DOI: 10.1534/genetics.107.086405

Note

Nearly Neutrality and the Evolution of Codon Usage Bias in Eukaryotic Genomes

Sankar Subramanian[1]

Allan Wilson Centre for Molecular Ecology and Evolution, Institute of Molecular BioSciences, Massey University, Auckland, 0632, New Zealand

Manuscript received December 23, 2007
Accepted for publication February 6, 2008

ABSTRACT

Here I show that the mean codon usage bias of a genome, and of the lowly expressed genes in a genome, is largely similar across eukaryotes ranging from unicellular protists to vertebrates. Conversely, this bias in housekeeping genes and in highly expressed genes has a remarkable inverse relationship with species generation time that varies by more than four orders of magnitude. The relevance of these results to the nearly neutral theory of molecular evolution is discussed.

THE nearly neutral theory of evolution proposed by Ohta (1992) predicts that the fate of a substantial fraction of mutations in a population is determined by both natural selection and random genetic drift. According to this theory, the eventual fixation of these mutations is determined by the product of effective population size (N_e) and selection coefficient (s). Therefore, the future of mutations with marginal fitness effects is largely governed by population size. A number of investigations including important studies such as the influence of population size on deleterious mutational load and on the evolution of genome complexity have confirmed this prediction (Keightley and Eyre-Walker 2000; Lynch and Conery 2003).

Similarly, selection on synonymous positions leading to bias in codon usage is also known to be weak (Akashi 1995, 1997; Llopart and Aguade 2000). Therefore codon usage bias is an ideal candidate to test the nearly neutral expectation of the population-size effects on weakly selected mutations. However, such studies are scarce, except for a few on a limited data set from closely related Drosophila species (Akashi 1995; Tamura *et al.* 2004). Although a recent study on a large number of prokaryotic genomes has been reported (Rocha 2004), there has been no previous attempt to examine the magnitude of codon usage bias among various eukaryotes ranging from unicellular protists to vertebrates and its relationship with their population size. The major reason for this limitation is because the magnitude of selection on synonymous codon usage could vary not only among the genomes of eukaryotes, but also among genes of a genome (Li 1997; Ohta 2002).

To examine this, I have assembled a data set of protein coding sequences from 20 eukaryotic species presumably having a wide range of population sizes (Figure 1 legend). First, I examined the variation in the codon usage bias of the whole genome as well as that of the housekeeping genes that are common to all these eukaryotes. In addition, using the gene expression data I investigated these patterns among the genes with very high and very low expression levels. I then examined whether such variations in the codon usage bias could largely be explained by the differences in population size. Since estimates of effective population sizes are hard to obtain, I have used species generation times as a proxy for population size because these two measures are well-known to correlate (Chao and Carr 1993; Ohta 1993; Keightley and Eyre-Walker 2000). Furthermore, studies in prokaryotes imply that the strength of selection for translational efficiency is itself correlated to generation time (Dong *et al.* 1996; Rocha 2004).

Codon usage bias was estimated using the modified effective codon number (ENC′) method, which accounts for base compositional difference caused by unequal rates of forward and reverse mutations (Novembre 2002). The genomic mean ENC′ was estimated using all the genes of a genome. Also the mean ENC′ was computed using the genes involved in translation, which largely consist of ribosomal genes, tRNA synthetases, initiation and elongation factors. These genes have been

[1] *Address for correspondence:* Allan Wilson Centre for Molecular Ecology and Evolution, Institute of Molecular BioSciences, Massey University, Private Bag 102 904, North Shore Mail Centre, Auckland, 0632, New Zealand. E-mail: s.sankarasubramanian@massey.ac.nz

