Roles of Solvent Accessibility and Gene Expression in Modeling Protein Sequence Evolution

Kuangyu Wang[1,*], Shuhui Yu[1,2,†,‡], Xiang Ji[1], Clemens Lakner[1,§], Alexander Griffing[1] and Jeffrey L. Thorne[1]

[1]Bioinformatics Research Center, North Carolina State University, Raleigh, NC, USA. [2]College of Life Science, Chongqing University, Chongqing, China. *First author. †Co-first author. ‡Current Address: University Library, Southwest University, Chongqing, China. §Current Address: European Molecular Biology Laboratory, Heidelberg, Germany.

ABSTRACT: Models of protein evolution tend to ignore functional constraints, although structural constraints are sometimes incorporated. Here we propose a probabilistic framework for codon substitution that evaluates joint effects of relative solvent accessibility (RSA), a structural constraint; and gene expression, a functional constraint. First, we explore the relationship between RSA and codon usage at the genomic scale as well as at the individual gene scale. Motivated by these results, we construct our framework by determining how probable is an amino acid, given RSA and gene expression, and then evaluating the relative probability of observing a codon compared to other synonymous codons. We come to the biologically plausible conclusion that both RSA and gene expression are related to amino acid frequencies, but, among synonymous codons, the relative probability of a particular codon is more closely related to gene expression than RSA. To illustrate the potential applications of our framework, we propose a new codon substitution model. Using this model, we obtain estimates of $2Ns$, the product of effective population size N, and relative fitness difference of allele s. For a training data set consisting of human proteins with known structures and expression data, $2Ns$ is estimated separately for synonymous and nonsynonymous substitutions in each protein. We then contrast the patterns of synonymous and nonsynonymous $2Ns$ estimates across proteins while also taking gene expression levels of the proteins into account. We conclude that our $2Ns$ estimates are too concentrated around 0, and we discuss potential explanations for this lack of variability.

KEYWORDS: protein evolution, protein structure, gene expression, codon usage, scaled selection coefficient, solvent accessibility

CITATION: Wang et al. Roles of Solvent Accessibility and Gene Expression in Modeling Protein Sequence Evolution. *Evolutionary Bioinformatics* 2015:11 85–96 doi: 10.4137/EBO.S22911.

RECEIVED: December 14, 2014. **RESUBMITTED:** February 4, 2015. **ACCEPTED FOR PUBLICATION:** February 9, 2015.

ACADEMIC EDITOR: Jike Cui, Associate Editor

TYPE: Original Research

FUNDING: CL, JLT, and SY were supported by NIH award GM070806. AG, KW, and JLT were supported by NSF award MCB 1021883. XJ was supported by NIH award GM090201. The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: thorne@statgen.ncsu.edu

COPYRIGHT: © the authors, publisher and licensee Libertas Academica Limited. This is an open-access article distributed under the terms of the Creative Commons CC-BY-NC 3.0 License.

Paper subject to independent expert blind peer review by minimum of two reviewers. All editorial decisions made by independent academic editor. Upon submission manuscript was subject to anti-plagiarism scanning. Prior to publication all authors have given signed confirmation of agreement to article publication and compliance with all applicable ethical and legal requirements, including the accuracy of author and contributor information, disclosure of competing interests and funding sources, compliance with ethical requirements relating to human and animal study participants, and compliance with any copyright requirements of third parties. This journal is a member of the Committee on Publication Ethics (COPE).

Published by Libertas Academica. Learn more about this journal.

Introduction

Proteins perform a vast array of functions within living organisms. As a result, natural selection affects both the structure of proteins and the regulation of their production. A major source of structural constraints arises because proteins require a stable and suitable three-dimensional structure to function. Mutations that destabilize proteins will be selected against.[1] Also, selective constraints in a protein vary according to structural locations. For example, Franzosa and Xia[2] found a strong, positive, and linear relationship between the ratio of nonsynonymous and synonymous rates and a measure of solvent exposure at the structural location. While structural constraints have mainly been investigated to explain nonsynonymous rate variation within proteins, functional constraints have been shown to explain a large amount of the variation in average nonsynonymous rate among proteins.[3–5] For example, Drummond et al.[4] showed that expression level explains roughly half the variation in average nonsynonymous rate among *Saccharomyces cerevisiae* protein-coding genes.

To reflect structural constraints when modeling protein evolution, the pioneering simulation work of Parisi and Echave[6] had a statistical potential to govern the amino acid replacement process. A similar approach was adopted by later inferential studies.[7–10] Statistical potentials can include terms related to solvent accessibility, pairwise distance interactions, torsion angles, and flexibility of the residues.[11]

Structurally constrained evolutionary models have tended to place less emphasis on codon usage. The assumption that synonymous mutations are selectively neutral has often been made.[7,8,10] This may be a reasonable first assumption, because so much natural selection depends on the protein sequence. However, synonymous change is also affected by natural selection. For example, Agashe et al.[12] found that synonymous mutations to a key-enzyme coding gene can decrease gene expression and fitness by more than 90% compared to wild type. Also, selection acting on gene expression was found to be the single dominant predictor, among all predictors considered, of the number of nonsynonymous substitutions per site in yeast.[4]

To explain the negative correlations between rates of coding sequence evolution and gene expression levels that have been inferred across a wide range of taxa, Drummond and Wilke[13] suggested that mistranslation-induced protein misfolding explains much of coding sequence evolution. Another study discovered that translationally optimal codons are associated with structurally sensitive sites.[14] These findings indicate that structural and functional constraints are coupled. While protein structures are increasingly considered in protein evolution models, less attention has been paid to the combined effect of gene expression and protein structure.

The overall codon usage in a genome can be dramatically different between species.[15] Chen et al.[16] concluded that different patterns of codon usage between species are determined primarily by mutational processes that act throughout the genome and only secondarily by selective forces acting on protein-coding sequences. From a selectionist point of view, a classic explanation for systematic variation across a genome is that certain preferred codons are translated more accurately and/or efficiently.[17–19] Strong evidence for this hypothesis has been found in several species.[20–24] However, 30% of bacterial species show no evidence of translational selection.[25] Understanding codon usage patterns continues to be an active area of research.

Relative solvent accessibility (RSA) is a summary of local structural environment at a protein location that aims to quantify the relative exposure of an amino acid in a globular protein to water molecules. It has been a widely used summary that is correlated with rates and patterns of protein evolution.[2] Here, we attempt to characterize the roles played by RSA and gene expression in modeling protein evolution. We propose a probabilistic framework aiming to combine effects of RSA and gene expression on amino acid usage as well as codon usage. We carry out hypothesis testing in order to address whether RSA tendencies vary among synonymous codons. To capture potentially species-specific association between RSA and codon usage, the testing procedure is conducted at the genomic scale as well as at the individual gene scale in two different species (*Mus musculus* and *Homo sapiens*). We assess the effects of gene expression on synonymous codon usage and the combined effects of RSA and gene expression in influencing amino acid usage. We do this via a multinomial logistic regression (MLR) approach that connects RSA and gene expression to evolution using correlation rather than causality. To demonstrate possible applications of our probabilistic framework, we propose a new protein evolutionary model that accounts for both the structural and functional contexts of a codon. Our model has a mutation-selection balance framework that incorporates the selective impacts of possible synonymous and nonsynonymous mutations on human protein-coding genes. The selective impacts of these mutations are assessed by our estimates of scaled selection coefficients (ie, twice the product of effective population size N and the relative fitness difference s between the mutant and wild-type allele).

