

The protein domains of vertebrate species in which selection is more effective have greater intrinsic structural disorder

Reviewed Preprint

v2 • July 16, 2024

Revised by authors

Reviewed Preprint

v1 • July 26, 2023

Catherine A Weibel, Andrew L Wheeler, Jennifer E James, Sara M Willis, Hanon McShea, Joanna Masel 

Department of Mathematics, University of Arizona, Tucson, Arizona 85721, USA • Department of Physics, University of Arizona, Tucson, Arizona 85721, USA • Department of Applied Physics, Stanford University, California, USA • Genetics Graduate Interdisciplinary Program, University of Arizona, Tucson, Arizona 85721, USA • Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona 85721, USA • Department of Ecology and Genetics, Evolutionary Biology Center, Uppsala University, Sweden • University Information Technology Services, University of Arizona, Tucson, Arizona 85721, USA • Department of Earth System Science, Stanford University

 https://en.wikipedia.org/wiki/Open_access

 Copyright information

Abstract

The nearly neutral theory of molecular evolution posits variation among species in the effectiveness of selection. In an idealized model, the census population size determines both this minimum magnitude of the selection coefficient required for deleterious variants to be reliably purged, and the amount of neutral diversity. Empirically, an “effective population size” is often estimated from the amount of putatively neutral genetic diversity and is assumed to also capture a species’ effectiveness of selection. A potentially more direct measure of the effectiveness of selection is the degree to which selection maintains preferred codons. However, past metrics that compare codon bias across species are confounded by among-species variation in %GC content and/or amino acid composition. Here we propose a new Codon Adaptation Index of Species (CAIS), based on Kullback-Leibler divergence, that corrects for both confounders. We demonstrate the use of CAIS correlations, as well as the Effective Number of Codons, to show that the protein domains of more highly adapted vertebrate species evolve higher intrinsic structural disorder

eLife assessment

This study develops a **useful** metric for quantifying codon usage adaptation - the Codon Adaptation Index of Species (CAIS). This metric permits direct comparisons of the strength of selection at the molecular level across species. The study is based on **solid** evidence, and the authors identify relationships between CAIS and the presence of disordered protein domains. Other correlations, such as the one between CAIS and body size, are weak and non-significant. In summary, the study introduces an interesting new approach to quantifying codon usage across species, which may be helpful in attempts to measure selection at the molecular level.

<https://doi.org/10.7554/eLife.87335.2.sa2>

Introduction

Species differ from each other in many ways, including mating system, ploidy, spatial distribution, life history, size, lifespan, genome size, mutation rate, selective pressure, and population size. These differences make the process of purifying selection more efficient in some species than others. Our understanding of both the causes and consequences of these differences is limited in part by a reliable metric with which to measure them. In the long term, the probability that a gene is fixed for one allele rather than another allele is given by the ratio of fixation and counter-fixation probabilities (Bulmer 1991 [↗](#)). In an idealized population of constant population size and no selection at linked sites, a mutation-selection-drift model describes how this ratio of fixation probabilities depends on the census population size N (Kimura 1962 [↗](#)), and hence gives the fraction of sites expected to be found in preferred vs. non-preferred states (**Figure 1** [↗](#)).

This reasoning has been extended to real populations by positing that species have an “effective” population size, N_e (Ohta 1973 [↗](#)). N_e is the census population size of an idealized population that reproduces a property of interest in the focal population. N_e is therefore not a single quantity per population, but instead depends on which property is of interest.

The amount of neutral polymorphism is the usual property used to empirically estimate N_e (Charlesworth 2009 [↗](#); Doyle et al. 2015 [↗](#); Lynch et al. 2016 [↗](#)). However, the property of most relevance to nearly neutral theory is instead the inflection point s at which non-preferred alleles become common enough to matter (**Figure 1** [↗](#)), and hence the degree to which highly exquisite adaptation can be maintained in the face of ongoing mutation and genetic drift (Kimura 1962 [↗](#); Ohta 1972 [↗](#); Ohta 1992 [↗](#)). While genetic diversity has been found to reflect some aspects of life history strategy (Romiguier et al. 2014 [↗](#)), there remain concerns about whether neutral genetic diversity and the limits to weak selection always remain closely coupled in non-equilibrium settings.

As a practical matter, N_e is usually calculated by dividing some measure of the amount of putatively neutral (often synonymous) polymorphism segregating in a population by that species’ mutation rate (Charlesworth 2009 [↗](#)). As a result, N_e values are only available for species that have both polymorphism data and accurate mutation rate estimates, limiting their use. Worse, N_e is not a robust statistic. In the absence of a clear species definition, polymorphism is sometimes calculated across too broad a range of genomes, substantially inflating N_e (Daubin & Moran 2004 [↗](#)); a poor sampling scheme can have the converse effect of deflating genetic diversity. Transient hypermutation (Plotkin et al. 2006 [↗](#)), which is common in microbes, causes further short-term inconsistencies in polymorphism levels. Perhaps most importantly, a recent bottleneck

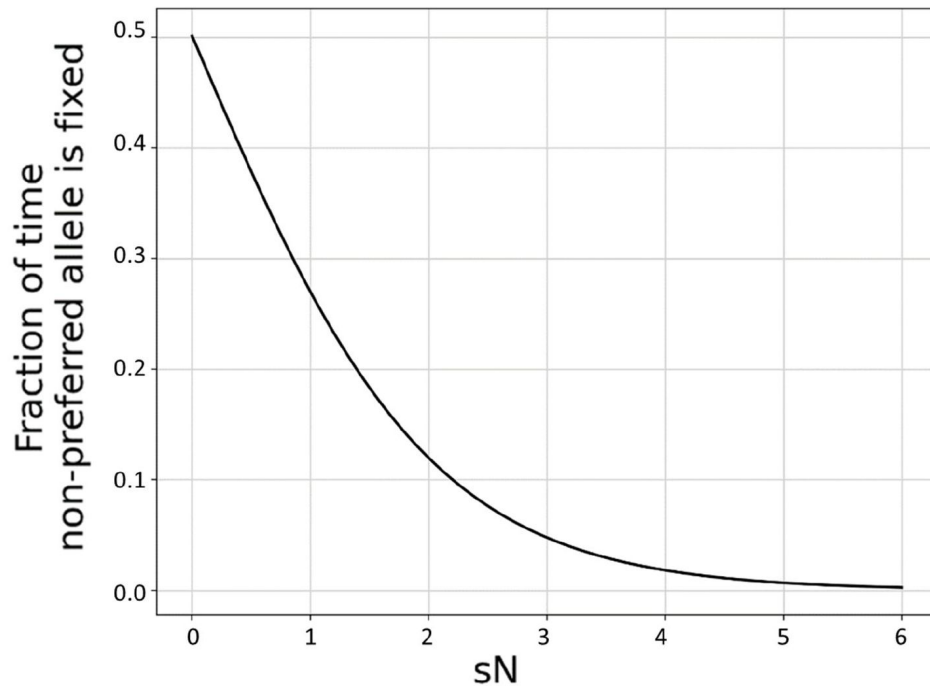


Figure 1.

The effectiveness of selection, calculated as the long-term ratio of time spent in fixed deleterious : fixed beneficial allele states given symmetric mutation rates, is a function of the product sN . Assuming a diploid Wright-Fisher population with $s \ll 1$, the probability of fixation of a new mutation $\pi(N, s) = \frac{1 - e^{-\frac{s}{2}}}{1 - e^{-Ns}}$, and the y-axis is calculated as $\pi(N, -s)/(\pi(N, -s) + \pi(N, s))$. s is held constant at a value of 0.001 and N is varied. Results for other small magnitude values of s are superimposable. For small sN , selection is ineffective at producing codon bias. For large sN , selection is highly effective. For only a relatively narrow range of intermediate values of sN , the degree of codon bias depends quantitatively on sN .

will deflate N_e based on the coalescence, even if too brief to lead to significant erosion of fine-tuned adaptations. But drift barrier theory concerns the level with which adaptation is fine-tuned, and so a better metric would capture that directly, rather than indirectly rely on neutral diversity.

An alternative approach to measure the efficiency of selection exploits codon usage bias, which is influenced by weak selection for factors such as translational speed and accuracy (Hershberg & Petrov 2008 [↗](#); Plotkin & Kudla 2011 [↗](#); Hunt et al. 2014 [↗](#)). The degree of bias in synonymous codon usage that is driven by selective preference offers a more direct way to assess how effective selection is at the molecular level in a given species (Li 1987 [↗](#); Bulmer 1991 [↗](#); Akashi 1996 [↗](#); Subramanian 2008 [↗](#)). Conveniently, it can be estimated from only a single genome, i.e., without polymorphism or mutation rate data for that species.

One commonly used metric, the Codon Adaptation Index (CAI) (Sharp & Li 1987 [↗](#); Sharp et al. 2010 [↗](#)) takes the average of Relative Synonymous Codon Usage (RSCU) scores, which quantify how often a codon is used, relative to the codon that is most frequently used to encode that amino acid in that species. While this works well for comparing genes within the same species, it unfortunately means that the species-wide strength of codon bias appears in the normalizing denominator (see equation 4 [↗](#) below and **Supplementary Figure 1A** [↗](#)). Paradoxically, this can make more exquisitely adapted species have lower rather than higher species-averaged CAI scores (**Supplementary Figure 1B** [↗](#); Rocha 2004 [↗](#); Botzman & Margalit 2011 [↗](#)).

To compare species using CAI, it has been suggested that instead of taking a genome-wide average, one should consider a set of highly expressed reference genes (Sharp et al. 2005 [↗](#); Vicario et al. 2007 [↗](#); Subramanian 2008 [↗](#); dos Reis & Wernisch 2009 [↗](#)). This approach assumes that the relative strength of selection on those reference genes (often a function of gene expression) remains approximately constant across the set of species considered (red distributions in **Figure 2** [↗](#)). Its use also requires careful attention to the length of reference genes (Urrutia & Hurst 2001 [↗](#); Doherty & McInerney 2013 [↗](#)), and some approaches also require information about tRNA gene copy numbers and abundances (dos Reis & Wernisch 2009 [↗](#)).

Since codon bias varies quantitatively within only a small range of sN (**Figure 1** [↗](#)), a promising approach is to measure the proportion of sites at which codon adaptation is effective. We posit that more highly adapted species have a higher proportion of both genes and sites subject to effective selection on codon bias (**Figure 2** [↗](#); Galtier et al. 2018 [↗](#)). Indeed, CAI might also rely in part on variation in the fraction of sites within the reference genes that is subject to effective selection as a function of species (**Figure 2** [↗](#), red). Here we take this logic further, considering all sites in a proteome-wide approach. Averaging across the entire proteome provides robustness to shifts in the expression level of or strength of selection on particular genes. The proteome-wide average depends on the fraction of sites whose selection coefficients exceed the “drift barrier” for that particular species (**Figure 2** [↗](#), blue threshold).

In estimating the effects of selection, it is critical to control for other causes of codon bias. In particular, species differ in their mutational bias with respect to %GC, and in the frequency of GC-biased gene conversion (Urrutia & Hurst 2001 [↗](#); Duret & Galtier 2009 [↗](#); Doherty & McInerney 2013 [↗](#); Figuet et al. 2014 [↗](#)). Here, we control for %GC, capturing species differences both in mutation and in gene conversion, by calculating the Kullback-Leibler divergence of the observed codon frequencies away from the codon frequencies that we would expect to see given the genomic %GC content of the species. Kullback-Leibler divergence measures the distance of an observed probability distribution from an expected reference distribution given %GC, capturing a measure of surprise (Kullback & Leibler 1951 [↗](#)). This method does not require us to specify preferred vs. non-preferred codons, and can thus also accommodate situations in which different genes have different codon preferences (Gingold et al. 2014 [↗](#); Cope et al. 2018 [↗](#)).

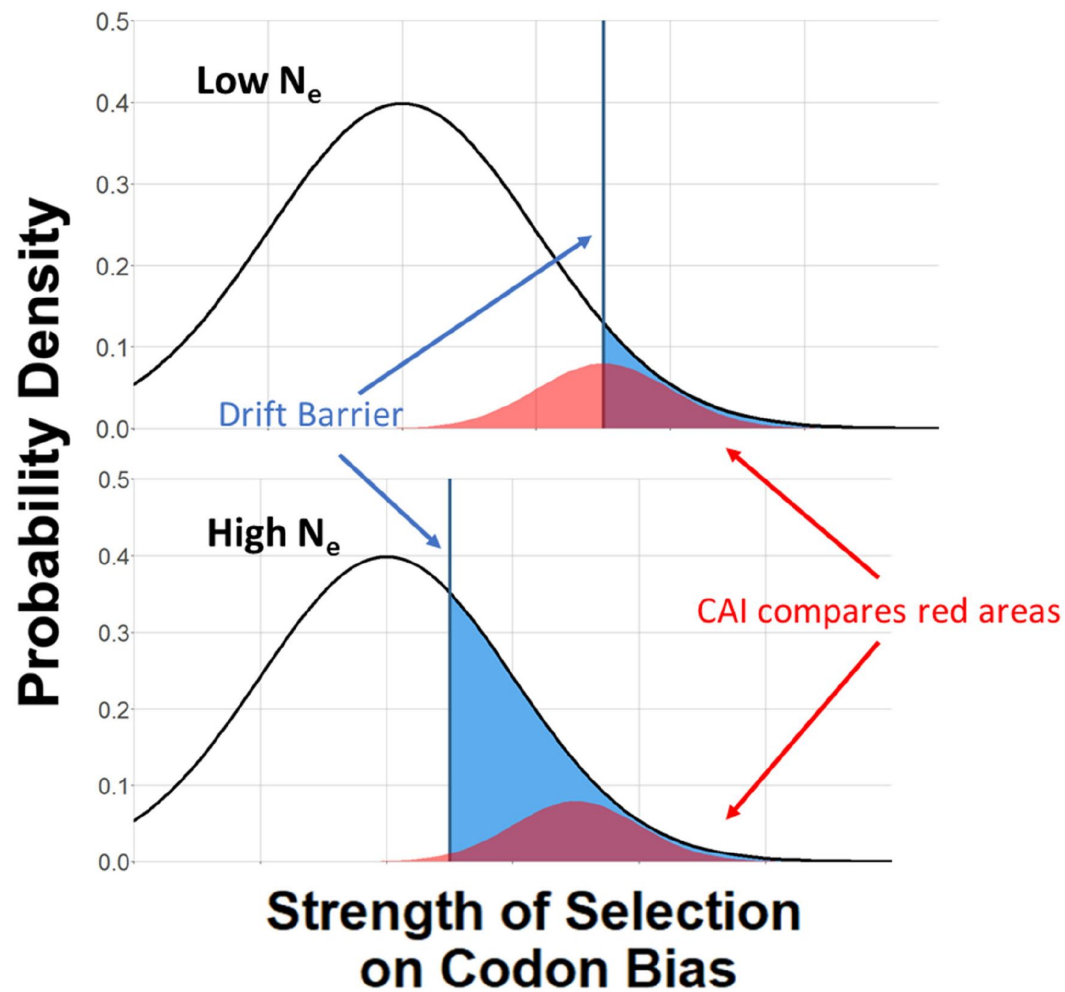


Figure 2.

