


# Molecular Evolution in Small Steps under Prevailing Negative Selection: A Nearly Universal Rule of Codon Substitution

Qingjian Chen <sup>1,†</sup>, Ao Lan<sup>1,†</sup>, Xu Shen<sup>1</sup>, and Chung-I Wu<sup>1,2,3,\*</sup>

<sup>1</sup>State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-sen University, Guangzhou, China

<sup>2</sup>CAS Key Laboratory of Genomic and Precision Medicine, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China

<sup>3</sup>Department of Ecology and Evolution, University of Chicago

<sup>†</sup>These authors contributed equally to this work.

\*Corresponding author: E-mail: ciwu@uchicago.edu.

Accepted: August 28, 2019

## Abstract

The widely accepted view that evolution proceeds in small steps is based on two premises: 1) negative selection acts strongly against large differences and 2) positive selection favors small-step changes. The two premises are not biologically connected and should be evaluated separately. We now extend a previous approach to studying codon evolution in the entire genome. Codon substitution rate is a function of the physicochemical distance between amino acids (AAs), equated with the step size of evolution. Between nine pairs of closely related species of plants, invertebrates, and vertebrates, the evolutionary rate is strongly and *negatively* correlated with a set of AA distances ( $\Delta_U$ , scaled to [0, 1]).  $\Delta_U$ , a composite measure of evolutionary rates across diverse taxa, is influenced by almost all of the 48 physicochemical properties used here. The new analyses reveal a crucial trend hidden from previous studies:  $\Delta_U$  is strongly correlated with the evolutionary rate ( $R^2 > 0.8$ ) only when the genes are predominantly under negative selection. Because most genes in most taxa are strongly constrained by negative selection,  $\Delta_U$  has indeed appeared to be a nearly universal measure of codon evolution. In conclusion, molecular evolution at the codon level generally takes small steps due to the prevailing negative selection. Whether positive selection may, or may not, follow the small-step rule is addressed in a companion study.

**Key words:** negative selection, amino acid distance, Fisher's Geometric Model, codon evolution.

## Introduction

Since the time of Darwin, biologists have accepted that evolution proceeds in small steps. An obvious explanation is mutational input: Large changes require many mutations and each only makes an incremental contribution. Approaching the issue from the angle of natural selection, R.A. Fisher formalized the selectionists' view of small-step evolution, known as Fisher's Geometric Model (FGM) (Fisher 1930). FGM uses the metaphor of climbing the adaptive peak in a multidimensional landscape and suggests that small changes are more likely to be advantageous than large ones. At the molecular level, the neutral theory also posits small-step evolution, which can be summarized by two rules (Kimura 1983). 1) Functionally less important molecules evolve faster than functionally important ones. 2) Variants that are functionally similar to the wildtype are more likely to be substituted than dissimilar ones. The two rules are mainly about escaping negative selection but, when applied to positive selection, would

converge with the FGM view. In this study, we focus on the second rule—coding-sequence evolution in small steps.

It is noteworthy that selectionists and neutralists appear to agree on small-step evolution, albeit with different emphases. In the neutralists' view, negative selection tolerates small-step changes while FGM postulates that positive selection favors small-step improvements. Because negative and positive selections are distinct forces driving different processes, small-step evolution can be considered two models in one.

In this study, the step size of evolution is represented by amino acid (AA) differences (or distances) (Grantham 1974; Dayhoff et al. 1978; Miyata et al. 1979; Henikoff and Henikoff 1992; Kumar et al. 2009; Adzhubei et al. 2013). One approach to AA distances attempts to identify physicochemical properties of AAs that can best explain long-term substitution patterns. Earlier methods by Grantham (Grantham 1974) and Miyata (Miyata et al. 1979) and the more recent ones

including SIFT (Kumar et al. 2009) and PolyPhen (Adzhubei et al. 2013) take this approach. The second approach attempts to identify substitution patterns directly from protein or DNA sequences and uses these evolutionary patterns as the proxy for AA distances. This second approach can be either AA based or codon based. For example, PAM (Dayhoff et al. 1978), LG (Le and Gascuel 2008), and BLOSUM (Henikoff and Henikoff 1992) are AA based, searching for long-term evolutionary patterns among the 190 (=20 × 19/2) pairwise comparisons. In contrast, the codon-based approach (Yang et al. 1998; Tang et al. 2004; Tang and Wu 2006) compares closely related species whose triplet codons differ by at most 1 bp. Among the 190 pairs, only 75 pairs can be exchanged by a 1-bp mutation. We shall take the codon-based approach to AA distances (Tang et al. 2004; Yang 2007).

The study, extending and revising the results of Tang et al. (2004) and Tang and Wu (2006), uses the tools developed therein (Tang et al. 2004; Tang and Wu 2006). The extensions are in three directions. First, in earlier studies, partial genomes are used. Given the partitions of AA changes into 75 kinds, the statistical resolution was barely adequate. These early releases of genomic data are also biased toward functionally important genes that bear distinct signatures of selection. Hence, whole-genome sequences obtained in the intervening years should be most useful. Second, it is important to densely sample within the same taxonomic rank (such as vertebrates, mammals, and primates) to test the consistency within the same phylum, class, or order. Third, and most important of all, previous studies did not separate the effects of positive and negative selection. For that reason, deviations from the general rule, potentially most informative about the working of the two opposing forces, have been ignored in previous studies.

With the AA distance as the step size of molecular evolution, the effects of negative and positive selection are separately analyzed in this and the companion report (Chen, He, et al. 2019). Here, we ask whether negative selection drives small-step evolution and whether there exist common rules at the codon level across a wide range of taxa.

## Materials and Methods

### Multiple Alignment Data

A multiple alignment file of 99 vertebrates with human for CDS regions was downloaded from the UCSC Genome Browser (<http://hgdownload.cse.ucsc.edu/goldenPath/hg38/multiz100-way/alignments/knownCanonical.exonNuc.fa.gz>). We then selected seven representative pairs of species across vertebrates (supplementary table S1, Supplementary Material online). There were two pairs of species from the Primates order, one was from the family *Hominidae* (*Homo sapiens* and *Pan troglodytes*) and the other from the family *Cercopithecidae* (*Macaca fascicularis* and *Callithrix jacchus*). The other five pairs of species were from the order *Rodentia* (*Mus musculus* and *Rattus norvegicus*), the order *Carnivora* (*Felis catus* and *Canis lupus*

*familiaris*), the order *Artiodactyla* (*Bos taurus* and *Ovis aries*), the class *Aves* (*Geospiza fortis* and *Taeniopygia guttata*), and the class *Reptilia* (*Chelonia mydas* and *Chrysemys picta bellii*). Additional technical details used here can also be found in Lin et al. (2018).

A multiple alignment file of 26 insects with *Drosophila melanogaster* was also downloaded from the UCSC Genome Browser (<http://hgdownload.cse.ucsc.edu/goldenPath/dm6/multiz27way/alignments/refGene.exonNuc.fa.gz>). The genome sequences of *D. melanogaster* and *Drosophila simulans* were then extracted from this file.

