Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Biochemical and Biophysical Research Communications 317 (2004) 957–964

BBRC

www.elsevier.com/locate/ybbrc

The 'effective number of codons' revisited[☆]

Anders Fuglsang*

Institute of Pharmacology, Danish University of Pharmaceutical Sciences, Universitetsparken 2, DK-2100 Copenhagen Ø, Denmark

Received 8 March 2004

Abstract

Frank Wright [Gene 87 (1990) 23] derived a formula for calculation of a quantity termed the 'effective number of codons' ($\hat{N}c$) based on codon homozygosities. This quantity is a number between 20 and 61 and tells to what degree the codon usage in a gene is biased, i.e., it approaches 20 codons for the extremely biased genes, and approaches 61 for the genes where all possible codons are used with no preference. Among the different measures of codon bias $\hat{N}c$ is considered the most useful and has found widespread use in papers dealing with codon usage phenomena. In this paper, the mathematical behaviours of codon homozygosities and $\hat{N}c$ are evaluated, using *Escherichia coli* as the model organism. The results indicate that the classical formula for calculation of $\hat{N}c$ could appropriately be substituted under circumstances, where there is bias discrepancy, i.e., when one amino acid (or more) within a degeneracy group is associated with strong codon bias while at the same time others in the same degeneracy group have little bias. An alternative estimator, termed $\hat{N}c^*$, is proposed and tested against $\hat{N}c$, and performs better when there is such bias discrepancy.

Keywords: Codon usage; Homozygosity; Genetic code; *Escherichia coli*; Resampling

Codon usage bias is a phenomenon that is defined by the fact that 61 codons encode 20 amino acids, making 18 of them degenerate. In all organisms studied so far, it has been shown that codon usage is directed by a complex set of factors. The composition of genomes is to a large extent determined by the mutational pressure that has acted on an organism, and this is generally reflected in the codon usage; organisms with a high GC-content (such as members of the *Streptomyces* genus) tend to use mainly GC-rich codons while the opposite is true for AT-rich species (such as *Campylobacter jejuni*) [1,2]. In addition, in many species selective factors, notably gene expressivity is a major determinant of codon usage [3], for a review, see [4], since codon usage in highly expressed genes has a preference for abundant tRNA species [5]. Some codon usage measures are based on this phenomenon, such as the 'codon adaptation index' [6]. However, these measures are species-dependent as optimal codons differ. As a measure of species-independent synonymous codon bias in genes, the 'effective number of codons' ($\hat{N}c$) was introduced by Wright in 1990 [7]. It is today one of the most widely used estimators for codon bias and has some advantages over other measures (reviewed in [8]). $\hat{N}c$ tells to what degree all 61 codons (with the standard genetic code) are in use in a gene. In extremely biased genes the effective number of codons can approach 20, while in unbiased genes it will approach 61.

To calculate the bias, one needs to quantify codon homozygosity ($\hat{F}$) for all amino acids having synonymous codons

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

where n is the total count for the amino acid in the gene, and p_i is the codon frequency for the ith synonymous codon for the particular amino acid. For the individual amino acid we can calculate the effective number of codons as

$$\hat{N}c(\text{aa}) = \frac{1}{\hat{F}_{\text{aa}}}. \tag{2}$$

[☆] *Abbreviations:* Nc, effective number of codons; aa, amino acid; *E.*, *Escherichia*; F, codon homozygosity.

[*] Fax: +45-35-30-60-20.
E-mail address: anfu@dfuni.dk.

0006-291X/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2004.03.138

The effective number of codons in the gene is then according to Wright calculated by:

$$\hat{N}c = 2 + \frac{9}{\hat{F}_2} + \frac{1}{\hat{F}_3} + \frac{5}{\hat{F}_4} + \frac{3}{\hat{F}_6}, \tag{3}$$

where $\bar{\hat{F}}_2$ is the average homozygosity for the amino acids having a degeneracy of two (histidine, glutamine, etc.) and so on. Wright also suggested solutions in cases where amino acids were missing. If, for example, a gene does not contain threonine, then $\bar{\hat{F}}_4$ will be the average of the codon homozygosities of glycine, valine, alanine, and proline:

$$\bar{\hat{F}}_4 = \frac{\hat{F}_{\text{pro}} + \hat{F}_{\text{gly}} + \hat{F}_{\text{ala}} + \hat{F}_{\text{val}}}{4}. \tag{4}$$

This is equivalent to assuming that the codon homozygosity for threonine in this situation equals the average codon homozygosity of the others within that degeneracy group. How good an approximation is this? In this paper, one of the goals is to test this assumption by evaluation of the correlation between estimated codon homozygosities and observed codon homozygosities where possible.

There is a chance that $\hat{N}c$, calculated through use of Eq. (3), will exceed 61. In that case, Wright recommends re-adjusting the result down to 61. He did not give any reason why this correction should not be applied to individual amino acids in stead. A hypothetical example: one could have a situation where the apparent number of effective alanine and glycine codons is 6 (arises when the codon homozygosity is 1/6, which is the case if three of the synonymous codons are present twice while the fourth synonymous codon is present three times in a gene) and with no other overshooters, but where the $\hat{N}c$ calculated the classical way (Eq. (3)) turns out to be 62. In this case, one should according to Wright re-adjust to 61. However, doesn't this still leave three of the effective codons unaccounted for? Also, Wright suggested, with little argumentation, using $\hat{F}_3 = (\bar{\hat{F}}_2 + \bar{\hat{F}}_4)/2$ when the isoleucine estimator was not possible to calculate. In this paper, these issues are subjected to further examination in an attempt to improve the accuracy of the estimate.