More highly adapted species (bottom) have a higher proportion of their sites subject to effective selection on codon bias (blue area). The CAI attempts to compare the intensity of selection (Figure 1, x-axis) in a subset of genes under strong selection (red areas). Given the narrow range of quantitative dependence of codon bias on sN shown in Figure 1, our new metric is intended to capture differences in the proportion of the proteome subject to substantial selection (blue areas).

An alternative metric, the Effective Number of Codons (ENC) originally quantified how far the codon usage of a sequence departs from equal usage of synonymous codons (Wright 1990 [↗](#)), with lower ENC values indicating greater departure. This approach creates a complex relationship with GC content (Fuglsang 2008 [↗](#)), and so ENC was later modified to correct for GC content (Novembre 2002 [↗](#)). However, a remaining issue with this modified ENC is that differences among species in amino acid composition might act as a confounding factor, even after controlling for GC content. Specifically, species that make more use of an amino acid for which there is stronger selection among codons (which is sometimes the case (Vicario et al. 2007 [↗](#))) would have higher codon bias, even if each amino acid considered on its own had identical codon bias irrespective of which species it is in. Confounding with amino acid frequencies has been shown to be a problem at the individual protein level (Cope et al. 2018 [↗](#)). Neither ENC (Fuglsang 2004 [↗](#); Fuglsang 2008 [↗](#)) nor the CAI (Sharp & Li 1987 [↗](#)) adequately control for differences in amino acid composition when applied across species. Despite early claims to the contrary (Wright 1990 [↗](#)), this problem is not easy to fix for ENC (Fuglsang 2004 [↗](#); Fuglsang 2008 [↗](#)).

Here, we extend the CAI, using the information-theory based Kullback-Leibler divergence, so that it corrects for both GC and amino acid composition (see Methods) to create a new Codon Adaptation Index of Species (CAIS). The availability of complete genomes allows both metrics to be readily calculated without data on polymorphism or mutation rate, without selecting reference genes, and without concerns about demographic history. Our purpose is to find an accessible metric that can quantify the limits to weak selection important to nearly neutral theory; this differs from past evaluations focused on comparing different genes of the same species and recapitulating “ground truth” simulations thereof (Sun et al. 2012 [↗](#); Zhang et al. 2012 [↗](#); Liu et al. 2018 [↗](#)). To demonstrate the usefulness of our method, we identify a novel correlation with intrinsic structural disorder (ISD), pointing to what else might be subject to weak selective preferences at the molecular level. While ENC can also identify subtle selection on ISD, CAIS can do so without the risk of confounding with amino acid frequencies.

Results

Both ENC and CAIS solve the GC confounding problem that plagues CAI

CAI is seriously confounded with GC content (Figure 3A [↗](#)). ENC is not confounded with GC content (Figure 3B [↗](#)), while CAIS has only a very weak correlation that is not significant after correction for multiple comparisons (Figure 3C [↗](#)).

Proteins in better adapted species evolve more structural disorder

As an example of how correlations with codon adaptation metrics can be used to identify weak selective preferences, we investigate protein ISD. Disordered proteins are more likely to be harmful when overexpressed (Vavouri et al. 2009 [↗](#)), and ISD is more abundant in eukaryotic than prokaryotic proteins (Schad et al. 2011 [↗](#); Xue et al. 2012 [↗](#); Basile et al. 2019 [↗](#)), suggesting that low ISD might be favored by more effective selection.

However, compositional differences among proteomes might not be driven by differences in how a given protein sequence evolves as a function of the effectiveness of selection. Instead, they might be driven by the recent birth of ISD-rich proteins in animals (James et al. 2021 [↗](#)), and/or by differences among sequences in their subsequent tendency to proliferate into many different genes (James et al. 2023 [↗](#)). To focus only on the effects of descent with modification, we use a linear mixed model, with each species having a fixed effect on ISD, while controlling for Pfam domain identity as a random effect. We note that once GC is controlled for, codon adaptation can

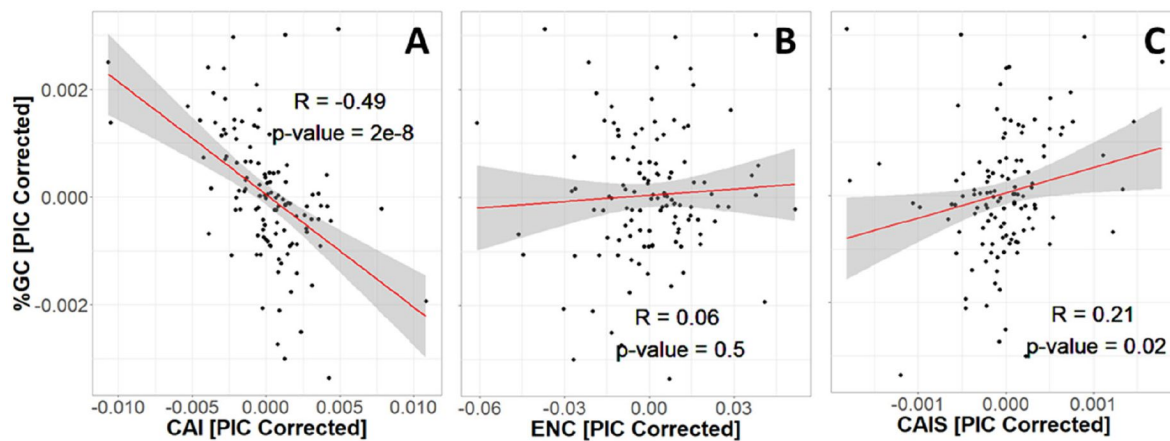


Figure 3.

CAI is seriously confounded with GC content, while ENC and CAIS are not. We control for phylogenetic confounding via Phylogenetic Independent Contrasts (PIC) (Felsenstein 1985 [\[1\]](#)); this yields an unbiased R^2 estimate (Rohlf 2006 [\[2\]](#)). Each datapoint is one of 118 vertebrate species with “Complete” intergenic genomic sequence (allowing for %GC correction) and TimeTree divergence dates (allowing for PIC correction). Red line shows unweighted $\text{lm}(y \sim x)$ with grey region as 95% confidence interval. Plots without PIC correction are shown in Supplementary Figure 2.

be assessed similarly in intrinsically disordered vs. ordered proteins (Gossmann et al. 2012 [↗](#)). Controlling for Pfam identity is supported, with standard deviation in ISD of 0.178 among Pfams compared to residual standard deviation of 0.058, and a p-value on the significance of the Pfam random effect term of 3×10^{-13} . Controlling in this way for Pfam identity, we then ask whether the fixed species effects on ISD are correlated with CAIS and with ENC.

Surprisingly, more exquisitely adapted species have more disordered protein domains (**Figure 4** [↗](#)). Results using ENC and CAIS are similar, with ENC having higher power; the correlation coefficient is 0.36 for CAIS compared to 0.50 for ENC, and the p-value for ENC is 3 orders of magnitude lower. We note, however, that amino acid frequencies strongly influence ISD (Theillet et al. 2013 [↗](#)). The CAIS correlation is more reliable than the ENC correlation because by construction, CAIS controls for differences in amino acid frequencies among species.

Different parts of the genome have different GC contents (Bernardi 2000 [↗](#); Eyre-Walker & Hurst 2001 [↗](#); International Human Genome Sequencing Consortium 2001 [↗](#)), primarily because the extent to which GC-biased gene conversion increases GC content depends on the local rate of recombination (Galtier et al. 2001 [↗](#); Meunier & Duret 2004 [↗](#); Duret et al. 2006 [↗](#); Duret & Galtier 2009 [↗](#)). We therefore also calculated a version of CAIS whose codon frequency expectations are based on local intergenic GC content. This performed worse (**Figure 4C** [↗](#)) than our simple use of genome-wide GC content (**Figure 4A** [↗](#)) with respect to the strength of correlation between CAIS and ISD. If GC-biased gene conversion is a more powerful force than weak selective preferences among codons, then local GC content will evolve more rapidly than codon usage (Kondrashov et al. 2010 [↗](#)). In this case, genome-wide GC may serve as an appropriately time-averaged proxy. It is also possible that the local non-coding sequences we used were too short (at 3000bp or more), creating excessive noise that obscured the signal.

Many vertebrates have higher recombination rates and hence GC-biased gene conversion near genes; in this case genome-wide GC content would misestimate the codon usage expected from the combination of mutation bias and GC-biased gene conversion in the vicinity of genes. If GC-biased gene conversion drove CAIS, we expect high $|\overline{\text{local GC}} - \overline{\text{global GC}}|$ to predict high CAIS. We do not see this relationship (**Figure 5** [↗](#)), suggesting that gene conversion strength is not a confounding factor impacting CAIS.

Younger animal-specific protein domains have higher ISD (James et al. 2021 [↗](#)). It is possible that selection in favor of high ISD is strongest in young domains, which might use more primitive methods to avoid aggregation (Foy et al. 2019 [↗](#); Bertram & Masel 2020 [↗](#)). To test this, we analyze two subsets of our data: those that emerged prior to the last eukaryotic common ancestor (LECA), here referred to as “old” protein domains, and “young” protein domains that emerged after the divergence of animals and fungi from plants. Young and old domains show equally strong trends of increasing disorder with species’ adaptedness (**Figure 6** [↗](#)).

Discussion

When different properties are each causally affected by a species’ exquisiteness of adaptation, this will create a correlation between the properties. We use codon adaptation as a reference property, such that correlations with codon adaptation indicate selection. To detect ISD as a novel property under selection, we used a linear mixed model approach that controls for Pfam identity as a random effect. This approach shows that the same Pfam domain tends to be more disordered when found in a well-adapted species (i.e. a species with a higher CAIS or ENC). This is true for both ancient and recently emerged protein domains.

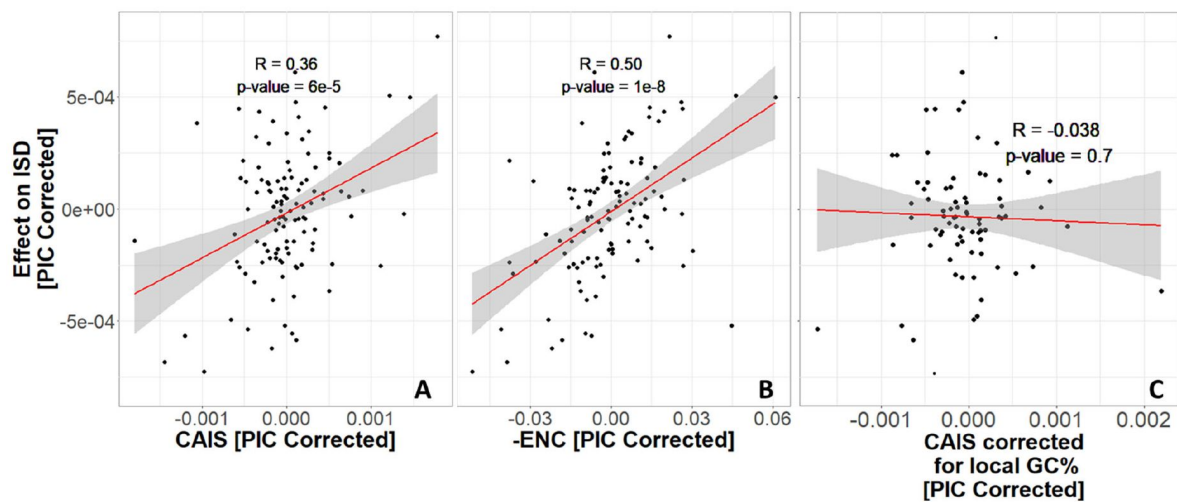


Figure 4.

Protein domains have higher ISD when found in more exquisitely adapted species. We show - ENC instead of ENC to more easily compare with CAIS. Each datapoint is one of 118 vertebrate species with “complete” intergenic genomic sequence available (allowing for %GC correction), and TimeTree divergence dates (allowing for PIC correction). “Effects” on ISD shown on the y-axis are fixed effects of species identity in our linear mixed model, after PIC correction. Red line shows unweighted $\text{lm}(y \sim x)$ with grey region as 95% confidence interval. Panels without PIC correction are presented in Supplementary Figure 3.

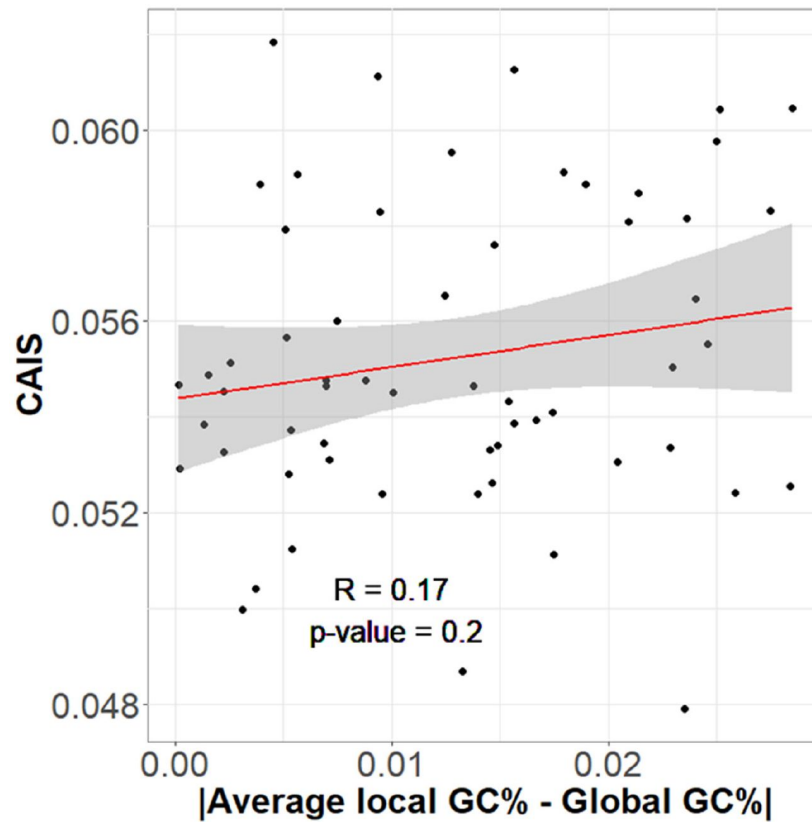


Figure 5.