To correct for multiple hits for Ki/Ks, two closely related species of each pair were chosen from the same family. The alignment files were filtered using the following criteria: 1) Filter out genes in uncanonical scaffolds. 2) Use the corresponding AA multiple alignment file to remove noncoding genes. 3) Filter redundant transcripts generated by alternative splicing, based on a gene-transcript id conversion table downloaded from Ensembl biomart. 4) Genes aligned with too many gaps (more than 5%) were removed. 5) Genes shorter than 90 bp were removed. 6) Substitutions related to CpG sites in the Homo–Pan pair were masked because CpG-related substitutions account for more than 30% of total substitutions in this pair. Hence, we masked CpG-related substitutions (CG => TG, CG => CA were replaced with NG => NG and CN => CN, respectively) using gorilla as an outgroup (supplementary fig. S3 and supplementary text III, Supplementary Material online).

We first obtained protein and DNA sequences of *Arabidopsis thaliana* and *Arabidopsis lyrata* from the Phytozome database (Goodstein et al. 2012). We then extracted gene alignments using the PAL2NAL program (Suyama et al. 2006) with the mRNA sequences and the corresponding protein sequences for 1:1 orthologous pairs aligned by MUSCLE (Edgar 2004) using default parameters.

### Ki/Ks Calculation

We used the subprogram *codeml* (codon-based maximum likelihood) from PAML to calculate each Ki/Ks (Yang 2007). The *codeml* considers codons as units of evolution. The substitution from codon *u* to codon *v* is

$$q_{uv} = \begin{cases} 0 & \text{if the two codons differ at more than one position} \\ \pi_v & \text{for synonymous transversion} \\ \kappa\pi_v & \text{for synonymous transition} \\ \omega_{ij}\pi_v & \text{for nonsynonymous transversion} \\ \omega_{ij}\kappa\pi_v & \text{for nonsynonymous transition} \end{cases}$$

Here,  $q_{uv}$  is the instantaneous rate from codon *u* to codon *v*.  $\kappa$  is the transition/transversion rate ratio,  $\pi_v$  is the equilibrium frequency of codon *v*.  $\omega_{ij}$  is the nonsynonymous/synonymous

rate ratio, where  $i$  ( $=aa_u$ ) and  $j$  ( $=aa_v$ ) are the two AAs involved. There are 75  $\omega_{ij}$ , analogous to  $Ki/Ks$  ( $i = 1:75$ ), of AA pairs whose codons differed by 1 bp. The method considering unequal substitution rate between different AAs has been summarized in Yang et al. (1998).

The controls in *codeml* are as follows:

1. *codenFreq* = 0. Assuming individual codon frequencies are equal. There are four codon frequencies models in common use, including F-equal, F1×4, F3×4, and F61 (see p. 48 of Yang [2006]). We opt for the simplest F-equal and discuss the influence of codon frequency models to the estimated  $Ki/Ks$  (see [supplementary figs. S5 and S6](#) and [supplementary text VI and VII, Supplementary Material online](#)).
2. *Model* = 0. Assuming the  $\omega_{ij}$  is constant in different braches.
3. *NSsites* = 0. Use the M0 (one ratio) model to estimate  $\omega$  for site models. Disregard the variation in  $\omega$  among sites.
4. *aaDist* = 7. Divide AA substitutions into several groups and estimate their  $\omega_{ij}$  separately. The AA groups are summarized in a file called OmegaAA.dat. The setting in OmegaAA.dat is as following:

```
" -1
//end of file.
"
```

Here, putting  $-1$  at the start of file, then the program will fit the "general model," assigning an independent  $\omega$  for each one-step AA pairs, which corresponding 75  $Ki/Ks$ . In the same time, the overall  $\omega$  ( $Ka/Ks$ ) is also given in the output. See the [supplementary table S4, Supplementary Material online](#), for the control file in *codeml*.

In the end, expected  $Ki/Ks$  values were calculated as  $E(Ki/Ks) = U_i \times Ka/Ks$  in each pair of species.

### Analysis of AA Distances

AA properties were extracted from table 2 of Gromiha et al. (1999). In total, 48 selected physicochemical, energetic, and conformational properties are given. The distances between AA pairs were defined by the differences in the raw values shown in their table 2. Various distance measures yielded similar results. The 48 distances of each AA pair were then scaled by the z-score method for the principal component analysis (PCA) and partial least square regression (PLSR) analysis. PCA and PLSR were performed using the R packages "factoextra" (Kassambara and Mundt 2017) and "pls" (Mevik et al. 2019), respectively. Exponential fitting was accomplished by using the function "nls" in R.

### The Relationship between $U_i$ and $Ka/Ks$

Orthologous genes for each paired species were divided into five categories with approximately equal number of nonsynonymous changes according to their  $Ka/Ks$  ranking. The  $Ka/Ks$

of each gene is calculated by *codeml* in PAML. Number of nonsynonymous changes of each gene is obtained by comparing the divergent sites between two species using our own codes. Genes with fewer than two substitutions were removed for further analysis. It is expected that  $Ka$  values increase when  $Ka/Ks$  values ascend and  $Ks$  values are stable in the five categories. This is true in most of cases. However, the top 20% group in Homo–Pan is an outlier, exhibiting a sharp decrease in  $Ks$ . This may due to the short divergence time between Homo and Pan. The  $Ks$  values of the five categories in the Homo–Pan comparison were replaced by genome-wide  $Ks$ . The Mus–Rattus comparison was removed in [figure 5](#) because  $U_i$  values were deduced from the comparison between Mus–Rattus and a pair of yeast (Tang et al. 2004). The 95% confidence intervals for  $R^2$  are obtained by bootstrapping.

## Results

By convention, the number of nonsynonymous changes per nonsynonymous site is designated  $Ka$  (Kimura 1983; Li et al. 1985) (or  $dN$  [Nei and Gojobori 1986]) and the corresponding number for synonymous changes is  $Ks$ .  $Ka$  is thus the aggregate measure of all nonsynonymous changes and can be decomposed into 75 classes of AA substitutions whose codons differ by only 1 bp. Each of these classes is labeled  $Ki$ ,  $i = 1, 75$ . Hence,

$$\frac{Ka}{Ks} = \sum_{i=1}^{75} w_i \frac{Ki}{Ks}, \quad (1)$$

where  $w_i$  is the weight reflecting the number of sites available for AA exchanges of the  $i$ th pair.  $Ka/Ks$  is hence the aggregate of 75  $Ki/Ks$ . Here, we implement a codon-based maximum-likelihood (*codeml*) method of the PAML packages to calculate each  $Ki/Ks$  (see Materials and Methods) (Yang et al. 1998; Yang 2007). The *codeml* algorithm is generally accepted as statistically rigorous and has been widely used in lieu of the counting methods with multiple-hits corrections (Li et al. 1985; Nei and Gojobori 1986; Li 1993; Comeron 1995; Ina 1995).

