J Math Chem (2008) 44:569–575
DOI 10.1007/s10910-007-9330-y

Mathematical relations between codon bias indexes and their applications

Jun Wang · Yi Zhang

Received: 9 March 2007 / Accepted: 10 October 2007 / Published online: 11 December 2007

Abstract Codon Adaptation Index (CAI), Effective Number of Codons ($\hat{N}_c$) as well as its modifications $\hat{N}_c{}^*$, $\hat{N}_c{}^{**}$ can be used to measure gene codon bias. In this article, we prove $\hat{N}_c{}^{**}$ is more efficient and unbiased than $\hat{N}_c{}^*$ and $\hat{N}_c$ by revisiting correlations of them with CAI in the level of individual amino acid's codon bias. Correlations are studied by mathematical expressions rather than statistic methods, because the latter unavoidably depend on the data set used. Additionally, the immediate cause of correlations of $\hat{N}_i$ with CAI (as well as those of RSCU with CAI) are also described in mathematical language. Perhaps, mathematics provides us a new way to study correlations between biological indexes.

Keywords Codon Adaptation Index · Effective Number of Codons · Correlation · Expression

1 Introduction

Most amino acids are encoded by more than one codon, which are called synonymous codons. Because of translational selection, mutational bias or gene expressivity, synonymous codons are not used with equal frequencies, and their usage varies among

J. Wang · Y. Zhang (✉)
Department of Applied Mathematics, Dalian University of Technology, Dalian 116024, P.R. China
e-mail: zhaqi1972@163.com

J. Wang
College of Advanced Science and Technology, Dalian University of Technology,
Dalian 116024, P.R. China

Y. Zhang
School of Sciences, Hebei University of Science and Technology, Shijiazhuang,
HeBei 050018, P.R. China

genes [1,2]. However, genes within an organism or cell type usually have almost common codon preference [1,3], it is called "genome hypothesis". In 2004, Chen et al. [4] pointed out that, in all three domains of life, mutational pressure primarily sets common codon preference, and selective pressure makes some small modifications later.

As the practical consequences of these findings, heterologous expression could be weakened when the studied gene contained low-usage codons such as AGA or AGG (coding for Arginine). On the other aspect, the problem could be solved by exchanging rare codons for more frequently used synonymous codons or by providing additional copies of the corresponding rare tRNAs. In this sense, it's significant to study the relationship between protein levels and codon usage in quantitative terms, which includes studies on the relationship between gene codon bias and mRNA levels. When Fuglsang [5] studied the correlation of mRNA levels with codon bias measures, he found the superiority of $\hat{N}_c^{**}$ over $\hat{N}_c$ and $\hat{N}_c^{*}$, but the reason still needs studying. As a further study on relationship between protein levels and codon usage, in this article, we shall answer this question by analyzing relations between codon bias indexes.

In measuring gene codon bias, Sharp and Li [6] proposed the "Codon Adaptation Index" (CAI) scheme. And in 1990, based on the "effective allele numbers" of population genetics, the "Effective Number of Codons" ($\hat{N}_c$) was introduced by Wright [7], which tells what degree all 61 codons are used in a gene, in extremely biased genes the effective number of codons can approach 20, while in unbiased genes it will approach 61. Recently, $\hat{N}_c^{*}$ [8], $\hat{N}_c^{**}$ [9] are also brought forward by Fuglsang as modifications of $\hat{N}_c$.

In each method, overall codon bias of a gene is a combination of individual amino acid's codon bias, measurement of which is specific for the method. And generally speaking, overall codon bias of a gene is positively correlated with individual amino acid's codon bias. So, the latter should underlie the former. Furthermore, considering all amino acids together may lead to missing some information [10]. On the other hand, correlations obtained by statistic methods are unavoidably dependent on the data used, hence not valid for all species. Taking into account the two aspects, we will study the correlations of CAI with $\hat{N}_c^{**}$ ($\hat{N}_c^{*}$, $\hat{N}_c$) in the level of individual amino acid's codon bias by mathematical expressions.

Let p_{ij} denote the absolute frequency of the j-th synonymous codon of i-th amino acid appeared in a gene, $p_{i\max}$ the maximum p_{ij} value for the i-th amino acid, k_i the degree of degeneracy of i-th amino acid, n_i is the total count for the i-th amino acid in the gene, and l is the number of codons of the gene studied.

(1) For $\hat{N}_c$ scheme, $\hat{N}_i$ denotes the synonymous codon bias of i-th amino acid, that is,

$$\hat{N}_i = \frac{n_i - 1}{n_i \sum_{j=1}^{k_i} p_{ij}^2 - 1}. \tag{1}$$

For $\hat{N}_c^{*}$ scheme, the synonymous codon bias of i-th amino acid is the same as $\hat{N}_i$ if the re-adjustment is ignored.

