Translational Selection Frequently Overcomes Genetic Drift in Shaping Synonymous Codon Usage Patterns in Vertebrates

Aoife Doherty[1] and James O. McInerney*[,1]

[1]Bioinformatics and Molecular Evolution Unit, Department of Biology, National University of Ireland Maynooth, Maynooth, Co. Kildare, Ireland

***Corresponding author:** E-mail: james.o.mcinerney@nuim.ie.

Associate editor: John Novembre

Abstract

Synonymous codon usage patterns are shaped by a balance between mutation, drift, and natural selection. To date, detection of translational selection in vertebrates has proven to be a challenging task, obscured by small long-term effective population sizes in larger animals and the existence of isochores in some species. The consensus is that, in such species, natural selection is either completely ineffective at overcoming mutational pressures and genetic drift or perhaps is effective but so weak that it is not detectable. The aim of this research is to understand the interplay between mutation, selection, and genetic drift in vertebrates. We observe that although variation in mutational bias is undoubtedly the dominant force influencing codon usage, translational selection acts as a weak additional factor influencing synonymous codon usage. These observations indicate that translational selection is a widespread phenomenon in vertebrates and is not limited to a few species.

Key words: **translational selection, mutational bias, vertebrates, synonymous codon usage, genetic drift.**

Nonrandom usage of synonymous codons is a widespread phenomenon that has been observed across all three domains of life. In the past, the exact codon encoding an amino acid was thought to have little physiological effect on the cell. However, recently it has been demonstrated that synonymous codon choice at a particular sequence position affects various cellular mechanisms, including protein folding (Zhou et al. 2009), exon splicing (Parmley and Hurst 2007), translational accuracy (Akashi 1994), translational efficiency (Bulmer 1991), and protein function (Hudson et al. 2011). In addition, understanding synonymous codon usage (SCU) has biomedical and biotechnological applications. For example, synonymous mutations are implicated in the progression of many common human diseases (Kimchi-Sarfaty et al. 2007; Sauna and Kimchi-Sarfaty 2011) and considerably alter transgene production rates (Gustafsson et al. 2004; Angov et al. 2008). Therefore, synonymous codon usage patterns, and the factors that influence them, are of interest for a variety of reasons.

What causes synonymous codon usage bias? Does such bias exist because it is necessary for efficient and accurate protein expression? Or simply because codons are subject to variable mutational pressure? The generally accepted selection–mutation–drift theory asserts that natural selection favors optimal over nonoptimal codons, whereas neutral processes (mutational bias and random genetic drift) allow nonoptimal codons to persist (Bulmer 1988, 1991). The selectionist component of this theory posits that there is co-adaptation of codon usage and cellular tRNA content to optimize translational accuracy and/or efficiency (Akashi 1994; Stoletzki and Eyre-Walker 2007; Drummond and Wilke 2008). The evidence for such co-adaptation is 2-fold. First, selective pressure is expected to be stronger for highly expressed genes than for lowly expressed genes. Highly biased synonymous codon usage is correlated with high expression level in several organisms (Gouy and Gautier 1982; Sharp and Li 1987; Duret and Mouchiroud 1999; Goetz and Fuglsang 2005). Second, in organisms for which there is information available, codon usage is biased toward those that match the most abundant cellular tRNA or bind those tRNAs with optimal binding strength (Ikemura 1985; Moriyama and Powell 1997; Kanaya et al. 1999; Duret 2000). However, the selective advantage offered by alternative synonymous codons is quite weak. Given that the effectiveness of selection on alternative synonymous codon choice is determined by a combination of effective population size and the selective advantage of a codon over its synonymous alternatives, translational selection is expected to primarily operate in large populations. In small populations, for example in vertebrates, such selective coefficients are thought to be unable to overcome random genetic drift (Rao et al. 2011). Furthermore, vertebrate genomes (particularly those of mammals and birds) are compositionally compartmentalized into isochores, a feature that greatly influences codon usage (Bernardi and Bernardi 1986; Bernardi 1995). Consistent with this logic, several multivariate analyses have revealed a single major force governing vertebrate synonymous codon usage that is strongly correlated with GC content at the third codon position (GC_3) and does not discriminate any aspect of gene function or gene expression level (Urrutia and Hurst 2001; Rao et al. 2011).