There have been a long series of studies that aim to group AA changes into classes, starting with Zuckerkandl and Pauling (1965), followed by Grantham (1974) and Miyata et al. (1979) (see [supplementary text I, Supplementary Material online](#)). In later studies, the classification is often along the line of radical versus conservative changes (Kr vs. Kc) (Smith 2003; Nabholz et al. 2013; Weber et al. 2014; Figuet et al. 2016). Because these classifications involve assigning AA changes into classes, the outcome has been reported to depend on the classification (Dagan et al. 2002; Hanada et al. 2007) (see [supplementary text II, Supplementary Material online](#), for details). The 75  $Ki/Ks$  classes, in contrast, involve no grouping as each pair of  $AA_i$  and  $AA_j$  is an elementary class. In the main text, we will show

that 1) Ki/Ks's are nearly universally correlated across taxa and 2) this high correlation reflects the physicochemical properties of AAs.

### The Correlation among Ki/Ks's across Taxa

We first calculate Ki/Ks between nine pairs of species using their whole-genome sequences (fig. 1A). Among them are one pair from plants (*Arabidopsis*), one pair of insects (*Drosophila*), and seven pairs of vertebrates with rodents (*Mus–Rattus*) representing the vertebrates. Each pair consists of two closely related species. Genome-wide Ka/Ks ratios of these nine pairs ranges from 0.12 to 0.27 (supplementary table S1, Supplementary Material online).

We then calculate the correlation in Ki/Ks between two pairs of distantly related taxa (fig. 1A). Pairwise correlations between *Arabidopsis* (*A. thaliana* vs. *A. lyrata*), *Drosophila* (*D. melanogaster* vs. *D. simulans*), and vertebrates (*Mus* vs. *Rattus*) are shown in figure 1B–D. The  $R^2$  values are 0.81 (vertebrates vs. insects), 0.84 (vertebrates vs. plants), and 0.90 (insects vs. plants). Such strong correlations over a large phylogenetic span suggest that AA substitutions follow a nearly universal rule. Although Ki/Ks within each taxon may be very different, their relative magnitudes remain the same across all taxa. The correlations of figure 1B–D are close to the results obtained by Tang et al. (2004), thus permitting further analyses based on the earlier results.

We now examine the correlation in a dense phylogenetic framework by comparing species pairs of vertebrate classes or mammalian orders (fig. 1E). Although one might have expected the correlation to be even greater in the lower taxonomic rank, the observations of figure 1E suggest otherwise. In fact, the  $R^2$  values are not strongly dependent on the phylogenetic distance. For example,  $R^2$  in the *Drosophila–Arabidopsis* comparison is higher than many of the 21 comparisons between vertebrates. In particular, the AA substitutions between hominoids and other vertebrates often yield  $R^2 < 0.8$ . It seems plausible that similar forces work reiteratively from taxa to taxa, yielding a degree of consistency across the phylogeny.

The high correlation among Ki/Ks values permits a generalized (or universal) Ki/Ks measure as proposed before (Tang et al. 2004). Here, we shall briefly introduce the measure, referred to as  $U_i$  (Tang et al. 2004). It ranges between 0.25 and 2.5 for  $i = 1–75$  and is the relative AA substitutions rate between 75 AAs.  $U_i$  is scaled such that the weighted mean is 1 across the 75 classes. For example,  $U_i = 2.5$  represents substitutions rate for this AA is 2.5 times higher than average substitutions rate.

For any species, the expected Ki/Ks is

$$E(\text{Ki/Ks}) = U_i \times \mathbf{Ka/Ks}, \quad (2)$$

where the boldface **Ka** and **Ks** denote the species' genome-wide value. As long as we know **Ka/Ks**,  $E(\text{Ki/Ks})$  can be

calculated for  $i = 1, 75$  by equation (2). Interestingly, the updated  $U_i$  in this expanded study is highly correlated with the  $U_i$  of Tang et al. (2004) with  $R = 0.97$  (supplementary fig. S1, Supplementary Material online). Given such a high correlation, we shall continue to use the published  $U_i$  scale. In doing so, the extensive published literature will remain comparable with future analyses.

### Ki/Ks in Relation to AA Distances

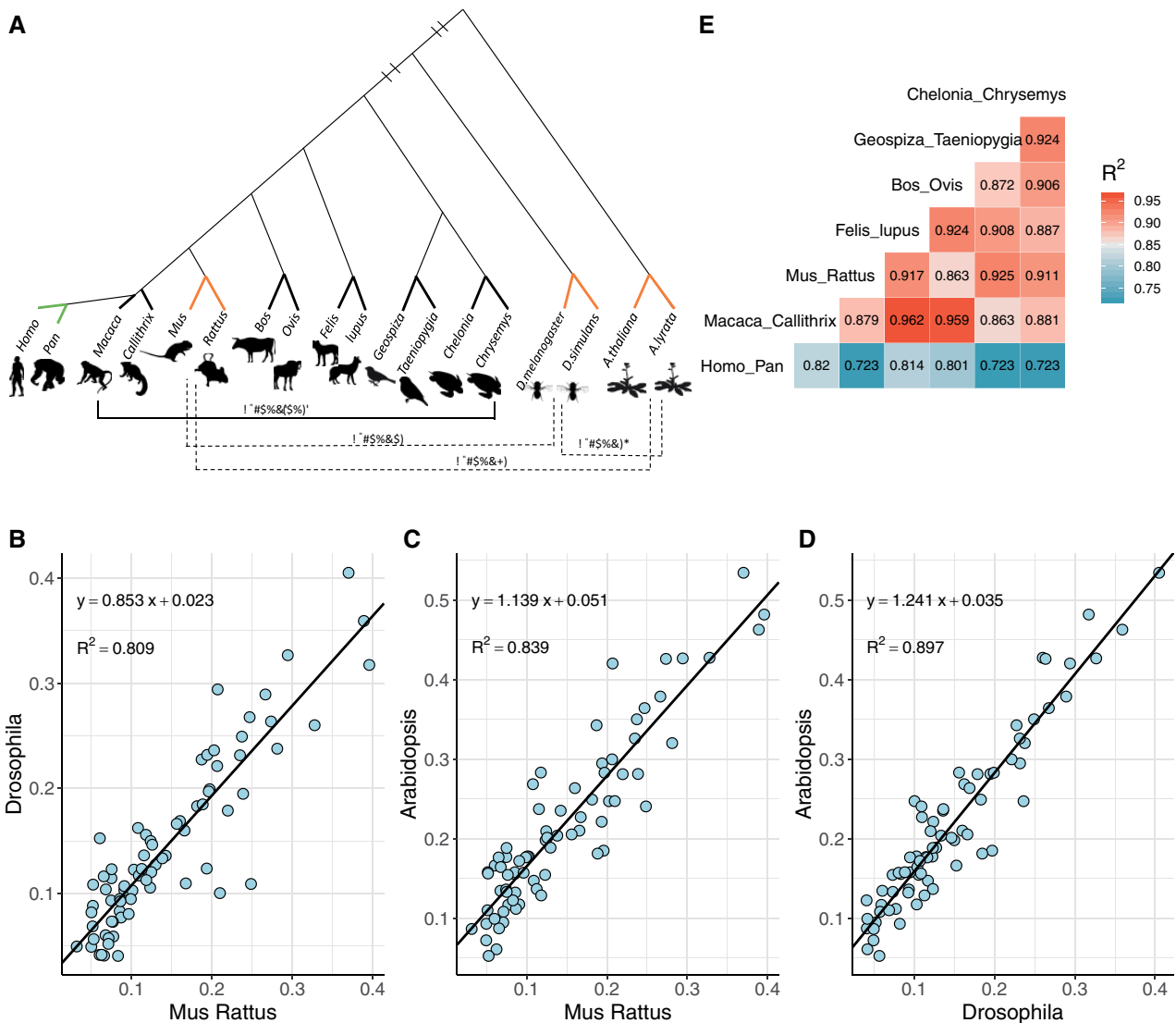
The high correlation across a large phylogenetic distance (fig. 1) suggests something as basic as physicochemical properties of AAs to be a major cause. It has been tantalizing to ask whether a small number of differences in these properties (AA distances, for short) might account for the evolution of protein sequences, thus reducing evolution to the simplest physical dimensions (Grantham 1974; Miyata et al. 1979).

