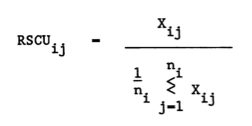
**Gribskov et al. (1984)**

In E. coli and S. cerevisiae a strong correlation between the frequency of a codon and the abundance of the corresponding tRNA was observed. In E. coli, genes of highly expressed proteins (e.g. ribosomal proteins) use codons corresponding to the most abundant tRNAs almost exclusively. This is thought to be due to a need for efficient translation. Proteins expressed at a low level use synonomous codons in rough proportion to the abundance of the corresponding tRNAs, resulting in a smaller, codon preference. In contrast, non-coding regions of E. coli DNA show no pronounced preference for any trinucleotide.

*(This difference of codon usage between genes and non-genes has been used to identify coding frames in DNA sequences)*

**Sharp & Li (1987)**

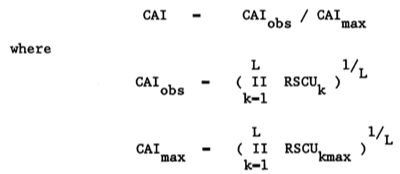
**CAI (Codon Adaptation Index)** = measure for assessing the degree of deviation from a postulated impartial pattern of usage. There is no intrinsic effect of gene length (L) on CAI, but values from short genes may be more variable due to sampling effects.

1. build a reference table of relative synonymous codon usage (RSCU) values from a set of reference genes (e.g. very highly expressed genes)

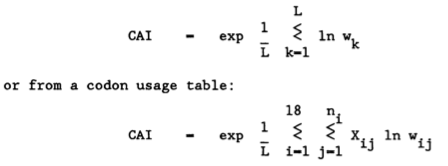
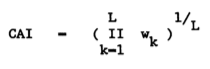
where Xij is the number of occurrences of the jth codon for the ith aminoacid, and ni is the number of alternative codons for the ith aminoacid.

Macintosh HD:Users:alecatz:Desktop:Screen Shot 2014-02-26 at 11.01.21.pngThe relative adaptiveness of a codon, Wij ,is then the frequency of use of that codon compared to the frequency of the optimal codon for that amino acid.

2. The Codon Adaptation Index (CAI) for a gene is then calculated as the geometric mean of the RSCU values corresponding to each of the codons used in that gene, divided by the maximum possible CAI for a gene of the same amino acid composition:



CAI can be also computed (saving computational time) as:

however due to real number underflow problems, a preferable version in the following:

*(they also used the CAI to predict how well suited a gene would be to the translational systems of a host using the RSCU values of the host to calculate CAI values for a heterologous gene).*

Problems with multicellular organsims (perhaps more recently clarified?):

codon usage not well understood, for ex. it appears that the mammalian genome comprises regions of quite different G+C content, and that local G+C content is an important influence on codon usage in any one gene. Also tRNA abundances are important selective constraints on codon usage, and in multicellular organisms tRNA populations vary among tissues. We also note that the only mammalian ribosomal protein genes for which DNA sequence data are available do not seem to show particularly high synonymous codon bias.

**Wright (1990)**

Synonimous Codon Usage (SCU) bias is species-specific, but there is also considerable variation among genes from a species. Two types of trend have been reported.

* The first type, observed in mammalian species (Ikemura, 1985), results from variation among genes in the G+C content at synonymous sites (i.e., **GC3s**, defined as the proportion of G+C content in the third codon position, excluding Met and Trp).
* A second type, observed in genes from unicellular species (e.g. E. coli and yeast) (Sharp and Li, 1987) and from D. melanogaster (Shields et al., 1988), exhibits a range from extreme SCU bias to minimal SCU bias. This type of SCU trend is not associated with GC3s content unless the 'preferred' codons themselves tend predominantly to contain (or not contain) G and/or C in the third position.

**Nc (Effective Number of Codons)** = measure that quantifies how far the codon usage of a gene departs from equal usage of synonymous codons. It can be easily calculated from codon usage data alone, and is independent of gene length and aminoacid composition. Nc is a good estimator and only underestimates SCU bias for gene lengths of less than 100 codons.

Nc statistics rationale:

An analogy can be drawn between the usage of synonymous codons for a particular aa and the frequencies of alleles at a locus. The Ne quantifies the number of alleles at a polymorphic locus by providing a figure for the number of equally frequent alleles that would produce the given level of homozygosity. The SCU bias of each aa can be described in this way. Summing the 'effective number of alleles' used by each of the 20 aa will then yield an Nc used in a gene. An extremely biased gene would use only 20 codons (i.e., one per aa), whereas an unbiased gene would tend to use all 61 codons equally (after correcting for aa usage).

aa can be subdivided into five SF types (1, 2, 3, 4, and 6) according to their respective number of synonymous codons.