Materials and methods

Choice of reference strain. In his original paper, Wright used *Escherichia coli* K12 as a reference organism. This is adopted here. *E. coli* K12 is fully sequenced and is in many other ways the best characterised microorganism. Furthermore, a recent paper describing use of effective number of codons for individual amino acids [9] was based on this strain.

Analysis of the behaviour of codon homozygosity. Using lysine as an example, it was tested how the codon homozygosity varies with varying content of lysine codons. This was done by generating simulated genes with 1–20 AAA codons and 1–20 AAG codons and subsequently calculating the codon homozygosity for lysine in the simulated genes. Plots of codon homozygosity as function of the counts of AAA and AAG codons were generated.

Test of homozygosity estimates. Wright recommended estimating the codon homozygosity in accordance with the example in Eq. (4) in cases where an amino acid was present in too low counts to allow calculation of its codon homozygosity. While the efficiency of this estimate cannot be tested directly, it was tested to what degree Eq. (4) estimates the codon homozygosity in cases where the actual codon homozygosity could also be calculated. This was carried out with the five amino acids having a fourfold degeneracy. The estimated codon homozygosities were then plotted against and the actual codon homozygosities. If the estimate is appropriate, then the plot should yield a straight line having a slope of one and an intercept of zero.

The special case of isoleucine. It is virtually always possible to calculate average codon homozygosities ($\bar{\hat{F}}$) for the group of twofold, fourfold, and sixfold degenerate amino acids, simply because it is very unlikely that not one single homozygosity ($\hat{F}$) can be calculated within each degeneracy group (bear in mind that, strictly, just one $\hat{F}$ value suffices in order to obtain an $\bar{\hat{F}}$ value). In this regard isoleucine may be problematic since it is the only amino acid that is encoded by three codons. Wright recommended without argumentation that the codon homozygosity for isoleucine be estimated as

$$\hat{F}_3 = \frac{\bar{\hat{F}}_2 + \bar{\hat{F}}_4}{2}. \tag{5a}$$

This looks intuitive, but it is quite unclear on what mathematical reasoning it is based. To explain my concern, consider the following example:

In the case of a completely unbiased gene, we have $\bar{\hat{F}}_2 = 0.5$ and $\bar{\hat{F}}_4 = 0.25$ (two and four codons, respectively). In such an unbiased gene we would according to Eq. (5a) arrive at $\hat{F}_3 = 0.375$, corresponding to 2.67 effective isoleucine codons (insertion in Eq. (2))—obviously this is problematic; the intuitive value must be 3, because the gene is completely unbiased, so all three isoleucine codons will (statistically) be used. Eq. (5a), though it looks intuitive, is thus incorrect.

The relationship between codon homozygosities is not linear (if there is a relationship), instead it is, with the argumentation above, the effective number of codons (F^{-1} estimates) that relate linearly. For $\bar{\hat{F}}_2^{-1} = 1$ (extreme bias, one codon effectively used for the twofold degenerate aa) the best estimate is $\hat{F}_3^{-1} = 1$ (corresponding to one isoleucine codon used). Similarly, if $\bar{\hat{F}}_2^{-1} = 2$ (no bias) we get $\hat{F}_3^{-1} = 3$ (three isoleucine codons in use). Then the actual estimate for $\hat{F}_3$ based on $\bar{\hat{F}}_2$ is (verify this by insertion):

$$\frac{1}{\hat{F}_3} = \frac{2}{\bar{\hat{F}}_2} - 1$$

$$\Updownarrow$$

$$\bar{\hat{F}}_3 = \left(\frac{2}{\bar{\hat{F}}_2} - 1\right)^{-1}. \tag{5b}$$

With the same argumentation it is easily shown that $\hat{F}_3$ can be estimated through $\bar{\hat{F}}_4$ by the equation:

$$\bar{\hat{F}}_3 = \left(\frac{2}{3\bar{\hat{F}}_4} + \frac{1}{3}\right)^{-1}. \tag{5c}$$

And also, $\hat{F}_3$ can be estimated through $\bar{\hat{F}}_6$ by the equation:

$$\bar{\hat{F}}_3 = \left(\frac{2}{5\bar{\hat{F}}_6} + \frac{3}{5}\right)^{-1}. \tag{5d}$$

Combining $\bar{\hat{F}}_2$ and $\bar{\hat{F}}_4$ in an estimate, as was probably the intention of Eq. (5a), yields:

$$\bar{\hat{F}}_3 = \frac{\left(\frac{2}{\bar{\hat{F}}_2} - 1\right)^{-1} + \left(\frac{2}{3\bar{\hat{F}}_4} + \frac{1}{3}\right)^{-1}}{2}. \tag{6}$$

Note how different this is compared to Eq. (5a). Ultimately, combining the three estimates in one, we arrive at

$$\frac{\bar{\hat{F}}_3 = \left(\frac{2}{\bar{\hat{F}}_2} - 1\right)^{-1} + \left(\frac{2}{3\bar{\hat{F}}_4} + \frac{1}{3}\right)^{-1} + \left(\frac{2}{5\bar{\hat{F}}_6} + \frac{3}{5}\right)^{-1}}{3.} \tag{7}$$

Note the inherent properties of these two expressions in contrast to Eq. (5a): for a completely biased gene, we have homozygosities of 1, so the resulting $\hat{\bar{F}}_3$ will be 1 also. For a completely unbiased gene, we have $\hat{\bar{F}}_2 = 0.5$, $\hat{\bar{F}}_4 = 0.25$, and $\hat{\bar{F}}_6 = 1/6$, making all three terms in the nominator equal to 3, whereby the resulting $\hat{F}_3$ is 3, as it should be.