Materials and Methods

Structural, sequence, and expression data. We collected protein structures, amino acid sequences, nucleotide sequences, and protein-coding gene expression data for two species, *H. sapiens* and *M. musculus*. Protein structures were obtained from the Protein Data Bank (PDB).[26] Only structures determined by X-ray crystallography with a resolution of 3.0 Å or better were employed. In addition, proteins were required to have lengths greater than 50 amino acids. In an effort to collect a relatively homogeneous data set to which a single model applies, we restrict our data to monomeric proteins with one chain. Also for the purposes of gathering a relatively homogeneous data set, membrane proteins and protein–DNA/RNA hybrid structures were excluded. To lessen problems arising from estimating parameters when proteins are correlated because of common ancestry, homologs were removed from our data by employing a 30% identity filter. Gene expression measurements were not used to determine which homologs were included or excluded.

Because the Consensus CDS (CCDS) database[27] stores a core set of human and mouse protein-coding regions that are associated with highly reliable annotation, it was used to identify nucleotide sequences that encode the proteins selected from PDB. To get CCDS IDs, protein PDB IDs were translated to the UniProt Knowledgebase (UniProtKB) IDs. The UniProt IDs were then mapped to Gene IDs. Finally, Gene IDs were converted to CCDS IDs. Proteins with CCDS IDs that could not be successfully identified were removed.

To match a protein in PDB with its corresponding protein-coding nucleotide sequence, we used the Smith–Waterman algorithm[28] of the water program from the EMBOSS tools.[29] This let us identify the longest ungapped region with an exact match between the nucleotide sequence and amino acid sequence. This process yielded 864 and 156 matches in *H. sapiens* and *M. musculus*, respectively (see PDB IDs in Supplementary files). For each site in the ungapped region in each protein structure, relative solvent accessibility (RSA) was calculated by the NACCESS software.[30] The presence of heteroatoms in some structures was ignored when calculating accessibilities.

RNA-seq data for *H. sapiens* and *M. musculus* were collected as part of a multi-species multi-organ gene expression study[31] and were downloaded from the Expression Atlas database.[32] To establish one-to-one mapping between the PDB ID of a protein and the Ensembl ID of a transcript, PDB IDs were translated to the UniProtKB IDs and then the UniProt IDs were mapped to Ensembl IDs via the ID mapping service hosted by the Protein Information Resource.[33] Following a previous study by Drummond and Wilke,[13] aggregate mRNA level was quantified as the geometric mean signal of the measurements from the six tissues in the Brawand et al data set. We log-transformed these geometric means in the following analysis. There are 241 human proteins and 60 mouse proteins in our data set that have both structure and gene expression information available.

A probabilistic framework for assessing the joint effect of RSA and gene expression on amino acid and codon usage. Our probabilistic framework is based on the idea that the propensity of a codon representing a given site of a protein can be predicted by a two-step process: First, estimate the probability of the corresponding amino acid by the observed frequency of that amino acid at similar structural and functional contexts in other proteins. Second, predict the relative probability of observing a codon conditional upon the amino acid, the structural context, and the functional context. For a codon C of a protein sequence, let A be the amino acid encoded by this codon. Here, R is the RSA at the site and represents the structural context, while E is the expression level of the gene and represents the functional context. The probability of obtaining C conditional upon the context is then

$$P(C \mid R,E) = P(C \mid A,R,E) \cdot P(A \mid R,E). \qquad (1)$$

Testing association between codon usage and RSA. Codon usage is correlated with gene expression.[19] Furthermore, amino acid usage is RSA dependent[2,34,35] and also influenced by gene expression.[36,37] Zhou et al.[14] *tested the association between codon optimality and solvent accessibility.* Instead of focusing on codon optimality that primarily reflects the influence of selection for translation speed and/or accuracy on codon usage, we test a slightly more general null hypothesis that, conditional upon the amino acid, codon usage is independent of RSA.

To do this, a k-sample Anderson–Darling (A-D) test[38] was applied to compare RSA distributions of different synonymous codons encoding an amino acid. Here, k equals the number of synonymous codons. For example, $k = 6$ if a test is conducted for serine. The A-D test is nonparametric and distribution free. To get the null distributions of A-D test statistics, we permuted RSA values among codons that specify the same amino acid type. By doing this, the association between amino acid usage and structural environments was not affected, but any association between codon usage and structural environments was eliminated. For each test, 10,000 simulated A-D test statistics were generated via this sort of permutation. The actual computation was done with the ad.test function in the R package kSamples.[38]

We can test whether RSA distributions of synonymous codons are different either at the genomic scale or at the individual gene scale. At the genomic scale, for each of the 18 amino acids that have more than one codon, a k-sample A-D test was conducted. Positions in all genes that have the same amino acid type were pooled together, and RSA values were permuted among these positions. In each test, the null distribution was formed by the 10,000 simulated A-D test statistics. To get the P-value, the test statistic computed from observed data is compared to the null distribution, and the proportion of the null distribution that exceeds the observed test statistic value is recorded. At the genomic scale, we have the power to detect even slight violations of the null hypothesis. A disadvantage of the genomic-level test is that permuted data sets do not retain the same codon usage patterns as the actual data. This means that association between codon usage and gene expression is disrupted in the permuted data sets.

When the hypothesis was tested at the gene scale, the resampling procedure only allows RSA values belonging to positions that have the same amino acid type to be permuted within a gene. Therefore, permuted data sets do not disrupt association between codon usage and gene expression because permuted genes retain the same codon usage for each gene as found in actual data. Under this restriction, 18 × 864 and 18 × 156 null distributions were generated, since we have 864 human proteins and 156 mouse proteins. To perform a single test of association between codon usage and RSA within each amino acid type, normalized test statistics (z-scores) were obtained for the amino acid type in each gene by subtracting the sample mean of the individual k-sample A-D test statistics for each combination of gene and amino acid type and then dividing the differences by the sample standard deviation. For each amino acid type, the 10,000 × 864 and 10,000 × 156 z-scores were summed across proteins so that 10,000 sum-of-z-scores (the combined null distribution) were obtained in human and in mouse. Finally, we compute P-values by comparing the observed sum-of-z-score for an amino acid with the corresponding null distribution. Although testing at the gene scale allows gene-specific codon bias to be removed, there is less power to reject the null hypothesis at the individual gene scale because each gene has a comparatively small number of codons for each amino acid.

Clustering amino acids according to RSA preference. The RSA of an amino acid residue in a protein is affected by tertiary structure of the protein, apart from the amino acid's physicochemical properties.[39] Before we proceed to quantitative modeling, we would like to find out how interchangeable two amino acids are in terms of RSA distribution and whether amino acids can be grouped together by their RSA distribution similarities. The two-sample Kolmogorov–Smirnov (KS) test is one of the most useful and general nonparametric methods for comparing two samples. In other contexts,[40–42] the KS test statistic has been used to define distances for the purpose of hierarchical clustering. Borrowing the idea from hypothesis testing, we define the distance between a pair of amino acids by the normalized two-sample KS test statistic (Equation 2) when contrasting their RSA distributions. We let n and n' be sample sizes for each amino acid type, and we obtain a test statistic:

$$\text{normalized KS statistic} = \frac{\sup_x |F_{1,n}(x) - F_{2,n'}(x)|}{\sqrt{\dfrac{n+n'}{nn'}}} \qquad (2)$$

where $F_{1,n}$ and $F_{2,n'}$ are the empirical RSA distribution functions of the two amino acids, respectively. Similar to hypothesis testing for codon usage, an amino acid distance matrix can be constructed either at the genomic scale or at the individual gene scale. For the latter, one distance matrix can be computed within each protein and all matrices can be combined by taking element-wise means. Complete-linkage clustering[43] is employed to find the grouping pattern in a distance matrix.

