January 2003

Codon Usage

Estuko N. Moriyama
University of Nebraska - Lincoln, emoriyama2@unl.edu

Published in *Encyclopedia of the Human Genome.*

Codon Usage

Etsuko N. Moriyama

University of Nebraska–Lincoln, Lincoln, Nebraska, USA; email: emoriyama2@unl.edu

Summary
The genetic codes have degeneracy; that is, most amino acids (18 out of 20 in the universal genetic code) are encoded by more than one codon. Codons encoding the same amino acid are called synonymous codons. Both in prokaryotic and eukaryotic genes, the synonymous codons are not used with equal frequencies.

Keywords: synonymous codon, translational selection, mutation bias, tRNA abundance, gene expression

Introduction

Unequal use of synonymous codons, or "codon usage bias" in short, has been found in many different organisms including both prokaryotes and eukaryotes. Patterns and degrees of codon usage bias vary not only among different organisms, but also among genes in the same genome. Both selective constraints and mutation bias seem to affect codon usage bias. Therefore, by examining codon usage bias we are able to detect changes in these two evolutionary forces between genomes or along one genome. For example, markedly different codon usage bias in one gene as compared with other surrounding genes might imply its foreign origin owing to horizontal transfer, a difference in functional constraints, or a difference in regional patterns of mutation. Further analyses would indicate which of these is the more likely explanation.

Functional constraints on synonymous codon usage are related to the level or pattern of gene expression. Therefore, examining codon usage bias may reveal some changes in functionality of the gene. Codon usage bias can also vary within a gene. For example, a study has shown that functionally important regions, such as deoxyribonucleic acid (DNA)-binding domains, tend to have stronger codon usage bias relative to other gene regions. Interspecies comparisons of codon usage bias can give us another level of information: genome-wide differences in codon usage bias between species would imply that evolutionary forces have been changed between the species. (See A0073.)

Several methods that are used to estimate codon usage bias are described in this article. The variation in codon usage bias among different organisms is then examined, followed by a discussion of the effect of selective constraints and mutation bias on codon usage bias in different genomes.

Measures of Codon Usage Bias

Several indices can be used to measure the degree of nonrandom usage of synonymous codons in a gene, of which a few representatives are described below.

Frequency of optimal codons

The frequency of optimal codons (F_{op}) is the simplest measure of species-specific codon usage bias.

$$F_{op} = \frac{X_{op}}{X_{op} + X_{non}}$$

where X_{op} and X_{non} are the numbers of "optimal" and "non-optimal" codons in a gene, respectively. Stop codons and codons for methionine, tryptophan, and other amino acids whose optimal codons are undetermined are excluded from the calculation.

Optimal codons were originally determined for *Escherichia coli* and *Saccharomyces cerevisiae* on the basis of availability of transfer ribonucleic acid (tRNA) and the nature of the codon-anticodon interaction (reviewed in Ikemura, 1992; see also "Causes of codon usage bias" below). These codons are considered to be translationally optimal and are found more often in genes that are expressed highly than in genes with low expression. Therefore, optimal codons can also be defined as those that occur in high-expression genes significantly more frequently than they occur in low-expression genes (e.g. see Stenico *et al.,* 1994).

Relative synonymous codon usage

The relative synonymous codon usage (RSCU) value for each codon is calculated as the observed number of occurrences divided by the number expected if all synonymous codons for an amino acid were used equally frequently. For synonymous codon i of an n-fold degenerate amino acid:

$$\mathrm{RSCU}_i = \frac{X_i}{\frac{1}{n}\sum_{i=1}^{n} X_i}$$

where X_i is the number of occurrences of codon i, and n is 1, 2, 3, 4, or 6.

