Available online at www.sciencedirect.com
SCIENCE DIRECT®
Biochemical and Biophysical Research Communications 327 (2005) 1–3
BBRC
www.elsevier.com/locate/ybbrc

Breakthroughs and Views

On the methodological weakness of 'the effective number of codons': a reply to Marashi and Najafabadi

Anders Fuglsang*

Department of Pharmacology, Institute of Pharmacology, Danish University of Pharmaceutical Sciences, 2 Universitetsparken, DK-2100 Copenhagen Ø, Denmark
TPR-Group ApS, 3 Puggaardsgade, DK-1573 Copenhagen V, Denmark

Received 7 November 2004
Available online 8 December 2004

Abstract

In a recent publication [Biochem. Biophys. Res. Commun. 317 (2004) 957] it was proposed that the 'effective number of codons' ($\hat{N}c$) in a gene should be calculated by summing the individual amino acid $\hat{N}c$'s using rounding whenever the codon homozygosities are lower than the reciprocal value of the number of members of the synonymous families. This led Marashi and Najafabadi to examine the consequences of individual re-adjustment when comparing observed $\hat{N}c$ with the expected Nc under assumptions of no selection, and C = G and A = T [Biochem. Biophys. Res. Commun. 324 (2004) 1]. Clearly, the present methodology has some weaknesses; in this work, I discuss these in relation to the observations by Marashi and Najafabadi, and finally an alternative method for the calculation of $\hat{N}c$ is introduced with the purpose of eliminating the need for re-adjustments.

Keywords: Codon bias; Homozygosity; Resampling; Genetic code; Re-adjustment

In all life forms known, the usage of synonymous codons is uneven, i.e., not all codons for a particular amino acid are used equally often. A central issue in the study of this phenomenon is the calculation of a codon bias parameter, i.e., a number that tells to what extent the codons used are restricted. Wright [1] came up with excellent and intuitive idea for this: since there are 20 amino acids and 61 sense codons (with the standard gene code), a number (called '$\hat{N}c$,' the effective number of codons) between 20 and 61 can be assigned to a pool of codons (a gene) reflecting to what extent the entire genetic code is used. This number will be 20 in the most extreme cases where the synonymous usage of codons is so restricted that just one codon for each of the 20 amino acids is used. Likewise, the number will be 61 if all codons within the families are equally used.

The inspiration for the calculation of $\hat{N}c$ came from population genetics where calculation of 'effective allele numbers' is sometimes used to characterize populations from a quantitative perspective. This is also the reason for the caret symbol; we can think of $\hat{N}c$ as an estimator of the effective number of codons in a (hypothetical) pool of codons from which a gene is sampled.

To estimate the effective number of codons for a particular amino acid with k synonymous codons, Wright gave the following formula:

$$\hat{N}c(\mathrm{aa}) = \frac{1}{\hat{F}_{\mathrm{aa}}}. \tag{1}$$

Here, the denominator is called the codon homozygosity and is given by:

$$\hat{F}(\mathrm{aa}) = \frac{\left(n\sum_{i=1}^{k} p_{\mathrm{i}}^2\right) - 1}{n-1}, \tag{2}$$

* Fax: +45 35306020.
E-mail address: anfu@dfuni.dk.

0006-291X/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2004.11.133

where n is the total count of codons for the amino acid, while the p_i values are the individual codon fractions.

To finally get the effective number of codons in a gene, the homozygosities in the degeneracy families are averaged and we have

$$\hat{N}c = 2 + \frac{9}{\hat{F}_2} + \frac{1}{\hat{F}_3} + \frac{5}{\hat{F}_4} + \frac{3}{\hat{F}_6}. \quad (3)$$

The addition of two stems from the fact that methionine and tryptophan only have one codon, thus by definition both have one codon in use. Note that sometimes an amino acid is not present in a gene under investigation, and in that case the codon homozygosity is not defined for this amino acid but still an average of codon homozygosities in the degeneracy group can be found on basis of the others, treating the missing one as though it had a homozygosity that was equal to the average of that of the others in the group. In practice, the calculation this way makes it possible for $\hat{N}c$ to exceed 61. Wright recommended simply to re-adjust the number of 61 in that case.

Using *Escherichia coli* as reference organisms I recently found that there is a poor correlation between specific codon homozygosities and the averaged homozygosities found by averaging the others in the group [2]. Therefore, it was suggested that Eq. (3) was sub-optimal. Instead, it was proposed to calculate each $\hat{N}c(\mathrm{aa})$ individually using Eqs. (1) and (2), and add them up one by one and doing individual re-adjustments where the calculated $\hat{N}c(\mathrm{aa})$ exceeds the number of synonyms

$$\hat{N}c^* = \hat{N}c(\mathrm{Ala}) + \hat{N}c(\mathrm{Arg}) + \cdots + \hat{N}c(\mathrm{Val}). \quad (4)$$

The drawback of this approach is that all amino acids must be present in a gene in sufficient numbers, meaning that $\hat{N}c^*$ in practice can be calculated for much fewer genes than $\hat{N}c$. Nevertheless, it was shown that $\hat{N}c^*$ is the superior estimator when there is a risk of 'bias discrepancy,' i.e., when the codon usage for one or more amino acids in a degeneracy group is restricted while the codon usage for one or more of the others is not.

