m$\widehat{N}_c'$

Novembre (2002) introduced a modified version of $\widehat{N}_c$ called ENCPrime (or $\widehat{N}_c'$) that gives CUB in a gene after filtering out expected CUB due to background nucleotide composition. As $\widehat{N}_c'$(Novembre 2002) is a modified version of $\widehat{N}_c$ (Wright 1990), an error in $\widehat{N}_c$ will obviously result error in $\widehat{N}_c'$ value. Keeping modifications by Fuglsang (2003, 2004, and 2005), Banerjee *et al.*(2005) and Sun *et al*. (2012) in the formula for $\widehat{N}_c$ in view, the more accurate formula for $\widehat{N}_c'$, designated as $m\widehat{N}_c'$(*m* stands for modified) can be written as follows:

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 $m\widehat{N}'_{c_{AA}}$ is calculated as follows:

$$m\widehat{N}'_{c_{AA}} = \frac{1}{F'_{AA}} \qquad \text{(Equation 9)}$$

Where $F'_{AA} = \frac{X^2 + 1}{k}$ (Equation 10)

and$X^2 = \sum_{i=1}^{k} \frac{(p_i - e_i)^2}{e_i}$ (Equation 11)

Here e_i is the expected usage of a codon calculated from the nucleotide composition.

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

$$m\widehat{N}_c' = 2 + \frac{9}{\bar{F}_2'} + \frac{1}{\bar{F}_3'} + \frac{5}{\bar{F}_4'} + \frac{3}{\bar{F}_6'} \qquad \text{(Equation 12)}$$

Here $\bar{F}_i'$ represents weighted average values of F_{AA}' for all the amino acids with degeneracy *i*. Instead of simple average, when weighted average is considered, contribution of each of the amino acid codons toward the final $m\widehat{N}_c'$ value will be proportionate to their codon abundance value. This will minimize the potential bias introduced by codon families with small *n* values (Sun *et al.* 2012). Further, it can be shown that, when the expected usage according to background nucleotide composition for a set of synonymous codon is uniform i.e. frequency of each of the synonymous codon is *1/k*, then $m\widehat{N}_c'$ reduced to $m\widehat{N}_c$.

Special adjustments:

(i) If equation 11 is undefined (e_i = 0), ignore the X^2 term while calculating F_{AA}' (Wright 1990).

(ii) If Ile codons are missing, F_3' should be computed as the average of$\bar{F}_2'$and $\bar{F}_4'$ (Wright 1990).

(iii) 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 $m\widehat{N}_c'$ (Wright 1990).

Note: Usually larger size genes with sufficient codons for all the amino acids are preferred in codon usage analysis. 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.

Novembre JA (2002) Accounting for background nucleotide composition when measuring codon usage bias. MolBiolEvol 19:1390–1394.

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.