We analyze 48 physicochemical, energetic, and conformational distances among AAs (fig. 2). As shown in supplementary table S2, Supplementary Material online, these properties are not all about each AA in isolation; many measurements (e.g., number of surrounding residues) depend on the context of protein sequences. These measurements are broadly depicted as “physicochemical,” as opposed to “functional.” We use  $U_i$  ( $i = 1, 75$  [Tang et al. 2004]) to represent Ki/Ks across species primarily given that  $U_i$  is highly correlated with each species' Ki/Ks (see fig. 4 below).

$U_i$  is fitted to each of the 48 AA distances via an exponential function (see Kimura 1983). Figure 2A–C shows regressions of  $U_i$  on AA distances of hydrophobicity, total nonbounded energy, and volume. Although hydrophobicity is generally thought to be an important attribute, the correlation is weak ( $R^2 = 0.09$ ). Instead, the volume difference of AA explains more of the  $U_i$  variation with  $R^2 = 0.27$ , consistent with the prior conclusion on the relative importance of AA's volume (Yang et al. 1998; Braun 2018). The spearman's correlation coefficient for each AA distance is given in figure 2D, which shows that 47 of the 48 measures are negatively correlated with  $U_i$  and  $P < 0.05$  is found for 39 of them (supplementary table S2, Supplementary Material online). Given the correlation with the evolutionary rate is nearly entirely negative for AA distances, the fitness effect is not likely to be narrowly distributed among a few AA distances.

Measures of these 48 AA properties are not randomly associated with the 75 AA pairs. Each pairwise comparison between the 48 measures is plotted for the 75 AA pairs to extract the correlation between these distances, designated  $R_d$ . The heatmap of all pairwise  $R_d$ 's is given in figure 3 where the inset shows one of the largest  $R_d$ 's between volume (property 4 of supplementary table S2, Supplementary Material online) and helical contact area (property 9).  $R_d = 0.94$  for this pair indicates that these two distances probably reflect two very similar AA characters. Along the diagonal, some clustering can be discerned but the correlation is



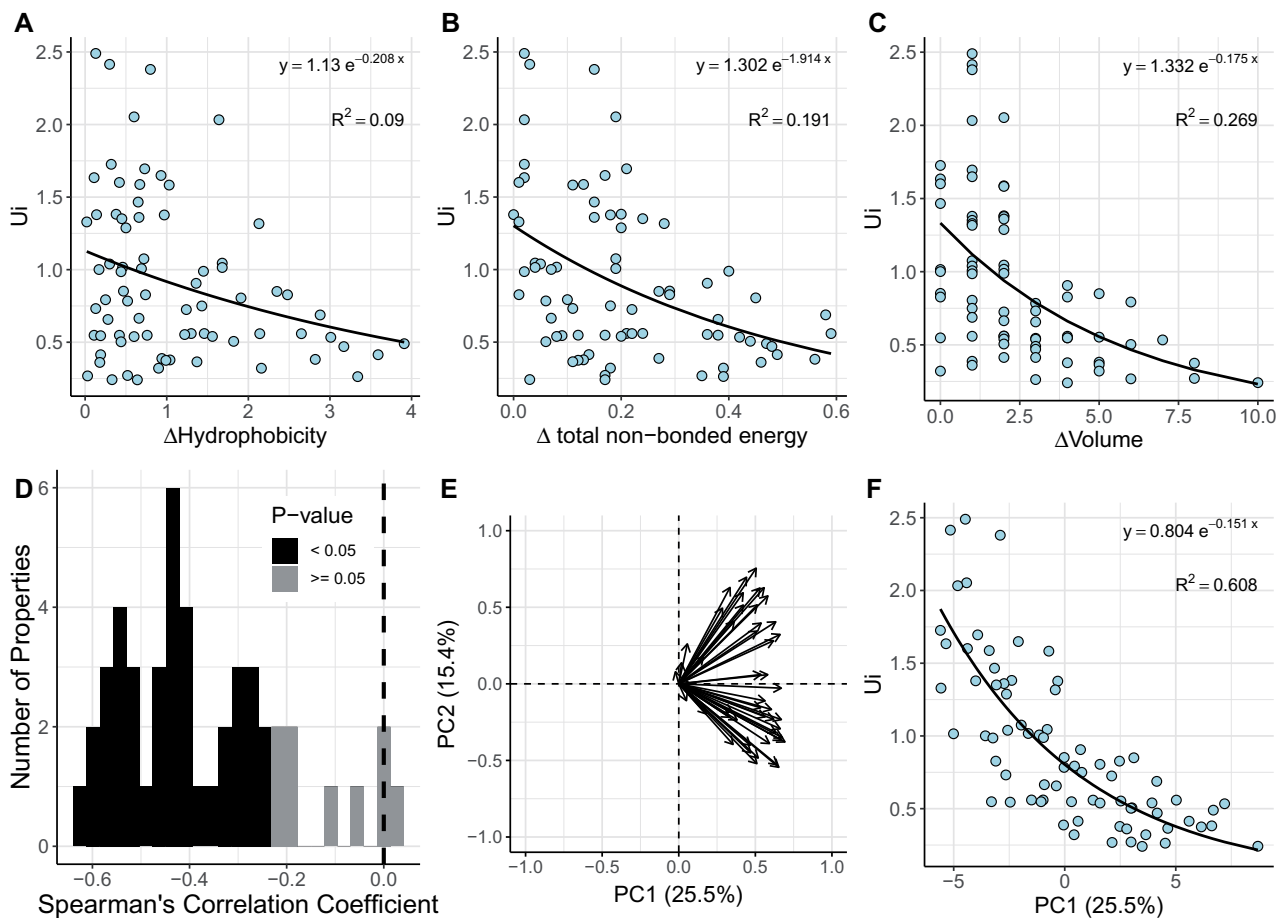


**FIG. 1.**—Correlation of Ki/Ks values between different taxa. (A) Phylogenetic Tree of 9 representative species pairs (full names are shown in [supplementary table S1, Supplementary Material](#) online).  $R^2$  values (pairwise correlation of Ki/Ks's) range from 0.86 to 0.96 in 6 vertebrates (horizontal solid line), except for *Homo–Pan* pairs. Three species pairs, *Rodent*, *Drosophila*, and *Arabidopsis* (bold orange lines), are selected to show species pairs with long evolutionary distance. (B–D) Scatter plots of Ki/Ks between *Rodent*, *Drosophila*, and *Arabidopsis*. The solid black lines are linear regression lines and the legends on the top left are the regression formulas and their  $R^2$ . (E) Pairwise  $R^2$  values of Ki/Ks among seven species pairs in vertebrates.

