$m\widehat{N}_c$

The standard version of the genetic code table defines a translational mapping from 61 sense codons to 20 amino acids. In this table, except Met and Trp, the other eighteen amino acids are encoded by two to six different codons. Different codons that encode the same amino acid are known as synonymous codons. Even though synonymous codons encode the same amino acids, it has been found that these codons are not used uniformly, a phenomenon observed in all genomes is known as codon usage bias (CUB). Though there are different measures have been proposed to measure CUB in a gene, the mathematical formula given by Wright, known as Effective number of codons ($\widehat{N}_c$, Wright 1900) is one of the most widely used, whose implementation is available in the codonW (Peden 1999) software.

The mathematical formula in $\widehat{N}_c$ (Wright 1990) is based on the principle of population genetics (Kimura and Crow1964). In spite of its wide acceptance, there are few pitfalls in the $\widehat{N}_c$ value calculation in case of short coding sequences having low abundance for some of the amino acids (Fuglsang 2003, 2004, 2005; Banerjee *et al*. 2005, Sun *et al.* 2012).In addition to this, the error in calculation is not same for all amino acids and is proportional to the codon degeneracy encoding the concerned amino acid. The flaw in the Wright's formula for $\widehat{N}_c$ can be explained with help of codon usage in a hypothetical coding sequence as follows. In the hypothetical coding sequence let us assume that, out of the 20 amino acids codon usage is uniform in all the five four-fold degenerate amino acids and abundance value of each of the codons is two. However, the codon usage in the remaining amino acids is highly biased and for each amino acid, codon abundance for one of the codon is 50 and the same is zero for the remaining synonymous codons. In this gene, contribution to the overall CUB from the four-fold degenerate amino acids is 20.0 and that from the remaining amino acids is 15.0. Therefore, the overall CUB for the gene is expected to be 35.0. However, when we calculate CUB of the gene using mathematical formula given by Wright (1990), the result we obtain is 50.0. As the value is below 61.0, 50.0 is accepted as the correct $\widehat{N}_c$ value for the coding sequence, which is not correct.

We have presented a modified version of $\widehat{N}_c$ ($m\widehat{N}_c$) considering suggestions from Fuglsang (2003, 2004, and 2005; Banerjee *et al.* 2005; Sun *et al*. 2012). In comparison to in $\widehat{N}_c$(Wright 1990), $m\widehat{N}_c$ correlates better with gene expression in several organisms.

The basic principle used in the mathematical models of $\widehat{N}_c$ is: first calculate "effective number of codons for individual amino acids" (Equation 1 and 2 shown below) and then combine theses values for all the 20 amino acids to obtain the "effective number of codons for the gene" (Equation 3).

For an amino acid *AA* with degeneracy *k*, i.e. with *k* number of synonymous codons, each with counts n_1, n_2,…, n_k, $n = \sum_{i=1}^{k} n_i$ and $p_i = n_i / n$, effective number of codons $\widehat{N}_{c_{AA}}$is calculated as follows:

$$m\widehat{N}_{c_{AA}} = \frac{1}{F_{AA}} \qquad \text{(Equation 1)}$$

$$\text{Where } F_{AA} = \sum_{i=1}^{k} p_i^2 \qquad \text{(Equation 2)}$$

Finally for standard genetic code the formula of $\widehat{N}_c$ for a gene can be given as:

$$m\widehat{N}_c = 2 + \frac{9}{\bar{F}_2} + \frac{1}{F_3} + \frac{5}{\bar{F}_4} + \frac{3}{\bar{F}_6} \qquad \text{(Equation 3)}$$

Here $\bar{F}_i$ represents **weighted** average values of $F_{AA}$for all the amino acids with degeneracy *i*.

Special adjustments:

(i) Ile codons are missing or rare:F_3 should be computed as the average of $\bar{F}_2$ and $\bar{F}_4$.
(ii) Further, if all the amino acids with degeneracy 2, 4 or 6 are completely missing or rare then probably the gene is too sort or exhibits extremely skewed amino acid usage and therefore do not compute $m\widehat{N}_c$.

Note 1: Usually larger size genes with sufficient codons for all the amino acids are preferred in codon usage analysis. We therefore put a caution in interpreting $m\widehat{N}_c$ values for smaller genes. We therefore suggest interpreting $m\widehat{N}_c$ values carefully for smaller genes.

Referrences:

Banerjee T, Gupta SK, Ghosh TC (2005) Towards a resolution on the inherent methodological weakness of the ‘‘effective number of codons used by a gene’’. *Biochem Biophys Res Commun* 330(4):1015–1018.

Fuglsang A (2003) The effective number of codons for individual amino acids: some codons are more optimal than others.Gene 320**:**185–190.

Fuglsang A (2004) The ‘effective number of codons’ revisited. *Biochem Biophys Res Commun* 317(3):957–964.

Fuglsang A (2005) On the methodological weakness of ‘the effective number of codons’: a reply to Marashi and Najafabadi.Biochem Biophys Res Commun 327(1):1–3.

Peden JF (1999) CodonW, PhD Thesis, University of Nottingham.

Sun X, Yang Q, Xia X (2012) An improved implementation of Effective Number of Codons (Nc). MolBiolEvol 30:191–196.

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