Evolution in small steps - A nearly-universal rule of negative selection on codons

Qingjian Chen[1#], Ao Lan[1#], Xu Shen[1] and Chung-I Wu[1,2,3*]
[1]State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-sen University, Guangzhou 510275, China.
[2]CAS Key Laboratory of Genomic and Precision Medicine, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China.
[3]Department of Ecology and Evolution, University of Chicago, Chicago Illinois 60637, USA
[#]These authors contributed equally to this work.
*Corresponding author:
Chung-I Wu, ciwu@uchicago.edu

Abstract
The widely accepted view that evolution proceeds in small steps is based on two assumptions: i) negative selection acts strongly against large differences (Kimura 1983); and ii) positive selection favors small-step changes (Fisher 1930). The two assumptions are not biologically connected and should be evaluated separately. To do so, we measure the substitution rates between codons, using the physico-chemical distance between amino acids (AAs) for step size. The distance used is a function of 48 physico-chemical properties. Between 9 pairs of closely-related species of plants, invertebrates and vertebrates, the evolutionary rate is indeed strongly and *negatively* correlated with the AA distance (Δ_U, scaled to [0, 1]). Remarkably, Δ_U is strongly correlated with the evolutionary rate ($R^2 > 0.8$) in almost all taxa, as long as the genes are under some negative selection. Δ_U is thus a nearly-universal measure of evolutionary constraint. We conclude that negative selection is the driving force of the small-step evolution, the strength of which is governed by Δ_U. The possibility that positive selection may not follow the small-step rule will be tested in the companion study (Chen and Wu).

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 multi-dimensional 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 selection are distinct forces driving different processes (Wang, et al. 2017), 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 amino acid-based or codon based. For example, PAM (Dayhoff, et al. 1978) and BLOSUM (Henikoff and Henikoff 1992) are AA-based, searching for long-term evolutionary patterns among the 190 (= 20 x 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).

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 and Wu). In this study, 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.

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 amino acid (AA) substitutions whose codons differ by only one bp. Each of these classes is labeled Ki, i=1, 75. Hence,

$$Ka = \sum_{i=1}^{75} w_i K_i \qquad \text{Eq. (1)}$$

where w_i is the weight reflecting the number of sites available for AA exchanges of the i-th pair. In the PAML software package (Yang 2007), Ki's are obtained first when calculating Ka by Eq. (1). Therefore, no additional calculation is needed. In the next two sections, we will show that i) Ki's are nearly universally correlated across taxa; and ii) This high correlation reflects the physicochemical properties of AAs.

I. The correlation among Ki's across taxa

We first calculate Ki 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.

Fig. 1 here (The big phylogeny)

We then calculate the correlation in Ki 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 Fig. 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. While Ki within each taxon may be very different, their relative magnitudes remain the same across all taxa.

We further examine the correlation in a smaller phylogenetic framework by comparing species pairs between vertebrate classes and between mammalian orders (Fig. 1E). While one might have expected the correlation to be even greater in the lower taxonomic rank, the observations given in Fig. 1E do not support this conjecture. In fact, when the evolutionary rate is taken into account (see below), the opposite is true. 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 $R^2 < 0.8$. Apparently, correlation strength does not depend on phylogenetic distance. Perhaps similar forces work reiteratively and independently in each taxon.

II. Ki in relation to AA properties

The high correlation across a large phylogenetic distance (Fig. 1) suggests something as basic as physicochemical properties of amino acids to be a major cause. We thus analyze 48 physicochemical, energetic, and conformational distances among AAs (Fig. 2). We use a composite scale of Ui (i = 1, 75 (Tang, et al. 2004)) to represent Ki across species, as Ui is highly correlated with each species' Ki (see Fig. 3 below).

Fig. 2 here (Physicochemical properties)

Ui is fitted to each of the 48 distance measures via an exponential function (see Kimura 1983)(Kimura 1983). Figs. 2A-C show regressions of Ui on hydrophobicity, total non-bounded energy, and volume. Although hydrophobicity is generally thought to be an important functional attribute, the correlation is weak ($R^2 = 0.09$). Instead, the volume of AAs explains more of the Ui variation with $R^2 = 0.27$. The correlation coefficient for each AA property is given in Fig. 2D which shows that 46 measures are negatively correlated with Ui, 38 with $P < 0.05$. Since only two of the 48 AA properties are uncorrelated with the evolutionary rate, no AA property stands out.