generally weak. There are nevertheless a few high-Rd pairs, shown as red squares along the diagonal. A subset of such red squares involving the top 13 AA distances of [supplementary table S2, Supplementary Material](#) online (P1–P13, with  $R > 0.5$ ) are most relevant. They form six clusters which are identifiable based on their Rd's—[P1, P11, P13], [P2, P8], P3, [P4, P7, P9, P10], [P5, P6], P12, where properties in brackets belong in the same cluster (see [fig. 3](#) and [supplementary table S2, Supplementary Material](#) online). Thus, the 48 properties may behave like half the number of uncorrelated

measures. It would then be more rigorous to do the PCA and PLSR analyses as done below.

By the PCA, we show that PC1 explains about 61% of the Ui variation ([fig. 2F](#)). [Figure 2E](#) further shows that the contributions to PC1 are distributed broadly among the 48 measures with “helical contact area” in the lead, contributing about 4%. As PC1 accounts for only 25.5% of the variance of AA distances, the remaining variance should also be informative. This can be seen when the PLSR is used. In PLSR, the first component accounts for 60.1% of the Ui variation using



**FIG. 2.**—The relationship between AA distances and  $U_i$ . Forty-eight physicochemical, energy, and conformational properties are used in the analysis. (A–C) Scatter plot of three selected AA distances against  $U_i$ , including hydrophobicity (A), total nonbonded energy (B), and volume (C). (D) Spearman's correlation coefficients between 48 AA distances and  $U_i$ . Among those distances, 47 distances are negatively correlated with  $U_i$ , where 39 of them are significantly (black,  $P$  value < 0.05). (E) Contribution of 48 AA distances to the first 2 principal components. Properties cluster into two categories in the second principal component but are indistinguishable in the first principal component. (F) Scatter plot of the first principal component (PC1) against  $U_i$ .

25.3% of the AA distance variance. The cumulative effect of the first five components accounts for 80% of the  $U_i$  variation, using <60% of the AA distance variance (supplementary fig. S1, Supplementary Material online). The results from PCA and PLSR are highly compatible.

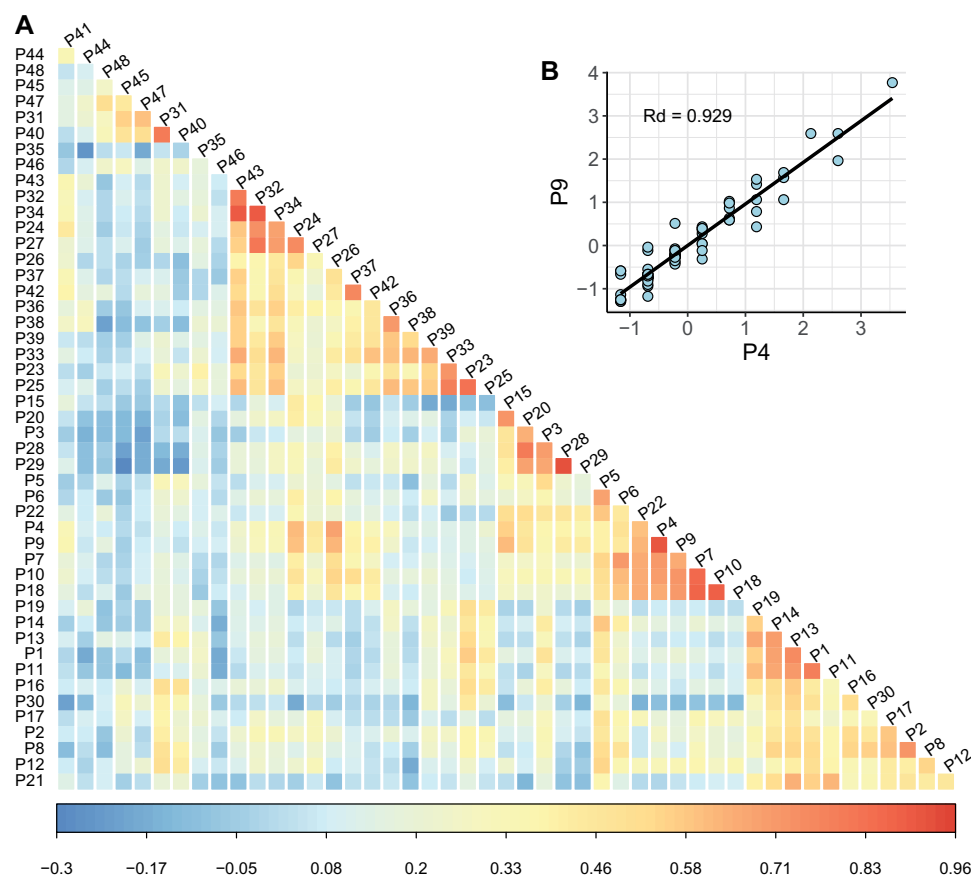
The overall results of figures 2 and 3 suggest that the physicochemical properties of AAs, or any combinations of such properties, are not likely to be fully predictive of the evolutionary rates of protein sequences. In the early days, a small number of such properties were found to capture a moderate amount of evolutionary rate variation (Miyata [Miyata et al. 1979] and Grantham [Grantham 1974]). It is natural to be hopeful that additional AA properties coupled with extensive DNA sequencing might be fully predictive of the evolutionary rate of protein sequences. The study shows that the optimism cannot be realized. Nevertheless, it is still possible to predict the evolutionary rate of most taxa because the AA substitution patterns are generalizable across diverse taxonomic ranks. Figures 2 and 3 show that such

general rules exist but they cannot be expressed in simple physicochemical terms. The evolutionary  $U_i$  measure remains the best predictor of the rate of AA substitutions when compared with Miyata and Grantham distances (see supplementary table S3 and supplementary text V, Supplementary Material online).

### The General Rule for Small-Step Evolution Expressed as $\Delta U(i)$

Given  $U_i$ , the AA distance of the  $i$ th pair can be more conveniently rescaled as  $\Delta_U(i) = (U_1 - U_i)/(U_1 - U_{75})$ , which falls in the range of [0, 1] with  $\Delta_U(1) = 0$  for the closest pair, [Ser-Thr], and  $\Delta_U(75) = 1$  for the most distant pair, [Asp-Tyr].