Codon adaptation index

The codon adaptation index (CAI) estimates the extent of bias toward codons that are known to be favored in highly expressed genes (Sharp and Li, 1987a). A "relative adaptedness" value, w_i, for codon i is calculated from its relative frequency of use in a species-specific reference set of very highly expressed genes.

$$w_i = \frac{\mathrm{RSCU}_i}{\mathrm{RSCU}_{\max}} = \frac{X_i}{X_{\max}}$$

where $\mathrm{RSCU}_{\max}$ and $X_{\max}$ are the RSCU and X values for the most frequently used codon for an amino acid. The CAI for a gene is then defined as the geometric mean of w values for codons in that gene:

$$\mathrm{CAI} = \left(\prod_{i=1}^{L} w_i\right)^{1/L} \quad \text{or} \quad \exp\left(\frac{1}{L}\sum_{i=1}^{L} \ln(w_i)\right)$$

where L is the number of codons in the gene excluding methionine, tryptophan, and stop codons. The CAI ranges from 0 for no bias (all synonymous codons are used equally) to 1 for the strongest bias (only optimal codons are used).

Effective number of codons

The effective number of codons (N,) is a general measure of bias from equal codon usage in a gene (Wright, 1990). Knowledge of the optimal codons or the "reference set" of highly expressed genes is not required. It is estimated as

$$N_c = 2 + \frac{9}{\bar{F}_2} + \frac{1}{\bar{F}_3} + \frac{5}{\bar{F}_4} + \frac{3}{\bar{F}_6}$$

where $\bar{F}_k$ (k =2,3,4 or 6) is the average of F_k values for k-fold degenerate amino acids. F_k for each of k-fold degenerate amino acids is estimated as

$$F_k = \frac{nS - 1}{n - 1}$$

where n is the total number of codons for that amino acid, and

$$S = \sum_{i=1}^{k} \left(\frac{n_i}{n}\right)^2$$

where n_i is the number of occurrences of the ith codon for this amino acid. N_c is analogous to the "effective number of alleles" (and $\bar{F}_k$ is the "average homozygosity" for k-allele loci) that is used in population genetics. It gives the number of equally used codons that would generate the same codon usage bias observed. N_c ranges from 20 for the strongest bias (where only one codon is used for each amino acid) to 61 for no bias (where all synonymous codons are used equally).

Scaled χ^2

Scaled χ^2 (χ^2/L) is another measure of bias from equal codon usage in a gene (Shields *et al.,* 1988). The χ^2 value calculated for a deviation from equal usage of codons within synonymous groups is divided by the total number of codons, L, in the gene. Tryptophan and methionine codons are excluded from the calculation.

The null hypothesis for N_c and χ^2/L is that synonymous codons are used equally. But if there is a bias in mutation pattern, then this will generate bias in synonymous codon usage. To see the codon usage bias separately from the mutation pattern, the deviation from synonymous codon usage predicted from a given mutation pattern needs to be calculated. Corrections can be made to incorporate a nonrandom mutation pattern into the methods mentioned above (e.g. see Akashi, 1995).

In general, codon bias values estimated by different methods are highly correlated, although χ^2/L has been found to be more affected by gene length than CAI and N_c. This effect is pronounced when gene length is short (Comeron and Aguadé, 1998; Moriyama and Powell, 1998).

Codon Usage Bias in Different Genomes

Table 1 shows the synonymous codon usage of five fourfold degenerate amino acids for four organisms: *E. coli, S. cerevisiae, Drosophila melanogaster,* and *Homo sapiens*. Clearly, these synonymous codons are not used equally. Pattern and bias in synonymous codon usage varies greatly among different organisms and also among genes in the same genome. Figure 1a–d shows the distribution of values for N_c. Note that N_c measures only the degree of codon bias from equal usage, and does not give us information regarding directionality of the bias. In Figure 1e–h, N_c values are plotted against the guanine plus cytosine content (G + C%) at third-codon positions in the genes. In *D. melanogaster* genes, G-ending and especially C-ending synonymous codons are used preferably (Table 1), as shown clearly in Figure 1g. In other words, genes with high codon bias (represented by lower N_c values) have a higher G +C% at third-codon positions. Codon usage bias in human genes seems to be driven in two opposite directions: toward adenine and thymine (AT)-richness and toward GC-richness (Figure 1h and Table 1). The human genome comprises a mosaic of long stretches of GC-rich and AT-rich regions the so-called "isochore" structure. This bidirectional codon usage in human genes is shown in Figure 1h.