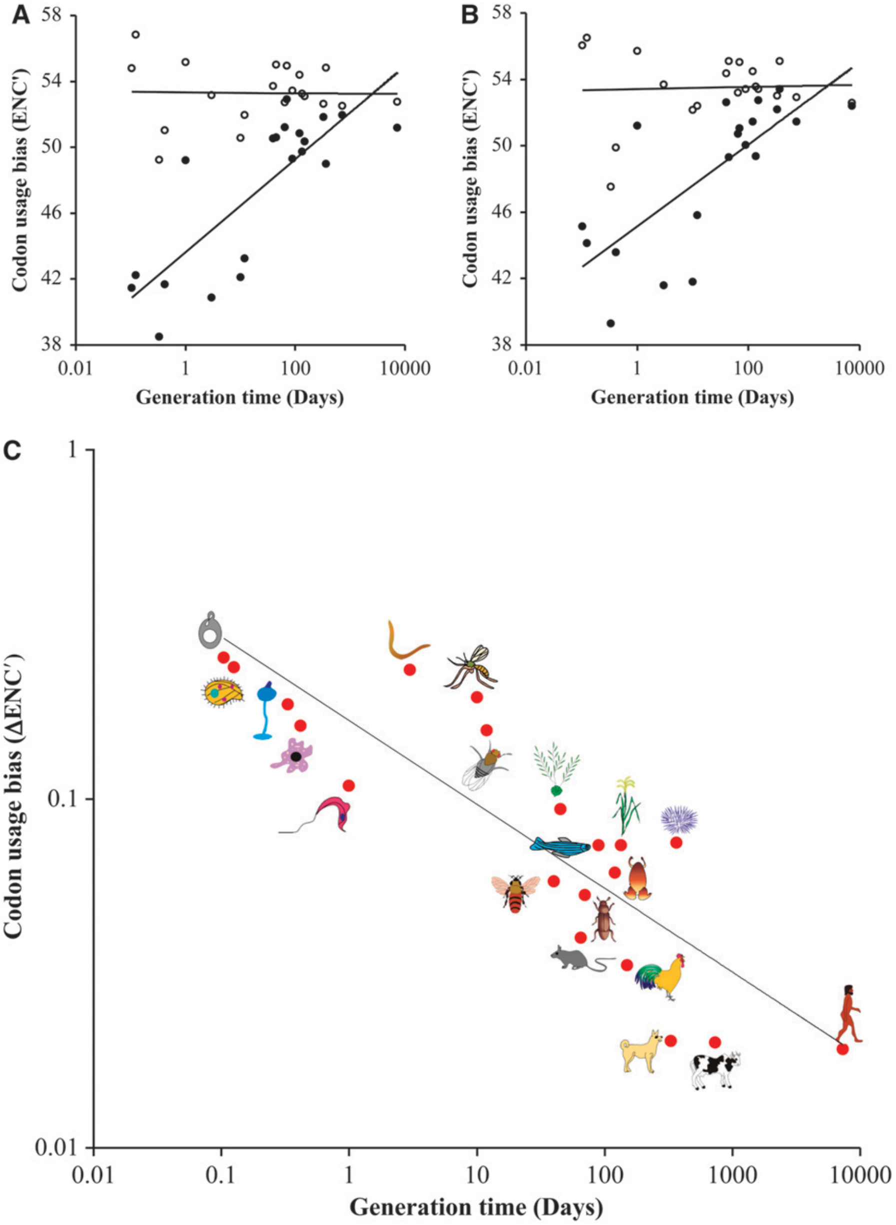

FIGURE 1.—Relationship between codon usage bias and generation time of eukaryotes. The protein-coding sequences of complete or nearly complete genomes of 20 eukaryotic species from various public data banks were obtained. The gene expression data in the form of expressed sequence tags (ESTs) were obtained from dbEST (http://www.ncbi.nlm.nih.gov) and using BLASTN the ESTs were matched to the respective genes using the method described before (DURET and MOUCHIROUD 2000). The species data set was chosen on the basis of the availability of a large number of genes as well as their corresponding gene expression data. Also the species were chosen to represent the major groups of eukaryotes and to get a wide distribution of generation times. Furthermore the choice of EST instead of microarray data (or other expression data) was purely based on its availability for all the species used in this study. To estimate the codon usage bias, the method ENC′ (NOVEMBRE 2002) was employed using the software ENC prime (http://home.uchicago.edu/~jnovembre/software/software.html). Although a recent report pointed out a drawback of the ENC′ method, this does not affect when the codon bias estimates are used in a relative manner such as in correlation (FUGLSANG 2006). The numbers of genes in the genomes, translational genes, lowly expressed genes (with 1EST), highly expressed genes (top 1%), and generation time (days) of the species used are as follows: *Anopheles gambiae* (4877, 50, 804, 39, 10); *Apis mellifera* (7854, 124, 911, 59, 40); *Arabidopsis thaliana* (26,536, 69, 2405, 70, 45); *Bos taurus* (18,895, 185, 1784, 48, 730); *Caenorhabditis elegans* (20,043, 136, 1674, 64, 3); *Canis familiaris* (19,599, 191, 2961, 77, 330); *Danio rerio* (23,482, 126, 2549, 41, 90); *Dictyostelium discoideum* (13,147, 102, 819, 29, 0.3); *Drosophila melanogaster* (13,982, 114, 1444, 63, 12); *Entamoeba histolytica* (9531, 128, 471, 11, 0.42); *Gallus gallus* (9518, 48, 2272, 57, 150); *Homo sapiens* (28,015, 226, 4515, 111, 7300); *Mus musculus* (30,079, 219, 3406, 105, 65); *Oryza sativa* (23,311, 276, 3980, 121, 135); *Saccharomyces cerevisia* (6687, 258, 1219, 27, 0.1); *Strongylocentrotus purpuratus* (17,472, 78, 1915, 63, 365); *Tetrahymena thermophila* (27,355, 120, 3655, 98, 0.13); *Tribolium castaneum* (9221, 49, 1456, 36, 70); *Trypanosoma cruzi* (15,546, 145, 2521, 41, 1); and *Xenopus tropicalis* (5477, 47, 371, 51, 120). The