The complexity of translating evolutionary rate and fitness back to biochemistry is further revealed by a Principal Component Analysis (PCA). Fig. 2F shows that PC1 explains about 61% of the Ui variation. Further details can be seen in Fig. 2E. It shows that the contributions to PC1 are distributed broadly among the 48 measures with "helical contact area" in the lead, contributing about 4%.

III. The general rule for small-step evolution based on Ki

The high correlation among Ki values permits a composite Ki measure. This composite measure, referred to as Ui [U for universal; Tang et al. (Tang, et al. 2004)]), ranges between 0.25 and 2.5 for i= 1 to 75. Ui is scaled such that the weighted mean is 1 across the 75 classes (Tang, et al. 2004). For any species, the expected Ki is

$$\mathrm{Exp(Ki)} = \mathrm{Ui} * \mathbf{Ka} \qquad \text{Eq. (2)}$$

where the boldface Ka denotes the species' genome-wide value. As long as we know **Ka**, Exp(Ki) can be calculated for i = 1, 75. In addition, the AA distance of the i-th pair can be defined as $\Delta_U(i) = (U_1\text{-Ui})/(U_1\text{-}U_{75})$, which is scaled to [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]. $\Delta_U(i)$ is used extensively in the companion study (Chen and Wu).

Fig. 3 here (Observed vs. expected Ki)

The observed vs. expected Ki for the nine pairs of species are shown in Fig. 3. The X-axis is Exp(Ki/Ks), which can be written as $[2.5 - 2.25\ \Delta U_i] * \mathbf{Ka/Ks}$ (see the legend of Fig. 3 and Table S1 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. 3A, B, and G). All three taxa show $R^2 > 0.9$ between Obs(Ki) and Exp(Ki). In the second group of five pairs of genomes of moderate quality, 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, which yields an R^2 of 0.627, far lower than the rest. There may be several explanations for this unusually low R^2. Since codon substitutions involving CpG sites have been removed from consideration (see Methods), this obvious explanation is ruled out (see Fig. S1). 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 (Fig. S2 and

Supplementary Note), 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.

IV. The rule of small-step evolution weakens when Ka/Ks is high

For each of the eight pairs of species, expect Mus-Rattus pair , we divide the genomes into five bins in the ascending order of Ka/Ks (see Methods). The Ka/Ks of each bin in each species, together with their R^2 values (Obs(Ki) vs. E(Ki)), are plotted in Fig. 4. There are two obvious 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 passes 0.4. Second, the human-chimpanzee comparison follows the same pattern but generally at a lower level.

Fig. 4 here (Correlation coefficient vs. Ka/Ks)

There may be two contrasting forces driving the trend displayed in Fig. 4 when Ka/Ks increases: weaker negative selection and/or stronger positive selection. Separating the two effects will be the subject of the accompanying study (Chen et al.). Without teasing apart the two separate influences, we can nevertheless draw a conclusion. The evolution of genes with Ka/Ks < 0.4, whereby negative selection prevails, takes small steps and follows a nearly universal rule. This rule is governed by the physicochemical properties of AAs.

Discussion

It is striking that the small-step evolution follows a simple and nearly universal rule. If 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 non-mammalian vertebrates, invertebrates, and plants. The relative magnitude remains nearly constant. This nearly universal pattern is observed as long as negative selection is the prevailing force. In most genes of most taxa, Ka/Ks < 0.4 and negative selection, eliminating more than half of AA mutations, is indeed the driving force.

Given the general pattern in such diverse taxa, negative selection at the codon level must be operating at a basic level of biochemistry. We show that the rule of negative selection depends on the properties of the AA residues. Interestingly, almost all measures of AA properties contribute to the fitness difference and even the most obvious properties like hydrophobicity and non-bonded energy contribute only a small fraction to the overall evolutionary rate. Collectively, the first principal component does explain more than 60% of the rate variation.

The simple evolutionary pattern associated with 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). Genetic disease associated with AA changes is such an example of negative selection (Adzhubei, et al. 2013). Since matrices relying on a few biochemical properties are not likely to capture most of the evolutionary pattern (see Fig. 2), it may be more informative to assess the evolutionary rate directly from DNA sequence data. In this perspective, the Δ_U measure of codon substitution should be particularly suited to investigating the strength of negative selection. Lastly, previous studies have reported unusual AA substitution patterns in extreme environments or under domestication, where positive selection might be prevalent (Lu, et al. 2006; Luo, et al. 2017; Xu, et al. 2017). The hypothesis that positive selection may follow rules distinct from those of negative selection will be tested in the companion study.

