Evolutionary Bioinformatics

Libertas Academica
FREEDOM TO RESEARCH

ORIGINAL RESEARCH

Relative Codon Adaptation Index, a Sensitive Measure of Codon Usage Bias

Soohyun Lee[1], Seyeon Weon[1,2], Sooncheol Lee and Changwon Kang

Department of Biological Sciences, Korea Advanced Institute of Science and Technology, 335 Gwahangno, Yuseong-gu, Daejeon 305-701, Korea. [1]These two authors contributed equally. [2]Present address: Bioinformatics Research Laboratories, Daejeon, Korea. Email: ckang@kaist.ac.kr

Abstract: We propose a simple, sensitive measure of synonymous codon usage bias, the Relative Codon Adaptation Index (rCAI), as a way to discriminate better between highly biased and unbiased regions, compared with the widely used Codon Adaptation Index (CAI). CAI is a geometric mean of the relative usage of codons in a gene, and is calculated using the codon usage table trained with a set of highly expressed genes. In contrast, rCAI is computed by subtracting the background codon usage trained with two noncoding frames of highly expressed genes from the codon usage in the coding frame. rCAI has higher signal-to-noise ratio than CAI, considering that non-coding frames would not show codon bias. Translation efficiency and protein abundance correlates comparably or better with rCAI than CAI or other measures such as 'effective number of codons' and 'SCUMBLE offsets'. Within overlapping coding regions, one of the two coding frames dominates in codon usage bias according to rCAI. Presumably, rCAI could substitute CAI in diverse applications.

Keywords: codon usage bias, codon adaptation index, translation efficiency, overlapping genes

Evolutionary Bioinformatics 2010:6 47–55

This article is available from http://www.la-press.com.

Introduction

It has long been known that synonymous codons are used with unequal frequencies in many organisms.[1] Such bias can be explained by several possible causes: heterogeneity of nucleotide base composition,[2] asymmetric mutation rates in the leading and lagging strands of DNA replication,[3,4] transcription effectiveness,[5,6] protein hydropathy,[7] and selection pressure for optimizing translational efficiency.[8] Although the major source of bias differs from species to species, the phenomenon of codon usage bias is universal across diverse taxa.[9–13] Codon usage bias is considered important in the study of molecular evolution[14] and expression of exogenous protein.[15,16]

Since Ikemura proposed a measure of codon usage bias, 'frequency of optimal codons' (f_{op}), associated with tRNA abundance,[17] a vast number of different codon bias measures[18–25] have been developed, indicating the biological relevance of codon bias. Among them, 'codon adaptation index' (CAI)[18] is one of the most widely used. CAI was originally proposed to provide a normalized estimate that can be used across genes and species, ranging from 0 to 1. The boundary values refer to the cases in which only the most frequent codons (CAI = 1) or only the least frequent codons (CAI = 0) are used within a gene.

CAI has been used as a simple and effective measure of the overall synonymous codon usage bias of a gene. Highly expressed genes, including ribosomal proteins and transcription and translation factors, tend to have high CAI.[26,27] CAI has been used extensively in diverse biological research: for measuring translation efficiency and predicting cellular protein levels,[28,29] for verifying high-throughput expression quantification techniques[30,31] and for optimizing DNA vaccines.[32]

CAI of a gene is computed as the geometric mean of the relative adaptiveness (w) of all the codons in the gene.[18] The relative adaptiveness w_j for codon of kind j that codes for amino acid i is defined as below:

$$w_j = X_{ij}/X_{imax},$$

where X_{ij} is the number of occurrences of codon j in the reference set of highly expressed genes and X_{imax} is the maximum X_{ij} for amino acid i. Practically in computer calculation, CAI is computed as follows:

$$\mathrm{CAI} = \exp\frac{1}{L}\sum_{k=1}^{L}\ln\, w_{c(k)},$$

where L is the number of codons in the gene and $w_{c(k)}$ is the w value for the k-th codon in the gene.

Intragenic variation of synonymous codon usage bias has been documented.[33] Some software programs provide an option of displaying local CAI values along a gene.[34,35] However, we observe that local CAI values are relatively noisy and that CAI still has a room for improvement in close examination of short regions. Motivated by this limitation, we propose a modified version of CAI that can apparently capture local signals more sensitively and produce less noise when applied to regions expecting little or no codon bias.