Logistic regression (LR).

If there is no significant association between codon usage and RSA, we can make the following simplification to Equation 1:

$$P(C \mid A, R, E) = P(C \mid A, E). \tag{3}$$

Since we are interested in understanding how gene expression (a predictor variable) affects probabilities of amino acid types (a categorical outcome variable), LR (in cases where there are exactly two synonymous codons) and MLR can be employed. The most frequent codon in each synonymous group is denoted as C_r, and the probability of any other synonymous codon C is modeled relative to it:

$$\log\left(\frac{P(C \mid A, E)}{P(C_r \mid A, E)}\right) = \alpha + \beta_C \cdot E \tag{4}$$

where α is the intercept and β_C is the slope for gene expression with codon C. When amino acids have exactly two synonymous codons, we use LR instead of MLR. Since methionine and tryptophan are encoded by a single codon, they are not included in this analysis.

To measure the joint effect of RSA and gene expression on amino acid usage quantitatively, we use both gene expression and RSA as independent variables and build a model similar to the one in Equation 4, but this time the probabilities are relative to A_r, the reference amino acid:

$$\log\left(\frac{P(A \mid R, E)}{P(A_r \mid R, E)}\right) = \alpha + \beta_{R,A} \cdot R + \beta_{E,A} \cdot \tag{5}$$

where $\beta_{R,A}$ and $\beta_{E,A}$ are the slopes for amino acid A of RSA and gene expression, respectively. The most frequent amino acid (in our case, leucine) is chosen as A_r. A likelihood ratio test can compare the full model in Equation 5 with reduced models in Equations 6 and 7.

$$\log\left(\frac{P(A \mid R, E)}{P(A_r \mid R, E)}\right) = \alpha + \beta_{R,A} \cdot \tag{6}$$

$$\log\left(\frac{P(A \mid R, E)}{P(A_r \mid R, E)}\right) = \alpha + \beta_{E,A} \cdot \tag{7}$$

A functionally and structurally constrained Markov model of codon substitution. Using the probabilistic framework that we have constructed, we develop a model of protein-coding DNA sequence evolution that explicitly accounts for structural and functional constraints.

The rate of change from codon i to codon j is denoted by q_{ij}. Our mutation-selection balance parameterization has each substitution rate be proportional to the mutation rate from i to j multiplied by the probability that a new mutation becomes fixed. We denote π_k as the stationary probability of nucleotide type k ($k \in \{A, C, G, T\}$) in the absence of natural selection and use π as the vector representing the four nucleotide frequencies. Let κ be the transition–transversion rate ratio. Assume a diploid population of effective size N individuals, and let $P(Z_{ij})$ be the probability that a new mutant allele j eventually gets fixed in the population that otherwise consists of $2N - 1$ alleles of type i. We set q_{ij} to 0 if i or j is a stop codon, or i and j differ by more than one nucleotide. When i and j differ by exactly a single codon position and codon j has nucleotide type h at this position, our model specifies the substitution rate q_{ij} from codon i to codon j as

$$q_{ij} = \begin{cases} u\pi_h \times 2N \times P(Z_{ij}) & \text{transversior} \\ u\pi_h\kappa \times 2N \times P(Z_{ij}) & \text{transition,} \end{cases} \tag{8}$$

where the scale factor u is defined by the requirement that the average substitution rate for a codon substitution process at stationarity is 1.

Following Choi et al.[44], we design our evolutionary model to have a stationary distribution that matches a desired target probability distribution. In this case, we have our evolutionary model yield a stationary distribution that is identical to the probability distribution of codons that we estimated via MLR from solvent accessibility and gene expression data of human proteins. Accordingly, we set $P(i|R,E)$ as the stationary probability for codon i. It can be calculated using Equations 1, 4, and 5.

Let $P_0(i|\pi)$ be the probability of codon i under a model where stop codons are lethal, but all other codons are equally fit. Adjusting for the three stop codons in the universal genetic code,

$$P_0(i \mid \pi) = \frac{\pi_{i_1}\pi_{i_2}\pi_{i_3}}{1 - \pi_T\pi_A\pi_A - \pi_T\pi_A\pi_G - \pi_T\pi_G\pi_A}. \tag{9}$$

According to the approximation introduced in Choi et al.[44],

$$\tau_{ij} = \frac{P(j \mid R, E) / P_0(j \mid \pi)}{P(i \mid R, E) / P_0(i \mid \pi)} \tag{10}$$

and

$$2N \times P(Z_{ij}) = \frac{\log(\tau_{ij})}{1 - 1/\tau_{ij}}, \tag{11}$$

yield a Markov model that is time reversible. One advantage of our model is that the number of parameters that are introduced in the effort to include RSA and gene expression is small. The trade-off is that our model will incur extra computation because RSA values can be different from site to site. We need to have one rate matrix for each site, although these rate matrices can share parameters such as κ.

We set the relative fitnesses of i and j, respectively, as 1 and 1 + s. For an effective population of size N diploid individuals with multiplicative fitnesses, the scaled selection coefficient $S = 2Ns$ can be assessed by Refs. 44–46:

$$S = 2N\,s = \frac{1}{2}\cdot\log\left(\tau_{ij}\right). \tag{12}$$

To get reliable estimates of coefficients for MLRs in Equations 4 and 5, a sufficient number of proteins with both PDB structure and RNA-Seq data are required. Owing to insufficient multi-organ RNA-seq data from mouse, we conducted the LR and trained the evolutionary model only for humans. For this analysis, we use nucleotide frequencies observed in the human genome (41% GC base pairs and 59% AT base pairs according to International Human Genome Sequencing[47]) as estimates of π. This way of estimating π stems from the fact that the majority of the human genome consists of DNA with no known biological function and that is putatively evolving without much impact from natural selection. We therefore set $\pi_A = \pi_T = 0.59/2 = 0.295$ and $\pi_C = \pi_G = 0.41/2 = 0.205$. This treatment could be improved by allowing π to vary among genes to account for regional genomic differences in mutation patterns, but it is not pursued in this study.

Results

After using the Benjamini–Hochberg (BH) method[48] to control for false discovery rate, no synonymous codon group in mouse was found to exhibit a significant correlation between codon usage and RSA. In contrast, 15 out of 18 synonymous codon groups in human were statistically significant (Table 1). For humans, the null hypothesis cannot be rejected in synonymous codon groups belonging to cysteine (C), phenylalanine (F), and histidine (H).

When the same hypothesis was tested at the individual gene level, the P-value results can be found in Table 2. After applying the BH method, no synonymous codon group was found to be statistically significant in either human or mouse, although P-values in human are consistently smaller than those in mouse.

Figure 1 shows the hierarchical clustering result of the amino acid distance matrix computed for combined human data. As would be expected, if natural selection is associated with solvent accessibility, the hierarchical clustering suggests that RSA tendencies of amino acids in naturally occurring proteins are correlated with the physicochemical properties of these amino acids. On the whole, the 20 amino acids are naturally separated into two groups. The first group contains

Table 1. Test at genomic scale of whether RSA tendencies vary among synonymous codons.