Materials and Methods

Multiple Alignment Data

A multiple alignment file of 99 vertebrates with human for CDS regions was downloaded from the UCSC Genome Brower (http://hgdownload.cse.ucsc.edu/goldenPath/hg38/multiz100way/alignments/knownCanonical.exonNuc.fa.gz). We then selected seven representative pairs of species across vertebrates (Table S1). 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. (Lin, et al. 2018) and Wang et al. (Wang, et al. 2018).

A multiple alignment file of 26 insects with *D. melanotaster* was also downloaded from the UCSC Genome Brower(http://hgdownload.cse.ucsc.edu/goldenPath/dm6/multiz27way/alignments/refGene.exonNuc.fa.gz). The genome sequences of *D.melanogaster* and *D.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 un-canonical scaffolds. **2)** Use the corresponding amino-acid multiple alignment file to remove non-coding genes. **3)** Filter redundant transcripts generated by alternative splicing, based on a gene-transcript id conversion table downloaded from Ensembl biomart (www.ensembl.org/biomart) . **4)** Genes aligned with too many gaps (more than 5%) were removed. **5)** Genes shorter than 90 bp were removed. **6)** Mutations related to CpG sites in the Homo-Pan pair were masked because CpG-related mutations account for more than 30% of total mutations in this pair. Hence, we masked CpG related mutations (CG => TG, CG => CA were replaced with NG => NG and CN => CN respectively) using gorilla as the outgroup (Fig. S1, Supplementary Note).

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 calculation

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

$$q_{uv} = \begin{cases} \mathbf{0}, & \text{if the two codons differ at more than one position} \\ \boldsymbol{\pi}_{\mathbf{v}}, & \text{for synonymous transversion} \\ \boldsymbol{\kappa}\boldsymbol{\pi}_{\mathbf{v}}, & \text{for synonymous transition} \\ \boldsymbol{\omega}_{\mathbf{ij}}\boldsymbol{\pi}_{\mathbf{v}}, & \text{for nonsynonymous transversion} \\ \boldsymbol{\omega}_{\mathbf{ij}}\boldsymbol{\kappa}\boldsymbol{\pi}_{\mathbf{v}}, & \text{for nonsynonymous transition} \end{cases}$$

Here, q_{uv} is the instantaneous rate from codon u to codon v. κ is the transition/transversion rate ratio, π_v is the equilibrium frequency of codon v. ω_{ij} is the non-synonymous/synonymous rate ratio, where i (= aa_u) and j (=aa_v) are the two amino acids involved. There are 75 ω_{ij}, analogous to Ki/Ks (i =1:75), of amino acid pairs whose codons differed by 1 bp. The method considering unequal substitution rate between different amino acids has been summarized in Yang et al. (1998)(Yang, et al. 1998).

The parameters in PAML are as follows : 1) *Model=0*. Use the site model to allow ω to vary among sites. 2) *codenFreq=0*. Assume individual codon frequencies are equal. 3) *NSsites=0*. Use the M0 (one ratio) model to estimate ω. 4) *aaDist=7*. Allow to estimate amino-acid substitution rates separately. Expected Ki/Ks values were calculated as Exp(Ki/**Ks**) = Ui***Ka/Ks** in each pair of species.

Analysis of amino acid physicochemical properties

Amino acid properties were extracted from Gromiha et al.(1999) (Gromiha, et al. 1999). There are 48 selected physicochemical, energetic, and conformational properties of AAs, measured using primary or secondary structure information. The distances between different AAs were simply defined by their absolute difference in each amino acid property. Various distance methods yielded similar results. PCA was performed using the "factoextra" R package (Kassambara and Mundt 2017). Exponential fitting was accomplished using the function 'nls' in R.

The relationship between Ui and Ka/Ks

Orthologous genes for each paired species were divided into five categories with approximately equal number of non-synonymous changes according to their Ka/Ks ranking. Genes with fewer than two mutations were removed. In an ideal case, Ka values increase when Ka/Ks values ascend and Ks values are constant 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. 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 Fig. 4 because Ui values were deduced from the comparison between Mus-Rattus and a pair of yeast (Tang, et al. 2004).

Acknowledgements

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 (91731301 and 31730046), the 985 Project (33000-18831107), and National Key Basic Research Program of China (2014CB542006).