The different estimates were plotted against the observed values for $\hat{F}_3$ and correlation analysis was performed in order to evaluate how good the estimates are.

An alternative formula for the effective number of codons. In his paper, Wright wrote (the theory has its basis in consideration of the effective number of alleles, N_e, for a given locus):

"The method here can be thought of as adding together the 'N_e' values for each of the 20 'loci'." Inspection of Eq. (3) tells us that this is actually not exactly what his formula does. It does not add together the 'N_e' (Nc) values for the '20 loci' (amino acids), rather it adds the average Nc for amino acids within the degeneracy groups. The concept of Eq. (3) is strictly contradictory to the quote given above; however, it appears more intuitive to me (and this is also the concept of the quote) to calculate the effective number of codons by addition of the individual effective numbers of codons:

$$\hat{N}c^* = \hat{N}c_{\text{ala}} + \hat{N}c_{\text{arg}} + \hat{N}c_{\text{asp}} + \cdots + \hat{N}c_{\text{val}}, \tag{8}$$

where each of the individual values is calculated according to Eq. (2), and where each individual Nc-value is adjusted if it exceeds the number of synonymous codons. Values for $\hat{N}c^*$ were calculated this way and compared to the classical $\hat{N}c$.

Test of $\hat{N}c^*$ and $\hat{N}c$ using simulated genes. The phenomenon of having strong bias for one or more amino acids within a degeneracy group while others members of the degeneracy group have little bias may introduce spurious deviation in the outcome of $\hat{N}c$. As a highly extreme and hypothetical example consider a situation where there are six serine and arginine codons in effective use (no bias at all) while there is complete bias of leucine with one effective codon. In this situation $\hat{\bar{F}}_6$ becomes 0.75 and the right-most part of Eq. (3) becomes 4, even though there are 13 leucine codons. This example is of course rather extreme, but the data revealed that this phenomenon is not uncommon (we shall hereafter call it 'bias discrepancy,' when or more one amino acids within a degeneracy group are associated with strong codon bias while at the same time others in the same degeneracy group have little bias). Therefore, different scenarios were derived in which the true effective number of codons was controlled through a multinomial distribution of codons with realistic codon probabilities. Simulated genes were used to test $\hat{N}c^*$ (with and without rounding) and $\hat{N}c$ under circumstances where there was no bias discrepancy and under circumstances where there is bias discrepancy. The simulated genes varied in length from 200 codons to 2500 codons, and had an amino acid composition corresponding to that actually observed in *E. coli*. The simulation algorithm implemented a random number generator written in accordance with the recommendation of Press et al. [11], and picked codons on basis of the codon probabilities of the different scenarios. Sequences were resampled 500 times and average and standard deviation of the estimators were calculated.

Data generation and statistics. The bioinformatic software used in this paper was programmed by the author of this paper. For software availability, see Section 3.5. The data presented were processed from the complete *E. coli* sequence (GenBank Accession No. NC_000913). Genes were only included if they had correct start and stop codons, no internal stop codons, and an intact frame. GraphPad Prism 3 (GraphPad, CA, USA) was used for correlation analysis. Linear regression was used when a linear relationship was expected; otherwise non-parametric correlation was used (Spearman's rank). The Kolmogorov–Smirnov test was run on the output of the resampling experiments in order to test if the data can be assumed Gaussian. A probability level corresponding to less than 5% chance was considered significant.

Results and discussion

Behaviour of codon homozygosity

Fig. 1 shows how codon homozygosity behaves, exemplified by varying lysine codon content in genes. The curve is generally applicable but it becomes more complex for amino acids having a higher degeneracy. It can be seen that the codon homozygosity tends to too low values when the individual codons are present in approximately equal counts, and especially when the overall counts are low. This suggests that rare amino acids, on which there is generally little difference in codon usage for the individual synonyms (for example, cysteine), will contribute more to $\hat{\bar{F}}$ than more abundant amino acids or amino acids where there is a marked difference in codon usage between the synonyms (for example, lysine), when calculating $\hat{N}c$ the classical way by use of Eq. (3). For a specific example dealing with this phenomenon, see later (the *lrp* gene).

