, whether differences among species in the effectiveness of selection can predict amino acid frequencies. Our first approach estimates net amino acid fluxes, i.e. the number of gains of an amino acid minus the number of losses, along a given phylogenetic branch. We compare recent fluxes in a shared set of sequences in rodent vs. in primate lineages, as we know that selection is more effective in the former (Eőry et al., 2010; Halligan et al., 2010; Keightley et al., 2005), corresponding to much larger census population sizes and greater neutral genetic diversity (Phifer-Rixey et al., 2012; Tenesa et al., 2007). Second, we quantify the degree to which the frequency of each amino acid depends on N_e, indicating the degree to which it is subtly preferred or avoided, such that selection requires high N_e for effectiveness. We do this using a recently developed metric, the Codon Adaptation Index of Species (CAIS; Weibel et al., 2023), and a broad range of vertebrate species.

The flux approach, implemented on a broad scale by Jordan et al. (2005) after a two-protein attempt by Zuckerkandl et al (1971), uses parsimony to score ancestral states, and thus amino acid gain and loss events. Its original implementation by Jordan et al. (2005) was subject to intense criticism because it assumed that all derived alleles were substitutions, when some instead could be polymorphisms – many of them slightly deleterious – segregating in the population (Hurst et al., 2006; McDonald, 2006). We address this critique in part by filtering out polymorphic sites to obtain a dataset of true substitutions (Figure 1A),

something it was not possible to do at the time Jordan et al. (2005) was published. To ensure the availability of polymorphism data, we restrict our analysis to two pairs of sister species (primates *Homo sapiens* and *Pan troglodytes*, and rodents *Mus musculus* and *Rattus norvegicus*).

However, due to the nature of parsimony-based methods, even our set of true substitutions will contain some smaller number of slightly deleterious substitutions, whose fixation may be transient and thus more likely to survive ascertainment bias when in the derived state than in the ancestral (McDonald, 2006) (Figure 1B). A similar argument applies with respect to equilibrium amino acid frequencies; a rare amino acid is comparable to a deleterious allele (McDonald, 2006).

Goldstein and Pollock (2006) noted another artifact which depends on the pattern of branch lengths, combined with rate heterogeneity (Figure 1C). In their simulations, the relative branch lengths between sister species and with the outgroup affected which types of substitutions were parsimony-informative and thus detectable. This bias in detection can produce the appearance of strong net flux, even in simulations of a stationary and time-reversible model of protein evolution, for which net flux is by definition zero. Because Jordan et al. (2005) selected their triplets of species on the basis of similar branch lengths (sister taxa pairwise amino acid divergence <15%, outgroup <25%), as do we, the same branch-length artifact will appear within any trends shared across different species triplets.

To subtract out both the universal bias toward ascertaining slightly deleterious substitutions, and the bias based on the branch length arrangements common to both our choices of species sets, we therefore statistically distinguish universal patterns (which likely reflect bias) from species-specific patterns (which may vary with the effectiveness of selection). We analyze universal trends under the hypothesis that they might reflect slightly deleterious mutations that do go to fixation (Fig. 1B), however transiently, and hence might be informative as to which mutation types are most often slightly deleterious.

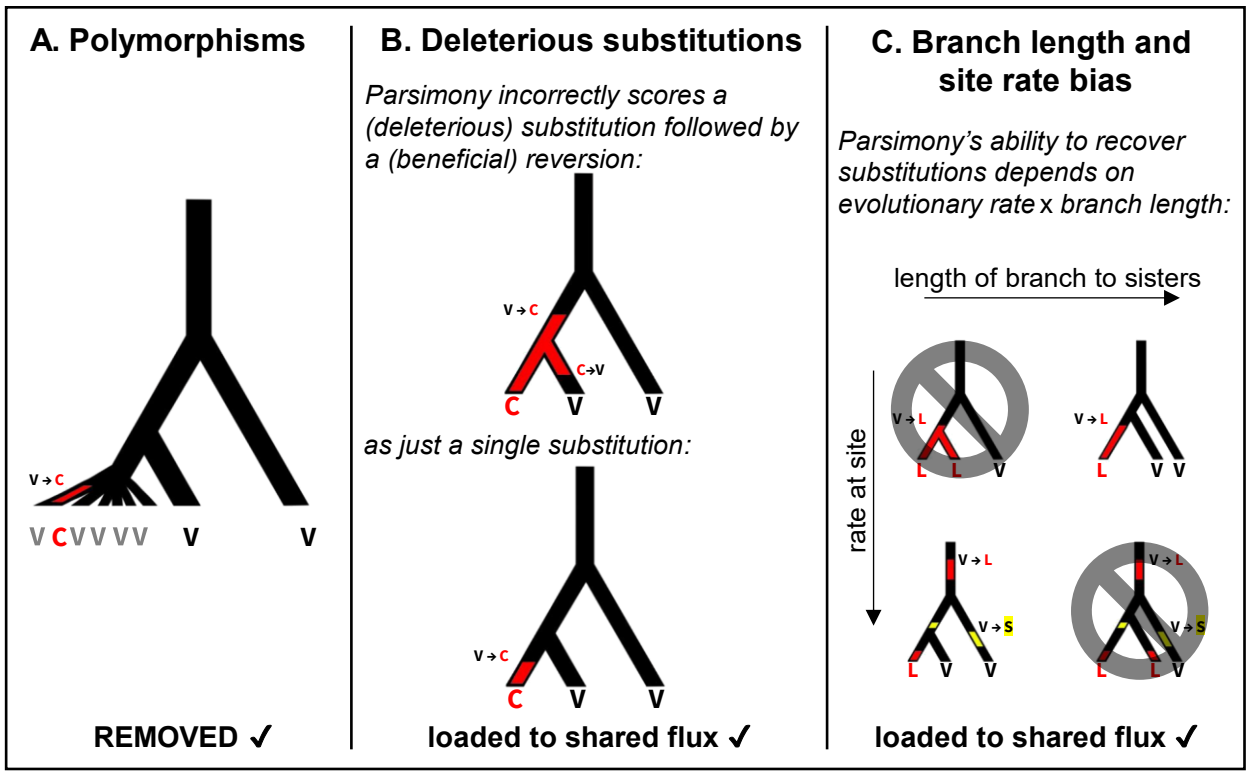

Figure 1: Artifacts of the net flux method and our approaches to them. Every color change along a branch represents a substitution at a focal site. (A) Polymorphisms were removed. (B and C) Deleterious

substitutions and branch length issues are expected to affect rodent and primate branches similarly. Flux that was seen in both rodent and primate branches was therefore subtracted, leaving a metric that describes how rodent flux differs from primate flux (Figure 3). V→C is used as an example of a slightly deleterious mutation prone to reversion, while V→L and V→S are used as examples of effectively neutral substitutions. On the left side of (C), when sister branch lengths are short, there will be greater sampling of fast-evolving sites, and when sister branch lengths are long, it is easier to detect slow-evolving sites, because they are less prone to multiple events, both within the long sister branches and in the still longer outgroup branch.

Our second approach looks at a broader range of vertebrate species, and scores an evolutionary outcome in the form of amino acid frequencies. We relate amino acid frequencies to a species' effectiveness of selection, measuring the latter by the Codon Adaptation Index of Species (CAIS; Weibel et al., 2023). CAIS captures the fraction of synonymous sites within the genome for which selection is able to overcome drift. Because the distribution of selection coefficients is likely to be similar for different vertebrate species, CAIS estimates N_e. Unlike other metrics of codon bias, it is associated with body size in vertebrates, providing validation (Weibel et al., 2023). CAIS is a measure of a species' codon bias that has been corrected for the %GC content of the species, and calculated with respect to a standardized set of amino acid frequencies.

Because each method has distinct assumptions and biases, agreement across both methods as to which amino acids are preferred in more effectively adapted species can be used to provide confidence in the results. Net flux measures recent change in amino acid frequencies in two taxa, and by focusing on a defined set of sites, avoids any danger posed by different sites being present in different taxa. However, it is vulnerable to parsimony biases, and to the possibility that the taxa differ in ways other than the limit to weak selection (e.g. %GC content, which is slightly higher in rodents than in primates). Our CAIS approach avoids the first problem and substantially reduces the second, by looking for correlations among end products arising from longer periods of evolution across more taxa. However, only our net flux approach and not CAIS can provide details not only about which amino acids are increasing vs. decreasing in frequency, but also which substitutions drive these net changes. Where both methods agree, we have greater confidence that we have detected a real evolutionary trend — namely the greater effectiveness of subtle forms of selection in high N_e species on amino acid frequencies.

Results

Patterns in net flux persist after removing polymorphisms

Following Jordan et al. (2005), we quantify net flux for each amino acid as (# gains - # losses) / (# gains + # losses), making values greater than 0 indicate a tendency towards gain. We did this for all sites with data for all four species, and that were resolvable by parsimony. Mice and rats have very similar net amino acid fluxes to each other (Supplementary Figure 1), as do humans and chimpanzees. This is expected, since %GC content, effective population size, and many other possible influences are similar within each closely related species pair. We therefore calculated "rodent" flux values by combining the amino acid gain and loss counts for mouse and rat branches, and "primate" flux values by combining the counts for human and chimpanzee branches (Figure 2A; numbers in Supplementary Table 1).

Our results correlate strongly with those of Jordan et al. (2005) (Figures 2B and 2C). However, our net flux values are generally of smaller magnitude than Jordan et al.'s (2005), particularly in primates (Figures 2B and 2C). The major difference in our method is to remove polymorphisms. We expect the primary effect of removing polymorphisms to be to remove slightly deleterious mutations (although we note that there are more to be removed for fast sites on short branches or slow sites on long branches, see Goldstein & Pollock, 2006). We interpret finding a weaker version of the same trends as evidence of a weaker version of the same bias toward slightly deleterious mutations. As discussed in the Introduction

(Figure 1B), fixation events scored by parsimony are also expected to be biased toward slightly deleterious mutations. This is because slightly deleterious mutations might sometimes reach fixation but will revert relatively quickly to a less deleterious state along subsequent branches, and therefore will be scored incorrectly.