Algorithms and Datasets

Relative CAI

The 'relative codon adaptation index' (rCAI) is calculated from the LNWD table, which is defined as below:

$$LNWD_{\mathrm{i}} = ln\ w_{1,i} - (ln\ w_{2,i} + ln\ w_{3,i})/2,$$

for codon of kind i, where $w_{1,i}$ is identical to w_i defined in the original CAI paper.[18] $w_{2,i}$ and $w_{3,i}$ are equivalent to $w_{1,i}$, except that they are trained from the phases +1 and +2 of the identical reference set composed of highly expressed genes.

rCAI is computed as below:

$$\mathrm{rCAI} = \exp\frac{1}{L}\sum_{k=1}^{L} LNWD_{c(k)}$$

where L is the number of codons in the gene and $LNWD_{c(k)}$ is the $LNWD$ value for the k-th codon in the gene. Based on our notation, CAI can be expressed as:

$$\mathrm{CAI} = \exp\frac{1}{L}\sum_{k=1}^{L} ln\ w_{1,c(k)}$$

Compiling protein coding sequences

In this study, we restricted our scope within prokaryotes. Genomic sequences (.fna) and gene locations

(.gff) obtained from the NCBI ftp site were used to extract protein coding sequences for *Escherichia coli* K12 M1655 (NC_000913) and *Lactococcus lactis* ssp. *lactis* IL1403 (NC_002662). A set of highly expressed genes predicted iteratively by Ramazzotti et al[34] were used to train the LNWD table for each of the two species mentioned above. We first checked that the positions listed in their results matched with the coding positions we compiled from the .gff files and used the corresponding extracted sequences to compute lnw and LNWD. The same reference set was used in calculation of rCAI and CAI for comparison.

Translation efficiency and protein abundance

Dressaire et al quantified absolute protein abundance based on the APEX method[30] along with mRNA abundance in *L. lactis*, under four different growth conditions.[31] Translation efficiencies could be approximated by dividing protein abundance by mRNA abundance, although they would be compounded by differential protein degradation. Using these data sets, we were able to obtain translation efficiencies for 171 *L. lactis* genes.

For *E. coli* genes, the APEX-based protein profiling data were provided by Lu et al[30] but the mRNA abundance levels were not. Thus, translation efficiency data could not be obtained for *E. coli* genes.

Correlation analysis

For testing correlation between rCAI and CAI, all genes annotated in the NCBI files were used (4379 for *E. coli* and 2321 for *L. lactis*). For testing correlation of rCAI or CAI with translation efficiency or protein abundance in *L. lactis*, the 171 *L. lactis* genes included in the Dressaire et al dataset[31] and the NCBI annotation were used. Correlation of rCAI or CAI with protein abundance in *E. coli* was analyzed similarly using the Lu et al dataset of 397 genes.[30] In estimation of correlation, $\log_{10}$ values of the translation efficiency and protein abundance were used. All statistical analyses were performed using Origin 7.5.

Comparison of rCAI and CAI on a genomic region

For comparison of individual genomic locations, we particularly selected 9 highly expressed *L. lactis* genes. We made sure that these genes were not part of the reference set, to avoid self-training artifact. Local rCAI and CAI were computed along the genomic regions containing these genes, using a sliding window of 25 codons at step size of 1 codon.

According to the original paper,[18] the number (X_{ij}) of any codon absent in the reference set was set to 0.5. We followed this in our computations of rCAI and CAI. Furthermore, to calculate the length of a gene (L), codons ATG and TGG were excluded for CAI, according to the original paper,[18] but not for rCAI. In the sliding-window approach, we did not exclude these codons from the length of window in calculations of either rCAI or CAI.

Gene-wise signal-to-noise ratio (SNR)

We measured SNR for rCAI and CAI for each gene in *L. lactis* and *E. coli*, assuming that the two noncoding frames represent noise. The median of local rCAI or CAI values within all 25-codon sliding windows in a coding frame was taken as signal intensity. The pooled median from the two noncoding frames was used as noise intensity. SNR was obtained by dividing signal intensity by noise intensity.

Effective number of codons (ENC)

rCAI was compared with ENC,[19] another widely used measure of codon usage bias, in terms of correlations with translation efficiency and protein abundance. Proteins of less than 100 amino acids were excluded from analysis, because their ENC was known to be inaccurate.[19] ENC was computable only for 170 *L. lactis* and 395 *E. coli* genes. A modified version of ENC, Nc**,[20] was also implemented and compared with rCAI. Nc** was computable only for 84 *L. lactis* and 320 *E. coli* genes, because it requires the presence of every amino acid in a protein.

SCUMBLE

The degree of codon bias affected by translation efficiency or gene expression level was estimated using the SCUMBLE (synonymous codon usage bias maximum likelihood estimation) method developed by Kloster and Tang.[25] A four-trend model was used, as suggested by the developers, for correlation analysis