Estimates of codon homozygosities

Where possible, the observed codon homozygosity for fourfold degenerate amino acids was calculated along

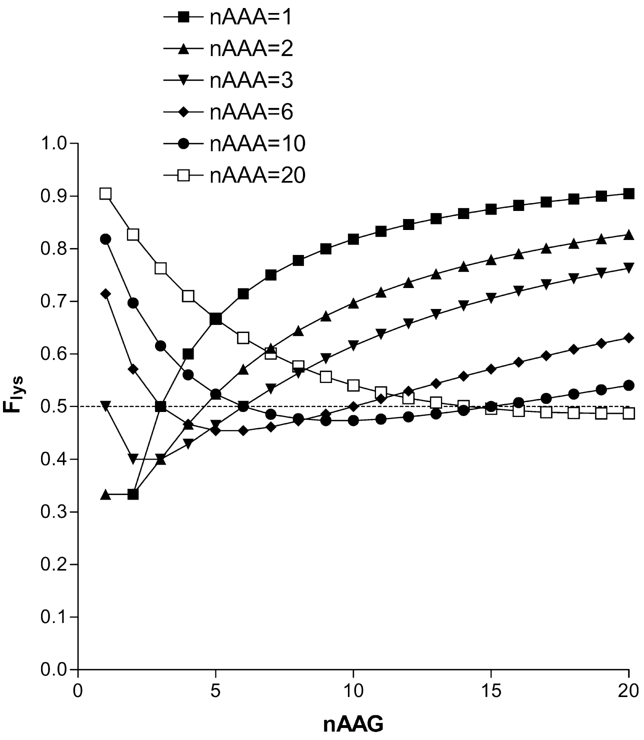

Fig. 1. Uncorrected codon homozygosity for a twofold degenerate amino acid (here exemplified by lysine), calculated through use of Eq. (1), in simulated genes. The dashed line discriminates values that need to be corrected; values below this line are too low since the standard genetic code only allows two codons. nAAA and nAAG are the counts of AAA and AAG codons, respectively, in the simulated gene. Values generally become too low when the counts of AAA and AAG are approximately equal. This curve is representative for all amino acids having a twofold degeneracy.

with the estimated values. Table 1 lists the results of these correlations. For good estimators, the real value of the slope should be close to one and the intercept close to zero.

Table 1
The estimate of codon homozygosity for the individual amino acids (based on average of homozygosities for the others) were correlated with the actual homozygosity, as $F_{\mathrm{aa,est.}} = aF_{\mathrm{aa}} + b$

Estimate	Slope, a	Intercept, b	r^2
F_{ala}	0.1944	0.2776	0.0351
F_{gly}	0.1728	0.2689	0.0579
F_{pro}	0.0903	0.2761	0.0754
F_{thr}	0.1252	0.2806	0.0487
F_{val}	0.1379	0.2914	0.0264

The correlation is weak. The estimation method is not very likely to give a good estimate. In all cases, the 95% confidence interval for the slope is below I, and the 95% confidence interval intercept above 0.

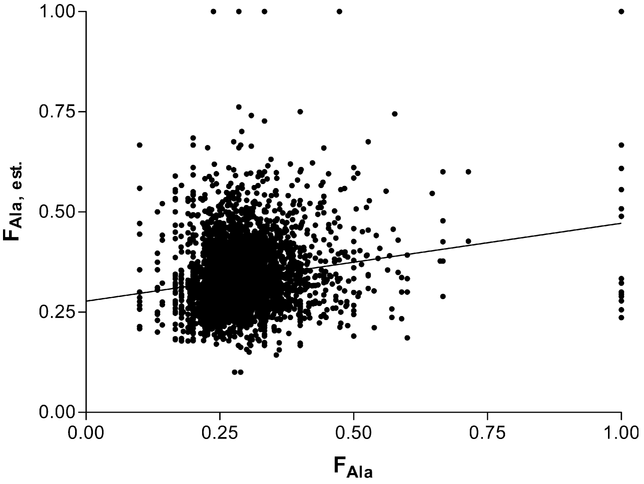

Fig. 2. Estimated homozygosity as function of actual homozygosity for alanine in genes of *E. coli*. The appropriateness of the estimation is tested by least-squares-fitting to a straight line according to $F_{\mathrm{ala,est.}} = aF_{\mathrm{ala}} + b$. Ideally, the best fit line should yield $a = 1$ and $b = 0$. The actual results are that $a = 0.1944$ and $b = 0.2776$, $r^2 = 0.0351$ (4245 data points). The 95% confidence intervals are $0.1624 < a < 0.2254$ and $0.2683 < b < 0.2870$. Because of the low correlation coefficient and the deviation from $a = 1$ and $b = 0$, alanine codon homozygosity is not well estimated by the method proposed by Wright.

This is not the case for any of the five homozygosities, and the correlation coefficients are very low. A graphical illustration of estimated alanine homozygosity versus observed homozygosity is given in Fig. 2. It can be concluded that in the absence of an amino acid, estimating it by averaging the homozygosities of the others is not a good method.

The special case of isoleucine deserves special attention. Table 2 lists the correlation observed with the different homozygosity estimates (Eqs. (5a)–(5d), (6), (7)) and the observed homozygosities. The table shows that the classical way of estimating $\hat{F}_3$ (Eq. (5a)) is rather inefficient as are the other estimates, too. Generally we observe very poor correlations, $r^2 < 0.1$ in all cases, the slopes differ significantly from one, and the intercepts differ significantly from zero.