The biggest difference between our analyses is that Jordan et al. (2005) scored polymorphisms in each reference genome as substitution events, whereas our analysis removed polymorphisms. We were also more conservative in scoring substitutions; we used two outgroup species rather than one (e.g., a substitution in mouse was only scored as such if it was different from the amino acid present in rat, human and chimpanzee). Amino acids whose dis/appearance tends to be slightly deleterious, and are therefore found disproportionately in polymorphisms relative to substitutions, will therefore spuriously appear as net losses/gains in Jordan et al. (2005). Our primate species pair is more closely related than our rodent species pair, i.e. it has fewer true substitutions. If we assume similar levels of polymorphisms in both, the smaller true divergence in primates means that removing polymorphism has a proportionately larger impact on primates fluxes than rodent fluxes.

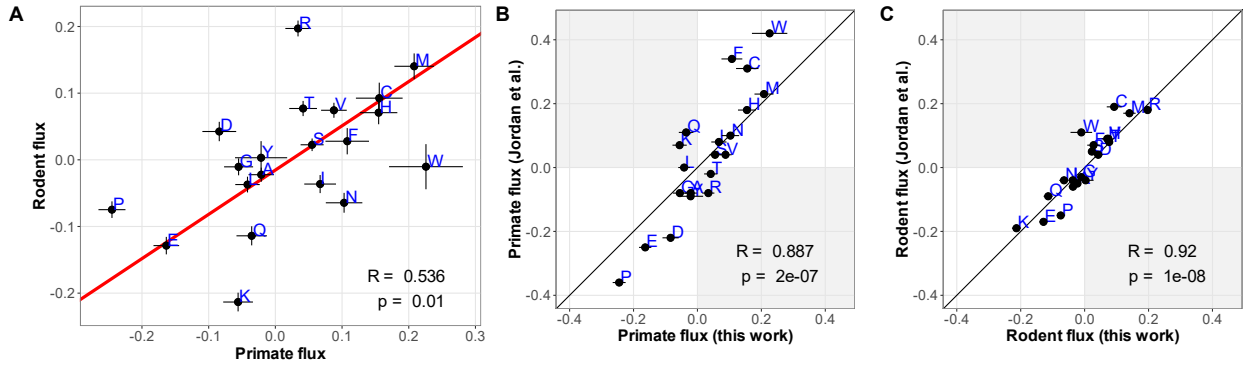

Figure 2: Net fluxes are similar for Rodents and Primates, and similar but with a tendency toward smaller magnitudes than those found by Jordan et al. (2005). Error bars show ± 1 standard error; R and P-values on figure panels are for Pearson's R correlation. (A) Primate fluxes vs. rodent fluxes. Red line shows the fit of a type II linear regression ($y = 0.664x - 0.0154$). Shared flux and species flux are calculated from this regression line, as discussed in the main text. (B and C) Flux values calculated in this study (i.e., with polymorphisms removed) vs. calculated by Jordan et al. (2005). Black lines show $y = x$, illustrating the greater magnitudes of values found by Jordan et al. (2005), especially in the case of primates. Shaded areas show sign reversal.

Fluxes shared by rodents and primates are likely driven by artifact

Artifacts in the method should be shared between rodents and primates (in addition to true universal trends, if any). In contrast, differences in net fluxes may reflect the more effective selection in rodents relative to the mutation bias that overwhelms selection in primates, as predicted by nearly neutral theory. By statistically isolating shared vs. species-specific patterns in flux, any universal trend or artifact should be reflected in the shared component, while the species-specific component will allow us to estimate the degree to which each of the 20 amino acids is favored vs. disfavored in more effectively adapted rodents relative to less well adapted primates. To do this, we fit a type-II linear regression model (Figure 2A, red line) and calculate the projection along the model line ("shared flux") for each amino acid, as well as the shortest distance from that line ("species flux"). The shared component thus reflects fluxes that are common to both rodents and primates, and the species component reflects subtle differences between the patterns of fluxes in the two groups. The sign we give to the species component allows it to be interpreted as a rodent-

driven component, i.e., a positive value indicates that the amino acid is gained in rodents to a greater degree (or lost to a lesser degree) than in primates.

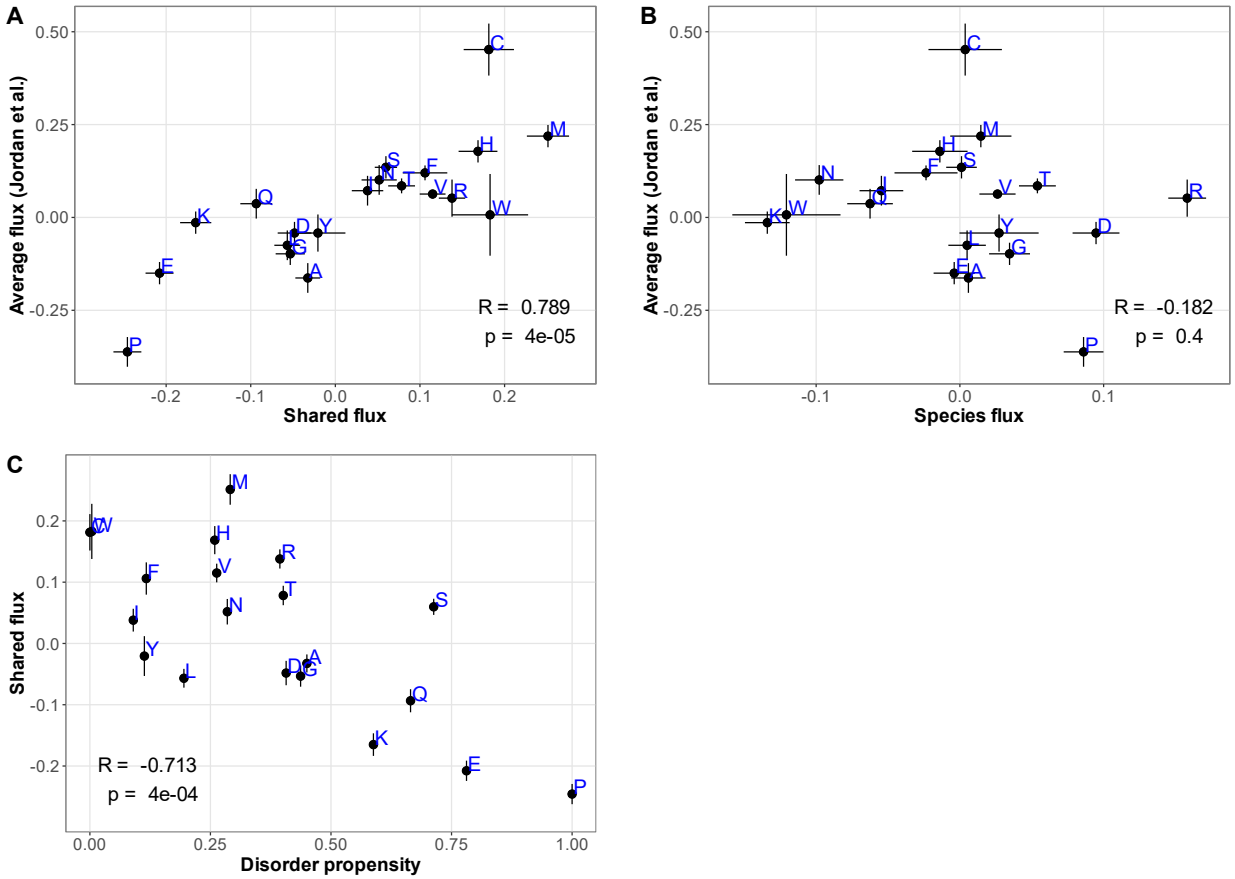

Figure 3: (A and B) The shared (but not the species) component of net flux correlates with Jordan et al.'s (2005) gain/loss rates and with amino acid disorder propensity. (C) Disorder propensity was measured by Theillet et al. (2013) as overrepresentation in disordered vs. well-folded proteins and protein regions. Error bars show ± 1 standard error; R and P-values on figure panels are for Pearson's correlation.

The shared component is large (Figure 2A). It has a strong positive relationship with Jordan et al.'s (2005) average flux (Figure 3A), while the species component has none (Figure 3B). Results are similar albeit slightly weaker for Jordan et al.'s (2005) more complex gain/loss rate metric. Amino acids that tend to be found in disordered regions of proteins are lost in both lineages (Figure 3C). We find similar results for the related protein metrics of stickiness (Pearson's R = 0.635, p = 0.003; Levy et al., 2012) and mean relative solvent accessibility (Pearson's R = 0.608, p = 0.004; Tien et al., 2013). The most explanatory of these after disorder propensity ($R^2 = 0.51$) is stickiness ($R^2 = 0.40$), an empirical measure of an amino acid's tendency to be in protein-protein interfaces rather than in non-interface surfaces (Levy et al., 2012). The species component of flux is not correlated with any of these properties. If we assume that the shared component of flux preferentially represents fast-evolving sites (Fig. 1C), we note that these are enriched in disordered regions, where slightly deleterious mutations will tend to decrease disorder through regression to the mean. Our observation that shared flux indicates a shift toward lower structural disorder is thus consistent with an interpretation that the shared component represents universal artifacts.

Two methods agree on which amino acids are preferred by selection

We hypothesize that rodent-specific fluxes reflect amino acids preferred by selection, due to more effective selection in rodents than in primates. However, it is also possible that the species component reflects idiosyncratic amino acid preferences in these two lineages, independently of any differences in the effectiveness of selection. To average over a larger range of vertebrate species to more reliably focus on variation among species in the effectiveness of selection, we assess the selective preference for a given amino acid as the slope of regression of its frequency on the Codon Adaptation Index of Species (CAIS; "CAIS slope"). CAIS slopes are not corrected for phylogeny, since phylogenetic correction makes units incomparable. To avoid the influence of differences in the composition of which proteins are present in which species, we tested whether controlling for Pfam identity made a difference, but it was not statistically supported (see Methods).