SYNONYMOUS CODON GROUPS	AA	HUMAN *P*-VALUE	MOUSE *P*-VALUE
GCT, GCC, GCA, GCG	A	0.0003*	0.2473
TGT, TGC	C	0.9381	0.2330
GAT, GAC	D	0.0115*	0.4074
GAA, GAG	E	<0.0001*	0.5510
TTT, TTC	F	0.4852	0.3127
GGT, GGC, GGA, GGG	G	0.0153*	0.9709
CAT, CAC	H	0.2765	0.3875
ATT, ATC, ATA	I	<0.0001*	0.0703
AAA, AAG	K	0.0012*	0.0320
TTA, TTG, CTT, CTC, CTA, CTG	L	0.0057*	0.0975
AAT, AAC	N	0.0001*	0.0188
CCT, CCC, CCA, CCG	P	0.0016*	0.6651
CAA, CAG	Q	0.0058*	0.1058
CGT, CGC, CGA, CGG, AGA, AGG	R	0.0338*	0.2547
TCT, TCC, TCA, TCG, AGT, AGC	S	<0.0001*	0.0326
ACT, ACC, ACA, ACG	T	0.0034*	0.3692
GTT, GTC, GTA, GTG	V	0.0255*	0.8705
TAT, TAC	Y	0.0082*	0.2580

Notes: The first column contains synonymous codon groups, and the second column shows the corresponding amino acid type. The P-values of individual k-sample A-D tests are shown in the final two columns. To control for false discovery at level $\alpha = 0.05$, the BH method[48] was applied to the human P-values and then the mouse P-values. An * indicates significance at $\alpha = 0.05$.

Table 2. Test at individual gene scale of whether RSA tendencies vary among synonymous codons.

SYNONYMOUS CODON GROUPS	AA	HUMAN *P*-VALUE	MOUSE *P*-VALUE
GCT, GCC, GCA, GCG	A	0.0941	0.2033
TGT, TGC	C	0.2529	0.5224
GAT, GAC	D	0.4482	0.2829
GAA, GAG	E	0.9285	0.9118
TTT, TTC	F	0.4127	0.9098
GGT, GGC, GGA, GGG	G	0.0071	0.4487
CAT, CAC	H	0.2685	0.3649
ATT, ATC, ATA	I	0.1844	0.6461
AAA, AAG	K	0.1098	0.7943
TTA, TTG, CTT, CTC, CTA, CTG	L	0.3073	0.4867
AAT, AAC	N	0.5365	0.0688
CCT, CCC, CCA, CCG	P	0.3841	0.2274
CAA, CAG	Q	0.4883	0.7631
CGT, CGC, CGA, CGG, AGA, AGG	R	0.0730	0.4402
TCT, TCC, TCA, TCG, AGT, AGC	S	0.0466	0.4255
ACT, ACC, ACA, ACG	T	0.1324	0.2816
GTT, GTC, GTA, GTG	V	0.0713	0.5674
TAT, TAC	Y	0.5540	0.5869

Notes: The first column contains synonymous codon groups, and the second column shows the corresponding amino acid type. The *P*-values of individual *k*-sample A-D tests are shown in the final two columns. To control for false discovery at level $\alpha = 0.05$, the BH method[48] was applied to the human *P*-values and then the mouse *P*-values. No tests for either human or mouse were significant at $\alpha = 0.05$.

mainly hydrophobic amino acids (C, M, W, A, Y, F, V, I, L) and the second group contains the other amino acids (D, Q, R, P, N, T, H, G, S, K, E). According to the dendrogram, there are a few smaller and tighter clusters within each group. For example, in the first group, amino acids L, I, V, and F have more similar RSA preference than other amino acids in this group. In the second group, the cluster that contains amino acids K and E shows strong distinct signal, while two

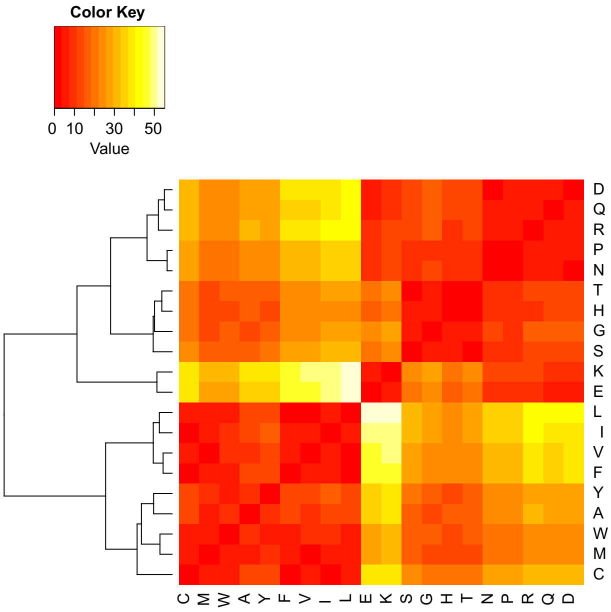

Figure 1. Distance matrix for amino acids using combined human data. Distance between two amino acid types is defined by the normalized KS test statistics (Equation 2). The complete-linkage method was used to perform hierarchical clustering.

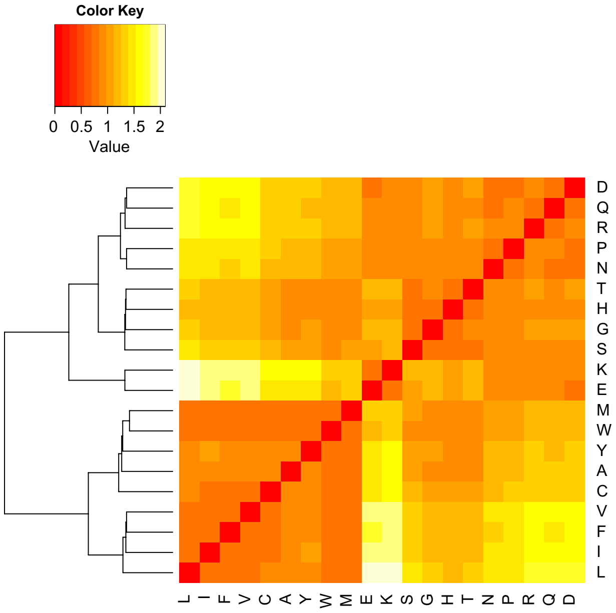

Figure 2. Distance matrix for amino acids using stratified human data. Distance between two amino acid types is defined by the normalized KS test statistics (Equation 2). The complete-linkage method was used to perform hierarchical clustering. Entries in the matrix are computed by taking the average of the distances between amino acids among all proteins.

other clusters (one formed by T, H, G, S and the other formed by D, Q, R, P, N) are also present.

As explained in the Materials and Methods section, we also constructed another distance matrix using data stratified by each gene. We computed one distance matrix for each protein in our data set and then calculated the average values for each cell in the distance matrix across all proteins. The resulting distance matrix is shown in Figure 2. The same clustering pattern was found as in Figure 1, although the signal is weaker because of the much smaller sample sizes in individual protein-coding genes.

MLR results. Using human data, the estimated coefficients of MLR models for synonymous codon groups with more than two codons are shown in Table 3. For groups with exactly two synonymous codons, estimated coefficients of LR models are in Table 4. The estimates of the slope parameter for log-scaled gene expression are all negative for LR models, and they are all statistically significant. Across all MLR models, 24 out of 32 slope estimates for gene expression are negative and statistically significant, while only one coefficient belonging to codon *TCG* in amino acid serine is positive and significant. Since we define the most frequent (also called

Table 3. Codons: MLR-estimated coefficients for gene expression using human data.