The reason for these deviations may lie in the differential usage of codons in the degeneracy classes. In a recent paper [9], I examined how translationally optimal codons as well as other codons are used in genes of *E. coli*, because selection of translational efficiency is know to play a major role in this bacterium and many other prokaryotes [5,11]. For example, there is only one optimal codon for proline, but there are two for alanine, glycine, and threonine, while valine is somewhat uncertain; there are three optimal codons according to Ikemura, but only one was identified, which showed a clear pattern of optimality in my recent paper. The point is, if codon usage is strongly determined by optimal codons and not all amino acids in a degeneracy group have an equal number of optimal codons, then this might account for some of the lacking correlations. I have run similar analysis on the genome of *Helicobacter pylori*, which is reported to be a species subjected to less translational selection on codon usage [12], but the correlations are not better (data not shown). This suggests that the problem with lack of correlation is a matter of the methodology rather than a species-specific codon usage phenomenon. The conclusion so far must be that codon homozygosities cannot be efficiently estimated in all cases. The lack of

Table 2
Correlation of isoleucine homozygosity estimates with the observed values according to Eqs. (5a)–(5d), (7), (8)

Estimate	Slope	Intercept	r^2
$\hat{F}_3 = \frac{\bar{\hat{F}}_2 + \bar{\hat{F}}_4}{2}$	0.0997	0.4003	0.0518
$\bar{\hat{F}}_3 = \left(\frac{2}{\bar{\hat{F}}_2} - 1\right)^{-1}$	0.0622	0.3740	0.0104
$\bar{\hat{F}}_3 = \left(\frac{2}{3\bar{\hat{F}}_4} + \frac{1}{3}\right)^{-1}$	0.1524	0.3432	0.0602
$\bar{\hat{F}}_3 = \left(\frac{2}{5\bar{\hat{F}}_6} + \frac{3}{5}\right)^{-1}$	0.2702	0.3734	0.0822
$\bar{\hat{F}}_3 = \frac{\left(\frac{2}{\bar{\hat{F}}2} - 1\right)^{-1} + \left(\frac{2}{3\bar{\hat{F}}4} + \frac{1}{3}\right)^{-1}}{2}$	0.1073	0.3586	0.0535
$\bar{\hat{F}}_3 = \frac{\left(\frac{2}{\bar{\hat{F}}2} - 1\right)^{-1} + \left(\frac{2}{3\bar{\hat{F}}4} + \frac{1}{3}\right)^{-1} + \left(\frac{2}{5\bar{\hat{F}}6} + \frac{3}{5}\right)^{-1}}{3}$	0.1634	0.3626	0.0942

correlation seems to justify a slightly different approach for calculation of the effective number of codons in a gene. To do so, one must deal with problematic genes, i.e., when the actual counts of codons otherwise seem to call for an estimate.

A problematic example: the lrp gene

The *lrp* gene of *E. coli* encodes a transcriptional regulator of branched chain amino acid metabolism. Using the standard method of Wright, $\hat{N}c$ can be calculated to be 60.3. However, if we inspect the gene more closely it becomes apparent that the codon counts in this gene are somewhat skewed, in that the counts of individual codons for a particular amino acid are quite similar, whereby uncorrected $\hat{N}c$ needs readjustment (as exemplified in Fig. 1). In fact, the Nc for 11 out of the 20 amino acids needs individual readjustment! Table 3 lists some of these. Note how alanine behaves quite different from the other fourfold degenerates; this gene is a good example of principle illustrated in Fig. 1, how low codon counts at approximately equal amount may yield a very low codon homozygosity, and conversely a high apparent value for the effective number of codons for individual amino acids. If we calculate the individual number of codons for every amino acid except for cysteine (and correct for overshooting) and add them (as suggested by Eq. (8)), we arrive at $\hat{N}c = 49.8$, but the gene only contains two cysteine codons, thus an Nc for cysteine cannot be included since the codon homozygosity becomes zero. Since Nc(Cys) must be a number between 1 and 2, it seems logical that $\hat{N}c^*$ then will be somewhere in the interval 50.8–51.8, however, it is not satisfactory to get an interval as result. I cannot devise a ready-to-run suggestion what to do in this situation, apart from the conservative approach of simply excluding those genes that do not allow proper calculation of the individual number of codons for all amino acids. This means that mainly the shorter genes would be discarded or unsuitable for the calculation. But, because Wright pointed out that $\hat{N}c$ calculated his way is overestimated for shorter genes, the limitations of $\hat{N}c^*$ may in this regard be similar to those of Nc.

Table 3
The *lrp* gene, selected uncorrected and corrected number of codons for individual amino acids

Amino acid	Uncorrected Nc (AA)	Corrected Nc (AA)
Ala	10	4
Gly	4	4
Pro	5	4
Thr	5.14	4
Val	5.25	4
Arg	2.45	2.45
Leu	3.20	3.20
Ser	9.33	6

Direct comparison of $\hat{N}c^*$ and $\hat{N}c$

It is thus proposed that $\hat{N}c^*$ could be an alternative to $\hat{N}c$. Fig. 3A shows how $\hat{N}c^*$ and $\hat{N}c$ correlate. The correlation is quite good: $r_s = 0.8249$, $P < 0.0001$.