The observed versus expected  $K_i/K_s$  for the nine pairs of species are shown in figure 4. The x axis is  $E(K_i/K_s)$ , which can be written as  $[2.5 - 2.25 \Delta_U(i)] \times K_a/K_s$  (see the legend of figure 4 and supplementary table S1, Supplementary Material



**Fig. 3.**—The pairwise correlation between 48 amino acid distances. (A) The heatmap of pairwise correlation between different amino acid measurements. P1–P48 are the AA measurements listed in [supplementary table S2, Supplementary Material](#) online. The color bar represents the spearman's correlation coefficients (Rd) between AA distances. (B) The inset figure above the diagonal shows the scatter plot (each dot being an AA pair) between the volume measurement (P4) and the helical contact area (P9) of each AA pair. This particular Rd is the second largest among all.

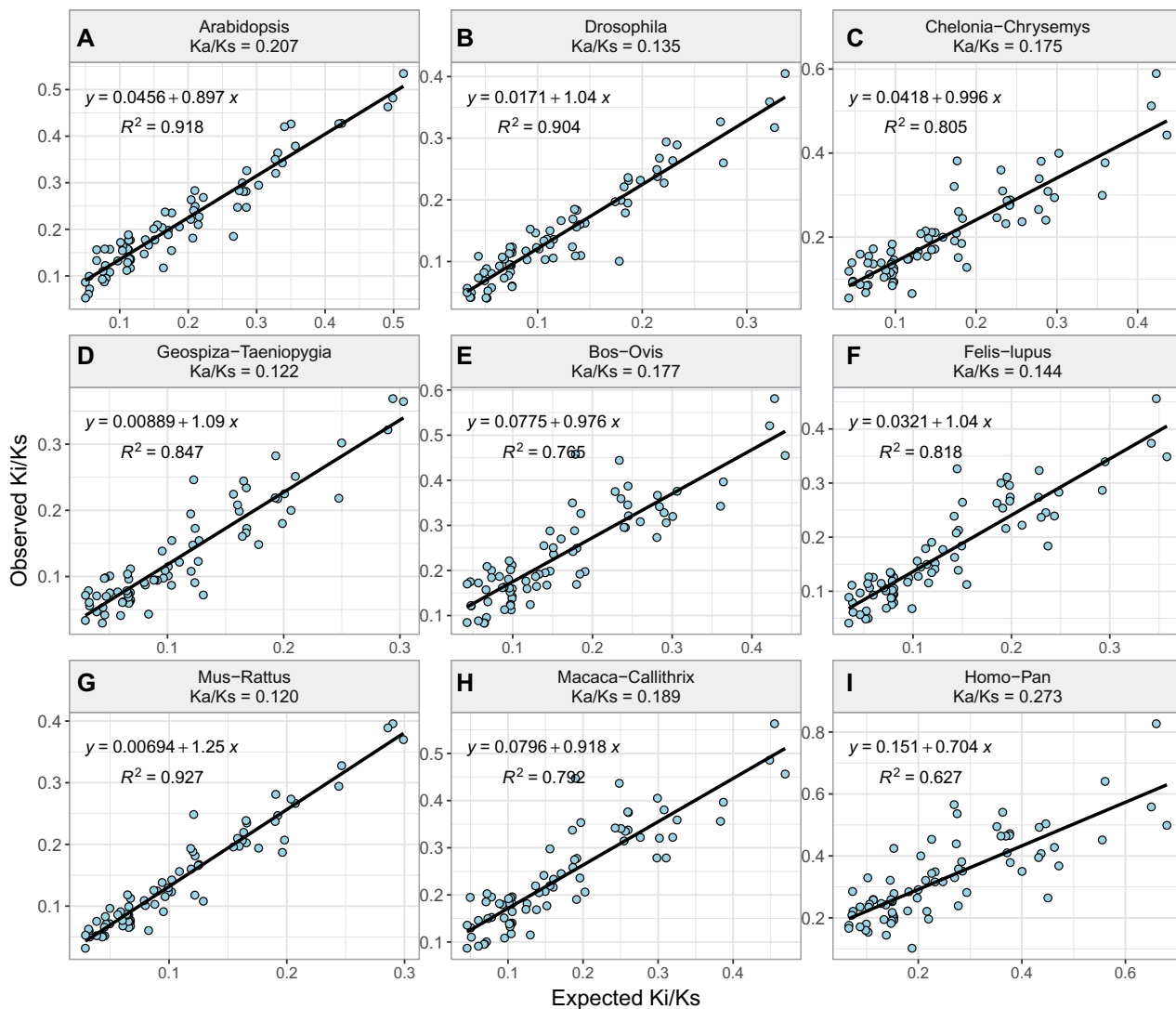
online, for details). There are three groups in this figure. The first group consists of three pairs of well-assembled genomes—*Arabidopsis*, *Drosophila*, and *Rodents* (fig. 4A, B, and G). All three taxa show  $R^2 > 0.9$  between Obs(Ki/Ks) and E(Ki/Ks). It is notable that the  $R^2$  improves significantly in *Drosophila*, rising from 0.706 to 0.904 in comparison with Tang et al. (2004), as the number of genes increases from 309 to 9,710. In the second group of five pairs of genomes of moderate quality (Fig. 4C–F, H),  $R^2$  ranges between 0.77 and 0.85. Taking into account the room for improving their quality, we conclude that these species follow the nearly universal pattern of codon substitutions.

The third group consists of the one exceptional case of human–chimpanzee comparison (Fig. 4I), which yields an  $R^2$  of 0.627, far lower than the rest. There may be several explanations for this unusually low  $R^2$ . Because codon substitutions involving CpG sites have been removed from consideration (see Materials and Methods), this obvious explanation is ruled out (see [supplementary fig. S3](#) and [supplementary text III, Supplementary Material](#) online). The second explanation is the unusual selective pressure in the human lineage

(Bustamante et al. 2005; Williamson et al. 2005; Subramanian 2011). However, because  $R^2$  along the human, chimpanzee, and gorilla lineages is, respectively, 0.572, 0.638, and 0.646 ([supplementary fig. S4](#) and [supplementary text IV, Supplementary Material](#) online), the human lineage does not stand out in this respect. Instead, we will show that hominoids as a group are unusual and many genes in their genomes have a high Ka/Ks ratio  $> 0.6$ .

### The Rule of Small-Step Evolution Is Governed by Strong Negative Selection

Between each pair of species, we divide the genomes into five bins in the ascending order of Ka/Ks (see Materials and Methods). It seems intuitively true that  $R^2$  would decrease as Ka/Ks increases. Imagine that, when Ka/Ks = 1 and under no negative selection, all observed Ki/Ks ratios would fluctuate around 1 and the correlation between the observed and expected Ki/Ks would be 0. Chen, He, et al. (2019) extend this intuition by an analytical model that incorporates negative



**FIG. 4.**—Correlations of expected and observed  $K_i/K_s$  for nine species pairs. (A–I) Scatter plots between expected and observed  $K_i/K_s$ . Because  $U_i = U_1 - (U_1 - U_{75}) \Delta_U(i)$ , the x axis label can be written as  $E(K_i/K_s) = U_i \times Ka/Ks = [2.5 - 2.25 \Delta_U(i)] \times Ka/Ks$ , which is a linear function of  $\Delta_U(i)$ . The boldface  $Ka/Ks$ , the genome-wide  $Ka/Ks$ , is the characteristic of each pair of species.

selection of variable strength. A decrease in  $R^2$  when  $Ka/Ks$  increases is confirmed by the model (see their fig. 1C).