In the original work by Wright, it was suggested to plot observed values of $\hat{N}c$ as function of the GC-content at third codon sites(s), as a possible means quantifying the degree of selection on codon usage. The idea is that if there is no selection on codon usage, and if we assume that G = C and A = T then the expected value of $\hat{N}c$ is given by a quite simple expression which varies between the amino acids. This led Marashi and Najafabadi [3] to study the difference between observed and expected $\hat{N}c$. Under some circumstances this difference disappears depending on the observed value of GC3s. Does re-adjustment then imply loss of information? Yes, in fact it does; for example, if gene A holds two lysine AAA codons and one AAG codon, then there will be 3.0 effective lysine codons before re-adjustment to $\hat{N}c(\mathrm{Lys}) = 2$. If gene B holds three AAA codons and two AAG codons, then there will be 2.5 effective lysine codons before we re-adjust to 2. After the re-adjustment we cannot distinguish these two genes on basis of their $\hat{N}c(\mathrm{Lys})$ but we certainly could before. Re-adjustment therefore results in a loss of information. The resampling experiments in [2] show that this is *generally* not a problem when there is a risk of bias discrepancy, that is, with overall amino acid usage like the one observed in *E. coli* rounding generally (but that does not mean specifically) results in more accurate estimates, but this has only been tested with *E. coli* as reference and only with the sum of all individual $\hat{N}c(\mathrm{aa})$'s.

I do have a few comments to the argumentation provided by Marashi and Najafabadi, who exemplify their concern with two examples (*cdsA* and *yegG*) from *E. coli* K12. However, one of the genes referred to, *yegG*, does at present not exist on the sequence (GenBank Accession No. NC_000913). Second, the expected value of $\hat{N}c$ is defined by considering $n \to \infty$, whereas in genes actual counts are very low, and as a consequence the uncertainty on $\hat{N}c$ is in practice high. It is therefore not surprising to find examples where the difference between observed and expected values is problematic. But that for sure does not mean that Marashi and Najafabadi are wrong. In fact, I tend to agree that we need to reconsider our way of quantifying codon bias.

Codon homozygosity—back to basics

We may have to re-evaluate the effective number of codons entirely. One disadvantage of the codon homozygosity calculated as in Eq. (2) becomes clear when we for consider examples of exactly equal observed usage of the two codons for a twofold degenerate amino acid. From Eq. (2) we can immediately see that $F \to 2^{(+)}$ for $n \to \infty$, and this is not very ideal; in fact, these situations are basically the reason for all of the re-adjustment discussion. In my opinion, the rationale for introducing Eq. (2) as appropriate estimator of codon homozygosity is somewhat unclear in the paper by Wright. In that paper, it was mentioned that the derivation was inspired by formulae for the effective number of alleles, when we simply think of codons as alleles. However, in classical population genetics, the homozygosity is simply the sum of frequencies [4]

$$\hat{F} = \sum_{i=1}^{k} p_i^2, \quad (5)$$

and arguably this could prove more appropriate to use in this context. This value approaches Wright's value for $n \to \infty$ (so is Eq. (2) in fact similar to Eq. (5) but intended to yield an unbiased estimate at low counts? I cannot tell!).