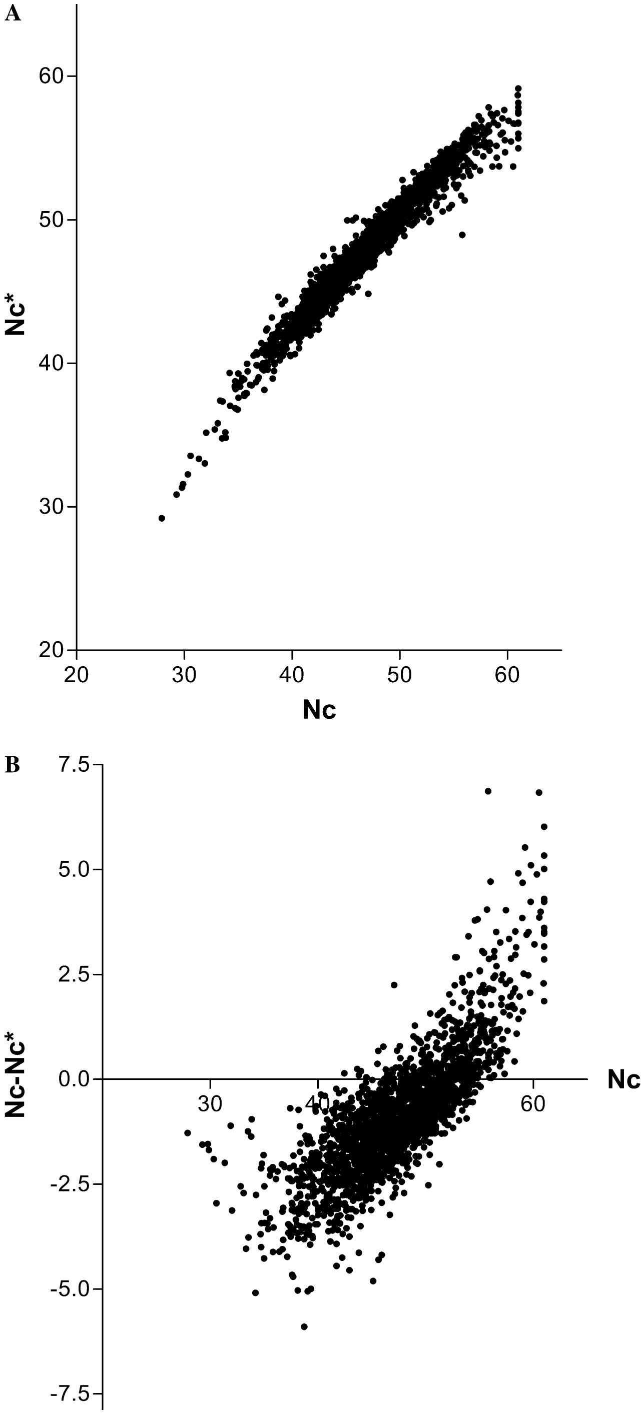

Fig. 3. Relationship between $\hat{N}c^*$ and $\hat{N}c$ calculated the classical way. There is a good correlation between the two parameters (upper curve, A): $r_s = 0.8249$ (1968 points). The lower curve (B) shows a plot of the difference between $\hat{N}c^*$ and $\hat{N}c$ versus $\hat{N}c$. $\hat{N}c$ is larger than $\hat{N}c^*$ at high values of $\hat{N}c$ and vice versa.

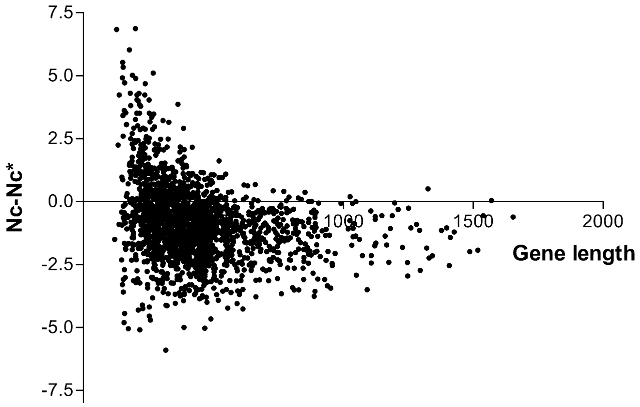

Fig. 4. Difference between $\hat{N}c$ and $\hat{N}c^*$ plotted against gene length measured in codons. The difference tends to decrease with increasing gene length, $r_s = -0.3041$, $P < 0.0001$ (1968 points).

The difference between $\hat{N}c$ and $\hat{N}c^*$ is plotted as function of $\hat{N}c$ in Fig. 3B. It can be seen that $\hat{N}c$ is larger than $\hat{N}c^*$ for genes having a low bias (high values of $\hat{N}c$). As mentioned the value of $\hat{N}c$ is generally high (overestimated) for shorter genes. Fig. 4 shows a plot of $\hat{N}c$–$\hat{N}c^*$ as function of gene length (codons). The figure is not exceedingly clear, but there is a tendency towards positive values of $\hat{N}c$–$\hat{N}c^*$ for short genes ($r_s = 0.3041$, $P < 0.0001$). Wright [7] found that $\hat{N}c$ overestimates at shorter lengths, so $\hat{N}c^*$ may from this point of view be a good alternative. One should however remember that $\hat{N}c^*$ can only be calculated if there are at least two codons for each amino acid with synonymous codons in the sequence.

