Improved Nc ($m\widehat{N}_c$) and ENCprime ($m\widehat{N}_c'$) measures

Citation: *S.S. Satapathy, A.K. Sahoo, S.K. Ray and T.C. Ghosh, "Codon degeneracy and amino acid abundance influence the measures of codon usage bias: improved Nc and ENCprime measures", Genes to cells (accepted 2017)*

The non-uniform usage of synonymous codons, a phenomenon known as codon usage bias (CUB) is common in all genomes. The mathematical model developed by Wright (1990) known as Effective number of codons ($\widehat{N}_c$) (Wright 1990) is used most widely by researcher in this field. The mathematical formula in $\widehat{N}_c$ (Wright 1990) is based on the principle of population genetics (Kimura & Crow 1964). The theoretical value of $\widehat{N}_c$ ranges from 61 (when all the synonymous codons are used uniformly) to 20 (when synonymous codon usages is maximally biased). The implementation of $\widehat{N}_c$ (Wright 1990) is available in the software CodonW (Peden 1999).

The basic principle used in the mathematical model of $\widehat{N}_c$ (Wright 1990) is: first to 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 $n_1, n_2, \ldots, n_k$, $n = \sum_{i=1}^{k} n_i$ and $p_i = n_i / n$, effective number of codons $\widehat{N}_{c_{AA}}$is calculated as follows:

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

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

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

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

Here $\bar{F}_i$ (*i*=2, 3, 4 and 6) represents average values of $F_{AA}$for all the amino acids with degeneracy *i*.

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. In table 1(a), the variation in $\widehat{N}_c$ value is explained with help of a 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 while the codon usage in the remaining amino acids is strongly biased. In this case the calculated $\widehat{N}_c$ value is 50, which is in the acceptable range of $\widehat{N}_c$ values. However, if we calculate effective number of codons manually for each amino acid and sum up the values, then the correct $\widehat{N}_c$ value will be 35.0, not 50.0. Therefore, the formula used in calculating $\widehat{N}_c$ value gives erroneous result when the usage of synonymous codons is uniform for low abundant amino acids.

Though the above error was pointed out earlier (Banerjee *et al*. 2005; Fuglsang 2003, 2004, 205; Sun at al. 2012), we would like to add further to this pitfall that the error magnitude is different from amino acid to another amino acid depending upon the degeneracy of the codons the amino acids are encoded for. To explain this, an example has been given in Table 1(b). By keeping the abundance value to two for each codon, we observed that the $\widehat{N}_c$ value for an amino acid calculated using the formula (Wright 1990) is not same as that calculated manually. For example, in case of Phe which is encoded by two fold degenerate codon the value is 3.0 instead of 2.0, similarly in case of Leu which is encoded by six fold degenerate codon the value is 11.0 instead of 6.0. So the error incurred into the $\widehat{N}_c$ value is not same for all amino acids and varies with the codon degeneracy of the concerned amino acid (Footnote, Table 1(b)).

ENCprime or $\widehat{N}_c'$ (Novembre 2002) is a modified version of $\widehat{N}_c$ (Wright 1990). $\widehat{N}_c'$ is based on the Pearson's X^2 statistics and describes the departure of the observed codon usage from some expected distribution. The method is basically a variant of the measure $\widehat{N}_c$ and it measures codon usage bias in a gene after filtering out expected codon usage calculated from the background nucleotide composition. $\widehat{N}_c'$ has been used extensively to study selection on codon usage bias in organisms. An incorrect $\widehat{N}_c$ automatically makes $\widehat{N}_c'$ also incorrect.

Keeping the above constraints, we are proposing improvement in the existing methods to calculate correctly the $\widehat{N}_c$ and $\widehat{N}_c'$ values. The implementations of the modified $\widehat{N}_c$ and $\widehat{N}_c'$ are represented as $m\widehat{N}_c$ and $m\widehat{N}_c'$ respectively.

Improvement in the mathematical formula used in $\widehat{N}_c$

The major source of error in the value of $\widehat{N}_c$ is due to the problem in the formula (Equation 2). In case of an amino acid with degeneracy *k*, if the codon usage

is highly uniform and synonymous codon abundances are comparatively less (as is the case in the above example), then the contribution of this amino acid to the overall CUB often exceeds *k*. In the mathematical model of $\widehat{N}_c$, Wright had suggested an *ad hoc* solution into it (Wright 1990). According to his observation, $\widehat{N}_c$ value can be greater than 61.0 if the observed codon usage pattern is more uniform than expected by chance. In this case the $\widehat{N}_c$ value is suggested to be revised to 61.0. However, this *ad hoc* solution can't be applied in the example given earlier, as the $\widehat{N}_c$ value is less than 61.0. Detailed analysis of the mathematical model by Fuglsang (2003, 2004, and 2005) and Banerjee *et al*. (2005) suggested replacing equation 2 by the equation 4 given below. With this replacement, the above problem can be mitigated.