The two different metrics of selective preferences among amino acids are in good agreement (Figure 4) – that is, amino acids that are gained more or lost less in rodents than in primates are *also* at higher frequencies in vertebrate species where selection among synonymous codons is more effective. They also reveal selective preferences between biophysically similar amino acids (pairs joined by arrows in Figure 4). The shared component of flux is not correlated with CAIS slopes.

We measured the relationship between flux and CAIS slope for two slightly different sets of CAIS slopes – those measured using amino acid frequency across all protein-coding sequences, and those measured using amino acid frequency for just the subset of the proteome that belongs to a protein structural family, or Pfam (Mistry et al., 2021). The relationship between flux and whole-proteome CAIS slopes is shown in Figure 4; for flux and Pfam CAIS slopes, we have a very similar relationship with Pearson's R = 0.678, p = 0.001. We focus on the whole proteome CAIS slopes because this produces smaller standard errors for amino acid frequencies than the Pfam CAIS slopes (Supplementary Figure 2). To make this comparison more explicit, we quantify the signal-to-noise ratio as the variance among the 20 amino acid point estimates, divided by the average of their squared standard errors (see Methods). By this measure, amino acid preference measured across the proteome is slightly better resolved (signal: noise of 23 vs. 20.7) than preference measured just in Pfams. Because the two are so similar, we do use Pfams to subset the data (see Figure 6).

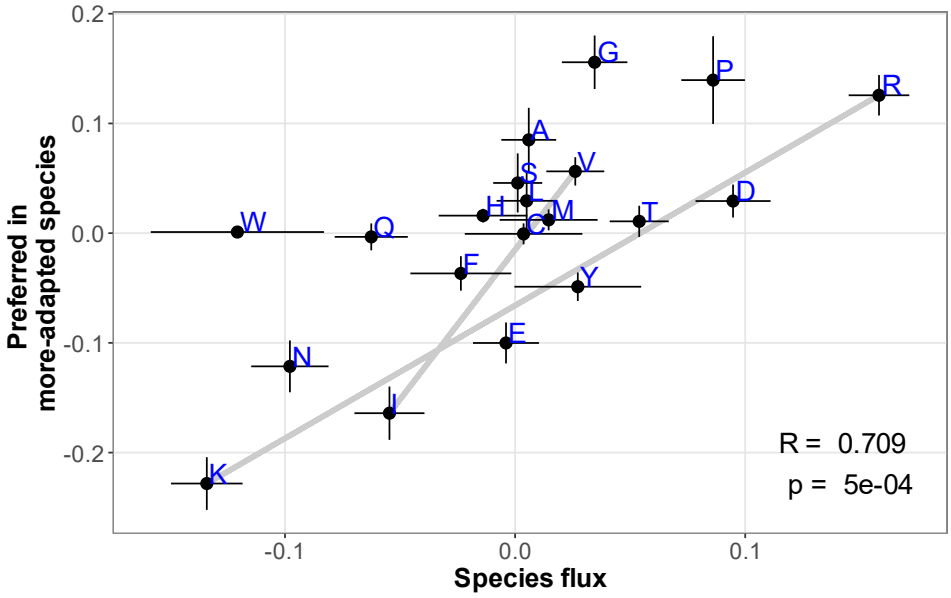

Figure 4: Two distinct methods with different biases recover similar selective preferences among amino acids. “Preferred in more-adapted species” indicates slopes from linear regressions of the proteome-wide amino acid frequency on CAIS. “Species flux” indicates the shortest distance of an amino acid to the model II fit shown in Figure 2. Error bars show ± 1 standard error; R and p-value on figure are for Pearson’s correlation. Connected pairs are biophysically similar, highly exchangeable amino acids for which flux and CAIS both show a strong preference.

Amino acids preferred in vertebrate domains are benign when expressed in random peptides

Our CAIS-based preferences show an interesting relationship to laboratory selection experiments. Kosinski et al., (2022) estimated the marginal fitness effect of each amino acid when expressed as part of a random peptide in the Escherichia coli experiments of Neme et al. (2017). Strikingly, we find that the amino acids preferred by subtle selection in existing vertebrate protein domains match those that have more benign effects when part of random peptides expressed in bacteria (Figure 5A).

Kosinski et al. (2022) found that 52% of the variance among these 20 marginal fitness effects can be explained by disorder propensity (Theillet et al., 2013) and mean molecular volume (Tsai et al., 1999b). We are not able to similarly partition predictors of CAIS-based preferences; correlations with CAIS are in the right direction, but statistically weakly significant or insignificant for disorder propensity (Figure 5B) and molecular volume (Figure 5C), as well as with the other 48% of marginal fitness value information not explained by disorder and volume (Figure 5D). Another possible predictor is the metabolic expense of amino acids. However, CAIS slopes are not correlated with amino acid costs under aerobic conditions in Saccharomyces cerevisiae (Raiford et al., 2008a), nor with amino acid nutritional essentiality in animals (Wu et al., 2014a) (Figure 5E; t-test $p = 0.2$).

The species component of flux does not show a correlation with any of the fitness-associated traits tested. In agreement with the tighter error bars seen in Figure 4 and a 1.6 times higher signal-to-noise ratio, this suggests greater statistical power from the CAIS correlation estimates than from the flux bias estimates. Thus, we focus on the former in the analyses in the next section.

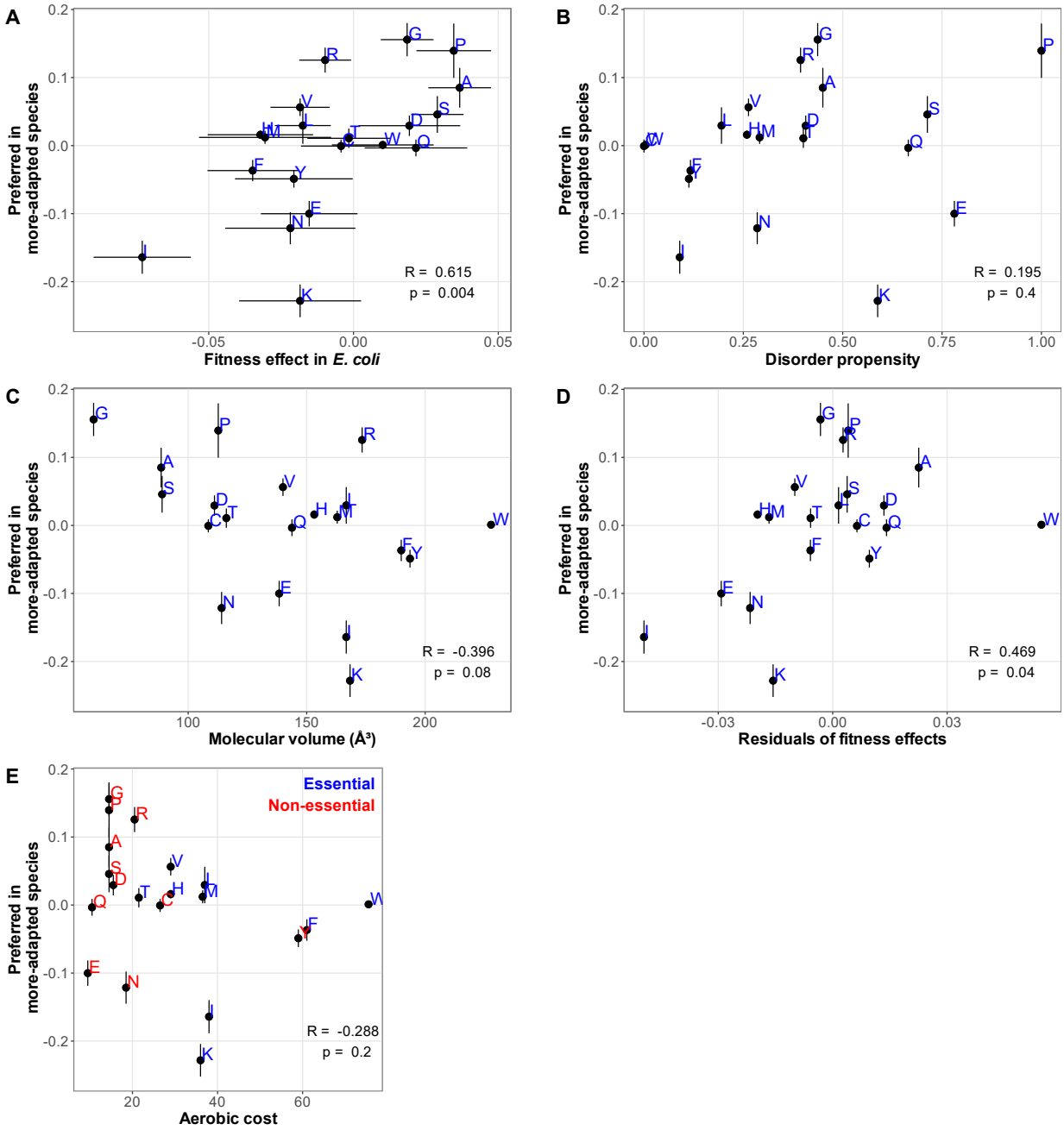

Figure 5: The amino acids preferred in vertebrate domains and rodents correlate with the estimated marginal effects of amino acids in random peptides, although it is unclear which component of these marginal effects is responsible. The y-axis "Preferred in more-adapted species" indicates CAIS slopes. Error bars show ± 1 standard error; R and P-values on figure panels are for Pearson's R correlation. (A) Amino acid fitness effects in *E. coli*, in units of the effect on genotype frequency per cycle of growth and dilution (Kosinski et al., 2022). (B) Disorder Propensity from Theillet et al. (2013), which Kosinski et al. (2022) found to be a predictor of amino acid fitness effects. (C) Mean molecular volume, units in Angstroms cubed (Tsai et al., 1999a), which Kosinski et al. (2022) found to be a predictor of amino acid fitness effects. (D) Residuals of amino acid fitness effects after controlling for volume and disorder propensity. (E) Biosynthetic cost in *S. cerevisiae* under aerobic conditions (Raiford et al., 2008b). Colors show essentiality in animals (Wu et al., 2014b).