CAIS is not correlated with the degree to which local genomic regions differ in their GC content from global GC content. If CAIS were driven by GC biased gene conversion, genomes with more heterogeneous %GC distributions should have higher CAIS scores.

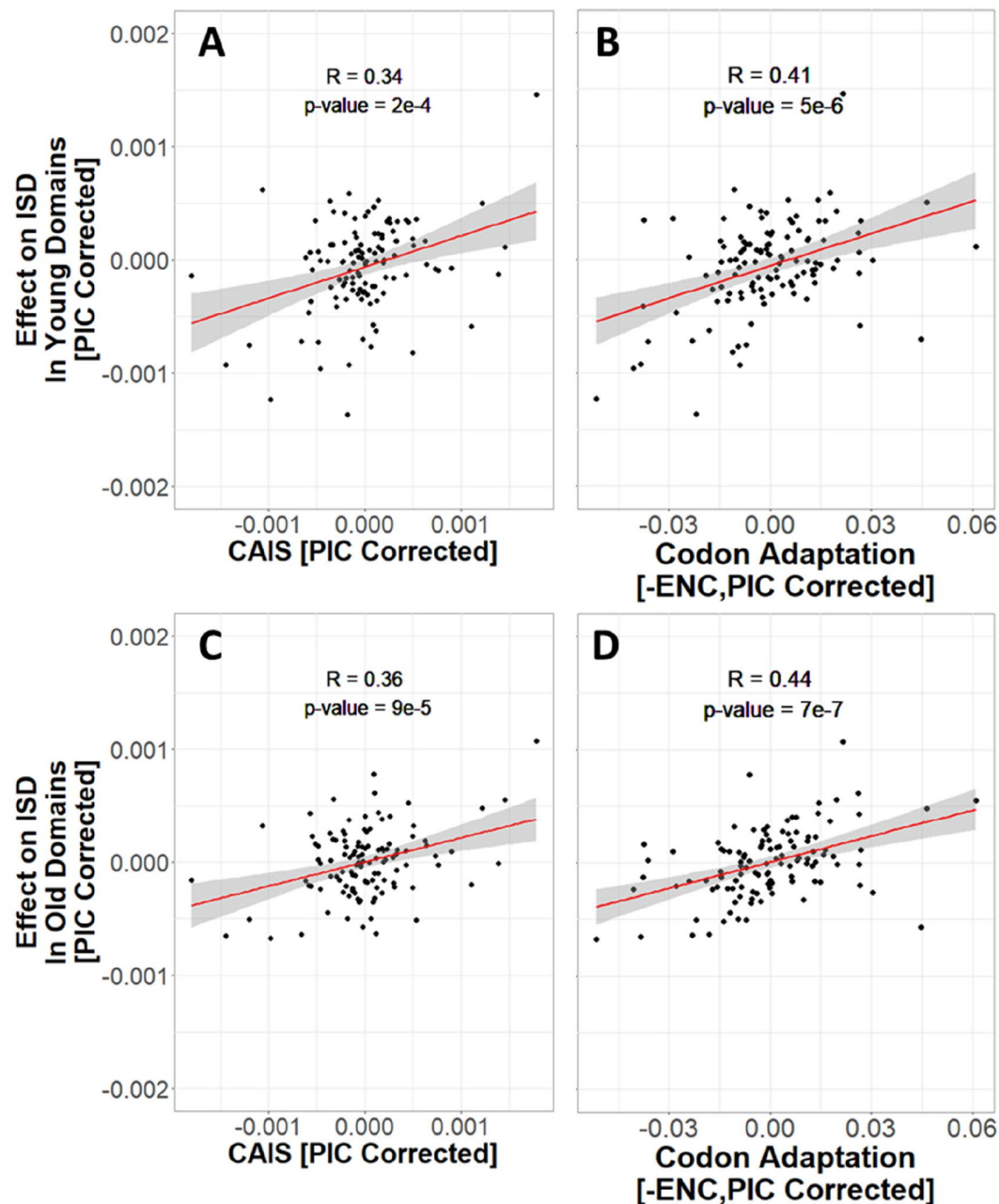


Figure 6.

More exquisitely adapted species have higher ISD in both ancient and recent protein domains. Age assignments are taken from James et al. (2021) <https://doi.org/10.1016/j.cub.2021.05.001>, with vertebrate protein domains that emerged prior to LECA classified as “old”, and vertebrate protein domains that emerged after the divergence of animals and fungi from plants as “young”. “Effects” on ISD shown on the y-axis are fixed effects of species identity in our linear mixed model. The same $n=118$ datapoints are shown as in Figures 3 and 4. Red line shows $\text{lm}(y \sim x)$, with grey region as 95% confidence interval. Panels without PIC correction are shown in Supplementary Figure 4. We show -ENC instead of ENC to more easily compare with CAIS.

It is important that no additional variable such as GC content or amino acid frequencies creates a spurious correlation by affecting both CAIS and our property of interest. For this reason, we define CAIS as the observed Kullback-Leibler divergence (Kullback & Leibler 1951 [↗](#)) from the codon usage expected given the GC content. The GC content pertinent to this expectation is primarily the product of mutation bias and GC-biased gene conversion (Romiguier & Roux 2017 [↗](#)), but potentially also of selection on individual nucleotide substitutions that is hypothesized to favor higher %GC (Long et al. 2018 [↗](#)). By controlling for %GC, we exclude all these forces from influencing CAIS or ENC. We thus capture the extent of adaptation in codon bias, including translational speed, accuracy, and any intrinsic preference for GC over AT that is specific to coding regions. These remaining codon-adaptive factors do not create a statistically convincing correlation between CAIS and GC (Figures 3C [↗](#)), nor between ENC and GC (Figure 3B [↗](#)), although CAI is strongly correlated with GC (Figure 3A [↗](#)). Notably, our new CAIS metric of codon adaptation controls for amino acid frequencies, rather than, like ENC, only GC content.

A direct effect of ISD on fitness agrees with studies of random ORFs in *Escherichia coli*, where fitness was driven more by amino acid composition than %GC content, after controlling for the intrinsic correlation between the two (Kosinski et al. 2022 [↗](#)). However, we have not ruled out a role for selection for higher %GC in ways that are general rather than restricted to coding regions, whether in shaping mutational biases (Smith & Eyre-Walker 2001 [↗](#); Hershberg & Petrov 2009 [↗](#); Hildebrand et al. 2010 [↗](#); Novoa et al. 2019 [↗](#); Forcelloni & Giansanti 2020 [↗](#)) or the extent of gene conversion, or even at the single nucleotide level in a manner shared between coding regions and intergenic regions (Long et al. 2018 [↗](#)).

A more complex metric could control for more than just GC content and amino acid frequencies. First vs. second vs. third codon positions have different nucleotide usage on average, but while correcting for this might be useful for comparing genes (Zhang et al. 2012 [↗](#)), correcting for it while comparing species might remove the effect of interest. Similarly, while it might be useful to control for dinucleotide and trinucleotide frequencies (Brbić et al. 2015 [↗](#)), to avoid circularity these would need to be taken from intergenic sequences, with care needed to avoid influence from unannotated protein-coding genes or even pseudogenes.

Note that if a species were to experience a sudden reduction in census population size, e.g. due to habitat loss, leading to less effective selection, it would take some multiple of the neutral coalescent time for CAIS to fully adjust. CAIS thus represents a relatively long-term historical pattern of adaptation. The timescales setting neutral polymorphism based N_e estimates are likely shorter, based on a single round of coalescence. It is possible that the reason that we obtained correlations when we controlled for genome-wide GC content, but not when we controlled for local GC content, is also that codon adaptation adjusts slowly relative to the timescale of fluctuations in local GC content.

Here we developed a new metric of species adaptedness at the codon level, capable of quantifying degrees of codon adaptation even among vertebrates. We chose vertebrates partly due to the abundance of suitable data, and partly as a stringent test case, given past studies finding limited evidence for codon adaptation (Kessler & Dean 2014 [↗](#)). It remains to be seen how CAIS scales among species with stronger codon adaptation. We restricted our analysis to only the best annotated genomes, in part to ensure the quality of intergenic %GC estimates, and in part limited by the feasibility of running linear mixed models with six million data points. The phylogenetic tree is well resolved for vertebrate species, with an overrepresentation of mammalian species. Despite the focus on vertebrates, we were able to discover new results regarding selection on ISD.

Our finding that more effective selection prefers higher ISD was unexpected, given that lower- N_e eukaryotes have more disordered proteins than higher- N_e prokaryotes (Ahrens et al. 2017 [↗](#); Basile et al. 2019 [↗](#)). However, this can be reconciled in a model in which highly disordered sequences are less likely to be found in high- N_e species, but the sequences that are present tend to

have slightly higher disorder than their low- N_e homologs. High ISD might help mitigate the trade-off between affinity and specificity in protein-protein interactions (Dunker et al. 1998 [↗](#); Huang & Liu 2013 [↗](#); Lazar et al. 2022 [↗](#)); non-specific interactions might be short-lived due to the high entropy associated with disorder, which specific interactions are robust to.

Codon adaptation metrics more directly quantify how species vary in their exquisiteness of adaptation, than do estimates of effective population size that are based on neutral polymorphism. Both CAIS and ENC can also be estimated for far more species because they do not require polymorphism or mutation rate data, nor tRNA gene copy numbers and abundances, but only a single complete genome. CAIS has the additional advantage of not being confounded with amino acid frequencies. This makes CAIS a useful tools for applying nearly neutral theory to protein evolution, as shown by our worked example of ISD.

Methods

Species

Pfam sequences and IUPRED2 estimates of Intrinsic Structural Disorder (ISD) predictions were taken from James et al. (2021) [↗](#), who studied species marked as “Complete” in the GOLD database, with divergence dates available in TimeTree (Kumar et al. 2017 [↗](#)). James et al. (2021) [↗](#) applied a variety of quality controls to exclude contaminants from the set of Pfams and assign accurate dates of Pfam emergence. Pfams that emerged prior to LECA are classified here as “old”, and Pfams that emerged after the divergence of animals and fungi from plants are classified as “young”, following annotation by James et al. (2021) [↗](#). Species list and other information can be found at <https://github.com/MasellLab/Codon-Adaptation-Index-of-Species> [↗](#).

Codon Adaptation Index

Sharp and Li (1987) [↗](#) quantified codon bias through the Codon Adaptation Index (CAI), a normalized geometric mean of synonymous codon usage bias across sites, excluding stop and start codons. We modify this to calculate CAI including stop and start codons, because of documented preferences among stop codons in mammals (Wangen & Green 2020 [↗](#)). While usually used to compare genes within a species, among-species comparisons can be made using a reference set of genes that are highly expressed (Sharp & Li 1987 [↗](#)). Each codon i is assigned a Relative Synonymous Codon Usage value:

$$RSCU_i = \frac{N_i}{\frac{1}{n_a} \sum_{j=1}^{n_a} N_j}, \quad (1)$$

where N_i denotes the number of times that codon i is used, and the denominator sums over all n_a codons that code for that specific amino acid. RSCU values are normalized to produce a relative adaptiveness values w_i for each codon, relative to the best adapted codon for that amino acid:

$$w_i \equiv \frac{RSCU_i}{RSCU_{max}}. \quad (2)$$

Let L be the number of codons across all protein-coding sequences considered. Then

$$CAI = [\prod_{i=1}^L w_i]^{\frac{1}{L}}. \quad (3)$$

To understand the effects of normalization, it is useful to rewrite this as:

$$CAI = \left[\prod_{i=1}^L \frac{RSCU_i}{RSCU_{max}} \right]^{\frac{1}{L}} = \frac{CAI_{raw}}{CAI_{max}}, \quad (4)$$

where CAI_{raw} is the geometric mean of the “unnormalized” or observed synonymous codon usages, and CAI_{max} is the maximum possible CAI given the observed codon frequencies.

GC Content

We calculated total %GC content (intergenic and genic) during a scan of all six reading frames across genic and intergenic sequences available from NCBI with access dates between May and July 2019 (described in James et al. (2023) [\[4\]](#)). Of the 170 vertebrates meeting the quality criteria of James et al. (2021) [\[4\]](#), 118 had annotated intergenic sequences within NCBI, so we restricted the dataset further to keep only the 118 species for which total GC content was available.

Codon Adaptation Index of Species (CAIS)

Controlling for GC bias in Synonymous Codon Usage

Consider a sequence region r within species s where each nucleotide has an expected probability of being G or C = g_r . For our main analysis, we consider just one region r encompassing the entire genome of a species s . In a secondary analysis, we break the genome up and use local values of g_r in the non-coding regions within and surrounding a gene or set of overlapping genes. To annotate the boundaries of these local regions, we first selected 1500 base pairs flanking each side of every coding sequence identified by NCBI annotations. Coding sequence annotations are broken up according to exon by NCBI. When coding sequences of the same gene did not fall within 3000 base pairs of each other, they were treated as different regions. When two coding sequences, whether from the same gene or from different genes, had overlapping 1500bp catchment areas, we merged them together. g_r was then calculated based on the non-coding sites within each region, including both genic regions such as promoters and non-genic regions such as introns and intergenic sequences.

