Estimating the strength of selection on synonymous codon usage

The frequency of optimal codons (F_{op}) reflects the balance between the optimal to non-optimal codons synonymous substitution rate (K_{on}) and the non-optimal to optimal codons synonymous substitution rate (K_{no}):

$$\text{Non-optimal codon} \underset{K_{on}}{\overset{K_{no}}{\rightleftharpoons}} \text{Optimal codon}$$

Substitution rates depend on the corresponding mutation rates (μ_{no}, μ_{on}) and fixation probabilities (P_{no}, P_{on}):

$$K_{no} = 2 N_e \mu_{no} P_{no}$$

$$K_{on} = 2 N_e \mu_{on} P_{on}$$

where N_e is the effective population size.

Fixation probabilities are given by:

$$P_{no} = \frac{1 - e^{-4 N_e f_0 s}}{1 - e^{-4 N_e s}} = \frac{1 - e^{-2s}}{1 - e^{-4 N_e s}} \overset{s \to 0}{=} \frac{2s}{1 - e^{-4 N_e s}} \quad \text{and} \quad P_{on} \overset{s \to 0}{=} \frac{-2s}{1 - e^{4 N_e s}}$$

Where s is the selection coefficient in favor of optimal codons and f_0 the allele frequency of a new arrival mutation ($f_0 = 1/2N_e$) .

At equilibrium, the frequency of optimal codons is given by:

$$Fop = \frac{K_{no}}{K_{no} + K_{on}}$$

which can be written as:

$$Fop = \frac{2 N_e \mu_{no} P_{no}}{2 N_e \mu_{no} P_{no} + 2 N_e \mu_{on} P_{on}}$$

$$Fop = \frac{\mu_{no} P_{no}}{\mu_{no} P_{no} + \mu_{on} P_{on}} = \frac{\frac{\mu_{no}}{\mu_{on}} \frac{2s}{1 - e^{-4 N_e s}}}{\frac{\mu_{no}}{\mu_{on}} \frac{2s}{1 - e^{-4 N_e s}} + \frac{-2s}{1 - e^{4 N_e s}}}$$

Let us note lambda, the ratio of mutation rates: $\lambda = \frac{\mu_{no}}{\mu_{on}}$

$$Fop = \frac{\lambda}{\lambda + \frac{-\left(1 - e^{-4N_e s}\right)}{1 - e^{4N_e s}}}$$

$$\frac{1}{Fop} = 1 + \frac{1}{\lambda} \times \frac{-\left(1 - e^{-4N_e s}\right)}{1 - e^{4N_e s}}$$

$$\frac{1}{Fop} - 1 = \frac{1}{\lambda} \times \frac{-\left(1 - e^{-4N_e s}\right)}{1 - e^{4N_e s}}$$

$$\frac{1 - Fop}{Fop} \times \lambda = \frac{-\left(1 - e^{-4N_e s}\right)}{1 - e^{4N_e s}}$$

$$\frac{Fop}{1 - Fop} \times \frac{1}{\lambda} = \frac{1 - e^{4N_e s}}{-\left(1 - e^{-4N_e s}\right)} \quad \textbf{(1)}$$

With the following simplification:

$$\frac{1 - e^{4N_e s}}{-\left(1 - e^{-4N_e s}\right)} = \frac{1 - e^{4N_e s}}{-\left(1 - \frac{1}{e^{4N_e s}}\right)} = \frac{e^{4N_e s} \times \left(1 - e^{4N_e s}\right)}{\left(1 - e^{4N_e s}\right)} = e^{4N_e s}$$

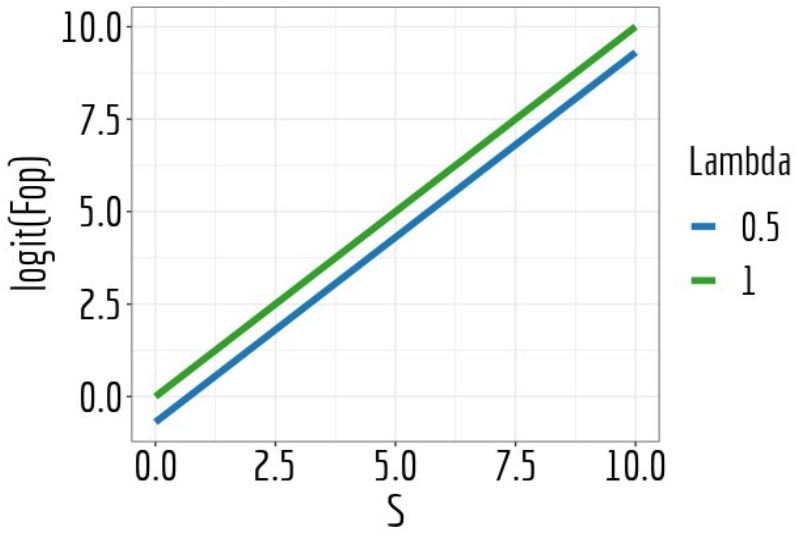
(1) → $\frac{Fop}{1 - Fop} \times \frac{1}{\lambda} = e^{4N_e s}$

Thus, the population-scaled selection coefficient (S=4 $N_e s$) is given by:

$$S = \log\left(\frac{Fop}{1 - Fop}\right) - \log(\lambda) = logit(Fop) - \log(\lambda)$$

Hence, we expect a linear correlation between *logit(Fop)* and *S*:

$$logit(Fop) = S + \log(\lambda)$$



If for weakly-expressed genes there is no selection, implied by the non-variation of *Fop* with gene expression, $S_{low\text{-}exp} \sim 0$:

$$logit\left(Fop_{low-\exp}\right) = 0 + \log\left(\lambda\right)$$

$$logit\left(Fop_{high-\exp}\right) = S_{high-\exp} + \log\left(\lambda\right)$$

$$S_{high-\exp} = logit\left(Fop_{high-\exp}\right) - logit\left(Fop_{lox-\exp}\right)$$