Recently, evidence has emerged to suggest that translational selection is detectable in a handful of vertebrate species

Letter

(Musto et al. 2001; Romero et al. 2003; Urrutia and Hurst 2003; Yang and Nielsen 2008). With the availability of many complete vertebrate genomes, it has become possible to systematically analyze the determinants of genome-wide codon bias in these species. The objective of the this work is to combine increased vertebrate genome sampling with a sensitive codon usage bias index and human expression data to address the null hypothesis that translational selection is unable to overcome genetic drift in vertebrates. Due to the genome sampling bias that is occurring at present, it is important to note that the majority of the vertebrates examined in this analysis are from the Class Mammalia. Data are available from the Dryad Digital Repository (datadryad.org, doi:10.5061/dryad.4k887).

For this analysis, we retrieved protein-coding sequences (CDSs) for 38 vertebrates from Ensembl version 66 (Flicek et al. 2012). After filtering, a dataset comprising 558,871 genes was retained. An expression level was assigned to each human gene as the highest expression level a transcript attained, according to the Su et al. (2004) data set. Putative vertebrate orthologs were identified for each human gene using a reciprocal best-hit BLAST search (Altschul et al. 1990). Highly and lowly expressed genes were identified as the 5% highest and lowest expressed human genes and their orthologs in other vertebrate species.

Correspondence analysis (CA) is an exploratory method to investigate two-way and multi-way tables that contain some measure of correspondence between the rows and columns. The most common kind of table of this type is the two-way frequency table. In biological terms, CA is a commonly implemented technique to examine how synonymous codon usage co-varies with other biological traits, such as expression level or GC content (Shields et al. 1988; Musto et al. 2001). In line with recommendations from Lafay et al. (2000), Lerat et al. (2000), and Perrière and Thioulouse (2002), we conducted two correspondence analyses for each species (on relative synonymous codon usage [RSCU] values and raw codon counts) using codonW (Peden 1997). The principal axes of each species' CA were correlated with expression level and GC_3 content to understand the factors that primarily contribute to codon usage variation in vertebrates. In the CA of RSCU values, the first axis accounts for ~37.36% on average and the next highest axis accounts for ~4.22% of the relative inertia, on average (supplementary table S1, Supplementary Material online). Similarly, for the CA carried out with raw codon counts, the first axis accounts for ~30.83% of the relative inertia on average, while the next highest axis accounts for ~6.97% on average (supplementary table S1, Supplementary Material online). The most important axis (axis 1) was plotted against axis 2 for each correspondence analysis, for each species (supplementary fig. S1, Supplementary Material online). Most of the genes fall in a single cloud around the origin, and highly expressed genes are scattered throughout this cloud. There was no correlation detected between gene expression and axis 1 in any species, or gene expression and axis 2, according to a Spearman correlation (Spearman correlation, <0.1) (supplementary table S2, Supplementary Material online). For each species, axis 1 co-ordinates strongly correlated with GC_3 (Spearman correlation, ~0.96–0.99) (supplementary table S2, Supplementary Material online). Thus, it is apparent that there is a single major source of variation in synonymous codon usage that correlates with the GC_3 of a gene. As third codon positions are subject to less selective constraint than first or second positions, the observation that GC_3 correlates with the majority of the variation in the first axis of each correspondence analysis is to be expected if synonymous codon usage bias were due to mutational bias.

However, multivariate analysis might not have the power to tease very small selective pressures apart from the obviously strong inter-gene variation in GC content. Therefore, we specifically searched for small signals that might be indicative of translational selection in each genome. Two premises are commonly invoked as evidence of translational selection. First, if selection acts to enhance protein translation efficiency and/or accuracy, such selection should be particularly pronounced in highly expressed genes. Second, there is expected to be a correlation between the set of preferred codons used in highly expressed genes with the most abundant cellular tRNAs. For a more detailed understanding of tRNAs and the relationship between codon and anticodon, the authors guide the reader to Agris et al. (2007) and Gustilo et al. (2008). For each species, the codon bias of each highly and lowly expressed gene was calculated using the CDC index that is implemented in the Compositional Analysis Toolkit v. 1.0 (Zhang et al. 2012). Subsequently, the average codon bias for each species' set of highly and lowly expressed genes was computed. We used the Mann–Whitney test to examine whether highly expressed genes were significantly more biased in synonymous codon usage than lowly expressed genes. It has previously been demonstrated that some codon usage indices are biased by gene length (Urrutia and Hurst 2001). To eliminate this possible confounding factor, a subset of highly and lowly expressed genes were extracted from each species in which each highly expressed gene was paired to a lowly expressed partner gene of exactly equal length. Using these gene sets, the calculation of average codon bias using the CDC index was repeated. Using the Wilcoxon test, we examined whether highly expressed genes were significantly more biased in their codon usage than lowly expressed genes. When gene length was explicitly accounted for, the average codon bias for each set of highly expressed genes was in the range of ~0.14–0.15 (average across all species: 0.143). Conversely, the average codon bias for each set of lowly expressed genes was primarily in the range of ~0.12–0.14 (average across all species: 0.135). The difference in codon bias distributions between highly and lowly expressed genes was statistically significant in 21 of the 38 species (table 1). A similar set of observations was made when gene length was not explicitly controlled for. The average codon bias was ~0.14–0.16 in highly expressed genes (average across all species: 0.156) and ~0.12–0.13 in lowly expressed genes (average across all species: 0.128). In this instance, the difference in codon bias levels between highly and lowly expressed genes was statistically significant for all 38 species (table 1). Although the difference in average