With no bias between C vs. G, nor between A vs. T, nor patterns beyond the overall composition taken one nucleotide at a time, the expected probability of seeing codon i in a triplet within r is

$$p_{i,r} = \frac{g_r^{k_{GC}}}{2} \left(1 - \frac{g_r}{2}\right)^{k_{AT}}, \quad (5)$$

where $k_{GC} + k_{AT} = 3$ total positions in codon i . The expected probability that amino acid a in region r is encoded by codon i is

$$E_{i,r} = \frac{p_{i,r}}{\sum_{j=1}^{n_a} p_{j,r}}. \quad (6)$$

We can then measure the degree to which the observed codon frequencies diverge from these expected probabilities using the Kullback–Leibler divergence. This gives a CAIS metric for a species s where $O_{i,s}$ is the observed frequency of codon i :

$$CAIS(s) = \sum_{i=1}^{64} O_{i,s} \log \left(\frac{O_{i,s}}{E_{i,s}} \right). \quad (7)$$

Controlling for Amino Acid Composition

Some amino acids may be more intrinsically prone to codon bias. We want a metric that quantifies effectiveness of selection (not amino acid frequency), so we re-weight CAIS on the basis of a standardized amino acid composition, to remove the effect of variation among species in amino acid frequencies.

Let F_a be the frequency of amino acid a across the entire dataset of 118 vertebrate genomes. We want to re-weight $O_{i,s}$ on the basis of F_a to ensure that differences in amino acid frequencies among species do not affect CAIS, while preserving relative codon frequencies for the same amino acid. We do this by solving for $\alpha_{a,s}$ so that

$$F_a = \alpha_{a,s} \sum_{j=1}^{n_a} O_{j,s}. \quad (8)$$

We then define $f'_{i,s} = \alpha_{a,s} O_{i,s}$ to obtain an amino acid frequency adjusted CAIS:

$$CAIS(S) = \sum_{i=1}^{64} f'_{i,s} \log \left(\frac{O_{i,s}}{E_{i,s}} \right). \quad (9)$$

The F_a values for our species set are at https://github.com/MasellLab/Codon-Adaptation-Index-of-Species/blob/main/CAIS_ENC_calculation/Total_amino_acid_frequency Vertebrates.txt. Use of the standardized set of amino acid frequencies F_a has only a small effect on computed CAIS values relative to using each vertebrate species' own amino acid frequencies (**Supplementary Figure 5**).

CAIS corrected for local intergenic GC content but not species-wide amino acid composition is

$$CAIS_{localGC}(s) = \left(\prod_{r=1}^G \sum_{i=1}^{64} O_{i,r} \log \left(\frac{O_{i,r}}{E_{i,r}} \right) \right)^{\frac{1}{L}}, \quad (10)$$

where $O_{i,r}$ is the number of times codon i appears in region r of species s , $E_{i,r}$ is the expected number of times codon i would appear in region r of species s given the local intergenic GC content, G is the number of regions, and $L = \sum_{r=1}^G \sum_{i=1}^{64} O_{i,r}$ is the total number of codons in the genome. Rewritten for greater computational ease:

$$CAIS_{localGC}(s) = e^{\frac{1}{L} \sum_{r=1}^G \ln \left(\sum_{i=1}^{64} O_{i,r} \log \left(\frac{O_{i,r}}{E_{i,r}} \right) \right)}. \quad (11)$$

Given the limited impact of amino acid frequency correction, we used **Equation 11** for the local GC results, but we could correct for amino acid composition by replacing the $O_{i,r}$ prefactor with $f'_{i,s}$, or even $f'_{i,r}$.

Novembre's Effective Number of Codons (ENC) controlled for total GC Content

The expected number of codons is based on the squared deviations $\hat{F}'_{n_a}$ of the frequencies of the codons for each amino acid a from null expectations:

$$X_a^2 = \sum_{i=1}^{n_a} \frac{N_a(O_i - E_i)^2}{E_i}, \quad (12)$$

where N_a is the total number of times that amino acid a appears. [Novembre \(2002\)](#) defines the corrected "F value" of amino acid a as

$$\hat{F}'_a = \frac{X_a^2 + N_a - n_a}{n_a(N_a - 1)} \quad (13)$$

and

$$ENC = 2 + \frac{9}{\hat{F}'_2} + \frac{1}{\hat{F}'_3} + \frac{5}{\hat{F}'_4} + \frac{3}{\hat{F}'_6}, \quad (14)$$

where each $\pi(N, s) = \frac{1 - e^{-\frac{s}{2}}}{1 - e^{-Ns}}$ is the average of the “F values” for amino acids with n_a synonymous codons. Past measures of ENC do not contain stop or start codons (Wright 1990 [↗](#); Novembre 2002 [↗](#); Fuglsang 2004 [↗](#)), but as we did for CAI and CAIS above, we include stop codons as an “amino acid” and therefore amend (14) to

$$ENC = 2 + \frac{9}{\overline{F}_{I_2}} + \frac{2}{\overline{F}_{I_3}} + \frac{5}{\overline{F}_{I_4}} + \frac{3}{\overline{F}_{I_6}}. \quad (15)$$

Statistical Analysis

All statistical modelling was done in R 3.5.1. Scripts for calculating CAI and CAIS were written in Python 3.7.

Phylogenetic Independent Contrasts

Spurious phylogenetically confounded correlations can occur when closely related species share similar values of both metrics. One danger of such pseudoreplication is Simpson’s paradox, where there are negative slopes within taxonomic groups, but a positive slope among them might combine to yield an overall positive slope. We avoid pseudoreplication by using Phylogenetic Independent Contrasts (PIC) (Felsenstein 1985 [↗](#)) to assess correlation. PIC analysis was done using the R package “ape” (Paradis & Schliep 2019 [↗](#)).

Data Availability

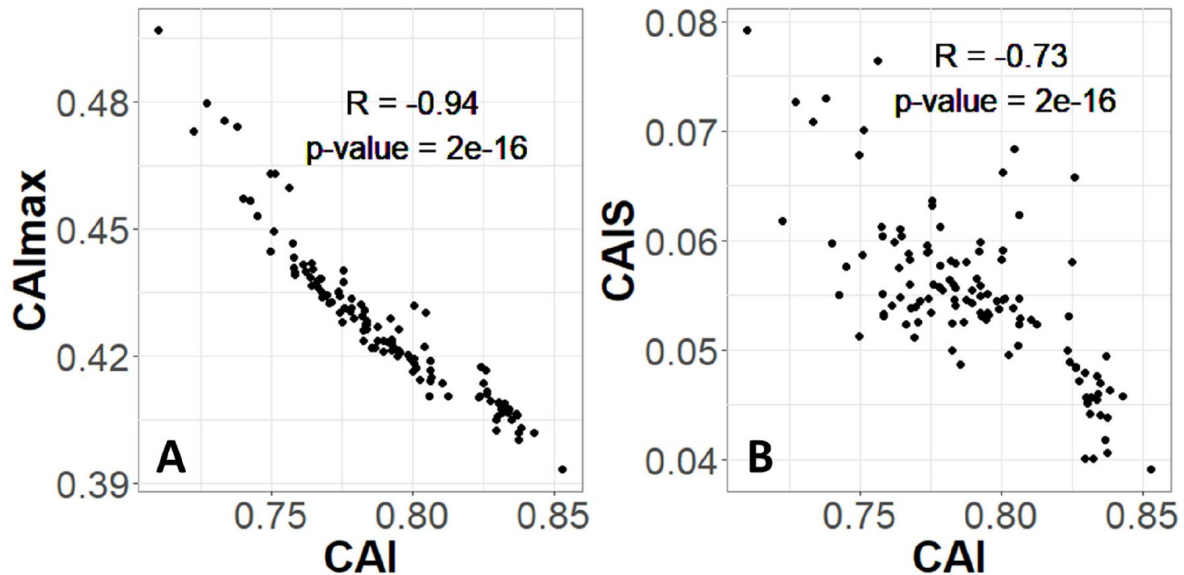
All data and code underlying this article are available in the public repository at <https://github.com/MasellLab/Codon-Adaptation-Index-of-Species> [↗](#).

Acknowledgements

This work was supported by the National Institutes of Health (R01 GM104040 and T32 GM132008), the John Templeton Foundation (60814), the Arnold and Mabel Beckman Foundation Scholars Program, the Western Alliance to Expand Student Opportunities (WAESO) Louis Stokes Alliance for Minority Participation (LSAMP) National Science Foundation (NSF) Cooperative Agreement (HRD-1101728), and UA/NASA Space Grant Undergraduate Research Internship program. We thank Luke Kosinski, David Liberles, and Sawsan Wehbi for helpful discussions, Paul Nelson for providing the genome-wide GC contents, and the University of Arizona Undergraduate Biology Research Program for training. We thank Gavin Douglas for writing a convenient end-to-end implementation of CAIS on the basis of our preprint, which can be found at https://github.com/gavinmdouglas/handy_pop_gen/blob/main/CAIS.py [↗](#), and for catching a minor bug in our code in time for us to correct it in the version of record. We thank the anonymous reviewers for constructive feedback, and Laurent Duret for helpful elaboration on the concerns of reviewer 1.

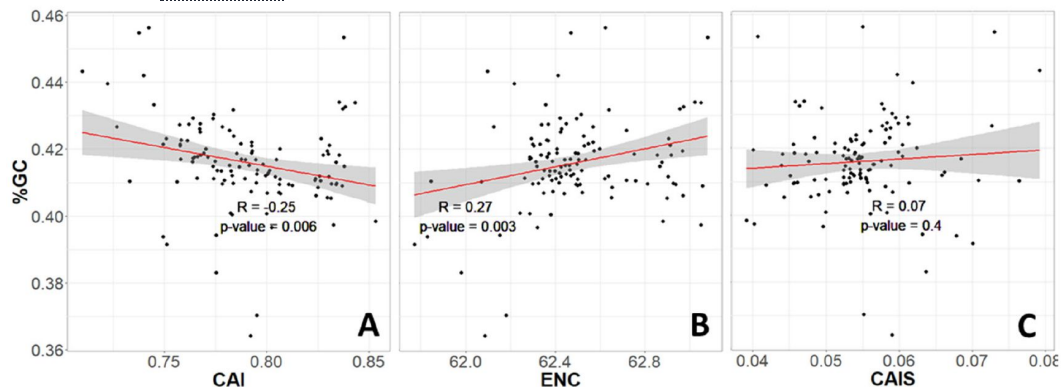
Supplementary Figure 1

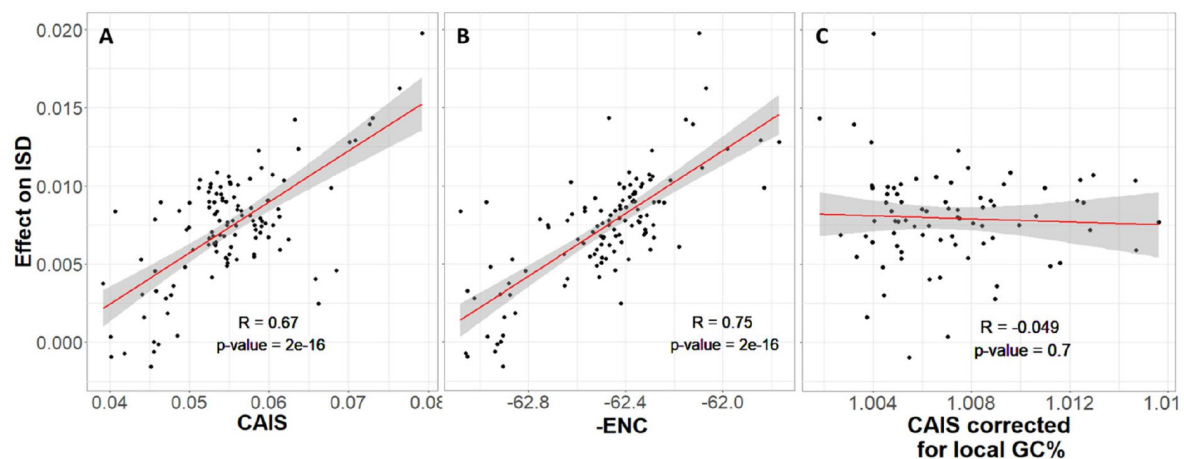
Codon Adaptation Index is not appropriate for species-wide effectiveness of selection measurements. Each CAI value shown is averaged over an entire species' proteome. A) The value of CAI is driven by its normalizing denominator term, CAI_{max} . B) As a result, CAI is inversely proportional to CAIS. Each datapoint is one of 118 vertebrate species with "Complete" intergenic genomic sequence available (allowing for %GC correction) and TimeTree divergence dates (allowing for PIC correction). P-values shown are for Pearson's correlation.



Supplementary Figure 2

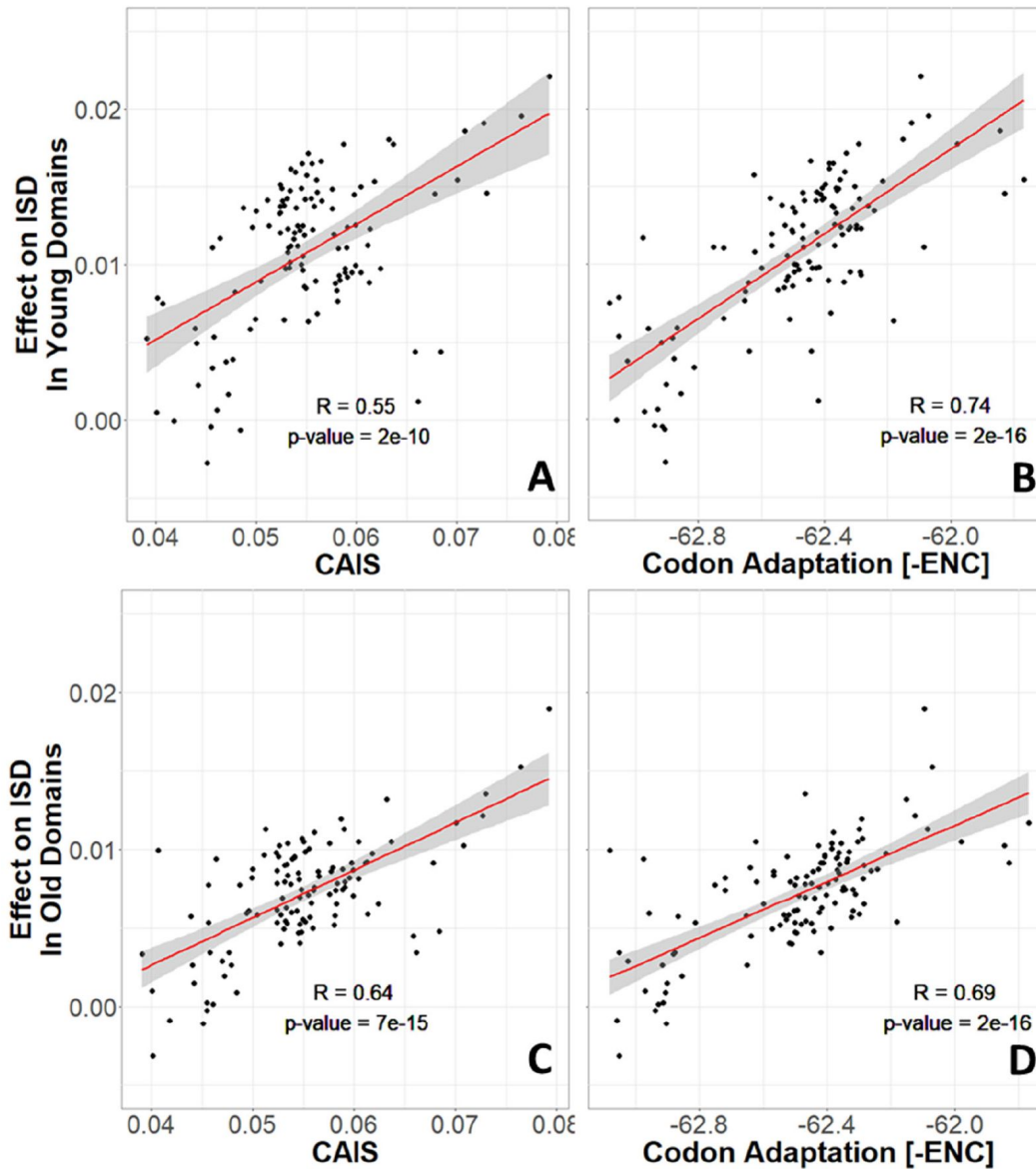
The same relationships are shown as in [Figure 3](#), here without correction for phylogenetic confounding. CAI and ENC both correlate with genomic GC, but CAIS does not. Red line shows $lm(y \sim x)$, with grey region as 95% confidence interval. We use PIC corrected results rather than these results because PIC correction removes non-independent errors to produce an unbiased R^2 estimate ([Rohlf 2006](#)).