Where $F_{AA} = \sum_{i=1}^{k} p_i^2$ (Equation 4)

Another important consideration in $\widehat{N}_c$ is the amino acid composition of the gene. Equation 3 used in calculating $\widehat{N}_c$ generates values within range of 20.0 to 61.0 assuming codon usages of the non-existing amino acids are similar to the existing amino acids. However it is now widely accepted that the codon usage bias widely differs among amino acids (Satapathy *et al.* 2016) and therefore it would be appropriate to calculate contributions towards $\widehat{N}_c$ value for different amino acids separately as given in equation 5. With this modification, $\widehat{N}_c$ value may fall below 20.0 when some of the amino acids are absent in a coding sequence.

$$\widehat{N}_c = \sum_{for\ all\ F_{AA} \neq 0} \frac{1}{F_{AA}} \qquad \text{(Equation 5)}$$

Implementation of $\widehat{N}_c$ (CodonW) and the modified implementation named as ($m\widehat{N}_c$, *m* stands for modified) are summarized in Supplementary material S1.

Improvement in the mathematical formula used in $\widehat{N}_c'$

Before presenting modifications in the $\widehat{N}_c'$, here we summarise the mathematical model for $\widehat{N}_c'$ (Novembre 2002) as follows:

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

$$\widehat{N}_{c_{AA}}' = \frac{1}{F'_{AA}} \qquad \text{(Equation 6)}$$

Where $F'_{AA} = \frac{X^2+n-k}{k(n-1)}$ (Equation 7)

and $X^2 = \sum_{i=1}^{k} \frac{n(p_i-e_i)^2}{e_i}$ (Equation 8)

Here e_i is the expected usage of a codon calculated from the nucleotide composition. Finally, for standard genetic code the formula of $\widehat{N}'_c$ for a gene can be given as:

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

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

Keeping modifications by Banerjee *et al*. (2005), Fuglsang (2003, 2004 and 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:

$$m\widehat{N}'_{c_{AA}} = \frac{1}{F'_{AA}} \quad \text{(Equation 10)}$$

Where $F'_{AA} = \frac{X^2+1}{k}$ (Equation 11)

And $X^2 = \sum_{i=1}^{k} \frac{(p_i-e_i)^2}{e_i}$ (Equation 12)

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:

$$m\widehat{N}'_c = \sum_{for\ all\ F'_{AA} \neq 0} \frac{1}{F'_{AA}} \quad \text{(Equation 13)}$$

Further, it can be shown that, when the expected usage according to background nucleotide composition for a set of synonymous codons is uniform i.e. frequency of each synonymous codon is *1*/*k*, then $m\widehat{N}'_c$ reduced to $m\widehat{N}_c$. Similar to

the case of $\widehat{N}_c$, $m\widehat{N}_c'$ value may fall below 20.0 when some of the amino acids are absent in a coding sequence.

Nucleotide composition of the gene itself may be used in calculation of expected codon usage while determining $m\widehat{N}_c'$

One of the important points in the calculation of $\widehat{N}_c'$ was consideration of back ground nucleotide composition. The expected codon usage may be calculated from the nucleotide composition of any of the four possibilities such as (i) coding sequence of the gene under consideration, (ii) whole genome sequence, (iii) all the coding sequences in the genome of the organism, and (iv) the inter-genic regions. To compare among the four possibilities we calculated $m\widehat{N}_c'$ taking into account all the above possibility for *Escherichia coli* genes. Our correlation analysis with gene expression data in *E. coli* suggested that the position specific nucleotide composition within the gene itself may be the best sequence to use in calculation of expected codon usage while determining $m\widehat{N}_c'$. Another advantage of this is that no additional reference sequence is required while calculating $m\widehat{N}_c'$.

References

Banerjee, T., Gupta, S.K. & Ghosh, T.C. (2005) Towards a resolution on the inherent methodological weakness of the ‘‘effective number of codons used by a gene’’. *Biochem. Biophys. Res. Commun.* **330**, 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**, 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**, 1–3.

Kimura, M. & Crow, J.F. (1964) The number of alleles that can be maintained in a finite population. *Genetics* **49**, 725–738.

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

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

Sun, X., Yang, Q. & Xia, X. (2012) An improved implementation of Effective Number of Codons (Nc). *Mol. Biol. Evol.* **30**, 191–196.