References

Adzhubei I, Jordan DM, Sunyaev SR. 2013. Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet 76:7-20.

Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT, Glanowski S, Tanenbaum DM, White TJ, Sninsky JJ, Hernandez RD, et al. 2005. Natural selection on protein-coding genes in the human genome. Nature 437:1153-1157.

Chen Q, Wu C-I. Molecular evolution in large steps - Codon substitutions under positive selection. unpublished.

Dayhoff M, Schwartz R, Orcutt B. 1978. A model of evolutionary change in proteins. In. Atlas of protein sequence and structure: National Biomedical Research Foundation Silver Spring, MD. p. 345-352.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792-1797.

Fisher RA. 1930. The genetical theory of natural selection. Oxford,: The Clarendon press.

Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, et al. 2012. Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res 40:D1178-1186.

Grantham R. 1974. Amino acid difference formula to help explain protein evolution. Science 185:862-864.

Gromiha MM, Oobatake M, Sarai A. 1999. Important amino acid properties for enhanced thermostability from mesophilic to thermophilic proteins. Biophys Chem 82:51-67.

Henikoff S, Henikoff JG. 1992. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A 89:10915-10919.

Kassambara A, Mundt F. 2017. Factoextra: extract and visualize the results of multivariate data analyses. R package version 1.0.5. In: https://cran.r-project.org/package=factoextra.

Kimura M. 1983. The neutral theory of molecular evolution: Cambridge University Press.

Kumar P, Henikoff S, Ng PC. 2009. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc 4:1073-1081.

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 2:150-174.

Lin T, Xu X, Ruan J, Liu S, Wu S, Shao X, Wang X, Gan L, Qin B, Yang Y, et al. 2018. Genome analysis of Taraxacum kok-saghyz Rodin provides new insights into rubber biosynthesis. National Science Review 5:78-87.

Lu J, Tang T, Tang H, Huang J, Shi S, Wu CI. 2006. The accumulation of deleterious mutations in rice genomes: a hypothesis on the cost of domestication. Trends Genet 22:126-131.

Luo H, Huang Y, Stepanauskas R, Tang J. 2017. Excess of non-conservative amino acid changes in marine bacterioplankton lineages with reduced genomes. Nature Microbiology 2:17091.

Miyata T, Miyazawa S, Yasunaga T. 1979. Two types of amino acid substitutions in protein evolution. J Mol Evol 12:219-236.

Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3:418-426.

Subramanian S. 2011. Fixation of Deleterious Mutations at Critical Positions in Human Proteins. Mol Biol Evol 28:2687-2693.

Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res 34:W609-612.

Tang H, Wu CI. 2006. A new method for estimating nonsynonymous substitutions and its applications to detecting positive selection. Mol Biol Evol 23:372-379.

Tang H, Wyckoff GJ, Lu J, Wu CI. 2004. A universal evolutionary index for amino acid changes. Mol Biol Evol 21:1548-1556.

Wang G-D, Shao X-J, Bai B, Wang J, Wang X, Cao X, Liu Y-H, Wang X, Yin T-T, Zhang S-J, et al. 2018. Structural variation during dog domestication: insights from gray wolf and dhole genomes. National Science Review:https://doi.org/10.1093/nsr/nwy1076.

Wang H-Y, Chen Y, Tong D, Ling S, Hu Z, Tao Y, Lu X, Wu C-I. 2017. Is the evolution in tumors Darwinian or non-Darwinian? National Science Review 5:15-17.

Williamson SH, Hernandez R, Fledel-Alon A, Zhu L, Nielsen R, Bustamante CD. 2005. Simultaneous inference of selection and population growth from patterns of variation in the human genome. Proc Natl Acad Sci U S A 102:7882-7887.
Xu S, He Z, Zhang Z, Guo Z, Guo W, Lyu H, Li J, Yang M, Du Z, Huang Y, et al. 2017. The origin, diversification and adaptation of a major mangrove clade (Rhizophoreae) revealed by whole-genome sequencing. National Science Review 4:721-734.
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 15:1600-1611.