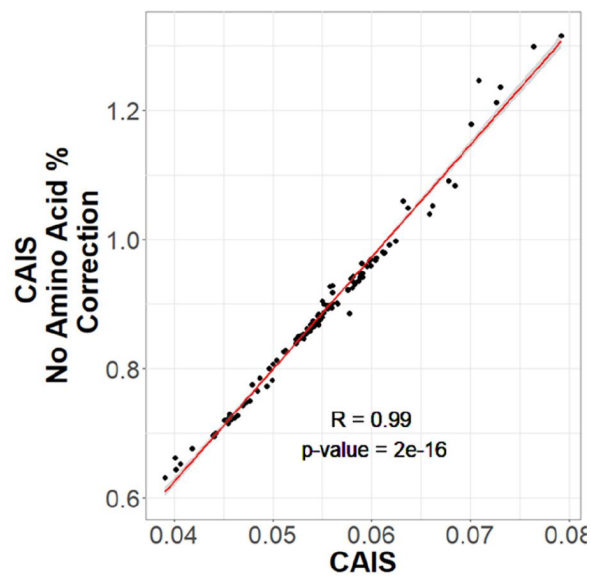
Supplementary Figure 3

The same relationships are shown as in [Figure 4](#), here without correction for phylogenetic confounding. As in [Figure 4](#), ISD of protein domains is higher in more highly adapted species, as measured by CAIS and ENC, but not by CAIS calculated with local GC% rather than genome-wide GC%. ISD is calculated as in [Figure 4](#). Red line shows $\text{Im}(y \sim x)$, with grey region as 95% confidence interval.



Supplementary Figure 4

Without correction for phylogenetic confounding, more highly adapted species have higher ISD in both young and old protein domains. Age assignments and ISD effects are calculated as in [Figure 6](#). same $n=118$ datapoints are shown as in [Figures 3-5](#). Red line shows $\text{lm}(y \sim x)$, with grey region as 95% confidence interval.



Supplementary Figure 5

Vertebrate CAIS values are not greatly affected by computation for a standardized amino acid composition vs. computation for the amino acid frequencies in the species in question.

References

- Ahrens JB, Nunez-Castilla J, Siltberg-Liberles J. (2017) **Evolution of intrinsic disorder in eukaryotic proteins** *Cellular and Molecular Life Sciences* **74**:3163–3174
- Akashi H. (1996) **Molecular Evolution Between *Drosophila melanogaster* and *D. simulans* Reduced Codon Bias, Faster Rates of Amino Acid Substitution, and Larger Proteins in *D. melanogaster*** *Genetics* **144**:1297–1307
- Basile W, Salvatore M, Bassot C, Elofsson A. (2019) **Why do eukaryotic proteins contain more intrinsically disordered regions?** *PLoS Computational Biology* **15**
- Bernardi G. (2000) **Isochores and the evolutionary genomics of vertebrates** *Gene* **241**:3–17
- Bertram J, Masel J. (2020) **Evolution rapidly optimizes stability and aggregation in lattice proteins despite pervasive landscape valleys and mazes** *Genetics* **214**:1047–1057
- Botzman M, Margalit H. (2011) **Variation in global codon usage bias among prokaryotic organisms is associated with their lifestyles** *Genome Biology* **12**
- Brbić M, Warnecke T, Kriško A, Supek F. (2015) **Global Shifts in Genome and Proteome Composition Are Very Tightly Coupled** *Genome Biology and Evolution* **7**:1519–1532
- Bulmer M. (1991) **The selection-mutation-drift theory of synonymous codon usage** *Genetics* **129**:897–907
- Charlesworth B. (2009) **Effective population size and patterns of molecular evolution and variation** *Nature Reviews Genetics* **10**:195–205
- Cope AL, Hettich RL, Gilchrist MA (2018) **Quantifying codon usage in signal peptides: Gene expression and amino acid usage explain apparent selection for inefficient codons** *Biochimica et Biophysica Acta (BBA) - Biomembranes* **1860**:2479–2485
- Daubin V, Moran NA (2004) **Comment on “The Origins of Genome Complexity”** *Science* **306**:978–978
- Doherty A, McInerney JO (2013) **Translational Selection Frequently Overcomes Genetic Drift in Shaping Synonymous Codon Usage Patterns in Vertebrates** *Molecular Biology and Evolution* **30**:2263–2267
- dos Reis M., Wernisch L. (2009) **Estimating Translational Selection in Eukaryotic Genomes** *Molecular Biology and Evolution* **26**:451–461
- Doyle JM, Hacking CC, Willoughby JR, Sundaram M, DeWoody JA (2015) **Mammalian genetic diversity as a function of habitat, body size, trophic class, and conservation status** *Journal of Mammalogy* **96**:564–572
- Dunker AK, Garner E, Guillot S, Romero P, Albrecht K, Hart J, Obradovic Z, Kissinger C, Villafranca JE (1998) **Protein disorder and the evolution of molecular recognition: theory, predictions and observations** *Pacific Symposium on Biocomputing* :473–484

- Duret L, Eyre-Walker A, Galtier N. (2006) **A new perspective on isochore evolution** *Gene* **385**:71–74
- Duret L, Galtier N. (2009) **Biased Gene Conversion and the Evolution of Mammalian Genomic Landscapes** *Annual Review of Genomics and Human Genetics* **10**:285–311
- Eyre-Walker A, Hurst LD (2001) **The evolution of isochores** *Nature Reviews Genetics* **2**:549–555
- Felsenstein J. (1985) **Phylogenies and the Comparative Method** *The American Naturalist* **125**:1–15
- Figuet E, Ballenghien M, Romiguier J, Galtier N. (2014) **Biased Gene Conversion and GC-Content Evolution in the Coding Sequences of Reptiles and Vertebrates** *Genome Biology and Evolution* **7**:240–250
- Forcelloni S, Giansanti A. (2020) **Evolutionary Forces and Codon Bias in Different Flavors of Intrinsic Disorder in the Human Proteome** *Journal of Molecular Evolution* **88**:164–178
- Foy SG, Wilson BA, Bertram J, Cordes MHJ, Masel J. (2019) **A Shift in Aggregation Avoidance Strategy Marks a Long-Term Direction to Protein Evolution** *Genetics* **211**:1345–1355
- Fuglsang A. (2004) **The ‘effective number of codons’ revisited** *Biochemical and Biophysical Research Communications* **317**:957–964
- Fuglsang A. (2008) **Impact of bias discrepancy and amino acid usage on estimates of the effective number of codons used in a gene, and a test for selection on codon usage** *Gene* **410**:82–88
- Galtier N, Piganeau G, Mouchiroud D, Duret L. (2001) **GC-Content Evolution in Mammalian Genomes: The Biased Gene Conversion Hypothesis** *Genetics* **159**:907–911
- Galtier N, Roux C, Rousselle M, Romiguier J, Figuet E, Glémin S, Bierne N, Duret L. (2018) **Codon Usage Bias in Animals: Disentangling the Effects of Natural Selection, Effective Population Size, and GC-Biased Gene Conversion** *Molecular Biology and Evolution* **35**:1092–1103
- Gingold H *et al.* (2014) **A Dual Program for Translation Regulation in Cellular Proliferation and Differentiation** *Cell* **158**:1281–1292
- Gossmann TI, Keightley PD, Eyre-Walker A. (2012) **The Effect of Variation in the Effective Population Size on the Rate of Adaptive Molecular Evolution in Eukaryotes** *Genome Biology and Evolution* **4**:658–667
- Hershberg R, Petrov DA (2008) **Selection on Codon Bias** *Annual Review of Genetics* **42**:287–299
- Hershberg R, Petrov DA (2009) **General Rules for Optimal Codon Choice** *PLoS Genetics* **5**
- Hildebrand F, Meyer A, Eyre-Walker A. (2010) **Evidence of Selection upon Genomic GC-Content in Bacteria** *PLoS Genetics* **6**
- Huang Y, Liu Z. (2013) **Do Intrinsically Disordered Proteins Possess High Specificity in Protein– Protein Interactions?** *Chemistry – A European Journal* **19**:4462–4467

- Hunt RC, Simhadri VL, Iandoli M, Sauna ZE, Kimchi-Sarfaty C. (2014) **Exposing synonymous mutations** *Trends In Genetics* **30**:308–321
- International Human Genome Sequencing Consortium (2001) **Initial sequencing and analysis of the human genome** *Nature* **409**:860–921
- James J, Nelson P, Masel J. (2023) **Differential retention of Pfam domains contributes to long-term evolutionary trends** *Molecular Biology and Evolution*
- James JE, Willis SM, Nelson PG, Weibel C, Kosinski LJ, Masel J. (2021) **Universal and taxon-specific trends in protein sequences as a function of age** *eLife* **10**
- Kessler MD, Dean MD (2014) **Effective population size does not predict codon usage bias in mammals** *Ecology and Evolution* **4**:3887–3900
- Kimura M. (1962) **On the probability of fixation of mutant genes in a population** *Genetics* **47**:713–719
- Kondrashov AS, Povolotskaya IS, Ivankov DN, Kondrashov FA (2010) **Rate of sequence divergence under constant selection** *Biology Direct* **5**
- Kosinski LJ, Aviles NR, Gomez K, Masel J. (2022) **Random Peptides Rich in Small and Disorder-Promoting Amino Acids Are Less Likely to Be Harmful** *Genome Biology and Evolution* **14**
- Kullback S, Leibler RA (1951) **On Information and Sufficiency** *The Annals of Mathematical Statistics* **22**:79–86
- Kumar S, Stecher G, Suleski M, Hedges SB (2017) **TimeTree: A Resource for timelines, timetrees, and divergence times** *Molecular Biology and Evolution* **34**:1812–1819
- Lazar T, Tantos A, Tompa P, Schad E. (2022) **Intrinsic protein disorder uncouples affinity from binding specificity** *Protein Science* **31**
- Li W-H. (1987) **Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons** *Journal of Molecular Evolution* **24**:337–345
- Liu SS, Hockenberry AJ, Jewett MC, Amaral LAN (2018) **A novel framework for evaluating the performance of codon usage bias metrics** *Journal of The Royal Society Interface* **15**
- Long H *et al.* (2018) **Evolutionary determinants of genome-wide nucleotide composition** *Nature Ecology & Evolution* **2**:237–240
- Lynch M, Ackerman MS, Gout J-F, Long H, Sung W, Thomas WK, Foster PL (2016) **Genetic drift, selection and the evolution of the mutation rate** *Nature Reviews Genetics* **17**:704–714
- Meunier J, Duret L. (2004) **Recombination Drives the Evolution of GC-Content in the Human Genome** *Molecular Biology and Evolution* **21**:984–990
- Novembre JA (2002) **Accounting for Background Nucleotide Composition When Measuring Codon Usage Bias** *Molecular Biology and Evolution* **19**:1390–1394

- Novoa EM, Jungreis I, Jaillon O, Kellis M. (2019) **Elucidation of Codon Usage Signatures across the Domains of Life** *Molecular Biology and Evolution* **36**:2328–2339
- Ohta T. (1972) **Population size and rate of evolution** *Journal of Molecular Evolution* **1**:305–314
- Ohta T. (1973) **Slightly Deleterious Mutant Substitutions in Evolution** *Nature* **246**:96–98
- Ohta T. (1992) **The Nearly Neutral Theory of Molecular Evolution** *Annual Review of Ecology and Systematics* **23**:263–286
- Paradis E, Schliep K. (2019) **ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R** *Bioinformatics* **35**:526–528
- Plotkin JB, Dushoff J, Desai MM, Fraser HB (2006) **Codon Usage and Selection on Proteins** *Journal of Molecular Evolution* **63**:635–653
- Plotkin JB, Kudla G. (2011) **Synonymous but not the same: the causes and consequences of codon bias** *Nature Reviews Genetics* **12**:32–42
- Rocha EPC (2004) **Codon usage bias from tRNA's point of view: Redundancy, specialization, and efficient decoding for translation optimization** *Genome Research* **14**:2279–2286
- Rohlf FJ (2006) **A Comment on Phylogenetic Correction** *Evolution* **60**:1509–1515
- Romiguier J *et al.* (2014) **Comparative population genomics in animals uncovers the determinants of genetic diversity** *Nature* **515**:261–263
- Romiguier J, Roux C. (2017) **Analytical Biases Associated with GC-Content in Molecular Evolution** *Frontiers in Genetics* **8**
- Schad E, Tompa P, Hegyi H. (2011) **The relationship between proteome size, structural disorder and organism complexity** *Genome Biology* **12**
- Sharp PM, Bailes E, Grocock RJ, Peden JF, Sockett RE (2005) **Variation in the strength of selected codon usage bias among bacteria** *Nucleic Acids Research* **33**:1141–1153
- Sharp PM, Emery LR, Zeng K. (2010) **Forces that influence the evolution of codon bias** *Philosophical Transactions of the Royal Society B: Biological Sciences* **365**:1203–1212
- Sharp PM, Li W-H. (1987) **The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications** *Nucleic Acids Research* **15**:1281–1295
- Smith NGC, Eyre-Walker A. (2001) **Synonymous Codon Bias Is Not Caused by Mutation Bias in G+C-Rich Genes in Humans** *Molecular Biology and Evolution* **18**:982–986
- Subramanian S. (2008) **Nearly Neutrality and the Evolution of Codon Usage Bias in Eukaryotic Genomes** *Genetics* **178**:2429–2432
- Sun X, Yang Q, Xia X. (2012) **An Improved Implementation of Effective Number of Codons (Nc)** *Molecular Biology and Evolution* **30**:191–196
- Theillet F-X, Kalmar L, Tompa P, Han K-H, Selenko P, Dunker AK, Daughdrill GW, Uversky VN (2013) **The alphabet of intrinsic disorder** *Intrinsically Disordered Proteins* **1**