sources of generation-time information are given in supplemental Table 1. (A) The correlation of the codon usage bias (ENC′) of all genes of the genomes (open circles) and that of the genes involved in translation (predominantly consist of ribosomal genes, tRNA syntetases, initiation and elongation factors) (solid circles) with generation time. *x*-axis is shown in log scale. Spearman's coefficient for the genome, $\rho = -0.15$, $P = 0.52$ and for translational genes, $\rho = 0.77$, $P = 0.0008$. (B) The relationship of the ENC′ estimated for the genes with low (open circles) and high (solid circles) expression levels (excluding the translational genes) with generation time. Spearman's coefficient for the lowly expressed genes, $\rho = -0.08$, $P = 0.74$ and for the highly expressed genes, $\rho = 0.74$, $P = 0.0014$. (C) The log–log relationship between ΔENC′ and species generation time. Here $\Delta\mathrm{ENC}' = (\mathrm{ENC}'_{\mathrm{L}} - \mathrm{ENC}'_{\mathrm{TH}})/\mathrm{ENC}'_{\mathrm{L}}$, where $\mathrm{ENC}'_{\mathrm{TH}}$ is the average codon usage bias of translational + highly expressed genes and $\mathrm{ENC}'_{\mathrm{L}}$ is that of low-expressed genes. Spearman's coefficient for all species $\rho = -0.87$, $P = 0.0002$ and for the vertebrate subset $\rho = -0.89$, $P = 0.029$. The best-fitting linear regression lines are shown.

chosen due to their essential functionality and ubiquitous presence in all eukaryotes as well as their expression in all tissues. Genomic ENC′ is mostly similar across eukaryotes and does not show any significant relationship with species generation times ($\rho = -0.15$, $P = 0.52$) (Figure 1A). Conversely, translational genes show an excellent correspondence with the generation time ($\rho = 0.77$, $P = 0.0008$). Since selection on synonymous sites is also known to be modulated by the level of gene expression, the average ENC′ was computed for genes with the highest level of expression (the top 1% of genes excluding translational genes) and for genes with the least expression level (1 EST). The relationships shown in Figure 1, A and B are qualitatively similar. The similarity between the mean ENC′s of the genomes and the lowly expressed genes suggests that the majority of the genes of the genomes have low codon usage bias. On the other hand the mean ENC′ of translational and highly expressed genes suggests that the magnitudes of selection on these two sets of genes are largely the same.