Table 1. CDC Scores for Each Set of Highly Expressed (HE) and Lowly Expressed (LE) Genes (gene length is explicitly controlled), and Correlation between CDC Scores of Highly Expressed Genes and tRNA Abundance for Each Species.

	CDC Scores of HE and LE Genes				Correlation between CDC Scores and tRNA Abundance							
	Number of Genes	CDC (HE)	CDC (LE)	*P* Value	Total No.	Two-Codon Amino Acid	Amino Acids Encoded by More Than Two Codons					
							Pro	Leu	Ser	Val	Arg	Gly
Anole	69	0.13	0.12	0.02	10	8	CCA	CUG				
Cat	38	0.14	0.15	0.48	8	8						
Chicken	72	0.13	0.12	0.29	10	9		CUG				
Chimpanzee	177	0.14	0.13	0.12	10	8				GUG		GGC
Cow	161	0.14	0.13	0.01	8	7				GUG		
Dog	166	0.14	0.13	0	10	8		CUG		GUG		
Finch	68	0.13	0.13	0.35	12	9		CUG	AGC			GGC
Gibbon	157	0.14	0.14	0.05	10	8				GUG		GGC
Gorilla	195	0.14	0.13	0.04	9	7				GUG		GGC
Guinea pig	151	0.14	0.13	0.41	12	8		CUG	AGC	GUG		GGC
Horse	171	0.14	0.13	0.02	12	9			AGC	GUG		GGC
Human	182	0.14	0.13	0.02	10	8				GUG		GGC
Macaque	175	0.15	0.14	0	9	7				GUG		GGC
Marmoset	161	0.14	0.13	0.02	11	8		CUG		GUG		GGC
Mouse	61	0.15	0.13	0	12	9		CUG		GUG		GGC
Mouse lemur	172	0.14	0.15	0.59	12	8		CUG	AGC	GUG		GGC
Opossum	133	0.14	0.12	0	13	7	CCU	CUG	UCU	GUG	AGA	GGC
Orangutan	156	0.14	0.14	0.01	11	8		CUG		GUG		GGC
Panda	156	0.14	0.13	0	10	8			AGC			GGC
Pig	130	0.14	0.14	0.31	9	7				GUG		GGC
Platypus	90	0.14	0.12	0.01	6	5					CGG	
Rabbit	143	0.13	0.12	0.04	13	9		CUG	AGC	GUG		GGC
Rat	141	0.15	0.13	0	11	8		CUG			AGG	GGC

NOTE.—The first five columns compare the CDC scores for highly and lowly expressed genes. Column 1 is the species name. Column 2 is the number of highly and lowly expressed genes whose CDC scores were compared once gene length was explicitly controlled for. Columns 3 and 4 are the average CDC score for the highly and lowly expressed genes, respectively. Column 5 is the *P* value, according to a Wilcoxon test. The last eight columns correlate CDC scores with tRNA abundance. Column 6 (named "Total No.") indicates the total number of cases (out of 20 amino acids) in which the preferred codon matched the most abundant tRNA gene for an amino acid in each set of highly expressed genes. Column 7 demonstrates the number of these cases in which the amino acid was encoded by exactly two codons. The remaining columns indicate, for those amino acids encoded by more than two codons, precisely which codons matched the most abundant tRNA genes.