(2) For $\hat{N}_c^{**}$ scheme, $\hat{N}_i^{**}$ denotes the synonymous codon bias of i-th amino acid, that is,

$$\hat{N}_i^{**} = \frac{1}{\sum_{j=1}^{k_i} p_{ij}^2}. \tag{2}$$

(3) We denote the codon bias of i-th amino acid implied by CAI scheme as $CAIAA_i$, that is

$$CAIAA_i = \sqrt[\ell]{\prod_{j=1}^{k_i} (w_{ij})^{c_{ij}}}, \tag{3}$$

where each $w_{ij} = \frac{p_{ij}}{p_{i\max}}$ iscalculated on a reference set consisting of some highly expressed genes, $c_{ij} = l \times p_{ij}$. Obviously, the CAI value of a gene is the product of $CAIAA_i$'s, corresponding to each kind of amino acids appeared in this gene. Suzuki et al. [11] indicated that, in multivariate analysis of codon usage data, the normalization scheme $\frac{p_{ij}}{p_{i\max}}$ can avoid bias derived from gene length, amino acid usage, and codon degeneracy, moreover, it can generate more PC's corresponding to variations in synonymous codon usage than others. So CAI scheme should be more efficient and unbiased in measuring codon bias. In [5] the superior efficiency of CAI was also shown by correlation of mRNA level with codon bias measure.

2 Correlations of CAI with $\hat{N}_c$, $\hat{N}_c^{*}$ and $\hat{N}_c^{**}$ in the level of individual amino acid's codon bias

CAI value of a gene is the "geometric mean" of the w_{ij}'s (obtained from reference set) corresponding to each of the codons appeared in the gene. Using "arithmetic mean" of the w_{ij}'s, we can obtain an alternative method of CAI, so-called "Codon Adaptation Index based on Arithmetic Mean" (denoted by CAIAM). The codon bias of i-th amino acid measured by CAIAM scheme is

$$CAIAMAA_i = \frac{\sum_{j=1}^{k_i} (c_{ij} \times w_{ij})}{l}. \tag{4}$$

And a gene's codon bias (also denoted by CAIAM) is the sum of $CAIAMAA_i$'s corresponding to each kind of amino acids appeared in this gene.

Does CAIAM scheme work? On the one hand, for i-th amino acid, $w_{ij} = 1$ if j-th codon is the optimal, and $w_{ij} < 1$ otherwise. The more times of optimal codons used, the more increased the values of CAIAA_i and $\mathrm{CAIAMAA}_i$, which means a rise in codon bias. On the other hand, CAI value is indicative of the level at which the gene is expressed, rather than dictating that level [6], suggesting the codon bias order among genes is the most important. The Spearman's rank correlation coefficient of CAI and CAIAM values of all *E. coil* genes is 0.9684 (shown in Fig. 1). It means that CAIAM is

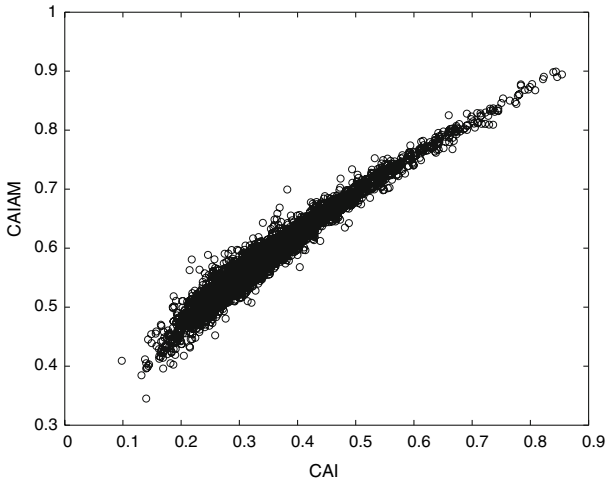

Fig. 1 A plot of CAI versus CAIAM for all *E. coli* genes. There is a strong positive correlation of the two parameters: $r_s = 0.9684$, $p < 10^{-12}$

almost equivalent to CAI in assessing codon bias order among genes. The *E. coli* K12 is taken from GenBank (accession number NC000913), and the w_{ij}'s are from [6].

To connect CAI (CAIAM) with $\hat{N}_c$, SCAIAM scheme is constructed as a simulant scheme of CAIAM. Codon bias of i-th amino acid measured by SCAIAM scheme is

$$SCAIAMAA_i = \frac{\sum_{j=1}^{k_i} \left(c_{ij} \times \frac{p_{ij}}{p_{i\max}} \right)}{l}. \tag{5}$$

And overall codon bias of a gene is the sum of $SCAIAMAA_i$'s corresponding to each kind of amino acids appeared in the gene.