Table 1 Codon usage of fourfold degenerate amino acids [a]

	Escherichia coli		*Saccharomyces cerevisiae*		*Drosophila melanogaster*		*Homo sapiens*	
Codon (amino acid)	(Low)[b]	(High)[b]	(Low)[b]	(High)[b]	(Low)[b]	(High)[b]	(GC high)[c]	(AT high)[c]
CCT (proline)	1126	619	1738	2642	7335	1650	3802	12908
CCC (proline)	811	125	1250	626	8922	**9297**	14360	4535
CCA (proline)	1129	926	1975	**6833**	12925	3016	3578	13197
CCG (proline)	1187	**4476**	1274	281	8289	4781	7549	1229
ACT (threonine)	1125	2039	2058	**6975**	9688	2446	1916	12568
ACC (threonine)	1409	**4507**	1584	5053	10567	**11929**	10958	5656
ACA (threonine)	1593	401	2732	2354	11978	1842	2509	13558
ACG (threonine)	1525	1162	1752	629	8208	4291	5347	1535
GTT (valine)	2090	**4125**	2110	**8662**	9759	3619	997	12282
GTC (valine)	1458	1694	1334	5559	6739	7911	7795	4595
GTA (valine)	1235	2100	2345	1187	6636	1289	789	8617
GTG (valine)	1737	3547	2233	1613	12057	**15033**	17975	9099
GCT (alanine)	1638	3471	1924	**11010**	10753	5972	4283	13839
GCC (alanine)	1927	2834	1638	5615	12847	**19828**	21774	6517
GCA (alanine)	2302	3364	2559	2644	11572	2509	3629	13160
GCG (alanine)	2092	**5086**	1527	625	7261	4331	8888	1297
GGT (glycine)	1864	**5922**	1813	**12841**	7670	6132	2560	9365
GGC (glycine)	1677	5468	1718	1873	9670	**14904**	18131	58830
GGA (glycine)	1553	317	2046	1005	11488	6844	2534	14804
GGG (glycine)	1337	659	1283	658	3817	871	8681	4717

[a] Sequence data are from genome resources at the National Center for Biotechnology Information. Only coding sequences longer than 100 codons are included. The total number of genes for each organism is 3,907 for *E. coli,* 6,115 for *S. cerevisiae,* 12,310 for *D. melanogaster* and 9,497 for *H. sapiens.*

[b] Cumulative codon usage of 10% of genes with the highest bias ("high") and of 10% of genes with the lowest bias ("low"). Codon usage bias is measured by CAI. The most frequently used synonymous codon in each fourfold degenerate amino acid for the "high" genes is shown in bold.

[c] Cumulative codon usage of 10% of genes with a high G +C% ("GC high") at the third-codon position and 10% of those with a high A +T% ("AT high") at the third-codon position.

Causes of Codon Usage Bias

Both selection and mutation bias can cause bias in synonymous codon usage. Codon usage bias of single cellular organisms (e.g. *E. coli* and *S. cerevisiae*) correlates significantly with the level of gene expression. For example, ribosomal protein genes are generally highly expressed, and their average codon usage bias is much higher than other genes (Figure 1, arrowheads). The most preferred synonymous codons in highly expressed genes correspond to the most abundant tRNAs (Ikemura, 1992). Codon usage bias is also inversely correlated with silent DNA divergence; that is, highly biased genes have fewer numbers of silent substitutions between species (Sharp and Li, 1987b). These data support the idea that selective constraints related to translational efficiency cause bias in synonymous codon usage. By using codons that correspond to the most abundant tRNAs and that have optimal interaction with anticodons, genes can be translated most efficiently. Such selective constraints are stronger in highly expressed genes. Translational selection seems to be responsible for codon usage bias also in some multicellular eukaryotes (e.g. *Drosophila* and *Caenorhabditis elegans;* Stenico *et al.,* 1994*;* Moriyama and Powell, 1997).

Some bacterial genomes have highly skewed base composition. For example, the bacterium *Mycoplasma capricolum* has a genomic G +C content of 25% and synonymous codon usage is very similar among the genes. In such bacterial genomes, mutation bias seems to have a predominant effect on codon usage bias. It should also be noted that, even in the genomes where translational selection has a principal role in determining codon usage bias, some genes under weaker selective constraints (e.g. genes with low expression) show the influence of mutation bias.