Since the high ENC′values of the genome or the lowly expressed genes suggest a minimal bias in codon usage, this could be used as a baseline to quantify the extent of codon usage bias in translational and highly expressed genes. Therefore the difference in the bias was estimated using the formula $\Delta\mathrm{ENC}' = (\mathrm{ENC}'_{\mathrm{L}} - \mathrm{ENC}'_{\mathrm{TH}})/\mathrm{ENC}'_{\mathrm{L}}$ (Rocha 2004), where $\mathrm{ENC}'_{\mathrm{TH}}$ is the average codon usage bias of translational + highly expressed genes and $\mathrm{ENC}'_{\mathrm{L}}$ is that of low-expressed genes (using the genomic ENC′ instead of that of low-expressed genes also produced similar results). Figure 1C shows a highly significant negative relationship between ΔENC′ and generation time ($\rho = -0.87$, $P = 0.0002$). This also held true when only vertebrates were considered ($\rho = -0.89$, $P = 0.029$). The average ΔENC′ of vertebrates, (invertebrates + plants) and protists was 0.038, 0.12, and 0.19, respectively, which suggests that the relative estimates of (invertebrates + plants) and of protists are approximately three and five times higher than that of vertebrates.

To examine the phylogenetic nonindependence of this result, independence contrast analysis (Felsenstein 1985) was performed using the CONTRAST package of PHYLIP (Felsenstein 2005). The concatenated alignment of 32 orthologous proteins (obtained through a reciprocal BLAST search) that are common for all the 20 species were used to construct a neighbor-joining tree (see supplemental Figure 1) and the branch lengths were used for the contrast analysis. I have also used a widely accepted eukaryotic tree topology (see supplemental Figure 2) and conducted this analysis using the software CAIC (Purvis and Rambaut 1995). The results from both analyses showed a highly significant relationship ($r = -0.72$, $P < 0.0005$) between the standardized contrasts of generation time and $\Delta ENC'$ (see supplemental Table 2). Furthermore, no significant relationship ($r = 0.23$, $P > 0.35$) between these contrasts and their variances was observed (see supplemental Table 2) (Garland *et al.* 1992; Purvis and Rambaut 1995). These results suggest that the correlation between generation time and codon usage bias observed in this study is independent and not influenced by the phylogenetic relationship of the species used.

The negative relationship between the $\Delta ENC'$ of translational/highly expressed genes and generation time implies that the selection coefficient (s) is much higher than the fixation probability of a neutral mutation ($s > 1/2N_e$) but it is small enough to be modulated by population size to fixation (Ohta 1992). This pattern could be explained in two ways on the basis of the variable(s) that correlates with generation time. If population size correlates with generation time, then this relationship could be explained by the differences in population sizes alone by assuming a similarity of s across all species. This seems unlikely because for nearly neutral mutations the absolute value $|N_e s|$ has to be between 1 and 2 (Ohta 2002) and thus the intermediate codon usage bias (as opposed to zero or complete codon usage bias) could vary only within this small window. Since the population sizes of the species used in this study vary by several orders of magnitude ($\sim 10^8$ unicellular eukaryotes and 10^4 for human), the observed wide range of $\Delta ENC'$ with respect to the species generation times could not explain the differences in N_e alone by assuming a constant s. An alternate and more likely possibility is that the strength of selection for translational efficiency also correlates with generation time and thus generation time here seems to represent $N_e s$ rather than N_e. For example, the effects of mildly deleterious mutations could delay the process of translation and this would affect species with short generation times drastically, as the mutants would be quickly outgrown by the wild types. Similarly, a slightly beneficial mutant would swiftly spread through the population of these species. However, fixation (or elimination) of such mutations is less effective in species with long generation times. Studies on the relationship between growth rates and codon usage bias in prokaryotes support this prediction (Dong *et al.* 1996; Rocha 2004).

Recent studies on mammals suggest selection on synonymous sites is caused by factors other than translational selection such as mRNA stability, alternative splicing, and micro RNA binding or the presence of exonic enhancers (Parmley and Hurst 2007), which might underestimate the absolute values of ENC′ for mammals. However, as these factors influence the low-expressed genes as well as translational and highly expressed genes, the relative ratio ΔENC′ is not affected due to the cancellation of their effects. Furthermore, some of these studies also suggest that only a very small proportion of synonymous positions are affected by these factors as the resultant reduction of divergence in these sites is marginal (1–8%) (Hurst 2006; Parmley *et al.* 2006).

