How are MILC and MELP computed?

The computation of the two statistics was originally described in:

> Supek F and Vlahovicek K: Comparison of codon usage measures and their applicability in prediction of microbial gene expressivity; *BMC Bioinformatics* (2005) 6:182, http://www.biomedcentral.com/1471-2105/6/182

However, two erroneous formulas were mistakenly included in the paper; since the *BMC Bioinformatics* journal does not publish errata, the authors decided to describe the correct procedure in this document.

It must be noted that the error does not affect the integrity of the paper regarding performance of MILC, MELP and other measures as all simulations were performed using a correct implementation of MILC in INCA 2.0.

The procedure to compute MILC follows; text quoted from paper with slight modification; errors marked in red, corrections marked in blue.

The individual contribution M_a of each amino acid a to the MILC statistic is calculated as

correct formula

$$M_a = 2\sum_c O_c \ln\frac{O_c}{E_c} = 2\sum_c O_c \ln\frac{f_c}{g_c}$$

formula in paper

$$M_a = \sum_c O_c \ln\frac{O_c}{E_c} = \sum_c O_c \ln\frac{f_c}{g_c} \qquad (1)$$

where O_c denotes the actual observed count of the codon c in a gene, and E_c stands for the expected count of the same codon. The O_c/E_c ratio is mathematically equal to, and can be replaced by f_c/g_c, where f_c is the frequency of the codon c in a gene, and g_c is the expected frequency of the same codon. The sum of f or g over all codons for each amino acid should equal 1. The total difference in codon usage is then assessed by the following formula:

$$MILC = \frac{\sum_a M_a}{L} - C \qquad (2)$$

The sum of contributions of all amino acids is divided by L, the gene length in codons, in attempt to compensate for the expected increase with total number of codons. However, such a statistic still depends on gene length, overestimating the overall amount of bias in shorter sequences due to sampling errors. The correction factor C in Equation 2 attempts to correct for this overestimation.

(... Equation 3 and accompanying text removed ...)

In a situation where all amino acids are present in a gene, the sampling errors will increase the MILC score by $41/L$. For the general case that allows missing amino acids, the correction factor C is calculated as:

correct formula

$$C = \frac{\sum_a (r_a - 1)}{L} - 0.5$$

formula in paper

$$C = \frac{\sum_a (r_a - 1)}{L} + 0.5 \qquad (4)$$

where r_a is the number of possible codons for the amino acid *a* - its degeneracy class. Only the amino acids actually present at least once in the sequence contribute to *C*, e.g. if a gene missed one of the four-fold amino acids, *C* would be 38/*L* - 0.5. When the observed frequencies match the expected codon distribution closely, MILC can assume negative values. In order to compensate, a constant of 0.5 is ~~added to~~ subtracted from the correction factor *C* (see Equation 4).

Regarding minimum sequence length, we recommend that only sequences of 80 codons or longer be analysed using MILC (or any other measure of codon usage); many researchers set this threshold to even higher values, such as 100.

Treatment of stop codons. In the paper, our general suggestion was to exclude stop codons from calculation; whether they're included or not should have only a minimal impact on the results. In INCA, I opted to include stop codons to be more consistent with the other supported measures (Karlin and Mrazek's B, MCB). If you want to compute MILC skipping the stop codons, you can use the *MILCnoStop* routine in INCAblocks 2.1.