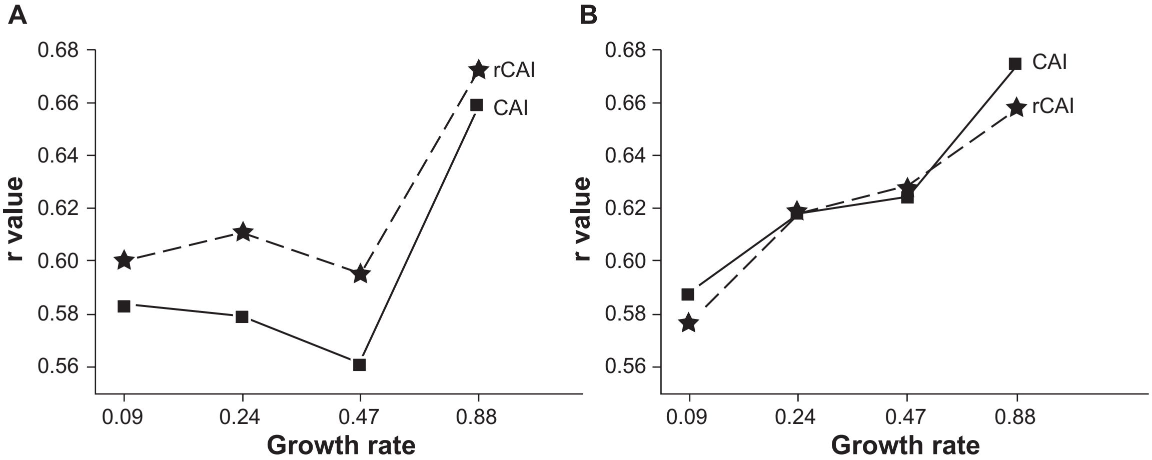

Figure 1. Comparison between rCAI and CAI with respect to correlation with translation efficiency. Pearson correlation coefficients (r) of gene-wise rCAI (asterisk) or CAI (square) were calculated with $\log_{10}$ of translation efficiency **A**) or with $\log_{10}$ of protein abundance **B**) obtained for 171 *L. lactis* genes under four separate growth conditions.[31]

using 165 *L. lactis* and 367 *E. coli* genes. By default, the SCUMBLE program removes genes shorter than 100 codons and further excludes those that cannot be explained by the predicted model.

Application of rCAI on overlapping genes

Overlapping gene pairs of histidine kinase and response regulator have been compiled for over 200 bacteria and studied extensively by Cock and Whitworth.[36] They observed that the property of an overlapping region was more similar to the upstream gene than its downstream counterpart, and proposed a scenario where an overlapping downstream gene has evolved by extending its 5' end into its upstream neighbor. From their collection, we chose a *Caulobacter crescentus* gene pair whose overlap was long enough to apply 25-codon windows. The sequence and annotation data (NC_002696) and reference set were obtained as in *E. coli* and *L. lactis*.

Results and Discussion

Correlation of rCAI with translation efficiency and protein abundance

In order to assess the qualification of rCAI as a measure of codon usage bias, we looked at the correlation between rCAI and translation efficiency, a direct effect of tRNA abundance.[37] For 171 *L. lactis* proteins, rCAI had better correlations with translation efficiency than CAI under all four growth conditions tested (Fig. 1A). For example, the correlation coefficients were estimated to be r = 0.675 for rCAI and r = 0.659 for CAI, using the highest growth rate

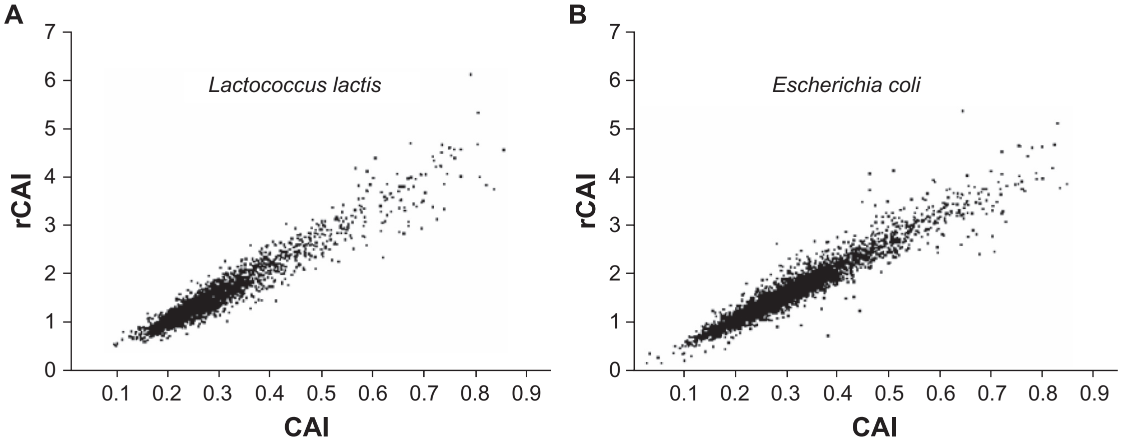

Figure 2. Genome-wide correlation between rCAI and CAI. Gene-wise rCAI and CAI values were calculated for all 2321 *L. lactis* **A**) or 4379 *E. coli* **B**) annotated genes. Pearson correlation coefficient (r) between rCAI and CAI values was 0.96 in both *L. lactis* and *E. coli*.