For genes with similar synonymous-codon-usage pattern to reference set, $\frac{p_{ij}}{p_{i\max}}$ is close to w_{ij}, so their $CAIAMAA_i$ and $SCAIAMAA_i$ should be strongly correlated. Since genes within an organism have almost common codon preference [1,3], the positive correlation of $CAIAMAA_i$ with $SCAIAMAA_i$ should be general. Practically, as shown in Table 1, for *E. coli* genes the positive correlations of $CAIAMAA_i$ with $SCAIAMAA_i$ are all strong indeed. Therefore $SCAIAMAA_i$ and $CAIAMAA_i$ will be correlated with $\hat{N}_i$ in similar manner. Moreover, when the equivalence between CAI and CAIAM is considered, the correlation of $SCAIAMAA_i$ with $\hat{N}_i$ may represent that of $CAIAA_i$ with $\hat{N}_i$. Theorem 1 should be understood in this sense.

Theorem 1 *We have the following equations:*

(i) *For* $\hat{N}_c$ *and* $\hat{N}_c{}^*$, $\hat{N}_i = \dfrac{n_i - 1}{(p_{i\max} \times SCAIAMAA_i) \times n_i - 1}$;

(ii) *For* $\hat{N}_c{}^{**}$, $\hat{N}_i{}^{**} \times SCAIAMAA_i = \dfrac{1}{p_{i\max}}$;

Proof (i) From definition we have

$$SCAIAMAA_i = \frac{\sum_{j=1}^{k_i}(c_{ij} \times \frac{p_{ij}}{p_{i\max}})}{l} = \frac{\sum_{j=1}^{k_i}(p_{ij})^2}{p_{i\max}}, \tag{6}$$

which implies that $\hat{N}_i = \dfrac{n_i - 1}{p_{i\max} \times SCAIAMAA_i \times n_i - 1}$. Equation (ii) can be proved in a similar way. □

(1) Theorem 1 points out there is a negative correlation between CAI and Effective Number of Codons. It can explain why the mRNA levels correlate with CAI *positively*, while correlate with $\hat{N}_c$, $\hat{N}_c{}^*$ and $\hat{N}_c{}^{**}$ *negatively* in [5].

(2) Theorem 1 shows that $\hat{N}_i{}^{**}$ is more closely related with $SCAIAMAA_i$ than $\hat{N}_i$. Because CAI is the most efficient one [5], we may believe $\hat{N}_c{}^{**}$ is more efficient than $\hat{N}_c{}^*$ and $\hat{N}_c$ in studying relationship between protein levels and codon usage.

(3) Because the normalization method used in CAI schemeleads to least bias [11], $\hat{N}_c{}^{**}$ should be less biased than $\hat{N}_c$ and $\hat{N}_c{}^*$.

3 Discussion

In [10] Fuglsang indicated that correlations of $\hat{N}_i$ with CAI (or those of RSCU with CAI) are not uniform, and he considered such differences can be explained by selectional advantage choice of codons provided in genes that are selectively biased. However, the immediate cause should be the relative relationship between p_{ij} and $\sum_{j=1}^{k_i} p_{ij}$ (or n_i).

(1) From Theorem 1 we see negative correlation of $\hat{N}_i$ with $SCAIAMAA_i$ (or $CAIAA_i$ for most genes) is mainly affected by $p_{i\max}$ and n_i in this way: the larger $p_{i\max}$ or smaller n_i, the stronger correlation. For example, as to *E. coli* genes, as indicated in [1], corresponding to tuf, r-pro., rpo, thr and trp gene classes, the usage of CTG (optimal codon of Leu) are 53, 79, 141, 55, 96 respectively, and the usage of GCT (optimal codon of Ala) are 24, 93, 30, 18 and 31 respectively, which means $p_{i\max}$ of Leuis much larger than that of Ala. Moreover, the n_i (618) of Leu is smaller than that (675) of Ala. Naturally, the negative correlation of $\hat{N}_{Leu}$ with CAI_{Leu} should bestronger than that of $\hat{N}_{Ala}$ with CAI_{Ala}. It may directly lead to a conclusion of Fuglsang: correlation of $\hat{N}_{Leu}$ with CAI is much strongerthan that of $\hat{N}_{Ala}$ with CAI for *E. coli* genes [10]. In this sense, the difference of correlations of $\hat{N}_i$ with CAI (shown in Table 1 of

Table 1 The spearman's rank correlation coefficients of $CAIAMAA_i$ and $SCAIAMAA_i$ of 20 kinds of amino acids for *E. coli* genes, where $p < 10^{-10}$

Amino acid	Gly	Glu	Asp	Val	Ala	Arg	Ser	Lys	Asn	Met
Correlation coefficient	0.8471	0.9202	0.9298	0.7228	0.8129	0.7814	0.6346	0.9535	0.6831	1.0000
Amino acid	Iso	Thr	Try	Cys	Tyr	Leu	Phe	Gln	His	Pro
Correlation coefficient	0.7556	0.6936	1.0000	0.9464	0.8111	0.5813	0.8528	0.8541	0.8664	0.6553