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

Table 1(a): $\widehat{N}_c$ values calculated for individual amino acids for a hypothetical gene based on Wright (1990) formula

Amino Acid	Codon	Count	Nc_{AA}	Amino Acid	Codon	Count	Nc_{AA}	Amino Acid	Codon	Count	Nc_{AA}	Amino Acid	Codon	Count	Nc_{AA}
Phe	UUU	50	**1.0**	**Ser**	UCU	50	**1.0**	**Tyr**	UAU	50	**1.0**	**Cys**	UGU	50	**1.0**
	UUC	0			UCC	0			UAC	0			UGC	0	
Leu	UUA	50	**1.0**		UCA	0		TER	UAA	-		TER	UGA	-	
	UUG	0			UCG	0			UAG	-		Trp	UGG	50	**1.0**
	CUU	0		**Pro**	CCU	2	**7.0**	**His**	CAU	50	**1.0**	**Arg**	CGU	50	**1.0**
	CUC	0			CCC	2			CAC	0			CGC	0	
	CUA	0			CCA	2		**Gln**	CAA	50	**1.0**		CGA	0	
	CUG	0			CCG	2			CAG	0			CGG	0	
Ile	AUU	50	**1.0**	**Thr**	ACU	2	**7.0**	**Asn**	AAU	50	**1.0**	**Ser**	AGU	0	
	AUC	0			ACC	2			AAC	0			AGC	0	
	AUA	0			ACA	2		**Lys**	AAA	50	**1.0**	**Arg**	AGA	0	
Met	AUG	50	**1.0**		ACG	2			AAG	0			AGG	0	
Val	GUU	2	**7.0**	**Ala**	GCU	2	**7.0**	**Asp**	GAU	50	**1.0**	**Gly**	GGU	2	**7.0**
	GUC	2			GCC	2			GAC	0			GGC	2	
	GUA	2			GCA	2		**Glu**	GAA	50	**1.0**		GGA	2	
	GUG	2			GCG	2			GAG	0			GGG	2	

Note: The table presents codon usage and $\widehat{N}_c$ values calculated for individual amino acids for a hypothetical gene. In this gene 4-fold degenerate amino acids are of low abundance but with highly uniform codon usage. Remaining amino acids are highly abundant and with highly biased codon usage. For this codon usage pattern, actual $\widehat{N}_c$ value calculated is 50.0, whereas the expected correct value is 35.0.

Table 1(b): $\widehat{N}_c$ values calculated for individual amino acids for a hypothetical gene based on Wright (1990) formula

Amino Acid	Codon	Count	Nc_{AA}	Amino Acid	Codon	Count	Nc_{AA}	Amino Acid	Codon	Count	Nc_{AA}	Amino Acid	Codon	Count	Nc_{AA}
Phe	UUU	2	**3.0**	**Ser**	UCU	2	**11.0**	**Tyr**	UAU	2	**3.0**	**Cys**	UGU	2	**3.0**
	UUC	2			UCC	2			UAC	2			UGC	2	
Leu	UUA	2	**11.0**		UCA	2		TER	UAA	-		TER	UGA	-	
	UUG	2			UCG	2			UAG	-		Trp	UGG	2	**1.0**
	CUU	2		**Pro**	CCU	2	**7.0**	**His**	CAU	2	**3.0**	**Arg**	CGU	2	**11.0**
	CUC	2			CCC	2			CAC	2			CGC	2	
	CUA	2			CCA	2		**Gln**	CAA	2	**3.0**		CGA	2	
	CUG	2			CCG	2			CAG	2			CGG	2	
Ile	AUU	2	**5.0**	**Thr**	ACU	2	**7.0**	**Asn**	AAU	2	**3.0**	**Ser**	AGU	2	
	AUC	2			ACC	2			AAC	2			AGC	2	
	AUA	2			ACA	2		**Lys**	AAA	2	**3.0**	**Arg**	AGA	2	
Met	AUG	2	**1.0**		ACG	2			AAG	2			AGG	2	
Val	GUU	2	**7.0**	**Ala**	GCU	2	**7.0**	**Asp**	GAU	2	**3.0**	**Gly**	GGU	2	**7.0**
	GUC	2			GCC	2			GAC	2			GGC	2	
	GUA	2			GCA	2		**Glu**	GAA	2	**3.0**		GGA	2	
	GUG	2			GCG	2			GAG	2			GGG	2	

Note: The table presents codon usage and $\widehat{N}_c$ values calculated for individual amino acids for a hypothetical gene. In this gene amino acids are of low abundance but with highly uniform codon usage. Though codon usages for different amino acids are same, calculated $\widehat{N}_c$ values are different. For example, $\widehat{N}_c$ for 2-fold, 3-fold, 4-fold and 6-fold amino acids are 3.0, 5.0, 7.0 and 11.0 respectively whereas the expected values are 2.0, 3.0, 4.0 and 6.0 respectively.