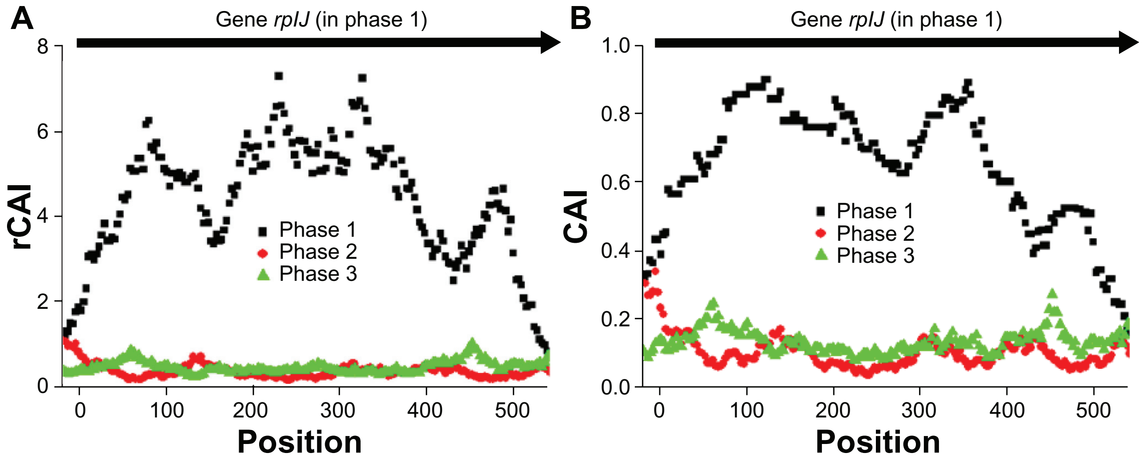

Figure 3. Signal-to-noise ratio of rCAI and CAI on coding regions. Distribution of local rCAI **A**) and CAI **B**) values on an *L. lactis* gene, *rplJ*. Phase 1 matches the coding frame. For phases 2 and 3, rCAI and CAI were calculated on triplets beginning at the +1 nucleotide and +2 nucleotide of the coding frame, respectively. The *rplJ* gene is shown in full length and starts at position 1 of the x axis.

dataset. The CAI values we obtained were slightly different from those used by Dressaire et al[31] (data not shown), presumably because different reference sets were used for training.

It has been known that gene expressivity, or protein abundance is also correlated with codon usage bias.[27] We observed that the correlation of rCAI with protein abundance was at a comparable level with that of CAI in *L. lactis* (Fig. 1B). *L. lactis* has been reported to exhibit relatively high codon usage bias among bacteria (strength of selected codon usage bias S = 2.288),[8] which supports the choice of this organism for assessment of rCAI.

For *E. coli*, we could not find a set of protein and mRNA abundance data measured under the same experimental conditions. Therefore, we could not obtain translation efficiency data for *E. coli* genes to estimate correlation with rCAI or CAI. Using just the protein profiling data from Lu et al,[30] however, correlation with protein abundance turned out to be similar between rCAI (r = 0.557) and CAI (r = 0.572). The strength of selected codon usage bias of *E. coli* K12 was S = 1.488.[8]

Whereas CAI is set to range from 0 to 1, rCAI ranges from 0 to an upper limit, which corresponds to an imaginary gene consisting only of the maximum-LNWD codons (Fig. 2). Nevertheless, rCAI and CAI were highly correlated with each other (r = 0.96) in both *L. lactis* (Fig. 2A) and *E. coli* (Fig. 2B).

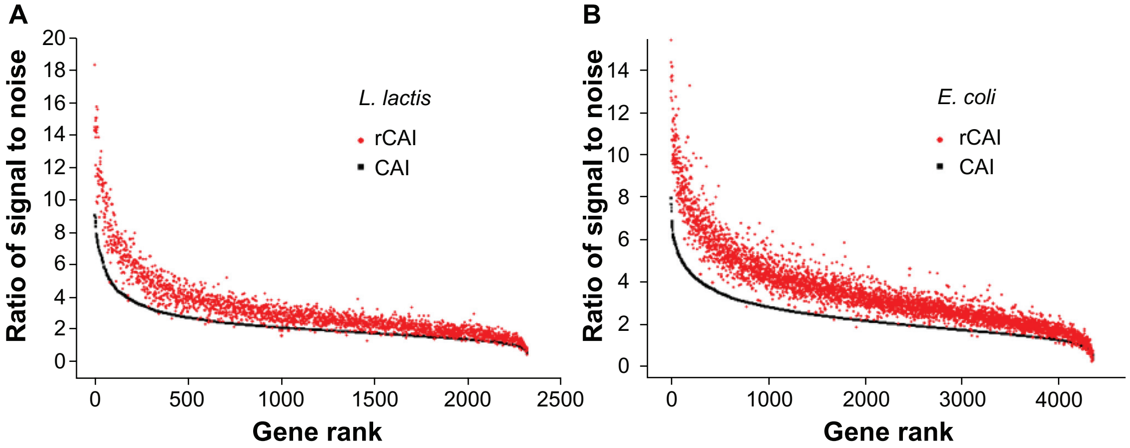

Figure 4. Genome-wide comparison between rCAI and CAI with respect to signal-to-noise ratio (SNR). SNR values of gene-wise rCAI (red circle) and CAI (black rectangle) were calculated for all 2321 *L. lactis* **A**) and 4379 *E. coli* **B**) annotated genes. Genes are ranked by CAI SNR.