Amino acid preferences are mostly similar across different kinds of protein domains

Because physical properties and evolutionary histories vary markedly across the proteome, we expected that selective preferences might vary among subsets of the proteome. To test this, we subsetted the proteome by transmembrane vs. non-transmembrane protein domains, and old vs. young protein domains (following the classification of James et al., 2021). Surprisingly, amino acid preferences are similar in ancient vs. recently born domains (Figure 6A). Partial exceptions are that preferences against lysine and glutamic acid, and for leucine and glycine, are stronger in young than in old Pfams.

Similarly, although transmembrane and non-transmembrane domains operate in different biophysical contexts, amino acid preferences are similar in transmembrane vs. non-transmembrane domains (Figure 6B). However, there are minor differences between the two with respect to leucine, valine, glycine, arginine, and phenylalanine. Phenylalanine frequency in transmembrane domains tends to be low in high-CAIS species, consistent with its role in increasing both membrane permeability and heteromeric interactions in transmembrane domains (Kwon et al., 2016; Perkins & Vaida, 2017), but has little dependence on CAIS when in non-transmembrane domains.

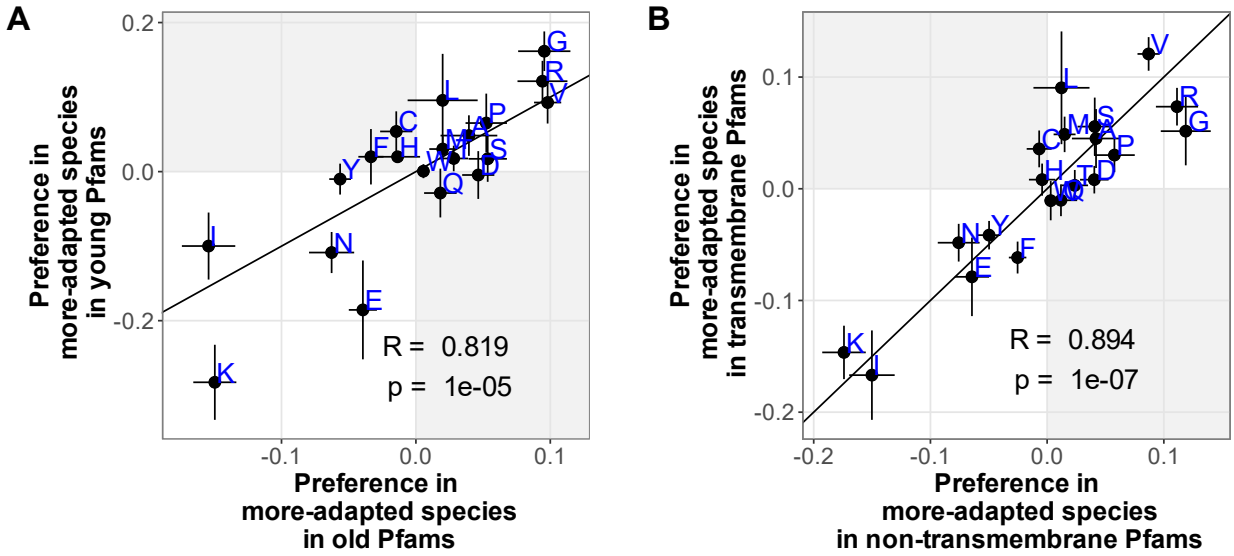

Figure 6: Amino acid preferences are similar in different groups of protein domains. Error bars show ± 1 standard error; R and P-values on figure panels are for Pearson's R correlation. Domain ages were taken from James et al. (2021), in which Pfam domains were assigned a date of evolutionary origin using TimeTree (Kumar et al., 2017). (A) Vertebrate protein domains that emerged prior to the last eukaryotic common ancestor (LECA) are categorized as "old", and protein domains that emerged after the divergence of animals and fungi from plants as "young". (B) Transmembrane vs nontransmembrane status was determined by James et al. (2021) using the program tmhmm (Krogh et al., 2001). In both panels, the black y=x lines indicate equal effects in the two sets of Pfam domains. Presence in an off-diagonal, shaded quadrant indicates that an amino acid trend has different signs for the two subsets of the data. Error bars indicate +/- one standard error.

Marked preferences between evolutionarily adjacent pairs of amino acids

We next focus on the causes of preferences between biophysically similar amino acids, which we expect to be amplified only when selection is highly effective. Rather than second guess which biophysical similarities are important (e.g. surface area, hydrophobicity, hydrogen bonding, pKa, rotameric states), we refer to empirically derived exchangeability matrices, in particular the BLOSUM (Henikoff & Henikoff, 1992) and PAM (Dayhoff, 1972) series matrices. Given the branch length range with our vertebrate dataset, we selected amino acid pairs that have exchangeability values greater than 1 in all BLOSUM and PAM matrices: these are [R,K], [V,I], [E,Q], and [M,L]. We further focus on the two pairs for which our two methods agree on the direction of preference: [R,K] and [V,I]. These pairs are connected by grey lines in Figure 4. In both cases, the strength of preference is large.

Arginine (R) is preferred over lysine (K). Notably, these are the only two amino acids with sidechains that are positively charged at neutral pH. Both amino acids are basic and aliphatic, with positively charged functional groups on an alkyl side chain, and both form salt bridges and other non-covalent interactions in protein tertiary structure. The flux data shows not only that rodents gain R and lose K, but more specifically that K residues are replaced with R (Figure 7, top left). Out of all ancestral rodent K residues that we observed substitutions away from, 48% became R (1452 of 3037 ancestral Ks), and 35% of derived rodent R residues came from K (1452 of 4126 derived Rs). This pattern is also found to a lesser degree in primates (113 of 526 and 705, or 21% and 16%, respectively; Figure 7 top right). The next two most common transitions from K in rodents are to the polar uncharged amino acids N (261 of 3037) and T (188 of 3037). In addition to K, R also comes from Q (609 of 4126), G (482 of 4126), S (274 of 4126), L (181 of 4126), and others (Figure 7 top; Supplementary Table 2).

Chemical differences between the two otherwise similar amino acids may explain the shift — K has the largest median hydrophobic water-accessible surface area in folded proteins, much larger than R (47 vs. 20 $Å^2$; Lins et al., 2003). I.e., at sites where an exposed K residue is replaced by an R, the exposed hydrophobic surface area is likely to be greatly reduced. This may increase the marginal stability of the folded protein. In further support of selection for stability, R can form more hydrogen bonds than K, which could have stabilizing effects enthalpically in the formation of secondary and tertiary structure, and entropically in the ordering of water molecules at the protein's surface. Taking into account protein-protein interactions rather than folding alone, we note that K is underrepresented in protein-protein interfaces relative to non-interface surfaces (Levy et al., 2012), while R is about equally likely to be in either. This may explain the paucity of X→K substitutions: introducing K can destroy an interface and thus be strongly deleterious. X→R substitutions might be far more numerous because R is tolerated equally well in multiple contexts.

The small hydrophobic amino acid valine (V) is preferred over isoleucine (I), with I→V substitutions again predominating in rodents but not primates (Figure 7 bottom). Isoleucine is very harmful in random peptides in *E. coli* (Kosinski et al., 2022), where it is more than twice as deleterious as the next most harmful amino acid. V and I have similar median hydrophobic surfaces areas (5 and 4 $Å^2$, respectively). However, V side chains have a smaller entropic penalty upon protein folding than I or the structurally similar alternative, L (-0.43 vs. -0.76 vs. -0.71 kcal/mol; Doig & Sternberg, 1995). Such a small but fundamental effect on stability, given other similarities, could explain the preference for V as an alternative to I. Considering the full set of transitions in Figure 7 (bottom), V shows general patterns of gain, with loss to M as an exception.

It is worth noting that in both [K,R] and [I,V] pairs, the less-preferred amino acid has an A or T nucleotide where the more-preferred one has a G or C. Rodents have slightly higher %GC than primates (42% vs. 41%; Weibel et al., 2023), so if the corresponding shift in mutation bias is still in process of

shifting amino acid frequencies, a recent increase in %GC could contribute to preferences between similar amino acids inferred from flux. However, we observed the same preferences using our CAIS approach, which is corrected for %GC.