Macintosh HD:Users:alecatz:Desktop:Screen Shot 2014-02-26 at 14.18.39.pngHomozygosity (F) can be calculated from the squared allele (codon) frequencies, where freq. pi = Ni/Naa (equal codon usage would be equivalent to minimum homozygosity):

Macintosh HD:Users:alecatz:Desktop:Screen Shot 2014-02-26 at 14.15.50.pngThe value of Ne is equivalent to the number of equally frequent alleles that would produce the particular level of homozygosity. Equal usage would yield (approximately) Ne= # codons for the aa, whereas usage of only one allele would give (approximately) Ne=1. (The Ne values will tend to the exact values as gene length increases.)

Macintosh HD:Users:alecatz:Desktop:Screen Shot 2014-02-26 at 14.23.12.pngNc is obtained by adding the contributions from each of the five SF types. These contributions consist of the number of members in the SF type divided by the average value of the aa in the SF type. Note that the contribution of the SF type with one codon (Met, Trp) is set equal to two.

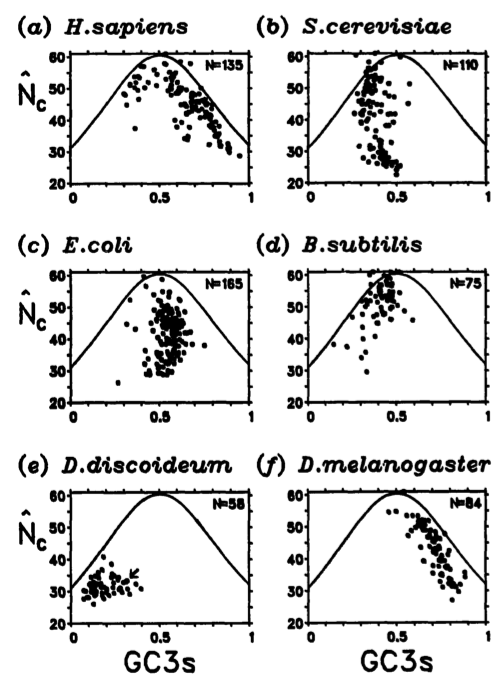
Adjustments need to be made if one or more aa are rarely used or absent. Rarely used aa are those aa for which either the numerator or denominator of F is 0; if this occurs they should be treated as absent. If Ile (SF type 3) is missing, F3 should be computed as the average of F2 and F4. If any of the other SF types are completely missing or 'rarely used' then the gene is probably too short (e.g., less than 61 codons) to accurately measure SCU bias, or exhibits extremely skewed aa usage. Nc could be greater than 61 (in case of more uniform usage than expected by chance), in which case it should be set to 61.

**Nc-plot: Nc vs G+C content at synonymous sites**.

The Nc-plot is particularly useful in the interpretation of SCU variation where the genome under study has a G+C content markedly different from 0.50.

Human genes have not been reported to show a relationship between SCU bias and level of gene expression. However, there is a very strong relationship between SCU bias and GC3s. The relationship between Nc and GC3s under Macintosh HD:Users:alecatz:Desktop:Screen Shot 2014-02-26 at 14.33.49.pngHo of no selection can be approximated by (where s = GC3s):

example of Nc-plot with Nc curve approximation:



**Stenico et al. (1994)**

**Frequency of Optimal Codons (Fop)** used in a gene = species-specific measure of bias towards those particular codons which appear to be translationally optimal in the particular species. The Fop is calculated as the number of occurrences of the optimal codons, divided by the total number of occurrences of the 18 informative aminoacids. Values can (in principle) range between 0 and 1; in C. elegans, where 21 optimal codons were found, the Fop value would be 21/59= 0.36 in a gene with uniform usage.

Correspondence Analysis (CA) of RSCU values was used to discriminate between high and low bias genes, then a chi-square test for codon occurrences in the 10% higher and lower bias genes was used to define the optimal codons, i.e. those used significantly more frequently in the high bias group (p<0.01). Correlation between the 1st axis of CA and Fop was higher than with Nc and GC3s, i.e. bias in optimal codon usage (as measured by Fop) is the most discriminative metrics between genes.

Comparison of the Fop and Nc values across genes indicate that the shortcomings of the latter measure pertain to the lowly biased genes: up to a point Nc values increase as Fop decreases, but then at very low values of Fop the Nc values begin to decrease again. These genes are under the weakest selection but are more biased (in the sense of deviation from uniform codon usage) than genes under some translational selection because mutational biases in the absence of selection do not lead to uniform codon usage.