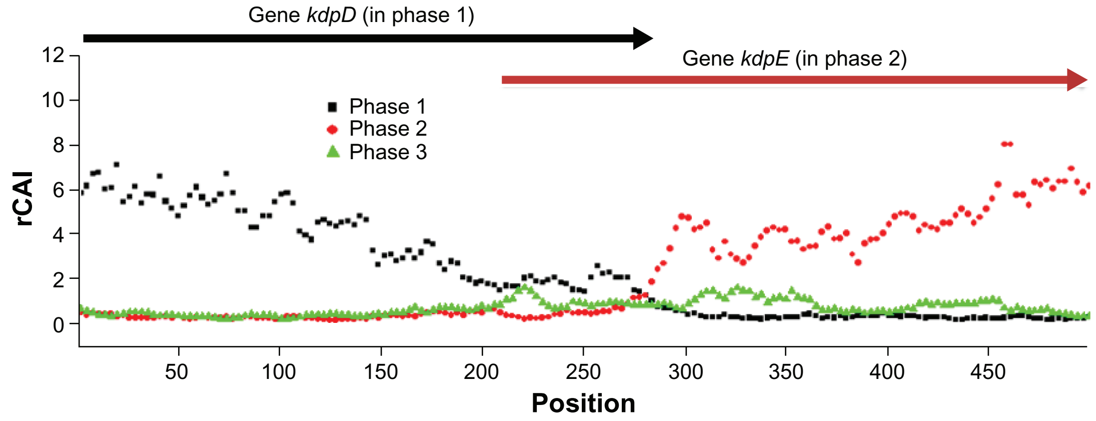

Figure 5. rCAI distribution across the overlapping region of two *C. crescentus* genes, *kdpD* and *kdpE*. Only parts of the two genes are shown with the overlapping region zoomed in. Phase 1 matches the coding frame of the upstream gene *kdpD*, and phases 2 and 3 start at the +1 and +2 nucleotides of phase 1, respectively. The coding frame of *kdpE* lies in phase 2. rCAI computed on phase 1 dominates over that of phase 2 in the region of overlap.

Distribution of rCAI and CAI on a genomic region

Figures 3A and 3B show local rCAI and CAI distributions, respectively drawn on the three sense phases of *L. lactis rplJ* gene, where phase 1 corresponds to the coding frame. *rplJ* is a highly expressed ribosomal protein gene but was not part of the reference set prepared by Ramazzotti et al.[34] The other 8 highly expressed genes we selected showed similar patterns (data not shown). As seen in Figure 3, rCAI showed more consistently lower values than CAI when applied to noncoding frames. This indicates that rCAI may pick up less false positives, when little or no codon usage bias is expected. In order to assess whether this trend is systematic, we measured SNR for each gene, assuming that the two noncoding frames represent noise. Figure 4 shows SNR values of rCAI and CAI of all *L. lactis* (Fig. 4A) and *E. coli* (Fig. 4B) genes, in the order of CAI SNR ranking. rCAI exhibits systematically higher SNR than CAI in these two species.

Application of rCAI on overlapping genes

It would be interesting to see the codon usage bias pattern on the regions where two or more genes overlap in different frames. Cock and Whitworth[36] studied mutability and codon frequencies within the two overlapping genes for sensor kinase and response regulator constituting the two-component systems in over 200 bacteria and documented that an overlapping region tends to be more similar to the upstream gene than downstream gene. We selected one of sufficiently long overlapping regions from their compilation to examine rCAI pattern. The region was overlapped by

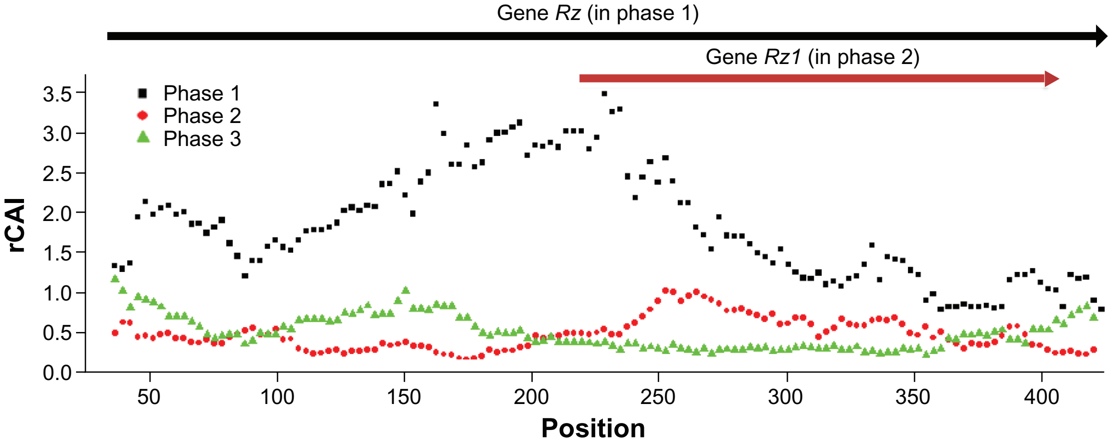

Figure 6. rCAI distribution across the overlapping region of two λ phage genes, *Rz* and *Rz1*. Phase 1 matches the coding frame of *Rz* shown in full length. *Rz1* is embedded in *Rz* in a different phase (phase 2). Within the *Rz1* gene, where two coding frames exist, codon usage bias was observed to be stronger in phase 1 than phase 2.