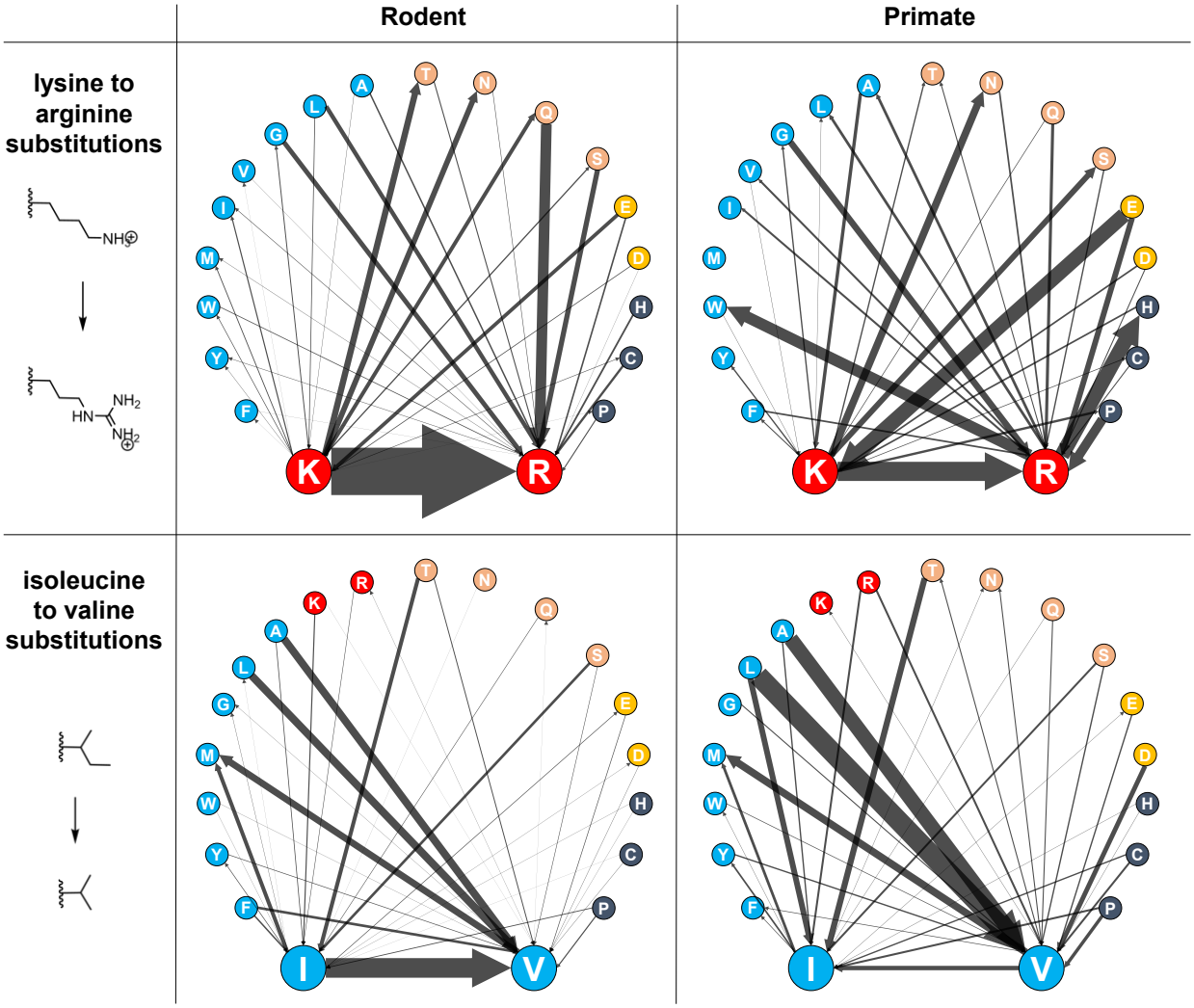

Figure 7: Net substitutions involving biophysically similar pairs of amino acids in rodent and primate lineages. Arrows represent all net substitutions involving K and R (top), or I and V (bottom). Edge weights represent net flux between each pair (i.e. the number of K→R substitutions minus the number of R→K substitutions), normalized according to the total number of substitutions in that species pair (52,431 for rodents and 9,313 for primates). To give a sense of scale, in rodents there is 1 net substitution from K to V, 179 from Q to R, and 793 from K to R. In primates there is 1 net V→K substitution, 9 Q→R, and 57 K→R. Colors are for visual aid only and represent amino acid chemical properties: hydrophobic amino acids are blue, positively charged are red, polar non-charged are pink, negatively charged are gold, and other, "special case," amino acids are charcoal. Note that the time non-reversibility shown should not be taken as a representation of the non-stationary amino acid frequencies – departures from stationarity are represented in Figure 4.

Discussion

We used two complementary methods, one measuring how amino acid fluxes differ between rodents vs. hominids, the other measuring amino acid frequency outcomes in a broader vertebrate species

set, to determine which amino acids are preferred when selection is more effective. The two methods broadly agree on which amino acids are preferred in vertebrates, with CAIS – the outcome metric – having better statistical resolution than flux. The most striking preferences are between pairs of biochemically similar amino acids, with selection preferring arginine over lysine, and valine over isoleucine. More preferred amino acids tend to have more benign effects when expressed as part of random peptides in *E. coli*.

The flux method has been much criticized, with contamination by non-fixed slightly deleterious mutations a particular concern. Although we partially addressed this concern by removing polymorphisms, our shared flux component remains similar to the artifact-laden flux patterns of Jordan et al. (2005). We think these shared patterns stem from slightly deleterious mutations and/or a phylogenetic tree that results in better ascertainment at faster evolving sites. Whatever is responsible, it correlates with intrinsic structural disorder, i.e., slightly deleterious mutations are enriched for substitutions that promote order, and/or substitutions in rapidly evolving disordered regions (making them less disordered) are subject to a higher ascertainment rate. While the flux method has weaknesses, it allowed us to analyze pairwise substitution rates between each of the twenty amino acids, revealing that within pairs of highly exchangeable amino acids, rodents exhibit a pronounced net flux from the less favored member of the pair to the more favored, in a manner not found in primates (Figure 7). The fluxes calculated here show a clear departure from assumptions of stationarity built into most models of amino acid substitution (Dang et al., 2022; Le et al., 2012; Le & Gascuel, 2008; Minh et al., 2021; Whelan & Goldman, 2001).

Inferring a non-stationary amino acid substitution model by maximum likelihood, rather than relying on parsimony, would take care of some of the artifacts we dealt with here, in particular those depicted in Figures 1b and 1c, although it is still necessary to filter out polymorphisms from the data, so that the model is not trained as though they were true substitutions. Unfortunately, non-stationary likelihood-based approaches to amino acid substitution models remain in their infancy, despite promising applications to nucleotide evolution (Kaehler et al., 2015, 2017). A major issue is the large number of parameters that must be fit, especially since branch heterogeneity (with respect to the amino acid substitution matrix, not merely the rate) is also required. Groussin et al. (2013) partially solved this problem by using a single exchangeability matrix across the complete tree, and varying only the vector of equilibrium amino acid frequencies among branches. However, fitting 19 parameters per branch would still be prohibitive. Groussin et al. (2013) used a correspondence analysis approach to instead vary only two parameters per branch, describing changes in GC content and optimal growth temperature, but falling short of a complete description of changes in amino acid content. We hope our findings help prompt future developments in this area.

A traditional view of structural biology might be skeptical of the value of information that is restricted to amino acid frequencies. However, it is striking the degree to which amino acid frequencies alone, absent information about amino acid ordering, are able to predict many properties of a polypeptide, including translational efficiency, folding kinetics, and local protein structure (Buhr et al., 2016; Saunders & Deane, 2010; Zarin et al., 2017, 2021). It is encouraging that the direction of selective preference within [K, R] and [I, V] pairs is consistent with expected selection on protein folding and stability, as supported by differences in median hydrophobic surface area, hydrogen bonding capacity, and particularly entropic penalty of folding. Note that the preferences for R over K and V over I is a statistical tendency; the opposite might still hold at a (minority of) individual sites. Because most proteins must fold (and unfold, and fold again), it is not surprising that the thermodynamics of folding would be subject to selection detectable at the level of amino acid frequencies. The specific biophysical hypotheses suggested by our observational work could be experimentally tested by aggregating data on arginine vs. lysine or valine vs. isoleucine mutations across a set of model proteins, where the effect of each mutation on marginal stability and kinetics was measured using multiplexed assays (Atsavapranee et al., 2021; Markin et al., 2021; Tsuboyama et al., 2023).

Flux measurements in extremophiles have yielded fascinatingly parallel results, with thermophily associated with the same trends as more effective selection. Lecocq et al. (2021) found that thermophiles in the archaeal order Methanococcales use R more often, and that K→R substitutions accompany transitions to higher optimal growth temperature while R→K substitutions accompany transitions to lower growth temperature. In thermophilic species of filamentous fungi, Van Noort et al. (2013) also found an excess of K→R substitutions using a similar method. Interestingly, Fontanillas et al. (2017) found an excess of R→K and V→I substitutions during a shift to cold adaptation of hydrothermal vent-dwelling polychaete worms, but did not find a corresponding excess of K→R or I→V substitutions in sister species that remained thermophiles throughout. The cold-adapted species had higher substitution rates and slightly higher dN/dS ratios (both <<<1), suggesting relaxed selection/reduced constraint relative to hot-adapted species.

In all of these taxa, the amino acid preferred in hot conditions is the same as preferred by highly effective selection in our findings. Proteins tend to evolve marginal stability at mutation-selection-drift balance (Bloom, Raval, et al., 2007), with greater stability improving robustness to mistranscription and mistranslation (Drummond et al., 2005; Drummond & Wilke, 2008, 2009; Serohijos et al., 2012; Wilke et al., 2005), and potentially also mutation (Bloom, Lu, et al., 2007). Both more effective selection and hotter temperature will cause this stability to be slightly less marginal.

Pinney et al. (2021) identified protein sites at which amino acid identity correlated with optimal growth temperature across the bacterial domain. Opposite to our work and the studies above, at these sites, K and I are strongly associated with high growth temperature in bacterial enzymes, while R and V show no association with temperature. Interestingly, hot-adapted enzymes are also enriched in I-I, K-K, and I-K spatial interactions, and K-E salt bridges are also significantly enriched in hot-adapted enzymes, while R-D salt bridges are enriched in cold-adapted enzymes (K-D and R-E salt bridges were not significant). Perhaps against a background that is generally depleted in K and I by adaptation to hot conditions, the importance of the remaining K and I interactions becomes more important. In other words, the few K and I residues remaining may have important site-specific functions.