Urrutia AO, Hurst LD (2001) **Codon Usage Bias Covaries With Expression Breadth and the Rate of Synonymous Evolution in Humans, but This Is Not Evidence for Selection** *Genetics* **159**:1191–1199

Vavouri T, Semple JI, Garcia-Verdugo R, Lehner B. (2009) **Intrinsic Protein Disorder and Interaction Promiscuity Are Widely Associated with Dosage Sensitivity** *Cell* **138**:198–208

Vicario S, Moriyama EN, Powell JR (2007) **Codon usage in twelve species of *Drosophila*** *BMC Evolutionary Biology* **7**

Wangen JR, Green R. (2020) **Stop codon context influences genome-wide stimulation of termination codon readthrough by aminoglycosides** *eLife* **9**

Wright F. (1990) **The ‘effective number of codons’ used in a gene** *Gene* **87**:23–29

Xue B, Dunker AK, Uversky VN (2012) **Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life** *Journal of Biomolecular Structure and Dynamics* **30**:137–149

Zhang Z, Li J, Cui P, Ding F, Li A, Townsend JP, Yu J. (2012) **Codon Deviation Coefficient: a novel measure for estimating codon usage bias and its statistical significance** *BMC Bioinformatics* **13**

Editors

Reviewing Editor

C Brandon Ogbunugafor

Yale University, New Haven, United States of America

Senior Editor

Detlef Weigel

Max Planck Institute for Biology Tübingen, Tübingen, Germany

Reviewer #2 (Public Review):

Assessment

This study develops a potentially useful metric for quantifying codon usage adaptation – the Codon Adaptation Index of Species (CAIS) – that is intended to allow for more direct comparisons of the strength of selection at the molecular level across species by controlling for interspecies variation in amino acid usage and GC content. As evidence to support their claim CAIS better controls for GC content and amino acid usage across species, they note that CAIS has only a weak positive correlation with GC% (that does not stand up to multiple hypothesis testing correction) while CAI has a clear negative correlation with GC%. Using CAIS, they find better adapted species have more disordered protein domains; however, excitement about these findings is dampened due to (1) this result is also observed using the effective number of codons (ENC) and

(2) concerns over the interpretation of CAIS as a proxy for the effectiveness of selection.

Public Review

Summary

The goal of the authors in this study is to develop a more reliable approach for quantifying codon usage such that it is more comparable across species. Specifically, the authors wish to estimate the degree of adaptive codon usage, which is potentially a general proxy for the strength of selection at the molecular level. To this end, the authors created the Codon Adaptation Index for Species (CAIS) that attempts to control for differences in amino acid usage and GC% across species. Using their new metric, the authors observe a positive relationship between CAIS and the overall “disorderedness” of a species protein domains. I think CAIS has the potential to be a valuable tool for those interested in comparing codon adaptation across species in certain situations. However, I have certain theoretical concerns about CAIS as a direct proxy for the efficiency of selection s_N when mutation bias changes across species.

Strengths

- (1) I appreciate that the authors recognize the potential issues of comparing CAI when amino acid usage varies and correct for this in CAIS. I think this is sometimes an under-appreciated point in the codon usage literature, as CAI is a relative measure of codon usage bias (i.e. only considers synonyms). However, the strength of natural selection on codon usage can potentially vary across amino acids, such that comparing mean CAI between protein regions with different amino acid biases may result in spurious signals of statistical significance.
- (2) The CAIS metric presented here is generally applicable to any species that has an annotated genome with protein-coding sequences. A significant improvement over the previous version is the implementation of software tool for applying this method.
- (3) The authors do a better job of putting their results in the context of the underlying theory of CAIS compared to the previous version.
- (4) The paper is generally well-written.

Weaknesses

- (1) The previously observed correlation between CAIS and body size was due to a bug when calculating phylogenetic independent contrasts. I commend the authors for acknowledging this mistake and updating the manuscript accordingly. I feel that the unobserved correlation between CAIS and body size should remain in the final version of the manuscript. Although it is disappointing that it is not statistically significant, the corrected results are consistent with previous findings (Kessler and Dean 2014).

- (2) I appreciate the authors for providing a more detailed explanation of the theoretical basis model. However, I remain skeptical that shifts in CAIS across species indicates shifts in the strength of selection. I am leaving the math from my previous review here for completeness.

As in my previous review, let's take a closer look at the ratio of observed codon frequencies vs. expected codon frequencies under mutation alone, which was previously notated as RSCUS in the original formulation. In this review, I will keep using the RSCUS notation, even though it has been dropped from the updated version. The key point is this is the ratio of observed and expected codon frequencies. If this ratio is 1 for all codons, then CAIS would be 0 based on equation 7 in the manuscript – consistent with the complete absence of selection on codon usage. From here on out, subscripts will only be used to denote the codon and it will be assumed that we are only considering the case of $r = \text{genome}$ for some species s .

$$RSCUS_i = \frac{O_i}{E_i}$$

I think what the authors are attempting to do is “divide out” the effects of mutation bias (as given by E_i), such that only the effects of natural selection remain, i.e. deviations from the expected frequency based on mutation bias alone represents adaptive codon usage. Consider Gilchrist et al. GBE 2015, which says that the expected frequency of codon i at selection-mutation-drift equilibrium in gene g for an amino acid with N_a synonymous codons is

$$E_{i,g} = \frac{e^{-\Delta M_i - \Delta \eta_i \phi_g}}{\sum_{j=1}^{N_a} e^{-\Delta M_j - \Delta \eta_j \phi_g}} \quad (1)$$

where ΔM is the mutation bias, $\Delta \eta$ is the strength of selection scaled by the strength of drift, and ϕ_g is the gene expression level of gene g . In this case, ΔM and $\Delta \eta$ reflect the strength and direction of mutation bias and natural selection relative to a reference codon, for which $\Delta M, \Delta \eta = 0$. Assuming the selection-mutation-drift equilibrium model is generally adequate to model of the true codon usage patterns in a genome (as I do and I think the authors do, too), the $E_{i,g}$ could be considered the expected observed frequency codon i in gene g

$E[O_i, g]$.

Let's re-write the $E_i = \frac{P_i}{\sum_{j=1}^{N_a} P_j}$ in the form of Gilchrist et al., such that it is a function of

mutation bias ΔM . For simplicity we will consider just the two codon case and assume the amino acid sequence is fixed. Assuming GC% is at equilibrium, the term g_r and $1 - g_r$ can be written as

$$g_r = \frac{\mu_{AT \rightarrow GC}}{\mu_{AT \rightarrow GC} + \mu_{GC \rightarrow AT}}$$

$$1 - g_r = \frac{\mu_{GC \rightarrow AT}}{\mu_{AT \rightarrow GC} + \mu_{GC \rightarrow AT}}$$

where $\mu_{x \rightarrow y}$ is the mutation rate from nucleotides x to y . As described in Gilchrist et al. MBE 2015 and Shah and Gilchrist PNAS 2011, the mutation bias

$\Delta M_{NNA, NNG} = \log\left(\frac{\mu_{AT \rightarrow GC}}{\mu_{GC \rightarrow AT}}\right)$. This can be expressed in terms of the equilibrium GC

content by recognizing that

$$\frac{g_r}{1 - g_r} = \frac{\mu_{AT \rightarrow GC}}{\mu_{GC \rightarrow AT}}$$

$$\Rightarrow \frac{g_r}{1 - g_r} = e^{\Delta M}$$

As we are assuming the amino acid sequence is fixed, the probability of observing a synonymous codon i at an amino acid becomes just a Bernoulli process.

$$p_i = g_r^x (1 - g_r)^{(1-x)}$$

If we do this, then

$$E_{NNA} = \frac{p_{NNA}}{p_{NNA} + p_{NNG}} \quad (2)$$

$$= \frac{1 - g_r}{g_r + (1 - g_r)} \quad (3)$$

$$= \frac{1}{\frac{g_r}{1 - g_r} + 1} \quad (4)$$

$$= \frac{1}{e^{\Delta M} + 1} \quad (5)$$

$$= \frac{e^{-\Delta M}}{1 + e^{-\Delta M}} \quad (6)$$

Recall that in the Gilchrist et al. framework, the reference codon has $\Delta M_{NNG, NNG} = 0 \Rightarrow e^{-\Delta M_{NNG, NNG}} = 1$

(1) Thus, we have recovered the Gilchrist et al. model from the formulation of E_i under the assumption that natural selection has no impact on codon usage and codon NNG is the pre-defined reference codon. To see this, plug in 0 for $\Delta\eta$ in equation (1).

We can then calculate the expected RSCUS using equation (1) (using notation $E[O_i]$) and equation (6) for the two codon case. For simplicity assume, we are only considering a gene of average expression (defined as $\phi_g = 1$). Assume in this case that NNG is the reference codon ($\Delta M_{NNG, NNG} = 0$).

$$E[RSCUS_{NNA}] = \frac{E[O_{NNA}]}{E_{NNA}} \quad (7)$$

$$= \frac{e^{-\Delta\eta_{NNA}} (e^{-\Delta M_{NNA}} + e^{-\Delta M_{NNG}})}{e^{-\Delta M_{NNA} - \Delta\eta_{NNA}} + e^{-\Delta M_{NNG} - \Delta\eta_{NNG}}} \quad (8)$$

$$= \frac{e^{-\Delta M_{NNA} - \Delta\eta_{NNA}} + e^{-\Delta M_{NNG} - \Delta\eta_{NNA}}}{e^{-\Delta M_{NNA} - \Delta\eta_{NNA}} + e^{-\Delta M_{NNG} - \Delta\eta_{NNG}}} \quad (9)$$

$$= \frac{e^{-\Delta M_{NNA} - \Delta\eta_{NNA}} + e^{-\Delta\eta_{NNA}}}{e^{-\Delta M_{NNA} - \Delta\eta_{NNA}} + 1} \quad (10)$$

This shows that the expected value of RSCUS for a two codon amino acid is expected to increase as the strength of selection $\Delta\eta$ increases, which is desired. Note that $\Delta\eta$ in Gilchrist et al. is formulated in terms of selection against a codon relative to the reference, such that a

negative value represents that a codon is favored relative to the reference. If $\Delta\eta = 0$ (i.e. selection does not favor either codon), then $E[\text{RSCUS}] = 1$. Also note that the expected RSCUS does not remain independent of the mutation bias. This means that even if sNe (i.e. the strength of natural selection) does not change between species, changes to the strength and direction of mutation bias across species could impact RSCUS. Assuming my math is right, I think one needs to be cautious when interpreting CAIS as representative of the differences in the efficiency of selection across species except under very particular circumstances.

Consider our 2-codon amino acid scenario. You can see how changing GC content without changing selection can alter the CAIS values calculated from these two codons. Particularly problematic appears to be cases of extreme mutation biases, where CAIS tends toward 0 even for higher absolute values of the selection parameter. Codon usage for the majority of the genome will be primarily determined by mutation biases,

with selection being generally strongest in a relatively few highly-expressed genes. Strong enough mutation biases ultimately can overwhelm selection, even in highly-expressed genes, reducing the fraction of sites subject to codon adaptation.