two *C. crescentus* genes, *kdpD* for histidine kinase and *kdpE* for response regulator (Fig. 5). We checked that the strength of selected codon usage bias for *C. crescentus* was moderately high (S = 1.152).[8]As shown in the rCAI plot of Figure 5, the upstream *kdpD* gene was clearly dominant over the downstream *kdpE* gene.

We applied rCAI on another overlapping gene pair, *Rz* and *Rz1*, in an *E. coli* bacteriophage λ (Fig. 6). It has been proposed that not all but many phages lacking intrinsic tRNAs, including λ, adapt to the codon usage bias of their host.[38,39] Therefore, we expected some degree of rCAI and CAI signals over λ phage genes, even though the lnw and LNWD tables were trained from the *E. coli* reference set. As seen in Figure 6, the upstream of the two frames seemed to dominate within the overlapping region, which in this case covered the entire *Rz1* gene. Accordingly, rCAI patterns of the overlapping coding regions evidently support for the upstream gene dominance proposed by Cock and Whitworth.[36]

Resistance to frameshift

It has been suggested that one of the selective constraints that affect formation of codon usage bias is resistance to frameshift caused by either mutation or mistranslation.[40] In other words, codon usage is adapted in a way that minimizes optimal codon usage in frame-shifted products. This is consistent with our observation that subtracting the background of shifted frames makes codon bias signals become clearer having higher SNR.

Simplicity

Although rCAI uses a background term, which is not used for CAI, rCAI does not require an additional training set, such as genes of low expression levels or randomly shuffled sequences. Given the reference set of genes used to calculate CAI, rCAI can be computed simply by using two additional phases. The procedure of computing LNWD involves computing lnw, as in the calculation of CAI, except that it is done on three different phases of the reference set.

Comparison of rCAI with other codon usage bias estimates

'Effective number of codons' (ENC) was proposed by Wright to describe codon bias without the reference of highly expressed genes.[19] ENC ranges from 20 (maximum bias) to 61 (no bias). In this study, ENC was computed for the *L. lactis* and *E. coli* genes, and the correlations of ENC were analyzed with translation efficiency and protein abundance. For example, ENC correlation with *L. lactis* translation efficiency was r = –0.338 for the highest growth rate dataset, being much worse than the rCAI (r = 0.675) or CAI (r = 0.659) correlation. Furthermore, a previous study also showed that CAI performs better than ENC in predicting mRNA expression level in *E. coli*.[24] In our analysis, ENC correlation with *E. coli* protein abundance was r = –0.417, which was also worse than the rCAI (r = 0.557) or CAI (r = 0.572) correlation.

As modified versions of ENC, Nc* and Nc** were sequentially introduced by Fuglsang.[20,22] The more recent Nc** was computed in this study. For the highest growth rate dataset of *L. lactis*, Nc** correlation with translation efficiency was r = –0.576, which was worse than the rCAI (r = 0.675) or CAI (r = 0.659) correlation. Likewise, Nc** correlation with *E. coli* protein abundance was –0.461, being further worse than the rCAI (r = 0.557) or CAI (r = 0.572) correlation. Accordingly, rCAI and CAI are better correlated with translation efficiency, protein abundance, or gene expressivity than ENC or NC**.

'Frequency of optimal codons' (F_{op}) suggested by Ikemura is the fraction of optimal codons (the most preferred by tRNA) to synonymous codons in a gene.[17] F_{op} may serve as the most direct measure of translation-related codon usage bias. Unfortunately, the use of F_{op} is limited due to lack of information on tRNA abundance in some organisms. According to the analysis by Goetz and Fuglsang, F_{op} is less correlated with expressivity than CAI,[24] and hence presumably than rCAI.