CDC index between highly and lowly expressed genes is quite small (generally ~0.01 difference in CDC scores between highly and lowly expressed genes), a slightly larger discrepancy (~0.02) was observed in rodents, marsupials, and monotremes. Rodents and marsupials have previously been shown to have higher effective population sizes than other mammals, such as primates (Hughes and Friedman 2009). This is compatible with the model in which species with a higher effective population size are more likely to display evidence of translational selection than those with smaller effective population sizes (Charlesworth 2009).

We asked whether the synonymous codons that were preferentially used in highly expressed genes tended to correlate with the most abundant cellular tRNAs. It has previously been demonstrated that the abundance of cellular tRNA is correlated with the number of tRNA genes in a genome (Ikemura 1981; Percudani et al. 1997; Kanaya et al. 1999). The number of tRNA genes for each codon for 23 of the species was retrieved from the Genomic tRNA database (retrieved 18 April 2012) (Chan and Lowe 2009). RSCU values for each amino acid were calculated for each set of highly and lowly expressed orthologs using codonW to identify the codon preferentially used to encode each amino acid in highly and lowly expressed genes. The preferred codon for each amino acid was defined as that with the highest RSCU value. Then, we examined whether the preferred codon for each amino acid matched the most abundant tRNA gene for that amino acid in each species. A correlation between tRNA abundance and codon preference was regularly observed in those amino acids that are encoded by two codons in all the species examined (table 1). For amino acids that are encoded by more than two codons, matches are still observed in all 23 species for leucine, valine, glycine, proline, serine, and arginine (table 1).

To our knowledge, these observations are the first to suggest that translational selection pervasively influences (albeit weakly) vertebrate synonymous codon usage and is not restricted to exceptional circumstances in a few species (Musto et al. 2001; Romero et al. 2003). Understanding the role of weak selective pressure in governing synonymous codon usage can provide vital insights into the interplay of selection, drift, mutation bias, and long-term effective population sizes. The consistency between the strength of translational

selection and long-term effective population size of the species in which signals of selection are observed might suggest that synonymous codon usage statistics might act as a proxy for estimates of long-term effective population sizes of species that are currently not well understood.

Supplementary Material

Supplementary tables S1 and S2 and figure S1 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/). All data will be uploaded to the Dryad database.

Acknowledgments

The authors sincerely thank Dr David Alvarez-Ponce for providing the curated Su et al. (2004) gene expression data set. The authors would also like to thank the Irish Centre for High End Computing and the NUI Maynooth HPC facility. This work was supported by an Irish Research Council for Science Engineering and Technology grant to A.D.

References

Agris P, Vendeix F, Graham W. 2007. tRNA's wobble decoding of the genome: 40 years of modification. *J Mol Biol.* 366:1–13.

Akashi H. 1994. Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. *Genetics* 136:927–935.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol.* 215:403–410.

Angov E, Hillier CJ, Kincaid RL, Lyon JA. 2008. Heterologous protein expression is enhanced by harmonizing the codon usage frequencies of the target gene with those of the expression host. *PLoS One* 3: e2189.

Bernardi G. 1995. The human genome: organization and evolutionary history. *Annu Rev Genet.* 29:445–476.

Bernardi G, Bernardi G. 1986. Compositional constraints and genome evolution. *J Mol Evol.* 24:1–11.

Bulmer M. 1988. Are codon usage patterns in unicellular organisms determined by selection-mutation balance? *J Evol Biol.* 1:15–26.

Bulmer M. 1991. The selection-mutation-drift theory of synonymous codon usage. *Genetics* 129:897–907.

Chan PP, Lowe TM. 2009. Gtrnadb: a database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Res.* 37: D93–D97.

Charlesworth B. 2009. Effective population size and patterns of molecular evolution and variation. *Nat Rev Genet.* 10:195–205.

Drummond DA, Wilke CO. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* 134:341–352.

Duret L. 2000. tRNA gene number and codon usage in the *C. elegans* genome are co-adapted for optimal translation of highly expressed genes. *Trends Genet.* 16:287–289.

Duret L, Mouchiroud D. 1999. Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. *Proc Natl Acad Sci U S A.* 96:4482.

Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S. 2012. Ensembl 2012. *Nucleic Acids Res.* 40:D84–D90.

Goetz RM, Fuglsang A. 2005. Correlation of codon bias measures with mRNA levels: analysis of transcriptome data from *Escherichia coli*. *Biochem Biophys Res Commun.* 327:4–7.