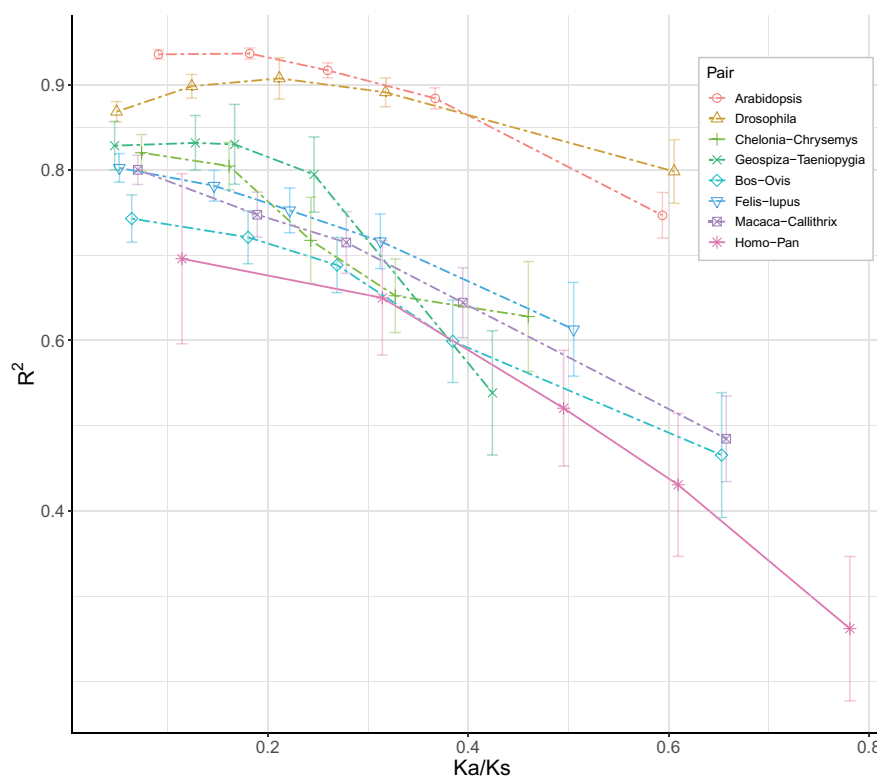
The  $R^2$  value (i.e., between Obs( $K_i/K_s$ ) versus E( $K_i/K_s$ )) of each bin is plotted against the  $Ka/Ks$  value of that bin (fig. 5). There are two decreasing trends. First,  $R^2$  is generally  $>0.6$  for gene groups with  $Ka/Ks < 0.4$ . The drop in  $R^2$  become steeper as  $Ka/Ks$  gets larger than 0.4. This important trend appears to escape notice in Tang et al. (2004). Second, the human–chimpanzee comparison follows the same pattern but generally at a lower level. Two opposing forces drive the trend of figure 5 with increasing  $Ka/Ks$ : weaker negative selection and/or stronger positive selection. Separating the two effects is the subject of Chen, He, et al. (2019). Nevertheless, when  $Ka/Ks < 0.4$  and negative selection dominates, we can draw the following conclusion: The evolution of genes under strong

negative selection takes small steps and follows a nearly universal rule. This rule is governed by the (broad sense) physicochemical properties of AAs.

### Discussion

The average pattern of molecular evolution is driven predominantly by negative selection. In most genes of most taxa,  $Ka/Ks < 0.3$ , which means the elimination of more than 70% of nonsynonymous mutations. The prevailing negative selection has led to a strikingly simple pattern: Codon evolution takes small steps and follows a nearly universal rule. By this rule, when an exchange between AA1 and AA2 is five times more likely than that between AA3 and AA4 in mammals, the same ratio would be preserved in nonmammalian vertebrates,





**FIG. 5.**—Relationship between  $R^2$  values (Obs(Ki/Ks) vs. E(Ki/Ks)) and selection strength (Ka/Ks). Orthologous genes in each pair of species are divided into five categories with equal nonsynonymous changes, according to their Ka/Ks ratio. The x axis is the average Ka/Ks ratio of each category in each species pair and the y axis is the  $R^2$  value (squared correlation coefficient) of their expected against observed Ki/Ks. The error bars represent the 95% confidence intervals for  $R^2$ .

invertebrates, and plants. The relative magnitude remains nearly constant.

Given the general pattern in such diverse taxa, negative selection at the codon level must be operating at a basic level of biochemistry (Weber and Whelan 2019). We show that the working of negative selection depends on the AA distances, almost all of which contribute to the fitness differences. Even the most obvious properties like hydrophobicity and non-bonded energy contribute only a small fraction to the overall evolutionary rate. Hence, the previous optimism that a small subset of AA distances may explain much of the variation in AA substitutions (Grantham 1974; Miyata et al. 1979) may be untenable.

The simple evolutionary pattern associated with the complex biochemistry provides an important lesson on the physicochemical basis of traits and diseases. There have been many proposals for measuring AA distance as an index of fitness difference (Grantham 1974; Dayhoff et al. 1978; Miyata et al. 1979; Henikoff and Henikoff 1992). Because matrices relying on a few biochemical properties are not likely to capture much of the evolutionary pattern (see fig. 2), it may be more informative to assess the evolutionary rate directly from DNA sequence evolution. In this perspective, the  $\Delta_U$

measure of codon substitution should be particularly suited to that task (supplementary table S3 and supplementary text V, Supplementary Material online).

In comparison with previous studies that link AA properties with molecular evolution at the codon level (Zuckerandl and Pauling 1965; Epstein 1967; Clarke 1970; Grantham 1974; Miyata et al. 1979; Kimura 1983; Tang et al. 2004), this study shows clearly the need to separate the effects of negative and positive selection. Indeed, some previous studies have reported unusual AA substitution patterns in extreme environments or under domestication, where positive selection could be prevalent (Lu et al. 2006; Luo et al. 2017; Xu et al. 2017; Chen, Shi, et al. 2019; He et al. 2019; Wang et al. 2017; Wen et al. 2018). Although the working of negative selection is, to some extent, predictable, positive selection may show very different patterns, which is addressed in the accompanying study (Chen, He, et al. 2019).

### Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

## Acknowledgments

We would like to thank Ziwen He, Haijun Wen, Hao Yang, Qipian Chen, and members of Wu Lab for discussions and advices. This work was supported by the [National Natural Science Foundation of China \(31730046 and 91731301\)](#) and the [985 Project \(33000-18841204\)](#).