Figures

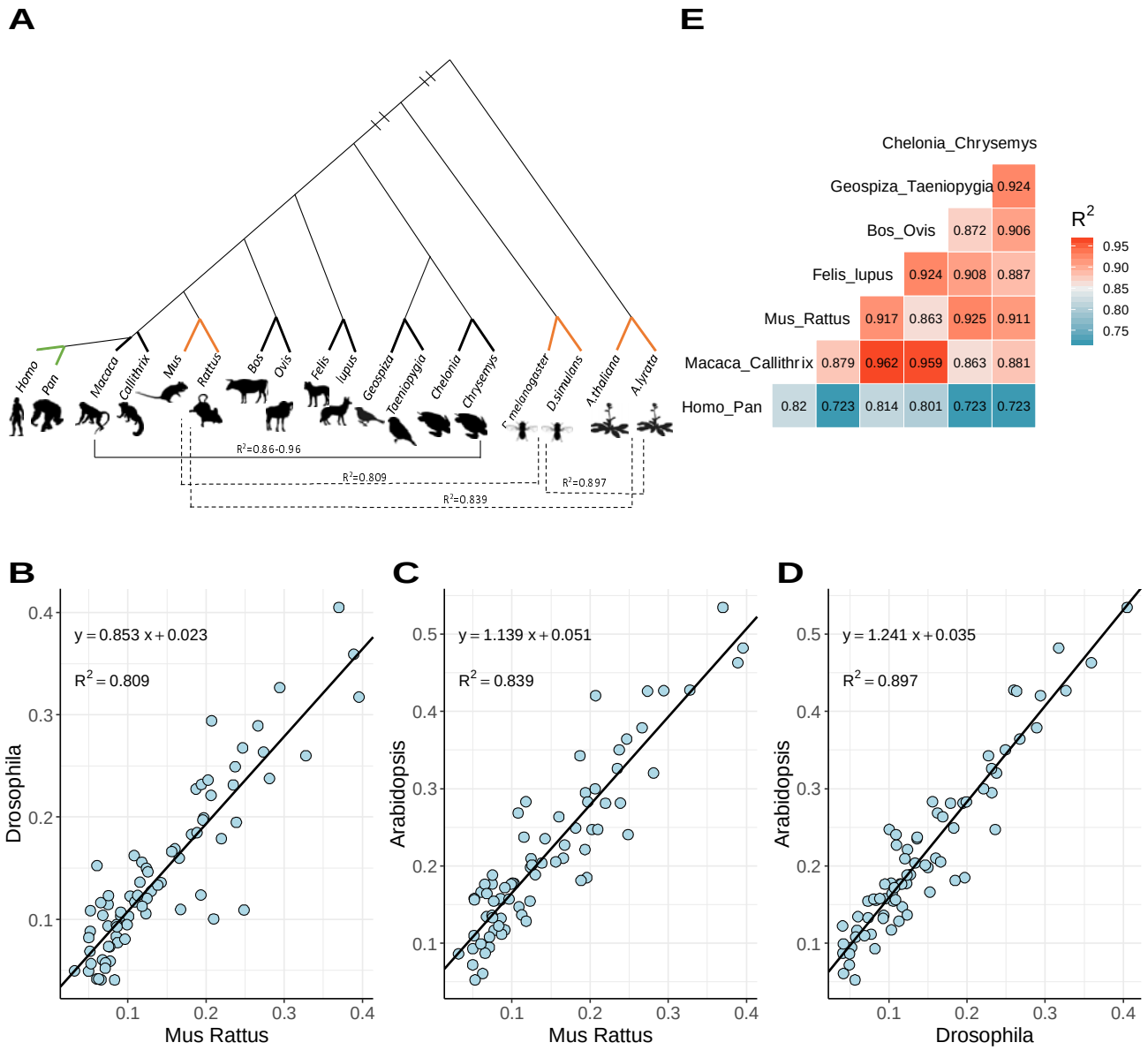

Figure 1| Correlation of Ki/Ks values between different taxa. (**A**) Phylogenetic Tree of 9 representative species pairs (full names are shown in Table S1). R^2 values (pairwise correlation of Ki's) range from 0.86 to 0.96 in 6 vertebrates (horizontal solid line), except for Homo-Pan pairs. Three species pairs, Rodent, Drosophila and Arabidopsis (bold orange lines), are selected to show species pairs with long evolutionary distance. (**B-D**) Linear regression results of Ki/Ks between Rodent, Drosophila and Arabidopsis. (**E**) Pairwise R^2 values of Ki/Ks among 7 species pairs in vertebrates.