Discernment of codon bias sources

'Correspondence analysis' (CA) powerfully discerns the major source of codon usage bias in a particular genome, and several versions of CA such as CA-AF, CA-RF, CA-RSCU and within-group CA (WCA) have been proposed.[23] CA extracts the most influential 'axes,' or directions from a multi-dimensional vector space, i.e. a contingency matrix of genes (in rows) and codons (in columns). By inspecting what CA axis separates genes, one can identify the major source of codon bias in the species. For example, highly expressed *E. coli* genes are separated by the CA axis correlated with gene expressivity, which is the major source of codon bias in *E. coli*.[41] In contrast, replication selection is the major source for codon bias in *Borrelia burgdorferi*, and the genes are separated into those on the leading strand and those on the lagging strand.[41]

Recently, a probabilistic model-based method, called SCUMBLE, was proposed to estimate the degrees of contribution by different sources ('trends') and their effects on a gene ('offsets' or β_i).[25] Where translational selection is the major source of bias, for example in *Saccharomyces cerevisiae*, the first offset (β_1, or the offset for the highest-variance trend) is highly correlated with gene expressivity. In contrast, the expressivity-correlated trend is not the major source of bias in *Helicobacter pylori*, as it is captured by β_3 (the offset for the third highest-variance trend) rather than β_1.

When SCUMBLE was applied on the *L. lactis* and *E. coli* datasets in this study, β_1 showed the highest absolute correlation coefficient with both protein abundance and translation efficiency, which are consequently presumed to be the major source of codon bias in these two organisms. On the highest growth rate dataset of *L. lactis*, the correlation of rCAI with translation efficiency or protein abundance ($r = 0.675$ and 0.658, respectively) was much better than that of β_1 ($r = -0.493$ and -0.583, respectively) computed using SCUMBLE in a four-trend model. On the *E. coli* dataset, β_1 showed a marginally better correlation with protein abundance ($r = -0.592$), compared with rCAI ($r = 0.557$) or CAI ($r = 0.572$).

Although CA and SCUMBLE are useful in discerning the sources of codon usage bias, it is not the most ideal for quantifying an individual kind(s) of codon bias. Firstly, which CA axis or trend represents a specific source is generally unknown and differs from species to species. Secondly, an axis or trend may possibly represent a mixed or partial effect. For example, in both *L. lactis* and *E. coli*, protein abundance was correlated with not only β_1 ($r = -0.583$ and -0.592, respectively) but also β_4 ($r = -0.469$ and -0.475, respectively), raising need for careful interpretation. We suggest that rCAI or CAI can complement such weakness of the CA- or SCUMBLE-based analysis.

Conclusions

CAI measures absolute codon usage bias by quantifying the similarity in the synonymous codon frequency between a given gene and a set of the most frequently translated genes. In contrast, the relative CAI, or rCAI, uses not only the most optimal frames but also potentially the least optimal frames, which are likely to occur in the noncoding frames of the maximally translated genes. Computationally, rCAI is nearly as easy to calculate as CAI, but shows higher discriminating power. As we demonstrated a usage of rCAI in codon bias analysis of overlapping genes, rCAI may provide a substitute for CAI in various applications.

Acknowledgements

We sincerely thank Drs. Tang and Kloster for kindly providing the code for SCUMBLE, and the anonymous reviewers of this manuscript for offering valuable suggestions. This work was supported by the Korea Healthcare Technology R&D Project [A080587].

Disclosures

This manuscript has been read and approved by all authors. This paper is unique and is not under consideration by any other publication and has not been published elsewhere. The authors and peer reviewers of this paper report no conflicts of interest. The authors confirm that they have permission to reproduce any copyrighted material.

References

1. Ikemura T. Codon usage and tRNA content in unicellular and multicellular organisms. *Mol Biol Evol.* 1985;2:13–34.
2. Sueoka N. On the genetic basis of variation and heterogeneity of DNA base composition. *Proc Natl Acad Sci U S A.* 1962;48:582–92.
3. Lobry JR. Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol Biol Evol.* 1996;13:660–5.
4. McLean MJ, Wolfe KH, Devine KM. Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes. *J Mol Evol.* 1998;47:691–6.
5. Xia X. Maximizing transcription efficiency causes codon usage bias. *Genetics.* 1996;144:1309–20.
6. Lao PJ, Forsdyke DR. Thermophilic bacteria strictly obey Szybalski's transcription direction rule and politely purine-load RNAs with both adenine and guanine. *Genome Res.* 2000;10:228–36.
7. D'Onofrio G, Jabbari K, Musto H, Bernardi G. The correlation of protein hydropathy with the base composition of coding sequences. *Gene.* 1999;238:3–14.
8. Sharp PM, Bailes E, Grocock RJ, Peden JF, Sockett RE. Variation in the strength of selected codon usage bias among bacteria. *Nucleic Acids Res.* 2005;33:1141–53.
9. Ermolaeva MD. Synonymous codon usage in bacteria. *Curr Issues Mol Biol.* 2001;3:91–7.