## Literature Cited

- Adzhubei I, Jordan DM, Sunyaev SR. 2013. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet*. Chapter 7:Unit7.20.
- Braun EL. 2018. An evolutionary model motivated by physicochemical properties of amino acids reveals variation among proteins. *Bioinformatics* 34(13):i350–i356.
- Bustamante CD, et al. 2005. Natural selection on protein-coding genes in the human genome. *Nature* 437(7062):1153–1157.
- Chen B, Shi Z, et al. 2019. Tumorigenesis as the paradigm of quasi-neutral molecular evolution. *Mol Biol Evol*. 36(7):1430–1441.
- Chen Q, He Z, et al. 2019. Molecular evolution in large steps—codon substitutions under positive selection. *Mol Biol Evol*. 36(9):1862–1873.
- Clarke B. 1970. Selective constraints on amino-acid substitutions during the evolution of proteins. *Nature* 228(5267):159–160.
- Comeron JM. 1995. A method for estimating the numbers of synonymous and nonsynonymous substitutions per site. *J Mol Evol*. 41:1152–1159.
- Dagan T, Talmor Y, Graur D. 2002. 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*. 19(7):1022–1025.
- Dayhoff M, Schwartz R, Orcutt B. 1978. A model of evolutionary change in proteins. In: Dayhoff M, editor. *Atlas of protein sequence and structure*. Silver Spring (MD): National Biomedical Research Foundation. p. 345–352.
- Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 32(5):1792–1797.
- Epstein CJ. 1967. Non-randomness of amino-acid changes in the evolution of homologous proteins. *Nature* 215(5099):355.
- Figuat E, et al. 2016. Life history traits, protein evolution, and the nearly neutral theory in amniotes. *Mol Biol Evol*. 33(6):1517–1527.
- Fisher RA. 1930. *The genetical theory of natural selection*. Oxford: The Clarendon press. p. 22–47.
- Goodstein DM, et al. 2012. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res*. 40(Database issue):D1178–D1186.
- Grantham R. 1974. Amino acid difference formula to help explain protein evolution. *Science* 185(4154):862–864.
- Gromiha MM, Oobatake M, Sarai A. 1999. Important amino acid properties for enhanced thermostability from mesophilic to thermophilic proteins. *Biophys Chem*. 82(1):51–67.
- Hanada K, Shiu SH, Li WH. 2007. The nonsynonymous/synonymous substitution rate ratio versus the radical/conservative replacement rate ratio in the evolution of mammalian genes. *Mol Biol Evol*. 24(10):2235–2241.
- He Z, et al. 2019. Speciation with gene flow via cycles of isolation and migration: insights from multiple mangrove taxa. *Natl Sci Rev*. 6(2):275–288.
- Henikoff S, Henikoff JG. 1992. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*. 89(22):10915–10919.
- Ina Y. 1995. New methods for estimating the numbers of synonymous and nonsynonymous substitutions. *J Mol Evol*. 40(2):190–226.
- Kassambara A, Mundt F. 2017. Factoextra: extract and visualize the results of multivariate data analyses. R package version. 1(5).
- Kimura M. 1983. *The neutral theory of molecular evolution*. Cambridge: Cambridge University Press. p. 103–103.
- Kumar P, Henikoff S, Ng PC. 2009. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc*. 4(7):1073–1081.
- Le SQ, Gascuel O. 2008. An improved general amino acid replacement matrix. *Mol Biol Evol*. 25(7):1307–1320.
- Li WH. 1993. Unbiased estimation of the rates of synonymous and non-synonymous substitution. *J Mol Evol*. 36(1):96–99.
- Li WH, Wu CI, Luo CC. 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol Biol Evol*. 2(2):150–174.
- Lin T, et al. 2018. Genome analysis of *Taraxacum kok-saghyz* Rodin provides new insights into rubber biosynthesis. *Natl Sci Rev*. 5(1):78–87.
- Lu J, et al. 2006. The accumulation of deleterious mutations in rice genomes: a hypothesis on the cost of domestication. *Trends Genet*. 22(3):126–131.
- Luo H, Huang Y, Stepanauskas R, Tang J. 2017. Excess of non-conservative amino acid changes in marine bacterioplankton lineages with reduced genomes. *Nat Microbiol*. 2:17091.
- Mevik B-H, Wehrens R, Liland KH. 2019. Pls: Partial Least Squares and Principal Component Regression. R package version. 2(7).
- Miyata T, Miyazawa S, Yasunaga T. 1979. Two types of amino acid substitutions in protein evolution. *J Mol Evol*. 12(3):219–236.
- Nabholz B, Uwimana N, Lartillot N. 2013. Reconstructing the phylogenetic history of long-term effective population size and life-history traits using patterns of amino acid replacement in mitochondrial genomes of mammals and birds. *Genome Biol Evol*. 5(7):1273–1290.
- Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol*. 3(5):418–426.
- Smith NG. 2003. Are radical and conservative substitution rates useful statistics in molecular evolution? *J Mol Evol*. 57(4):467–478.
- Subramanian S. 2011. Fixation of deleterious mutations at critical positions in human proteins. *Mol Biol Evol*. 28(9):2687–2693.
- Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res*. 34(Web Server issue):W609–W612.
- Tang H, Wu CI. 2006. A new method for estimating nonsynonymous substitutions and its applications to detecting positive selection. *Mol Biol Evol*. 23(2):372–379.
- Tang H, Wyckoff GJ, Lu J, Wu CI. 2004. A universal evolutionary index for amino acid changes. *Mol Biol Evol*. 21(8):1548–1556.
- Wang H-Y, et al. 2017. Is the evolution in tumors Darwinian or non-Darwinian? *Natl Sci Rev* 5(1):15–17.
- Wen H, He X, Wu C-I, Wang H-Y. 2018. On the low reproducibility of cancer studies. *Natl Sci Rev* 5(5):619–624.
- Weber CC, Nabholz B, Romiguier J, Ellegren H. 2014. Kr/Kc but not dN/dS correlates positively with body mass in birds, raising implications for inferring lineage-specific selection. *Genome Biol*. 15(12):542.
- Weber CC, Whelan S. 2019. Physicochemical amino acid properties better describe substitution rates in large populations. *Mol Biol Evol*. 36(4):679–690.
- Williamson SH, et al. 2005. Simultaneous inference of selection and population growth from patterns of variation in the human genome. *Proc Natl Acad Sci U S A*. 102(22):7882–7887.
- Xu S, et al. 2017. The origin, diversification and adaptation of a major mangrove clade (Rhizophoraceae) revealed by whole-genome sequencing. *Natl Sci Rev*. 4(5):721–734.
- Yang Z. 2006. *Computational Molecular Evolution*. Oxford: Oxford University Press. p. 48–49.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 24:1586–1591.

Yang Z, Nielsen R, Hasegawa M. 1998. Models of amino acid substitution and applications to mitochondrial protein evolution. *Mol Biol Evol.* 15(12):1600–1611.

Zuckerkandl E, Pauling L. 1965. Evolutionary Divergence and Convergence in Proteins. In: Bryson V, Vogel HJ, editors. *Evolving Genes and Proteins*. New York: Academic Press. p. 97–166.

**Associate editor:** Kirk Lohmueller