AA	REF CODON	NON-REF CODON	INTERCEPT COEF	EXPRESSION COEF	INTERCEPT *P*-VALUE	EXPRESSION *P*-VALUE
A	GCC	GCA	−0.3091	−0.1650	0.0002	<0.0001
		GCG	−1.6278	0.0331	<0.0001	0.3824
		GCT	−0.0943	−0.1363	0.2237	<0.0001
G	GGC	GGA	0.1279	−0.1639	0.1294	<0.0001
		GGG	−0.3834	−0.0039	<0.0001	0.8958
		GGT	−0.5043	−0.0909	<0.0001	0.0076
I	ATC	ATA	−0.6378	−0.2427	<0.0001	<0.0001
		ATT	0.0974	−0.1437	0.2063	<0.0001
L	CTG	CTA	−1.3565	−0.1966	<0.0001	<0.0001
		CTC	−0.5859	−0.0597	<0.0001	0.0089
		CTT	−0.6472	−0.2047	<0.0001	<0.0001
		TTA	−1.1326	−0.2292	<0.0001	<0.0001
		TTG	−0.9156	−0.1259	<0.0001	<0.0001
P	CCC	CCA	0.1390	−0.1315	0.1450	0.0001
		CCG	−1.4229	0.0621	<0.0001	0.1675
		CCT	0.1333	−0.1263	0.1619	0.0001
R	CGG	AGA	0.3689	−0.1976	0.0012	<0.0001
		AGG	0.2032	−0.1231	0.0754	0.0013
		CGA	0.0133	−0.2373	0.9170	<0.0001
		CGC	0.0172	−0.0124	0.8795	0.7315
		CGT	−0.6141	−0.0978	<0.0001	0.0428
S	AGC	AGT	−0.2386	−0.1159	0.0157	0.0012
		TCA	−0.3185	−0.1327	0.0018	0.0004
		TCC	−0.1431	0.0058	0.1140	0.8499
		TCG	−1.7424	0.1118	<0.0001	0.0167
		TCT	−0.0936	−0.1069	0.3207	0.0016
T	ACC	ACA	−0.0842	−0.0641	0.3423	0.0341
		ACG	−1.1397	0.0214	<0.0001	0.5834
		ACT	−0.1699	−0.1289	0.0705	0.0001
V	GTG	GTA	−0.9888	−0.2038	<0.0001	<0.0001
		GTC	−0.6873	−0.0132	<0.0001	0.6153
		GTT	−0.6146	−0.1727	<0.0001	<0.0001

Notes: The first column denotes amino acid types for synonymous codon groups. Within each synonymous codon group, the most frequent codon for each amino acid is chosen as the reference category, and these codons are shown in the second column. The non-reference codons are listed in the third column. The fourth and fifth columns contain the estimated coefficients for intercept and gene expression in Equation 4. The corresponding *P*-values of the estimated coefficients can be found in the last two columns.

Table 4. Codons: LR-estimated coefficients for gene expression using human data.

AA	REF CODON	NON-REF CODON	INTERCEPT COEF	EXPRESSION COEF	INTERCEPT *P*-VALUE	EXPRESSION *P*-VALUE
C	TGC	TGT	0.1474	−0.1109	0.1622	0.0035
D	GAC	GAT	0.0650	−0.0985	0.3275	<0.0001
E	GAG	GAA	0.1872	−0.1632	0.0018	<0.0001
F	TTC	TTT	0.1959	−0.1009	0.0091	0.0001
H	CAC	CAT	−0.1216	−0.1132	0.2212	0.0018
K	AAG	AAA	−0.0871	−0.1144	0.1714	<0.0001
N	AAC	AAT	0.0934	−0.1340	0.2320	<0.0001
Q	CAG	CAA	−0.6138	−0.1920	<0.0001	<0.0001
Y	TAC	TAT	−0.1045	−0.0688	0.2184	0.0213

Notes: The first column denotes amino acid types for synonymous codon groups. Within each synonymous codon group, the most frequent codon for each amino acid is chosen as the reference category, and these codons are shown in the second column. The non-reference codons are listed in the third column. The fourth and fifth columns contain the estimated coefficients for intercept and gene expression in Equation 4. The corresponding *P*-values of the estimated coefficients can be found in the last two columns.

preferred) codon to be the reference category in both MLR and LR cases, these estimates coincide with findings from earlier studies[21,24,49] that the probability of observing the most frequent codon increases when gene expression level is higher (with one exception).

A likelihood ratio test suggests that the full model for amino acid probability (Equation 5) is significantly better than the model considering only RSA (Equation 6; *P*-value ≐ 4.1e–09) and the model considering only the level of gene expression (Equation 7; *P*-value ≐ 2.2e–16). Table 5 summarizes the maximum-likelihood estimates of the parameters and their *P*-values in the full model. With the exception of three amino acids (cysteine, phenylalanine, and isoleucine), slope estimates for RSA of the non-reference amino acids are

Table 5. Amino acids: MLR-estimated coefficients for RSA and gene expression using human data.

AA	INTERCEPT		RSA		EXPRESSION	
	COEF	*P*-VALUE	COEF	*P*-VALUE	COEF	*P*-VALUE
A	−0.9449	<0.0001	0.0208	<0.0001	0.0378	0.0065
C	−1.2789	<0.0001	−0.0080	<0.0001	−0.0255	0.2123
D	−1.9803	<0.0001	0.0479	<0.0001	0.0320	0.0292
E	−2.0101	<0.0001	0.0523	<0.0001	0.0478	0.0006
F	−0.7462	<0.0001	−0.0034	0.0105	−0.0020	0.9009
G	−1.3812	<0.0001	0.0354	<0.0001	0.0309	0.0283
H	−2.0176	<0.0001	0.0322	<0.0001	−0.0128	0.5008
I	−0.5001	<0.0001	−0.0057	<0.0001	−0.0282	0.0634
K	−2.2285	<0.0001	0.0537	<0.0001	0.0526	0.0003
M	−1.9446	<0.0001	0.0096	<0.0001	0.0381	0.0813
N	−2.0827	<0.0001	0.0455	<0.0001	−0.0246	0.1393
P	−2.0514	<0.0001	0.0455	<0.0001	0.0295	0.0571
Q	−2.1544	<0.0001	0.0464	<0.0001	0.0176	0.2721
R	−1.9454	<0.0001	0.0441	<0.0001	0.0373	0.0136
S	−1.3001	<0.0001	0.0366	<0.0001	−0.0126	0.3740
T	−1.5064	<0.0001	0.0329	<0.0001	0.0305	0.0421
V	−0.5067	<0.0001	0.0034	0.0020	0.0304	0.0263
W	−1.9197	<0.0001	0.0022	0.2615	−0.0159	0.5218
Y	−1.2309	<0.0001	0.0109	<0.0001	0.0019	0.9128

Notes: The first column shows amino acid types in their one-letter format. The second and third columns contain estimated values and their corresponding *P*-values of the intercept (see Equation 5). The fourth and fifth columns show estimated coefficients and their *P*-values of RSA in MLR. The last two columns show coefficient estimates and *P*-values of log-transformed gene expression. The most frequent amino acid leucine (L) is not included in the table because it is chosen as the reference category for MLR.

positive. This means that the relative probability of observing these amino acids rather than the reference amino acid leucine is higher as RSA increases, holding gene expression identical. To depict a full picture of the joint effect of RSA and gene expression, heatmaps present the predicted probabilities of each amino acid (Fig. 3 and Supplementary Fig. 1). While RSA plays a dominant role in determining the region of high probabilities for amino acids such as leucine and proline, the effect of gene expression is evident in cases such as histidine and serine (Fig. 3).