10. Lloyd AT, Sharp PM. Evolution of codon usage patterns: the extent and nature of divergence between Candida albicans and Saccharomyces cerevisiae. *Nucleic Acids Res.* 1992;20:5289–95.
11. Kotlar D, Lavner Y. The action of selection on codon bias in the human genome is related to frequency, complexity, and chronology of amino acids. *BMC Genomics.* 2006;7:67.
12. Powell JR, Moriyama EN. Evolution of codon usage bias in Drosophila. *Proc Natl Acad Sci U S A.* 1997;94:7784–90.
13. Duret L, Mouchiroud D. Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. *Proc Natl Acad Sci U S A.* 1999;96:4482–7.
14. Novembre JA. Accounting for background nucleotide composition when measuring codon usage bias. *Mol Biol Evol.* 2002;19:1390–4.
15. Kleber-Janke T, Becker WM. Use of modified BL21(DE3) Escherichia coli cells for high-level expression of recombinant peanut allergens affected by poor codon usage. *Protein Expr Purif.* 2000;19:419–24.
16. Kane JF. Effects of rare codon clusters on high-level expression of heterologous proteins in Escherichia coli. *Curr Opin Biotechnol.* 1995;6: 494–500.
17. Ikemura T. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system. *J Mol Biol.* 1981;151:389–409.
18. Sharp PM, Li WH. The codon Adaptation Index-a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* 1987;15:1281–95.
19. Wright F. The 'effective number of codons' used in a gene. *Gene.* 1990; 87:23–9.
20. Fuglsang A. On the methodological weakness of 'the effective number of codons': a reply to Marashi and Najafabadi. *Biochem Biophys Res Commun.* 2005;327:1–3.
21. Wall DP, Herbeck JT. Evolutionary patterns of codon usage in the chloroplast gene rbcL. *J Mol Evol.* 2003;56:673–88; discussion 89–90.
22. Fuglsang A. The 'effective number of codons' revisited. *Biochem Biophys Res Commun.* 2004;317:957–64.
23. Suzuki H, Brown CJ, Forney LJ, Top EM. Comparison of correspondence analysis methods for synonymous codon usage in bacteria. *DNA Res.* 2008;15:357–65.
24. Goetz RM, Fuglsang A. Correlation of codon bias measures with mRNA levels: analysis of transcriptome data from Escherichia coli. *Biochem Biophys Res Commun.* 2005;327:4–7.
25. Kloster M, Tang C. SCUMBLE: a method for systematic and accurate detection of codon usage bias by maximum likelihood estimation. *Nucleic Acids Res.* 2008;36:3819–27.
26. Post LE, Nomura M. DNA sequences from the str operon of Escherichia coli. *J Biol Chem.* 1980;255:4660–6.
27. Gouy M, Gautier C. Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Res.* 1982;10:7055–74.
28. Futcher B, Latter GI, Monardo P, McLaughlin CS, Garrels JI. A sampling of the yeast proteome. *Mol Cell Biol.* 1999;19:7357–68.
29. Gygi SP, Rochon Y, Franza BR, Aebersold R. Correlation between protein and mRNA abundance in yeast. *Mol Cell Biol.* 1999;19:1720–30.
30. Lu P, Vogel C, Wang R, Yao X, Marcotte EM. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat Biotechnol.* 2007;25:117–24.
31. Dressaire C, Gitton C, Loubiere P, Monnet V, Queinnec I, Cocaign-Bousquet M. Transcriptome and proteome exploration to model translation efficiency and protein stability in Lactococcus lactis. *PLoS Comput Biol.* 2009;5:e1000606.
32. Ruiz LM, Armengol G, Habeych E, Orduz S. A theoretical analysis of codon adaptation index of the Boophilus microplus bm86 gene directed to the optimization of a DNA vaccine. *J Theor Biol.* 2006;239:445–9.
33. Qin H, Wu WB, Comeron JM, Kreitman M, Li WH. Intragenic spatial patterns of codon usage bias in prokaryotic and eukaryotic genomes. *Genetics.* 2004;168:2245–60.
34. Ramazzotti M, Brilli M, Fani R, Manao G, Degl'innocenti D. The CAI Analyser Package: inferring gene expressivity from raw genomic data. *In Silico Biol.* 2007;7:507–26.
35. Puigbo P, Bravo IG, Garcia-Vallve S. CAIcal: a combined set of tools to assess codon usage adaptation. *Biol Direct.* 2008;3:38.
36. Cock PJ, Whitworth DE. Evolution of gene overlaps: relative reading frame bias in prokaryotic two-component system genes. *J Mol Evol.* 2007;64:457–62.
37. Sorensen MA, Kurland CG, Pedersen S. Codon usage determines translation rate in Escherichia coli. *J Mol Biol.* 1989;207:365–77.
38. Kunisawa T, Kanaya S, Kutter E. Comparison of synonymous codon distribution patterns of bacteriophage and host genomes. *DNA Res.* 1998;5:319–26.
39. Lucks JB, Nelson DR, Kudla GR, Plotkin JB. Genome landscapes and bacteriophage codon usage. *PLoS Comput Biol.* 2008;4:e1000001.
40. Huang Y, Koonin EV, Lipman DJ, Przytycka TM. Selection for minimization of translational frameshifting errors as a factor in the evolution of codon usage. *Nucleic Acids Res.* 2009;37:6799–810.
41. Perriere G, Thioulouse J. Use and misuse of correspondence analysis in codon usage studies. *Nucleic Acids Res.* 2002;30:4548–55.