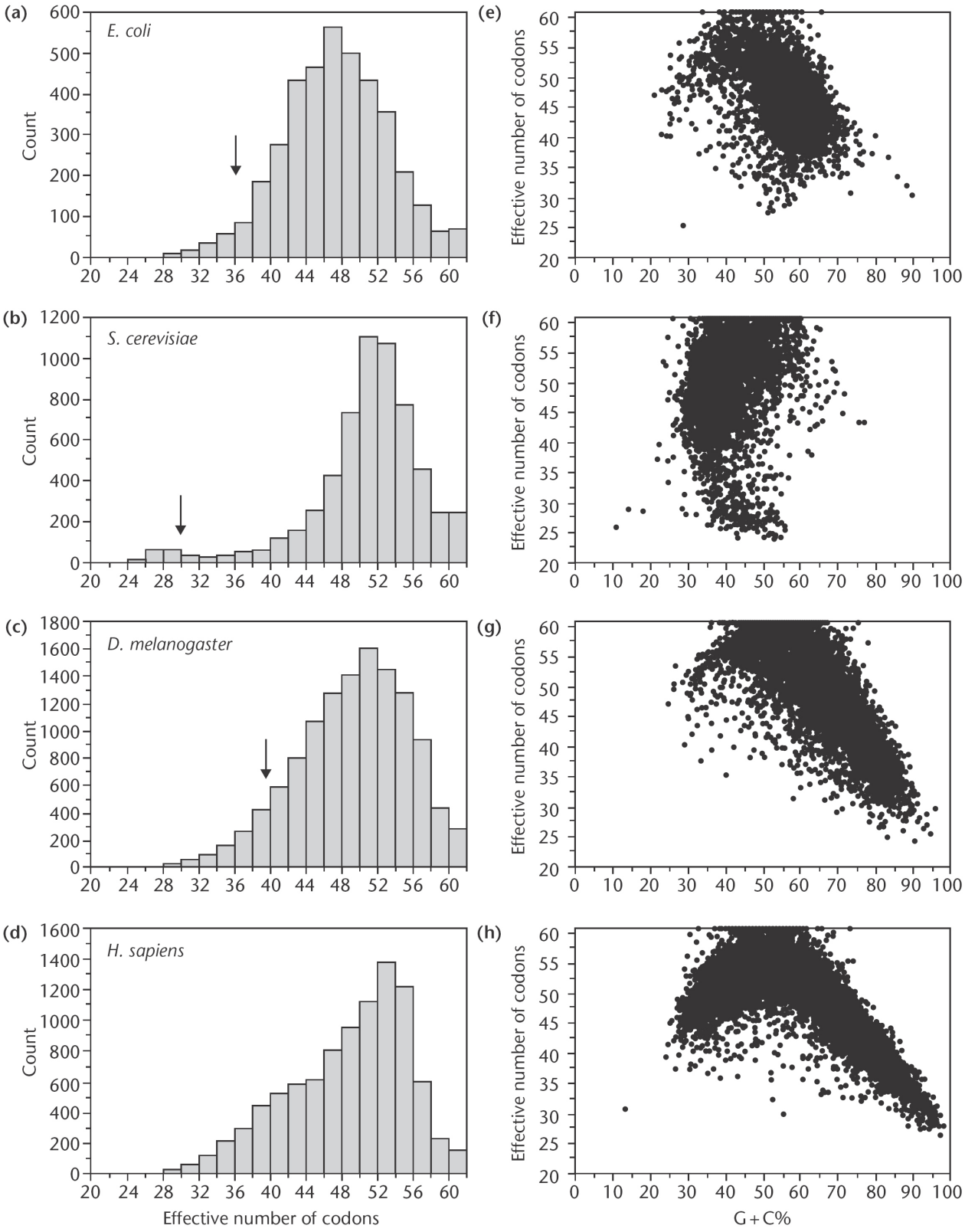

Figure 1 Codon usage bias in different organisms. (a, e) *Escherichia coli,* (b, f) *Saccharomyces cerevisiae,* (c, g) *Drosophila melanogaster* and (d, h) *Homo sapiens.* Codon usage bias is measured by the "effective number of codons" (N_c). "G +C%" is the G +C content at third-codon positions. The sources for the sequence data are given in Table 1. The average N_c for each organism is 47.3 for *E. coli,* 50.6 for *S. cerevisiae,* 49.2 for *D. melanogaster* and 48.7 for *H. sapiens.* Arrowheads indicate the average N_c for ribosomal protein genes (including predicted genes) with more than 100 codons. The number of ribosomal protein genes included (and the average N_c) for each organism is 39 (36.3) for *E. coli,* 102 (29.9) for *S. cerevisiae* and 94 (39.3) for *D. melanogaster.*

The "isochore" structure—that is, large-scale compositional heterogeneity—is found in mammalian and avian genomes. Not only silent sites in coding regions but also introns and flanking regions in the gene have a similar base (G + C%) composition (Aota and Ikemura, 1986). Therefore, the translational selective constraints described above cannot explain the compositional bias found in these warm-blooded vertebrates. Isochores are not only heterogeneous in base composition, they are also related to genome organization. Gene density and recombination rates are higher in GC-rich isochores. The distribution of repetitive elements differs between GC-rich and AT-rich isochores. Several hypotheses based on both mutation bias and selection have been put forth to explain the isochore structure. However, there has not been decisive explanation how mutation bias and/or selection can generate such compositional heterogeneity along the genome..)

References

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

Aota S-I and Ikemura T (1986) Diversity in G + C content at the third position of codons in vertebrate genes and its cause. *Nucleic Acids Research* 14: 6345–6355.

Comeron JM and Aguadé M (1998) An evaluation of measure of synonymous codon usage bias. *Journal of Molecular Evolution* 47: 268–274.

Ikemura T (1992) Correlation between codon usage and tRNA content in microorganisms. *In:* Hatfield DL, Lee BJ and Pirtle RM (eds) *Transfer RNA in Protein Synthesis,* pp. 87–111. Boca Raton, FL: CRC Press.