Simulation results

I shall here present data for a scenario with a typical bias level, that is, 40.5 effective codons. In Table 4 the codon probabilities for the simulation experiments are given. They represent two scenarios, one without bias discrepancy and one with bias discrepancy. In the scenario with no bias discrepancy all twofold degenerate aa have $Nc = 1.5$, isoleucine, $Nc = 2$, the fourfold degenerate aa have $Nc = 2.5$, and the sixfold degenerate have $Nc = 3.5$ (these numbers sum up to 40.5 effective codons). In the scenario with bias discrepancy, five of the twofold degenerate aa have $Nc = 1.67$, the others in that group have $Nc = 1.3$. Isoleucine has $Nc = 2$. Three of the fourfold degenerate aa have $Nc = 3.23$, the two others in that group have $Nc = 1.4$. Two of the sixfold degenerate aa have $Nc = 4.5$, the last aa in that group has 1.5 Fig. 5A shows how $\hat{N}c$ and $\hat{N}c^*$ behave as function of gene length when there is no bias discrepancy. It can be seen that $\hat{N}c$ is the best estimator under these circumstances. All estimators tend to overestimate the actual value at short gene lengths. It can also be seen that using homozygosity rounding improves $\hat{N}c^*$. Fig. 5B shows a similar graph for the scenario with bias discrepancy. It can be seen that $\hat{N}c^*$

Table 4
Codon frequencies in two scenarios used for simulation and testing of the estimators of the effective number of codons

Codon	Amino acid	p, without bias discrepancy	p, with bias discrepancy
TTC	Phe	0.788675	0.726285
TTT	Phe	0.211325	0.273715
AAA	Lys	0.788675	0.8669
AAG	Lys	0.211325	0.1331
AGC	Ser	0.481627	0.381832
AGT	Ser	0.103675	0.123634
TCA	Ser	0.103675	0.123634
TCC	Ser	0.103675	0.123634
TCG	Ser	0.103675	0.123634
TCT	Ser	0.103675	0.123634
TGG	Trp	1	1
TAC	Tyr	0.788675	0.726285
TAT	Tyr	0.211325	0.273715
CTA	Leu	0.481627	0.812164
CTC	Leu	0.103675	0.037567
CTG	Leu	0.103675	0.037567
CTT	Leu	0.103675	0.037567
TTA	Leu	0.103675	0.037567
TTG	Leu	0.103675	0.037567
CCA	Pro	0.58541	0.460852
CCC	Pro	0.138197	0.179716
CCG	Pro	0.138197	0.179716
CCT	Pro	0.138197	0.179716
AGA	Arg	0.481627	0.381832
AGG	Arg	0.103675	0.123634
CGA	Arg	0.103675	0.123634
CGC	Arg	0.103675	0.123634
CGG	Arg	0.103675	0.123634
CGT	Arg	0.103675	0.123634
CAA	Gln	0.788675	0.726285
CAG	Gln	0.211325	0.273715
GTA	Val	0.58541	0.460852
GTC	Val	0.138197	0.179716
GTG	Val	0.138197	0.179716
GTT	Val	0.138197	0.179716
GCA	Ala	0.58541	0.460852
GCC	Ala	0.138197	0.179716
GCG	Ala	0.138197	0.179716
GCT	Ala	0.138197	0.179716
GGA	Gly	0.58541	0.840097
GGC	Gly	0.138197	0.053301
GGG	Gly	0.138197	0.053301
GGT	Gly	0.138197	0.053301
GAC	Asp	0.788675	0.8669
GAT	Asp	0.211325	0.1331
GAA	Glu	0.788675	0.726285
GAG	Glu	0.211325	0.273715
ATA	Ile	0.666667	0.666667
ATC	Ile	0.166667	0.166667
ATT	Ile	0.166667	0.166667
ATG	Met	1	1
ACA	Thr	0.58541	0.840097
ACC	Thr	0.138197	0.053301
ACG	Thr	0.138197	0.053301
ACT	Thr	0.138197	0.053301
TGC	Cys	0.788675	0.8669
TGT	Cys	0.211325	0.1331
CAC	His	0.788675	0.8669
CAT	His	0.211325	0.1331
AAC	Asn	0.788675	0.726285
AAT	Asn	0.211325	0.273715

Both scenarios correspond to a total of 40.5 effective codons.