Scaled selection coefficients. By applying Equation 12, we estimated the scaled selection coefficient $S = 2N s$ of all possible point mutations at each site for the 241 protein-coding human genes. The estimated values fall into two groups defined by nonsynonymous and synonymous substitutions. Figure 4 depicts how the distribution of scaled selection coefficients for genes changes with the corresponding gene expression levels. We can see from the figure that when mutations are being selected against ($S < 0$), both nonsynonymous and synonymous mutations are inferred to be more deleterious as the gene expression level increases. The figure also shows that the degree of selective advantage for beneficial mutations gets larger with increasing gene expression. As expected, Figure 4 indicates that nonsynonymous mutations are influenced by gene expression to a lesser degree than synonymous ones.

Discussion

We developed a probabilistic framework that simultaneously considers RSA (a structural constraint) and gene expression (a functional constraint). By design, the two-step construction process of this framework (Equation 1) captures codon usage bias at the nucleotide level as well as structural and functional dependence of amino acids.

Hypothesis testing at a genomic scale reveals a potentially species-specific difference in the relationship between synonymous codon choice and RSA. In humans, the null hypothesis that RSA is independent of synonymous codon choice can be rejected. In mouse, the result is not statistically significant, possibly because of insufficient data and possibly because the relationship between RSA and synonymous codon usage is less strong. At the individual gene level, we test the same hypothesis using permutation and resampling. Although we fail to reject the null hypothesis in both human and mouse, we notice that the association between RSA and codon usage

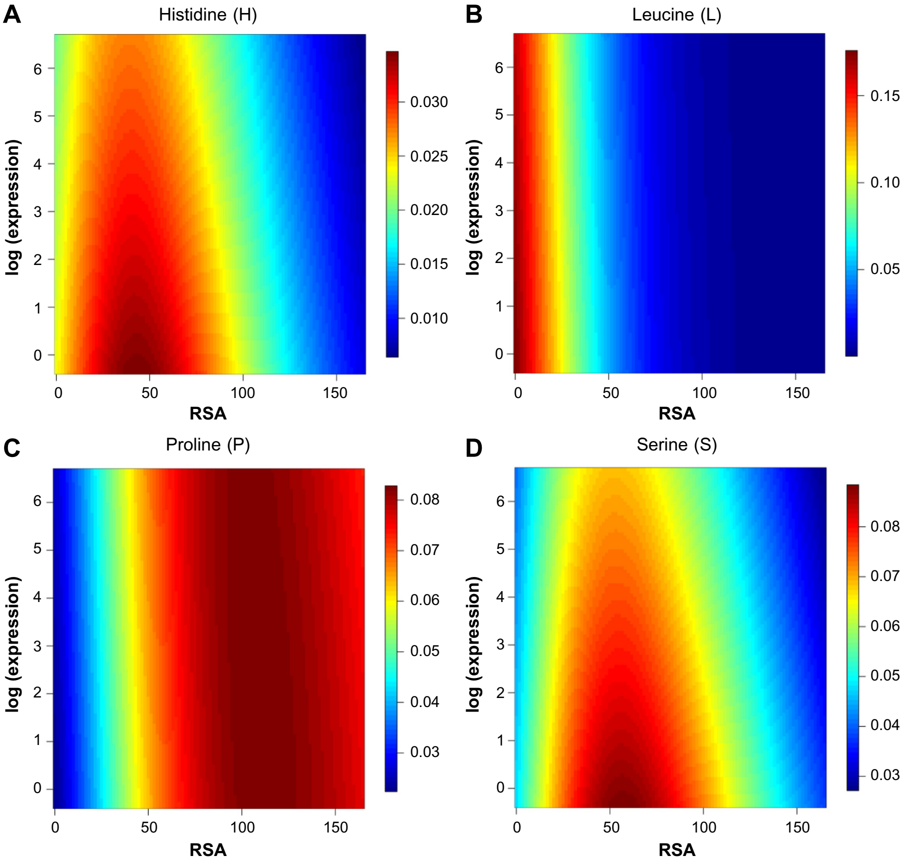

Figure 3. Predicted probabilities of four amino acids across possible ranges of RSA and gene expression. Probabilities for each amino acid were calculated from the estimated RSA and gene expression coefficients (see Table 5). The horizontal axis covers the possible range of RSA, and the vertical axis is for gene expression. Red regions indicate relatively high probabilities, while blue regions show low probabilities. The four amino acids depicted here (histidine, leucine, proline, and serine) were selected for the diversity of heatmap diagrams that they represent. Heatmaps for the other 16 amino acids can be found in Supplementary Figure 1.

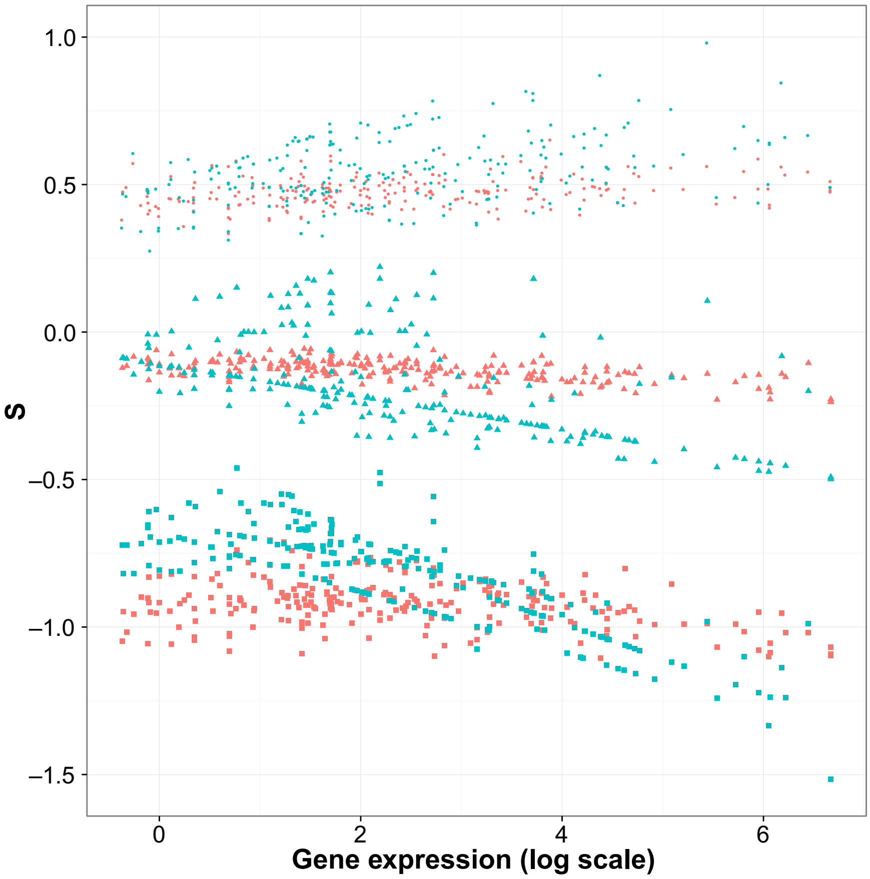

Figure 4. Comparison of $S = 2N\,s$ estimates between nonsynonymous and synonymous point mutations to human genes. For each protein-coding human gene, estimates of S were obtained for possible nonsynonymous (red) and synonymous (blue) point mutations at each site. The x-axis represents the logarithm of the geometric mean across tissues of expression measurements for a gene. The y-axis represents inferred scaled selection coefficients (S). For each gene, the scaled selection coefficient distribution among point mutations is summarized by the 5th percentile (square shape), the 50th percentile (triangular shape), and the 95th percentile (round shape).

is stronger in human than in mouse. This is consistent with our finding at the genomic scale. Because we derive one test statistic from each gene and combine the test statistics to get an overall test statistic and its null distribution, we are treating each protein equally. However, because of small sample size, the test statistics calculated in some proteins may not be informative and can contribute stochastic noise to the combined test statistic. Improved test statistics might yield different conclusions.