Moriyama EN and Powell JR (1998) Gene length and codon usage bias in *Drosophila melanogaster, Saccharomyces cerevisiae* and *Escherichia coli. Nucleic Acids Research* **26:** 3188–3193.

Moriyama EN and Powell JR (1997) Codon usage bias and tRNA abundance in *Drosophila. Journal of Molecular Evolution* 45: 514–523.

National Center for Biotechnology Information.
ftp://ncbi.nhn.nih.gov/refseq/
ftp://ncbi.nlm.nih.gov/genbank/genomes/D_melanogaster/Scaffolds/LARGE/

Sharp PM and Li W-H (1987a) The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Research* 15: 1281–1295.

Sharp PM and Li W-H (1987b) The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. *Molecular Biology and Evolution* 4: 222–230.

Shields DC, Sharp PM. Higgins DG and Wright F (1988) 'Silent' sites in *Drosophila* genes are not neutral: evidence of selection among synonymous codons. *Molecular Biology and Evolution* 5: 704–716.

Stenico M, Lloyd AT and Sharp PM (1994) Codon usage in *Caenorhabditis elegans:* delineation of translational selection and mutational biases. *Nucleic Acids Research* 22: 2437–2446.

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

Further reading

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

Akashi H (2001) Gene expression and molecular evolution. *Current Opinion in Genetics & Development* **11**: 660–666.

Eyre-Walker A (1999) Evidence of selection on silent site base composition in mammals: potential implications for the evolution of isochores and junk DNA. *Generics* 152: 675–683.

Powell JR and Moriyama EN (1997) Evolution of codon usage bias in *Drosophila. Proceedings of the National Academy of Sciences of the USA* **94**: 7784–7790.

Urrutia AO and 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.

Glossary

Synonymous codons. Codons that encode the same amino acid.

Codon usage bias. The nonrandom usage of synonymous codons.

Optimal codon. Synonymous codons that correspond to the most abundant tRNAs and that have optimal interaction with anticodons. Such codons are translated most efficiently.

Isochore. A long stretch of DNA that is homogeneous in base composition. Isochores are found in mammalian and avian (warm-blooded vertebrate) genomes. Both GC-rich and AT-rich isochores exist.