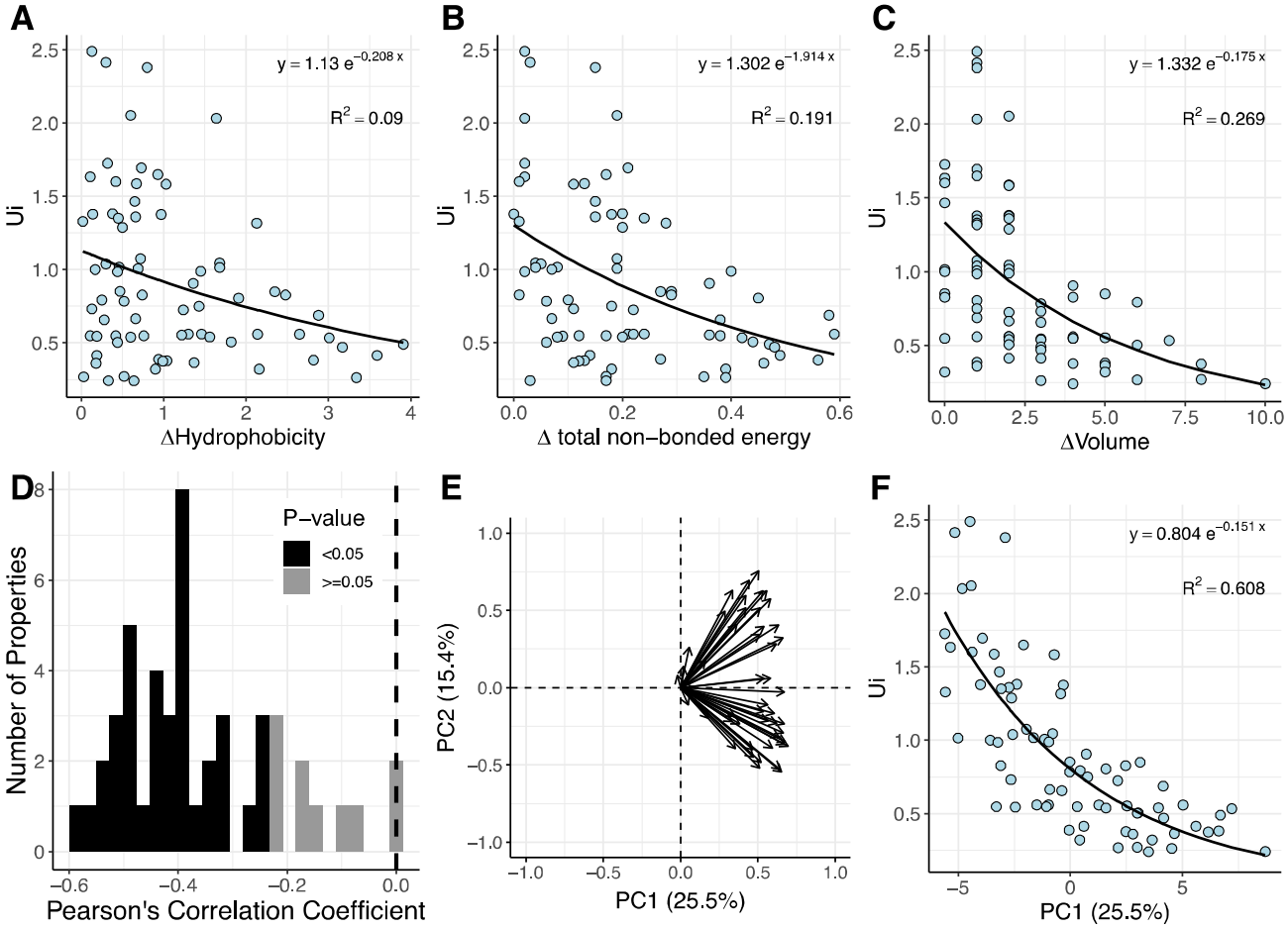

Figure 2 |The relationship between physicochemical AA distance and Ui . 48 physicochemical, energy and conformational properties are used in the analysis. (**A-C**) Scatter plot of three selected properties against Ui , including hydrophobicity(**A**), total non-bounded energy(**B**) and volume(**C**). (**D**) Pearson's correlation coefficients between 48 properties and Ui. Among those properties, 46 properties are negatively correlated with Ui, where 38 of them are significantly (black, p-value < 0.05). (**E**) Contribution of 48 properties to the first two principal components. Properties cluster into two categories in the second principal component but are indistinguishable in the first principal component. (**F**) Scatter plot of first principal component (PC1) against Ui.