In the first step of constructing our probabilistic framework, we used MLR and LR to model the probabilities of different synonymous codons, given gene expression. Slope estimates suggest that, in general, it is more likely to observe the preferred codon in highly expressed genes. This is consistent with previous findings.[21,24,49] RSA is ignored in this step because our analyses indicate that the correlation between RSA and codon usage is small to nonexistent. This does not mean that there is no structure constraint on codon usage. Protein structure may lead to selection pressure on synonymous codon choice through an interaction between the translation process and protein folding.[50] Codon usage is also subject to other constraints because it can affect splicing and/or mRNA stability.[51]

To predict amino acid types, we incorporate both RSA and gene expression as independent variables in our MLR model. Our likelihood ratio test confirms that the full model with both variables fits data significantly better than reduced models. Although the impact of gene expression on amino acid probabilities is weaker than the impact of RSA, it is still significant. The heatmaps for amino acids such as histidine and serine show that these amino acids reach their highest probabilities at intermediate RSA values (Fig. 3).

We observed a trend that $2N\,s$ estimates for both nonsynonymous and synonymous point mutations become more negative as gene expression increases (Fig. 4). This observation is biologically reasonable, and a few mechanisms might contribute to it. Highly expressed proteins have been found to evolve slower because of the stronger selection pressure that they experience.[4,13,52] Many highly expressed proteins are functionally important or even essential. For these proteins, the cell cannot afford their function to be disturbed or their abundance to become low.[53]

Some aspects of the scaled selection coefficients are qualitatively reasonable. For example, 64.2% of the possible point mutations yield negative values of $2N s$. This is in keeping with the expectation that if protein-coding genes are affected by natural selection, more possible point mutations should be deleterious than advantageous. The qualitative pattern remains when we restrict consideration to synonymous point mutations. Of these mutations, 64.0% yield negative values. Among nonsynonymous mutations, 64.3% of possible point mutations are inferred to be deleterious.

However, the distribution of nonsynonymous scaled selection coefficient estimates is clearly unrealistic in that it is too tightly clustered around the value of 0. Because some possible nonsynonymous mutations are lethal or at least highly deleterious, there should be some scaled selection coefficient values that are far below 0. We do not observe this expected long lower tail of the distribution of scaled selection coefficients for nonsynonymous point mutations (eg, see the 5% nonsynonymous values in Fig. 4). Because some nonsynonymous mutations should be extremely deleterious, we also expect the average over all possible nonsynonymous mutations to be substantially below zero. We do not observe this. In fact, the average estimates of $2N s$ for nonsynonymous mutations is about −0.162, and this is only slightly less than the −0.158 value represented by the average $2N s$ estimate among synonymous point mutations. Similar shortcomings have been noted for distributions of scaled selection coefficients that have been inferred via a mutation-selection balance model of molecular evolution.[44,54] Much of the mismatch between our expectations and our estimates is likely because of flaws of our evolutionary model. Specifically, the rates in our model depend on the RSA and gene expression covariates, but other aspects of phenotype are clearly very relevant to natural selection and are not captured by our model.

An additional possible weakness of our mutation-selection model and other mutation-selection models is the assumption that each new mutation is fixed or lost before the next one occurs. Our inferential framework does not accommodate the possibility that fitness-affecting genetic variants at one locus interfere with the survival or loss of fitness-affecting variants at other loci. This Hill–Robertson phenomenon[55] is potentially important to consider when connecting interspecific models of sequence change to population genetics.[56]

A basic assumption of our probabilistic framework is that sites in a protein-coding sequence are independent. This assumption is commonly made, and it allows simplification of computation, even though it is clear that sites in a protein sequence do not evolve independently. It would be challenging yet worthwhile for future evolutionary studies to include structural constraints related to site dependence in a coding sequence.

Acknowledgments

We thank four anonymous reviewers for their comments.

Author Contributions

Conceived and designed the experiments: KW, SY, XJ, JLT. Analyzed the data: KW, SY, XJ, CL, AG. Wrote the first draft of the manuscript: KW, JLT. Contributed to the writing of the manuscript: KW, JLT. Agree with manuscript results and conclusions: All authors. Jointly developed the structure and arguments for the paper: All authors. Made critical revisions and approved final version: KW, JLT. All authors reviewed and approved of the final manuscript.

Supplementary Material

Supplementary Figure 1. Predicted probabilities of 16 amino acids across possible ranges of RSA and gene expression.

REFERENCES

1. DePristo MA, Weinreich DM, Hartl DL. Missense meanderings in sequence space: a biophysical view of protein evolution. *Nat Rev Genet*. 2005;6(9):678–87.
2. Franzosa EA, Xia Y. Structural determinants of protein evolution are context-sensitive at the residue level. *Mol Biol Evol*. 2009;26(10):2387–95.
3. Pál C, Papp B, Hurst LD. Highly expressed genes in yeast evolve slowly. *Genetics*. 2001;158(2):927–31.
4. DrummondDA,BloomJD,AdamiC,WilkeCO,ArnoldFH.Whyhighlyexpressed proteins evolve slowly. *Proc Natl Acad Sci USA*. 2005;102(40):14338–43.
5. Drummond DA, Raval A, Wilke CO. A single determinant dominates the rate of yeast protein evolution. *Mol Biol Evol*. 2006;23(2):327–37.
6. Parisi G, Echave J. Structural constraints and emergence of sequence patterns in protein evolution. *Mol Biol Evol*. 2001;18(5):750–6.
7. Robinson DM, Jones DT, Kishino H, Goldman N, Thorne JL. Protein evolution with dependence among codons due to tertiary structure. *Mol Biol Evol*. 2003;20(10):1692–704.
8. Rodrigue N, Philippe H, Lartillot N. Assessing site-interdependent phylogenetic models of sequence evolution. *Mol Biol Evol*. 2006;23(9):1762–75.
9. Kleinman CL, Rodrigue N, Bonnard C, Philippe H, Lartillot N. A maximum likelihood framework for protein design. *BMC Bioinformatics*. 2006;7(1):326.
10. Rodrigue N, Kleinman CL, Philippe H, Lartillot N. Computational methods for evaluating phylogenetic models of coding sequence evolution with dependence between codons. *Mol Biol Evol*. 2009;26(7):1663–76.
11. Kleinman CL, Rodrigue N, Lartillot N, Philippe H. Statistical potentials for improved structurally constrained evolutionary models. *Mol Biol Evol*. 2010;27(7):1546–60.
12. Agashe D, Martinez-Gomez NC, Drummond DA, Marx CJ. Good codons, bad transcript: large reductions in gene expression and fitness arising from synonymous mutations in a key enzyme. *Mol Biol Evol*. 2013;30(3):549–60.
13. Drummond DA, Wilke CO. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell*. 2008;134(2):341–52.
14. Zhou T, Weems M, Wilke CO. Translationally optimal codons associate with structurally sensitive sites in proteins. *Mol Biol Evol*. 2009;26(7):1571–80.
15. Sharp PM, Averof M, Lloyd AT, Matassi G, Peden JF. DNA sequence evolution: the sounds of silence. *Philos Trans R Soc Lond B Biol Sci*. 1995;349(1329):241–7.
16. Chen SL, Lee W, Hottes AK, Shapiro L, McAdams HH. Codon usage between genomes is constrained by genome-wide mutational processes. *Proc Natl Acad Sci USA*. 2004;101(10):3480–5.
17. Jia W, Higgs PG. Codon usage in mitochondrial genomes: distinguishing context-dependent mutation from translational selection. *Mol Biol Evol*. 2008;25(2):339–51.
18. Higgs PG, Ran W. Coevolution of codon usage and tRNA genes leads to alternative stable states of biased codon usage. *Mol Biol Evol*. 2008;25(11):2279–91.
19. Hershberg R, Petrov DA. Selection on codon bias. *Annu Rev Genet*. 2008;42(1):287–99.
20. Post LE, Nomura M. Nucleotide sequence of the intercistronic region preceding the gene for RNA polymerase subunit alpha in *Escherichia coli*. *J Biol Chem*. 1979;254(21):10604–6.
21. Ikemura T. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. *J Mol Biol*. 1981;151(3):389–409.
22. Ikemura T. Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes: differences in synonymous codon choice patterns of yeast and *Escherichia coli* with reference to the abundance of isoaccepting transfer RNAs. *J Mol Biol*. 1982;158(4):573–97.