The results of this study reveal the relative magnitude of codon usage bias in eukaryotes modulated by their population sizes and also explain the reduction of this bias in species such as vertebrates.

The author is grateful to David Lambert and acknowledges the support from Center of Research Excellence Fund, the Marsden Fund, and Massey University. I thank Adam Eyre-Walker and an anonymous reviewer for their valuable suggestions. I also thank John Waugh and Bill Peacock for the careful reading of the manuscript.

LITERATURE CITED

Akashi, H., 1995 Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in Drosophila DNA. Genetics **139:** 1067–1076.

Akashi, H., 1997 Codon bias evolution in Drosophila. Population genetics of mutation-selection drift. Gene **205:** 269–278.

Chao, L., and D. E. Carr, 1993 The molecular clock and the relationship between population-size and generation time. Evolution **47:** 688–690.

Dong, H., L. Nilsson and C. G. Kurland, 1996 Co-variation of tRNA abundance and codon usage in Escherichia coli at different growth rates. J. Mol. Biol. **260:** 649–663.

Duret, L., and D. Mouchiroud, 2000 Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. Mol. Biol. Evol. **17:** 68–74.

Felsenstein, J., 1985 Phylogenies and the comparative method. Am. Nat. **125:** 1–15.

Felsenstein, J., 2005 *PHYLIP (Phylogeny Inference Package) Version 3.6*. University of Washington, Seattle.

Fuglsang, A., 2006 Accounting for background nucleotide composition when measuring codon usage bias: brilliant idea, difficult in practice. Mol. Biol. Evol. **23:** 1345–1347.

Garland, T., P. H. Harvey and A. R. Ives, 1992 Procedures for the analysis of comparative data using phylogenetically independent contrasts. Syst. Biol. **41:** 18–32.

Hurst, L. D., 2006 Preliminary assessment of the impact of microRNA-mediated regulation on coding sequence evolution in mammals. J. Mol. Evol. **63:** 174–182.

Keightley, P. D., and A. Eyre-Walker, 2000 Deleterious mutations and the evolution of sex. Science **290:** 331–333.

Li, W.-H., 1997 *Molecular Evolution*. Sinauer Associates, Sunderland, MA.

Llopart, A., and M. Aguade, 2000 Nucleotide polymorphism at the RpII215 gene in Drosophila subobscura: weak selection on synonymous mutations. Genetics **155:** 1245–1252.

Lynch, M., and J. S. Conery, 2003 The origins of genome complexity. Science **302:** 1401–1404.

Novembre, J. A., 2002 Accounting for background nucleotide composition when measuring codon usage bias. Mol. Biol. Evol. **19:** 1390–1394.

Ohta, T., 1992 The nearly neutral theory of molecular evolution. Annu. Rev. Ecol. Syst. **23:** 263–286.

Ohta, T., 1993 An examination of the generation-time effect on molecular evolution. Proc. Natl. Acad. Sci. USA **90:** 10676–10680.

Ohta, T., 2002 Near-neutrality in evolution of genes and gene regulation. Proc. Natl. Acad. Sci. USA **99:** 16134–16137.

Parmley, J. L., and L. D. Hurst, 2007 How do synonymous mutations affect fitness? BioEssays **29:** 515–519.

Parmley, J. L., J. V. Chamary and L. D. Hurst, 2006 Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. Mol. Biol. Evol. **23:** 301–309.

Purvis, A., and A. Rambaut, 1995 Comparative analysis by independent contrasts (CAIC): an Apple Macintosh application for analysing comparative data. Comput. Appl. Biosci. **11:** 247–251.

Rocha, E. P. C., 2004 Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient decoding for translation optimization. Genome Res. **14:** 2279–2286.

Tamura, K., S. Subramanian and S. Kumar, 2004 Temporal patterns of fruit fly (Drosophila) evolution revealed by mutation clocks. Mol. Biol. Evol. **21:** 36–44.

Communicating editor: S. G. Hamish