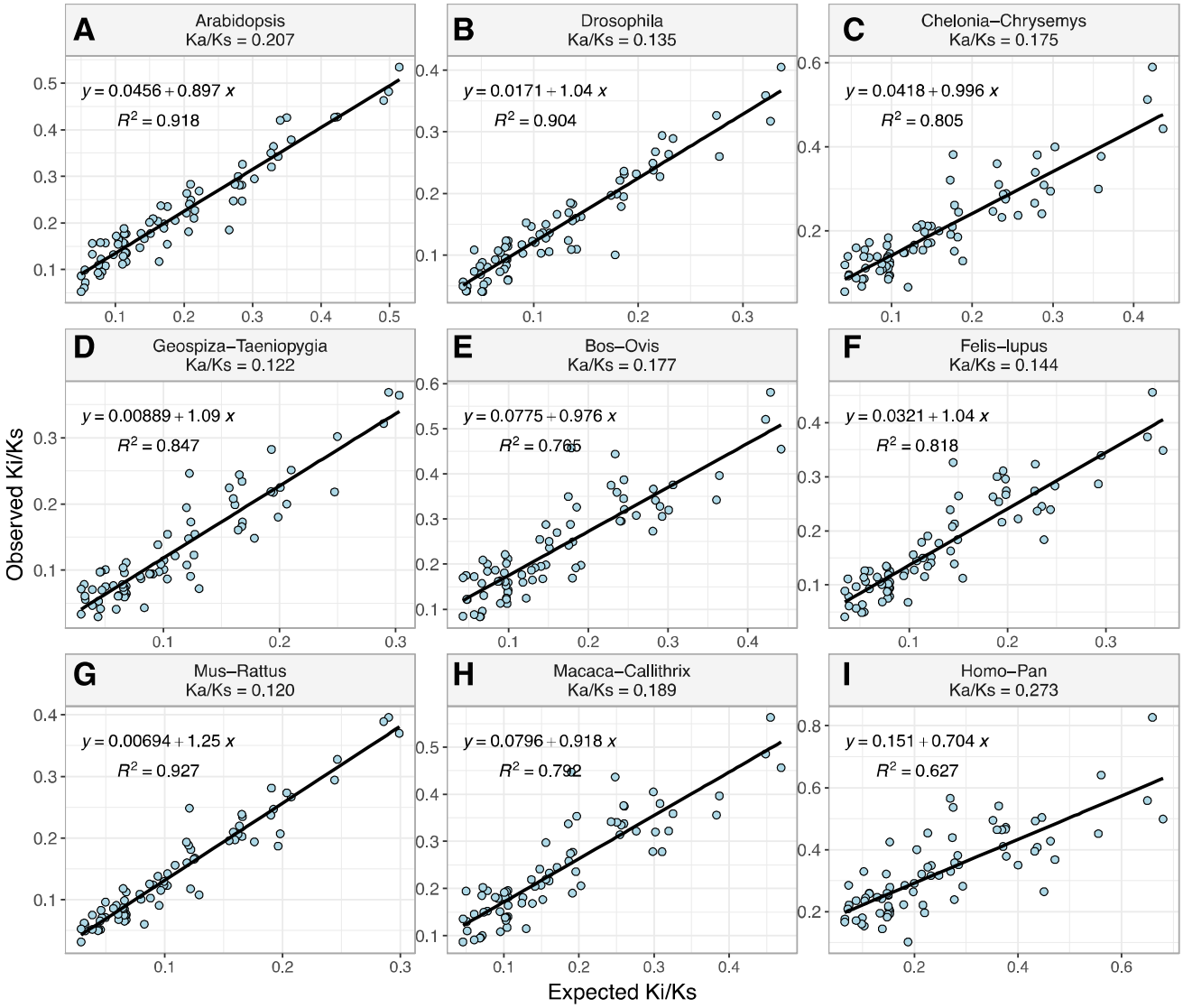

Figure 3 | Correlations of expected and observed Ki/Ks for nine species pairs. Since $U_i = U_1 - (U_1 - U_{75})\ \Delta U_i$, the X-axis label can be written as Exp(Ki/**Ks**) = Ui***Ka/Ks** = [2.5 - 2.25 ΔU_i]***Ka/Ks**, which is a linear function of ΔU_i. The boldface **Ka/Ks**, the genome-wide Ka/Ks, is the characteristic of each pair of species.