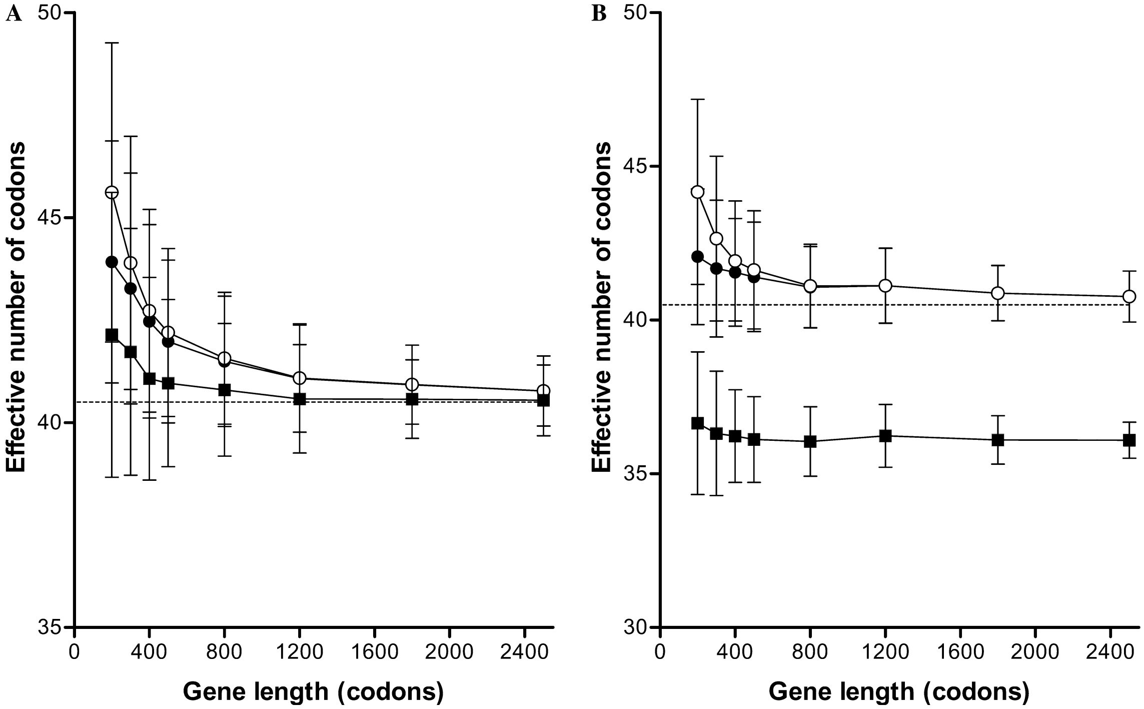

Fig. 5. (A) Behaviour of $\hat{N}c$ and $\hat{N}c^*$ on simulated genes, where there is no bias discrepancy and where the true number of effective codons is held constant at 40.5 (dashed line). Points represent averages of 500 individual resamplings and bars represent standard deviations. Symbols: ■, $\hat{N}c$; ○, $\hat{N}c^*$ calculated without rounding; ●, $\hat{N}c^*$ calculated with rounding. $\hat{N}c$ is a better estimator under these circumstances. (B) Similar to (A), but with bias discrepancy. $\hat{N}c^*$ (with rounding (●)) is a better estimator under these circumstances. Note that $\hat{N}c$ because of the bias discrepancy converges towards 36 instead of 40.5 (see text for details).

outperforms $\hat{N}c$ at these conditions, where $\hat{N}c^*$ asymptotically approaches a value of 36.0 effective codons. This can easily be verified by insertion of the codon probabilities from Table 4 in Eq. (1), followed by insertion of the homozygosities in Eq. (3), remembering that $F \rightarrow \sum p^2$ when $n \rightarrow \infty$ as is the case with the multinomial distribution used for simulation. So strictly speaking $\hat{N}c$ only converges towards the correct value when there is no bias discrepancy, otherwise this estimator has an intrinsic methodological error.

Conclusion remarks

This study has demonstrated that the classical way of calculating the effective number of codons in a gene is associated with some disadvantages, and that it could alternatively be calculated through addition of individual number of codons for individual amino acids, yielding the alternative quantity $\hat{N}c^*$. Which formula to use is very difficult to give a recommendation about, but should depend on the individual gene and purpose of the study. Generally speaking, Wright's $\hat{N}c$ applies to more sequences than $\hat{N}c^*$ because the latter does not accept homozygosity estimates by averaging in cases where individual amino acids are absent. On the other hand, the resampling results show that the estimate provided by $\hat{N}c^*$ is better than $\hat{N}c$ in cases where there are bias discrepancies. Data such as those presented in Fig. 2 suggest that there is 'sometimes' a bias discrepancy, and if 'sometimes' here means 'most often' then $\hat{N}c^*$ (with rounding) should probably be used, while Wright's method must be considered better if 'sometimes' means 'rarely.' It is therefore a future objective to establish if bias discrepancy is a common phenomenon in all life forms. A mathematical way to quantify observed bias discrepancy could pave the way forward; it might be possible to define a limit that tells when $\hat{N}c^*$ should preferentially be used instead of $\hat{N}c$.

References

[1] S.A. Gray, M.E. Konkel, Codon usage in the A/T-rich bacterium *Campylobacter jejuni*, Adv. Exp. Med. Biol. 473 (1999) 231–235.

[2] F. Wright, M.J. Bibb, Codon usage in the G + C-rich *Streptomyces* genome, Gene 113 (1992) 55–65.

[3] M. Gouy, C. Gautier, Codon usage in bacteria: correlation with gene expressivity, Nucleic Acids Res. 10 (1982) 7055–7074.

[4] H. Akashi, Gene expression and molecular evolution, Curr. Opin. Genet. Dev. 11 (2001) 660–666.

[5] T. Ikemura, Codon usage and tRNA content in unicellular and multicellular organisms, Mol. Biol. Evol. 2 (1985) 13–34.

[6] P.M. Sharp, W.-H. Li, The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications, Nucleic Acids Res. 15 (1987) 1281–1295.

[7] F. Wright, The ‘effective number of codons’ used in a gene, Gene 87 (1990) 23–29.

[8] J.M. Comeron, M. Aguade, An evaluation of measures of synonymous codon usage bias.382, J. Mol. Evol. 47 (1998) 268–274.

[9] A. Fuglsang, The effective number of codons for individual amino acids: some codons are more optimal than others, Gene 320 (2003) 185–190.

[10] W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, Numerical Recipes in C—The Art of Scientific Computing, second ed., Cambridge University Press, New York, USA, 1998.

[11] S. Kanaya, Y. Yamada, Y. Kudo, T. Ikemura, Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis, Gene 238 (1999) 143–155.

[12] B. Lafay, J.C. Atherton, P.M. Sharp, Absence of translationally selected synonymous codon usage bias in *Helicobacter pylori*, Microbiology 146 (2000) 851–860.