Table 2 Correlations of RSCU with CAI as well as β_{ij}'s of 14 codons (identified as optimal codons by Ikemura in [1]) for *E. coli* genes, the correlations are from [10]

Optimal codon	CGT (Arg)	CGC (Arg)	CTG (Leu)	GCT (Ala)	GCA (Ala)	GCG (Ala)	GGT (Gly)
Correlation of RSCU with CAI	0.3951	0.2392	0.7604	0.04266	−0.1021	0.07965	0.2778
β_{ij}	0.5612	0.3621	0.6719	0.2904	0.2074	0.2919	0.4679
Optimal codon	GGC (Gly)	CCG(pro)	ACT (Thr)	ACC (Thr)	GTT (Val)	GTA (Val)	GTG (Val)
Correlation of RSCU with CAI	0.3701	0.5007	0.1182	0.4428	0.0551	−0.00611	0.0748
β_{ij}	0.4177	0.6964	0.2844	0.5269	0.3962	0.2226	0.2491

[10]) should be derived from the variation in relative relationship between their $p_{i\,\max}$ and n_i.

(2) Set $\beta_{ij} = \dfrac{p_{ij}}{\sum_{j=1}^{k_i} p_{ij}}$. For 14 codons identified as optimal ones by Ikemura in [1], correlations of RSCU with CAI as well as β_{ij}'s are listed in Table 2.

β_{ij} and the correlation of RSCU with CAI are strongly correlated ($r_s = 0.9121$). Noticeably, GCA and GTA, which are identified as non-optimal codons in [10] because of the negative correlations of their RSCU with CAI, have two smallest β_{ij}'s; while codons identified as optimal ones all have larger β_{ij}'s. In this sense, β_{ij} is a translation of the correlation of RSCU with CAI. Then, based on the definition of β_{ij}, the difference of correlations of RSCU with CAI should be derived from the variation in the relative relationship between p_{ij} and $\sum_{j=1}^{k_i} p_{ij}$. In a word, selectional advantage choice of codons seems to influence the correlations of $\hat{N}_i$ with CAI (or those of RSCU with CAI) through the relative relationship between p_{ij} and $\sum_{j=1}^{k_i} p_{ij}$ (or n_i). It may suggest that the relative relationship between p_{ij} and $\sum_{j=1}^{k_i} p_{ij}$ (or n_i) is the stand of studying gene expressivity, selectional pressure and codon bias.

Finally, above results may imply that, biological principles often have mathematical roots, and studying the roots is significant for us to understand and apply these biological principles in practice.

Acknowledgements The author thanks all the anonymous referees for their valuable suggestions and support. This work was partially supported by the National Natural Science Foundation of China and the Natural Science Foundation of Liaoning Province of China.

References

1. T. Ikemura, Codon usage and tRNA content in unicellular and multicellular organisms. Mol. Biol. Evol. **2**, 13–34 (1985)
2. P.M. Sharp, M. Stenico, J.F. Peden, A.T. Lloyd, Codon usage mutational bias, translational selection, or both? Biochem. Soc. Trans. **21**, 835–841 (1993)
3. R. Grantham, C. Gautier, M. Gouy, R. Mercier, A. Pave, Codon catalog usage and the genome hypothesis. Nucleic Acids. Res **8**, 49–62 (1980)
4. S.L. Chen, W. Lee, A.K. Hottes, L. Shapiro, H.H. MacAdams, Codon usage between genome is constrained by genome-wide mutational processes. Proc. Natl. Acad. Sci. USA. **101**, 3480–3485 (2004)

5. A. Fuglsang, Correlation of codon bias measures levels: analysis of transcriptome data from *Escherichia Coli*. Biochem. Biophys. Res. Commun. **327**, 4–7 (2005)
6. P.M. Sharp, W.H. Li, The codon adaptation index—a measure of directional synonymous codon usage bias, and its applications. Nucleic Acids Res. **15**, 1281–1296 (1987)
7. F. Wright, The 'effective number of codons' used in a gene. Gene **87**, 23–29 (1990)
8. A. Fuglsang, The effective number of codons revisited. Biochem. Biophys. Res. Commun. **317**, 957–964 (2004)
9. A. Fuglsang, On the methodological weakness of the effective number of codons: a reply to Marashi and Najafabadi. Biochem. Biophys. Res. Commun. **327**, 1–3 (2005)
10. A. Fuglsang, The effective number of codons for individual amino acids: some codons are more optimal than others. Gene **320**, 185–190 (2003)
11. H. Suzuki, R. Saito, M. Tomita, A problem in multivariate analysis of codon usage data and a possible solution. FEBS Lett, **579**(28), 6499–6504 (2005)