23. Moriyama EN, Powell JR. Codon usage bias and tRNA abundance in drosophila. *J Mol Evol*. 1997;45(5):514–23.
24. Duret L. tRNA gene number and codon usage in the *C. elegans* genome are co-adapted for optimal translation of highly expressed genes. *Trends Genet*. 2000;16(7):287–9.
25. Sharp PM, Bailes E, Grocock RJ, Peden JF, Sockett RE. Variation in the strength of selected codon usage bias among bacteria. *Nucleic Acids Res*. 2005; 33(4):1141–53.
26. Berman HM, Westbrook J, Feng Z, et al. The protein data bank. *Nucleic Acids Res*. 2000;28(1):235–42.
27. Pruitt KD, Harrow J, Harte RA, et al. The consensus coding sequence (CCDS) project: identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res*. 2009;19(7):1316–23.
28. Smith TF, Waterman MS. Identification of common molecular subsequences. *J Mol Biol*. 1981;147(1):195–7.
29. Rice P, Longden I, Bleasby A. EMBOSS: the European molecular biology open software suite. *Trends Genet*. 2000;16(6):276–7.
30. Hubbard S, Thornton J. *Naccess v2.1.1*. 1996. Available at: http://www.bioinf.manchester.ac.uk/naccess.
31. Brawand D, Soumillon M, Necsulea A, et al. The evolution of gene expression levels in mammalian organs. *Nature*. 2011;478(7369):343–8.
32. Petryszak R, Burdett T, Fiorelli B, et al. Expression atlas update – a database of gene and transcript expression from microarray- and sequencing-based functional genomics experiments. *Nucleic Acids Res*. 2014;42(D1):D926–32.
33. Wu CH, Yeh L-SL, Huang H, et al. The protein information resource. *Nucleic Acids Res*. 2003;31(1):345–7.
34. Ramsey DC, Scherrer MP, Zhou T, Wilke CO. The relationship between relative solvent accessibility and evolutionary rate in protein evolution. *Genetics*. 2011;188(2):479–88.
35. Yeh S-W, Liu J-W, Yu S-H, Shih C-H, Hwang J-K, Echave J. Site-specific structural constraints on protein sequence evolutionary divergence: local packing density versus solvent exposure. *Mol Biol Evol*. 2013;31(1):135–9.
36. Herbeck JT. Gene expression level influences amino acid usage, but not codon usage, in the tsetse fly endosymbiont Wigglesworthia. *Microbiology*. 2003;149(9):2585–96.
37. Schaber J, Rispe C, Wernegreen J, et al. Gene expression levels influence amino acid usage and evolutionary rates in endosymbiotic bacteria. *Gene*. 2005; 352:109–17.
38. Scholz FW, Stephens MA. K-sample Anderson–Darling tests. *J Am Stat Assoc*. 1987;82(399):918–24.
39. Kim H, Park H. Prediction of protein relative solvent accessibility with support vector machines and long-range interaction 3D local descriptor. *Proteins*. 2004;54(3):557–62.
40. Pauwels EJ, Frederix G. Image segmentation by nonparametric clustering based on the Kolmogorov–Smirnov distance. In: Vernon D, ed. *Computer Vision–ECCV 2000, Number 1843 in Lecture Notes in Computer Science*. Berlin: Springer; 2000:85–99.
41. Volkovich Z, Kirzhner V, Barzily Z. *Genome Clustering: From Linguistic Models to Classification of Genetic Texts*. Springer Science & Business Media; 2010.
42. Zhang X, Boutros M. A novel phenotypic dissimilarity method for image-based high-throughput screens. *BMC Bioinformatics*. 2013;14(1):336.
43. Sorensen T. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. In *Biologiske Skrifter*. I kommission hos E. Munksgaard. 1948.
44. Choi SC, Redelings BD, Thorne JL. Basing population genetic inferences and models of molecular evolution upon desired stationary distributions of DNA or protein sequences. *Philos Trans R Soc B Biol Sci*. 2008;363:3931–39.
45. Halpern AL, Bruno WJ. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Mol Biol Evol*. 1998;15(7):910–7.
46. Yang Z, Nielsen R. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Mol Biol Evol*. 2008;25(3):568–79.
47. Consortium IHGS. Initial sequencing and analysis of the human genome. *Nature*. 2001;409(6822):860–921.
48. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B Stat Methodol*. 1995;57(1):289–300.
49. Hiraoka Y, Kawamata K, Haraguchi T, Chikashige Y. Codon usage bias is correlated with gene expression levels in the fission yeast schizosaccharomyces pombe. *Genes Cells*. 2009;14(4):499–509.
50. Scherrer MP, Meyer AG, Wilke CO. Modeling coding-sequence evolution within the context of residue solvent accessibility. *BMC Evol Biol*. 2012;12(1):179.
51. Chamary JV, Parmley JL, Hurst LD. Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat Rev Genet*. 2006;7(2):98–108.
52. Serohijos AWR, Rimas Z, Shakhnovich EI. Protein biophysics explains why highly abundant proteins evolve slowly. *Cell Reports*. 2012;2(2):249–56.
53. Sikosek T, Bornberg-Bauer E, Chan HS. Evolutionary dynamics on protein bistability landscapes can potentially resolve adaptive conflicts. *PLoS Comput Biol*. 2012;8(9):e1002659.
54. Thorne JL, Choi SC, Yu J, Higgs PG, Kishino H. Population genetics without intraspecific data. *Mol Biol Evol*. 2007;24(8):1667–77.
55. Hill WG, Robertson A. The effect of linkage on limits to artificial selection. *Genet Res*. 1966;8(03):269–94.
56. Cartwright RA, Lartillot N, Thorne JL. History can matter: non-Markovian behavior of ancestral lineages. *Syst Biol*. 2011;60(3):276–90.