Amino acid frequency differences across the tree of life (Basile et al., 2019; Bogatyreva et al., 2006) have been previously attributed to differences in metabolism (Krick et al., 2014), environmental amino acid availability (Moura et al., 2013), environment and body temperature (Cambillau & Claverie, 2000; Farias & Bonato, 2003; Glyakina et al., 2007; McDonald, 2010; Pinney et al., 2021; Suhre & Claverie, 2003; Wang & Lercher, 2010), and nucleotide composition (Brbić et al., 2015). We found that differences in amino acid frequencies also depend on the effectiveness of selection on the species as a whole. It is likely that the effectiveness of selection also affects other traits, and CAIS can be used to investigate these.

Proteome-wide amino acid frequencies are shaped by some combination of protein domain birth, death, and tinkering. Differential domain loss and duplication (James et al., 2023) shifts animal proteomes toward amino acid compositions favoring lower intrinsic structural disorder (James et al., 2021). This is counterbalanced by *de novo* birth (Wilson et al., 2017). Our work here instead focuses on descent with modification, which is likely to operate on faster timescales. This focus is strictly enforced in the flux analysis by considering only homologous sites shared by all four species. Interpretation of outcome results is also suggestive of descent with modification, given that controlling for Pfam identity was not statistically supported for individual amino acids, and that similar results were previously found for intrinsic structural disorder in which this control was included (Weibel et al., 2023).

There are many metrics one can use to order the 20 amino acids. Here we provide another one more (regression slope with CAIS, which has better statistical resolution than flux measurements), one that captures the subtle preferences of selection. We compare our new metric to disorder propensity, size,

metabolic cost, tendency to benign effects in *E. coli*, and trends with respect to domain age. There could be other correlates – notably preferences related to growth temperature. Much can be learned about proteins, and their evolution, through these comparisons, and finding points of overlap or lack thereof.

Amino acids have distinctive biophysical properties, and selective preference among these may be universal. The concordance of our mammalian study with previous results on random peptides in *E. coli* supports the idea that for amino acid composition, "anything found to be true of *E. coli* must also be true of Elephants" (Monod & Jacob, 1961). Sometimes the demands on a position within a protein are different, e.g. for alpha helix vs. beta sheet vs. intrinsic disorder, for surface vs. buried, for acid vs. base, or for lesser vs. greater rotameric degrees of freedom, but given a particular demand, there are subtle preferences among amino acids that can deliver similar things. This is what we have quantified. This is nearly neutral theory, as applied to the essence of biochemistry.

Materials and Methods

Amino acid flux

In this analysis, we used the 65 amniota vertebrates multiple genome alignment, available on Ensembl (https://useast.ensembl.org/info/genome/compara/mlss.html?mlss=2041) (Yates et al., 2020). This alignment was generated using the Mercator-Pecan pipeline, which effectively handles rearrangements and duplications that have occurred during species' evolutionary history by organizing genomes into segments to create a homology map, prior to alignment (Paten et al., 2008). We downloaded the aligned coding regions for our four focal species from the Ensembl REST API server (https://rest.ensembl.org/, Ensembl 98, accessed during the months of November and December 2019) using our own Python scripts, using the mouse reference genome to retrieve alignment blocks based on mouse coding regions. To be included in our dataset, genomic regions for all four of our focal species had to be present in the multiple genome alignment.

The locations of polymorphic sites to be excluded were taken from the Ensembl variation database (Hunt et al., 2018). If a polymorphism reached a threshold minor allele frequency of 5% in any of the populations in which it was reported, that site was masked in the multiple species alignment. We note that due to the larger number of studies on human populations, there are more human polymorphisms recorded in Ensembl, and thus masked in our data, and fewer for rats and chimpanzees.

Aligned DNA sequences were then translated, and informative substitutions were identified, using our own Python scripts. To be included in our analysis, an amino acid site had to be identical in three species and different in the fourth, i.e., inferred by parsimony to occur on one of the branches between mouse-rat or human-chimp. For each species in our quartet, the normalized gain/loss ratio (flux) for each amino acid was calculated by the same method as Jordan et al. (2005), where flux = (# derived - #ancestral)/(#derived + # ancestral). Net flux values and the shared and species-specific components are given in Supplementary Table 1.

CAIS

We use the CAIS calculations for 117 vertebrate species from Weibel et al., (2023). In that paper, each species was assigned a fixed effect on the quantity of interest (in our case amino acid frequency), while controlling for Pfam identity as a random effect. While previously supported for intrinsic structural disorder, in our current analysis of amino acid frequencies, controlling for Pfam identity was not supported ($p > 0.05$ for all 20 amino acids). We therefore instead use the simple fractional content of each amino acid in the data for each species. We did this both for the whole proteome, and for the set of annotated Pfam sequences, and found the proteome-wide measurements have a higher signal-to-noise ratio (see below). In both cases, we then performed 20 weighted linear regressions to predict amino acid frequencies from each species' CAIS value, and then used the regression slopes to summarize the degree of preference for each amino acid, together with the associated standard errors on the slopes (Supplementary Table 3). Linear models were implemented using the R package lme4 (Bates et al., 2015). Positive slopes indicate that species with more effective selection (high CAIS) prefer that amino acid, negative that they avoid it. We did not, as in previous work (Weibel et al., 2023), perform regression using the method of Phylogenetically Independent Contrasts (PIC), because PIC regression slopes have inconsistent units that cannot be compared across the 20 models of each of the 20 amino acid frequencies.

Signal-to-noise

A signal-to-noise ratio for both CAIS slopes and net flux was calculated as the variance among the 20 amino acid values over the average of the 20 associated squared errors, e.g. $\frac{\sum(x-\mu)^2/19}{\sum err^2/20}$ where x is the CAIS slope or net flux value for each amino acid, μ is their mean, and *err* is the standard error for the CAIS or species flux estimate for each amino acid. The signal-to-noise ratio for CAIS slopes in the proteome-wide models (23.0) was slightly higher than for the Pfam models (20.7), and the magnitudes of the signals also tended to be higher in the proteome correlations (0.01 for proteome vs. 0.006 for Pfam vs. 0.01; Supplementary Figure 2). We therefore use the proteome-wide data by default in our CAIS analyses.

Pfam subsets

In the vertebrate genomes, Pfams that emerged prior to the last eukaryotic common ancestor (LECA) are identified as "old", and Pfams that emerged after the divergence of animals and fungi from plants are identified as "young", as annotated by James et al. (2021). Pfams were assigned as either transmembrane or non-transmembrane by James et al. (2021) using the program Tmhmm (Krogh et al., 2001). Amino acid frequencies were calculated for each Pfam subset in each species.

Data availability

Full substitution tables used to calculate fluxes and CAIS slopes, as well as the values themselves and all code, calculations, and supplementary data are available at https://github.com/cyclase/aa_flux. All statistical modelling was done in R 3.5.1, 3.6.3, 4.2.1, or 4.3.1.

Acknowledgments

JM and CW were supported by the John Templeton Foundation (60814). CW was supported by the Arnold and Mabel Beckman Foundation Scholars Program, the Western Alliance to Expand Student Opportunities (WAESO), Louis Stokes Alliance for Minority Participation (LSAMP) National Science Foundation (NSF) Cooperative Agreement (HRD-1101728), and the UA/NASA Space Grant Undergraduate Research Internship program. HM was supported by an NSF Graduate Research Fellowship. SW was supported in part by NASA Astrobiology Program ICAR grant [80NSSC21K0592] to Betul Kacar.

We thank Andrew Wheeler, Catherine DeBrule Stark, Em Horst, Bennett Kapili, Minh Bui, and Luke Kosinski for helpful discussions.

Citations

Akashi, H. (1996). Molecular Evolution Between Drosophila melanogaster and D. simulans Reduced Codon Bias, Faster Rates of Amino Acid Substitution, and Larger Proteins in D. melanogaster. *Genetics*, *144*(3), 1297–1307. https://doi.org/10.1093/genetics/144.3.1297

Atsavapranee, B., Stark, C. D., Sunden, F., Thompson, S., & Fordyce, P. M. (2021). Fundamentals to function: Quantitative and scalable approaches for measuring protein stability. *Cell Systems*, *12*(6), 547–560. https://doi.org/10.1016/j.cels.2021.05.009

Basile, W., Salvatore, M., Bassot, C., & Elofsson, A. (2019). Why do eukaryotic proteins contain more intrinsically disordered regions? *PLOS Computational Biology*, *15*(7), e1007186. https://doi.org/10.1371/journal.pcbi.1007186

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Bloom, J. D., Lu, Z., Chen, D., Raval, A., Venturelli, O. S., & Arnold, F. H. (2007). Evolution favors protein mutational robustness in sufficiently large populations. *BMC Biology*, *5*(1), 29. https://doi.org/10.1186/1741-7007-5-29

Bloom, J. D., Raval, A., & Wilke, C. O. (2007). Thermodynamics of neutral protein evolution. *Genetics*, *175*(1), 255–266. https://doi.org/10.1534/genetics.106.061754

Bogatyreva, N., Finkelstein, A., & Galzitskaya, O. (2006). Trend of amino acid composition of proteins of different taxa. *Journal of Bioinformatics and Computational Biology*, *4*, 597–608. https://doi.org/10.1142/S0219720006002016

Brbić, M., Warnecke, T., Kriško, A., & Supek, F. (2015). Global Shifts in Genome and Proteome Composition Are Very Tightly Coupled. *Genome Biology and Evolution*, *7*(6), 1519–1532. https://doi.org/10.1093/gbe/evv088

Buhr, F., Jha, S., Thommen, M., Mittelstaet, J., Kutz, F., Schwalbe, H., Rodnina, M. V., & Komar, A. A. (2016). Synonymous Codons Direct Cotranslational Folding toward Different Protein Conformations. *Molecular Cell*, *61*(3), 341–351. https://doi.org/10.1016/j.molcel.2016.01.008

Cambillau, C., & Claverie, J.-M. (2000). Structural and Genomic Correlates of Hyperthermostability *. *Journal of Biological Chemistry*, *275*(42), 32383–32386. https://doi.org/10.1074/jbc.C000497200

Dang, C. C., Minh, B. Q., McShea, H., Masel, J., James, J. E., Vinh, L. S., & Lanfear, R. (2022). nQMaker: Estimating Time Nonreversible Amino Acid Substitution Models. *Systematic Biology*, syac007. https://doi.org/10.1093/sysbio/syac007

Dayhoff, M. O. (1972). *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation.