**Chiapello et al. (1998)**

Factorial Correspondence Analysis (FCA) to determine relationship between codon usage and the physiological pattern of expression of Arabidopsis genes (N=815). The main trend seems dependant on the GC content:

* at one extreme lie genes with a highly G/C biased codon usage. This group contains mainly photosynthetic and housekeeping genes, which are known to encode the most abundant proteins of the vegetal cell.
* At the other extreme lie genes with a weaker A/T-biased codon usage, which exhibits most of the time a strong tissue-specific pattern of expression in relation, for example, to stress conditions.

In the absence of data concerning relative abundances of tRNAs, it is impossible to conclude that the 21 codons (G/C rich) preferentially used in the most abundant plant proteins are translationally optimal.

**Karlin et al. (1998)**

**Macintosh HD:Users:alecatz:Desktop:Screen Shot 2014-03-03 at 15.29.25.pngB(F|C) = codon usage differences (codon bias)** of one group of genes (or a single gene) relative to a second group of genes.

Let C be a family of genes with codon frequencies c(x, y, z) for codon xyz normalized such that for each amino acid codon family SUM(x,y,z) = a^c(x,y,z) = 1, where the sum extends over all codons (x, y, z) translated to amino acid a. Let F={f(x, y, z)} indicate the corresponding codon frequencies for the gene family F, again normalized to 1 in each amino acid codon family. We assess the codon usage difference of the gene family F relative to the gene family C by the formula above, where {pa(F)} is the set of amino acid frequencies of the genes of F.

max value of B(F|C) = 2 (but rarely above 0.5); differences between two gene families generally range from 0.05 to 0.30.

C is the standard to which different gene groups F (1), F (2), . . . ,F (r) are compared (for ex. a specific gene class, or an average gene meaning that C is the set of all genes of the genome). When C is the set of all genes, then B(F|C) = B(F|all) measures the codon usage difference of the class of genes F from the average gene and we refer to B(F|all ) as the **codon bias (CB)** of F.

(The authors also introduce a measure of **codon pair bias (dicodon bias)**, not reported here).

Explanations for codon bias have generally involved combinations of selection and mutational pressures; optimizations of translation rate and accuracy in relation to tRNA abundances are considered among primary selection forces. Some authors proposed that for the high expression level of a gene, high translation initiation rate is important, whereas high elongation rate is less important. An additional selective pressure (putatively active on mRNA) consists of preventing formation of secondary structure, which competes with constraints optimizing elongation speed.

About GC bias: Introns are relatively AT rich and genes in AT-rich regions putatively disengage the template strand more easily fomenting more rapid transcription. On the other hand, codons of more S3 types (high site 3 G+C frequencies) conceivably are more stable during transcription and translation. On this basis smaller genes appear to congregate in higher G+C isochores ensuring greater translational fidelity (for ex. in humans).

About rare codons: use of repetitive rare codons might reduce translation rate and induce translation pauses, allowing protein domains and suitable secondary structures to fold into native structural conformation. On this point there are differences in prokaryotic and eukaryotic translation mechanisms, e.g. the important role of chaperonins in prokaryotes but not in eukaryotes and the potentially important activity of co-translational folding in eukaryotes but not in prokaryotes.

Special facts about E.coli: tRNA synthetases show a different codon bias from other highly expressed genes. Perhaps there is greater need for these enzymes to fold properly and therefore a greater codon pair bias for directing the kinetics of translation (e.g. pause sites) to facilitate correct folding. Also, in general middle and final third (3’ end) of genes have the same levels of codon bias (quite similar codon frequencies), whereas the 5’ possess significantly different codon frequencies.

**Duret & Mouchiroud (1999)**

The authors found a correlation between codon usage and expression levels of Arabidopsis genes (N=2,917), showing that this is not due to a mutational bias. Also, a strong negative correlation between codon usage and protein length was found. This effect is not due to a smaller size of highly expressed proteins. Thus, for a same-expression pattern, the selective pressure on codon usage appears to be lower in genes encoding long rather than short proteins.

Powell and Moriyama hypothesized that this length effect could be explained by selection for translation rate. Assume for example that a non-optimal codon requires twice as long to incorporate an amino acid as does the optimal codon. Mutations will have a greater relative effect in smaller genes than in longer ones: in a short gene with 100 codons, such mutation would increase translation time by 1%, whereas the same mutation in a gene with 1,000 codons would increase translation time by only 0.1%. Thus, such mutations are more likely to be counter selected in short genes than in long ones.

However, several arguments suggest that this model is not realistic. It seems that the initiation rate is the limiting step in protein translation, not the elongation rate. The effect of codon usage on protein synthesis is then thought to be indirect: the use of optimal codons increase the elongation rate and thus reduces the time the ribosome is bound to the mRNA; this leads to an increase in the pool of free ribosomes and hence to an increase in the translation initiation rate of all mRNA species. Therefore the use of optimal codons in a given gene should increase the production of all proteins in a cell, not only its own protein product. This way the selection on codon usage should be independent of protein length. Selection to minimize the cost of proofreading is also expected to be independent of protein length.