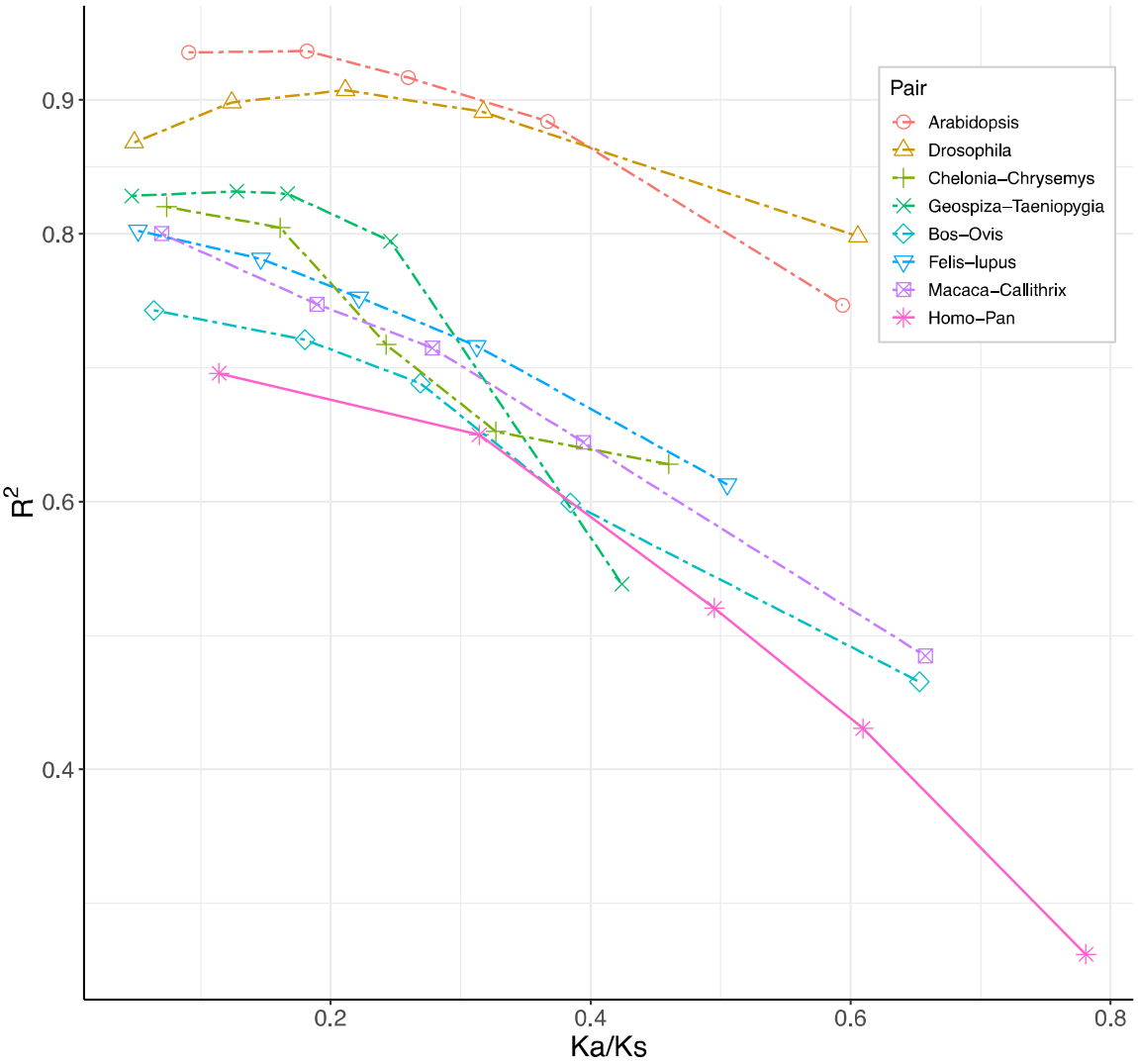

Figure 4 | Relationship between R^2 values(Obs(Ki) vs. Exp(Ki)) and selection strength(Ka/Ks). Orthologous genes in each pairs of species are divided them into 5 categories with equal non-synonymous changes, according to their Ka/Ks ratio. The x-axis is average Ka/Ks ratio of each category in each species pair and the y-axis is the R^2 value(squared correlation coefficient) of their expected against observed Ki/Ks.