Peer review image 1.

```
library(tidyverse)

## -- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
## v dplyr      1.1.4      v readr      2.1.5
## v forcats    1.0.0      v stringr   1.5.1
## v ggplot2    3.5.1      v tibble    3.2.1
## v lubridate  1.9.3      v tidyr     1.3.1
## v purrr      1.0.2

## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()     masks stats::lag()
## | Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

# Notation largely based on Gilchrist et al. GBE 2015
calculateCAIS <-
function(gc,dEta,phi)
{
  at <- 1 - gc
  dM <- log(gc/at) expected.w.selection.AT <- exp(-dM - dEta*phi)/(1 + exp(-dM - dEta*phi))
  expected.mut.only.AT <- exp(-dM)/(1 + exp(-dM)) expected.w.selection.GC <- 1 -
  expected.w.selection.AT expected.mut.only.GC <- 1 - expected.mut.only.AT
  CAIS.AT <- (expected.w.selection.AT * log(expected.w.selection.AT/expected.mut.only.AT)) CAIS.GC <-
  (expected.w.selection.GC * log(expected.w.selection.GC/expected.mut.only.GC)) cais.df <- data.frame(GC=gc,
  dEta = dEta,
                                Phi = phi,
                                Exp.w.selection.AT=expected.w.selection.AT,
                                Exp.mut.only.AT = expected.mut.only.AT,
                                Exp.w.selection.GC= expected.w.selection.AT,
                                Exp.mut.only.AT= expected.mut.only.AT, CAIS =
                                CAIS.AT + CAIS.GC)

  return(cais.df)
}

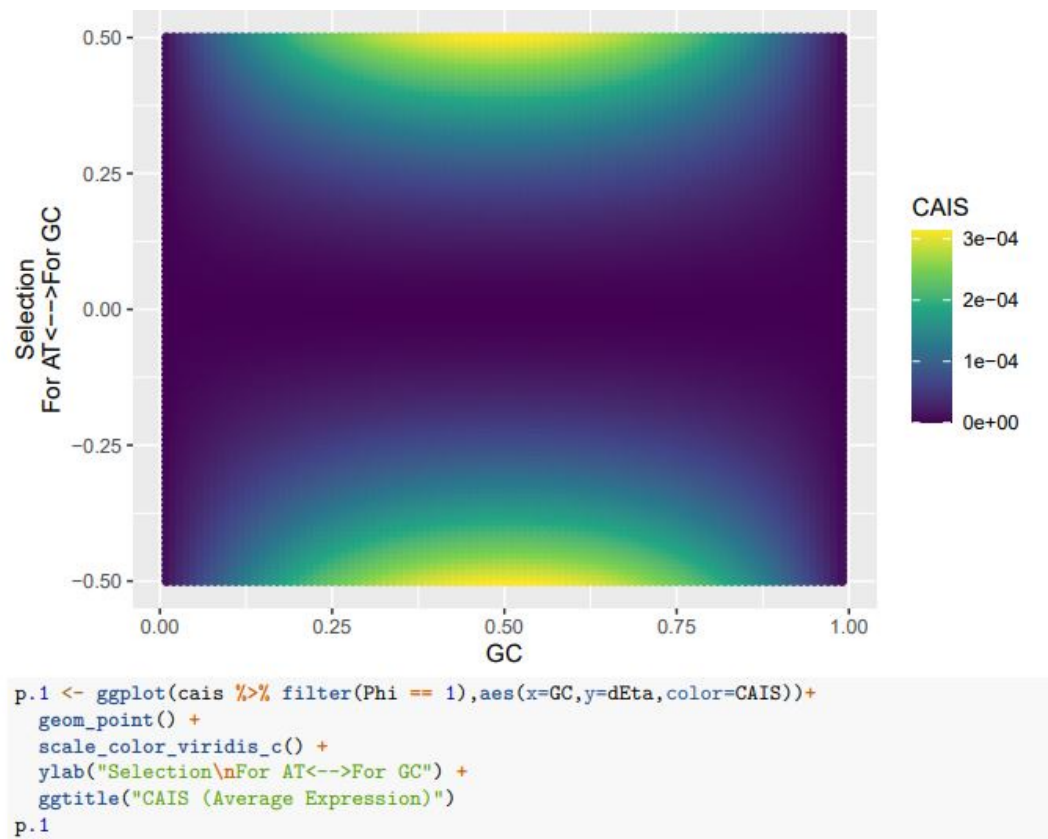
gc <- seq(0.01,0.99,by=0.01) dEta <- seq(-0.5,0.5,by=0.01) # a
function of s*Ne phi <- c(0.1,1,10)
simulate.param <- crossing(gc=gc,dEta=dEta,phi=phi)

cais <- pmap(simulate.param,calculateCAIS) %>%
  bind_rows()

p.0.1 <- ggplot(cais %>% filter(Phi == 0.1),aes(x=GC,y=dEta,color=CAIS))+ geom_point() +
  scale_color_viridis_c() +
  ylab("Selection\nFor AT<-->For GC") + ggtitle("CAIS (Low
  Expression)")
p.0.1
```

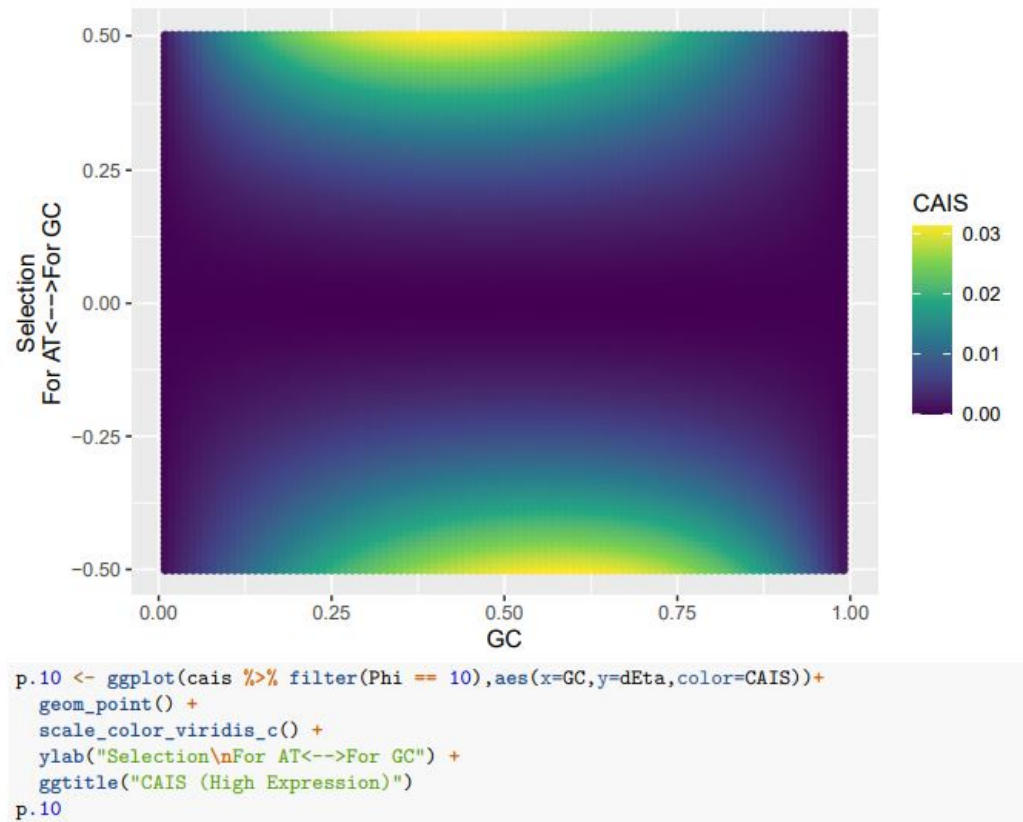
Peer review image 2.

CAIS (Low Expression)



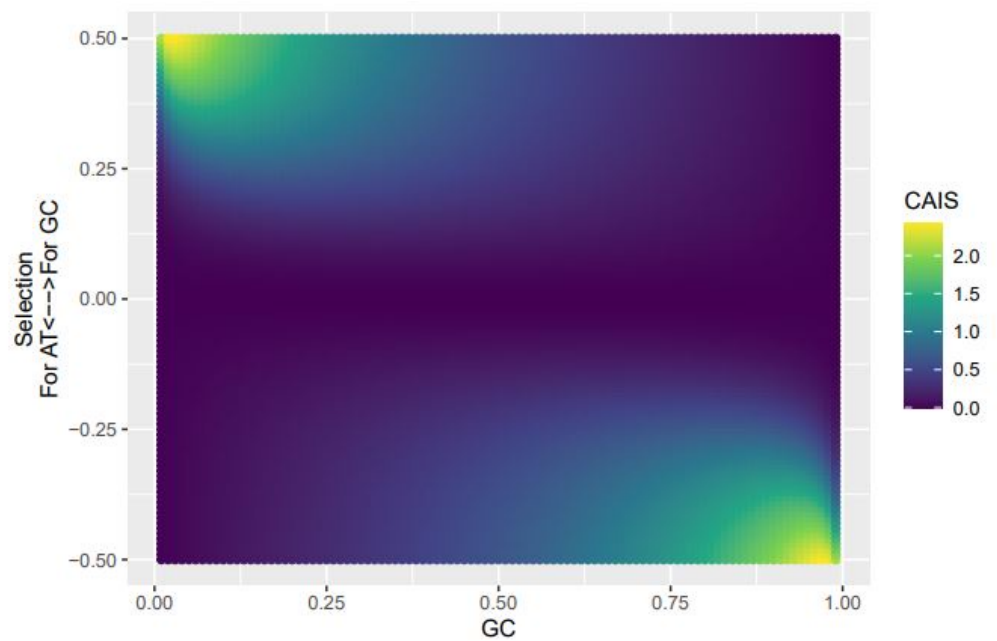
Peer review image 3.

CAIS (Average Expression)



Peer review image 4.

CAIS (High Expression)



If we treat the expected codon frequencies as genome-wide frequencies, then we are basically assuming this genome made up entirely of a single 2-codon amino acid with selection on codon usage being uniform across all genes. This is obviously not true, but I think it shows some of the potential limitations of the CAIS approach. Based on these simulations, CAIS seems best employed under specific scenarios. One such case could be when it is known that mutation bias varies little across the species of interest. Looking at the species used in this manuscript, most of them have a GC content around 0.41, so I suspect their results are okay (assuming things like GC-biased gene conversion are not an issue). Outliers in GC content probably are best excluded from the analysis.

Although I have not done so, I am sure this could be extended to the 4 and 6 codon amino acids. One potential challenge to CAIS is the non-monotonic changes in codon frequencies observed in some species (again, see Shah and Gilchrist 2011 and Gilchrist et al. 2015).

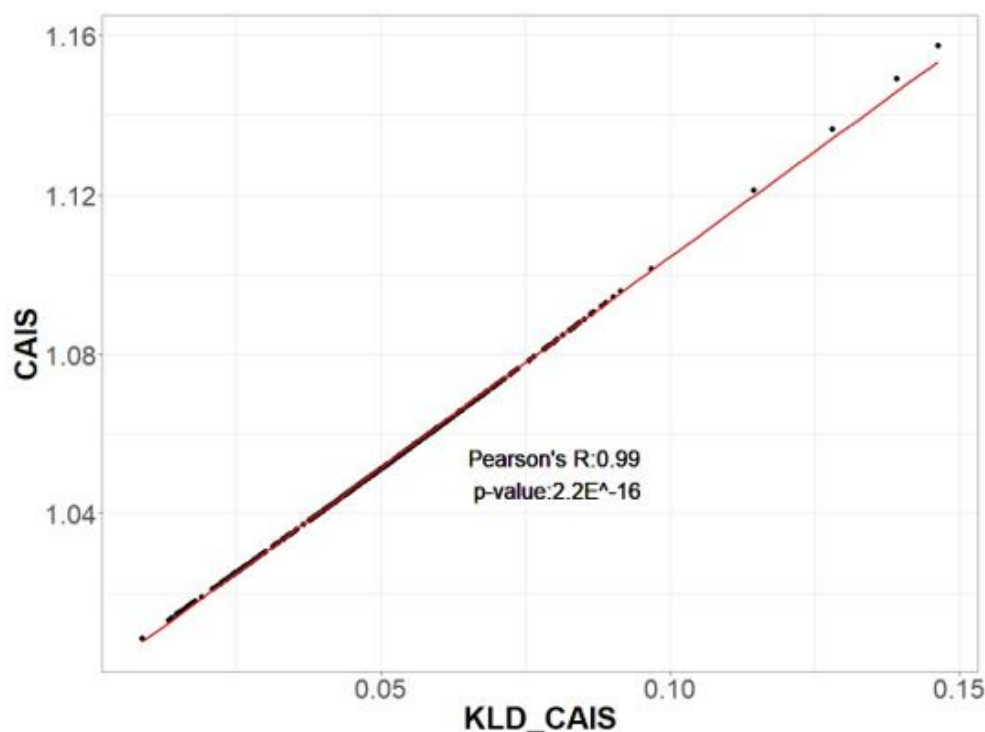
<https://doi.org/10.7554/eLife.87335.2.sa1>

Author response:

The following is the authors' response to the original reviews.

In addition to our responses to reviewer suggestions below, a minor bug in the calculation of CAIS was brought to our attention by a reader of our preprint. We have corrected this bug and rerun analyses, whose results became slightly stronger as noise was removed. While we were doing that, someone pointed out to us that our equations were almost the same as Kullback-Leibler divergence, which explains why our metric performed so well. We have made the numerically trivial (see before vs. after figure below) mathematical change to use Kullback-Leibler divergence instead, and now have a better story, with a solid basis in information theory, as to why CAIS works.

Author response image 1.



Unfortunately, we discovered a second bug that caused our PIC correction code to fail to perform the needed correction for phylogenetic confounding. The previously reported correlation between CAIS (or ENC) with body mass no longer survives PIC-correction. We have therefore removed this analysis from the manuscript. Our story now stands more on the theoretical basis of CAIS and ENC than on the post facto validation than it previously did. We now also present CAIS and ENC on a more equal footing. ENC results are slightly stronger, while CAIS has the complementary advantage of correcting for amino acid frequencies.

The work involved in these changes, as well as some of the responses to reviews below, justifies changing the second author into a co-first author, and adding an additional coauthor (Hanon McShea) who discovered the second bug.

Reviewer #1 (Public Review):

In this manuscript, the authors propose a new codon adaptation metric, Codon Adaptation Index of Species (CAIS), which they present as an easily obtainable proxy for effective population size. To permit between-species comparisons, they control for both amino acid frequencies and genomic GC content, which distinguishes their approach from existing ones. Having confirmed that CAIS negatively correlates with vertebrate body mass, as would be expected if small-bodied species with larger effective populations experience more efficient selection on codon usage, they then examine the relationship between CAIS and intrinsic structural disorder in proteins.

The idea of a robust species-level measure of codon adaptation is interesting. If CAIS is indeed a reliable proxy for the effectiveness of selection, it could be useful to analyze species without reliable life history- or mutation rate data (which will apply to many of the genomes becoming available in the near future).

A key question is whether CAIS, in fact, measures adaptation at the codon level. Unfortunately, CAIS is only validated indirectly by confirming a negative correlation with body mass. As a result, the observations about structural disorder are difficult to evaluate.

As discussed in the preamble above, we have replaced the body mass validation with a stronger theoretical basis in information theory.

A potential problem is that differences in GC between species are not independent of life history. Effective population size can drive compositional differences due to the effects of GC-biased gene conversion (gBGC). As noted by Galtier et al. (2018), genomic GC correlates negatively with body mass in mammals and birds. It would therefore be important to examine how gBGC might affect CAIS, and to what extent it could explain the relationship between CAIS and body mass.

Suppose that gBGC drives an increase in GC that is most pronounced at 3rd codon positions in high recombination regions in small-bodied species. In this case, could observed codon usage depart more strongly from expectations calculated from overall genomic GC in small vertebrates compared to large ones? The authors also report that correcting for local intergenic GC was unsuccessful, based on the lack of a significant negative relationship with body mass (Figure 3D). In principle, this could also be consistent with local GC providing a relatively more appropriate baseline in regions with high recombination rates. Considering these scenarios would clarify what exactly CAIS is capturing.

Figure 3 (previously Supplementary Figures S5A and S5B) shows that CAIS is negligibly correlated with %GC (not robust to multiple comparisons correction), and ENC not at all. We believe this is evidence against the possibility brought up by the reviewer, i.e. that N_e might affect gBGC (and hence global %GC). This relationship, if present, could act as a confounding effect, but it is not present within our species dataset.