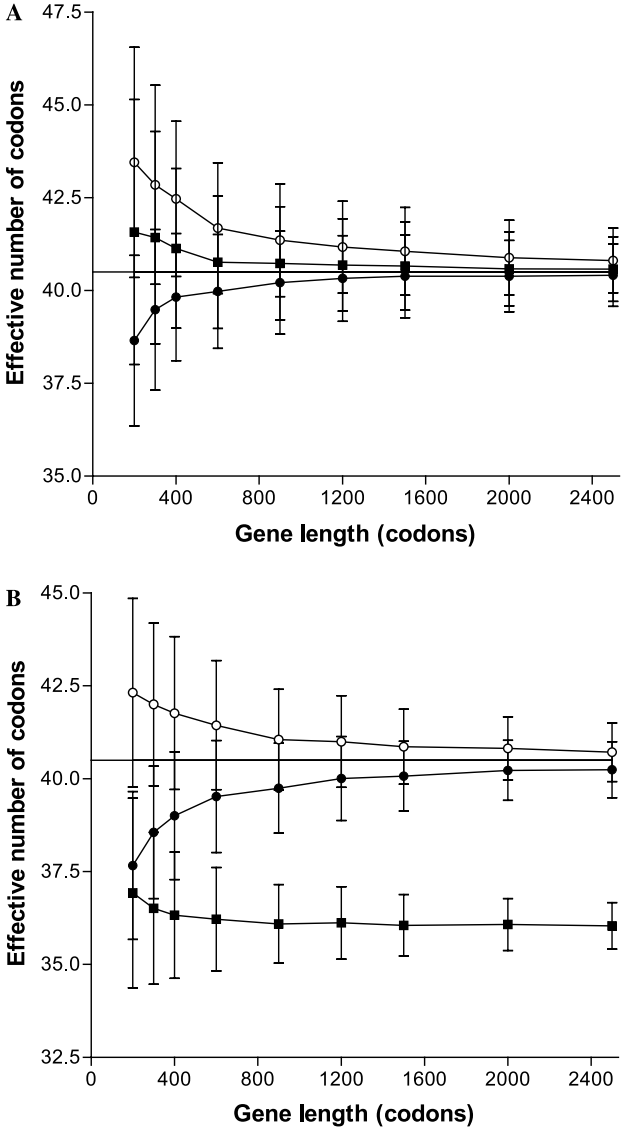

Fig. 1. (A) Simulation results of $\hat{N}c^{**}$ (●) versus $\hat{N}c^{*}$ (○) and Wright's $\hat{N}c$ (■). The points are averages and error bars represent standard deviations on the basis of 500 resamplings per point. The test conditions are the same as for Fig. 5A in [2], i.e., $Nc = 40.5$ (horizontal line) and no bias discrepancy. Under these circumstances Wright's $\hat{N}c$ is superior. $\hat{N}c^{*}$ underestimates the codon bias, whereas $\hat{N}c^{**}$ overestimates the codon bias. (B) With bias discrepancy, Wright's $\hat{N}c$ converges towards an incorrect value (for an explanation, see [2]). Again, $\hat{N}c^{*}$ underestimates the codon bias, whereas $\hat{N}c^{**}$ overestimates the codon bias. The standard deviation for $\hat{N}c^{**}$ is lower than for $\hat{N}c^{*}$ and Wright's $\hat{N}c$ in both figures.

The general formula for the effective number of codons for a particular amino acid then simply becomes

$$\hat{N}c(\mathrm{aa}) = \frac{1}{\sum_{i=1}^{k} p_i^2}. \tag{6}$$

We can see that re-adjustment will no longer be an issue since $\hat{F} \geqslant \mathrm{k}^{-1}$ always holds, and consequently the problem with even proportions is solved.

So the new formula for the effective number of codons ($\hat{N}c^{**}$) in a gene ends up being:

$$\hat{N}c^{**} = 2 + \sum_{\mathrm{aa}=1}^{18} \left(\sum_{i=1}^{k(\mathrm{aa})} p_i^2 \right)^{-1}. \tag{7}$$

This new codon bias parameter has been tested against Wright's $\hat{N}c$ and $\hat{N}c^{*}$, using the same test procedure as that reported in [2]. That is, the true value of Nc was kept constant at 40.5, and the test involved 500 resamplings under conditions of no bias discrepancy (Fig. 1A) and strong bias discrepancy (Fig. 1B), respectively. As Fig. 1A shows, Wright's $\hat{N}c$ is superior to $\hat{N}c^{*}$ and $\hat{N}c^{**}$ when there is no bias discrepancy. However, under circumstances of high bias discrepancy it can be seen that $\hat{N}c^{*}$ is the superior method. Both $\hat{N}c^{*}$ and $\hat{N}c^{**}$ converge towards the true value, while $\hat{N}c$ approaches a value much lower than the true value (for an explanation of this phenomenon, see [2]). The only advantage of $\hat{N}c^{**}$ lies in the fact that it has a lower standard deviation. It can be seen that $\hat{N}c^{*}$ underestimates the codon bias while $\hat{N}c^{**}$ overestimates it. Averaging $\hat{N}c^{*}$ and $\hat{N}c^{**}$ therefore gives the better estimate. There is thus plenty of room for further improvement in the current methodology, and this viewpoint is fully in line with that of Marashi and Najafabadi.

References

[1] F. Wright, The 'effective number of codons' used in a gene, Gene 87 (1990) 23–29.
[2] A. Fuglsang, The 'effective number of codons' revisited, Biochem. Biophys. Res. Commun. 317 (2004) 957–964.
[3] S.A. Marashi, H.S. Najafabadi, How reliable re-adjustment is: correspondence regarding A. Fuglsang, "The 'effective number of codons' revisited", Biochem. Biophys. Res. Commun. 324 (2004) 1–2.
[4] M. Kimura, J.F. Crow, The number of alleles that can be maintained in a finite population, Genetics 49 (1964) 725–738.