Doig, A. J., & Sternberg, M. J. E. (1995). Side-chain conformational entropy in protein folding. *Protein Science*, *4*(11), 2247–2251. https://doi.org/10.1002/pro.5560041101

Drummond, D. A., Bloom, J. D., Adami, C., Wilke, C. O., & Arnold, F. H. (2005). Why highly expressed proteins evolve slowly. *Proceedings of the National Academy of Sciences*, *102*(40), 14338–14343. https://doi.org/10.1073/pnas.0504070102

Drummond, D. A., & Wilke, C. O. (2008). Mistranslation-Induced Protein Misfolding as a Dominant Constraint on Coding-Sequence Evolution. *Cell*, *134*(2), 341–352. https://doi.org/10.1016/j.cell.2008.05.042

Drummond, D. A., & Wilke, C. O. (2009). The evolutionary consequences of erroneous protein synthesis. *Nature Reviews Genetics*, *10*(10), Article 10. https://doi.org/10.1038/nrg2662

Eőry, L., Halligan, D. L., & Keightley, P. D. (2010). Distributions of Selectively Constrained Sites and Deleterious Mutation Rates in the Hominid and Murid Genomes. *Molecular Biology and Evolution*, *27*(1), 177–192. https://doi.org/10.1093/molbev/msp219

Farias, S. T., & Bonato, M. C. M. (2003). Preferred amino acids and thermostability. *Genetics and Molecular Research*, *2.4*, 383–393.

Fontanillas, E., Galzitskaya, O. V., Lecompte, O., Lobanov, M. Y., Tanguy, A., Mary, J., Girguis, P. R., Hourdez, S., & Jollivet, D. (2017). Proteome Evolution of Deep-Sea Hydrothermal Vent Alvinellid Polychaetes Supports the Ancestry of Thermophily and Subsequent Adaptation to Cold in Some Lineages. *Genome Biology and Evolution*, *9*(2), 279–296. https://doi.org/10.1093/gbe/evw298

Galtier, N., Roux, C., Rousselle, M., Romiguier, J., Figuet, E., Glémin, S., Bierne, N., & Duret, L. (2018). Codon Usage Bias in Animals: Disentangling the Effects of Natural Selection, Effective Population Size, and GC-Biased Gene Conversion. *Molecular Biology and Evolution*, *35*(5), 1092–1103. https://doi.org/10.1093/molbev/msy015

Glyakina, A. V., Garbuzynskiy, S. O., Lobanov, M. Yu., & Galzitskaya, O. V. (2007). Different packing of external residues can explain differences in the thermostability of proteins from thermophilic and mesophilic organisms. *Bioinformatics*, *23*(17), 2231–2238. https://doi.org/10.1093/bioinformatics/btm345

Goldstein, R. A., & Pollock, D. D. (2006). Observations of Amino Acid Gain and Loss during Protein Evolution Are Explained by Statistical Bias. *Molecular Biology and Evolution*, *23*(7), 1444–1449. https://doi.org/10.1093/molbev/msl010

Groussin, M., Boussau, B., & Gouy, M. (2013). A branch-heterogeneous model of protein evolution for efficient inference of ancestral sequences. *Systematic Biology*, *62*(4), 523–538. https://doi.org/10.1093/sysbio/syt016

Halligan, D. L., Oliver, F., Eyre-Walker, A., Harr, B., & Keightley, P. D. (2010). Evidence for Pervasive Adaptive Protein Evolution in Wild Mice. *PLOS Genetics*, *6*(1), e1000825. https://doi.org/10.1371/journal.pgen.1000825

Henikoff, S., & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America*, *89*(22), 10915–10919. https://doi.org/10.1073/pnas.89.22.10915

Hunt, S. E., McLaren, W., Gil, L., Thormann, A., Schuilenburg, H., Sheppard, D., Parton, A., Armean, I. M., Trevanion, S. J., Flicek, P., & Cunningham, F. (2018). Ensembl variation resources. *Database*, *2018*, bay119. https://doi.org/10.1093/database/bay119

Hurst, L. D., Feil, E. J., & Rocha, E. P. C. (2006). Causes of trends in amino-acid gain and loss. *Nature*, *442*(7105), E11–E12. https://doi.org/10.1038/nature05137

James, J. E., Nelson, P., & Masel, J. (2023). Differential retention of Pfam domains creates long-term evolutionary trends. *Molecular Biology and Evolution*, *2023*. https://doi.org/10.1101/2022.10.27.514087

James, J. E., Willis, S. M., Nelson, P. G., Weibel, C., Kosinski, L. J., & Masel, J. (2021). Universal and taxon-specific trends in protein sequences as a function of age. *eLife*, *10*, e57347. https://doi.org/10.7554/eLife.57347

Jordan, I. K., Kondrashov, F. A., Adzhubei, I. A., Wolf, Y. I., Koonin, E. V., Kondrashov, A. S., & Sunyaev, S. (2005). A universal trend of amino acid gain and loss in protein evolution. *Nature*, *433*(7026), 633–638. https://doi.org/10.1038/nature03306

Kaehler, B. D., Yap, V. B., & Huttley, G. A. (2017). Standard Codon Substitution Models Overestimate Purifying Selection for Nonstationary Data. *Genome Biology and Evolution*, *9*(1), 134–149. https://doi.org/10.1093/gbe/evw308

Kaehler, B. D., Yap, V. B., Zhang, R., & Huttley, G. A. (2015). Genetic Distance for a General Non-Stationary Markov Substitution Process. *Systematic Biology*, *64*(2), 281–293. https://doi.org/10.1093/sysbio/syu106

Keightley, P. D., Lercher, M. J., & Eyre-Walker, A. (2005). Evidence for Widespread Degradation of Gene Control Regions in Hominid Genomes. *PLOS Biology*, *3*(2), e42. https://doi.org/10.1371/journal.pbio.0030042

Kosinski, L., Aviles, N., Gomez, K., & Masel, J. (2022). Random peptides rich in small and disorder-promoting amino acids are less likely to be harmful. *Genome Biology and Evolution*, evac085. https://doi.org/10.1093/gbe/evac085

Krick, T., Verstraete, N., Alonso, L. G., Shub, D. A., Ferreiro, D. U., Shub, M., & Sánchez, I. E. (2014). Amino Acid Metabolism Conflicts with Protein Diversity. *Molecular Biology and Evolution*, *31*(11), 2905–2912. https://doi.org/10.1093/molbev/msu228

Krogh, A., Larsson, B., von Heijne, G., & Sonnhammer, E. L. L. (2001). Predicting transmembrane protein topology with a hidden markov model: Application to complete genomes. *Journal of Molecular Biology*, *305*(3), 567–580. https://doi.org/10.1006/jmbi.2000.4315

Kumar, S., Stecher, G., Suleski, M., & Hedges, S. B. (2017). TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Molecular Biology and Evolution*, *34*(7), 1812–1819. https://doi.org/10.1093/molbev/msx116

Kwon, M. J., Park, J., Jang, S., Eom, C. Y., & Oh, E. S. (2016). The conserved phenylalanine in the transmembrane domain enhances heteromeric interactions of syndecans. *Journal of Biological Chemistry*, *291*(2), 872–881. https://doi.org/10.1074/jbc.M115.685040

Le, S. Q., Dang, C. C., & Gascuel, O. (2012). Modeling Protein Evolution with Several Amino Acid Replacement Matrices Depending on Site Rates. *Molecular Biology and Evolution*, *29*(10), 2921–2936. https://doi.org/10.1093/molbev/mss112

Le, S. Q., & Gascuel, O. (2008). An improved general amino acid replacement matrix. *Molecular Biology and Evolution*, *25*(7), 1307–1320. https://doi.org/10.1093/molbev/msn067

Lecocq, M., Groussin, M., Gouy, M., & Brochier-Armanet, C. (2021). The Molecular Determinants of Thermoadaptation: Methanococcales as a Case Study. *Molecular Biology and Evolution*, *38*(5), 1761–1776. https://doi.org/10.1093/molbev/msaa312

Levy, E. D., De, S., & Teichmann, S. A. (2012). Cellular crowding imposes global constraints on the chemistry and evolution of proteomes. *Proceedings of the National Academy of Sciences*, *109*(50), 20461–20466. https://doi.org/10.1073/pnas.1209312109

Lins, L., Thomas, A., & Brasseur, R. (2003). Analysis of accessible surface of residues in proteins. *Protein Science*, *12*(7), 1406–1417. https://doi.org/10.1110/ps.0304803

Lynch, M., & Conery, J. S. (2003). The Origins of Genome Complexity. *Science*, *302*(5649), 1401–1404. https://doi.org/10.1126/science.1089370

Markin, C. J., Mokhtari, D. A., Sunden, F., Appel, M. J., Akiva, E., Longwell, S. A., Sabatti, C., Herschlag, D., & Fordyce, P. M. (2021). Revealing enzyme functional architecture via high-throughput microfluidic enzyme kinetics. *Science*, *373*(6553), eabf8761. https://doi.org/10.1126/science.abf8761

McDonald, J. H. (2006). Apparent Trends of Amino Acid Gain and Loss in Protein Evolution Due to Nearly Neutral Variation. *Molecular Biology and Evolution*, *23*(2), 240–244. https://doi.org/10.1093/molbev/msj026

McDonald, J. H. (2010). Temperature Adaptation at Homologous Sites in Proteins from Nine Thermophile–Mesophile Species Pairs. *Genome Biology and Evolution*, *2*, 267–276. https://doi.org/10.1093/gbe/evq017

