Um(g) (Satapathy et al 2014)

In the genetic code table synonymous codons are arranged systematically as these are usually different from each other at the 3rd position. A set of four codons with degeneracy only at the 3rd position is grouped as a family box in this table. There are eight such family boxes in genetic code table: one for each amino acid having four or more synonymous codons. Third position in these codons is known as four-fold degenerate site (FDS). The FDS in coding sequences is important for studying the effect of any selection pressure on codon usage bias because nucleotide substitution *per se* is not under any such pressure at the site due to the unaltered amino acid sequence in a protein. Um(g), the unevenness measure of a gene *g*. is an estimation of the variation of nucleotide frequency at the FDS across the eight family boxes Um(g) is a good indicator of selection pressure on codon usage bias in genomes with higher G+C%.

Method for Um(g)

The consequence of CUB on the variation of nucleotide frequency at the FDS is explained easily with the two hypothetical examples given in Table 1. In case I, the frequency of a nucleotide across the column does not vary since the pattern of CUB is same in the different families. In case II, the variation in the frequency of a nucleotide across a column is more due to the difference in the patterns of CUB in the different families.

Table 1. Hypothetical examples showing effects of codon usage bias on non-uniform frequency of a nucleotide at four-fold degenerate site

Amino Acid	Case I				Case II			
	U	C	A	G	U	C	A	G
Ala_{GCN}	0.05	0.45	0.05	0.45	0.00	0.90	0.00	0.10
Arg_{CGN}	0.05	0.45	0.05	0.45	0.00	0.50	0.40	0.10
Gly_{GGN}	0.05	0.45	0.05	0.45	0.50	0.10	0.00	0.40
Leu_{CUN}	0.05	0.45	0.05	0.45	0.00	0.10	0.00	0.90
Pro_{CCN}	0.05	0.45	0.05	0.45	0.00	0.50	0.40	0.10
Ser_{UCN}	0.05	0.45	0.05	0.45	0.40	0.40	0.10	0.10
Thr_{ACN}	0.05	0.45	0.05	0.45	0.00	0.10	0.00	0.90
Val_{GUN}	0.05	0.45	0.05	0.45	0.40	0.10	0.00	0.50

Case I: CUB is there but frequency of a nucleotide is invariable along the columns.

Case II: CUB is there but it is increasing variation of a nucleotide frequency along the columns.

The variation of a nucleotide frequency at FDS across the eight FBs in a gene *g* is calculated using the equation:

$$Um(g) = \left[\frac{1}{4} \sum_{z \in \{A,C,G,T\}} \left\{ \frac{1}{n} \sum_{xy \in \{AC,CC,CG,CT,GC,GG,GT,TC\}} | SCF^{F}_{xyz} - M_z | \right\} \right] \Big/ k$$

where Um(g) stands for the Unevenness measure of a gene *g*, which is defined as the variation in the frequencies of nucleotide at FDS. SCF^{F}_{xyz} stands for synonymous codon frequency (SCF) of a codon 'xyz' within a family box (F). There are 32 SCF values in 8 FBs. For example, the calculation of SCF for the codon ACA is given as follows:

$$SCF^{F}_{ACA} = \frac{X_{ACA}}{\sum_{N \in \{A,C,G,T\}} X_{ACN}}$$

M_z is the arithmetic mean of SCF^{F}_{xyz} values among all FB codons with nucleotide '*z*' at the third position. The absolute difference of the SCF values from the mean has been used to find the frequency variation. We preferred using the absolute difference rather than the square of the difference. Because we have less number of SCF values and a single large squared difference value might abruptly affect the Um(g) value. This also made the calculation simple.

The divisor '*n*' represents the number of FBs considered in Um(g). The maximum value of '*n*' is 8. Family boxes with total number of codons less than four were not considered for calculating the Um(g).The value of '*n*' therefore, varies in the different genes depending on the available FB codons. However in larger genes, this limitation can be avoided.

Um(g) is the arithmetic mean of the four average deviations divided by a constant '*k*'. The value of '*k'* is the theoretically determined maximum Um(g) value. The Um(g) value is maximum when only one codon out of the four FB codons is used. The SCF value is 1.0 for the selected codon and for the other three codons, the value is 0.0. The selection pressure on codons in different family boxes is different, so that maximum variation in the frequency at FDS is generated similar to the case II in Table I. The value of '*k*' is 0.375 when the value of '*n*' is 8. The '*k*' values have been calculated as 0.367, 0.361, 0.360 and 0.375 when the '*n*' values are 7, 6, 5 and 4 respectively. In order to fit the Um(g) value in the range [0.0, 1.0], it is scaled by a factor '*k*'. The Um(g) value 1.0 implies maximum frequency variation whereas 0.0 implies no variation.

References

Satapathy S.S., Powdel B.R., Dutta M., Buragohain A.K. and Ray S.K., Selection on GGU and CGU codons in the high expression genes in bacteria, J Mol. Evol., 78:13–23. 2014.