Selection on codon usage has been clearly demonstrated in several unicellular organisms:

* 1. codon usage is biased toward ‘‘preferred’’ codons that generally correspond to the most abundant tRNA species.
* 2. there is a positive correlation between codon-usage bias and the level of gene expression.
* 3. the rate of synonymous substitution between species is inversely correlated with codon usage bias, implying greater purifying selection on silent changes in highly biased genes.

The selective differences between alternative synonymous codons are probably very small. Thus, in genes with low expression levels, or in species with small population sizes, selection is not sufficient to overcome genetic drift, and codon usage is essentially shaped by mutation patterns.

In multicellular eukaryotes, gene expression and tRNA abundance can be tissue and developmental stage-specific and are difficult to quantify. However, the action of natural selection on codon usage has been established in Drosophila melanogaster: the limited data available show a relationship between codon preference and tRNA abundance (9, 10); negative correlations between codon usage bias and silent substitution rate have been also observed (9, 11, 12); finally, anecdotal (??) evidence suggests a relationship between codon-usage bias and gene expression level: genes known to be expressed at a high level, such as those encoding ribosomal proteins or glycolytic enzymes, show a greater-than-average codon bias. Optimal codons probably confer fitness benefits by enhancing translation efficiency. However, it is not yet clear whether codon usage affects primarily the elongation rate, the cost of proofreading, or the accuracy of translation.

**Jansen et al. (2003)**

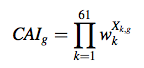
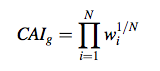
The authors verified if CAI (Sharp&Li 1987) and codon usage (Karlin 1998) performances in predicting gene expression levels could be improved using updated genome-wide data, however it appears that the original small reference gene set were already enough for reliable estimates.

**Alternative formalism for CAI and link with codon usage:**

**Macintosh HD:Users:alecatz:Desktop:Screen Shot 2014-03-04 at 15.22.15.png**The “relative adaptiveness” for a set of highly expressed genes G, where faa,i is the frequency of codon i (which encodes amino acid aa), and faa,max the frequency of the codon most often used for encoding amino acid aa in a set of highly expressed genes G.

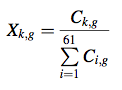
Values range between 0-1, with 0 indicating that a codon is not present in G, and 1 that occurs most often in G for a given amino acid.

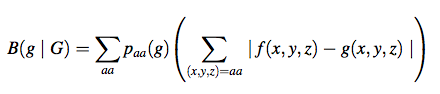
The CAI of a gene g is the geometric average of the relative adaptiveness of all codons in a gene sequence, which can be written in both ways:



* wi is the relative adaptiveness of the ith codon in a gene with N codons
* wk is the relative adaptiveness of the kth out of the 61 codons in the genetic code (excluding stop codons)

Xk,g is the fraction of codon k among the total number of codons in gene g, where Ck,g is the number of times codon k appears in the gene:



The “codon bias” of a gene g relative to a gene set G:

Macintosh HD:Users:alecatz:Desktop:Screen Shot 2014-03-04 at 15.32.10.pngwhere paa(f) is the fraction of amino acid aa in gene g; f(x, y, z) the frequency of a codon triplet (x, y, z) in gene g normalized such that f(x, y, z) = 1 if (x, y, z) is the most common synonymous codon; g(x, y, z) is the corresponding normalized codon frequency in gene set G. Equivalent to:

**Dos Reis et al. (2004)**

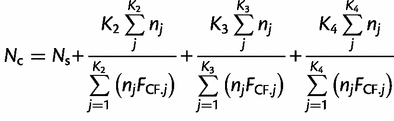
* previous studies on unicellular organisms (yeast, E.coli) established that highly expressed genes use a subset of optimal codons in accordance with their respective major isoacceptor tRNA levels. The evolutionary force responsible for this was coined “translational selection”.
* however, when attempts were made to extend the findings to higher eukaryotes, the situation became confusing. Some organisms (ex. human) seem to have their codon usage determined solely by genomic GC content or isochore composition, while others (D. melanogaster, C. elegans) seem to present an intermediate degree of selection partly determining their codon usage. A mutation– selection balance theory of synonymous codon usage was developed to explain these observations.

Here the authors developed an algorithm showing that tRNA gene redundancy and genome size are interacting forces that ultimately determine the action of translational selection, and that an optimal genome size exists for which this kind of selection is maximal. Accordingly, genome size also presents upper and lower boundaries beyond which selection on codon usage is not possible. Coevolution of genome size and tRNA genes explains the observed patterns in translational selection in all living organisms.

(is this still the accepted theory on the subject?)

**Sun et al. (2012)**

Improved Nc:



Macintosh HD:Users:alecatz:Downloads:graphic-3.gif