Minh, B. Q., Dang, C. C., Vinh, L. S., & Lanfear, R. (2021). QMaker: Fast and Accurate Method to Estimate Empirical Models of Protein Evolution. *Systematic Biology*, *70*(5), 1046–1060. https://doi.org/10.1093/sysbio/syab010

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., Tosatto, S. C. E., Paladin, L., Raj, S., Richardson, L. J., Finn, R. D., & Bateman, A. (2021). Pfam: The protein families database in 2021. *Nucleic Acids Research*, *49*(D1), D412–D419. https://doi.org/10.1093/nar/gkaa913

Monod, J., & Jacob, F. (1961). Teleonomic mechanisms in cellular metabolism, growth, and differentiation. *Cold Spring Harbor Symposia on Quantitative Biology*, *26*, 389–401. https://doi.org/10.1101/sqb.1961.026.01.048

Moura, A., Savageau, M. A., & Alves, R. (2013). Relative Amino Acid Composition Signatures of Organisms and Environments. *PLOS ONE*, *8*(10), e77319. https://doi.org/10.1371/journal.pone.0077319

Neme, R., Amador, C., Yildirim, B., McConnell, E., & Tautz, D. (2017). Random sequences are an abundant source of bioactive RNAs or peptides. *Nature Ecology & Evolution*, *1*(6), 1–7. https://doi.org/10.1038/s41559-017-0127

Ohta, T. (1973). Slightly Deleterious Mutant Substitutions in Evolution. *Nature*, *246*(5428), 96–98. https://doi.org/10.1038/246096a0

Paten, B., Herrero, J., Beal, K., Fitzgerald, S., & Birney, E. (2008). Enredo and Pecan: Genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Research*, *18*(11), 1814–1828. https://doi.org/10.1101/gr.076554.108

Perkins, R., & Vaida, V. (2017). Phenylalanine Increases Membrane Permeability. *Journal of the American Chemical Society*, *139*(41), 14388–14391. https://doi.org/10.1021/jacs.7b09219

Phifer-Rixey, M., Bonhomme, F., Boursot, P., Churchill, G. A., Piálek, J., Tucker, P. K., & Nachman, M. W. (2012). Adaptive Evolution and Effective Population Size in Wild House Mice. *Molecular Biology and Evolution*, *29*(10), 2949–2955. https://doi.org/10.1093/molbev/mss105

Pinney, M. M., Mokhtari, D. A., Akiva, E., Yabukarski, F., Sanchez, D. M., Liang, R., Doukov, T., Martinez, T. J., Babbitt, P. C., & Herschlag, D. (2021). Parallel molecular mechanisms for enzyme temperature adaptation. *Science*, *371*(6533), eaay2784. https://doi.org/10.1126/science.aay2784

Raiford, D. W., Heizer, E. M., Miller, R. V., Akashi, H., Raymer, M. L., & Krane, D. E. (2008a). Do Amino Acid Biosynthetic Costs Constrain Protein Evolution in Saccharomyces cerevisiae? *Journal of Molecular Evolution*, *67*(6), 621–630. https://doi.org/10.1007/s00239-008-9162-9

Raiford, D. W., Heizer, E. M., Miller, R. V., Akashi, H., Raymer, M. L., & Krane, D. E. (2008b). Do amino acid biosynthetic costs constrain protein evolution in Saccharomyces cerevisiae? *Journal of Molecular Evolution*, *67*(6), 621–630. https://doi.org/10.1007/s00239-008-9162-9

Saunders, R., & Deane, C. M. (2010). Synonymous codon usage influences the local protein structure observed. *Nucleic Acids Research*, *38*(19), 6719–6728. https://doi.org/10.1093/nar/gkq495

Serohijos, A. W. R., Rimas, Z., & Shakhnovich, E. I. (2012). Protein Biophysics Explains Why Highly Abundant Proteins Evolve Slowly. *Cell Reports*, *2*(2), 249–256. https://doi.org/10.1016/j.celrep.2012.06.022

Subramanian, S. (2008). Nearly neutrality and the evolution of codon usage bias in eukaryotic genomes. *Genetics*, *178*(4), 2429–2432. https://doi.org/10.1534/genetics.107.086405

Suhre, K., & Claverie, J.-M. (2003). Genomic correlates of hyperthermostability, an update. *Journal of Biological Chemistry*, *278*(19), 17198–17202. https://doi.org/10.1074/jbc.M301327200

Tenesa, A., Navarro, P., Hayes, B. J., Duffy, D. L., Clarke, G. M., Goddard, M. E., & Visscher, P. M. (2007). Recent human effective population size estimated from linkage disequilibrium. *Genome Research*, *17*(4), 520–526. https://doi.org/10.1101/gr.6023607

Theillet, F.-X., Kalmar, L., Tompa, P., Han, K.-H., Selenko, P., Dunker, A. K., Daughdrill, G. W., & Uversky, V. N. (2013). The alphabet of intrinsic disorder. *Intrinsically Disordered Proteins*, *1*(1), e24360. https://doi.org/10.4161/idp.24360

Tien, M. Z., Meyer, A. G., Sydykova, D. K., Spielman, S. J., & Wilke, C. O. (2013). Maximum Allowed Solvent Accessibilites of Residues in Proteins. *PLOS ONE*, *8*(11), e80635. https://doi.org/10.1371/journal.pone.0080635

Tsai, J., Taylor, R., Chothia, C., & Gerstein, M. (1999a). The packing density in proteins: Standard radii and volumes. *Journal of Molecular Biology*, *290*(1), 253–266. https://doi.org/10.1006/jmbi.1999.2829

Tsai, J., Taylor, R., Chothia, C., & Gerstein, M. (1999b). The packing density in proteins: Standard radii and volumes11Edited by J. M. Thornton. *Journal of Molecular Biology*, *290*(1), 253–266. https://doi.org/10.1006/jmbi.1999.2829

Tsuboyama, K., Dauparas, J., Chen, J., Laine, E., Mohseni Behbahani, Y., Weinstein, J. J., Mangan, N. M., Ovchinnikov, S., & Rocklin, G. J. (2023). Mega-scale experimental analysis of protein folding stability in biology and design. *Nature*, *620*(7973), Article 7973. https://doi.org/10.1038/s41586-023-06328-6

van Noort, V., Bradatsch, B., Arumugam, M., Amlacher, S., Bange, G., Creevey, C., Falk, S., Mende, D. R., Sinning, I., Hurt, E., & Bork, P. (2013). Consistent mutational paths predict eukaryotic thermostability. *BMC Evolutionary Biology*, *13*(1), 7. https://doi.org/10.1186/1471-2148-13-7

Wang, G.-Z., & Lercher, M. J. (2010). Amino acid composition in endothermic vertebrates is biased in the same direction as in thermophilic prokaryotes. *BMC Evolutionary Biology*, *10*(1), 263. https://doi.org/10.1186/1471-2148-10-263

Weibel, C., Wheeler, A. L., James, J. E., Willis, S. M., & Masel, J. (2023). A new codon adaptation metric predicts vertebrate body size and tendency to protein disorder. *eLife*, *12*. https://doi.org/10.7554/eLife.87335

Whelan, S., & Goldman, N. (2001). A General Empirical Model of Protein Evolution Derived from Multiple Protein Families Using a Maximum-Likelihood Approach. *Molecular Biology and Evolution*, *18*(5), 691–699. https://doi.org/10.1093/oxfordjournals.molbev.a003851

Wilke, C. O., Bloom, J. D., Drummond, D. A., & Raval, A. (2005). Predicting the Tolerance of Proteins to Random Amino Acid Substitution. *Biophysical Journal*, *89*(6), 3714–3720. https://doi.org/10.1529/biophysj.105.062125

Wilson, B. A., Foy, S. G., Neme, R., & Masel, J. (2017). Young genes are highly disordered as predicted by the preadaptation hypothesis of de novo gene birth. *Nature Ecology & Evolution*, *1*(6), Article 6. https://doi.org/10.1038/s41559-017-0146

Wu, G., Bazer, F. W., Dai, Z., Li, D., Wang, J., & Wu, Z. (2014a). Amino Acid Nutrition in Animals: Protein Synthesis and Beyond. *Annual Review of Animal Biosciences*, *2*(1), 387–417. https://doi.org/10.1146/annurev-animal-022513-114113

Wu, G., Bazer, F. W., Dai, Z., Li, D., Wang, J., & Wu, Z. (2014b). Amino acid nutrition in animals: Protein synthesis and beyond. *Annual Review of Animal Biosciences*, *2*, 387–417. https://doi.org/10.1146/annurev-animal-022513-114113

Yates, A. D., Achuthan, P., Akanni, W., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., Azov, A. G., Bennett, R., Bhai, J., Billis, K., Boddu, S., Marugán, J. C., Cummins, C., Davidson, C., Dodiya, K., Fatima, R., Gall, A., … Flicek, P. (2020). Ensembl 2020. *Nucleic Acids Research*, *48*(D1), D682–D688. https://doi.org/10.1093/nar/gkz966

Zarin, T., Strome, B., Peng, G., Pritišanac, I., Forman-Kay, J. D., & Moses, A. M. (2021). Identifying molecular features that are associated with biological function of intrinsically disordered protein regions. *eLife*, *10*, e60220. https://doi.org/10.7554/eLife.60220

Zarin, T., Tsai, C. N., Nguyen Ba, A. N., & Moses, A. M. (2017). Selection maintains signaling function of a highly diverged intrinsically disordered region. *Proceedings of the National Academy of Sciences*, *114*(8), E1450–E1459. https://doi.org/10.1073/pnas.1614787114

Zuckerkandl, E., Derancourt, J., & Vogel, H. (1971). Mutational trends and random processes in the evolution of informational macromolecules. *Journal of Molecular Biology*, *59*(3), 473–490. https://doi.org/10.1016/0022-2836(71)90311-1