Note that we expect our genomic-GC-based codon usage expectations to reflect unchecked gBGC in an average genomic region, independently of whether that species has high or low N_e . Our working model is that non-selective forces, include gBGC as well as conventional mutation biases, vary among species, and that they rather than selection determine each species' genome-wide %GC. By correcting for genome-wide %GC, CAIS and ENC correct for both mutation bias and gBGC, in order to isolate the effects of selection.

This argument, based on an average genomic region, is vulnerable to gene-rich genomic regions having differentially higher recombination rates and hence GC-biased gene conversion. However, we do not see the expected positive correlation between $|\text{local GC} - \text{global GC}|$ and CAIS (see new Figure 5), again suggesting that gene conversion strength is not a confounding factor acting on CAIS.

Given claims about "exquisitely adapted species", the case for using CAIS as a measure of codon adaptation would also be stronger if a relationship with gene expression could be demonstrated. RSCU is expected to be higher in highly expressed genes. Is there any evidence that the equivalent GC-controlled measure behaves similarly?

Correlations with gene expression are outside the scope of the current work, which is focused on producing and exploiting a single value of codon adaptation per species. It is indeed possible that our general approach of using Kullback-Leibler divergence to correct for genomic %GC could be useful in future work investigating differences among genes.

The manuscript is overall easy to follow, though some additional context may be helpful for the general reader. A more detailed discussion of how this work compares to the approach taken by Galtier et al. (2018), which accounted for GC content and gBGC when examining codon preferences, would be appropriate, for example. In addition, it would have been useful to mention past work that has attempted to explicitly quantify selection on codon usage.

One key difference between our work and that of Galtier et al. 2018 is that our approach does not rely on identifying specific codon preferences as a function of species. Our approach might therefore be robust to scenarios where different genes have different codon preferences (see Gingold et al. 2014 <https://doi.org/10.1016/j.cell.2014.08.011>). At a high level, our results are in broad agreement with those of Galtier et al., 2018, who found that gBGC affected all animal species, regardless of N_e , and who like us, found that the degree of selection on codon usage depended on N_e .

Reviewer #2 (Public Review):

Summary

The goal of the authors in this study is to develop a more reliable approach for quantifying codon usage such that it is more comparable across species. Specifically, the authors wish to estimate the degree of adaptive codon usage, which is potentially a general proxy for the strength of selection at the molecular level. To this end, the authors created the Codon Adaptation Index for Species (CAIS) that controls for differences in amino acid usage and GC% across species. Using their new metric, the authors find a previously unobserved negative correlation between the overall adaptiveness of codon usage and body size across 118 vertebrates. As body size is negatively correlated with effective population size and thus the general strength of natural selection, the negative correlation between CAIS and body size is expected. The authors argue this was previously unobserved due to failures of other popular metrics such as Codon Adaptation Index (CAI) and the Effective Number of Codons (ENC) to adequately control for differences in amino acid usage and GC content across species. Most surprisingly, the authors also find a positive relationship between CAIS and the overall "disorderedness" of a species protein domains. As some of these results are unexpected, which is acknowledged by the authors, I think it would be particularly beneficial to work with some simulated datasets. I think CAIS has the potential to be a valuable tool for those interested in comparing codon adaptation across species in certain situations. However, I have certain theoretical concerns about CAIS as a direct proxy for the efficiency of selection sN_e when the mutation bias changes across species.

Strengths

(1) I appreciate that the authors recognize the potential issues of comparing CAI when amino acid usage varies and correct for this in CAIS. I think this is sometimes an under-appreciated point in the codon usage literature, as CAI is a relative measure of codon usage bias (i.e. only considers synonyms). However, the strength of natural selection on codon usage can potentially vary across amino acids, such that comparing mean CAI between protein regions with different amino acid biases may result in spurious signals of statistical significance (see Cope et al. Biochemica et Biophysica Acta - Biomembranes 2018 for a clear example of this).

We now cite Cope et al. as an example of how amino acid composition can act as a confounding factor.

(2) The authors present numerous analysis using both ENC and mean CAI as a comparison to CAIS, helping given a sense of how CAIS corrects for some of the issues with these other metrics. I also enjoyed that they examined the previously unobserved relationship between codon usage bias and body size, which has bugged me ever since I saw Kessler and Dean 2014. The result comparing protein disorder to CAIS was particularly interesting and unexpected.

Unfortunately, our previous PIC correction code was buggy, and in fact the relationship with body size does not survive PIC correction (although it is strong prior to PIC correction). We have therefore removed it from the paper. However, the more novel result on protein disorder remains strong.

(3) The CAIS metric presented here is generally applicable to any species that has an annotated genome with protein-coding sequences.

Weaknesses

(1) The main weakness of this work is that it lacks simulated data to confirm that it works as expected. This would be particularly useful for assessing the relationship between CAIS and the overall effect of protein structure disorder, which the authors acknowledge is an unexpected result. I think simulations could also allow the authors to assess how their metric performs in situations where mutation bias and natural selection act in the same direction vs. opposite directions. Additionally, although I appreciate their comparisons to ENC and mean CAI, the lack of comparison to other popular codon metrics for calculating the overall adaptiveness of a genome (e.g. dos Reis et al.'s $\$S\$$ statistic, which is a function of tRNA Adaptation Index (tAI) and ENC) may be more appropriate. Even if results are similar to $\$S\$$, CAIS has a noted advantage that it doesn't require identifying tRNA gene copy numbers or abundances, which I think are generally less readily available than genomic GC% and protein-coding sequences.

The main limitation of dos Reis's test in our view is that, like the better versions of CAI, it requires comparable orthologs across species. See also the discussion below re the benefits of proteome-wide approach. We now also note the advantage of not needing tRNA gene copy numbers and abundances.

Simulated datasets would be great, but we think it a nice addition rather than must-have, in particular because we are skeptical about whether our understanding of all relevant processes is good enough such that simulations would add much to our more heuristic argument along the lines of Figure 2. E.g. the complications of Gingold et al. 2014 cited above are pertinent, but incorporating them would make simulations quite involved. Instead, we now have a stronger theoretical justification for CAIS grounded in information theory. We have significantly expanded discussion of Figure 2 to give a clearer idea of the conceptual underpinnings of CAIS and ENC.

The authors mention the selection-mutation-drift equilibrium model, which underlies the basic ideas of this work (e.g. higher $\$N_e\$$ results in stronger selection on codon usage), but a more in-depth framing of CAIS in terms of this model is not given. I think this could be valuable, particularly in addressing the question "are we really estimating what we think we're estimating?"

Let's take a closer look at the formulation for RSCUS. From here on out, subscripts will only be used to denote the codon and it will be assumed that we are only considering the case of $r = \text{genome for some species } s$.

$$RSCUS_i = \frac{O_i}{E_i}$$

I think what the authors are attempting to do is "divide out" the effects of mutation bias (as given by E_i), such that only the effects of natural selection remain, i.e. deviations from the expected frequency based on mutation bias alone represent adaptive codon usage. Consider Gilchrist et al. MBE 2015, which says that the expected frequency of codon i at selection-mutation-drift equilibrium in gene g for an amino acid with N_a synonymous codons is

$$E_{i,g} = \frac{e^{-\Delta M_i - \Delta \eta_i \phi_g}}{\sum_{j=1}^{N_a} e^{-\Delta M_j - \Delta \eta_j \phi_g}} \quad (1)$$

where ΔM is the mutation bias, $\Delta \eta$ is the strength of selection scaled by the strength of drift, and ϕ_g is the gene expression level of gene g . In this case, ΔM and $\Delta \eta$ reflect the strength and direction of mutation bias and natural selection relative to a reference codon, for which $\Delta M, \Delta \eta = 0$. Assuming the selection-mutation-drift equilibrium model is generally adequate to model of the true codon usage patterns in a genome (as I do and I think the authors do, too), the $E_{i,g}$ could be considered the expected observed frequency codon i in gene g

$E[O_i, g]$.

Let's re-write the $E_i = \frac{p_i}{\sum_{j=1}^{N_a} p_j}$ in the form of Gilchrist et al., such that it is a function of

mutation bias ΔM . For simplicity we will consider just the two codon case and assume the amino acid sequence is fixed. Assuming GC% is at equilibrium, the term g_r and $1 - g_r$ can be written as

$$g_r = \frac{\mu_{AT \rightarrow GC}}{\mu_{AT \rightarrow GC} + \mu_{GC \rightarrow AT}}$$

$$1 - g_r = \frac{\mu_{GC \rightarrow AT}}{\mu_{AT \rightarrow GC} + \mu_{GC \rightarrow AT}}$$

where $\mu_{x \rightarrow y}$ is the mutation rate from nucleotides x to y . As described in Gilchrist et al. MBE 2015 and Shah and Gilchrist PNAS 2011, the mutation bias

$\Delta M_{NNA, NNG} = \log\left(\frac{\mu_{AT \rightarrow GC}}{\mu_{GC \rightarrow AT}}\right)$. This can be expressed in terms of the equilibrium GC

content by recognizing that

$$\frac{g_r}{1 - g_r} = \frac{\mu_{AT \rightarrow GC}}{\mu_{GC \rightarrow AT}}$$

$$\Rightarrow \frac{g_r}{1 - g_r} = e^{\Delta M}$$

As we are assuming the amino acid sequence is fixed, the probability of observing a synonymous codon i at an amino acid becomes just a Bernoulli process.

$$p_i = g_r^x (1 - g_r)^{(1-x)}$$

If we do this, then

$$E_{NNA} = \frac{p_{NNA}}{p_{NNA} + p_{NNG}} \quad (2)$$

$$= \frac{1 - g_r}{g_r + (1 - g_r)} \quad (3)$$

$$= \frac{1}{\frac{g_r}{1 - g_r} + 1} \quad (4)$$

$$= \frac{1}{e^{\Delta M} + 1} \quad (5)$$

$$= \frac{e^{-\Delta M}}{1 + e^{-\Delta M}} \quad (6)$$

Recall that in the Gilchrist et al. framework, the reference codon has $\Delta M_{NNG, NNG} = 0 \Rightarrow e^{-\Delta M_{NNG, NNG}} = 1$. Thus, we have recovered the Gilchrist et al. model from the formulation of E_i under the assumption that natural selection has no impact on codon usage and codon NNG is the pre-defined reference codon. To see this, plug 0 for $\Delta\eta$ in equation (1)..

We can then calculate the expected RSCUS using equation (1) (using notation $E[O_i]$) and equation (6) for the two codon case. For simplicity assume, we are only considering a gene of average expression (defined as $\phi_g = 1$). Assume in this case that NNG is the reference codon ($\Delta M_{NNG, \Delta\eta NNG} = 0$).

$$E[RSCUS_{NNA}] = \frac{E[O_{NNA}]}{E_{NNA}} \quad (7)$$

$$= \frac{e^{-\Delta\eta_{NNA}} (e^{-\Delta M_{NNA}} + e^{-\Delta M_{NNG}})}{e^{-\Delta M_{NNA} - \Delta\eta_{NNA}} + e^{-\Delta M_{NNG} - \Delta\eta_{NNG}}} \quad (8)$$

$$= \frac{e^{-\Delta M_{NNA} - \Delta\eta_{NNA}} + e^{-\Delta M_{NNG} - \Delta\eta_{NNA}}}{e^{-\Delta M_{NNA} - \Delta\eta_{NNA}} + e^{-\Delta M_{NNG} - \Delta\eta_{NNG}}} \quad (9)$$

$$= \frac{e^{-\Delta M_{NNA} - \Delta\eta_{NNA}} + e^{-\Delta\eta_{NNA}}}{e^{-\Delta M_{NNA} - \Delta\eta_{NNA}} + 1} \quad (10)$$

This shows that the expected value of RSCUS for a two-codon amino acid is expected to increase as the strength of selection $\Delta\eta$ increases, which is desired. Note that $\Delta\eta$ in Gilchrist et al. is formulated in terms of selection *against* a codon relative to the reference, such that a negative value represents that a codon is favored relative to the reference. If $\Delta\eta = 0$ (i.e. selection does not favor either codon), then $E[RSCUS] = 1$. Also note that the expected RSCUS does not remain independent of the mutation bias. This means that even if s_N (i.e. the strength of natural selection) does not change between species, changes to the strength and direction of mutation bias across species could impact RSCUS. Assuming my math is right, I think one needs to be cautious when interpreting CAIS as representative of the differences in the efficiency of

selection across species except under very particular circumstances. One such case could be when it is known that mutation bias varies little across the species of interest. Looking at the species used in this manuscript, most of them have a GC content ranging around 0.41, so I suspect their results are okay.

Although I have not done so, I am sure this could be extended to the 4 and 6 codon amino acids.

We thank Reviewer 2 for explicitly laying out the math that was implicit in our Figures 1 and 2. While we keep our more heuristic presentation, our revised manuscript now more clearly acknowledges that the per-site codon adaptation bias depicted in Figure 1 has limited sensitivity to s^*N_e . The reason that we believe our approach worked despite this, is that we think the phenomenon is driven by what is shown in Figure 2. I.e., where N_e makes a difference is by determining the proteome-wide fraction of codons subject to significant codon adaptation, rather than by determining the strength of codon adaptation at any particular site or gene. We have made multiple changes to the texts to make this point clearer.

Another minor weakness of this work is that although the method is generally applicable to any species with an annotated genome and the code is publicly available, the code itself contains hard-coded values for GC% and amino acid frequencies across the 118 vertebrates. The lack of a more flexible tool may make it difficult for less computationally-experienced researchers to take advantage of this method.

Genome-wide %GC values are hard-coded because they were taken from the previous study of James et al. (2023) <https://doi.org/10.1093/molbev/msad073>. As summarized in the manuscript, genome-wide %GC was a byproduct of a scan of all six reading frames across genic and intergenic sequences available from NCBI with access dates between May and July 2019. The more complicated code used to calculate the intergenic %GC, and the code used to calculate amino acid frequencies is located at <https://github.com/MaselLab/CodonAdaptation-Index-of-Species>. Luckily, someone else just wrote a simpler end to end pipeline for us, on the basis of our preprint. We now note this in the Acknowledgements, and link to it: https://github.com/gavinmdouglas/handy_pop_gen/blob/main/CAIS.py.

<https://doi.org/10.7554/eLife.87335.2.sa0>