Gouy M, Gautier C. 1982. Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Res.* 10:7055–7074.

Gustafsson C, Govindarajan S, Minshull J. 2004. Codon bias and heterologous protein expression. *Trends Biotechnol.* 22:346–353.

Gustilo E, Vendeix F, Agris P. 2008. tRNA's modification brings order to gene expression. *Curr Opin Microbiol.* 11:134–140.

Hudson NJ, Gu Q, Nagaraj SH, Ding YS, Dalrymple BP, Reverter A. 2011. Eukaryotic evolutionary transitions are associated with extreme codon bias in functionally-related proteins. *PLoS One* 6: e25457.

Hughes AL, Friedman R. 2009. More radical amino acid replacements in primates than in rodents: support for the evolutionary role of effective population size. *Gene* 440:50–56.

Ikemura T. 1981. 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.* 151:389.

Ikemura T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. *Mol Biol Evol.* 2:13–34.

Kanaya S, Yamada Y, Kudo Y, Ikemura T. 1999. Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAS: gene expression level and species-specific diversity of codon usage based on multivariate analysis. *Gene* 238: 143–155.

Kimchi-Sarfaty C, Oh JM, Kim IW, Sauna ZE, Calcagno AM, Ambudkar SV, Gottesman MM. 2007. A "silent" polymorphism in the MDR1 gene changes substrate specificity. *Science* 315:525–528.

Lafay B, Atherton JC, Sharp PM. 2000. Absence of translationally selected synonymous codon usage bias in *Helicobacter pylori*. *Microbiology* 146:851–860.

Lerat E, Biémont C, Capy P. 2000. Codon usage and the origin of P elements. *Mol Biol Evol.* 17:467–468.

Moriyama EN, Powell JR. 1997. Codon usage bias and tRNA abundance in *Drosophila*. *J Mol Evol.* 45:514–523.

Musto H, Cruveiller S, D'Onofrio G, Romero H, Bernardi G. 2001. Translational selection on codon usage in *Xenopus laevis*. *Mol Biol Evol.* 18:1703.

Parmley JL, Hurst LD. 2007. Exonic splicing regulatory elements skew synonymous codon usage near intron-exon boundaries in mammals. *Mol Biol Evol.* 24:1600–1603.

Peden J. 1997. CodonW. Dublin: Trinity College.

Percudani R, Pavesi A, Ottonello S. 1997. Transfer RNA gene redundancy and translational selection in *Saccharomyces cerevisiae*. *J Mol Biol.* 268:322–330.

Perrière G, Thioulouse J. 2002. Use and misuse of correspondence - analysis in codon usage studies. *Nucleic Acids Res.* 30: 4548–4555.

Rao Y, Wu G, Wang Z, Chai X, Nie Q, Zhang X. 2011. Mutation bias is the driving force of codon usage in the *Gallus gallus* genome. *DNA Res.* 18:499–512.

Romero H, Zavala A, Musto H, Bernardi G. 2003. The influence of translational selection on codon usage in fishes from the family cyprinidae. *Gene* 317:141–147.

Sauna ZE, Kimchi-Sarfaty C. 2011. Understanding the contribution of synonymous mutations to human disease. *Nat Rev Genet.* 12: 683–691.

Sharp PM, Li WH. 1987. The codon Adaptation Index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* 15:1281–1295.

Shields DC, Sharp PM, Higgins DG, Wright F. 1988. "Silent" sites in *Drosophila* genes are not neutral: evidence of selection among synonymous codons. *Mol Biol Evol.* 5:704–716.

Stoletzki N, Eyre-Walker A. 2007. Synonymous codon usage in *Escherichia coli*: selection for translational accuracy. *Mol Biol Evol.* 24:374–381.

Su AI, Wiltshire T, Batalov S, Lapp H, Ching K, Block D, Zhang J, Soden R, Hayakawa M, Kreimen G. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A.* 101:6062–6067.

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.

Urrutia AO, Hurst LD. 2003. The signature of selection mediated by expression on human genes. *Genome Res.* 13:2260–2264.

Yang Z, Nielsen R. 2008. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Mol Biol Evol.* 25:568–579.

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* 13:43.

Zhou T, Weems M, Wilke CO. 2009. Translationally optimal codons associate with structurally sensitive sites in proteins. *Mol Biol Evol.* 26:1571–1580.